Introduction aux méthodes statistiques en ingénierie

C. BASSIM ET BRYAN LEE

eCampus Ontario – Libre accès

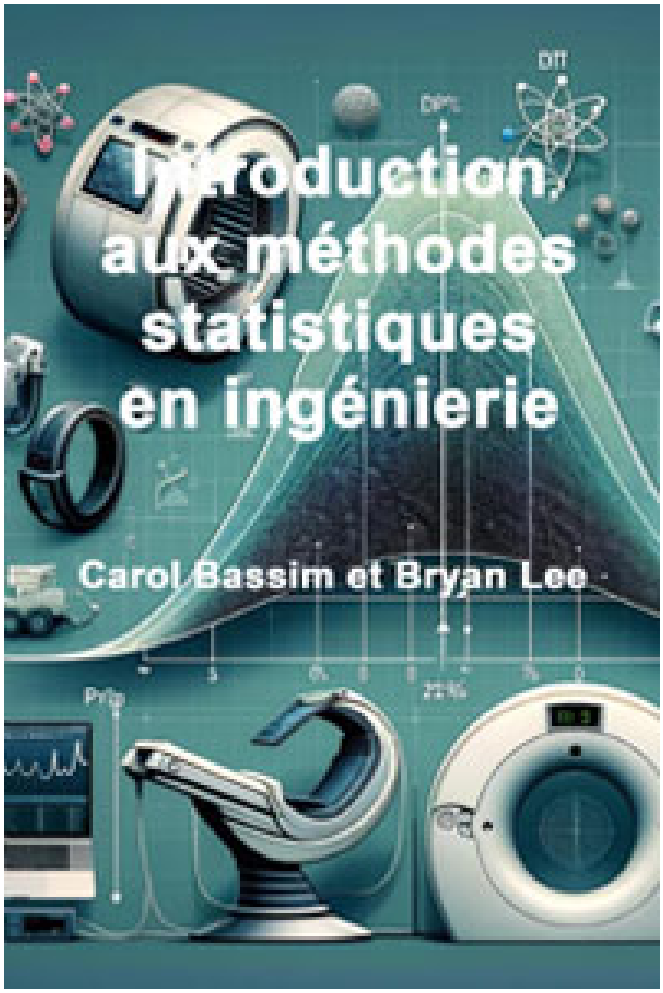
Table des matières

Introduction aux méthodes statistiques en ingénierie

Financé par le gouvernement de l'Ontario.

Les opinions exprimées dans cette publication sont celles de l'auteur ou des auteurs et ne reflètent pas nécessairement celles du gouvernement de l'Ontario ou du Consortium ontarien pour l'apprentissage en ligne.

Page de titre



Financé par le Gouvernement de l'Ontario
Les avis exprimés dans cette publication sont les avis du ou des auteur(s) et ne reflètent pas forcément celui du Gouvernement de l'Ontario ou du Consortium ontarien pour l'apprentissage en ligne

À propos de cet ouvrage

Bienvenue dans le monde stimulant et transformateur des statistiques d'ingénierie, où la théorie mathématique et l'innovation convergent afin de façonner l'avenir de l'ingénierie, de la technologie, de l'environnement et des soins de santé. Ce manuel en accès libre est conçu à l'intention des étudiants de premier cycle en tant qu'introduction. Il vous permet d'acquérir les connaissances fondamentales et les compétences pratiques nécessaires pour prospérer dans le domaine dynamique de l'ingénierie et des spécialisations de la discipline.

Pourquoi des statistiques en ingénierie?

L'ingénierie est au premier plan de l'innovation technologique et du vécu de l'humanité.

Exploration de divers domaines

Ce manuel vous initiera à divers domaines de l'ingénierie et vous fera découvrir l'utilité des méthodes statistiques dans ces différents domaines. Ce manuel et les ressources connexes renferment des exemples pratiques, des études de cas et des exercices représentant des cas réels pour vous transmettre des compétences pratiques. De la théorie aux applications concrètes, ce manuel va parcourir les outils et méthodes descriptifs et analytiques des statistiques en insistant sur leur application pratique pour la résolution de problème d'ingénierie. Vous ne découvrirez pas seulement la théorie statistique : vous verrez aussi comment appliquer ces concepts pour concevoir des expériences, analyser des données et contrôler des processus dans des contextes d'ingénierie.

Ressources informatiques et en libre accès pour une découverte

Nous vous encourageons à tirer pleinement parti du fait que ce manuel est publié en accès libre, ce qui vous permet d'explorer l'application des statistiques aux systèmes d'ingénierie. Quelle que soit le domaine d'ingénierie qui vous passionne, ce manuel vous servira de guide inestimable dans votre parcours. Les tutoriels de traitement statistique sur GitHub offrent des exemples interactifs et concrets de programmation en statistiques et permettent d'apprendre en explorant et en créant des simulations. Le référentiel GitHub se trouve ici : GitHub : Introduction aux méthodes statistiques en ingénierie.

Attributions et nouveautés

La première version du manuel s'inspire beaucoup du texte de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte et liés tout au long du document.

Le professeur émérite de l'Université de l'Iowa Stephen Vardeman et le professeur émérite de l'Université de Miami J. Marcus Jobe (ISU PhD, 1984) ont placé leur livre *Basic Engineering Data Collection and Analysis*, publié à l'origine par Duxbury/Thompson Learning/Cengage, en téléchargement libre sous licence CC BY-NC-SA internationale 4.0, par l'intermédiaire de l'Iowa State University Digital Press. Ce livre est disponible ici et est affecté du DOI suivant : https://doi.org/10.31274/isudp.2023.127

Le manuel Basic Engineering Data Collection and Analysis est essentiellement une révision/deuxième édition de l'ouvrage *Statistics for Engineering Problem Solving* de Vardeman, qui a remporté, en 1994, le Meriam/Wiley Distinguished Author Award de l'American Society for Engineering Education.

Cette ressource s'appuie également sur les ressources statistiques fondamentales de « Process Improvement Using Data ». Il s'agit d'un leg inestimable créé et mis à disposition en tant que ressource d'éducation libre par Kevin Dunn pendant son mandat à l'Université McMaster entre 2012 et 2016. L'ouvrage de Kevin n'a pas été d'une valeur inestimable que pour ce texte, mais aussi pour de nombreux éducateurs et éducatrices du monde entier, car il rend les statistiques d'ingénierie et la science des données accessibles, compréhensibles et applicables. Il est disponible ici. Cette ressource est sous licence CC BY-SA 4.0.

Elle s'appuie également sur le précieux ouvrage intitulé « Introductory Statistics » d'OpenStax par Barbara Illowsky et Susan Dean : Introductory Statistcis. CC BY-NC-SA 4.0.

Nous avons ici synthétisé ces ressources (et de nombreuses autres) dans une optique de prise en charge du traitement statistique et d'ingénierie en les adaptant particulièrement aux spécialisations de l'ingénierie.

Les Jupyter Notebooks et le langage de programmation Python sont pris en charge dans ce cadre en tant qu'expérience pratique et d'apprentissage actif, en associant ce texte aux principes FAIR des ressources en accès libre, ce qui signifie qu'elles sont Faciles à trouver, Accessibles, Interopérables et Réutilisables.

Un parcours important

Tout au long de votre parcours d'apprentissage, rappelez-vous que le but de l'ingénierie n'est pas que de résoudre des problèmes, mais bien d'améliorer des vies. Votre travail peut avoir un impact significatif sur les gens et changer le monde. Ensemble, nous allons entreprendre un parcours de transformation, où statistiques et innovation vont de pair.

Découvrons le carrefour passionnant entre l'ingénierie, les statistiques et la technologie, et créons l'avenir ensemble!

Objectifs d'apprentissage

Objectifs d'apprentissage

Les étudiant.e.s:

1. maîtriseront les principes de base des statistiques d'ingénierie
2. seront en mesure de mettre en œuvre des techniques d'analyse de données adaptées aux scénarios d'ingénierie
3. développeront des compétences pratiques en Python à l'aide de tutoriels et de simulations
4. appliqueront leurs connaissances des statistiques à des problèmes d'ingénierie réels.

.

Importance pour l'ingénierie

Ces objectifs d'apprentissage sont essentiels aux ingénieur.e.s. En effet, ils fournissent une base solide en matière d'analyse statistique et d'analyse des données, en plus de fournir des compétences en programmation Python. S'ils atteignent ces objectifs, les étudiant.e.s seront en mesure de résoudre des problèmes d'ingénierie complexes qui nécessitent de prendre des décisions basées sur les données et des analyses statistiques.

Parties, modules et chapitres

Les parties spécifiques suivantes et leurs objectifs d'apprentissage, tels qu'ils sont enseignés dans les parties, modules et chapitres, sont conformes aux objectifs globaux mentionnés ci-dessus.

Partie 1 : Explorer les données

- Reconnaître et distinguer les différents termes clés.
- Appliquer différents types de méthodes d'échantillonnage pour la collecte des données.
- Comprendre le rôle des statistiques en ingénierie.
- Appliquer des compétences informatiques en matière de statistiques à des fins d'exploration des données.
- Nettoyer les données pour les préparer à l'analyse statistique.

Partie 2 : Résumer, visualiser et communiquer via les données

- Apprendre à effectuer un tracé des données et à les communiquer efficacement
- Afficher les données sous forme de graphiques et interpréter les graphiques.
- Reconnaître, décrire et calculer des mesures de position et de répartition des données.

Partie 3 : Probabilités et variables aléatoires discrètes

- Comprendre et utiliser la terminologie des probabilités.
- Calculer les probabilités en utilisant les règles d'addition et de multiplication.
- Construire et interpréter des tableaux de contingence, des diagrammes de Venn et des schéma en arbre.
- Reconnaître et comprendre les fonctions de distribution de probabilité discrètes.
- Calculer et interpréter l'espérance mathématique.
- Appliquer correctement les distributions de probabilité discrètes.

Partie 4 : Variables aléatoires continues et distribution normale de la probabilité

- Reconnaître et comprendre les fonctions de densité de probabilité continue.
- Appliquer correctement les distributions de probabilité continues.
- Reconnaître et appliquer la distribution de probabilité normale.

Partie 5 : Statistiques inférentielles et test d'hypothèses à l'aide d'échantillons

- Appliquer et interpréter le théorème limite central pour les moyennes.
- Décrire les tests d'hypothèses et faire la distinction entre les types d'erreurs de test d'hypothèses.
- Réaliser et interpréter des tests d'hypothèses pour les paramètres de population.
- Réaliser et interpréter des tests d'hypothèses pour deux paramètres de population.
- Comprendre et appliquer les méthodes non paramétriques pour comparer des distributions.
- Calculer et interpréter les intervalles de confiance pour les paramètres de population.
- Déterminer les tailles d'échantillons requises pour les intervalles de confiance.
- Comprendre et pouvoir expliquer la valeur p et les conclusions du test statistique.
- Choisir en tout confiance le test statistique à exécuter.

Partie 6 : Inférence pour les études mutli-échantillons non structurées et ANOVA.

- Interpréter la distribution de la probabilité F.
- Exécuter et interpréter l'analyse de la variance à un facteur et des essais de variances.
- Appliquer des méthodes d'intervalle de confiance individuels et simultanés pour l'analyse de la variance à un facteur.

Partie 7 : Méthode des moindres carrés et analyse de régression linéaire simple

- Discuter des concepts de régression linéaire et de corrélation.
- Calculer et analyser des diagrammes de dispersion, calculer des coefficients de corrélation, et identifier les valeurs aberrantes.
- Tirer des conclusions sur des modèles simples de régression linéaire et communiquer ces conclusions avec assurance.
- Trouver la droite de régression de modèles établis et créer de nouveaux modèles à partir des données.

Partie 8 : Analyse de régression linéaire multiple

- Appliquer l'analyse de régression linéaire multiple.
- Apprendre l'ajustement et l'élaboration de modèles pour la régression linéaire multiple.
- Présentation du plan complet d'expériences factorielles.

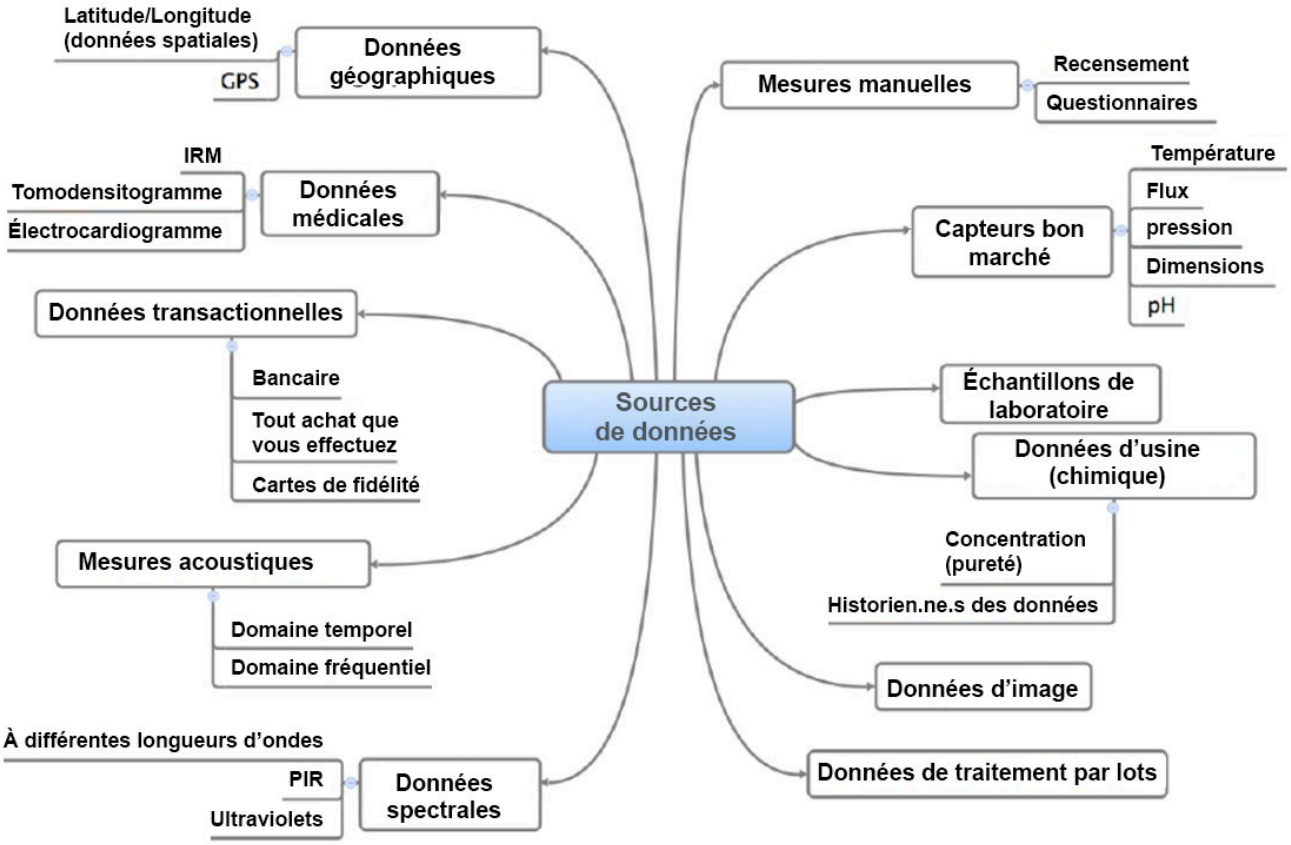
Partie 9 : Plan d'expériences

- Appliquer et mettre en œuvre un plan d'expérience.
- Appliquer des plans factoriels complets et fractionnés.
- Comprendre et utiliser les méthodologies de surfaces de réponse et les méthodes d'optimisation.

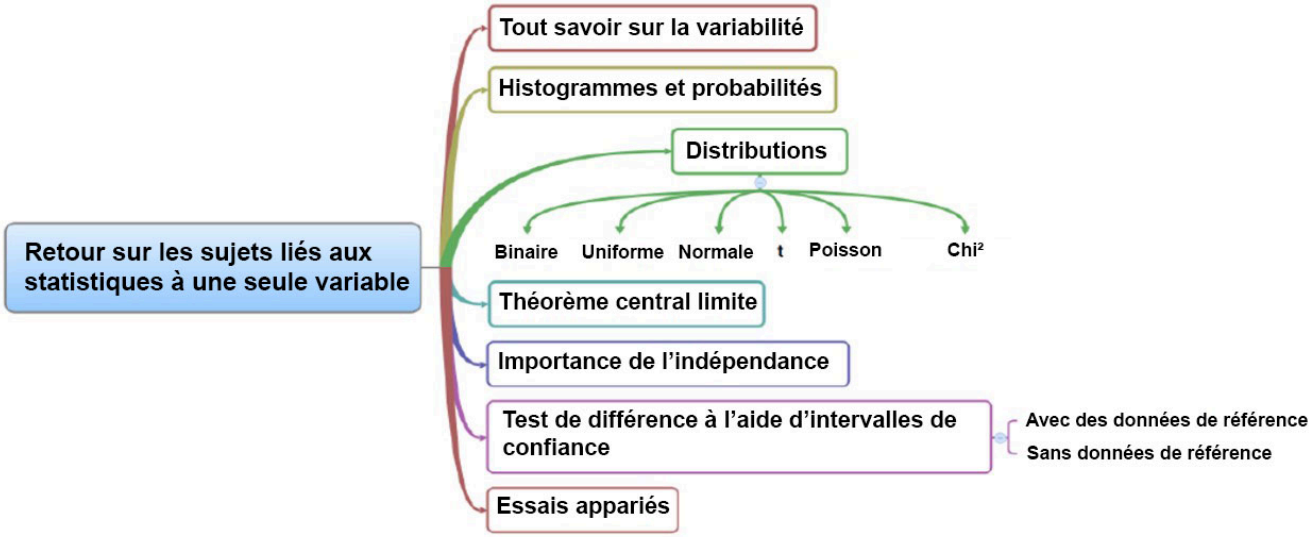
.

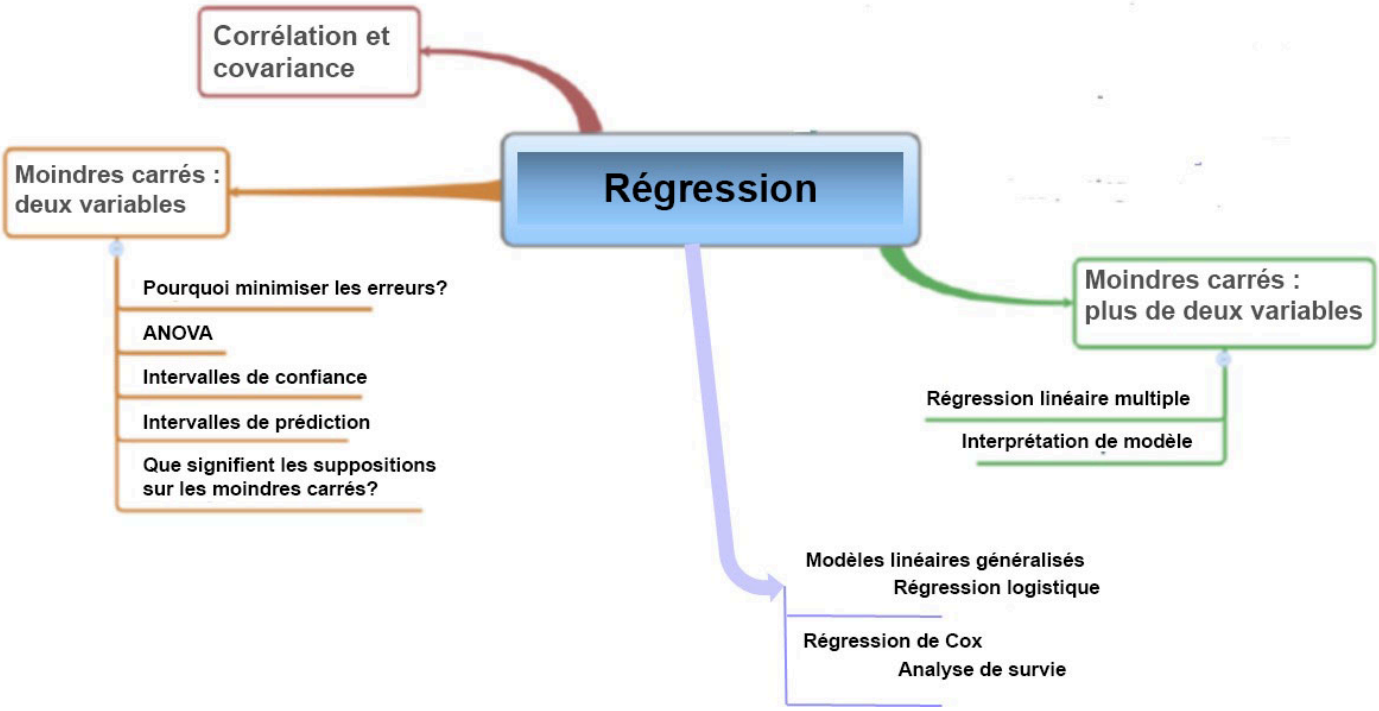
Dans l'ensemble, ces modules et objectifs d'apprentissage donnent aux étudiant.e.s en génie les connaissances et compétences en statistiques requises pour exceller dans leur domaine et leur permettre de prendre des décisions basées sur les données pour résoudre les problèmes d'ingénierie avec efficacité.

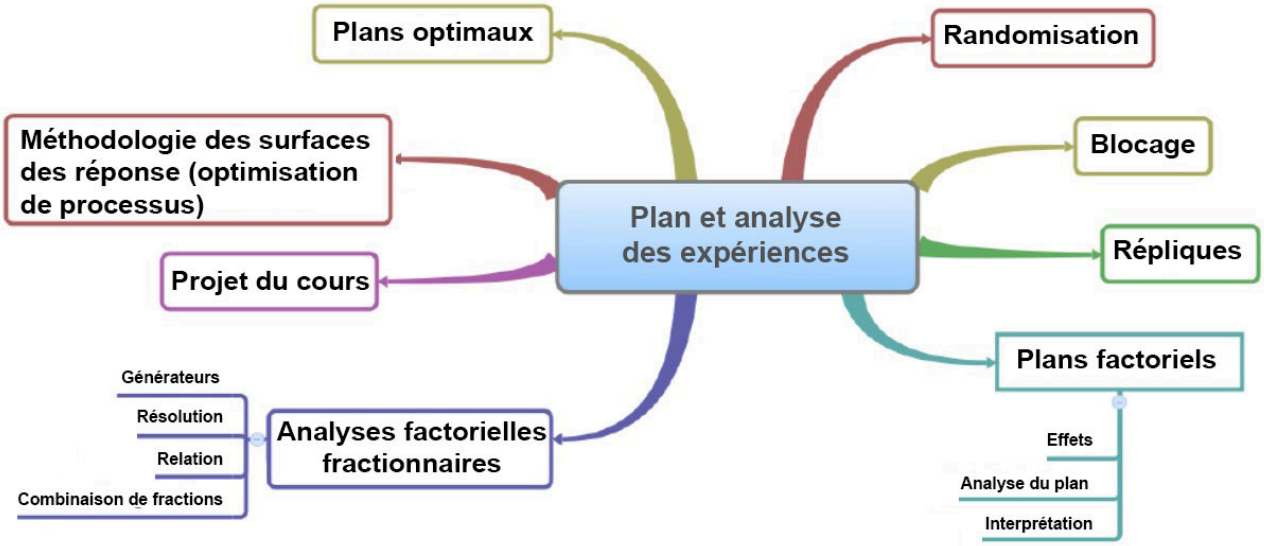
Aperçu graphique des modules d'apprentissage


Sources de données
Données géographiques
Latitude/Longitude (données spatiales)
GPS
Données médicales
IRM
Tomodensitogramme
Électrocardiogramme
Données transactionnelles
Bancaire
Tout achat que vous effectuez
Cartes de fidélité
Mesures acoustiques
Domaine temporel
Domaine fréquentiel
Données spectrales
À différentes longueurs d'ondes
PIR
Ultraviolets
Mesures manuelles
Recensement
Questionnaires
Capteurs bon marché
Température
Flux
pression
Dimensions
pH
Échantillons de laboratoire
Données d'usine (chimique)
Concentration (pureté)
Historien.ne.s des données
Données d'image
Données de traitement par lots


Visualisation des données
Utilisation de couleurs
Tableaux
Diagrammes de dispersion
Graphiques de séries temporelles
Graphiques à barres
Diagramme à boîtes


Retour sur les sujets liés aux statistiques à une seule variable
Tout savoir sur la variabilité
Histogrammes et probabilités
Distributions
Binaire
Uniforme
Normale
t
Poisson
Chi²
Théorème central limite
Importance de l'indépendance
Test de différence à l'aide d'intervalles de confiance
Avec des données de référence
Sans données de référence
Essais appariés





Source : Cet aperçu graphique des modules d'apprentissage est tiré de « Process Improvement Using Data », de Kevin Dunn. Cette ressource est disponible ici; tous les droits d'auteur lui appartiennent, et le contenu est partagé sous licence CC BY-SA 4.0.

Installation et révision de Python

Pour tirer le meilleur parti de cette ressource, il est fortement recommandé d'utiliser une routine de calcul statistique capable de lire le code Python. Nous vous recommandons d'utiliser Jupyter Lab ou Jupyter Notebook avec la distribution Anaconda. Voir les instructions ci-dessous pour l'installation sur les différents systèmes d'exploitation.

Étapes de l'installation:

1. Accéder à la page Web Anaconda et télécharger le fichier de configuration approprié.
    - Lien vers le site de téléchargement d'Anaconda : https://www.anaconda.com/download#downloads
2. Suivre les instructions appropriées.
    - Windows : https://docs.anaconda.com/free/anaconda/install/windows/
    - Mac : https://docs.anaconda.com/free/anaconda/install/mac-os/
    - Linux : https://docs.anaconda.com/free/anaconda/install/linux/
3. En cas de problème avec l'installation de Python, vous pouvez utiliser Google Colab : https://colab.google/
    - Il suffit d'avoir un compte Gmail pour se connecter et utiliser Jupyter Notebook sur un navigateur Internet.

Une résumé de Python se trouve sur le site GitHub du cours, Getting Started with Python.

Lien vers le référentiel GitHub

Référentiel GitHub

Le référentiel GitHub se trouve ici : GitHub : Introduction aux méthodes statistiques en ingénierie.

Les tutoriels de traitement statistique sur GitHub offrent des exemples interactifs et concrets de programmation en statistiques et permettent d'apprendre en explorant et en créant des simulations. Le référentiel GitHub se trouve ici : GitHub : Introduction aux méthodes statistiques en ingénierie.

Ce référentiel contient les fichiers Python de Jupyter Notebooks à utiliser lors de ce cours. **Il est recommandé de télécharger les fichiers sur votre ordinateur et de les exécuter localement.** Cependant, il est également possible de travailler sur Jupyter Notebooks en mode interactif à partir des modules de cours sans utiliser quoi que ce soit d'autre; trouver le badge BinderHub dans la section ReadMe du référentiel et cliquer dessus.

Ces liens interactifs sont également incorporés dans le texte de cette ressource pour vous permettre de travailler sur les exemples donnés lors de la révision des concepts de chaque module. Des référentiels spéciaux sur GitHub ont été créés à cette fin.

1.0.1 Introduction à l'exploration des données

Photographie datée de 1912 de Karl Pearson (auteur inconnu – Google Books – Nock, Albert Jay (1912-03). « A New Science And Its Findings ». The American Magazine LXXIII: 579. The Phillips Publishing Co.., Domaine public, https://commons.wikimedia.org/w/index.php?curid=4578734, et Google Books: Karl Pearson, The Grammar of Science, Adam et Charles Black, 1911, Londres : https://www.google.com/books/edition/The_Grammar_of_Science/9mISAAAAIAAJ?hl=en&gbpv=1&dq=grammar+of+science&printsec=frontcover, Domaine public.

Karl Pearson, pionnier controversé des mathématiques et de la biostatistique né en Angleterre en 1857, a profondément influencé le domaine des statistiques. Son livre « The Grammar of Science », publié pour la première fois en 1892, est un pilier de la philosophie scientifique; il peut être considéré comme un lien entre les statistiques et l'ingénierie dans la mesure où il met l'accent sur l'importance des méthodes statistiques pour comprendre et décrire les phénomènes naturels. Cette perspective trouve un écho particulier dans le domaine de l'ingénierie, lequel repose en grande partie sur l'observation, la mesure, la description, la communication technique et l'application créative – des aspects clés de la méthode scientifique qui s'appuient fortement sur le raisonnement statistique.

Les statistiques et les méthodes statistiques sont essentielles dans les domaines de l'ingénierie et de l'ingénierie biomédicale, car elles jouent un rôle crucial dans la conception, l'analyse et l'interprétation des données. Ces domaines reposant de plus en plus sur la technologie et les données, la littératie statistique et la capacité à utiliser « la grammaire de la science » deviennent essentielles pour les ingénieur.e.s en biomédecine.

Principaux points à retenir

Ce cours porte sur l'exploitation des données et sur la description et la communication de leur incertitude à l'aide de méthodes statistiques.

Ces méthodes sont essentielles dans le domaine de la santé et nécessaires pour créer, tester et comprendre l'impact des nouvelles technologies biomédicales, qui produisent d'énormes quantités de données. Contrairement à ce qui est d'usage en mathématiques pures, dans le monde réel, les données contiennent toujours des erreurs et des variations. Les statistiques facilitent la prise de décision éclairées dans ce contexte d'incertitude inhérente, une compétence essentielle dans divers domaines tels que l'économie, la santé, le commerce et l'ingénierie.

Les statistiques comprennent deux grands domaines : les méthodes descriptives, qui résument les données d'un échantillon, et les méthodes inférentielles, qui tirent des conclusions sur une population plus grande. L'exploration, le nettoyage et la catégorisation des données sont essentiels pour choisir la bonne méthode d'analyse statistique. Il est fondamental de pouvoir comprendre la tendance globale et la variation des données et de pouvoir en parler; c'est là qu'interviennent des mesures comme la moyenne, le mode, l'écart-type et l'écart interquartile.

Cette partie du cours se concentre sur les concepts fondamentaux de la statistique. Elle présente l'utilisation de l'informatique statistique et certains concepts fondamentaux de la science des données qui permettent d'appliquer des méthodes statistiques aux données. La science des données est le domaine interdisciplinaire des statistiques, de l'informatique scientifique, de la science et de l'ingénierie. Son objectif est d'extraire des connaissances à partir de données et d'en faire usage. Dans ce cours, nous utiliserons les JupyterLab Notebooks basés sur Python comme outil de calcul statistique pour explorer les concepts statistiques et les mettre en application.

Objectifs d'apprentissage

Objectifs d'apprentissage de la partie 1

- Distinguer les statistiques descriptives et inférentielles et comprendre leurs applications dans des contextes d'ingénierie.
- Comprendre les échantillons statistiques et les techniques d'échantillonnage de base.
- Connaître et comprendre la planification d'expériences et les plans d'expériences en ingénierie.
- Identifier, classer et utiliser différents types de données statistiques (catégoriques, classées, discrètes et continues).
- Revoir les fondements du nettoyage et de la préparation des données pour les explorer.

Objectifs d'apprentissage de la partie 1 – Tutoriels Jupyter Notebooks

- Ouvrir et utiliser un tutoriel JupyterLab Notebook et lire un jeu de données simple.
- Utiliser le calcul statistique pour nettoyer et préparer les données.

La partie 1 de ce cours établit les bases pour tout ce qui suit – elle contient une feuille de route pour l'étude des statistiques en ingénierie. Elle définit le sujet, décrit son importance, introduit des termes de base et aborde la

question importante des mesures. Enfin, elle se penche sur le rôle des modèles mathématiques dans la réalisation des objectifs des statistiques en ingénierie.

1.0.2 Sources de la partie 1

Cette première version de la partie 1 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs. ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte et liés tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn et sont partagées sous licence CC BY-SA 4.0. Le chapitre sur la variabilité provient directement de cette ressource et est la propriété intellectuelle de Kevin Dunn.

1.1.1 Méthodes statistiques en ingénierie

En règle générale, le rôle des ingénieur.e.s est de concevoir, de construire, d'utiliser ou d'améliorer des systèmes et des produits physiques. Ce travail repose sur des théories mathématiques et physiques acquises dans un programme de premier cycle en génie. Au fur et à mesure que l'ingénieur.e acquiert de l'expérience, il peut se fier à son jugement en plus des principes quantitatifs et scientifiques. Mais avec l'évolution de la technologie et l'arrivée de nouveaux systèmes et produits, l'ingénieur.e se trouve inévitablement confronté à des questions pour lesquelles la théorie et son expérience ne lui sont pas d'une grande aide. Que faire dans ce cas?

Il est possible de faire appel à des consultants de manière ponctuelle, mais la plupart du temps, il faut se débrouiller tout seul pour comprendre le fonctionnement du système. Pour ce faire, il est nécessaire de **collecter et d'interpréter des données**. Sans formation sur la collecte et l'analyse des données, les tentatives pourraient être désorganisées ou mal conçues, ce qui entraîne une perte de temps et de ressources, d'autant plus que les conclusions peuvent être erronées (ou inutilement floues). Pour éviter cela, il faut disposer d'une trousse d'outils comprenant les meilleurs principes et méthodes possibles de collecte et d'interprétation des données. Ces outils, ce sont les **méthodes statistiques pour l'ingénierie.**

L'objectif des statistiques en ingénierie est de fournir les concepts et les méthodes nécessaires lorsqu'on se trouve face à un problème exigeant un jugement indépendant ou une innovation. Elles fournissent les principes d'acquisition et de traitement des données empiriques nécessaires pour comprendre et manipuler les systèmes d'ingénierie.

DÉFINITION 1.1.1.1. Statistiques d'ingénierie

Les statistiques d'ingénierie représentent l'étude de la meilleure façon de

1. collecter des données
2. résumer ou de décrire les données d'ingénierie
3. tirer des inférences formelles et des conclusions pratiques fondées sur des données d'ingénierie, tout en reconnaissant la réalité de la variation.

Pour mieux comprendre cette définition, il est utile de voir comment les statistiques interviennent dans un problème réel.

Exemple 1.1.1.1. Traitement thermique des engrenages.

L'article « Statistical Analysis: Mack Truck Gear Heat Treating Experiments » de P. Brezler (Heat Treating, novembre, 1986) décrit une application simple des statistiques d'ingénierie. Un ingénieur des procédés a dû répondre à la question suivante : « Comment les engrenages doivent-ils être chargés dans un four de cémentation en continu afin de minimiser les déformations pendant le traitement thermique? » Diverses personnes avaient des opinions partiellement informées sur la façon de procéder, notamment sur la question de savoir si les engrenages devaient être empilés ou suspendus à des tiges traversant les alésages, Mais personne ne connaissait vraiment les conséquences de l'empilement ou de la suspension.

Collecte des données

Pour répondre à cette question, l'ingénieur a décidé d'obtenir des faits en recueillant des données sur le faux-rond de la face de poussée (une mesure de la distorsion de l'engrenage) des engrenages empilés et des engrenage suspendus. Le choix des modalités précises de cette collecte de données a nécessité une réflexion approfondie. Il pouvait y avoir des différences entre les lots de matières premières des engrenages, les machinistes et les machines qui produisaient les engrenages, les conditions à différents moments et positions dans le four, les personnes et les appareils de mesure qui produisaient les mesures finales de faux-rond, etc. L'ingénieur ne voulait pas que ces différences soient confondues avec les différences entre les deux techniques de chargement ou qu'elles compliquent inutilement le tableau. Pour l'éviter, il fallait faire preuve de prudence.

En fait, l'ingénieur a mené de main de maître une étude bien réfléchie. Le tableau 1.1.1.1 répertorie les valeurs de faux-ronds pour 38 engrenages empilés et 39 engrenages suspendus après le traitement thermique. Sous forme brute, ces valeurs n'évoquent pas grand chose. Comme elles ne sont pas organisées, on ne peut pas comprendre le tableau 1.1.1.1 au premier coup d'œil. Les données devaient être résumées.

Résumé des données

L'une des actions menées a consisté à calculer des résumés numériques des données. Par exemple, l'ingénieur de procédé a calculé les moyennes de faux-rond suivantes :

engrenages empilés = 12,6

engrenages suspendus = 17,9

Visualisation

Ensuite, il a résumé les données graphiquement, comme l'illustre la figure 1.1.1.1.

Variation

Grâce à ces résumés, certains faits deviennent évidents. D'abord, les valeurs de faux-ronds varient, même pour une méthode de chargement donnée. La variabilité est un fait omniprésent de la vie, et toute méthodologie statistique le reconnaît explicitement. Dans le cas des engrenages, il ressort de la figure 1.1.1.1 qu'il y a un peu plus de variation dans les engrenages suspendus que dans les engrenages empilés. Mais malgré la variabilité qui complique la comparaison entre les méthodes

de chargement, la figure 1.1.1.1 et les moyennes des deux groupes montrent que le faux-ronds des engrenages empilés est généralement inférieur à celui des faux-ronds des engrenages suspendus. Dans quelle mesure? Calculons la différence :

Faux-rond moyen, engrenages suspendus − faux-rond moyen, engrenages empilés = 5,3

Mais quelle est la précision de ce calcul? Les faux-ronds varient. Peut-on être sûr que la différence observée dans les moyennes

actuelles réapparaîtrait lors d'un autre essai? Ou est-ce simplement du bruit? L'empilement des engrenages coûte plus cher que leur suspension. Peut-on déterminer si ces dépenses supplémentaires sont justifiées?

Représentation des inférences à partir des données

Ces questions soulignent la nécessité d'utiliser des méthodes d'inférence statistique formelle à partir des données et de traduire ces inférences en conclusions pratiques. Les méthodes présentées dans ce texte peuvent, par exemple, être utilisées pour étayer les affirmations suivantes concernant l'empilement et la suspension d'engrenages :

- On peut être sûr à environ 90% qu'à long terme et dans les conditions de l'étude, la différence entre les

moyennes est comprise entre

3,2 et 7,4

- On peut être sûr à 95 % que 95 % des faux-ronds des engrenages empilés dans des conditions telles que celles de l'étude de l'ingénieur se situent entre

3,0 et 22,2

- On peut être sûr à 95 % que 95% des faux-ronds des engrenages suspendus dans des conditions telles que celles de l'étude de l'ingénieur se situent entre

0,8 et 35,0

Voilà des quantifications formelles de ce qui a été appris lors de l'étude des engrenages empilés et suspendus. Pour utiliser ces affirmations concrètement, l'ingénieur.e de procédé a dû les combiner avec d'autres informations, comme les conséquences d'un certain niveau de faux-rond et les coûts de suspension et d'empilement des engrenages. Il a aussi dû faire preuve d'un bon jugement technique. En fin de compte, l'amélioration du faux-rond était suffisamment importante pour justifier une dépense supplémentaire, et la méthode de l'empilement a été instaurée.

Engrenages empilés	Engrenages suspendus
5, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11, 11, 11, 11, 11, 11, 12, 12, 12, 12, 13, 13, 13, 13, 14, 14, 14, 15, 15, 15, 15, 16, 17, 17, 18, 19, 27	7, 8, 8, 10, 10, 10, 10, 11, 11, 11, 12, 13, 13, 13, 15, 17, 17, 17, 17, 18, 19, 19, 20, 21, 21, 21, 22, 22, 22, 23, 23, 23, 23, 24, 27, 27, 28, 31, 36

Tableau 1.1.1.1. *Faux-rond de la face de poussée (0,0001 po)*

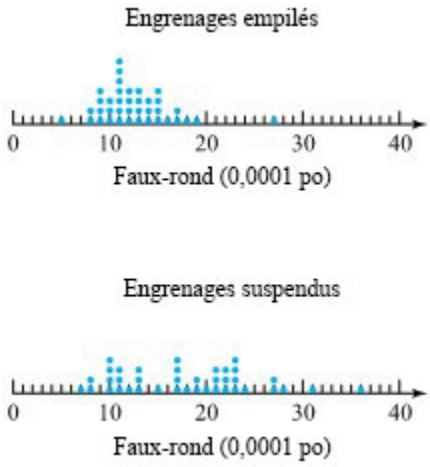


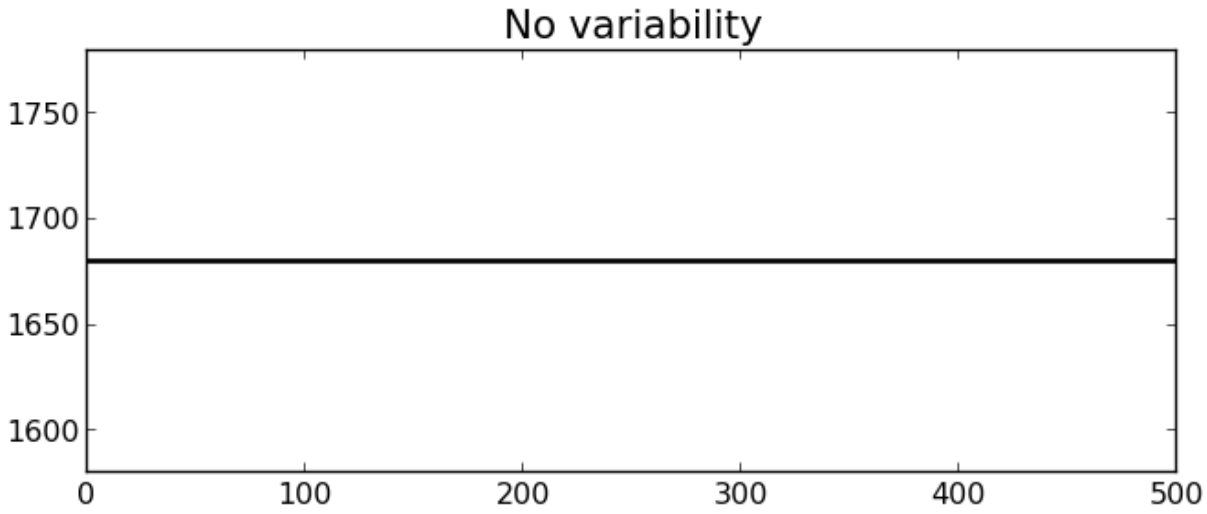
Figure 1.1.1.1. Diagramme de dispersion des faux-ronds

Cet exemple démontre comment les statistiques ont contribué à résoudre le problème d'un ingénieur. Tout au long de cet ouvrage, nous allons insister sur le fait que les sujets abordés ne sont pas des fins en soi, mais plutôt des méthodes qu'on peut utiliser pour travailler efficacement.

1.1.2 Variabilité

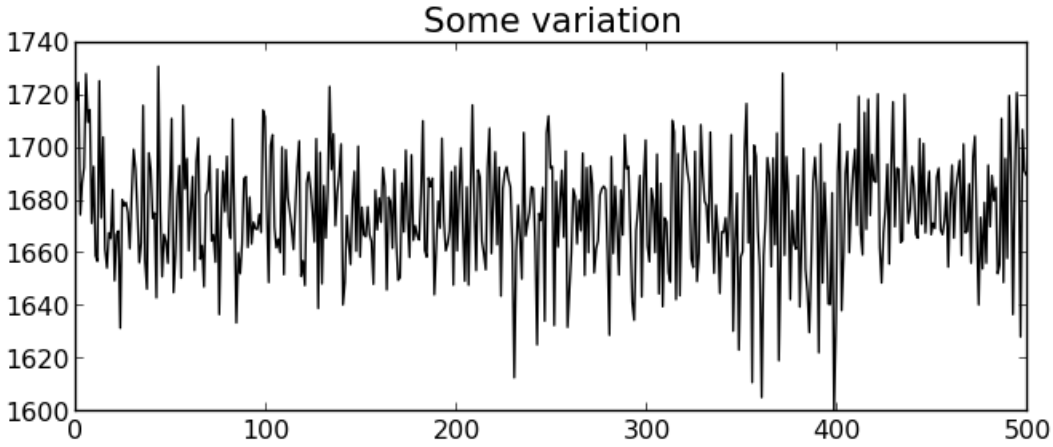
QU'EST-CE QUE LA VARIABILITÉ?

La vie est plutôt ennuyeuse sans variabilité, et ce cours (ainsi que presque tout le domaine des statistiques) serait inutile si les choses ne variaient pas naturellement.

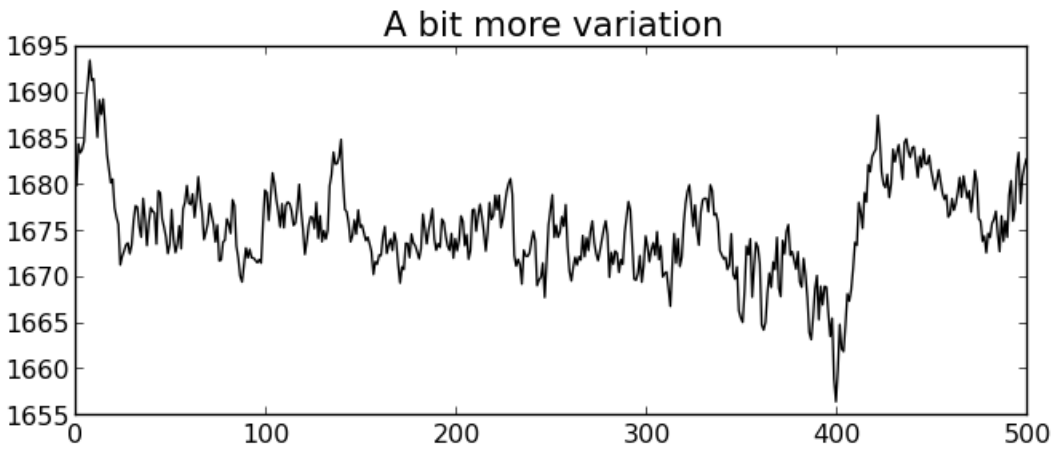


Heureusement les données consignées liées à nos processus et systèmes sont très variables :

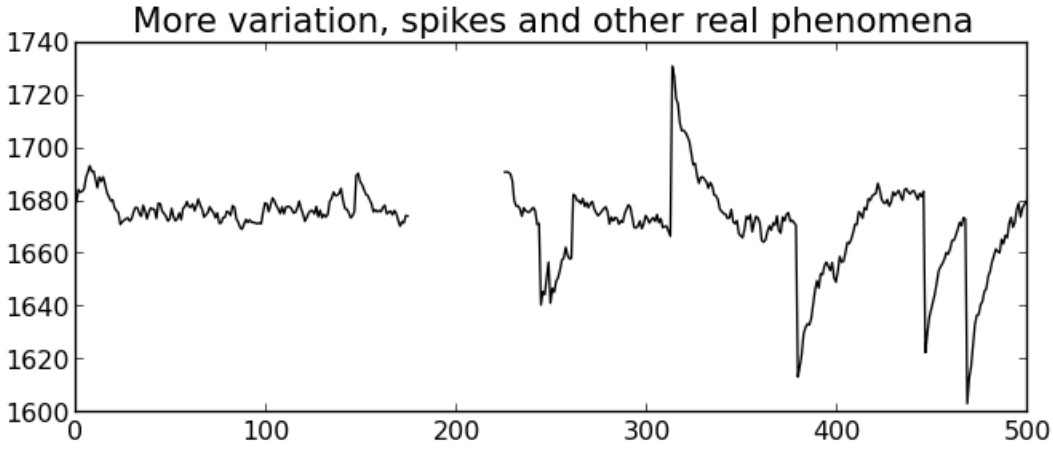
- Les propriétés des matières premières et des intrants sont loin d'être constantes.
- Il y a des sources inconnues, souvent dénommées « *erreur* » ou « *bruit* », dont l'explication dépasse notre compréhension du processus en question.



- Variabilité des mesures et de l'échantillonnage : dérive des capteurs, pics, bruit, décalages de recalibrage, erreurs dans l'analyse de l'échantillon, et équipement de laboratoire.



- Perturbations de la production :
    - modification des conditions extérieures, comme la température ambiante ou l'humidité, ou
    - bris ou usure de pièces d'équipement nécessitant un remplacement.



1.1.3 Types d'études et de méthodes statistiques

Lorsqu'on entreprend de collecter des données, il faut décider de son niveau d'activité. Doit-on tourner des boutons et manipuler des variables de procédés, ou doit-on se contenter d'observer et de consigner les caractéristiques qui ressortent?

DÉFINITION 1.2.3.1. Étude d'observation

Une étude d'observation est une étude dans laquelle l'ingénieur.e.s joue un rôle essentiellement passif. On observe un processus ou un phénomène et on consigne les données, mais sans intervenir.

DÉFINITION 1.2.3.2. Étude expérimentale

une étude expérimentale (ou plus simplement, une expérience) est une étude dans laquelle l'ingénieur.e joue un rôle actif. On manipule les variables de processus, et l'environnement de l'étude est contrôlé.

La plupart des études statistiques comprennent des volets d'observation et d'expérimentation; et ces deux définitions doivent être considérées comme les extrémités opposées idéalisées d'un continuum. Sur ce continuum, l'extrémité expérimentale offre généralement les moyens les plus efficaces et les plus fiables de collecter des données d'ingénierie. Il est généralement beaucoup plus rapide de manipuler les variables du processus et d'observer la réaction du système
aux modifications plutôt que d'observer de manière passive, en espérant remarquer un élément intéressant ou révélateur.

Déduction de la causalité

En outre, il est beaucoup plus facile et sûr de déduire la causalité d'une expérience, à partir d'une étude d'observation. Les systèmes réels sont complexes. Il est possible d'observer plusieurs exemples de bons fonctionnements d'un processus et de noter qu'ils ont tous été liés à
des circonstances X sans pour autant supposer que ces circonstances en sont la cause. Il peut y avoir des variables importantes en arrière-plan qui changent et qui sont la véritable raison du bon fonctionnement du système. Ces variables dites « cachées » peuvent régir à la fois le fonctionnement du processus et les circonstances X. Il se peut aussi que de nombreuses variables changent de manière aléatoire, sans avoir d'impact appréciable sur le système et que, par hasard, au cours d'une période d'observation limitée, certaines d'entre elles produisent les circonstances X au même moment où le système fonctionne bien. Dans un cas comme dans l'autre, les efforts pour recréer les circonstances X en espérant que les choses fonctionneront correctement seront des efforts inutiles.

En revanche, dans une expérience où l'environnement est largement régulé, à l'exception de quelques variables qu'on modifie délibérément, la déduction de la causalité est beaucoup plus forte. Si les circonstances de l'étude s'accompagnent systématiquement de résultats favorables, il est raisonnable de penser qu'elles en sont à l'origine.

Exemple 1.1.3.1. Granulation de l'hexamine en poudre

Cyr, Ellson et Rickard ont voulu réduire la portion de pastilles de combustible non conformes produites lors de la compression de poudre d'hexamine brute dans une machine à pastillage. De nombreux facteurs sont susceptibles d'influencer le pourcentage de pastilles non conformes, dont la vitesse de la machine, le remplissage de la filière, le pourcentage de paraffine ajouté à l'hexamine, la température ambiante, l'humidité lors de la fabrication, la teneur en humidité, la composition « neuve » ou « réutilisée » du mélange à granuler, et la rugosité de la goulotte dans laquelle pénètrent les pastilles fraîchement fabriquées. Il était impossible d'établir une corrélation entre ces nombreux facteurs et les performances du processus par le biais d'une observation passive.

Cependant, les étudiant.e.s ont pu réellement progresser en menant une expérience. Ils ont sélectionné trois des facteurs qui semblaient les plus importants et les ont modifiés tout en maintenant les autres facteurs aussi constants que possible. Les changements importants observés dans le pourcentage de pastilles de combustible acceptables ont été attribués à juste titre à l'influence des variables du système qui avaient été manipulées.

Outre la distinction entre les études statistiques observationnelles et expérimentales, il est utile de distinguer les études en fonction de l'étendue de l'application prévue des résultats. Vous trouverez ci-dessous la définition de deux termes, popularisés par feu W. E. Deming :

DÉFINITION 1.1.3.3. Étude énumérative

Une étude énumérative est une étude dans laquelle il existe un groupe particulier, bien défini et fini d'objets à étudier. Les données sont collectées sur tous ces objets ou sur une partie des objets, et les conclusions visent à s'appliquer uniquement à ces objets.

DÉFINITION 1.1.3.4. Étude analytique

Une étude analytique est une étude dans laquelle un processus ou un phénomène fait l'objet d'une recherche à un point dans l'espace et dans le temps, avec l'espoir que les données collectées seront représentatives du comportement du système à d'autres endroits et à d'autres moments dans les mêmes conditions. Dans ce type d'étude, il existe rarement, voire jamais, de groupe d'objets particulier bien défini auquel les conclusions sont censées se limiter.

En ingénierie, la plupart des études appartiennent à la seconde catégorie, même si certaines applications importantes impliquent un traitement énumératif. L'un de ces exemples est le test de fiabilité des composants critiques – par exemple, pour une utilisation dans une navette spatiale. Ce qui intéresse alors les ingénieur.e.s, ce sont les composants en question et leur performance, et non un problème plus large tel que « le comportement de tous les composants de ce type ». L'échantillonnage d'acceptation (dans lequel on vérifie un lot entrant avant de le réceptionner) représente un autre type significatif d'étude énumérative. Cependant, comme indiqué, la plupart des études liées à l'ingénierie sont de nature analytique.

Suite de l'exemple 1.1.3.1

Les étudiant.e.s qui se penchaient sur la machine de fabrication de pastilles ne s'intéressaient pas à un lot de pastilles particulier, mais plutôt à la question du fonctionnement efficace de la machine. Ils espéraient (ou supposaient tacitement) que ce qu'ils avaient appris sur la fabrication de pastilles resterait valide ultérieurement, au moins dans des conditions identiques à celles où ils ont mené leurs travaux. Leur étude expérimentale était de nature analytique.

Les deux définitions suivantes sont nécessaires, surtout en ce qui a trait aux études énumératives.

DÉFINITION 1.1.3.5. Population

Une population représente un groupe complet d'objets sur lesquels on souhaite collecter des informations dans le cadre d'une étude statistique.

DÉFINITION 1.1.3.6. Échantillon

Un échantillon est un groupe d'objets sur lesquels on souhaite collecter des données. Dans une étude énumérative, l'échantillon est un sous-ensemble de la population (et peut parfois englober la population complète).

La figure 1.1.3.1 illustre la relation entre une population et un échantillon. Si une caisse de 100 pièces de machine est livrée sur un quai de chargement et que 5 pièces sont examinées afin de vérifier l'acceptabilité du lot, les 100 pièces constituent la population d'intérêt et les 5 pièces forment un échantillon (unique) de taille 5 de la population. (Notez l'utilisation des mots ici: il y a un échantillon et non cinq).

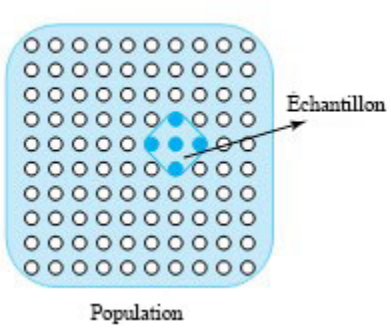


Figure 1.1.3.1. Population et échantillon

La définition des mots « population » et « échantillon » est souvent élargie de plusieurs façons. D'une part, il est courant de les utiliser pour désigner non seulement les objets étudiés, mais aussi les valeurs des données associées à ces objets. Par exemple, si on l'on considère les valeurs de dureté Rockwell associées à 100 pièces de machine dans une caisse, les 100 valeurs de dureté peuvent être désignées par le terme population (de nombres). Les cinq valeurs de dureté qui correspondent aux pièces examinées (l'échantillon d'acceptation) peuvent être désignées sous le terme d'échantillon tirée de cette population.

Suite de l'exemple 1.1.3.1

Cyr, Ellson et Rickard ont identifié huit ensembles différents de conditions expérimentales pour tester le fonctionnement de la machine à pastilles. Plusieurs cycles de production de pastilles ont été exécutés dans chaque ensemble de conditions, et chacun d'eux a produit son propre pourcentage de pastilles conformes. Ces huit ensembles de pourcentages peuvent désignés comme
huit échantillons (de nombres) différents.

Soit dit en passant, même si aucune population concrète à proprement parler ne fait l'objet d'une recherche dans le cadre d'une étude analytique, il est courant de faire référence à une population conceptuelle dans ce cas. Des expressions telles que « la population composée de tous les objets qui pourraient être produits dans ces conditions » sont parfois utilisées. Cela peut être source de confusion, mais il s'agit d'un usage courant, étayé par le fait que les mêmes mathématiques sont généralement utilisées pour tirer des conclusions dans des contextes énumératifs et analytiques.

TYPES DE MÉTHODES STATISTIQUES

Il y a deux grandes méthodes statistiques pour analyser des données : les statistiques descriptives et les statistiques inférentielles. Les statistiques descriptives résument les données d'un échantillon, par exemple en utilisant sa moyenne et son écart-type. Elles seront le sujet principal de la partie 2 de ce cours. Les statistiques inférentielles permettent de tirer des conclusions à partir de données extraites d'un échantillon soumis à des variations aléatoires. Les statistiques inférentielles s'appuient sur un modèle de probabilité pour décrire le processus à partir duquel les données ont été obtenues, ce que nous verrons dans les parties 3 et 4. Les données sont ensuite utilisées pour tirer des conclusions sur le processus en estimant les paramètres du modèle et en faisant des prédictions basées sur le modèle. Les tests statistiques

inférentiels formels seront abordés dans la partie 5 de ce cours. La figure 1.1.2.2 montre comment les statistiques descriptives et inférentielles sont liées.

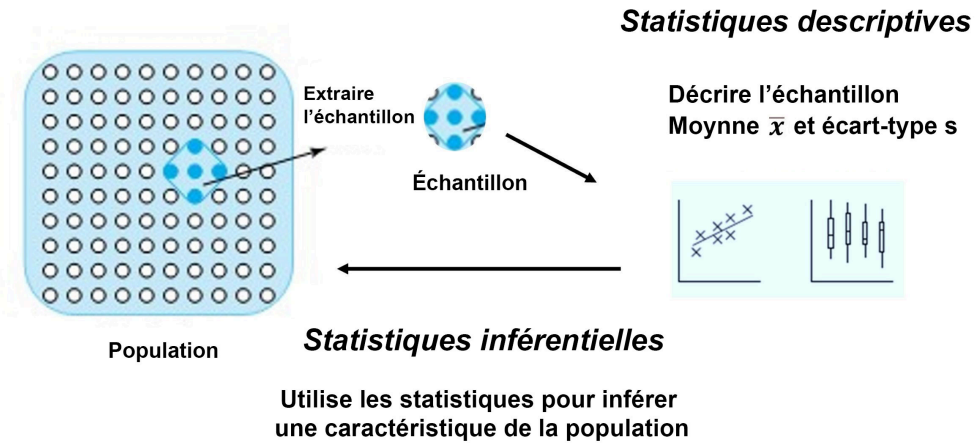


Figure 1.1.2.2. Lien entre les statistiques descriptives et inférentielles.

1.1.4 Échantillonnage

ÉCHANTILLONNAGE DANS LES ÉTUDES ÉNUMÉRATIVES

Une étude énumérative possède une population d'éléments concrète et identifiable. Ce chapitre décrit la sélection d'un échantillon des éléments à inclure dans une recherche statistique.

L'utilisation d'un échantillon pour représenter une population (généralement plus grande) confère des avantages évidents. Par exemple, il peut être très facile d'examiner certaines caractéristiques d'un échantillon de 30 composants électriques, tandis que le recensement (une étude comprenant tous les membres de la population) du lot de 10 000 peut s'avérer impossible. Parfois, l'essai est destructif et l'étude rend l'article inutilisable. Parfois, la rapidité d'exécution et la qualité des données d'une recherche par échantillonnage dépassent de loin tout ce qui pourrait être obtenu dans le cadre d'un recensement. Si l'étude prend beaucoup de temps, la technique de collecte de données peut se relâcher ou être négligée. Une quantité modérée de données collectées sous étroite supervision et utilisées immédiatement peut être très efficace – souvent plus efficace que les données d'une étude qui pourrait sembler plus complète, mais qui, en réalité, prendrait trop de temps.

Si un échantillon doit être utilisé pour représenter une population, le choix de cet échantillon devient primordial. L'échantillon doit représenter la population d'une manière ou d'une autre. La question traitée ici concerne la manière d'y arriver.

Les méthodes systématiques et fondées sur le jugement peuvent, dans certaines circonstances, produire des échantillons qui représentent fidèlement les caractéristiques importantes d'une population. Si un lot d'articles est fabriqué dans un ordre connu, il peut être raisonnable de sélectionner, par exemple, chaque vingtième article pour l'inclure dans une étude de statistiques d'ingénierie. Il peut aussi être pertinent de forcer l'échantillon à être équilibré, c'est-à-dire que chaque opérateur, chaque machine et chaque lot de matière première (par exemple) figure dans l'échantillon. Une personne ayant beaucoup d'expérience peut aussi être en mesure d'observer une population physique et d'en extraire un échantillon représentatif de manière assez fiable.

Mais ces méthodes de sélection d'échantillons peuvent poser des problèmes. Les humains sont sujets à des idées préconçues et à des préjugés conscients et inconscients, et par conséquent, les échantillons reposant sur le jugement peuvent ne pas refléter leur population. Les méthodes systématiques peuvent échouer gravement en présence de phénomènes cycliques imprévisibles. (Par exemple, supposons que l'on examine tous les 20^{e} articles d'un lot selon l'ordre dans lequel ils sortent de la chaîne de production. Supposons en outre que les articles soient traités à un moment donné sur une machine dotée de cinq têtes similaires, chacune effectuant la même opération sur un article sur cinq. L'examen de chaque 20^{e} article ne donne une image du comportement que d'une seule des têtes. Les quatre autres têtes pourraient être vraiment mal réglées, et il n'y aurait aucun moyen de s'en rendre compte).

Au-delà de ces problèmes, les méthodes d'échantillonnage systématiques ou basées sur le jugement ne permettent pas de quantifier les propriétés de manière utile. Il n'existe aucune manière efficace d'extraire des informations d'échantillons sélectionnées via ces méthodes et d'en tirer des conclusions fiables sur les marges d'erreur probables. La méthode présentée ci-après éviter les faiblesses des échantillonnages systématiques ou basés sur le jugement.

DÉFINITION 1.1.4.1. Échantillon aléatoire simple

Un échantillon aléatoire simple de taille n dans une population est un échantillon sélectionné de telle manière que chaque collection de n éléments de la population a, *a priori*, la même probabilité de composer l'échantillon.

La façon la plus simple d'envisager l'échantillonnage aléatoire simple est sans doute de dire qu'il équivaut, sur le plan conceptuel, à tirer n billets d'un chapeau qui contient un billet pour chaque membre de la population.

Exemple 1.1.4.1. Échantillonnage aléatoire des résident.e.s d'un dortoir

C. Black a réalisé une étude partiellement énumérative et partiellement expérimentale afin de comparer les temps de réaction des étudiant.e.s dans deux conditions d'éclairage différentes. Il a décidé de créer un échantillon aléatoire simple en recrutant 20 étudiant.e.s sélectionné.e.s au hasard dans son dortoir mixte. En fait, la méthode de sélection qu'il a utilisée consistait en une table de chiffres dits aléatoires. Aujourd'hui, il pourrait 'hui utiliser un générateur de nombres aléatoires à l'aide d'un logiciel de calcul statistique. Mais il aurait tout aussi bien pu écrire les noms de toutes les personnes vivant sur son palier sur des billets de taille uniforme, les mettre dans un bol, les mélanger soigneusement, fermer les yeux et en piger 20.

Méthodes mécaniques, tables de chiffres aléatoires et échantillons aléatoires simples

Pour sélectionner un échantillon aléatoire simple, on peut utiliser des méthodes mécaniques ou des méthodes utilisant des chiffres « aléatoires ». L'efficacité des méthodes mécaniques repose sur la symétrie et le mélange minutieux dans un dispositif physique de randomisation. En d'autres termes, les billets dans le chapeau doivent être de la même taille et bien mélangés avant que la sélection de l'échantillon ne commence.

La première loterie de conscription américaine pour la guerre du Vietnam est un cas célèbre où l'on n'a pas pris les précautions nécessaires pour garantir le bon fonctionnement d'un dispositif mécanique de randomisation. Les anniversaires étaient censés se voir attribuer les numéros de priorité 1 à 366 de manière « aléatoire ». Toutefois, il est apparu après coup que les boules représentant les dates de naissance avaient été placées dans un bac un mois à la suite de l'autre, et que le bac avait été mal mélangé. Lors du tirage des boules, les dates de naissance situées vers la fin de l'année ont reçu une part disproportionnée
des numéros les plus petits. Selon la terminologie actuelle, les cinq première dates de la corbeille ne doivent pas être considérées comme un simple échantillon aléatoire de taille 5. Les exploitants de jeux de hasard s'assurent (par la collecte de données appropriées) que leurs dispositifs mécaniques fonctionnent de manière aléatoire.

L'utilisation de chiffres aléatoires pour l'échantillonnage repose implicitement sur le caractère réellement aléatoire de la méthode utilisée pour générer les chiffres. Généralement, ces méthodes reposent sur des processus physiques aléatoires, comme la désintégration radioactive, ou des générateurs de nombres pseudo-aléatoires (des algorithmes numériques récursifs compliqués). Jusqu'à récemment, il était d'usage de consigner ces chiffres dans des tables imprimées.

Logiciel de statistique et échantillons aléatoires

Avec la démocratisation des ordinateurs personnels, les tables de chiffres aléatoires sont devenues complètement obsolètes. Désormais, on peut utiliser un logiciel statistique ou un tableur pour générer des nombres aléatoires au moment où on en a besoin.

Remarques sur l'échantillonnage aléatoire

Quelle que soit la mise en œuvre de la définition 1.1.4.1, plusieurs commentaires sur la méthode s'imposent. Tout d'abord, il convient d'admettre que l'échantillonnage aléatoire simple ne répond à l'objectif initial de fournir des échantillons représentatifs qu'en moyenne ou à long terme. Il est possible que certains échantillons ainsi sélectionnés ne soient absolument pas représentatifs de la population. Par exemple, un échantillon aléatoire simple de 20 essieux sur 80 pourrait en fait être composé des essieux ayant les plus petits diamètres. Mais cela ne se produit pas souvent. En moyenne, un échantillon aléatoire simple donnera une image fidèle de la population. La définition 1.1.4.1 énonce une méthode, et non une garantie de succès pour une application donnée de la méthode.

Ensuite, il convient également d'admettre qu'il n'existe aucune garantie qu'il sera facile de procéder à la sélection physique d'un échantillon aléatoire simple. Imaginez s'il fallait prendre cinq fours à micro-ondes précis sur un lot de 1 000 fours stockés dans un entrepôt. Ce serait probablement une tâche très désagréable que de localiser et de rassembler les cinq fours correspondant à des numéros de série choisis au hasard pour, par exemple, les transporter vers un
laboratoire d'essais.

Mais les avantages conférés par l'échantillonnage aléatoire simple compensent largement ses inconvénients. Premièrement, il s'agit d'un méthode objective d'échantillonnage. En l'utilisant, on se protège des biais humains conscients et inconscients. Deuxièmement, la méthode introduit des probabilités dans le processus de sélection d'une manière qui se révèle gérable. Par conséquent, la qualité des informations provenant d'un échantillon aléatoire simple peut être quantifiée. Ainsi, on peut utiliser les méthodes d'inférence statistique formelle, de même que les conclusions qui en découlent (« Je suis sûr à 95 % que… »).

1.1.5 Types de données

Les ingénieur.e.s gèrent de nombreux types de données. Il st souvent utile de les classer selon la mesure dans laquelle elles sont intrinsèquement numériques.

DÉFINITION 1.1.5.1. Données catégoriques

Les données qualitatives ou catégoriques sont les valeurs des caractéristiques fondamentalement non numériques associées aux éléments d'un échantillon. Elles peuvent parfois être ordonnées, mais il faut les agréger et les dénombrer pour produire des valeurs numériques significatives.

Considérons à nouveau un échantillon de cinq pièces de machine tiré d'une caisse de 100 pièces. S'il est possible de classer chaque pièce dans l'une des catégories (ordonnées) 1) conformes, 2) à retravailler, et 3) à jeter, et que l'on connaît les classifications des cinq pièces, on dispose alors de cinq points de données qualitatives. Si l'on dénombre trois pièces conformes, une à retravailler et une à jeter, on se retrouve alors avec un résumé numérique des données catégoriques.

Les données numériques s'opposent aux données catégoriques.

DÉFINITION 1.1.5.2. Données numériques

Les données quantitatives ou numériques sont les valeurs des caractéristiques numériques associées aux éléments d'un échantillon. Il s'agit généralement de compter le nombre d'occurrences d'un phénomène d'intérêt ou de mesurer une propriété physique des éléments.

En reprenant l'exemple des pièces de machine en caisse, les valeurs de dureté de Rockwell des cinq pièces sélectionnées pourraient constituer un ensemble de données (de mesure) quantitatives. Le nombre de défauts visibles sur une surface usinée pour chacune des cinq pièces sélectionnées constituerait un ensemble de données (de dénombrement) quantitatives.

Il est parfois pratique de faire comme si la précision des mesures était infinie. Sous cette hypothèse, les variables mesurées sont continues dans le sens où elles peuvent prendre n'importe laquelle des valeurs appartenant à une plage continue. Par exemple, on peut supposer que la dureté de Rockwell d'une pièce de machine se situe n'importe où dans l'intervalle $(0, \infty)$, mais il ne s'agit bien sûr que d'une idéalisation. En réalité, toutes les mesures sont effectuées à l'unité la plus proche (quelle que soit cette unité). Cela devient d'autant plus évident que les instruments de mesure sont de plus en plus souvent équipés d'écrans numériques. En réalité, lorsqu'on les examine d'assez près, toutes les données numériques (qu'elles soient mesurées ou comptées) sont discrètes, en ce sens qu'elles ne peuvent prendre que certaines valeurs, et non n'importe quelle valeur sur un continuum.

Bien que la plage $(0, \infty)$ soit mathématiquement utile et tout à fait adéquate à des fins pratiques, l'ensemble

réel des valeurs possibles pour la dureté de Rockwell mesurée d'une pièce de machine ressemble probablement davantage à {0,1, 0,2, 0,3,...} qu'à (0, ∞).

Il est généralement convenu que les données de mesure sont préférables aux données catégoriques ou de dénombrement. Les méthodes statistiques pour les mesures sont plus simples et plus éclairantes que les méthodes pour les données qualitatives et les dénombrements. En outre, de bonnes mesures nous renseignent généralement beaucoup plus que les données qualitatives. Toutefois, il faut parfois tenir compte du fait que les mesures peuvent prendre plus de temps (et donc coûter plus cher) que la collecte de données qualitatives.

Exemple 1.1.5.1. Mesures de la masse des pastilles

En préliminaire à leur étude expérimentale sur le processus de granulation (abordé dans l'exemple 1.1.3.1), Cyr, Ellson et Rickard ont recueilli des données sur un certain nombre d'aspects du comportement de la machine, dont la masse des pastilles produites dans des conditions d'utilisation normales. Étant donné que la majorité des non conformités résulte d'un détachement du matériau au cours de la production, la masse de la pastille est un indicateur de la performance du système. Les spécifications indiquaient que la masse devait être comprise entre 6,2 et 7,0 g.

Des données sur 200 pastilles ont été recueillies. Les étudiant.e.s auraient pu se contenter d'observer et de noter si une pastille donnée avait une masse conforme aux spécifications, produisant ainsi des données qualitatives, mais ils ont plutôt pris le temps de mesurer la masse des pastilles à 0,1 g près, recueillant ainsi des données de mesure. La figure 1.1.5.1 illustre le résumé de leurs constatations.

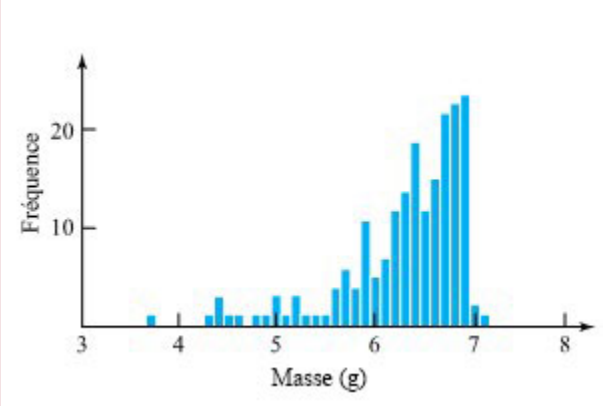


Figure 1.1.5.1 Mesures de la masse des pastilles

Remarquez qu'il est possible de récupérer les informations de conformité à partir des mesures : environ 28,5% (57 sur 200) des pastilles avaient des masses qui ne répondaient pas aux spécifications. Mais la figure 1.1.5.1 ne se limite pas à cela. La forme de la distribution peut donner des indications sur le fonctionnement de la machine et sur les conséquences potentielles de simples de modifications du processus de granulation. Par exemple, notez l'aspect tronqué de la figure. La queue avant des données ne ressemble en rien à la queue arrière. Les étudiant.e.s ont déduit que c'était dû au fait qu'après avoir été placée dans une matrice, la poudre passe sous une palette qui élimine l'excès de matière avant qu'un vérin ne la comprime dans la matrice. La quantité initialement distribuée dans une matrice donnée peut avoir une distribution assez symétrique en forme de monticule, mais la palette introduit probablement la caractéristique tronquée de l'affichage.

De plus, à partir des données numériques de la figure 1.1.5.1, on peut trouver un pourcentage de masses de pastilles dans n'importe quel intervalle d'intérêt, et pas seulement dans l'intervalle [6.2, 7.0]. En déplaçant mentalement la figure vers la droite, il est même possible de projeter les effets probables d'une augmentation de la taille des matrices dans des proportions variables.

Dans les études d'ingénierie, il est courant d'avoir plusieurs variables d'intérêt. Les définitions suivantes présentent des termes utiles pour préciser le nombre de variables impliquées et leur relation.

DÉFINITION 1.1.5.3. Données à une seule variable

Les données à une seule variable apparaissent lorsqu'on observe une seule caractéristique de chaque élément de l'échantillon.

DÉFINITION 1.1.5.4. Données à plusieurs variables

Les données à une plusieurs variable apparaissent lorsqu'on observe plusieurs caractéristiques de chaque élément de l'échantillon. Il y a un cas particulier concernant les**données à deux variables**.

DÉFINITION 1.1.5.5. Mesures répétées

Lorsque on obtient des données à plusieurs variables en mesurant plusieurs fois une caractéristique essentiellement identique (par exemple, avec des instruments différents ou à des moments différents), on parle de données à mesures répétées. Dans le cas particulier des réponses à deux variables, on parle de données appariées.

Il est important de reconnaître les données à plusieurs variables. Le fait de disposer de valeurs de dureté de Rockwell pour cinq des 100 pièces en caisse de machines et de déterminer le pourcentage de carbone pour cinq autres pièces n'est pas du tout équivalent au fait de disposer à la fois de valeurs de dureté et de teneur en carbone pour un échantillon unique de cinq pièces. Dans le premier cas, il y a deux échantillons de cinq points de données à une seule variable, tandis qu'il n'y a qu'un seul échantillon de cinq points de données à deux variables dans le second. La seconde situation est préférable à la première, car elle
permet d'analyser et de tirer parti de toute éventuelle relation entre les variables « dureté » et « pourcentage de carbone ».

Exemple 1.1.5.2. Mesures de distorsion appariée

Dans le scénario de chargement du four évoqué à l'exemple 1.1.1.1, les mesures de faux-ronds radiaux ont été en fait effectuées sur tous les (38 + 39 =) 77 engrenages avant et après le traitement thermique. (Le tableau 1.1 ne donne que les données après traitement.) On disposait donc de deux échantillons (de tailles respectives 38 et 39) de données appariées. Ainsi, on pouvait (si on le souhaitait) analyser la corrélation entre la distorsion après traitement et la distorsion avant traitement.

1.1.6 Mesure : Importance et difficultés

La réussite des études d'ingénierie statistique dépend de la capacité à effectuer des mesures. Pour des propriétés physiques, comme la longueur, la masse, la température, etc., les méthodes de mesure sont très courantes et évidentes. Souvent, ces propriétés suffisent à caractériser adéquatement le comportement d'un système d'ingénierie. Mais lorsque c'est impossible, il faut définir précisément ce qui est à observer dans le système, puis faire preuve d'ingéniosité pour créer une méthode de mesure adaptée.

Exemple 1.1.6.1. Mesure de la fragilité

Dans le cadre d'un projet de fin d'études en génie métallurgique, il fallait aider un fabricant à améliorer les performances d'une pièce métallique en forme de pointe. Dans l'application à laquelle elle était destinée, cette pièce devait être solide mais très fragile : lorsqu'elle rencontrait un obstacle sur son chemin, elle devait se briser plutôt que de se plier, car la flexion
aurait causé d'autres dommages à la machine dans laquelle la pièce fonctionnait. Pendant qu'ils planifiaient une étude statistique visant à déterminer les variables de fabrication qui affectent la performance des pièces, les étudiant.e.s ont réalisé que l'entreprise ne disposait pas d'un bon moyen pour évaluer la performance des pièces. Pour mener à bien leur étude, ils ont mis au point un appareil de mesure qui ressemblait grossièrement à l'illustration de la figure 1.1.7.1. Un bras oscillant avec une grande masse à son extrémité était placé en position horizontale, puis on le libérait pour qu'il aille frapper une pièce d'essai solidement
fixée en position verticale au bas de sa trajectoire. L'angle maximal au-delà de la verticale formé par le bras après l'impact avec la pièce constituait une mesure efficace de la fragilité.

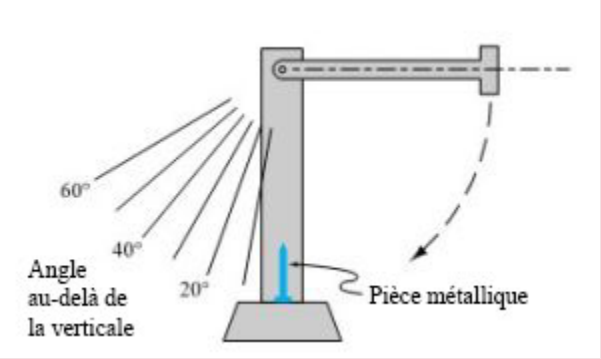


Figure 1.1.6.1. Appareil de mesure de la fragilité

Exemple 1.1.6.2. Mesure de la solidité d'un joint en bois

Dimond et Dix voulaient mener une étude visant à comparer la solidité d'un joint pour une combinaison de trois essences différentes et de trois colles. Comme ils ne disposaient pas d'un appareil d'essai de solidité, ils en ont inventé un. Pour tester un joint, ils ont suspendu un grand récipient à l'une des pièces de bois et y ont versé de l'eau jusqu'à ce le joint se brise sous le poids de l'eau. Connaissant le volume d'eau versé dans le récipient et la densité de l'eau, ils ont pu déterminer la force nécessaire pour briser le joint.

Qu'on utilise une technologie disponible sur le marché ou qu'il faille fabriquer un nouveau dispositif, un certain nombre de questions concernant les mesures doivent être prises en compte, dont la validité, les variations et les erreurs de mesure, l'exactitude et la précision.

DÉFINITION 1.1.6.1. Validité

Une mesure ou une méthode de mesure est considérée comme valide si elle représente de manière utile et correcte une caractéristique importante d'un objet ou d'un système.

On ne saurait trop insister sur l'importance de se poser la question de la validité des mesures avant de se lancer dans une étude d'ingénierie statistique. La collecte des données d'ingénierie coûte de l'argent, et trop souvent, on consacre des ressources considérables à recueillir des données qui, finalement, n'aident pas vraiment à résoudre le problème en question.

Erreur de mesure

La section 1.1.1.1 a souligné qu'en utilisant des données, on est rapidement confronté au fait que la variation est omniprésente. Une partie de ces variations est due au fait que les objets étudiés ne sont jamais exactement les mêmes. Mais une autre partie des variations découle du fait que les processus de mesure ont également leur propre variabilité inhérente. Si on a une échelle de mesure suffisamment précise, il sera impossible – quels que soient les efforts déployés – d'obtenir plusieurs fois la même mesure pour un même objet. Il est naïf d'attribuer toute variation dans les mesures répétées à une mauvaise technique ou à un manque de rigueur. (Bien entendu, une mauvaise technique et un manque de rigueur peuvent accroître la variation des mesures au-delà de ce qui est inévitable).

Un exercice suggéré par W. J. Youden dans son ouvrage *Experimentation and Measurements* illustre bien la réalité des erreurs de mesure. Essayez de mesure l'épaisseur du papier de ce livre en utilisant la technique suivante : Ouvrez le livre à une page située vers le début et à une page située vers la fin. Saisissez fermement la pile entre les deux pages
entre le pouce et l'index et mesurez l'épaisseur de la pile (à 0,1 mm près) à l'aide d'une règle d'écolier. En divisant l'épaisseur de la pile par le nombre de feuilles de la pile et en consignant le résultat à 0,0001 mm près, vous obtiendrez la mesure de l'épaisseur.

Exemple 1.1.6.3. Mesures d'épaisseur du papier d'un livre

Vous trouverez ci-dessous dix mesures de l'épaisseur du papier, tirées de l'ouvrage Statistics for Experimenters de Box, Hunter et Hunter, effectuées au cours d'un semestre par les étudiant.e.s en ingénierie Wendel et Gulliver.

Wendel : 0,0807,0,0826, 0,0854, 0,0817, 0,0824,
0,0799, 0,0812, 0,0807, 0,0816, 0,0804
Gulliver : 0,0972, 0,0964, 0,0978, 0,0971, 0,0960,
0,0947, 0,1200, 0,0991, 0,0980, 0,1033

La figure 1.1.6.2 présente un graphique de ces données et révèle clairement que même des mesures répétées par une même personne sur un même livre varient, et que les schémas de variation pour deux personnes différentes peuvent être tout à fait différents. (Les valeurs de Wendel sont à la fois plus petites et plus cohérentes que celles de Gulliver.)

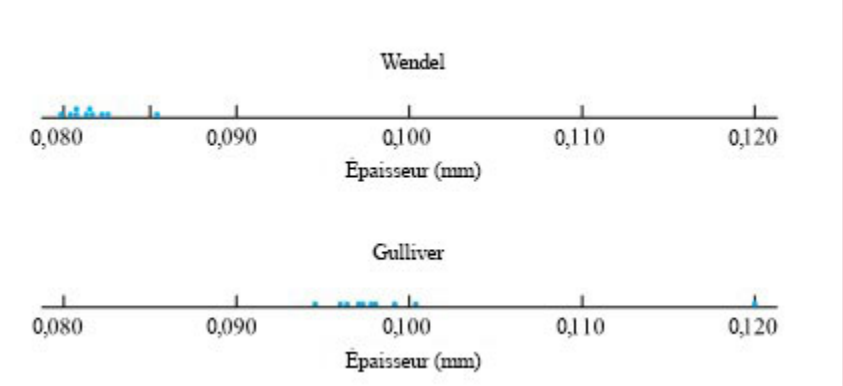


Figure 1.1.6.2. Diagrammes de dispersion des mesures d'épaisseur de papier

La variabilité qui est inévitable dans les mesures peut être considérée comme ayant des composantes internes et externes.

DÉFINITION 1.1.7.2. Précision

Un système de mesure est dit précis s'il produit de faibles variations lors de mesures répétées du même objet.

La précision correspond à la cohérence interne d'un système de mesure. Généralement, il est possible de l'améliorer en n'opérant que quelques petites modifications dans la configuration du système.

Suite de l'exemple 1.1.6.3.

En ignorant la possibilité qu'une propriété du livre de Gulliver soit à l'origine de la plus grande dispersion de ses valeurs par rapport à celles de Wendel, il apparaît que la technique de mesure de Wendel était plus précise que celle de Gulliver. Il aurait probablement été possible d'améliorer la précision des mesures de Gulliver et de Wendel en leur donnant un trombone et un micromètre. Le trombone exerce une pression relativement constante sur les piles de pages à mesurer, éliminant ainsi la subjectivité et les variations liées à la saisie ferme de la pile entre le pouce et l'index. Pour déterminer l'épaisseur de la pile, le micromètre est un instrument nettement plus précis que la règle.

La précision des mesures est importante, mais pour de nombreux objectifs, elle n'est pas suffisante.

Définition 1.1.7.3 Exactitude
un système de mesure est dit exact (ou parfois, non biaisé) si, en moyenne, il produit la vraie valeur d'une quantité mesurée.

L'exactitude représente la conformité d'un système de mesure à une norme externe quelconque. Il s'agit d'une propriété qui peut généralement être modifiée sans nécessiter de changement physique important à la méthode de mesure. La calibration (aussi appelée étalonnage) d'un système peut être aussi simple que de comparer les mesures du système à un étalon, d'élaborer un schéma de conversion approprié, puis d'utiliser les valeurs converties au lieu des valeurs brutes du système.

Suite de l'exemple 1.1.6.3.

On ne sait pas ce que la méthode de mesure standard du secteur aurait donné pour l'épaisseur du papier dans la copie du texte de Wendel, mais supposons qu'une valeur de 0,0850 mm/feuille soit appropriée. Le fait que les mesures de Wendel aient été en moyenne de 0,0817 mm/feuille suggère que son exactitude future pourrait être améliorée en procédant comme précédemment, mais en multipliant tout chiffre obtenu par le rapport 0,0850/0,0817, soit 1,04.

Au Canada, les ensembles de référence des mesures physiques sont établis par Mesures Canada. Aux États-Unis, c'est le National Institute of Standards and Technology qui s'en charge. C'est un travail important. Des appareils de mesure mal calibrés peuvent être suffisants pour comparer les conditions locales, mais pour établir les valeurs des quantités dans un sens absolu, ou pour que les valeurs locales puissent être utilisées en d'autres lieux et à d'autres moments, il est essentiel d'étalonner les systèmes de mesure par rapport à un étalon constant. Un millimètre aujourd'hui en Ontario doit correspondre à un millimètre la semaine dernière en Colombie-Britannique.

Exactitude et études statistiques

La possibilité de biais ou d'inexactitude dans les systèmes de mesure a au moins deux implications importantes pour la planification des études d'ingénierie statistique. La première, c'est que les appareils doivent être surveillés et recalibrés au besoin. Le phénomène bien connu de la dérive des instruments peut ruiner une étude statistique autrement irréprochable. La deuxième, c'est que dans la mesure du possible, il faut utiliser un seul système pour toutes les mesures. Si on a recours à plusieurs appareils et plusieurs personnes, il devient difficile de déterminer si les différences observées sont attribuables aux variables à l'étude, ou aux instruments et aux gens. S'il faut absolument utiliser plusieurs appareils, ceux-ci doivent être calibrés par rapport à un étalon (ou au moins les uns par rapport aux autres). L'exemple suivant illustre le rôle du facteur humain dans les différences.

Exemple 1.1.6.4. Différences d'utilisation d'une jauge entre plusieurs technicien.ne.s

Cowan, Renk, Vander Leest et Yakes ont travaillé avec une entreprise sur la surveillance d'une dimension critique

d'une pièce métallique de haute précision produite sur un tour à commande numérique. Ils ont constaté une variation importante et initialement inexplicable de cette dimension entre les différentes équipes de l'usine. Cette variation a finalement été attribuée
non pas à une différence réelle entre les pièces d'une équipe à l'autre, mais à une instabilité du système de mesure de l'entreprise. Toutes les équipes utilisaient la même jauge pour mesurer la dimension critique, mais pas de la même manière. L'entreprise a donc dû former les technicien.ne.s pour leur montrer à utiliser la jauge d'une seule et unique façon standardisée.

Pour illustrer la différence entre précision et exactitude, prenons l'exemple d'une cible. Si on tire sur la cible, on peut être sur la cible ou en dehors (exactitude), et les tirs peuvent être groupés ou non (précision ou imprécision). La figure 1.1.7.2 illustre cette analogie.

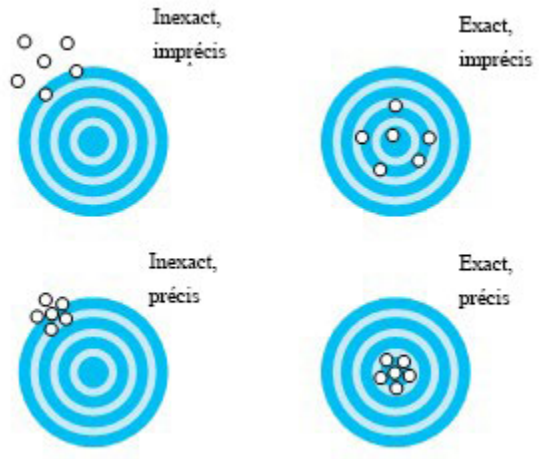


Figure. 1.1.6.2. Analogie entre la mesure et le tir sur cible.

La prise de bonnes mesures est un travail difficile, mais sans elle, la collecte de données est vaine. Pour progresser, il faut obtenir des mesures valides, effectuées par des méthodes suffisamment précises et exactes pour pouvoir identifier les changements importants dans le comportement du système. En règle générale, l'inexactitude et l'imprécision des mesures doivent être inférieures d'un ordre de grandeur à la variation de la réponse causée par ces changements.

1.1.7 Modèles mathématiques, réalité et analyse des données

Il est possible d'apprendre les bases de la statistique et les méthodes statistiques de l'ingénierie sans comprendre les mathématiques sous-jacentes. Les statistiques contiennent une bonne quantité de mathématiques que la plupart des ingénieur.e.s trouveront raisonnablement compréhensibles, bien que peu familières et initialement déroutantes. Mais si on adopte une approche mathématique au contexte d'apprentissage, on pave la voir vers une utilisation approfondie et améliorée des méthodes statistiques en ingénierie. C'est aussi une bonne manière d'appliquer la théorie mathématique apprise dans une application pratique. Il semble donc judicieux d'essayer de mettre en perspective le contenu mathématique du livre dès le début. Cette section porte sur les relations entre les mathématiques, le monde physique et les statistiques d'ingénierie.

Modèles mathématiques et réalité

Les mathématiques sont une construction et un outil. Bien qu'elles présentent un intérêt en soi pour certaines personnes, les ingénieur.e.s les abordent généralement du point de vue de leur utilité pour décrire et prédire le comportement des systèmes physiques. En effet, les théories mathématiques sont des guides dans toutes les disciplines de génie moderne.

Tout au long de ce texte, nous utiliserons fréquemment l'expression « modèle mathématique ».

DÉFINITION 1.1.7.1. Modèle mathématique

Un modèle mathématique est une description ou un résumé des principales caractéristiques d'un système ou d'un phénomène réel, sous forme de symboles, d'équations, de nombres, etc.

Les modèles mathématiques ne sont pas la réalité, mais ils peuvent être des descriptions extrêmement efficaces de la réalité. Cette efficacité repose sur deux propriétés quelque peu opposées d'un modèle mathématique: 1) son degré de simplicité, et 2) sa capacité de prédiction. Les modèles mathématiques les plus puissants sont ceux qui sont à la fois simples et qui produisent de bonnes prédictions. La simplicité d'un modèle permet d'opérer dans son cadre en utilisant des hypothèses de base pour tirer des conséquences mathématiques, lesquelles forment des prédictions sur le comportement du processus. Lorsque ces prédictions sont empiriquement correctes, c'est que le modèle est un outil efficace.

Les lois de Newton sont un exemple remarquable de modélisation mathématique efficace. Par exemple, la simple affirmation mathématique que l'accélération gravitationnelle est constante,

$$a = g$$

permet, après une opération mathématique facile (une intégration), de prédire qu'après un temps$t$, un objet initialement au repos en chute libre se déplacera à la vitesse

$$v = gt$$

Une deuxième intégration permet de prédire qu'après un temps t, le même objet aura parcouru la distance

$$d = \frac{1}{2}gt^2$$

L'avantage est que, dans la plupart des cas, ces prédictions simples sont très adéquates. Elles correspondent bien à ce qui est observé de manière empirique et peuvent être prises en compte lorsqu'on conçoit, construit, exploite ou améliore des processus physiques ou des produits.

Mais alors, quel rôle la notion de modélisation mathématique joue-t-elle dans les statistiques d'ingénierie? Elle joue plusieurs rôles, en fait. D'une part, la collecte et l'analyse des données sont essentielles pour ajuster ou estimer les paramètres des modèles mathématiques. Pour illustrer cela, reprenons l'exemple du corps en chute

Modèles mathématiques appliqués aux statistiques

libre. Si l'on postule que l'accélération due à la gravité est constante, il reste ensuite à définir la valeur numérique de cette constante. Il faut évaluer le paramètre g avant d'utiliser le modèle à des fins pratiques. Pour cela, il faut collecter des données.

Souvent, au niveau postsecondaire, le premier laboratoire de physique consiste à évaluer empiriquement la valeur de g. La méthode la plus couramment utilisée consiste à laisser tomber une masse le long d'un fil vertical passant par un trou en son centre et à laisser un courant électrique de 60 Hz former un arc entre le fil et un autre fil, brûlant légèrement une bande de papier intercalée entre les deux fils à chaque cycle. La figure 1.1.7.1 illustre ce genre de montage. Les marques de brûlure sont indiquent la position de la masse à des intervalles de $\frac{1}{60}$ d'une seconde. Le tableau 1.1.7.1 répertorie les mesures de ces positions. (Le ruban a été fourni par Frank Peterson, du département de physique et d'astronomie de l'ISU.) Le tracé des positions de la masse dans le tableau à intervalles égaux produit le tracé approximativement quadratique illustré à la figure 1.1.7.2. Pour obtenir la valeur de g, il suffit de calculer la courbe de régression. La méthode d'ajustement de courbe par moindres carrés donne une valeur de 9,79$m/_{sec^2}$ pour g, très proche de la valeur communément admise de 9,8 $m/_{sec^2}$.

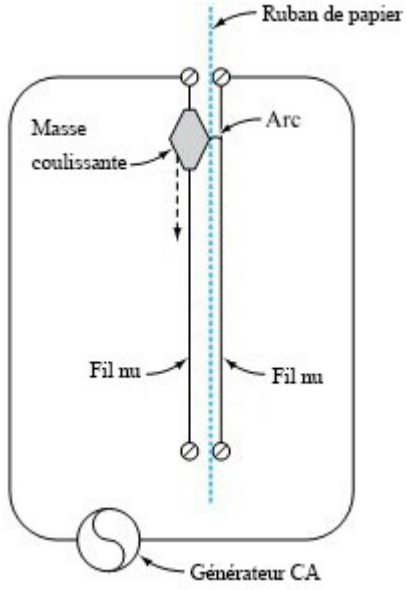


Figure 1.1.7.1. Appareil de mesure de la valeur g

Numéro de point	Déplacement (mm)	Numéro de point	Déplacement (mm)
1	0,8	13	223,8
2	4,8	14	260,0
3	10,8	15	299,2
4	20,1	16	340,5
5	31,9	17	385,0
6	45,9	18	432,2
7	63,3	19	481,8
8	83,1	20	534,2
9	105,8	21	589,8
10	131,3	22	647,7
11	159,5	23	708,8
12	190,5		

Tableau 1.1.7.1. Mesure du déplacement de la masse en chute libre

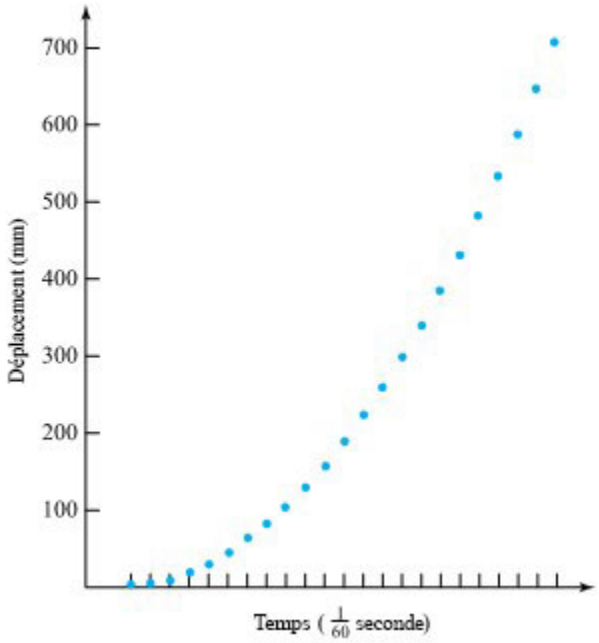


Figure 1.1.7.2. Position de la masse en chute libre

Notez que (au moins, avant Newton) les données du tableau 1.1.7.1 auraient pu être utilisées d'une autre manière. La forme parabolique du tracé de la figure 1.1.7.2 aurait pu suggérer la forme du modèle décrivant le mouvement d'un corps en chute libre. En observant attentivement le tracé de la position en fonction du temps, on devrait conclure qu'il existe une relation approximativement quadratique entre la position et le temps (et, à partir de là, faire deux dérivées pour conclure que l'accélération gravitationnelle est à peu près constante). Ce manuel regorge d'exemples montrant à quel point il peut être utile d'utiliser des données pour, d'une part, identifier des formes potentielles de modèles empiriques et, d'autre part, pour évaluer les paramètres de ces modèles (ce qui nous permettra de les utiliser pour faire des prédictions).

Cette discussion s'est concentrée sur le fait que les statistiques fournissent la matière première pour développer des modèles mathématiques réalistes de systèmes réels. Mais il existe un autre interaction essentielle entre les statistiques et les mathématiques. La théorie mathématique des probabilités fournit un cadre permettant de quantifier l'incertitude associée aux inférences liées aux données.

DÉFINITION 1.1.7.2. Probabilité

La probabilité est la théorie mathématique servant à décrire les situations et les phénomènes que l'on qualifierait familièrement d'aléatoires.

Si, par exemple, cinq étudiant.e.s obtiennent les cinq valeurs expérimentales de g suivantes :

9,78, 9,82, 9,81, 9,78, 9,79

on se demande naturellement comment utiliser ces données pour énoncer à la fois une meilleure valeur pour g et une certaine mesure de précision pour cette valeur. La théorie des probabilités sert à résoudre ces

questions. Le contenu du chapitre 3 montre que les considérations de probabilité permettent d'utiliser la moyenne de la classe de 9,796 pour estimer la valeur $g$et d'y attacher une précision de l'ordre de ± 0,02 m/sec^2.

Les mathématiques des probabilités constituent un sujet à part entière, aussi ce texte ne fournira-t-il qu'une introduction minimale au sujet. Mais il ne faut pas perdre de vue que les probabilités et statistiques ne sont pas synonymes. La probabilité est plutôt une branche des mathématiques et une matière utile en soi. Elle se retrouve dans un cours de statistique en tant qu'outil parce que la variation que l'on observe dans les données réelles est étroitement liée, d'un point de vue conceptuel, à la notion de hasard modélisée par la théorie des probabilités.

1.1.8 Taxonomie des variables dans un modèle

La nature multidimensionnelle du monde représente l'une des difficiles réalités de la modélisation statistique et de la planification d'expériences. On voudrait souvent mener des expériences pour comprendre les systèmes observés mais non expérimentaux et leur performance, ce qui nous permettrait d'en décrire de nombreuses caractéristiques et d'identifier les nombreuses variables qui les affectent. À la lumière de cette complexité, il convient de disposer d'une terminologie précise pour faciliter la réflexion et la discussion.

DÉFINITION 1.1.8.1. Variable de réponse

Une variable de réponse dans une expérience est une variable surveillée pour caractériser la performance ou le comportement du système. Il s'agit de la variable dépendante dans le modèle du système.

DÉFINITION 1.1.8.2. Variable d'entrée

Pour les données existantes recueillies autrement que par une expérience, une variable d'entrée d'un système se comporte comme la variable qui influence le modèle, ou la variable indépendante d'intérêt dans le modèle du système.

Dans les études expérimentales, la variable d'entrée est une variable supervisée (ou gérée) au cours de l'expérience, sur laquelle on exerce un certain contrôle en sélectionnant un ou plusieurs réglages à utiliser dans le cadre de l'étude. Lorsqu'une variable supervisée est maintenue constante (à un seul réglage), on dit alors qu'il s'agit d'une variable contrôlée; lorsqu'une variable supervisée prend plusieurs valeurs dans le cadre d'une étude, on dit qu'il s'agit d'une variable expérimentale.

Toutefois, on observe aussi certaines des variables qui ne sont ni des réponses principales, ni des variables gérées au cours d'une expérience.

DEFINITION 1.1.8.3 Variable d'accompagnement

Une variable d'accompagnement (ou variable concomitante) dans une expérience est une variable qui est identifiée et comprise dans l'analyse, mais qui n'est ni une variable de réponse principale, ni une variable d'entrée. Une telle variable peut changer en réaction à des variables d'entrée ou à des causes inconnues. En outre, cette variable peut avoir ou non une impact sur une variable de réponse.

La figure 1.1.8.1 illustre les définitions 1.1.8.1 à 1.1.8.3. Dans cette figure, le processus physique de la boîte noire produit d'une manière ou d'une autre des valeurs de réponse au cours d'une expérience. Les « boutons » du processus représentent les variables gérées. Les variables concomitantes flottent dans l'environnement de l'expérience sans en être l'objet principal.

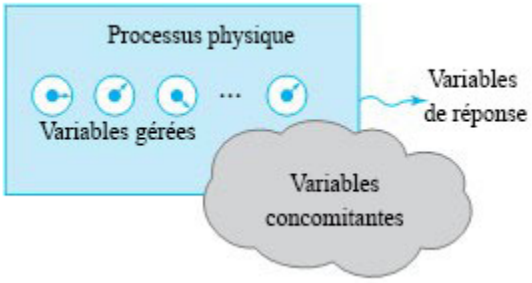


Figure 1.1.8.1. Les variables dans une expérience ("Basic Engineering Data Collection and Analysis" de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.)

L'identification des variables susceptibles d'affecter la réponse du système nécessite une connaissance approfondie du processus étudié. Les ingénieur.e.s qui n'ont pas d'expérience pratique d'un système peuvent parfois apporter des idées tirés de leur expérience avec des systèmes similaires et de la théorie de base mais il est également sage (et, dans la plupart des cas, essentiel) d'inclure dans une équipe de projet plusieurs personnes qui ont une connaissance directe du processus en question et de s'entretenir longuement avec celles et ceux qui travaillent régulièrement avec le système.

En général, l'identification des facteurs potentiellement importants dans une étude d'ingénierie statistique est une activité de groupe, réalisée dans le cadre de séances de remue-méninges. Il est donc utile de disposer d'outils pour mettre de l'ordre dans ce qui serait sinon un processus inefficace et désorganisé. Un outil qui s'est révélé efficace est connu sous le nom de diagramme de cause et d'effet, de diagramme en arête de poisson ou de diagramme d'Ishikawa. La figure 1.1.9.2 représente un modèle de diagramme en arête de poisson pour un système. Dans une analyse des causes profondes, la méthode 5M (ou 8M) est l'un des cadres les plus courants de l'analyse des causes profondes. (Contributeurs Wikipedia. [2023b, 3 décembre.] Diagramme de causes et effets. Wikipedia. https://fr.wikipedia.org/wiki/Diagramme_de_causes_et_effets).

Si on ne prend pas le temps de réfléchir à ces variables de manière organisée, il est souvent difficile de dresser une liste complète des facteurs importants d'un système réel complexe.

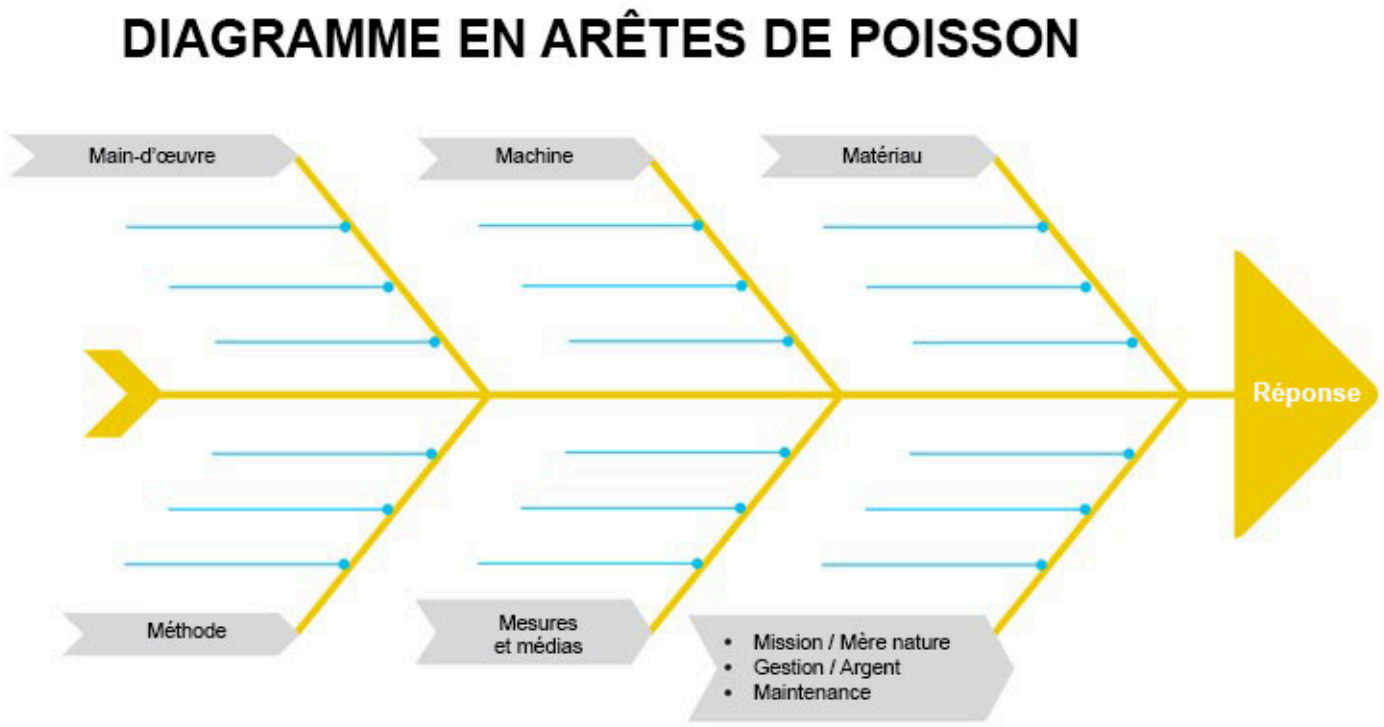


Figure 1.1.8.2. Diagramme en arête de poisson d'un système.

1.1.9 Tutoriel 1 – Exploration des données à l'aide de Python

À ce stade, il est recommandé de travailler sur l'exercice du tutoriel 1 qui se trouve sur le référentiel GitHub. Cet exercice vous montrera à importer et à manipuler des données avec Python.

Il est fortement recommandé de consulter le fichier du Jupyter Notebook intitulé Reading Data into Python & Data Cleaning. Vous pouvez les trouver dans la section « How do I do X in Python? ».

2.0.1 Résumer, visualiser et communiquer des données – Introduction

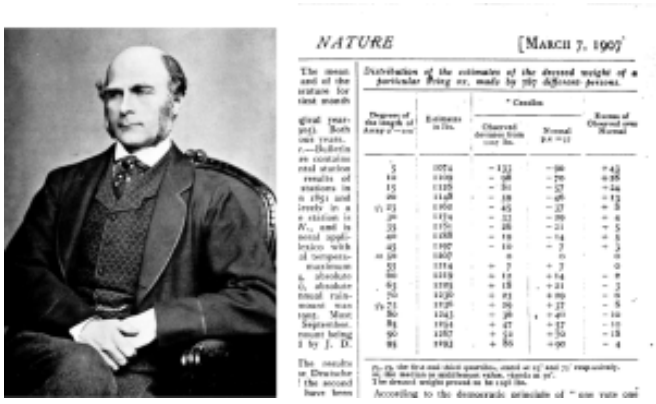


Figure 2.0.1.1. Sir Francis Galton, photographie probablement prise dans les années 1850 ou au début des années 1860, image tirée de Wikipedia https://en.wikipedia.org/wiki/Francis_Galton#/media/File:Francis_Galton_1850s.jpg et de l'article de Nature Vox Populi 1907, image tirée de https://galton.org/cgi-bin/searchImages/search/essays/pages/galton-1907-vox-populi_1.htm.

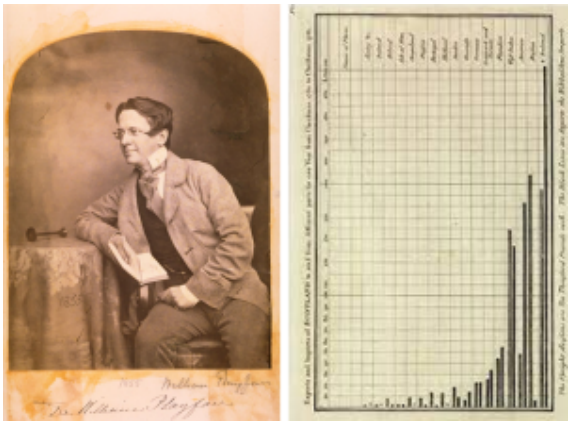


Figure 2.0.1.2. William Playfair, images tirées de Wikipedia https://en.wikipedia.org/wiki/William_Playfair.

Le polymathe britannique Francis Galton (1822-1911; figure 2.0.1.1) a été un pionnier dans l'utilisation de statistiques synthétiques. Fasciné par les mesures et la quantification, il a développé des concepts statistiques novateurs (bien que profondément problématiques) pour y répondre. Il a notamment fait une observation intelligente concernant la médiane lors d'un concours d'estimation du poids d'un bœuf lors d'une foire de bétail. Intrigué par la diversité des estimations, Galton a analysé les données et a constaté que si les estimations individuelles variaient considérablement, la médiane des estimations était étonnamment proche du poids réel du bœuf. Cette découverte a mis en évidence l'efficacité de la médiane en tant que mesure de la tendance centrale, notamment en ce qui concerne sa robustesse face aux valeurs aberrantes et aux données asymétriques. Cette avancée a été publiée dans la revue Nature en 1907.

William Playfair (1759-1832; figure 2.0.1.2) est considéré comme le fondateur des méthodes graphiques en statistiques, et notamment des diagrammes topologiques, à barres, de zones et en secteurs. Il a révolutionné la manière de présenter les données et a démontré que les graphiques pouvaient communiquer des informations plus efficacement que les tableaux de données. Une fois qu'on les a décrites et résumées à l'aide de statistiques descriptives, les données peuvent être présentées sous de nombreuses formes de visualisation graphiques, ce qui permet d'en extraire des conclusions.

Le besoin et la croissance des visualisations de données démontrent le rôle critique des graphiques statistiques en tant qu'outils efficaces pour comprendre la distribution et la forme des données. Contrairement à une simple collection de chiffres, les graphiques fournissent une représentation visuelle qui permet de discerner plus facilement les groupes de données, les tendances et les valeurs aberrantes. Il s'agit d'une pratique largement utilisée dans divers médias et industries pour comparer et communiquer les données rapidement et efficacement.

Principaux points à retenir

Les graphiques fournissent une représentation visuelle des données et permettent de communiquer et de décrypter des statistiques descriptives.

Nous nous concentrons sur les méthodes graphiques fondamentales comme les histogrammes, les diagrammes

en barres, les diagrammes en boîte, les séries temporelles et les nuages de points (aussi appelés diagrammes de dispersion). Les applications pratiques de ces concepts sont démontrées par des exercices utilisant des tutoriels Jupyter Notebook basés sur Python. Nous conclurons en soulignant les principes de l'excellence graphique et l'importance de créer des graphiques informatifs, véridiques et visuellement utiles.

Dans l'ensemble, ce module fournit un bon mélange de concepts théoriques, d'applications pratiques et d'outils de calcul statistique essentiels pour maîtriser la communication graphique des données dans les statistiques en génie biomédical.

Objectifs d'apprentissage

Objectifs d'apprentissage du module 2 :

- Découvrir les résumés statistiques descriptifs basés sur la tendance centrale et la répartition des données.
- Apprendre à construire et à interpréter différents types de graphiques tels que les histogrammes, les diagrammes à barres et les diagrammes en boîte.
- Comprendre comment les statistiques descriptives résument et décrivent les caractéristiques d'un ensemble de données par le biais de visualisations.
- Créer et interpréter une visualisation appropriée des données et comprendre l'utilité des techniques graphiques pour découvrir et résumer des tendances et des comparaisons dans les données.
- Comprendre comment utiliser des graphiques simples de séries temporelles pour visualiser les caractéristiques importantes de données temporelles.
- Appliquer les principes de l'excellence graphique et de la présentation efficace des données.

Objectifs d'apprentissage du module 2 – Tutoriels Jupyter Notebooks:

- Utiliser un logiciel statistique pour résumer, visualiser et interpréter des données.
- Apprendre à créer des tracés de base en utilisant les bibliothèques de traçage de Python.

2.0.2 Sources de la partie 2

Cette première version de la partie 2 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs. ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte et liés tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de ce travail sont la propriété intellectuelle de Kevin Dunn, et sont partagées à travers la licence CC BY-SA 4.0.

2.1.1 Introduction aux données quantitatives et aux quantiles

Les données d'ingénierie sont toujours variables. Si les mesures sont suffisamment précises, même des conditions de processus supposément constantes produisent des réponses différentes. Par conséquent, il ne faut pas s'attarder aux valeurs individuelles des données, mais plutôt aux tendances ou à la distribution de ces réponses. Synthétiser des données consiste à décrire les caractéristiques principales de leur distribution. Ce chapitre présente des méthodes utiles pour ce faire.

TRAITEMENT GRAPHIQUE ET TABULAIRE ÉLÉMENTAIRE DES DONNÉES QUANTITATIVES

La vaste majorité du temps, le premier pas dans l'analyse des données, c'est de représenter les données adéquatement sous forme de graphique ou de tableau. En effet, lorsqu'on n'a que quelques échantillons, une bonne image ou un bon tableau peut souvent suffire à en dire long sur les données. Les chapitres suivants traitent de l'utilité des diagrammes en points, des diagrammes à tiges et à feuilles, des tableaux de fréquences, des histogrammes, des diagrammes de dispersion et des organigrammes d'exploitation.

QUANTILES ET OUTILS GRAPHIQUES CONNEXES

Après cette présentation de quelques méthodes élémentaires de synthèse des données sous forme de graphiques et de tableaux, nous aborderons les les concepts de quantiles d'une distribution et les utiliserons pour réaliser d'autres représentations graphiques utiles.

2.1.2 Diagrammes à points et diagrammes à tiges et à feuilles

Lorsqu'une étude produit une quantité faible ou modérée de données quantitatives à une seule variable, un diagramme à points (qu'on peut facilement créer avec un crayon et du papier) est souvent très révélateur. Ces diagrammes présentent chaque observation sous la forme d'un point placé, sur une droite numérique, à une position correspondant à sa valeur.

Exemple 2.1.2.1. Représentation du faux-rond des engrenages

Au module 1.1, nous avons abordé un problème de traitement thermique portant sur la distorsion des engrenages empilés et des engrenages suspendus. Cette figure est reproduite ici à la figure 2.1.2.1. Il s'agit de deux diagrammes à points, l'un montrant les valeurs de faux-rond de la face de poussée pour les engrenages empilés, et l'autre, les valeurs correspondantes pour les engrenages suspendus. On voit clairement que les valeurs des engrenages empilés sont généralement plus petites et moins dispersées que celles des engrenages suspendus.

.

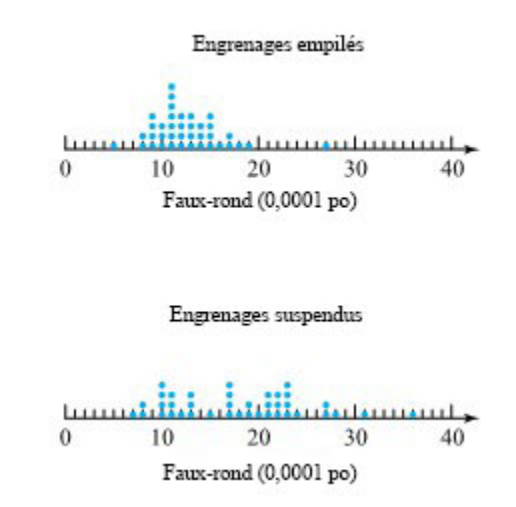


Figure 2.1.2.1. Diagramme de points des faux-ronds.

Exemple 2.1.2.2. Pénétration de balle à 200 grains

Sale et Thom ont comparé la profondeur de pénétration de plusieurs types de balles de calibre 0,45 tirées dans du bois de chêne à une distance de 15 pieds. Le tableau 2.1.2.1 répertorie les profondeurs de pénétration (exprimées en mm de la surface de la cible jusqu'à l'arrière des balles) pour deux types de balles. La figure 2.1.2.2 présente une paire de diagrammes à points correspondants.

Profondeur de pénétration des balles (mm)	
Balles chemisées, 230 grains	Balles chemisées, 200 grains
40,50, 38,35, 56,00, 42,55, 38,35, 27,75, 49,85, 43,60, 38,75, 51,25, 47,90, 48,15, 42,90, 43,85, 37,35, 47,30, 41,15, 51,60, 39,75, 41,00	63,80, 64,65, 59,50, 60,70, 61,30, 61,50, 59,80, 59,10, 62,95, 63,55, 58,65, 71,70, 63,30, 62,65, 67,75, 62,30, 70,40, 64,05, 65,00, 58,00

Tableau 2.1.2.1. Profondeur de pénétration des balles

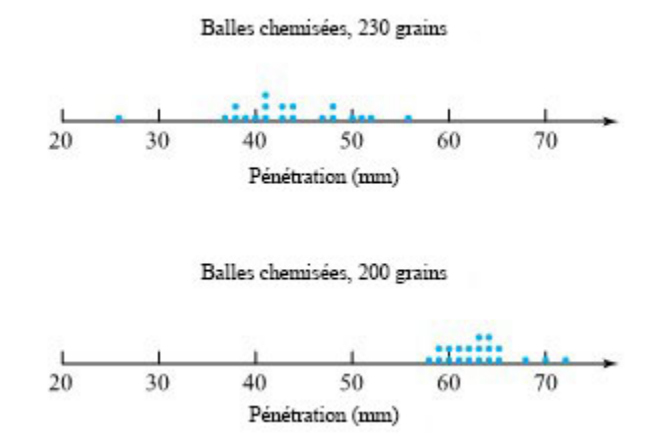


Figure 2.1.2.2. Diagrammes à points des profondeurs de pénétration

Les diagrammes à points montrent que la pénétration des balles de 200 grains est à la fois plus importante et plus uniforme que celle des balles de 230 grains. (Les étudiant.e.s avaient prédit des pénétrations plus importantes pour les balles plus légères sur la base d'une plus grande vitesse initiale et d'une plus petite surface d'action du frottement. Les différences d'uniformité de la pénétration n'étaient ni prévues ni expliquées.)

Les diagrammes à points donnent une idée générale d'un ensemble de données, mais ils ne permettent pas toujours de récupérer les valeurs utilisées pour les créer. Un diagramme à tiges et à feuilles contient à peu près les mêmes informations visuelles qu'un diagramme à points, tout en préservant exactement les valeurs d'origine. Ce type de diagramme se construit en utilisant les derniers chiffres de chaque point de données pour indiquer où il se situe.

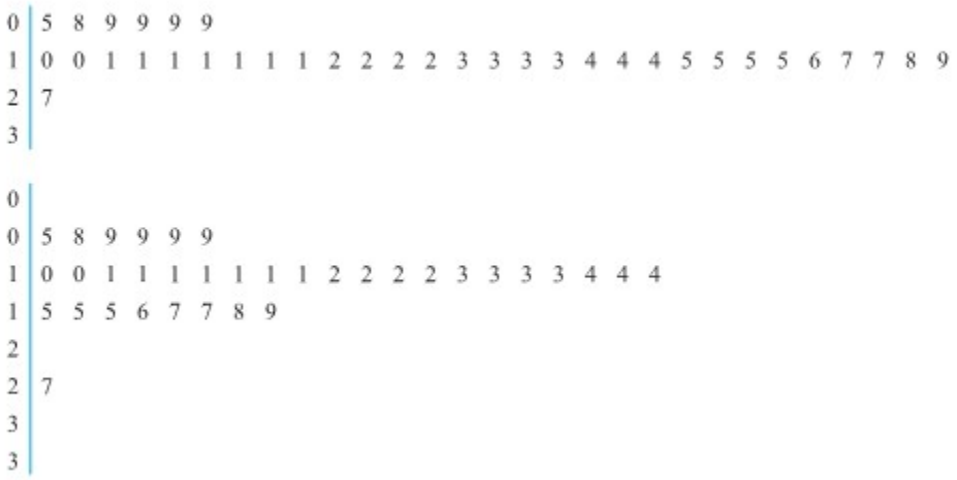


Figure 2.1.2.3. Diagrammes à tiges et à feuilles du faux-ronds des engrenages empilés

Exemple 2.1.2.1 Représentation du faux-rond des engrenages (suite)

La figure 2.1.2.3 présente deux diagrammes à tiges et à feuilles possibles pour le faux-ronds de la face de poussée des engrenages empilés. Dans les deux cas, le premier chiffre de chaque observation est représenté par le nombre situé à gauche de la ligne verticale (la « tige ») du diagramme. Les nombres situés à droite de la ligne verticale constituent les « feuilles » et donnent le deuxième chiffre des faux-ronds observés. Le deuxième diagramme est un peu plus détaillé que le premier : il indique les positions des feuilles « $0-4$ » et « $5-9$ » pour chaque premier chiffre possible, au lieu d'une seule feuille « $0-9$ » pour chaque premier chiffre.

Exemple 2.1.2.2 Pénétration des balles de 200 grains, suite

La figure 2.1.2.4 présente deux diagrammes à tiges et à feuilles possibles pour la pénétration de balles de 200 grains (données du tableau 2.1.2.1). Sur ces diagrammes, il était pratique d'utiliser les deux chiffres à gauche de la virgule pour la tige et les deux chiffres à droite de la virgule pour les feuilles. Le premier diagramme a été réalisé en consignant les valeurs des feuilles directement à partir du tableau (de gauche à droite et de haut en bas). Le deuxième diagramme est meilleur : on l'a obtenu en ordonnant les valeurs qui composent chaque feuille. À noter que les deux diagrammes donnent essentiellement la même impression visuelle que le deuxième diagramme à points de la figure 2.2.1.2.

58	.65, .00	58	.00, .65
59	.50, .80, .10	59	.10, .50, .80
60	.70	60	.70
61	.30, .50	61	.30, .50
62	.95, .65, .30	62	.30, .65, .95
63	.80, .55, .30	63	.30, .55, .80
64	.65, .05	64	.05, .65
65	.00	65	.00
66		66	
67	.75	67	.75
68		68	
69		69	
70	.40	70	.40
71	.70	71	.70

Figure 2.1.2.4. Diagramme à tiges et à feuilles de la profondeur de pénétration de (200 grains)

Pour comparer deux jeux de données, on peut accoler deux diagrammes à tiges et à feuilles.

Exemple 2.1.2.1. Diagrammes accolés des données de faux-rond (suite)

La figure 2.1.2.5 présente les diagrammes à tiges et à feuilles dos à dos des données du tableau 2.1.2.1. Cette disposition montre clairement les différences d'emplacement et de répartition des deux ensembles de données.

Faux-ronds des engrenages empilés		Faux-ronds des engrenages suspendus
9 9 9 9 8 8 5	0	7 8 8
4 4 4 3 3 3 3 2 2 2 2 1 1 1 1 1 1 0 0 0	1	0 0 0 0 1 1 1 2 3 3 3
9 8 7 7 6 5 5 5 5	1	5 7 7 7 7 8 9 9
	2	0 1 1 1 2 2 2 3 3 3 3 4
7	2	7 7 8
	3	1
	3	6

Figure 2.1.2.5. Diagrammes à tiges et à feuilles des faux-ronds

2.1.3 Tableaux de fréquences et histogrammes

Les diagrammes de dispersion et les diagrammes tige et feuille sont des outils utiles lorsque l'on étudie un ensemble de données, mais ils ne sont pas utilisés fréquemment dans les présentations et les rapports. Ce sont les tableaux de fréquences et les histogrammes qui sont le plus souvent utilisés dans ces contextes plus formels. Pour établir un tableau de fréquences, il faut d'abord regrouper les données en un nombre approprié d'intervalles de même longueur. Ensuite, on enregistre le nombre de points tombant dans chaque intervalle. Enfin, il est possible d'y ajouter les fréquences, les fréquences relatives et les fréquences relatives cumulées.

Exemple 2.1.3.1. Faux-rond des engrenages empilés

Le tableau 2.1.3.1 constitue une table des fréquences pour le faux-rond des engrenages empilés. Les valeurs de fréquence relative sont obtenues en divisant les entrées dans la colonne des fréquences par 38, le nombre total de points de données. Les entrées de la colonne de fréquence relative cumulées se calculent en divisant le total d'une classe donnée et de toutes les classes précédentes par le nombre total de points de données. (Sauf arrondissement, il s'agit de la somme des fréquences relatives sur la même ligne et au-dessus d'une fréquence relative cumulative). La colonne de points indique le même type d'informations sur la forme de la distribution que ce qu'on obtient d'un diagramme de dispersion ou d'un diagramme tige et feuilles.

Table de fréquences pour les faux-ronds des engrenages empilés

Faux-rond (0,0001 po)	Comptage	Fréquence	Fréquence relative	Fréquence relative cumulée
5–8	|||	3	0,079	0,079
9–12	𝍸 𝍸 𝍸 |||	18	0,474	0,553
13–16	𝍸 𝍸 ||	12	0,316	0,868
17–20	||||	4	0,105	0,974
21–24		0	0,0	0,974
25–28	|	1	0,026	1,000
		38	1,000	

Tableau 2.1.3.1. Tableau de fréquences pour le faux-rond de la face de poussée des engrenages empilés.

Choix d'un intervalle d'un tableau de fréquences

Le choix des intervalles à utiliser dans un tableau de fréquences est une subjectif (deux personnes ne choisiront pas forcément les mêmes intervalles). Cependant, il faut prendre en compte quelques points simples. Tout d'abord, pour éviter que la colonne des points ne fausse l'impression de la forme de la distribution, il convient d'utiliser des intervalles de même longueur. En outre, pour des raisons esthétiques, il est préférable de choisir des nombres ronds pour les extrémités des intervalles. Étant donné que la réduction des données brutes en tables implique généralement une agrégation (et donc une certaine perte d'informations), plus le nombre d'intervalles utilisés est élevé, plus les informations présentées dans le tableau sont détaillées. En contrepartie, pour qu'un tableau de fréquences synthétise réellement les données, il ne doit pas être encombré d'un trop grand nombre d'intervalles.

Après avoir monté un tableau de fréquences, on utilise généralement l'organisation fournie qu'il fournit pour

générer un histogramme. Un histogramme (de fréquence ou de fréquence relative) est un type de diagramme à barres utilisé pour représenter la forme d'une distribution de points de données.

Exemple 2.1.2.2. Pénétration de balles de 200 grains (suite)

Le tableau 2.1.3.2 est un tableau de fréquences pour les profondeurs de pénétration des balles de 200 grains, et la figure 2.1.3.1 présente cette table sous forme d'histogramme.

Table de fréquences : profondeurs de pénétration (200 grains)

Profondeur de pénétration (mm)	Comptage	Fréquence	Fréquence relative	Fréquence relative cumulée
58,00–59,99	𝍸	5	0,25	0,25
60,00–61,99	III	3	0,15	0,40
62,00–63,99	𝍸 I	6	0,30	0,70
64,00–65,99	III	3	0,15	0,85
66,00–67,99	I	1	0,05	0,90
68,00–69,99		0	0	0,90
70,00–71,99	II	2	0,10	1,00
		20	1,00	

Tableau 2.1.3.2. Tableau de fréquences pour les profondeurs de pénétration de balles de 200 grains.

.

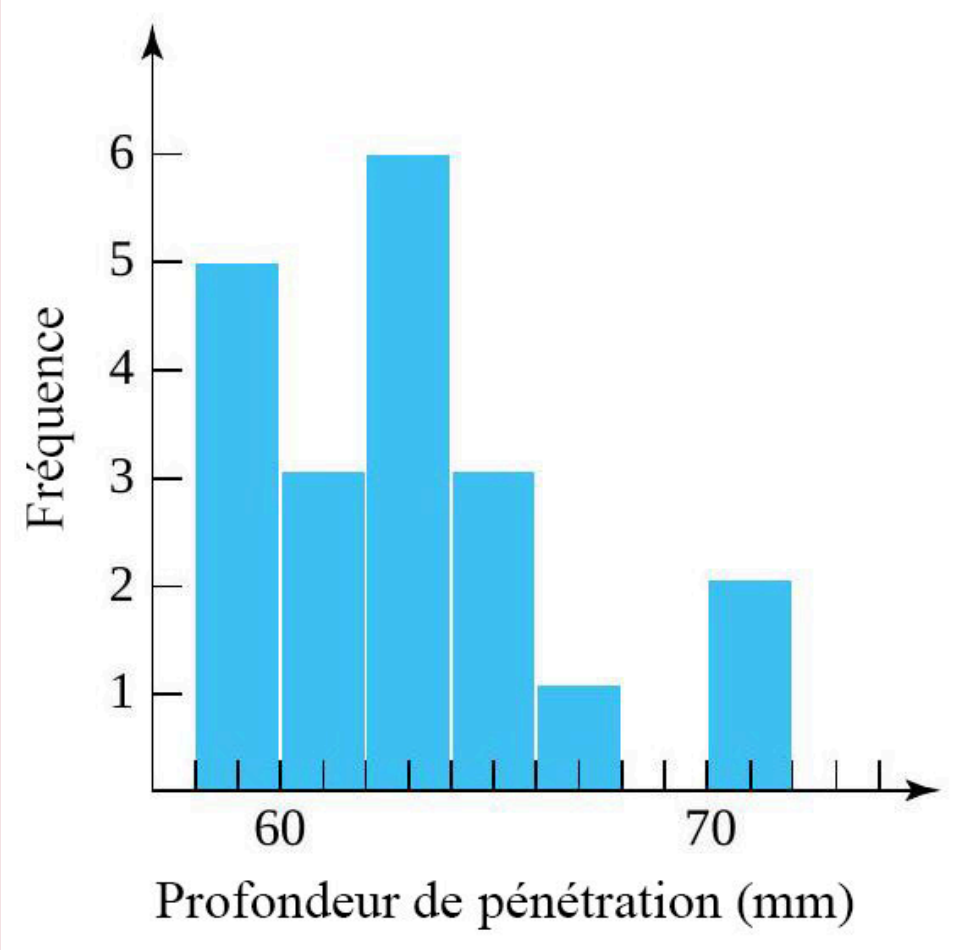


Figure 2.1.3.1. Histogramme de profondeur de pénétration des balles de 200 grains.

L'échelle verticale de la figure 2.1.3.1 est une échelle de fréquence, et l'histogramme est un histogramme de fréquence. En passant à la fréquence relative sur l'échelle verticale, on peut produire un histogramme de fréquence relative.

LIGNES DIRECTRICES POUR L'ÉLABORATION D'HISTOGRAMMES

Lors de l'élaboration de la figure 2.1.3.1, on a veillé à:

1. (continuer à) utiliser des intervalles de longueur égale;
2. afficher l'ensemble de l'axe vertical à partir de zéro;
3. éviter de franchir un axe;
4. maintenir l'échelle uniforme sur un axe donné;
5. centrer les barres de hauteur appropriée aux points médians des intervalles (de profondeur de pénétration).

Ces lignes directrices produisent un graphique dans lequel des zones fermées égales correspondent à des nombres égaux de points de données. En outre, la position des points de données est clairement indiquée par la position de la barre sur l'axe horizontal. Si on ne respecte pas ces lignes directrices, le diagramme à barres ne

représentera pas un bon portrait des données qui le composent. La figure 2.1.3.2 illustre les termes utilisés pour décrire quelques formes de distribution courantes lorsqu'on crée ou qu'on utilise des diagrammes de dispersion, des diagrammes tige et feuilles et des histogrammes.

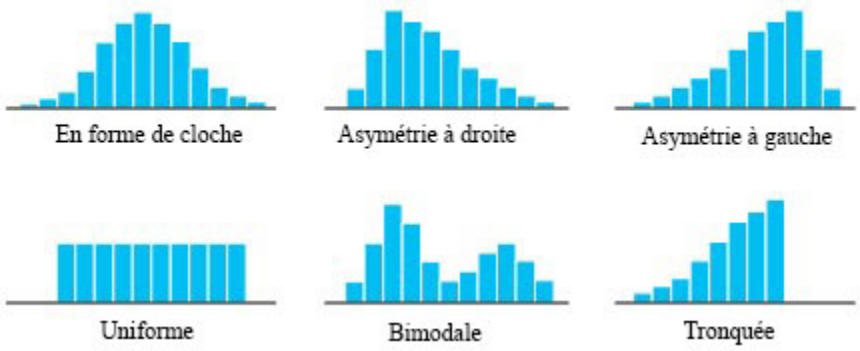


Figure 2.1.3.2. Formes de distribution.

Les méthodes graphiques et tabulaires vues jusqu'à présent sont d'une simplicité trompeuse. Lorsqu'utilisées régulièrement et intelligemment, elles constituent de puissants outils d'ingénierie. Les informations relatives à la localisation, à la répartition et à la forme qui sont représentées si clairement sur un histogramme peuvent donner des indications importantes sur le fonctionnement du processus physique qui génère les données. Elles peuvent également suggérer les mécanismes physiques sous-jacents du processus.

Exemples d'interprétations d'ingénierie de la forme de la distribution

Si, par exemple, les données relatives aux diamètres des vérins métalliques usinés achetés à un fournisseur produisent un histogramme résolument bimodal (ou multimodal, avec plusieurs bosses nettes), cela suggère que l'usinage a été effectué sur plusieurs machines, ou par plusieurs opérateurs, ou à différents moments. La conséquence pratique d'un tel usinage varié est une distribution des diamètres qui présente plus de variations que celles typiques d'une production provenant d'une seule machine, d'un seul opérateur et d'une seule configuration. Par ailleurs, si l'histogramme est tronqué, cela peut suggérer que le lot de vérins a été entièrement inspecté et trié afin d'éliminer tous les vérins présentant des diamètres excessifs. Qui plus est, en indiquant les spécifications (exigences) pour le diamètre du vérin sur l'histogramme, on peut obtenir une image comme celle de la figure 2.1.3.3. Il devient alors évident que le tour qui usine les vérins doit être ajusté afin d'augmenter le diamètre typique, mais il est aussi manifeste que la variation du processus de base est si importante que cet ajustement ne parviendra pas à rendre tous les diamètres conformes aux spécifications. Avec cette constatation et sa connaissance des conséquences économiques de la non-conformité des pièces aux spécifications, l'ingénieur.e peut évaluer intelligemment d'autres possibilités d'action : trier toutes les pièces entrantes, exiger du fournisseur qu'il utilise un équipement plus précis, chercher un nouveau fournisseur, etc.

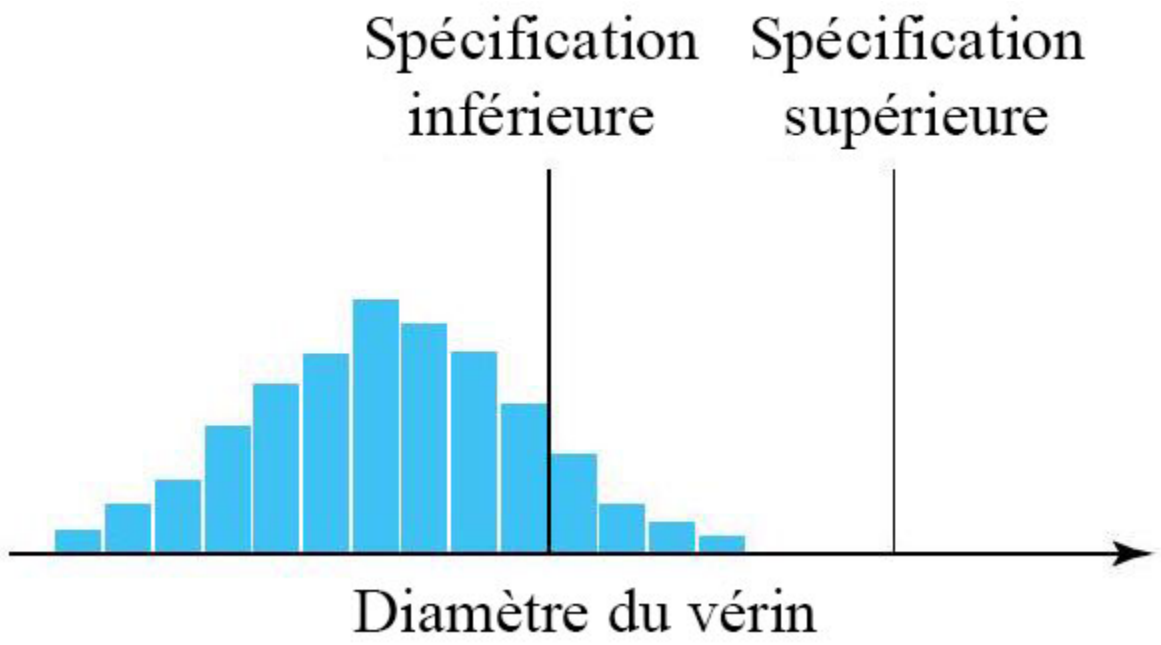


Figure 2.1.3.3. Histogramme indiquant les spécifications.

L'étude de la forme d'un ensemble de données est utile non seulement parce qu'elle peut donner un aperçu des mécanismes physiques, mais aussi parce que la forme peut être importante pour déterminer l'adéquation des méthodes d'inférence statistique formelle, comme celles que nous verrons plus tard. Une méthodologie appropriée pour une forme de distribution peut ne pas l'être pour une autre.

2.1.4 Diagrammes de dispersion et cartes de contrôle

Les diagrammes en points, les diagrammes à tiges et à feuilles, les tableaux de fréquences et les histogrammes sont des outils à une seule variable. Mais les données d'ingénierie comprennent souvent plusieurs variables, et dans ce cas, on s'intéresse généralement aux relations entre ces variables. Il est courant de produire un nuage de points bidimensionnel de paires de données, ce qui est un moyen simple et efficace d'illustrer les relations potentielles entre deux variables.

Exemple 2.1.4.1. Couple de serrage des boulons sur une plaque avant

Brenny, Christensen et Schneider ont mesuré le couple nécessaire pour desserrer six boulons distincts retenant la plaque avant d'un type de composant d'équipement lourd. Le tableau 2.1.4.1 répertorie les couples (en pi lb) requis pour les boulons numéros 3 et 4, respectivement, sur 34 composants différents. La figure 2.1.4.1 illustre un diagramme de dispersion des données à deux variables du tableau 2.1.4.1. Dans cette figure, s'il y avait plus d'un point au même endroit, on a indiqué le nombre de points à cet endroit.

Couple requis pour desserrer deux boulons (pi lb)

Composant	Couple du boulon 3	Couple du boulon 4	Composant	Couple du boulon 3	Couple du boulon 4
1	16	16	18	15	14
2	15	16	19	17	17
3	15	17	20	14	16
4	15	16	21	17	18
5	20	20	22	19	16
6	19	16	23	19	18
7	19	20	24	19	20
8	17	19	25	15	15
9	15	15	26	12	15
10	11	15	27	18	20
11	17	19	28	13	18
12	18	17	29	14	18
13	18	14	30	18	18
14	15	15	31	18	14
15	18	17	32	15	13
16	15	17	33	16	17
17	18	20	34	16	16

Tableau 2.1.4.1. Couple requis pour desserrer deux boulons de plaque avant (pi lb)

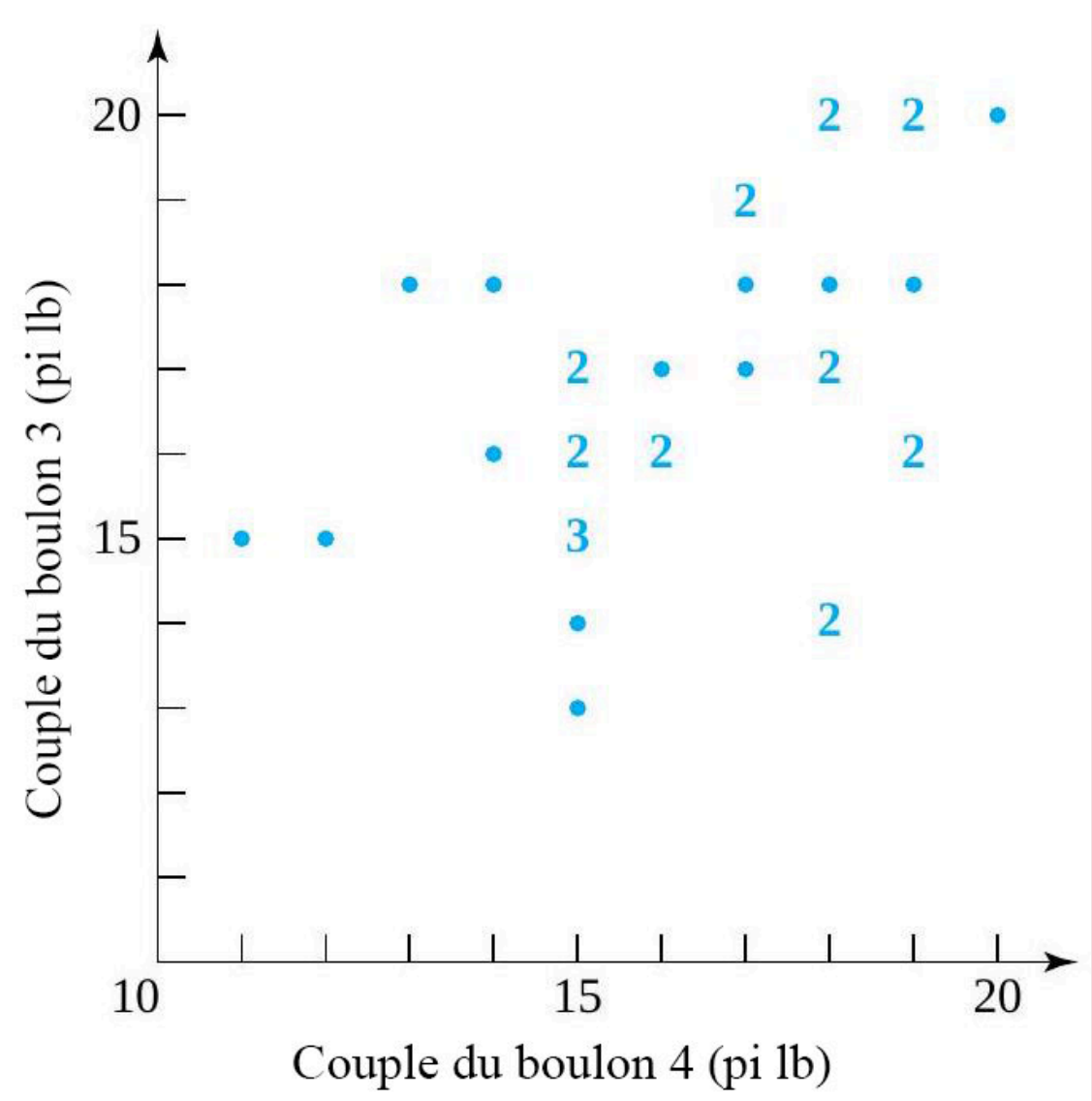


Figure 2.1.4.1. Diagramme de dispersion du couple des boulons 3 et 4.

Le graphique suggère au moins faiblement qu'un couple élevé au boulon 3 s'accompagne d'un couple élevé au boulon 4. Dans la pratique, c'est une bonne chose, car sinon, la plaque pourrait subir des forces différentielles indésirables. Il est également tout à fait raisonnable que les couples des boulons 3 et 4 soient liés, puisque les boulons ont été serrés par les différentes têtes d'une même clé pneumatique fonctionnant à partir d'une seule source d'air comprimé. Il est logique que les variations de la pression atmosphérique affectent de la même manière le serrage des boulons dans les deux positions, produisant ainsi le schéma « valeurs élevées ensemble » et « valeurs faibles ensemble » de la figure 2.1.4.1.

L'exemple précédent illustre le fait que les relations observées sur les diagrammes de dispersion suggèrent une cause physique commune pour le comportement des variables et peuvent aider à révéler cette cause.

CARTE DE CONTRÔLE

Généralement, sur un diagramme de dispersion, la variable sur l'axe des abscisses est le temps. Un diagramme de dispersion dans lequel des données à une seule variable sont représentées en fonction de l'ordre chronologique d'observation est appelé carte de contrôle ou diagramme de tendance. Les cartes de contrôle sont l'un des outils statistiques les plus utiles. L'observation de tendances sur une carte de contrôle amène à réfléchir aux variables du processus qui ont changé en même temps que la tendance, ce qui peut aider à mieux comprendre comment le comportement du processus est affecté par les variables qui changent au fil du temps.

.

Exemple 2.1.4.2. Diamètre des pièces consécutives usinées sur un tour

Williams et Markowski ont étudié un processus d'ébauche de tournage du diamètre extérieur de la bague extérieure d'un joint homocinétique. Le tableau 2.1.4.2 répertorie les diamètres (en pouces au-dessus du diamètre nominal) pour 30 joints consécutifs usinés sur un même tour automatique. La figure 2.1.4.2 illustre à la fois un diagramme en points et une carte de contrôle pour les données dans le tableau. Conformément aux pratiques standard, les points consécutifs de la carte de contrôle ont été reliés par des segments de ligne.

.

Joint	Diamètre (pouces au-dessus de la valeur nominale)	Joint	Diamètre (pouces au-dessus de la valeur nominale)
1	−0,005	16	0,015
2	0,000	17	0,000
3	−0,010	18	0,000
4	−0,030	19	−0,015
5	−0,010	20	−0,015
6	−0,025	21	−0,005
7	−0,030	22	−0,015
8	−0,035	23	−0,015
9	−0,025	24	−0,010
10	−0,025	25	−0,015
11	−0,025	26	−0,035
12	−0,035	27	−0,025
13	−0,040	28	−0,020
14	−0,035	29	−0,025
15	−0,035	30	−0,015

Tableau 2.1.4.2. 30 diamètres extérieurs consécutifs usinés sur un tour

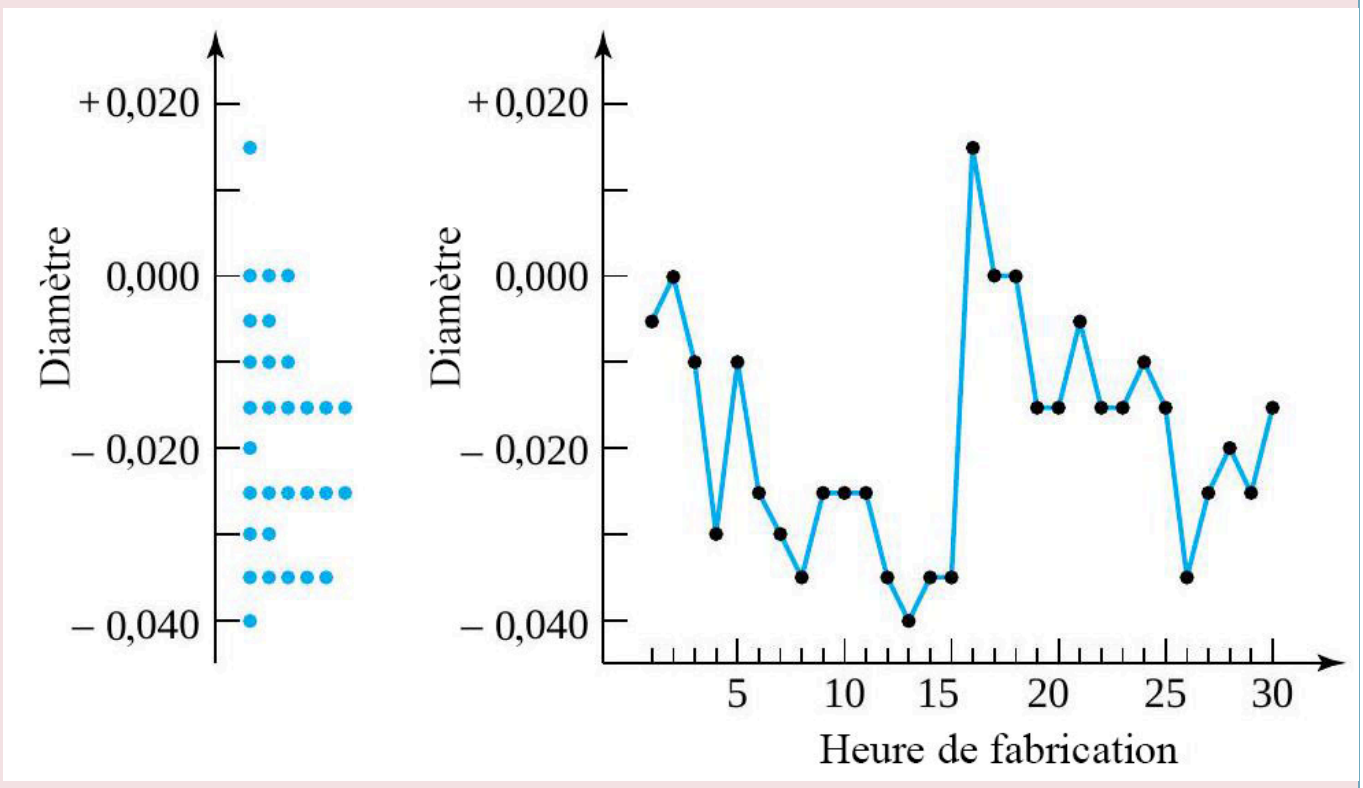


Figure 2.1.4.2. Diagramme de dispersion et carte de contrôle des diamètres extérieurs consécutifs.

Ici, le diagramme de dispersion ne donne pas vraiment d'indice quant aux mécanismes physiques qui ont généré les données, mais les informations sur le temps ajoutées dans la carte de contrôle sont révélatrices. Au fil du temps, les diamètres extérieurs tendent à diminuer jusqu'à la pièce 16, où l'on observe un saut important, suivi à nouveau d'une tendance à la diminution générale du diamètre au fil du temps. En fait, après avoir vérifié les registres de production, Williams et Markowski ont constaté qu'on avait arrêté le tour pour le laisser refroidir entre les pièces 15 et 16. Le saut visible sur la carte de contrôle est probablement lié au comportement du système hydraulique du tour. À froid, le système hydraulique ne pousse probablement pas aussi bien l'outil de coupe dans la pièce à tourner. Par conséquent, les pièces tournées deviennent plus petites au fur et à mesure que le tour se réchauffe. Afin d'obtenir des pièces plus proches de la valeur nominale, on peut augmenter le diamètre visé d'environ 0,020 po et n'usiner les pièces qu'après avoir laissé le tour chauffer.

2.1.5 Quantiles et diagrammes quantile

La plupart des lecteurs connaissent le concept de percentile (ou rang centile), une notion surtout vue dans le contexte des résultats des examens scolaires. Par exemple, si une personne a obtenu une note la plaçant au 80^{e} rang centile, environ 80% des personnes qui ont passé l'examen ont obtenu une moins bonne note, et 20% ont obtenu une meilleure note. Ce concept est également utile pour décrire des données d'ingénierie. Toutefois, comme il est souvent plus pratique de travailler en termes de fractions entre 0 et 1 plutôt qu'en termes de pourcentages entre 0 et 100, on utilisera une terminologie légèrement différente : on parlera de « quantiles » plutôt que de rang centiles. Après avoir soigneusement défini les quantiles d'un ensemble de données, on les utilise pour créer divers outils utiles de statistiques descriptives : diagrammes quantile, diagrammes en boîte, diagrammes $Q-Q$, et diagrammes normaux (un type de diagramme $Q-Q$ théorique).

En gros, pour un nombre p compris entre 0 et 1 , le quantile p d'une distribution est un nombre tel qu'une fraction p de la distribution se trouve à gauche, et une fraction $1-p$, à droite. Toutefois, en raison du caractère discret des ensembles finis de données, il est nécessaire d'indiquer exactement ce que l'on veut dire par là. La définition 1 donne la convention qui sera utilisée dans ce texte.

Définition 3.1.5.1 Quantile p

Pour un ensemble de données composé de n valeurs ordonnées $x_1 \leq x_2 \leq \cdots \leq x_n$,

1. Si $p = \dfrac{i-0{,}5}{n}$ pour un entier positif $i \leq n$, le quantile p de l'ensemble de données est

$$Q(p) = Q\left(\frac{i-.5}{n}\right) = x_i$$

(Le i^{e} point le plus petit des données est appelé quantile $\dfrac{i-0{,}5}{n}$.)

2. Pour tout nombre p compris entre $\dfrac{0{,}5}{n}$ et $\dfrac{n-.5}{n}$ qui n'est pas de la forme $\dfrac{i-0{,}5}{n}$ avec i entier, le quantile p de l'ensemble de données s'obtient par interpolation linéaire entre les deux valeurs de $Q\left(\dfrac{i-0{,}5}{n}\right)$ avec les valeurs $\dfrac{i-0{,}5}{n}$ correspondantes qui entourent p.

Dans les deux cas, le quantile p est dénoté $Q(p)$.

La définition 2.1.5.1 donne $Q(p)$ pour tous les p compris entre $0{,}5/n$ et $(n-0{,}5)/n$. Pour trouver $Q(p)$ pour une telle valeur de p, on isole i dans $p = (i-0{,}5)/n$, ce qui donne

Index (i) du point de données ordonnées au quantile $Q(p)$

$$i = np + 0{,}5$$

et on trouve le « $(np+.5)^{e}$ point de données ordonnées ».

Exemple 2.1.5.1. Quantiles de force de rupture à sec de serviettes en papier

Lee, Sebghati et Straub ont mené une étude sur la force de rupture de plusieurs marques de serviettes en papier. Le

tableau 3.1.5.1 répertorie dix force de rupture (en grammes) rapportées par les étudiant.e.s pour une serviette standard. En ordonnant les données de force et en calculant les valeurs de $\frac{i-0{,}5}{10}$, il est facile de trouver les quantiles d'ordre $0{,}05,\ 0{,}15,\ 0{,}25, \ldots,\ 0{,}85$ et 0,95 de la répartition de la force de rupture, comme illustré au tableau 2.1.5.2.

Résistances à la rupture de dix serviettes en papier

Essai	Résistance à la rupture (g)
1	8 577
2	9 471
3	9 011
4	7 583
5	8 572
6	10 688
7	9 614
8	9 614
9	8 527
10	9 165

Tableau 2.1.5.1.

.

Quantiles de la distribution de la résistance à la rupture d'une serviette en papier

i	$\frac{i-0{,}5}{10}$	i^e point de données le plus petit, $x_i = Q\left(\frac{i-0{,}5}{10}\right)$
1	0,05	$7\ 583 = Q(0,05)$
2	0,15	$8\ 527 = Q(0,15)$
3	0,25	$8\ 572 = Q(0,25)$
4	0,35	$8\ 577 = Q(0,35)$
5	0,45	$9\ 011 = Q(0,45)$
6	0,55	$9\ 165 = Q(0,55)$
7	0,65	$9\ 471 = Q(0,65)$
8	0,75	$9\ 614 = Q(0,75)$
9	0,85	$9\ 614 = Q(0,85)$
10	0,95	$10\ 688 = Q(0,95)$

Tableau 2.1.5.2.

Étant donné qu'il y a $n = 10$ points de données, chacun d'eux compte pour 10% de l'ensemble de données. Appliquons la convention (1) de la définition 3.1.5.1 pour trouver le quantile d'ordre 0,35 (par exemple). Les trois points de données les plus petits et la moitié du quatrième plus petit sont considérés comme se trouvant à gauche du nombre souhaité, et les six points de données les plus grands et la moitié du septième plus grand sont considérés comme se trouvant à droite. Ainsi, le quatrième point de données le plus petit doit être le quantile d'ordre 0,35, comme le montre le tableau 2.1.5.2.

Pour illustrer la convention (2) de la définition 1, calculons les quantiles d'ordre 0,5 et 0,93 de la distribution de la force. Étant donné que 0,5 est à $\frac{.5-.45}{.55-.45} = .5$ unité à mi-chemin entre 0,45 et 0,55, l'interpolation linéaire donne :

$$Q(0{,}5) = (1 - 0{,}5)Q(0{,}45) + 0{,}5Q(0{,}55) = 0{,}5(9\ 011) + 0{,}5(9\ 165) = 9\ 088 \text{ g}$$

Puis, comme 0,93 est à $\frac{.93 - .85}{.95 - .85} = .8$ unité à mi-chemin entre 0,85 et 0,93, l'interpolation linéaire donne :

$$Q(0{,}93) = (1 - 0{,}8)Q(0{,}85) + 0{,}8Q(0{,}95) = 0{,}2(9\ 614) + 0{,}8(10\ 688) = 10\ 473{,}2 \text{ g}$$

Certaines valeurs rondes de p donne des quantités $Q(p)$ qui portent des noms spéciaux.

DÉFINITION 2.1.5.2 Médiane

Définition 2 $Q(0{,}5)$ est la médiane de la distribution.

DÉFINITION 2.1.5.3 Premier et troisième quartiles

Définition 3 $Q(0{,}25)$ et $Q(.75)$ sont respectivement le premier et le troisième quartiles d'une distribution.

Exemple 2.1.5.1 Quantiles de force de rupture à sec de serviettes en papier (suite)

Si l'on se réfère à nouveau au tableau 2.1.5.2 et à la valeur de $Q(\ 0{,}5\ )$ précédemment calculée, pour la distribution de la force de rupture, on a :

$$\begin{aligned} \text{Médiane} &= Q(\ 0{,}5) = 9\ 088 \text{ g} \\ 1\text{er quartile} &= Q(\ 0{,}25) = 8\ 572 \text{ g} \\ 3\text{ed quartile} &= Q(\ 0{,}75) = 9\ 614 \text{ g} \end{aligned}$$

On peut représenter les quantiles à l'aide d'un diagramme.

DÉFINITION 2.1.5.4 Diagramme quantile

Un diagramme quantile est un graphique de $Q(p)$ en fonction de p. Pour un ensemble de données

ordonnées de taille n contenant les valeurs $x_1 \leq x_2 \leq \cdots \leq x_n$, on obtient ce graphique en traçant les points $\left(\frac{i-.5}{n}, x_i\right)$ puis en reliant les points consécutifs par des segments de droite.

C'est la convention (2) de la définition 2.1.5.1, qui demande une interpolation linéaire, qui fait qu'on ajoute des segments de droite au diagramme de quantiles.

Exemple 2.1.5.1. Quantiles de force de rupture à sec de serviettes en papier (suite)

Si l'on se réfère à nouveau au tableau 2.1.5.2 pour les quantiles de la distribution de la force de rupture, il est clair qu'un diagramme quantile pour ces données impliquera de tracer puis de relier les paires ordonnées consécutives suivantes.

(.05, 7,583)	(.15, 8,527)	(.25, 8,572)
(.35, 8,577)	(.45, 9,011)	(.55, 9,165)
(.65, 9,471)	(.75, 9,614)	(.85, 9,614)
(.95, 10,688)		

Ce graphique se trouve à la figure 2.1.5.1.

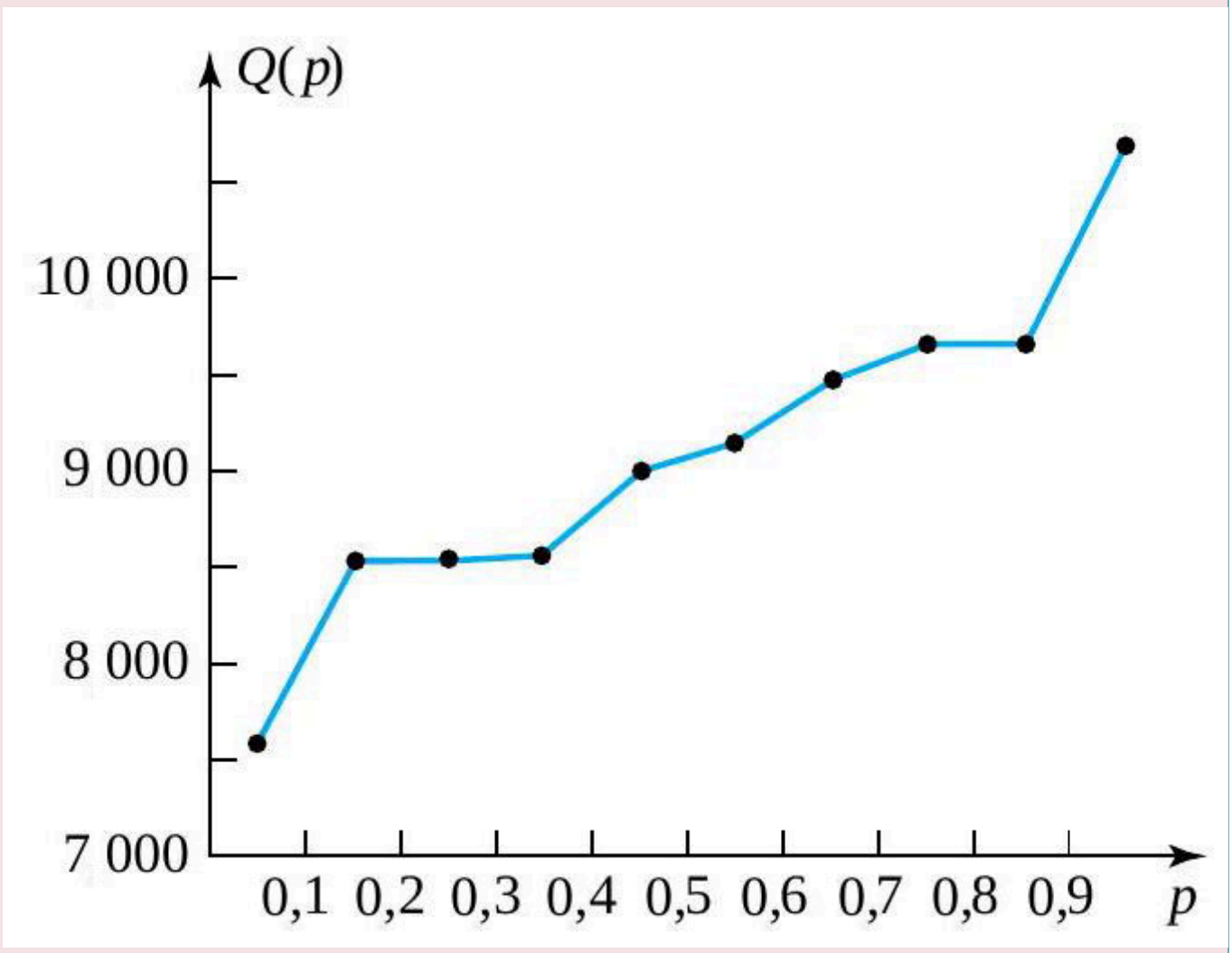


Figure 2.1.5.1. Diagramme quantile des force de rupture de serviettes en papier.

Un diagramme quantile permet d'effectuer de lisser quelque peu les données irrégulières. (On suppose tacitement que pour le mécanisme de génération des données à l'étude, si on augmentait la taille de l'échantillon, on obtiendrait un diagramme quantile plus lisse.)

2.1.6 Diagrammes en boîtes

La principale condition préalable pour élaborer des diagrammes en boîtes, un type de graphique qui s'ajoute aux diagrammes de dispersion et aux histogrammes, est de maîtriser la notion de quantiles. Le diagramme en boîtes contient un peu moins d'informations, mais il présente l'avantage qu'on peut en placer plusieurs sur une même page pour les comparer.
Il existe plusieurs conventions courantes pour l'élaboration de diagrammes en boîtes. Celle utilisée ici est illustrée de manière générique à la figure 2.1.6.1. On dessine une boîte qui se rend du premier au troisième quartile, et on ajout une ligne à la médiane. Ensuite, on calcule l'écart interquartile :

DÉFINITION 2.1.6.1. Écart interquartile : EI

$$EIQ = Q(\ 0{,}75) - Q(\ 0{,}25)$$

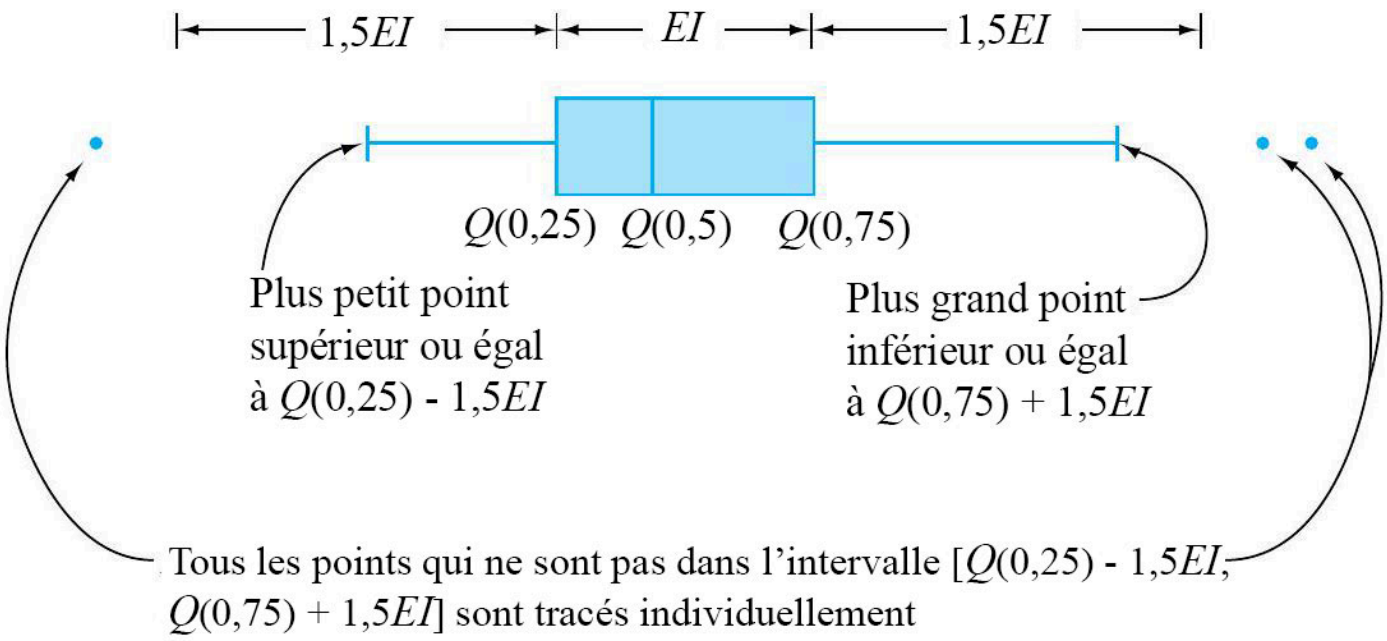


Figure 2.1.6.1. Diagramme en boîtes générique.

puis, on détermine le plus petit point des données situé dans un intervalle de 1,5 EI de $Q(.25)$ et le plus grand point des données située dans un intervalle de 1,5 EI de $Q(.75)$. On trace des lignes (les « moustaches ») allant de la boîte à ces valeurs. En général, la plupart des points se situent dans l'intervalle $[Q(.25) - 1.5IQR, Q(.75) + 1.5IQR]$. Ceux qui ne le sont pas sont ajoutés individuellement, ce qui indique qu'il s'agit de valeurs aberrantes ou inhabituelles.

Exemple 2.1.6.2. Quantiles de force de rupture à sec de serviettes en papier (suite)

Créons un diagramme en boîte pour les données sur la force de rupture des serviettes en papier. Pour commencer,

$$Q(0{,}25) = 8\ 572 \text{ g}$$
$$Q(0{,}5) = 9\ 088 \text{ g}$$
$$Q(0{,}75) = 9\ 614 \text{ g}$$

Donc

$$EIQ = Q(0{,}75) - Q(0{,}25) = 9\ 614 - 8\ 572 = 1\ 042 \text{ g}$$

et

$$1{,}5\ EIQ = 1\ 563 \text{ g}$$

d'où

$$Q(.75) + 1.5IQR = 9,614 + 1,563 = 11,177 \text{ g}$$

et

$$Q(0{,}25) - 1{,}5IQR = 8\ 572 - 1\ 563 = 7\ 009 \text{ g}$$

Comme tous les points de données se trouvent dans la plage $7,009$ g to $11,177$ g, le diagramme en boîtes ressemble donc à la figure 2.1.6.2.

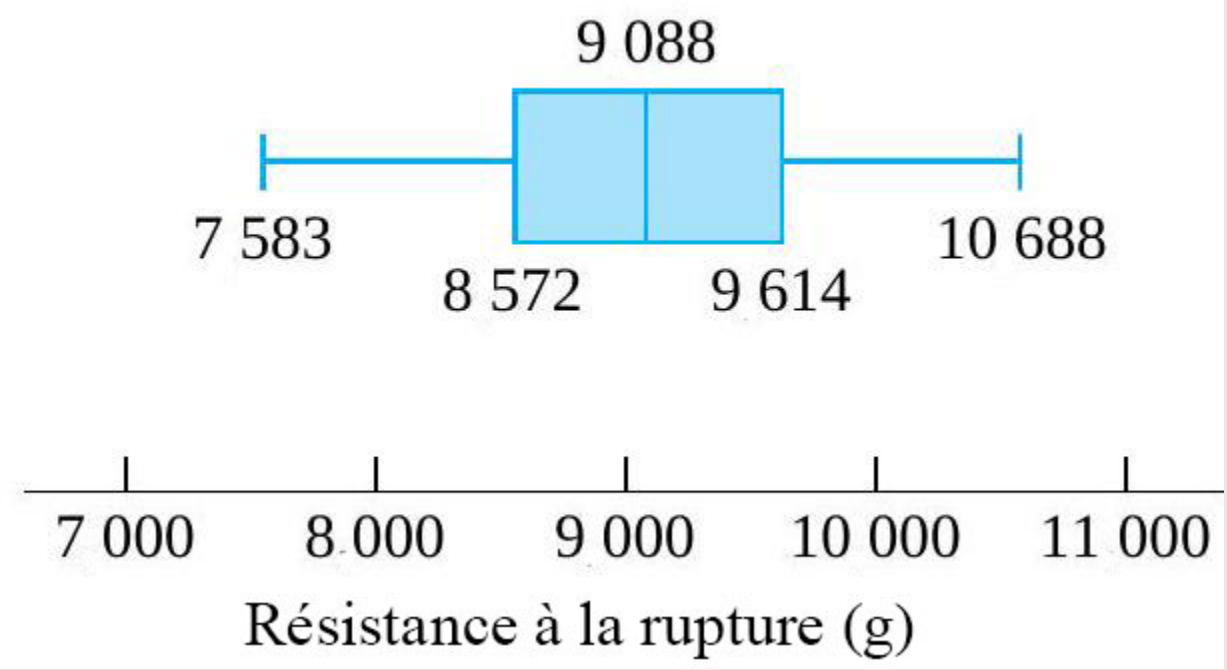


Figure 2.1.6.2. Diagramme en boîtes de la force de rupture de serviettes en papier.

Un diagramme en boîtes illustre la distribution par l'intermédiaire de la boîte, qui englobe le 50% du milieu de la distribution, et les moustaches. Certains éléments de la forme de la distribution sont indiqués par la symétrie (ou l'asymétrie) de la boîte et des moustaches. En outre, un espace entre l'extrémité d'une moustache et un point individuel rappelle qu'il n'y a aucune autre valeur dans cet intervalle.
Pour comparer plusieurs échantillons efficacement, on peut juxtaposer leurs diagrammes de boîtes.

Exemple 2.1.6.3 Profondeur de pénétration des balles (suite)

Le tableau 2.1.6.1 répertorie les informations brutes nécessaires pour trouver les quantiles $\frac{i-.5}{20}$ des deux

distributions de profondeur de pénétration des balles présentées à la section précédente. Pour les profondeurs de pénétration des balles de 230 grains, l'interpolation donne

$$Q(0,25) = 0,5Q(0,225) + 0,5Q(0,275) = 0,5(38,75) + 0,5(39,75) = 39,25 \text{ mm}$$
$$Q(0,5) = 0,5Q(0,475) + 0,5Q(0,525) = 0,5(42,55) + 0,5(42,90) = 42,725 \text{ mm}$$
$$Q(0,75) = 0,5Q(0,725) + 0,5Q(0,775) = 0,5(47,90) + 0,5(48,15) = 48,025 \text{ mm}$$

Donc

$$\begin{aligned} EIQ &= 48,025 - 39,25 = 8,775 \text{ mm} \\ 1,5\,EIQ &= 13,163 \text{ mm} \\ Q(0,75) + 1,5\,EIQ &= 61,188 \text{ mm} \\ Q(0,25) - 1,5\,EIQ &= 26,087 \text{ mm} \end{aligned}$$

Pour les profondeurs de pénétration des balles de 200 grains, l'interpolation donne

$$\begin{aligned} Q(0,25) &= 60,25 \text{ mm} \\ Q(0,5) &= 62,80 \text{ mm} \\ Q(0,75) &= 64,35 \text{ mm} \\ Q(0,75) + 1,5EIQ &= 70,50 \text{ mm} \\ Q(0,25) - 1,5EIQ &= 54,10 \text{ mm} \end{aligned}$$

Quantiles de la distribution de profondeur de pénétration des balles

i	$\frac{i-.5}{20}$	i^e point de données le plus petits (230 grains) $= Q(\frac{i-.5}{20})$	i^e point de données le plus petits (200 grains) $= Q(\frac{i-.5}{20})$
1	0,025	27,75	58,00
2	0,075	37,35	58,65
3	0,125	38,35	59,10
4	0,175	38,35	59,50
5	0,225	38,75	59,80
6	0,275	39,75	60,70
7	0,325	40,50	61,30
8	0,375	41,00	61,50
9	0,425	41,15	62,30
10	0,475	42,55	62,65
11	0,525	42,90	62,95
12	0,575	43,60	63,30
13	0,625	43,85	63,55
14	0,675	47,30	63,80
15	0,725	47,90	64,05
16	0,775	48,15	64,65
17	0,825	49,85	65,00
18	0,875	51,25	67,75
19	0,925	51,60	70,40
20	0,975	56,00	71,70

Tableau 2.1.6.1.

La figure 2.1.6.3 illustre les diagrammes en boîte placées côte à côte sur la même échelle. On constate que les balles de 200 grains ont une profondeur de pénétration plus importante et plus régulière, et on remarque qu'il y a un point

particulièrement extrême dans l'ensemble de données de 200 grains. En outre, les longueurs relatives des moustaches indiquent une certaine asymétrie (rappelons la terminologie introduite précédemment pour discuter de la forme de la distribution) dans les données. Et tout ça, dans un graphique très épuré et compact. Il serait possible d'ajouter beaucoup d'autres boîtes à la figure 2.1.6.3 (pour comparer d'autres types de balles) sans alourdir le graphique.

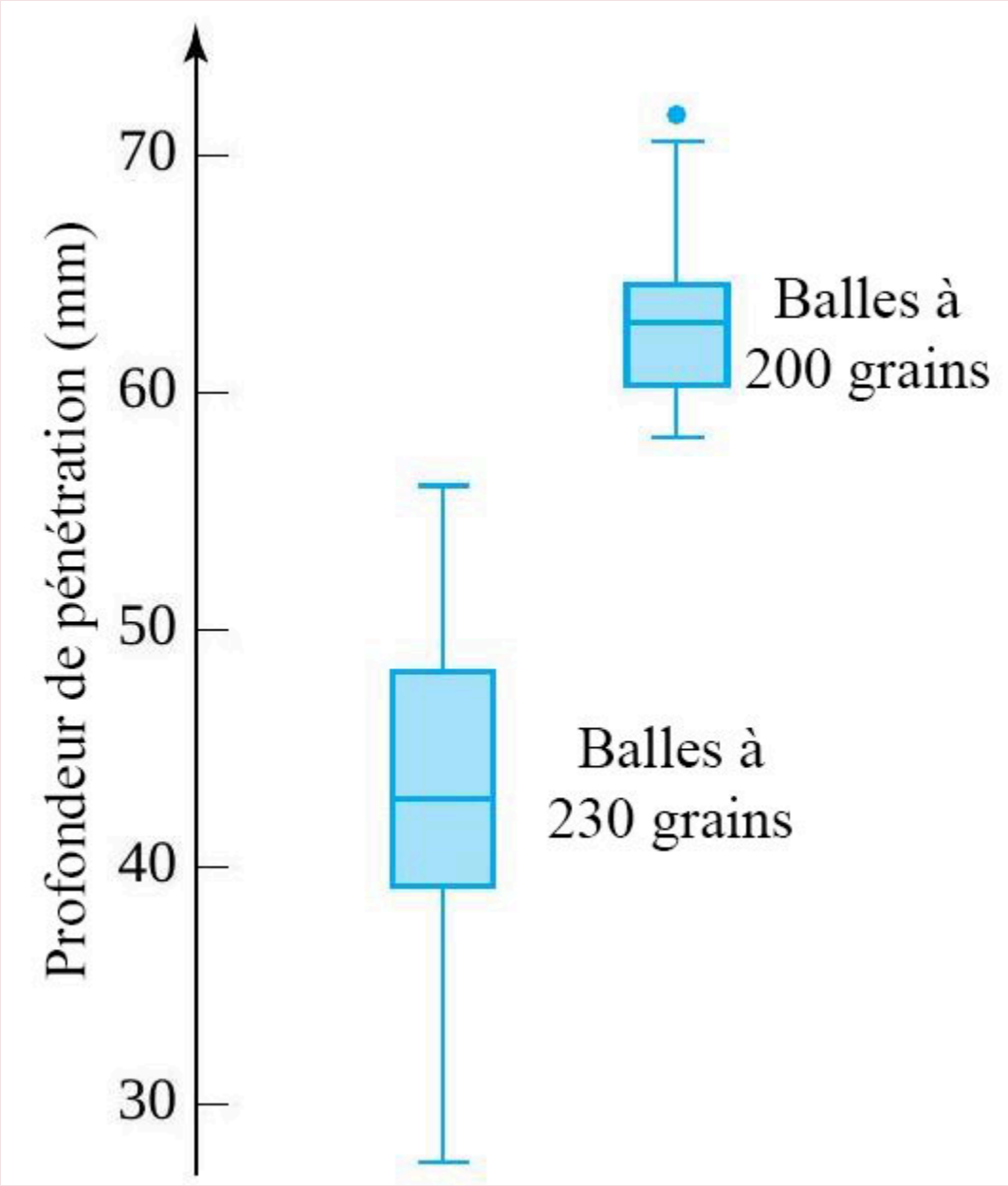


Figure 2.1.6.3. Diagrammes en boîte côte à côte pour la profondeur de pénétration des balles

2.1.7 Diagrammes Q-Q et comparaison des formes de distribution

Il est souvent important de comparer les formes de deux distributions. Pour ce faire, on peut y aller approximativement, avec des histogrammes, mais pour plus de précision, on peut représenter les fonctions quantile des deux distributions sur un même graphique, sachant que « forme égale » équivaut à « fonctions quantile linéairement proportionnelles ». Ce type de diagramme s'appelle **diagramme quantile-quantile** ou, pour faire court, diagramme $\mathbf{Q} - \mathbf{Q}$.
Considérons les deux petits ensembles de données artificielles présentés dans le tableau 2.1.7.1. Les diagrammes à points de ces deux ensembles de données sont présentés à la figure 2.1.71. Les deux ensembles de données ont la même forme. Pourquoi? Pour considérer l'égalité des formes, on peut noter que :

2.1.7.1
la i^{e} plus petite valeur de l'ensemble de données $2 = 2(i^{\text{e}}$ plus petite valeur de l'ensemble de données $1) + 1$

Ensuite, en reconnaissant les valeurs de données ordonnées en tant que quantiles et en laissant Q_1 et Q_2 représenter les fonctions de quantiles des deux ensembles de données, on voit clairement à la figure 2.1.7.1 que

2.1.7.2 $$Q_2(p) = 2Q_1(p) + 1$$

Deux petits ensembles de données artificielles	
Ensemble de données 2	Ensemble de données 2
3, 5, 4, 7, 3	15, 7, 9, 7, 11

Tableau 2.1.7.1.

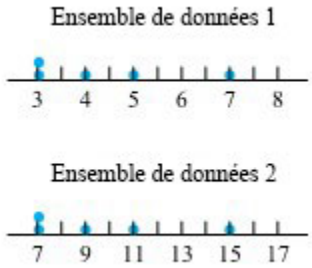


Figure 2.1.7.1 Diagrammes à points de deux petits ensembles de données.

.

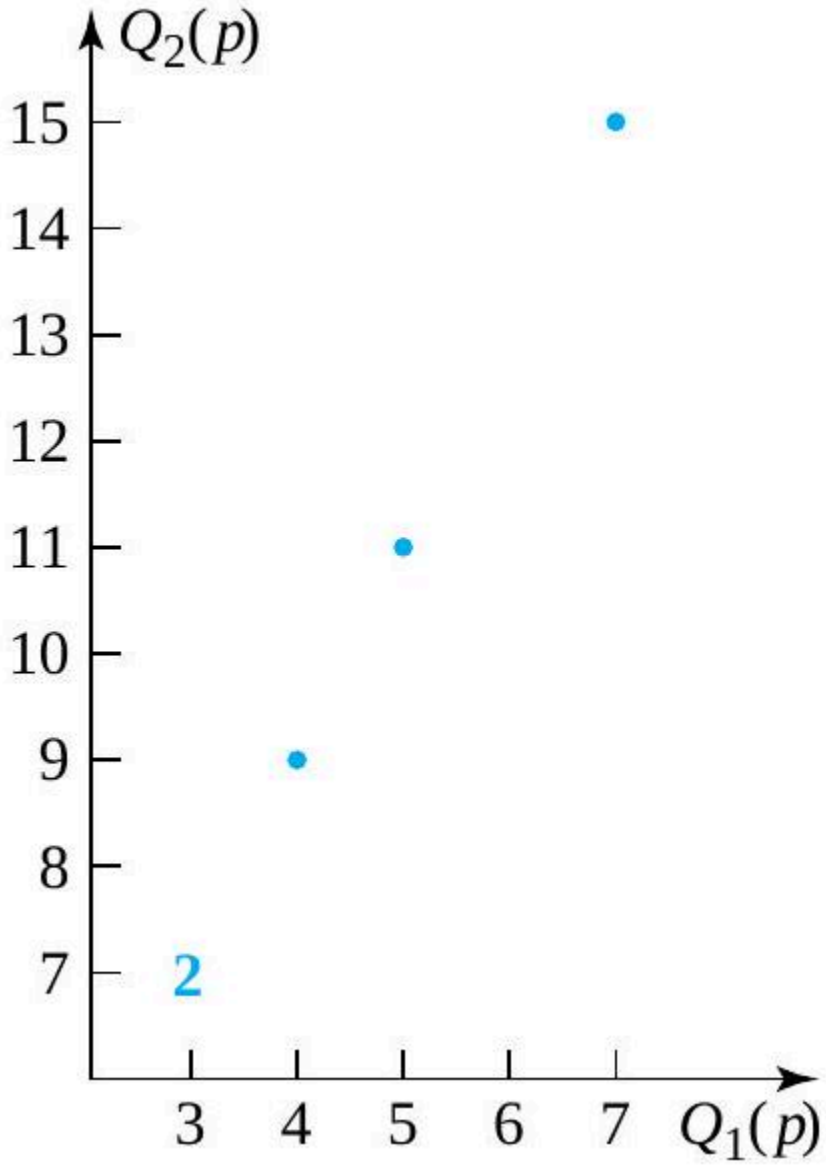


Figure 2.1.7.2. Diagramme Q-Q pour les données du tableau 2.1.7.1.

En d'autres termes, les fonctions de quantile des deux ensembles de données sont linéairement proportionnelles. En observant les figures 2.1.7.1 et 2.1.7.2, il est évident que le graphique des points :

$$\left(Q_1\left(\frac{i-0{,}5}{5}\right),\quad Q_2\left(\frac{i-0{,}5}{5}\right)\right)$$

(pour $i = 1, 2, 3, 4, 5$) devrait être exactement linéaire. La figure 2.1.7.2 illustre cela – en fait la figure 2.1.7.2 est un diagramme $Q - Q$ pour les ensembles de données du tableau 2.1.7.1.

DÉFINITION 2.1.7.1. Diagramme $Q - Q$

Un diagramme $Q - Q$ de deux ensembles de données avec des fonctions quantile respectives Q_1 et Q_2 est un diagramme de paires ordonnées $(Q_1(p), Q_2(p))$ pour les valeurs appropriées de p. Lorsque les deux ensembles de données sont de même taille n, les valeurs de p utilisées pour élaborer le diagramme seront $\frac{i - 0{,}5}{n}$, avec $i = 1, 2, \ldots, n$. Lorsque les ensembles de données sont de taille inégale, les valeurs de p utilisées pour élaborer le diagramme seront $\frac{i - 0{,}5}{n}$ avec $i = 1, 2, \ldots, n$, où n correspond à la taille de l'ensemble le plus petit.

ÉTAPES D'ÉLABORATION D'UN DIAGRAMME Q-Q

Pour élaborer le diagramme $Q - Q$ de deux ensembles de données de taille égale :

1. On classe les données de la plus petite à la plus grande.
2. On associe les données correspondantes des deux ensembles.
3. On représente graphiquement les paires ordonnées en utilisant les données du premier ensemble pour les abscisses et celles du second pour les ordonnées.

Lorsque l'on traite des ensembles de données de taille inégale, on associe les valeurs ordonnées du petit ensemble aux quantiles du grand ensemble obtenus par interpolation.

Un diagramme $Q - Q$ raisonnablement linéaire indique que les deux distributions ont des formes similaires. Lorsqu'il y a des écarts significatifs par rapport à la linéarité, le caractère de ces écarts révèle la manière dont les formes diffèrent.

Exemple 2.1.7.1. Pénétration des balles (suite)

Retournons à la profondeur de pénétration des balles. Le tableau précédent a fourni la matière première nécessaire à la réalisation d'un diagramme $Q - Q$. Il suffit d'associer les profondeurs de chaque ligne de ce tableau et de les tracer pour obtenir le tracé de la figure 2.1.7.3.

Dans l'ensemble, le diagramme de dispersion de la figure 2.1.7.3 n'est pas très linéaire. Toutefois, les points des valeurs n^{o} 2 à 13 de chaque ensemble de données semblent assez linéaires, ce qui indique que les extrémités inférieures des deux distributions ont des formes similaires (sauf pour leur bout).

L'espace horizontal entre les 13^{e} et 14^{e} points indique que l'écart entre $43{,}85\ \text{mm}$ et $47{,}30\ \text{mm}$ (pour les données des balles à 230 grains) est disproportionné par rapport à l'écart entre 63,55 et $63{,}80\ \text{mm}$ (pour les données des balles à 200 grains). Cela laisse supposer qu'il existe une différence physique fondamentale dans les mécanismes ayant causé la dispersion des données de profondeur des balles à 230 grains. Les statistiques peuvent révéler ce genre d'indice, mais pour expliquer les causes, il faut faire appel à des spécialistes de la balistique ou des matériaux.

En raison de l'écart marqué par rapport à la linéarité produit par le premier point $(27{,}75, 58{,}00)$, il existe également une différence importante dans la forme des extrémités inférieures des deux distributions. Pour remettre ce point en ligne avec le reste des points tracés, il faudrait le déplacer vers la droite (augmenter la plus petite donnée des balles à

230 grains) ou vers le bas (diminuer la plus petite observation des balles à 200 grains). En d'autres termes, par rapport à la distribution des balles à 200 grains, la distribution des balles à 230 grains présente une longue queue inférieure. (Ou, autrement dit, par rapport à la distribution des balles à 230 grains, la distribution des balles à 200 grains a une queue inférieure courte.) À noter que la différence de forme était déjà évidente dans le diagramme en boîte de la figure précédente. Encore une fois, il faudrait un spécialiste pour expliquer cette différence dans les formes de distribution.

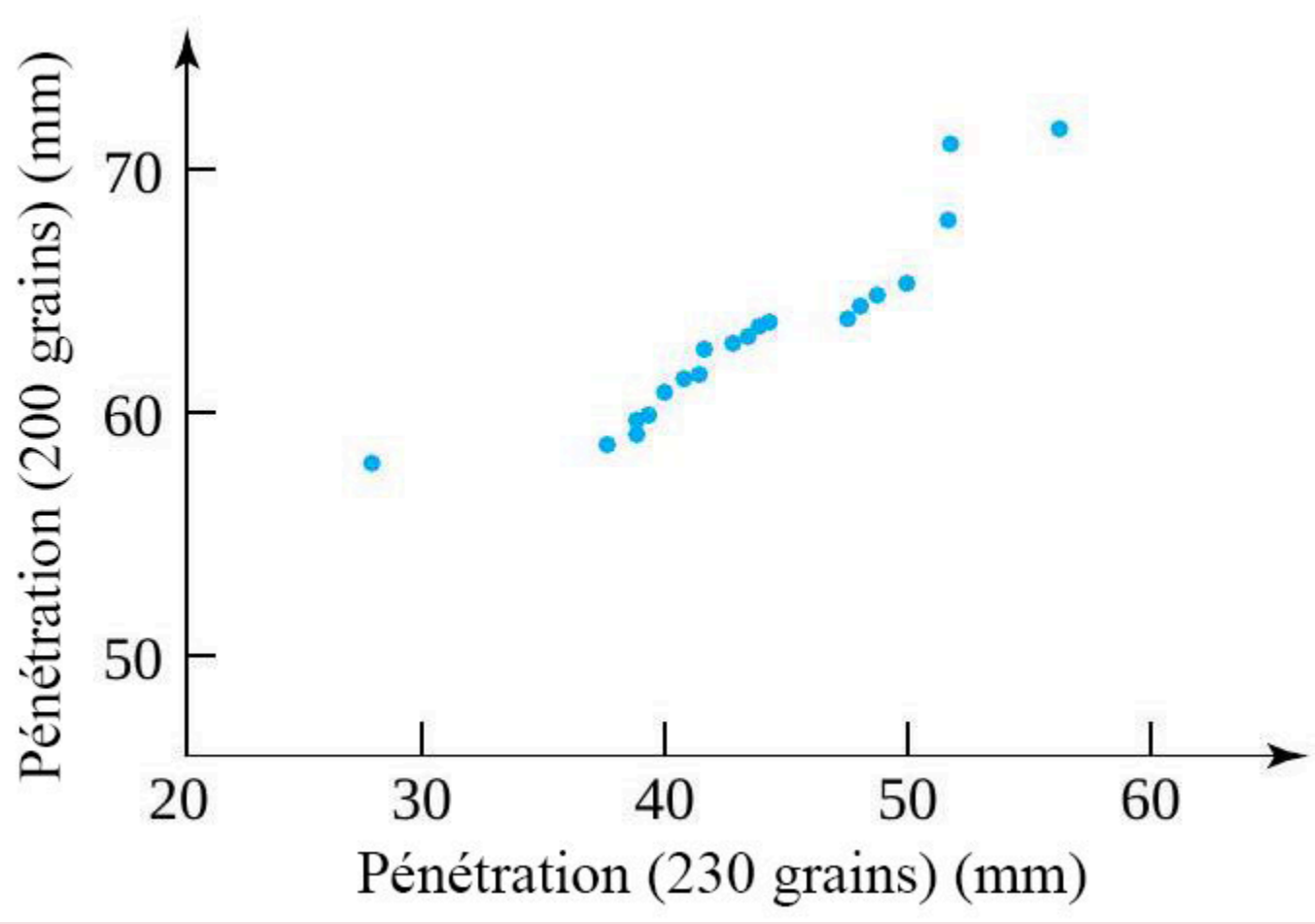


Figure 2.1.7.3. Diagrammes Q-Q de la profondeur de pénétration des balles

Le plus facile pour expliquer le concept de diagramme Q-Q (un outil très pratique pour comparer des jeux de donnés), c'est de voir une application où l'on compare des données empiriques. Mais le diagramme Q-Q est vraiment utile lorsqu'on l'applique à une fonction quantile qui représente un ensemble de données et à une seconde qui représente une *distribution théorique*.

DÉFINITION 2.1.7.2 Diagramme Q-Q théorique

Un diagramme Q-Q théorique, ou diagramme de probabilité, pour un ensemble de données de taille n et une distribution théorique, sont les fonctions quantiles sont respectivement Q_1 et Q_2, est un diagramme de paires ordonnées (Q_1(p), Q_2(p)) pour des valeurs appropriées de p. Dans cet ouvrage, les valeurs de p prennent la forme $\frac{i-0,5}{n}$, avec i = 1, 2,..., n.

Soit Q ($\frac{i-0,5}{n}$) le i^{e} point du petit ensemble de données, le
diagramme Q-Q théorique est un diagramme de points dans lequel les abscisses correspondent aux données expérimentales, et les ordonnées, aux quantiles de la distribution théorique. Autrement dit, on utilise les données ordonnées $x_1 \leq x_2 \leq \ldots \leq x_n$ pour tracer les points

2.1.7.3 Paires ordonnées d'un diagramme de probabilité

$$\left(x_i, Q_2\left(\frac{i-.5}{n}\right)\right)$$

Un tel diagramme permet de poser la question suivante: « L'ensemble des données a-t-il une forme similaire à la distribution théorique? »

TRACÉ NORMAL

Le diagramme théorique $Q-Q$ le plus connu est celui de la distribution normale (ou gaussienne), la distribution en forme de cloche bien. Le tableau 2.1.7.2 donne quelques quantiles de cette distribution. Pour trouver $Q(p)$ pour $p = .01, .02, \ldots, .98, .99$, on repère la ligne correspondant au premier chiffre après la décimale et la colonne correspondant au deuxième chiffre après la décimale. (Par exemple, $Q(.37) = -.33$.) Pour approximer les valeurs du tableau 2.1.7.2, on peut utiliser la relation suivante :

2.1.7.3 Approximation des quantiles normaux standards

$$Q(p) \approx 4.9\left(p^{.14} - (1-p)^{.14}\right)$$

À ce stade, le tableau 2.1.7.2 semble sortir de nulle part. Nous expliquerons comment l'obtenir à la partie 4, mais pour l'instant, contentons-nous de dire que les quantiles du tableau correspondent à une distribution normale. Imaginons que chaque entrée du tableau 2.1.7.2 corresponde à un point de données dans un ensemble de taille $n = 99$. Le tableau 2.1.7.3 présente une table de fréquences pour ces 99 points de données. La colonne Comptage du tableau 2.1.7.3 montre clairement la forme de cloche.
Les quantiles normaux standard peuvent servir à tracer un diagramme $Q-Q$ théorique afin d'évaluer la forme en cloche d'un ensemble de données. Le diagramme obtenu est appelé **diagramme normal (de probabilité)**.

Quantiles normaux standards

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0		−2,33	−2,05	−1,88	−1,75	−1,65	−1,55	−1,48	−1,41	−1,34
0,1	−1,28	−1,23	−1,18	−1,13	−1,08	−1,04	−0,99	−0,95	−0,92	−0,88
0,2	−0,84	−0,81	−0,77	−0,74	−0,71	−0,67	−0,64	−0,61	−0,58	−0,55
0,3	−0,52	−0,50	−0,47	−0,44	−0,41	−0,39	−0,36	−0,33	−0,31	−0,28
0,4	−0,25	−0,23	−0,20	−0,18	−0,15	−0,13	−0,10	−0,08	−0,05	−0,03
0,5	0,00	0,03	0,05	0,08	0,10	0,13	0,15	0,18	0,20	0,23
0,6	0,25	0,28	0,31	0,33	0,36	0,39	0,41	0,44	0,47	0,50
0,7	0,52	0,55	0,58	0,61	0,64	0,67	0,71	0,74	0,77	0,81
0,8	0,84	0,88	0,92	0,95	0,99	1,04	1,08	1,13	1,18	1,23
0,9	1,28	1,34	1,41	1,48	1,55	1,65	1,75	1,88	2,05	2,33

Tableau 2.1.7.2. Quantiles normaux standards

Table de fréquences des quantiles normaux standards

Valeur	Comptage	Fréquence
−2,80 to −2,30	|	1
−2,29 to −1,79	||	2
−1,78 to −1,28	𝍸 ||	7
−1,27 to −0,77	𝍸 𝍸 ||	12
−0,76 to −0,26	𝍸 𝍸 𝍸 ||	17
−0,25 to 0,25	𝍸 𝍸 𝍸 𝍸 |	21
0,26 to 0,76	𝍸 𝍸 𝍸 ||	17
0,77 to 1,27	𝍸 𝍸 ||	12
1,28 to 1,78	𝍸 ||	7
1,79 to 2,29	||	2
2,30 to 2,80	|	1

Tableau 2.1.7.3. Table de fréquences des quantiles normaux standards

.

Exemple 2.1.7.2. Résistance d'une serviette en papier (suite)

Revenons au test de résistance de la serviette en papier et voyons si les données suivent une distribution en forme de cloche. Le tableau 2.1.7.4 a été établi à partir du tableau original et du tableau 2.1.7.2; il fournit les informations nécessaires pour produire le diagramme $Q - Q$ théorique de la figure 2.1.7.4.

Malgré la petite taille de l'ensemble de données, le diagramme de la figure 2.1..4 semble relativement linéaire, et l'ensemble de données est donc raisonnablement en forme de cloche. La conséquence pratique de cette observation est qu'il est alors possible d'utiliser les modèles de probabilité normale, dont il sera question au chapitre 4, pour décrire la résistance des serviettes en papier. Ces modèles pourraient servir à faire des prévisions de résistance, et les méthodes d'inférence statistique formelle basées sur ces modèles pourraient servir à analyser les données de résistance.

Quantiles de résistance à la rupture et quantiles normaux standards

i	$\frac{i-0,5}{10}$	Quantile $\frac{i-0,5}{10}$ de résistance à la rupture	Quantile $\frac{i-0,5}{10}$ normal standard
1	0,05	7 583	−1,65
2	0,15	8 527	−1,04
3	0,25	8 572	−0,67
4	0,35	8 577	−0,39
5	0,45	9 011	−0,13
6	0,55	9 165	0,13
7	0,65	9 471	0,39
8	0,75	9 614	0,67
9	0,85	9 614	1,04
10	0,95	10 688	1,65

Tableau 2.1.7.4. Quantiles de résistance et quantiles normaux standards

.

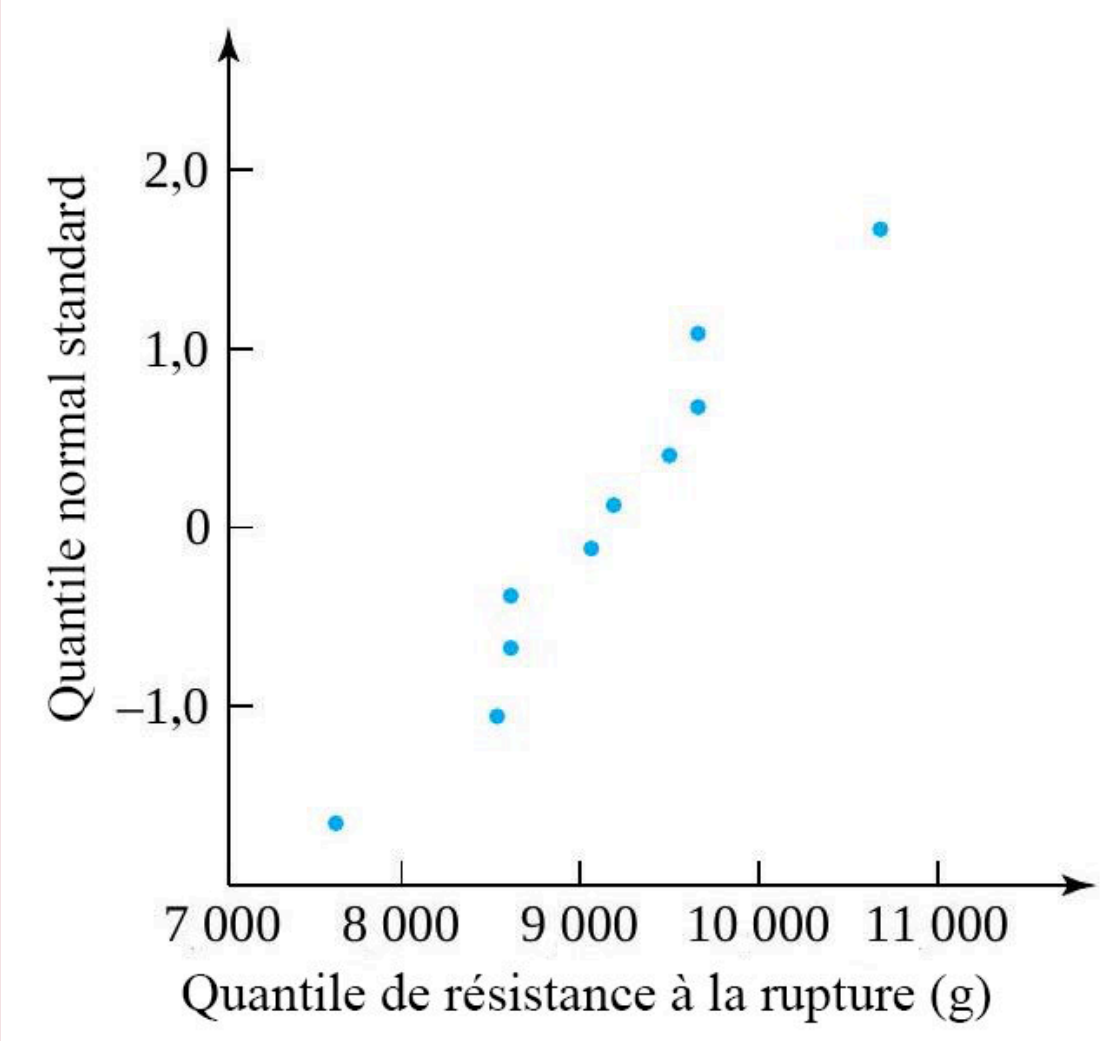


Figure 2.1.7.3. Diagramme Q-Q théorique de la résistance des serviettes en papier.

Pour produire des graphiques normaux, on peut utiliser un papier graphique spécial qu'on appelle papier de probabilité normale (ou simplement papier de probabilité). Au lieu de tracer des points sur du papier millimétré ordinaire en utilisant les positions verticales du tableau 2.1.7.2, sur du papier de probabilité, on trace les points en utilisant les positions verticales de la forme $\frac{i-0,5}{n}$. La figure 2.1.7.4 montre les données de résistance de l'exemple 2.1.7.2 tracées sur du papier de probabilité. À noter que ce tracé est pratiquement identique à celui de la figure 2.1.7.2.

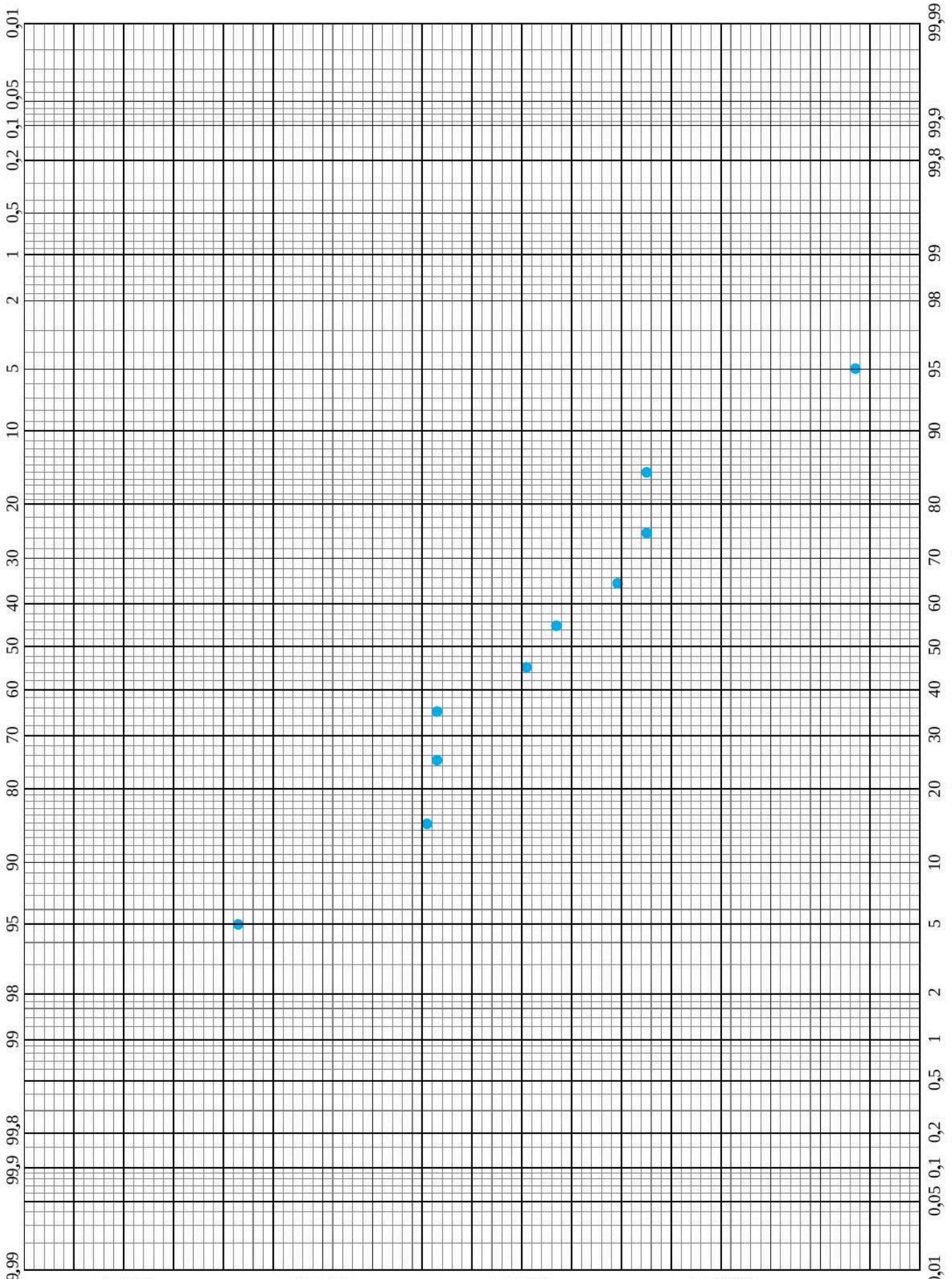

0,01 0,05 0,1 0,2 0,5 1 2 5 10 20 30 40 50 60 70 80 90 95 98 99 99,8 99,9 99,99
99,99 99,9 99,8 99 98 95 90 80 70 60 50 40 30 20 10 5 2 1 0,5 0,2 0,1 0,05 0,01

Figure 2.1.7.4. Tracé normal pour la résistance des serviettes en papier (sur papier de probabilité; image de Keuffel and Esser Company).

Les tracés normaux ne sont pas le seul type de tracés $Q - Q$ théoriques utiles en ingénierie. De nombreux autres types de distributions théoriques sont importants, et chacun d'entre eux peut servir à produire des tracés $Q - Q$ théoriques. Ce point est abordé plus en détail dans d'autres modules, mais l'introduction du tracé *Q-Q[/latex]permet d'insister sur le lien entre les graphiques de probabilités et les graphiques [latex]Q-Q[/latex] empiriques.*

2.2.1 Mesures de position

Pour la plupart des gens, le concept de « moyenne » évoque quelque chose de représentatif, ou le « centre », d'un ensemble de données. Les températures peuvent varier d'un endroit à l'autre dans un haut fourneau, mais la température moyenne donne une idée de la température « centrale » ou représentative. Les notes obtenues lors d'un examen peuvent varier, mais on est toujours content d'être au-dessus de la moyenne.
Le mot « moyenne », tel qu'il est utilisé dans le langage courant, correspond en fait à diverses significations techniques. La première est la médiane, $Q(0{,}5)$, qui a été présentée dans la dernière section. La médiane divise un ensemble de données en deux. Dans un histogramme bien conçu, environ la moitié de l'aire des barres se situe de part et d'autre de la médiane. En tant que mesure du centre, elle est totalement insensible aux effets de quelques observations extrêmes ou aberrantes. Par exemple, le petit ensemble de données

$$2, 3, 6, 9, 10$$

a une médiane de 6 , et cela reste vrai même si la valeur 10 est remplacée par $10\ 000\ 000$ et si la valeur 2 est remplacée par $-200,000$.
La section précédente a utilisé la médiane comme valeur centrale dans l'élaboration des diagrammes en boîte. Mais la médiane n'a pas le sens technique le plus souvent attaché à la notion de moyenne dans les analyses statistiques. Il est plus courant d'utiliser la moyenne (arithmétique).

DÉFINITION 2.2.1.1. Moyenne arithmétique.

La moyenne (arithmétique) d'un échantillon de données quantitatives, par exemple , $x_1, x_2, \ldots, x_n$, correspond à

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La moyenne est parfois appelée premier moment ou centre de masse d'une distribution, par analogie avec la mécanique. Si on place une masse unitaire le long de la droite numérique à la position de chaque valeur d'un ensemble de données – le point d'équilibre de la distribution de masse se situe à $\bar{x}$.

Exemple 2.2.1.1. Perte des rouleaux de papier

Hall, Luethe, Pelszynski et Ringhofer ont travaillé avec une entreprise qui découpe du papier à partir de grands rouleaux achetés en gros auprès de plusieurs fournisseurs. L'entreprise souhaitait déterminer la quantité de perte (en poids) sur les rouleaux provenant des différentes sources. Le tableau 2.2.1.1 présente les données relatives au pourcentage de perte que les étudiant.e.s ont obtenues pour six et huit rouleaux de papier, respectivement, achetés auprès de deux sources différentes.

Les médianes et les moyennes des deux ensembles de données sont faciles à obtenir. Pour le fournisseur 1,

$$Q(0{,}5) = 0{,}5(0{,}65) + 0{,}5(0{,}92) = 0{,}785\% \text{ de perte}$$

et

$$\bar{x} = \frac{1}{6}(0{,}37 + 0{,}52 + 0{,}65 + 0{,}92 + 2{,}89 + 3{,}62) = 1{,}495\% \text{ de perte}$$

Pour le fournisseur 2,

Q(0,5) = 0,5(1,47) + 0,5(1,58) = 1,525 \% \text { de pertes }

et

$$\bar{x} = \frac{1}{8}(0{,}89 + 0{,}99 + 1{,}45 + 1{,}47 + 1{,}58 + 2{,}27 + 2{,}63 + 6{,}54)$$
$$= 2{,}228\% \text{ de perte}$$

Pourcentage de perte par poids sur les rouleaux de papier

Fournisseur 1	Fournisseur 2
0,37, 0,52, 0,65, 0,92, 2,89, 3,62	0,89, 0,99, 1,45, 1,47, 1,58, 2,27, 2,63, 6,54

Tableau 2.2.1.1.

La figure 2.2.1.1 illustre des diagrammes à points sur lesquels on a indiqué la médiane et la moyenne. Remarquez que les médianes et les moyennes des deux fournisseurs montrent que les pertes du fournisseur 2 sont plus importantes que celles du fournisseur 1. Notez également qu'il existe une différence substantielle entre les valeurs médianes et moyennes pour un fournisseur donné. Dans les deux cas, la moyenne est nettement supérieure à la médiane correspondante. Cela reflète la nature asymétrique à droite des deux ensembles de données. Dans les deux cas, le centre de masse de la distribution est fortement tiré vers la droite par quelques valeurs extrêmement élevées.

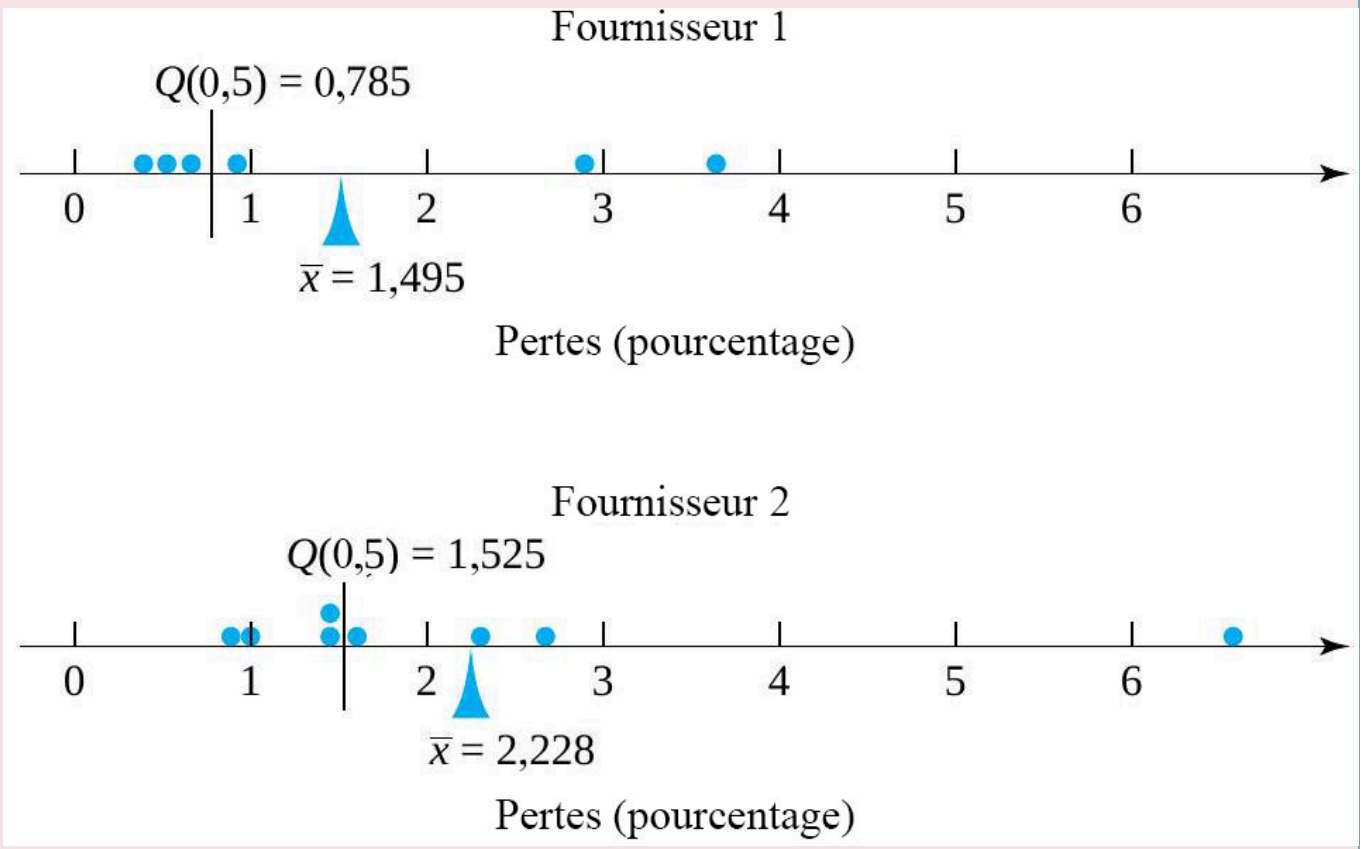


Figure 2.2.1.1. Diagrammes à points des pourcentages de pertes

L'exemple 2.2.1.1 montre clairement que, contrairement à la médiane, la moyenne est une mesure centrale qui peut être fortement influencée par quelques valeurs extrêmes. Certaines personnes disent parfois que, pour cette raison, l'une ou l'autre des deux mesures est « meilleure » – une affirmation qui n'a aucun sens. Ni l'une ni l'autre n'est meilleure; il s'agit simplement de mesures ayant des propriétés différentes. Et ces différences, les personnes averties qui lisent des statistiques doivent les garder à l'esprit. Par exemple, le salaire « moyen » des employé.e.s d'une entreprise qui paie neuf personnes 10 000 $ par an et son président 110 000 $ par an peut être décrit comme 10 000 $ par an (médiane) ou 20 000 $ par an (moyenne).

2.2.2 Mesures de dispersion

Quantifier la variation d'un ensemble de données peut être aussi important que d'en mesurer sa position. Dans le secteur manufacturier, par exemple, si une caractéristique des pièces sortant d'une machine donnée est mesurée et consignée, la dispersion des données obtenues donne des informations sur la précision et la capacité intrinsèques de la machine. La position des données obtenues est souvent fonction de la configuration de la machine ou du réglage des boutons d'ajustement. Les réglages peuvent être modifiés assez facilement, mais l'amélioration de la précision intrinsèque de la machine nécessite généralement des dépenses d'investissement pour un nouvel équipement ou la remise en état d'un équipement existant.
Bien que nous n'avons pas insisté sur ce point dans le module 2.1, on peut utiliser l'écart interquartile, $EIQ = Q(\ 0{,}75) - Q(\ 0{,}25)$ pour représenter la dispersion d'une distribution. L'écart interquartile mesure la répartition de la moitié centrale d'une distribution. Il est donc insensible à quelques valeurs extrêmes éventuelles. Une mesure apparentée est l'étendue, qui indique la dispersion de l'ensemble de la distribution.

DÉFINITION 2.2.2.1. Étendue
L'étendue d'un ensemble de données constitué de valeurs ordonnées $x_1 \leq x_2 \leq \cdots \leq x_n$ est

$$R = x_n - x_1$$

Notez l'utilisation des mots ici. Le mot étendue peut être utilisé comme verbe pour dire « Les données s'étendent de 3 à 21 ». Mais pour utiliser le mot comme un substantif, on dit « L'étendue est de $(21 - 3) = 18$ ». Étant donné que l'étendue ne dépend que des valeurs du plus petit point et du plus grand point d'un ensemble de données, elle est nécessairement très sensible aux valeurs extrêmes (ou aberrantes). Parce qu'elle est facile à calculer, elle a longtemps été populaire dans les milieux industriels, notamment en tant qu'outil de contrôle statistique de la qualité.
Cependant, la plupart des méthodes d'inférence statistique formelle sont basées sur une autre mesure de la répartition de la distribution. La notion d'« écart moyen au carré» ou d'« écart quadratique moyen » est utilisée pour obtenir des mesures appelées **variance** et **écart-type**, respectivement.

DÉFINITION 2.2.2.2. Variance et écart-type d'un échantillon
La **variance de l'échantillon** d'un ensemble de données composé des valeurs $x_1, x_2, \ldots, x_n$ est

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

L'**écart-type de l'échantillon,** s, est la racine carrée positive de la variance de l'échantillon.

Sauf pour le remplacement de $n - 1$ par n dans le diviseur, s^2 est la distance au carré moyenne des points de données par rapport à la valeur centrale $\bar{x}$. Par conséquent, s^2 est non négatif, et ne vaut 0 que si tous les points

de données sont exactement identiques. Les unités de s^2 sont le carré des unités des données d'origine. La racine carrée de s^2 (pour obtenir s) produit une mesure de la dispersion exprimée dans les unités d'origine.

Example 2.2.2.1. Pertes des rouleaux de papier (suite)

La dispersion des deux ensembles de pourcentages de pertes répertoriés dans le tableau 2.2.1.1 peuvent être exprimés dans l'un des termes précédents. Pour le fournisseur 1,

$$Q(.25) = .52$$
$$Q(.75) = 2.89$$

et par conséquent,

$$EIQ = 2{,}89 - 0{,}52 = 2{,}37\% \text{ de perte}$$

De même,

$$R = 3{,}62 - 0{,}37 = 3{,}25\% \text{ de perte}$$

En outre,

$$\begin{aligned} s^2 =& \frac{1}{6-1}\left((0{,}37 - 1{,}495)^2 + (0{,}52 - 1{,}495)^2 + (0{,}65 - 1{,}495)^2 + (0{,}92 - 1{,}495)^2\right. \\ &\left. + (2{,}89 - 1{,}495)^2 + (3{,}62 - 1{,}495)^2\right) \\ =& 1{,}945(\% \text{ de perte})^2 \end{aligned}$$

de sorte que

$$s = \sqrt{1{,}945} = 1{,}394\% \text{ de perte}$$

Des calculs similaires appliqués aux données du fournisseur 2 donnent les valeurs

$$EIQ = 1{,}23\% \text{ de perte}$$

et

$$R = 6{,}54 - 0{,}89 = 5{,}65\% \text{ de perte}$$

En outre,

$$\begin{aligned} s^2 =& \frac{1}{8-1}\left((0{,}89 - 2{,}228)^2 + (0{,}99 - 2{,}228)^2 + (1{,}45 - 2{,}228)^2 + (1{,}47 - 2{,}228)^2\right. \\ &\left. + (1{,}58 - 2{,}228)^2 + (2{,}27 - 2{,}228)^2 + (2{,}63 - 2{,}228)^2 + (6{,}54 - 2{,}228)^2\right) \\ =& 3{,}383(\% \text{ de perte})^2 \end{aligned}$$

donc

$$s = 1{,}839\% \text{ de perte}$$

Le fournisseur 2 a un EI plus petit, mais des valeurs de R et s plus grandes. Ceci est cohérent avec la figure 2.2.1.1 : la partie centrale de la distribution du fournisseur 2 est très dense, mais le point extrême rend la variabilité globale plus importante pour le second fournisseur que pour le premier.

Le calcul des variances d'échantillon que nous venons d'illustrer vise simplement à renforcer le fait que s^2 représente une sorte de moyenne du carré de l'écart. Bien entendu, la façon la plus judicieuse de trouver les variances d'échantillon dans la pratique est d'utiliser une calculatrice électronique de poche avec une fonction de variance préprogrammée, ou un logiciel statistique.

2.2.3 Statistiques et paramètres

À ce stade, il est important de présenter un peu plus de terminologie de base. Le jargon et les notations des distributions d'échantillons sont quelque peu différents de ceux des distributions de population (et des distributions théoriques).

DÉFINITION 2.2.3.1. Statistiques et paramètres

Les synthèses numériques des données d'échantillons sont nommées **statistiques** d'échantillon. Les synthèses numériques de distribution de population ou théoriques sont nommées **paramètres** (de population ou de modèle). Généralement, on utilise les lettres latines pour les statistiques, et les lettres grecques pour les paramètres.

Prenons l'exemple de la moyenne. La définition 2.2.1.1 porte spécifiquement sur le calcul pour un échantillon. Si un ensemble de données représente une population entière, il est courant d'utiliser la lettre grecque minuscule mu (μ) pour représenter la moyenne de la population et de noter :

Moyenne de la population 2.2.3.2.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Si on compare cette expression à celle de la définition 2.2.1.1, on remarque que le symbole utilisé pour la moyenne est différent, et qu'on utilise N plutôt que n. Il est courant de désigner la taille d'une population par N et la taille d'un échantillon par n.
Un autre exemple de l'utilisation suggérée par la définition 2.2.3.1 est celui de la variance et de l'écart-type. La définition 2.2.1.2 porte spécifiquement sur la variance et sur l'écart-type de l'échantillon. Si un ensemble de données représente une population entière, il est courant d'utiliser la lettre grecque sigma minuscule au carré ((σ^2)) pour représenter la variance de la population et pour définir:

Variance de la population 2.2.3.3.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

La racine carré positive de σ^2 correspond donc à l'écart-type de la population, σ.
Ce manuel s'écarte de la convention de symbolisme romain/grec exposée dans la définition 2.2.3.1 sur un point : la notation des quantiles. Nous utilisons $Q(p)$ pour représenter le quantile p d'une distribution, qu'elle soit extraite d'un échantillon, d'une population ou d'un modèle théorique.

2.2.4 Diagrammes de statistiques synthétiques en fonction du temps et de facteurs

Pendant l'analyse préliminaire des données, il est souvent utile de représenter des mesures numériques sur divers graphiques. Notamment, les diagrammes de statistiques synthétiques en fonction du temps sont souvent révélateurs.

Exemple 2.2.4.1 Contrôle d'une dimension critique de pièces usinées (suite)

Cowan, Renk, Vander Leest et Yakes ont travaillé avec une entreprise qui fabrique des pièces métalliques de précision. Une dimension critique de l'une de ces pièces a été contrôlée en sélectionnant et en mesurant occasionnellement cinq pièces consécutives, puis en traçant la moyenne et l'étendue de l'échantillon. Le tableau 2.2.4.1 répertorie les valeurs $\bar{x}$ et R pour 25 échantillons consécutifs de cinq pièces. Les valeurs indiquées sont exprimées en multiples de 0,0001 po.

La figure 2.2.4.1 illustre un diagramme des moyennes et des étendues en fonction de l'ordre d'observation. Si on considère tout d'abord le diagramme des étendues, aucune tendance forte n'est évidente, ce qui suggère que la variation de base à court terme est stable pour cette dimension. La précision du processus et des mesures est stable dans le temps. Le diagramme des moyennes, cependant, suggère un changement physique. Les dimensions moyennes du deuxième quart du 27 octobre (échantillons 9 à 15) sont nettement plus petites que le reste des moyennes. Il s'est avéré que les pièces produites par cette équipe n'étaient pas systématiquement différentes des autres. En fait, la personne qui a effectué les mesures pour les échantillons 9 à 15 a utilisé la jauge d'une manière fondamentalement différente de celle des autres employé.e.s. La distribution des valeurs $\bar{x}$ a été généré par cette modification de la technique de mesure.

.

Moyennes et plages des dimensions critiques pour les échantillons de n= 5 pièces

Échantillon	Date	Heure		$\bar{x}$	R	Échantillon	Date	Heure		$\bar{x}$	R
1	10/27	7:30	AM	3509,4	5	14		10:15		3504,4	4
2		8:30		3509,2	2	15		11:15		3504,6	3
3		9:30		3512,6	3	16	10/28	7:30	AM	3513,0	2
4		10:30		3511,6	4	17		8:30		3512,4	1
5		11:30		3512,0	4	18		9:30		3510,8	5
6		12:30	PM	3513,6	6	19		10:30		3511,8	4
7		1:30		3511,8	3	20		6:15	PM	3512,4	3
8		2:30		3512,2	2	21		7:15		3511,0	4
9		4:15		3500,0	3	22		8:45		3510,6	1
10		5:45		3502,0	2	23		9:45		3510,2	5
11		6:45		3501,4	2	24		10:45		3510,4	2
12		8:15		3504,0	2	25		11:45		3510,8	3
13		9:15		3503,6	3						

Tableau 2.2.4.1

.

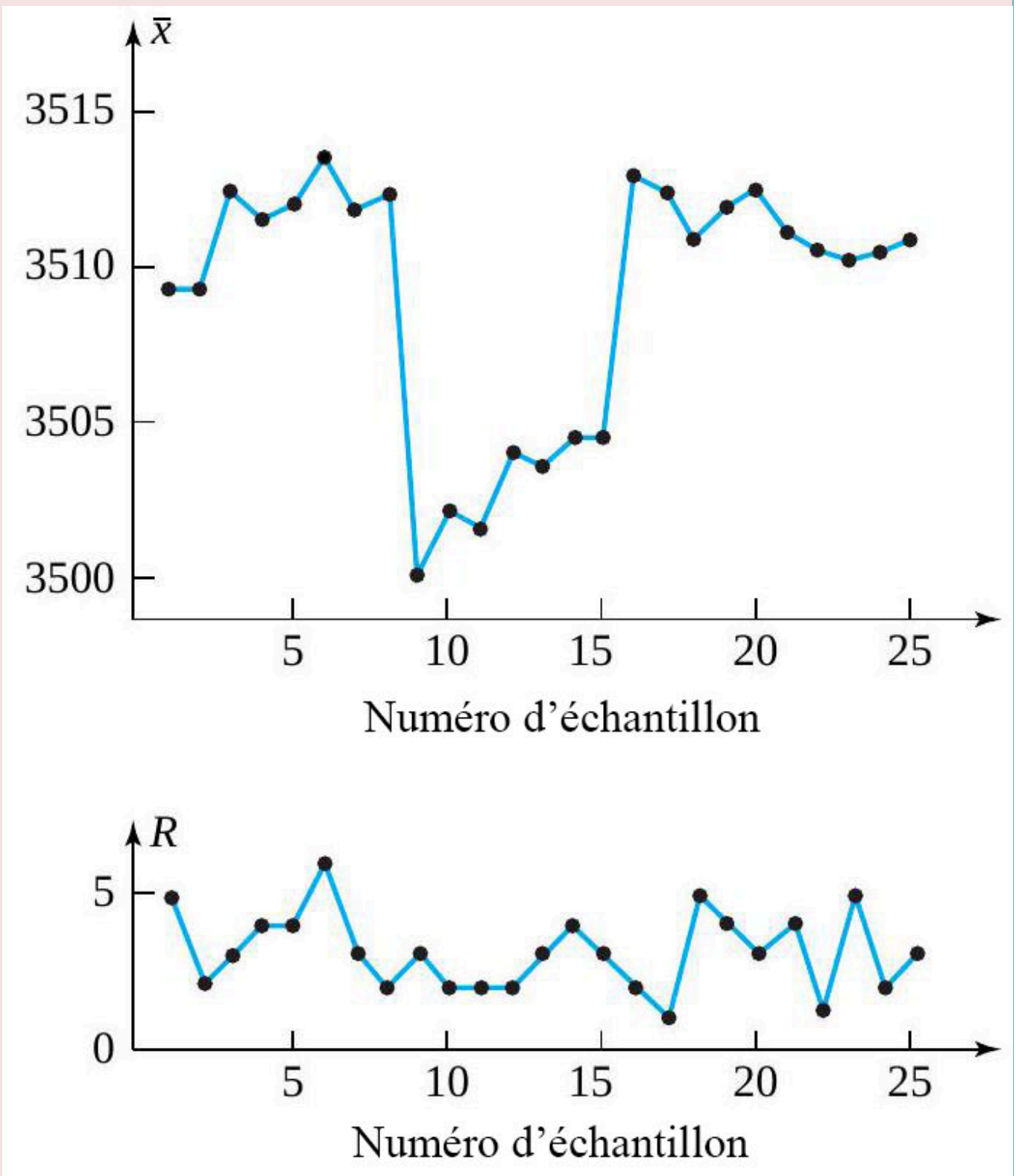


Figure 2.2.4.1 Tracés de `Formula does not parse` *\bar{x}* `Formula does not parse` *et* `Formula does not parse` *E* `Formula does not parse` *en fonction du temps.*

TERMINOLOGIE ET CAUSES DES TENDANCES SUR LES GRAPHIQUES EN FONCTION DU TEMPS

Les tendances révélées par le tracé des statistiques de l'échantillon en fonction du temps devraient signaler qu'il faut chercher une cause physique (et généralement, une solution). Les variations systématiques ou cycliques dans un diagramme de moyennes peuvent souvent être liées à des variables de processus qui vont et viennent de manière plus ou moins régulière. Il s'agit par exemple de variables saisonnières ou quotidiennes telles que la température ambiante, ou encore la rotation des jauges ou des appareils. Une instabilité ou une variation supérieure à celle de la précision de l'équipement peut parfois être due à l'hétérogénéité des matières premières ou à un réglage excessif de l'équipement. Le changement de la moyenne d'un processus peut être causée par l'introduction de nouvelles machines, de nouvelles matières premières, la formation des employé.e.s, l'usure des outils, etc. Les combinaisons de plusieurs schémas de variation sur un même diagramme d'une statistique synthétique en fonction du temps peuvent parfois être attribués à des changements dans l'étalonnage des mesures, comme dans l'exemple 2.2.4.1. Ils sont aussi parfois produits par des différences constantes dans les machines ou les flux de matières premières.

TRACÉS EN FONCTION DES VARIABLES DU PROCESSUS

Les diagrammes de statistiques synthétiques en fonction du temps ne sont pas les seuls qui sont utiles. Les graphiques des variables du processus peuvent également être très instructifs.

.

Exemple 2.2.4.2 Graphique de la résistance moyenne au cisaillement des joints de bois.

Dans leur étude sur la résistance des joints de bois collés, Dimond et Dix ont obtenu les valeurs indiquées dans le tableau 2.2.4.2 comme résistances moyennes (sur trois essais de cisaillement) pour chaque combinaison de trois essences de bois et de trois colles. La figure 2.2.4.2 illustre un tracé révélateur de ces $3 \times 3 = 9\,\bar{x}$ différents.

Le diagramme montre clairement que les propriétés de collage du pin et du sapin sont assez similaires, les joints en pin étant en moyenne de 40 à 45 lb plus résistants. Pour ces deux bois tendres, la Cascamite semble légèrement supérieure à la colle de menuisier, toutes deux permettant de réaliser de bien meilleurs assemblages que la colle blanche. Les propriétés de collage du chêne (un bois dur) sont très différentes de celles du pin et du sapin. En fait, les colles ont des effets exactement opposés sur la résistance des joints de chêne. Tout ceci est clairement illustré par le diagramme simple de la figure 2.2.4.2.

.

Résistances moyennes du joint pour les neuf combinaisons essence/colle

Essence	Colle	$\bar{x}$ Résistance moyenne du joint au cisaillement (lb)
Pin	blanche	131,7
Pin	menuisier	192,7
Pin	Cascamite	201,3
Sapin	blanche	92,0
Sapin	menuisier	146,3
Sapin	Cascamite	156,7
Chêne	blanche	257,7
Chêne	menuisier	234,3
Chêne	Cascamite	177 7

Tableau 2.2.4.2

.

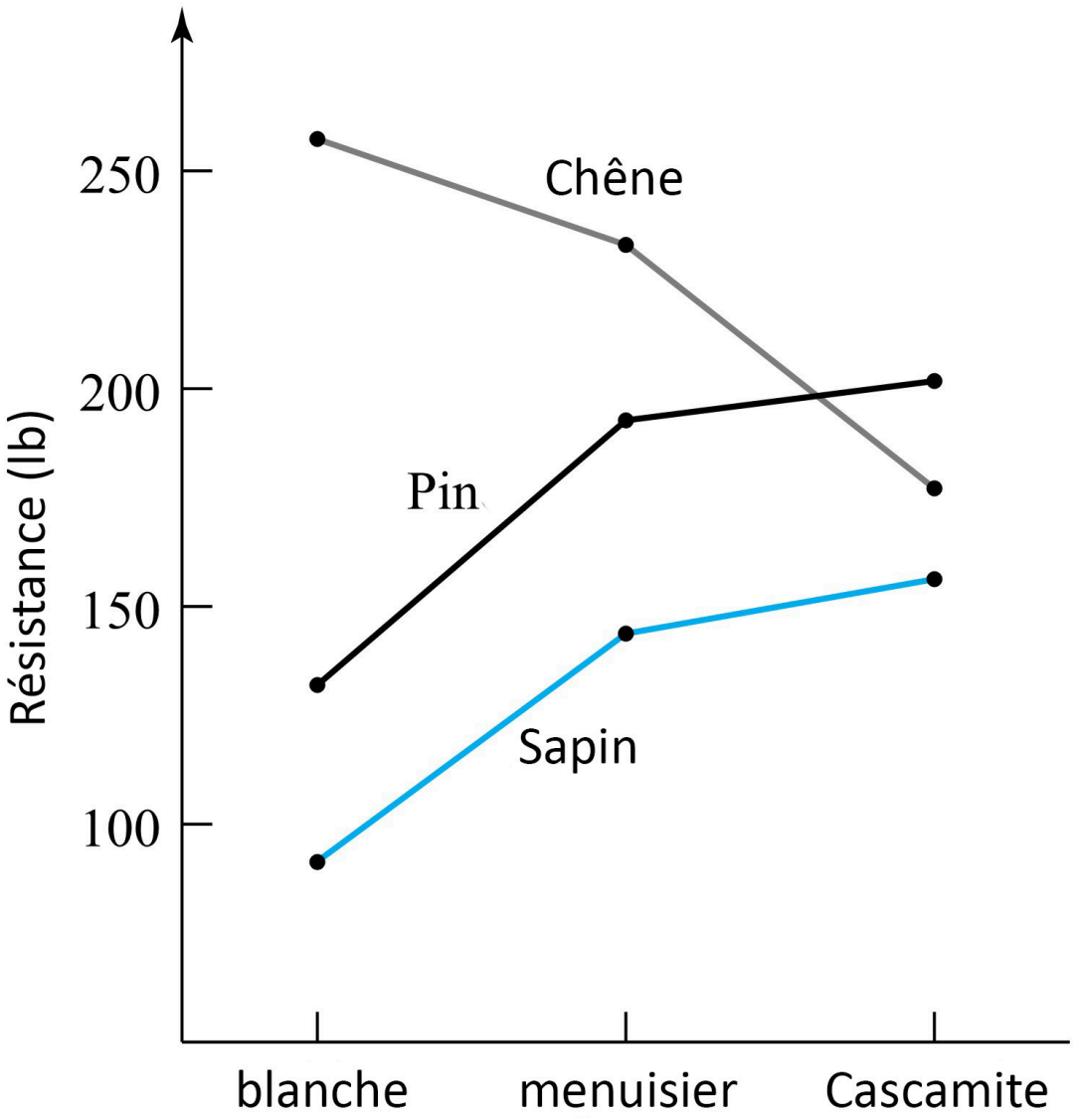


Figure 2.2.4.2 Graphique de la résistance moyenne des joints en fonction du type de colle pour trois essences de bois.

Les deux exemples précédents ont illustré l'utilité de tracer les statistiques d'un échantillon en fonction du temps ou d'une variable expérimentale.

2.2.5 Diagramme à barres et graphiques de données qualitatives ou de dénombrement

Les techniques présentées jusqu'à présent dans ce chapitre concernent principalement l'analyse des données de mesure. Comme nous l'avons mentionné dans la partie 1, les données de mesure (ou données de variables) sont généralement préférables, si on peut les obtenir, aux données chiffrées et aux données qualitatives (ou données d'attributs). Néanmoins, les données qualitatives ou de dénombrement sont parfois les principales informations disponibles. Il est donc intéressant d'étudier leur synthèse et leur visualisation.

Souvent, une étude produit plusieurs valeurs de $\hat{p}$ ou $\hat{u}$ qui doivent être comparées. Les diagrammes en barres et les diagrammes simples à deux variables peuvent être d'une grande aide pour résumer ces résultats.

Example 2.2.5.1. Classification des connecteurs de câble selon leur défaut.

Delva, Lynch et Stephany ont travaillé avec un fabricant de connecteurs de câbles. Des échantillons de 100 connecteurs d'une conception donnée ont été prélevés chaque jour pendant 30 jours de production, et chaque connecteur échantillonné a été inspecté conformément à un ensemble de règles (opérationnelles) bien définies. Sur la base des informations fournies par les inspections, chaque connecteur inspecté a pu être classé dans l'une des cinq catégories mutuellement exclusives suivantes :

Catégorie A : présente des défauts « très graves »
Catégorie B : présente des défauts « graves » mais pas « très graves »
Catégorie C : présente des défauts « modérément graves » mais pas « graves » ni « très graves »
Catégorie D : présente seulement des défauts « mineurs »
Catégorie E : ne présente aucun défaut

Le tableau 2.2.5.1 indique le nombre de connecteurs échantillonnés classés dans les quatre premières catégories (les quatre catégories de défauts) au cours de la période de 30 jours. Ensuite, en s'appuyant sur le fait que <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/a22cadfd31e1dcf72c7e432150c2fd42.png » alt= »30 \times 100 = 3 000 Formula does not parse connecteurs ont été inspectés au cours de cette période,

Formula does not parse \begin{aligned}& \hat{p}_{\mathrm{A}} = 3 / 3000 = 0,0010 \\& \hat{p}_{\mathrm{B}} = 0 / 3000 = 0,0000 \\& \hat{p}_{\mathrm{C}} = 11 / 3000 = 0,0037 \\& \hat{p}_{\mathrm{D}} = 1 / 3000 = 0,0003\end{aligned}" title="30 \times 100 = 3 000 Formula does not parse connecteurs ont été inspectés au cours de cette période,

Formula does not parse \begin{aligned}& \hat{p}_{\mathrm{A}} = 3 / 3000 = 0,0010 \\& \hat{p}_{\mathrm{B}} = 0 / 3000 = 0,0000 \\& \hat{p}_{\mathrm{C}} = 11 / 3000 = 0,0037 \\& \hat{p}_{\mathrm{D}} = 1 / 3000 = 0,0003\end{aligned}" class="latex mathjax">

Notez qu'ici $\hat{p}_{\mathrm{E}} = 1 - (\hat{p}_{\mathrm{A}} + \hat{p}_{\mathrm{B}} + \hat{p}_{\mathrm{C}} + \hat{p}_{\mathrm{D}})$, car les catégories A à E constituent un ensemble de catégories exhaustives et mutuellement exclusives, de sorte que la somme des $\hat{p}$ doit valoir 1.

Nombre de connecteurs classés dans quatre catégories de défauts

Catégorie	Nombre de connecteurs échantillonnés
A	3
B	0
C	11
D	1

Tableau 2.2.5.1.

La figure 2.2.5.1 est un diagramme à barres des fractions de connecteurs des catégories A à D. Elle montre clairement que la plupart des connecteurs présentant des défauts appartiennent à la catégorie C, celle des défauts modérément graves, mais ni graves ni très graves. Ce diagramme à barres présente le comportement d'une variable catégorique.

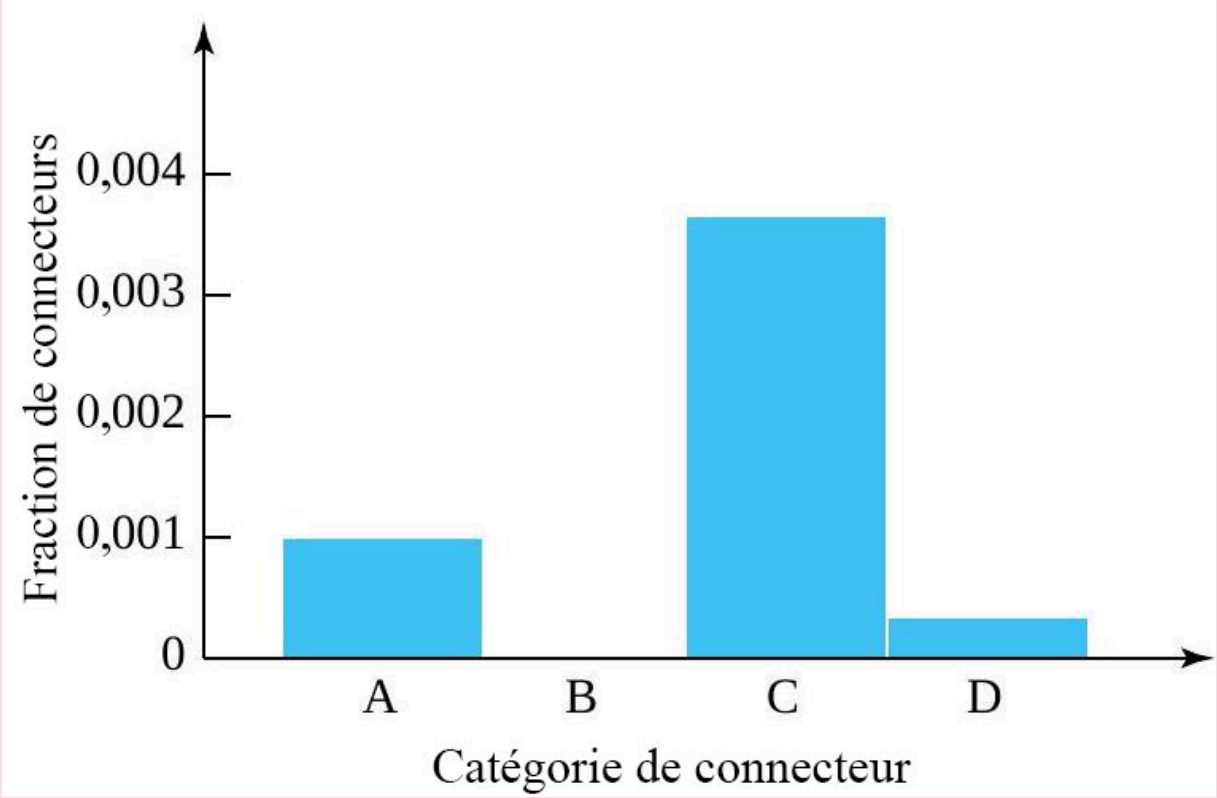


Figure 2.2.5.1. Diagramme à barres des défauts des connecteurs.

Example 2.2.5.2. Fabrication d'outils pneumatiques.

Kraber, Rucker et Williams ont travaillé avec un fabricant d'outils pneumatiques. Chaque outil produit est minutieusement inspecté avant d'être expédié. Les étudiant.e.s ont recueilli des données sur plusieurs types de problèmes découverts lors de l'inspection finale. Le tableau 2.2.5.2 indique le nombre d'outils présentant ces problèmes dans une série de 100 outils.

Nombre et fractions d'outil présentant divers problèmes

Problème	Nombre d'outils	$\hat{p}$
Fuite de type 1	8	0,08
Fuite de type 2	4	0,04
Fuite de type 3	3	0,03
Pièce 1 manquant	2	0,02
Pièce 2 manquant	1	0,01
Pièce 3 manquant	2	0,02
Pièce 4 défaillante	1	0,01
Pièce 5 défaillante	2	0,02
Pièce 6 défaillante	1	0,01
Mauvaise pièce 7	2	0,02
Mauvaise pièce 8	2	0,02

Tableau 2.2.5.2.

Ce tableau est une synthèse de données qualitatives à plusieurs variables. Les catégories énumérées dans le tableau 2.2.5.2 ne sont pas mutuellement exclusives; un outil donné peut être compté dans plusieurs catégories. Au lieu de représenter différentes valeurs possibles d'une seule variable catégorique (comme c'était le cas avec les catégories de connecteurs dans l'exemple 2.2.5.1), ces catégories sont construites selon deux conditions possibles (présence ou absence), et leur valeur correspond au

dénombrement des présences. Par exemple, pour les types de fuites 1, $\hat{p} = 0{,}08$, donc la proportion d'outils ne présentant pas le type de fuite 1 est $1 - \hat{p} = 0{,}92$. Le total des valeurs $\hat{p}$ n'est pas forcément égal à la fraction des outils problématiques lors de l'inspection finale. Un outil défectueux donné peut être comptabilisé dans plusieurs valeurs $\hat{p}$.

La figure 2.2.5.2 représente un diagramme à barres des informations sur les problèmes d'outils figurant dans le tableau 2.2.5.1. Elle montre que les fuites sont les problèmes les plus fréquents sur cette série de production.

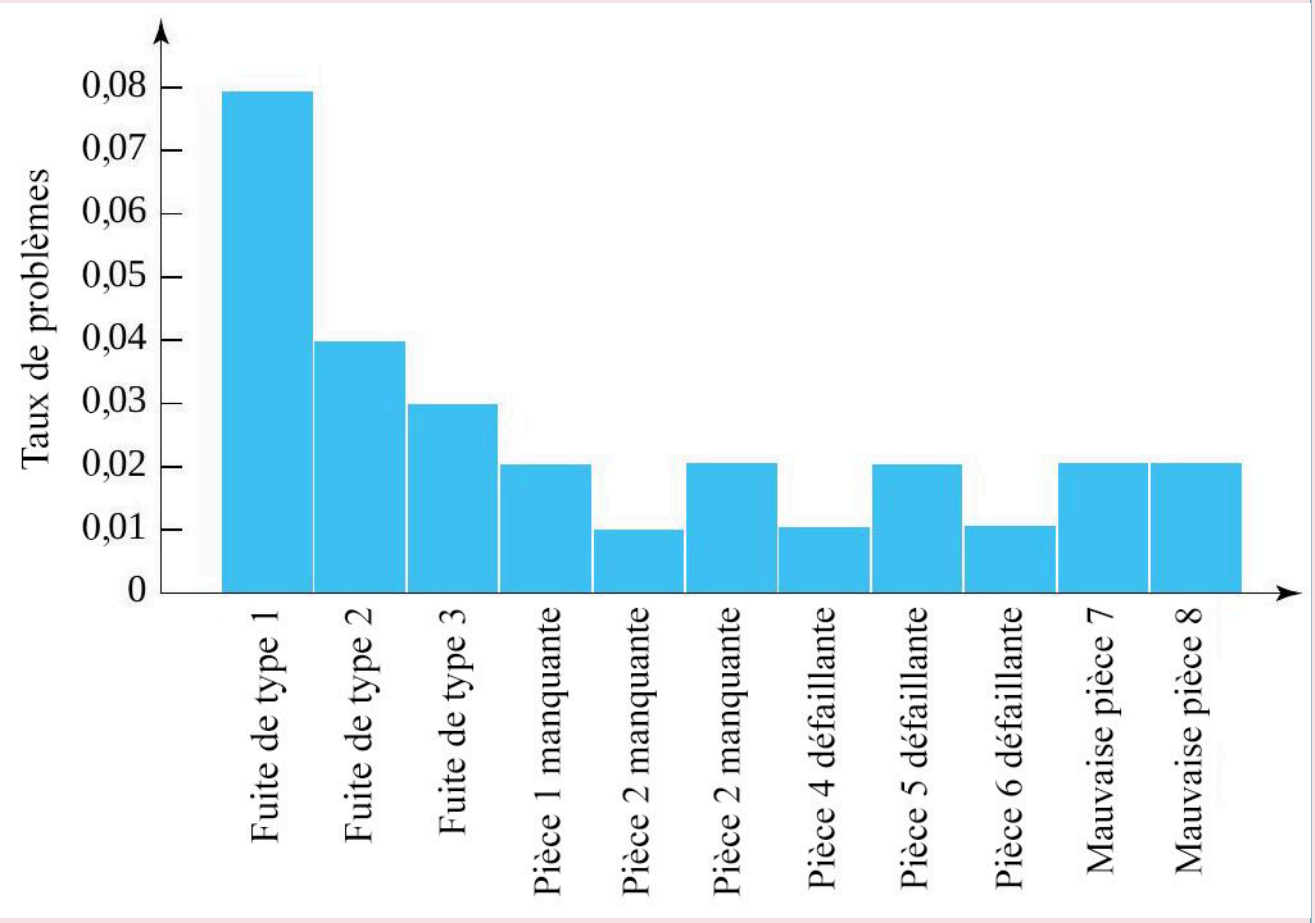


Figure 2.2.5.2. Diagramme à barres des problèmes d'assemblage.

Les figures 2.2.5.1 et 2.2.5.2 illustrent toutes deux des diagrammes à barres, mais elles diffèrent considérablement. La première montre le comportement d'une seule variable catégorique (ordonnée), à savoir la classe de connecteur. La seconde concerne le comportement de 11 variables catégoriques présence-absence différentes, comme la fuite de type 1, la pièce 3 manquante, etc. La forme de la figure 2.2.5.1 peut avoir une certaine signification, car les catégories A à D sont classées par ordre décroissant de gravité des défauts, et c'est cet ordre qui a été utilisé dans la figure. Mais la forme de la figure 2.2.5.2 est essentiellement arbitraire, puisque l'ordre des catégories de problèmes d'outils est lui-même arbitraire. D'autres ordres tout aussi sensés donneraient des formes tout à fait différentes.

2.2.6 Statistiques synthétiques et calcul statistique

Les synthèses de données numériques présentées dans ce chapitre sont relativement simples. Pour de petits ensembles de données, elles peuvent être calculées assez facilement à l'aide d'une simple calculatrice de poche. Toutefois, pour les grands ensembles de données ou lorsqu'on voudra peut-être faire d'autre calculs ou produire des graphiques, un logiciel statistique peut être pratique.

Ce Jupyter Notebook utilisant Python est accessible et peut être consulté et téléchargé sur le site GitHub du cours ou sur le site Special GitHub Site for Part 2.

Vous pouvez également ouvrir un environnement informatique interactif pour travailler avec le Jupyter Notebook utilisant Python à travers un site Binder en passant par le site GitHub de l'exemple de la partie 2. Cliquez sur ce site Binder pour accéder au site Binder pour l'exemple (accessible à l'adresse https://mybinder.org/v2/gh/Statistical-Methods-for-Engineering/Special-GitHub-Site-Part-2-Example-Percent-Waste-by-Weight-on-Bulk-Paper-Rolls/HEAD).

La capture d'écran ci-dessous montre comment utiliser le module statistique de Python pour Jupyter Notebook afin de produire des statistiques synthétiques pour les ensembles de données sur les pourcentages de pertes dont il a été question dans ce chapitre. La moyenne, la médiane et l'écart-type de correspondent à ce qui a été obtenu dans l'exemple. Cependant, le premier et le troisième quartiles ne correspondent pas exactement à ceux trouvés précédemment. (Les bibliothèques Python numpy et pandas utilisent des conventions légèrement différentes de celles présentées dans le chapitre 2 pour ces quantités.)

:	Fournisseur_1
nombre	6,000000
moyenne	1,495000
écart-type	1,394457
min.	0,370000
25 %	0,552500
50 %	0,785000
75 %	2,397500
max	3,620000

Il y a de nombreux outils informatiques de haute qualité pour les statistiques, dont Python, mais aussi JMP, SAS, SPSS, SYSTAT, SPLUS, MINITAB, MATLAB, R, etc. Tout.e ingénieur.e devrait en avoir un sur son ordinateur de travail. Malheureusement, ce n'est pas toujours le cas et les ingénieur.e.s supposent souvent qu'un logiciel de feuille de calcul standard (potentiellement avec des plugiciels tiers) constitue un substitut viable. C'est souvent vrai, mais pas toujours. Les ingénieur.e.s modernes ont besoin des statistiques informatiques et d'un certain niveau de compétence en sciences des données.

.

:	Fournisseur_2
nombre	8,000000
moyenne	2,227500
écart-type	1,839229
min.	0,890000
25%	1,335000
50%	1,525000
75%	2,360000
max	6,540000

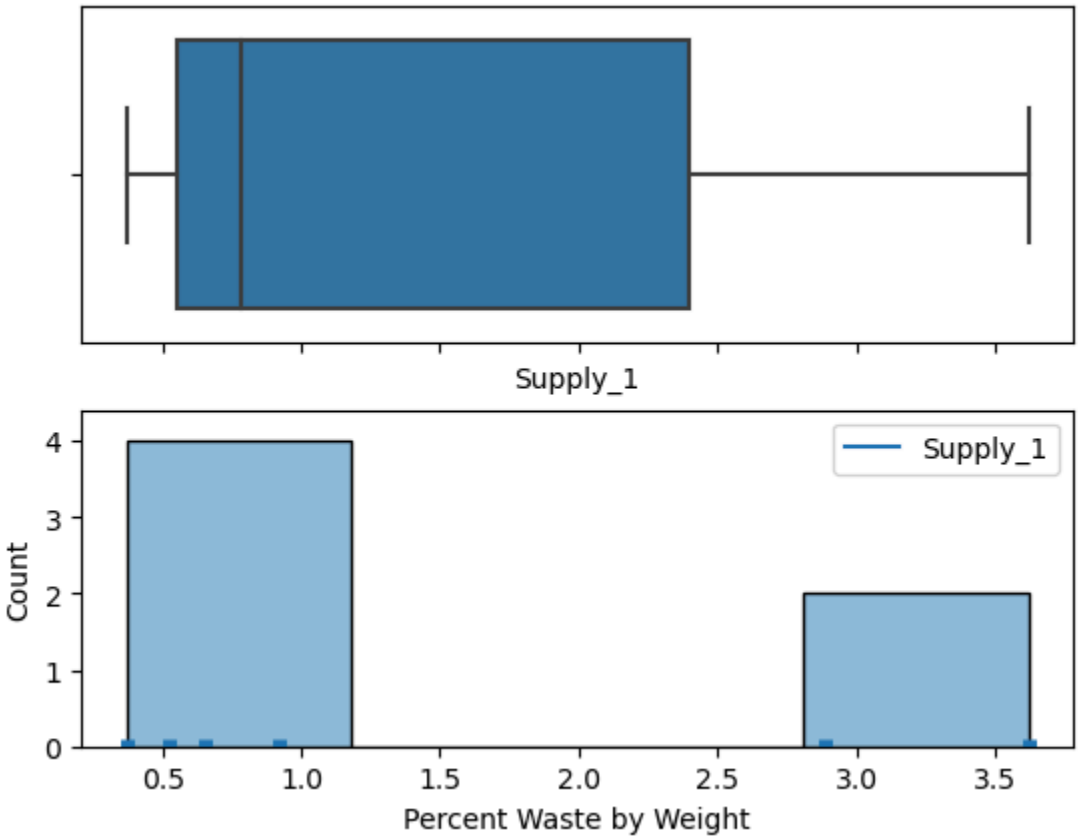


Figure 2.2.6.1.

La figure 2.2.6.1 est un diagramme en boîtes et un histogramme du fournisseur 1 de l'exemple. Consultez le Jupyter Notebook et commencez à synthétiser ces données et à les représenter graphiquement.

2.2.7 Tutoriel 2 – Nettoyage des données, synthèse et graphiques dans Python

À ce stade, il est recommandé de travailler sur l'exercice du tutoriel 2 qui se trouve sur le référentiel GitHub associé. Cet exercice vous permettra de vous familiariser avec le nettoyage des données et la création de graphiques simples dans Python.

Il est fortement recommandé de consulter le fichier du Jupyter Notebook intitulé Reading Data into Python & Data Cleaning. Vous pouvez les trouver dans la section « How do I do X in Python? ».

3.0.1 Introduction aux probabilités et aux variables aléatoires

Figure 3.1.0.1. Blaise Pascal, Palais de Versailles, CC BY 3.0 <https://creativecommons.org/licenses/by/3.0>, via Wikimedia Commons. Pierre de Fermat, https://commons.wikimedia.org/wiki/File:Pierre_de_Fermat.jpg

La naissance des probabilités en tant que discipline pratique et scientifique est principalement attribuée aux efforts conjoints de Blaise Pascal (1623-1662) et de Pierre de Fermat (1601-1665). Leur collaboration a débuté à la suite d'un problème de jeu posé par le Chevalier de Méré en 1654. Dans leur correspondance datant de 1654, ils se sont penchés sur un problème de jeu connu sous le nom de « problème des points », également appelé « problème des partis ». Ce problème consiste essentiellement à élaborer une méthode équitable pour distribuer les mises lorsqu'une partie se termine prématurément, sans vainqueur définitif. Dans leur correspondance, Pascal et Fermat ont non seulement abordé ces problèmes spécifiques, mais ils ont également jeté les bases de la théorie des probabilités.

Dans les modules précédents, nous avons exploré comment utiliser les statistiques descriptives et la visualisation des données pour synthétiser les données. Après la description des données, il est souvent essentiel de décrire le processus à l'origine des données, en particulier lorsqu'on essaie de prédire les performances à long terme d'un processus à partir d'un échantillon de données limité. Cette approche comporte intrinsèquement un certain degré d'incertitude, étant donné qu'on doit s'appuyer sur les données de l'échantillon.

Les variables aléatoires constituent un outil fondamental pour quantifier et gérer l'incertitude inhérente à divers processus et expériences. Ces variables, qui peuvent être discrètes ou continues, supposent des résultats numériques basés sur le caractère aléatoire des phénomènes observés. Une variable aléatoire décrit les résultats d'une observation statistique ou d'une expérience, et les valeurs d'une variable aléatoire peuvent varier à chaque répétition d'une expérience.

Principaux points à retenir

- Une variable aléatoire discrète est une variable aléatoire avec un ensemble fini de résultats possibles (données ponctuelles).
- Une variable aléatoire continue est une variable aléatoire avec un intervalle de résultats possibles (données continues).

Les variables aléatoires sont le résultat d'une observation ou d'une expérience. Les probabilités sont essentiellement la « meilleure estimation » de l'issue d'un événement aléatoire en vue de prendre une décision. La théorie des probabilités repose sur la prise de décision basée sur la « supposition » la plus éclairée. La nécessité de faire des suppositions éclairées sur des résultats présentant une incertitude inhérente est fréquente dans divers domaines. Par exemple, les politicien.ne.s utilisent les sondages pour estimer leurs chances de gagner une élection, les médecins choisissent des traitements en fonction des résultats attendus, les joueurs choisissent des jeux en fonction des chances perçues de gagner, et les choix de carrière sont souvent influencés par la perception des débouchés. Les probabilités jouent un rôle fondamental dans l'application des statistiques à l'ingénierie, car elles fournissent un cadre permettant de comprendre et d'interpréter les données statistiques. On calcule constamment les probabilités pour ensuite affiner la meilleure « estimation ».

Principaux points à retenir

Les probabilités statistiques fournissent un cadre pour décrire et analyser les phénomènes aléatoires et l'incertitude, et nous donnent la meilleure « estimation » possible.

Objectifs d'apprentissage

Objectifs d'apprentissage du module 3.1 :

- Comprendre les variables aléatoires dans le contexte d'une observation ou d'une expérience statistique.
- Démontrer une compréhension des fréquences relatives à long terme.
- Comprendre les propriétés et la terminologie des probabilités.
- Comprendre les concepts d'événements mutuellement exclusifs et indépendants.
- Appliquer les règles d'addition et de multiplication pour calculer les probabilités d'événements multiples.
- Reconnaître le rôle des statistiques inférentielles dans le domaine plus large des statistiques

3.0.2 Sources de la partie 3

Cette première version de la partie 3 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs. ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte et liés tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de ce travail sont la propriété intellectuelle de Kevin Dunn, et sont partagées à travers la licence CC BY-SA 4.0.

3.1.1 Probabilité d'événements aléatoires

PROBABILITÉ

Les probabilités sont le cadre mathématique concernant les événements d'une activité particulière et les descriptions numériques des chances qu'ils se produisent.. Il s'agit d'une mesure, à laquelle on attribue un nombre compris en 0 et 1 (inclusivement), qui est associée à la certitude des résultats.

Commençons par réviser un peu de terminologie liée aux probabilités :

Terminologie des probabilités :

- Une *expérience* est un processus (une activité particulière, une expérience, un phénomène ou une opération planifiée réalisée dans des conditions contrôlées) qui produit une *observation*.
- Un *résultat* est le résultat *mutuellement exclusif* des observations possibles d'une expérience.
- Par « résultats *mutuellement exclusifs* », on entend qu'un seul des résultats possibles peut être observé.
- L'ensemble de tous les résultats possibles forme ce que l'on appelle un *espace échantillon* (aussi appelé univers).
- Un *événement* est un sous-ensemble de l'espace échantillon.
- Un *essai* est une exécution unique d'une expérience.

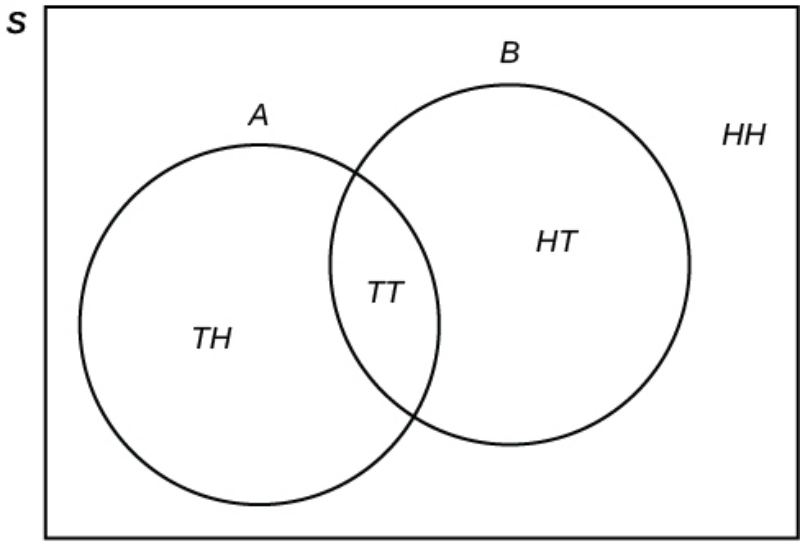
ÉVÉNEMENTS ALÉATOIRES

Le hasard et l'incertitude existent dans toutes les expériences, dans la vie quotidienne et partout dans le monde, ainsi que dans toutes les disciplines de la médecine, de la science et de l'ingénierie. Une expérience aléatoire est une expérience dans laquelle le résultat existe mais n'est ni prédéterminé, ni connu. Un événement aléatoire est donc le sous-ensemble de l'espace échantillon d'une expérience aléatoire. Jouer à pile ou face est un exemple d'expérience aléatoire, puisque le résultat de pile ou face est incertain. L'espace échantillon peut être représenté par une liste de l'ensemble, un diagramme de Venn, un schéma en arbre ou un tableau de contingence. Ces méthodes peuvent être utiles lorsqu'on commence à attribuer et à calculer des probabilités pour des événements multiples.

Nous utiliserons des lettres majuscules pour désigner un ensemble et énumérerons tous les résultats entre crochets. Par exemple, pour définir un espace échantillon de l'expérience aléatoire de type pile ou face: $S = \{H, T\}$ où *P* = pile et *F* = face sont les résultats. L'espace échantillon correspondant à lancer une fois deux pièces de monnaie s'exprime comme suit : S = {(PP),(PF),(FP),(FF)}. Nous utiliserons également des lettres majuscules pour désigner un événement, comme A et B. Par exemple, on peut définir l'événement A comme le fait d'obtenir pile sur la première pièce, et l'événement B, comme le fait d'obtenir pile sur la deuxième pièce. Cela se traduirait par $A = \{TT, TH\}$ et $B = \{TT, HT\}$. Les diagrammes sont utiles pour représenter ensemble les opérations de plusieurs événements.

Diagramme de Venn

Un diagramme de Venn est la représentation visuelle d'un espace échantillon et d'événements sous forme de cercles ou d'ovales montrant leurs intersections. Dans l'exemple ci-dessus, où l'on tire deux fois à pile ou face, on a l'événement A et l'événement B, et le résultat HH n'est ni dans A ni dans B. Le diagramme de Venn est alors représenté comme suit :



Pile ou face avec deux pièces normales : $A = \{PP, PF\}$ *et* $B = \{PP, FP\}$. *Par conséquent,* $A \text{ ET } B$ = *A \text{ INTERSECTION } B =* $A \cap B = \{PP\}$.
$A \text{ OU } B = A \text{ UNION } B = A \cup B = \{PF, PP, FP\}$.

Schéma en arbre

Un schéma en arbre est une représentation d'un espace échantillon et d'événements sous la forme d'un « arbre » dont les branches sont marquées par les résultats possibles.

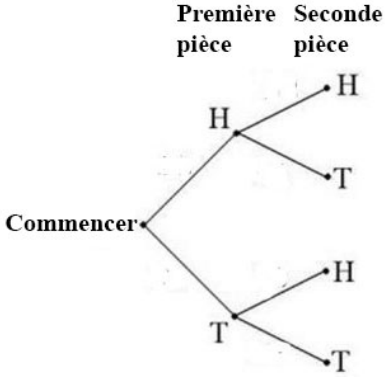


Tableau de contingence

Les tableaux de contingence classent les résultats et les événements. Ces tableaux contiennent des lignes et

des colonnes qui affichent des fréquences à deux variables de données catégoriques. Les événements conjoints se produisent ensemble dans une cellule, comme, pour l'exemple ci-dessus, les événements conjoints de $A = \{TT, TH\}$ et de $B = \{TT, HT\}$ = TT. Les événements marginaux sont ceux qui figurent en marge du tableau et qui se produisent pour un seul événement, sans tenir compte des autres événements du tableau. Dans cet exemple, le tableau contient un événement marginal A auquel sont associés les événements conjoints $A = \{TT, TH\}$.

	2nd coin		
1st coin	Head	Tail	Total
Head	HH	HT	HH,HT
Tail	TH	TT	TH,TT
Total	HH,TH	HT,TT	S

THÉORIE DES ENSEMBLES

Les événements des expériences aléatoires étant des ensembles, nous passerons en revue quelques notions de base de la théorie des ensembles :

Soient A et B des événements d'un espace échantillon.

- Si un élément A appartient à un ensemble B, on l'indique par le symbole $\in$.
- L'ensemble vide est représenté par le symbole $\emptyset$.
- A et B sont disjoints (ou mutuellement exclusifs) si $A \cap B = \emptyset$.
- A est un sous-ensemble de B si chaque élément de A se trouve également dans B. On écrit alors $A \subset B$.

Pour les événements multiples, on peut donc énoncer :

Soient A et B des événements d'un espace échantillon.

- $A \cap B$ est l'ensemble des résultats qui se trouvent à la fois dans A et B.
- $A \cup B$ est l'ensemble des résultats qui ne sont dans A ou dans B ou dans les deux.
- Le complément de A est $\overline{A} = S - A$. Par conséquent, $\overline{A}$ est l'ensemble de résultats qui ne sont pas dans A.

THÉORIE DES PROBABILITÉS

L'utilité des probabilités est d'attribuer des chances d'occurrence raisonnables à des événements possibles dans le cadre d'expériences aléatoires. Avant d'examiner quelques applications pratiques, voyons comment on peut interpréter la théorie des probabilités dans le cadre d'expériences aléatoires.

Les probabilités représentent une quantification du degré de conviction personnelle subjective qu'un événement se produira dans le cadre d'une expérience aléatoire. L'approche subjective la plus courante consiste à utiliser les probabilités bayésiennes, mais cela dépasse le cadre de ce cours. Pour simplifier, on peut considérer que la probabilité est la proportion d'occurrences d'un événement favorable par rapport au nombre total de résultats possibles dans un espace échantillon à probabilité uniforme. Une autre interprétation repose sur la quantification des résultats objectifs d'expériences aléatoires. Cette approche fréquentiste des probabilités est à la base de la plupart des cours d'introduction aux statistiques et d'une grande partie des méthodes statistiques. C'est ce cadre que nous utiliserons pour exploiter le caractère aléatoire des expériences aléatoires.

Les probabilités fréquentistes énoncent que la probabilité d'un événement aléatoire est la fréquence relative de l'événement lorsque l'expérience est répétée indéfiniment. Cette interprétation est souvent énoncée comme étant la fréquence relative d'une expérience « à la longue » ou « à long terme ». Étant donné un événement A dans un espace échantillon, la fréquence relative de A est le rapport $\frac{m}{n}$, où m est le nombre de résultats d'occurrences de l'événement A, et n le nombre total de résultats de l'expérience. L'approche fréquentiste affirme qu'au fur et à mesure que le nombre d'essais augmente, la variation de la fréquence relative diminue. La probabilité est donc la *valeur limite* des fréquences relatives correspondantes. On peut déterminer la fréquence relative soit en réalisant des expériences réelles et en trouvant une probabilité empirique (ou estimée), soit en reconnaissant le modèle théorique de l'expérience et en adoptant une probabilité théorique basée sur les événements de l'espace échantillon.

Dans le cas d'un espace échantillon où les résultats sont équiprobables :

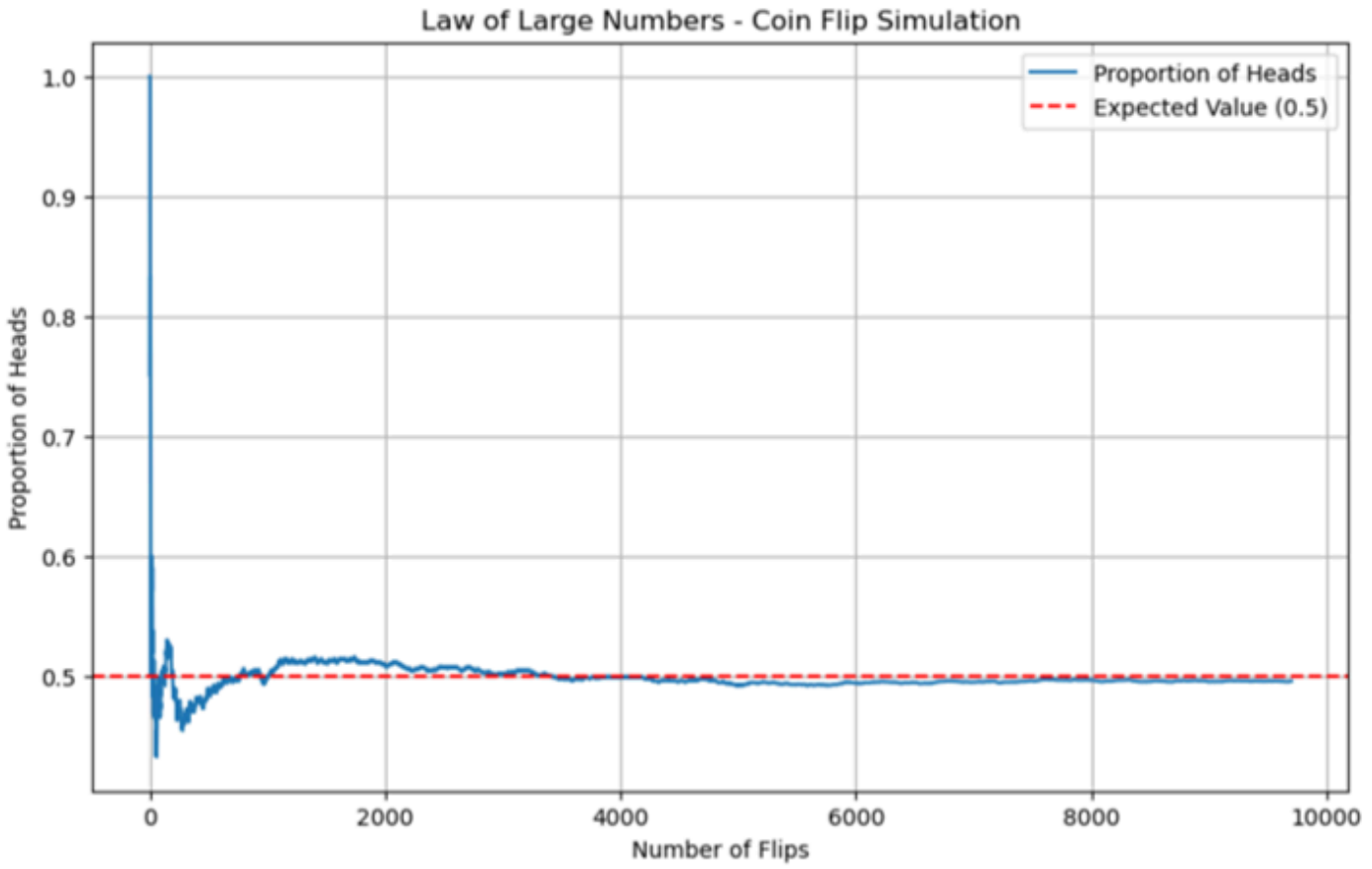
Si les résultats d'un espace échantillon fini S ont tous la même probabilité, alors pour tout événement A :

$$P[A] = \frac{\text{le nombre de résultats pour l'événement A}}{\text{le nombre de résultats dans S}}$$

Le terme équiprobable signifie que les résultats d'une expérience ont tous la même probabilité de se produire. Par exemple, si on tire à pile ou face avec une pièce de monnaie, il est tout aussi probable d'obtenir pile (*P*) que face (*F*). On peut donc compter le nombre de résultats pour l'événement A = obtenir une fois face, et le diviser par le nombre total de résultats dans l'espace échantillon. Si on lance deux pièces, l'espace échantillon est {*PP*, *PF*, *FP*, *FF*}. Deux résultats répondent à cette condition, soient {*PF*, *FP*}, donc $P(A) = \frac{2}{4} = 0{,}5$.

Ce texte utilisera la convention de notation selon laquelle un P majuscule suivi d'une expression ou d'une phrase entre crochets signifie « la probabilité » de cette expression. P(A) est donc la probabilité du résultat A.

À long terme, la fréquence relative du tirage à pile ou face s'approchera de la probabilité théorique de 0,5. Comme il n'y a que deux résultats possibles, cette probabilité empirique de succès, en tant que fréquence relative expérimentale, convergera vers la probabilité théorique. La loi des grands nombres dit qu'à mesure que le nombre d'essais augmente, les valeurs de l'échantillon tendent à converger vers le résultat attendu. Autrement dit, la proportion de faces dans un « grand » nombre de tirages à pile ou face « devrait être » d'environ 0,5. Plus précisément, la proportion de faces après *n* lancés convergera vers 0,5 quand *n* tend vers l'infini. Même si les résultats ne suivent pas de modèle prédéterminé, dans l'ensemble, la fréquence relative observée à long terme s'approchera de la probabilité théorique.



Voir le Jupyter Notebook dans le référentiel GitHub pour une simulation de cette démonstration de jeu de pile ou face à long terme : CoinTossSimulation.

3.1.2 Probabilité et indépendance des événements

PROBABILITÉ D'ÉVÉNEMENTS ALÉATOIRES

L'objectif de la théorie des probabilité est d'attribuer un nombre entre 0 et 1 pour mesurer la probabilité d'un événement aléatoire. Par exemple, si l'expérience consiste à tirer à pile ou face, on peut considérer l'événement A comme le fait d'obtenir au maximum une fois face. La *probabilité de l'événement* A s'écrit $P(A)$ et correspond à un nombre compris entre zéro et un (inclusivement) qui décrit la proportion de fois où l'on s'attend à ce que l'événement se produise sur le long terme. $P(A) = 0$ signifie que l'événement A ne peut jamais se produire. $P(A) = 1$ signifie que l'événement A se produit toujours. $P(A) = 0.5$ signifie que l'événement A a la même probabilité de se produire ou de ne pas se produire. Par exemple, si vous tirez à pile ou face de manière répétée (20, 2 000 ou 20 000 fois), la fréquence relative de faces tends vers 0,5 (soit la probabilité d'obtenir face).

Passons en revue les axiomes de probabilité qui serviront à construire les règles de probabilité que nous utiliserons dans ce cours :

Un système de probabilités est une affectation de nombres (probabilités) $P(A)$, à des événements A de telle sorte que :

- Pour chaque événement A, $P(A)$ est un nombre réel non négatif compris entre 0 et 1 inclusivement. Autrement dit, $0 \leq P(A) \leq 1$.
- La probabilité de l'espace échantillon S est de 1 et la probabilité de l'ensemble vide est de 0. Autrement dit, $P(S) = 1$ et $P(\varnothing) = 0$.
- Les probabilités sont dénombrablement additives pour des événements disjoints. Autrement dit, $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

PROBABILITÉ CONDITIONNELLE ET INDÉPENDANCE DES ÉVÉNEMENTS

L'idée d'attribuer des probabilités à un événement en fonction de la valeur d'un autre événement est un concept essentiel pour comprendre pour les statistiques. Attribution conditionnelle de probabilités d'événements :

Pour l'événement A et l'événement B, si l'événement B a une probabilité non nulle, la probabilité conditionnelle de A étant donné B est

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$

« $P(A|B)$ » signifie « probabilité de A, étant donné B ».

Souvent, l'événement A et l'événement B sont dépendants, ce qui veut dire que les probabilités conditionnelles s'appliquent et que les valeurs numériques de P(A|B) et P(A) sont différentes. Concrètement, cela signifie qu'on peut changer la probabilité de A si on sait que B s'est produit. Dans les cas où il n'y a pas de différence, on parle d'indépendance.

Deux événements A et B sont **indépendants** si le fait de savoir que l'un s'est produit n'affecte pas les chances que l'autre se produise. Par exemple, les résultats de deux lancers d'un dé standard sont des événements indépendants. Le résultat du premier jet ne modifie pas la probabilité du résultat du deuxième jet. Deux événements sont indépendants si l'un des éléments suivants est vrai :

Si A et B sont des événements avec une probabilité non-nulle dans l'espace échantillon S, et qu'ils sont indépendants, les identités suivantes s'appliquent :

- $P(A \cap B) = P(A)P(B)$.
- $P(A|B) = P(A)$.
- $P(B|A) = P(B)$.

Les probabilités des événements obéissent à des règles qui découlent de l'application des axiomes des probabilités et de l'application de l'indépendance, et peuvent être illustrées comme suit :

Pour A et B des événements dans un espace échantillon S :

- Pour tout événement A, P(A) = 1 − P($\overline{A}$).
- La règle additive énonce que pour toute paire d'événements A et B : P (A ou B) = P (A) + P(B) − P(A et B)
- Pour des événements *disjoints* (mutuellement exclusifs) A et B, la règle additive se simplifie : P(A ou B) = P(A) + P(B)
- La règle de multiplication énonce que pour P(B) > 0, P(A et B) = P(A | B) · P(B).
- Pour des événements *indépendants* A et B, la règle de multiplication se simplifie : P(A et B) = P(A) · P(B).

On peut maintenant étendre la définition de l'indépendance à l'indépendance mutuelle de plusieurs événements. L'indépendance de plus de deux événements élargit la notion d'indépendance : le fait de savoir quelque chose sur certains de ces événements ne donne aucune information probabiliste sur les autres. L'indépendance mutuelle s'étend à toutes les collections d'événements dans l'espace échantillon. La notion d'indépendance mutuelle sera très importante pour l'attribution de probabilités aux événements des expériences aléatoires.

Les événements $A_1, A_2, \ldots, A_n \subset S$ sont mutuellement indépendants si pour tout sous-ensemble $A_{i_1}, \ldots, A_{i_k}$:

- $P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \ldots \cdot P(A_{i_k})$

ÉCHANTILLONNAGE ALÉATOIRE ET INDÉPENDANCE

L'échantillonnage peut être effectué **avec remplacement** ou **sans remplacement**.

- **Avec remplacement :** Chaque membre d'une population est remplacé après avoir été sélectionné, et il a la possibilité d'être sélectionné plus d'une fois. Lorsque l'échantillonnage est effectué avec remplacement, les événements sont considérés comme indépendants, ce qui signifie que le résultat de la première sélection ne modifiera pas les probabilités de la deuxième sélection.

- **Sans remplacement :** Lorsque l'échantillonnage est effectué sans remplacement, chaque membre d'une population ne peut être sélectionné qu'une seule fois. Dans ce cas, les probabilités de la deuxième sélection sont affectées par le résultat de la première sélection. Les événements sont considérés comme dépendants (ou non indépendants).

3.1.2.1. Échantillonnage à partir d'un jeu de cartes bien mélangé

Soit un jeu de 52 cartes bien mélangées. Il se compose de quatre couleurs : le trèfle, le carreau, le cœur et le pique. Chaque couleur comporte 13 cartes :$1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, J (valet), Q (dame), K (roi) de cette couleur.

1. **Échantillonnage avec remplacement :**Supposons qu'on sélectionne trois cartes avec remplacement. La première carte choisie sur les 52 cartes est la Q de pique. On remet la carte, on mélange le paquet et on pige une deuxième carte du jeu de 52 cartes. Il s'agit du 10 de trèfle. On remet la carte, on mélange le paquet et on pige une troisième carte du jeu de 52 cartes. Cette fois, la carte est à nouveau la Q de pique. Les sélections sont {Q de pique, 10 de trèfle, Q de pique}. La Q de pique a été pigée deux fois. Chaque sélection est faite parmi le jeu complet (52 cartes).
2. **Échantillonnage sans remplacement :**Supposons qu'on sélectionne trois cartes sans remplacement. La première carte choisie sur les 52 cartes est le K de cœur. On met cette carte de côté, puis on choisit une deuxième carte parmi les 51 cartes restantes du jeu. Il s'agit du 3 de carreau. On met cette carte de côté, puis on choisit une troisième carte parmi dans les 50 cartes restantes du jeu. La troisième carte est un J de pique. Les sélections sont {K de cœur, 3 de carreau, J de pique}. Comme les cartes ont été pigées sans remplacement, on ne peut pas piger deux fois la même carte.

3.1.3 Variables aléatoires et distributions de probabilités

CARACTÈRE ALÉATOIRE ET VARIATION

Nous avons vu que la nature aléatoire représente l'élément fondamental du hasard, comme dans le cas du jeu de pile ou face, mais elle peut également représenter l'incertitude, comme dans le cas d'une erreur de mesure. Après avoir introduit le concept d'événements et d'expériences aléatoires dans le chapitre précédent, considérons maintenant qu'une expérience consiste à prendre une mesure numérique issue d'une expérience d'ingénierie. Les données de mesures comportent généralement une part de hasard et sont soumises à des influences fortuites. Dans l'échantillonnage statistique et les études de fréquence, le hasard est introduit par les techniques d'échantillonnage. Le hasard est également introduit par l'erreur de mesure. Il peut y avoir d'autres sources de hasard, dont les nombreuses petites causes non identifiées qui influent sur la mesure du phénomène aléatoire. Dans les contextes analytiques, les changements dans les conditions du système font varier les réponses mesurées, ce qui est le plus souvent attribué au hasard.

Quel que soit le soin apporté à la conception et à la réalisation d'une expérience, des variations se produisent souvent en raison de ces phénomènes fortuits. L'objectif est donc de comprendre, de quantifier et de modéliser la variation, puis de l'exploiter dans nos analyses afin de tirer des conclusions basées sur les données qui restent valides malgré cette variation.

VARIABLES ALÉATOIRES ET DISTRIBUTION DES PROBABILITÉS

Une **variable aléatoire** est une formalisation mathématique, ou fonction, d'un événement qui dépend d'une expérience aléatoire sous-jacente. Il s'agit d'une variable associée à une variable réelle qui attribue une valeur numérique à chaque résultat possible de l'expérience.

Dans la plupart des cas, une variable aléatoire X est une fonction faisant correspondre un espace échantillon (un espace de mesure de la probabilité) à des nombres réels (un espace mesurable) :

$$X : S \to \mathbb{R}$$

Ainsi, il est possible de créer la distribution mathématique d'une variable aléatoire conforme aux axiomes de probabilité. Cette distribution fournit la mesure de probabilité $[0, 1] \subseteq \mathbb{R}$ associée à chacune des valeurs possibles de la variable aléatoire. Les variable aléatoires sont exprimées sous la forme de lettres latines majuscules, souvent celles de la fin de l'alphabet, comme X, Y, Z, T.

Pour l'exemple simple de pile ou face, on utilise une fonction qui fait correspondre aux valeurs de l'espace échantillon de $S = \{H, T\}$ une valeur mesurable de l'espace $\{-1, 1\}$, où 1 correspond à P et -1 correspond à F, en utilisant la variable aléatoire X pour représenter la mesure aléatoire de l'expérience.

Une fois défini l'espace échantillon de S par la variable aléatoire correspondante X, il est désormais possible de se poser la question : « Quelle est la probabilité que la valeur de X soit égale à +1? ». C'est la probabilité de l'événement = E = x = +1, qu'on note $P(X = 1)$.

Consigner toutes les probabilités des sorties d'une variable aléatoire X permet d'obtenir la distribution de probabilité de X. Une **distribution de probabilité** est la fonction mathématique qui définit les probabilités que

l'événement (le sous-ensemble défini de l'espace échantillon) se produise. Elle définit donc l'expérience aléatoire en termes de l'événement.

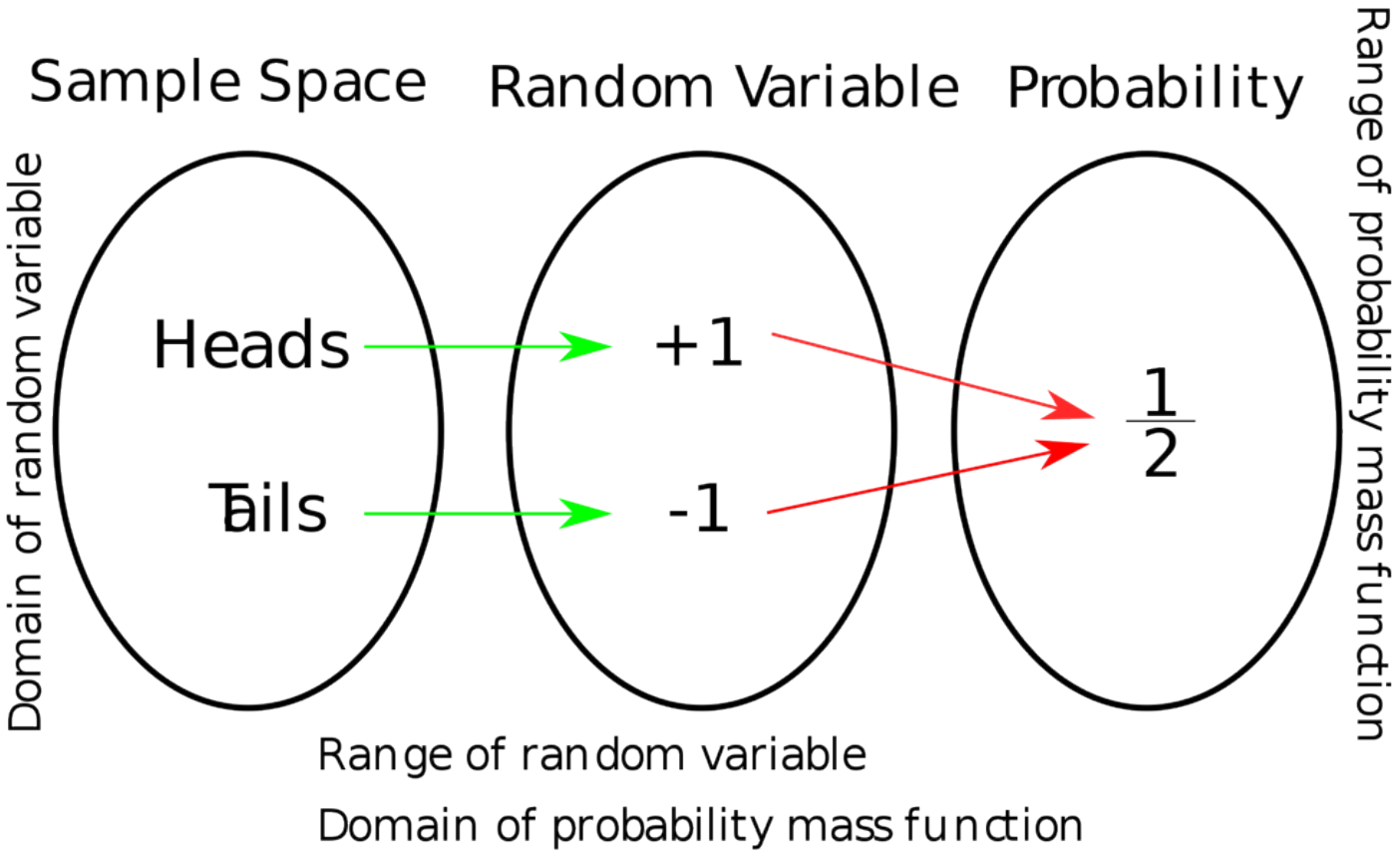


Figure 3.1.3.1. Une variable aléatoire est une fonction qui relie tous les résultats possibles d'une expérience aléatoire à des valeurs réelles. Cette figure montre comment le résultat d'un tirage à pile ou face est représenté sous la forme d'une variable aléatoire discrète utilisée pour définir une fonction de masse de probabilité.

Pour l'exemple du pile ou face, si X est la variable aléatoire utilisée pour définir le résultat aléatoire de l'expérience, la distribution de probabilité de X prend alors la valeur 0,5 (ou 1/2) pour X = Pile, et 0,5 pour X = Face.

Principaux points à retenir

Révision des termes relatifs aux variables aléatoires et aux distributions de probabilités :

- Variable aléatoire : à partir de valeurs d'un espace échantillon, attribue des probabilités en fonction de la probabilité de l'événement expérimental.

- Événement : ensemble des valeurs possibles (résultats) d'une variable aléatoire qui se produit avec une certaine probabilité,
- Distribution de probabilité : fonction qui fournit la probabilité d'occurrence des événements pour l'expérience, ou $P(X \in E)$ pour un événement.

3.1.4 Fonctions de distribution cumulative

FONCTION DE DISTRIBUTION CUMULATIVE

Les distributions de probabilité peuvent être définies de différentes manières selon la description de la variable aléatoire utilisée, mais elles peuvent toujours être définies par une fonction de distribution cumulative (FDC; aussi appelée « fonction de répartition »). Cette fonction décrit la probabilité que la variable aléatoire ne dépasse pas une valeur donnée – autrement dit, $P(X \leq x)$.

Chaque distribution de probabilité reposant sur des valeurs réelles est définie par une fonction continue à droite et non décroissante $F: \mathbb{R} \to [0,1]$ telle que $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to \infty} F(x) = 1$. Toute fonction possédant ces quatre propriétés est une FDC : pour chaque fonction de ce type, on peut définir une variable aléatoire qui a cette fonction pour fonction de distribution cumulative.

Définition 3.1.4.1. Fonction de distribution cumulative (FDC)

La fonction de probabilité cumulative d'une variable aléatoire X est une fonction F(x) qui, pour chaque nombre x, donne la probabilité que X prenne cette valeur ou une valeur plus petite. Symboliquement :

$$F(x) = P[X \leq x]$$

3.1.5 Variables aléatoires discrètes et variables aléatoires continues

VARIABLES ALÉATOIRES DISCRÈTES

Nous avons déjà fait la distinction entre les données discrètes et continues au module 1, lorsque nous avons exploré les données et les statistiques descriptives. Cette terminologie s'applique au contexte actuel et inspire deux autres définitions.

Il existe deux types de variables aléatoires :

- Une variable aléatoire discrète est une variable qui ne peut prendre que certaines valeurs isolées (plutôt qu'un continuum de valeurs).
- Une variable aléatoire continue est une variable qui peut être idéalisée comme pouvant prendre n'importe laquelle des valeurs d'un intervalle continu.

Les variables aléatoires qui sont essentiellement des variables de dénombrement relèvent clairement de la première définition et sont discrètes. On pourrait soutenir que toutes les variables de mesure sont discrètes, puisque toutes les mesures sont effectuées « à l'unité près », mais pour des raisons pratiques, nous continuerons à utiliser les définitions des types de données et traiterons les valeurs numériques comme des valeurs continues. Nous étudierons les distributions de probabilités continues dans le module suivant.

Rappelez-vous que nous utilisons la convention de notation selon laquelle un P majuscule suivi d'une expression ou d'une phrase entre crochets signifie « la probabilité » de cette expression. Dans cette notation, une fonction de probabilité pour X, le résultat d'un tirage à pile ou face, qui, selon notre définition, est une variable aléatoire discrète, est une fonction f telle que

$$f(x) = P[X = x]$$

Autrement dit, « *f (x)* est la probabilité que (la variable aléatoire) *X* prend la valeur *x* » , soit = 0,5 dans le cas où x = Pile ou x = Face.

3.1.6 Synthèse des modèles de probabilité

MODÈLES DE PROBABILITÉS

Comme nous l'avons vu précédemment, les variables aléatoires constituent un outil fondamental pour quantifier et gérer l'incertitude inhérente à divers processus et expériences. Les probabilités d'une variable aléatoire sont généralement déterminées à partir d'un modèle qui décrit l'expérience aléatoire. Les concepts-clé à comprendre sont l'espérance mathématique, la variance et l'écart-type, qui représentent respectivement le résultat moyen, la variation et la mesure de dispersion des valeurs potentielles d'une variable aléatoire. Ces paramètres forment la distribution de probabilité d'une variable aléatoire et constituent une description des probabilités associées aux valeurs possibles de la variable aléatoire. Ces distributions de probabilités sont essentielles en ingénierie pour modéliser, prédire et contrôler le comportement des systèmes, ce qui permet de prendre des décisions éclairées dans des conditions d'incertitude et de risque.

Principaux points à retenir

La distribution de probabilité d'une variable aléatoire est une description des probabilités associées aux valeurs possibles de cette variable aléatoire.

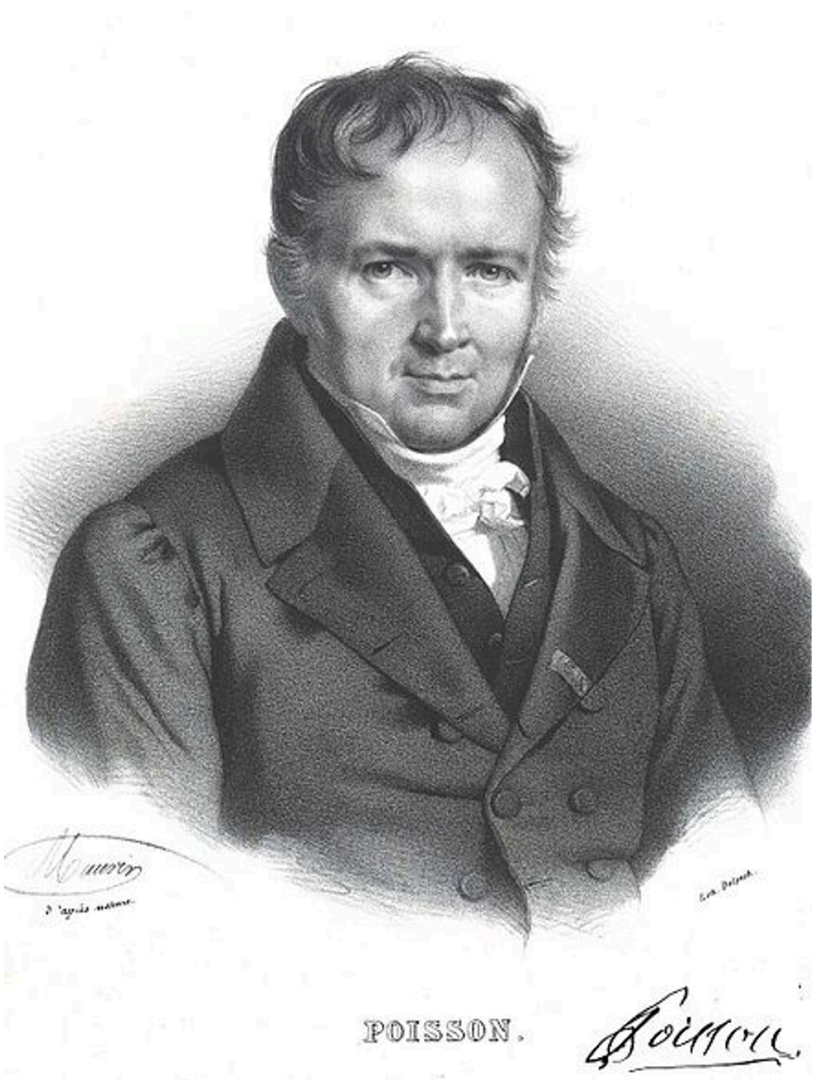
Une distribution de probabilité est une description mathématique des probabilités d'événements (les sous-ensembles de résultats possibles de l'expérience). En termes simples, une fonction de distribution de probabilité est un modèle théorique qu'on essaie de définir pour trouver la meilleure estimation des probabilités.

Principaux points à retenir

- Les distributions de probabilité sont des outils ou des modèles théoriques qui facilitent la résolution des problèmes de probabilité.

Ces distributions de probabilités sont des outils ou des modèles théoriques qui facilitent la résolution des problèmes de probabilités. Chaque distribution a ses propres suppositions, caractéristiques et paramètres. Apprendre à les reconnaître permet de distinguer les différentes distributions et de choisir le meilleur modèle à utiliser. En reconnaissant la distribution de probabilité d'une variable aléatoire identifiée, il est possible de caractériser et d'exploiter le hasard et la variabilité afin de déterminer la probabilité que tel ou tel événement se produise. Ces outils permettent d'évaluer au mieux les résultats expérimentaux futurs et inconnus en choisissant l'événement le plus probable. Cette « meilleure estimation » permet à son tour de formuler des prédictions basées sur le choix d'un modèle et l'analyse d'un échantillon de données.

3.2.0 Introduction aux distributions de probabilités discrètes



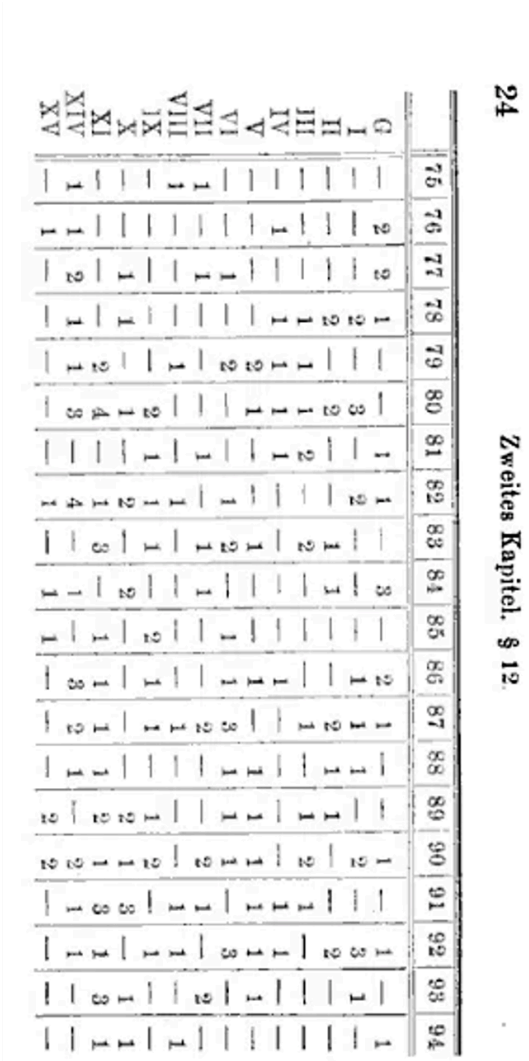


Figure 3.2.0.1. Siméon Poisson : François-Séraphin Delpech, domaine public, via Wikimedia Commons https://upload.wikimedia.org/wikipedia/commons/0/0d/Sim%C3%A9onDenisPoisson.jpg. Ladislaus von Bortkiewicz, Das Gesetz der kleinen Zahlen [The law of small numbers] (Leipzig, Allemagne : B.G. Teubner, 1898). Bortkiewicz présente la distribution de Poisson, p. 23–2

La distribution de Poisson, qui doit son nom au mathématicien français Siméon Denis Poisson (1781-1840; figure 3.2.0.1), est une distribution de probabilité discrète qui exprime la probabilité qu'un nombre donné d'événements se produisent dans un intervalle de temps ou d'espace fixe, en supposant que ces événements se produisent avec un taux moyen constant connu et indépendamment du temps écoulé depuis le dernier événement. Une application historique célèbre de la distribution de Poisson est son utilisation pour analyser l'incidence des décès dus aux coups de sabot dans la cavalerie prussienne. Cet exemple est souvent cité pour illustrer la puissance et l'utilité de la distribution de Poisson dans la modélisation d'événements rares et aléatoires dans divers domaines.

Pour voir comment la distribution de Poisson a été utilisée afin de modéliser la mortalité due aux coups de sabot dans la cavalerie prussienne, consulter l'activité Jupyter Notebook sur GitHub : Poisson Distribution and the Prussian Cavalry.

VARIABLES ALÉATOIRES DISCRÈTES

Comme nous l'avons vu, pour une variable aléatoire discrète, il suffit de spécifier une fonction de masse de probabilité attribuant une probabilité à chaque résultat ou événement possible. Pour la fonction de masse de probabilité, on définit une fonction de distribution cumulative évaluant la probabilité que la variable aléatoire prenne une valeur inférieure ou égale à une valeur donnée.

Principaux points à retenir

- Pour les variables aléatoires discrètes, une fonction de masse de probabilité définit la probabilité d'un événement à partir d'une expérience aléatoire.

Ces distributions de probabilités sont des outils ou des modèles théoriques qui facilitent la résolution des problèmes de probabilités. Chaque distribution a ses propres suppositions, caractéristiques et paramètres. Apprendre à les reconnaître permet de distinguer les différentes distributions et de choisir le meilleur modèle à utiliser. Parmi les fonctions de probabilité discrètes les plus courantes, citons les fonctions binomiale, géométrique, hypergéométrique et de Poisson.

Objectifs d'apprentissage

Objectifs d'apprentissage du module 3.2 :

- Reconnaître les variables aléatoires discrètes et les appliquer aux probabilités empiriques et théoriques.
- Reconnaître et comprendre les fonctions de distribution de probabilité discrètes et leurs suppositions.
- Calculer et interpréter l'espérance mathématique et les paramètres de distribution de la fonction de masse de probabilité.
- Comprendre la fonction de distribution cumulative et l'appliquer aux calculs.
- Reconnaître la distribution de probabilité binomiale et l'appliquer de manière appropriée.
- Reconnaître la distribution de probabilité de Poisson et l'appliquer de manière appropriée.

- Reconnaître la distribution géométrique des probabilités et l'appliquer de manière appropriée.
- Reconnaître la distribution de probabilité hypergéométrique et l'appliquer de manière appropriée.

3.2.1 Fonction de masse de probabilité d'une variable aléatoire discrète

VARIABLE ALÉATOIRE DISCRÈTE

Revoyons la définition d'une **variable aléatoire discrète** vue au module précédent :

> Une variable aléatoire discrète est une variable qui ne peut prendre que certaines valeurs isolées (plutôt qu'un continuum de valeurs).

En outre, une variable aléatoire est imprévisible, et sa valeur n'est pas connue avant une expérience aléatoire. Par conséquent, décrire ou modéliser la variable aléatoire consiste à énumérer toutes les valeurs potentielles et les probabilités qui leur sont associées.

> **DÉFINITION 3.2.1.1. Distribution de probabilité**
> La spécification d'une distribution de probabilités d'une variable aléatoire consiste à donner l'ensemble des valeurs possibles et à, d'une manière ou d'une autre, attribuer de manière cohérente des chiffres compris entre 0 et 1 – appelés probabilités – qui correspondent à la probabilité que telle ou telle valeur se réalise.

L'outil le plus souvent utilisé pour décrire une distribution de probabilité discrète est la fonction de masse de probabilité.

> **DÉFINITION 3.2.1.2. Fonction de masse de probabilité**
> La fonction de probabilité d'une variable aléatoire discrète X pouvant prendre les valeurs x_1,x_2,..., est une fonction non-négative f(x) telle que f (x_i) donne la probabilité que X prenne la valeur x_i.

Rappel : P(X) ou P[X] signifie « la probabilité de [l'expression ou la phrase X] ». Par conséquent, la fonction de probabilité (fonction de masse de probabilité) de X est la fonction f telle que :

f(x) = P[X = x]

Autrement dit, « f(x) est la probabilité que (la variable aléatoire) X prenne la valeur x. »

Exemple 3.2.1.1. Retour sur le couple des boulons

Variable aléatoire de l'exigence de couple

Reprenons l'exemple du chapitre 2, où Brenny, Christensen et Schneider ont mesuré le couples des boulons de la plaque avant d'un composant d'équipement lourd.

Soit Z le prochain couple mesuré pour le boulon 3 (arrondi à l'entier le plus proche), que nous traiterons comme une variable aléatoire discrète. Il faut maintenant trouver une fonction de probabilité plausible pour Z. Les fréquences relatives des mesures de couple enregistrées sur le boulon 3 permettent d'établir la distribution des fréquences relative :

Distribution de la fréquence relative pour le couple mesuré des boulons 3

z, Couple (pi lb)	Fréquence	Fréquence relative
11	1	1/34 ≈ 0,02941
12	1	1/34 ≈ 0,02941
13	1	1/34 ≈ 0,02941
14	2	2/34 ≈ 0,05882
15	9	9/34 ≈ 0,26471
16	3	3/34 ≈ 0,08824
17	4	4/34 ≈ 0,11765
18	7	7/34 ≈ 0,20588
19	5	5/34 ≈ 0,14706
20	1	1/34 ≈ 0,02941
	34	1

Tableau 3.2.1.1.

Ce tableau montre, par exemple, que pendant la période de collecte des données, environ 15 % des couples mesurés étaient de 19 pi lb. S'il est raisonnable de croire que le système qui a produit les données de ce tableau produira le couple du prochain boulon 3, il est alors logique d'établir la fonction de probabilité de Z selon les fréquences relatives de ce tableau.

Fonction de probabilité pour Z

Couple z	Probabilité $f(z)$
11	0,03
12	0,03
13	0,03
14	0,06
15	0,26
16	0,09
17	0,12
18	0,20
19	0,15
20	0,03

Tableau 3.2.1.2.

Autrement dit, on peut utiliser la distribution de probabilité spécifiée dans le prochain tableau. (En passant des fréquences relatives du premier tableau aux valeurs proposées pour f(z) dans le second tableau, les chiffres ont été arrondis de manière légèrement arbitraire pour que les valeurs de probabilité soient exprimées avec deux décimales et que leur total soit d'exactement 1,00.)

Distribution de masse de probabilité d'une valeur donnée sélectionnée aléatoirement dans une population

L'adéquation de la fonction de probabilité du tableau ci-dessus pour décrire Z dépend essentiellement de la stabilité physique du processus de serrage des boulons. Mais il y a une deuxième façon dont les fréquences relatives peuvent devenir des choix évidents pour les probabilités. Par exemple, considérons les 34 couples du tableau 3.2.1.1 comme une population, dans laquelle n = 1 élément doit être échantillonné de manière aléatoire, et soit Y la valeur de couple sélectionnée.

Dans ce cas, la fonction de probabilité du tableau 3.2.1.2 est également approximativement appropriée pour Y. Ce point n'est pas aussi important dans cet exemple spécifique qu'il l'est en général : lorsqu'il faut choisir une valeur hasard dans une population, une distribution de probabilité appropriée est une distribution équivalente à la distribution de fréquence relative de la population.

Principaux points à retenir

La **distribution de probabilité** d'une variable aléatoire répertorie toutes les valeurs possible de la variable aléatoire et la probabilité que la variable prenne chaque valeur. Elle décrit la manière dont les probabilités sont distribuées sur les valeurs de la variable aléatoire. S'il faut choisir une valeur au hasard dans une population, une distribution de probabilité appropriée est une distribution équivalente à la distribution de fréquence relative de la population.

Propriétés des fonctions de probabilité mathématiquement valides

La fonction de probabilité présentée dans le tableau 3.2.1.2 possède deux propriétés nécessaires à la cohérence mathématique d'une distribution de probabilité discrète. Les valeurs de f(z) sont toutes comprises dans l'intervalle [0, 1], et leur somme est égale à 1. Les probabilités négatives ou supérieures à 1 n'auraient aucun sens, étant donné qu'une probabilité de 1 indique une certitude d'occurrence, et une probabilité de 0 indique une impossibilité. Ainsi, selon le modèle du tableau 3.2.1.2, puisque la somme des valeurs de f(z) est égale à 1, l'occurrence de l'une des valeurs {11, 12, 13, 14, 15, 16, 17, 18, 19, 20} ft lb est certaine.

Une fonction de probabilité f(x) donne la probabilité d'occurrence de chacune des valeurs. Si on veut trouver la probabilité que X prenne n'importe laquelle des valeurs dans un sous-ensemble de ses valeurs possibles, on additionne les probabilités des valeurs du sous-ensemble.

Exemple 3.2.1.2. Retour sur le couple des boulons (suite)

Reprenons la fonction f(z) définie dans le tableau 3.2.1.2 pour trouver

P[Z > 17] = P [le couple suivant dépasse 17]

La somme des entrées de f (z) correspondant aux valeurs possibles supérieures à 17 ft lb donne

P[Z > 17] = f(18) + f(19) + f(20) = 0,20 + 0,15 + 0,03 = 0,38

La probabilité que le couple suivant soit supérieur à 17 ft lb est d'environ 38%.

Si, par exemple, les spécifications indiquent que le couple doit être compris entre 16 et 21 ft lb, la probabilité que le prochain couple mesuré soit conforme est :

P[16 ≤ Z ≤ 21] = f(16) + f(17) + f(18) + f(19) + f(20) + f(21)

= 0,09 + 0,12 + 0,20 + 0,15 + 0,03 + 0,00

= 0,59

Dans l'exemple de la mesure du couple, la fonction de probabilité est donnée sous forme de tableau. Dans d'autres cas, il est possible de trouver une équation pour f(x).

Exemple 3.2.1.3. Numéro de série d'outil aléatoire

La dernière étape du processus d'assemblage des outils pneumatiques étudié par Kraber, Rucker et Williams consistait à apposer une plaque de numéro de série sur l'outil terminé. Imaginez que vous vous rendiez à la fin de la chaîne de montage à 9 h 00 lundi prochain et que vous observiez le premier numéro de série apposé après 9 h 00.

Soit

W = le dernier chiffre du numéro de série observé

Supposons que les numéros de série des outils commencent par un code spécifique au modèle d'outil et se terminent par des numéros consécutifs reflétant le nombre d'outils du modèle en question qui ont été produits. La symétrie de cette situation suggère que les valeurs de W (w = 0, 1,...,9) sont équiprobables. Autrement dit, une fonction de probabilité plausible pour W prend la forme suivante :

$$f(w) = \begin{cases} 0{,}1 & \text{pour } w = 0, 1, 2, \ldots, 9 \\ 0 & \text{sinon} \end{cases}$$

3.2.2 Fonction de distribution cumulative

FONCTION DE DISTRIBUTION CUMULATIVE

Il existe une autre façon de spécifier une distribution de probabilité discrète : la fonction de distribution cumulative (FDC, aussi appelée fonction de probabilité cumulative).

Rappelez-vous la définition d'une FDC.

$F(x) = P[X \leq x]$

Puisque (pour les distributions discrètes) les probabilités sont calculées en additionnant les valeurs de f(x), pour une distribution discrète,

DÉFINITION 3.2.2.1. Fonction de distribution cumulative d'une variable discrète X

F(x)=\sum_{z \leq x}^{ } f(z)

La somme est calculée sur les valeurs possibles inférieures ou égales à x. Dans ce cas discret, le graphique de F(x) sera un diagramme en marches d'escalier avec des sauts situés aux valeurs possibles et de taille égale aux probabilités associées à ces valeurs.

Exemple 3.2.2.1. Retour sur le couple des boulons

Suite de l'exemple des variables de couple de la section 3.2.1

Les valeurs de la fonction de probabilité et de la fonction de probabilité cumulative pour la variable de couple Z sont indiquées dans le tableau 3.2.1.1. Les valeurs de F(z) pour d'autres z sont également faciles à obtenir. Par exemple,

F (10,7) = P [Z ≤ 10,7] = 0

$$F(16{,}3) = P[Z \leq 16{,}3] = P[Z \leq 16] = F(16) = 0{,}50$$

$$F(32) = P[Z \leq 32] = 1.00$$

La figure 3.2.2.1 présente un diagramme de la fonction de probabilité cumulative de Z. On y voit la forme en escalier caractéristique des fonctions de probabilité cumulative des distributions discrètes.

Valeurs de la fonction de probabilité et de la fonction de probabilité cumulative pour Z

z, couple	$f(z) = P[Z = z]$	$F(z) = P[Z \leq z]$
11	0,03	0,03
12	0,03	0,06
13	0,03	0,09
14	0,06	0,15
15	0,26	0,41
16	0,09	0,50
17	0,12	0,62
18	0,20	0,82
19	0,15	0,97
20	0,03	1,00

Tableau 3.2.2.1. *Valeurs de la fonction de probabilité et de la fonction de probabilité cumulative de Z*

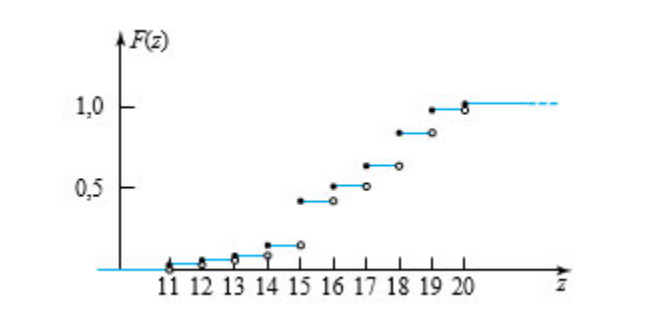


Figure 3.2.2.1. *Graphique de la fonction de probabilité de Z.*

Les informations sur une distribution discrète véhiculée par sa fonction de probabilité cumulative sont équivalentes à celles véhiculées par la fonction de probabilité correspondante. La version cumulative est parfois préférable pour les tableaux, parce que les problèmes d'arrondis sont plus graves lorsqu'on additionne plusieurs termes f(x) que lorsqu'on prend la différence de deux valeurs de F(x) pour obtenir une probabilité associée à une séquence consécutive de valeurs possibles, et parce qu'elle est plus facile à comprendre.

3.2.3 Probabilité exprimée avec deux décimales

EXPRESSION DES PROBABILITÉS

Les probabilités sont généralement exprimées avec deux décimales, comme dans le tableau 3.2.1.2. On peut les calculer à beaucoup plus de décimales, mais les probabilités finales ne sont généralement indiquées qu'avec deux décimales. En effet, les chiffres exprimés à plus de deux décimales ont tendance à paraître trop impressionnants et à être pris trop au sérieux par les non-initiés. Prenons l'exemple de la déclaration suivante: « Il y a une probabilité de 0,097328 pour que le moteur d'appoint tombe en panne » lors du lancement d'un missile donné. Cette valeur peut représenter le résultat de manipulations mathématiques très minutieuses et être correcte à six décimales près dans le contexte du modèle mathématique utilisé pour l'obtenir. Mais il est peu probable que le modèle utilisé soit une description suffisamment correcte de la réalité physique pour justifier une telle précision apparente. La précision à deux décimales est à peu près ce qui est justifié dans la plupart des applications techniques des probabilités simples.

3.2.4 Moyenne ou espérance mathématique et écart-type de distributions de probabilités discrètes

RÉSUMÉ DES DISTRIBUTIONS DE PROBABILITÉS DISCRÈTES

Presque tous les outils utilisés pour décrire les distributions de fréquences relatives (empiriques) dans les modules 1 et 2 sur l'exploration, la synthèse et la visualisation des données ont des versions qui peuvent décrire des distributions (théoriques) de probabilités.

Pour une variable aléatoire discrète dont les valeurs possibles sont espacées également, l'histogramme de probabilité donne une image de la forme de la distribution de la variable. On produit cet histogramme en centrant une barre de hauteur f(x) sur chaque x possible. Les histogrammes de probabilité des variables aléatoires Z et W dans les exemples 3.2.1 sont illustrés à la figure 3B.4.1. L'interprétation de ces histogrammes de probabilité est similaire à celle des histogrammes de fréquence relative, à ceci près que l'aire représente la probabilité (théorique) au lieu des fractions (empiriques) d'ensembles de données.

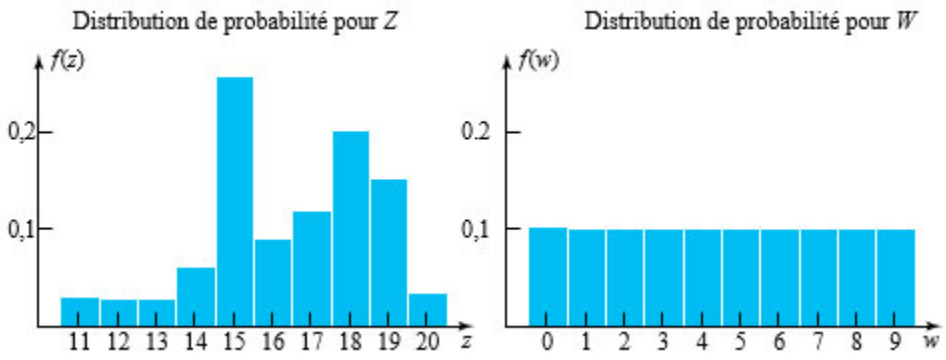


Figure 3.2.4.1. Histogrammes de probabilité pour Z et W (exemples 3.2.1.1 et 3.2.1.2)

Il est utile d'avoir une notion de valeur moyenne pour une variable aléatoire discrète (ou sa distribution de probabilité).

DÉFINITION 3.2.4.1. Moyenne d'une variable aléatoire discrète

La moyenne ou espérance mathématique d'une variable aléatoire discrète X (parfois appelée moyenne de sa distribution de probabilité) est

$$EX = \sum_x x f(x)$$

E(X) correspond à « l'espérance mathématique de X », qu'on note parfois μ.

Rappelez-vous que μ représente à la fois la moyenne d'une population et la moyenne d'une distribution de probabilité, comme nous l'avons vu avec les distributions empiriques.

Exemple 3.2.4.1. Exemple du couple des boulons (suite)

En reprenant l'exemple des couples de serrage des boulons, l'espérance mathématique (ou moyenne théorique) du prochain couple est :

$$EZ = \sum_z zf(z)$$

= 11(0,03) + 12(0,03) + 13(0,03) + 14(0,06) + 15(0,26) + 16(0,09) + 17(0,12) + 18(0,20) + 19(0,15) + 20(0,03)

= 16,35 ft lb

Cette valeur est essentiellement la moyenne arithmétique des couples de serrage du boulon 3 énumérés précédemment. Ce type d'accord motive l'utilisation du symbole μ, vu pour la première fois dans le module 2, comme synonyme de E(Z).

On peut interpréter la moyenne d'une distribution de probabilité discrète comme étant le point d'équilibre de la distribution, comme on l'avait fait pour la moyenne arithmétique d'un ensemble de données. Si on place des masses (ponctuelles) de valeur f(x) en des points x le long d'une droite des nombres, E(X) est le centre de masse de cette distribution.

Exemple 3.2.4.2. Exemple des numéros de série (suite)

Si on reprend l'exemple du numéro de série et la deuxième partie de la figure 3.2.4.1, pour que l'interprétation de l'espérance mathématique comme point d'équilibre tienne la route, il vaudrait mieux que E(W) soit égal à 4,5. Heureusement,

E(W) = 0(0,1) + 1(0,1) + 2(0,1) +···+8(0,1) + 9(0,1) = 45(0,1) = 4,5

Il était pratique de mesurer la dispersion d'un ensemble de données (ou de sa distribution de fréquence relative) à l'aide de la variance et de l'écart-type. Il est également utile d'avoir des notions de dispersion pour une distribution de probabilité discrète.

DÉFINITION 3.2.4.2. Variance d'une variable aléatoire discrète X

La variance d'une variable aléatoire discrète X (ou la variance de sa distribution) est :

$$\text{Var } X = \sum (x - EX)^2 f(x) \quad \left(= \sum x^2 f(x) - (EX)^2 \right)$$

L'écart-type de X est $\sqrt{VarX}$. On utilise souvent la notation σ^2 à la place de Var(X), et σ à la place de $\sqrt{VarX}$.

La variance d'une variable aléatoire est sa distance au carré attendue (ou moyenne) par rapport au centre de sa distribution de probabilité. L'utilisation de σ^2 pour représenter à la fois la variance d'une population et la variance d'une distribution de probabilité est motivée par les mêmes raisons que la double utilisation du symbole μ.

Exemple 3.2.4.3. Exemple du couple des boulons (suite)

Les calculs nécessaires pour obtenir l'écart type du couple de serrage sont présentés dans le tableau 3.2.4.1. Donc :
$\sigma = \sqrt{\text{Var}(Z)} = \sqrt{4{,}6275}$ = 2,15 ft lb

À l'exception d'une petite différence due aux arrondissements lors de la création du tableau 3.2.1.2, cet écart-type de la variable aléatoire Z est numériquement le même que l'écart-type de la population associé des couples du boulon 3 dans le tableau 2.1.4.1. (Encore une fois, ceci est cohérent avec l'équivalence entre la distribution de fréquence relative de la population et la distribution de probabilité de Z).

Calculs de Var(Z)

z	$f(z)$	$(z - 16{,}35)^2$	$(z - 16{,}35)^2 f(z)$
11	0,03	28,6225	0,8587
12	0,03	18,9225	0,5677
13	0,03	11,2225	0,3367
14	0,06	5,5225	0,3314
15	0,26	1,8225	0,4739
16	0,09	,1225	0,0110
17	0,12	0,4225	0,0507
18	0,20	2,7225	0,5445
19	0,15	7,0225	1,0534
20	0,03	13,3225	0,3997
			Var(Z)= 4,6275

Tableau 3.2.4.1. Calculs de Var(Z)

Exemple 3.2.4.4. Exemple des numéros de série (suite)

Pour illustrer l'autre méthode de calcul de la variance donnée dans la définition 3.2.4.2, essayons de calculer la variance et de l'écart type de la variable numéro de série W. Le tableau 3.2.4.2 indique le calcul de $\sum w^2 f(w)$.

Calculations for $\sum w^2 f(w)$

w	$f(w)$	$w^2 f(w)$
0	.1	0.0
1	.1	.1
2	.1	.4
3	.1	.9
4	.1	1.6
5	.1	2.5
6	.1	3.6
7	.1	4.9
8	.1	6.4
9	.1	8.1
		28.5

Tableau 3.2.4.2.

D'où

$$\mathrm{Var}(W) = \sum w^2 f(w) - (E(W))^2 = 28{,}5 - (4{,}5)^2 = 8{,}25$$

de sorte que

$$\sqrt{Var(W)} = 2{,}87$$

En comparant les deux histogrammes de probabilité de la figure précédente, on remarque que la distribution de W semble plus dispersée que celle de Z. Heureusement, cela se prouve mathématiquement :

$\sqrt{\mathbf{Var}(W)} = 2{,}87 > 2{,}15 = \sqrt{\mathbf{Var}(Z)}$ 2,15=\sqrt{\operatorname{Var}(Z)} » title= »\sqrt{\operatorname{Var}(W)}=2,87>2,15=\sqrt{\operatorname{Var}(Z)} » class= »latex mathjax »>

3.2.5 Distribution binomiale

Les distributions de probabilités discrètes sont parfois développées à partir de l'expérience passée d'un phénomène physique particulier (comme dans l'exemple 1). Cependant, il est parfois possible de constituer un ensemble d'hypothèses mathématiques facilement manipulables et susceptibles de décrire une variété de situations réelles. Lorsqu'il est possible de les manipuler pour obtenir des distributions génériques, ces distribution peuvent être utilisées pour modéliser de nombreux phénomènes aléatoires. L'une de ces hypothèses est celle d'**essais de succès et d'échec indépendants et identiques**.

De nombreuses situations d'ingénierie impliquent la répétition du même scénario « succès-échec », où :

1. Il y a une probabilité constante de réussite à chaque répétition du scénario (appelée probabilité p).
2. Les répétitions sont indépendantes en ce sens que la connaissance du résultat de l'une d'entre elles ne modifie pas la probabilité des autres.

Parmi les exemples de ce type, on peut citer la vérification de conformité d'articles fabriqués consécutivement; le respect de la limite de vitesse à un poste de contrôle routier; et la mesure du travail de plusieurs personnes dans deux espaces de configuration différente afin de voir si elles travaillent mieux dans la configuration A ou la configuration B.

Dans ce contexte, il existe deux types génériques de variables aléatoires pour lesquelles il est facile de dériver des distributions de probabilité appropriées. Le premier est le cas d'un décompte des répétitions sur les n qui aboutissent à un résultat de type « succès ». Autrement dit, soit la variable :

Variables aléatoires binomiales

X = le nombre de succès dans n essais succès-échec identiques indépendants <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/955a246b99acabb5e7c8a45cef7cc01b.png » alt= »X[\latex] suit une **distribution binomiale (n, p)** .

DÉFINITION 3.2.5.1. Définition de la distribution binomiale La distribution binomiale Formula does not parse(n, p)" title="X[\latex] suit une **distribution binomiale (n, p)** .

DÉFINITION 3.2.5.1. Définition de la distribution binomiale La distribution binomiale Formula does not parse (n, p)" class="latex mathjax"> est une distribution de probabilité discrète avec une fonction de probabilité

$$f(x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & \text{pour } x = 0, 1, \ldots, n \\ 0 & \text{sinon} \end{cases}$$

avec n un entier positif et <img src="https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/75d5ea78867a7472b54c906e926ba4c2.png" alt="0<p<1" title="0<p.

L'équation 3.2.5.1 est entièrement plausible. Elle contient un facteur de p pour chaque essai produisant un succès et un facteur de $(1-p)$ pour chaque essai produisant un échec. Le terme $n!/x!(n-x)!$ est un décompte du nombre de combinaisons dans lesquelles il est possible de voir x succès en n essais. Le nom distribution $bi-$ nomiale trouve son origine dans le fait que les valeurs $f(0), f(1), f(2), \ldots, f(n)$ sont les termes du développement de

$$(p + (1-p))^n$$

selon le théorème binomial.

Prenons le temps de tracer les histogrammes de probabilité de quelques distributions binomiales. De fait, si <img src="https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/314dc267dffc559dc055d95deb5a1e9b.png" alt="p<0,5" title="p, l'histogramme résultant est asymétrique à droite. Si $p > .5$0,5" title="p>0,5" class="latex mathjax">, l'histogramme résultant est asymétrique à gauche. L'asymétrie augmente à mesure que p s'éloigne de 0,5 et diminue à mesure que n augmente. La figure 3.2.5.1 illustre quatre histogrammes de probabilité binomiale.

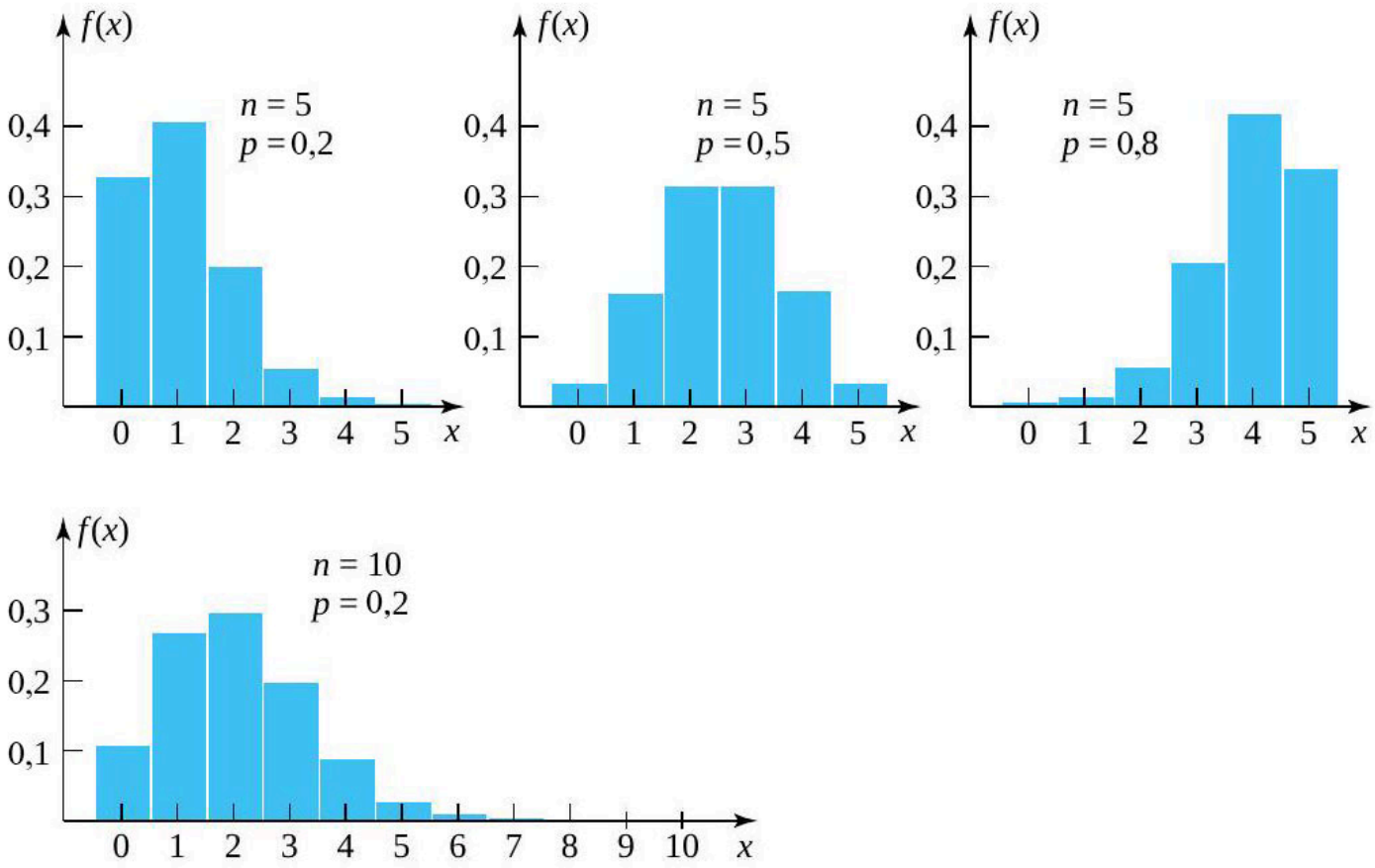


Figure 3.2.5.1. Quatre histogrammes de probabilité binomiale

Exemple 3.2.5.1. Distribution binomiale et nombre d'arbres réusinables

Prenons l'exemple d'une étude de performance d'un procédé de tournage d'arbres en acier. Au début de cette étude, environ 20% des arbres ont été classés comme « réusinables ». Supposons que $p = 0{,}2$ soit une probabilité crédible qu'un arbre soit réusinable. Supposons en outre qu'on inspecte $n = 10$ arbres et qu'on s'intéresse à la probabilité qu'au moins deux d'entre eux soient classifiés comme réusinable.

En adoptant un modèle d'essais indépendants et identiques de réussite et d'échec pour l'état des arbres,

$$U = \text{ nombre d'arbres réusinables sur un échantillon de 10}$$

est une variable aléatoire binomiale avec $n = 10$ et $p = 0{,}2$. Donc :

$$\begin{aligned} P[\text{ au moins deux arbres réusinables }] &= P[U \geq 2] \\ &= f(2) + f(3) + \cdots + f(10) \\ &= 1 - (f(0) + f(1)) \\ &= 1 - \left(\frac{10!}{0!10!}(0{,}2)^0(0{,}8)^{10} + \frac{10!}{1!9!}(0{,}2)^1(0{,}8)^9 \right) \\ &= 0{,}62 \end{aligned}$$

\begin{aligned}P[\text{ au moins deux arbres réusinables }] &= P[U \geq 2] \\ &= f(2)+f(3)+\cdots+f(10) \\ &= 1-(f(0)+f(1)) \\ &= 1-\left(\frac{10!}{0!10!}(0{,}2)^0(0{,}8)^{10}+\frac{10!}{1!9!}(0{,}2)^1(0{,}8)^9\right) \\ &= 0{,}62\end{aligned}

(Ici, nous avons employé une astuce très utile et couramment utilisée qui consiste à éviter de calculer neuf fois la fonction de probabilité binomiale en reconnaissant que la somme des $f(u)$ doit être égale à 1.)

La probabilité 0,62 est seulement aussi fiable que les hypothèses sous-jacentes. Si le modèle d'essais succès/échec identiques et indépendants ne convient pas pour décrire la réalité physique, la valeur de 0,62 est mathématiquement correcte, mais elle n'est peut-être pas représentative de la réalité. Par exemple, disons qu'en raison de l'usure de l'outil, il est courant de voir 40 arbres conformes aux spécifications, puis 10 arbres réusinables, puis après un changement d'outil, 40 arbres conformes aux spécifications, et ainsi de suite. Dans ce cas, la distribution binomiale serait une très mauvaise description de la situation de U, et le nombre 0,62 est alors peu pertinent. (L'hypothèse de l'indépendance des essais serait inappropriée dans cette situation.)

Distribution binomiale et échantillonnage aléatoire simple

Il y a un contexte important dans lequel le modèle d'essais succès/échec indépendants et identiques n'est pas exactement approprié, mais pour lequel la distribution binomiale peut encore convenir à des fins pratiques : la description des résultats d'un échantillonnage aléatoire simple à partir d'une population dichotomique. Supposons une population de taille N qui contient une fraction p d'objets de type A et une fraction $(1-p)$ d'objets de type B. Si on sélectionne un échantillon aléatoire simple de n articles, alors la variable

$$X = \text{ the number of type A items in the sample}$$

n'est pas, à strictement parler, une variable aléatoire binomiale.x Mais si n est petit par rapport à N (disons, moins de 10%) et que p n'est pas trop extrême (c'est-à-dire, n'est pas proche de 0 ou 1), X suit approximativement une distribution binomiale (n, p).

Exemple 3.2.5.2. Échantillonnage aléatoire simple à partir d'un lot de pastilles d'hexamine

Lors d'une expérience sur une machine à granuler, Greiner, Grimm, Larson et Lukomski ont trouvé une combinaison de réglages de la machine qui leur a permis de produire 66 pastilles conformes sur un lot de 100 pastilles. Considérons ce lot de 100 pastilles comme une population d'intérêt et sélectionnons un échantillon aléatoire simple de taille $n = 2$.

Si l'on définit la variable aléatoire

$$V = \text{ nombre de pastilles conformes dans l'échantillon de taille } 2$$

la distribution de probabilité la plus naturelle pour V est obtenue comme suit. Les valeurs possibles pour V sont 0,1 et 2.

$$\begin{aligned} f(0) =& P[V=0] \\ =& P[\text{ la première pastille sélectionnée} \\ & \text{n'est pas conforme et la seconde ne l'est pas non plus }] \\ f(2) =& P[V=2] \\ =& P[\text{ la première pastille sélectionnée n'est pas} \\ & \text{conforme et la seconde ne l'est pas non plus }] \\ f(1) =& 1-(f(0)+f(2)) \end{aligned}$$

Ensuite, on fait le raisonnement suivant :« À long terme, la première sélection produira une pastille non conforme environ 34 fois sur 100. Si on considère uniquement les cas où cela se produit, à long terme, la sélection suivante produira également une pastille non conforme environ 33 fois sur 99 ». Cela revient à dire que $f(0)$ vaut

$$f(0) = \frac{34}{100} \cdot \frac{33}{99} = .1133$$

De la même façon,

$$f(2) = \frac{66}{100} \cdot \frac{65}{99} = .4333$$

et par conséquent

$$f(1) = 1 - (0{,}1133 + 0{,}4333) = 1 - 0{,}5467 = 0{,}4533$$

Manifestement, V ne peut être considérée comme le résultat d'essais parfaitement indépendants. Par exemple, le fait de savoir que la première pastille sélectionnée était conforme ferait passer la probabilité que la deuxième soit également conforme de $\frac{66}{100}$ à $\frac{65}{99}$. Néanmoins, à des fins pratiques, V peut être considérée comme essentiellement binomiale avec $n = 2$ et $p = .66$. Pour s'en convaincre, il suffit de noter que

$$\frac{2!}{0!2!}(.34)^2(.66)^0 = .1156 \approx f(0)$$
$$\frac{2!}{1!1!}(.34)^1(.66)^1 = .4488 \approx f(1)$$
$$\frac{2!}{2!0!}(.34)^0(.66)^2 = .4356 \approx f(2)$$

Comme n est petit par rapport à N, p et que p n'est pas trop extrême, la distribution binomiale est une description correcte d'une variable issue d'un échantillonnage aléatoire simple.

Moyenne et variance d'une distribution binomiale (n, p)

Le calcul de la moyenne et de la variance des variables aléatoires binomiales est vraiment simplifié par le fait que lorsque les formules présentées dans ce module sont utilisées avec l'expression des probabilités binomiales de l'équation 3.2.5.1, on obtient des formules simples. Soit X une variable nominale binomiale (n, p) :

DÉFINITION 3.2.5.2. Moyenne d'une distribution binomiale (n,p)

$$\mu = EX = \sum_{x=0}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = np$$

De plus,

DÉFINITION 3.2.5.3. Variance d'une distribution binomiale (n,p)

$$\sigma^2 = \operatorname{Var} X = \sum_{x=0}^{n} (x - np)^2 \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} = np(1-p)$$

Exemple 3.2.5.3. Usinage d'arbres en acier.

En revenant à l'usinage des arbres en acier, supposons qu'une distribution binomiale avec $n = 10$ et $p = 0{,}2$ décrit adéquatement la variable

$$U = \text{ nombre d'arbres réusinables sur un échantillon de 10}$$

En utilisant les formules 3.2.5.2 et 3.2.5.3,

$$E(U) = (10)(0{,}2) = 2 \text{ arbres}$$

$$\sqrt{\text{Var}(U)} = \sqrt{10(0{,}2)(\,0{,}8)} = 1{,}26 \text{ arbres}$$

3.2.6 Distribution de Poisson

Il est souvent important de noter le nombre total d'occurrences d'un phénomène relativement rare dans un intervalle de temps ou d'espace où il peut y avoir plusieurs occurrences. Par exemple, une caisse de tuiles de sol peut présenter un grand nombre d'imperfections; dans un intervalle d'une seconde, il y a potentiellement un grand nombre de messages qui peuvent arriver dans un centre de référence en vue d'être réacheminés; un échantillon de verre de 1 cc contient potentiellement un grand nombre d'imperfections.
Il faut donc des distributions de probabilités pour décrire le décompte aléatoire du nombre d'occurrences d'un phénomène relativement rare dans un intervalle de temps ou d'espace donné. Dans ce genre de situation, la distribution théorique la plus fréquemment utilisée est de loin la distribution de Poisson.

DÉFINITION 3.2.6.1. Distribution de Poisson (λ)

La distribution de Poisson (λ) est une distribution de probabilité discrète avec la fonction de probabilité

$$f(x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & \text{pour } x = 0, 1, 2, \ldots \\ 0 & \text{sinon} \end{cases}$$

où $\lambda > 0$0" title= »\lambda>0" class= »latex mathjax »>.

La forme de l'équation 3.2.6.1 peut sembler un peu rébarbative au premier abord, mais il s'agit d'un modèle qui a des origines mathématiques raisonnables, qui est gérable et qui s'est révélé empiriquement utile dans de nombreux contextes d'« événements rares ». Une façon d'arriver à l'équation (3.2.6.1) est de penser à un très grand nombre d'essais indépendants (possibilités d'occurrence), où la probabilité de succès (occurrence) lors d'un seul essai est très faible et où le produit du nombre d'essais et de la probabilité de succès est λ. On obtient alors la distribution binomiale $\left(n, \frac{\lambda}{n}\right.$. En fait, si n est grand, la fonction de probabilité binomiale $\left(n, \frac{\lambda}{n}\right)$ se rapproche de celle donnée dans la définition 3.2.6.1. On peut donc considérer que la distribution de Poisson pour les dénombrements résulte d'un mécanisme qui présente de nombreuses possibilités d'occurrences (à très faible probabilité) ou de non-occurrences indépendantes dans un intervalle de temps ou d'espace donné.
La distribution de Poisson est asymétrique à droite avec $x = 0, 1, 2, \ldots$, et l'histogramme de probabilité atteint un sommet près deλ. La figure 3.2.6.1 présente deux histogrammes de probabilité de Poisson différents.

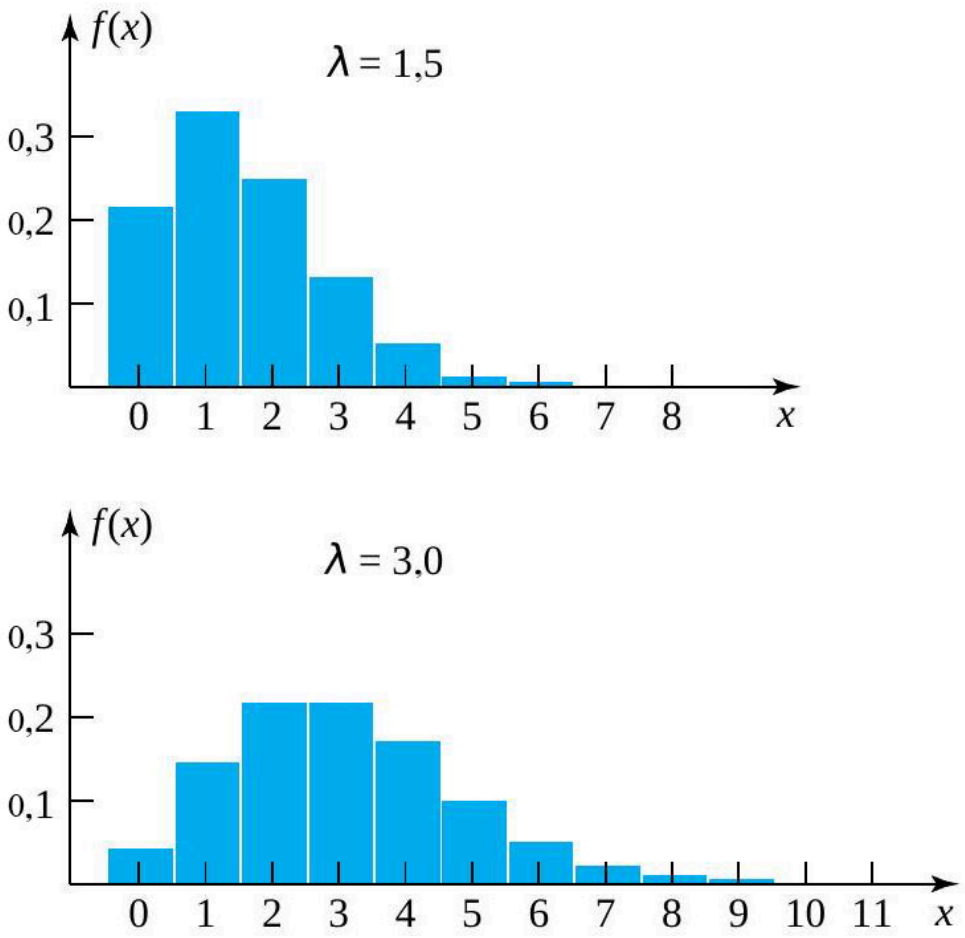


Figure 2.3.6.1. Deux histogrammes de probabilité de Poisson.

λ est à la fois la moyenne et la variance de la distribution de Poisson (λ). Autrement dit, si X suit une distribution de Poisson (λ), alors

DÉFINITION 3.2.6.2. Moyenne de la distribution de Poisson (λ)

$$\mu = EX = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda$$

et

DÉFINITION 3.2.6.3. Variance de la distribution de Poisson (λ)

$$\operatorname{Var} X = \sum_{x=0}^{\infty} (x-\lambda)^2 \frac{e^{-\lambda}\lambda^x}{x!} = \lambda$$

La définition 3.2.6.2 est utile pour déterminer quelle distribution de Poisson pourrait être utile pour décrire une situation d'« événements rares » donnée.

Exemple 3.2.6.1. Distribution de Poisson et décomptes de particules α

En 1910, Rutherford et Geiger ont publié dans le *Philosophical Magazine* un célèbre ensemble de données décrivant le nombre de particules α émises par une petite barre de polonium et entrant en collision avec un écran placé près de la barre au cours de 2 608 périodes de 8 minutes chacune. La distribution des fréquences relatives de Rutherford et Geiger a une moyenne de 3,87 et une forme remarquablement similaire à celle de la distribution de probabilité de Poisson avec une moyenne de $\lambda = 3.87$.

Si on reproduit l'expérience de Rutherford/Geiger, on peu raisonnablement décrire

$$S = \text{the number of } \alpha\text{-particles striking the screen in an additional 8-minute period}$$

par la fonction de probabilité

$$f(s) = \begin{cases} \frac{e^{-3.87}(3.87)^s}{s!} & \text{pour } s = 0, 1, 2, \ldots \\ 0 & \text{sinon} \end{cases}$$

En utilisant un tel modèle, on obtient (par exemple)

P [au moins 4 particules sont enregistrées]

$$\begin{aligned} &= P[S \geq 4] \\ &= f(4) + f(5) + f(6) + \cdots \\ &= 1 - (f(0) + f(1) + f(2) + f(3)) \\ &= 1 - \left(\frac{e^{-3.87}(3.87)^0}{0!} + \frac{e^{-3.87}(3.87)^1}{1!} + \frac{e^{-3.87}(3.87)^2}{2!} + \frac{e^{-3.87}(3.87)^3}{3!} \right) \\ &= 0{,}54 \end{aligned}$$

Exemple 3.2.6.2. Entrées dans la bibliothèque de l'université

Stork, Wohlsdorf et McArthur ont recueilli des données sur le nombre d'étudiant.e.s entrant dans la bibliothèque de l'ISU à différentes périodes au cours d'une semaine. Leurs données indiquent qu'entre 12 h 00 et 12 h 10 du lundi au mercredi, une moyenne d'environ 125 étudiant.e.s sont entré.e.s. Soit

$$M = \text{ nombre d'étudiants qui entreront dans la bibliothèque de l'ISU entre 12 h 00 et 12 h 01 mardi prochain}$$

Si on utilise une distribution de Poisson pour décrire M, λ semble être

$$\lambda = \frac{125 \text{ étudiant.e.s}}{10 \text{ minutes}}(1 \text{ minute }) = 12{,}5 \text{ étudiant.e.s}$$

Ainsi,

$$E(M) = \lambda = 12{,}5 \text{ étudiant.e.s}$$
$$\sqrt{\text{Var}(M)} = \sqrt{\lambda} = \sqrt{12{,}5} = 3{,}54 \text{ étudiant.e.s}$$

La probabilité que, par exemple, entre 10 et 15 étudiant.e.s (inclusivement) arrivent à la bibliothèque entre 12 h 00 et 12 h 01 serait évaluée comme suit :

$$\begin{aligned} P[10 \leq M \leq 15] &= f(10) + f(11) + f(12) + f(13) + f(14) + f(15) \\ &= \frac{e^{-12{,}5}(12{,}5)^{10}}{10!} + \frac{e^{-12{,}5}(12{,}5)^{11}}{11!} + \frac{e^{-12{,}5}(12{,}5)^{12}}{12!} \\ &\quad + \frac{e^{-12{,}5}(12{,}5)^{13}}{13!} + \frac{e^{-12{,}5}(12{,}5)^{14}}{14!} + \frac{e^{-12{,}5}(12{,}5)^{15}}{15!} \\ &= 0{,}60 \end{aligned}$$

3.2.7 Utilisation de Python pour les distributions de probabilités discrètes

Si vous souhaitez manipuler des distributions de probabilités discrètes avec Python, il est fortement recommandé de consulter les fichiers Jupyter Notebook **Normal Probability & Confidence Intervals**. Vous pouvez les trouver dans la section « How do I do X in Python? ». Le fichier « Discrete Probability Distributions » sera particulièrement utile.

4.0.1 Introduction aux variables aléatoires continues et aux distributions de probabilités continues

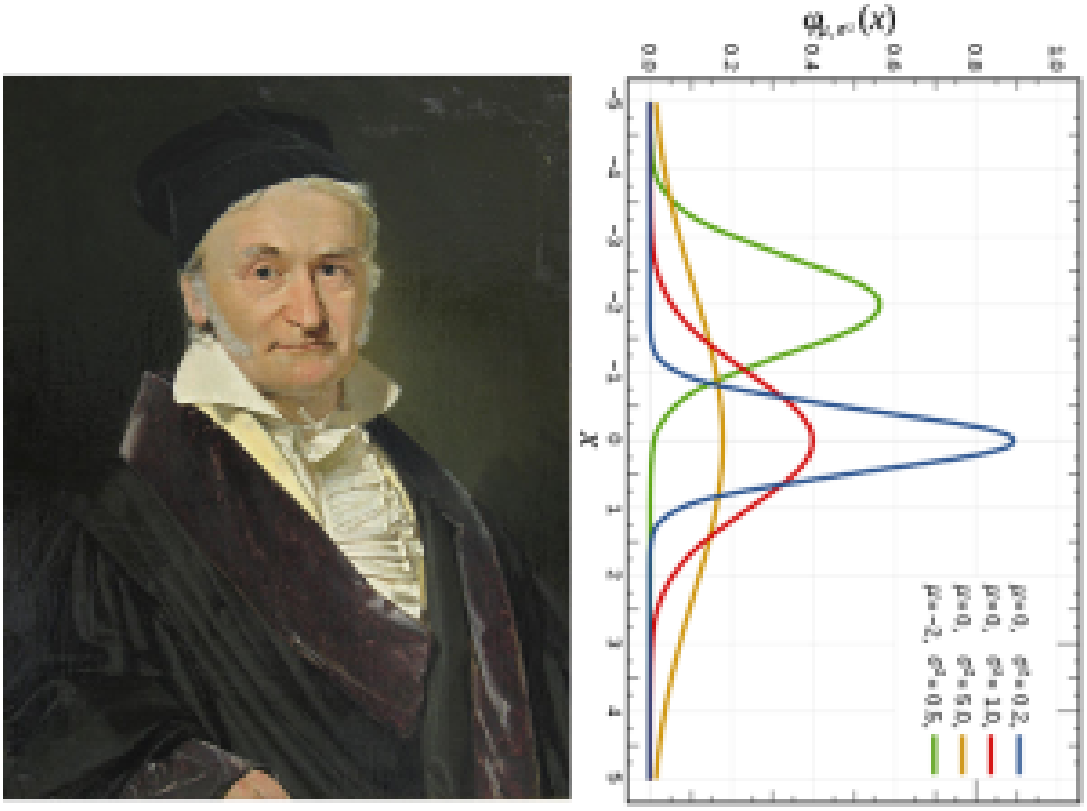


Figure 4.1.0.1. Friedrich Gauss: https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

Reconnu comme le « Prince des mathématiciens », le mathématicien allemand Carl Friedrich Gauss (1777-1855; figure 4.1.0.1) occupe une place prépondérante dans l'histoire des statistiques et des mathématiques. Gauss a apporté des contributions prodigieuses dans divers domaines, mais ses travaux sur les statistiques et la théorie des probabilités sont remarquables. Il est surtout connu pour avoir élaboré la méthode des moindres carrés et la distribution normale, également connue sous le nom de distribution gaussienne ou de courbe en cloche, essentielle pour l'analyse statistique dans divers domaines, des sciences sociales aux sciences naturelles en passant par l'ingénierie. La distribution normale est une distribution de probabilité symétrique qui décrit la manière dont une variable aléatoire continue peut être distribuée. Sa courbe caractéristique en forme de cloche apparaît lorsqu'un ensemble de données présente une fréquence élevée de valeurs proches de la moyenne, les fréquences diminuant progressivement au fur et à mesure que les valeurs s'éloignent de la moyenne. Elle est omniprésente parce qu'elle modélise naturellement de nombreux phénomènes du monde réel et parce que de nombreux processus et expériences aléatoires tendent à produire des données qui suivent une distribution normale. Son importance réside dans sa capacité à fournir un cadre simple, mais puissant, pour comprendre et interpréter les ensembles de données, ce qui en fait une pierre angulaire de l'analyse statistique.

VARIABLES ALÉATOIRES CONTINUES

Il est souvent plus facile de considérer qu'une variable aléatoire n'est pas discrète mais plutôt continue, dans le

sens où elle peut prendre n'importe laquelle des valeurs d'un intervalle continu. Les outils utilisés pour décrire les distributions de probabilités continues diffèrent de ceux de la dernière section. Pour commencer, nous allons la notion de fonction de densité de probabilité et sa relation avec la fonction
de probabilité cumulative pour une variable aléatoire continue, puis nous allons voir comment on peut l'utiliser pour calculer la moyenne et la variance d'une distribution continue. Nous passerons ensuite en revue quelques distributions utiles : la distribution uniforme, la distribution exponentielle et la distribution de Weibull. Ensuite, nous nous attarderons sur la distribution continue la plus importante et la plus courante, utile dans les applications techniques de la théorie des probabilités : la distribution normale.

4.0.2 Sources de la partie 4

Cette première version de la partie 4 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte, et on trouve des liens pour y accéder tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn et sont partagées sous licence CC BY-SA 4.0.

4.1.1 Fonction de densité de probabilité et fonction de probabilité cumulative

FONCTION DE DENSITÉ DE PROBABILITÉ

Les méthodes utilisées pour spécifier et décrire les distributions de probabilité ont des parallèles en physique mécanique qui sont particulièrement utiles dans le cas des distributions de probabilité continues. En mécanique, les propriétés d'une distribution de masse continue sont liées à la densité (potentiellement variable) de la masse dans la région de l'espace qu'elle occupe. La quantité de masse d'une région donnée s'obtient en intégrant la densité de masse sur cette région.

En théorie des probabilités, le concept qui correspond à la densité de masse en mécanique, c'est la densité de probabilité. Pour spécifier une distribution de probabilité continue, il faut décrire « l'épaisseur » de la probabilité dans les différentes parties de l'ensemble des valeurs possibles. La définition formelle est la suivante :

DÉFINITION 4.1.1.1. Fonction de densité de probabilité (FDP)

EXPRESSION 4.1.1.1.

Une fonction de densité de probabilité pour une variable aléatoire continue X est une fonction non négative f(x) telle que :

$$\int_{-\infty}^{\infty} f(x)dx$$

et telle que pour tout a ≤ b, P [a ≤ X ≤ b] correspond à :

EXPRESSION 4.1.1.2.

$$P(a \le X \le b) = \int_{a}^{b} f(x)dx$$

.

La figure 4.1.1.1 illustre une fonction de densité de probabilité (FDP) générique. Comme on peut le voir, le diagramme d'une distribution de probabilité continue est une courbe. Conformément aux équations de définition de la FDP, le graphique de f(x) ne descend pas sous l'axe des x, l'aire totale sous la courbe y = f(x) est égale à 1, et l'aire sous la courbe d'un intervalle donné correspond à la probabilité que la valeur se trouve dans cet intervalle. La fonction f(x) est définie de sorte que l'aire sous la courbe soit une probabilité. La probabilité maximale étant égale à 1, l'aire maximale est également égale à 1.

La courbe est la FDP. On utilise le symbole $f(x)$ pour

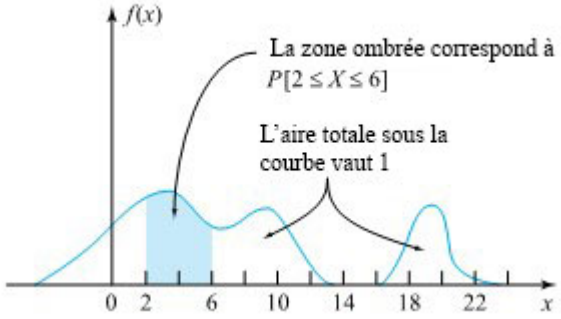


Figure 4.1.1.1. Une fonction de densité de probabilité générique.

représenter la courbe. $f(x)$ est la fonction qui correspond au graphique; on utilise la fonction de densité $f(x)$ pour faire le graphique de la distribution de probabilité. L'aire sous la courbe représente la probabilité.

Analogie de la mécanique pour la densité de probabilité (suite)

Comme en mécanique, si *f(x)* est effectivement la « densité de probabilité » à *x*, alors la probabilité dans un petit intervalle *dx* autour de *x* est approximativement *f(x) dx*. (En mécanique, si *f(x)* est la densité de masse à *x* , alors la masse dans un petit intervalle *dx* autour de *x* est approximativement *f(x) dx*.) Pour calculer la probabilité entre *a* et *b*, il faut donc additionner les valeurs de *f(x) dx*. $\int_a^b f(x)dx$ est exactement la limite de $\sum$ *f(x) dx* lorsque *dx* tend vers zéro. (En mécanique, $\int_a^b f(x)dx$ est la masse entre *a* et *b*.) L'expression dans la définition de la FDP et l'expression 4.1.1.2 sont donc raisonnables.

Pour une variable aléatoire continue X, P(X = a) = 0

Il y a un détail concernant les distributions de probabilité continuées qui peut sembler contre-intuitif à première vue : la probabilité qu'une variable aléatoire continue prenne une valeur précise (par exemple, *a*). Tout comme la masse en un seul point d'une distribution de masse continue est nulle, P(X = *a*) = 0 pour une variable aléatoire continue X. Cela découle de l'expression 4.1.1.2, car :

$P(a \leq X \leq b) = \int_a^b f(x)dx = 0$

Une conséquence de cette curiosité mathématique est que lorsque l'on travaille avec des variables aléatoires continues, il n'est pas nécessaire de se préoccuper de savoir si les signes d'inégalité que l'on écrit sont des signes d'inégalité stricts. Autrement dit, si X est continue :

$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$

FONCTION DE DISTRIBUTION CUMULATIVE

Précédemment, nous avons donné une définition parfaitement générale de la fonction de distribution cumulative d'une variable aléatoire, et nous l'avons précisée dans le contexte des variables discrètes. On peut maintenant utiliser l'équation 4.1.1.2 pour exprimer la fonction de distribution cumulative d'une variable aléatoire continue en termes d'une intégrale de sa densité de probabilité. Soit *X* une variable aléatoire continue de densité de probabilité *f*(*x*) :

DÉFINITION 4.1.1.3. Fonction de distribution cumulative (FDC) d'une variable continue X
EXPRESSION 4.1.1.3

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

F(x) est calculée à partir de l'intégrale de *f(x)*, en appliquant le théorème fondamentale du calcul à l'équation 4.1.1.3.

Autre relation entre F(x) et f(x)
EXPRESSION 4.1.1.4

$$\frac{d}{dx} F(x) = f(x)$$

Autrement dit, *f(x)* est calculé à partir de la dérivée de *F(x)*.

L'aire sous la courbe de la FDP est donnée par la FDC – ces deux fonctions ne sont pas identiques. La FDC est utilisée pour évaluer les probabilités et peut être trouvée en utilisant la géométrie, des formules, la technologie statistique ou des tables de probabilité.

Distributions de probabilités continues

Il existe de nombreuses distributions de probabilités continues. Lorsqu'on utilise une distribution de probabilité continue pour modéliser la probabilité, il faut choisir la distribution la plus adaptée au contexte. Dans ce module, nous étudierons la distribution uniforme, la distribution exponentielle et la distribution de Weibull, puis nous nous concentrerons sur la distribution la plus importante pour un cours d'introduction aux statistiques : la distribution normale.

Propriétés des distributions continues

La fonction de densité de probabilité (FDP) est utilisée pour décrire les probabilités des variables aléatoires continues. L'aire sous la courbe de densité entre deux points correspond à la probabilité que la variable se situe entre ces deux valeurs. En d'autres termes, l'aire sous la courbe entre les points *a* et *b* est égale à P(a<x<b). La fonction de distribution cumulative (FDC) donne la probabilité sous la forme d'une aire. Si *X* est une variable aléatoire continue, la fonction de densité de probabilité (FDP) *f*(*x*) permet de dessiner le graphique de la distribution de probabilité. L'aire totale sous le graphique de *f*(*x*) est égale à 1. L'aire sous le graphique de *f*(*x*) et entre les valeurs *a* et *b* donne la probabilité P(a<x<b). Ceci est illustré à la figure 4.1.1.2.

.

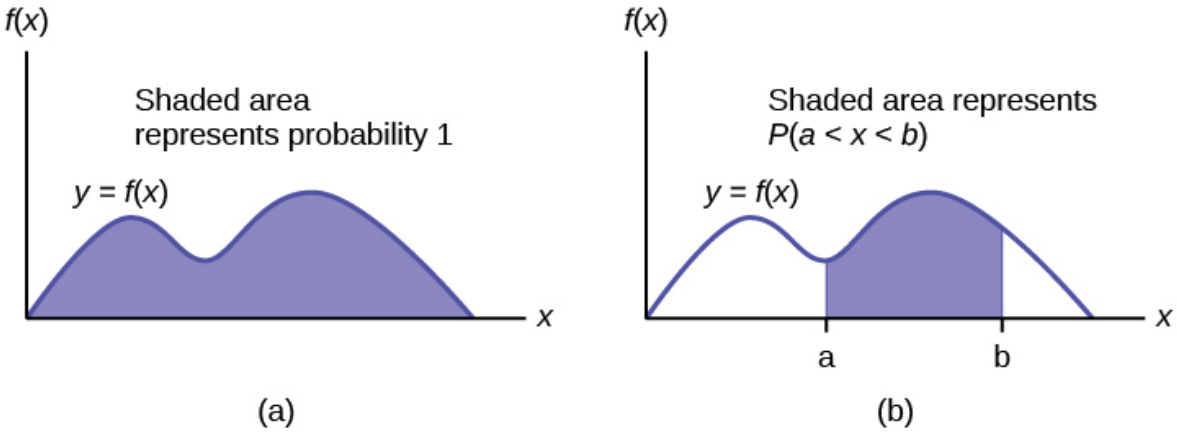


Figure 4.1.1.2. Le graphique de gauche représente une courbe de densité générique y = f(x). L'aire (ombrée) entre la courbe et l'axe des x (ombrée) est égale à 1, ce qui montre que tous les résultats possibles sont représentés par la courbe. Le graphique de droite montre la même courbe de densité. Les lignes verticales x = a et x = b s'étendent de l'axe à la courbe, et l'aire (ombrée) entre les lignes correspond à la probabilité qu'une valeur x se situe entre a et b.

La fonction de distribution cumulative (FDC) de *X* est définie par $P(X \leq x)$. C'est une fonction de *x* qui donne la probabilité que la variable aléatoire est inférieure ou égale à *x*.

- Les résultats sont mesurés et non comptés.
- L'aire sous la courbe est égale à 1.
- La probabilité est calculée pour des intervalles de valeurs x plutôt que pour des valeurs individuelles x.
- *P(c<x<d)* est la probabilité que la variable aléatoire X se trouve dans l'intervalle entre les valeurs c et d. *P(c<x<d)* est l'aire sous la courbe, entre c et d.
- $P(x = c) = 0$. La probabilité que x prenne une valeur précise est nulle. L'aire sous la courbe entre $x = c$ et $x = c$ a une largeur nulle et est donc nulle.$= 0$ La probabilité étant égale à l'aire, la probabilité est donc également nulle.
- *P(c<x<d)* est égale à $P(c \leq x \leq d)$, car la probabilité est égale à l'aire.

4.1.2 Moyenne et variance des distributions continues

MOYENNE ET VARIANCE DES DISTRIBUTIONS CONTINUES

Le graphique de la densité de probabilité f(x) est une sorte d'histogramme idéalisé. Il offre le même type d'interprétations visuelles que celles vues pour les histogrammes de fréquence relative et les histogrammes de probabilité. En outre, il est possible de définir une moyenne et une variance pour une distribution de probabilité continue. Ces synthèses numériques s'utilisent de la même manière que la moyenne et la variances pour décrire des ensembles de données et des distributions de probabilités discrètes.

DÉFINITION 4.1.2.1. Moyenne d'une variable aléatoire continue X
EXPRESSION 4.1.2.1.
La moyenne, ou espérance mathématique, d'une variable aléatoire continue X (parfois appelée moyenne de sa distribution de probabilité) est:

$$EX = \int_{-\infty}^{\infty} x f(x) dx.$$

Comme pour les variables aléatoires discrètes, on utilise parfois μ plutôt que E(X).

L'équation 4.1.2.1 est parfaitement plausible d'au moins deux points de vue. Premièrement, la probabilité dans un petit intervalle autour de x de longueur dx est approximativement f(x) dx. Ainsi, en multipliant par x et en additionnant, on obtient $\sum$xf(x) dx, et l'équation 4.1.2.1 correspond exactement à la limite de cette somme lorsque dx tend vers zéro. Deuxièmement, en mécanique, le centre de masse d'une distribution de masse continue est de la forme donnée dans l'équation 4.1.2.1, à l'exception de la division par la masse totale, qui pour une distribution de probabilité est 1.

La « continuisation » de la formule de la variance d'une variable aléatoire discrète produit une définition de la variance d'une variable aléatoire continue.

DÉFINITION 4.1.2.2. Variance d'une variable aléatoire continue X
EXPRESSION 4.1.2.2.
La variance d'une variable aléatoire continue X (parfois appelée variance de sa distribution de probabilité) est:

$$\operatorname{Var} X = \int_{-\infty}^{\infty} (x - EX)^2 f(x) dx \quad \left(= \int_{-\infty}^{\infty} x^2 f(x) dx - (EX)^2 \right)$$

L'écart-type de X est $\sqrt{VarX}$. On utiliser souvent la notation ^{2} à la place de Var(X), et le symbole σ est utilisé à la place de $\sqrt{VarX}$.

4.1.3 Distribution normale de probabilités

DISTRIBUTION NORMALE DE PROBABILITÉS

Bien qu'il existe plusieurs distributions continues couramment appliquées aux problèmes d'ingénierie, la distribution normale est particulièrement importante. De manière formelle, la distribution normale se définit comme suit :

DÉFINITION 4.1.3.1. Distribution normale
EXPRESSION 4.1.3.1

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

pour tout *x*, avec $\sigma > 0$.

Ce n'est pas nécessairement évident, mais l'équation 4.1.3.1 donne une densité de probabilité légitime, dans la mesure où l'aire totale sous la courbe y = f(x) est égale à 1. En outre, il est également vrai que :

Moyenne et variance d'une distribution normale

$$EX = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = \mu$$

et

$$\operatorname{Var} X = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = \sigma^2$$

Paramètres de la distribution normale

La distribution normale a deux paramètres (deux mesures numériques descriptives de la distribution théorique), la moyenne μ et la variance ^{2} (on se rappelle que l'écart-type = $\sqrt{\sigma^2}$ = σ). La figure 4.1.3.1 illustre la notation de la distribution normale standard et indique que la forme de la distribution dépend de ces paramètres. Comme l'aire sous la courbe doit être égale à 1, un changement dans l'écart-type *σ* entraîne une modification de la forme de la courbe; la courbe devient plus large ou plus étroite selon si *σ* augmente ou diminue, respectivement.

Une modification de μ entraîne une translation du graphique vers la gauche ou la droite. En vertu de ces deux paramètres, il existe un nombre infini de distributions de probabilités normales.

Les paramètres μ et ^{2} utilisés dans la définition 4.1.3.1 sont, respectivement, la moyenne et la variance (telles que définies dans les définitions 4.1.2.1 et 4.1.2.2) de la distribution. La figure 4.13.2 représente un graphique de la densité de probabilité donnée par l'équation 4.1.3.1. La courbe en forme de cloche représentée ici est symétrique par rapport à x = μ et présente des points d'inflexion à x = μ ± σ.

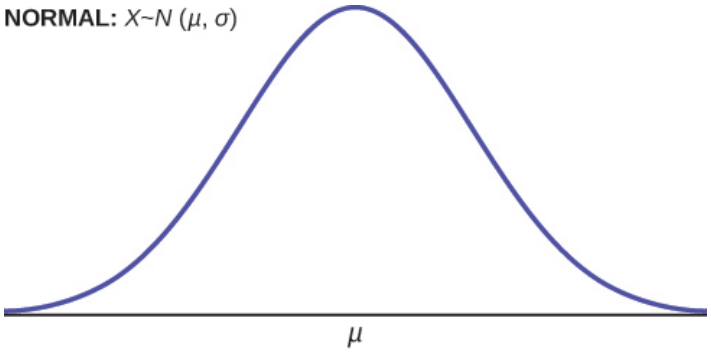


Figure 4.1.3.1. Notation de la distribution normale standard.

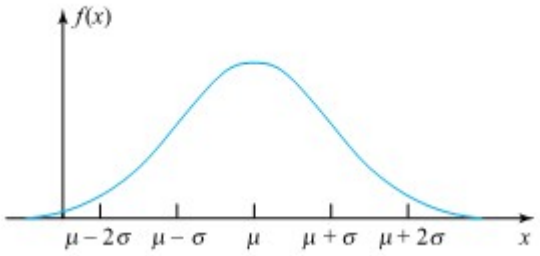


Figure 4.1.3.2. Graphique d'une fonction de densité de probabilité normale

L'équation 4.1.3.1 a plusieurs origines théoriques, mais c'est également une forme qui se révèle empiriquement utile dans une grande variété d'applications. En théorie, les probabilités des distributions normales peuvent être trouvées directement par intégration en utilisant l'équation 4.1.3.1. En effet, les calculatrices de poche préprogrammées pour effectuer l'intégration numérique permettent de vérifier certains des calculs dans les exemples qui suivent, en utilisant directement les équations 4.1.1.2 et 4.1.3.1. Nous utiliserons également le calcul statistique pour les trouver à l'aide d'une formule. Mais les méthodes d'évaluation des intégrales par primitives utilisées en première année de calcul échoueront lorsqu'il s'agira des distributions normales. Ces fonctions n'ont pas de primitives exprimables en termes de fonctions élémentaires. On utilise plutôt des tables de probabilités basées sur une version spécialisée de la distribution normale : la distribution normale réduite.

4.1.4 Distribution normale réduite

DISTRIBUTION NORMALE RÉDUITE

L'utilisation de tables pour évaluer les probabilités normales repose sur la relation suivante : si X est distribuée normalement avec une moyenne μ et une variance σ^2, alors

EXPRESSION 4.1.4.1.

$$P[a \leq X \leq b] = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

où la seconde inégalité est obtenue par un changement de variable :

$$z = \frac{x-\mu}{\sigma}$$

La cote *z* est une valeur normalisée mesurée en unités de l'écart-type. Par exemple, si la moyenne d'une distribution normale est de 5 et que l'écart-type est de 2, la valeur 11 se situe à trois écarts-types au-dessus (ou à droite) de la moyenne. Le calcul est effectué comme suit: $x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$, et la cote *z* est égale à trois : z = (11-5)/2 = 3. La cote *z* indique de combien d'écarts-types la valeur *x* est supérieure (à droite) ou inférieure (à gauche) à la moyenne *μ*. Les valeurs de *x* supérieures à la moyenne ont une cote *z* positive, tandis que les valeurs de *x* inférieures à la moyenne ont une cote *z* négative. Si *x* est égal à la moyenne, alors *x* a une cote *z* de 0.

L'équation 4.1.4.1 implique une intégrale de la densité normale avec μ = 0 et σ = 1. Le changement de variable $z = \frac{x-\mu}{\sigma}$z=\frac{x-\mu}{\sigma} produit la distribution Z ~ N(0,1). Cela signifie que la valeur x dans l'équation donnée provient d'une distribution normale avec une moyenne de 0 et un écart-type de 1. Autrement dit, toutes les probabilités normales peuvent être réduites à cette distribution normale spéciale. Ainsi, la distribution normale réduite est une distribution normale des valeurs normalisées à l'aide des cotes z.

DÉFINITION 4.1.4.2. DISTRIBUTION NORMALE RÉDUITE

EXPRESSION 4.1.4.2.

La distribution normale avec μ = 0 et σ = 1 est appelée distribution normale réduite.

$$\int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Cote z d'une valeur x d'une variable aléatoire normale (μ, σ²)

L'expression 4.1.4.2 montre comment utiliser la fonction de probabilité cumulative normale réduite pour calculer des probabilités normales générales. Pour une variable *X* normale, (μ, σ^2) et une valeur x associée à X on obtient la cote Z comme suit :

EXPRESSION 4.1.4.3.

$$z = \frac{x - \mu}{\sigma}$$

Ensuite, on consulte la table normale réduite en utilisant z à la place de x.

Relation entre les probabilités normales (μ, σ²) et les probabilités normales réduites : probabilité normale réduite cumulative

La relation entre les probabilités normales (μ, σ²) et les probabilités normales réduite est illustrée à la figure 4.1.4.1

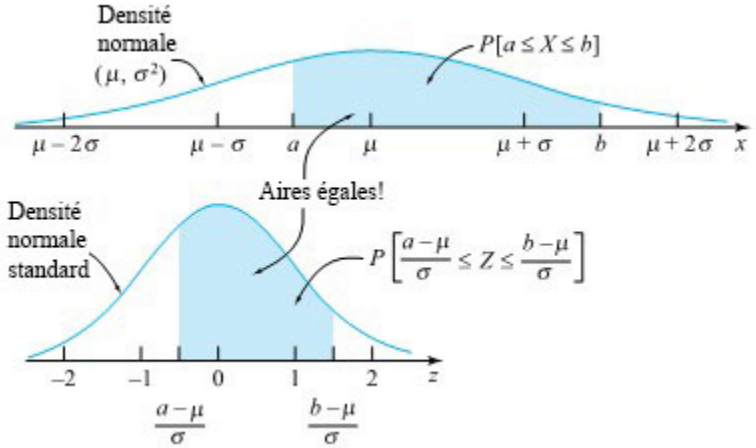


Figure 4.1.4.1. Illustration de la relation entre les probabilités normales (μ, σ^2) et les probabilités normales réduites.

Une fois que l'on a compris que les probabilités de toutes les distributions normales peuvent être obtenues en tabulant les probabilités de la distribution normale réduite seulement, il est relativement simple d'utiliser des techniques d'intégration numérique pour produire une table normale réduite. Dans ce texte, nous utiliserons les valeurs données dans la table A1.1 *Table de probabilités de la loi normale centrée réduite*, à l'annexe 1. (Il existe d'autres formats.) La table A1.1 est une table de la fonction de probabilité normale réduite cumulative. En d'autres termes, pour les valeurs z situées sur les marges de la table, les entrées dans le corps de la table correspondent à :

EXPRESSION 4.1.4.4

$$\Phi(z) = F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

où $\Phi(z)$ est utilisé pour représenter la fonction de probabilité normale réduite cumulative, au lieu de la lettre F, plus générique.

Relation entre la fonction de probabilité normale réduite cumulative et la fonction quantile normale réduite

Symboliquement, avec $\Phi(z)$, la fonction de probabilité cumulative normale réduite, et $Q_z(p)$, la fonction quantile normale réduite, on a :

EXPRESSION 4.1.4.5.

$$\left.\begin{aligned} \Phi\left(Q_z(p)\right) &= p \\ Q_z(\Phi(z)) &= z \end{aligned}\right\}$$

Les relations 4.1.4.5 signifient que $Q_z(p)$ et $\Phi(z)$ sont des fonctions inverses. (En fait, la relation $Q = F^{-1}$ n'est pas une propriété particulière de la distribution normale réduite; cette identité tient pour toutes les distributions continues).

EXEMPLES

Exemple 4.1.4.1. Probabilités normales réduite

Supposons que Z soit une variable aléatoire normale réduite. Nous allons trouver des probabilités pour Z à l'aide de la table A1.1 Table des probabilités normales réduite, en annexe. En consultant la table, on constate que :

Probabilité cumulative d'une valeur de Z

(La valeur de la table est de 0,9608, mais pour tenir la promesse faite à la section 3.2.3, nous l'avons arrondie à deux décimales, soit 0,96.)

Il suffit de consulter la table deux fois et de faire une soustraction pour obtenir :

Probabilité entre deux valeurs de Z : P [0,57 < Z < 1,32] = P [Z < 1,32] − P [Z ≤ 0,57]

$= \Phi(1.32) - \Phi(0.57)$

= 0,9066 − 0,7157

= 0,19

Il suffit de consulter la table une seule fois et de faire une soustraction pour obtenir une probabilité de queue droite :

Probabilité de queue droite d'une valeur Z

P [Z > −0,89] = 1 − P [Z ≤−0,89] = 1 − 0,1867 = 0,81

Pour ces exemples, nous avons trouvé les valeurs de Z dans les marges, puis utilisé les valeurs dans le corps de la table, mais on peut inverser ce processus. En effet, il suffit de localiser une probabilité dans le corps de la table pour trouver la cote z correspondante dans les marges. Par exemple, considérons la position d'une cote z telle que

P [−z < Z < z] = 0,95.

Il y aura donc une probabilité de $\frac{1-0.95}{2}$ = 0,025 que z se trouve dans la queue droite de la distribution normale réduite. Autrement dit, on aura $\Phi(z) = 0{,}975$. En repérant 0,975 dans le corps de la table, on constate que z = 1,96.

Cela revient à trouver le quantile 0,975 de la distribution normale réduite, ce qui nous permet de comprendre et de décrire les quantiles normaux réduits.

La figure 4.1.4.2 illustre tous les calculs de cet exemple.

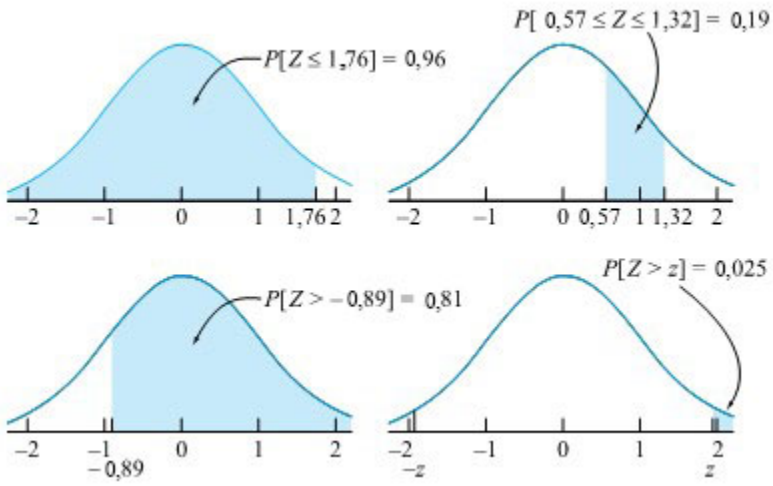


Figure 4.1.4.2. Probabilités normales réduites pour l'exemple 4.7.1.

Exemple 4.1.4.2. Poids nets des pots d'aliments pour bébés

Dans son article « Computer Assisted Net Weight Control » (Quality Progress, juin 1983), J. Fisher traite du remplissage (en grammes) de pots de nourriture. L'article présente un histogramme raisonnablement en forme de cloche des poids nets individuels des pots de nourriture pour bébé. La moyenne des valeurs représentées est d'environ 137,2 g ,et l'écart-type est d'environ 1,6 g. Le poids déclaré (ou étiqueté) sur les pots de ce produit est de 135,0 g.

Supposons qu'il soit adéquat de modéliser

W = le poids du prochain pot

par une distribution normale avec μ = 137,2 et σ = 1,6. Supposons en outre qu'on s'intéresse à la probabilité que le prochain pot rempli soit inférieur au poids déclaré (c'est-à-dire P [W < 135,0]). En utilisant l'expression 4.1.4.3, on trouve la cote z correspondant à w = 135,0 :

$$z = \frac{135{,}0 - 137{,}2}{1{,}6} = -1{,}38$$

Ensuite, à l'aide de la table A1.1 en annexe, on trouve que :

<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/e376f007fdb232013207d123f987e4b9.png » alt= »P[W<135,0]=\Phi(-1,38)=0,08" title= »P[W

Selon ce modèle, le risque d'obtenir un niveau de remplissage inférieur à la valeur nominale est d'environ 8 %.

En guise de deuxième exemple, considérons la probabilité que W se situe à moins d'un gramme de la valeur nominale (P [134,0 < W < 136,0]). Avec l'équation 4.1.4.3, on trouve les cotes z de w_1 = 134,0 et w_2 = 136,0 :

$$z_1 = \frac{134{,}0 - 137{,}2}{1{,}6} = -2{,}00$$

$$z_2 = \frac{136.0 - 137.2}{1.6} = -.75$$

Par conséquent,

Les deux probabilités précédentes et leurs contreparties normales réduite sont illustrées à la figure 4.1.4.3.

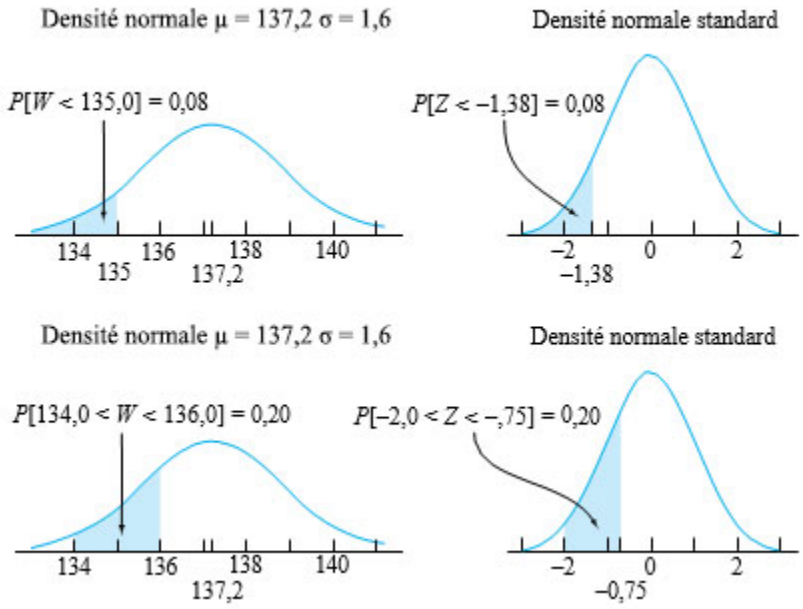


Figure 4.1.4.3. Probabilités normales pour l'exemple 4.1.4.2.

Exemple 4.1.4.3. Poids nets des pots de nourriture pour bébé (suite)

Pour les calculs de cet exemple, nous sommes partis des quantités du côté droit de l'équation 4.1.4.3 et des marges de la table A1.1 pour trouver les probabilités associées à diverses valeurs de W. Une variante importante de ce processus consiste à passer du corps de la table aux marges pour obtenir z, puis d'utiliser deux des trois variables du côté droit de l'équation 4.1.4.3 pour trouver la troisième.

Par exemple, supposons qu'il soit facile d'ajuster la cible du processus de remplissage (c'est-à-dire la moyenne μ de W) et que l'on veuille réduire la probabilité que le prochain pot soit inférieur au poids déclaré de 135,0 à 0,01 en augmentant μ. Quelle est la valeur minimale de μ qui permet d'atteindre cet objectif (en supposant que σ constant à 1,6 g)?

La figure 4.1.4.4 montre ce qu'il faut faire : on doit choisir μ telle que w = 135,0 corresponde au quantile 0,01 de la distribution normale de moyenne μ et d'écart-type σ = 1,6. En consultant la table A1.1, il est facile de déterminer que le quantile 0,01 de la distribution normale réduite est :

$$z = Q_z(0{,}01) = -2{,}33$$

Ainsi, selon l'équation 4.1.4.3, on obtient

$$-2{,}33 = \frac{135{,}0 - \mu}{1{,}6}$$

d'où μ = 138,7 g.

Il faut donc augmenter le remplissage nominal d'environ 138,7 – 137,2 = 1,5 g.

En pratique, la réduction de P [W < 135,0] s'accompagne d'une augmentation des coûts, puisque les pots contiennent en moyenne beaucoup plus que leur contenu nominal. Dans certaines applications, ce type de coût sera prohibitif. Il faudra adopter une autre approche : réduire la variation du niveau de remplissage en achetant un équipement plus précis. Selon l'équation 4.1.4.3, au lieu d'augmenter μ, on pourrait envisager de payer le coût associé à la réduction de σ. Vous pouvez vérifier qu'une réduction de σ à environ 0,94g produirait également P [W < 135,0] = 0,01 sans aucun changement dans μ.

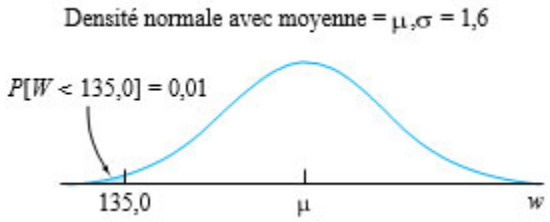


Figure 4.1.4.4. Distribution normale et P[W < 135,0] = 0,01

Comme l'illustrent ces exemples, l'équation 4.1.4.3 est la relation fondamentale utilisée dans les problèmes impliquant des distributions normales. D'une manière ou d'une autre, on obtient trois des quatre variables de l'équation sont spécifiées, et on cherche la quatrième.

4.1.5 La règle empirique

LA RÈGLE EMPIRIQUE

Si X est une variable aléatoire suivant une distribution normale avec une moyenne de μ et un écart-type de σ, alors la règle empirique énonce ce qui suit :

- Environ 68% des valeurs de x se situent entre $-\sigma$ et $+\sigma$ de la moyenne μ (à moins d'un écart-type de la moyenne).
- Environ 95% des valeurs de x se situent entre -2σ et $+2\sigma$ de la moyenne μ (à moins de deux écarts-types de la moyenne).
- Environ 99,7% des valeurs de x se situent entre -3σ et $+3\sigma$ de la moyenne μ (à moins de trois écarts-types de la moyenne). Autrement dit, presque toutes les valeurs x se situent à moins de trois écarts types de la moyenne.
- Les cotes z pour $+\sigma$ et $-\sigma$ sont +1 et -1, respectivement.
- Les cotes z pour $+2\sigma$ et -2σ sont +2 et -2, respectivement.
- Les cotes z pour $+3\sigma$ et -3σ sont +3 et -3, respectivement.

La règle empirique est également connue sous le nom de règle 68-95-99,7. Elle est illustrée à la figure 4.1.5.1.

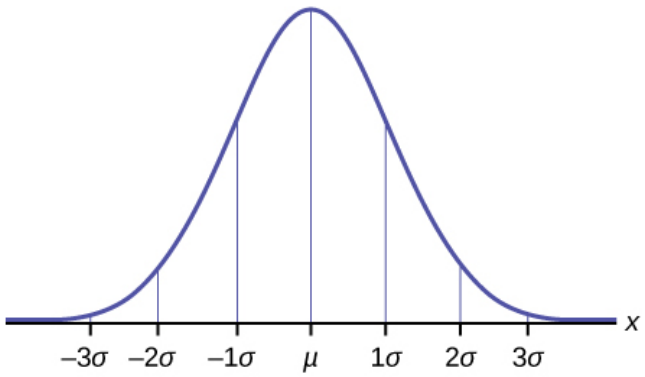


Figure 4.1.5.1. La règle empirique.

4.1.6 Tutoriel 3 – Distributions normales de probabilités

À ce stade, il est recommandé de travailler sur l'exercice du tutoriel 3 qui se trouve sur le référentiel GitHub. Cet exercice vous permettra de vous familiariser avec le calcul en Python des probabilités à l'aide de la distribution normale standard.

Il est fortement recommandé de consulter les fichiers Normal Probability & Confidence Intervals de Jupyter Notebook. Vous pouvez les trouver dans la section « How do I do X in Python? ». Le fichier « Standard Normal Distribution in Python » sera particulièrement utile.

4.2.0 Distributions conjointes et indépendance – Introduction

La plupart des applications des probabilités dans le domaine des statistiques d'ingénierie impliquent non pas une mais plusieurs variables aléatoires. Dans certains cas, l'application comporte intrinsèquement plusieurs variables. Il est alors logique de considérer que plusieurs variables du processus sont soumises à des influences aléatoires et d'évaluer les probabilités associées quand on les combine. Prenons l'exemple de l'assemblage d'une bague de roulement d'un diamètre intérieur nominal de 1,00 po sur une tige d'un diamètre nominal de 0,99 po. Si :

X = le diamètre intérieur de la bague de roulement

Y = le diamètre de la tige

on peut s'intéresser à

$$P [X < Y] = P [\text{il y a une interférence dans l'assemblage}]$$

qui implique les deux variables.

Mais même lorsqu'une situation est ne comporte qu'une seule variable, on utilise pratiquement toujours des échantillons de taille plus grande que 1. Les n valeurs de données d'un échantillon sont généralement considérées comme soumises à des causes aléatoires, et il faut modéliser leur comportement simultané. Les méthodes vues jusqu'à présent ne peuvent traiter qu'une seule variable aléatoire à la fois. Pour créer des méthodes permettant de décrire plusieurs variables aléatoires simultanément, l faut les généraliser.

Des livres entiers sont consacrés à divers aspects de la modélisation simultanée de nombreuses variables aléatoires. Cette section ne peut donner qu'une brève introduction au sujet. Nous discuterons ici du cas relativement simple de variables aléatoires discrètes conjointes, des fonctions de probabilité conjointes et marginales, des distributions conditionnelles et de l'indépendance. Ces sujets seront abordés en s'appuyant des exemples à deux variables simples.

Les concepts de fonctions de densité de probabilité conjointe et marginale, de distributions conditionnelles et d'indépendance des variables aléatoires continues conjointes ne sont pas abordés dans ce cours, mais ils sont analogues ce que nous verrons ici.

4.2.1 Distributions conjointes

DESCRIPTION DES VARIABLES ALÉATOIRES DISCRÈTES CONJOINTES

Pour plusieurs variables discrètes, on spécifie généralement les probabilités à l'aide d'une **fonction de probabilité conjointe**, qui se définit comme suit :

DÉFINITION 4.2.1.1. Fonction de probabilité conjointe
EXPRESSION 4.2.1.1

Une fonction de probabilité conjointe pour les variables aléatoires discrètes X et Y est une fonction non-négative $f(x, y)$, qui calcule la probabilité que X prenne la valeur x et que Y prenne la valeur y (simultanément). Autrement dit,

$$f(x, y) = P[X = x \text{ et } Y = y]$$

Exemple 4.2.1.1. Distribution de probabilité conjointe de deux couples de boulons

Reprenons l'étude de Brenny, Christensen et Schneider sur la mesure, à l'entier le plus proche, du couple des boulons de la plaque avant d'un composant d'équipement lourd. Soient

$X =$ le prochain couple enregistré pour le boulon 3

et

$Y =$ le prochain couple enregistré pour le boulon 4

Les données présentées dans le tableau et la figure précédents suggèrent, par exemple, qu'une valeur raisonnable de $P[X = 18 \text{ et } Y = 18]$ pourrait être $\frac{1}{34}$, la fréquence relative de cette paire dans l'ensemble de données. De même, les valeurs

$$P[X = 18 \text{ et } Y = 17] = \frac{2}{34}$$
$$P[X = 14 \text{ et } Y = 9] = 0$$

correspondent également aux fréquence relatives observées.

Si l'on est prêt à accepter que l'ensemble des fréquences relatives définies par les données des étudiant.e.s correspondent aux probabilités de X et de Y, ces probabilités peuvent être rassemblées de manière pratique dans un tableau à deux dimensions spécifiant une fonction de probabilité conjointe pour X et Y. Cette fonction est illustrée dans le tableau 4.2.1.1. (Pour alléger le tableau, les entrées « 0 » ont été laissées en blanc).

$f(x, y)$ pour le problème du couple des boulons

$y \setminus x$	11	12	13	14	15	16	17	18	19	20
20								2/34	2/34	1/34
19							2/34			
18			1/34	1/34			1/34	1/34	1/34	
17					2/34	1/34	1/34	2/34		
16				1/34	2/34	2/34			2/34	
15	1/34	1/34			3/34					
14					1/34			2/34		
13					1/34					

Tableau 4.2.1.1.

PROPRIÉTÉS D'UNE FONCTION DE PROBABILITÉ CONJOINTE POUR X ET Y

La fonction de probabilité présentée sous forme de table dans le tableau 4.2.1.1 possède deux propriétés requises pour qu'elle soit mathématiquement cohérente : les valeurs de $f(x, y)$ sont toutes comprises dans l'intervalle $[0, 1]$, et leur somme vaut 1. Pour obtenir la probabilité d'une configuration d'intérêt $X\,Y$ donnée, il suffit d'additionner les valeurs correspondantes de $f(x, y)$.

Exemple 4.2.1.2 Exemple de couples de boulons (suite)

Utilisons la distribution conjointe du tableau 4.2.1.1 pour évaluer

$$P[X \geq Y],$$
$$P[|X - Y| \leq 1],$$
$$\text{et } P[X = 17]$$

Commençons par $P[X \geq Y]$, la probabilité que le couple du boulon 3 soit au moins aussi important que le couple du boulon 4. La figure 4.2.1.1 indique par des astérisques les combinaisons possibles de x et y qui satisfont ce critère, En se référant au tableau 4.2.1.1 et en additionnant les entrées correspondant aux cellules contenant des astérisques, on obtient :

$$\begin{aligned} P[X \geq Y] =& f(15,13) + f(15,14) + f(15,15) + f(16,16) \\ &+ f(17,17) + f(18,14) + f(18,17) + f(18,18) \\ &+ f(19,16) + f(19,18) + f(20,20) \\ =& \frac{1}{34} + \frac{1}{34} + \frac{3}{34} + \frac{2}{34} + \cdots + \frac{1}{34} = \frac{17}{34} \end{aligned}$$

Un raisonnement similaire permet d'évaluer $P[|X - Y| \leq 1]$-la probabilité que les couples des boulons 3 et 4 se situent à 1 pi lb l'un de l'autre. La figure 4.2.1.2 illustre les combinaisons de x et y avec une différence absolue de 0 ou 1. Ensuite, on additionne les probabilités correspondant à ces combinaisons :

$$P[|X-Y| \leq 1] = f(15,14) + f(15,15) + f(15,16) + f(16,16) + f(16,17) + f(17,17) + f(17,18) + f(18,17) + f(18,18) + f(19,18) + f(19,20) + f(20,20) = \frac{18}{34}$$

y \ x	11	12	13	14	15	16	17	18	19	20
20										*
19									*	*
18								*	*	*
17							*	*	*	*
16						*	*	*	*	*
15					*	*	*	*	*	*
14				*	*	*	*	*	*	*
13			*	*	*	*	*	*	*	*

Figure 4.2.1.1. Combinaisons des couples des boulons 3 et 4 pour lesquels $x \geq y$

.

y \ x	11	12	13	14	15	16	17	18	19	20
20									*	*
19								*	*	*
18							*	*	*	
17						*	*	*		
16					*	*	*			
15				*	*	*				
14			*	*	*					
13		*	*	*						

Figure 4.2.1.2. Combinaisons des couples des boulons 3 et 4 pour lesquels $|x - y| \leq 1$.

Enfin, $P[X = 17]$, la probabilité que le couple mesuré sur le boulon 3 soit 17 pi lb , s'obtient en additionnant la colonne $x = 17$ dans le tableau 4.2.1.1. Autrement dit,

$$\begin{aligned} P[X = 17] &= f(17, 17) + f(17, 18) + f(17, 19) \\ &= \frac{1}{34} + \frac{1}{34} + \frac{2}{34} \\ &= \frac{4}{34} \end{aligned}$$

OBTENIR UNE FONCTION DE PROBABILITÉ MARGINALE À PARTIR D'UNE FONCTION DE PROBABILITÉS CONJOINTES À DEUX VARIABLES

Dans les problèmes à deux variables comme celui-ci, on peut additionner les colonnes d'un tableau à deux entrées de $f(x, y)$ pour obtenir les valeurs de la fonction de probabilité de X, $f_X(x)$. On peut également additionner les lignes du même tableau pour obtenir les valeurs de la fonction de probabilité de Y, $f_Y(y)$. On peut alors inscrire ces sommes dans les marges du tableau à double entrée, d'où l'appellation « distributions marginales ». L'encadré qui suit définit la terminologie utilisée dans le cas d'un problème à deux variables discrètes.

DÉFINITION 4.2.1.2. Fonction de probabilité marginale
EXPRESSION 4.2.1.2

Les fonctions de probabilité individuelle des variables aléatoires discrètes X et Y obéissant à une fonction de probabilité conjointe $f(x, y)$ sont désignées sous le terme **fonctions de probabilité marginale**. Elles sont obtenues en additionnant les valeurs de $f(x, y)$ avec toutes les valeurs possibles de l'autre variable. Autrement dit, la fonction de probabilité marginale de X est

$$f_X(x) = \sum_y f(x, y)$$

et la fonction de probabilité marginale pour Y est

$$f_Y(y) = \sum_x f(x, y)$$

.

Example 4.2.1.3, suite.

Le tableau 4.2.1.2 est une copie du tableau 4.2.1.1, auquel on a ajouté les probabilités marginales de X et Y. En séparant les marges du tableau à double entrée, on obtient des tableaux de probabilités marginales dans le format habituel. Par exemple, la fonction de probabilité marginale de Y est présentée séparément dans le tableau 4.2.1.3 .

Probabilités jointes et marginales pour X et Y

y \ x	11	12	13	14	15	16	17	18	19	20	$f_Y(y)$
20								2/34	2/34	1/34	5/34
19							2/34				2/34
18			1/34	1/34			1/34	1/34	1/34		5/34
17					2/34	1/34	1/34	2/34			6/34
16				1/34	2/34	2/34			2/34		7/34
15	1/34	1/34			3/34						5/34
14					1/34			2/34			3/34
13					1/34						1/34
$f_X(x)$	1/34	1/34	1/34	2/34	9/34	3/34	4/34	7/34	5/34	1/34	

Tableau 4.2.1.2.

Fonction de probabilité marginale pour Y

y	$f_Y(y)$
13	1/34
14	3/34
15	5/34
16	7/34

Tableau 4.2.1.3.

L'obtention de fonctions de probabilité marginales à partir de fonctions de probabilité conjointes soulève la question logique de savoir si le processus peut être inversé. Autrement dit, si $f_X(x)$ et $f_Y(y)$ sont connues, y a-t-il alors une seule option pour $f(x, y)$? De fait, non. La figure 4.2.1.3 montre deux distributions conjointes à deux variables très différentes qui possèdent néanmoins les mêmes distributions marginales. La différence marquée entre les distributions de la figure 4.2.1.3 est liée au comportement conjoint, plutôt qu'individuel, de X et de Y.

Distribution 1

y \ x	1	2	3	
3	0,4	0	0	0,4
2	0	0,4	0	0,4
1	0	0	0,2	0,2
	0,4	0,4	0,2	

Distribution 2

y \ x	1	2	3	
3	0,16	0,16	0,08	0,4
2	0,16	0,16	0,08	0,4
1	0,08	0,08	0,04	0,2
	0,4	0,4	0,2	

Figure 4.2.1.3. Deux distributions conjointes différentes avec les mêmes distributions marginales.

4.2.2 Distributions conditionnelles et indépendance

DISTRIBUTIONS CONDITIONNELLES ET INDÉPENDANCE DES VARIABLES ALÉATOIRES DISCRÈTES

Lorsqu'on travaille avec plusieurs variables aléatoires, il est souvent utile de réfléchir à ce que l'on attend de l'une des variables compte tenu des valeurs prises par toutes les autres. Par exemple, dans l'exercice du couple de serrage du boulon (X), un technicien qui vient de desserrer le boulon 3 et qui a mesuré le couple à la valeur 15 pi lb devrait avoir des attentes pour le couple du boulon 4 (Y) quelque peu différentes, à la lumière de la distribution marginale du tableau 4.2.1.3. Après tout, si on reprend les données du tableau 4.2.2.1, la distribution de la fréquence relative des couples des boulons 4 pour les composants dont le couple du boulon 3 est de 15 pi lb est similaire aux valeurs du tableau 4.2.2.1. D'une certaine manière, le fait de savoir que $X = 15$ devrait modifier la distribution de probabilité de Y pour que la distribution de fréquence relative corresponde au tableau 4.2.2.1 plutôt qu'à la distribution marginale du tableau 4.1.1.3.

Distribution de la fréquence relative pour le couple du boulon 4 (couple du boulon 3 = 15 pi lb)

y, couple (pi lb)	Fréquence relative
13	1/9
14	1/9
15	3/9
16	2/9
17	2/9

Tableau 4.2.2.1

La théorie des probabilités tient compte de cette notion de « distribution d'une variable lorsqu'on connaît les valeurs des autres » à travers le concept de distribution conditionnelle. La version à deux variables est définie ci-après.

DÉFINITION 4.2.2.1. Fonction de probabilité conditionnelle de X étant donné que Y=y
EXPRESSION 4.2.2.1

Pour des variables aléatoires discrètes X et Y avec une fonction de probabilité conjointe $f(x, y)$, la fonction de probabilité conditionnelle de X étant donné que $Y = y$ est la fonction de x suivante :

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{\sum_x f(x, y)}$$

La fonction de probabilité conditionnelle de Y étant donné que $X = x$ est la fonction de y suivante :

Formula does not parse

En comparant les définitions 4.2.1.1 et 4.2.2.1, on obtient :

Fonction de probabilité conditionnelle de X étant donné que Y=y **4.2.2.2**

$$f_{X|Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}$$

et

Fonction de probabilité conditionnelle pour Y étant donné que X=x **4.2.2.3**

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

Calcul de distributions conditionnelles à partir d'une fonction de probabilité conjointe

Les équations 4.2.2.2 et 4.2.2.3 sont parfaitement logiques. La première indique qu'à partir d'une fonction $f(x, y)$ répertoriée dans un tableau à deux entrées et en ne considérant que la ligne $Y = y$, la distribution conditionnelle appropriée pour X est indiquée par les probabilités de cette ligne (les valeurs de $f(x, y)$), qu'on divise par leur somme $(f_Y(y) = \sum_x f(x, y)$) pour les renormaliser (faire en sorte qu'elles totalisent 1). De même, l'équation 4.2.2.3 indique que si l'on considère uniquement la colonne $X = x$, la distribution conditionnelle appropriée pour Y est donnée par les probabilités de cette colonne divisées par leur somme.

Exemple 4.2.2.1. Couples des boulons (suite)

Pour illustrer l'utilisation des équations 4.2.2.2 et 4.2.2.3, considérons quelques-unes des distributions conditionnelles associées à la distribution conjointe des couples des boulons 3 et 4, en commençant par la distribution conditionnelle de

Y étant donné que $X = 15$.

À partir de l'équation 4.2.2.3,

$$f_{Y|X}(y \mid 15) = \frac{f(15, y)}{f_X(15)}$$

En se référant au tableau 4.2.1.2, la probabilité marginale associée à $X = 15$ est $\frac{9}{34}$. Ainsi, en divisant les valeurs dans la colonne $X = 15$ de ce tableau par $\frac{9}{34}$, on obtient la distribution conditionnelle pour Y, qui est présentée dans le tableau 4.2.2.2. Si l'on compare ce résultat au tableau 4.2.1.4, on constate que l'équation 4.2.2.3 produit une distribution conditionnelle conforme à l'intuition.

Fonction de probabilité conditionnelle pour Y lorsque $X = 15$

y	$f_{Y\|X}(y \mid 15)$
13	$\left(\frac{1}{34}\right) \div \left(\frac{9}{34}\right) = \frac{1}{9}$
14	$\left(\frac{1}{34}\right) \div \left(\frac{9}{34}\right) = \frac{1}{9}$
15	$\left(\frac{3}{34}\right) \div \left(\frac{9}{34}\right) = \frac{3}{9}$
16	$\left(\frac{2}{34}\right) \div \left(\frac{9}{34}\right) = \frac{2}{9}$
17	$\left(\frac{2}{34}\right) \div \left(\frac{9}{34}\right) = \frac{2}{9}$

Tableau 4.2.2.2

Considérons ensuite $f_{Y|X}(y \mid 18)$f_{Y \mid X}(y \mid 18) :

$$f_{Y|X}(y \mid 18) = \frac{f(18, y)}{f_X(18)}$$

Le tableau 4.2.1.2 nous donne la distribution conditionnelle de Y étant donné que $X = 18$, présentée dans le tableau 4.2.2.3. Les tableaux 4.2.2.2 et 4.2.4.3 confirment que les distributions conditionnelles de Y étant donné que $X = 15$ et étant donné que $X = 18$ sont très différentes. Par exemple, si on sait que $X = 18$, on s'attend à ce que Y soit plus grand que lorsque $X = 15$.

.

Fonction de probabilité conditionnelle pour Y lorsque $X = 18$

y	$f_{Y\|X}(y \mid 18)$
14	2/7
17	2/7
18	1/7
20	2/7

Tableau 4.2.2.3.

Pour s'assurer que la signification de l'équation 4.2.2.2 est également claire, considérons la distribution conditionnelle du couple du boulon 3 (X) étant donné que le couple du boulon 4 est de 20 $(Y = 20)$. Dans cette situation,

l'équation 4.2.2.2 donne :

$$f_{X|Y}(x \mid 20) = \frac{f(x, 20)}{f_Y(20)}$$

(Les probabilités conditionnelles pour X sont les valeurs de la ligne $Y = 20$ du tableau 4.2.1.2 divisées par la valeur marginale de $Y = 20$.) $f_{X|Y}(x \mid 20)$ est répertoriée dans le tableau 4.2.2.4.

Fonction de probabilité conditionnelle pour X lorsque $Y = 20$

x	$f_{X\|Y}(x \mid 20)$
18	$\left(\frac{2}{34}\right) \div \left(\frac{5}{34}\right) = \frac{2}{5}$
19	$\left(\frac{2}{34}\right) \div \left(\frac{5}{34}\right) = \frac{2}{5}$
20	$\left(\frac{1}{34}\right) \div \left(\frac{5}{34}\right) = \frac{1}{5}$

Tableau 4.2.2.4.

L'exemple du couple des boulons présente la particularité que les distributions conditionnelles de Y étant donné les valeurs possibles pour X sont différentes. En outre, ces distributions ne sont généralement pas identiques à la distribution marginale de Y. X. X fournit des informations à propos de Y, en ce sens que selon sa valeur, il existe différentes évaluations de probabilité pour Y. Comparez cette situation à l'exemple suivant.

Exemple 4.2.2.2. Échantillonnage aléatoire de deux couples du boulon 4

Supposons que les couples de 34 boulons 4 obtenus par Brenny, Christensen et Schneider et figurant dans le tableau 4.2.2.5 soient inscrits sur des bouts de papier et placés dans un chapeau. Supposons en outre que les papiers soient mélangés, qu'on en choisisse un, qu'on note le couple correspondant et qu'on replace le papier dans le chapeau. Ensuite, on mélange les papiers, on en sélectionne un autre, et on note le deuxième couple. Soient les deux variables aléatoires suivantes :

$$U = \text{ la valeur du premier couple sélectionné}$$

et

$$V = \text{ valeur du deuxième couple sélectionné}$$

Couple requis pour desserrer deux boulons (pi lb)

Composant	Couple du boulon 3	Couple du boulon 4	Composant	Couple du boulon 3	Couple du boulon 4
1	16	16	18	15	14
2	15	16	19	17	17
3	15	17	20	14	16
4	15	16	21	17	18
5	20	20	22	19	16
6	19	16	23	19	18
7	19	20	24	19	20
8	17	19	25	15	15
9	15	15	26	12	15
10	11	15	27	18	20
11	17	19	28	13	18
12	18	17	29	14	18
13	18	14	30	18	18
14	15	15	31	18	14
15	18	17	32	15	13
16	15	17	33	16	17
17	18	20	34	16	16

Tableau 4.2.2.5.

On comprend intuitivement que, contrairement aux situations de X et Y de l'exemple 4.2.2.1, les variables U et V ne fournissent aucune information l'une sur l'autre. Quelle que soit la valeur de U, la distribution de fréquence relative des couples du boulon 4 dans le chapeau est correcte comme distribution de probabilité (conditionnelle) pour V, et inversement. En d'autres termes, non seulement U et V partagent la distribution marginale commune du tableau 4.2.2.6, mais il est également vrai que pour tout u et tout v, on a :

4.2.2.4 $$f_{U|V}(u \mid v) = f_U(u)$$

et

4.2.2.5 $$f_{V|U}(v \mid u) = f_V(v)$$

Les équations 4.2.2.4 et 4.2.2.5 indiquent que les probabilités marginales du tableau 4.2.2.6 servent également de probabilités conditionnelles. Elles précisent également comment les probabilités conjointes des U and V doivent être structurées, puisqu'en réécrivant le côté gauche de l'équation 4.2.2.4 à l'aide de l'expression 4.2.2.2, on obtient :

$$\frac{f(u,v)}{f_V(v)} = f_U(u)$$

Autrement dit :

4.2.2.6 $$f(u,v) = f_U(u) f_V(v)$$

(La même logique appliquée à l'équation 4.2.2.5 conduit également à l'équation 4.2.2.6). L'expression 4.2.2.6 indique que les valeurs de probabilité conjointe pour U et V s'obtiennent en multipliant les probabilités marginales correspondantes. Le tableau 4.2.2.7 donne la fonction de probabilité conjointe pour U et V.

Fonction de probabilité marginale courante pour U et V

u ou v	$f_U(u)$ ou $f_V(v)$
13	1/34
14	3/34
15	5/34
16	7/34
17	6/34
18	5/34
19	2/34
20	5/35

Tableau 4.2.2.6.

Probabilités jointes pour U et V

v \ u	13	14	15	16	17	18	19	20	$f_V(v)$
20	$\frac{5}{(34)^2}$	$\frac{15}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{35}{(34)^2}$	$\frac{30}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{10}{(34)^2}$	$\frac{25}{(34)^2}$	5/34
19	$\frac{2}{(34)^2}$	$\frac{6}{(34)^2}$	$\frac{10}{(34)^2}$	$\frac{14}{(34)^2}$	$\frac{12}{(34)^2}$	$\frac{10}{(34)^2}$	$\frac{4}{(34)^2}$	$\frac{10}{(34)^2}$	2/34
18	$\frac{5}{(34)^2}$	$\frac{15}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{35}{(34)^2}$	$\frac{30}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{10}{(34)^2}$	$\frac{25}{(34)^2}$	5/34
17	$\frac{6}{(34)^2}$	$\frac{18}{(34)^2}$	$\frac{30}{(34)^2}$	$\frac{42}{(34)^2}$	$\frac{36}{(34)^2}$	$\frac{30}{(34)^2}$	$\frac{12}{(34)^2}$	$\frac{30}{(34)^2}$	6/34
16	$\frac{7}{(34)^2}$	$\frac{21}{(34)^2}$	$\frac{35}{(34)^2}$	$\frac{49}{(34)^2}$	$\frac{42}{(34)^2}$	$\frac{35}{(34)^2}$	$\frac{14}{(34)^2}$	$\frac{35}{(34)^2}$	7/34
15	$\frac{5}{(34)^2}$	$\frac{15}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{35}{(34)^2}$	$\frac{30}{(34)^2}$	$\frac{25}{(34)^2}$	$\frac{10}{(34)^2}$	$\frac{25}{(34)^2}$	5/34
14	$\frac{3}{(34)^2}$	$\frac{9}{(34)^2}$	$\frac{15}{(34)^2}$	$\frac{21}{(34)^2}$	$\frac{18}{(34)^2}$	$\frac{15}{(34)^2}$	$\frac{6}{(34)^2}$	$\frac{15}{(34)^2}$	3/34
13	$\frac{1}{(34)^2}$	$\frac{3}{(34)^2}$	$\frac{5}{(34)^2}$	$\frac{7}{(34)^2}$	$\frac{6}{(34)^2}$	$\frac{5}{(34)^2}$	$\frac{2}{(34)^2}$	$\frac{5}{(34)^2}$	1/34
$f_U(u)$	1/34	3/34	5/34	7/34	6/34	5/34	2/34	5/34	

Tableau 4.2.2.7.

INDÉPENDANCE DES OBSERVATIONS DANS LES ÉTUDES STATISTIQUES

L'exemple 4.2.2.2 suggère qu'on peut formaliser la notion intuitive que pour des variables aléatoires non liées, les distributions conditionnelles sont toutes égales aux distributions marginales correspondantes. De manière équivalente, on peut dire que les probabilités conjointes sont les produits des probabilités marginales correspondantes. Formellement, dans ce genre de cas, on parle de variables aléatoires **indépendantes**. La définition pour le cas à deux variables est la suivante.

DÉFINITION 4.2.2.7. Indépendance des variables aléatoires
EXPRESSION 4.2.2.7

Les variables aléatoires discrètes X et Y sont dites indépendantes si leur fonction de probabilité conjointe $f(x, y)$ est le produit de leurs fonctions de probabilité marginales respectives. Autrement dit, l'indépendance signifie que

$$f(x, y) = f_X(x) f_Y(y) \quad \text{pour tout } y$$

Si l'équation 4.2.2.7 n'est pas valide, les variables X et Y sont dite dépendantes.

(L'équation 4.2.2.7 implique que les distributions conditionnelles sont toutes égales à leurs fonctions marginales correspondantes, de sorte que la définition correspond bien à sa motivation de « non-relation ».)
Les variables U et V de l'exemple 4.2.2.2 sont indépendantes, tandis que les variable X et Y de l'exemple 4.2.2.1 sont dépendantes. En outre, les deux distributions conjointes illustrées à la figure 4.2.1.3 donnent un exemple de distribution conjointe fortement dépendante (la première) et de distribution conjointe indépendante (la seconde) qui ont les mêmes fonctions marginales.
La notion d'indépendance est fondamentale. Les variables indépendantes *simplifient énormément les calculs*. L'hypothèse d'indépendance entre les observations est souvent appropriée lorsqu'on recueille des données d'ingénierie dans un contexte analytique en prenant soin de minimiser toutes les causes physiques évidentes d'effets de report susceptibles d'influencer les observations successives. De même, dans les contextes énumératifs, les échantillons aléatoires simples relativement petits (par rapport à la taille de la population) produisent des observations qui peuvent généralement être considérées comme au moins approximativement indépendantes.

Exemple 4.2.2.3. Exemple du couple des boulons (suite)

Imaginons à nouveau qu'on a inscrit les couples de boulons sur des bouts de papier dans un chapeau. La méthode de sélection du couple décrite précédemment pour produire U and V n'est pas un échantillonnage aléatoire simple. L'échantillonnage aléatoire simple tel que défini dans la partie 1 est un échantillonnage sans remplacement, et non la méthode d'échantillonnage avec remplacement utilisée pour produire U et V. En effet, si le premier papier n'est pas remplacé avant que le second ne soit sélectionné, les probabilités du tableau 4.2.2.7 ne décrivent pas U et V. Par exemple, si aucun remplacement n'est effectué, puisqu'un seul papier est étiqueté 13 pi lb , il faut clairement que

$$f(13,13) = P[U = 13 \quad \text{et} \quad V = 13] = 0$$

et non

$$f(13,13) = \frac{1}{(34)^2}$$

contrairement à ce qui est indiqué dans le tableau 4.2.2.7. En d'autres termes, si aucun remplacement n'est effectué, il est clair qu'il faut utiliser

$$f_{V|U}(13 \mid 13) = 0$$

plutôt que la valeur

$$f_{V|U}(13 \mid 13) = f_V(13) = \frac{1}{34}$$

ce qui serait approprié si l'échantillonnage était effectué avec remplacement. L'échantillonnage aléatoire simple ne conduit pas à des observations exactement indépendantes.

Mais supposons qu'au lieu de contenir 34 papiers, le chapeau contienne 100×34 papiers, en suivant la fréquence relative du tableau 4.2.2.6. Ainsi, même si l'échantillonnage est effectué sans remplacement, les probabilités développées précédemment pour U et V (et placées dans le tableau 4.2.2.7) restent aux moins approximativement valides. Par exemple, avec 3 400 papiers et en utilisant un échantillonnage sans remplacement, on a :

$$f_{V|U}(13 \mid 13) = \frac{99}{3,399}$$

Ensuite, comme

$$f_{V|U}(v \mid u) = \frac{f(u,v)}{f_U(u)}$$

on a :

$$f(u,v) = f_{V|U}(v \mid u) f_U(u)$$

sans remplacement, le calcul

$$f(13,13) = \frac{99}{3\ 399} \cdot \frac{1}{34}$$

est exact. Mais ce qu'il faut retenir, c'est que

$$\frac{99}{3,399} \approx \frac{1}{34}$$

et par conséquent,

$$f(13,13) \approx \frac{1}{34} \cdot \frac{1}{34}$$

Pour cette situation hypothétique où la taille de la population $N = 3,400$ est beaucoup plus grande que la taille de l'échantillon $n = 2$, l'indépendance est une description approximative appropriée des observations obtenues à l'aide d'un échantillonnage aléatoire simple.

Il y a d'autres termes pour décrire les variables indépendantes qui suivent la même distribution marginale.

VARIABLES ALÉATOIRES INDÉPENDANTES ET IDENTIQUEMENT DISTRIBUÉES

DÉFINITION 4.2.2.8. Variables indépendantes et identiquement distribuées.
Si les variables aléatoires $X_1, X_2, \ldots, X_n$ ont toutes la même distribution marginale et sont indépendantes, on dit qu'elles sont indépendantes et identiquement distribuées (iid).

Par exemple, la distribution conjointe de U et V donnée dans le tableau 4.2.2.7 indique que U et V sont des variables aléatoires iid.

Observations pouvant être modélisées comme des variables iid

Les exemples standard en statistiques de variables aléatoires iid sont les mesures successives d'un processus stable et les résultats d'un échantillonnage aléatoire avec
remplacement à partir d'une population unique. La question de savoir si un modèle iid est approprié dans une application statistique donnée dépend donc du fait que le mécanisme de génération de données étudié peut ou non être considéré comme conceptuellement équivalent à ces modèles.

4.2.3 Moyenne et variance des combinaisons linéaires de variables aléatoires

La section précédente a présenté les mathématiques utilisées pour modéliser simultanément plusieurs variables aléatoires. Une utilisation importante de ces outils en ingénierie concerne l'analyse des résultats de systèmes qui sont fonctions d'entrées aléatoires. Cette section porte sur la manière dont la variation d'une variable aléatoire de sortie dépend
des variables utilisées pour la produire. Nous nous concentrerons sur l'utilisation de combinaisons linéaires de variables aléatoires.

Distribution d'une fonction de variables aléatoires

Le problème examiné dans cette section est le suivant. Soient une distribution conjointe pour les variables aléatoires $X, Y, \ldots, Z$ et une fonction $g(x, y, \ldots, z)$. L'objectif est de prédire le comportement de la variable aléatoire

4.2.3.1 $$U = g(X, Y, \ldots, Z)$$

Dans certains cas très simples, il est possible de trouver exactement la distribution dont U hérite de $X, Y, \ldots, Z$

Exemple 4.2.3.1 Distribution du jeu entre deux pièces assemblées dont les dimensions sont déterminées de manière aléatoire

Supposons qu'une plaque d'acier d'une épaisseur nominale de 0,15 po doit reposer dans une rainure d'une largeur nominale de 0,155 po usinée sur la surface d'un bloc d'acier. Un lot de plaques a été fabriqué et les épaisseurs ont été mesurées, produisant la distribution de fréquence relative du tableau 4.2.3.1; une distribution de fréquence relative pour les largeurs de rainure mesurées sur un lot de blocs usinés est donnée dans le tableau 4.2.3.2.

Si une plaque est choisie au hasard et qu'un bloc est choisi séparément au hasard, la distribution conjointe naturelle pour les variables aléatoires

X = l'épaisseur de la plaque

Y = la largeur de la rainure

est indépendante, avec la distribution marginale de X donnée dans le tableau 4.2.3.1 et la distribution marginale de Y donnée dans le tableau 4.2.3.2. En d'autres termes, le tableau 4.2.3.3 donne une fonction de probabilité conjointe plausible pour X et Y.

Distribution de fréquence relative pour l'épaisseur des plaques

Épaisseur de la plaque (po)	Fréquence relative
0,148	0,4
0,149	0,3
0,150	0,3

Tableau 4.2.3.1

Distribution de fréquence relative de largeur des fentes

Largeur de fente (po)	Fréquence relative
0,153	0,2
0,154	0,2
0,155	0,4
0,156	0,2

Tableau 4.2.3.2

Une variable dérivée de X et Y qui présente un intérêt potentiel substantiel est le jeu dans l'assemblage plaque/bloc,

$$U = Y - X$$

Remarquez qu'en prenant les extrêmes représentés dans les tableaux 4.2.3.1 et 4.2.3.2, U doit se trouver dans la plage comprise entre (0,153 – 0,150 =) 0,003 po et *(0,156 – 0,148 =) 0,008 po. Cependant, on peut obtenir beaucoup plus d'informations. En examinant le tableau 4.2.3.3, on constate que les diagonales des entrées (du bas à gauche au haut à droite) correspondent toutes à la même valeur de Y – X. En additionnant les probabilités sur ces diagonales, on obtient la distribution de U donnée dans le tableau 4.2.3.4.

Probabilités marginales et jointes pour X et Y

$y \backslash x$	0,148	0,149	0,150	$f_Y(y)$
0,156	0,08	0,06	0,06	0,2
0,155	0,16	0,12	0,12	0,4
0,154	0,08	0,06	0,06	0,2
0,153	0,08	0,06	0,06	0,2
$f_X(x)$	0,4	0,3	0,3	

Tableau 4.2.3.3

Fonction de probabilité pour le jeu $U = Y - X$

u	$f(u)$
0,003	0,06
0,004	$0{,}12 = 0{,}06 + 0{,}06$
0,005	$0{,}26 = 0{,}08 + 0{,}06 + 0{,}12$
0,006	$0{,}26 = 0{,}08 + 0{,}12 + 0{,}06$
0,007	$0{,}22 = 0{,}16 + 0{,}06$
0,008	0,08

Tableau 4.2.3.4

L'exemple 4.2.3.1 implique une distribution conjointe discrète très simple et une fonction g très simple, à savoir, g(x, y) = y – x. En général, il n'est pas possible en pratique de trouver une solution complète et exacte pour la distribution de U = g(X, Y,...,Z). Heureusement, pour de nombreuses applications des probabilités en ingénierie, les solutions approximatives et/ou partielles suffisent à répondre aux questions d'intérêt pratique.

MOYENNE ET VARIANCE D'UNE COMBINAISON LINÉAIRE DE VARIABLES ALÉATOIRES

Pour les besoins de l'ingénierie, il suffit souvent de connaître la moyenne et la variance de U données par l'équation 4.2.3.1. (Autrement dit, nul besoin de connaître la distribution complète de U.) Lorsque c'est le cas et que g est linéaire, il existe des formules explicites pour trouver ces valeurs.

.

PROPOSITION 4.2.3.2

Soient $X, Y, \ldots, Z$ sont n variables aléatoires indépendantes, et $a_0, a_1, a_2, \ldots, a_n, n+1$ constantes. La variable aléatoire $U = a_0 + a_1 X + a_2 Y + \cdots + a_n Z$ a alors une moyenne de

4.2.3.3 $$EU = a_0 + a_1 EX + a_2 EY + \cdots + a_n EZ$$

et une variance de

.

4.2.3.4 $$VarU = a_1^2 \operatorname{Var} X + a_2^2 \operatorname{Var} Y + \cdots + a_n^2 VarZ$$

.

L'équation 4.2.3.3 est valide indépendamment du fait que les variables $X, Y, \ldots, Z$ soient dépendantes ou indépendantes. Par contre, l'équation 4.2.3.4 repose sur l'indépendance des variables, mais il existe une façon de la généraliser aux cas où les variables sont dépendantes. Cependant, la forme de l'équation 4.2.3.4 donnée ici est adéquate pour les besoins actuels.

.

Un type d'application dans lequel la proposition 4.2.3.2 est immédiatement utile est celui des problèmes de tolérances géométriques, auquel cas on a $a_0 = 0$, et les autres a_i.

.

Exemple 4.2.3.2 Jeu d'une plaque d'acier.

Reprenons la situation du jeu nécessaire pour insérer une plaque d'acier dans une rainure usinée sur un bloc d'acier. Soient X, Y, et U l'épaisseur de la plaque, la largeur de la rainure et le jeu (respectivement). Les moyennes et variances pour ces variables peuvent alors être calculées. Nous vous invitons à vérifier que

.

$$\begin{aligned} E(X) &= 0{,}1489 \quad \text{et} \quad \operatorname{Var}(X) = 6{,}9 \times 10^{-7} \\ E(Y) &= 0{,}1546 \quad \text{et} \quad \operatorname{Var}(Y) = 1{,}04 \times 10^{-6} \end{aligned}$$

.

Comme

.

$$U = Y - X = (-1)X + 1Y$$

.

Il est possible d'appliquer la proposition 4.2.3.2 et de conclure que

.

$$\begin{aligned} E(U) &= -1E(X) + 1E(Y) = -\ 0{,}1489 + \ 0{,}1546 = \ 0{,}0057\,\text{po} \\ \operatorname{Var}(U) &= (-1)^2 6{,}9 \times 10^{-7} + (1)^2 1{,}04 \times 10^{-6} = 1{,}73 \times 10^{-6} \end{aligned}$$

.

Ainsi,

.

$$\sqrt{\operatorname{Var}(U)} = \ 0{,}0013\ \text{po}$$

.

Il vaut la peine de vérifier que la moyenne et l'écart-type du jeu obtenu à l'aide de la proposition 4.2.3.2 concordent avec les valeurs obtenues à l'aide de la distribution de U figurant dans le tableau 4.2.3.4 et les formules de la moyenne et de la variance données dans la partie 3. L'avantage d'utiliser la proposition 4.2.3.2 est que si on a seulement besoin de EU et de $\operatorname{Var}(U)$, ce n'est pas nécessaire de passer par l'étape intermédiaire consistant à obtenir la distribution de U. Les

calculs effectués à l'aide de la proposition 4.2.3.2 n'utilisent que les caractéristiques des distributions marginales.

.

VARIABLES ALÉATOIRES $X_1, X_2, \ldots, X_n$ MODÉLISANT DES SÉLECTIONS ALÉATOIRES (AVEC REMPLACEMENT) DANS UNE MÊME POPULATION

Une autre utilisation particulièrement importante de la proposition 4.2.3.2 concerne n variables aléatoires iid pour lesquelles les coefficients a_i valent tous $\frac{1}{n}$. Autrement dit, lorsque les variables aléatoires $X_1, X_2, \ldots, X_n$ sont conceptuellement équivalentes à des sélections aléatoires (avec remplacement) dans une même population, la proposition 4.2.3.2 indique comment la moyenne et la variance de la variable aléatoire

.

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

.

sont liées aux paramètres de la population μ et σ^2. Pour les variables indépendantes $X_1, X_2, \ldots, X_n$ de même moyenne μ et de même variance σ^2, la proposition 4.2.3.2 énonce que :

.

4.2.3.5 Moyenne d'une moyenne de n variables aléatoires iid

.

$$E\bar{X} = \frac{1}{n}EX_1 + \frac{1}{n}EX_2 + \cdots + \frac{1}{n}EX_n = n\left(\frac{1}{n}\mu\right) = \mu$$

et

4.2.3.6 Variance d'une moyenne de n variables aléatoires iid

.

<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/a7fa64aca13a7c86d0d0875dfb4350c6.png » alt= »\begin{aligned}\operatorname{Var(\bar{X})} & =\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_1)} +\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_2)} +\cdots+\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_{n})} \\& =n\left(\frac{ 1}{n}\right)^2 \sigma^2=\frac{\sigma^2}{n}\end{aligned}

.

Étant donné que σ^2 / n » title= »\begin{aligned}\operatorname{Var(\bar{X})} & =\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_1)} +\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_2)} +\cdots+\left(\frac{ 1}{n}\right)^2 \operatorname{Var(X_{n})} \\& =n\left(\frac{ 1}{n}\right)^2 \sigma^2=\frac{\sigma^2}{n}\end{aligned}

.

Étant donné que σ^2 / n » class= »latex mathjax »> diminue en n\frac{ 1}{n}title, les équations 4.2.3.5 et 4.2.3.6 présente l'image rassurante de $\bar{X}$ ayant la distribution de probabilité centrée sur la moyenne de la population μ, avec une dispersion qui diminue au fur et à mesure que la taille de l'échantillon augmente.

.

Les relations 4.2.3.5 et 4.2.3.6, qui décrivent parfaitement le comportement aléatoire de $\bar{X}$[/latex] *sous échantillonnage aléatoire avec remplacement, sont aussi des descriptions approximatives du comportement de* [latex]$\bar{X}$ dans le cadre d'une échantillonnage aléatoire simple dans des contextes énumératifs. (Rappelons la discussion sur l'indépendance approximative des observations résultant d'un échantillonnage aléatoire simple dans une

grande population.)

4.2.4 Théorème central limite

EFFET CENTRAL LIMITE

L'une des statistiques les plus fréquemment utilisées dans les applications d'ingénierie est la moyenne de l'échantillon. Nous avons déjà évoqué les équations pour la moyenne et la variance de la distribution de probabilité d'une moyenne d'échantillon ou d'une observation unique lorsque le modèle de variables iid s'applique. L'un des faits les plus utiles de la probabilité appliquée est que si la taille de l'échantillon est raisonnablement grande, il est également possible d'approximer la forme de la distribution de probabilité de $\bar{X}$, quelle que soit la forme de la distribution sous-jacente des observations individuelles. Autrement dit, le fait suivant est avéré :

.

Proposition 4.2.4.1 Théorème central limite

Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires iid (avec une moyenne μ et une variance σ^2). Pour des échantillons à grand n, la variable $\bar{X}$ est approximativement normalement distribuée. (En d'autres termes, on peut approximer les probabilités de $\bar{X}$ avec une distribution normale de moyenne μ et de variance σ^2/n.)

.

La preuve de la proposition 4.2.4.1 dépasse le cadre de ce manuel, mais on peut en saisir intuitivement la notion à l'aide d'un exemple.

Exemple 4.2.4.1 Effet central limite et moyenne d'un échantillon de numéros de série d'outils (suite)

Reprenons l'exemple de la section 3.2.1.2 concernant le dernier chiffre du numéro de série d'outils pneumatiques sélectionnés de manière essentiellement aléatoire. Supposons que

.

$$W_1 = \text{le dernier chiffre du numéro de série observé lundi prochain à 9 h.}$$

.

$$W_2 = \text{le dernier chiffre du numéro de série observé le lundi suivant à 9 h.}$$

.

On peut raisonnablement supposer que les variables aléatoires W_1, W_2 sont indépendantes, chacune avec la fonction de probabilité marginale :

.

4.2.4.1 $$f(w) = \begin{cases} 0{,}1 & \text{si } w = 0, 1, 2, \ldots, 9 \\ 0 & \text{sinon} \end{cases}$$

.

Cette fonction de probabilité marginale est illustrée à la figure 4.2.4.1

.

Grâce à ce modèle, il est très facile de déduire que $\bar{W} = \frac{1}{2}(W_1 + W_2)$ a la fonction de probabilité donnée au tableau 4.2.4.1 et illustrée à la figure 4.2.4.2.

.

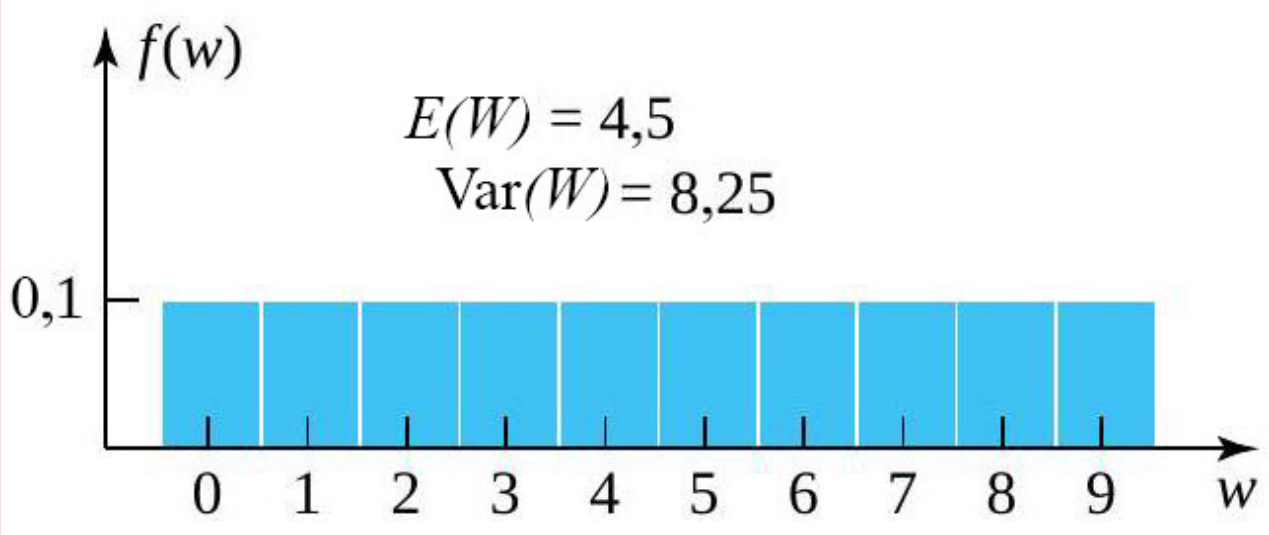


Figure 4.2.4.1 Histogramme de probabilité de W

.

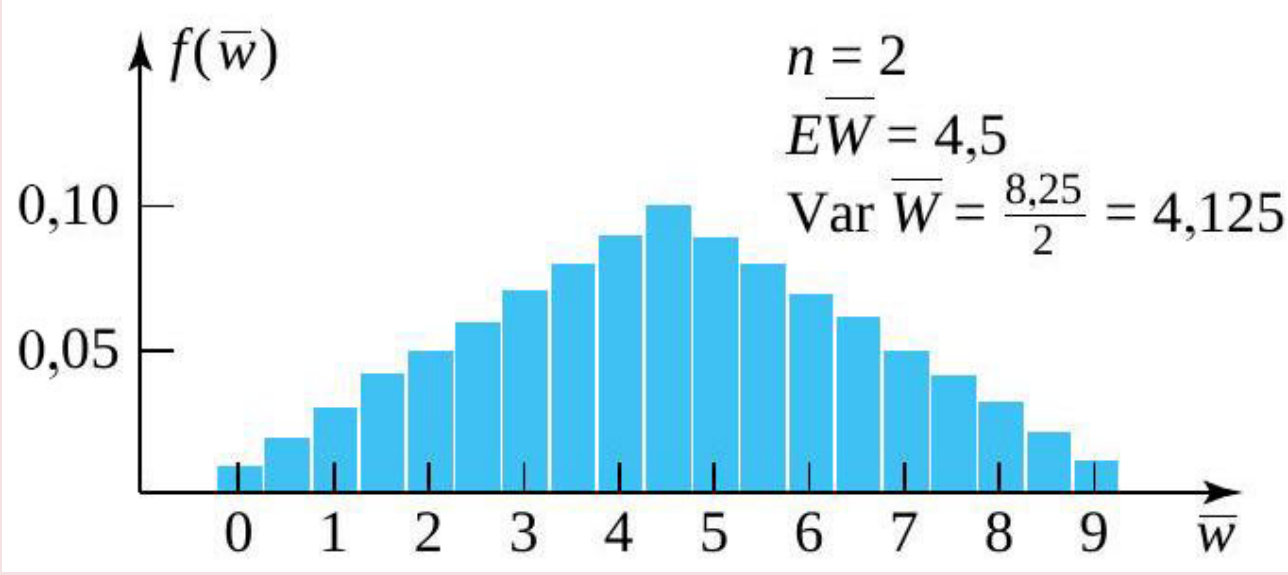


Figure 4.2.4.2 Histogramme de probabilité de $\bar{W}$*, avec* $n = 2$

.

Fonction de probabilité pour $\overline{W}$ lorsque $n = 2$

$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$
0,0	0,01	2,0	0,05	4,0	0,09	6,0	0,07	8,0	0,03
0,5	0,02	2,5	0,06	4,5	0,10	6,5	0,06	8,5	0,02
1,0	0,03	3,0	0,07	5,0	0,09	7,0	0,05	9,0	0,01
1,5	0,04	3,5	0,08	5,5	0,08	7,5	0,04		

Tableau 4.2.4.1

.

En comparant les figures 4.2.4.1 et 4.2.4.2, il est clair que même pour une distribution sous-jacente complètement plate et uniforme de W et une taille d'échantillon de $n = 2$, la distribution de probabilité de $\bar{W}$ commence à prendre une forme de cloche – à tout le moins, plus que la distribution sous-jacente. La raison en est claire. Plus on s'éloigne de la moyenne ou de la valeur centrale de $\bar{W}$, moins il y a de combinaisons de w_1 et w_2 qui peuvent produire une valeur donnée de $\bar{w}$. Par exemple, pour que $\bar{W} = 0$, il faut que $W_1 = 0$ et $W_2 = 0$ – autrement dit, il faut non pas une, mais deux valeurs extrêmes. En revanche, il existe 10 combinaisons différentes de w_1 et w_2 qui produisent $\bar{W} = 4{,}5$.

.

Il est possible d'utiliser le même type de logique que celle qui a conduit au tableau 4.2.4.1 pour produire des distributions de probabilités exactes pour $\bar{W}$ avec de grandes tailles d'échantillons n. Mais ce travail est fastidieux, et pour indiquer plus ou moins comment l'effet central limite prend le dessus au fur et à mesure que n grossit, il suffit d'approximer la distribution de $\bar{W}$ en simulant un échantillon de grande taille. À cette fin, regardons l'histogramme de fréquence (figure 4.2.4.3) de 1 000 ensembles de valeurs pour les variables iid $W_1, W_2, \ldots, W_8$ (avec une distribution marginale qui a été simulée et chaque ensemble pondéré pour produire 1 000 valeurs simulées de $\bar{W}$ avec $n = 8$. Remarquez le caractère en forme de cloche du graphique. (La moyenne simulée de $\bar{W}$ était de $4.508 \approx 4.5 = E\bar{W} = EW$, alors que la variance de $\bar{W}$ était de $1{,}025 \approx 1{,}013 = \text{Var}(\bar{W}) = 8{,}25/8$, en étroite concordance avec les formules.)

.

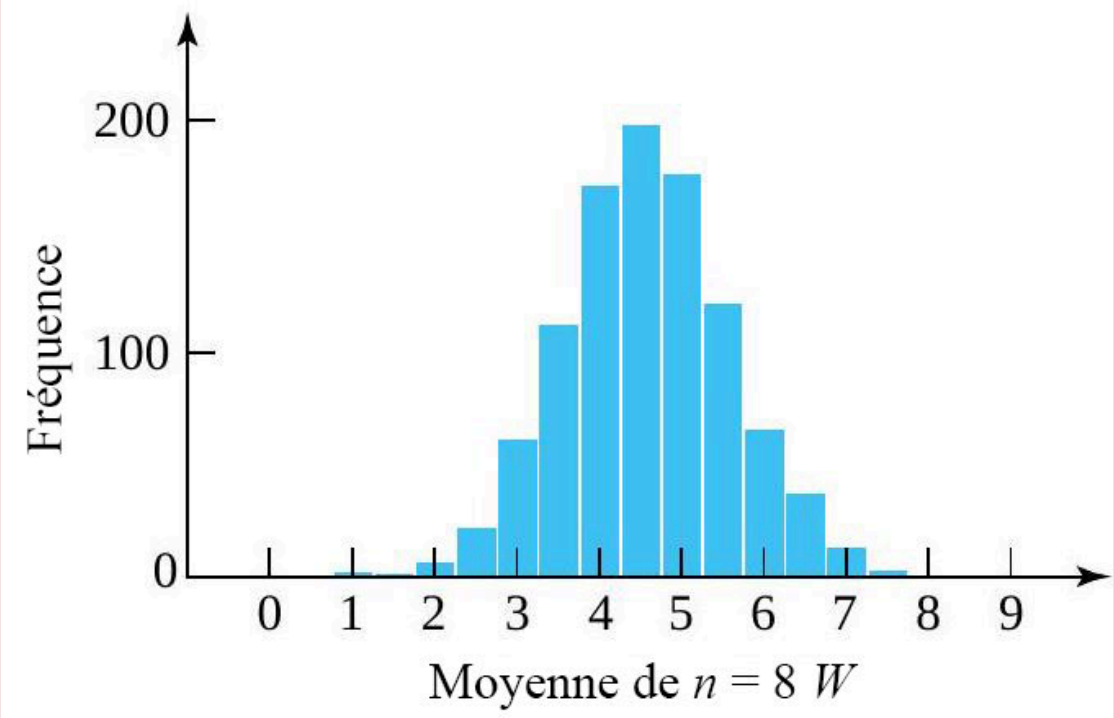


Figure 4.2.4.3 Histogramme des 1 000 valeurs simulées de $\bar{W}$ *avec* $n = 8$.

.

Taille de l'échantillon et effet central limite

Ce qui constitue un « grand échantillon n » dans la proposition 4.2.4.1 n'est pas évident. En réalité, la taille de l'échantillon nécessaire pour que $\bar{X}$ puisse être considérée comme essentiellement normale dépend de la forme de la distribution sous-jacente des observations individuelles. Les distributions sous-jacentes ayant des formes résolument non normales requièrent des valeurs un peu plus élevées de n. Mais dans la plupart des applications d'ingénierie, $n \geq 25$ est généralement suffisant pour que $\bar{X}$ soit essentiellement normale pour la majorité des mécanismes de génération de données. (Les exceptions sont celles qui sont sujettes à la production occasionnelle de valeurs très éloignées de la réalité.) En effet, comme le suggère l'exemple 4.2.4.2, dans de nombreux cas $\bar{X}$ est essentiellement normale pour des tailles d'échantillon très inférieures à 25.

.

L'utilité pratique de la proposition 4.2.4.1 est que, dans de nombreux contextes, il suffit d'une table normale pour évaluer les probabilités des moyennes d'échantillon.

.

Exemple 4.2.4.2 Exigence en matière de délai de vente de timbres.

Supposons qu'il y ait des exigences concernant le délai de vente des timbres, et que nous voulions observer $n = 100$ durées de service excessives pour obtenir :

.

$\bar{S} =$ le temps moyen de l'échantillon (au-dessus du seuil de $7{,}5\text{sec}$) nécessaire pour réaliser les 100 prochaines ventes de timbres.

.

Supposons en outre que nous voulions approximer $P[\bar{S} > 17]$17] » title= »P[\bar{S}>17] » class= »latex mathjax »>.

.

Nous supposerons qu'un modèle iid avec une distribution de probabilité marginale exponentielle $\alpha = 16.5$ est plausible pour les temps de service excessifs individuels S. Ainsi, on obtient

.

$$E(\bar{S}) = \alpha = 16{,}5 \text{ sec}$$

.

et

.

$$\sqrt{\text{Var}(\bar{S})} = \sqrt{\frac{\alpha^2}{100}} = 1{,}65 \text{ sec}$$

.

pour $\bar{S}$, selon nos équations. En outre, en tenant compte du fait que $n = 100$ est grand, la table de probabilité normale peut être utilisée pour calculer les probabilités approximatives de $\bar{S}$. La figure 4.2.4.4 illustre une distribution approximative pour $\bar{S}$ et l'aire correspondant à $P[\bar{S} > 17]$17] » title= »P[\bar{S}>17] » class= »latex mathjax »>.

.

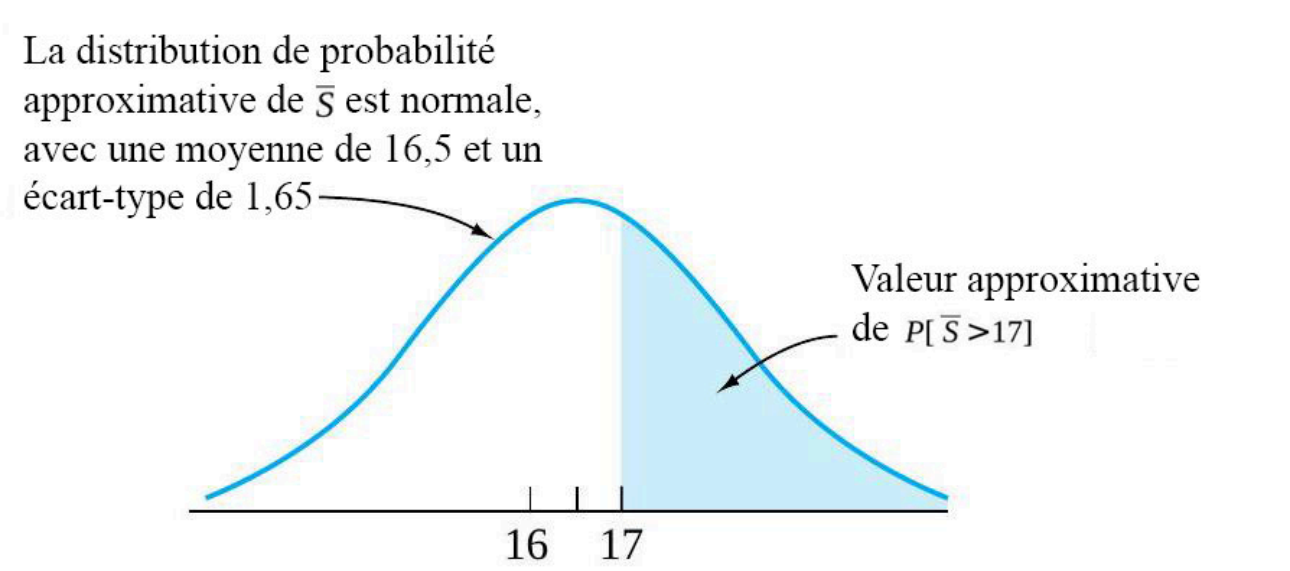


Figure 4.2.4.4 Distribution de probabilité approximative pour $\bar{S}$ *et* $P[\bar{S} > 17]$.

.

Comme toujours, il faut obtenir les cotes z avant de consulter la table normale standard. Dans ce cas, la moyenne et l'écart-type à utiliser sont (respectivement) $16{,}5 \text{ sec}$ et $1{,}65\text{sec}$. Les cotes z valent donc :

.

$$z = \frac{17 - 16.5}{1.65} = .30$$

.

Ainsi :

.

$$P[\bar{S} > 17] \approx P[Z > 0{,}30] = 1 - \Phi(0{,}30) = 0{,}38$$17] \approx P[Z>0,30]=1-\Phi(0,30)=0,38" title= »P[\bar{S}>17] \approx P[Z>0,30]=1-\Phi(0,30)=0,38" class= »latex mathjax »>

.

COTE Z D'UNE MOYENNE D'ÉCHANTILLON

La cote z calculée dans l'exemple est une application de l'équation générale suivante :

.

4.2.4.1 Cote z calculée pour la moyenne d'un échantillon

$$z = \frac{\bar{x} - E\bar{X}}{\sqrt{\operatorname{Var} \bar{X}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

.

Cette application est appropriée lorsqu'on utilise le théorème central limite pour approximer les probabilités de la moyenne d'un échantillon. L'équation 4.2.4.1 est pertinente parce que, comme l'indique la proposition 4.2.4.1, $\bar{X}$ est approximativement normale si n est grand, auquel cas il y a des équations pour obtenir sa moyenne et son écart-type.

5.0.1 Introduction à l'inférence statistique formelle

L'inférence statistique formelle utilise la théorie des probabilités pour quantifier la fiabilité de conclusions basées sur des données. Cette partie présente la logique impliquée dans plusieurs types généraux d'inférence statistique formelle. Les méthodes spécifiques les plus courantes pour les études statistiques à un et deux échantillons sont ensuite examinées.

La partie commence par une introduction à l'estimation des intervalles de confiance, en utilisant le cas important de l'inférence de la moyenne d'un grand échantillon. Le thème des tests d'hypothèse est ensuite abordé, toujours dans le cas de l'inférence de la moyenne d'un grand échantillon. Ces notions générales ayant été expliquées, les sections suivantes traitent de l'intervalle de confiance standard pour un et deux échantillons et des méthodes de test d'hypothèse pour les moyennes, puis les variances, et enfin les proportions. Enfin, les thèmes importants de la tolérance et des intervalles de prédiction sont introduits.

5.0.1 Sources de la partie 5

La partie 5 de cette ressource éducative libre est principalement adaptée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs tirés du chapitre 6, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres.

Le professeur émérite de l'Université de l'Iowa Stephen Vardeman et le professeur émérite de l'Université de Miami J. Marcus Jobe (ISU PhD, 1984) ont placé leur livre *Basic Engineering Data Collection and Analysis*, publié à l'origine par Duxbury/Thompson Learning/Cengage, en téléchargement libre sous licence CC BY-NC-SA internationale 4.0, par l'intermédiaire de l'Iowa State University Digital Press. Ce manuel est disponible à l'adresse

https://www.iastatedigitalpress.com/plugins/books/127/

et est affecté du DOI suivant

https://doi.org/10.31274/isudp.2023.127

Le manuel Basic Engineering Data Collection and Analysis est essentiellement une révision/deuxième édition de l'ouvrage *Statistics for Engineering Problem Solving* de Vardeman, qui a remporté, en 1994, le Meriam/Wiley Distinguished Author Award de l'American Society for Engineering Education.

Le module 5.3 est issu de « Statistics for Research Students », de Erich C. Fein, John Gilmour, Tanya Machin et Liam Hendry – https://usq.pressbooks.pub/statisticsforresearchstudents/

Partie 9 : Statistiques non paramétriques

5.1.1 Intervalles de confiance de la moyenne d'un grand échantillon

INTERVALLES DE CONFIANCE DE LA MOYENNE D'UN GRAND ÉCHANTILLON

De nombreuses applications importantes des statistiques dans le domaine de l'ingénierie s'inscrivent dans le moule standard suivant. Les valeurs des paramètres d'un processus de génération de données sont inconnues, et l'objectif est d'utiliser des données pour

1. trouver un intervalle de valeurs susceptible de contenir un paramètre inconnu (ou une fonction d'un ou plusieurs paramètres) et
2. quantifier la « probabilité » que l'intervalle couvre la valeur correcte.

.

Par exemple, un équipement qui distribue des aliments pour bébés dans des pots peut produire un niveau de remplissage moyen μ inconnu. Il peut être important de déterminer un intervalle basé sur des données susceptible de contenir μ et d'évaluer la fiabilité de l'intervalle. Ou encore, une machine qui usine des filetages sur des boulons en U peut présenter une variation inhérente dans la longueur des filetages, que l'on peut décrire en termes d'écart-type σ. Le but de la collecte de données pourrait alors être de produire un intervalle de valeurs probables pour σ en précisant le degré de confiance de l'intervalle. Ou encore, deux méthodes différentes de fonctionnement d'une machine de granulation peuvent avoir des propensions inconnues différentes à produire des granulés défectueux (disons, p_1 et p_2). Il pourrait falloir utiliser les données pour trouver un intervalle pour $p_1 - p_2$ et fournir le niveau de confiance associé à cet intervalle.

.

DÉFINITION 5.1.1.1 Intervalle de confiance

Un intervalle de confiance pour un paramètre (ou une fonction d'un ou de plusieurs paramètres) est un intervalle numérique basé sur des données et considéré comme susceptible de contenir le paramètre (ou la fonction d'un ou de plusieurs paramètres), avec un certain niveau de confiance (ou de fiabilité) établi selon la théorie des probabilités.

Cette section traite de la façon dont les fondements de la probabilité conduisent à des formules simples pour trouver les intervalles de confiance de la moyenne μ d'un grand échantillon. Le cas inhabituel où l'écart-type σ est connu est traité en premier. Un raisonnement parallèle permet ensuite d'obtenir une formule pour la situation beaucoup plus courante où σ n'est pas connu. La section se termine par des discussions sur trois questions pratiques liées à l'application des intervalles de confiance.

INTERVALLE DE CONFIANCE POUR μ IMPLIQUANT σ (N = GRAND)

L'exemple 4.1.4.3 concernait un processus de remplissage physiquement stable, dont l'écart-type de poids net était de σ = 1,6 g. Étant donné que, pour un grand n, la moyenne d'échantillon des variables aléatoires iid est approximativement normale, cet exemple a montré que pour n = 47 et

$\bar{x}$ = le poids net de remplissage moyen de l'échantillon de 47 pots remplis par le procédé (g)

il y a environ 80 % de chances que $\bar{x}$ soit à 0,3 gramme de μ. Ce fait est illustré à nouveau à la figure 5.1.1.1.

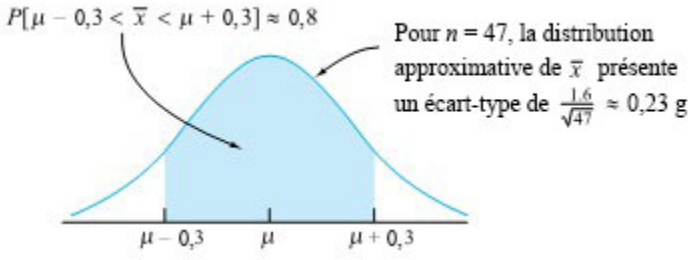


Figure 5.1.1.1 Distribution de probabilité approximative pour \bar{x} avec n = 47

Conventions de notation

À ce stade, nous devons nous consacrer à la notation. Dans la partie 4, les lettres majuscules ont été soigneusement utilisées pour dénoter les variables aléatoires, et les lettres minuscules correspondantes, pour les valeurs possibles ou observées. Mais ici, nous utilisons un symbole minuscule, $\bar{x}$ comme variable aléatoire de la moyenne de l'échantillon. Il s'agit là d'un usage statistique assez courant, qui correspond au type de convention utilisé dans les parties précédentes. Nous allons donc abandonner le respect strict de la convention des majuscules introduite à la partie 4. Les variables aléatoires sont souvent symbolisées par des lettres minuscules, et les mêmes symboles sont utilisés pour les valeurs observées. La convention des majuscules de la partie 4 est particulièrement utile pour apprendre les bases de la probabilité. Mais une fois ces bases maîtrisées, il est courant d'abuser de la notation et de déterminer à partir du contexte s'il s'agit d'une variable aléatoire ou de sa valeur observée.

.

La façon la plus courante d'analyser un graphique comme celui de la figure 5.1.1.1 est de le voir comme la probabilité que

> **5.1.1.1** <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/c637a12264a4b56921ee3431df9c1b86.png » alt= »\mu-0,3<\bar{x}<\mu+0,3" title= »\mu-0,3<\bar{x}

Autrement dit, il s'agit de la probabilité que $\bar{x}$ se situe dans un intervalle centré sur μ et de longueur 2(0,3) = 0,6. Mais on peut aussi voir cela comme la probabilité qu'un intervalle centré sur $\bar{x}$ et de longueur 0,6 comprenne μ. Algébriquement, l'inégalité 5.1.1.1 est équivalente à

> **5.1.1.2** <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/0b4f461aa5a2a1418b4d9a7b53ac7b90.png » alt= »\bar{x}-0,3<\mu<\bar{x}+0,3" title= »\bar{x}-0,3<\mu

Cependant, cette nouvelle inégalité change l'éclairage du problème. Le fait que l'expression 5.1.1.2 ait environ 80 % de chances d'être vraie chaque fois qu'un échantillon de 47 poids de remplissage est prélevé suggère que l'intervalle aléatoire

5.1.1.3 $(\bar{x} - 0{,}3, \bar{x} + 0{,}3)$

peut être utilisé comme intervalle de confiance pour μ, avec une fiabilité ou une confiance associée de 80 %.

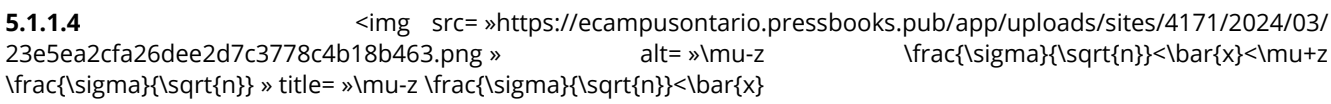
Exemple 5.1.1.1 Intervalle de confidence pour le poids moyen d'un procédé de remplissage

Supposons que pour un échantillon de n = 47 pots, on a $\bar{x}$= 138,2 g. L'expression 5.1.1.3 suggère alors l'intervalle dont les extrémités sont

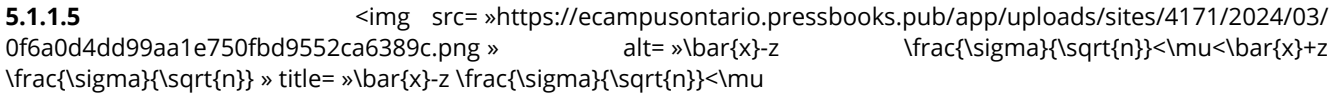
$$138{,}2 \text{ g} \pm 0{,}3 \text{ g}$$

(c'est-à-dire l'intervalle de 137,9 g à 138,5 g) correspond à l'intervalle de confiance de 80 % pour le poids de remplissage moyen du processus.

Ce n'est pas difficile de généraliser la logique qui a conduit à l'expression 5.1.1.3. Chaque fois qu'un modèle iid est approprié pour les éléments d'un grand échantillon, le théorème central limite implique que la moyenne de l'échantillon $\bar{x}$ est approximativement normale avec une moyenne μ et un écart-type de σ/√n. Ainsi, si pour p > 0,5, z est le quantile p de la distribution normale standard, la probabilité que

5.1.1.4 

vaut approximativement 1 − 2(1 − p). Mais l'inégalité 5.1.1.4 peut être réécrite comme suit :

5.1.1.5 

et considérée comme l'éventualité que l'intervalle aléatoire avec les extrémités

EXPRESSION 5.1.1.6 Bornes de l'intervalle de confiance de μ (σ connu, n = grand)

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}}$$

comprend la moyenne μ inconnue. Ainsi, un intervalle avec les extrémités 5.1.1.6 est un intervalle de confiance approximatif pour μ, avec un niveau de confiance 1 − 2(1 − p).

Dans une application, le z de l'équation 5.1.1.6 est choisi de sorte que la probabilité normale réduite entre -z et z corresponde à un niveau de confiance souhaité. La table A1.1 de l'annexe (Table de probabilités de la loi normale centrée réduite) peut être utilisée pour valider les entrées du tableau 5.1.1.1. (Ce tableau donne les valeurs de z à utiliser dans l'expression 5.1.1.6 pour quelques niveaux de confiance courants).

z pour calculer les intervalles bilatéraux à grand n de μ

Confiance souhaitée	z
80%	1,28
90%	1,645
95%	1,96
98%	2,33
99%	2,58

Tableau 5.1.1.1

Exemple 5.1.1.2 Intervalle de confiance pour l'écart moyen par rapport à la valeur nominale dans une opération de meulage

Dib, Smith et Thompson ont étudié un processus de meulage utilisé dans la reconstruction des moteurs automobiles. La variabilité naturelle à court terme associée aux diamètres des manetons des vilebrequins de moteurs meulés par le procédé était de l'ordre de $\sigma = 0{,}7 \times 10^{-4}$ po. Supposons que le processus de meulage des manetons puisse être considéré comme physiquement stable sur des séries de 50 manetons ou moins. Si 32 diamètres de maneton consécutifs présentent un écart moyen par rapport à la valeur nominale de $\bar{x} = -0{,}16 \times 10^{-4}$ po, on peut utiliser l'expression 5.1.1.6 pour établir un intervalle de confiance pour l'écart moyen du processus actuel par rapport à la valeur nominale. Considérons un niveau de confiance de 95 %. En consultant le tableau 5.1.1.1 (ou en réalisant autrement que le quantile $p = 0{,}975$ de la distribution normale standard correspond à 1,96), on utilise $z = 1{,}96$ dans l'équation 5.1.1.6. (On utilise $p = 0{,}975$ puisque $0{,}95 = 1 - 2(1 - 0{,}975)$.) Ainsi, l'intervalle de confiance de 95 % pour l'écart moyen du procédé actuel par rapport au diamètre nominal du maneton a pour bornes

$$-.16 \times 10^{-4} \pm (1.96)\frac{.7 \times 10^{-4}}{\sqrt{32}}$$

soit

$$-0{,}40 \times 10^{-4} \text{ po} \quad \text{et} \quad 0{,}08 \times 10^{-4} \text{ po}$$

Un intervalle comme celui-ci pourrait avoir une importance technique pour déterminer la pertinence de procéder à un ajustement de l'objectif du processus. L'intervalle comprend à la fois des valeurs positives et négatives. Par conséquent, même si $\bar{x} < 0$, les informations dont nous disposons ne sont pas suffisamment précises pour nous permettre de déterminer avec certitude dans quelle direction le processus de meulage doit être ajusté. Ces données, jumelées au fait que les ajustements potentiels de la machine sont probablement plus grossiers que l'erreur d'ajustement la plus probable $\bar{x} = -0{,}16 \times 10^{-4}$ po, suggère fortement de ne pas modifier l'objectif du processus.

INTERVALLE DE CONFIDENCE GÉNÉRALEMENT APPLICABLE POUR μ (N = GRAND)

Bien que l'expression 5.1.1.6 fournisse un intervalle de confiance mathématiquement correct, la présence de σ limite fortement son utilité pratique. Il est inhabituel de devoir estimer une moyenne μ en connaissant déjà σ (qu'on pourrait alors insérer dans une équation). Ces situations se produisent principalement dans des situations de fabrication comme celles des exemples 5.1.1.1 et 5.1.1.2. Une expérience passée considérable peut parfois donner une valeur raisonnable à σ, tandis que les dérives des processus physiques au fil du temps peuvent remettre en question la valeur actuelle de μ.

Heureusement, on peut modifier le raisonnement qui a conduit à l'expression 5.1.1.1 pour obtenir une équation d'intervalle de confiance pour μ qui ne dépend que des caractéristiques de l'échantillon. L'argument conduisant à l'expression 5.1.1.6 repose sur le fait que pour des n grands, $\bar{x}$ est approximativement normale avec une moyenne μ et un écart-type σ/√n. Autrement dit, la variable

5.1.1.7 $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

est approximativement normale réduite. La présence de σ dans l'équation 5.1.1.7 est ce qui conduit à sa présence dans l'expression 5.1.1.6 pour l'intervalle de confiance. Mais une légère généralisation du théorème central limite garantit que si n est grand,

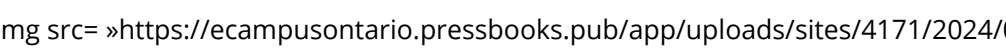
5.1.1.8 $$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

est également approximativement normale réduite. Cette fois-ci, il n'y a pas de σ.

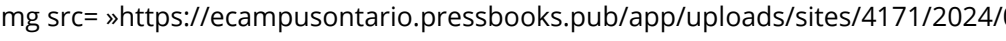
En partant du fait que (lorsqu'un modèle iid pour les observations est approprié et que n est grand) la variable 5.1.1.8 est approximativement normale réduite, le raisonnement est le même que précédemment. Pour un z positif,

$$-z < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < z$$

équivaut à



qui à son tour, équivaut à



Ainsi, l'intervalle dont le centre aléatoire est $\bar{x}$ et la longueur aléatoire est 2zs/√n – autrement dit, avec des bornes aléatoires

EXPRESSION 5.1.1.9 Intervalle de confiance pour μ (n = grand)

$$\bar{x} \pm z\frac{s}{\sqrt{n}}$$

peut être utilisé comme intervalle de confiance approximatif de μ. La valeur de z doit être choisie de sorte que la probabilité normale standard entre -z et z corresponde au niveau de confiance voulu.

Example 5.1.1.3 Couples de rupture et défaillances des disques durs

Dans l'article « The Case of the Derailed Disk Drives » (Mechanical Engineering, 1988), F. Willett évoque une étude réalisée pour isoler la cause de la « défaillance du code de clignotement A » dans un modèle de disque dur Winchester. Les données de cet article ont été reprises à la figure 5.1.1.2, qui présente les couples de rupture (po oz) nécessaires pour desserrer le dispositif d'interruption du lecteur sur l'arbre du moteur pas à pas pour 26 disques durs renvoyés au fabricant pour une défaillance du code de clignotement A. Pour ces données, $\bar{x}$ = 11,5 po oz et s = 5,1 po oz.

Si l'on considère que les disques durs qui ont produit les données de la figure 5.1.1.2 représentent la population des lecteurs sujets à une défaillance du code de clignotement A, il semble raisonnable d'utiliser un modèle iid et l'expression 5.1.1.9 pour estimer le couple de rupture moyen de la population. On choisit d'utiliser un intervalle de confiance de 90 % pour μ; le tableau 5.1.1.1 donne alors z = 1.645.En utilisant l'expression 5.1.1.9, on obtient les bornes

$$11{,}5 \pm 1{,}645\frac{5{,}1}{\sqrt{26}}$$

(c'est-à-dire 9,9 et 13,1 po oz).

L'intervalle montre que le couple de rupture moyen pour les lecteurs présentant une défaillance du code de clignotement A était nettement inférieur à la valeur cible de 33,5 po oz fixée par l'usine, ce qui a été essentiel pour trouver et éliminer un défaut de conception dans les lecteurs.

0	0 2 3
0	7 8 8 9 9
1	0 0 0 1 1 2 2 2 3
1	5 5 6 6 7 7 7 9
2	0
2	

Figure 5.1.1.2 Couples nécessaires pour desserrer 26 indicateurs d'interrupteurs

QUELQUES COMMENTAIRES SUR LES INTERVALLES DE CONFIANCE

Les expressions 5.1.1.6 et 5.1.1.9 ont été utilisées pour établir des déclarations de confiance du type « μ est compris entre a et b ». Mais souvent, une déclaration telle que « μ est au moins égal à c » ou « μ n'est pas supérieur à d » aurait une plus grande valeur pratique. Par exemple, un.e ingénieur.e automobile pourrait déclarer: « L'émission moyenne de monoxyde d'azote pour ce moteur est au maximum de 5 ppm ». Un.e ingénieur.e en génie civil peut aussi vouloir faire une déclaration telle que « la résistance moyenne à la compression des échantillons de ce type de béton est d'au moins 4 188 psi ». En d'autres termes, les problèmes pratiques d'ingénierie se prête parfois mieux à des intervalles de confiance unilatéraux.

Établissement d'intervalles de confiance unilatéraux

L'élaboration de formules pour les intervalles de confiance unilatéraux ne pose pas de réel problème. Si vous disposez d'une formule bilatérale viable, il suffit de faire ce qui suit :

1. Remplacer la limite inférieure par $-\infty$ ou la limite supérieure par $+\infty$.

2. Ajuster le niveau de confiance déclaré à la hausse de manière appropriée (ce qui implique généralement de diviser le « niveau de confiance » par 2).

Cette méthode fonctionne non seulement avec les expressions 5.1.1.6 et 5.1.1.9, mais aussi avec les autres intervalles de confiance bilatéraux présentés dans cette partie.

Exemple 5.1.1.4 (suite)

Pour le couple moyen de rupture des disques durs défectueux, définissions un intervalle de confiance unilatéral à 90 % pour μ de la forme (-∞, #), avec # un nombre approprié. En d'autres termes, cherchons une limite de confiance supérieure (#) de 90 % pour μ.

Pour obtenir un intervalle de confiance unilatéral à 90 %, on part d'un intervalle de confiance bilatéral à 80 % pour μ, et on remplace la limite inférieure par -∞. Ainsi, en utilisant l'expression 5.1.1.9, la limite de confiance supérieure à 90 % pour le couple de rupture moyen est la suivante :

$$\bar{x} + 1{,}28\frac{s}{\sqrt{n}} = 11{,}5 + 1{,}28\frac{5{,}1}{\sqrt{26}} = 12{,}8 \text{ po oz}$$

De manière équivalente, l'intervalle de confiance unilatéral à 90 % pour μ est (-∞, 12,8).

Le chiffre de 12,8 po oz est inférieur à la limite supérieure de 13,1 po oz de l'intervalle bilatéral à 90 % trouvé précédemment (et plus proche de la moyenne de l'échantillon). Dans le cas unilatéral, -∞ est déclaré comme limite inférieure, de sorte qu'il n'y a pas de risque de produire un intervalle contenant uniquement des nombres plus grands que l'inconnue μ. Il est donc possible d'utiliser une limite supérieure plus petite que celle de l'intervalle bilatéral correspondant.

Interprétation des intervalles de confiance

Un deuxième problème lié à l'application des intervalles de confiance est la compréhension correcte de la signification technique du terme confiance. Malheureusement, il est facile de se méprendre à ce sujet. Il est donc important d'exposer avec soin ce que la confiance
signifie et ne signifie pas.

Avant de sélectionner un échantillon et d'utiliser l'expression 5.1.1.6 ou 5.1.1.9, la signification d'un niveau de confiance est évidente. En choisissant un niveau de confiance (bilatéral) de 90 % et donc z = 1,645 pour l'expression 5.1.1.9, avant la sélection de l'échantillon et les calculs, « il y a environ 90 % de chances de trouver un intervalle qui comprend μ ». Sous forme de symboles, cela pourrait être exprimé comme suit :

<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/
b4b1dca0965666b2b65d21cdce2c71b6.png » alt= »P\left[\bar{x}-1,645 \frac{s}{\sqrt{n}}<\mu<\bar{x}+1,645
\frac{s}{\sqrt{n}}\right] \approx 0,90" title= »P\left[\bar{x}-1,645 \frac{s}{\sqrt{n}}<\mu

Mais comment envisager un niveau de confiance après la sélection de l'échantillon? En fait, c'est là une toute autre question. Une fois les chiffres introduits dans l'expression 5.1.1.6 ou 5.1.1.9, les dés sont déjà jetés et l'intervalle numérique est soit correct, soit erroné. La difficulté pratique est que si on ne peut déterminer lequel des deux est le cas, il n'est plus logique d'associer une probabilité à l'exactitude de l'intervalle. Par exemple, cela n'aurait aucun sens d'examiner à nouveau l'intervalle bilatéral obtenu dans l'exemple 5.1.1.3 et d'essayer de dire quelque chose comme « il existe une probabilité de 90 % que μ soit compris entre 9,9 et 13,1 po oz ». μ n'est pas une variable aléatoire; il s'agit d'une quantité fixée (bien qu'inconnue) qui se situe – ou pas – entre 9,9 et 13,1. Il n'y a plus de probabilité dans la situation à discuter.

Que signifie donc que (9,9, 13,1) est un intervalle de confidence à 90 % pour μ? Qu'on le veuille ou non, l'expression « 90 % de confiance » fait davantage référence à la méthode utilisée pour obtenir l'intervalle (9,9, 13,1) qu'à l'intervalle lui-même. Pour déterminer l'intervalle, une méthodologie a été utilisée qui produirait des

intervalles numériques se situant entre μ dans environ 90 % des applications répétées. Mais la validité de cet intervalle en particulier dans cette application est inconnue et n'est pas quantifiable en termes de probabilité. Une personne qui, au cours de sa vie, définit de nombreux intervalles de confiance de 90 % peut s'attendre à avoir un « taux de réussite au cours de sa vie » d'environ 90 %. Mais la validité d'une application donnée n'est généralement pas connue.

Voici un bref énoncé pour résumer ces discussions.

DÉFINITION 5.1.1.2 Interprétation des intervalles de confiance

Dire qu'un intervalle numérique (a, b) est (par exemple) un intervalle de confiance à 90 % pour un paramètre, c'est dire qu'en l'obtenant, on a appliqué des méthodes de collecte de données et de calcul qui produiraient des intervalles comprenant le paramètre dans environ 90 % des applications répétées. Le fait que l'intervalle (a, b) en question comprenne – ou pas – le paramètre est inconnu et ne peut être décrit en termes de probabilité.

On peut penser que la définition 5.1.1.2 donne une signification assez faible au niveau de la fiabilité associée aux intervalles de confiance. Néanmoins, il s'agit de l'interprétation correcte, et c'est tout ce à quoi on peut s'attendre rationnellement. Et bien que l'interprétation correcte puisse sembler peu attrayante au départ, les méthodes d'intervalle de confiance se sont révélées d'une grande utilité pratique.

TAILLE DE L'ÉCHANTILLON POUR ESTIMER μ

Pour conclure cette introduction aux intervalles de confiance, notons que les expressions 5.1.1.6 et 5.1.1.9 peuvent donner des réponses quantitatives rudimentaires à la question « Quelle doit être la taille de n? ». En utilisant l'expression 5.1.1.9, par exemple, si on connaît 1) le niveau de confiance souhaité, 2) l'estimation du pire cas pour l'écart-type de l'échantillon, et 3) la précision d'estimation souhaitée pour μ, il est facile de déterminer la taille de l'échantillon correspondant. Autrement dit, supposons que le niveau de confiance souhaité dicte la valeur de z utilisée dans l'expression 5.1.1.9, que s est la valeur probable du pire cas pour l'écart type de l'échantillon, et qu'on souhaite avoir une ou des limites de confiance de la forme $\bar{x} \pm \Delta$. Soit

$$\Delta = z\frac{S}{\sqrt{n}}$$

On isole n, ce qui donne :

$$n = \left(\frac{zS}{\Delta}\right)^2$$

Exemple 5.1.1.3 suite

Supposons que, dans le cas du problème des disque durs, les ingénieur.e.s veulent approfondir l'analyse des données de la figure 5.1.1.2 en procédant à des essais sur de nouveaux lecteurs après les avoir soumis à des conditions de température accélérée (élevée) dans le but de comprendre le mécanisme à l'origine de la création des faibles couples

de rupture. Supposons en outre que le couple de rupture moyen des lecteurs soumis à des contraintes de température doit être estimé avec un intervalle de confiance bilatéral de 95 % et que la variabilité du couple attendue n'excède pas la valeur s = 5,1 po oz obtenue à partir des lecteurs retournés. On souhaite obtenir une précision d'estimation de ±1 po oz. En utilisant la partie ± de l'expression 5.1.1.9 et en se reportant au tableau 5.1.1.1, l'exigence est la suivante :

$$1 = 1.96\frac{5.1}{\sqrt{n}}$$

En isolant n, on obtient :

$$n = \left(\frac{(1.96)(5.1)}{1}\right)^2 \approx 100$$

Il faudrait donc étudier n = 100 disques durs soumis à des contraintes de température. Si ce nombre n'est pas pratique, les calculs indiquent au moins que le fait de descendre en dessous de cette taille d'échantillon entraînera une réduction de la confiance ou de la précision associée à l'intervalle final (à moins que la variabilité associée aux nouveaux disques soit inférieure à celle des disques retournés).

Il y a deux raisons pour lesquelles les calculs effectués dans l'exemple précédent ne constituent pas une réponse absolue à la question de la taille de l'échantillon. La première, c'est que leur validité dépend de la justesse de s, l'estimation de l'écart-type de l'échantillon. Si s est sous-estimé, n sera trop petit. (De même, si on surestime s par excès de prudence, l'échantillon sera inutilement grand.) La deuxième, c'est que l'expression 5.1.1.9 repose sur l'hypothèse d'un grand échantillon. Si les calculs donnent un n inférieur à environ 25 ou 30, il faudra augmenter la taille pour garantir la validité de l'expression 5.1.1.9.

5.1.2 Tests d'hypothèse pour la moyenne d'un grand échantillon

OBJECTIF DES TESTS D'HYPOTHÈSE

Le chapitre précédent a illustré la façon dont les probabilités peuvent permettre d'estimer un intervalle de confiance. Ce chapitre présente en parallèle les tests d'hypothèse.

Les tests d'hypothèse consistent à utiliser des données pour évaluer quantitativement la plausibilité d'une valeur d'essai d'un paramètre (ou d'une fonction d'un ou plusieurs paramètres). Cette valeur d'essai représente généralement le statu quo ou l'estimation pré-expérimentale. Par exemple, en ingénierie des procédés, on peut utiliser des tests d'hypothèse pour évaluer la plausibilité que le procédé actuel utilisé pour remplir des pots de nourriture pour bébés produise effectivement un remplissage moyen correspondant à la valeur idéale de 138 g. Ou encore, deux méthodes différentes de fonctionnement d'une machine de granulation peuvent avoir des propensions inconnues à produire des pastilles défectueuses (disons, p_1 et p_2), et le test d'hypothèse servir à évaluer la plausibilité de $p_1 - p_2 = 0$ (c'est-à-dire, que les deux méthodes sont aussi efficaces l'une que l'autre).

Cette section explique comment les notions de base des probabilités conduisent à des expressions simples pour les tests d'hypothèse concernant la moyenne μ d'une grand échantillon. Elle présente la terminologie des tests d'hypothèses dans le cas où l'écart-type σ est connu. Ensuite, un format en cinq étapes pour résumer les tests d'hypothèses est présenté. On considère ensuite le cas, plus général, des tests d'hypothèse pour μ lorsque σ n'est pas connu. La section se termine par deux discussions sur des questions pratiques liées à l'application de la logique des tests d'hypothèses.

TESTS D'HYPOTHÈSE POUR μ IMPLIQUANT σ (N = GRAND)

Rappelez-vous l'exemple 4.1.4.3, qui concernait un processus de remplissage physiquement stable, dont l'écart-type de poids net était de σ = 1,6 g. Supposons en outre qu'avec un poids déclaré (sur l'étiquette) de 135 g, les ingénieurs de processus ont fixé un poids net moyen cible de 135 + 3σ = 139,8 g. Enfin, supposons que lors d'un contrôle de routine du processus de remplissage destiné à détecter tout changement de la moyenne du processus par rapport à sa valeur cible, un échantillon de n = 25 pots a produit $\bar{x}$ = 139.0 g. Qu'est-ce que cette valeur nous apprend sur la plausibilité que la moyenne actuelle du processus respecte bien l'objectif de 139,8 g?

Le théorème centra limite peut être invoqué ici. Si effectivement la moyenne actuelle du processus est de 139,8 g, $\bar{x}$ a une distribution approximativement normale avec une moyenne de 139,8 g et un écart-type σ/√n = 1,6/√25 = 0,32 g, comme le montre la figure 5.1.2.1, qui montre aussi la valeur observée de $\bar{x}$ = 139,0 g.

La figure 5.1.2.2 illustre l'image normale standard qui correspond à la figure 5.1.2.1. Elle repose sur le fait que si la moyenne actuelle du processus est conforme à l'objectif de 139,8 g, alors le fait que $\bar{x}$ soit approximativement normal avec une moyenne μ et un écart-type σ/√n = 0,32 g implique que

5.1.2.1

$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 139.8}{.32}$$

est approximativement normale réduite. La valeur $\bar{x}$ = 139,0 g observée à la figure 5.1.2.1 correspond à la valeur de z = −2,5 à la figure 5.1.2.2.

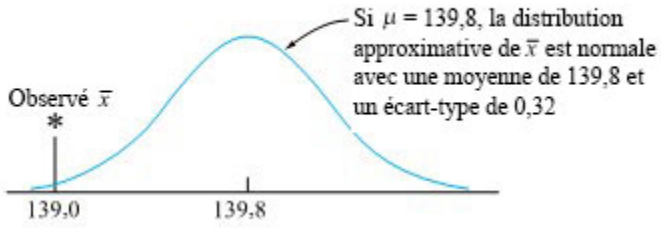


Figure 5.1.2.1 Distribution de probabilité approximative pour \bar{x} avec μ = 139,8 g. et la valeur observée de \bar{x} = 139,0 g.

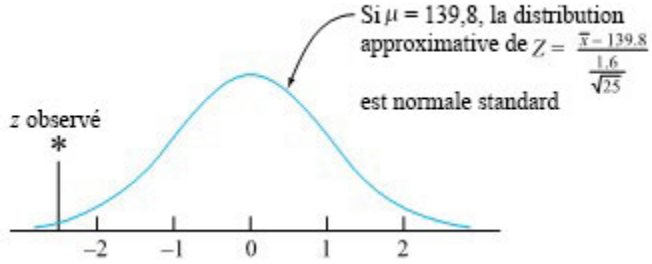


Figure 5.1.2.2 Distribution normale réduite de la figure 5.1.2.1

La figure 5.1.2.1 et la figure 5.1.2.2 montrent clairement que si la moyenne du processus est conforme à l'objectif de 139,8 g (et que les chiffres sont donc corrects), on a observé une valeur de $\bar{x}$ (ou une valeur de z associée) assez extrême ou rare. Bien sûr, il arrive que des phénomènes extrêmes ou rares se produisent, mais la valeur de$\bar{x}$ (et de z) observée semble plutôt infirmer l'hypothèse que le processus se déroule comme prévu.

Les chiffres suggèrent même un moyen de quantifier leur propre invraisemblance, soit de calculer la probabilité associée aux valeurs de $\bar{x}$ (ou Z) au moins aussi extrême que celle observée. L'expression « au moins aussi extrême » doit maintenant être définie par rapport à l'objectif initial de la collecte de données, à savoir détecter une valeur de μ inférieure ou supérieure à la valeur cible. Les valeurs de $\bar{x}$ ≤ 139,0 g (z ≤ −2,5) sont aussi extrêmes que celle observée, et ce serait également le cas des valeurs de $\bar{x}$ ≥ 140,6 g (z ≥ 2,5). (Dans le premier cas, $\bar{x}$ suggérerait une valeur de μ inférieure à la cible; dans le second, une valeur supérieure.) En d'autres termes, l'invraisemblance d'être sur la cible peut être quantifiée en notant que si c'était le cas, seule une fraction de

$$\Phi(-2.5) + (1 - \Phi(2.5)) = .01$$

de tous les échantillons produirait la valeur $\bar{x}$ (ou z) aussi extrême que celle
observée. Cela revient à dire que les données infirment assez fortement l'hypothèse selon laquelle le processus respecte sa cible.

L'argument qui vient d'être présenté est une application de la logique typique du test d'hypothèse. Afin de rendre le fil de pensée évident, il est utile d'en isoler certains éléments sous forme de définition. La définition 5.1.2.1 aborde cette tâche en reformulant l'objectif général des tests d'hypothèse.

DÉFINITION 5.1.2.1 Test d'hypothèse statistique

Un test d'hypothèse statistique consiste à utiliser des données pour évaluer quantitativement la plausibilité d'une valeur d'essai d'un paramètre (ou d'une fonction d'un ou plusieurs paramètres).

Logiquement, les tests d'hypothèses commencent par la spécification de l'essai ou de la valeur hypothétique. Il existe une terminologie et une notation pour énoncer cette valeur.

DÉFINITION 5.1.2.2 Hypothèse nulle

Une hypothèse nulle est un énoncé de la forme

Paramètre = #

ou

Fonction de paramètres = #

(pour un nombre #) qui constitue la base d'investigation dans un test d'hypothèse. Une hypothèse nulle est généralement formulée pour représenter le statu quo ou l'estimation pré-expérimentale du paramètre (ou de la fonction du ou des paramètres). On la note généralement H_0.

La notion d'hypothèse nulle est vraiment centrale dans les tests d'hypothèse. La partie « nulle » de l'expression « hypothèse nulle » fait référence au fait que les hypothèses nulles sont des déclarations d'absence de différence, ou autrement dit, d'égalité. Par exemple, dans le contexte du remplissage des pots, l'usage standard serait d'écrire

5.1.2.2 $$\mathrm{H}_0 : \mu = 139.8$$

ce qui signifie qu'il n'y a pas de différence entre μ et la valeur cible de 139,8 g.

Après avoir formulé une hypothèse nulle, il faut préciser quels types d'écarts par rapport à cette hypothèse sont intéressants.

DÉFINITION 5.1.2.3 Hypothèse alternative

Une hypothèse alternative est un énoncé qui s'oppose à l'hypothèse nulle. Il spécifie les formes d'écart par rapport à l'hypothèse nulle qui doivent être prises en compte. L'hypothèse alternative se note généralement H_a. Elle est de la même forme que l'hypothèse nulle correspondante, à l'exception du signe d'égalité qui est remplacé par ≠, > ou <.

Souvent, l'hypothèse alternative repose sur les soupçons et ou les espoirs de l'ingénieur.e quant à la situation réelle, ce qui équivaut à une sorte d'hypothèse de recherche qu'on espère démontrer. Par exemple, si on teste ce qui est censé être un dispositif permettant d'améliorer la consommation d'essence d'une automobile, on pourrait émettre l'hypothèse nulle « pas de changement à la consommation d'essence » et l'hypothèse alternative « réduction de la consommation ».

Les définitions 5.1.2.2 et 5.1.2.3 impliquent toutes deux que, dans le cas d'un test portant sur une moyenne unique, les trois paires possibles d'hypothèses nulle et alternative sont les suivantes :

$$\mathrm{H}_0 : \mu = \# \quad \mathrm{H}_0 : \mu = \# \quad \mathrm{H}_0 : \mu = \#$$
$$\mathrm{H}_\mathrm{a} : \mu > \# \quad \mathrm{H}_\mathrm{a} : \mu < \# \quad \mathrm{H}_\mathrm{a} : \mu \neq \#$$

\# & \mathrm{H}_{\mathrm{a}}: \mu\# & \mathrm{H}_{\mathrm{a}}: \mu

Dans l'exemple du remplissage des pots, il faut détecter à la fois la possibilité de sous-remplissage (μ < 139,8 g) et la possibilité de surremplissage (μ > 139,8 g). Par conséquent, l'hypothèse alternative suivante est appropriée :

5.1.2.3 $\mathrm{H}_\mathrm{a} : \mu \neq 139.8$

Une fois que les hypothèses nulle et alternative ont été établies, il faut définir soigneusement la manière dont les données seront utilisées pour évaluer la plausibilité de l'hypothèse nulle. Cela implique de spécifier une statistique à calculer, une distribution de probabilité appropriée si l'hypothèse nulle est vraie, et les types de valeurs observées qui infirmeront l'hypothèse nulle.

DÉFINITION 5.1.2.4 Statistique de test

Une statistique de test est la forme particulière de synthèse des données numériques utilisée dans un test d'hypothèse. La formule de la statistique de test implique généralement le nombre apparaissant dans l'hypothèse nulle.

DÉFINITION 5.1.2.5 Distribution nulle

La distribution nulle (ou distribution de référence) d'une statistique de test est la distribution de probabilité décrivant la statistique de test, à condition que l'hypothèse nulle soit effectivement vraie.

Les valeurs de la statistique de test considérées comme mettant en doute la validité de l'hypothèse nulle sont spécifiées après avoir examiné la forme de l'hypothèse alternative. En gros, on identifie les valeurs qui ont plus de chances de se produire si la réalité correspond à l'hypothèse alternative plutôt qu'à l'hypothèse nulle.

La discussion sur le processus de remplissage a oscillé entre l'utilisation de $\bar{x}$ et sa version normalisée z donnée par l'équation 5.1.2.1 pour les statistiques de test. L'équation 5.1.2.1 est une forme spécialisée de la statistique de test générale pour μ (avec n grand et σ connu) :

5.1.2.4 $$Z = \frac{\bar{x} - \#}{\frac{\sigma}{\sqrt{n}}}$$

pour le cas actuel, où, selon l'hypothèse, μ = 139,8 g, avec n = 25 et σ = 1,6. Il est plus pratique d'envisager la statistique de test pour ce type de problème sous la forme normalisée présentée dans l'équation 5.1.2.4 plutôt que sous la forme $barx$. En utilisant la forme 5.1.2.4, la distribution de référence sera toujours la même, à savoir la distribution normale réduite.

En poursuivant avec l'exemple du remplissage, notons que si au lieu de l'hypothèse nulle 5.1.2.2, c'est l'hypothèse alternative 5.1.2.3 qui prévaut, on aura tendance à observer des valeurs de $\bar{x}$ soit beaucoup plus grandes, soit beaucoup plus petites que 139,8 g. L'équation 5.1.2.4 permettra alors de transformer les valeurs de $\bar{x}$ en des valeurs observées de Z grandes ou petites (c'est-à-dire de grands nombres négatifs dans ce cas) – c'est-à-dire de grandes valeurs de |z|. De telles valeurs observées rendent l'hypothèse nulle peu plausible.

Après avoir indiqué comment les données seront utilisées pour juger de la plausibilité de l'hypothèse nulle, il reste à les collecter, à les introduire dans l'expression de la statistique de test et, en utilisant la valeur calculée et la distribution de référence, à parvenir à une évaluation quantitative de la plausibilité de H_0. Il existe un terme pour cela.

DÉFINITION 5.1.2.6 Valeur p

Le seuil de signification observé, ou valeur p, d'un test d'hypothèse est la probabilité attribuée, par la distribution de référence, à l'ensemble des valeurs possibles de la statistique du test qui sont au moins aussi extrêmes que celle observée (en termes de mise en doute de l'hypothèse nulle).

Les petites valeurs p infirment H_0

.

Plus le seuil de signification observé est faible, plus la preuve de la validité de l'hypothèse nulle est forte. Dans le cadre de l'opération de remplissage, la valeur observée de la statistique de test étant z = −2,5,

le seuil de signification observé associé vaut

$$\Phi(-2.5) + (1 - \Phi(2.5)) = .01$$

ce qui constitue une preuve assez forte contre la possibilité que la moyenne du processus soit conforme à l'objectif.

5.1.3 Modèle de synthèse de tests d'hypothèse en cinq étapes

MODÈLE DE SYNTHÈSE DE TESTS D'HYPOTHÈSE EN CINQ ÉTAPES

Il est utile de définir un modèle étape par étape pour organiser les comptes rendus des tests d'hypothèse. Celui qui sera utilisé dans ce manuel comprend les cinq étapes suivantes :

Étape 1 Énoncer l'hypothèse nulle.

Étape 2 Énoncer l'hypothèse alternative.

Étape 3 Énoncer les critères du test, c'est-à-dire, donner la formule de la statistique de test (en introduisant uniquement une valeur supposée de l'hypothèse nulle, mais aucune information sur l'échantillon) et la distribution de référence. Ensuite, indiquer, en termes généraux, les valeurs observées de la statistique de test
qui constitueront la preuve contre l'hypothèse nulle.

Étape 4 Montrer les calculs basés sur l'échantillon.

Étape 5 Signaler le seuil de signification observé et, dans la mesure du possible, expliquer ce qu'il signifie pour le problème d'ingénierie.

Exemple 5.1.3.1 Test d'hypothèse concernant le niveau moyen de remplissage

Le modèle de test d'hypothèse en cinq étapes peut être utilisé pour résumer la discussion précédente sur le processus de remplissage.

1. H_0 : μ = 139,8 g.
2. H_a : μ ≠ 139,8 g.
3. La statistique de test est la suivante :

$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}}$$

La distribution de référence est la distribution normale réduite, et de grandes valeurs observées |z| constitueront une preuve contre H_0.

4. L'échantillon donne

$$z = \frac{139.0 - 139.8}{\frac{1.6}{\sqrt{100}}} = -2.5$$

5. Le seuil de signification observé est :

$$\begin{aligned} &P[\text{une variable normale réduite} \leq -2{,}5] \\ &\quad + P[\text{une variable normale réduite} \geq 2{,}5] \\ &\quad = P[|\text{une variable normale réduite}| \geq 2{,}5] \\ &\quad = 0{,}01 \end{aligned}$$

Il s'agit d'un appui raisonnablement faible en faveur de l'hypothèse nulle. Par conséquent, il s'agit d'un appui raisonnablement solide que le niveau de remplissage moyen n'est pas conforme à la cible.

5.1.4 Tests d'hypothèse pour moyennes généralement applicables (n = grand)

La méthode de test d'hypothèse utilisée pour mener la discussion jusqu'ici est facile à expliquer et à comprendre, mais son utilisation pratique est limitée en raison de la présence du paramètre σ dans la statistique 5.1.2.4. Comme indiqué au chapitre 5.1.1, il existe peu de contextes d'ingénierie dans lesquels il est nécessaire de faire des déductions concernant μ tout en connaissant la valeur de σ correspondante. Heureusement, en raison du même fait de la théorie des probabilités qui a permis de produire une formule d'intervalle de confiance pour μ sans σ dans le cas d'un grand échantillon, il est également possible de réaliser des tests d'hypothèse pour μ sans avoir à fournir σ dans le cas d'un grand échantillon.

Pour les observations qui peuvent être décrites comme essentiellement équivalentes à des sélections aléatoires avec remplacement dans une population unique avec une moyenne μ et une variance σ^2, si n est grand,

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

est approximativement normale réduite. Cela signifie que pour un grand n, pour tester

$$\mathrm{H}_0 : \mu = \#$$

dans la plupart des cas, il suffira d'appliquer la logique déjà présentée mais avec la statistique

EXPRESSION 5.1.4.1 Statistique de test pour μ (n = grand)

$$Z = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

à la place de la statistique 5.1.2.4.

Exemple 5.1.4.1. Test d'hypothèse et défaillance de disque dur (suite)

Prenons le cas d'une défaillance du code de clignotement A d'un disque dur. Les couples de rupture ajustés en usine sur la connexion de l'indicateur d'interrupteur à l'arbre du moteur pas à pas étaient en moyenne de 33,5 po oz, et on soupçonnait que la défaillance du code de clignotement A était associée à un couple de rupture réduit. Rappelons qu'un échantillon de n = 26 disques durs défectueux présentait des couples de rupture (indiqués à la figure 5.1.2.2) de $\bar{x}$ = 11,5 po oz et s = 5,1 po oz.

Si on souhaite déterminer dans quelle mesure les données infirment la possibilité que les disques présentant une défaillance du code de clignotement A ont un couple de rupture moyen égal à la valeur moyenne définie en usine de 33,5 po oz. Le modèle de test d'hypothèse en cinq étapes peut être utilisé.

1. H_0: μ = 33,5.
2. H_a: μ < 33,5.

(Ici, l'hypothèse alternative est directionnelle et se résume à une hypothèse de recherche basée sur les soupçons de l'ingénieur.e concernant la relation entre la défaillance du disque et le couple de rupture.)

3. La statistique de test est la suivante :

$$Z = \frac{\bar{x} - 33.5}{\frac{s}{\sqrt{n}}}$$

La distribution de référence est normale réduite, et de petites valeurs de z observées constitueront une preuve contre la validité de H_0. (Les moyennes inférieures à 33,5 tendent à produire des valeurs $\bar{x}$ petites – c'est-à-dire, des grandes valeurs de z négatives.)

4. L'échantillon donne

$$z = \frac{11{,}5 - 33{,}5}{\frac{5{,}1}{\sqrt{26}}} = -22{,}0$$

5. Le seuil de signification observé est :

P [une variable normale standard < −22,0] ≈ 0

L'échantillon fournit des preuves irréfutables que les disques défectueux ont un couple de rupture moyen inférieur au niveau défini en usine.

Il est important de ne pas faire un saut logique trop important et de ne pas conclure à tort que c'est là la solution complète au véritable problème d'ingénierie. Les disques durs renvoyés pour une défaillance du code de clignotement A ont un couple de rupture inférieur aux normes, mais en l'absence de preuve du contraire, il est possible qu'ils ne soient pas différents, à cet égard, des disques durs non défaillants actuellement utilisés. Et même si la réduction du couple de rupture est en cause, une solution réelle nécessite l'identification et la prévention du mécanisme physique qui en est à l'origine. Nous ne disons pas que le test d'hypothèse manque d'importance; nous vous rappelons seulement qu'il ne s'agit que de l'un des nombreux outils utilisés pour effectuer le travail d'ingénieur.e.

5.1.5 Test d'hypothèse et décision statistique

La logique à laquelle fait référence ce chapitre s'applique parfois à un contexte de prise de décision, où l'on fait appel à des données pour orienter le choix entre deux options concurrentes. Dans de tels cas, un cadre décisionnel explicite accompagne généralement l'analyse statistique formelle, ce qui entraîne le recours à une terminologie et à des schémas de pensée supplémentaires.

Dans certains contextes décisionnels, on peut envisager que les deux avenues sont liées à l'hypothèse nulle et à l'hypothèse alternative. Par exemple, dans le scénario du remplissage de pots, H_0 : μ = 139,8 g pourrait correspondre à la décision « ne pas toucher à la procédure », et H_a : μ ≠ 139,8 pourrait correspondre à la décision « ajuster la procédure ». Dans ce genre de contexte, il y a deux types d'erreurs distincts qui peuvent se produire.

DÉFINITION 5.1.5.1 Erreur de type 1

Lorsqu'on recoure à un test d'hypothèse dans le cadre d'une prise de décision, le fait de se prononcer en faveur de H_a alors que la réalité correspond plutôt à H_0 constitue une erreur de type 1.

DEFINITION 5.1.5.2 Erreur de type 2

Lorsqu'on utilise un test d'hypothèse dans le cadre d'une prise de décision, le fait de se prononcer en faveur de H_0 alors que la réalité correspond plutôt à H_a constitue une erreur de type 2.

Le contenu de ces deux définitions est représenté dans le tableau 2 × 2 illustré à la figure 5.1.5.1. Dans l'exemple du remplissage de pots, l'ajustement d'une procédure correcte constituerait une erreur de type 1. En revanche, une erreur de type 2 consisterait à ne pas ajuster une procédure hors cible.

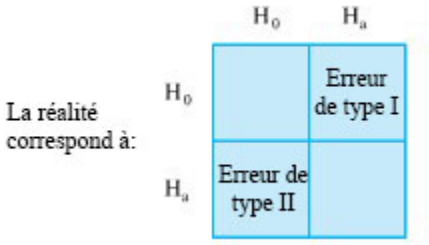


Figure 5.1.5.1. Les quatre possibilités dans un problème de décision

On utilise les tests d'hypothèse pour parvenir à une décision en choisissant une valeur critique, et si le seuil

de signification observé se révèle inférieur à la valeur critique (rendant ainsi l'hypothèse nulle peu plausible), on tranche en faveur de H_a. Autrement, on applique l'option correspondant à H_0. La valeur critique relative au seuil de signification observé représente finalement la probabilité initiale qu'on se prononce en faveur de H_a, probabilité calculée sous réserve de la véracité de H_0. Ce concept fait l'objet d'une terminologie spécifique.

DÉFINITION 5.1.5.3 Seuil de signification

Lorsqu'un test d'hypothèse est utilisé dans un contexte décisionnel, la valeur critique qui sépare les seuils de signification élevés pour lesquels H_0 sera validée des seuils de signification inférieurs pour lesquels H_0 se verra rejetée en faveur de H_a, se nomme probabilité d'erreur de type 1, ou seuil de signification. On utilise généralement le symbole α pour désigner la probabilité d'erreur de type 1.

En pratique, α est généralement un petit nombre (comme 0,1, 0,05 ou même 0,01), ce qui permet d'introduire une certaine inertie en faveur de H0 dans le processus de prise de décision. (Une telle pratique garantit la faible fréquence des erreurs de type 1. Mais, parallèlement, il en résulte une asymétrie dans le traitement de H_0 et de H_a qui s'avère parfois injustifiée.)

La définition 5.1.5.2 et la figure 5.1.5.1 établissent clairement que les erreurs de type 1 ne constituent pas la seule possibilité peu souhaitable. Il faut également prendre en considération la possibilité d'erreurs de type 2.

DÉFINITION 5.1.5.4 Erreur de type 2

Lorsque des tests d'hypothèse sont utilisés dans un contexte décisionnel, la probabilité – calculée en supposant qu'une valeur particulière du paramètre décrite par H_a est valide – que le seuil de signification observé se révèle supérieur à α (c'est-à-dire que H_0 ne soit pas rejetée) se nomme probabilité d'erreur de type 2. On utilise généralement le symbole β
pour représenter la probabilité d'erreur de type 2. Il existe un concept appelé « puissance statistique », qui correspond à β.

Dans la plupart des méthodes d'essai étudiées dans cet ouvrage, le calcul de β excède la portée de l'introduction limitée aux probabilités fournie à la partie 4. Mais la tâche peut être effectuée pour la situation simple où l'on connaît σ évoquée lors de l'introduction du thème des tests d'hypothèse. De plus, quelques calculs de ce type vous procureront une certaine intuition représentative de ce qui prévaut généralement, du moins sur le plan qualitatif.

Example 5.1.5.1 (suite)

Considérons à nouveau la procédure de remplissage et le test H_0 et $H_a \neq$. Cette fois, imaginons que, demain, on recourra à un test d'hypothèse avec n = 25

afin de déterminer s'il faut modifier ou non la procédure. Les probabilités d'erreur de type 2, calculées en présumant que μ = 139,5 et μ = 139,2 pour les tests utilisant α = 0,05 et α = 0,2, seront comparées.

On se penche d'abord sur α = 0,05. La décision sera prise en faveur de H_0 si la valeur p est supérieure à 0,05. En d'autres termes, la décision est favorable à l'hypothèse nulle si la valeur observée de Z donnée dans l'équation correspond à

$$|z| < 1,96$$

ce qui revient à dire

$$139,8-1,96(0,32)<\bar{x}<139,8+1,96(0,32)$$

Autrement dit, on veut déterminer si :

> **5.1.5.1**<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/fe8f117cedd54f9dd050c132cca373d1.png » alt= »139,2<\bar{x}<140,4" title= »139,2<\bar{x}

Si μ décrite parH_a (avec $\mathrm{H_a}: \mu \neq 139.8$) représente la véritable moyenne de la procédure,$\bar{x}$ n'est pas approximativement normale avec une moyenne de 139,8 et un écart-type de 0,32, mais plutôt approximativement normale avec une moyenne de μ et un écart-type de 0,32. Ainsi, pour une telle valeur de μ, l'inégalité 5.1.5.1 et la définition 5.1.5.4 démontrent que le β correspondant représente la probabilité que la distribution normale correspondante attribue à la possibilité que 139,2 < $\bar{x}$ < 140,4. Ceci est illustré dans la figure 5.1.5.2 pour les deux moyennes μ = 139,5 et μ = 139,2.

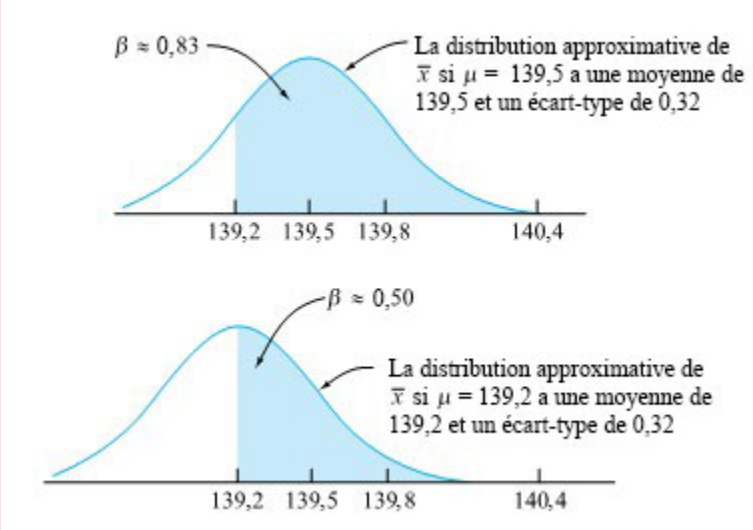


Figure 5.1.5.2 Distributions de probabilités approximatives de \bar{x} pour deux valeurs différentes de μ décrites par H_a et les β correspondants, lorsque α = 0,05

Pour calculer les cotes z correspondant à $\bar{x}$ = 139,2 et $\bar{x}$ = 140,4 avec des moyennes de 139,5 et de 139,2 et un écart-type de 0,32, il suffit de consulter une table de distribution normale standard, ce qui a été fait pour fournir les deux β de la figure 5.1.5.2.

Le raisonnement parallèle pour la situation avec α = 0,2 se fait comme suit : la décision sera prise en faveur de H_0 si la valeur p est supérieure à 0,2. En d'autres termes, la décision sera favorable à H_0 si |z| < 1,28, soit si :

$$139,4 < \bar{x} < 140,2$$

Si μ décrite par H_a représente la véritable moyenne de la procédure, $\bar{x}$ est approximativement normale avec une moyenne μ et un écart type de 0,32. Ainsi, le β correspondant représente la probabilité que cette distribution normale attribue à la possibilité que 139,4 < $\bar{x}$ < 140,2. La figure 5.1.5.3 l'illustre pour les deux moyennes μ = 139,5 et μ = 139,2, avec les probabilités d'erreur de type 2 correspondantes β = 0,61 et β = 0,27.

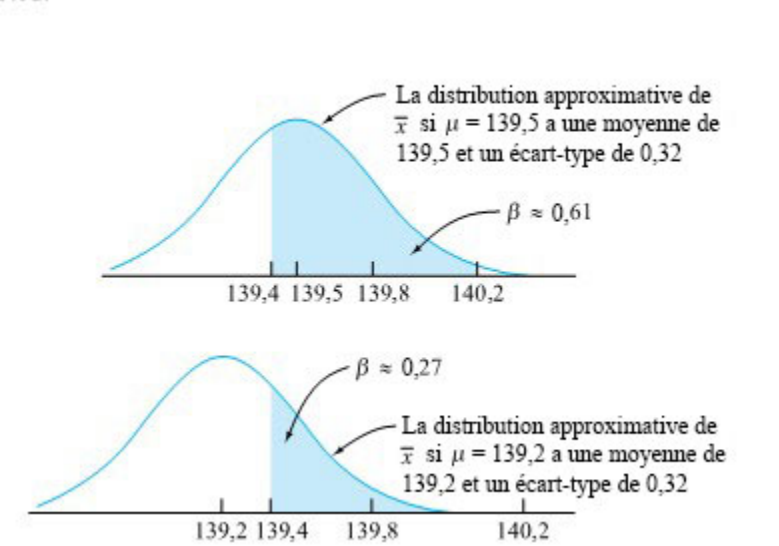


Figure 5.1.5.3. Figure 2 Distributions de probabilités approximatives de \bar{x} pour deux valeurs différentes de μ décrites par H_a et les β correspondants, lorsque α = 0,05

Les calculs représentés par les deux figures sont résumés dans le tableau 5.1.5.1. On peut remarquer deux caractéristiques dans ce tableau. Premièrement, les valeurs de β pour α = 0,05 sont plus élevées que celles pour α = 0,2. Si l'on ne veut courir qu'un risque de 5 % de décider (à tort) d'ajuster une procédure qui atteint sa cible, le prix à payer est d'avoir une plus grande probabilité de ne pas reconnaître que la procédure est hors cible. Deuxièmement, les valeurs de β pour μ = 139,2 sont plus petites que les valeurs β pour μ = 139,5. Plus la procédure de remplissage s'éloigne de la cible, plus il est probable qu'on détecte cet écart par rapport à la cible.

Probabilités d'erreur de type II (β), $n = 25$

		μ	
		139,2	139,5
α	0,05	0,50	0,83
	0,2	0,27	0,61

Tableau 5.1.5.1 Valeurs de β

Le récit présenté dans le tableau 5.1.5.1 s'applique de manière qualitative à toutes les utilisations de tests d'hypothèse dans des contextes de prise de décision. Plus H_0 s'éloigne de la vérité, plus le β correspondant diminue. En outre, de petits α entraînent de grands β, et vice-versa.

Effet de la taille de l'échantillon sur βs

Il y a un autre élément de ce panorama qui joue un rôle important dans la détermination des probabilités d'erreur : la taille de l'échantillon. Pour un α donné, si la taille d'un échantillon peut être augmentée, les β correspondants peuvent être réduits. Reprenons les calculs de l'exemple précédent, en supposant cette fois que n = 100 au lieu de 25. Le tableau 5.1.5.2 illustre les probabilités d'erreur de type 2 censées en résulter; en

comparant avec le tableau 5.1.5.1, on voit l'effet de la taille de l'échantillon dans l'exemple de la procédure de remplissage.

Probabilités d'erreur de type II (β), $n = 100$			
		μ	
		139,2	139,5
α	0,05	0,04	0,53
	0,2	0,01	0,28

Tableau 5.1.5.2 Valeurs de β

Analogie entre un test d'hypothèse et un procès criminel

Pour mieux comprendre la logique communément appliquée aux tests d'hypothèse dans le cadre d'un processus de prise de décision, on peut faire une analogie avec un procès criminel. Dans un procès criminel, deux hypothèses s'opposent :

H_0 : la personne accusée est innocente

H_a : la personne accusée est coupable

Des preuves, dont le rôle se rapproche de celui des données utilisées dans les tests, sont recueillies et utilisées pour trancher entre les deux hypothèses. Deux types d'erreurs éventuelles surviennent lors d'un procès criminel : la possibilité de condamner une personne innocente (erreur de type 1) et la possibilité d'acquitter une personne coupable (erreur de type 2). Un procès criminel constitue une situation où les deux types d'erreur ont des conséquences résolument différentes et où les deux hypothèses subissent un traitement asymétrique. La présomption *a priori* est favorable à H_0, l'innocence de la personne accusée. Afin de maintenir un faible risque de fausse condamnation (c'est-à-dire un α faible), des preuves accablantes s'avèrent nécessaires pour obtenir une condamnation, à l'instar de l'utilisation d'un α faible dans les tests, qui nécessite des valeurs extrêmes de la statistique de test pour infirmer H_0. L'une des conséquences de cette procédure dans les procès criminels, c'est qu'il existe une probabilité substantielle qu'un individu coupable soit acquitté, au même titre que les faibles α produisent des β élevés dans les contextes de test d'hypothèse.

Ce parallèle entre les tests d'hypothèse et les procès criminels peut se révéler utile, à condition de ne pas en abuser. Il ne s'applique pas à toutes les utilisations des tests d'hypothèse, et rares sont les scénarios d'ingénierie suffisamment simples pour se réduire à un simple choix entre H_0 et H_a. Les applications judicieuses des tests d'hypothèse ne constituent souvent que des étapes de « l'évaluation de la preuve » dans le cadre d'une tâche aux nombreuses facettes, fondée sur des données, nécessaire à la résolution d'un problème d'ingénierie. De plus, même lorsqu'un problème réel peut se réduire à la simple question de choisir entre H_0 et H_a, cela ne veut pas dire que la logique dicte forcément qu'il faut choisir un α petit. Dans certains contextes en ingénierie, les conséquences d'une erreur de type 2 sur le plan pratique font en sorte qu'une prise de décision rationnelle doit trouver un équilibre entre un petit α et un petit β, deux possibilités qui s'opposent l'une à l'autre.

5.1.6 Signification statistique, estimation et importance pratique

QUELQUES COMMENTAIRES CONCERNANT LES TESTS D'HYPOTHÈSE ET L'ESTIMATION

L'estimation d'un l'intervalle de confiance et le test d'hypothèse sont les deux formes d'inférence statistique formelle les plus couramment utilisées. À la lumière de leur présentation, il convient de faire quelques observations comparatives sur leur utilité pratique et, ce faisant, de faire état d'une *orientation en matière d'estimation* dont il sera question dans la majeure partie du traitement de l'inférence formelle dans le reste de cet ouvrage.

La plupart du temps, on se demande « Quelle est la valeur du paramètre? » plutôt que « Le paramètre équivaut-il à une valeur hypothétique? » Pour répondre à la première question, c'est à l'estimation de l'intervalle de confiance et non aux tests d'hypothèse qu'il faudra recourir. Un intervalle de confiance pour la moyenne du couple de rupture de 9,9 à 13,1 po oz indique quelles valeurs de μ semblent plausibles. Un infime seuil de signification observé dans le test de H_0 : μ = 33,5 indique seulement que les données s'opposent manifestement à la possibilité que μ = 33,5, mais il ne donne pas d'indice sur la valeur probable de μ.

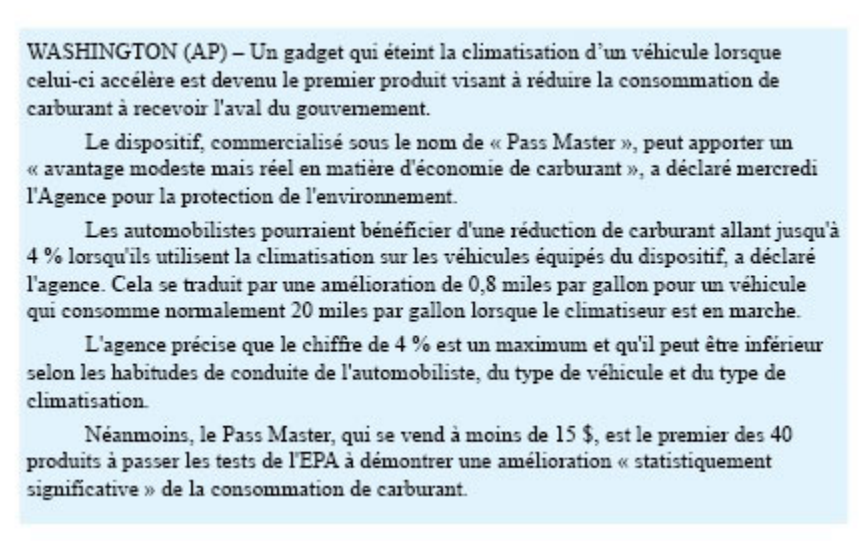
Signification statistique et importance pratique

Le fait que les tests d'hypothèse ne donnent aucune indication utile sur les valeurs plausibles des paramètres peut être occulté par une interprétation maladroite du langage semi-standard. Par exemple, il est courant dans certains domaines d'appeler les valeurs p inférieures à 0,05 « statistiquement significatives » et celles inférieures à 0,01 « hautement significatives ». Le piège dans ce type d'utilisation est que « significatif » peut être compris, à tort, comme synonyme de « grande conséquence pratique » et que la valeur p peut être interprétée, à tort, comme une mesure de l'écart entre un paramètre et la valeur énoncée dans l'hypothèse nulle. L'une des raisons pour lesquelles cette interprétation est faussée, c'est que le seuil de signification observé dans un test dépend non seulement de l'écart entre H_0 et la réalité, mais aussi de la taille de l'échantillon. Avec un échantillon de taille suffisante, tout écart par rapport à H0 peut être considéré comme « hautement significatif », qu'il ait une importance pratique ou non.

Exemple 5.1.6.1 Signification statistique et importance pratique dans le cadre d'un essai réalisé par un organisme de réglementation

L'article de presse de la figure 5.1.6.1 illustre parfaitement les points précédents. Le fabricant du Pass Master a effectué suffisamment de tests physiques de consommation d'essence (avec un n suffisamment élevé) pour produire une valeur p inférieure à 0,05 afin de tester l'hypothèse nulle de l'absence d'amélioration du kilométrage. Autrement dit, il a obtenu un résultat « statistiquement significatif ».

Cependant, l'amélioration du kilométrage réel rapportée reste « faible mais réelle », puisqu'elle s'élève à environ 0,8 mpg. Le fait que cette amélioration revête une importance pratique ou non reste une question nettement distincte du résultat du test d'hypothèse. Si on dispose d'un intervalle de confiance pour la moyenne d'amélioration du kilométrage, on se trouve en meilleure posture pour juger de l'importance pratique
que si on n'a qu'une valeur p inférieure à 0,05.



WASHINGTON (AP) – Un gadget qui éteint la climatisation d'un véhicule lorsque celui-ci accélère est devenu le premier produit visant à réduire la consommation de carburant à recevoir l'aval du gouvernement.

Le dispositif, commercialisé sous le nom de « Pass Master », peut apporter un « avantage modeste mais réel en matière d'économie de carburant », a déclaré mercredi l'Agence pour la protection de l'environnement.

Les automobilistes pourraient bénéficier d'une réduction de carburant allant jusqu'à 4 % lorsqu'ils utilisent la climatisation sur les véhicules équipés du dispositif, a déclaré l'agence. Cela se traduit par une amélioration de 0,8 miles par gallon pour un véhicule qui consomme normalement 20 miles par gallon lorsque le climatiseur est en marche.

L'agence précise que le chiffre de 4 % est un maximum et qu'il peut être inférieur selon les habitudes de conduite de l'automobiliste, du type de véhicule et du type de climatisation.

Néanmoins, le Pass Master, qui se vend à moins de 15 $, est le premier des 40 produits à passer les tests de l'EPA à démontrer une amélioration « statistiquement significative » de la consommation de carburant.

Figure 5.1.6.1 Article du Lafayette Journal and Courier, page D-3, 28 août 1980. Copyright 1980 de l'Associated Press. Réimprimé avec l'autorisation de l'Associated Press dans l'ouvrage de Stephen B. Vardeman et J. Marcus Jobe, Basic Engineering Data Collection and Analysis (figure 6.8 du chapitre 6).

Exemple 5.1.6.2 (suite)

Pour illustrer l'effet de la taille de l'échantillon sur le seuil de signification observé, reprenons l'exemple du couple de rupture et examinons deux échantillons hypothétiques, l'un de n = 25 et l'autre de n = 100, mais tous deux donnant $\bar{x}$ = 32,5 po oz et s = 5,1 po oz.

Pour les essais H_0 : μ = 33,5 avec H_a : μ<33,5, le premier échantillon hypothétique donne :

$$z = \frac{32{,}5 - 33{,}5}{\frac{5{,}1}{\sqrt{25}}} = -0{,}98$$

avec un seuil de signification observé associé de

$$\Phi(-0{,}98) = 0{,}16$$

Le deuxième échantillon hypothétique donne

$$z = \frac{32.5 - 33.5}{\frac{5.1}{\sqrt{100}}} = -1.96$$

avec une valeur p correspondante de

$$\Phi(-1{,}96) = 0{,}02$$

La taille du deuxième échantillon étant plus importante, celui-ci démontre plus clairement que la moyenne du couple de démarrage est inférieure à 33,5 po oz. Mais la meilleure supposition en fonction des données concernant la différence entre μ et 33,5 est $\bar{x}$ dans les deux cas. Or, c'est précisément l'ampleur de la différence entre μ et 33,5 po oz qui revêt une importance primordiale en ingénierie.

En outre, il importe de savoir qu'en plus de sa fonction principale, qui consiste à fournir un intervalle de valeurs plausibles pour un paramètre, l'intervalle de confiance apporte également des informations relatives au test d'hypothèse. Par exemple, un intervalle de confiance à 95 % pour un paramètre contient toutes les valeurs du

paramètre pour lesquelles les tests d'hypothèse effectués à l'aide des données disponibles produiraient des valeurs p supérieures à 5 %. (Les valeurs non couvertes par l'intervalle auraient des valeurs p associées inférieures à 5 %.)

Exemple 6.1.6.3 (suite)

Au chapitre 5.1.1, il a été démontré que l'intervalle de confiance unilatéral à 90 % pour la moyenne du couple de rupture des disques durs défectueux est de (-∞, 12,8). Cela signifie que pour toute valeur # supérieure à 12,8 po oz, un test d'hypothèse de H_0 : μ = # avec H_a : μ < # produirait une valeur p inférieure à 0,1. Ainsi, il apparaît clairement que le seuil de signification observé correspondant à l'hypothèse nulle H_0 : μ = 33,5 est inférieur à 0,1. (En fait, comme il a été vu plus haut dans ce chapitre, la valeur p est de 0 à la deuxième décimale.) En termes plus simples, l'intervalle (-∞, 12,8) est encore loin de contenir 33,5 po oz, ce qui rend une telle valeur de μ peu plausible.

La réflexion menée ici pourrait bien soulever la question suivante : « Dans la pratique, à quoi les tests d'hypothèse peuvent-ils servir? » Voici quelques réponses pertinentes à cette question :

1. D'une certaine manière, les valeurs p peuvent servir à évaluer dans quelle mesure les données disponibles sont peu probantes. Un seuil de signification élevé signifie qu'il faut obtenir plus d'informations pour parvenir à un jugement décisif.

2. Parfois, la loi impose l'utilisation de tests d'hypothèse dans le cadre d'une démonstration de conformité ou d'efficacité. (C'était le cas dans l'exemple 5.1.6.2, où la commercialisation du Pass Master exigeait une démonstration légale de la réduction de la consommation d'essence.)

3. Dans certains cas, l'utilisation de tests d'hypothèses dans un cadre de prise de décision se révèle nécessaire et pertinente. (L'échantillonnage pour acceptation en est un exemple : à partir d'informations tirées d'un échantillon d'articles provenant d'un lot volumineux, on doit décider si on réceptionne le lot ou non.)

4. À titre de preuves supplémentaire et de compléments aux rapports ou aux résultats de publications scientifiques.

Ainsi, lorsque les tests d'hypothèses sont correctement interprétés et utilisés, ils trouvent leur place dans la pratique de l'ingénierie. Par conséquent, bien que le reste de cet ouvrage mette l'accent sur l'estimation plutôt que sur les tests d'hypothèse, on ne peut négliger les méthodes utilisées dans le cadre des tests d'hypothèse.

5.2.0 Inférence sur les moyennes à partir d'un et de deux échantillons – Introduction

La partie 5 a présenté les concepts de base de l'estimation de l'intervalle de confiance et des tests d'hypothèse. Il existe des milliers de méthodes propres à ces deux outils. Cet ouvrage ne peut en aborder qu'une petite partie, soit celles qui sont les mieux connues et les plus utiles en ingénierie. Les sections suivantes examinent les plus élémentaires d'entre elles, dont certaines qui s'appliquent aux études à un ou deux échantillons, en commençant dans la présente section par les méthodes d'inférence formelle pour les moyennes.

Les inférences pour une moyenne unique, basées non pas sur les grands échantillons de la partie 5 mais sur de petits échantillons, sont traitées dans un premier temps. Au cours de ce processus, il est indispensable d'introduire les distributions t (ou distributions de Student). On présentera ensuite des méthodes d'inférence formelle pour les données appariées. La section se termine par l'étude des méthodes de comparaison de deux moyennes tirées d'échantillons indépendants, avec n = grand et n = petit.

5.2.1 Inférence pour une moyenne unique sur un petit échantillon

La principale limitation pratique à l'utilisation des méthodes évoquées dans les deux chapitres précédents, c'est que n doit être grand. Cette restriction résulte du fait que, si ce n'est pas le cas, il est impossible de conclure que la variable

5.2.1.1 $$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

est approximativement normale réduite. Donc, si on utilise machinalement la formule de l'intervalle de confiance pour un grand n

5.2.1.2 $$\bar{x} \pm z\frac{S}{\sqrt{n}}$$

sur un petit échantillon, il s'avérera impossible d'évaluer le niveau de confiance réel. Autrement dit, pour un petit n, l'utilisation de z = 1,96 dans l'équation 5.2.1.2 ne génère généralement pas d'intervalles de confiance à 95 %. Sans condition supplémentaire, il n'y a ni moyen de savoir quelle confiance est associée à z = 1,96 ni moyen de savoir comment choisir z de façon à obtenir un niveau de confiance de 95 %.

Il y a un cas particulier et important pour lequel un raisonnement parallèle à celui de la partie 5 permet d'obtenir des méthodes d'inférence pour la moyenne d'un petit échantillon : lorsqu'on peut modéliser les observations en tant que variables aléatoires normales iid. Le cas des observations normales est pratique car, même si la variable 5.2.1.1 n'est pas normale réduite, sa distribution est connue et illustrée sous forme de tableau. Il s'agit de la distribution t de Student.

DÉFINITION 5.2.1.1 La distribution t de Student

La distribution t (de Student) avec un paramètre de degrés de liberté v est une distribution de probabilité continue ayant une densité de probabilité

EXPRESSION 5.2.1.3

f(t)=\frac{\Gamma\left(\frac{v+1}{ 2}\right)}{\Gamma\left(\frac{v}{ 2}\right) \sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1) / 2}alttitlef(t)=\frac{\Gamma\left(\frac{v+1}{ 2}\right)}{\Gamma\left(\frac{v}{ 2}\right) \sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-(v+1) / 2}title

pour tout t.

Si une variable aléatoire présente une densité de probabilité donnée par l'équation 5.2.1.3, on dit qu'elle suit une distribution t_v .

.

Le terme « Student » employé dans la définition 5.2.1.1 est en fait le nom de plume du premier statisticien à avoir découvert l'équation 5.2.1.3. Cette expression présente un caractère plutôt impressionnant. Le présent ouvrage ne prévoit aucun calcul direct à l'aide de cette formule. Toutefois, il s'avère utile de l'avoir à disposition pour

esquisser quelques densités de probabilité t, afin de se faire une idée de leur forme. La figure 5.2.1.1 illustre les densités t pour les degrés de liberté v = 1, 2, 5, et 11, ainsi que la densité normale réduite.

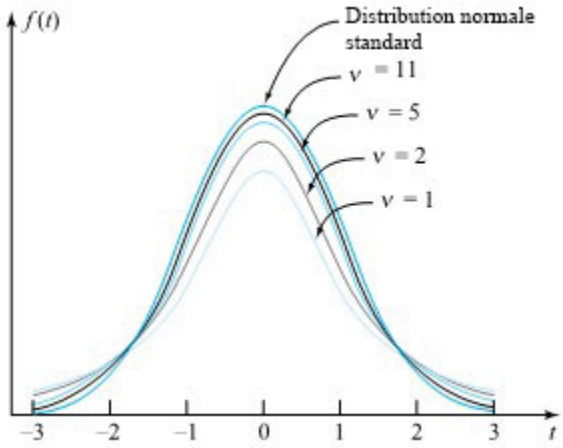


Figure 5.2.1.1 Densités de probabilité t pour v = 1, 2, 5 et 11, et densité normale réduite

Distributions t et distribution normale réduite

Le message véhiculé par la figure 5.2.1.1 est que les densités de probabilité t ont la forme d'une cloche et sont symétriques par rapport à t = 0. Elles présentent un aspect plus plat que la densité normale réduite, mais s'en rapprochent de plus en plus à mesure que v augmente. En fait, dans la plupart des cas, pour des v supérieurs à environ 30, la distribution t avec ν degrés de liberté et la distribution normale standard sont impossibles à distinguer.

Généralement, pour calculer une probabilité associée à une distribution t, on n'utilise pas l'expression 5.2.1.3, car il n'existe pas d'antidérivée simple pour $f(t)$. On utilise plutôt des tables (ou des logiciels statistiques) pour évaluer les quantiles communs de la distribution t et ainsi obtenir des limites approximatives sur les types de probabilités nécessaires pour les tests d'hypothèse. La table A1.3 de l'annexe 1 des tables statistiques représente un tableau typique de quantiles t. Les colonnes représentent les probabilités cumulatives et les lignes, les valeurs du paramètre des degrés de liberté, v. Le corps de la table énumère les quantiles correspondants. À noter que la dernière ligne du tableau est une ligne « v =∞ » (c.-à-d. distribution normale réduite).

Exemple 5.2.1.1 Utilisation des tables de quantiles de distribution t

Soit T une variable aléatoire ayant une distribution t avec v = 5 degrés de liberté. Trouvons d'abord le quantile 0,95 de la distribution de T, puis voyons ce que le tableau A1.3 révèle sur P [T < -1,9] et ensuite sur P [|T | > 2,3].

Tout d'abord, si l'on examine la ligne v = 5 du tableau A1.3 sous la probabilité cumulative 0,95, on trouve 2,015 dans le corps du tableau. Autrement dit, Q(0,95) = 2,015 ou (de manière équivalente) P [T ≤ 2,015] = 0,95.

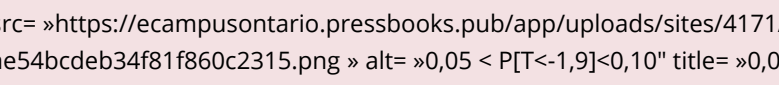
On notera alors que par symétrie,

P[T<-1,9]=P[T>1,9]=1-P[T \leq 1,9]alttitleP[T<-1,9]=P[T>1,9]=1-P[T \leq 1,9]title

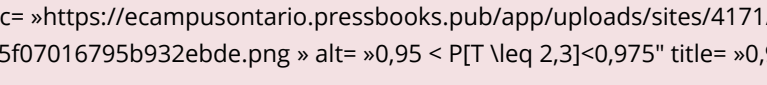
En regardant la ligne v = 5 du tableau A1.3, on constate que 1,9 se situe entre les quantiles 0,90 et 0,95 de la distribution t_5. Autrement dit,

<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/e1890f6402fcf9259a7b0f0f04a0da1d.png » alt= »0,90 < P[T \leq 1,9] \leq 0,95" title= »0,90

Finalement :



Ensuite, en regardant la ligne v = 5 du tableau A1.3, on constate que 2,3 se situe entre les quantiles 0,95 et 0,975 de la distribution t_5. Autrement dit,



donc

Les trois calculs de cet exemple apparaissent à la figure 5.2.1.2.

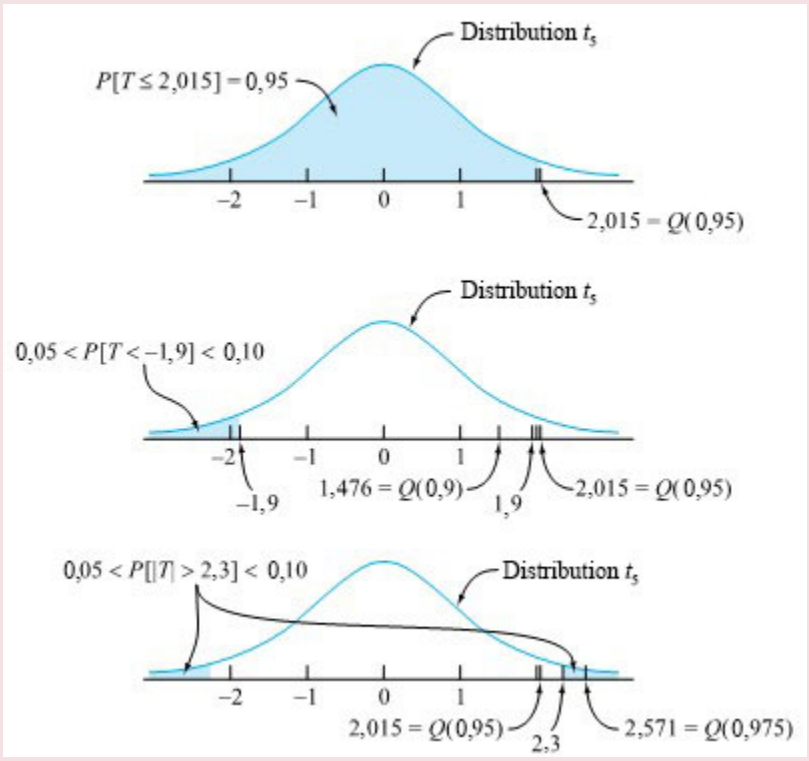


Figure 5.2.1.2 Les trois calculs de probabilité t_5 de l'exemple 5.2.1.1.

La relation entre les expressions 5.2.1.3 et 5.2.1.1 qui permet de développer des méthodes d'inférence à petit n pour les observations normales se résume au fait que si un modèle normal iid convient,

5.2.1.4 $$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

suit une distribution t avec v = n − 1 degrés de liberté. (Ce résultat cohérent avec le principe de base utilisé dans les deux chapitres précédents : pour un grand n, v est grand, donc la distribution t_v devient approximativement normale réduite; et pour un grand n, la variable 5.2.1.4 a déjà été traitée comme étant approximativement normale réduite.)

Puisque la variable 5.2.1.4 peut, dans des circonstances favorables, être traitée comme une variable aléatoire altt_{n-1}alttitlet_{n-1}title, il est possible de reprendre exactement ce qui a été fait à la partie 5 pour trouver

des méthodes permettant d'établir des intervalles de confiance et d'effectuer des tests d'hypothèse. Autrement dit, si on peut considérer qu'un mécanisme de génération de données équivaut fondamentalement à tirer des observations indépendantes d'une distribution normale unique, un intervalle de confiance bilatéral pour μ comporte des des bornes

EXPRESSION 5.2.1.5 Bornes de distribution normale de confiance pour μ

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

où t est choisie de sorte que la distribution t_{n-1} attribue une probabilité correspondant au niveau de confiance souhaité à l'intervalle compris entre -t et t. De même, l'hypothèse nulle

$$\mathrm{H}_0 : \mu = \#$$

peut être testée à l'aide de la statistique

EXPRESSION 5.2.1.6 Statistique de test de la distribution normale pour μ

$$T = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

et d'une distribution t_{n-1} de référence.

Sur le plan opérationnel, la seule différence entre les méthodes d'inférence exposées ici et les méthodes à grand échantillon des deux chapitres précédents est qu'au lieu d'utiliser les quantiles et les probabilités de la distribution normale réduite, on utilise ceux de la distribution t_{n-1}. *Sur le plan conceptuel*, les propriétés de confiance et de signification nominales ne sont pertinentes en pratique que sous la condition supplémentaire d'une distribution sous-jacente raisonnablement normale. Avant d'utiliser les expressions 5.2.1.5 et 5.2.1.6, il est recommandé d'examiner la pertinence du modèle de distribution normale.

Exemple 5.2.1.2 Bornes de confiance pour la moyenne de la durée de vie d'un ressort sur un petit échantillon

Une partie d'un ensemble de données de W. Armstrong (figurant dans *Analysis of Survival Data*, de Cox and Oakes) fournit le nombre de cycles jusqu'à la rupture de 10 ressorts du même type sous une contrainte de 950 N/mm^2. Ces observations relatives à la durée de vie des ressorts se retrouvent dans le tableau 5.2.1.1 en unités de 1 000 cycles.

Cycles de défaillances de dix ressorts sous une contrainte de 950 N/mm² (10^3 cycles)
Durée de vie du ressort
225, 171, 198, 189, 189 135, 162, 135, 117, 162

Tableau 5.2.1.1. Dans ce scénario, on pourrait se demander : « Quelle est la durée de vie moyenne d'un ressort soumis à une contrainte de 950 N/mm^2? » Comme il n'y a que de n = 10 observations, la méthode pour les grands échantillons vue au module 5.1 ne s'applique pas. Seule la méthode indiquée par l'expression 5.2.1.5 peut potentiellement être utilisée, et pour qu'elle convienne, les durées de vie doivent être distribuées normalement.

Faute de disposer d'expérience pertinente en ingénierie des matériaux, il apparaît difficile de spéculer *a priori* sur la justesse d'un modèle de durée de vie normale dans ce contexte. Néanmoins, il reste possible d'examiner les données du tableau 5.2.1.1 pour vérifier si elles présentent un écart important par rapport à la distribution normale. La figure 5.2.1.3 illustre un tracé normal des données, qui suggère que les données semblent être normalement distribuées.

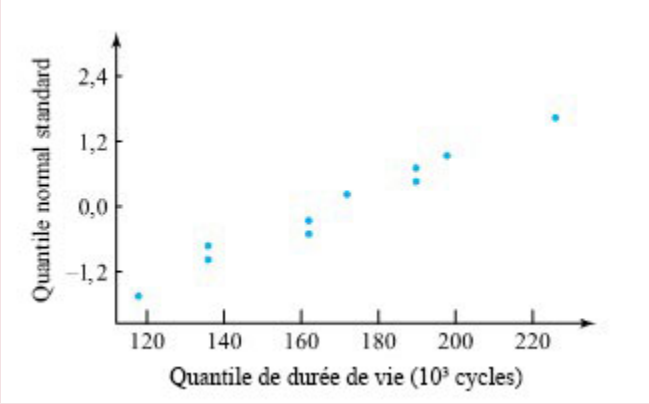


Figure 5.2.1.3 Tracé normal de la durée de vie d'un ressort

Pour les 10 durées de vie, $\bar{x}$ = 168,3(× 10^3 cycles) et s = 33,1(×10^3 cycles). Ainsi, pour estimer la durée de vie moyenne du ressort, ces valeurs peuvent être utilisées dans l'expression 5.2.1.5, avec une valeur de t convenablement choisie. En utilisant, par exemple, un niveau de confiance de 90 % et un intervalle bilatéral, t correspond au quantile 0,95 de la distribution t avecυ degrés de liberté. Autrement dit, on utilise la distribution t_9 et on choisit t > 0 de sorte que

$$P[-t < \text{une variable aléatoire } t_9 < t] = 0,90$$

En consultant le tableau A1.3, le choix t = 1,833 s'impose. Ainsi, l'intervalle de confiance bilatéral de 90 % pour μ comporte les bornes

$$168,3 \pm 1,833\frac{33,1}{\sqrt{10}}$$

soit

$$168.3 \pm 19.2$$

c.-à-d.

$$149,1 \times 10^3 \text{ cycles et } 187,5 \times 10^3 \text{ cycles}$$

VÉRIFICATION DES TRACÉS NORMAUX

Comme l'illustre l'exemple 5.2.1.2, produire un tracé normal des données pour vérifier sommairement la plausibilité d'une distribution normale sous-jacente constitue une pratique valable, utilisée à maintes reprises dans ce manuel. Cependant, il faut éviter de s'attendre à plus que la méthode ne le justifie. Il vaut assurément mieux l'utiliser plutôt que de présumer, sans preuve, qu'une distribution est normale, ce qui pourrait ne pas être le cas. Mais il faut aussi reconnaître que lorsqu'elle est utilisée sur de petits échantillons, la méthode fournit rarement des indications décisives quant à la pertinence d'un modèle normal. Les petits échantillons issus de distributions normales ne présenteront souvent que des tracés normaux d'apparence légèrement linéaire. Parallèlement, les petits échantillons provenant de distributions assez anormales peuvent souvent présenter des courbes normales relativement linéaires. En somme, en raison de la variabilité de l'échantillonnage, les petits échantillons ne fournissent pas beaucoup d'informations sur la forme de la distribution sous-jacente. Tout ce que l'on peut attendre d'un tracé normal préliminaire sur un petit échantillon, tel que celui de l'exemple 5.2.1.2, c'est un avertissement en cas d'écart flagrant par rapport à la normalité –une distribution sous-jacente beaucoup plus lourde dans les queues qu'une distribution normale (c'est-à-dire produisant plus de valeurs extrêmes qu'une forme normale ne le ferait).

TESTS D'HYPOTHÈSES POUR μ SUR PETITS ÉCHANTILLONS

L'exemple 5.2.1.2 montre comment utiliser la formule de l'intervalle de confiance 5.2.1.5, mais pas comment faire un test d'hypothèse (équation 5.2.1.6). Puisque la méthode sur petit échantillon s'apparente exactement à la méthode sur grand échantillon du module 5.1 (en utilisant la distribution t plutôt que la distribution normale réduite), et que la source d'où proviennent les données n'indique aucune valeur de μ qui pourrait d'emblée constituer l'hypothèse nulle, l'utilisation de la méthode correspondant à l'expression 5.2.1.6 ne sera pas illustrée à ce stade.

5.2.2 Comparaisons de deux moyennes sur un grand échantillon (basées sur des échantillons indépendants)

Tournons-nous maintenant vers les méthodes pouvant être utilisées pour comparer deux moyennes tirées de deux échantillons distincts « sans lien de parenté », en commençant par les méthodes pour les grands échantillons.

Exemple 5.2.2.1 Comparaison des propriétés d'empilement de morceaux moulés et concassés d'un solide

Une entreprise voulait trouver une géométrie fonctionnelle pour les pièces moulées d'un solide. Une des comparaisons effectuées portait sur le poids de pièces versées dans un contenant donné, en s'attardant sur la différence entre les pièces moulées selon une certaine géométrie et es pièces irrégulières obtenues par concassage. Une série de 24 tentatives de remplissage de morceaux moulés et concassés du solide a permis d'obtenir les données (en grammes) présentées à la figure 5.2.2.1 sous de diagrammes à tiges et à feuilles juxtaposés.

On remarque que, bien que la figure présente le même nombre de masses de pièces moulées que de masses de pièces concassés, les deux types d'échantillons sont nettement différents. Cette situation ne se compare en rien à celles de différence appariée traitées dans un autre chapitre, ce qui suggère d'utiliser une autre méthode d'inférence statistique.

Moulage		Concassage
7,9	11	
4,5, 3,6, 1,2	12	
9,8, 8,9, 7,9, 7,1, 6,1, 5,7, 5,1	12	
2,3, 1,3, 0,0	13	
8,0, 7,0, 6,5, 6,3, 6,2	13	
2,2, 0,1	14	
	14	
2,1, 1,2, 0,2	15	
	15	
	16	1,8
	16	5,8, 9,6
	17	1,3, 2,0, 2,4, 3,3, 3,4, 3,7
	17	6,6, 9,8
	18	0,2, 0,9, 3,3, 3,8, 4,9
	18	5,5, 6,5, 7,1, 7,3, 9,1, 9,8
	19	0,0, 1,0
	19	

Figure 5.2.2.1 Diagrammes à tige et à feuille juxtaposés des masses d'empilement pour les pièces moulées et les pièces concassées.

Dans des situations comme celle de l'exemple 5.2.2.1, il est utile de noter les paramètres et les statistiques par des indices – par exemple, en prenant μ_1 et μ_2 pour représenter les moyennes distributionnelles sous-jacentes correspondant aux première et deuxième conditions et $\bar{x}_1$ et $\bar{x}_2$ pour représenter les moyennes de l'échantillon correspondantes. Or, si les deux mécanismes de génération de données correspondent essentiellement et conceptuellement à un échantillonnage avec remplacement à partir de deux distributions, la partie 4 indique

que $\bar{x}_1$ a une moyenne μ_1 et une variance σ_1^2/n_1, et que$\bar{x}_2$ a une moyenne $\bar{x}_2$ et une variance σ_2^2/n_2. La différence entre les moyennes des échantillons $\bar{x}_1 - \bar{x}_2$ est une statistique naturelle à utiliser pour comparer μ_1 et μ_2. Toujours selon la partie 4, s'il paraît raisonnable de percevoir les deux échantillons comme étant choisis séparément ou indépendants, cette variable aléatoire a l'espérance mathématique :

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

et la variance

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Si, en outre,n_1 et n_2 sont grands (de sorte que $\bar{x}_1$ et $\bar{x}_2$ sont toutes deux approximativement normales), $\bar{x}_1$ – $\bar{x}_2$ est approximativement normale. Ainsi,

EXPRESSION 5.2.2.1

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

suit une distribution de probabilité approximativement normale.

Comme la variable 5.2.2.1 est approximativement normale réduite, on peut obtenir un intervalle de confiance et à des méthodes de test d'hypothèse pour $\mu_1 - \mu_2$ en utilisant une logique exactement parallèle à celle des parties « σ connu » du module 5.1. Mais dans la pratique, il s'avère beaucoup plus utile de commencer par une expression sans σ_1 ni σ_2. Heureusement, si n_1 et n_2 sont grands, non seulement la variable 5.2.2.1 est approximativement normale réduite, mais il en va de même pour

EXPRESSION 5.2.2.2

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Ainsi, la logique standard du module 5.1 démontre que l'intervalle de confiance bilatéral de la différence des moyennes $\mu_1 - \mu_2$, basé sur deux grands échantillons indépendants, a les bornes suivantes :

EXPRESSION 5.2.2.3 Bornes de confiance pour $\mu_1 - \mu_2$ (n = grand)

$$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

où z est choisie de sorte que la probabilité que la distribution normale standard attribue à l'intervalle entre -z et z correspond à la confiance souhaitée. Et la logique exposée au module 5.2 démontre que, dans les mêmes conditions,

$$\mathrm{H}_0 : \mu_1 - \mu_2 = \#$$

peut être testée à l'aide de la statistique

EXPRESSION 5.2.2.4 Statistique de test pour $\mu_1 - \mu_2$ (n = grand)

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

et d'une distribution normale réduite de référence.

Exemple 5.2.2.2 suite.

Dans le problème du moulage, on s'attendait *a priori* à ce que les pièces concassées s'empilent mieux que les pièces moulées (qui, autrement, conviennent mieux). Mesurons la signification statistique de la différence entre les poids moyens et établissons un intervalle de confiance unilatéral à 95 % pour cette différence (ce qui revient à affirmer que le
de la différence de poids moyen du concassé moins le poids moyen du moulé équivaut au moins à un certain nombre).

La taille des échantillons (n_1 = n_2 = 24) se situe à la limite de ce que l'on peut qualifier de grand. Il aurait été préférable d'avoir quelques observations de plus pour chaque type, mais faute de quoi, on utilisera la méthode des expressions 5.2.2.3 et 5.2.2.4, tout en faisant preuve de réserve à l'égard des résultats si ces derniers conduisaient à une décision « serrée » au sens de l'ingénierie ou des affaires.

En étiquetant arbitrairement la condition 1 « concassé » et la condition 2 « moulé » et en calculant à partir des données de la figure 5.2.2.2 que alt\bar{x}_1alttitle\bar{x}_1title = 179,55 g, s_1 = 8,34 g, $\bar{x}_2$ 132,97 g et s_2 = 9,31 g, le modèle de test d'hypothèse en cinq étapes conduit au récapitulatif suivant :

1. $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
2. $\mathrm{H}_\mathrm{a} : \mu_1 - \mu_2 > 0$0" title= »\mathrm{H}_{\mathrm{a}}: \mu_1-\mu_2>0" class= »latex mathjax »>

(L'hypothèse de recherche retenue ici est que la moyenne du concassé surpasse la moyenne du moulé, de sorte que la différence, prise dans cet ordre, est positive.)

3. La statistique de test est la suivante :

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

La distribution de référence est normale réduite, et de grandes valeurs |z| observées constitueront une preuve contre H_0 et en faveur de H_a.

4. Les échantillons donnent

$$z = \frac{179.55 - 132.97 - 0}{\sqrt{\frac{(8.34)^2}{24} + \frac{(9.31)^2}{24}}} = 18.3$$

5. Le seuil de signification observé correspond à P[une variable normale réduite ≥ 18,3] ≈ 0. Les données indiquent de manière irréfutable que $\mu_1 \mu_2$ 0" title= »\mu_2> 0" class= »latex mathjax »> : le poids moyen d'empilement des pièces concassées surpasse celui des pièces moulées.

En ce qui à trait à l'intervalle de confiance unilatéral pour $\mu_1 - \mu_2$, il convient de noter que seule la borne inférieure donnée dans l'équation 5.2.2.3 sera utilisée. Par conséquent, z = 1,645 conviendra. Autrement dit, avec une confiance de 95 %, on peut conclure que la différence entre les moyennes (concassé moins moulé) surpasse

$$(179{,}55 - 132{,}97) - 1{,}645\sqrt{\frac{(8{,}34)^2}{24} + \frac{(9{,}31)^2}{24}}$$

Autrement dit, elle surpasse :

$$46{,}58 - 4{,}20 = 42{,}38 \text{ g}$$

Formulé autrement, l'intervalle de confiance unilatéral à 95 % pour $\mu_1 - \mu_2$ correspond à

$$(42{,}38, \infty)$$

Les étudiant.e.s éprouvent parfois un certain malaise face au choix arbitraire qu'implique l'étiquetage des deux conditions dans une étude à deux échantillons. En réalité, les deux options peuvent être utilisés. Pour autant qu'on respecte ce choix tout au long du raisonnement, il n'affectera aucunement les conclusions tirées dans le monde réel. Dans l'exemple 5.5.2.2, si la condition « moulé » est désignée par le numéro 1 et la condition « concassé » par le numéro 2, l'intervalle de confiance la moyenne « moulé » moins la moyenne « concassé » est la suivante :

$$(-\infty, -42{,}38)$$

Concrètement, cet intervalle a exactement le même sens que celui de l'exemple.

Rappelons que les présentes méthodes s'appliquent lorsque des mesures uniques sont réalisées sur chaque élément de deux échantillons différents. Ceci contraste avec les questions relatives aux données appariées (où il y a des observations à deux variables sur un seul échantillon); nous reviendrons à ce cas plus tard.

5.2.3. Comparaisons de deux moyennes sur un petit échantillon (basée sur des échantillons indépendants suivant une distribution normale)

Les dernières méthodes d'inférence présentées dans cette section correspondent à la différence entre deux moyennes dans les cas où au moins l'une des deux tailles d'échantillon n_1 et n_2 est petite. Toute la discussion se limitera à des cas où les observations sont normales. En fait, les méthodes les plus directes sont réservées aux cas où, en plus, les deux écarts-types sous-jacents sont comparables. Nous commencerons celles-là.

VÉRIFICATION GRAPHIQUE DE LA PLAUSIBILITÉ DU MODÈLE

Un moyen de vérifie sommairement la plausibilité des suppositions du modèle « distributions normales, même variance » consiste à effectuer un tracé normal de deux échantillons sur le même ensemble d'axes, en vérifiant non seulement la linéarité approximative, mais aussi l'égalité approximative de la pente.

Exemple 5.2.3.1 (suite)

Les données de W. Armstrong sur la durée de vie des ressorts (figurant dans l'ouvrage de Cox et Oakes) concernent non seulement la longévité des ressorts sous une contrainte de 950 mm^2 mais aussi sous une contrainte de 900 mm^2. Le tableau 5.2.3.1 reprend les données de 950 mm^2 précédentes et y ajoute celles pour la contrainte de 900 mm^2.

Durées de vie du ressort sous deux contraintes différentes (10^3 cycles)

Contrainte de 950 N/mm²	Contrainte de 900 N/mm²
225, 171, 198, 189, 189	216, 162, 153, 216, 225
135, 162, 135, 117, 162	216, 306, 225, 243, 189

Tableau 5.2.3.1

La figure 5.2.3.1 montre des tracés normaux pour les deux échantillons sur un seul ensemble d'axes. Compte tenu du type de variation de la linéarité et de la pente que présentent les tracés normaux pour des échantillons de cette taille (n = 10) issus d'une distribution normale unique, la figure 5.2.3.1 ne constitue nullement une preuve solide contre la pertinence d'un modèle de « variances égales, distributions normales » pour la durée de vie des ressorts.

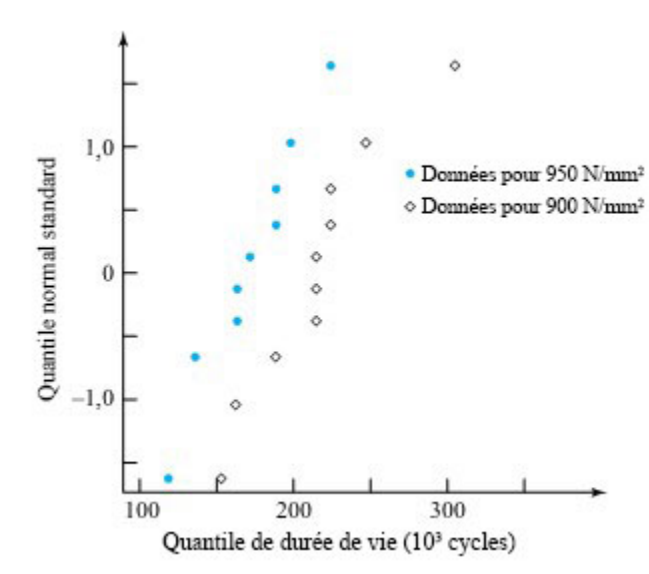


Figure 5.2.3.1 Tracés normaux de la durée de vie des ressorts sous deux contraintes différentes

VARIANCE PONDÉRÉE D'UN ÉCHANTILLON

Si on suppose que $\sigma_1 = \sigma_2$, la valeur commune se nomme σ, et il apparaît logique que s_1 et s_2 se rapprochent tous deux de σ. Cela suggère qu'il faudrait les combiner pour obtenir une estimation unique de la variation réelle. Il s'avère que la convention mathématique impose une méthode particulière de combinaison ou de *pondération* des s individuels afin d'obtenir une estimation unique de σ.

DÉFINITION Variance pondérée d'un échantillon sp^2

EXPRESSION 5.2.3.1

Si deux échantillons numériques de tailles respectives n_1 et n_2 produisent des variances d'échantillon respectives s_1^2 et s_2^2, la variance pondérée de l'échantillon, sp^2 est la moyenne pondérée de s_1^2 et s_2^2 où les coefficients de pondération correspondent aux tailles d'échantillon moins 1. Autrement dit,

$$s_{\mathrm{P}}^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

L'écart type pondéré de l'échantillon s_P est égal à la racine carrée de sp^2 .

s_P est une sorte de moyenne de s_1 et de s_2 qui se trouve forcément entre s_1 et s_2. Sa forme exacte est davantage dictée par souci de convention mathématique que par une intuition logique.

Exemple 5.2.3.2 (suite)

Dans le cas de la durée de vie des ressorts, en choisissant arbitrairement de désigner 900 mm^2 la condition 1 et 950 mm^2 la condition 2, on obtient s_1 = 42,9 \cdot 10^3 cycles et s_2 cycles. En regroupant les deux variances de l'échantillon par l'équation 5.2.3.1, on obtient :

$$s_{\mathrm{P}}^2 = \frac{(10-1)(42{,}9)^2 + (10-1)(33{,}1)^2}{(10-1)+(10-1)} = 1{,}468\left(10^3 \text{ cycles}\right)^2$$

Puis, en prenant la racine carrée, on trouve :

$$s_{\mathrm{P}} = \sqrt{1\ 468} = 38{,}3\ \left(10^3 \text{ cycles}\right)$$

Selon l'argument conduisant aux méthodes d'inférence à grand échantillon pour μ_1 – μ_2, la quantité

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

a été brièvement examinée. Lorsque σ_1 = σ_2 = σ, cette variable peut être réécrite comme suit :

5.2.3.3

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

On peut exploiter le fait que la variable 5.2.3.3 est normale réduite pour produire des méthodes d'estimation des intervalles de confiance et de tests d'hypothèse. Or, leur utilisation nécessiterait le paramètre σ. Par conséquent, plutôt que de commencer par l'expression 5.2.3.3, il est courant de remplacer σ dans l'expression (5.2.3.3) par s_P et de commencer par la quantité

5.2.3.4

$$T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\mathrm{P}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

La variable 5.2.3.4 a été construite expressément pour que, selon les suppositions du modèle actuel, elle suive une distribution de probabilité connue et représentée dans un tableau : la distribution t avec v = (n_1 – 1) + (n_2 – 1) = n_1 + n_2 degrés de liberté. (On peut remarquer que les n_1 degrés de liberté associés au premier échantillon s'ajoutent aux n_2 degrés de liberté associés au second pour produire $n_1 n_2$ degrés de liberté au total.) Ainsi, toujours au moyen du type de raisonnement développé dans les modules 5.1 et 5.2, on peut obtenir des méthodes d'inférence pour alt\mu_1-alttitle\mu_1-titleμ_2. Autrement dit, un intervalle de confiance bilatéral pour la différence $\mu_1 \mu_2$, basé sur des échantillons indépendants provenant de distributions normales de même variance, aura pour bornes

EXPRESSION 5.2.3.5 Bornes confiance pour $\mu_1 mu_2$ suivant des distributions normales avec σ_1 = σ_2

$$\bar{x}_1 - \bar{x}_2 \pm t s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

dans laquelle on choisit t de sorte que la probabilité que la distribution$t_{n_1+n_2-2}$ attribue à l'intervalle entre -t et t correspond à la confiance souhaitée. Dans les mêmes conditions, l'hypothèse

$$\mathrm{H}_0 : \mu_1 - \mu_2 = \#$$

peut être testée à l'aide de la statistique

EXPRESSION 5.2.3.6 Statistique de test pour$\mu_1 mu_2$ suivant des distributions normales avec σ_1 = σ_2

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \#}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

et d'une distribution $t_{n_1+n_2-2}$ de référence.

Exemple 5.2.3.3 (suite)

Reprenons le cas de la durée de vie des ressorts pour illustrer l'inférence pour deux moyennes avec des petits échantillons. Tout d'abord, testons l'hypothèse d'une moyenne de durée de vie égale, avec l'hypothèse alternative que la contrainte plus faible conduit à une durée de vie plus longue. Ensuite, établissons un intervalle de confiance bilatéral à 95 % pour la différence entre les moyennes de durée de vie.

En continuant à désigner la contrainte de $mm2$ condition 1 et la contrainte de mm^2 condition 2, à partir du tableau 5.3.3.1, on obtient $\bar{x}_1$ et $\bar{x}_2$ et s_P = 38,3 (comme nous l'avons vu précédemment). Ainsi, le modèle de test d'hypothèse en cinq étapes donne ceci :

1. $\mathrm{H}_0 : \mu_1 - \mu_2 = 0$
2. $\mathrm{H}_a : \mu_1 - \mu_2 > 0$.0 . » title= »\mathrm{H}_{\mathrm{a}}: \mu_1-\mu_2>0 . » class= »latex mathjax »>
(Par raisonnement physique, on s'attend à ce que la condition 1 produise des durées de vie plus longues.)
3. La statistique de test est la suivante :

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

La distribution de référence est t, avec 10 + 10 – 2 = 18 degrés de liberté, et un grand t observé constituera une preuve contre H_0.

4. Les échantillons donnent

$$t = \frac{215{,}1 - 168{,}3 - 0}{38{,}3\sqrt{\frac{1}{10} + \frac{1}{10}}} = 2{,}7$$

5. Le seuil de signification observé $P[\text{une variable aléatoire } t_{18} \geq 2,7]$ se situe entre 0,01 et 0,005, ce qui constitue une preuve solide que la contrainte faible est associée à une durée de vie plus élevée en moyenne.

Par la suite, si on utilise l'expression 5.2.3.5 pour produire un intervalle de confiance bilatéral à 95 %, t correspond au quantile 0,975 de la distribution $t_1 8$. Les bornes de l'intervalle de confiance pour \μ_1 – μ_2 sont les suivantes :

$$(215\ 1 - 168{.}3) \pm 2{.}101(38{.}3)\sqrt{\frac{1}{10} + \frac{1}{10}}$$

soit :

$$46.8 \pm 36.0$$

ou encore :

$$10{,}8 \times 10^3 \text{ cycles et } 82{,}8 \times 10^3 \text{ cycles}$$

Les données du tableau 5.2.3.1 fournissent suffisamment d'informations pour confirmer qu'une contrainte plus forte entraîne une réduction de la moyenne de durée de vie des ressorts. Mais bien que l'ampleur apparente de cette réduction lors du passage de mm^2 (condition 1) à mm^2 (condition 2) soit de 46,8 10³ cycles, la variabilité présente dans les données est suffisamment grande (et la taille des échantillons suffisamment petite) pour que seule une précision de ±36,0 · 10³ cycles puisse être rattachée à cette différence.

INFÉRENCE POUR μ_1 – μ_2 SANS LA SUPPOSITION $\sigma_1 = \sigma_2$ (N = PETIT)

Il n'existe pas de réponse pleinement satisfaisante quant à la manière de réaliser l'inférence pour $\mu_1 - \mu_2$ lorsque l'on ne peut pas supposer que $\sigma_1 \sigma_2$. La méthode la plus répandue (mais approximative) pour résoudre ce problème est celle de Satterthwaite, qui se rapproche de la formule pour les grands échantillons (voir la section 5.2.1). Autrement dit, si les bornes de la section 5.2.1 ne conviennent pas lorsque altn_1alttitlen_1title ou n_2 est petit (elles ne produisent pas de niveaux de confiance réels à proximité du niveau nominal), il faut les modifier. Soit

EXPRESSION 5.3.3.7 « Degrés de liberté estimés » de Satterthwaite

$$\hat{v} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{(n_1-1)n_1^2} + \frac{s_2^4}{(n_2-1)n_2^2}}$$

Pour un niveau de confiance souhaité, supposons que $\hat{t}$ est telle que la distribution $t\hat{v}$ avec $\hat{v}$alt\hat{v}alttitle\ hat{v}title degrés de liberté attribue la probabilité correspondante à l'intervalle entre $-\hat{t}$ et $\hat{t}$. Ainsi, les deux bornes

EXPRESSION 5.2.3.8 Bornes de confiance pour $\mu_1\mu_2$ avec la distribution normale (approximative) de Satterthwaite

$$\bar{x}_1 - \bar{x}_2 \pm \hat{t}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

peuvent servir de bornes de confiance pour μ_1 – μ_2 avec un niveau de confiance approximativement égal à celui souhaité. (On peut utiliser une seul des bornes 5.2.3.8 pour obtenir un intervalle de confiance unilatéral en divisant le niveau de « non-confiance » par deux.)

Exemple 5.2.3.4 (suite)

Armstrong a collecté des données sur la durée de vie des ressorts soumis à d'autres contraintes que 900 et 950 mm^2 utilisés jusqu'à présent dans cet exemple. Dix ressorts testés à 850 mm^2 présentaient des durées de vie correspondant à $\bar{x}$ (chacune en 10^3 cycles) et un tracé normal relativement linéaire. Mais si on examine ensemble les données de 850, 900 et 950 mm^2, on voit clairement que les durées de vie deviennent plus courtes et plus uniformes à mesure qu'on augmente la contrainte. Si on compare les moyennes des durées de vie sous les contraintes de 850 et 950 mm^2, la supposition d'une variance constante semble douteuse.

On peut alors examiner ce que la méthode de Satterthwaite (expression 5.2.3.8) donne comme bornes de confiance bilatérales approximatives à 95 % pour la des moyennes à 850 et à 950 $mm2$. L'équation 5.2.2.7 donne :

$$\hat{v} = \frac{\left(\frac{(57{,}9)^2}{10} + \frac{(33{,}1)^2}{10}\right)^2}{\frac{(57{,}9)^4}{9(100)} + \frac{(33{,}1)^4}{9(100)}} = 14{,}3$$

En arrondissant les « degrés de liberté » à l'entier inférieur, le quantile 0,975 de la distribution $t_1 4$ est donc 2,145. Les bornes à 95 % de l'expression 5.3.3.8 pour la différence ($mm^2 mm^2$) des moyennes de durées de vie ($\mu_9 50$}) sont donc les suivantes :

$$348.1 - 168.3 \pm 2.145\sqrt{\frac{(57.9)^2}{10} + \frac{(33.1)^2}{10}}$$

soit :

$$179.8 \pm 45.2$$

ou encore :

$$134{,}6 \times 10^3 \text{ cycles} \quad \text{et} \quad 225{,}0 \times 10^3 \text{ cycles}$$

REMARQUES AU SUJET DES MÉTHODES SUR PETITS ÉCHANTILLONS

Les méthodes exposées dans cette section sont les dernières méthodes d'inférence standard pour les moyennes de un ou deux échantillons. Nous étudierons maintenant une méthode parallèle pour les variances. Toutefois,

avant de passer à la prochaine section, il convient de faire un dernier commentaire sur les méthodes pour les les petits échantillons.

Nous avons vu qu'à proprement parler, les propriétés nominales (en ce qui concerne les probabilités de couverture pour les intervalles de confiance et les déclarations de valeur p des tests d'hypothèse) des méthodes pour les petits échantillons reposent sur l'hypothèse que les distributions sous-jacentes sont exactement normales et, dans le cas des méthodes 5.2.3.5 et 5.2.3.6, que les variances sont exactement égales. D'autre part, lorsqu'on a utilisé ces méthodes, des vérifications plutôt rudimentaires des tracés de probabilité ont été utilisées à des fins de vérification (seulement) du caractère à peu près plausible des modèles. Selon la théorie statistique conventionnelle, les méthodes pour petits échantillons exposées ici présentent un degré de fiabilité considérable, sauf en cas d'écarts flagrants par rapport aux suppositions du modèle. Autrement dit, tant que les suppositions du modèle représentent à peu près la réalité, les niveaux de confiance nominaux et les valeurs p resteront raisonnablement corrects. (Par exemple, une méthode d'intervalle de confiance nominale de 90 % peut en réalité correspondre à un intervalle de confiance de 80 %, mais pas à un intervalle de confiance de 20 %.) Par conséquent, l'utilisation des graphiques que nous avons faite ici représente généralement une mesure de précaution adéquate contre l'application injustifiée des méthodes d'inférence pour les moyennes de petits échantillons.

5.2.4 Inférence pour les variances de deux échantillons

INFÉRENCE POUR LE RAPPORT DE DEUX VARIANCES (BASÉE SUR DES ÉCHANTILLONS INDÉPENDANTS SUIVANT UNE DISTRIBUTION NORMALE)

Pour passer d'une inférence pour variance unique à une inférence permettant de comparer deux variances, il faut introduire une nouvelle famille de distributions de probabilité : les distributions F de Fisher-Snedecor.

DÉFINITION 5.2.4.1 Distribution F de Fisher-Snedecor

EXPRESSION 5.2.4.1

La distribution F F de Fisher-Snedecor avec paramètres de degrés de liberté du numérateur et du dénominateur v_1 et v_2 désigne une distribution de probabilité continue ayant pour densité de probabilité

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{v_1+v_2}{2}\right)\left(\frac{v_1}{v_2}\right)^{v_1/2} x^{(v_1/2)-1}}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)\left(1+\frac{v_1 x}{v_2}\right)^{(v_1+v_2)/2}} & \text{pour } x > 0 \\ 0 & \text{sinon} \end{cases}$$

0 \\ 0 & \text { sinon }\end{cases} » title= »f(x)= \begin{cases}\frac{\Gamma\left(\frac{v_1+v_2}{2}\right)\left(\frac{v_1}{v_2}\right)^{v_1 / 2} x^{\left(v_1 / 2\right)-1}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right)\left(1+\frac{v_1 x}{v_2}\right)^{\left(v_1+v_2\right) / 2}} & \text { pour } x>0 \\ 0 & \text { sinon }\end{cases} » class= »latex mathjax »>

Si une variable aléatoire présente une densité de probabilité donnée par la formule (5.2.4.1), on dit qu'elle a une distribution F_{v_1,v_2}.

Comme le montre la Figure 5.2.4.1, les distributions F de Fisher-Snedecor sont des distributions à forte asymétrie droite, dont le maximum se situe à un argument légèrement inférieur à 1. Globalement, plus les valeurs v_1 et v_2 sont basses, plus la distribution F de Fisher-Snedecor concernée est asymétrique et étalée.

Utilisation des tables de distribution F (de Fisher-Snedecor), Tableau A1.5

Utiliser directement la formule (5.2.4.1) pour trouver des probabilités pour la distribution F nécessite d'employer des méthodes d'intégration numérique. Pour appliquer la distribution F en inférence statistique, on utilise plutôt soit un logiciel de statistiques, soit des tables très abrégées de quantiles de distribution F. Les tableaux A1.5 en annexe sont des tables de quantiles F. Ils présentent, pour une valeur de p donnée, les quantiles p de la distribution F pour différentes combinaisons de v_1 (le degré de liberté du numérateur) et v_2 (le degré de liberté du dénominateur). Les valeurs de v_1 sont données dans l'en-tête de la table, et les valeurs de v_2, dans la colonne de gauche.

Les tableaux A1.5 ne donnent que les quantiles p pour p supérieur à 0,5, mais les quantiles de distribution F pour p inférieur à 0,5 sont souvent également utiles. Plutôt que de créer des tables pour ces valeurs, la pratique

la plus courante est d'utiliser une astuce de calcul. En effet, le rapport entre les quantiles F_{v_1,v_2} et F_{v_2,v_1} permet de déterminer les quantiles pour de faibles p. Soit Q_{v_1,v_2} la fonction quantile F_{v_1,v_2} et Q_{v_2,v_1} la fonction quantile pour la distribution F_{v_2,v_1}, alors

EXPRESSION 5.2.4.2 Rapport entre quantiles F_{v_1,v_2} et F_{v_2,v_1}

$$Q_{v_1,v_2}(p) = \frac{1}{Q_{v_2,v_1}(1-p)}$$

L'encadré (5.2.4.2) indique qu'il est possible d'obtenir un point de pourcentage de distribution inférieur en prenant l'inverse d'un point de pourcentage de distribution supérieur correspondant, en inversant les degrés de liberté.

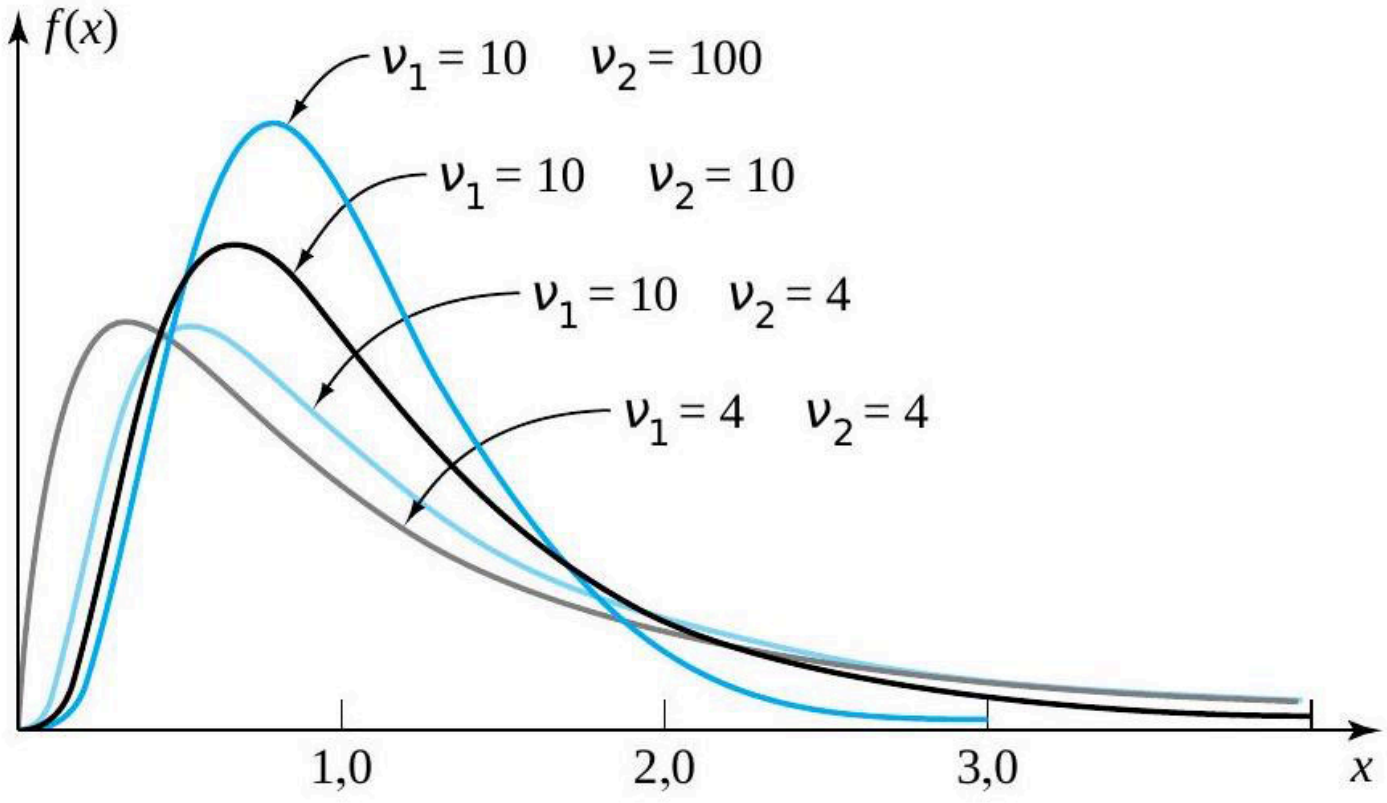


Figure 5.2.4.1 Quatre densités de probabilité F différentes

Exemple 5.2.4.1 Utilisation des tables de quantiles de distribution

Supposons que V soit une variable aléatoire $F_{3,5}$. Cherchons les quantiles .95 et .01 de la distribution de V, puis regardons ce que le tableau A1.5 révèle sur $P[V >$ » title= »P[V> » class= »latex mathjax »> 4,0] et <img

src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/109e3ebc4e9893167fd000040c2f0f7b.png » alt= »P[V<0,3] » title= »P[V.

En consultant directement la table des quantiles pour $p = .95$, colonne $v_1 = 3$, ligne $v_2 = 5$, on trouve dans un premier temps le nombre 5,41. Autrement dit, $Q(.95) = 5.41$, ce qui équivaut à dire que <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/e87e037cb2d9dc42f354bdabcc0770e6.png » alt= »P[V<5,41]=0,95" title= »P[V.

Pour trouver le quantile $p = \ 0{,}01$ de la distribution $F_{3,5}$, il faut utiliser l'expression (5.2.4.2), soit :

$$Q_{3,5}(.01) = \frac{1}{Q_{5,3}(.99)}$$

de sorte qu'en utilisant la colonne $v_1 = 5$ et la ligne $v_2 = 3$ de la table de quantiles $F.99$, on obtient :

$$Q_{3,5}(\ 0{,}01) = \frac{1}{28{,}24} = \ 0{,}04$$

En considérant ensuite que $P[V > 4{,}0]$4,0] » title= »P[V>4,0] » class= »latex mathjax »>, on constate (en utilisant la colonne [/latex]v_{1}=3[/latex] et la ligne $v_2 = 5$ du tableau A1.5) que 4,0 se situe entre les quantiles 0,90 et 0,95 de la distribution $F_{3,5}$. C'est à dire que

0,90<P[V≤4.0]<0,95

de sorte que

0,05<P[V>4,0]<0,10

Enfin, si l'on considère que <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/109e3ebc4e9893167fd000040c2f0f7b.png » alt= »P[V<0,3] » title= »P[V, notons qu'aucune des entrées des tableaux A1.5 n'est inférieure à 1,00. Ainsi, pour placer la valeur 3 dans la distribution $F_{3,5}$, il faut localiser sa réciproque, 3,33 ($= 1/.3$), dans la distribution $F_{5,3}$, puis utiliser l'expression (5.2.4.2). En utilisant les colonnes $v_1 = 5$ et les lignes $v_2 = 3$ des tableaux A1.5, on constate alors que 3,33 se situe entre les quantiles 0,75 et 0,90 de la distribution $F_{5,3}$. Donc d'après l'expression (5.2.4.2), 0,3 se situe entre les quantiles 0,1 et 0,25 de la distribution $F_{3,5}$, et

0,10<P[V<0,3]<0,25

Cet effort pour déterminer les faibles quantiles de distribution F est une conséquence des normes de création des tables, et non une particularité propres aux distributions F. Pour faciliter les choses et trouver les quantiles F plus simplement, il est notamment possible d'utiliser un logiciel de statistiques standard ou une calculatrice scientifique.

La distribution F est utilisée ici, car un fait de probabilité lie le comportement des rapports des variances d'échantillons indépendants (basés sur des échantillons suivant une distribution normale) aux variances σ_1^2 et σ_2^2 des distributions sous-jacentes. Autrement dit, lorsque s_1^2 et <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/4271702ef47127321ec8a369ddc787e9.png » alt= »s_^ Formula does not parse proviennent d'échantillons indépendants suivant une distribution normale, la variable

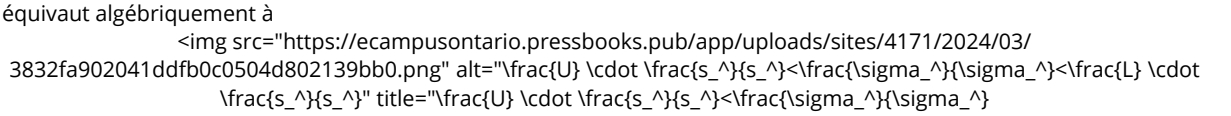
5.2.4.3 Formula does not parseF=\frac{s_^}{\sigma_^} \cdot \frac{\sigma_^}{s_^}" title="s_^ Formula does not parse proviennent d'échantillons indépendants suivant une distribution normale, la variable

> **5.2.4.3** Formula does not parseF=\frac{s_^}{\sigma_^} \cdot \frac{\sigma_^}{s_^}" class="latex mathjax">
>
> suit une distribution F_{n_1-1,n_2-1}. (s_1^2 a $n_1 - 1$ degrés de liberté associés et figure dans le numérateur de cette expression, tandis que s_2^2 a $n_2 - 1$ degrés de liberté associés et figure dans le dénominateur, motivant le langage introduit à la définition 5.2.4.1)
>
> C'est exactement dont nous avons besoin pour produire des méthodes d'inférence formelles pour le rapport σ_1^2/σ_2^2. Il est par exemple possible de choisir L et U, les bons quantiles F, de sorte que la probabilité que la variable (5.2.4.3) se situe entre

L et U corresponde au niveau de confiance souhaité. (L et U sont typiquement choisis de manière à « répartir le manque de confiance » entre les queues F_{n_1-1,n_2-1} supérieure et inférieure.) Mais

$$L<\frac{s_1^2}{\sigma_1^2}\cdot\frac{\sigma_2^2}{s_2^2}<U$$

équivaut algébriquement à



Autrement dit, lorsqu'un mécanisme de génération de données peut être considéré comme fondamentalement équivalent à un échantillonnage aléatoire indépendant suivant deux distributions normales, l'intervalle de confiance bilatéral pour σ_1^2/σ_2^2 a pour bornes

5.2.4.4 Limites de confiance de distribution normale pour σ_1^2/σ_2^2

$$\frac{s_1^2}{U\cdot s_2^2} \quad \text{et} \quad \frac{s_1^2}{L\cdot s_2^2}$$

où L et U (quantiles F_{n_1-1,n_2-1}) sont tels que la probabilité F_{n_1-1,n_2-1} assignée à l'intervalle (L, U) correspond au niveau de confiance souhaité.

De plus, il y a une méthode de test d'hypothèse évidente pour σ_1^2/σ_2^2. Sous réserve des limites de modélisation nécessaires pour que la méthode d'intervalle de confiance s'applique,

5.2.4.5

$$\mathrm{H}_0 : \frac{\sigma_1^2}{\sigma_2^2} = \#$$

peut être testée à l'aide de la statistique

5.2.4.6 Variable à tester suivant une distribution normale pour σ_1^2/σ_2^2

$$F = \frac{s_1^2/s_2^2}{\#}$$

et d'une distribution de référence F_{n_1-1,n_2-1}. (Le choix de $\# = 1$ dans les encadrés (5.2.4.5) et (5.2.4.6) correspond à une hypothèse nulle où les variances sont égales. C'est le seul choix communément utilisé en pratique.)

Valeurs P de test $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \#$

Les valeurs p pour les hypothèses alternatives unilatérales $\mathrm{H_a} : \sigma_1^2/\sigma_2^2 <$ # et $\mathrm{H_a} : \sigma_1^2/\sigma_2^2 > \#$\#" title="\mathrm{H}_{\mathrm{a}}: \sigma_1^2 / \sigma_2^2>\#" class="latex mathjax"> sont (respectivement) les queues de distribution F_{n_1-1,n_2-1} gauche et droite au-delà des valeurs observées de la variable à tester. Pour les hypothèses bilatérales alternatives $\mathrm{H_a} : \sigma_1^2/\sigma_2^2 \neq \#$, la convention standard est de reporter deux fois la probabilité F_{n_1-1,n_2-1} à droite de la fonction étudiée si $f > 1$1" title="f>1" class="latex mathjax">, et de reporter deux fois la probabilité F_{n_1-1,n_2-1} à gauche de la fonction f étudiée si f<1.

Exemple 5.2.4.2 Comparaison de l'uniformité des mesures d'indice de dureté de deux types d'acier

Condon, Smith et Woodford ont mené des essais de dureté sur des échantillons d'acier au carbone à 4%. Une partie de leurs données figurent dans le tableau 5.2.4.1, où sont représentées les mesures de dureté de Rockwell pour dix échantillons provenant d'un lot d'acier à traitement thermique et cinq échantillons provenant d'un lot d'acier laminé à froid.

Comparons l'uniformité des mesures de dureté pour ces deux types d'acier (plutôt que la dureté moyenne, comme on l'a fait au chapitre 5.2.3). La figure 5.2.4.2 présente les diagrammes de dispersion de chaque échantillon et suggère que la variabilité associée

aux échantillons d'acier à traitement thermique est plus élevée que celle associée aux échantillons d'acier laminé à froid. Les deux tracés normaux de la Figure 5.2.4.3 ne démontrent aucun problème clair avec l'hypothèse que les variables suivent une distribution normale.

Traitement thermique	Laminage à froid
32,8, 44,9, 34,4, 37,0, 23,6, 29,1, 39,5, 30,1, 29,2, 19,2	21,0, 24,5, 19,9, 14,8, 18,8

Tableau 5.2.4.1 Mesures de dureté de Rockwell pour échantillons de deux types d'acier

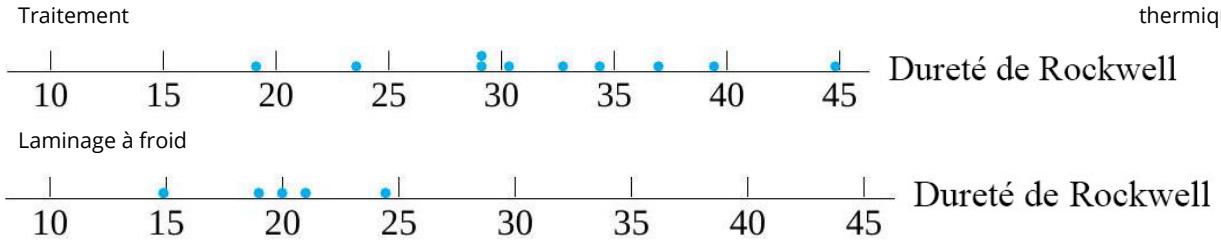


Figure 5.2.4.2 Diagrammes de dispersion présentant la dureté de l'acier à traitement thermique et de l'acier laminé à froid

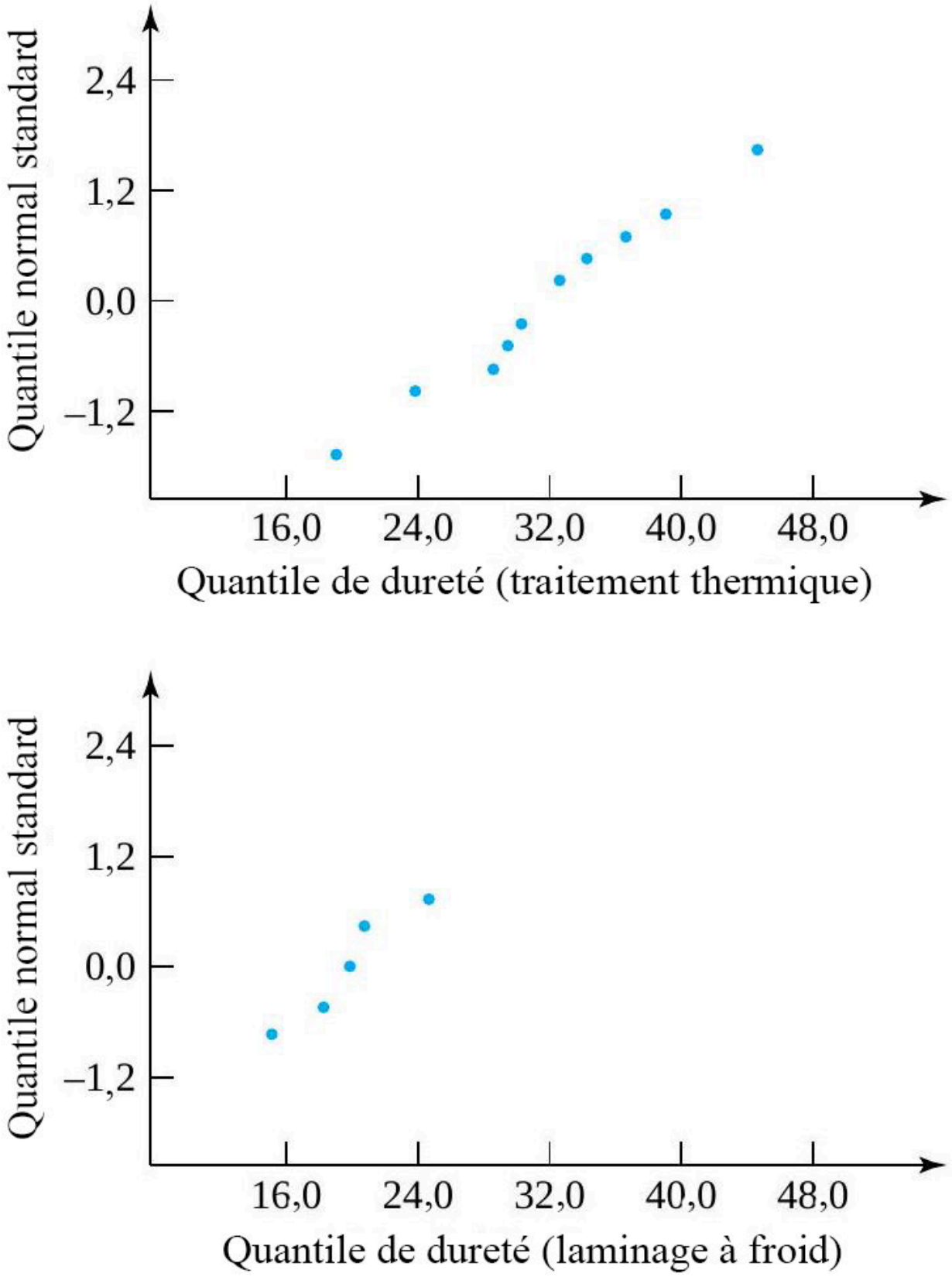


Figure 5.2..4.3 Tracés de dureté normaux pour acier à traitement thermique et acier laminé à froid

Choisissons arbitrairement ensuite d'appeler condition numéro 1 le traitement thermique et condition numéro 2 le laminage à froid, on a$s_1 = 7.52$ et $s_2 = 3{,}52$; un test d'hypothèse d'égalité des variances en cinq étapes reposant sur la variable (5.2.4.6) se présente comme suit :

1. $]\mathrm{H}_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1$
2. $\mathrm{H}_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$

(Si une raison matérielle justifie le choix d'une autre hypothèse unilatérale, les auteurs n'en ont pas connaissance.)

3. La variable à tester est la suivante :

$$F = \frac{s_1^2}{s_2^2}$$

.

La distribution de référence est la distribution $F_{9,4}$, et H_0 sera infirmée si la valeur observée de f est soit grande, soit petite.

.

4. Les échantillons donnent

.

$$f = \frac{(7{,}52)^2}{(3{,}52)^2} = 4{,}6$$

.

5. Comme f est supérieure à 1, pour l'alternative bilatérale, la valeur p est la suivante :

.

$$2P\left[\text{ une variable aléatoire } F_{9,4} \geq 4{,}6\right]$$

.

D'après les tableaux A1.5, 4,6 se situe entre les quantiles 0,9 et 0,95 de la distribution $F_{9,4}$, donc le niveau de signification étudié se situe entre 0,1 et 0,2. Il semble donc peu probable (bien que pas impossible) que le traitement thermique et le laminage à froid aient des variabilités égales.

.

Afin de préciser la taille relative des variabilités, les racines carrées des valeurs dans l'expression 5.2.4.6 peuvent être utilisées pour obtenir un intervalle de confiance bilatéral de 90 % pour le rapport σ_1 / σ_2. Comme le quantile 0,95 de la distribution $F_{9,4}$ vaut 6,0 et que le quantile 0,95 de la distribution $F_{4,9}$ vaut 3,63, le quantile .05 de la distribution $F_{9,4}$ est $\dfrac{1}{3{,}63}$. Ainsi, l'intervalle de confiance de 90 % pour le rapport d'écarts-types σ_1 / σ_2 a pour bornes

.

$$\sqrt{\frac{(7{,}52)^2}{6{,}0(3{,}52)^2}} \quad \text{et} \quad \sqrt{\frac{(7{,}52)^2}{(1/3{,}63)(3{,}52)^2}}$$

.

En faisant le calcul, on obtient :

.

0,87 et 4,07

.

Le fait que l'intervalle (0,87, 4,07) couvre des valeurs à la fois inférieures et supérieures à 1 indique que les données ne permettent pas de déterminer précisément laquelle des deux variabilités est la plus élevée.

L'une des plus importantes applications en ingénierie des méthodes d'inférence représentées par ces expressions, c'est de pouvoir comparer la précision intrinsèque de diverses pièces d'équipement ou de diverses méthodes de fonctionnement d'une pièce d'équipement donnée.

Exemple 4.2.4.3 Comparaison de l'uniformité de fonctionnement de deux massicots

Abassi, Afinson, Shezad et Yeo ont travaillé avec une entreprise de découpe de feuilles de papier à partir de rouleaux. L'uniformité

de la longueur des feuilles est importante, car meilleure elle est, plus la longueur moyenne des feuilles peut se rapprocher de la valeur nominale sans générer de sous-dimensionnement, réduisant ainsi le gaspillage pour l'entreprise.
Les étudiant.e.s ont comparé l'uniformité des feuilles coupées à l'aide de deux massicots, l'un étant muni d'un frein manuel, et l'autre d'un frein automatique. Cette comparaison se base sur les écarts-types estimés de longueur de feuille coupée par les deux machines – précisément le genre de données qu'on utilise pour
poser le cadre des inférences formelles de cette section. Les étudiant.e.s ont estimé que $\sigma_{manuel}/\sigma_{automatique}$ était de l'ordre de 1,5. Selon leurs calculs, il faudrait tout au plus deux ans pour que l'entreprise récupère les coûts requis pour équiper tous les massicots de freins automatiques.

AVERTISSEMENT CONCERNANT LES INFÉRENCES POUR VARIANCE

Si l'on veut faire preuve de rigueur, les méthodes de cette section ne s'appliquent qu'aux distributions normales. Il est pertinent de se demander à quel point cette condition est essentielle pour qu'on puisse recourir à ces méthodes d'inférence pour une ou deux variances. À la fin du module 5.2.3, nous avions mentionné que les méthodes pour les moyennes tiennent relativement la route en cas de violation modérée des hypothèses du module. Malheureusement, ce n'est pas le cas des méthodes pour les variances présentées ici.

Pour ces méthodes, si les données ne suivent pas une distribution normale, le niveau de confiance nominal et les valeurs p peuvent facilement induire en erreur. Par conséquent, il est essentiel d'examiner minutieusement les données (comme on l'a fait avec les tracés normaux dans les exemples) pour justifier l'utilisation des méthodes de la présente section. Les tracés normaux n'étant typiquement pas très révélateurs à moins que l'échantillon concerné ne soit de taille moyenne ou grande, les inférences formelles pour variances seront plus fiables si les échantillons (d'apparence normale) sont de taille moyenne ou grande.

L'importance de la distribution normale en matière de fonctionnement prévisible des méthodes de cette section n'est pas la seule raison de préférer des échantillons de grande taille pour les inférences pour variances. L'expérience révèle rapidement que même s'ils suivent une loi normale, les petits échantillons sont souvent inadéquats pour répondre aux questions pratiques sur les variances. Les intervalles de confiance F pour les variances et le rapport des variances de petits échantillons peuvent être tellement grands qu'ils n'ont guère d'utilité pratique. De fait, il faut généralement utiliser de grands échantillons pour résoudre les problèmes de variances dans le monde réel. Cet avis n'est pas une admission de défaut de ces méthodes; il s'agit simplement d'une mise en garde et d'une explication sur le fait que pour être véritablement parlantes, les variances nécessitent plus de données que les moyennes (par exemple).

5.2.5 Inférence pour moyenne de différences appariées

Les méthodes d'estimation de l'intervalle de confiance et de test d'hypothèse trouvent une application importante dans les données appariées. En ingénierie, il est courant de prendre deux mesures similaires sur le même échantillon d'un objet physique, mais à une heure différente ou à un autre endroit. L'objectif dans ce cas est souvent voir s'il y a un écart constant entre les deux mesures.

Exemple 5.2.5.1 Comparaison des mesures des bords avant et arrière sur un produit façonné en bois

Drake, Hones et Mulholland ont travaillé avec une entreprise sur le contrôle du fonctionnement d'une fraise à détourer en bout dans une usine fabriquant des produits en bois. Ils ont mesuré une dimension critique d'un certain nombre de pièces d'un type donné lorsqu'elles sortaient de la machine. Les mesures des bords avant et arrière ont été prises sur chaque pièce. La conception de la pièce en question prévoit que les bords avant et arrière doivent tous deux présenter une valeur cible de 0,172 po. Le tableau 5.2.5.1 indique les mesures de bord avant et de bord arrière prises sur cinq pièces consécutives.

Dans cette situation, le fait que les dimensions des bords avant et arrière correspondent était au moins aussi essentiel à l'assemblage que le fait chaque dimension soit conforme à la valeur nominale de 0,172 po. Il s'agissait donc d'une situation d'appariement de données, dans laquelle l'une des préoccupations était la possibilité d'un écart constant entre les dimensions de bords avant et arrière. (Cet écart pouvait être dû à un mauvais réglage machine ou à une mauvaise utilisation de la machine.)

Pièce	Mesure du bord avant (po)	Mesure du bord arrière (po)
1	0,168	0,169
2	0,170	0,168
3	0,165	0,168
4	0,165	0,168
5	0,170	0,169

Tableau 5.2.5.1 Dimensions des bords avant et arrière sur cinq pièces usinées

Dans des situations comme celle de l'exemple 5.2.5.1, une méthode simple pour rechercher un potentiel écart constant entre données appariées consiste à d'abord réduire les deux mesures sur chaque objet physique à leur différence. Les méthodes d'estimation d'intervalle de confiance et de test d'hypothèse étudiées peuvent ensuite être appliquées aux différences. Par conséquent, après avoir réduit les données appariées aux différences d1, d2, ...,d_n, si n (le nombre de données appariées) est élevé, les bornes de l'intervalle de confiance pour la différence moyenne sous-jacente μ_d sont

5.2.5.1 Limites de confiance pour μ_d, pour un grand échantillon

$$\bar{d} \pm z\frac{s_d}{\sqrt{n}}$$

où s_d est l'écart-type de l'échantillon d1, d2, ...,d_n. De même, l'hypothèse nulle

5.2.5.2 $$\mathrm{H}_0 : \mu_d = \#$$

peut être testée à l'aide de la variable

5.2.5.3 Variable à tester pour μ_d, grand échantillon

$$Z = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}}$$

et d'une distribution normale standard de référence.

Si n est petit, pour trouver des méthodes d'inférence formelle, il doit être plausible que les différences suivent une distribution normale. Si tel est le cas, l'intervalle de confiance pour μ_d a pour bornes

5.2.5.4 Limites de confiance pour μ_d suivant une distribution normale

$$\bar{d} \pm t\frac{s_d}{\sqrt{n}}$$

et l'hypothèse nulle (5.2.5.2) peut être testée à l'aide de la variable

5.2.5.5 Variable à tester pour μ_d suivant une distribution normale

$$T = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}}$$

et d'une distribution de référence t_{n-1}.

Suite de l'exemple 5.2.5.2

Pour illustrer cette méthode des différences appariées, testons l'hypothèse nulle H_0 : μ_d = 0 avec un intervalle de confiance de 95 %, pour tout écart constant entre les dimensions de bords avant et arrière, μ_d, à l'aide des données du tableau 5.2.5.1

Commençons par réduire les n = 5 observations d'appariement du tableau 5.2.5.1 aux différences

d = dimension du bord avant − dimension du bord arrière

figurant dans le tableau 5.2.5.2. La figure 5.2.5.1 représente un tracé normal des n = 5 différences du tableau 5.2.5.2. Un peu d'expérimentation avec les tracés normaux des échantillons simulés de taille n = 5 suivant une distribution normale suffit à mettre en évidence le fait que le manque de linéarité sur la figure 5.2.5.1 n'est en aucun cas atypique des données normales. Si l'on ajoute à cela le fait que les distributions normales décrivent souvent très bien les dimensions usinées de pièces produites en série, on pourrait en conclure que les méthodes représentées par les expressions 5.2.5.4 et 5.2.5.5 sont de mise dans cet exemple.

Les différences du tableau 5.2.5.2 indiquent $\hat{d}$ =−0,0008 po et d_d = 0,0023 po. Effectuons un test d'hypothèse en cinq étapes pour voir s'il est plausible d'affirmer que l'écart est constant :

1. H0 : μd = 0.
2. Ha : μd = 0.

(Il n'y a *a priori* aucune raison d'opter pour une hypothèse alternative unilatérale.)

3. La variable à tester est la suivante :

$$T = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

La distribution de référence sera la distribution t avec v = n − 1 = 4 degrés de liberté. Une grande valeur de |t| permettra d'infirmer H_0 et de confirmer H_a.

4. L'échantillon donne

$$t = \frac{-.0008}{\frac{.0023}{\sqrt{5}}} = -.78$$

5. Le niveau de signification observé est P [|une variable aléatoire t_4 | ≥0,78], ce qui, d'après le tableau A1.2, est supérieur à 2(0,10) = 0,2. Les données ne se montrent pas en faveur d'une différence systématique entre les mesures des bords avant et arrière.

D'après le tableau A1.2 pour le quantile 0,975 de la distribution t_4, le multiplicateur à utiliser dans l'expression pour un niveau de confiance à 95 % est t = 2,776. Cela signifie qu'un intervalle de confiance bilatéral à 95 % pour la différence moyenne entre les dimensions des bords avant et arrière a pour bornes

$$-0{,}0008 \pm 2{,}776\frac{0{,}0023}{\sqrt{5}}$$

soit :

−0,0008 po ± 0,0029 po

ou encore :

−0,0037 po et 0,0021 po

Cet intervalle de confiance pour μ_d sous-entend (étant donné que 0 fait partie de l'intervalle calculé) que le niveau de confiance observé dans le cadre du test de H_0 : μ_d = 0 est supérieur à 0,05 (= 1 − 0,95). En d'autres termes, le calcul de l'IC ci-dessus montre bien que l'imprécision indiquée par le signe plus ou moins de l'expression est suffisamment grande pour concevoir que la différence perçue, $\bar{d}$ =−0,0008, n'est que le résultat de la variabilité d'échantillonnage.

Pièce	d = différence de dimensions (po)	
1	− 0,001	(= 0,168 − 0,169)
2	0,002	(= 0,170 − 0,168)
3	− 0,003	(= 0,165 − 0,168)
4	− 0,003	(= 0,165 − 0,168)
5	0,001	(= 0,170 − 0,169)

Tableau 5.2.5.2 Cinq différences dans les mesures de bords avant et arrière

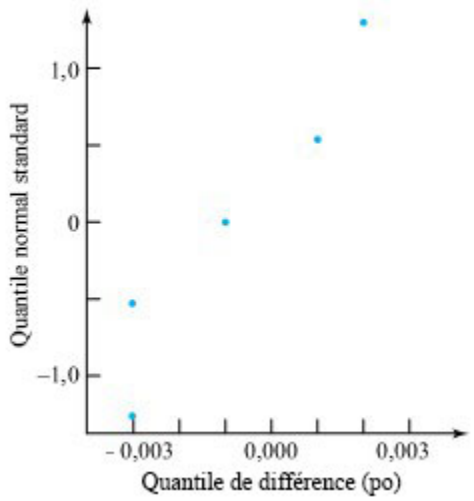


Figure 5.2.5.1 Tracé normal de n = 5 différences

Inférence d'un grand échantillon pour μ_d

L'Exemple 5.2.5.2 traite un problème sur un petit échantillon. Nous ne présentons pas d'exemple pour n élevé. Un tel exemple ne consisterait qu'à refaire ce qui a déjà été vu. En effet, étant donné que pour un n élevé, la distribution t avec v = n − 1 degré de liberté devient essentiellement une distribution normale standard, on pourrait même suivre l'exemple 5.2.5.2 pour n élevé et n'avoir aucun problème de logique. Ce n'est donc plus nécessaire de s'attarder à la méthode des différences appariées.

Données appariées ou non appariées

Le problème des données appariées (qui contiennent des observations à deux variables sur un seul échantillon) marque un contraste avec le problème précédent où les méthodes s'appliquent à une mesure unique faite sur chaque élément de deux échantillons différents. Dans le cas de l'usinage du bois de l'exemple 5.2.5.2, les données sont appariées puisque les mesures des bords avant et arrière ont été effectuées sur chaque pièce. Si les mesures du bord avant étaient prises sur un groupe de pièces et les mesures du bord arrière sur un autre, il faudrait alors faire une une analyse de deux échantillons (et non une analyse de différences appariées).

5.2.6 Tutoriel 4A - Statistiques inférentielles et tests t

À ce stade, il est recommandé de travailler sur l'exercice du Tutoriel 4A qui se trouve sur le référentiel GitHub associé. Cet exercice vous apprendra comment mener des tests t en utilisant le langage Python.

Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur la mise à l'essai d'hypothèses. Vous pouvez les trouver dans la section « How do I do X in Python? ». Les fichiers « T-tests » et « P-Values » vous seront particulièrement utiles. En outre, le fichier « Confidence Intervals – Difference of Means » vous sera utile si vous cherchez à calculer les intervalles lorsque vous comparez des groupes multiples. Si vous voulez calculer la taille de l'échantillon ou si vous devez effectuer des calculs de puissance, le fichier « Sample Size & Power Calculations » vous servira également.

5.3.0 Introduction aux modèles non paramétriques

L'objectif de ce module est d'aborder le principe de statistiques non paramétriques. Les statistiques non paramétriques sont des types de variables à tester, auxquelles des formules sont associées, qui peuvent servir à estimer les associations entre deux variables ou plus, sans baser ces associations sur les changements de moyenne. La moyenne arithmétique peut être fortement influencée par des valeurs extrêmes ou une dispersion anormale. Si une collection de données ne semble pas suivre une *distribution normale*, ou si on peut être raisonnablement sûr que la distribution réelle de valeurs variables dans une population n'est *pas* normale, les statistiques non paramétriques peuvent servir à mieux estimer les associations entre variables.

5.3.1 Méthodes non paramétriques

MÉTHODES NON PARAMÉTRIQUES

Que faire lorsque les hypothèses vues dans les précédentes leçons (tests t, corrélation, etc.) ne sont pas fondées? Il existe des tests qui s'utilisent lorsqu'un certain nombre d'hypothèses, requises par les tests usuels comme les tests t et les corrélations, ne sont pas fondées (p. ex., en cas de distribution non normale ou d'échantillons de petite taille). Ces tests – appelés tests non paramétriques – utilisent le même type de comparaisons, mais sur la base d'hypothèses différentes.

HYPOTHÈSES PARAMÉTRIQUES

La statistique paramétrique est une branche des statistiques qui suppose que les données échantillonnées proviennent d'une population qui suit les paramètres et hypothèses qui s'appliquent dans la majorité, voire la totalité, des cas. La plupart des méthodes statistiques élémentaires connues sont paramétriques; nous en avons évoqué bon nombre d'entre elles, et l'article Wikipédia Parametric Statistics traite aussi du sujet.

HYPOTHÈSES PARAMÉTRIQUES ET DISTRIBUTION NORMALE

La distribution normale est une hypothèse courante pour de nombreux tests, notamment le test t, l'ANOVA et la régression. Rappelez-vous les tests paramétriques que nous avons évoqués ici respectaient les hypothèses de distribution normale suivantes : asymétrie et aplatissement des variables nuls ou faibles, et indépendance des termes d'erreur des variables.

Ces hypothèses permettent de déduire que la population suit une distribution normale.

MÉTHODES NON PARAMÉTRIQUES

Les méthodes statistiques ne nécessitant pas de faire des hypothèses de distribution quant aux données sont appelées méthodes non paramétriques. Le terme « non paramétrique » ne s'applique pas aux données, mais bien aux méthodes utilisées pour analyser ces dernières. Ces tests utilisent des rangs pour analyser les différences. Les méthodes non paramétriques peuvent être utilisées pour différents types de comparaisons ou de modèles.

HYPOTHÈSES NON PARAMÉTRIQUES

1. Les tests non paramétriques reposent sur des hypothèses concernant l'échantillonnage (en particulier, sur son caractère généralement aléatoire).
2. En fonction du test non paramétrique utilisé, il y a des hypothèses sur la dépendance ou l'indépendance des échantillons, mais il n'y a pas d'hypothèse sur la distribution des scores dans la population.

TESTS NON PARAMÉTRIQUES ET NIVEAU DE MESURE

Les variables qui suivent des niveaux de mesure catégoriques peuvent nécessiter des tests non paramétriques.

Pensons à l'autonomie, la compétence et le revenu : de telles variables suivraient-elles toujours une distribution

normale? Pour le revenu, par exemple, on peut s'attendre à ce que la distribution soit asymétrique, étant donné qu'il y a une faible minorité de gens qui ont un revenu extrêmement élevé.

MOYENNE ET MÉDIANE

Lorsqu'une distribution est fortement asymétrique, la moyenne est affectée par le grand nombre de valeurs (relativement) aberrantes. Par exemple, lorsque l'on mesure quelque chose comme le revenu, où les personnes au salaire très élevé sont rares, mais celles aux revenus moyens et faibles très nombreuses, le « milieu » de la distribution est considérablement « décentré ». Dans ces cas, il est plus pertinent d'utiliser la médiane (la valeur « centrale » – celle qui sépare la population en deux parts égales).

TAILLE DE L'ÉCHANTILLON

La taille de l'échantillon est un autre élément à prendre en compte pour choisir entre les tests paramétriques et les tests non paramétriques. Souvent, les scientifiques souhaitent utiliser un certain type de test paramétrique, mais leur échantillon est trop petit. Et souvent, dans ce genre de cas, on ne peut pas effectuer les tests de normalité en raison de la faiblesse de l'échantillon, qui ne donne pas de résultat interprétable. Si en plus les données ne suivent pas une distribution normale, on peut décider d'utiliser des tests non paramétriques.

VALEURS ABERRANTES

Comme il l'a été dit dans les chapitres précédents, les tests paramétriques reposent sur la continuité des données de la variable dépendante. Ces données doivent suivre une distribution normale et ne pas présenter de fausses valeurs aberrantes. Cependant, quelques tests non paramétriques peuvent fonctionner sur des données ordinales (classées) pour la variable dépendante. Ces tests pourraient aussi ne pas être affectés par les données aberrantes ou qui ne suivent pas une distribution normale. Chaque test paramétrique possède ses propres critères; il est donc conseillé de vérifier les hypothèses pour chaque test.

5.3.2 Choix du test d'hypothèse approprié

CHOIX DES TESTS D'HYPOTHÈSE APPROPRIÉS

DES CONSIDÉRATIONS MULTIPLES

Pour décider s'il faut utiliser des statistiques non paramétrique, on doit déterminer si la meilleure représentation du centre de la distribution des données est la moyenne ou la médiane. S'il se trouve que c'est la médiane, les tests non paramétriques sont *a priori* les plus adéquats, même avec un grand échantillon. Si l'échantillon est petit, les statistiques non paramétriques peuvent convenir dans un cas comme dans l'autre.

DIFFÉRENTS TESTS

Tous les tests paramétriques de différence que nous avons évoqués plus tôt possèdent un équivalent non paramétrique, qu'on peut utiliser lorsque les données ne suivent pas une distribution normale ou que l'échantillon est petit.

5.3.3 Comparaison de deux conditions indépendantes : le test U de Mann-Whitney

Le test U de Mann-Whitney est celui qui se prête le mieux à une analyse des différences entre deux groupes. Ce test permet d'analyser les différences entre les médianes, de même que l'ampleur de ces différences. Exemple : Y a-t-il une différence de nombre médian d'ami.e.s Facebook entre les hommes et les femmes? Par contre, si on voulait comparer deux conditions liées, le test à utiliser serait celui des rangs signés de Wilcoxon.

Ranks

	Gender	N	Mean Rank
FacebookFriends	Male	82	159.46
	Female	285	191.06
	Total	367	

Test Statistics

	FacebookFriends
Chi-Square	5.65
df	1
Asymp. Sig.	.017

INTERPRÉTATION DU TEST U DE MANN-WHITNEY

La différence est statistiquement significative – voir la valeur p, en bleu. Il faut également signaler la valeur du chi² et les degrés de liberté. Les rangs médians indiquent que les femmes ont plus d'ami.e.s Facebook que les hommes.

CONCLUSION

Les résultats du test U de Mann-Whitney indiquent que les femmes rapportent plus d'ami.e.s Facebook (médiane = 191,06) que les hommes (médiane = 159,46), et ce résultat est statistiquement significatif (U = 5,65, p = 0,017).

5.3.4 Test de Wilcoxon pour échantillons appariés

Le test des rangs signés de Wilcoxon est idéal pour analyser les différences internes d'un groupe. Ce test permet d'analyser les différences de cotes, de même que l'ampleur de ces différences.

Exemple : niveau de soutien social perçu rapportés par un groupe d'Australiens avant et après avoir participé à un programme de renforcement des compétences sociales.

Rangs

		N	Rang moyen	Somme des rangs
Soutien (avant) - Soutien (après)	Rangs négatifs	259	184,30	47732,50
	Rangs positifs	68	86,70	5895,50
	Égalités	40		
	Total	367		

Statistiques du test

	Soutien (avant) - Soutien (après)
Z	-12,24
Sig. asympt. (bilatérale)	0,000

INTERPRÉTATION DU TEST DE WILCOXON

En utilisant le même exemple que pour le module de test t, comparons le niveau de soutien social perçu rapportés par un groupe d'Australiens avant et après avoir participé à un programme de renforcement des compétences sociales. En rouge la cote Z; en vert la valeur p. On voit qu'il y a une différence entre les médianes avant et après le test. On peut déduire de la cote Z négative et du nombre de rangs positifs au temps 2 que les cotes se sont améliorées entre les deux coups de sonde.

CONCLUSION

Exemple de conclusion : Le test des rangs signés de Wilcoxon a permis d'indiquer que les rangs médians en matière de soutien social étaient statistiquement plus élevés après le test qu'avant (Z = -12,24, p <0,001).

5.3.5 Différences entre plusieurs groupes indépendants : le test de Kruskal-Wallis

TEST H DE KRUSKAL-WALLIS POUR TROIS ÉCHANTILLONS INDÉPENDANTS OU PLUS

Pour examiner les différences entre trois groupes ou plus, le test idéal est le test H de Kruskal-Wallis. Ce test permet d'analyser les différences entre les médianes, de même que l'ampleur de ces différences. Il analyse le principal effet sur la variable, à l'image d'une analyse de variance. Exemple : Y a-t-il une différence entre les niveaux de détresse psychologique médians des personnes qui travaillent à temps plein, à temps partiel ou occasionnellement? Si l'on voulait comparer les différences entre plusieurs groupes liés, le test à utiliser serait l'ANOVA de Friedman.

Ranks

	Are you employed?	N	Mean Rank
MentalDistress	Full-time	161	157.01
	Part-time	83	185.11
	Casual	123	218.59
	Total	367	

Test Statistics

	MentalDistress
Chi-Square	23.53
df	2
Asymp. Sig.	.000

INTERPRÉTATION DU TEST H DE KRUSKAL-WALLIS

Comme le montre la valeur p (en vert), la différence est statistiquement significative. Dans ce test, il faut aussi fournir la valeur du chi² ainsi que le nombre de degrés de liberté. Les rangs médians indiquent que c'est le personnel occasionnel qui présente les scores de détresse psychologique les plus élevés. Il est important de noter que des tests de suivi sont indispensables pour les différences entre groupes individuels (comme les tests U de Mann-Whitney), à l'image des tests post-hoc pour ANOVA.

CONCLUSION

Le test H de Kruskal-Wallis a démontré qu'il y avait une différence statistique importante entre les niveaux de détresse psychologique, $\chi 2(2) = 23{,}53$, $p < 0{,}001$, pour le personnel à temps plein (médiane = 157,01), pour le personnel à temps partiel (médiane = 185,11) et pour le personnel occasionnel (médiane = 218,58).

5.3.6 Tutoriel 4 – Tests non paramétriques

À ce stade, il est recommandé de faire l'exercice du Tutoriel 4 qui se trouve sur le référentiel GitHub. Cet exercice vous apprendra comment effectuer un test non paramétrique en Python.

Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur les tests d'hypothèses. On les retrouve dans le chapitre « How do I do X in Python? ». Les fichiers « Non-Parametric Tests » vous seront particulièrement utiles.

6.0.1 Introduction au modèle normal à un facteur

Les études d'ingénierie statistique produisent souvent des échantillons prélevés non pas dans un ou deux ensembles de conditions, mais dans des ensembles de conditions nombreux et variés. Par conséquent, bien que les méthodes d'inférence abordées dans la partie 5 soient un bon début, elles ne constituent pas pour autant une « boîte à outils statistique » complète permettant de résoudre tous les problèmes d'ingénierie; des méthodes d'inférence formelle adaptées aux études multi-échantillons sont également nécessaires.

.

Cette section amorce une présentation de telles méthodes. D'abord, elle rappelle l'utilité de certains des outils graphiques simples présentés dans la partie 2 pour effectuer des comparaisons informelles dans les études multi-échantillons. Elle introduit ensuite le modèle de « distributions normales de même variance ». Puis, elle explique le rôle des résidus dans l'évaluation de la pertinence de ce modèle dans une application, en insistant sur leur importance. Elle présente ensuite l'idée de combiner plusieurs variances d'échantillons pour produire une estimation unique et globale de la variance de référence. Enfin, elle ouvre une discussion sur la manière dont on peut utiliser les résidus standardisés lorsque les tailles d'échantillons varient considérablement.

6.0.2 Sources de la partie 6

Cette première version de la partie 6 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte, et on trouve des liens pour y accéder tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn et sont partagées sous licence CC BY-SA 4.0.

6.1.1 Comparaison graphique de plusieurs échantillons de données de mesure

Toute analyse réfléchie de plusieurs échantillons de données de mesures d'ingénierie doit commencer par la mise en graphique de ces données. Quand les échantillons sont de petite taille, le plus simple, c'est de juxtaposer des diagrammes de dispersion. Quand les échantillons sont de taille moyenne à grande (disons au moins six points de données par échantillon, à peu près), il vaut mieux juxtaposer des diagrammes en boîte.

Exemple 1 Comparaison de la résistance à la compression de huit formules de béton différentes

Armstrong, Babb et Campen ont mené des tests de résistance à la compression sur 16 formules de béton différentes. Une partie de leurs données figurent dans le Tableau 7.1, où huit formules différentes sont représentées. (Les seules différences entre les formules 1 à 8 sont leur rapport eau/ciment. La formule 1 présentait le plus faible rapport eau/ciment, rapport augmentant avec les numéros de formule, en suivant la progression 0,40, 0,44, 0,49, 0,53, 0,58, 0,62, 0,66, 0,71. Évidemment, connaître ces rapports eau/ciment sous-entend qu'une analyse d'ajustement de courbe sur la base de ces données pourrait se révéler utile, mais laissons cette possibilité de côté pour le moment.)

.

Placer les diagrammes de dispersion côte à côte pour ces huit échantillons de tailles $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = n_7 = n_8 = 3$ revient à établir un diagramme de dispersion de la résistance à la compression en fonction du numéro de la formule. Ce diagramme est présenté à la figure 6.1.1.1. Ce qui ressort globalement de la figure 6.1.1.1, c'est que les moyennes de résistance à la compression sont nettement différentes d'une formule à l'autre, mais que leurs variabilités sont à peu près comparables.

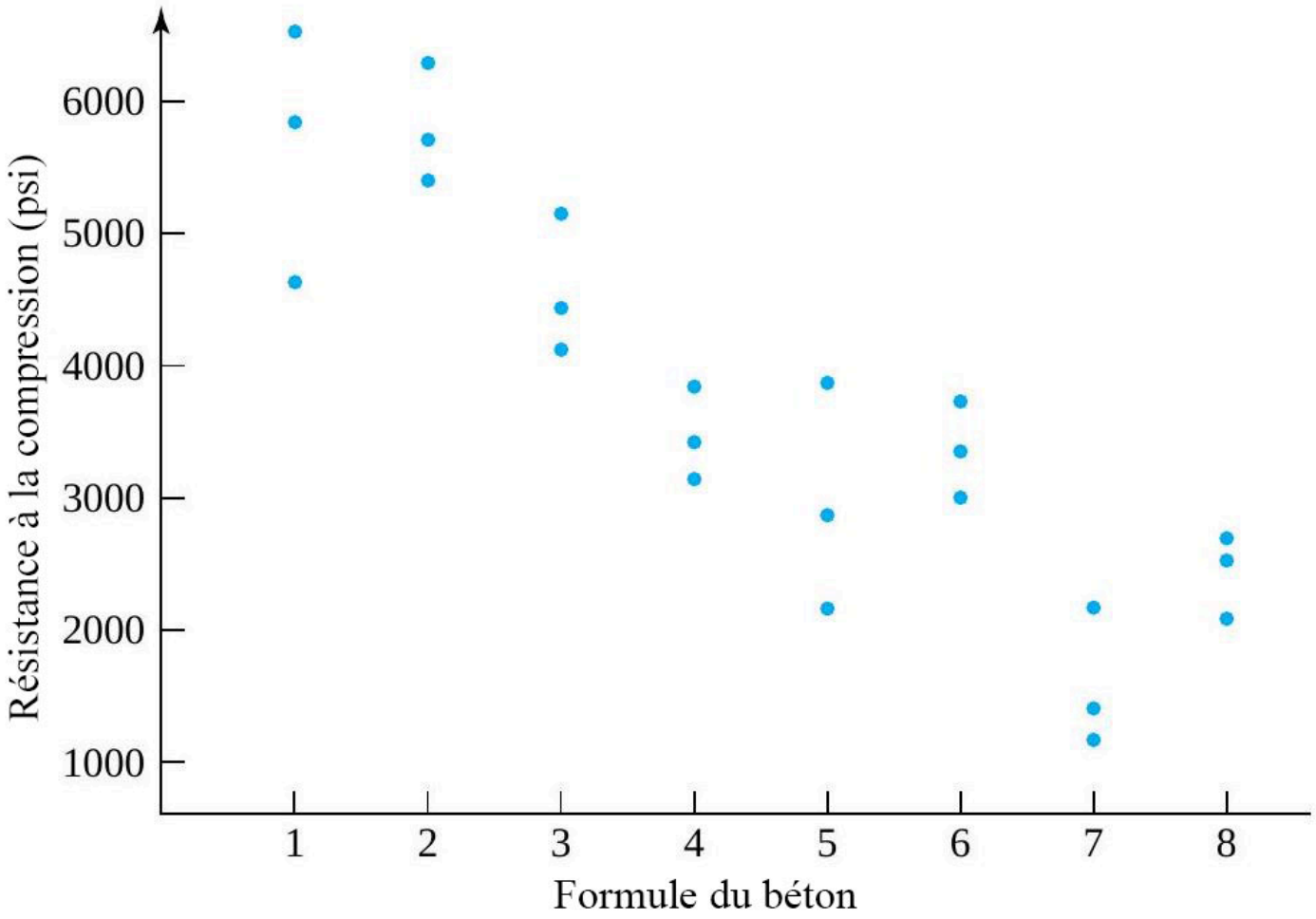


Figure 6.1.1.1 Diagrammes de dispersion côte à côte pour huit échantillons de résistance à la compression

Résistances à la compression de 24 échantillons de béton

Échantillon	Formule du béton	Résistance à la compression pendant 28 jours (psi)
1	1	5 800
2	1	4 598
3	1	6 508
4	2	5 659
5	2	6 225
6	2	5 376
7	3	5 093
8	3	4 386
9	3	4 103
10	4	3 395
11	4	3 820
12	4	3 112
13	5	3 820
14	5	2 829
15	5	2 122
16	6	2 971
17	6	3 678
18	6	3 325
19	7	2 122
20	7	1 372
21	7	1 160
22	8	2 051
23	8	2 631
24	8	2 490

Table 6.1.1.1 Résistances à la compression de 24 échantillons de béton

Exemple 6.1.1.2 Comparaison des constantes de ressort empiriques pour trois types de ressorts différents

Hunwardsen, Springer et Wattonville ont mené des tests sur trois types de ressorts en acier différents. Ils ont déterminé de manière expérimentale les constantes de ressort pour $n_1 = 7$ ressorts de type 1 (conception 4 po avec constante de ressort théorique de 1,86), $n_2 = 6$ ressorts de type 2 (conception 6 po avec constante de ressort théorique de 2,63) et $n_3 = 6$ ressorts de type 3 (conception 4 po avec constante de ressort théorique de 2,12), en utilisant une charge de $8{,}8\ \text{lb}$. Les valeurs expérimentales figurent dans le tableau 6.1.1.2.

Ces échantillons sont tout juste assez grands pour créer des diagrammes en boîte révélateurs. La figure 6.6.1.2 propose une représentation de ces données sous forme de diagrammes en boîte placés côte à côte. Ce qui ressort surtout de la figure 6.6.1.2, c'est que les constantes empiriques diffèrent considérablement, entre les ressorts de 6 po et les deux types de ressorts de $4\ \text{po}$, mais qu'aucune différence entre les deux types de ressorts 4 po n'est évidente. Évidemment, les informations du tableau 6.1.1.2 peuvent également être présentées sous forme de diagramme de dispersion côte à côte, comme dans la figure 6.1.1.3.

Constante empirique de ressort		
Ressorts de type 1	Ressorts de type 2	Ressorts de type 3
1,99, 2,06, 1,99 1,94, 2,05, 1,88 2,30	2,85, 2 74, 2,74 2,63, 2 74, 2,80	2,10, 2 01, 1,93 2,02, 2 10, 2,05

Tableau 6.1.1.2 Constantes empiriques des ressorts

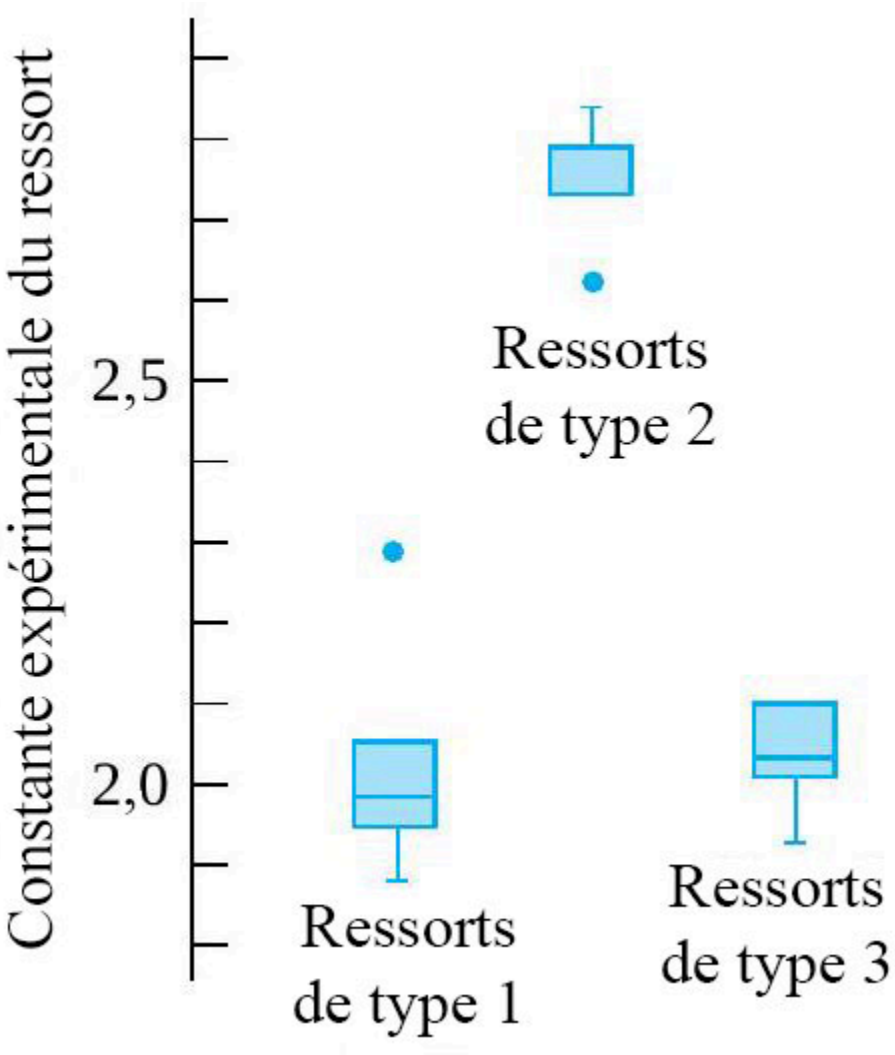


Figure 6.1.1.2 Diagrammes en boîte côte à côte des constantes empiriques de ressorts de trois types

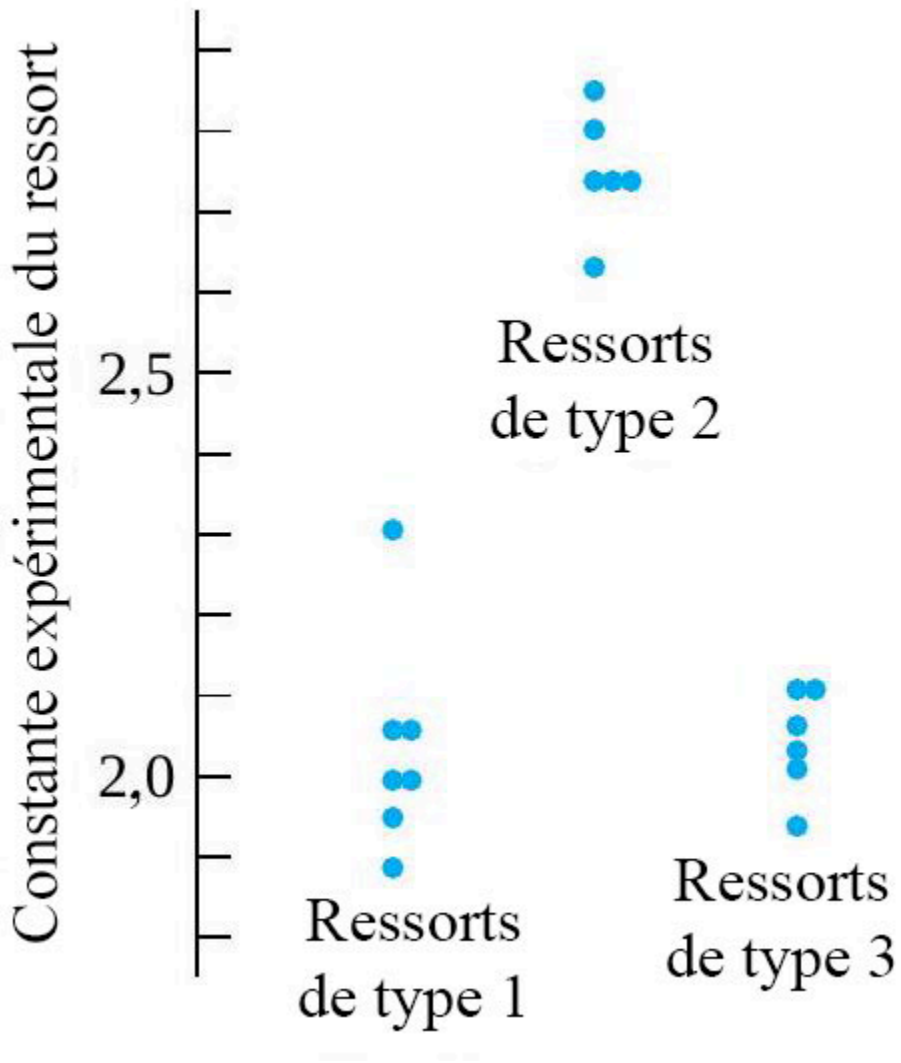


Figure 6.2.1.3 Diagrammes de dispersion côte à côte pour trois échantillons de constantes de ressorts empiriques

Les méthodes d'inférence statistique formelle ont pour but d'affiner et de quantifier les impressions que l'on peut avoir en réalisant une analyse descriptive de données. Mais en observant intelligemment les graphiques et en appliquant correctement les méthodes d'inférence formelle, il est rare d'obtenir des résultats complètement différents. En effet, les méthodes d'inférence formelle proposées ici pour des études multi-échantillons simples et non structurées sont confirmatoires – dans des cas comme ceux des exemples 1 et 2, ces méthodes ont pour but de confirmer ce qui ressort clairement d'une observation descriptive ou exploratoire des données.

6.1.2 Modèle multi-échantillons (normal) à un facteur, valeurs ajustées et résidus

HYPOTHÈSES DE MODÈLE NORMAL À UN FACTEUR

La partie 5 a beaucoup mis l'accent sur le fait que, pour faire des inférences relatives à un ou deux échantillons, il faut adopter un modèle de génération de données qui soit à la fois plausible et gérable. Il en va de même pour le cas présent, et les méthodes d'inférence standard pour études multi-échantillons non structurées reposent sur une extension naturelle du modèle utilisé à la section 5.3 pour comparer les moyennes de deux petits échantillons. Aux fins de la présente discussion, on supposera que les$r$ échantillons, de tailles respectives $n_1, n_2, \ldots, n_r$ sont indépendants et suivent des distributions normales, avec une variance commune de σ^2. Tout comme dans la section 5.3, où la version $r = 2$ de ce modèle à un facteur (contrairement aux modèles à plusieurs facteurs) a amené des méthodes d'inférence pratiques pour $\mu_1 - \mu_2$, cette version générale permettra d'utiliser de nombreuses méthodes d'inférence pratiques pour les études utilisant r échantillons. La figure 6.1.2.1 présente plusieurs distributions normales différentes ayant le même écart-type. Essentiellement, elle représente la source des réponses mesurées lorsque l'on applique les méthodes de ce chapitre.

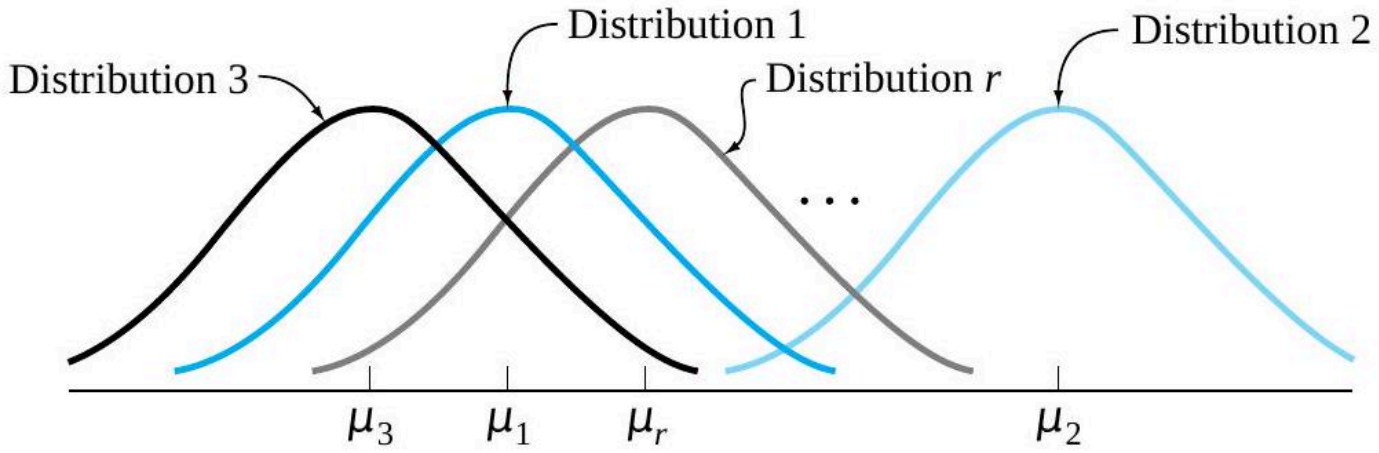


Figure 6.1.2.1 Distributions normales ayant le même écart-type

.

Outre la traduction en mots du modèle à un facteur et la représentation graphique de la figure 6.1.2.1, une traduction du modèle en symboles peut être utile. La présente section et les trois suivantes utilisent la notation

.

$$y_{ij} = \text{ l'observation } j \text{ dans l'échantillon } i$$

.

Le modèle d'équation utilisée pour spécifier le modèle à un facteur est alors :

.

6.1.2.1 Énoncé du modèle à un facteur en symboles

$$y_{ij} = \mu_i + \epsilon_{ij}$$

.

où μ_i est la i^{e} moyenne sous-jacente et les quantités $\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \epsilon_{22}, \ldots, \epsilon_{2n_2}, \ldots, \epsilon_{r1}, \epsilon_{r2}, \ldots, \epsilon_{rn_r}$ sont des variables aléatoires normales indépendantes de moyenne 0 et de variance σ^2. (Ici, les moyennes $\mu_1, \mu_2, \ldots, \mu_r$ et la variance σ^2 sont typiquement des paramètres inconnus.)

.
L'équation 6.1.2.1 présente exactement ce que véhiculent la figure 6.1.2.1 et la traduction en mots des hypothèses à un facteur : elle indique qu'une donnée de l'échantillon i se compose de la moyenne sous-jacente correspondante, à laquelle s'ajoute le bruit aléatoire suivant :
.

$$e_{ij} = y_{ij} - \mu_i$$

.
Il s'agit de la contrepartie théorique d'une notion empirique que nous aborderons plus tard, lorsque nous parlerons des moindres carrés dans un contexte de droite de régression. Il sera alors pertinent de décomposer les données en valeurs ajustées et en résidus correspondants.
.
Dans le cas présent, étant donné qu'on écarte volontairement toute structure reliant les r échantillons, il peut être difficile de savoir comment appliquer les notions de valeurs ajustées et de résidus. Toutefois, il est probable que
.

$$\hat{y}_{ij} = \text{ la valeur ajustée correspondant à } y_{ij}$$

.
corresponde en contexte à la moyenne du i^{e} échantillon
.

i^{e} moyenne d'échantillon $$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

Soit :
.

6.1.2.2 Valeurs ajustées pour le modèle à un facteur

$$\hat{y}_{ij} = \bar{y}_i$$

.
À la lumière de l'équation 6.1.2.2 qui fournit les valeurs ajustées d'une étude à r échantillons, le schéma établi indique que les résidus sont les différences entre valeurs observées et moyennes d'échantillon. Par conséquent, avec :
.

$$e_{ij} = \text{ la valeur résiduelle correspondant à } y_{ij}$$

.
on obtient :
.

6.1.2.3 Résidus pour modèle à un facteur

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$$

.
En réorganisant l'équation 6.1.2.3, on obtient
.

6.1.2.4 $$y_{ij} = \hat{y}_{ij} + e_{ij} = \bar{y}_i + e_{ij}$$

.

qui est la contrepartie empirique de l'énoncé théorique 6.1.2.1. En réalité, en combinant les équations 6.1.2.1 et 6.1.2.4, on obtient

.

6.1.2.5 $$y_{ij} = \mu_i + \epsilon_{ij} = \bar{y}_i + e_{ij}$$

.

Voici un exemple spécifique de schéma de pensée que l'on retrouve dans toutes les méthodes d'analyse courantes fondées sur la distribution normale pour les études multi-échantillons. Traduite en mots, l'équation (6.1.2.5) donnerait :

.

6.1.2.6 Observation = réponse déterministe + bruit = valeur ajustée + résidu

.

L'équation 6.1.2.6 est un paradigme présentant une approche unifiée permettant d'aborder la majorité des méthodes d'analyse présentées dans le reste de cet ouvrage.

.

Les décompositions 6.1.2.5 et 6.1.2.6 suggèrent que :

.

1. Les valeurs ajustées $\left(\hat{y}_{ij} = \bar{y}_i\right)$ sont censées représenter approximativement la part déterministe de la réponse du système (μ_i).

.

2. Les résidus (e_{ij}) sont donc censés représenter approximativement le bruit correspondant dans la réponse (ϵ_{ij}).

.

Le fait que les termes ϵ_{ij} de l'équation 6.1.2.1 sont supposés être des variables aléatoires normales indépendantes et identiquement distribuées (iid) $\left(0, \sigma^2\right)$ laisse alors supposer que les e_{ij} devraient au moins à peu près ressembler à un échantillon aléatoire suivant une distribution normale.

.

Le tracé normal d'un ensemble complet de résidus est donc un moyen de vérifier la pertinence du modèle à un facteur. Pour étudier la justesse des hypothèses du modèle, on peut aussi regarder les tracés de résidus en fonction : 1) des valeurs ajustées; 2) de l'ordre d'observation; ou 3) de toute autre variable pouvant se révéler intéressante, en espérant ne voir que des dispersions aléatoires.

.

Ces types de tracés combinant les résidus de tous les r échantillons sont souvent très utiles en pratique. Si r est très grand, les contraintes budgétaires concernant le coût total de la collecte de données restreignent souvent la taille $n_1, n_2, \ldots, n_r$ des échantillons à être relativement petite. Cela rend vaine toute étude des hypothèses du modèle de distribution normale de variance unique à l'aide (par exemple) de tracés normaux, échantillon par échantillon. (Évidemment, quand tous les $n_1, n_2, \ldots, n_r$ sont de taille décente, l'approche échantillon par échantillon peut être efficace.)

.

Exemple 6.1.2.1 (suite)

Revenons à notre étude sur la résistance du béton et étudions la pertinence du modèle 6.1.2.1 dans ce cas. Commençons par observer la figure 6.1.1.1 Comme on a pu le remarquer plus haut, elle donne visuellement l'impression

qu'au moins la partie « de même variance » des hypothèses du modèle à un facteur est plausible. Il est ensuite logique de calculer quelques statistiques synthétiques et de les examiner, notamment les écarts-types des échantillons. Le tableau 6.1.2.1 présente les tailles, moyennes et écarts-types des échantillons d'après les données du tableau 6.1.1.1.

.

Au premier abord, il peut sembler étrange que, dans ce tableau, s_1 soit plus de trois fois plus grand que s_8. Mais les échantillons sont si petits ($r = 8$ échantillons de taille 3 suivant une distribution normale) que ce n'est pas si inhabituel de voir un rapport de l'ordre de 3,2 entre le plus grand et le plus petit écart-type. À noter que d'après les tables F (table A1.5), même si on avait seulement deux écarts-types seulement sont impliqués (plutôt que huit), un rapport de variances de $(965.6/302.5)^2 \approx 10.2$ donnerait, pour les échantillons de taille 3, une valeur p située entre 0,10 et 0,20 pour le test de l'hypothèse nulle de variances égales, avec une hypothèse alternative bilatérale. Les écarts-types des échantillons du tableau 6.1.2.1 n'indiquent pas foncièrement que le modèle à un facteur n'est pas adapté.

.

Les échantillons étant trop petits, c'est peine perdu que de tenter de conclure quoi que ce soit d'utile sur huit tracés normaux distincts. Il est toutefois possible de tirer quelques informations en calculant et en traçant l'ensemble des $8 \times 3 = 24$ résidus. Le tableau 6.1.2.2 présente certains calculs nécessaires pour obtenir les résidus des données du tableau 6.1.1.1 (en utilisant les valeurs ajustées figurant dans le tableau 6.1.2.1 en tant que moyennes d'échantillons) sont présentés dans. Les figures 6.1.2.2 et 6.1.2.3 représentent respectivement un tracé de résidus en fonction de $y\,(e_{ij}$ en fonction de $\bar{y}_{ij}$et un tracé normal des 24 résidus.

.

La figure 6.1.2.2 n'indique pas que σ semble dépendre de μ (ce qui violerait la restriction de « variance constante »). Le tracé de la figure 6.1.2.3 est plutôt linéaire, ne présentant ainsi aucun obstacle clair quant à l'hypothèse d'une distribution normale. Dans l'ensemble, après examen des données brutes et des résidus, l'analyse des données du tableau 6.1.1.1 sur la base du modèle 6.1.2.1 semble parfaitement appropriée.

.

Synthèse des statistiques de l'étude sur la résistance du béton

i, Formule du béton	n_i, Taille de l'échantillon	$\bar{y}_i$, Moyenne de l'échantillon (psi)	s_i, Écart-type de l'échantillon (psi)
1	$n_1 = 3$	$\bar{y}_1 = 5\ 635{,}3$	$s_1 = 965{,}6$
2	$n_2 = 3$	$\bar{y}_2 = 5\ 753{,}3$	$s_2 = 432{,}3$
3	$n_3 = 3$	$\bar{y}_3 = 4\ 527{,}3$	$s_3 = 509{,}9$
4	$n_4 = 3$	$\bar{y}_4 = 3\ 442{,}3$	$s_4 = 356{,}4$
5	$n_5 = 3$	$\bar{y}_5 = 2\ 923{,}7$	$s_5 = 852{,}9$
6	$n_6 = 3$	$\bar{y}_6 = 3\ 324{,}7$	$s_6 = 353{,}5$
7	$n_7 = 3$	$\bar{y}_7 = 1\ 551{,}3$	$s_7 = 505{,}5$
8	$n_8 = 3$	$\bar{y}_8 = 2\ 390{,}7$	$s_8 = 302{,}5$

Tableau 6.1.2.1

.

Exemple de calculs de résidus dans l'étude de la résistance du béton

Échantillon	i, Formule du béton	y_{ij}, Résistance à la compression (psi)	$\hat{y}_{ij} = \bar{y}_i$, Valeur ajustée	e_{ij}, Résidus
1	1	5 800	5 635,3	164,7
2	1	4 598	5 635,3	−1 037,3
3	1	6 508	5 635,3	872,7
4	2	5 659	5 753,3	−94,3
5	2	6 225	5 753,3	471,7
⋮	⋮	⋮	⋮	⋮
22	8	2 051	2 390,7	−339,7
23	8	2 631	2 390,7	240,3
24	8	2 490	2 390,7	99,3

Tableau 6.1.2.2

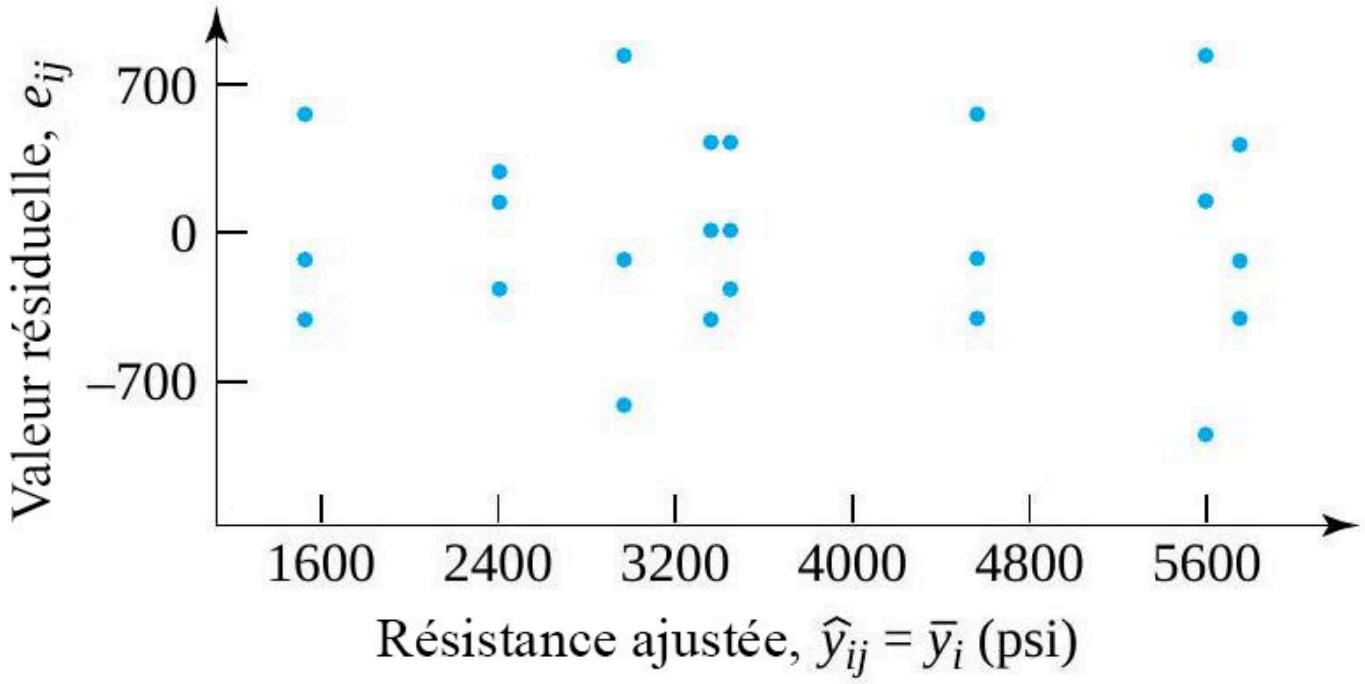


Figure 6.1.2.2 Tracé des résidus en fonction de la réponse ajustée pour la résistance à la compression

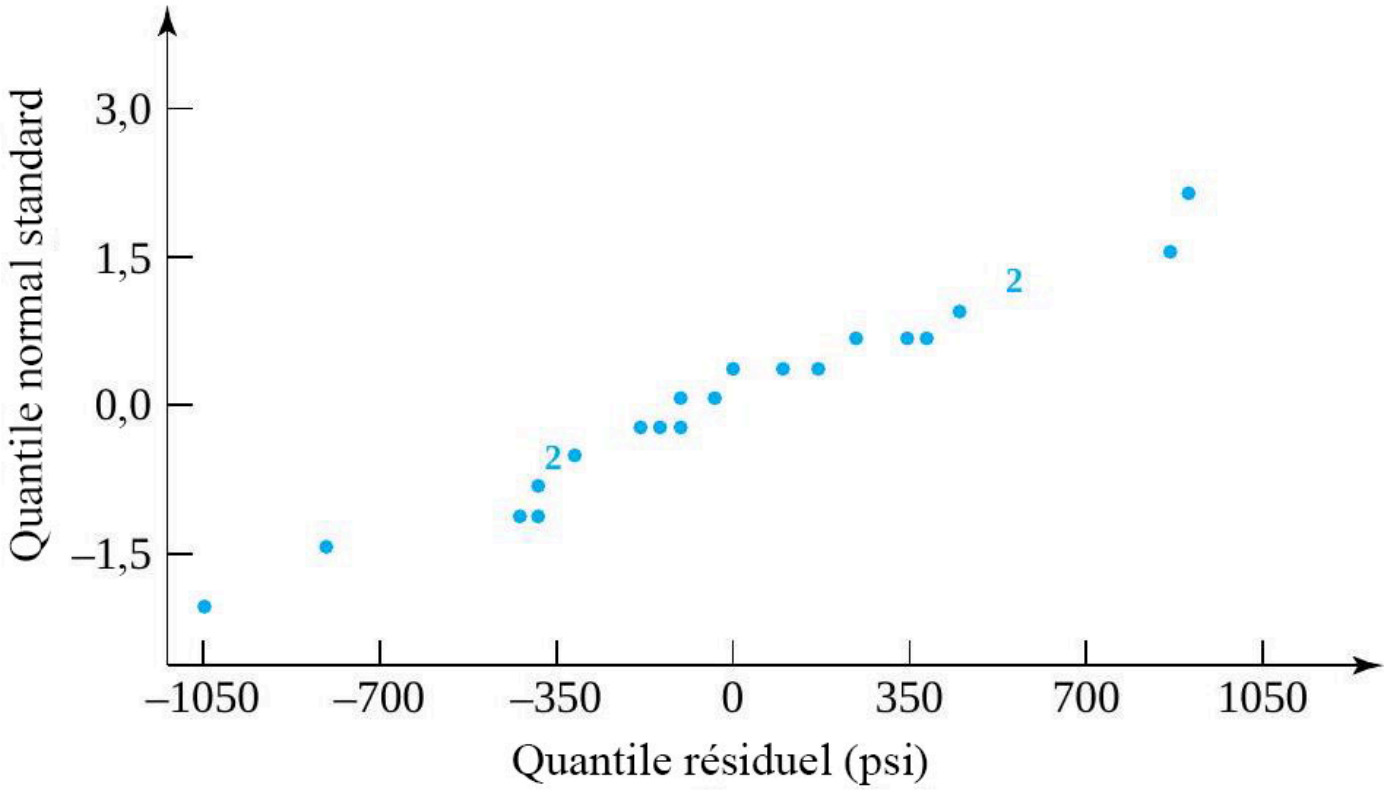


Figure 6.1.2.3 Tracé normal des résidus pour la résistance à la compression

.

Exemple 6.1.2.2 Test sur les ressorts (suite)

Les données expérimentales sur les ressorts peuvent également être analysées en tenant compte de l'utilisation potentielle du modèle normal à un facteur 6.1.1.1. Les figures 6.1.1.2 et 6.1.1.3 indiquent des variabilités comparables pour les constantes de ressort expérimentales, pour les r = 3 types de ressorts différents. La valeur élevée du ressort de type 1 suscite le doute concernant cette opinion et la description de « distribution normale » des constantes expérimentales de type 1 (en raison de la position de cette valeur aberrante sur le diagramme en boîtes). Le tableau 6.1.2.3 présente des statistiques synthétiques de ces échantillons.

.

Sans cette valeur extrême de 2,30, l'écart-type du premier échantillon serait 0,068, ce qui est totalement en adéquation avec ceux des autres échantillons. Mais même le rapport entre la plus grande et la plus petite variance d'échantillons (à savoir $(.134/.064)^2 = 4.38$) ne suffit pas pour abandonner la description de modèle à un facteur des constantes de ressort. (On constate dans les tables F pour $v_1 = 6$ et $v_2 = 5$ que 4,38 se trouve entre les quantiles 0,9 et 0,95 de la distribution $F_{6,5}$. Par conséquent, même s'il y avait seulement deux échantillons et non trois, un rapport de variance de 4,38 donnerait une valeur p située entre 0,1 et 0,2 pour le test (bilatéral) d'égalité des variances.) Avant de laisser la constante empirique du ressort de type 1 de 2,30 mener à l'abandon du très utile modèle 6.1.2.1, il est judicieux de regarder ce qui se passe d'un peu plus près.

.

Les tailles d'échantillon $n_1 = 7$ et $n_2 = n_3 = 6$ sont suffisamment grandes pour qu'il soit pertinent d'observer les

tracés normaux des données constantes des ressorts échantillon par échantillon. La figure 6.1.2.4 présente ces tracés, réalisés sur les mêmes axes. De plus, l'utilisation des valeurs ajustées $(\bar{y}_i)$ du tableau 6.1.2.3, dont les données originales proviennent du tableau 6.1.1.2, produit 19 résidus, comme l'illustre en partie le tableau 6.1.2.4. Les figures 6.1.2.5 et 6.1.2.6 montrent ensuite respectivement un tracé de résidus en fonction des réponses ajustées, et un tracé normal de l'ensemble des 19 résidus.

.

Mais les figures 6.1.2.5 et 6.1.2.6 attirent elles aussi l'attention sur la constante empirique de ressort de type 1 la plus élevée. En comparaison avec les autres valeurs mesurées, 2,30 est tout simplement une valeur trop élevée (et produit ainsi un résidu trop important par rapport aux autres) pour qu'on puisse réellement appliquer le modèle 6.1.2.1 aux données des constantes de ressort. Hormis si, en vérifiant les fiches techniques originales, on découvrait que la valeur 2,30 était une grossière erreur de calcul ou de mesure (ce qui pourrait soit être corrigé, soit justifier d'omettre cette valeur), il semble que l'utilisation du modèle 6.1.2.1 avec les r = 3 types de ressorts pourrait produire des inférences aux propriétés réelles (et inconnues) assez différentes de leurs propriétés nominales.

.

On peut évidemment se limiter à étudier les ressorts de type 2 et 3. Il n'y a rien dans les deuxième et troisième échantillons qui rende le modèle de « distributions normales de même variance » indéfendable pour ces deux types de ressorts. Mais le schéma de variation des ressorts de type 1 semble être clairement différent de celui des ressorts de types 2 et 3, et ce modèle à un facteur n'est donc pas adéquat si l'on tient compte des trois types.

Synthèse des statistiques des constantes empiriques des ressorts

i, Type de ressort	n_i	$\bar{y}_i$	s_i
1	7	2,030	0,134
2	6	2,750	0,074
3	6	2,035	0,064

Tableau 6.1.2.3 Statistiques synthétiques des constantes de ressort empiriques

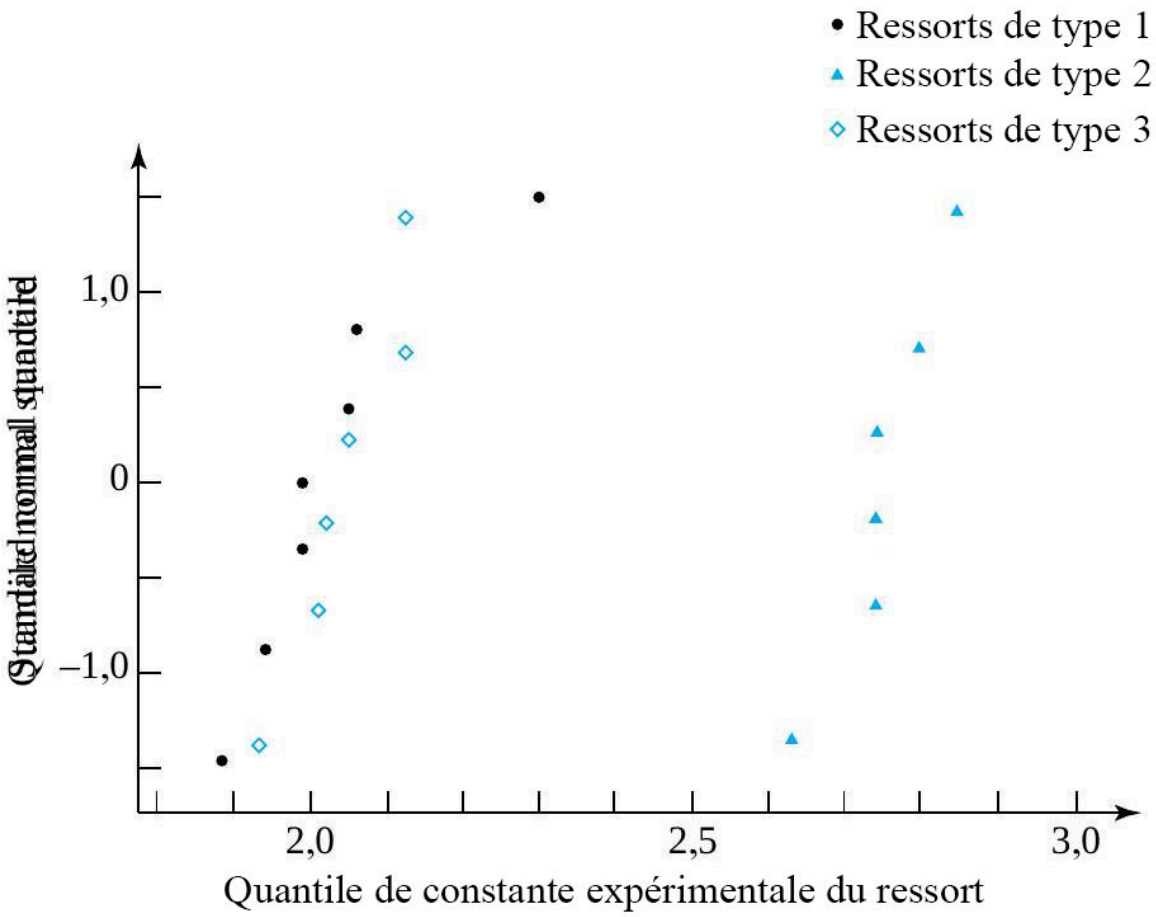


Figure 6.1.2.4 Tracés normaux des constantes de ressort empiriques pour trois types de ressorts

Exemple de calculs de valeurs résiduelles dans l'étude des constantes de ressort

i, Type de ressort	j, Numéro d'observation	y_{ij}, Constante de ressort	$\hat{y}_{ij} = \bar{y}_i$, Moyenne de l'échantillon	e_{ij}, Résidus
1	1	1,99	2,030	− 0, 040
⋮	⋮	⋮	⋮	⋮
1	7	2,30	2,030	0, 270
2	1	2,85	2,750	0, 100
⋮	⋮	⋮	⋮	⋮
2	6	2,80	2,750	0, 050
3	1	2,10	2,035	0, 065
⋮	⋮	⋮	⋮	⋮
3	6	2,05	2,035	0, 015

Table 6.1.2.4 Exemple de calculs de résidus dans l'étude des constantes des ressorts

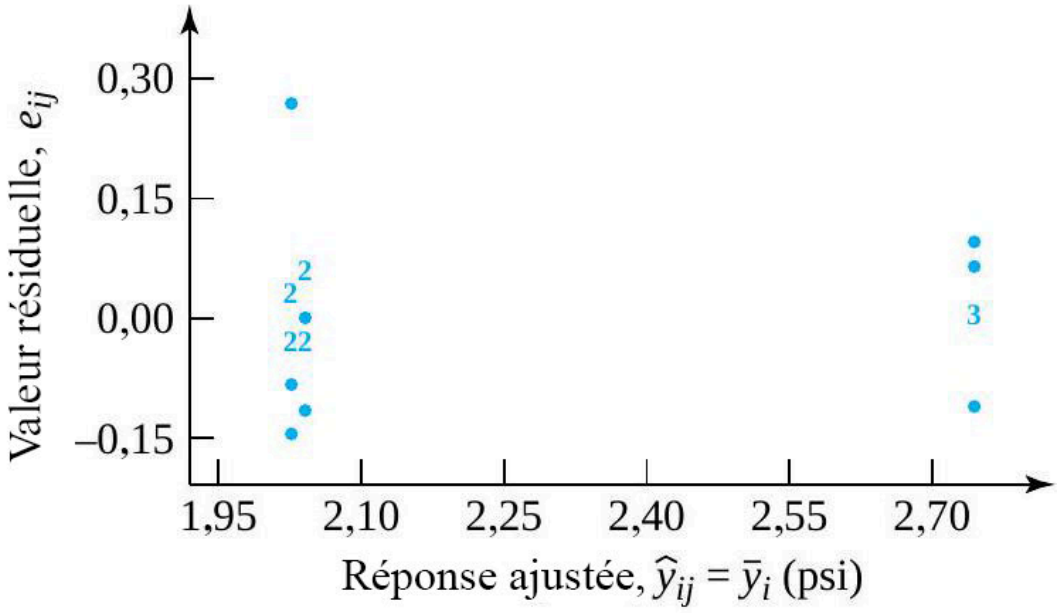


Figure 6.1.2.5 Tracé des résidus en fonction des réponses ajustées pour les constantes de ressort empiriques

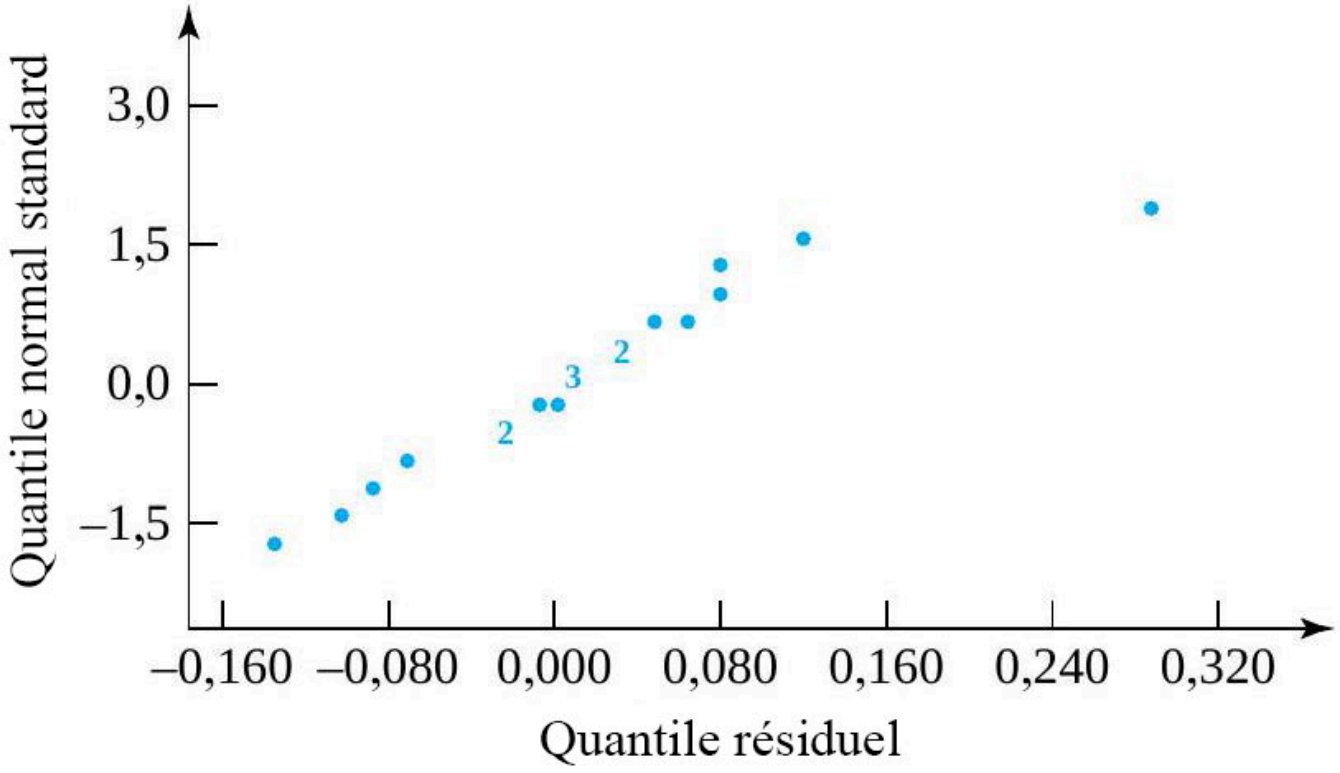


Figure 6.1.2.6 Tracé normal des résidus des constantes de ressort

6.1.3 Estimation de la variance pondérée pour les études multi-échantillons

Le modèle de « distributions normales de même variance » 6.1.2.1 présente un paramètre fondamental : σ, l'écart-type associé aux réponses des conditions $1, 2, 3, \ldots, r$. À l'image de ce qui a été fait dans le cas $r = 2$ de la partie 5, il est tout à fait typique, dans les études multi-échantillons, de regrouper les variances d'échantillon r pour parvenir à une seule estimation de σ dérivée de tous les r échantillons.

DÉFINITION Écart-type pondéré

.

EXPRESSION 6.1.3.1

.

Si r échantillons numériques de tailles respectives $n_1, n_2, \ldots, n_r$ produisent des variances d'échantillons $s_1^2, s_2^2, \ldots, s_r^2$, la **variance pondérée**, s_P^2 est la moyenne pondérée des variances d'échantillons, où les coefficients de pondération équivalent aux tailles des échantillons moins 1 :

.

$$s_\mathrm{P}^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 + \cdots + (n_r - 1)\, s_r^2}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_r - 1)}$$

.

L'écart-type pondéré des échantillons, s_P, correspond à la racine carrée de s_p^2.

.

La définition 6.1.3.1 élargit simplement celle donnée dans la partie 5 pour l'adapter aux cas à plus de deux échantillons. Comme c'était le cas pour s_p avec deux échantillons, s_p doit se situer entre les valeurs maximum et minimum des s_i; c'est une sorte de compromis pratique pour cette valeur sur le plan mathématique.

.

L'équation 6.1.3.1 peut être réécrite sous de nombreuses formes, toutes équivalentes. Avant tout, soit

.

> le nombre total d'observations dans le cadre d'une étude à r échantillons
>
> $n = \sum_{i=1}^{r} n_i = \text{ the total number of observations in all } r \text{ samples}$

.

Il est courant de réécrire le dénominateur à la droite de l'équation 6.1.3.1 comme suit :

.

$$\sum_{i=1}^{r} (n_i - 1) = \sum_{i=1}^{r} n_i - \sum_{i=1}^{r} 1 = n - r$$

.

Comme la variance de l'échantillon i vaut :

.

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

.

Le numérateur à la droite de l'équation 6.1.3.1 est :

.

> 6.1.3.2 et 6.1.3.3

$$\sum_{i=1}^{r} (n_i - 1) \left(\frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$
$$= \sum_{i=1}^{r} \sum_{j=1}^{n_i} e_{ij}^2$$

.

> Autres formules pour s_P^2
>
> .
>
> On peut donc définir s_{P}^2 en fonction du côté droit de l'équation 6.1.3.2 ou 6.1.3.3 divisé par $n - r$.

Exemple 6.1.3.1 Résistance à la compression (suite)

Pour les données de résistance à la compression, chacune des valeurs $n_1, n_2, \ldots, n_8$ vaut 3, et les valeurs s_1 à s_8 sont données dans le tableau 6.1.2.1. Donc en utilisant l'équation 6.1.3.1,

.

Formula does not parse

.

et par conséquent

.

$$s_{\mathrm{P}} = \sqrt{338\ 213} = 581{,}6\ \text{psi}$$

.

On estime que si un grand nombre d'échantillons de n'importe laquelle des formules 1 à 8 était testé, on obtiendrait un écart-type des résistances à la compression de l'ordre de 582 psi.

Signification de s_P

s_{P} est une estimation de la variation intrinsèque ou de référence d'une variable étudiée pour un ensemble de conditions fixe, calculée en supposant que cette variation de référence est constante dans les conditions dans lesquelles les échantillons ont été prélevés. Lorsque cette supposition est raisonnable, l'idée de la pondération est de combiner plusieurs estimations de petits échantillons (non fiables séparément) pour obtenir une seule estimation relativement plus fiable. C'est une mesure fondamentale très appliquée dans de nombreuses méthodes efficaces d'inférence formelle.

Limites de confiance pour la variance du modèle à un facteur

Parfois, en plus d'une estimation de σ^2 basée sur des données, on a également besoin d'un intervalle de confiance. Avec les restrictions du modèle 6.2.1.1, la variable

.

$$\frac{(n-r)s_{\mathrm{P}}^2}{\sigma^2}$$

.

suit une distribution χ^2_{n-r}. Ainsi, de la même façon que pour la dérivation de la partie 5, l'intervalle de confiance bilatéral pour

$< imgsrc =$" $https : //ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/3250dba6ba9d84059174d0c9a2d8c5fb.png$ " $alt =$" σ^2 a pour bornes

.

6.1.3.4 $\dfrac{(n-r)s_{\mathrm{P}}^2}{U}$ et $\dfrac{(n-r)s_{\mathrm{P}}^2}{L}$ " $title =$" σ^2 a pour bornes

.

6.1.3.4

\frac{(n-r) s_{\mathrm{P}}^2}{U} \text { et } \frac{(n-r) s_{\mathrm{P}}^2}{L}" class="latex mathjax"></div> <div>où et sont telles que la probabilité assignée à l'intervalle correspond au niveau de confiance souhaité. Et bien sûr, on peut obtenir un intervalle unilatéral en utilisant uniquement l'une des bornes 6.1.3.4 et en choisissant une valeur ou telle que la probabilité assignée à l'intervalle [latex](0, U)
ou (L, ∞) correspond au niveau de confiance souhaité.

Exemple 6.1.3.2 (suite)

Dans le cas de la résistance à la compression du béton, utilisons l'équation 6.1.3.4 pour obtenir un intervalle de confiance bilatéral de 90% pour σ. Étant donné que $n-r=16$ degrés de liberté sont associés à s_{P}^2, on consulte le tableau A1.4 pour les quantiles 0,05 et 0,95 de la distribution χ_{16}^2. On lit alors respectivement 7,962 et 26,296. Ainsi, l'intervalle de confiance pour σ^2 a pour bornes

.

$$\frac{16(581,6)^2}{26,296} \quad \text{et} \quad \frac{16(581,6)^2}{7,962}$$

.

L'intervalle de confiance bilatéral de 90% pour σ a pour bornes

.

$$\sqrt{\frac{16(581,6)^2}{26,296}} \quad \text{et} \quad \sqrt{\frac{16(581,6)^2}{7,962}}$$

.

soit :

.

$$453,7 \text{ psi et } 824,5 \text{ psi}$$

6.2.0 Intervalles de confiance d'études multi-échantillons – Introduction

La partie 5 illustre combien les intervalles de confiance pour moyennes et différences de moyennes peuvent être pratiques dans les études à un ou deux échantillons. Estimer une moyenne individuelle et comparer une paire de moyennes est tout aussi important quand il y a r échantillons que lorsqu'il n'y en a qu'un ou deux. Les méthodes de la partie 5 peuvent être appliquées aux études à r échantillons en se focalisant tout simplement sur un ou deux échantillons à la fois.

Mais étant donné que les échantillons sont souvent petits dans les études multi-échantillons, cette stratégie d'inférence se révèle souvent relativement peu éclairante. Dans le cadre des hypothèses du modèle à un facteur abordées à la section précédente, il est possible de fonder les méthodes d'inférence sur l'écart-type pondéré, s_p. Dans ce contexte, ces méthodes ont en effet tendance à être plus instructives que les équations de la partie 5 telles quelles. Cette section tient d'abord compte de l'estimation de l'intervalle de confiance d'une seule moyenne et de la différence entre deux moyennes suivant le modèle de « distributions normales de même variance » avant de formuler quelques commentaires abordant les notions de niveaux de confiance individuels et simultanés.

6.2.1 Intervalles pour moyennes et comparaison de moyennes

Le principal inconvénient à appliquer les équations de la partie 5 à des échantillons multiples est typiquement que la petite taille des échantillons se traduit par un petit nombre de degrés de liberté et de grandes valeurs de t dans la partie ± des formules d'intervalle – par conséquent, on se retrouve avec de grands intervalles. Mais grâce aux hypothèses du modèle à un facteur, il est possible d'obtenir des formules d'intervalles de confiance qui ont tendance à produire des intervalles plus petits.

.

C'est-à-dire, dans un développement similaire à celui de la partie 5, et en suivant le modèle normal à un facteur, que

.

$$T = \frac{\bar{y}_i - \mu_i}{\frac{s_P}{\sqrt{n_i}}}$$

.

présente une distribution t_{n-r}. Par conséquent, l'intervalle de confiance bilatéral pour la i^e moyenne, μ_i, a pour bornes

.

6.2.1.1 Limites de confiance pour μ_i basées sur le modèle à un facteur

.

$$\bar{y}_i \pm t\frac{s_P}{\sqrt{n_i}}$$

.

où le niveau de confiance associé correspond à la probabilité assignée à l'intervalle entre $-t$ et t dans la distribution t_{n-r}. Il s'agit de la même formule qu'à la partie 5, hormis le fait que s_P a remplacé s_i et que les degrés de liberté sont passés de $n_i - 1$ à $n - r$.

.

De la même manière, pour les conditions i et i', la variable

.

$$T = \frac{\bar{y}_i - \bar{y}_{i'} - (\mu_i - \mu_{i'})}{s_P\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}}$$

.

suit une distribution t_{n-r}. Par conséquent, l'intervalle de confiance bilatéral pour $\mu_i - \mu_{i'}$ a pour bornes

6.2.1.2 Limites de confiance pour $\mu_i - \mu_{i'}$ basées sur le modèle à un facteur

.

$$\bar{y}_i - \bar{y}_{i'} \pm ts_P\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

.

où le niveau de confiance associé correspond à la probabilité assignée à l'intervalle entre $-t$ et t dans la distribution t_{n-r}. L'équation 6.2.1.2 reprend essentiellement la formule de la partie 5, hormis le fait que s_P est calculé sur la base de r échantillons plutôt que deux, et qu'il y a $n - r$ degrés de libertés plutôt que $n_i + n_{i'} - 2$.

.

Évidemment, utiliser uniquement une borne de la formule 6.2.1.1 ou 6.2.1.2 donne un intervalle de confiance unilatéral, pour lequel le niveau de confiance associé correspond à la probabilité t_{n-r} assignée à l'intervalle $(-\infty, t)$ (avec $t > 0$0" title= »t>0" class= »latex mathjax »>). L'avantage des formules 6.2.1.1 et 6.2.1.2 (lorsqu'on peut les appliquer), en comparaison avec les formules correspondantes de la partie 5, c'est que pour un niveau de confiance donné, elles ont tendance à produire des intervalles plus courts.

.

Exemple 6.2.1.1 Intervalles de confiance pour moyennes et différences de moyennes des résistances à la compression du béton (suite)

Reprenons l'étude de résistance à la compression du béton d'Armstrong, Babb et Campen. Créons d'abord un intervalle de confiance bilatéral de 90% pour la résistance moyenne à la compression d'une seule formule de béton, puis un intervalle de confiance bilatéral de 90% pour la différence de résistance moyenne de deux formules. Étant donné que $n = 24$ et $r = 8$, il y a $n - r = 16$ degrés de liberté associés à $s_{\mathrm{P}} = 581{,}6$. Le quantile 0,95 de la distribution t_{16}, à savoir 1,746, peut alors être utilisé dans les deux formules 6.2.1.1 et 6.2.1.2.

.

Penchons-nous d'abord sur l'estimation d'une seule résistance moyenne à la compression; puisque tous les n_i valent 3, la partie ± de la formule 6.2.1.1 donne :

.

$$t\frac{s_{\mathrm{P}}}{\sqrt{n_i}} = 1{,}746\frac{581{,}6}{\sqrt{3}} = 586{,}3 \text{ psi}$$

.

La précision de ± 586.3 psi pourrait être rattachée à n'importe laquelle des moyennes d'échantillon du Tableau 6.2.1.1 comme estimation de la résistance moyenne de la formule correspondante. Par exemple, comme $\bar{y}_3 = 4\ 527{,}3$ psi, l'intervalle de confiance bilatéral de 90% pour μ_3 a pour bornes

.

$$4,527.3 \pm 586.3$$

.

soit :

.

$$3\ 941{,}0 \text{ psi et } 5\ 113{,}6 \text{ psi}$$

.

De la même manière, estimons la différence entre deux moyennes de résistances à la compression avec un niveau de confiance de 90%. Là encore, puisque tous les n_i valent 3, la partie ± de la formule 6.2.1.2 donne :

.

$$ts_{\mathrm{P}}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} = 1{,}746(581{,}6)\sqrt{\frac{1}{3} + \frac{1}{3}} = 829{,}1 \text{ psi}$$

.

La précision de ± 829.1 psi pourrait être rattachée à n'importe quelle différence entre les moyennes d'échantillons du tableau 6.2.1.1 comme estimation des résistances moyennes de la différence des formules correspondante. Par exemple, étant donné que $\bar{y}_3 = 4\ 527{,}3$ psi et que $\bar{y}_7 = 1\ 551{,}3$ psi , l'intervalle de confiance bilatéral de 90% pour $\mu_3 - \mu_7$ a pour bornes

.

$$(4,527.3 - 1,551.3) \pm 829.1$$

.

soit :

.

$$2\ 146{,}9 \text{ psi} \quad \textbf{et} \quad 3\ 805{,}1 \text{ psi}$$

Résistance moyenne des échantillons de béton selon la formule

Formule du béton	Résistance moyenne de l'échantillon
1	5 635,3
2	5 753,3
3	4 527,3
4	3 442,3
5	2 923,7
6	3 324,7
7	1 551,3
8	2 390,7

Tableau 6.2.1.1 Résistance moyenne des échantillons de béton selon la formule

.

L'utilisation de $n - r = 16$ degrés de liberté dans l'exemple 6.2.1.1 plutôt que de $n_i - 1 = 2$ et de $n_i + n_{i'} - 2 = 4$ reflète la baisse de l'incertitude associée à l'utilisation de s_{P} pour estimer σ plutôt que l'utilisation de s_i et de s_{P}, comme on le faisait avec deux échantillons. Cette réduction de l'incertitude vient avec une condition : elle n'est valide que pour les modèles de variances égales.

6.2.2 Niveaux de confiance individuels et simultanés

Cette section a présenté des intervalles de confiance divers et variés pour les études multi-échantillons. Nombre d'entre eux peuvent être utilisés dans le cadre d'une application donnée, peut-être même plusieurs à la fois. Par exemple, même dans le cadre relativement simple de l'exemple 6.2.1.1. (étude d'absorption de la serviette en papier), il serait raisonnable de vouloir des intervalles de confiance pour toutes les valeurs suivantes :

.

$$\mu_1, \mu_2, \mu_3, \mu_1 - \mu_2, \mu_1 - \mu_3, \mu_2 - \mu_3, \text{ and } \mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$$

.

Étant donné qu'il faut souvent donner plusieurs estimations de confiance dans les études multi-échantillons, il est important de réfléchir au sens de ces niveaux de confiance et de ne pas oublier qu'ils sont rattachés à un seul intervalle. Autrement dit, si on donne plusieurs intervalles de confiance de 90%, ce chiffre de 90% s'applique individuellement à chaque intervalle : on est « à 90% sûr » du premier intervalle, mais aussi « à 90% sûr » du deuxième, et également « à 90% sûr » du troisième, et ainsi de suite. Il est très difficile de savoir comment obtenir un taux de fiabilité conjoint ou simultané pour les intervalles (par exemple une probabilité a priori que tous les intervalles s'appliquent), mais il est assez évident que ce taux doit être inférieur à 90%. Le niveau de confiance simultané ou conjoint (le taux de fiabilité global) à associer à un groupe d'intervalles n'est en effet généralement pas facile à déterminer, mais il est typiquement inférieur (et parfois même bien inférieur) au niveau de confiance de chaque intervalle individuel.

.

Maintenant que la différence entre niveaux de confiance simultanés et individuels établie, il existe au moins trois approches possibles. L'option la plus évidente est de créer des intervalles de confiance individuels et de veiller à bien les interpréter comme tels (sachant que plus le nombre
d'intervalles augmente, plus il est probable qu'un ou plusieurs de ces intervalles ne couvrent pas les quantités qu'ils sont censés couvrir).

.

Une autre façon de résoudre le problème de confiance simultanée plutôt qu'individuelle est d'employer des niveaux individuels très élevés pour chaque intervalle, puis d'utiliser une inégalité assez sommaire, l'inégalité de Bonferroni, pour trouver au moins la valeur minimum de confiance simultanée associée au groupe d'intervalles. Cette inégalité indique que si k intervalles de confiance ont des niveaux de confiance associés $\gamma_1, \gamma_2, \ldots, \gamma_k$, le niveau de confiance simultané ou conjoint valable pour tous les k intervalles (disons γ) satisfait à

.

> **6.2.2.1** Inégalité de Bonferroni
>
> .
>
> $$\gamma \geq 1 - ((1 - \gamma_1) + (1 - \gamma_2) + \cdots + (1 - \gamma_k))$$

.

(Cette inégalité dit essentiellement que la « non-confiance » $(1 - \gamma)$ conjointe des k intervalles n'est pas supérieure à la somme des k « non-confiances » individuelles. Par exemple, cinq intervalles de niveaux de confiance individuels de 99% présentent un niveau de confiance conjoint ou simultané d'au moins 95 %.)

.

La troisième façon d'aborder le problème du niveau de confiance simultané est de développer et d'utiliser des méthodes qui, pour certains ensembles spécifiques de quantités inconnues, donnent des intervalles de niveau de confiance simultané connu. Des livres entiers sont consacrés à ce genre de méthodes d'inférence simultanée.

La section suivante aborde l’une des plus simples et plus connues.

6.2.3 Méthodes d'intervalles de confiance simultanés

Comme l'explique la section 6.2.2, dans le cadre d'une étude multi-échantillons, on peut créer plusieurs types d'intervalles de confiance pour des moyennes ou des combinaisons linéaires de moyennes. On y a aussi parlé du problème des niveaux de confiance individuels plutôt que simultanés, mais le seul moyen mentionné pour obtenir un niveau simultané était d'utiliser l'inégalité de Bonferroni.

.

Cette section présente une méthode permettant de créer divers intervalles de confiance et de maintenir un niveau de confiance simultané au cours du processus. Il s'agit de la méthode de Tukey pour l'estimation d'intervalles de confiance simultanés de toutes les différences deux à deux entre les moyennes sous-jacentes.

MÉTHODE DE TUKEY

Souvent, dans les études à r échantillons, on s'intéresse aussi aux différences entre les $\frac{r(r-1)}{2}$ paires de réponses moyennes μ_i et $\mu_{i'}$. La section 6.2 affirme qu'estimer une différence unique entre les réponses de moyenne $\mu_i - \mu_{i'}$ est possible à l'aide d'un intervalle ayant pour bornes

.

> 6.2.3.1 $\bar{y}_i - \bar{y}_{i'} \pm t s_{\mathrm{P}} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$

.

où le niveau de confiance associé est individuel. Mais si, par exemple, $r = 8$, il y a 28 différentes comparaisons deux à deux de moyennes sous-jacentes à prendre en compte (μ_1 et μ_2, μ_1 et $\mu_3, \ldots, \mu_1$ et μ_8, μ_2 et $\mu_3, \ldots$, et μ_7 et μ_8). Si on veut garantir un niveau de confiance simultané raisonnable pour toutes ces comparaisons en utilisant l'inégalité rudimentaire de Bonferroni, il faut un très haut niveau de confiance individuel pour les intervalles 6.2.3.1. Par exemple, avec 28 intervalles, il faut un niveau de confiance de $99{,}82\%$ pour pouvoir garantir un niveau de confiance simultané de 95%.

.

Une meilleure approche pour définir les limites de confiance simultanées pour toutes les différences $\mu_i - \mu_{i'}$ est de remplacer t dans la formule 6.2.3.1 par un multiplicateur obtenu expressément afin de fournir un niveau de confiance simultané pour l'estimation de ce genre de différences. C'est J. Tukey qui a expliqué qu'il était possible de trouver de tels multiplicateurs en utilisant les quantiles de distributions d'intervalles studentisés. Les tables A5A et A5B donnent les valeurs des constantes q^*, de sorte que l'ensemble d'intervalles bilatéraux ayant pour bornes

.

> **6.2.3.2 Limites de confiance simultanées bilatérales de Tukey pour toutes les différences entre r moyennes**
>
> $$\bar{y}_i - \bar{y}_{i'} \pm \frac{q^*}{\sqrt{2}} S_{\mathrm{P}} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

.

présente un niveau de confiance simultané d'au moins 95% ou 99% (selon si on lit $Q(.95)$ dans la table A5A ou $Q($

0,99) dans la table A5B) pour l'estimation de toutes les différences $\mu_i - \mu_{i'}$. Si toutes les tailles d'échantillon $n_1, n_2, \ldots, n_r$ sont les mêmes, le niveau de confiance simultané nominal de 95% ou 99% est exact, tandis que si les tailles sont différentes, la valeur réelle est au moins égale à la valeur nominale.

.

Pour pouvoir appliquer la méthode de Tukey, il faut trouver, dans la table A5 (par interpolation, si nécessaire), la colonne correspondant au nombre d'échantillons ou de moyennes à comparer, ainsi que la ligne correspondant aux degrés de liberté associés à s_P, (à savoir $v = n - r$).

.

Exemple 6.2.3.1 Résistances à la compression (suite)

La figure 6.2.3.1 représente le graphique de la résistance moyenne à la compression de huit échantillons, avec des barres d'erreur dérivées des limites de confiance simultanées.

Calculons les intervalles de confiance pour les différences de résistances à la compression selon l'échantillon. Si on veut obtenir un intervalle de confiance individuel bilatéral de 95 % pour une différence $\mu_i - \mu_{i'}$ donnée, la formule 6.2.3.1 montre que les bornes adéquates sont

.

$$\bar{y}_i - \bar{y}_{i'} \pm 2{,}120(581{,}6)\sqrt{\frac{1}{3} + \frac{1}{3}}$$

.

soit :

.

$$\bar{y}_i - \bar{y}_{i'} \pm 1\ 006{,}7\text{psi}$$

.

En revanche, si on souhaite estimer toutes les différences de résistances moyennes à la compression avec un niveau de confiance simultané de 95 %, l'équation 6.2.3.2 dit que les intervalles bilatéraux de Tukey ont pour bornes

.

$$\bar{y}_i - \bar{y}_{i'} \pm \frac{4.90}{\sqrt{2}}(581.6)\sqrt{\frac{1}{3} + \frac{1}{3}}$$

.

soit :

.

$$\bar{y}_i - \bar{y}_{i'} \pm 1\ 645{,}4\ \text{psi}$$

.

(Le chiffre 4,90 correspond à la valeur à la colonne $r = 8$ et à la ligne $v = 16$ de la table A5A.)

.

Compte tenu du fait que le niveau de confiance associé aux deuxièmes intervalles est simultané, les intervalles de Tukey sont plus larges que ceux indiqués par la première formule.

.

La partie ± de l'équation finale est moins du double de la partie ± de l'expression précédente. Ainsi, sur la figure 6.3.2.1, il n'est pas nécessaire que les barres d'erreur entourant deux moyennes données ne se chevauchent pas pour pouvoir affirmer qu'on peut détecter une différence entre ces moyennes. Il suffit plutôt que la différence entre deux moyennes d'échantillon soit au moins égale à la partie ± de la formule 6.2.3.2, soit 1 645,4 psi dans le cas présent.

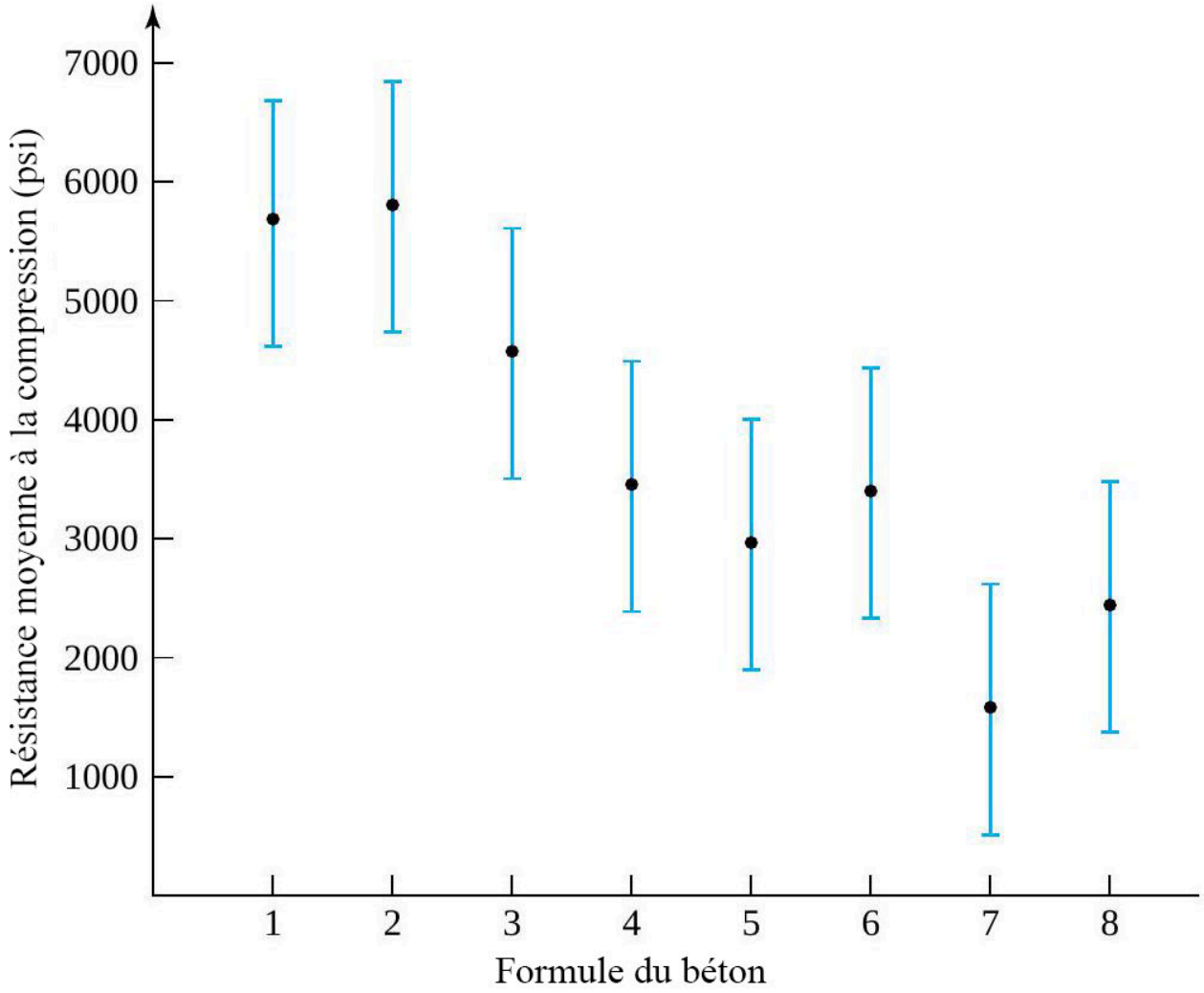


Figure 6.2.3.1 Graphique de la résistance moyenne à la compression de huit échantillons, avec barres d'erreur dérivées des limites de confiance simultanées.

6.3.0 ANOVA – Introduction

Dans ce cours, l'approche de l'inférence dans les études multi-échantillons a jusque-là été complètement axée sur les intervalles. Il existe toutefois aussi des méthodes de test d'hypothèses qui se prêtent aux études multi-échantillons. Cette section examine certaines de ces méthodes, ainsi que les problèmes soulevés par leur utilisation. Elle commence par quelques commentaires généraux concernant les tests d'hypothèse dans les études à r échantillons. Elle abordera ensuite le test d'analyse de variance à un facteur (ANOVA) pour l'égalité de r moyennes. Le tableau d'ANOVA à un facteur sera présenté, ainsi que son organisation et ce qu'il sous-entend.

6.3.1 Tests d'hypothèse et études multi-échantillons

De même que de nombreuses quantités peuvent être à estimer dans une étude multi-échantillon, il peut y avoir de nombreux tests d'hypothèse à effectuer. On peut par exemple souhaiter obtenir des valeurs p pour des hypothèses comme :

.

6.3.1 .1 $\mathrm{H}_0 : \mu_3 = 7$

6.3.1.2 $\mathrm{H}_0 : \mu_3 - \mu_7 = 0$

6.3.1.3 $\mathrm{H}_0 : \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0$

.

Les méthodes d'intervalle de confiance abordées à la section 6.2 ont des méthodes équivalentes pour les tests d'hypothèse qui, comme les trois ci-dessus, impliquent une combinaison linéaire des moyennes $\mu_1, \mu_2, \ldots, \mu_r$.

.

En général (sous le modèle à un facteur standard), soit :

.

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r$$

.

Alors, l'hypothèse

6.3.1.4 $\mathrm{H}_0 : L = \#$

.

peut être testée à l'aide de la statistique de test suivante :

.

6.3.1.5

.

$$T = \frac{\hat{L} - \#}{s_{\mathrm{P}}\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_r^2}{n_r}}}$$

.

et d'une distribution t_{n-r} de référence. Pour adapter cet énoncé aux hypothèses de type 6.3.1.1 à 6.3.1.3, il suffit de choisir les c_i et # correctement.

.

Mais la méthode de test d'hypothèses le plus souvent associée au modèle normal à un facteur ne convient pas aux hypothèses de type 6.3.1.4. La méthode la plus courante repose plutôt sur l'hypothèse selon laquelle les r moyennes sous-jacentes sont égales. En symboles, cela équivaut à :

.

6.3.1.6 $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_r$

.

Si on travaille à partir des hypothèses du modèle à un facteur, l'hypothèse 6.3.1.6 revient à affirmer que les r distributions sous-jacentes sont essentiellement les mêmes – autrement dit, « il n'y a aucune différence entre les

processus ».

.

L'hypothèse 6.3.1.6 peut être considérée comme l'égalité simultanée des $\frac{r(r-1)}{2}$ paires de moyennes – c'est-à-dire qu'elle est équivalente à l'affirmation selon laquelle, simultanément,

.

$$\mu_1 - \mu_2 = 0, \quad \mu_1 - \mu_3 = 0, \quad \ldots, \quad \mu_1 - \mu_r = 0$$
$$\mu_2 - \mu_3 = 0, \quad \ldots, \quad \text{et} \quad \mu_{r-1} - \mu_r = 0$$

.

Cet énoncé n'est pas sans rappeler ce qui a été dit sur les intervalles de confiance simultanés de la section précédente (à savoir la méthode de Tukey). D'ailleurs, l'une des façons de juger de la signification statistique d'un ensemble de données à r échantillons, en ayant pour référence l'hypothèse 6.3.1.6, est d'appliquer la méthode de Tukey d'estimation d'intervalles simultanés et de noter si tous les intervalles de différences de moyennes incluent la valeur 0. Si c'est le cas pour tous, la valeur p associée est supérieure à 1 moins le niveau de confiance simultané. Sinon, la valeur p associée est inférieure à 1 moins le niveau de confiance simultané. (Par exemple, si les intervalles simultanés de 95% incluent tous 0, alors aucune différence entre les moyennes n'est définitivement établie et la valeur p correspondante est supérieure à 0,05.)

.

Nous admettons privilégier l'estimation plutôt que les tests d'hypothèse. Conséquence : une tendance à déduire une approximation de la valeur p pour l'hypothèse 6.3.1.6 en utilisant la méthode de Tukey. Mais une méthode de test plus connue pour l'hypothèse 6.3.1.6 mérite également d'être abordée : l'analyse de la variance à un facteur.

.

Il semble étrange, à ce stade, qu'un test de moyennes porte un nom qui mette apparemment l'accent sur la variance. Ce terme se justifie par le fait que ce test est lié à une façon très pratique de considérer la répartition de la variabilité globale d'une variable de réponse. Il s'agit du test F de l'ANOVA à un facteur.

6.3.2 Test F de l'ANOVA à un facteur

La méthode standard de test de l'hypothèse 6.3.2.6

.

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

.

pour l'absence de différence entre r moyennes s'opposant à l'hypothèse
H_a : not H_0

.

est essentiellement basée sur la comparaison entre la mesure de la variabilité des moyennes d'échantillons et la variance d'échantillons pondérée, s_p^2. Certaines conventions de notation supplémentaires sont nécessaires pour pouvoir présenter pleinement cette méthode.

.

Dans la suite de cet ouvrage, il sera souvent plus pratique d'utiliser des notations globales pour les mesures de la partie 2 (moyennes et variances d'échantillons) appliquées aux données des études multi-échantillons, en laissant de côté le nombre total d'échantillons (r). La lettre n dénuée d'indice a déjà été utilisée pour remplacer $n_1 + n_2 + \cdots + n_r$, le nombre d'observations disponibles, en ignorant le nombre total d'échantillons (r). Cette convention sera désormais expressément étendue aux statistiques calculées à partir des n réponses. Pour plus de clarté, cela sera énoncé sous forme de définition.

.

DÉFINITION 6.3.2.1 Convention de notation pour les études multi-échantillons

Dans les études multi-échantillons, les symboles représentant des tailles ou des statistiques d'échantillons qui apparaîtront sans indice ni point seront considérées comme tenant compte de toutes les réponses disponibles, en combinant tous les échantillons.

.

Ainsi, n représentera le nombre total de points de données (même dans une étude à r échantillons), $\bar{y}$ la moyenne du grand échantillon de la réponse y, et s^2 une variance du grand échantillon calculée en combinant tous les échantillons.

.

Pour les besoins présents (construire une variable pour tester l'hypothèse 6.3.1.6), il faut utiliser $\bar{y}$, la moyenne du grand échantillon. Il est important de reconnaître que $\bar{y}$ et

.

6.3.2.1 Moyenne (non pondérée) des moyennes de r échantillons

$$\bar{y}_{\cdot} = \frac{1}{r}\sum_{i=1}^{r} \bar{y}_i$$

.

ne sont pas forcément identiques, à moins que tous les échantillons aient la même taille. En effet, lorsque la taille des échantillons varie, $\bar{y}$ est la moyenne arithmétique (non pondérée) des valeurs de données brutes y_{ij}, et la moyenne pondérée des moyennes d'échantillons $\bar{y}_i$. En contrepartie, $\bar{y}$ est la moyenne arithmétique (non pondérée) des moyennes d'échantillons $\bar{y}_i$, et la moyenne pondérée des valeurs de données brutes y_{ij}. Par exemple, dans le cas très simple où $r = 2, n_1 = 2$ et $n_2 = 3$,

.

$$\bar{y} = \frac{1}{5}(y_{11} + y_{12} + y_{21} + y_{22} + y_{23}) = \frac{2}{5}\bar{y}_1 + \frac{3}{5}\bar{y}_2$$

.

alors que

.

$$\bar{y}_{\cdot} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{6}y_{21} + \frac{1}{6}y_{22} + \frac{1}{6}y_{23}$$

.

et qu'en général, $\bar{y}$ et $\bar{y}$ ne sont pas identiques.

.

D'après l'hypothèse 6.3.1.6 ($\mu_1 = \mu_2 = \cdots = \mu_r, \bar{y}$ est une estimation naturelle de la moyenne commune. (Toutes les distributions sous-jacentes sont les mêmes; les données disponibles sont donc considérées, à juste titre, non pas comme r échantillons différents, mais plutôt comme un échantillon unique de taille n.) Les différences $\bar{y}_i - \bar{y}$ sont donc des indicateurs de différences potentielles parmi les μ_i. Il est pratique de résumer la taille de ces différences $\bar{y}_i - \bar{y}$ en prenant une sorte de somme de leurs carrés :

.

6.3.2.2 $$\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2$$

.

On peut considérer la statistique 6.3.2.2 soit comme une somme pondérée des quantités $(\bar{y}_i - \bar{y})^2$, soit comme une somme non pondérée, avec un terme de la somme pour chaque point de données brutes, donc n_i termes $(\bar{y}_i - \bar{y})^2$. La quantité 6.3.2.2 est une mesure de la variation inter-échantillons des données. Pour un ensemble donné de tailles d'échantillons, plus cette quantité est grande, plus la variation entre les moyennes d'échantillons $\bar{y}_i$ est grande.

.

Pour construire une variable à tester pour l'hypothèse 6.3.1.6, il suffit de diviser la mesure 6.3.2.2 par $(r-1)s_{\mathrm{p}}^2$. On obtient alors :

.

6.3.2.3 Statistique de test de l'ANOVA à un facteur pour l'égalité de r moyennes

$$F = \frac{\frac{1}{r-1}\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2}{s_{\mathrm{p}}^2}$$

.
Le fait est que si $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ est vraie, d'après les hypothèses du modèle à un facteur, cette statistique présente une distribution $F_{r-1,n-r}$. L'hypothèse d'égalité de r moyennes peut donc être testée à l'aide de l'équation 6.3.2.3 et d'une distribution $F_{r-1,n-r}$ de référence, où les grandes valeurs observées de F permettent d'infirmer H_0 et de confirmer H_a : pas H_0.
.

Exemple 6.3.2.1 Étude sur la compression du béton (suite)

Reprenons l'exemple de l'étude de résistance à la compression du béton d'Armstrong, Babb et Campen. Nous avions $\bar{y} = 3\ 693{,}6$. et les 8 moyennes d'échantillons $\bar{y}_i$ présentent diffèrent de cette valeur par les écarts donnés dans le tableau 6.3.2.1.

.
Puis, étant donné que pour tout i, $n_i = 3$, dans cette situation :

.

$$\begin{aligned}\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 =& 3(1,941.7)^2 + 3(2,059.7)^2 + \cdots \\ &+ 3(-2,142.3)^2 + 3(-1,302.9)^2\end{aligned}$$

.

$$= 47\ 360\ 780(\text{psi})^2$$

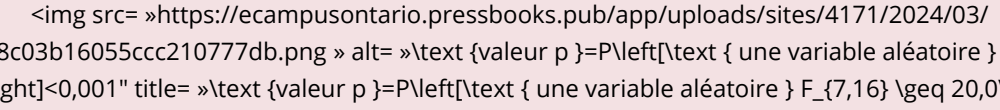
.
Pour pouvoir utiliser ce chiffre pour juger de la signification statistique, on standardise à l'aide de l'équation 6.3.2.3 afin d'obtenir la valeur observée de la statistique de test :

.

$$f = \frac{\frac{1}{8-1}(47\ 360\ 780)}{(581{,}6)^2} = 20{,}0$$

.
Il est très facile de vérifier dans les tables A1.5 (les tables F) que 20.0 est supérieur au quantile 0,999 de la distribution $F_{7,16}$. On a donc :

.



.
Ainsi, les données permettent largement de prouver que $\mu_1, \mu_2, \ldots, \mu_8$ ne sont pas toutes égales.

Moyenne de l'échantillon et des écarts-types par rapport à $\bar{y}$ dans l'étude de la résistance du béton

i, Formule	$\bar{y}_i$	$\bar{y}_i - \bar{y}$
1	5 635,3	1 941,7
2	5 753,3	2 059,7
3	4 527,3	833,7
4	3 442,3	−251,3
5	2 923,7	−769,9
6	3 324,7	−368,9
7	1 551,3	−2 142,3
8	2 390,7	−1 302,9

Tableau 6.3.2.1 Moyennes d'échantillon et écarts par rapport à \bar{y} dans l'étude de la résistance du béton

.

Pour des raisons purement pédagogiques, le test de l'ANOVA à un facteur a été présenté après les méthodes d'inférence axées sur les intervalles dans les études à r échantillons. Mais si elle doit être mise en pratique, la méthode de test d'hypothèse intervient typiquement chronologiquement avant l'estimation. En d'autres termes, le test de l'ANOVA peut permettre de déterminer en amont si les données disponibles conviennent pour différencier les moyennes de façon concluante, ou s'il faut plus de données.

6.3.3 Identité et tableau d'ANOVA à un facteur

La statistique de test de l'ANOVA est associée à une forte intuition quant à la répartition de la variabilité observée, en raison d'une identité algébrique énoncée ci-dessous sous la forme d'une proposition.

.

Proposition 6.3.3.1

.

Identité d'ANOVA à un facteur

.

Pour toute combinaison de n nombres y_{ij} :

.

6.3.3.1 $$(n-1)s^2 = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 + (n-r)s_{\mathrm{P}}^2$$

.

ou, de manière équivalente :

.

Deuxième version de l'identité d'ANOVA à un facteur :

.

6.3.3.2 $$\sum_{i,j}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

.

La proposition 6.3.3.1 permet de commencer à éclaircir le terme « analyse de variance ». Elle indique que la mesure globale de variabilité de la réponse y, à savoir

.

$$(n-1)s^2 = \sum_{i,j}(y_{ij} - \bar{y})^2$$

.

peut être divisée ou décomposée algébriquement en deux parties. La première,

.

$$\sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2$$

.

peut être considérée comme une mesure de la variation entre les échantillons ou les « traitements », et l'autre,

.

$$(n-r)s_{\mathrm{P}}^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

.
mesure la variation intra-échantillon (il s'agit en réalité de la somme des carrés de l'erreur résiduelle). Dans la statistique F 6.3.2.3, conçue pour tester $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_r$, le numérateur se rapporte à la première des deux, et le dénominateur à la deuxième. Utiliser la statistique F de l'ANOVA revient presque à analyser la variabilité brute de y.

.
Reconnaissant leur importance dans le calcul de la statistique F de l'ANOVA à un facteur et leur utilité en tant que statistiques descriptives à part entière, on attribue généralement un nom spécifique et une abréviation aux trois sommes (des carrés) figurant dans les équations 6.3.3.1 et 6.3.3.2, énoncés ci-dessous sous forme de définition :

.

DÉFINITION 6.3.3.1 Somme totale des carrés SCTot

Dans une étude multi-échantillons, la somme des carrés des différences entre les valeurs des données brutes et la moyenne du grand échantillon, $(n-1)s^2$, est appelée somme totale des carrés et notée SCTot.

DÉFINITION 6.3.3.2 Somme des carrés des traitements SCTr

Dans une étude multi-échantillons non structurée, la somme $\sum n_i(\bar{y}_i - \bar{y})^2$ est appelée somme des carrés des traitements et notée SCTr.

DÉFINITION 6.3.3.3 Somme des carrés d'erreur résiduelle SCE

Dans une étude multi-échantillons, la somme des carrés des résidus, $\sum(y - \hat{y})^2$ (qui équivaut à $(n-r)s_{\mathrm{p}}^2$ dans un cas non structuré) est appelée somme des carrés d'erreur résiduelle et notée SCE.

.

Dans la nouvelle notation présentée dans ces définitions, la proposition 6.3.3.1 énonce que dans un cadre multi-échantillons non structuré :

.

6.3.3.3 Troisième version de l'identité de l'ANOVA à un facteur

$$SSTot = SSTr + SSE$$

.

Créer une table d'analyse de la variance permet d'une part de faciliter l'organisation du calcul de la statistique F de l'équation 6.3.2.3 et, d'autre part, de consolider et d'élargir l'intuition au sujet de la variance fournie par les équations 6.3.3.1, 6.3.3.2 et 6.3.3.3. Il existe de nombreuses formes de tables d'analyse de la variance correspondant à des analyses multi-échantillons diverses et variées. La plus judicieuse à utiliser dans le cas présent est celle représentée sous forme symbolique dans le tableau 6.3.3.1.

.

Les intitulés de colonnes du tableau 6.3.3.1 sont Source (de la variation), Sum des Carrés *SS* (de la source), degrés de liberté *df* (de la source), carré de la moyenne *MS* (de la source), et F (pour le test d'hypothèse de la contribution de la source dans la variabilité globale observée). Dans la colonne Source du tableau, les entrées sont Traitements, Erreur et Total. Mais le terme « traitements » peut parfois être remplacé par « inter (échantillons) », et « Erreur » par « intra (échantillons) » ou « résiduel ». La somme des deux premières entrées de la colonne SC (*SS*) doit correspondre à la troisième, comme indiqué par l'équation 6.3.3.3. De même, la somme des degrés de liberté pour les traitements et l'erreur donne le nombre total de degrés de liberté, $(n-1)$. À noter que les entrées de la colonne df sont respectivement liées au numérateur et au dénominateur de la statistique de test dans l'équation 6.3.2.3. Les rapports entre les sommes des carrés et les degrés de liberté sont appelés carrés de moyennes; ici, le carré de la moyenne pour les traitements (MSTr) et le carré de la moyenne pour l'erreur (MSE). Dans le cas présent, il faut vérifier que $MSE = s_{\text{P}}^2$ et que $MSTr$ est le numérateur de la statistique F de l'équation 6.3.2.3. Le rapport apparaissant dans la colonne F est donc la valeur observée de F pour le test $\text{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_r$.

.

Forme générale de la table d'analyse de la variance à un facteur

Table d'analyse de la variance (pour tester H_0: $\mu_1 = \mu_2 = \cdots = \mu_r$)

Source	*SS*	*df*	*MS*	*F*
Traitements	*SSTr*	$r-1$	$SSTr/(r-1)$	$MSTr/MSE$
Erreur	*SSE*	$n-r$	$SSE/(n-r)$	
Total	*SSTot*	$n-1$		

Tableau 6.3.3.1 Forme générale du tableau d'analyse de la variance à un facteur (ANOVA)

Exemple 6.3.3.1 Étude sur la résistance du béton (suite)

Revenons encore une fois à l'étude de résistance du béton. En retournant voir les données brutes du tableau 6.1.1.1, on constate que $\bar{y} = 3\ 693{,}6$, donc

.

$$\begin{aligned}\text{SSTot} =& (n-1)s^2 \\ =& (5\ 800 - 3\ 693{,}6)^2 + (4\ 598 - 3\ 693{,}6)^2 + (6\ 508 - 3\ 693{,}6)^2 \\ &+ \cdots + (2\ 631 - 3\ 693{,}6)^2 + (2\ 490 - 3\ 693{,}6)^2 \\ =& 52\ 772\ 190(\text{psi})^2\end{aligned}$$

.

En outre, comme dans la section 6.1.1.1, $s_{\text{P}}^2 = 338\ 213{,}1(\text{psi})^2$ et $n - r = 16$, on a donc :

.

$$SSE = (n-r)s_{\text{P}}^2 = 5\ 411\ 410(\text{psi})^2$$

.

Et d'après ce qu'on a pu voir précédemment dans la présente section :

.

$$SSTr = \sum_{i=1}^{r} n_i(\bar{y}_i - \bar{y})^2 = 47,360,780$$

.

On insère ces résultats et les valeurs de degré de liberté adéquates dans la forme générale de la table d'analyse de la variance à un facteur, créant ainsi la table de l'étude de résistance à la compression du béton (tableau 6.3.3.2).

.

À noter que, comme promis par le principe d'identité d'ANOVA à un facteur, la somme des carrés des traitements et la somme des carrés de l'erreur résiduelle égalent la somme totale des carrés. De plus, le tableau 6.3.3.2 constitue une synthèse pratique du processus de test, permettant de trouver en un coup d'œil la valeur observée de F, les degrés de liberté et les valeurs de $s_{\text{P}}^2 = MSE$.

Table d'analyse de la variance à un facteur pour l'étude de résistance du béton

Table d'analyse de la variance (pour tester $H_0: \mu_1 = \mu_2 = \cdots = \mu_8$)				
Source	*SS*	*df*	*MS*	*F*
Traitements	47 360 780	7	6 765 826	20,0
Erreur	5 411 410	16	338 213	
Total	52 772 190	23		

Tableau 6.3.3.2 Tableau d'analyse de la variance (ANOVA) à un facteur pour l'étude de résistance du béton

6.3.4 Calcul de l'ANOVA avec Python

Les calculs présentés dans la partie 6 ne sont en aucun cas impossibles à effectuer « à la main », mais la façon la plus précise de le faire est d'utiliser un logiciel statistique.

En utilisant les Jupyter Notebooks et Python, il est possible de consulter les figures, le tableau d'ANOVA et le résultat d'une analyse à un facteur des données de résistance à la compression, qui illustrent la majorité des points abordés dans la partie 6. **Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur les tests d'hypothèses.** Vous pouvez les trouver dans la section « How do I do X in Python? ». Le fichier « ANOVA » sera particulièrement utile.

Pour une discussion interactive étape par étape de cet exemple et des résultats, consulter l'exemple sur l'ANOVA de la partie 6 sur GitHub à partir du site Binder.

Ou rendez-vous sur le site GitHub : https://github.com/Statistcial-Methods-for-Engineering/Special-GitHub-Site-Part-6-ANOVA-and-Compression-Strength-Example, *Special GitHub Site Part 6: ANOVA and Compression Strength Example*, puis cliquez sur le lien Binder pour lancer le tutoriel interactif.

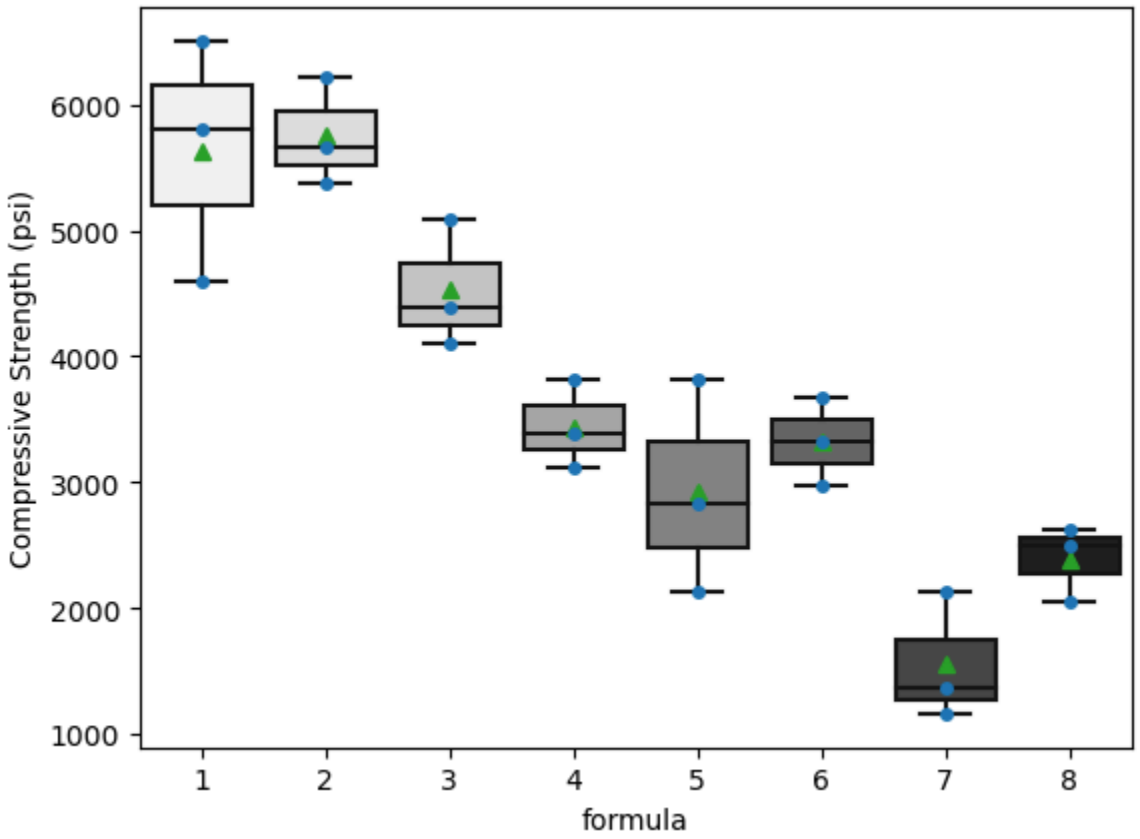


Figure 6.3.4.1 Diagramme en boîtes montrant la résistance de compression de huit formules

	df	sum_sq	mean_sq	F	PR(>F)
formula	7.0	4.736078e+07	6.765826e+06	20.004625	8.550775e-07
Residual	16.0	5.411409e+06	3.382131e+05	NaN	NaN

Tableau 6.3.4.1 Tableau ANOVA pour l'exemple de résistance à la compression

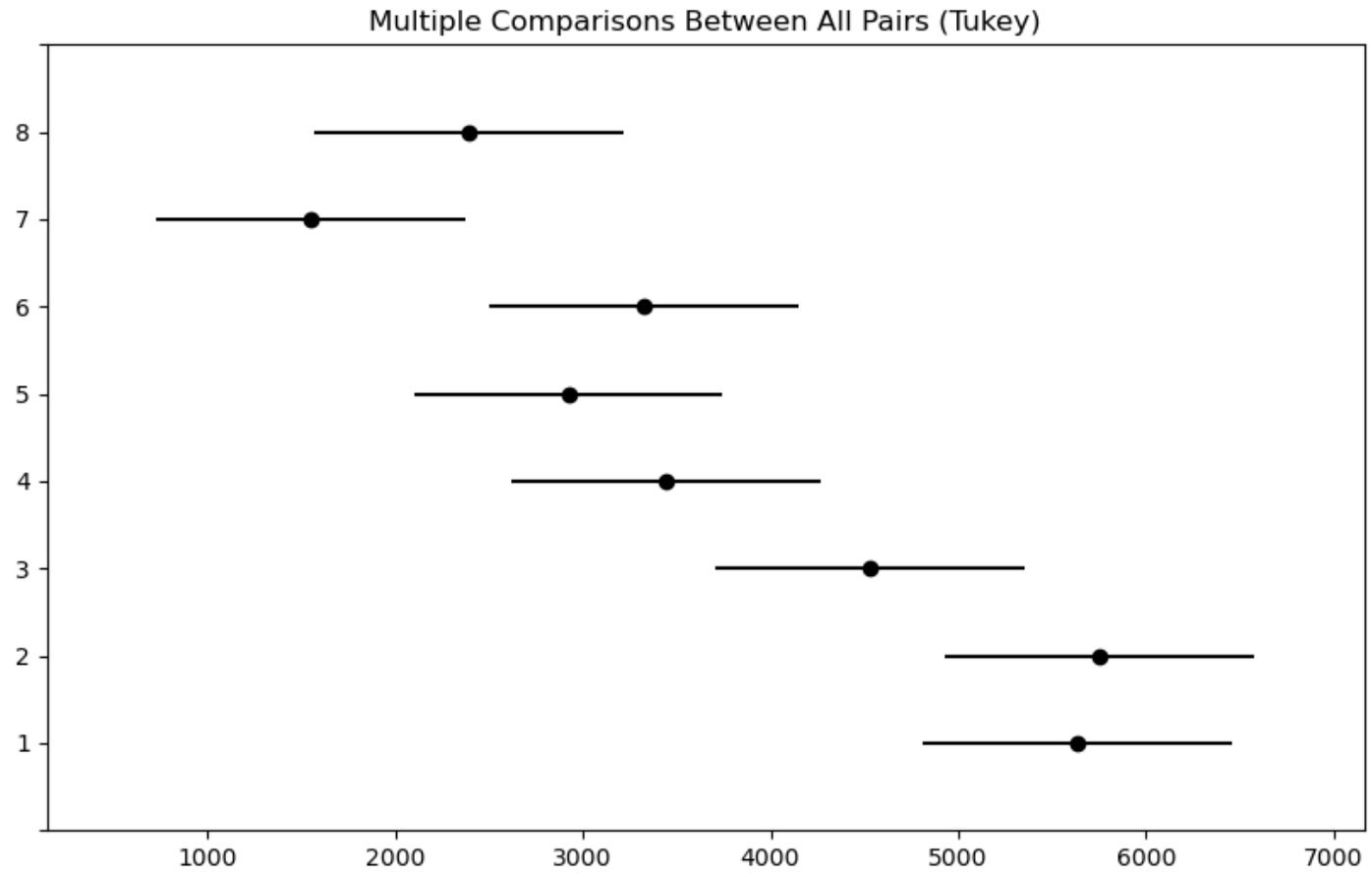


Figure 6.3.4.2 Comparaison multiple de paires (méthode de Tukey) pour comparaisons simultanées

Et comme toujours, ce Jupyter Notebook est consultable sur le site de cours GitHub, avec des tutoriels et d'autres exemples, dans la partie 6 ANOVA : IntroEngStatsMethods_GitHub Site.

7.0.1 Moindres carrés et analyse de régression linéaire simple – Introduction

Cette partie part d'un nouveau principe : tenir compte de plusieurs variables à la fois. Les outils d'intervalles de confiance et de visualisation présentés dans les sections précédentes seront tout de même utilisés pour interpréter les modèles des moindres carrés analytiquement et visuellement.

Les sections qui suivent, qui portent sur les plans d'expériences et leur analyse, reposent sur le modèle des moindres carrés, abordé ici.

La matière de cette section sera utile chaque fois qu'il faudra interpréter et quantifier un rapport entre deux variables ou plus. Exemples de quantifications de ce type pouvant être étudiées :

- Collègue : Quel rapport y a-t-il entre le rendement de notre fermentation d'acide lactique par lots et la pureté du saccharose?
- Vous : On peut prédire ce rendement à partir de la pureté du saccharose avec une erreur de plus ou moins 8 %.
- Collègue : Et qu'en est-il du rapport entre le rendement et la pureté du glucose?
- Vous : Sur l'ensemble de nos données historiques, on n'en constate aucun.

.

- *Ingénieur.e 1* : L'équation théorique de l'indice de fusion est liée de manière non linéaire à la viscosité.
- *Ingénieur.e 2* : Le modèle linéaire ne montre pas cela, mais la capacité de prédiction du modèle s'améliore légèrement si on utilise une transformation non linéaire avec le modèle des moindres carrés.

.

- *Responsable RH* : On utilise un modèle de régression des moindres carrés pour calculer la progression salariale du personnel. Ce modèle dépend du niveau de formation et du nombre d'années d'expérience. Que signifient les coefficients de modèle?

7.0.2 Sources

Cette première version de la partie 7 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte, et il y a des liens pour y accéder tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn et sont partagées sous licence CC BY-SA 4.0.

7.1.0 Introduction aux moindres carrés : description de la relation entre des données quantitatives à deux variables

Les données à deux variables résultent souvent du fait qu'une variable quantitative expérimentale *x* dépend de plusieurs paramètres, ce qui produit plusieurs d'échantillons d'une variable de réponse *y*. Aux fins de la réduction des données, de l'interpolation, de l'extrapolation limitée, ou encore de l'optimisation ou de l'ajustement du processus, il est extrêmement utile d'avoir une équation reliant *y* à *x*. Une équation linéaire de la forme

EXPRESSION 7.1.0.1

$$y \approx \beta_0 + \beta_1 x$$

qui relie y à x est l'équation la plus simple potentiellement utile à envisager après avoir fait un nuage de points (x, y) simple.

Dans ce chapitre, la méthode des moindres carrés est utilisée pour tracer une droite de régression correspondant aux données (x, y). L'exactitude de cette régression est évaluée à l'aide de la corrélation de l'échantillon et du coefficient de détermination. La représentation graphique des résidus est présentée comme une méthode importante pour approfondir l'étude des problèmes éventuels liés à la droite de régression. Une discussion sur certaines mises en garde pratiques et sur l'utilisation de logiciels statistiques pour les droites de régression suit.

7.1.1 : Application de la méthode des moindres carrés

Exemple 7.1.1.1 : Pression de pressage et densité des échantillons d'un composé céramique

Benson, Locher et Watkins ont étudié les effets de différentes pressions de pressage sur la densité d'échantillons cylindriques fabriqués par pressage à sec d'un composé céramique. Un mélange d'Al_2O_3, d'alcool polyvinylique et d'eau a été préparé, séché pendant une nuit, broyé et tamisé pour obtenir des grains d'une taille de 100 mesh. Ceux-ci ont été pressés dans des cylindres à des pressions allant de 2 000 psi à 10 000 psi, puis la densité des cylindres a été calculée. Les données obtenues sont présentées dans le tableau 7.1.1.1, et un nuage de points simple de ces données est présenté dans la figure 7.1.1.1.

x, Pression (psi)	y, Densité (g/cc)
2 000	2,486
2 000	2,479
2 000	2,472
4 000	2,558
4 000	2,570
4 000	2,580
6 000	2,646
6 000	2,657
6 000	2,653
8 000	2,724
8 000	2,774
8 000	2,808
10 000	2,861
10 000	2,879
10 000	2,858

Tableau 7.1.1.1 : Pressions de pressage et densité de l'échantillon résultant

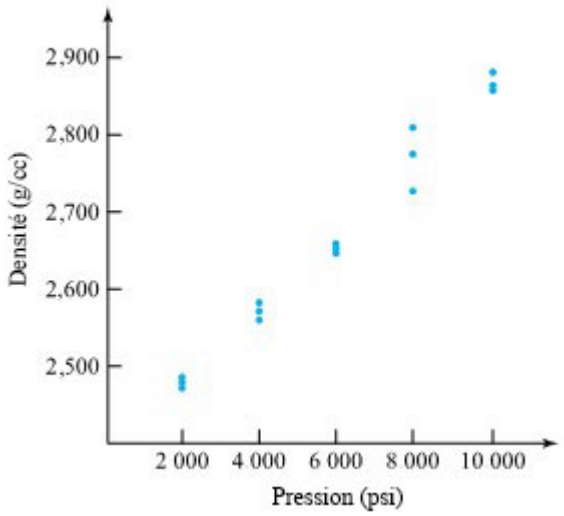


Figure 7.1.1.1 : Nuage de points de la densité en fonction de la pression de pressage

Il est très facile d'imaginer une ligne droite passant par les points de la figure 7.1.1.1 . Cette droite pourrait alors être utilisée pour illustrer comment la densité varie en fonction de la pression. La méthode des moindres carrés permet de choisir la « meilleure » droite pour décrire les données.

DÉFINITION Méthode des moindres carrés

EXPRESSION 7.1.1.1

La méthode des moindres carrés pour ajuster une équation pour y à un ensemble de données de n points consister à trouver les paramètres de l'équation qui minimisent la somme

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

où $y_1, y_2, \ldots, y_n$ sont les réponses observées et $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ sont les réponses correspondantes prédites (ou ajustées) par l'équation.

Selon la définition 7.1.1.1, dans le cadre de l'ajustement d'une droite aux données (x, y), il faut choisir une pente et une ordonnée à l'origine de manière à minimiser la somme des carrés des distances verticales entre les points (x, y) et la droite en question. Cette notion est illustrée de façon générique à la figure 7.1.1.2 pour un ensemble fictif de cinq données. (C'est la somme des carrés des cinq différences indiquées qu'il faut minimiser.)

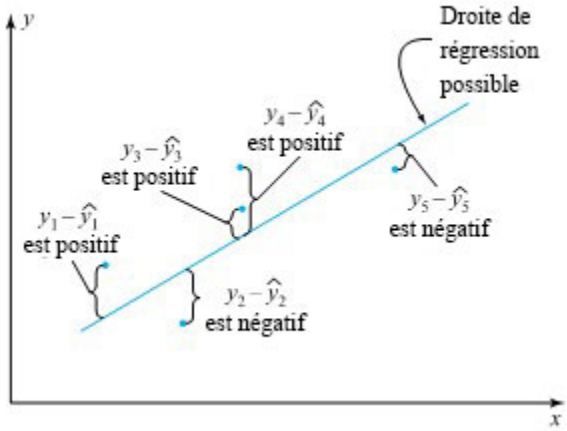


Figure 7.1.1.2 Cinq points de données (x, y) et une possible droite de régression.

Selon la forme de l'équation (7.1.0.1), l'équation pour la régression d'une droite est

$$\hat{y} = \beta_0 + \beta_1 x$$

Par conséquent, l'expression à minimiser en choisissant la pente (β_1), et l'ordonnée à l'origine (β_0) est la suivante :

7.1.1.2 $$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

La minimisation de la fonction de deux variables $S(\beta_0, \beta_1)$ est un exercice de calcul différentiel. On met les dérivées partielles de S par rapport à β_0 et à β_1 égales à zéro, puis on résout le système de deux équations pour obtenir β_0 et β_1. Les équations ainsi obtenues sont

7.1.1.3 $$n\beta_0 + \left(\sum_{i=1}^{n} x_i\right) \beta_1 = \sum_{i=1}^{n} y_i$$

et

7.1.1.4
$$\left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 = \sum_{i=1}^{n} x_i y_i$$

Pour des raisons obscures, les équations 7.1.1.3 et 7.1.1.4 sont parfois appelées équations normales (dans le sens de « perpendiculaire ») pour la régression d'une droite. Ce sont deux équations linéaires à deux inconnues qu'on peut résoudre assez facilement pour β_0 et β_1 (à condition qu'il y ait au moins deux x_i différents dans l'ensemble de données). En résolvant les équations 7.1.1.3 et 7.1.1.4, on obtient les valeurs de Formula does not parse et β_1 suivantes :

Pente de la droite des moindres carrés, β_1 **7.1.1.5**
$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

et

Ordonnée à l'origine de la droite des moindres carrés, β_0 **7.1.1.6**
$$b_0 = \bar{y} - b_1 \bar{x}$$

Soulignons ici la convention de notation. La pente et l'ordonnée à l'origine qui minimisent $S(\beta_0, \beta_1)$ sont désignées non pas par les paramètres β, mais par les paramètres b_1 et b_0.

Remarque concernant l'expression (7.1.1.5) : la pratique assez courante qui a été suivie (et l'abus de notation de la somme) consiste à ne pas indiquer la variable de sommation (i) ni son intervalle (de 1 à n).

Exemple 7.1.1.2 (suite)

Il est possible de vérifier que les données du tableau 7.1.1.1 donnent les résultats sommaires suivants :

$$\begin{aligned}
\sum x_i &= 2,000 + 2,000 + \cdots + 10,000 = 90,000,\\
\text{so } \bar{x} &= \frac{90,000}{15} = 6,000\\
\sum (x_i - \bar{x})^2 &= (2,000 - 6,000)^2 + (2,000 - 6,000)^2 + \cdots +\\
\sum y_i &= 2.486 + 2.479 + \cdots + 2.858 = 40.005,\\
\text{so } \bar{y} &= \frac{40.005}{15} = 2.667\\
\sum (y_i - \bar{y})^2 &= (2.486 - 2.667)^2 + (2.479 - 2.667)^2 + \cdots +\\
&\quad (2.858 - 2.667)^2 = .289366\\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= (2,000 - 6,000)(2.486 - 2.667) + \cdots +\\
&\quad (10,000 - 6,000)(2.858 - 2.667) = 5,840
\end{aligned}$$

Les équations 7.1.1.5 et 7.1.1.6 donnent alors la pente et l'ordonnée à l'origine des moindres carrés, b_1 et b_0 :

$$b_1 = \frac{5\ 840}{120\ 000\ 000} = 0{,}0000486\overline{(\text{g/cc})/\text{psi}} \qquad \}) / \mathrm{psi}$$

et

$$b_0 = 2{,}667 - (0{,}0000486\overline{)}(6,000) = 2{,}375\ \text{g/cc}$$

La figure 7.1.1.3 montre la droite des moindres carrés

$$\hat{y} = 2.375 + .0000487x$$

tracée sur le nuage de points (x, y) tiré du tableau 7.1.1.1.

Interprétation de la pente de la droite des moindres carrés

Il convient de noter que la pente sur ce graphique, b_1 ≈ 0,0000487 (g/cm^3)/psi, correspond physiquement à l'augmentation (approximative) de y (densité) qui accompagne une augmentation d'une unité (1 psi) de x (pression).

Interprétation de l'ordonnée à l'origine et extrapolation prudente

L'ordonnée à l'origine du tracé, b_0 = 2,375 g/cm^3, positionne la droite verticalement et est la valeur à laquelle la droite coupe l'axe des y. Cependant, il ne faut probablement pas l'interpréter comme la densité qui correspondrait à une pression de x = 0 psi. Le fait est que la relation raisonnablement linéaire constatée pour des pressions comprises entre 2 000 et 10 000 psi pourrait bien ne pas s'appliquer à des pressions plus grandes ou plus petites Considérer b_0 comme la densité obtenue lorsque la pression est 0 revient à extrapoler en dehors de la plage de données utilisée pour obtenir la droite de régression, ce qu'il faut toujours faire avec une extrême prudence.

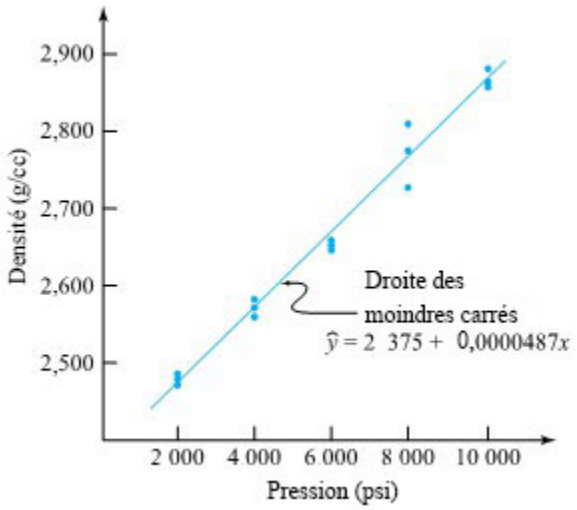


Figure 7.1.1.3 Nuage de points des données de pression/densité et droite des moindres carrés.

Comme l'indique la définition 7.1.1.1, la valeur de y sur la droite des moindres carrés correspondant à un x donné peut être appelée une valeur ajustée ou prédite. Elle peut être utilisée pour représenter le comportement probable de y pour ce x.

Exemple 7.1.1.3 (suite)

Trouvons la densité type correspondant à une pression de 4 000 psi et à une pression de 5 000 psi. Tout d'abord, si l'on considère que x = 4 000, une manière simple de représenter un y type est de
noter que pour les trois points de données à x = 4 000,

$$\bar{y} = \frac{1}{3}(2{,}558 + 2{,}570 + 2{,}580) = 2{,}5693 \text{ g/cc}$$

et donc d'utiliser cette valeur comme valeur représentative. Mais en supposant que y soit effectivement approximativement linéairement proportionnelle à x, la valeur ajustée

$$\hat{y} = 2{,}375 + 0{,}0000486(4\ 000) = 2{,}5697\ \mathrm{g/cc}$$

pourrait être encore plus représentative de la densité moyenne pour une pression de 4 000 psi.

Interpolation

En examinant la situation, on constate qu'il n'y a pas de données pour x = 5 000 psi. La seule chose que l'on puisse faire pour représenter la densité à cette pression est de se demander si l'interpolation est raisonnable d'un point de vue physique. Si c'est le cas, la valeur ajustée

$$\hat{y} = 2.375 + .0000486(5,000) = 2.6183\ \mathrm{g/cc}$$

peut être utilisée pour représenter la densité à une pression de 5 000 psi.

7.1.2 Corrélation d'échantillon et coefficient de détermination

CORRÉLATION

Visuellement, la droite des moindres carrés de la figure 7.1.1.3 semble bien correspondre aux points indiqués. Cependant, il serait utile de disposer de méthodes permettant de quantifier la qualité de cette régression. L'une de ces méthodes est la corrélation d'échantillon.

DÉFINITION Corrélation (linéaire) d'échantillon
EXPRESSION 7.1.2.1

La corrélation (linéaire) d'échantillon entre x et y dans un échantillon de n paires de données (x_i, y_i) est

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Interprétation de la corrélation d'échantillon

La corrélation d'échantillon se situe toujours entre -1 et 1 inclusivement. Elle est égale à -1 ou à 1 uniquement lorsque tous les points de données (x , y) se situent sur une même ligne droite. En comparant les équations 7.1.1.5 et 7.1.2.1, on constate que $r = b_1 \left(\sum (x_i - \bar{x})^2 / \sum (y_i - \bar{y})^2 \right)^{1/2}$, ce qui indique que b_1 *et* r *ont le même signe.* Ainsi, une corrélation d'échantillon de -1 signifie que y diminue de façon linéaire lorsque x augmente, tandis qu'une corrélation d'échantillon de +1 signifie que y augmente de façon linéaire lorsque x augmente.

Les ensembles de données réelles sont rarement en corrélation parfaite (+1 ou -1). La valeur de r est généralement comprise entre -1 et 1. Mais en se basant sur les faits relatifs à son comportement, on considère r comme une mesure de la force d'une relation linéaire apparente : un r proche de +1 ou -1 est interprété comme indiquant une relation linéaire relativement forte, et un r proche de 0, comme indiquant une absence de relation linéaire. Le signe de r indique si y tend à augmenter ou à diminuer lorsque x augmente.

Exemple 7.1.2.2 (suite)

Pour les données relatives à la pression et à la densité, les données récapitulatives de l'exemple donnent :

$$r = \frac{5\ 840}{\sqrt{(120\ 000\ 000)(0{,}289366)}} = 0{,}9911$$

Cette valeur de r est proche de +1 et indique clairement la forte relation linéaire positive mise en évidence dans les figures 7.1.1.1 et 7.1.1.3.

COEFFICIENT DE DÉTERMINATION

DÉFINITION Coefficient de détermination
EXPRESSION 7.1.2.2

Pour une équation de régression d'un ensemble de données de n points obtenue par la méthode des moindres carrés produisant des valeurs ajustées y, le coefficient de détermination vaut :

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Interprétation de R^2

R^2 peut être interprétée comme *la fraction de la variation brute de y prise en compte par l'équation ajustée,* à condition que l'équation ajustée comprenne une constante, $\sum (y_i - \bar{y})^2 \geq \sum (y_i - \hat{y}_i)^2$. De plus, $\sum (y_i - \bar{y})^2$ est une mesure de la variabilité brute de y, tandis que $\sum (y_i - \hat{y}_i)^2$ est une mesure de la variation de y restante après la régression de l'équation. La différence non négative $\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$ est donc une mesure de la variabilité de y prise en compte dans le processus de régression. R^2 exprime cette différence sous forme de fraction (de la variation brute totale).

Exemple 7.1.2.2 (suite)

En utilisant la droite de régression, on peut trouver les valeurs $\hat{y}$ pour tous les n = 15 points de données de l'ensemble de données d'origine. Ces valeurs sont indiquées dans le tableau 7.1.2.1.

x, Pression	$\hat{y}$, Densité ajustée
2 000	2,4723
4 000	2,5697
6 000	2,6670
8 000	2,7643
10 000	2,8617

Tableau 7.1.2.1 Valeurs de densité ajustées

Ensuite, en se reportant à nouveau au tableau 7.1.1.1,

$$\begin{aligned}\sum (y_i - \hat{y}_i)^2 =& (2{,}486 - 2{,}4723)^2 + (2{,}479 - 2{,}4723)^2 + (2{,}472 - 2{,}4723)^2 \\ &+ (2{,}558 - 2{,}5697)^2 + \cdots + (2{,}879 - 2{,}8617)^2 \\ &+ (2{,}858 - 2{,}8617)^2 \\ =& \, 0{,}005153\end{aligned}$$

De plus, puisque $\sum (y_i - \bar{y})^2 = 0,289366$, l'équation 7.1.2.2 donne :

$$R^2 = \frac{0,289366 - 0,005153}{0,289366} = 0,9822$$

Ainsi, la droite de régression représente plus de 98 % de la variabilité brute de la densité, réduisant la variation « inexpliquée » de 0,289366 à 0,005153.

R^2 en tant que corrélation quadratique

Le coefficient de détermination a une deuxième interprétation utile. Pour les équations dont les paramètres sont linéaires (qui sont les seules pris en compte ici et qui seront abordées en détail ultérieurement), R^2 s'avère être une corrélation quadratique entre les valeurs observées y_i et les valeurs ajustées $\hat{y}_i$. (Dans la régression linéaire –le cas qui nous intéresse en ce moment –, les valeurs $\hat{y}_i$ sont parfaitement corrélées avec les valeurs x_i, R^2 est donc également la corrélation quadratique entre les valeurs y_i et x_i.)

Exemple 7.1.2.3 (suite)

Pour les données relatives à la pression et à la densité, la corrélation entre x et y est la suivante :

$$r = 0,9911$$

Puisque $\hat{y}$ est parfaitement corrélé avec x, c'est aussi la corrélation entre $\hat{y}$ et y. Notons également que

$$r^2 = (\ 0,9911)^2 = \ 0,9822 = R^2$$

R^2 est bien la corrélation d'échantillon quadratique entre y et $\hat{y}$.

7.1.3 Calcul et utilisation des résidus

Lorsqu'on ajuste une équation à un ensemble de données, on espère que l'équation extrait le message principal des données, ne laissant derrière elle que la variation de y qui est ininterprétable (non prédite par l'équation ajustée). En d'autres mots, on espère que les y_i ressembleront aux $\hat{y}_i$, sauf pour de petites fluctuations qui ne peuvent être expliquées que par des variations aléatoires. Une façon d'évaluer si ce point de vue est raisonnable consiste à calculer et à tracer les **résidus**.

DÉFINITION Résidus
EXPRESSION 7.1.3.1

Si l'ajustement d'une équation ou d'un modèle à un ensemble de données comportant des réponses $y_1, y_2, \ldots, y_n$ donne des valeurs ajustées $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, les résidus correspondants ont alors les valeurs suivantes :

$$e_i = y_i - \hat{y}_i$$

Si une équation ajustée raconte toute l'histoire d'un ensemble de données, alors ses résidus devraient être dépourvus de toute tendance. Ainsi, lorsqu'ils sont tracés en fonction de l'ordre chronologique d'observation, des valeurs des variables expérimentales, des valeurs ajustées ou de toute autre valeur raisonnable, les tracés doivent être dispersés de manière aléatoire. Lorsque ce n'est pas le cas, les tendances peuvent elles-mêmes suggérer ce qui n'a pas été pris en compte lors de l'ajustement ou la manière dont la synthèse des données pourrait être améliorée.

Exemple 7.1.3.1 Résistance à la compression de cylindres en béton de cendres volantes en fonction de la quantité d'additif de phosphate d'ammonium

Pour illustrer de façon exagérée le point précédent, examinons l'ajustement simpliste d'une droite sur certaines données réalisé par B. Roth, qui a étudié la résistance à la compression de cylindres en béton de cendres volantes. Ces cylindres ont été fabriqués en utilisant des quantités variables de phosphate d'ammonium comme additif. Une partie de ses données est présentée dans le tableau 7.1.3.1. Les valeurs de phosphate d'ammonium sont exprimées en pourcentage du poids de la quantité de cendres volantes utilisée.

Concentrations en additifs et résistances à la compression des cylindres de cendres volantes

x, phosphate d'ammonium (%)	y, résistance à la compression (psi)	x, phosphate d'ammonium (%)	y, résistance à la compression (psi)
0	1221	3	1609
0	1207	3	1627
0	1187	3	1642
1	1555	4	1451
1	1562	4	1472
1	1575	4	1465
2	1827	5	1321
2	1839	5	1289
2	1802	5	1292

Tableau 7.1.3.1. Concentrations en additif et résistances à la compression des cylindres en béton de cendres volantes

En utilisant les formules 7.1.1.5 et 7.1.1.6, il est possible de montrer que la droite des moindres carrés qui passe par les points (x, y) du tableau 7.1.3.1 est :

7.1.3.2

$$\hat{y} = 1498{,}4 - 0{,}6381x$$

Une substitution directe dans l'équation 7.1.3.2 donne les valeurs $\hat{y}_i$ et les résidus $e_i = y_i - \hat{y}_i$, comme indiqué dans le tableau 7.1.3.2. La figure 7.1.3.1 montre les résidus de cette régression linéaire en fonction de x.

x	y	$\hat{y}$	$e = y - \hat{y}$	x	y	$\hat{y}$	$e = y - \hat{y}$
0	1221	1498,4	−277,4	3	1609	1496,5	112,5
0	1207	1498,4	−291,4	3	1627	1496,5	130,5
0	1187	1498,4	−311,4	3	1642	1496,5	145,5
1	1555	1497,8	57,2	4	1451	1495,8	−44,8
1	1562	1497,8	64,2	4	1472	1495,8	−23,8
1	1575	1497,8	77,2	4	1465	1495,8	−30,8
2	1827	1497,2	329,8	5	1321	1495,2	−174,2
2	1839	1497,2	341,8	5	1289	1495,2	−206,2
2	1802	1497,2	304,8	5	1292	1495,2	−203,2

Tableau 7.1.3.2 Résidus d'un ajustement d'une droite aux données de cendres volantes

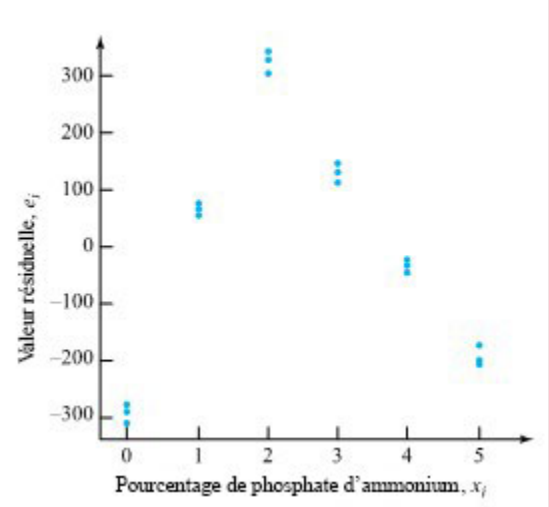


Figure 7.1.3.1 montre les résidus de la régression linéaire des données de cendres volantes en fonction de x.

Le schéma curviligne de la figure 7.1.3.1, qui se caractérise par un mouvement de croissance et de décroissance, n'est pas typique d'une dispersion aléatoire. Quelque chose a été omis dans la droite de régression des données de Roth. La figure 7.1.3.2 est un simple nuage de points des données de Roth. (Dans la pratique, on devrait faire ce genre de graphique avant d'ajuster une courbe aux données.) Ce nuage de points montre clairement que la relation entre la quantité de phosphate d'ammonium et la résistance à la compression est non linéaire. En fait, une fonction quadratique serait beaucoup plus appropriée pour ajuster les données du tableau 7.1.3.1.

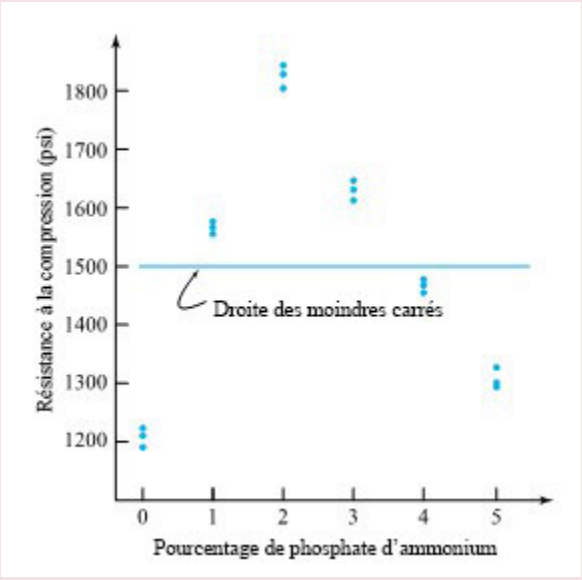


Figure 7.1.3.2 Nuage de points des données de cendres volantes.

Interprétation des tendances des tracés résiduels

La figure 7.1.3.3 présente quelques tendances qui peuvent être observées dans les tracés des résidus en fonction de diverses variables. Le graphique 1 de la figure 7.1.3.3 montre une tendance du tracé des résidus en fonction de l'ordre chronologique d'observation. Cette tendance suggère qu'une variable changeant dans le temps agit sur y et n'a pas été prise en compte dans les valeurs d'ajustement. Par exemple, la dérive instrumentale (lorsqu'un instrument affiche une valeur plus élevée à la fin d'une analyse qu'au début) pourrait produire une tendance comme celle du graphique 1. Le graphique 2 montre un modèle en forme d'éventail sur un tracé des résidus par rapport aux valeurs ajustées. Une telle tendance indique que les valeurs élevées sont ajustées (et très probablement produites ou mesurées) de manière moins uniforme que les valeurs faibles. Le graphique 3 montre des résidus correspondant à des observations faites par l'ingénieur.e 1 qui sont dans l'ensemble plus faibles que celles faites par l'ingénieur.e 2. On peut en déduire que le travail de l'ingénieur.e 1 est plus précis que celui de l'ingénieur.e 2.

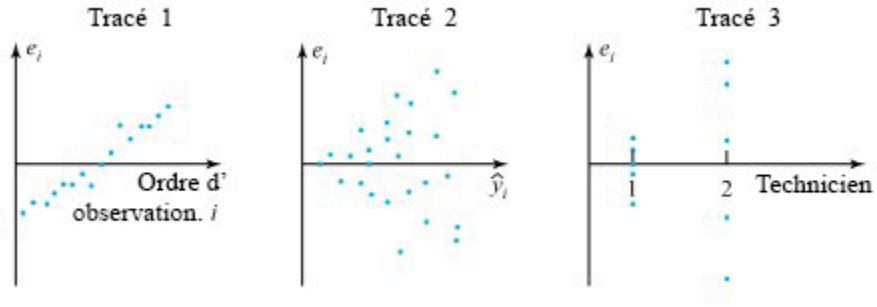


Figure 7.1.3.3 Tendances dans les graphiques de résidus.

Tracé normal des résidus

Une autre façon utile de tracer les résidus est de les représenter sous forme de graphique normal. L'idée est que la forme de la distribution normale est caractéristique de la variation aléatoire et que le tracé normal des résidus est un moyen de vérifier si cette forme de distribution s'applique à ce qui reste des données après l'ajustement d'une équation ou d'un modèle.

Exemple 7.1.3.2 (suite)

Le tableau 7.1.3.3 donne les résidus pour l'ajustement d'une droite aux données relatives à la pression et à la densité. Les résidus e_i ont été traités comme un échantillon de 15 nombres et ont été tracés de façon normale (en utilisant les méthodes que nous avons présentées précédemment) pour donner la figure 7.1.3.4.

La partie centrale du graphique de la figure 7.1.3.4 est assez linéaire, indiquant une distribution des résidus généralement en forme de cloche. Mais le point tracé correspondant au résidu le plus élevé, et probablement celui correspondant au résidu le plus faible, ne suivent pas le modèle linéaire établi par les autres. Comparés aux autres, ces résidus semblent plutôt importants.

Le tableau 7.1.3.3 et le nuage de points de la figure 7.1.1.3 montrent que ces résidus élevés proviennent tous deux de la pression de 8 000 psi. De plus, l'écart pour les trois densités à cette valeur de pression semble en effet considérablement plus important que pour les autres valeurs de pression. Le tracé normal laisse supposer que la tendance de la variation à 8 000 psi est véritablement différente de celle observée aux autres pressions. Il est possible qu'un mécanisme physique de compression différent ait agi à 8 000 psi par rapport aux autres pressions. Cependant, il est plus probable qu'il y ait eu

un problème de technique de laboratoire, d'enregistrement ou de matériel d'essai lorsque les essais à 8 000 psi ont été effectués.

Quoi qu'il en soit, le tracé normal des résidus permet de mettre en évidence une particularité des données du tableau 7.1.1.1 qui mérite d'être approfondie et, peut-être, de faire l'objet d'une nouvelle collecte de données.

x, Pression	y, Densité	$\hat{y}$	$e = y - \hat{y}$
2 000	2,486	2,4723	0,0137
2 000	2,479	2,4723	0,0067
2 000	2,472	2,4723	−0,0003
4 000	2,558	2,5697	−0,0117
4 000	2,570	2,5697	0,0003
4 000	2,580	2,5697	0,0103
6 000	2,646	2,6670	−0,0210
6 000	2,657	2,6670	−0,0100
6 000	2,653	2,6670	−0,0140
8 000	2,724	2,7643	−0,0403
8 000	2,774	2,7643	0,0097
8 000	2,808	2,7643	0,0437
10 000	2,861	2,8617	−0,0007
10 000	2,879	2,8617	0,0173
10 000	2,858	2,8617	−0,0037

Tableau 7.3.3.3 Résidus de l'ajustement linéaire aux données de pression et de la densité.

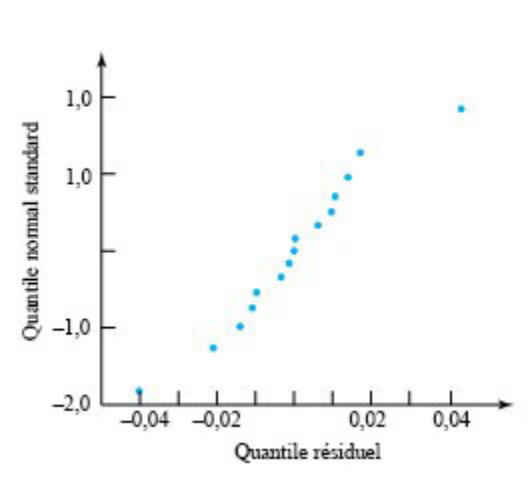


Figure 7.1.3.4 Tracé normal des résidus d'un ajustement linéaire aux données de pression et de la densité.

7.1.4 Mises en garde relatives à l'utilisation de la régression linéaire des moindres carrés

Les méthodes présentées dans cette section sont des outils très utiles lorsqu'elles sont appliquées judicieusement. Il convient de formuler quelques remarques supplémentaires afin de prévenir certaines erreurs de logique.

r exprime uniquement l'association linéaire

La première mise en garde concerne la corrélation. Il convient de rappeler que r exprime uniquement la relation linéaire entre x et y. Il est parfaitement possible d'avoir une forte relation non linéaire entre x et y tout en ayant une valeur de r proche de 0. En fait, notre deuxième exemple illustre parfaitement cette situation. La résistance à la compression est fortement liée à la teneur en phosphate d'ammonium, mais r = −0,005, soit très proche de 0, pour l'ensemble des données du tableau 7.1.3.1.

Corrélation et causalité

La deuxième mise en garde est en fait une reformulation d'une mise en garde implicite faite au début de cette discussion : la corrélation n'indique pas nécessairement un lien de causalité. Il est possible d'observer une forte corrélation entre x et y dans une étude d'observation sans pour autant que x soit à l'origine de y ou vice versa. Il se peut qu'une autre variable (par exemple, z) régisse le système étudié et provoque des changements simultanés dans x et y.

L'influence des observations extrêmes

La dernière mise en garde est que r, R^2 et la régression des moindres carrés peuvent être considérablement perturbés par quelques données aberrantes. Par exemple, la figure 7.1.4.1 indique l'âge et la taille de 36 étudiant.e.s d'un cours de statistiques élémentaires. Quand les gens entrent à l'université, il n'y a plus vraiment de relation utile entre l'âge et la taille. Néanmoins, la corrélation entre l'âge et la taille est de 0,73. Cette valeur assez importante est obtenue essentiellement en raison d'un seul point de données. Si l'on retire de l'ensemble des données le point correspondant à l'étudiant de 30 ans qui mesure 6 pieds 8 pouces, la corrélation tombe à 0,03.

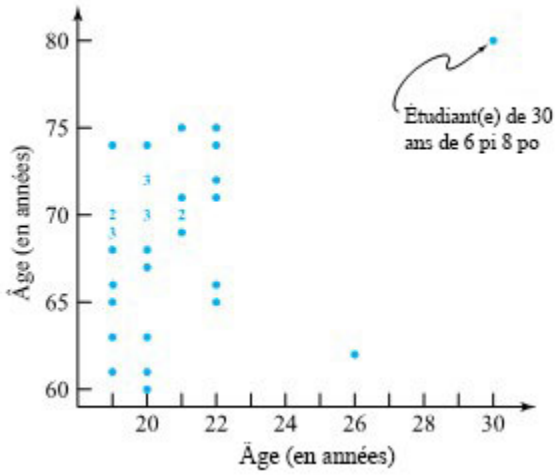


Figure 7.1.4.1 Nuage de points de l'âge et de la taille de 36 étudiant.e.s.

Pour éviter d'être induit en erreur par ce type de phénomène, il faut prendre l'habitude de représenter les données d'autant de façons différentes qu'il est nécessaire pour se faire une idée de la façon dont elles sont structurées. Même un simple diagramme en boîte des données relatives à l'âge ou à la taille aurait permis de dire que les données relatives à l'étudiant de 30 ans de la figure 7.1.4.1 sont aberrantes. On aurait alors pu supposer que ce point de données influençait fortement r et toute courbe pouvant de régression obtenue par la méthode des moindres carrés.

7.1.5 Utilisation du calcul statistique

Les exemples présentés dans cette section ont sans doute donné l'impression que les calculs étaient effectués « à la main ». En réalité, ces calculs sont presque toujours effectués à l'aide d'un logiciel d'analyse statistique. La régression d'une droite par la méthode des moindres carrés est généralement effectuée au moyen d'un programme de régression. La plupart du temps, ces programmes calculent aussi R^2 et disposent d'une option qui permet de calculer et de tracer les résidus.

Ce cours utilise la programmation en Python et Jupyter Notebooks comme plateforme de calcul statistique, mais il existe de nombreuses autres plateformes. Des captures d'écran annotées sont souvent incluses pour montrer comment Python formate et affiche ses résultats.

La capture d'écran 7.1.5.1, qui provient de notre site GitHub, présente une analyse des données de pression et de densité de l'exemple du module 7.1.1. Cette analyse reprend largement ce que nous avons fait au cours de cette partie. Elle peut être consultée ou téléchargée (comme d'habitude) sous la partie 7 d'Intro Statistical Methods for Engineering ou sur le site GitHub spécial pour la partie 7.

Vous pouvez également ouvrir un environnement informatique interactif pour travailler avec le Jupyter Notebook en utilisant Python sur le site Binder grâce au site GitHub spécial pour l'exemple de la partie 7. Cliquez ICI pour aller au site Binder (qui se trouve à).

La bibliothèque Statsmodels de Python que nous utilisons offre à ses utilisateurs bien plus de possibilités d'analyse pour la régression des courbes des moindres carrés que ce qui a été discuté jusqu'à présent. Ainsi, votre compréhension de la capture d'écran sera limitée. Vous devriez cependant être en mesure de repérer les valeurs des principales
statistiques de synthèse présentées ici.

```
The regression equation is
density = 2.375 + 4.867e-05 *pressure
```

```
                  Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.981
Dependent Variable: density          AIC:                -73.0762
Date:               2024-01-30 15:06 BIC:                -71.6601
No. Observations:   15               Log-Likelihood:     38.538
Df Model:           1                F-statistic:        717.1
Df Residuals:       13               Prob (F-statistic): 9.31e-13
R-squared:          0.982            Scale:              0.00039636
-----------------------------------------------------------------
              Coef.    Std.Err.      t       P>|t|    [0.025   0.975]
-----------------------------------------------------------------
Intercept     2.3750     0.0121   197.0079   0.0000   2.3490   2.4010
pressure      0.0000     0.0000    26.7780   0.0000   0.0000   0.0001
-----------------------------------------------------------------
Omnibus:                 2.101        Durbin-Watson:           1.682
Prob(Omnibus):           0.350        Jarque-Bera (JB):        0.427
Skew:                    0.137        Prob(JB):                0.808
Kurtosis:                3.780        Condition No.:           15556
=================================================================
```

ANOVA table

```
df    sum_sq   mean_sq          F        PR(>F)
```

```
pressure   1.0  0.284213  0.284213  717.060422  9.306841e-13
Residual  13.0  0.005153  0.000396         NaN           NaN
```

	pressure	density	Fit	StDev Fit	Residual	St Resid
0	2000	2.486	2.472333	0.008903	0.013667	0.767491
1	2000	2.479	2.472333	0.008903	0.006667	0.374386
2	2000	2.472	2.472333	0.008903	-0.000333	-0.018719
3	4000	2.558	2.569667	0.006296	-0.011667	-0.617705
4	4000	2.570	2.569667	0.006296	0.000333	0.017649
5	4000	2.580	2.569667	0.006296	0.010333	0.547110
6	6000	2.646	2.667000	0.005140	-0.021000	-1.091834
7	6000	2.657	2.667000	0.005140	-0.010000	-0.519921
8	6000	2.653	2.667000	0.005140	-0.014000	-0.727889
9	8000	2.724	2.764333	0.006296	-0.040333	-2.135495
10	8000	2.774	2.764333	0.006296	0.009667	0.511813
11	8000	2.808	2.764333	0.006296	0.043667	2.311982
12	10000	2.861	2.861667	0.008903	-0.000667	-0.037439
13	10000	2.879	2.861667	0.008903	0.017333	0.973403
14	10000	2.858	2.861667	0.008903	-0.003667	-0.205912

7.1.6 Tutoriel 5 – Corrélation et covariance

À ce stade, il est recommandé de faire l'exercice du tutoriel 5 qui se trouve sur le référentiel GitHub. Cet exercice vous apprendra à calculer la covariance et la corrélation en utilisant le langage Python.

Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur la régression linéaire simple. Vous pouvez les trouver dans la section « How do I do X in Python? ». Les fichiers « Correlation & Covariance » vous seront particulièrement utiles.

7.2.0 Introduction aux méthodes d'inférence de la régression linéaire simple liées à la régression d'une droite selon la méthode des moindres carrés (régression linéaire simple)

Nous avons commencé l'étude des méthodes d'inférence pour les études multi-échantillons en abordant d'abord les méthodes qui n'utilisent pas explicitement la structure relative à plusieurs échantillons, et nous terminerons le cours en discutant de celles qui sont orientées vers l'analyse de la structure factorielle. Dans ce module, nous examinerons principalement les méthodes d'inférence pour les études à multi-échantillons lorsque les facteurs en cause sont intrinsèquement quantitatifs et qu'il est raisonnable de penser qu'il existe une relation fonctionnelle approximative entre les valeurs des variables du système, d'entrée ou indépendantes et les réponses observées du système. Autrement dit, ce chapitre présente et applique des méthodes d'inférence aux cas de régression des droites abordés dans le module 7.1.

Nous commencerons par examiner la situation la plus simple, à savoir celle où une variable de réponse *y* est liée de façon approximativement linéaire à une seule variable d'entrée quantitative *x*. Dans un tel contexte, on peut donner des formules explicites et illustrer en termes concrets les possibilités offertes par les méthodes d'inférence pour les analyses de régression. Nous aborderons l'analyse de régression multiple (régression de courbe et de surface) dans le prochain module.

Nous commencerons par examiner la situation la plus simple, à savoir celle où une variable de réponse y est liée de façon approximativement linéaire à une seule variable d'entrée quantitative x. Nous présentons tout d'abord le modèle de régression linéaire simple (normal) et expliquons comment estimer la variance de la réponse dans ce contexte. Ensuite, nous examinons les résidus normalisés. Puis, nous abordons le taux de variation $(\Delta y/\Delta x)$, ainsi que l'inférence pour la réponse moyenne à une valeur x donnée. Nous discuterons ensuite des intervalles de prévision et de tolérance pour les réponses à une valeur donnée de x et présenterons des concepts liés à ANOVA pour la situation actuelle. Enfin, nous montrerons comment les logiciels statistiques permettent d'exécuter rapidement les calculs présentés dans cette partie.

7.2.1 Modèle de régression linéaire simple, estimation de la variance correspondante et résidus normalisés

À la partie 6, nous avons vu que le modèle à un facteur (même variance, distribution normale) est la base de probabilité la plus courante des méthodes d'inférence pour les études multi-échantillons. Il était représenté par les symboles

7.2.1.1 $$y_{ij} = \mu_i + \epsilon_{ij}$$

où les moyennes $\mu_1, \mu_2, \ldots, \mu_r$ étaient considérées comme r paramètres non limités. Dans le cas d'une inférence basée sur les paires de données $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ dont le nuage de points est approximativement linéaire, il faut à nouveau imposer une restriction au modèle à un facteur 7.2.1.1. En mots, les hypothèses du modèle seront qu'il y a des distributions normales sous-jacentes pour la réponse y avec une variance commune σ^2, mais que les moyennes $\mu_{y|x}$ varient de façon linéaire en fonction de x. En symboles, il est courant d'écrire que pour $i = 1, 2, \ldots, n$:

Modèle de régression linéaire simple (normal) 7.2.1.2

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

où les i sont des variables aléatoires (non observables) indépendantes et identiquement distribuées (iid) suivant une distribution normale (0,σ^2), les x_i sont des constantes connues, et β_0, β_1 et σ^2 sont des paramètres inconnus du modèle (constantes fixes). Le modèle 7.2.1.2 est couramment appelé modèle de régression linéaire simple (normal).

Si on considère que les différentes valeurs de x dans un ensemble de données (x, y) le « séparent » en divers échantillons de y, l'expression 7.2.1.2 est la formulation du modèle 7.2.1.1 dans lequel les moyennes (auparavant non limitées) satisfont à la relation linéaire $\mu_{y|x} = \beta_0 + \beta_1 x$. La figure 7.2.1.1 est une représentation graphique du modèle « distribution normale, variance constante, moyenne linéaire (en x) ».

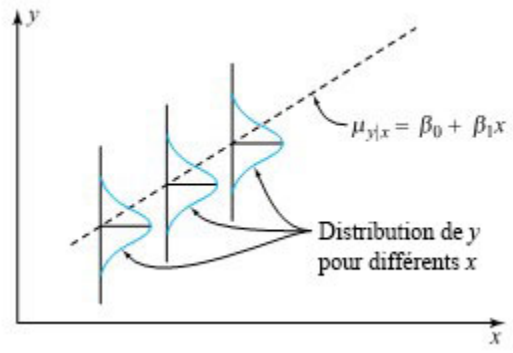


Figure 7.2.1.1 Représentation graphique du modèle de régression linéaire simple

Les inférences sur les quantités comprenant les valeurs x représentées dans les données (comme la réponse moyenne à un x donné ou la différence entre les réponses moyennes à deux valeurs différentes de x) seront généralement plus précises lorsqu'on peut utiliser des méthodes basées sur le modèle 7.2.1.2 plutôt que les méthodes générales de la partie 6 et d'ANOVA. Et dans la mesure où le modèle 7.2.1.2 décrit le comportement du système pour des valeurs de x non incluses dans les données, un tel modèle permet de faire des inférences limitées d'interpolation et d'extrapolation limitées sur x.

Le module 7.1 aborde en détail l'utilisation des moindres carrés dans l'ajustement de la relation approximativement linéaire

7.2.1.3

$$y \approx \beta_0 + \beta_1 x$$

à un ensemble de données (x, y). À présent, nous pouvons constater que le module 7.1 peut être considéré comme une présentation de la régression et de l'utilisation des résidus dans la vérification du modèle pour la régression linéaire simple (équation 7.2.1.2). Plus particulièrement, les estimations de β_1 et de $null\beta_0$ sont associées au modèle de régression linéaire simple, que nous montrons à nouveau ici :

Pente de la droite des moindres carrés b_1 7.2.1.3

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

et

Ordonnée à l'origine de la droite des moindres carrés b_0 7.2.1.4

$$b_0 = \bar{y} - b_1 \bar{x}$$

ainsi que les valeurs ajustées correspondantes

Valeurs ajustées pour la régression linéaire simple 7.2.1.5

$$\hat{y}_i = b_0 + b_1 x_i$$

et les résidus

Résidus pour la régression linéaire simple 7.2.1.6

$$e_i = y_i - \hat{y}_i$$

De plus, les résidus (ou erreurs) de l'équation 7.2.1.6 peuvent servir à estimer σ^2. Comme toujours, la somme des résidus au carré est divisée par un nombre approprié de degrés de liberté. La définition ci-dessous est celle d'une régression linéaire simple ou d'une variance d'échantillon d'une régression de droite, que nous appellerons l'erreur quadratique moyenne de la régression de droite (MSE_{LF}, de l'anglais *mean squared error* et *line fitting*).

ERREUR QUADRATIQUE MOYENNE DU MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE D'UNE DROITE

DÉFINITION Erreur quadratique moyenne du modèle de régression linéaire simple d'une droite (MSE_{LF})

EXPRESSION 7.2.1.7

$$MSE_{\text{LF}} = s^2_{\text{LF}} = \frac{1}{n-2}\sum(y-\hat{y})^2 = \frac{1}{n-2}\sum e^2$$

est l'erreur quadratique moyenne de la régression d'une droite (MSE_{LF}). Il s'agit de l'ajustement (par régression linéaire simple) de la variance de l'erreur de l'échantillon (s^2_{LF}).

Elle est associée aux $\nu = n-2$ degrés de liberté et à l'erreur type du modèle de régression d'une droite ($sqrtMSE_{\text{LF}}$, une estimation de l'écart-type de la variable de réponse ($\sqrt{s^2_{\text{LF}}}$).

DÉFINITION Erreur type du modèle de régression linéaire simple d'une droite ($\sqrt{MSE_{\text{LF}}}$)

EXPRESSION 7.2.1.8

$$\sqrt{MSE_{\text{LF}}} = \sqrt{s^2_{\text{LF}}} = s_{\text{LF}}$$

s_{LF} est une estimation de la variation de base σ^2 si le modèle de régression linéaire simple 7.2.1.2 décrit bien le système étudié.

Si ce n'est pas le cas, s_{LF} aura tendance à surestimer σ. La comparaison de s_{LF} à s_p (l'écart-type de l'échantillon groupé) est donc une autre façon de déterminer si le modèle 7.2.1.2 est approprié. Un s_{LF} beaucoup plus élevé que s_p laisse penser que le modèle de régression linéaire ne convient pas.

Example 7.2.1.1 Inférence dans l'étude sur le pressage de la poudre céramique (suite du module 7.1)

L'exemple principal de cette section sera l'étude de Benson, Locher et Watkins sur la pression et la densité, qui a été abondamment utilisée dans le module 7.1 pour illustrer l'analyse descriptive des données (x, y). Le tableau 7.2.1.1 présente à nouveau les n = 15 paires de données (x, y) (présentées pour la première fois dans le tableau 7.1.1.1), où

x = la pression utilisée (psi)

y = la densité obtenue (g/cc)

pour le pressage à sec d'un composé céramique en cylindres. La figure 7.2.1.1 est un nuage de points des données.

Rappelons également que d'après le calcul de R^2, les données du tableau 7.2.1.1 donnent les valeurs ajustées du tableau 7.1.1.2, et que

$$\sum(y-\hat{y})^2 = 0{,}005153$$

Ainsi, pour les données de pression et de densité, on obtient (par l'équation 7.2.1.7) que

$$s^2_{\text{LF}} = \frac{1}{15-2}(0{,}005153) = 0{,}000396(\text{g/cc})^2$$

donc

$$s_{\mathrm{LF}} = \sqrt{0{,}000396} = 0{,}0199 \text{ g/cc}$$

Si l'on accepte la pertinence du modèle 7.2.1.2 dans cet exemple, pour toute pression fixe, l'écart-type des densités associées à de nombreux cylindres fabriqués à cette pression serait d'environ 0,02 g/cc.

Les données initiales de cet exemple peuvent être considérées comme organisées en r = 5 échantillons distincts de taille m = 3, un pour chacune des pressions de 2 000 psi, 4 000 psi, 6 000 psi, 8 000 psi et 10 000 psi. Ce raisonnement mène à une autre estimation de σ, soit s_P. Le tableau 7.2.1.2 donne les valeurs de $\hat{y}$ et de s pour les cinq échantillons.

Les écart-types des échantillons du tableau 7.2.1.2 peuvent être utilisés de la manière habituelle pour calculer s_P. Donc (à partir de l'expression de la partie 5),

$$s_{\mathrm{P}}^2 = \frac{(3-1)(.0070)^2 + (3-1)(.0110)^2 + \cdots + (3-1)(.0114)^2}{(3-1) + (3-1) + \cdots + (3-1)} = 0{,}000424(\mathrm{g/cc})^2$$

d'où

$$s_{\mathrm{P}} = \sqrt{s_{\mathrm{P}}^2} = 0{,}0206 \text{ g/cc}$$

Lorsqu'on compare s_{LF} et s_P, rien n'indique que ces valeurs soient mal ajustées.

x, Pression (psi)	y, Densité (g/cc)
2 000	2,486
2 000	2,479
2 000	2,472
4 000	2,558
4 000	2,570
4 000	2,580
6 000	2,646
6 000	2,657
6 000	2,653
8 000	2,724
8 000	2,774
8 000	2,808
10 000	2,861
10 000	2,879
10 000	2,858

Tableau 7.2.1.1 : Pressions de pressage et densité de l'échantillon résultant

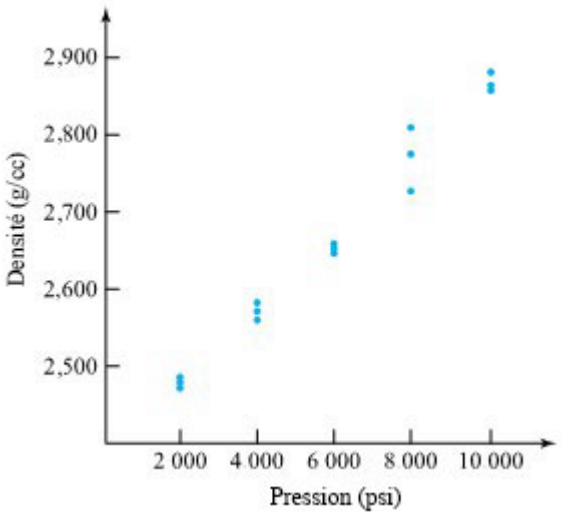


Figure 7.2.1.2. Nuage de points de la densité en fonction de la pression de pressage

x, Pression (psi)	$\bar{y}$, Moyenne de l'échantillon	s, Écart-type de l'échantillon
2 000	2,479	0,0070
4 000	2,569	0,0110
6 000	2,652	0,0056
8 000	2,769	0,0423
10 000	2,866	0,0114

Tableau 7.2.1.2 Moyenne et écart-type d'échantillon pour la densité obtenue avec cinq pressions de pressage différentes

Le module 7.1 comprend un tracé des résidus (équation 7.2.1.6) pour les données de pression et de densité (et notamment, un tracé normal à la figure 7.1.3.4). Même si les résidus (bruts) de l'équation 7.2.1.6 sont les plus faciles à calculer, la plupart des programmes de régression sur le marché proposent également de calculer les résidus normalisés. C'est même l'option privilégie par certains logiciels.

RÉSIDUS NORMALISÉS

Dans les analyses d'ajustement de courbes et de surfaces, la variance des résidus dépend de x.
La normalisation avant le tracé est un moyen d'éviter de confondre une tendance sur un tracé de résidus qui peut être expliquée par ces différentes variances avec une tendance qui indique des problèmes avec le modèle de base. Selon le modèle 7.2.1.2, pour un x donné avec la réponse correspondante y,

7.2.1.7

$$\mathrm{Var}(y-\hat{y}) = \sigma^2\left(1-\frac{1}{n}-\frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}\right)$$

Ainsi, à la lumière de l'équation 7.2.1.7 et des discussions sur la normalisation, la paire de données (x_i, y_i) est le résidu normalisé pour la régression linéaire simple.

Résidus normalisés pour la régression linéaire simple 7.2.1.8

$$e_i^* = \frac{e_i}{s_{\mathrm{LF}}\sqrt{1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{\sum(x-\bar{x})^2}}}$$

La méthode plus avancée pour examiner les résidus du modèle 7.2.1.2 consiste donc à tracer les valeurs de l'équation 7.2.1.8 plutôt que les résidus bruts (de l'équation 7.2.1.6).

Exemple 7.2.1.2 (suite)

Examinons comment les résidus normalisés pour l'ensemble de données de pression et de densité sont liés aux résidus bruts. En sachant que

\sum(x-\bar{x})^2=120 000 000

et que les valeurs x_i des données initiales ne comprenaient que les pressions de 2 000 psi, 4 000 psi, 6 000 psi, 8 000 psi et 10 000 psi, il est facile d'obtenir les valeurs nécessaires de la racine dans le dénominateur de l'expression 7.2.1.8. Ces valeurs se trouvent dans le tableau 7.2.1.3.

x	$\sqrt{1-\frac{1}{15}-\frac{(x-6\,000)^2}{120\,000\,000}}$
2 000	0,894
4 000	0,949
6 000	0,966
8 000	0,949
10 000	0,894

Tableau 7.2.1.3 Calculs des résidus normalisés dans l'étude de la pression et de la densité

Les données du tableau 7.2.1.3 montrent, par exemple, qu'il faut s'attendre à ce que les résidus correspondant à x = 6 000 psi soient (en moyenne) environ 0,966/0,894 = 1,08 fois plus importants que les résidus correspondant à x = 10 000 psi. Il faut diviser les résidus bruts par s_{FL} multiplié par la valeur correspondante de la deuxième colonne du tableau 7.2.1.3 pour les mettre sur un pied d'égalité. Le tableau 7.2.1.4 donne les résidus bruts (tirés du module 7.1) et leurs contreparties normalisées.

x	e	Valeurs résiduelles normalisées
2 000	0,0137, 0,0067, —0,0003	0,77, 0,38, —0,02
4 000	—0,0117, 0,0003, 0,0103	—0,62, 0,02, 0,55
6 000	—0,0210, —0,0100, —0,0140	—1,09, —0,52, —0,73
8 000	—0,0403, 0,0097, 0,0437	—2,13, 0,51, 2,31
10 000	—0,0007, 0,0173, —0,0037	—0,04, 0,97, —0,21

Tableau 7.2.1.4 Résidus bruts et normalisés de l'étude de la pression et de la densité

Dans le cas présent, étant donné que les valeurs 0,894, 0,949 et 0,966 sont assez comparables, la normalisation au moyen de l'équation 7.2.1.8 ne modifie pas beaucoup les conclusions relatives à la justesse du modèle. Par exemple, les figures 7.2.1.3 et 7.2.1.4 sont des tracés normaux des résidus bruts et des résidus normalisés (respectivement). À toutes fins pratiques, ils sont identiques. Par conséquent, toutes les conclusions (comme celles formulées dans le module 7.1) sur la justesse du modèle étayées par la figure 7.2.1.3 sont également étayées par la figure 7.2.1.4, et vice-versa.

Cependant, dans d'autres situations(en particulier lorsqu'un ensemble de données contient quelques valeurs de x très extrêmes), la normalisation peut comporter des dénominateurs plus variés pour l'équation 7.2.1.8 que ceux du tableau 7.2.1.3, ce qui aurait une incidence sur les résultats d'une analyse des résidus.

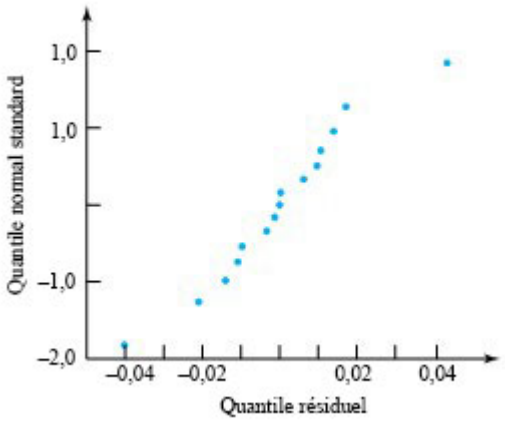


Figure 7.2.1.3 Tracé normal des résidus d'un ajustement linéaire aux données de pression et de la densité.

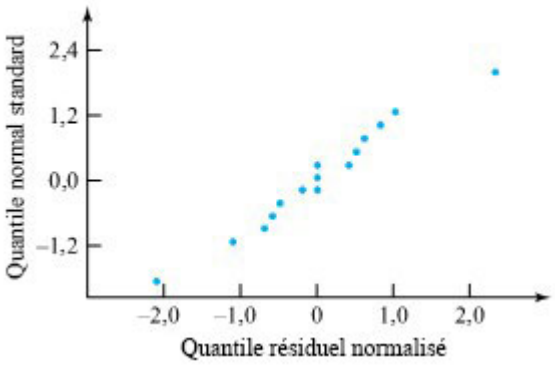


Figure 7.2.1.4 Tracé normal des résidus normalisés d'un ajustement linéaire aux données de pression et de la densité.

7.2.2 Inférence du paramètre de pente

Dans les applications du modèle de régression linéaire simple (équation 7.2.1.1) où x représente une variable qui peut être physiquement manipulée, le paramètre de pente β_1 est d'un intérêt fondamental. Il représente le taux de variation de la réponse moyenne par rapport à x et détermine l'incidence d'une variation de x sur la sortie du système. L'inférence de β_1 est assez simple, parce que b_1 (la pente de la droite des moindres carrés) hérite des propriétés de distribution du modèle. Ainsi, selon le modèle 7.2.1.1, b_1 suit une distribution normale avec

$$Eb_1 = \beta_1$$

et

7.2.2.1
$$\operatorname{Var} b_1 = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

ce qui signifie que

$$Z = \frac{b_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum (x - \bar{x})^2}}}$$

suit une distribution normale réduite. D'une manière similaire à de nombreux arguments des parties 5 et 6, cela explique le fait que la quantité

7.2.2.2
$$T = \frac{b_1 - \beta_1}{\frac{s_{\mathrm{LF}}}{\sqrt{\sum (x - \bar{x})^2}}}$$

suit une distribution t_{t-2}. Les arguments standard de la partie 5 appliqués à l'expression 7.2.2.2 montrent alors que

7.2.2.3
$$\mathrm{H}_0 : \beta_1 = \#$$

peut être testée à l'aide de la statistique de test (aussi appelée variable de décision)

7.2.2.4 Statistique de test pour $\mathrm{H}_0 : \beta_1 = \#$

$$T = \frac{b_1 - \#}{\frac{s_{\mathrm{LF}}}{\sqrt{\sum (x - \bar{x})^2}}}$$

et une distribution de référence t_{n-2}. Plus important encore, dans le cadre du modèle de régression linéaire simple (équation 7.2.1.2), un intervalle de confiance bilatéral pour β_1 peut être établi avec les bornes suivantes :

7.2.2.5 Limites de confiance pour la pente β_1

$$b_1 \pm t \frac{s_{\mathrm{LF}}}{\sqrt{\sum (x - \bar{x})^2}}$$

où le niveau de confiance associé correspond à la probabilité assignée à l'intervalle entre $-t$ et t dans la distribution t_{n-2}. Un intervalle de confiance unilatéral est établi de la manière habituelle, en utilisant une seule borne dans l'équation 7.2.2.5.

Exemple 7.2.2.1 Étude sur le pressage de la poudre (suite)

Dans l'étude du pressage de poudre, nous avons vu au module 7.1 que la pente de la droite des moindres carrés passant par les données de pression et de densité est la suivante :

$$b_1 = 0{,}0000486\overline{6}(\text{g/cc})/\text{psi}$$

Un intervalle de confiance bilatéral de 95 % pour β_1 peut être établi en utilisant le quantile 0,975 de la distribution $t_1 3$ dans l'équation 7.2.2.5. On peut donc utiliser les bornes

$$.0000486\overline{6} \pm 2.160\frac{.0199}{\sqrt{120,000,000}}$$

soit

$$.0000486\overline{6} \pm .0000039$$

ou

$$0{,}0000448(\text{g/cc})/\text{psi} \text{ et } 0{,}0000526(\text{g/cc})/\text{psi}$$

Un intervalle de confidence comme celui-ci de β_1 peut être converti en un intervalle de confiance pour une différence de réponses moyennes pour deux valeurs différentes de x. Selon le modèle 7.2.1.2, deux valeurs de x qui diffèrent de Δx ont des réponses moyennes qui diffèrent de $\beta_1 \Delta x$. Pour obtenir un intervalle de confidence pour la différence des réponses moyennes, il suffit alors de multiplier les bornes de l'intervalle de confidence pour β_1 par Δx. Par exemple, puisque 8 000 – 6 000 = 2 000, la différence entre les densités moyennes à 8 000 psi et à 6 000 psi a un intervalle de confidence de 95 % avec les bornes

$$2\ 000(\ 0{,}0000448)\text{g/cc} \text{ et } 2\ 000(\ 0{,}0000526)\text{g/cc}$$

soit

$$0{,}0896\ \text{g/cc} \text{ et } 0{,}1052\ \text{g/cc}$$

POINTS À PRENDRE EN CONSIDÉRATION POUR LA SÉLECTION DES VALEURS DE X

L'équation 7.2.2.5 indique la précision de la pente de la droite des moindres carrés. Il est utile de se demander comment cette précision est liée aux caractéristiques de l'étude qu'on peut potentiellement contrôler. Les équations 7.2.2.1 et 7.2.2.5 indiquent toutes deux que plus $\sum(x - \bar{x})^2$ est élevée (c'est-à-dire plus les valeurs x_i sont dispersées), plus b_1 est une estimation précise de la pente sous-jacente β_1. Ainsi, pour estimer β_1, dans les études où x représente la valeur d'une variable de système contrôlable, il convient de choisir les paramètres de x ayant la plus grande variance d'échantillon possible. (En fait, si l'on dispose de n observations à utiliser et que l'on peut choisir des valeurs de x n'importe où dans un intervalle $[a, b]$, en prendre $\frac{n}{2}$ à $x = a$ et $\frac{n}{2}$ à $x = b$ donne la meilleure précision possible pour estimer la pente β_1.)

Toutefois, ce conseil (éloigner les x_i) doit être pris avec un grain de sel. La relation approximativement linéaire (équation 7.2.1.2) peut ne s'appliquer qu'à une plage limitée de valeurs de x. Si on veut obtenir une bonne estimation de la pente, il est évidemment insensé de choisir des valeurs expérimentales de x au-delà des limites où l'on peut raisonnablement s'attendre à ce que l'équation 7.2.1.2 s'applique. Il est également important de comprendre que l'estimation précise de β_1 reposant sur les hypothèses du modèle 7.2.1.2 n'est pas le seul élément à prendre en considération lors de la planification de la collecte des données. Il est généralement aussi important d'être en mesure de dire quand la forme linéaire de l'équation 7.2.1.2 est inappropriée. Pour cela, il faut recueillir des données pour un plusieurs valeurs différentes de x, et non pas simplement pour les valeurs les plus

faibles et les plus élevées
possibles.

7.2.3 Inférence pour la moyenne de la réponse d'un système pour une valeur particulière de x

Au chapitre 6, nous avons abordé le problème de l'estimation de la moyenne de y avec les niveaux du ou des facteurs considérés. À présent, le problème analogue est celui de l'estimation de la réponse moyenne pour une valeur fixée de la variable du système x,

7.2.3.1 $$\mu_{y|x} = \beta_0 + \beta_1 x$$

L'approximation naturelle (et basée sur des données) de la moyenne dans l'équation 7.2.3.1 est la valeur y correspondante tirée de la droite des moindres carrés. La notation

7.2.3.2 Estimateur de $\mu_{y|x} = \beta_0 + \beta_1 x$

$$\hat{y} = b_0 + b_1 x$$

sera utilisée pour cette valeur sur les droites des moindres carrés. (Et ce, malgré le fait que la valeur de l'équation 7.2.3.2 peut ne pas être une valeur ajustée au sens où cette expression a été le plus souvent utilisée jusqu'à présent. Il n'est pas nécessaire que x soit égal à $x_1, x_2, \ldots, x_n$ pour que les expressions 7.2.3.1 et 7.2.3.2 soient valables.) Le modèle de régression linéaire simple (équation 7.2.1.2) donne des propriétés de distribution simples pour $\hat{y}$, qui mènent à des méthodes d'inférence pour $\mu_{y|x}$.

Selon le modèle 7.2.1.2, $\hat{y}$ suit une distribution normale avec

$$E\hat{y} = \mu_{y|x} = \beta_0 + \beta_1 x$$

et

7.2.3.3 $$\operatorname{Var} \hat{y} = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right)$$

(Dans l'expression 7.2.3.3, on a un peu abusé de la notation. Les indices i de la somme dans $\sum (x - \bar{x})^2$ ont été supprimés. Cette sommation porte sur les n valeurs x_i de l'ensemble de données original. D'autre part, dans le terme $(x - \bar{x})^2$ du numérateur de l'expression 7.2.3.3, le x considéré n'est pas nécessairement égal à l'un des $x_1, x_2, \ldots, x_n$. Il s'agit plutôt de la valeur de la variable du système à laquelle la réponse moyenne doit être estimée.) Ainsi,

$$Z = \frac{\hat{y} - \mu_{y|x}}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}}$$

suit une distribution normale standard, ce qui implique que

7.2.3.4 $$T = \frac{\hat{y} - \mu_{y|x}}{s_{\mathrm{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}}$$

suit une distribution t_{n-2}. Les arguments standard de la partie 5 appliqués à l'expression 7.2.3.4 montrent alors que

7.2.3.5 $$\mathrm{H}_0 : \mu_{y|x} = \#$$

peut être testée à l'aide de la statistique de test (aussi appelée variable de décision)

7.2.3.6 Statistique de test pour $H_0 : \mu_{y|x} = \#$

$$T = \frac{\hat{y} - \#}{s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x-\bar{x})^2}}}$$

et d'une distribution de référence t_{n-2}. De plus, dans le cadre du modèle de régression linéaire simple (équation 7.2.1.2), un intervalle de confidence individuel bilatéral pour $\mu_{y|x}$ peut être établi avec les bornes suivantes :

7.2.3.7 Limites de confiance pour la réponse moyenne, $\mu_{y|x} = \beta_0 + \beta_1 x$

$$\hat{y} \pm t s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

où le niveau de confiance associé correspond à la probabilité assignée à l'intervalle entre $-t$ et t dans la distribution $t_(n - 2)$. Un intervalle de confiance unilatéral s'obtient de la manière habituelle, en utilisant une seule borne dans l'équation 7.2.3.7.

Exemple 7.2.3.1 (suite)

Revenons à l'étude de la pression et de la densité. Établissons des intervalles de confidence individuels à 95 % pour les densités moyennes des cylindres obtenus d'abord à 4 000 psi, puis à 5 000 psi.

En commençant avec x = 4 000 psi, l'estimation correspondante de la densité moyenne est la suivante :

$$\hat{y} = 2{,}375 + 0{,}0000486(4\ 000) = 2{,}5697 \text{ g/cc}$$

De plus, d'après l'équation 7.2.3.7 et le fait que le quantile 0,975 de la distribution $t_1 3$ vaut 2,160, une précision de plus ou moins

$$2{,}160(0{,}0199)\sqrt{\frac{1}{15} + \frac{(4\ 000 - 6\ 000)^2}{120\ 000\ 000}} = 0{,}0136 \text{ g/cc}$$

peut être associée à la valeur de 2,5697 g/cc. Autrement dit, les bornes d'un intervalle de confidence bilatéral à 95 % pour la densité moyenne dans la condition de 4 000 psi sont les suivantes :

$$2{,}5561 \text{ g/cc} \quad \text{et} \quad 2{,}5833 \text{ g/cc}$$

À x = 5 000 psi, l'estimation correspondante de la densité moyenne est la suivante :

$$\hat{y} = 2.375 + .0000486(5,000) = 2.6183 \text{ g/cc}$$

D'après l'équation 7.2.3.7, une précision de plus ou moins

$$2{,}160(0{,}0199)\sqrt{\frac{1}{15} + \frac{(5\ 000 - 6\ 000)^2}{120\ 000\ 000}} = 0{,}0118 \text{ g/cc}$$

peut être associée à la valeur de 2,6183 g/cc. Autrement dit, les bornes d'un intervalle de confidence bilatéral à 95 % pour la densité moyenne dans la condition de 5 000 psi sont les suivantes :

$$2{,}6065 \text{ g/cc} \quad \text{et} \quad 2{,}6301 \text{ g/cc}$$

Il convient de comparer les valeurs plus ou moins des deux intervalles de confidence trouvés. L'intervalle pour x = 5 000 psi est plus court et donc plus parlant que l'intervalle pour x = 4 000 psi. L'origine de cette divergence devrait être claire, du moins après examen de l'équation 7.2.3.7. Selon les données de l'étude,
$\bar{x}$ = 6 000 psi. x = 5 000 psi est plus proche de $\bar{x}$ que x = 4 000 psi. Ainsi, $(x - \bar{x})^2$ (et donc la longueur de l'intervalle) est plus petit pour x = 5 000 psi que pour x = 4 000 psi.

Le phénomène observé dans l'exemple précédent, à savoir que la longueur d'un intervalle de confidence pour $\mu_{y|x}$ augmente à mesure que l'on s'éloigne de $\bar{x}$, est important. De plus, il a une signification intuitivement plausible pour la planification d'expériences lors desquelles on s'attend à une relation approximativement linéaire entre y et x, où x est une variable contrôlée. S'il y a un intervalle de valeurs de x sur lequel on veut obtenir une bonne précision dans l'estimation des réponses moyennes, il est logique de centrer les efforts de collecte de données sur cet intervalle.

Inférence pour l'ordonnée à l'origine β_0

Une bonne utilisation des équations 7.2.3.5, 7.2.3.6 et 7.2.3.7 donne des méthodes d'inférence pour le paramètre β_0 du modèle 7.2.1.2, l'ordonnée à l'origine de la relation linéaire (équation 7.2.3.1).β_0 Ainsi, en fixant x = 0 dans les équations 7.2.3.5, 7.2.3.6 et 7.2.3.7, on obtient des tests et des intervalles de confidence pour β_0. Cependant, à moins que x = 0 soit une valeur réalisable pour la variable d'entrée et que la région où la relation linéaire (équation 7.2.3.1) est une description raisonnable de la réalité physique comprenne x = 0, l'inférence pour β_0 seul est rarement d'un intérêt pratique.

LIMITES DE CONFIANCE BILATÉRALES SIMULTANÉES POUR TOUTES LES MOYENNES $\mu_{y|x}$

7.2.3.8 Intervalle de confiance à 95 % de la réponse moyenne

$$(b_0 + b_1 x) \pm \sqrt{2f}\, s_{\mathrm{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

où, la confiance simultanée associée est la probabilité $F_{2,n-2}$ attribuée à l'intervalle [0, f] (avec f positif).

Bien entendu, par « toutes les moyennes $\mu_{y|x}$ », on veut en vérité dire « pour toutes les réponses moyennes dans un intervalle où le modèle de régression linéaire simple 7.2.1.2 est une description adéquate de la relation entre x et y ». Comme c'est toujours le cas pour l'ajustement des courbes et des surfaces, il est risqué d'extrapoler en dehors de la plage de valeurs de x pour laquelle on dispose de données (et même, dans une certaine mesure, d'interpoler dans cette plage). Toute extrapolation doit être étayée par une expertise dans le domaine, afin de prouver qu'elle est justifiable.

Il peut être quelque peu difficile de saisir la signification d'une valeur de confiance simultanée applicable à tous les intervalles possibles de la forme 7.2.3.8. Jusqu'à présent, les niveaux de confiance étudiés l'ont été pour des ensembles finis d'intervalles. La meilleure façon de comprendre l'ensemble théoriquement infini d'intervalles donné par l'équation 7.2.3.8 est probablement de définir une région du plan (x, y) qu'on suppose contenir la droite $\mu_{y|x} = \beta_0 + \beta_1 x$. La figure 7.2.3.1 représente une région de confiance typique obtenue par l'équation 7.2.3.8. Il y a une région indiquée autour de la droite des moindres carrés dont l'étendue verticale augmente avec la distance par rapport à $\bar{x}$ et qui couvre, pour le niveau de confiance donné, la droite décrivant la relation entre x et $\mu_{y|x}$.

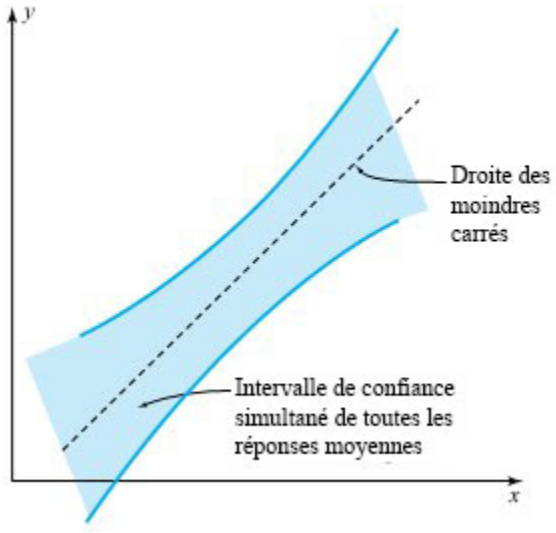


Figure 7.2.3.1 Région du plan (x, y) définie par des intervalles de confiance simultanés pour toutes les valeurs de $\mu_{y|x}$.

Exemple 7.2.3.2 (suite)

Au moyen de l'équation 7.2.3.8, on peut trouver les intervalles de confiance simultanés à 95 % pour la densité moyenne des cylindres obtenus dans les cinq conditions d'expérimentation réelles.

Puisque il faut utiliser l'équation 7.2.3.8 avec $v_1 = 2$ et $v_2 = 13$ degrés de liberté, il convient d'utiliser des limites simultanées de la forme

$$\hat{y} \pm \sqrt{2(3{,}81)}s_{\mathrm{LF}}\sqrt{\frac{1}{15} + \frac{(x - 6\ 000)^2}{120\ 000\ 000}}$$

Cela peut également se comparer à l'utilisation de la méthode P-R de la partie 6 pour le calcul simultané de l'intervalle de confiance à 95 %.

Tout d'abord, la formule (de la partie 6) montre qu'avec n − r = 15 − 5 = 10 degrés de liberté pour s_P et r = 5 conditions étudiées, les limites de confiance bilatérales simultanées à 95 % pour les cinq densités moyennes sont de la forme suivante :

$$\bar{y}_i \pm 3.103\frac{s_{\mathrm{P}}}{\sqrt{n_i}}$$

En l'occurrence,

$$\bar{y}_i \pm 3{,}103\frac{0{,}0206}{\sqrt{3}}$$

soit

$$\bar{y}_i \pm 0{,}0369\ \mathrm{g/cc}$$

Le tableau 7.2.3.1 présente les cinq intervalles résultant de l'utilisation de deux méthodes d'intervalles de confiance simultanées, ainsi que les intervalles individuels de l'équation 7.2.3.7.

Deux faits ressortent de ce tableau. Premièrement, les intervalles résultant de l'équation 7.2.3.8 sont un peu plus larges que les intervalles individuels correspondants donnés par l'équation 7.2.3.7. Deuxièmement, il est également clair que l'utilisation des hypothèses du modèle de régression linéaire simple plutôt que les hypothèses à un facteur plus générales

de la partie 6 peut conduire à des intervalles de confidence simultanés plus courts et à des déductions techniques réelles plus nettes.

x, Pression	$\mu_{y\|x}$ Densité moyenne (Méthode P-R)	$\mu_{y\|x}$ Densité moyenne (d'après l'équation 7.2.3.8)	$\mu_{y\|x}$ Densité moyenne (d'après l'équation 7.2.3.7)
2 000 psi	2,4790 ± 0,0369 g/cc	2,4723 ± 0,0246 g/cc	2,4723 ± 0,0136 g/cc
4 000 psi	2,5693 ± 0,0369 g/cc	2,5697 ± 0,0174 g/cc	2,5697 ± 0,0118 g/cc
6 000 psi	2,6520 ± 0,0369 g/cc	2,6670 ± 0,0142 g/cc	2,6670 ± 0,0111 g/cc
8 000 psi	2,7687 ± 0,0369 g/cc	2,7643 ± 0,0174 g/cc	2,7643 ± 0,0118 g/cc
10 000 psi	2,8660 ± 0,0369 g/cc	2,8617 ± 0,0246 g/cc	2,8617 ± 0,0136 g/cc

Tableau 7.2.3.1 Intervalles de confiance simultanés (et individuels) à 95 % pour la densité moyenne des cylindres

7.2.4 Intervalles de prédiction et de tolérance

L'inférence pour $\mu_{y|x}$ est une réponse à la question qualitative : « Si on garde la variable d'entrée x constante, à quelles réponses peut-on s'attendre de la part du système? » Il s'agit d'une réponse exprimée sous forme de moyenne ou de moyenne à long terme, mais parfois, il est plus pratique d'avoir une réponse exprimée sous forme de réponses individuelles. Dans ces cas, il est utile de savoir que les hypothèses du modèle de régression linéaire simple 7.2.1.2 donnent leurs propres équations spécifiques pour les intervalles de prédiction et de tolérance.

Le fait de base qui rend possible les intervalles de prédiction dans les hypothèses de l'équation 7.2.1.2 est que si y_{n+1} est une observation supplémentaire, provenant de la distribution des réponses correspondant à un x donné, et que $\hat{y}$ est la valeur ajustée correspondante à ce x (selon les n paires de données d'origine), alors

$$T = \frac{y_{n+1} - \hat{y}}{s_{\text{LF}}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}}}$$

suit une distribution t_{n-2}. Habituellement, ce fait conduit à la conclusion que, dans le cadre du modèle 7.2.1.2, l'intervalle bilatéral dont les bornes sont

7.2.4.1 Limites de prédiction de la régression linéaire simple pour un y supplémentaire à un x donné

$$\hat{y} \pm t s_{\text{LF}}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x-\bar{x})^2}}$$

peut être utilisé comme intervalle de prédiction pour une observation supplémentaire y à une valeur donnée de la variable d'entrée x. La confiance de prédiction associée est la probabilité que la distribution t_{n-2} attribue à l'intervalle entre -t et t. Les intervalles unilatéraux s'obtiennent de la manière habituelle, en utilisant une seule borne dans l'équation 7.2.4.1 et en ajustant le niveau de confiance en conséquence.

Il est possible non seulement de dériver des formules d'intervalles de prédiction à partir des hypothèses du modèle de régression linéaire simple, mais aussi de définir des formules relativement simples pour les limites de tolérance unilatérales approximatives. Autrement dit, les intervalles

7.2.4.2 Intervalle de tolérance unilatéral pour la distribution y à x

$$(\hat{y} - \tau s_{\text{LF}}, \infty)$$

et

7.2.4.3 Autre intervalle de tolérance unilatéral pour la distribution y à x

$$(-\infty, \hat{y} + \tau s_{\text{LF}})$$

peuvent être utilisés comme intervalles de tolérance unilatéraux pour une fraction p de la distribution sous-jacente des réponses correspondant à une valeur donnée de la variable du système x, à condition que le paramètre τ soit choisi de manière appropriée (en fonction des données, de p, de x et du niveau de confiance souhaité).

7.2.4.4 Rapport de $\sqrt{\text{Var}\,\hat{y}}$ sur σ pour la régression linéaire simple

$$A = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

sera adopté comme multiplicateur (p. ex., dans l'équation précédente pour passer d'une estimation de σ à une estimation de l'écart-type de $\hat{y}$). Puis, pour situer une fraction p des réponses au point x avec un niveau de confiance d'approximativement γ, le τ approprié dans l'intervalle de l'équation 7.2.4.2 ou 7.2.4.3 est le suivant :

7.2.4.5 Multiplicateur à utiliser dans les limites de tolérance

$$\tau = \frac{Q_z(p) + AQ_z(\gamma)\sqrt{1 + \frac{1}{2(n-2)}\left(\frac{Q_z^2(p)}{A^2} - Q_z^2(\gamma)\right)}}{1 - \frac{Q_z^2(\gamma)}{2(n-2)}}$$

Exemple 7.2.4.1 (suite)

Pour illustrer l'utilisation des équations d'intervalles de prédiction et de tolérance dans le contexte d'une régression linéaire simple, prenons l'exemple d'une limite inférieure de prédiction de 90 % pour une donnée de densité supplémentaire avec une pression de 4 000 psi. Puis, trouvons une limite de tolérance inférieure de 95 % pour 90 % des nombreuses densités de cylindre supplémentaires à cette même pression.

En commençant par le problème de prédiction, l'équation 7.2.4.1 montre qu'une limite de prédiction appropriée est la suivante :

$$2.5697 - 1.350(.0199)\sqrt{1 + \frac{1}{15} + \frac{(4,000 - 6,000)^2}{120,000,000}} = 2.5796 - .0282$$

soit

2,5514 g/cc

Si, au lieu de prévoir une seule densité supplémentaire pour x = 4 000 psi, il faut trouver 90 % des densités supplémentaires correspondant à une pression de 4 000 psi, il convient d'établir une limite de tolérance. Utilisons d'abord l'équation 7.2.4.4 :

$$A = \sqrt{\frac{1}{15} + \frac{(4\ 000 - 6\ 000)^2}{120\ 000\ 000}} = 0{,}3162$$

Ensuite, pour une confiance de 95 %, l'équation 7.4.4.5 donne :

$$\tau = \frac{1{,}282 + (0{,}3162)(1{,}645)\sqrt{1 + \frac{1}{2(15-2)}\left(\frac{(1{,}282)^2}{(0{,}3162)^2} - (1{,}645)^2\right)}}{1 - \frac{(1{,}645)^2}{2(15-2)}} = 2\ 149$$

Enfin, la limite inférieure de tolérance d'environ 95 % pour 90 % des densités produites avec une pression de 4 000 psi vaut (selon l'équation 7.2.4.2) :

2,5697-2,149(0,0199)=2,5697-0,0428

soit

2,5269 g/cc

MISES EN GARDE CONCERNANT LES INTERVALLES DE PRÉDICTION ET DE

TOLÉRANCE DANS LA RÉGRESSION

Étant donné que l'ajustement des courbes facilite l'interpolation et l'extrapolation, il est essentiel de faire preuve de prudence dans l'interprétation des intervalles de prédiction et de tolérance. Toutes les mises en garde concernant l'interprétation des intervalles de prédiction et de tolérance soulevées dans la partie 5 s'appliquent également à la présente situation. Ici, il faut être encore plus prudent en raison du fait que les intervalles peuvent être calculés pour des valeurs de x pour lesquelles on ne dispose d'aucune donnée. Si on veut utiliser les équations 7.2.4.1, 7.2.4.2 et 7.2.4.3 pour une valeur de x autre que $x_1, x_2, \ldots, x_n$, il doit être plausible que le modèle 7.2.1.2 décrive le comportement du système non seulement pour les valeurs de x pour lesquelles on dispose de données, mais aussi pour la nouvelle valeur de x. Et même lorsque ce modèle est « plausible », l'application des équations 7.2.4.1, 7.2.4.2 et 7.2.4.3 à de nouvelles valeurs de x doit être traitée avec prudence. Si ce jugement (non vérifié) s'avère erroné, le niveau de confiance nominal n'a aucune pertinence pratique.

7.2.5 Régression linéaire simple et ANOVA

La partie 6 a illustré comment, pour les études non structurées, la répartition de la somme totale des carrés en éléments interprétables fournit à la fois 1) une intuition et la quantification concernant l'origine de la variation observée et 2) la base d'un test F indiquant « aucune différence entre les moyennes ». Il s'avère que quelque chose de similaire est possible dans des contextes de régression linéaire simple.

Dans le contexte non structuré de la partie 6, il était utile de nommer la différence entre SCTot (somme totale des carrés) et SCE (somme des carrés d'erreur résiduelle). La convention correspondante pour l'ajustement des courbes et des surfaces est énoncée ci-dessous sous forme de définition.

DÉFINITION SOMME DES CARRÉS DE LA RÉGRESSION (SCR)

EXPRESSION 7.2.5.1

Dans les analyses de régression de courbes et de surface pour des études multi-échantillons, la différence

$$SSR = SSTot - SSE$$

est appelée la somme des carrés de la régression (SCR).

La différence de la définition 7.2.5.1 a en général la forme d'une somme de carrés de quantités appropriées. Dans le cas présent (ajustement d'une droite par les moindres carrés), il s'agit de :

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Sans utiliser la terminologie particulière de la définition 7.2.5.1 (ce manuel a déjà largement utilisé SCR = SCTot – SCE), un examen de la définition 7.1.2.2 (le coefficient de détermination R^2) dans la partie 7.1 et des définitions de la partie 6 montrera que dans le contexte de courbes et de surfaces de régression :

7.2.5.1 Coefficient de détermination pour une régression linéaire simple dans la notation de la somme des carrés

$$R^2 = \frac{SSR}{SSTot}$$

En d'autres termes, SCR est le numérateur du coefficient de détermination défini à la définition 7.1.2.2 (partie 7.1). Il est généralement considéré comme la partie de la variabilité brute de y qui est prise en compte dans le processus de régression de la courbe ou de la surface.

SCR et SCE ne fournissent pas seulement une répartition attrayante de SCTot, mais aussi les éléments pour effectuer un test F de

7.2.5.2 $$H_0 : \beta_1 = 0$$

et de

7.2.5.3 $\mathrm{H}_a : \beta_1 \neq 0$

Selon le modèle 7.2.1.2, l'hypothèse 7.2.5.2 peut être testée à l'aide de la statistique

7.2.5.4 Statistique de test F pour $\mathrm{H}_0 : \beta_1 = 0$

$$F = \frac{SSR/1}{s^2_{\mathrm{LF}}} = \frac{SSR/1}{SSE/(n-2)}$$

et d'une distribution $F_{1,n-2}$ de référence, où des valeurs élevées observées de la statistique de test constituent une preuve contre H_0.

Précédemment dans cette section, l'hypothèse nulle générale $\mathrm{H}_0 : \beta_1 = \#$ a été testée à l'aide de la variable t. Il est donc raisonnable d'examiner la relation entre le test F des équations 7.2.5.2, 7.2.5.3 et 7.2.5.4 et le test t précédent. L'hypothèse nulle $\mathrm{H}_0 : \beta_1 = 0$ est un cas particulier de l'hypothèse $\mathrm{H}_0 : \beta_1 = \#$. C'est la version la plus fréquemment testée de l'hypothèse, puisqu'elle peut (dans certaines limites) être interprétée comme l'hypothèse nulle selon laquelle la réponse moyenne ne dépend pas de x. C'est parce que si l'hypothèse 7.2.5.2 est vraie dans le cadre du modèle de régression linéaire simple 7.2.1.2, on a $\mu_{y|x} = \beta_0 + 0 \cdot x = \beta_0$, qui ne dépend pas de x. (En fait, une meilleure interprétation d'un test de l'hypothèse 7.2.5.2 est qu'il s'agit de voir si un terme linéaire en x augmente de manière significative la capacité à modéliser la réponse y après avoir tenu compte d'une réponse moyenne globale.)

Si on examine ensuite les hypothèses 7.2.5.2 et 7.2.5.3, il peut sembler que la version # = 0 des équations du module 7 représente deux méthodes de test différentes. Or, elles sont équivalentes. La statistique 7.2.5.4 est le carré de la version # = 0 de la statistique, et les niveaux de signification (bilatéraux) observés basés sur la valeur et la distribution t_{n-2} sont les mêmes que les niveaux de signification observés basés pour la statistique 7.2.5.2 et pour $F_{1,n-2}$. Ainsi, le test F indiqué ici est redondant, compte tenu de la discussion précédente, mais il est présenté ici en raison de sa relation avec les idées d'ANOVA de la partie 6 et parce qu'il offre une généralisation naturelle importante dans des situations plus complexes d'ajustement de courbes et de surfaces. (Cette généralisation, qui sera abordée dans la partie 8, n'est pas équivalente à un test t.)

La répartition de SCTot en ses parties, SCR et SCE et le calcul de la staitistique 7.2.5.4 peuvent être organisés sous la forme d'un tableau d'ANOVA. Le tableau 7.2.5.1 présente le format général utilisé dans cet ouvrage dans le contexte de la régression linéaire simple.

Table d'analyse de la variance (pour tester $\mathrm{H}_0 : \beta_1 = 0$)

Source	*SS*	*df*	*MS*	*F*
Régression	*SSR*	1	$SSR/1$	MSR/MSE
Erreur	*SSE*	$n-2$	$SSE/(n-2)$	
Total	*SSTot*	$n-1$		

Table 7.2.5.1 Forme générale du tableau d'ANOVA pour la régression linéaire simple

Exemple 7.2.5.1 (suite)

Reprenons la discussion sur l'exemple de la pression et de densité du module 7.1.1. Nous avions alors

$$\mathrm{SSTot} = \sum (y - \bar{y})^2 = 0{,}289366$$

et

$$SSE = \sum (y - \hat{y})^2 = 0{,}005153$$

Par conséquent,

$$SSR = SSTot - SSE = 0{,}289366 - 0{,}005153 = 0{,}284213$$

et le tableau 7.2.5.1 pour le présent exemple correspond au tableau 7.2.5.2.

Le niveau de signification observé pour tester $[latex]\mathrm{H}_0 : \beta_1 = 0$ est

$$P\left[\text{une variable aléatoire } F_{1,13} > 717\right] < ,001_{717\right]717\right]}$$

et l'on dispose de preuves très solides contre la possibilité que β_1. Un terme linéaire en pression contribue de manière importante à la description de la densité du cylindre. Ce résultat est tout à fait cohérent avec l'analyse précédente axée sur les intervalles, qui a donné des limites de confidence de 95 % pour β_1 de

$$0{,}0000448(\mathrm{g/cc})/\mathrm{psi} \text{ et } 0{,}0000526(\mathrm{g/cc})/\mathrm{psi}$$

Cet intervalle ne contient pas 0.

La valeur de R^2 (trouvée la première fois dans le module 7) peut également être facilement obtenue à partir des entrées du tableau 7.2.5.2 et de la relation 7.2.5.1.

Table d'analyse de la variance (pour tester $\mathbf{H_0 : \beta_1 = 0}$)

Source	*SS*	*df*	*MS*	*F*
Régression	*SSR*	1	*SSR*/1	*MSR*/*MSE*
Erreur	*SSE*	$n-2$	$SSE/(n-2)$	
Total	*SSTot*	$n-1$		

Tableau 7.2.5.2 Tableau d'ANOVA pour les données de pression et de densité

7.2.6 Calculs statistiques pour la régression linéaire simple : exemple de la pression et de la densité

Les logiciels statistiques simplifient un grand nombre des calculs nécessaires à l'application des méthodes décrites dans cette section. Aucune des méthodes de cette section n'est complexe au point de nécessiter absolument l'utilisation d'un tel logiciel, mais il est utile d'envisager d'en utiliser un dans le contexte de la régression linéaire simple. Apprendre où trouver, sur une capture d'écran typique, les diverses données sommaires correspondant aux calculs effectués dans cette section permet de repérer d'importantes statistiques de synthèse pour les analyses compliquées de courbes et de surfaces présentées dans le prochain chapitre.

La capture d'écran 7.2.6.1 provient d'une analyse en Python dans JupyterLab Notebook des données de l'exemple portant sur la pression et la densité. Le Notebook se trouve sur notre site GitHub à l'adresse suivante : Intro Statistal Methods for Engineering GitHub Site, sous *Part 7A*.

Il est également possible de le consulter et de le télécharger sur le site GitHub spécial pour le chapitre 7.

Vous pouvez également ouvrir un environnement informatique interactif pour travailler avec le Jupyter Notebook utilisant Python à travers un site Binder en passant par le site GitHub de l'exemple de la partie 2. Cliquez ici pour aller sur le site Binder (qui se trouve à l'adresse https://mybinder.org/v2/gh/Statistical-Methods-for-Engineering/Special-GitHub-Site-Part-2-Example-Percent-Waste-by-Weight-on-Bulk-Paper-Rolls/HEAD).

.

La capture d'écran montre un exemple typique des résumés d'analyses de régression produites par les logiciels statistiques. Le renseignement le plus élémentaire est, bien entendu, l'équation de régression. Vient ensuite un sommaire d'un tableau donnant les coefficients estimés (b_0 et b_1), leurs écarts-types estimés et les rapports t (pour tester si les coefficients β sont égaux à 0). La capture d'écran donne les valeurs d'échelle = MSE_{LF} = ${s_{LF}}^2$ et R^2. Nous montrons également la capture d'écran d'un tableau d'ANOVA. Les niveaux de signification observés sont indiqués pour différentes statistiques de test. Le tableau d'ANOVA est suivi d'un tableau des valeurs de y, de y ajusté, de l'écart-type de y ajusté, des résidus et des résidus normalisés correspondant aux n points de données. Le programme de régression de Statsmodels en Python dispose d'une option qui permet de demander des valeurs ajustées, des intervalles de confidence pour $\mu_{y|x}$ et des intervalles de prévision pour x valeurs d'intérêt. L'aperçu des captures d'écran se termine par les renseignements qui suivent pour la valeur x = 5 000.

Nous vous recommandons de comparer les renseignements figurant sur la capture d'écran 7.2.6.1 avec les divers résultats obtenus dans les exemples du chapitre 7 et de vérifier que ces valeurs sont comparables. Nous poursuivrons notre apprentissage des autres éléments au chapitre 8.

```
The regression equation is
density = 2.375 + 4.867e-05 *pressure
```

```
                  Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.981
Dependent Variable: density          AIC:                -73.0762
Date:               2024-01-30 15:06 BIC:                -71.6601
No. Observations:   15               Log-Likelihood:     38.538
Df Model:           1                F-statistic:        717.1
Df Residuals:       13               Prob (F-statistic): 9.31e-13
R-squared:          0.982            Scale:              0.00039636
------------------------------------------------------------------
              Coef.    Std.Err.      t       P>|t|    [0.025   0.975]
------------------------------------------------------------------
Intercept     2.3750     0.0121   197.0079   0.0000   2.3490   2.4010
pressure      0.0000     0.0000    26.7780   0.0000   0.0000   0.0001
```

```
------------------------------------------------------------------------
Omnibus:                 2.101        Durbin-Watson:           1.682
Prob(Omnibus):           0.350        Jarque-Bera (JB):        0.427
Skew:                    0.137        Prob(JB):                0.808
Kurtosis:                3.780        Condition No.:           15556
========================================================================
```

ANOVA table

```
df    sum_sq   mean_sq           F        PR(>F)
pressure   1.0  0.284213  0.284213  717.060422  9.306841e-13
Residual  13.0  0.005153  0.000396         NaN           NaN
```

```
    pressure  density       Fit  StDev Fit  Residual  St Resid
0       2000    2.486  2.472333   0.008903  0.013667  0.767491
1       2000    2.479  2.472333   0.008903  0.006667  0.374386
2       2000    2.472  2.472333   0.008903 -0.000333 -0.018719
3       4000    2.558  2.569667   0.006296 -0.011667 -0.617705
4       4000    2.570  2.569667   0.006296  0.000333  0.017649
5       4000    2.580  2.569667   0.006296  0.010333  0.547110
6       6000    2.646  2.667000   0.005140 -0.021000 -1.091834
7       6000    2.657  2.667000   0.005140 -0.010000 -0.519921
8       6000    2.653  2.667000   0.005140 -0.014000 -0.727889
9       8000    2.724  2.764333   0.006296 -0.040333 -2.135495
10      8000    2.774  2.764333   0.006296  0.009667  0.511813
11      8000    2.808  2.764333   0.006296  0.043667  2.311982
12     10000    2.861  2.861667   0.008903 -0.000667 -0.037439
13     10000    2.879  2.861667   0.008903  0.017333  0.973403
14     10000    2.858  2.861667   0.008903 -0.003667 -0.205912
```

Predicted new value

	mean	mean_se	mean_ci_lower	mean_ci_upper	obs_ci_lower	obs_ci_upper
0	2,618333	0,005452	2,606554	2,630112	2,573739	2,662927

7.2.7 Tutoriels 6 et 7 – Régression linéaire simple

À ce stade, il est recommandé de faire l'exercice du tutoriel 6 et l'exercice du tutoriel 7 qui se trouvent sur le référentiel GitHub. Le tutoriel 6 vous montrera comment interpréter les différents résultats obtenus lors du traitement d'un modèle MCO avec Python et comment réaliser ce calcul à la main. Le tutoriel 7 vous montrera comment construire un modèle MCO en Python.

Il est fortement recommandé de consulter les fichiers Jupyter Notebook Simple Linear Regression. Vous pouvez les trouver dans la section « How do I do X in Python? ». Les fichiers « Ordinary Least Squares Regression » concernant la régression par les moindres carrés ordinaires et « Goodness of Fit » concernant la justesse de l'ajustement seront particulièrement utiles.

8.0.1 Introduction à la régression multiple et logistique

Les principes de la régression linéaire simple posent les bases de méthodes de régression plus sophistiquées, utilisées dans un grand nombre de domaines complexes. Dans cette section, nous explorerons la régression multiple, qui présente la possibilité d'utiliser plus d'un prédicteur. Les idées de base présentées dans la partie 7 sur la régression linéaire simple seront généralisées pour produire un outil d'ingénierie puissant : la régression linéaire multiple, présentée dans cette section.

La régression multiple est une extension de la régression simple à deux variables au cas où il n'existe qu'une réponse mais de nombreux prédicteurs (notés x_1, x_2, x_3, ...). La méthode est justifiée par des scénarios dans lesquels plusieurs variables peuvent être associées simultanément à une réponse.

8.0.2 Sources de la partie 8

Cette première version de la partie 8 est majoritairement tirée de « Basic Engineering Data Collection and Analysis » de Stephen B. Vardeman et J. Marcus Jobe, un ouvrage placé sous licence CC BY-NC-SA 4.0.

Les modifications apportées concernent la réécriture de certains passages et l'ajout de quelques éléments originaux mineurs, ainsi que le formatage pour la plateforme Pressbook et l'adaptation de la numérotation et de l'imbrication des chapitres. Les Jupyter Notebooks basés sur Python ont été adaptés à partir des exemples du texte, et on trouve des liens pour y accéder tout au long du document.

Cette ressource s'appuie également sur le document « Process Improvement Using Data », disponible ici. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn et sont partagées sous licence CC BY-SA 4.0.

Le contenu des chapitres 8.2.1.1 et 8.2.2.2 est issu de l'ouvrage « Quantitative Research Methods for Political Science, Public Policy and Public Administration : 4th Edition With Applications in R », de *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, et Josie Davis,* consultable à l'adresse https://bookdown.org/ripberjt/qrmbook/. Cet ouvrage est partagé en vertu d'une licence Creative Commons Attribution 4.0 International (CC BY 4.0).

8.1.0 Introduction à la régression linéaire multiple : ajustement des courbes et des surfaces par les moindres carrés

La partie 8.1 couvre dans un premier temps l'ajustement de courbes définies par des polynômes et d'autres fonctions présentant des paramètres linéaires pour les couples de données (x, y). Puis vient l'ajustement des surfaces à des données où la réponse y dépend de plusieurs variables $x_1, x_2, \ldots, x_k$. Dans les deux cas, la discussion soulignera l'utilité de R^2 et des tracés des résidus et abordera la question du choix entre différentes équations de régression. Enfin, nous présenterons quelques mises en gardes pour l'application.

8.1.1 Ajustement des courbes par les moindres carrés

Dans la partie 7.1, une droite permettait de représenter les couples de données pression/densité de manière satisfaisante. Mais dans l'étude sur le béton de cendres volantes, une droite ne convenait pas aux couples de données phosphate d'ammonium/résistance à la compression. Cette section traite dans un premier temps de la possibilité d'ajuster des courbes plus complexes qu'une ligne droite à des données (x, y). À titre d'exemple, une étude sera menée pour trouver une équation plus pertinente pour décrire les données relatives au béton de cendres volantes.

.
L'équation linéaire

.

8.1.1.1 $$y \approx \beta_0 + \beta_1 x$$

.
se généralise naturellement par l'**équation polynomiale**

.

8.1.1.2 $$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

.
L'ajustement par les moindres carrés de l'équation 8.1.1.2 pour un ensemble de n paires est conceptuellement à peine plus difficile que l'ajustement de l'équation 8.1.1.1. La fonction de $k + 1$ variables

.

$$\begin{aligned} y &\approx \frac{1}{2} g \left(t_0 + \frac{1}{60}(x - 1) \right)^2 \\ &= \frac{g}{2} \left(\frac{x}{60} \right)^2 + g \left(t_0 - \frac{1}{60} \right) \left(\frac{x}{60} \right) + \frac{g}{2} \left(t_0 - \frac{1}{60} \right)^2 \\ &= \frac{g}{7200} x^2 + \frac{g}{60} \left(t_0 - \frac{1}{60} \right) x + \frac{g}{2} \left(t_0 - \frac{1}{60} \right)^2 \end{aligned}$$

.
doit être minimisée. En mettant les dérivés partielles de $S(\beta_0, \beta_1, \ldots, \beta_k)$ égales à 0, on obtient l'ensemble des **équations normales** pour ce problème des moindres carrés, généralisant ainsi la paire d'équations de la partie 7.1. Il existe $k + 1$ équations linéaires avec $k + 1$ inconnues $\beta_0, \beta_1, \ldots, \beta_k$. Généralement, il existe un ensemble de solutions unique $b_0, b_1, \ldots, b_k$, qui minimise $S(\beta_0, \beta_1, \ldots, \beta_k)$.

.

Exemple 8.1.1.1 Retour sur les données de cendres volantes

Revenons à l'étude de B. Roth sur les cendres volantes et au tableau 7.1.3.1. Une équation quadratique pourrait être plus représentative des données qu'une équation linéaire. Utilisons le modèle d'équation 8.1.1.2 avec $k = 2$:

8.1.1.3 $$y \approx \beta_0 + \beta_1 x + \beta_2 x^2$$

pour les données du tableau 7.1.3.1. Les captures d'écran 8.1.1.1 et 8.1.1.2 montrent les résultats de Jupyter Notebook

(basé sur Python) pour ce modèle de régression. (Après avoir saisi les valeurs x et y du tableau 8.1.1.2 en deux colonnes dans le data frame, une colonne supplémentaire a été créée en élevant les valeurs de x au carré, créant ainsi la variable x_sqr). Ce Jupyter Notebook basé sur le langage Python est disponible sur le site GitHub du cours.

.

Ce Notebook peut également être consulté depuis le site interactif Binder Site, sur le site GitHub spécial « Fly_Ash Data Example ».

.

L'équation de régression est
y = 1,243e+03 + 382,7 x + -76,66 x_sqr

.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0,849
Dependent Variable:	y	AIC:	212,5036
Date:	2024-02-08 14:22	BIC:	215,1747
Nbr d'observations :	18	Log-Likelihood:	-103,25
Df Model:	2	F-statistic:	48,78
Df Residuals:	15	Prob (F-statistic):	2,73e-07
R-squared:	0,867	Scale:	6747,1

	Coef.	Err. std	t	P>|t|	[0,025	0,975]
Intercept	1242,8929	42,9816	28,9169	0,0000	1151,2798	1334,5059
x	382,6655	40,4297	9,4650	0,0000	296,4916	468,8394
x_sqr	-76,6607	7,7616	-9,8770	0,0000	-93,2041	-60,1173

Omnibus:	2,696	Durbin-Watson:	0,822
Prob(Omnibus):	0,260	Jarque-Bera (JB):	1,446
Skew:	0,386	Prob(JB):	0,485
Kurtosis:	1,845	Nbr de condition :	38

Capture d'écran 8.1.1.1 Ajustement quadratique des données sur les cendres volantes

.

	df	sum_sq	mean_sq	F	PR(>F)
x	1,0	21,376190	21,376190	0,003168	9,558562e-01
x_sqr	1,0	658208,892857	658208,892857	97,554600	5,879309e-08
Residual	15,0	101206,230952	6747,082063	NaN	NaN

Capture d'écran 8.1.1.2 Tableau ANOVA pour l'ajustement quadratique des données sur les cendres volantes.

.

L'équation quadratique d'ajustement est

.

$$\hat{y} = 1242.9 + 382.7x - 76.7x^2$$

.

Sur la figure 8.1.1.1, on a superposé la courbe de régression au diagramme en nuage de points des données (x, y). Bien que la parabole ne soit pas représentative des données de Roth d'une manière tout à fait satisfaisante, elle suit beaucoup mieux la tendance des données que la droite tracée précédemment.

.

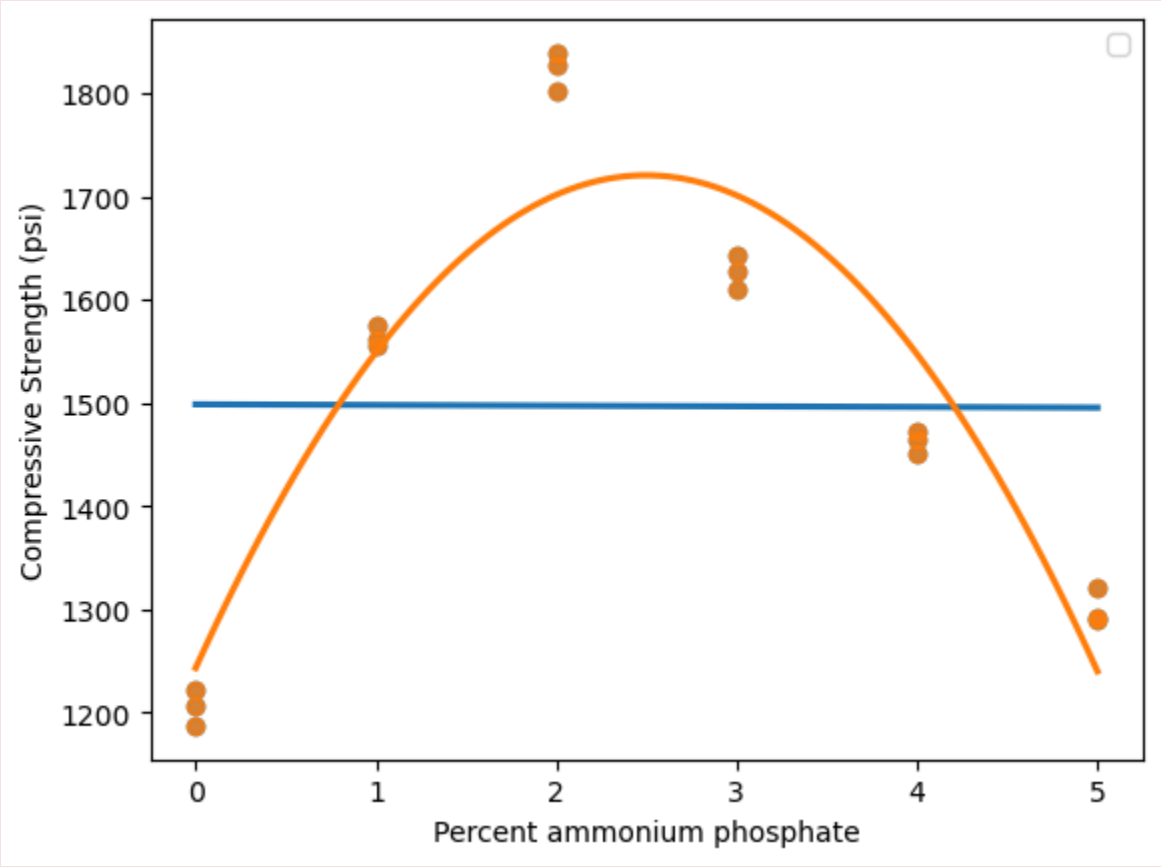


Figure 8.1.1.1 Diagramme en nuage de points, courbe de régression linéaire simple (bleu) et courbe de régression parabolique pour l'exemple des cendres volantes.

.

À la partie précédente, nous avons vu que lorsqu'on représente des données (x, y) par une droite, il est utile de quantifier la qualité de la régression linéaire au moyen de R^2. On peut aussi utiliser le coefficient de détermination lors d'une régression avec un polynôme de la forme de l'équation 8.1.1.2. Rappelons une fois de plus que selon la définition 3,

8.1.1.3 DÉFINITION et expression du coefficient de détermination

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

est la fraction de la variabilité totale en y prise en compte par la courbe de régression. Il est possible de calculer l'équation 8.1.1.3 à la main, mais la manière la plus simple de déterminer R^2 est bien entendu d'utiliser un outil de calcul statistique informatique.

Exemple 8.1.1.2 (suite)

Les captures d'écran ci-dessus montrent que l'équation$\hat{y} = 1242{,}9 + 382{,}7x - 76.7x^2$ donne $R^2 = .867$. Donc 86.7%de la variabilité totale concernant la résistance à la compression est prise en compte par l'équation de régression quadratique. Le coefficient de corrélation entre les valeurs de résistance observées y_i et les valeurs de résistance ajustées $\hat{y}_i$ est $+\sqrt{.867} = .93$.

.

En comparant ce qui a été fait dans cette section à ce qui a été fait dans la partie 7.1, il est intéressant de noter que pour l'ajustement des données sur les cendres volantes par une droite, la valeur de R^2 était de -0,005 (à trois décimales). La régression quadratique constitue une amélioration remarquable par rapport à la régression linéaire pour représenter ces données.

.

Il est naturel de se demander « Et si on utilisait une version cubique de l'équation 8.1.1.2? » Les captures d'écran 8.1.1.3 et 8.1.1.4 présentent quelques résultats issus d'un calcul réalisé pour explorer cette possibilité, et la figure 8.1.1.2 présente la courbe de régression cubique superposée à un nuage de points des données. Les valeurs de x ont été élevées au carré et au cube pour obtenir les valeurs de x, x^2 et x^3 à utiliser pour la régression pour chaque valeur de y.

.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0,942
Dependent Variable:	y	AIC:	196,0175
Date:	2024-02-08 15:34	BIC:	199,5790
Nbr d'observations :	18	Log-Likelihood:	-94,009
Df Model:	3	F-statistic:	93,13
Df Residuals:	14	Prob (F-statistic):	1,73e-09
R-squared:	0,952	Scale:	2588,5

	Coef.	Err. std	t	P>|t|	[0,025	0,975]
Intercept	1188,0503	28,7856	41,2724	0,0000	1126,3113	1249,7892
x	633,1133	55,9134	11,3231	0,0000	513,1910	753,0356
x_sqr	-213,7672	27,7869	-7,6931	0,0000	-273,3642	-154,1701
x_cube	18,2809	3,6491	5,0098	0,0002	10,4544	26,1073

Omnibus:	0,872	Durbin-Watson:	1,068
Prob(Omnibus):	0,647	Jarque-Bera (JB):	0,717
Skew:	0,438	Prob(JB):	0,699
Kurtosis:	2,565	Nbr de condition :	324

Capture d'écran 8.1.1.3 Régression cubique pour les données sur les cendres volantes.

.

	df	sum_sq	mean_sq	F	PR(>F)
x	1,0	21,376190	21,376190	0,008258	9,288806e-01
x_sqr	1,0	658208,892857	658208,892857	254,277093	2,259920e-10
x_cube	1,0	64966,535185	64966,535185	25,097658	1,910288e-04
Residual	14,0	36239,695767	2588,549698	NaN	NaN

Capture d'écran 8.1.1.4 Tableau ANOVA pour l'ajustement cubique des données sur les cendres volantes.

.

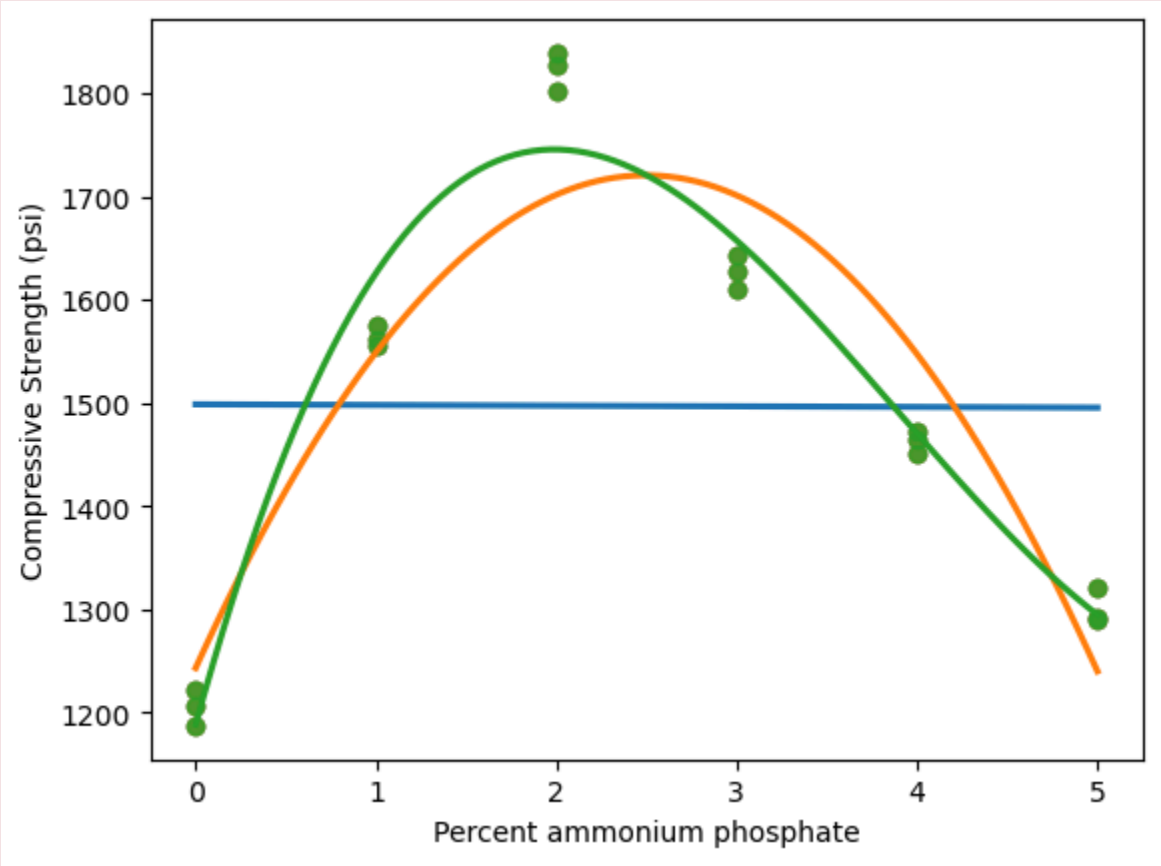


Figure 8.1.1.2 Diagramme en nuage de points et courbe d'ajustement par régression cubique (vert) pour les données sur les cendres volantes.

La valeur obtenue pour R^2 en utilisant une équation du troisième degré est de 0,952, soit un peu plus qu'en utilisant une équation du second degré. Mais la figure 8.1.1.2 montre clairement que même un polynôme du troisième degré ne permet pas d'obtenir une représentation totalement satisfaisante de ces données. Les courbes de régression quadratique et cubique présentées sur les figures 8.1.1.2 et 8.1.1.1 ne s'ajustent pas de manière satisfaisante avec les données de x = 2%. Malheureusement, il s'agit de la zone où la résistance à la compression est la plus importante – exactement la zone présentant le plus grand intérêt d'un point de vue pratique.

Cet exemple illustre le fait que R^2 n'est pas le seul élément à prendre en compte pour évaluer la pertinence du polynôme d'un modèle de régression, et qu'il faut également examiner les graphiques. Les diagrammes en nuage de points de y en fonction de x et les courbes de régression superposées peuvent être utiles, mais les graphiques

des résidus aussi. Ceci peut être illustré avec un ensemble de données où la relation attendue entre y et x est presque parfaitement quadratique.

Exemple 8.1.1.3 Analyse des données lors du lâcher d'une masse

Considérons à nouveau les données issues de la détermination expérimentale de l'accélération due à la gravité (avec la masse en acier), qui sont présentées dans la partie 1, et qui sont reproduites ici dans les deux premières colonnes du tableau 8.1.1.1. Rappelons que les positions y ont été enregistrées à intervalles de $\frac{1}{60}\text{sec}$ à partir d'un instant t_0 inconnu (inférieur à $\frac{1}{60}\text{sec}$) après que la masse soit lâche. Étant donné que la physique newtonienne prévoit que le déplacement de la masse vaut

.

$$\text{déplacement} = \frac{gt^2}{2}$$

.

on s'attend à ce que

.

8.1.1.4

.

$$\begin{aligned} y &\approx \frac{1}{2}g\left(t_0 + \frac{1}{60}(x-1)\right)^2 \\ &= \frac{g}{2}\left(\frac{x}{60}\right)^2 + g\left(t_0 - \frac{1}{60}\right)\left(\frac{x}{60}\right) + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \\ &= \frac{g}{7200}x^2 + \frac{g}{60}\left(t_0 - \frac{1}{60}\right)x + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \end{aligned}$$

.

C'est-à-dire que y est supposé avoir une relation approximativement quadratique avec$x$ et, effectivement, le graphique des paires (x, y) de la figure de la partie 1 semble présenter ce caractère.

Une courte parenthèse : cette expression montre que si la régression des données du tableau 8.1.1.1 est réalisée avec une équation du second degré par la méthode des moindres carrés, on obtient l'expression

.

8.1.1.5 $$\hat{y} = b_0 + b_1 x + b_2 x^2$$

.

et la valeur expérimentale de g (en mm/sec^2) sera égale à $7200b_2$. C'est par cette méthode que la valeur $9{,}79\ \text{m/sec}^2$, présentée à la section 1.4 a été obtenue.

.

$$\hat{y} = 0{,}0645 - 0{,}4716x + 1{,}3597x^2$$

.

(d'où $g \approx 9790\ \text{mm/sec}^2$) avec R^2 valant 1,0, à six décimales près. Dans le cas de cette régression, les résidus peuvent être déterminés en utilisant la définition 8.1.1.3. Ils sont présentés dans le tableau 8.1.1.1. La figure 8.1.1.3 présente un graphique normal des résidus. Il est raisonnablement linéaire et ne présente donc rien de remarquable (à l'exception d'une légère suggestion selon laquelle le ou les deux résidus les plus importants ne sont peut-être pas aussi extrêmes

qu'on pourrait le prévoir, une situation qui ne présente pas d'explication physique apparente).

.

Données, valeurs ajustées, et valeurs résiduelles pour un ajustement quadratique du déplacement de la masse

x, Numéro de point	y, Déplacement	y, Déplacement ajusté	e, valeur résiduelle
1	0,8	0,95	−0, 15
2	4,8	4,56	0, 24
3	10,8	10,89	−0, 09
4	20,1	19,93	0, 17
5	31,9	31,70	0, 20
6	45,9	46,19	−0, 29
7	63,3	63,39	−0, 09
8	83,1	83,31	−0, 21
9	105,8	105,96	−0, 16
10	131,3	131,32	−0, 02
11	159,5	159,40	0, 10
12	190,5	190,21	0, 29
13	223,8	223,73	0, 07
14	260,0	259,97	0, 03
15	299,2	298,93	0, 27
16	340,5	340,61	−0, 11
17	385,0	385,01	−0, 01
18	432,2	432,13	0, 07
19	481,8	481,97	−0, 17
20	534,2	534,53	−0, 33
21	589,8	589,80	0, 00
22	647,7	647,80	−0, 10
23	708,8	708,52	0, 28

Tableau 8.1.1.1 Données, valeurs ajustées, et résidus pour une régression quadratique du déplacement de la masse.

.

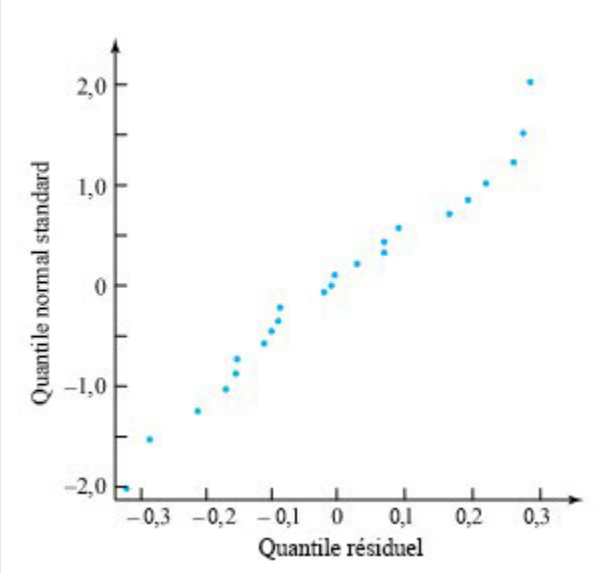


Figure 8.1.1.3 Graphique normal des résidus d'une régression quadratique pour le déplacement de la masse

.

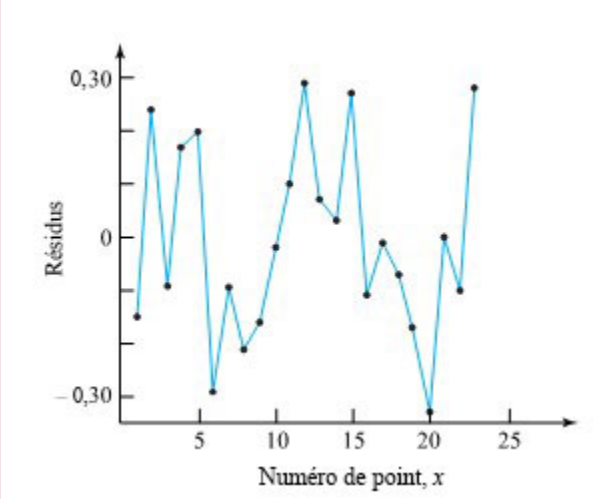


Figure 8.1.1.4 Graphique des résidus pour le déplacement de la masse en fonction du numéro d'observation

.

Toutefois, un graphique des résidus en fonction de x (observations chronologiques) est intéressant. Ce type de graphique est présenté sur la figure 8.1.1.4; les points successifs ont été reliés par des segments de droite. La figure 8.1.1.4 laisse supposer l'existence d'une tendance cyclique dans les résidus. Les valeurs des déplacements observées sont à tour de rôle trop élevées, puis trop basses, puis trop élevées, etc. Il serait intéressant de regarder d'autres rubans expérimentaux pour voir si ce modèle cyclique apparaît de manière constante avant de se pencher sérieusement sur son origine. Mais si la tendance suggérée par la figure 8.1.1.4 réapparaissait systématiquement, cela indiquerait que quelque chose dans le mécanisme qui génère un courant à 60 Hz peut entraîner des cycles alternativement un peu plus courts et

un peu plus longs que $\frac{1}{60}\text{sec}$. Conséquence pratique de cette observation : si une mesure plus précise de altgalttitlegtitle était envisagée, il faudrait prendre en compte la régularité de la variation du courant AC.

QUE FAIRE SI UN POLYNÔME NE PERMET PAS LA RÉGRESSION DE DONNÉES (x, y)?

Les exemples 8.1.1.2 et 8.1.1.3 illustrent (respectivement) une représentation partiellement satisfaisante, puis une autre très satisfaisante, d'un ensemble de données (x, y) au moyen d'une équation polynomiale. Naturellement, des situations telles que celle de l'exemple 8.1.1.3 se présentent parfois, et il est raisonnable de s'interroger sur ce que l'on peut en tirer. Il convient de garder à l'esprit deux choses simples.

.

D'une part, bien qu'un polynôme puisse ne pas représenter de manière satisfaisante la relation entre x et y en tous points, il peut tout à fait être pertinent de manière locale, c'est-à-dire pour une plage relativement restreinte des valeurs de x. Par exemple dans l'étude des cendres volantes, la représentation quadratique de la résistance à la compression comme une fonction du pourcentage de phosphate d'ammonium n'est pas appropriée pour la plage 0 à 5%. Mais la région autour de 2% ayant été identifiée comme une zone d'intérêt particulier, il serait pertinent de mener une étude de suivi en se concentrant (par exemple) sur les données entre 1,5 % et 2,5 % de phosphate d'ammonium.2.5% Il est tout à fait possible qu'une régression quadratique réalisée uniquement pour la plage de données $1.5 \leq x \leq 2.5$ soit satisfaisante et utile pour une synthèse de l'étude de suivi.

.

D'autre part, les termes $x, x^2, x^3, \ldots, x^k$ de l'équation 8.1.1.2 peuvent être remplacés par n'importe quelle fonction (connue) de x et ce que nous avons dit ici restera inchangé pour l'essentiel. Cela nous amène à considérer la transformation de termes pour parvenir à une régression plus satisfaisante.

8.1.2 Transformations

TRANSFORMATIONS POUR LA RÉGRESSION LINÉAIRE

La seconde observation formulée dans le chapitre 8.1.1, concernant les cas où un modèle de régression ne semble pas être satisfaisant, indique que les termes $x, x^2, x^3, \ldots, x^k$ de l'équation 8.1.12 peuvent être remplacés par n'importe quelle fonction (connue) de x sans changer le reste de l'analyse. Les équations normales consisteront toujours en $k+1$ équations linéaires de paramètres $\beta_0, \beta_1, \ldots, \beta_k$, et un programme de régressions linéaire multiple produira toujours les valeurs de moindres carrés $b_0, b_1, \ldots, b_k$. Cela peut toujours être très utile quand il existe des raisons théoriques de supposer que x et y sont liées par une fonction simple mais non linéaire. Par exemple, l'équation de Taylor pour la durée de vie des outils est de la forme

.

$$y \approx \alpha x^{\beta}$$

.

où y est la durée de vie de l'outil (par exemple, en minutes) et x la vitesse de coupe appliquée (par exemple, en m/min). En prenant le logarithme des deux côtés de l'équation, on obtient :

.

$$\ln(y) \approx \ln(\alpha) + \beta \ln(x)$$

.

Il s'agit d'une équation linéaire qui relie $\ln(y)$ et la variable $\ln(x)$, et dont les paramètres $\ln(\alpha)$ et β sont linéaires. Ainsi, à partir d'un ensemble de données (x, y), on détermine les valeurs empiriques de α et β en faisant ce qui suit :

.

1. Prendre les logarithmes des x et des y.
2. Traçer la droite de régression (équation 8.1.1.2).
3. Utiliser $\ln(\alpha)$ comme valeur β_0 (et ainsi αavec $\exp(\beta_0)$) et β comme valeur β_1.

.

TRANSFORMATIONS DE VARIABLES EN MODÉLISATION

Ce cours est une introduction à l'un des thèmes principaux de l'analyse statistique pour l'ingénierie : la découverte et l'utilisation d'une structure simple dans des situations complexes. Cela peut parfois être fait en réexprimant des variables au moyen d'une autre échelle (non linéaire) de mesure que celle qui viendrait à l'esprit en premier lieu. Cela signifie que, parfois, une structure peut ne pas sembler simple en utilisant les échelles de mesure initiales, mais qu'elle peut le devenir après qu'on a transformé une ou plusieurs variables. Cette section présente plusieurs exemples de situations où les transformations sont utiles. Ce faisant, quelques commentaires sur les types de transformations couramment utilisés et sur les raisons plus spécifiques de leur utilisation sont proposés.

.

Transformations et échantillons uniques

Comme discuté dans les parties 3 et 4, il existe plusieurs distributions théoriques standard. Lorsqu'une de ces

distributions peut être utilisée pour décrire une réponse y, tout ce qu'on sait à propos de ce modèle peut servir à faire des prédictions et inférences concernant y. Toutefois, lorsqu'aucune forme de distribution standard ne semble décrire y, il est néanmoins possible de définir une fonction $g(y)$ qui, elle, correspond à une distribution standard.$g(\cdot)$

Exemple 8.1.2.1 Temps de détection

Elliot, Kibby, et Meyer ont mené une étude sur les interventions dans un centre de réparation automobile. Ils ont collecté des données sur ce qu'ils appelaient le « temps de détection » nécessaire pour diagnostiquer les réparations que le personnel allait recommander aux propriétaires des véhicules. Trente temps de détection de ce type (exprimés en minutes) sont présentés sur la figure 8.1.2.1, sous forme d'un diagramme à tige et à feuilles.

.

Les données du diagramme à tige et à feuilles semblent être un peu asymétrique à droite. La plupart des méthodes d'inférence statistique courantes sont basées sur l'hypothèse qu'un mécanisme de génération de données produira à long terme des données symétriques et en forme de cloche; il serait donc critiquable d'utiliser de telles méthodes pour tirer des conclusions et réaliser des prédictions sur les temps de détection dans ce garage. Toutefois, supposons que les temps de détection peuvent être transformés de manière à suivre une distribution en forme de cloche. Les méthodes standard pourraient être utilisées pour tirer des conclusions sur les temps de détections transformés, qui pourraient ensuite être traduits (en inversant la transformation) en conclusions applicables aux temps de détections bruts.

.

Une transformation commune pour raccourcir la queue droite d'une distribution est la transformation logarithmique $g(y) = \ln(y)$. Pour illustrer son utilisation dans le cas présent, des graphiques normaux des temps de détection et du logarithme des temps de détections sont présentés à la figure 8.1.2.2. Ces graphiques montrent qu'Elliot, Kibby et Meyer n'auraient pas pu raisonnablement appliquer les méthodes d'inférence standard aux temps de détection, mais qu'ils auraient pu les utiliser avec les logarithmes des temps de détection. Le second graphique normal est beaucoup plus linéaire que le premier.

.

0	4 3
0	6 5 5 5 6 6 8 9 8 8
1	4 0 3 0
1	7 5 9 5 6 9 5
2	0
2	9 5 7 9
3	2
3	6

Figure 8.1.2.1 Diagramme à tige et à feuilles des temps de détection.

.

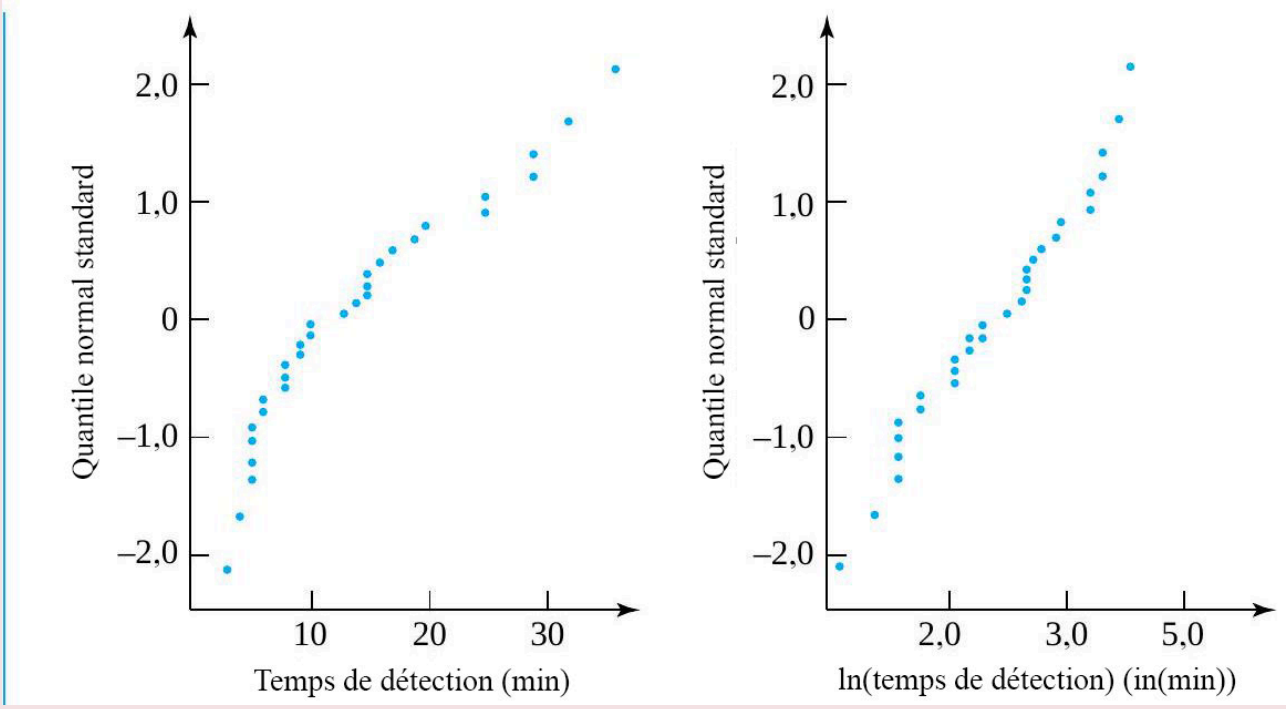


Figure 8.1.2.2 Graphiques normaux des temps de détection et du logarithme des temps de détection.

.
Dans l'exemple précédent, la transformation logarithmique était utile pour réduire l'asymétrie de la distribution de la réponse. Dans les études d'ingénierie statistique, les transformations de puissance constituent un autre type de transformation couramment employé pour modifier la forme d'une distribution de réponse.
.

8.1.2.1 Transformations de puissance $$g(y) = (y - \gamma)^{\alpha}$$

.
Dans la transformation 8.1.2.1, γ constitue généralement une valeur seuil, qui correspond à une valeur minimum de la réponse. α détermine la forme de base du graphique qui représente $g(y)$ en fonction de y. Si $\alpha > 1$1" title= »\alpha>1" class= »latex mathjax »>, la transformation 8.1.2.1 a tendance à allonger la queue droite de la distribution de y. Si <img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/bbfc5cb85d863c86329fad1cca04a7de.png » alt= »0<\alpha<1" title= »0<\alpha, la transformation a tendance à réduire la queue droite de la distribution de y; la réduction devient plus prononcée à mesure que α s'approche de 0, mais elle n'est jamais aussi prononcée que lorsqu'on utilise une transformation logarithmique
.

8.1.2.2 Transformation logarithmique $$g(y) = \ln(y - \gamma)$$

.

Transformations et échantillons multiples

Comparer plusieurs ensembles de conditions d'un processus constitue l'un des problèmes fondamentaux de l'analyse statistique en ingénierie. Il est préférable d'effectuer la comparaison en utilisant une échelle où les échantillons présentent des variabilités comparables, pour au moins deux raisons. Premièrement, les comparaisons sont alors simplement réduites à la comparaison entre les moyennes des réponses. Deuxièmement, les propriétés des méthodes standard d'inférence statistique ne sont souvent bien comprises que lorsque la variabilité des réponses est comparable pour les différents ensembles de conditions.
.
Lorsque la variabilité d'une réponse n'est pas comparable pour des ensembles de conditions différents, une transformation peut parfois être appliquée à l'ensemble des observations pour remédier à ce problème. Cette possibilité d'**effectuer une transformation pour stabiliser la variance** existe lorsque la variance de la réponse est essentiellement une fonction de la moyenne de la réponse. Certains calculs théoriques suggèrent l'utilisation des consignes suivantes comme point de départ pour rechercher une transformation de stabilisation de la variance appropriée :
.

1. Si l'écart-type de la réponse est approximativement proportionnel à la moyenne de la réponse, essayer une transformation logarithmique.
.
2. Si l'écart-type de la réponse est approximativement proportionnel à la puissance δ de la moyenne de la réponse, essayer la transformation 8.1.2.1 avec $\alpha = 1 - \delta$.

.
Quand il y a plusieurs échantillons (et plusieurs valeurs $\bar{y}$ et s), une manière empirique de déterminer si les consignes 1) ou 2) ci-dessus peuvent être utiles consiste à tracer $\ln(s)$ en fonction de $\ln(\bar{y})$, puis à voir s'il existe une linéarité approximative. Si tel est le cas, la consigne 1) est appropriée quand la pente est approximativement égale à 1, tandis qu'une pente de $\delta \neq 1$ indique qu'il faut utiliser la consigne 2) et suggère la valeur à utiliser.

8.1.3 Ajustement des surfaces par les moindres carrés

Il n'y a qu'un pas entre l'idée d'effectuer une régression au moyen d'une droite ou d'une courbe polynomiale et le fait de réaliser que les mêmes méthodes peuvent être utilisées pour faire la synthèse des effets de plusieurs variables quantitatives $x_1, x_2, \ldots, x_k$ sur une réponse y. D'un point de vue géométrique, le problème consiste à effectuer une régression avec une équation du type

.

8.1.3.1 $$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

.

et ainsi ajuster la surface correspondante aux données par la méthode des moindres carrés. C'est ce qui est représenté en trois dimensions sur la figure 8.1.3.1 pour un cas où $k = 2$, avec six points de données (x_1, x_2, y) et une surface d'ajustement possible de la forme 8.1.3.1. Pour faire la régression d'un ensemble de n points de données $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ avec une équation de la forme 8.1.3.1 en utilisant la méthode des moindres carrés, il faut minimiser la fonction de $k + 1$ variables

.

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2$$

.

en choisissant les coefficients $\beta_0, \beta_1, \ldots, \beta_k$. En mettant les dérivées partielles par rapport aux coefficients β égales à 0, on obtient des équations normales, généralisant ainsi les équations de régression linéaire. La résolution de ces $k + 1$ équations linéaires à $k + 1$ inconnues $\beta_0, \beta_1, \ldots, \beta_k$ constitue la première étape d'une régression linéaire multiple. Les coefficients de régression $b_0, b_1, \ldots, b_k$ qui en sont issus minimisent $S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$.

.

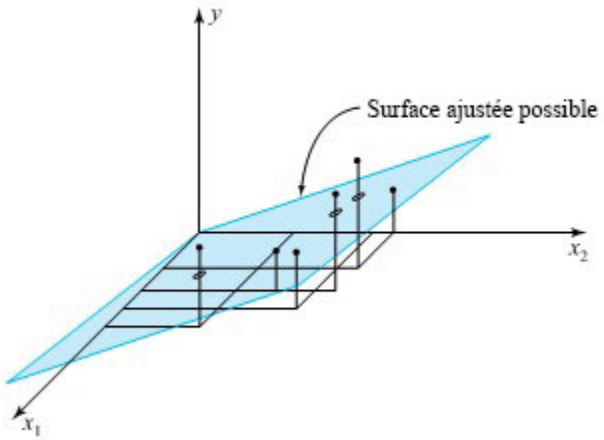


Figure 8.1.3.1 Six points de données (x_1, x_2, y) et une surface d'ajustement possible.

.

Exemple 8.1.3.1 Ajustement par une surface et données de Brownlee sur les pertes dans la cheminée

Le tableau 8.1.3.1 contient une partie d'un ensemble de données sur le fonctionnement d'une usine d'oxydation de l'ammoniac en acide nitrique qui a été publié pour la première fois dans l'ouvrage de Brownlee « Statistical Theory and Methodology in Science and Engineering ». Durant le fonctionnement de l'usine, l'oxyde nitrique produit est absorbé dans une cheminée d'extraction à contre-courant.

.

La variable de flux d'air, x_1, correspond au taux de fonctionnement de l'installation. La variable de concentration en acide, x_3, correspond au pourcentage en circulation moins 50, multiplié par 10. La variable de réponse, y, correspond à 10 fois le pourcentage d'ammoniac entrant qui s'échappe de la cheminée sans être absorbé (essentiellement, c'est une mesure inverse de l'efficacité globale de l'installation). Afin de comprendre, prévoir, et si possible optimiser les performances de l'usine, il serait utile d'avoir une équation décrivant la manière dont y dépend de x_1, x_2, et x_3. L'ajustement des surfaces par les moindres carrés constitue une méthode pour obtenir ce type d'équation empirique.

.

La capture d'écran 8.1.3.1 présente les résultats d'un Jupyter Notebook en Python exécuté pour produire une équation de régression de la forme

.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

.

Données de perte dans la cheminée de Brownlee

i, Numéro d'observation	x_{1i}, Flux d'air	x_{2i}, Température d'admission de l'eau de refroidissement	x_{3i}, Concentration en acide	y_i, Perte dans la cheminée
1	80	27	88	37
2	62	22	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	18	89	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	89	8
13	50	18	86	7
14	50	19	72	8
15	50	19	79	8
16	50	20	80	9
17	56	20	82	15

Tableau 8.1.3.1 Données de Brownlee sur les pertes dans la cheminée

.

L'équation générée par le programme est

.

8.1.3.2 $$\hat{y} = -37.65 + .80x_1 + .58x_2 - .07x_3$$

.

avec $R^2 = 0{,}975$. Les coefficients de cette équation peuvent être vus comme le taux de variation des pertes dans la cheminée en fonction des variables individuelles x_1, x_2, et x_3, si on garde les autres variables constantes. Par exemple, $b_1 = .80$ représente l'augmentation des pertes dans la colonne y qui accompagne une augmentation de flux d'air x_1 d'une unité si la température de l'eau x_2 et la concentration en acide x_3 sont constantes. Les signes des coefficients indiquent si y tend à augmenter ou à diminuer avec l'augmentation des valeurs de x correspondantes. Par exemple, le fait que b_1 soit positif indique que plus l'usine fonctionne à un rythme élevé, plus y tend à avoir une valeur élevée (ce qui signifie que l'usine fonctionne de manière moins efficace). La valeur importante de R^2 est un premier indicateur de l'efficacité de l'équation 8.1.3.2 pour représenter les données.

.

L'équation de régression est
cheminée = -37,65 + 0,80 air + 0,58 eau + -0,07 acide.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0,969
Dependent Variable:	stack	AIC:	59,3440
Date:	2024-02-08 18:15	BIC:	62,6769
Nbr d'observations :	17	Log-Likelihood:	-25,672
Df Model:	3	F-statistic:	169,0
Df Residuals:	13	Prob (F-statistic):	1,16e-10
R-squared:	0,975	Scale:	1,5693

	Coef.	Err. std	t	P>|t|	[0,025	0,975]
Intercept	-37,6525	4,7321	-7,9569	0,0000	-47,8754	-27,4295
air	0,7977	0,0674	11,8282	0,0000	0,6520	0,9434
eau	0,5773	0,1660	3,4786	0,0041	0,2188	0,9359
acid	-0,0671	0,0616	-1,0886	0,2961	-0,2001	0,0660

Omnibus:	0,830	Durbin-Watson:	1,572
Prob(Omnibus):	0,660	Jarque-Bera (JB):	0,523
Skew:	-0,408	Prob(JB):	0,770
Kurtosis:	2,731	Nbr de condition :	1644

Capture d'écran 8.1.3.1 : Régression multiple des données sur les pertes dans la cheminée.

.

	df	sum_sq	mean_sq	F	PR(>F)
air	1,0	775,482188	775,482188	494,160440	9,969916e-12
eau	1,0	18,492672	18,492672	11,784083	4,452801e-03
acide	1,0	1,859634	1,859634	1,185015	2,961071e-01
Residual	13,0	20,400800	1,569292	NaN	NaN

Capture d'écran 8.1.3.2 : Tableau ANOVA des données sur les pertes dans la cheminée pour la régression multiple.

Objectif de la régression multiple

Bien que les techniques de régression de données à plusieurs variables au moyen d'équations de la forme 8.1.3.1 soient relativement simples, le choix et l'interprétation des équations appropriées ne sont pas aussi évidents. Lorsqu'un grand nombre de variables x sont prises en compte, le nombre d'équations potentielles de la forme 8.1.3.1 est considérable. Et pour ne rien arranger, il n'existe pas de méthode totalement satisfaisante pour représenter graphiquement plusieurs variables $(x_1, x_2, \ldots, x_k, y)$ et « voir » la qualité de la régression. Tout ce que nous pouvons faire à ce stade est d'offrir le conseil général de rechercher l'équation de régression la plus simple permettant un ajustement adéquat aux données, puis de fournir des exemples de la manière dont R^2 et le tracé des résidus peuvent constituer des outils utiles pour résoudre les difficultés qui se présentent.

Exemple 8.1.3.2 (suite)

Dans le cas de l'usine de production d'azote, il est pertinent de se demander si les trois variables, x_1, x_2, et x_3, sont nécessaires pour représenter de manière adéquate la variation de y observée. Par exemple, l'évolution des pertes dans la cheminée pourrait être expliquée de manière appropriée en utilisant uniquement une ou deux des trois variables x. Cela aurait plusieurs conséquences pratiques importantes en matière d'ingénierie. Premièrement, dans un tel cas, le processus d'oxydation pourrait être décrit au moyen d'une version simple ou **parcimonieuse** de l'équation 8.1.3.1. Et si une variable ne s'avère pas nécessaire pour prédire y, alors des économies liées à sa mesure peuvent être réalisées. Ou alors, si une variable ne semble pas avoir de réel impact sur y (parce qu'il ne paraît pas essentiel de l'inclure dans l'équation décrivant le comportement de y), il doit être possible de l'ajuster pour des motifs purement économiques, sans crainte de dégrader l'efficacité du processus.

.

Pour déterminer si un sous-ensemble de x_1, x_2, et x_3 permet effectivement d'expliquer le comportement des pertes dans la cheminée, des valeurs de R^2 ont été calculées pour les équations basées sur tous les sous-ensembles possibles de x_1, x_2 et x_3, et elles ont été regroupées dans le tableau 8.1.3.2. Ce tableau montre que, par exemple, 95% de la variabilité totale en y peut être prise en compte à l'aide d'une équation linéaire comprenant uniquement la variable de débit d'air x_1 . L'utilisation de x_1 et de la variable de température de l'eau x_2 permet de rendre compte de 97.3% de la variabilité totale des pertes dans la cheminée. Inclure x_3, la variable de concentration en acide, dans une équation comprenant déjà x_1 et x_2, ne fait passer la valeur de R^2 que de 0,973 à 0,975.

.

R^2 pour les équations de perte dans la cheminée

Équation de régression	R^2
$y \approx \beta_0 + \beta_1 x_1$	950
$y \approx \beta_0 + \beta_2 x_2$	695
$y \approx \beta_0 + \beta_3 x_3$	165
$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$	973
$y \approx \beta_0 + \beta_1 x_1 + \beta_3 x_3$	952
$y \approx \beta_0 + \beta_2 x_2 + \beta_3 x_3$	706
$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$	975

Tableau 8.1.3.2

.

Si l'objectif est d'obtenir une équation de régression simple et correctement ajustée aux données pour les pertes dans la cheminée, le tableau 8.1.3.2 semble indiquer qu'il faut d'abord tenir compte de x_1, puis éventuellement de x_2. À la lumière de ces valeurs de R^2, il semble inutile d'inclure un terme en x_3 dans l'équation de y. Rétrospectivement, ceci est tout à fait cohérent avec le comportement de l'équation de régression 8.1.3.1 : x_3 varie entre 72 et 93 dans la série de données originale, ce qui signifie que la valeur de$\hat{y}$ n'évolue globalement que de

.

$$.07(93 - 72) \approx 1.5$$

.

par rapport à une variation de x_3. (Il faut se rappeler que $.07 = b_3 =$ pente de y en fonction de x_3 dans l'équation de régression.) Une valeur de 1,5 est relativement peu élevée comparée à la plage des valeurs de y observées.

.

Une fois que les valeurs de R^2 ont été utilisées pour identifier les simplifications possibles de l'équation

.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

.

elles peuvent et doivent être intégrées à une analyse approfondie des résidus, avant d'être validées en tant que données de synthèse. À titre d'exemple, considérons une équation de régression en x_1 et x_2. Un programme de régression linéaire multiple peut être utilisé pour produire l'équation de régression suivante :

.

8.1.3.3 $$\hat{y} = -42{,}00 - 0{,}78x_1 + 0{,}57x_2$$

.

(Il faut noter que les valeurs de b_0, b_1, et b_2 de l'équation 8.1.3.3 diffèrent légèrement des valeurs de l'équation 8.1.3.2. En effet, l'équation 8.1.3.3 n'a pas été obtenue en reprenant simplement l'équation 8.1.3.2 et en supprimant le dernier terme. Généralement, les valeurs des coefficients b changent en fonction des variables x qui sont incluses ou non dans la régression.)

Le fait de supprimer une variable dans une équation de régression en modifie généralement les coefficients.

.

Les résidus issus de l'équation 8.1.3.3 peuvent être calculés et représentés de différentes manières potentiellement utiles. La figure 8.1.3.2 présente un graphique normal des résidus et trois autres graphiques des résidus en fonction dex_1, x_2, et $\hat{y}$, respectivement. Les graphiques de

la figure 8.1.3.2 n'apportent pas d'information très significative, sauf peut-être que l'ensemble de données présente une valeur de x_1 inhabituellement élevée ainsi qu'une valeur de $\hat{y}$ inhabituellement élevée (qui elle-même correspond à la valeur élevée de x_1). Toutefois, le tracé des résidus en fonction de x_1 présente une configuration curviligne « croissante-décroissante-croissante » permettant de suggérer l'ajout d'un terme en x_1^2 à l'équation de régression 8.1.3.3.

.

Il convient de vérifier que l'utilisation d'une équation de régression du type

.

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$$

.

pour les données du tableau 8.1.3.1 donne approximativement

.

8.1.3.4 $$\hat{y} = -15{,}409 - 0{,}069x_1 + 0{,}528x_2 + 0{,}007x_1^2$$

.

avec $R^2 = .980$ et des résidus présentant une tendance encore moins marquée que ceux de l'équation de régression 8.1.3.3. Et on remarquera que le signe d'une courbure identifié sur le graphique des résidus en fonction de x_1 pour l'équation 8.1.3.3 n'apparaît pas sur le graphique équivalent pour l'équation 8.1.3.4. Il est intéressant de noter, à travers cet exemple, que l'équation de régression 8.1.3.4 présente une meilleure valeur R^2 que l'équation de régression 8.1.3.2, malgré le fait que l'équation 8.1.3.2 implique la variable de processus

width= »546" height= »454" /> Figure 8.1.3.2 Plots of residuals from a two-variable equation fit to the stack loss data (yˆ =−42.00 − .78×1 + .57×2)[/caption] . Equation (8.1.3.4) is somewhat more complicated than equation (8.1.3.3). But because it still really only involves two different input **Formula does not parse**x" class="latex mathjax"> . Figure 8.1.3.2 Plots of residuals from a two-variable equation fit to the stack loss data (yˆ =−42.00 − .78x1 + .57x2)et qu'elle élimine également la légère tendance observée sur le graphique des résidus de l'équation 8.1.3.3 en fonction de x_1, elle semble être un choix intéressant pour présenter une synthèse des données sur les pertes dans la cheminée. La figure 8.1.3.3 présente un nuage de points 3D des valeurs x_1 et x_2 issues de l'équation de régression 8.1.3.4. La figure 8.1.3.4 présente une vue 2D de la surface d'ajustement définie par l'équation 8.1.3.4. La légère courbure du tracé résulte du terme en x_1^2 qui apparaît dans l'équation 8.1.3.4. Étant donné que x_1 varie entre 50 et 62 et x_2 varie entre 17 et 24 pour la majorité des données, le graphique démontre que sur ces plages de valeurs, x_1 semble influencer les pertes dans la cheminée davantage que x_2. Cette conclusion est cohérente avec la réflexion menée autour du tableau 8.1.3.2.

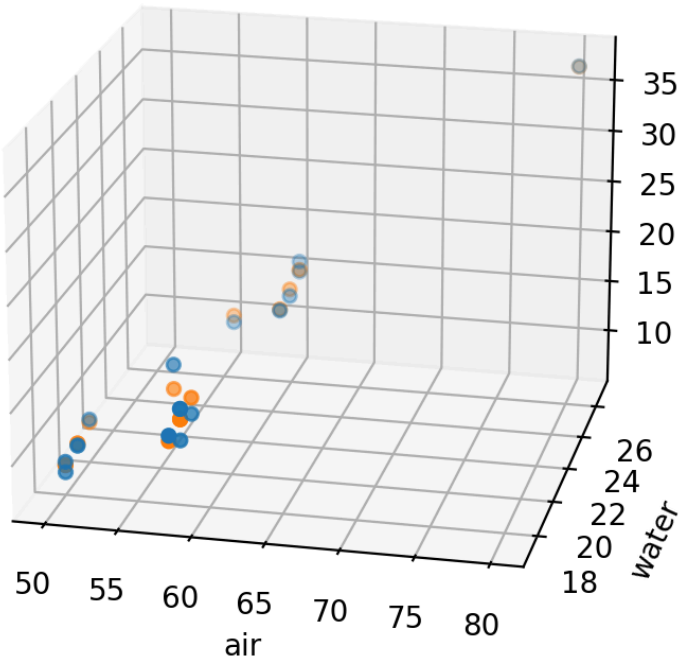


Figure 8.1.3.3 Nuage de points 3D des données issues de l'équation de régression 8.1.3.4.

.

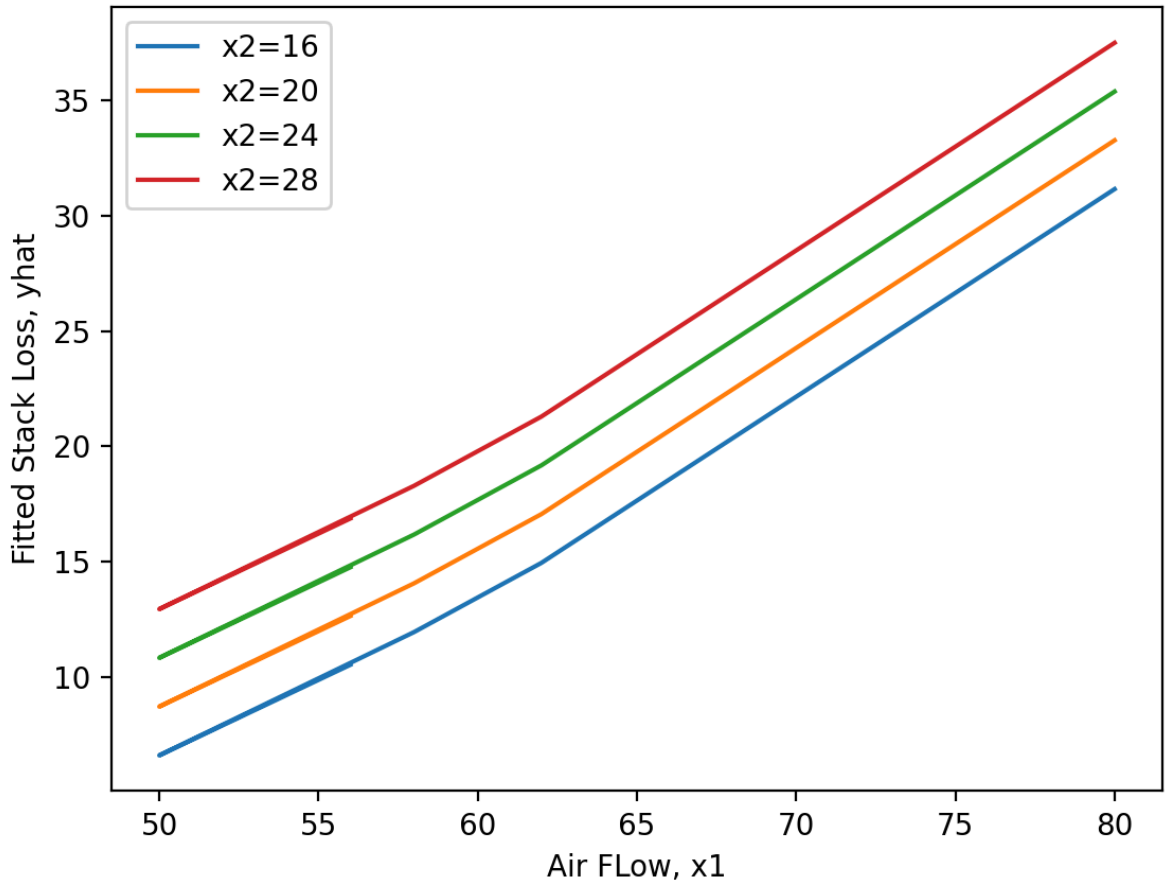


Figure 8.1.3.4 Courbes des pertes dans la cheminée issues de l'équation de régression 8.3.1.4

8.1.4 Tracés résiduels communs en régression multiple

Les tracés résiduels utilisés dans l'exemple 8.1.3 sont typiques. Les voici :

.

1. tracés résiduels normaux

.

2. tracés résiduels en fonction de toutes les variables x

.

3. tracés résiduels en fonction de $\hat{y}$

.

4. tracés résiduels en fonction de l'ordre chronologique d'observation

.

5. tracés résiduels en fonction de variables (p. ex., numéro de la machine ou de l'opérateur) non utilisées dans l'équation d'ajustement mais potentiellement importantes

.

Tous ces types de tracés peuvent être utilisés pour évaluer la qualité de l'ajustement des surfaces à des données à plusieurs variables, et ils ont tous le potentiel de révéler quelque chose de nouveau sur l'ensemble de données ou sur le processus qui les a générées.

8.1.5 Interactions

Précédemment dans cette section, il a été question du fait qu'un « terme en x » dans les équations d'ajustement par les moindres carrés peut être une fonction connue (par exemple, un logarithme) d'une variable de base du processus. En fait, il est souvent utile de permettre à un « terme x » d'être fonction de plusieurs variables de base du processus, comme on le voit dans l'exemple suivant.

Exemple 8.1.5.1 : Rapport portance/traînée pour une configuration à trois surfaces

P. Burris a étudié les effets des positions relatives de l'aile d'un plan canard (une surface portante avant) et de l'empennage sur le rapport portance/traînée pour une configuration à trois surfaces. Une partie de ses données est présentée dans le tableau 8.1.5.1, où

.

$$x_1 = \text{le placement du plan canard en pouces au-dessus du plan de l'aile principale}$$

.

$$x_2 = \text{l'emplacement de l'empennage en pouces au-dessus du plan de l'aile principale}$$

.

(Les positions avant-arrière des trois surfaces sont restées constantes tout au long de l'étude.)

.

Un ajustement direct de l'équation par les moindres carrés

.

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

.

à ces données produit un R^2 de seulement 0,394. Même l'ajout de termes au carré à x_1 et x_2, c'est-à-dire l'ajustement de l'équation

.

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

.

augmente R^2 à seulement 0,513. Cependant, la capture d'écran 8.1.5.1 montre que l'ajustement de l'équation

.

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

.

donne $R^2 = 0{,}641$, pour la fonction suivante :

.

8.1.5.1 $$\hat{y} = 3.4284 + .5361x_1 + .3201x_2 - .5042x_1x_2$$

.

Rapports portance/traînée pour 9 combinaisons de positions plan canard/empennage

x_1, Position du plan canard	x_2, Position de l'empennage	y, Rapport portance/traînée
−1,2	−1,2	0,858
−1,2	0,0	3,156
−1,2	1,2	3,644
0,0	−1,2	4,281
0,0	0,0	3,481
0,0	1,2	3,918
1,2	−1,2	4,136
1,2	0,0	3,364
1,2	1,2	4,018

Tableau 8.1.5.1

.

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0,425
Dependent Variable:	y	AIC:	23,8681
Date:	2024-02-09 12:32	BIC:	24,6570
No. Observations:	9	Log-Likelihood:	-7,9341
Df Model:	3	F-statistic:	2,971
Df Residuals:	5	Prob (F-statistic):	0,136
R-squared:	0,641	Scale:	0,61449

	Coef.	Err. std	t	P>|t|	[0,025	0,975]
Intercept	3,4284	0,2613	13,1208	0,0000	2,7568	4,1001
x1	0,5361	0,2667	2,0103	0,1006	-0,1494	1,2216
x2	0,3201	0,2667	1,2004	0,2837	-0,3654	1,0057
x1:x2	-0,5042	0,2722	-1,8523	0,1232	-1,2038	0,1955

Omnibus:	1,710	Durbin-Watson:	2,194
Prob(Omnibus):	0,425	Jarque-Bera (JB):	0,496
Skew:	0,574	Prob(JB):	0,780
Kurtosis:	2,928	Nbr de condition :	1

Capture d'écran 8.1.5.1 : Régression multiple des données de rapport portance/traînée

.

	df	sum_sq	mean_sq	F	PR(>F)
x1	1,0	2,483267	2,483267	4,041195	0,100611
x2	1,0	0,885504	0,885504	1,441043	0,283737
x1:x2	1,0	2,108304	2,108304	3,430991	0,123185
Residual	5,0	3,072441	0,614488	NaN	NaN

Capture d'écran 8.1.5.2 : Tableau d'analyse de la variance pour la régression multiple des données sur le rapport portance/traînée

.

L'équation de régression est

.

$$y = 3.43 + 0.536x1 + 0.320x2 - 0.504x1 * x2$$

.

(Après avoir lu les valeurs x_1, x_2, et y du tableau 8.1.5.1 dans les colonnes, les produits $x_1 x_2$ ont été créés et y ajustés aux trois variables prédictives x_1, x_2, et $x_1 x_2$ afin de créer cette capture d'écran.)

.

La figure 8.1.5.1 montre la nature de la surface ajustée 8.1.5.1. L'élévation du plan canard (augmentation de x_1) semble avoir des effets visiblement différents sur y en fonction de la valeur de x_2 (la position de l'empennage). (Il semble que le canard et l'empennage ne doivent pas être alignés, c'est-à-dire que x_1 ne doit pas être proche de x_2. Pour maximiser la fonction, x_1 doit être petit si x_2 est grand, et vice-versa.) C'est le terme mixte $x_1 x_2$ dans l'équation 8.1.5.1 qui permet aux courbes de réponse d'avoir des caractères différents pour différentes valeurs de x_2. Sans ce terme, les tranches de la surface ajustée $(x_1, x_2, \hat{y})$ seraient parallèles pour différentes valeurs de x_2, comme dans la situation du module 8.1.4.

.

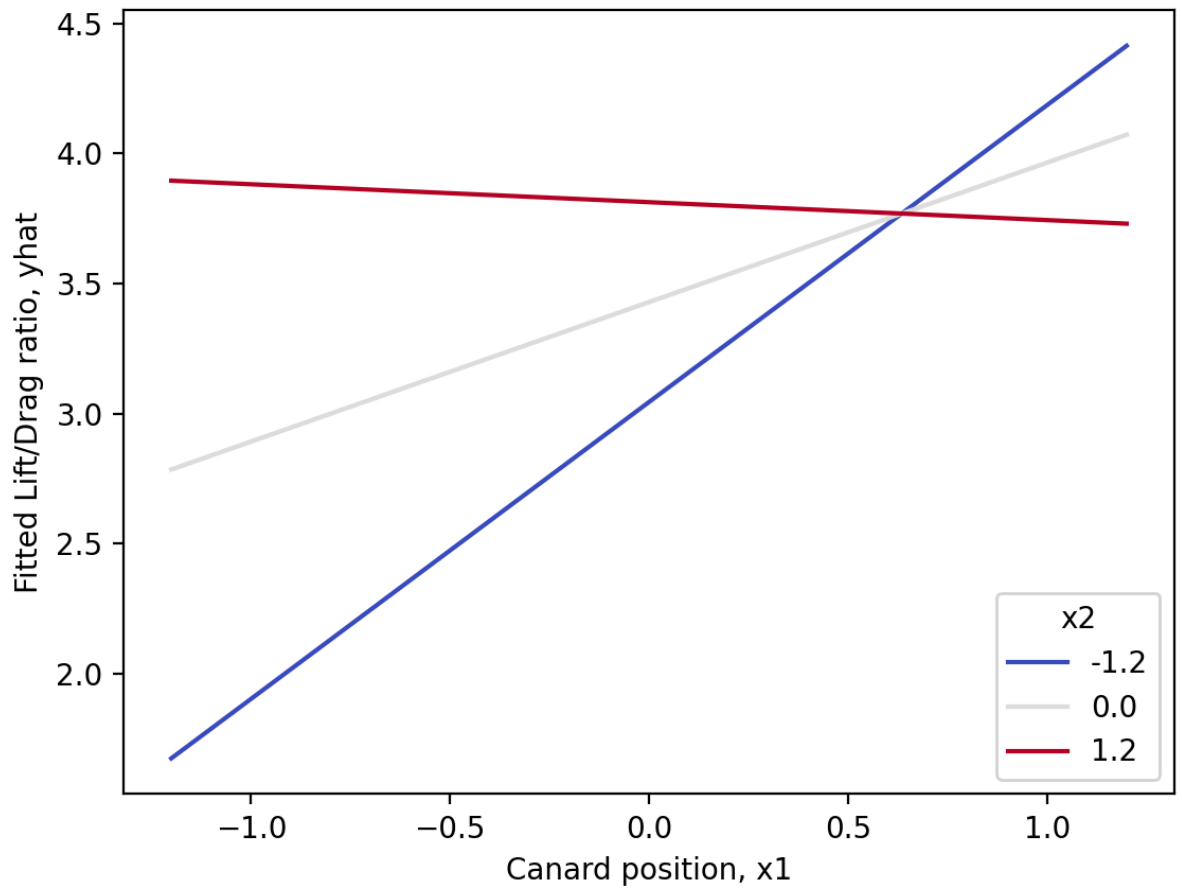


Figure 8.1.5.1 : Tracés de la portance/traînée ajustée de l'équation 8.1.5.1

.

Bien que le principal point d'intérêt de cet exemple ait été présenté, il convient probablement de mentionner qu'on peut faire mieux que l'équation 8.1.5.1 pour ajuster les données du tableau 8.1.5.1. La figure 8.1.5.2 montre un tracé résiduel de cette équation en fonction de la position du plan canard x_1; on peut observer une forte tendance curvilinéaire. En fait, pour l'équation ajustée

.

8.1.5.2 $$\hat{y} = 3.9833 + .5361x_1 + .3201x_2 - .4843x_1^2 - .5042x_1x_2$$

.

on a $R^2 = 0{,}754$ et des résidus qui semblent généralement aléatoires. En traçant les courbes $\hat{y}$ en fonction de x_1 pour plusieurs valeurs de x_2, on peut voir que l'équation ajustée 8.1.5.2 produit des tranches paraboliques non parallèles de la surface ajustée $(x_1, x_2, \hat{y})$, au lieu des tranches linéaires non parallèles observées sur la figure 8.1.5.1.

.

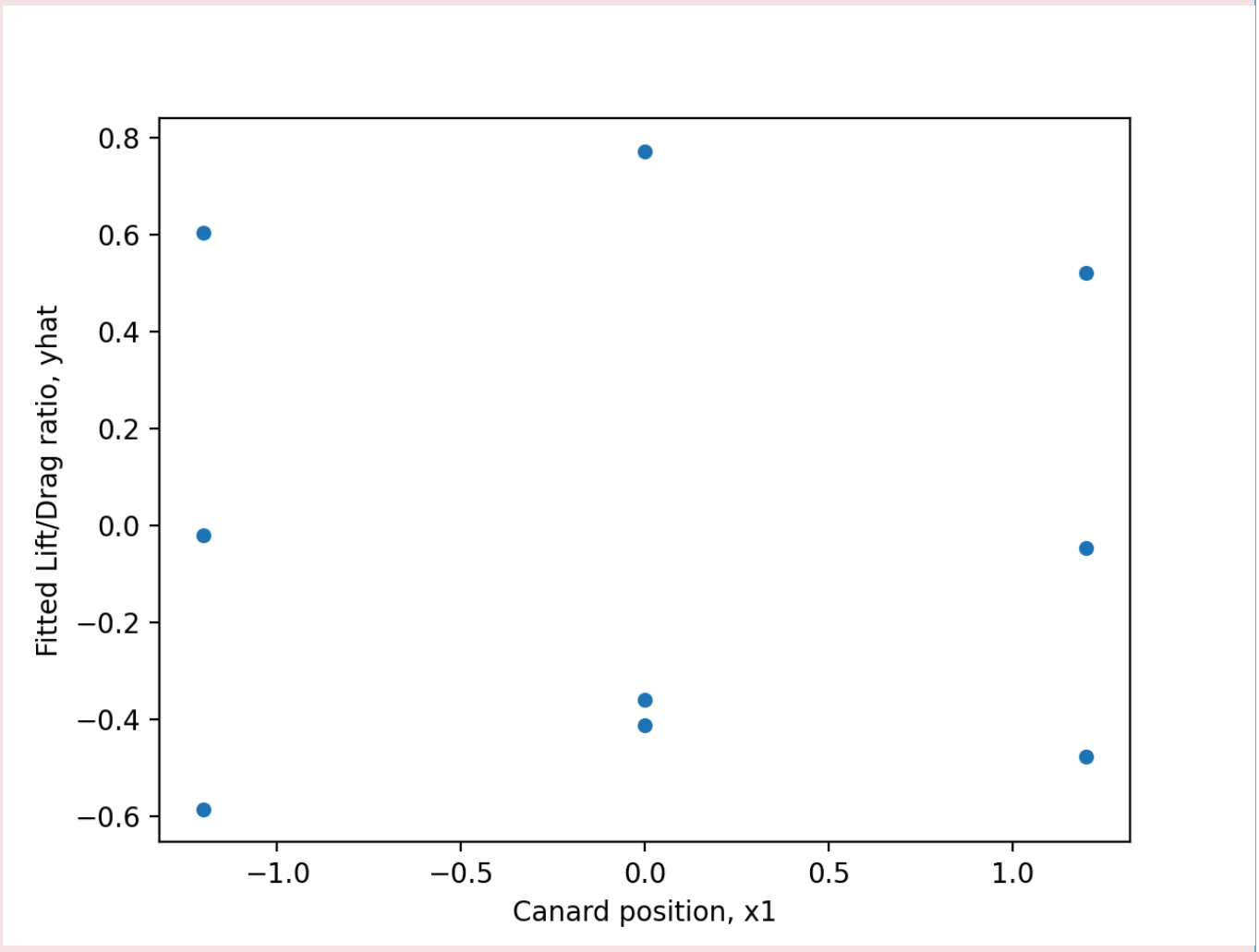


Figure 8.1.5.2 : Tracé résiduel de l'équation 8.1.5.1 en fonction de x1.

.

Cet exemple est disponible dans le Python Jupyter Notebook sur le site GitHub du cours.

Vous pouvez également utiliser le lien Binder suivant pour réviser cet exemple dans un environnement interactif (site GitHub spécial pour l'exemple 8.1.5) : consulter l'exemple 8.1.5 sur Binder.

8.1.6 Quelques précautions additionnelles : extrapolation, valeurs aberrantes et parcimonie

L'ajustement de courbes et de surfaces par la méthode des moindres carrés est d'une importance considérable pour l'ingénierie, mais il faut l'utiliser traité avec soin et discernement. Avant de laisser le sujet jusqu'au module 8.2, où nous verrons les méthodes d'inférence formelle qui lui sont associées, il convient de procéder à quelques mises en garde supplémentaires.

EXTRAPOLATION

Tout d'abord, il est nécessaire de mettre en garde contre les dangers d'une extrapolation substantiellement en dehors de l'étendue des données $(x_1, x_2, \ldots, x_k, y)$. On peut raisonnablement compter sur une équation ajustée pour décrire la relation entre y et un ensemble particulier de valeurs de $x_1, x_2, \ldots, x_k$ uniquement si ces valeurs sont semblables à celles utilisés pour créer l'équation. L'ajustement de surface présente un défi : lorsque plusieurs variables x différentes sont impliquées, il est difficile de dire si un vecteur $(x_1, x_2, \ldots, x_k, y)$ donné est une extrapolation « substantielle ». Tout ce qu'on peut faire, c'est vérifier qu'il est proche d'un point $(x_1, x_2, \ldots, x_k, y)$ de l'ensemble sur toutes les coordonnées. Il ne suffit pas qu'un point ait une valeur x_1 proche de celle qui nous intéresse, puis qu'un autre point ait une valeur x_2 proche de celle qui nous intéresse, etc. Par exemple, le fait d'avoir des données avec $1 \le x_1 \le 5$ et $10 \le x_2 \le 20$ ne signifie pas que la paire (x_1, x_2) (3, 15) est nécessairement proche de n'importe quelle paire de l'ensemble de données. Ce fait est illustré à la figure 8.1.6.1 pour un ensemble fictif de valeurs (x_1, x_2).

.

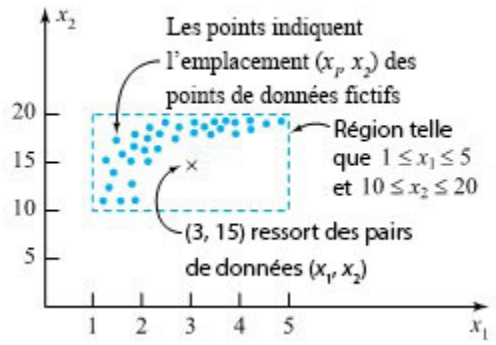


Figure 8.1.6.1 : Graphique hypothétique de paires (x_1, x_2)

.

L'INFLUENCE DES VECTEURS DE DONNÉES ABERRANTES

L'ajustement des courbes et des surfaces par les moindres carrés peut être fortement affecté par quelques données aberrantes ou extrêmes, ce qui constitue un autre piège potentiel. On peut essayer de déceler ces points en examinant les graphiques et en comparant les ajustements réalisés avec et sans le ou les points suspects.

Exemple 8.1.6.1 Données sur la perte dans la cheminée (suite)

On a remarqué à la figure 8.1.3.2 que l'ensemble de données de l'usine d'azote contient un point avec une valeur x_1 extrême. La figure 8.1.6.2 est un nuage de points des paires (x_1, x_2) pour les données du tableau 8.1.3.1. Elle montre que, selon la plupart des normes qualitatives, l'observation 1 du tableau 8.1.3.1 est inhabituelle ou aberrante.

.

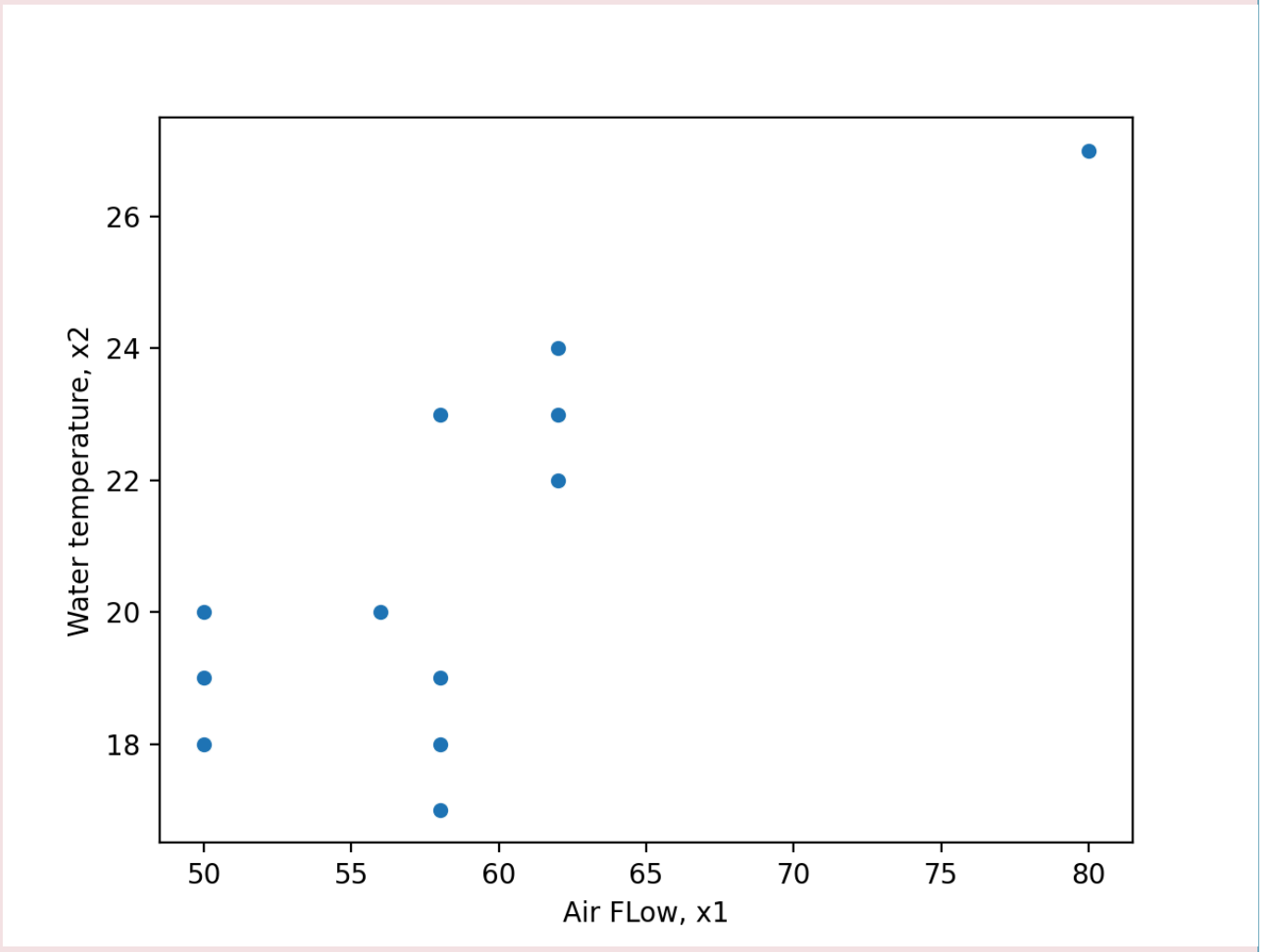


Figure 8.1.6.2

.

Si on refait l'ajustement de l'équation en utilisant uniquement les 16 derniers points de données du tableau 8.1.3.1, on obtient l'équation

.

$$\hat{y} = -56.797 + 1.404x_1 + .601x_2 - .007x_1^2 \quad \textbf{8.1.6.1}$$

.

et $R^2 = 0{,}942$. On pourrait envisager d'utiliser l'équation 8.1.6.1 pour prédire les pertes dans la cheminée et de limiter l'attention à x_1 dans l'étendue de 50 à 62. Il est toutefois possible de vérifier que, même si certains des coefficients

(les b) des équations 8.1.3.4 et 8.1.6.1 diffèrent considérablement, les deux équations produisent des valeurs de $\hat{y}$ comparables pour les 16 points de données pour lesquels x_1 se situe entre 50 et 62. En fait, la plus grande différence entre les valeurs ajustées est d'environ 0,4. Par conséquent, étant donné que le point 1 du tableau 8.1.3.1 ne modifie pas radicalement les prédictions faites à l'aide de l'équation ajustée, il est logique de le garder, d'adopter l'équation (8.1.3.4) et de l'utiliser pour décrire les pertes dans la cheminée pour les paires (x_1, x_2) à l'intérieur du modèle de dispersion de la figure 8.1.6.2.

LE RISQUE DE SURAJUSTEMENT

Il faut également souligner que la notion de simplicité (parcimonie) de l'équation n'est pas seulement importante pour des raisons de simplicité d'interprétation et de réduction des coûts liés à l'utilisation de l'équation. Elle est également importante dans la mesure où elle permet d'obtenir une interpolation régulière tout en évitant de **surajuster** un ensemble de données. Prenons l'exemple de données artificielles, généralement linéaires (x, y). On pourrait trouver un polynôme de degré $k = 10$ à travers chacun de ces points. Mais dans la plupart des problèmes physiques, une telle courbe serait beaucoup moins efficace pour prédire les valeurs de y correspondant à d'autres valeurs de x qu'une simple ligne ajustée. Un polynôme du 10^{e} degré serait surajusté.

MODÈLES EMPIRIQUES ET INGÉNIERIE

Pour conclure cette section, examinons comment les méthodes abordées ici s'inscrivent dans le cadre général de l'utilisation de modèles pour résoudre des problèmes d'ingénierie. Il faut reconnaître que les théories de la physique, de la chimie, des matériaux, etc. produisent rarement des équations des formes simples présentées ici. Parfois, des équations pertinentes de ces théories peuvent être réécrites sous ces formes, comme dans le cas de l'équation de Taylor pour la durée de vie des outils, présentée précédemment dans cette section. Mais la majorité des applications d'ingénierie des méthodes de cette section concernent la foule de problèmes pour lesquels il n'existe pas de théorie physique simple et bien connue, et pour lesquels une simple description empirique de la situation serait utile. Dans de tels cas, l'ajustement des courbes et des surfaces par les moindres carrés peut offrir un aperçu ou une estimation éclairée permettant d'établir des descriptions empiriques approximatives de la relation entre une réponse y et les variables d'entrée du système $x_1, x_2, \ldots, x_k$.

8.1.7 Informatique statistique avec Python

Plusieurs des fichiers Jupyter Notebook en Python qui ont été utilisés dans cette partie sur la régression linéaire multiple (RLM) peuvent être consultés et téléchargés sur le site GitHub du cours ou sur les sites GitHub spéciaux pour la partie 8 :

Site GitHub spécial pour 8.1.1 Données sur les cendres volantes, ou aller à l'environnement de programmation Binder du module 8.1.1 Cendres volantes sur le site Binder.

Site GitHub spécial pour 8.1.3 Données sur les pertes dans la cheminée, ou aller à l'environnement de programmation Binder : module 8.1.3 Perte dans la cheminée sur le site Binder.

Site GitHub spécial pour 8.1.5 Données sur la portance/traînée, ou aller à l'environnement de programmation Binder : module 8.1.5 Portance/traînée sur le site Binder.

8.1.8 Tutoriel 8 – Transformations

À ce stade, il est recommandé de travailler sur l'exercice du tutoriel 8 qui se trouve sur le référentiel GitHub. Cet exercice vous apprendra à transformer des données non linéaires afin qu'elles puissent être utilisées avec des modèles linéaires à l'aide de la syntaxe Python.

Il est fortement recommandé de consulter les fichiers Jupyter Notebook sur la régression linéaire simple. Vous pouvez les trouver dans la section « How do I do X in Python? ». Le fichier « Transformations » sera particulièrement utile.

8.1.9 Transition de la régression linéaire simple à la régression linéaire multiple avec Python

La régression linéaire multiple s'appuie sur la régression linéaire simple d'un point de vue conceptuel, mais la génération et l'interprétation des résultats dans Python diffèrent quelque peu.

Pour faciliter cette tâche, il est fortement recommandé de consulter les fichiers Jupyter Notebook sur la régression linéaire multiple. Vous pouvez les trouver dans la section « How do I do X in Python? ». Les fichiers « Transitioning from Simple to Multiple Linear Regression » et « Multiple Linear Regression » seront particulièrement utiles.

8.2.1 Variables catégoriques, variables indépendantes et variables muettes

Jusqu'à présent, nous avons considéré des modèles de moindres carrés ordinaires (MCO) qui incluent des variables mesurées sur des échelles d'intervalle (ou, à la rigueur et avec prudence, sur des échelles ordinales). Cette approche est satisfaisante lorsqu'on dispose de variables pour lesquelles on peut développer des mesures d'intervalles (ou ordinales) valides et fiables. Mais en ingénierie, il est fréquent de devoir tenir compte de concepts qui ne se prêtent pas facilement à une mesure par intervalles, notamment dans de nombreux cas où une variable est dichotomique (p. ex., présence/absence). Dans d'autres cas, il s'agit d'inclure un concept de nature essentiellement nominale, de sorte qu'une observation peut être classée dans un sous-ensemble, mais non mesurée sur une échelle de type « élevé/faible » ou « plus/moins ». Dans de telles situations, on peut utiliser ce que l'on appelle généralement une variable muette, aussi connue sous les noms de variable indicatrice, variable booléenne ou variable catégorique.

.

Qu'est-ce qu'une « variable muette »?

- Une variable dichotomique, qui peut prendre les valeurs 0 ou 1.
- Une valeur de 1 représente la présence d'une qualité, un 0 son absence,
- Les 1 sont comparés aux 0, qui constituent le « groupe de référence ».
- Les variables muettes sont souvent considérées comme une approximation d'une variable qualitative.

.

Les variables muettes permettent de tester les différences de valeur globale de Y pour différents groupes nominaux dans les données. Ce type de test est similaire à un test d'écart moyen pour les groupes identifiés par la variable muette. Les variables muettes permettent de comparer un groupe inclus (les 1) et un groupe omis (les 0). Il est donc important d'indiquer clairement quel groupe est omis et sert de « catégorie de comparaison ».

.

Il arrive souvent qu'il y ait plus de deux groupes représentés par un ensemble de catégories nominales. Dans ce cas, la variable consistera en deux ou plusieurs variables muettes, avec des codes $0/1$ pour chaque catégorie, à l'exception du groupe de référence (qui est omis). Voici quelques exemples de variables catégoriques qui peuvent être représentées dans une régression multiple par des variables muettes :
– groupes de traitements expérimentaux et de contrôle (traitement $= 1$, contrôle $= 0$)
– genre (homme $= 1$, femme $= 0$ ou vice versa)
– race et ethnicité (une variable muette pour chaque groupe, avec un groupe de référence omis)
– lot de produits (une variable muette pour chaque lot de produits avec un lot de référence omis)
– réglage de machine (une variable muette pour chaque type avec un type de référence omis)

.

La valeur du coefficient de la variable muette représente la différence estimée de Y entre le groupe de la variable muette et le groupe de référence. Comme la différence estimée est la moyenne de toutes les observations Y, il faut voir la variable muette comme un changement de la valeur de l'ordonnée à l'origine (A) pour le groupe « muet », ce qui est illustré dans la figure 8.2.1.1. Dans ce graphique, la valeur de Y est fonction de X_1 (une variable continue) et de X_2 (une variable muette). Lorsque X_2 est égal à 0 (le cas de référence), c'est la droite de régression du haut qui s'applique. Lorsque $X_2 = 1$, la valeur de Y est réduite à la droite inférieure. En résumé, on peut estimer que X_2 a un coefficient de régression partiel négatif, comme en témoigne la différence de hauteur entre les deux droites de régression.

.

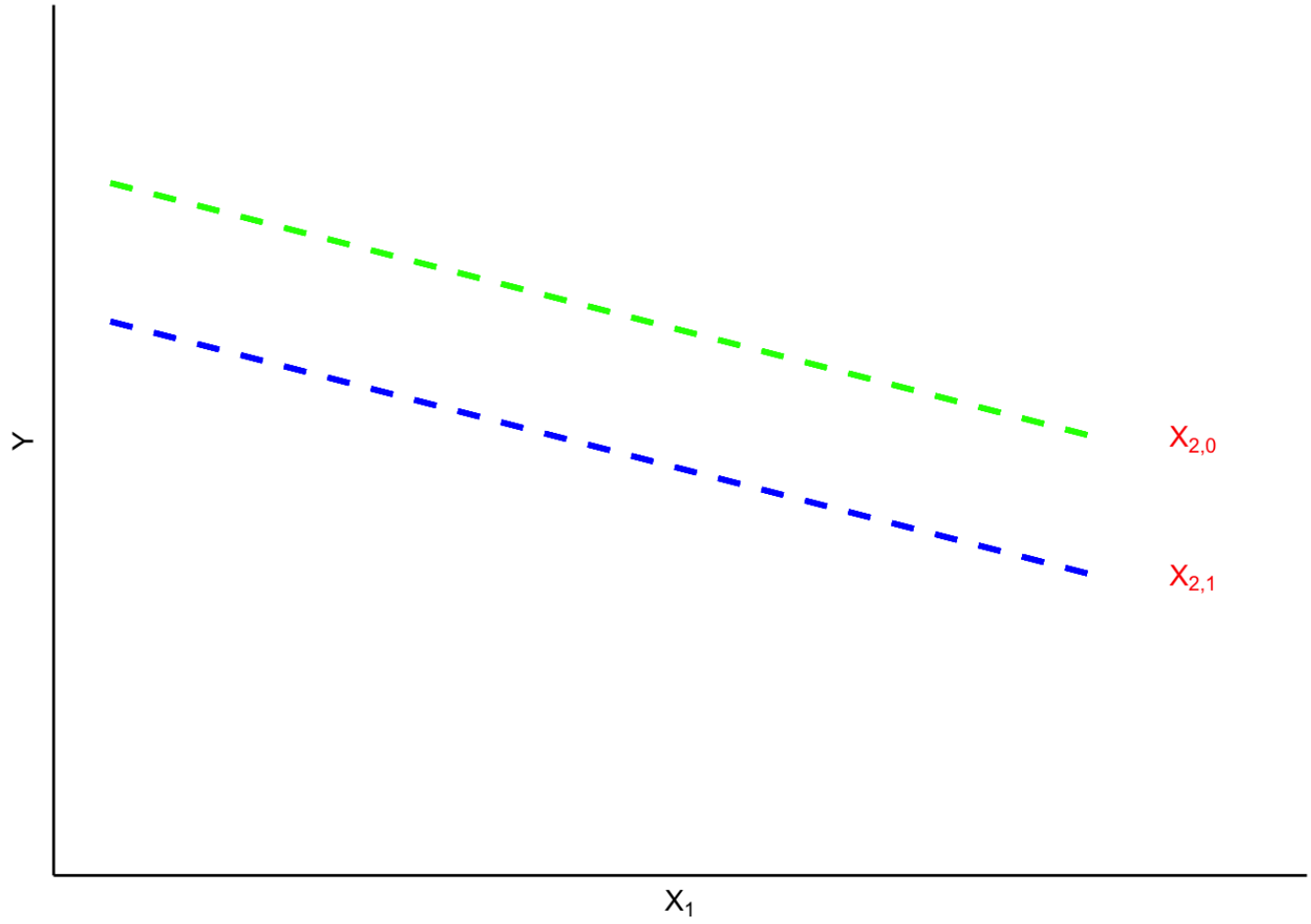


Figure 8.2.1.1 : Variables muettes et ordonnée à l'origine

.

Dans le cas de catégories nominales multiples (par exemple, la région), la procédure est la suivante : choisir la catégorie qui sera le groupe de référence, puis créer une variable muette pour chacune des autres catégories. Par exemple, pour coder un cas à quatre régions (Nord, Sud, Est et Ouest), on peut désigner le Sud comme groupe de référence, puis créer des variables muettes pour les trois autres régions. Ensuite, toutes les observations du Nord obtiendraient une valeur de 1 dans la variable muette Nord, et des 0 dans toutes les autres. De même, les observations relatives à l'Est et à l'Ouest recevraient un 1 dans leur catégorie muette respective et des 0 partout ailleurs. Les observations de la région Sud se verraient attribuer des valeurs de 0 dans les trois catégories. L'interprétation des coefficients de régression partielle pour chacune des trois variables muettes serait alors la différence en Y estimée entre les observations du Nord, de l'Est et de l'Ouest et celles du Sud.

.

EFFETS D'INTERACTION ET VARIABLES MUETTES

Les variables muettes peuvent également être utilisées pour estimer la manière dont l'effet d'une variable diffère

dans les sous-ensembles de cas. Ces types d'effets sont généralement appelés « interactions ». Lorsqu'il y a interaction, l'effet d'un X dépend de la valeur d'un autre. Typiquement, les modèles MCO sont additifs, c'est-à-dire qu'on additionne les B pour prédire Y :

.

$$Y_i = ABX_1BX_2BX_3BX_4E_i.$$

..

Cependant, un modèle d'interaction a un effet multiplicatif où deux des variables indépendantes sont multipliées :

$$Y_i = ABX_1BX_2BX_3 * BX_4E_i.$$

.

Une « variable muette de pente » est un type particulier d'interaction dans lequel une variable muette a une interaction avec (est multipliée par) une variable d'échelle (ordinale ou supérieure). Supposons, par exemple, que l'on ait émis l'hypothèse que les effets de l'idéologie politique sur la perception des risques liés aux changements climatiques sont différents pour les hommes et pour les femmes. Les hommes sont peut-être plus susceptibles que les femmes d'intégrer systématiquement l'idéologie dans la perception des risques liés au changement climatique. Dans un tel cas, une variable muette $(0 =$ femmes, $1 =$ hommes$)$ pourrait être associée à l'idéologie $(1 =$ fortement à gauche, $7 =$ fortement à droite) pour prédire le niveau de risque perçu des changements climatiques ($0 =$ aucun risque, $10 =$ risque extrême). Si l'interaction hypothétique était correcte, on observerait une tendance comme celle illustrée à la figure 8.2.1.2.

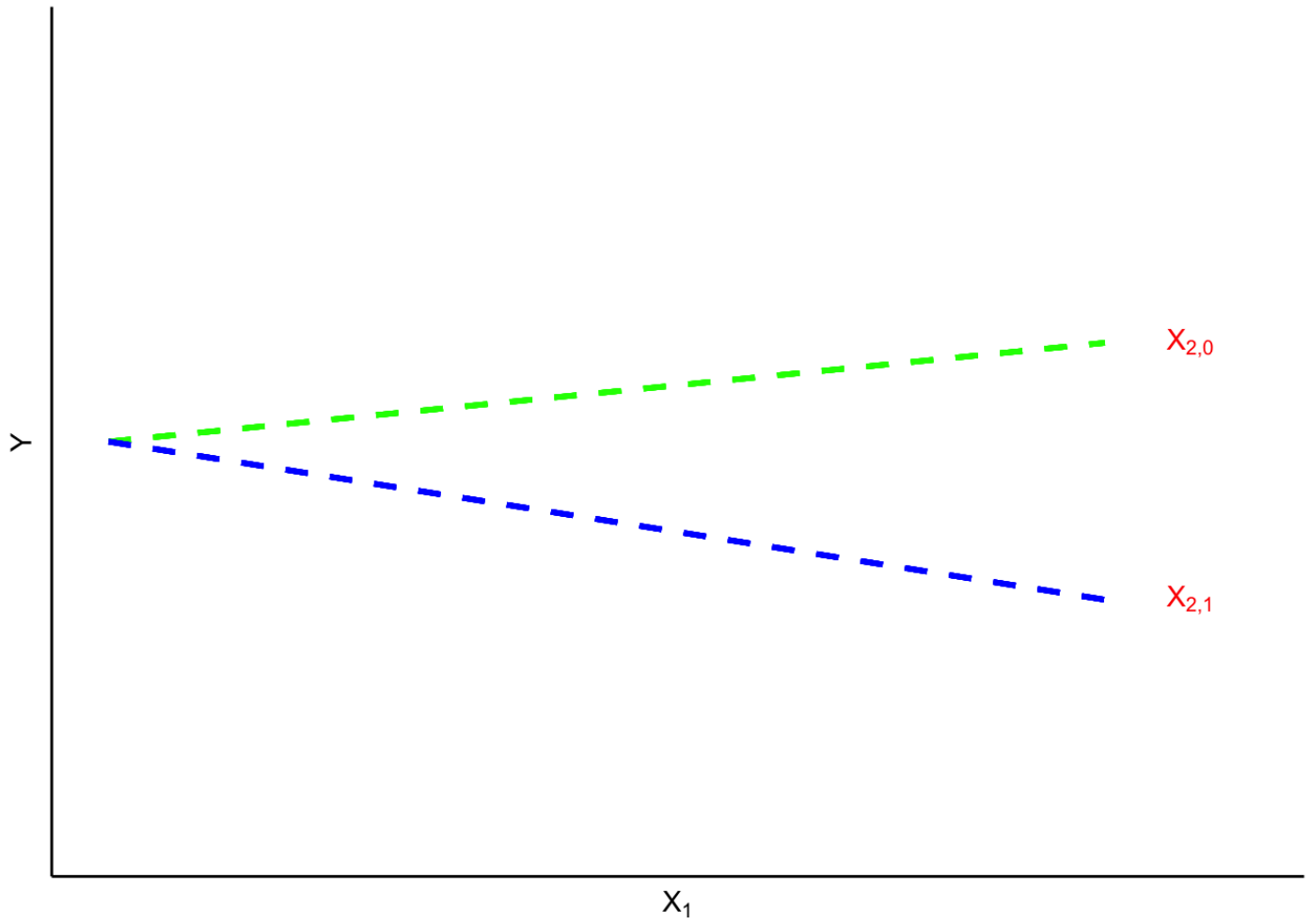


Figure 8.2.1.2 : Illustration d'une interaction de pente

.

En somme, les variables muettes augmentent considérablement la flexibilité des modèles MCO. Elles permettent d'inclure des variables catégoriques et de tester des hypothèses sur les interactions entre les groupes et d'autres variables indépendantes au sein du modèle. Ce type de flexibilité est l'une des raisons pour lesquelles les modèles MCO sont largement utilisés dans le domaine des sciences sociales et de l'analyse politique.

Sources

Le contenu des chapitres 8.2.1.1 et 8.2.2.2 est issu de l'ouvrage Quantitative Research Methods for Political Science, Public Policy and Public Administration : 4th Edition With Applications in R, de *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, et Josie Davis,* consultable à l'adresse https://bookdown.org/ripberjt/qrmbook/.

8.2.2 Algèbre matricielle et régression multiple

L'algèbre matricielle est largement utilisée pour effectuer des régressions multiples, car elle permet de représenter l'analyse de régression de manière compacte et intuitive. Par exemple, un modèle de régression multiple estimé en notion scalaire s'exprime comme suit : $Y = A + BX_1 + BX_2 + BX_3 + E$. En utilisant la notation matricielle, la même équation peut être exprimée sous une forme plus compacte et (étonnamment!) plus intuitive : $y = Xb + e$.

.

En plus, la notation matricielle est flexible puisqu'elle peut traiter n'importe quel nombre de variables indépendantes. Les opérations effectuées sur la matrice modélisant X sont réalisées simultanément sur toutes les variables indépendantes. Enfin, vous verrez que l'expression matricielle est largement utilisée dans les présentations statistiques des résultats de l'analyse des MCO. Pour toutes ces raisons, on commence donc par développer la régression multiple sous forme de matrice.

.

FONDEMENTS DE L'ALGÈBRE MATRICIELLE

Une matrice est un tableau rectangulaire de nombres disposés en lignes et en colonnes. Comme nous l'avons dit, les opérations effectuées sur les matrices le sont simultanément sur tous les éléments de la matrice. Cette section présentera les notions de base de l'algèbre matricielle nécessaires pour comprendre l'expression de la régression multiple sous forme matricielle.

Fondements des matrices

Les nombres composant une matrice sont appelés « éléments ». Les éléments d'une matrice peuvent être identifiés par leur position dans une ligne et une colonne, désignées par $A_{r,c}$. Dans l'exemple suivant, m fait référence à la ligne de la matrice et n fait référence à la colonne.

.

$$A_{m,n} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

.

Par conséquent, dans la matrice suivante,

$$A = \begin{bmatrix} 10 & 5 & 8 \\ -12 & 1 & 0 \end{bmatrix}$$

.

$a_{2,3} = 0$ et $a_{1,2} = 5$.

.

Vecteurs

Un vecteur est une matrice composée d'une seule colonne ou d'une seule ligne. Par exemple :
.

$$A = \begin{bmatrix} 6 \\ -1 \\ 8 \\ 11 \end{bmatrix}$$

.
ou
.

$$A = \begin{bmatrix} 1 & 2 & 8 & 7 \end{bmatrix}$$

Opérations matricielles

Plusieurs « opérations » peuvent être effectuées avec et sur les matrices. La plupart d'entre elles peuvent être calculées à l'aide de Python; nous utiliserons donc cet exemple tout au long du parcours.
Consultez le référentiel spécial GitHub sur le site interactif Binder pour un tutoriel sur la régression linéaire multiple à l'aide d'un ensemble de données de fils qui vous guidera à travers les concepts et l'utilisation des opérations matricielles pour ajuster le modèle.
Comme toujours, ce référentiel peut être téléchargé sur le site GitHub du cours.

Sources

Le contenu des chapitres 8.2.1.1 et 8.2.2.2 est issu de l'ouvrage « Quantitative Research Methods for Political Science, Public Policy and Public Administration : 4th Edition With Applications in R », de *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, et Josie Davis,* consultable à l'adresse https://bookdown.org/ripberjt/qrmbook/.

9.0.2 Sources de la partie 9

La partie 9 de cette ressource éducative libre est composée de textes adaptés de *Process Improvement Using Data*, de Kevin Dunn, utilisés sous licence CC BY-SA 4.0.

Bien que la majorité des textes ait été réécrite avec de nouveaux exemples, le chapitre 5 « Design and Analysis of Experiments » de l'ouvrage de Dunn a été une source d'inspiration importante et certains passages ont été tirés de ce chapitre. La mise en forme pour la plateforme Pressbooks et l'adaptation de la numérotation et de l'imbrication des chapitres ont aussi été effectuées. Des parties de cet ouvrage sont la propriété intellectuelle de Kevin Dunn.

9.0.1 Introduction aux plans d'expériences

Jusqu'à présent, la majeure partie de cette ressource a été consacrée à la détermination des corrélations. Ce chapitre aborde plutôt la question de la causalité, c'est-à-dire la détermination de la relation de *cause à effet*. Pour confirmer l'existence d'un lien *causal* entre les facteurs et un résultat mesurable, il faut perturber et modifier le système. En dépit de l'appellation « plans d'expériences », on doit toutefois noter que ces principes ne s'appliquent pas uniquement aux travaux de laboratoire ou à la recherche appliquée. Les principes de ce module ont une grande portée et peuvent être appliqués à des systèmes aussi simples que la préparation de biscuits qu'à des scénarios complexes comme l'amélioration des processus dans un hôpital ou une usine.

9.1.1 Plans d'expériences : Introduction

9.1.1 PLANS ET ANALYSES D'EXPÉRIENCES EN CONTEXTE

Ce module explique comment perturber volontairement un système pour mieux le comprendre. Les principes présentés dans les modules précédents, en particulier ceux axés sur les tests d'hypothèse et la régression linéaire, seront appliqués ici.

9.1.2 TERMINOLOGIE

Dans un souci de clarté, voici la terminologie commune qui sera utilisée dans cette section (tableau 9.1.2.1) lors de la discussion sur les plans d'expérience.

Tableau 9.1.2.1 Terminologie des plans d'expériences

Terme	Définition
Expérience	Modifier un système et utiliser les informations qui en résultent pour l'améliorer
Objectif	Une amélioration visée
Résultat	Le résultat mesurable de l'expérience
Facteur	Une chose qui peut être activement modifiée pour influencer le résultat
Niveaux	Échelle en fonction des facteurs

Objectifs et résultats

Supposons que l'objectif soit d'améliorer le rendement d'un seul lot de biscuits dans une recette. Un tel objectif pourrait être d'augmenter le nombre de biscuits et le résultat mesuré serait donc le nombre de biscuits. Ou encore, l'objectif pourrait être d'améliorer les propriétés esthétiques des biscuits; le résultat serait alors la couleur des biscuits (par exemple, blanc, brun, brun doré). D'autres exemples sont donnés dans le tableau 9.1.2.2 ci-dessous :

Tableau 9.1.2.2 : Exemples d'objectifs et de résultats pour la préparation de biscuits

Objectif	Résultat mesuré	Résultat quantitatif ou qualitatif
Augmenter le nombre de biscuits	Nombre de biscuits	Quantitatif
Améliorer les propriétés esthétiques des biscuits	Couleur des biscuits	Qualitatif
Réduire le temps de cuisson	Temps de cuisson	Quantitatif
Améliorer le goût	Évaluations des juges de dégustation	Qualitatif

Chaque expérience a généralement un objectif qui combine un résultat et la nécessité d'ajuster ce résultat. Cet objectif peut être d'augmenter, de diminuer ou de maintenir un facteur. Les résultats doivent toujours être mesurables, mais ils peuvent être quantitatifs ou qualitatifs. Sans résultats, aucune analyse n'est possible!

Facteurs

Les facteurs sont l'aspect central des plans d'expériences : ce sont les variables que l'on peut modifier pour

influencer le résultat. Pour réaliser une expérience, il faut modifier au moins un facteur. Comme pour tous les types de données, il existe des facteurs numériques et catégoriques, et la plupart des expériences comportent les deux.

Si l'on reprend l'exemple de la préparation des biscuits, voici quelques facteurs potentiels :

1. la quantité de sucre utilisée dans la recette → *facteur numérique*
2. le type de lait utilisé (lait d'avoine ou d'amande) → *facteur catégorique*
3. le temps passé à mélanger → *facteur numérique*
4. l'utilisation d'un mélangeur automatique ou manuel → *facteur catégorique*

Les facteurs numériques sont quantifiés par la mesure et, de ce fait, il existe un certain ordre implicite. Si l'on prend l'exemple de la quantité de sucre, deux tasses sont supérieures à une tasse. En revanche, les facteurs catégoriques ont un nombre limité de valeurs. Le choix entre le lait d'avoine et le lait d'amande ne fait l'objet d'aucun ordre implicite. On notera toutefois que de nombreux facteurs catégoriques pourraient être convertis en variables numériques continues. Par exemple, la teneur en calcium du lait d'avoine et du lait d'amande peut être différente, soit respectivement 300 et 400 mg de calcium/tasse.

Niveaux

Dans la forme la plus simple du plan d'expériences, chaque facteur ne comporte que deux niveaux, comme dans l'exemple précédent. Les exemples ci-dessus représentent tous des facteurs à deux niveaux : deux tasses ou une tasse de sucre, mélangeur automatique ou manuel, 300 ou 400 mg de calcium/tasse. Le choix des niveaux pour une expérience est une décision importante et repose généralement sur une certaine expertise ou connaissance du système. Dans les expériences relativement complexes, les facteurs peuvent avoir trois, quatre ou même encore plus de niveaux. Ce module se concentre sur les plans à deux niveaux par facteur, car les plans à deux ou trois niveaux par facteur sont les plus courants.

Le choix des niveaux est important. Voici quelques bonnes pratiques pour choisir la fourchette de niveaux :

- La fourchette de niveaux doit être suffisante pour mettre en évidence une différence dans les résultats (mais si elle est trop large, elle risque de ne pas correspondre à un modèle linéaire).
- Il ne faut pas utiliser de valeurs extrêmes pour commencer.
- Il faut perturber le système, mais pas de manière trop granulaire.
- Sans connaissances préalables, une fourchette de 25 % de la norme de fonctionnement est un bon point de départ.

L'exécution d'une expérience se nomme un « essai ». Si l'on exécute huit expériences, on peut dire qu'il y a huit essais dans l'ensemble d'expériences.

9.1.3 EXEMPLE DE PLANS D'EXPÉRIENCES

Supposons que nous gérons une boulangerie et que nous voulons augmenter les profits. Nous proposons de mener une expérience pour déterminer la solution optimale. Dans ce cas, l'analyse a été simplifiée à deux facteurs seulement. Les chapitres suivants aborderont les méthodes permettant de réduire le nombre de facteurs d'une expérience. Ce problème peut être résumé comme suit :

Exemple 9.1.3.1 : Exemple de plan d'expériences

Objectif : Augmenter le profit
Résultat : Profit réalisé en une journée lors de la vente de biscuits
Facteurs : Luminosité du commerce et prix du produit (voir tableau 9.1.3.1 pour les niveaux)

9.1.3.1 Niveaux du plan d'expériences pour l'exemple des biscuits

Facteur	Niveau bas	Niveau élevé
Lumière	Luminosité faible (50 %)	Luminosité forte (75 %)
Prix	7,79 $	8,49 $

Pour mener une expérience, on doit envisager toutes les combinaisons possibles de facteurs. Ces données sont généralement présentées dans un tableau appelé **tableau d'ordre standard.** Les tableaux standards sont généralement donnés avec des valeurs discrètes/codées de 0, -1, 1, etc.

Tableau 9.1.3.2 : Exemple de tableau d'ordre standard

Expérience	Lumière	Prix
1	-1 (faible)	-1 (bas)
2	1 (forte)	-1 (bas)
3	-1 (faible)	1 (élevé)
4	1 (forte)	1 (élevé)

Comme le montre le tableau 9.1.3.2, cet ordre nous aide à déterminer toutes les combinaisons possibles de facteurs qui pourraient être utilisés dans l'expérience. Certains logiciels statistiques sont également conçus pour recevoir des données préparées de cette manière. Si l'on menait ces expériences, le tableau deviendrait le tableau 9.1.3.3, où le profit est le résultat mesuré. Notez la colonne « Essai ». Il est impératif que les expériences soient menées dans un ordre aléatoire afin d'éviter les effets des perturbations (voir la section 9.2.2).

Tableau 9.1.3.3 : Essais expérimentaux pour le plan d'expériences des biscuits

Expérience	Essai	Niveau de luminosité	Niveau de prix	Profit
1	2	Luminosité faible (50 %)	Bas (7,79 $)	490 $
2	1	Luminosité forte (75 %)	Bas (7,79 $)	570 $
3	4	Luminosité faible (50 %)	Élevé (8,49 $)	370 $
4	3	Luminosité forte (75 %)	Élevé (8,49 $)	450 $

La figure 9.1.3.1 permet de visualiser ce tableau et d'en extraire quelques résultats.
Le passage d'une luminosité faible à une luminosité forte augmente les profits, en moyenne, de 80 $:

- En passant d'un éclairage faible à un éclairage fort, la différence de profit au prix bas est de (570 $ – 490)$) = 80 $.
- En passant d'un éclairage faible à un éclairage fort, la différence de profit à un prix élevé est de (450 $ – 370 $) = 80 $.

L'augmentation du prix de 7,79 $ à 8,49 $ diminue le profit, en moyenne, de 120 $:

- Augmenter le prix de 7,79 $ à 8,49 $ sous une lumière faible donne (370 $ – 490 $) = -120 $.
- Augmenter le prix de 7,79 $ à 8,49 $ sous une lumière forte donne (450 $ – 570 $) = -120 $.

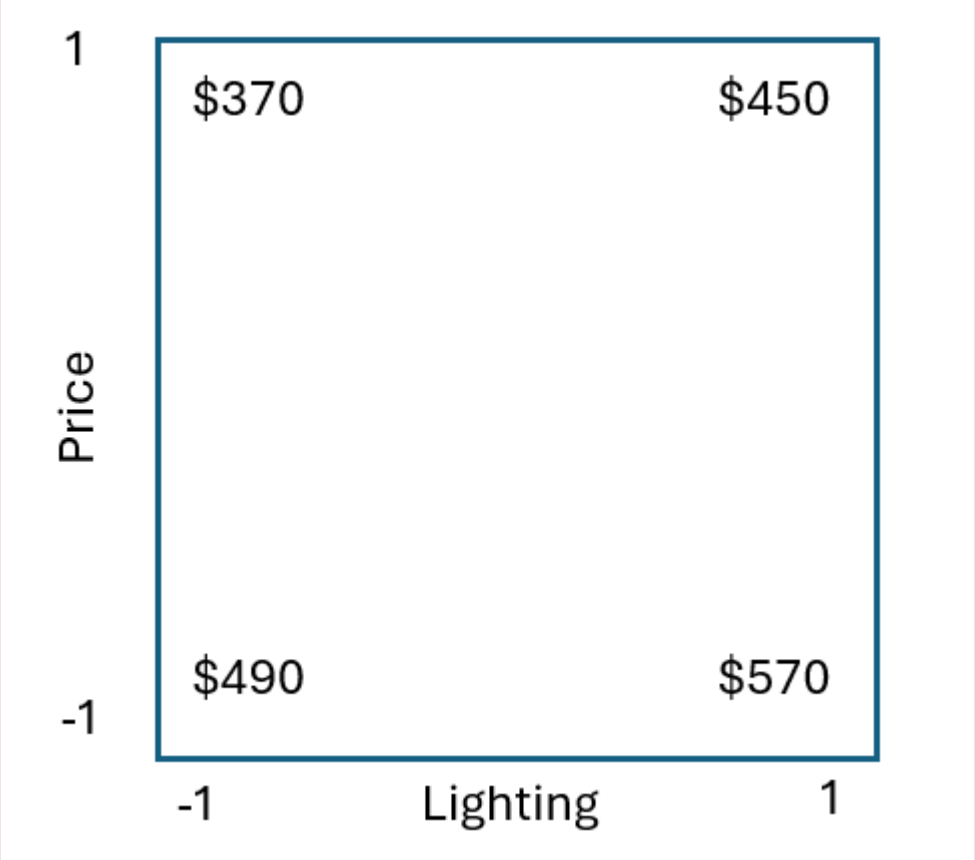


Figure 9.1.3.1 : Graphique visualisant le tableau d'ordre standard. Le profit est indiqué pour les différentes combinaisons de luminosité et de prix.

L'utilisation de plans d'expériences nous permet d'examiner les interactions entre ces facteurs. Plus précisément, on pourrait alors tracer des courbes de niveau entre les différents points de données, ce qui nous permettrait d'obtenir le « centre » (tous les points de données potentiels à l'intérieur du carré) et pas seulement le périmètre. De plus, ce processus peut être étendu à plusieurs facteurs.

9.1.4 POURQUOI UTILISER LES PLANS D'EXPÉRIENCES?

Lorsqu'on envisage d'exécuter un plan d'expériences, on se demande souvent : est-ce vraiment nécessaire? Pour de nombreux systèmes, il existe un grand nombre de données historiques, alors pourquoi ne pas les utiliser? Données existantes = données historiques = données potentiellement fortuites. En l'absence de documentation détaillée, on ne peut pas supposer que les données ont été correctement manipulées. Par conséquent, seules les corrélations établies dans les données peuvent être vérifiées avec certitude. L'expérimentation planifiée est le seul moyen de s'assurer que les événements corrélés sont causaux! En outre, sans plan d'expériences, les expériences sont généralement menées en utilisant des méthodes d'essai et d'erreur, ce qui signifie qu'on ne

modifie qu'un seul facteur à la fois. Les méthodes de plans d'expériences permettent d'atteindre la solution optimale plus rapidement, de manière plus efficace et plus structurée que les méthodes d'essai et d'erreur. Ce point sera expliqué plus en détail dans les chapitres suivants.

9.1.2 Plans d'expériences : Analyse

Comme pour toute expérience, une analyse est nécessaire avant de pouvoir se prononcer sur les résultats. Ce chapitre aborde les méthodes d'analyse des plans d'expériences en utilisant les connaissances acquises dans les modules sur la régression.

9.1.5 ANALYSE DE PLANS D'EXPÉRIENCES

Supposons que nous sommes des ingénieur.e.s en biomatériaux cherchant à améliorer la conception d'un implant dentaire. Nous considérons les effets de la rugosité de la surface et de l'angle de contact avec l'eau sur la viabilité d'un biomatériau potentiel pour cette application. Pour que l'implant soit utile, il doit encourager la croissance de grandes quantités de cellules osseuses (ostéoblastes) à sa surface. Comme dans l'exemple 9.1.3, les tableaux ci-dessous présentent les niveaux (tableau 9.1.5.1), l'ordre standard (tableau 9.1.5.2) et les résultats expérimentaux (tableau 9.1.5.3). Cet exemple peut être résumé comme suit :

Exemple 9.1.5.1 : Analyse de plans d'expériences

Objectif : Augmenter la viabilité de l'implant dentaire
Résultat : Viabilité cellulaire à la surface du matériau prospectif
Facteurs : Rugosité de la surface et angle de contact avec l'eau (voir tableau 9.1.5.1 pour les niveaux)

Tableau 9.1.5.1 : Niveaux du plan d'expériences pour l'implant dentaire

Facteur	Niveau bas	Niveau élevé
Rugosité de la surface	300 µm	350 µm
Angle de contact avec l'eau	50°	100°

Tableau 9.1.5.2 : Tableau d'ordre standard pour l'implant dentaire

Expérience	Rugosité de la surface	Angle de contact avec l'eau
1	-1	-1
2	1	-1
3	-1	1
4	1	1

Tableau 9.1.5.3 : Essais expérimentaux pour les plans d'expériences de l'implant dentaire

Expérience	Essai	Rugosité de la surface	Angle de contact avec l'eau	Viabilité (unités arbit
1	4	Faible (300 µm)	Faible (50°)	31
2	1	Élevée (350 µm)	Faible (50°)	70
3	2	Faible (300 µm)	Élevé (100°)	56
4	3	Élevée (350 µm)	Élevé (100°)	82

À partir de ces quatre essais, on obtient également un point médian, la moyenne, qui est de 59,75 u.a. On peut ainsi découvrir manuellement les principaux effets de la rugosité et de l'angle de contact avec l'eau (voir la figure 9.1.5.1).

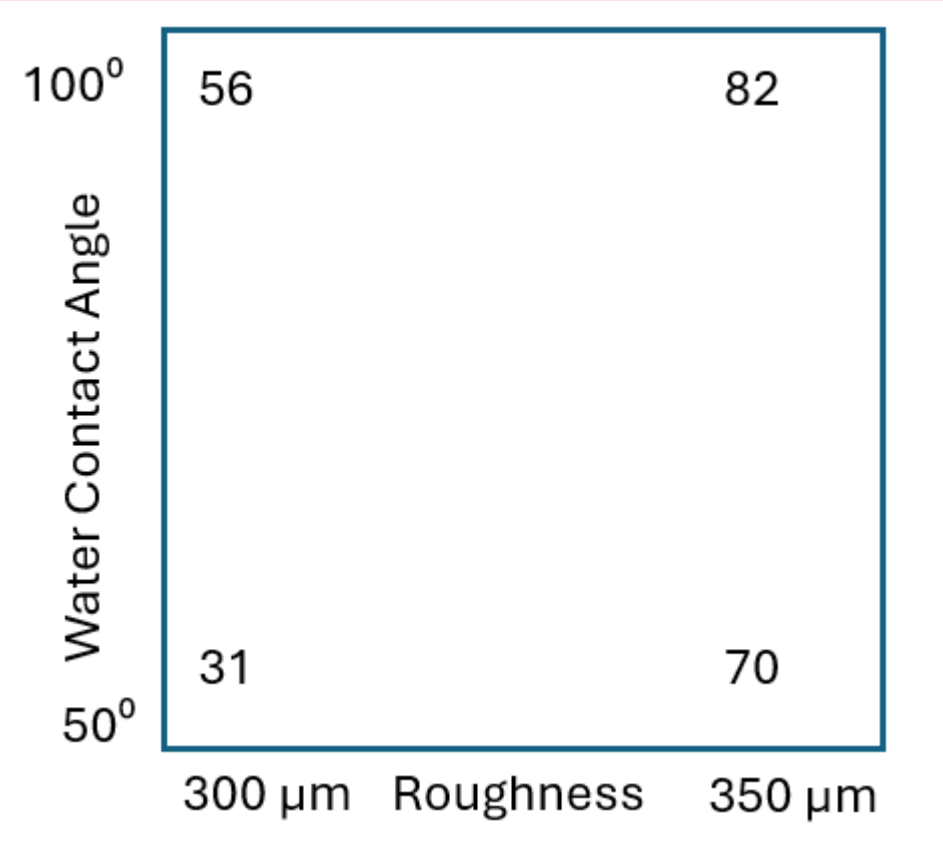


Figure 9.1.5.1 : Graphique visualisant le tableau d'ordre standard. La viabilité cellulaire est indiquée pour les différentes combinaisons de rugosité et d'angle de contact avec l'eau.

Rugosité de la surface

Le passage de 300 à 350 µm de rugosité augmente la viabilité cellulaire, en moyenne, de 32,5 u.a. par 50 µm :

- en passant de 300 à 350 µm de rugosité, la différence de viabilité cellulaire à un angle de contact avec l'eau de 50° donne : (70 – 31) = 39 u.a.
- en passant de 300 à 350 µm de rugosité, la différence de viabilité cellulaire à un angle de contact avec l'eau de 100° donne : (82 – 56) = 26 u.a.

Angle de contact avec l'eau

L'augmentation de l'angle de contact avec l'eau de 50 à 100° diminue la viabilité cellulaire, en moyenne, de 18,5 u.a. par 50° :

- en changeant l'angle de contact avec l'eau de 50 à 100°, la différence de viabilité cellulaire à une rugosité de 300 µm donne : (56 – 31) = 25 u.a.
- en passant d'un angle de contact avec l'eau de 50 à 100°, la différence de viabilité cellulaire à une rugosité de 350 µm donne : (82 – 70) = 12 u.a.

Dans la plupart des logiciels statistiques, ces effets sont considérés comme étant la moitié de ce que nous venons de calculer ci-dessus. Cette différence s'explique par le fait que nous avons codé les niveaux comme si nous allions de -1 à 1, alors que ces niveaux sont considérés mathématiquement comme étant compris entre 0 et 1. Ainsi, les demi-effets rapportés sont :

La rugosité de la surface augmente la viabilité cellulaire, en moyenne, de 16,25 u.a. par 25 µm.

L'angle de contact avec l'eau augmente la viabilité cellulaire, en moyenne, de 9,25 u.a. par 25°.

En utilisant les moindres carrés ordinaires, on peut déterminer que le modèle MCO pour ce système est :

$y = 59{,}75 + 16{,}25x_1 + 9{,}25x_2$

où y correspond à la viabilité cellulaire, x_1 à la rugosité de la surface et x_2 à l'angle de contact avec l'eau.

9.1.6 INTERACTIONS

Comme pour la régression linéaire, il faut tenir compte des interactions doivent dans les plans d'expériences. Rappelons que l'on parle d'interactions lorsque l'effet d'un facteur dépend du niveau d'un autre facteur.

En utilisant l'exemple de l'implant dentaire de la section 9.1.5, des diagrammes d'interaction peuvent être générés pour la rugosité et l'angle de contact avec l'eau (figure 9.1.6.1). Le fait que les deux lignes ne sont pas parallèles indique clairement qu'il existe une interaction entre la rugosité et l'angle de contact avec l'eau. (Les ingénieur.e.s en biomatériaux le savent bien!) En fait, toute interaction doit être symétrique : si la rugosité interagit avec l'angle de contact avec l'eau, l'angle de contact avec l'eau interagit avec la rugosité dans la même mesure.

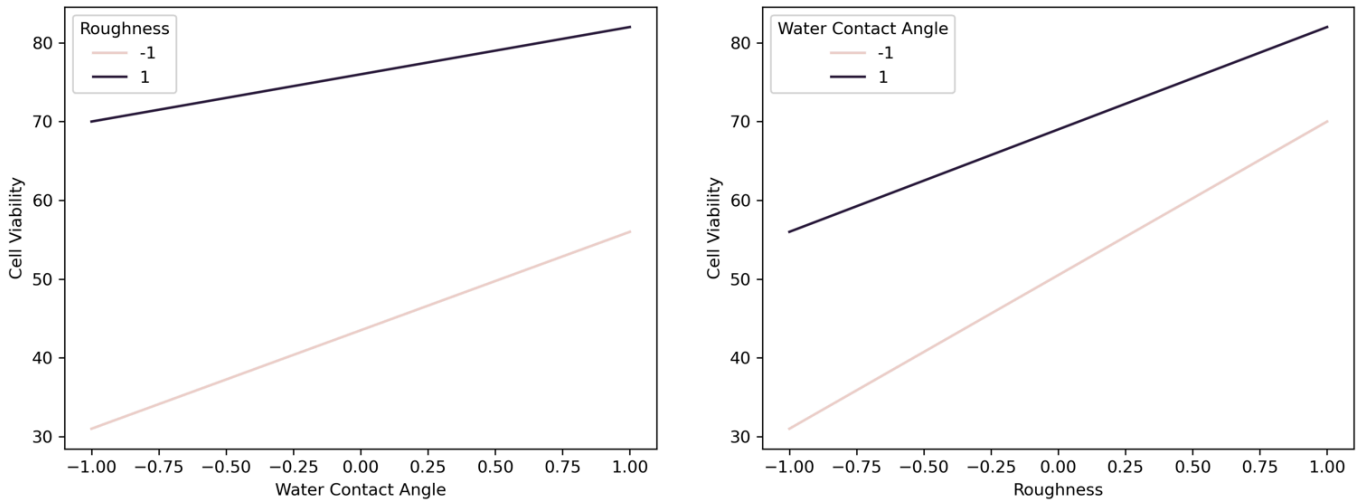


Figure 9.1.6.1 : Diagrammes d'interaction de la rugosité de surface et de l'angle de contact avec l'eau pour l'exemple du plan d'expériences de l'implant dentaire

Si l'on souhaitait calculer les termes d'interaction à la main, on obtiendrait les résultats suivants :

Rugosité de la surface

- à un angle de contact élevé avec l'eau : 82 – 56 = 26 u.a.
- à un angle de contact faible avec l'eau : 70 – 31 = 39 u.a.
- (26 – 39)/2 = -6,5

Angle de contact avec l'eau

- à rugosité élevée : 82 – 70 = 12 a.u.
- à rugosité faible : 56 – 31 = 25 u.a.
- (26 – 39)/2 = -6,5

Terme d'interaction moyen = -6,5/2 = -3,25 u.a.

Rappelons que l'on divise à nouveau par deux parce que nous avons codé les niveaux comme si on allait de -1 à 1, mais que ces niveaux sont considérés mathématiquement comme étant 0 et 1.

En incluant le terme d'interaction, on peut créer le modèle MCO suivant :

$$\mathbf{y = 59{,}75 + 16{,}25x_1 + 9{,}25x_2 - 3{,}25x_3}$$

où y est la viabilité cellulaire, x_1 la rugosité de la surface, x_2 l'angle de contact avec l'eau et x_3 le terme d'interaction.

9.1.7 QUELLE EST LA PROCHAINE ÉTAPE?

Ces expériences ne sont que l'ébauche d'une réflexion qui vous aidera à comprendre votre système. Si vous voulez vraiment l'optimiser, il faudra mener d'autres expériences. Par conséquent, cela nous amène à la question suivante : « Quelle est la prochaine étape? »

Pour être en mesure de répondre à cette question, il faut faire évoluer les niveaux de nos facteurs dans une direction qui optimise notre objectif. Dans le cas de l'exemple utilisé aux sections 9.1.5 et 9.1.6, il s'agirait de modifier les niveaux de rugosité de surface et d'angle de contact avec l'eau dans une direction qui, selon nous, permettra d'améliorer la viabilité des cellules. On peut mieux visualiser ce processus à l'aide d'un graphique des courbes de niveau (figure 9.1.7.1). À partir de ce graphique, nos prochaines expériences pour cet implant dentaire se situeraient dans la partie *supérieure droite* du graphique (c'est-à-dire un angle de contact plus élevé et une rugosité de surface plus importante). L'utilisation de courbes de niveau est utile pour les systèmes à deux ou trois facteurs, mais avec l'augmentation de la complexité, il n'est plus possible de les visualiser. Il faut alors trouver le vecteur pointant dans la direction à suivre pour augmenter le résultat mesuré.

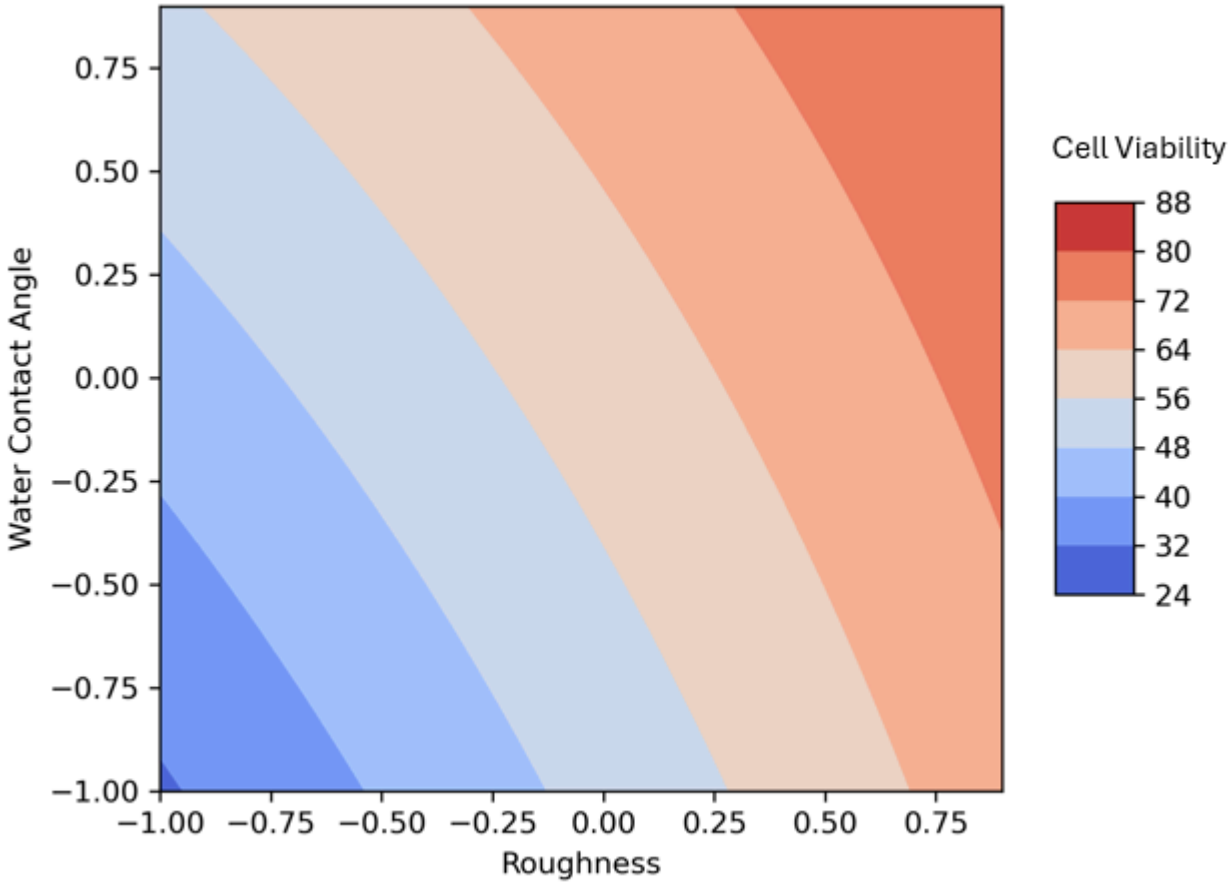


Figure 9.1.7.1 : Graphique des courbes de niveau montrant les interactions entre la rugosité de la surface et l'angle de contact avec l'eau sur la viabilité cellulaire.

9.1.3 Tutoriel 9 – Plan d'expériences

À ce stade, il est recommandé de faire l'exercice du tutoriel 9 qui se trouve sur le référentiel GitHub associé. Cet exercice vous apprendra à importer correctement les données d'un tableau d'ordre standard afin de calculer un modèle MCO en Python.

Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur le plan d'expériences. Vous pouvez les trouver dans la section « How do I do X in Python? ». Le fichier « Full Factorial Example » sera particulièrement utile.

9.2.1 Plans d'expériences : plans factoriels complets

9.2.1 PLANS FACTORIELS COMPLETS

Comme illustré aux sections 9.1.3 et 9.1.5, on peut utiliser un plan d'expériences pour étudier les effets de plusieurs facteurs simultanément. Cette approche est plus efficace pour recueillir des informations sur un système.

En fin de compte, il faut déterminer le nombre d'expériences nécessaires. Selon le nombre de facteurs (k) et le nombre correspondant de niveaux (X), le nombre d'expériences dans un plan factoriel est donné par : $\mathbf{X^k}$.

Dans l'exemple des biscuits de la section 9.1.3, il y avait deux facteurs (luminosité et prix), et chaque facteur avait deux niveaux. Par conséquent, il y avait 2^2 expériences, soit 4 expériences dans ce plan factoriel. Naturellement, cette méthode peut être étendue à trois, quatre ou cinq facteurs (ou même plus), ce qui donne respectivement 8, 16 et 32 expériences en supposant que chaque facteur a deux niveaux. On parle alors de *plans factoriels complets*.

9.2.2 APPLICATION DE LA RÉGRESSION LINÉAIRE AUX PLANS FACTORIELS

Supposons maintenant que l'on applique un modèle de régression linéaire à un plan factoriel pour lequel on dispose de quatre paramètres à estimer et de quatre points de données. Autrement dit, il n'y aura *aucun degré de liberté* et, par conséquent, il n'y aura aucune erreur résiduelle. Cela signifie qu'on ne peut pas effectuer de tests d'hypothèse sur les paramètres ni générer d'intervalles de confiance. Dans la section 9.2.3, nous verrons comment ajuster le plan de manière à obtenir des erreurs résiduelles et à pouvoir calculer les tests d'hypothèse souhaités.

Exemple 9.2.2.1 : Application de la régression linéaire aux plans factoriels

Pour l'instant, en utilisant l'exemple de la section 9.1.5 (voir tableau 9.2.2.1), on peut générer le modèle de régression des moindres carrés suivant pour l'échantillon :

$$y_i = b_0 + b_R x_R + b_W x_W + b_{RW} x_{RW} + e$$

Tableau 9.2.2.1 : Essais expérimentaux pour le plan d'expériences des implants dentaires

Expérience	Essai	Rugosité de la surface	Angle de contact avec l'eau	Viabilité cellulaire unités arbitraires (u.
1	4	−(300 µm)	−(50°)	31
2	1	+(350 µm)	−(50°)	70
3	2	−(300 µm)	+(100°)	56
4	3	+(350 µm)	+(100°)	82

Pour conceptualiser cet ensemble d'expériences, on peut utiliser les matrices ci-dessous (où $\mathbf{x_R}$ = = la rugosité de surface et $\mathbf{x_W}$ = = l'angle de contact avec l'eau) :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & R- & W- & (R-W-) \\ 1 & R+ & W- & (R+W-) \\ 1 & R- & W+ & (R-W+) \\ 1 & R+ & W+ & (R+W+) \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

$$\begin{bmatrix} 31 \\ 70 \\ 56 \\ 82 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

On peut résoudre ce système à l'aide de nos connaissances en régression linéaire. Comme le système est orthogonal, la matrice $(\mathbf{X^T X})$ n'a que des valeurs non nulles sur la diagonale. Par conséquent :

$$\mathbf{b} = (\mathbf{X^T X})^{-}\mathbf{1}(\mathbf{X^T y})$$

$$X^T X = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

La formule n'est pas analysée

Formula does not parse

L'équation résultante, $\mathbf{y = 59.75 + 16.25x_R + 9.25x_W - 3.25x_{RW}}$, peut être interprétée de la même manière que précédemment. Par exemple, une augmentation d'une unité de la rugosité correspond à une augmentation de 16,25 u.a. de la viabilité cellulaire. Cette méthode explique également pourquoi il avait fallu diviser par 2 une deuxième fois précédemment, puisque ce coefficient représente l'effet du passage de la rugosité de surface de 0 à 1, soit de 325 à 350 µm. Il en va de même pour l'angle de contact avec l'eau. Enfin, le terme d'interaction diminue la viabilité cellulaire de 3,25 unités si la rugosité de la surface et l'angle de contact avec l'eau sont au même niveau (tous deux élevés ou tous deux faibles).

9.2.3 DÉTERMINATION DE LA SIGNIFICATION STATISTIQUE

Comme mentionné dans la section précédente, en l'absence de degrés de liberté, aucun test d'hypothèse ou intervalle de confiance ne peut être généré pour les effets principaux ou les termes d'interaction.

Dans le cas d'un plan factoriel complet, il y a plusieurs choix :

1. Effectuer une série complète de répétitions.
2. Ajouter des points centraux.
3. Supprimer les facteurs de faible ampleur ou sans intérêt.
4. Utiliser un motif de confusion ou un plan fractionnaire.

1) Effectuer une série complète de répétitions

Avec des ressources et un temps infinis, ce serait la méthode la plus simple, car on aurait alors plus d'expériences que de paramètres. Cela permettrait d'obtenir les degrés de liberté nécessaires pour calculer l'erreur-type de tous les coefficients du modèle. Toutefois, il s'agit généralement d'une solution inefficace qui monopolise beaucoup de ressources. Il existe de meilleurs choix, mais c'est toujours une option. Une fois que l'on dispose de degrés de liberté, il est possible de déterminer quels coefficients sont peu significatifs et de les supprimer pour obtenir des degrés de liberté supplémentaires.

2) Ajouter des points centraux

Les points centraux sont des paramètres à mi-chemin entre les niveaux d'un facteur donné. En utilisant l'exemple des biomatériaux, on pourrait mener un essai au point central correspondant à 325 µm de rugosité de surface et à 75° d'angle de contact avec l'eau. Cette opération peut être effectuée autant de fois que souhaité, car l'ajout de ces éléments ne modifie pas l'orthogonalité de X et ajoute des degrés de liberté pour faciliter le calcul de l'erreur standard. L'ajout de points centraux est toujours une option viable, car il ne nécessite pas autant d'essais qu'un ensemble de répétitions. Comme dans le cas d'une répétition complète, une fois qu'on a des degrés de liberté, on peut identifier les coefficients qui sont significatifs ou non, puis en supprimer pour obtenir des degrés de liberté supplémentaires. En notation matricielle, on obtiendrait ceci en effectuant trois répétitions :

Exemple 9.2.2.2 Ajout de répétitions pour obtenir des degrés de liberté

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & R- & W- & R-W- \\ 1 & R+ & W- & R+W- \\ 1 & R- & W+ & R-W+ \\ 1 & R+ & W+ & R+W+ \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

3) Supprimer les facteurs de faible ampleur ou sans intérêt

Avec un plan factoriel complet, on peut également choisir de supprimer un coefficient, même si on ne dispose pas d'intervalles de confiance pour étayer ce choix. Si le coefficient est de l'ordre de 0,00001 mais qu'on travaille dans un contexte pratique où les changements dans le système ont des valeurs de l'ordre de 100 ou 1000, il n'est peut-être pas pratique de conserver le coefficient (même s'il est statistiquement significatif), étant donné qu'il n'a que peu d'intérêt pratique ou clinique. La suppression des coefficients de cette manière doit être effectuée avec prudence : il faut connaître le système et tenir compte du contexte pour éviter de commettre une erreur. Comme dans les deux exemples précédents, on obtiendra des degrés de liberté pour calculer l'erreur standard.

Les diagrammes de Pareto (voir figure 9.2.3.1) sont un moyen de visualiser ce concept. En classant les coefficients par ordre décroissant d'importance (à l'exclusion de l'ordonnée à l'origine) et en les représentant sous la forme d'un diagramme à barres, il est possible de déterminer rapidement les coefficients qui ont le plus de répercussions sur le système.

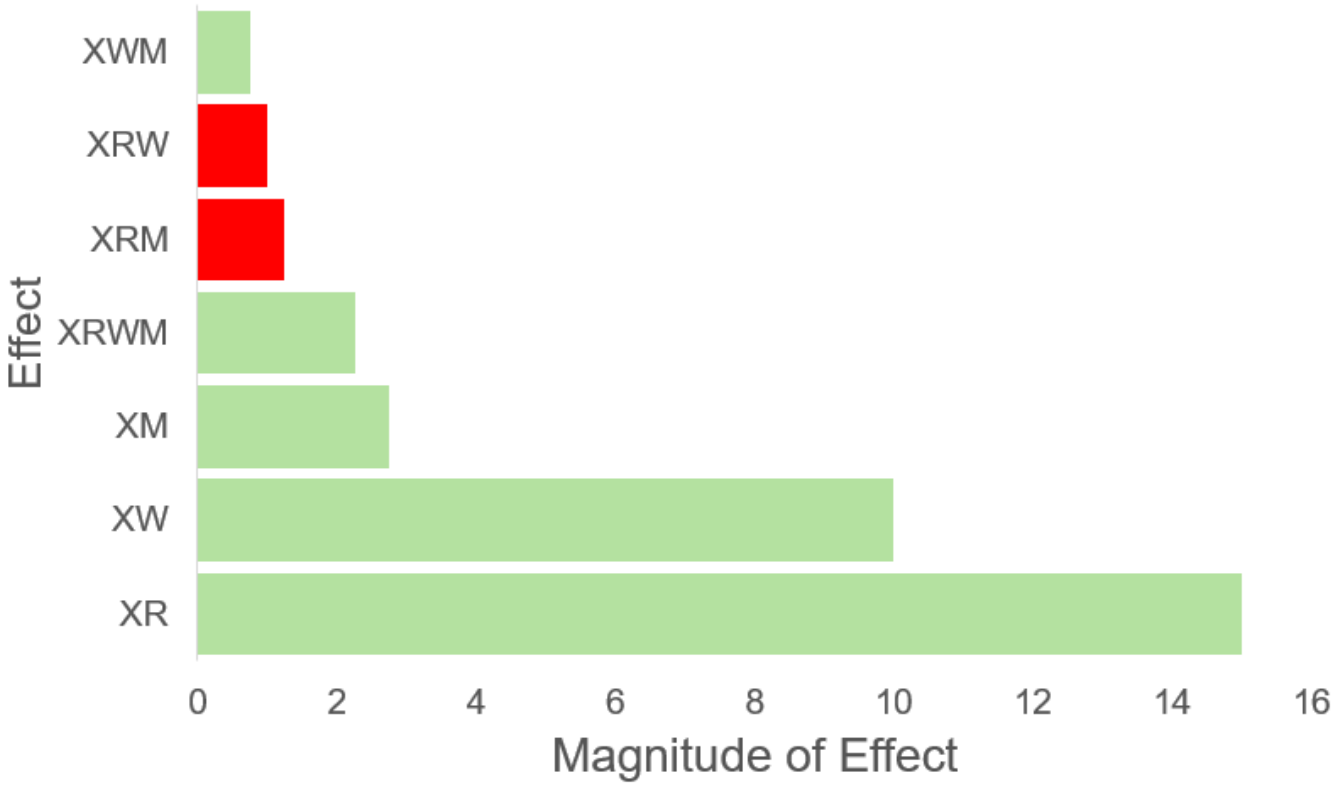


Figure 9.2.3.1 : Diagramme de Pareto pour les plans d'expériences d'implants dentaires

La figure 9.2.3.1 permet de repérer rapidement les coefficients qui ont le plus de répercussions sur les résultats. L'utilisation de couleurs peut également nous permettre de déterminer rapidement quels coefficients ont un effet positif (vert) et lesquels ont un effet négatif (rouge).

4) Utiliser un motif de confusion ou un plan fractionnaire

Voir les sections 9.2.6 et 9.2.7, car il s'agit d'un concept essentiel pour les plans fractionnaires.

9.2.4 AUGMENTER LE NOMBRE DE FACTEURS

Tous les exemples utilisés jusqu'à présent concernaient des situations à deux facteurs. L'augmentation du nombre de facteurs ajoute à la complexité, mais les méthodes et les mathématiques sous-jacentes restent les mêmes.

Exemple 9.2.4.1 : Augmenter le nombre de facteurs

Reprenons l'exemple des biomatériaux et considérons maintenant un troisième facteur, le matériau, car les implants dentaires de l'entreprise ont été conçus pour utiliser soit du titane, soit de l'acier inoxydable. Les tableaux des niveaux (tableau 9.2.4.1), de l'ordre standard (tableau 9.2.4.2) et des résultats expérimentaux (tableau 9.2.4.3) sont présentés ci-dessous.

Tableau 9.2.4.1 : Niveaux du plan d'expériences pour les implants dentaires

Facteur	Niveau faible	Niveau élevé
Rugosité de la surface	300 µm	350 µm
Angle de contact avec l'eau	50°	100°
Matériau	Titane	Acier inoxydable

Tableau 9.2.4.2 : Tableau d'ordre standard pour l'implant dentaire

Expérience	Rugosité de la surface	Angle de contact avec l'eau	Matériau
1	-1	-1	-1
2	1	-1	-1
3	-1	1	-1
4	1	1	-1
5	-1	-1	1
6	1	-1	1
7	-1	1	1
8	1	1	1

Tableau 9.2.4.3 : Essais expérimentaux pour le plan d'expériences des implants dentaires

Expérience	Essai	Rugosité de la surface	Angle de contact avec l'eau	
1	8	-1 (300 µm)	-1 (50°)	-
2	5	+1 (350 µm)	-1 (50°)	-
3	2	-1 (300 µm)	+1 (100°)	-
4	6	+1 (350 µm)	+1 (100°)	-
5	1	-1 (300 µm)	-1 (50°)	ino
6	7	+1 (350 µm)	-1 (50°)	ino
7	3	-1 (300 µm)	+1 (100°)	ino
8	4	+1 (350 µm)	+1 (100°)	ino

Le modèle matriciel correspondant est (où $\mathbf{x_M}$ = = coefficient de matériau) :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$\begin{bmatrix}31\\70\\56\\82\\42\\67\\61\\91\end{bmatrix} = \begin{bmatrix}1&-1&-1&-1&1&1&1&-1\\1&1&-1&-1&-1&-1&1&1\\1&-1&1&-1&-1&1&-1&1\\1&1&1&-1&1&-1&-1&-1\\1&-1&-1&1&1&-1&-1&1\\1&1&-1&1&-1&1&-1&-1\\1&-1&1&1&-1&-1&1&-1\\1&1&1&1&1&1&1&1\end{bmatrix}\begin{bmatrix}b_0\\b_R\\b_W\\b_M\\b_{RW}\\b_{RM}\\b_{WM}\\b_{RWM}\end{bmatrix} + \begin{bmatrix}e_0\\e_R\\e_W\\e_M\\e_{RW}\\e_{RM}\\e_{WM}\\e_{RWM}\end{bmatrix}$$

L'équation résultante est donc la suivante :

$$\mathbf{y = 62{,}5 + 15x_R + 10x_W + 2{,}75x_M - 1x_{RW} - 1{,}25x_{RM} + 0{,}75x_{WM} + 2{,}25x_{RWM}}$$

9.2.2 Plans d'expériences : perturbations et blocage

9.2.5 COMPRENDRE LES PERTURBATIONS

Toute expérience comporte des éléments externes qui peuvent avoir ou auront une influence sur les résultats. Ces éléments se nomment des perturbations. En tant que scientifiques ou ingénieur.e.s, il nous appartient de concevoir des expériences qui réduisent autant que possible l'impact des perturbations.

D'une manière générale, on peut classer les perturbations comme suit :

- connues ou inconnues
- contrôlables ou incontrôlables
- mesurables ou non mesurables

Dans un monde idéal, toutes les perturbations seraient connues, contrôlables et mesurables, mais ce n'est presque jamais le cas. La température ambiante, un changement inattendu sur le marché boursier, le choix d'un.e machiniste, etc. sont des facteurs qui ne peuvent pas être contrôlés ou même planifiés, et c'est la raison pour laquelle la randomisation est si importante. La randomisation permet d'éviter que des perturbations n'affectent systématiquement le résultat.

Une méthode courante pour gérer les perturbations consiste à concevoir l'expérience de manière à en tenir compte. Si la perturbation est contrôlée et maintenue constante pour toutes les expériences, il ne s'agit plus d'une perturbation puisque son effet s'annule. L'appariement peut également annuler l'effet des perturbations en utilisant le même sujet/les mêmes échantillons pour les raisons indiquées à la **partie 5**. On peut classer les facteurs en fonction de leur capacité à être contrôlés ou mesurés (tableau 9.2.5.1). Les covariables sont des paramètres susceptibles de modifier le résultat, mais qui ne nous intéressent pas. La température ambiante en est un exemple. Pour de nombreuses expériences, la température ne présente pas d'intérêt majeur, mais elle pourrait influencer le résultat. Le blocage sera abordé à la section 9.2.6.

		Measurable	
		Yes	No
Controllable	Yes	Factors	Blocking
	No	Covariates	Disturbances

Tableau 9.2.5.1 : Tableau montrant comment classer les facteurs selon qu'ils sont mesurables ou contrôlables.

9.2.6 BLOCAGE (ET CONFUSION)

Grâce à un plan astucieux, le blocage nous permet de minimiser l'impact d'une perturbation sur l'interprétation du système. Le blocage est utilisé lorsqu'on sait qu'il y a des perturbations, mais qu'on ne peut pas les contrôler. La

solution est de confondre délibérément l'effet de la perturbation avec un autre effet du système qui devrait être faible (ou insignifiant).

Prenons l'exemple d'un système à trois facteurs, A, B et C. Dans les plans factoriels, les termes d'interaction d'ordre supérieur ont généralement une incidence très faible sur le résultat, ce qui en fait des coefficients intéressants à confondre avec une perturbation. De fait, il sera impossible de faire la différence entre l'effet d'interaction de ABC et la perturbation. On pourrait aussi dire que le coefficient correspondant est : $\mathbf{b_{ABC}} =$ **= effet d'interaction de ABC + perturbation**.

Ce concept peut être combiné avec des essais expérimentaux pour utiliser un processus appelé blocage. Normalement, avec trois facteurs, il y aurait 2^3 expériences. Mais avec le blocage, on divise les essais en deux, de sorte que la moitié des essais s'effectue à ABC+, et l'autre moitié, à ABC-.

Exemple 9.2.6.1 Blocage (et confusion)

Prenons l'exemple d'une expérience de marketing pour une application mobile, dont le résultat mesuré est le nombre d'achats dans l'application 60 jours après le marketing. Les trois facteurs sont la promotion (A), le message envoyé (B) et le prix (C). Cependant, on constate rapidement que certaines personnes de l'étude auront des iPhone, tandis que d'autres auront des téléphones Android. Le type de téléphone des utilisateurs correspond aux critères d'un facteur mesurable, mais non contrôlable. Le tableau 9.2.6.1 montre comment cela se conceptualise.

Tableau 9.2.6.1 : Tableau d'ordre standard pour l'expérience sur l'application mobile

Expérience	A (Promotion)	B (Message)	C (Prix)	AB	AC	BC
1	–	–	–	+	+	+
2	+	–	–	–	–	+
3	–	+	–	–	+	–
4	+	+	–	+	–	–
5	–	–	+	+	–	–
6	+	–	+	–	+	–
7	–	+	+	–	–	+
8	+	+	+	+	+	+

Il y a inévitablement une certaine confusion, car l'effet du terme d'interaction ABC et du type de téléphone ne peut être séparé. Cependant, ce compromis est bénéfique, car les effets principaux et les interactions à deux facteurs peuvent être interprétés sans biais en supposant que la perturbation a été maintenue constante.

9.2.3 Plans d'expériences : plans fractionnaires

9.2.7 PLANS FRACTIONNAIRES

Avec 2^k essais, il est évident qu'en augmentant le nombre de facteurs (k), la quantité de ressources nécessaires augmentera rapidement. Il est donc nécessaire de discuter des méthodes permettant de réduire la quantité de travail à effectuer et d'économiser des ressources. Cela s'applique surtout aux scénarios de sélection ou d'évaluation d'un nouveau système, comme une exploration à l'échelle du laboratoire, la fabrication d'un nouveau produit ou même la résolution d'un problème.

Ce concept repose sur l'utilisation de la notion de confusion, introduite précédemment à la section 9.2.6. En confondant les facteurs les uns avec les autres, on peut réduire le nombre d'essais nécessaires, ce qui permet de diviser par deux la quantité de travail requise. Une expérience de 2^k essais peut devenir une expérience de 2^{k-1} essais grâce à ce principe. C'est ce qu'on appelle un plan fractionné. Ce principe fonctionne parce qu'on s'intéresse généralement davantage aux effets principaux et parce que les interactions n'ont qu'une signification pratique limitée (surtout lorsqu'il y a trois facteurs ou plus).

Exemple 9.2.7.1 Plans fractionnaires

Comme le montre le tableau 9.2.7.1, on peut prendre une expérience de 2^3 (huit) essais et la réduire de moitié à quatre essais en confondant un facteur avec l'interaction des deux autres facteurs. On écrit les deux premiers facteurs comme des facteurs normaux, mais le troisième facteur prend la forme du produit des deux premiers facteurs.

Tableau 9.2.7.1 Essais expérimentaux pour le système 2^{3-1} où le facteur C est confondu avec l'interaction de AB

Essai	A	B	C = AB
1	–	–	+
2	+	–	–
3	–	+	–
4	+	+	+

Maintenant, il est important de se demander quelles sont les conséquences de cette approche.

1) Ça réduit le travail de moitié! On ne saurait trop insister sur ce point. On a réduit la quantité de ressources utilisées et amélioré l'efficacité du processus, surtout si on considère que les premières expériences ne permettront pas de trouver les paramètres optimaux et qu'il faudra mener plusieurs expériences pour les déterminer (voir la section 9.3).

2) Il y a maintenant plusieurs facteurs confondus. Chacun des effets principaux ($\mathbf{ABC}$) est désormais confondu avec un terme d'interaction.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

$$y_i = b_0 + b_A x_a + b_B x_B + b_C x_C + b_{AB} x_{AB} + b_{AC} x_{AC} + b_{BC} x_{BC} + b_{ABC} x_{ABC} + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_A \\ b_B \\ b_C \\ b_{AB} \\ b_{AC} \\ b_{BC} \\ b_{ABC} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

Il devrait être évident que ce système est maintenant *sous-déterminé* puisqu'il y a huit inconnues, mais seulement quatre équations (comme il n'y a eu que quatre essais). En outre, la matrice X n'est plus orthogonale. Pour y remédier, il faut faire exactement ce qui a été dit plus haut, à savoir confondre les effets principaux avec les termes d'interaction. Ceci est illustré ci-dessous :

Formula does not parse

Par exemple, on dirait maintenant que l'effet principal de $\mathbf{A}$ est confondu avec l'interaction de $\mathbf{BC}$, puisque le coefficient du modèle est la somme de ces deux effets. On peut donc affirmer que $\mathbf{A}$ est un alias de $\mathbf{BC}$, que $\mathbf{B}$ est un alias de $\mathbf{AC}$, que $\mathbf{C}$ est un alias de $\mathbf{AB}$ et que l'ordonnée à l'origine est un alias de l'interaction à trois facteurs $\mathbf{ABC}$. Ceci peut être exprimé par les équations ci-dessous.

$\mathbf{b_0 + b_{ABC} \rightarrow I + ABC}$
$\mathbf{b_A + b_{BC} \rightarrow A + BC}$
$\mathbf{b_B + b_{AC} \rightarrow B + AC}$
$\mathbf{b_C + b_{AB} \rightarrow C + AB}$

10.8 GÉNÉRATEURS

Pour un système à trois facteurs, il est assez simple de déterminer comment confondre les facteurs. Toutefois, lorsqu'il y a plus de facteurs, c'est beaucoup plus compliqué. Pour simplifier ce processus, on peut utiliser des générateurs.

Pour un système à quatre facteurs (2^4), il y aurait les facteurs $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ et $\mathbf{D}$. Les facteurs $\mathbf{A}$ à $\mathbf{C}$ seraient considérés comme normaux, mais le facteur $\mathbf{D}$ s'écrirait $\mathbf{D = ABC}$. C'est ce qu'on appelle la *relation génératrice*.

Pour travailler avec une *relation génératrice*, il faut connaître certaines règles :

1. L'ordonnée à l'origine $\mathbf{I}$ est une colonne de uns.
2. La multiplication d'un facteur par lui-même donne l'identité (ou l'ordonnée à l'origine) : $\mathbf{AxA = I}$
3. Un facteur multiplié par l'identité (ou l'ordonnée à l'origine [une colonne de uns]) est égal à lui-même : $\mathbf{AxI = A}$
4. Grâce à un peu d'algèbre, on peut également établir la relation de définition $\mathbf{I = ABCD}$. Prenons la relation génératrice et multiplions les deux côtés par $\mathbf{D}$. On obtient alors Formula does not parse. Or, $\mathbf{D \times D = I}$, d'où $\mathbf{I = ABCD}$.

En multipliant un effet principal par la relation de définition, on peut rapidement déterminer le facteur avec lequel il est aliasé (ou confondu). Par exemple, pour la moitié 2^{4-1}, on peut voir que $\mathbf{A}$ est aliasé avec $\mathbf{BCD}$ par l'équation suivante :

Formula does not parse

On sait que pour la moitié 2^{3-1} , la relation génératrice est $\mathbf{I = ABC}$, ce qui indique que $\mathbf{B}$ est aliasé avec $\mathbf{AC}$, puisque $\mathbf{BxI = BxABC = AC}$

9.2.9 RÉSOLUTION

La confusion et l'utilisation de plans fractionnés entraînent un compromis en ce qui concerne la résolution, la mesure dans laquelle un ou des effets principaux estimés sont aliasés avec des interactions estimées à deux niveaux, trois niveaux ou plus. La résolution correspond à l'ordre de la plus petite interaction confondue avec un effet principal, plus un. Le tableau de compromis ci-dessous (figure 9.2.9.1) permet de visualiser ce phénomène.

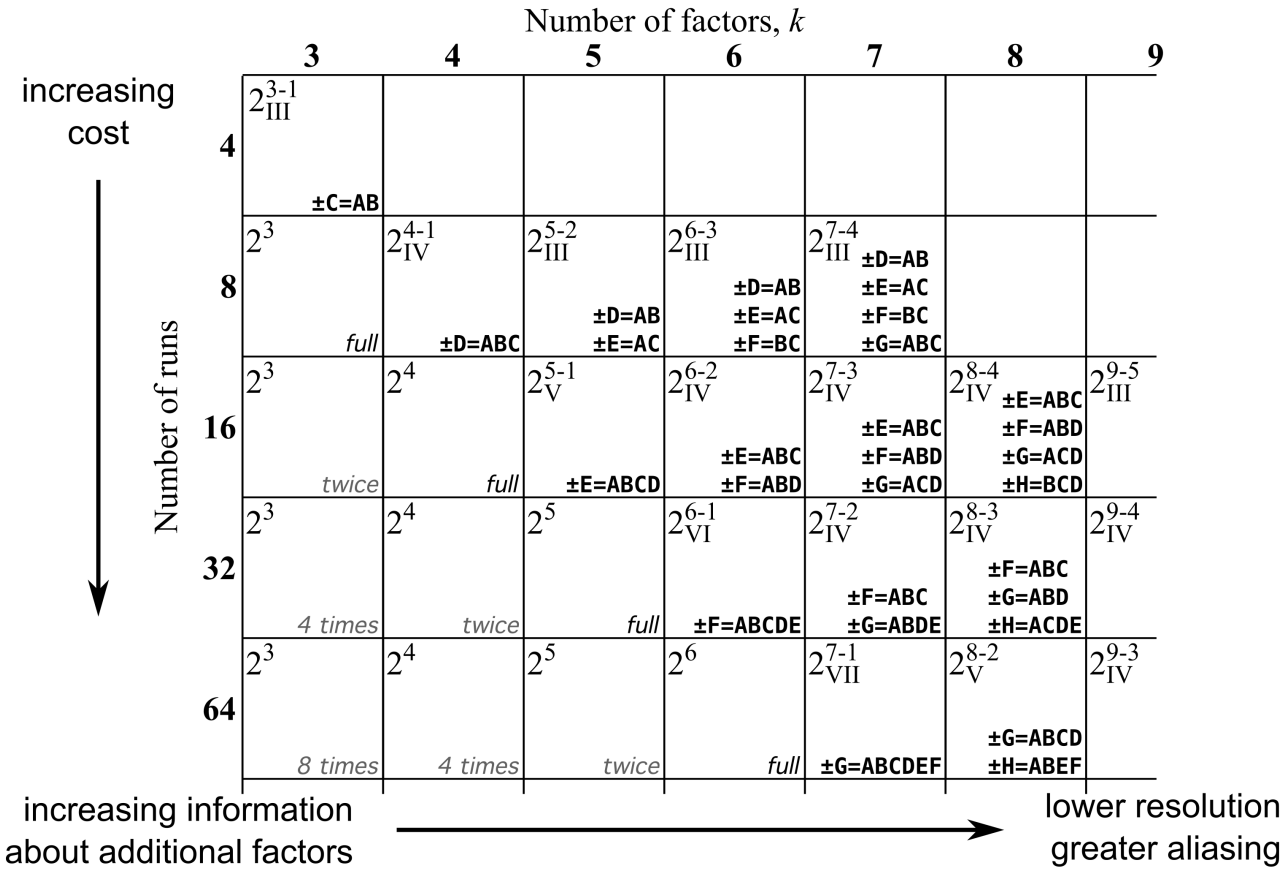


Figure 9.2.9.1 Tableau des compromis pour les plans d'expérience montrant comment la résolution et l'aliassage sont liés.

Prenons l'exemple de : $\mathbf{2^{4-1}_{IV}}$. Ici, les chiffres romains $\mathbf{IV}$ indiquent le niveau de résolution du plan. Ce nombre est équivalent au nombre de facteurs présents dans la relation de définition. Puisque $\mathbf{I = ABCD}$ pour une expérience $\mathbf{2^{4-1}_{IV}}$ on dit qu'il s'agit d'un plan de résolution $\mathbf{IV}$.

De manière générale :

- les plans de résolution III sont bons pour le tri
- les plans de résolution IV sont bons pour la caractérisation
- les plans de résolution V sont bons pour l'optimisation

Il convient de noter qu'aucun de ces plans ne comporte de confusion entre les effets principaux.

Caractéristiques uniques des plans de résolution III, IV et V

Plans de résolution III :

- Effets principaux confondus avec des interactions à deux facteurs

Plans de résolution IV :

- Effets principaux confondus avec des interactions à trois facteurs.
- Interactions à deux facteurs confondues entre elles.

Plans de résolution V :

- Pas d'aliassage entre les effets principaux et les interactions à deux facteurs.
- Interactions à deux facteurs sont aliasées avec des interactions à trois facteurs.

9.3.1 Plan d'expériences : Optimisation et méthodologie des surfaces de réponse

9.3.1. OPTIMISATION

Le but ultime des expériences est d'optimiser un système. Les plans factoriels ou fractionnels conviennent aux essais initiaux lorsqu'on dispose d'un nombre limité d'informations. Après quoi, on peut mener une série d'expériences pour s'assurer de remplacer progressivement les expériences factorielles par des plans qui se rapprochent des conditions optimales. Cette procédure se nomme la méthode des surfaces de réponse (MSR).

MSR pour une seule variable

Examinons d'abord l'effet d'un seul facteur $\mathbf{x_1}$ sur la réponse, $\mathbf{y}$. Cet exemple servira à illustrer le principe général du processus de surface de réponse.

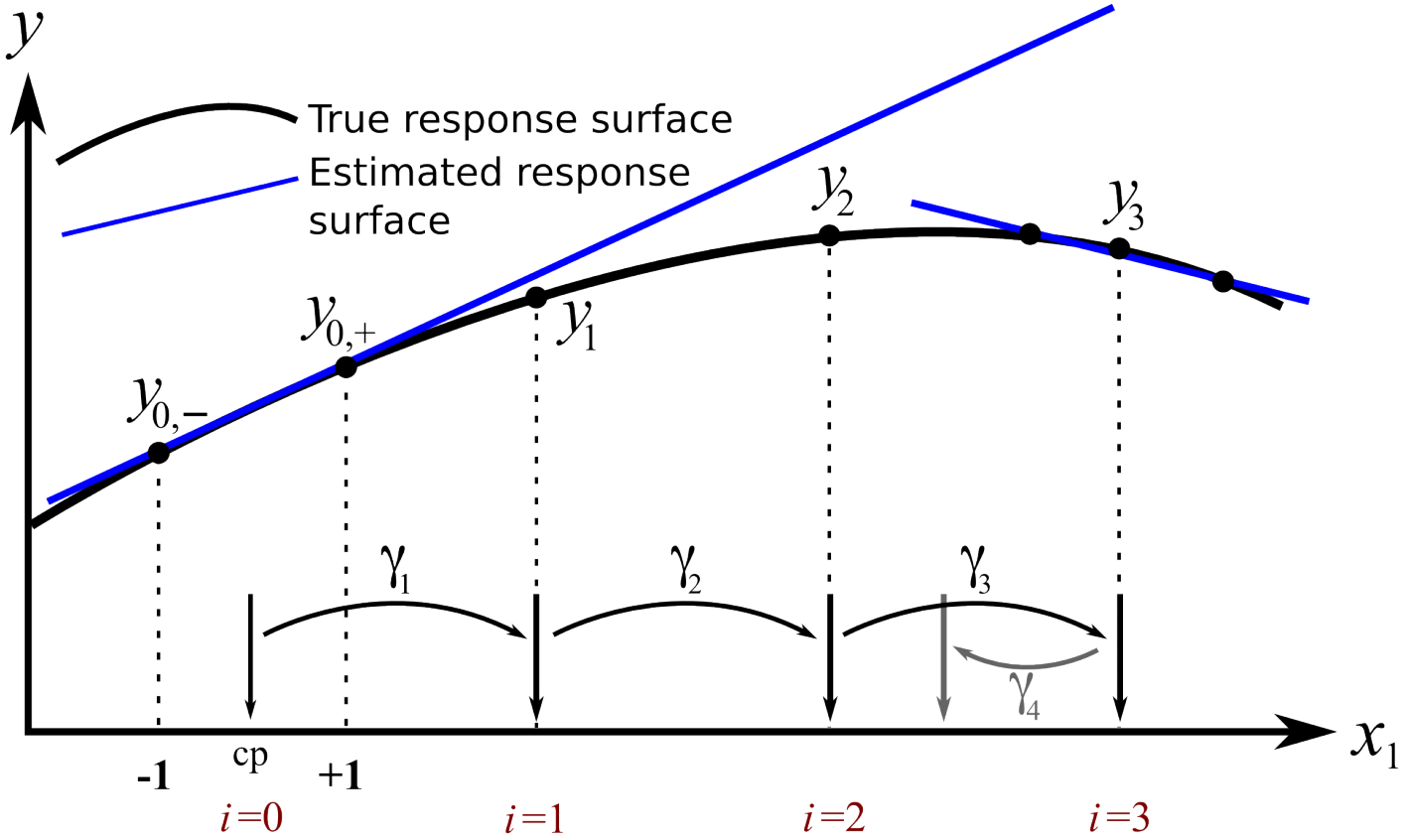


Figure 9.3.1.1 Tracé illustrant la méthode des surfaces de réponse avec un seul facteur.

Le point $\mathbf{i = 0}$ sert de point de référence initial (pc = point de centrage). Puis, on exécute une expérience à deux niveaux, au-dessous (à – 1) et au-dessus (à 1) de ce point de référence, et on obtient des valeurs de réponse correspondantes de $\mathbf{y_{0,-}}$ et $\mathbf{y_{0,+}}$. À partir de là, on peut estimer la droite de régression et la suivre dans la direction croissante $\mathbf{y}$. Notez que l'inclinaison de la pente d'une droite tangente correspond à la trajectoire de

pente maximale. On effectue un pas de γ_1 unités le long de $\mathbf{x_1}$, puis on mesure la réponse, $\mathbf{y_1}$. Puisque la variable de la réponse a augmenté, on poursuit dans cette direction.

On effectue un nouveau pas, cette fois-ci deγ_2 unités dans la direction où $\mathbf{y}$ est croissant. On mesure la réponse, $\mathbf{y_2}$, qui est encore croissante. Ce résultat nous encourage à effectuer un nouveau pas de γ_3 . Les pas γ_i doivent être suffisamment grands pour causer une variation de la réponse lors d'un nombre raisonnable d'expériences, sans toutefois être si grands qu'ils nous feraient passer à côté d'un optimum.

Le nouveau point, $\mathbf{y_3}$, a environ la même valeur que $\mathbf{y_2}$, ce qui indique qu'on a atteint un plateau. À ce stade, on peut se lancer dans une démarche exploratoire et réajuster la tangente (qui est maintenant de pente inverse). On peut aussi utiliser les points de données accumulés pour faire une régression non linéaire. Dans les deux cas, on peut alors estimer un nouveau pas de progression de γ_4 pour se rapprocher de l'optimum.

Cette approche convient lorsque la réponse ne dépend que d'un seul facteur. Cependant, dans la plupart des systèmes, la réponse est influencée par plusieurs facteurs, ce qui nous oblige à adapter cette méthode pour trouver les optimums du système.

9.3.2. OPTIMISATION D'UN SYSTÈME À DEUX VARIABLES

Supposons qu'on cherche à optimiser un bioréacteur dont le rendement est affecté par deux facteurs, la température **T** et la concentration en substrat **S**. Or, le résultat qui nous intéresse dans ce contexte, c'est le profit total, qui prend en compte les coûts énergétiques, les coûts des matières premières et autres facteurs pertinents. La figure 9.3.2.1 illustre en gris clair des courbes (hypothétiques) de profit. En pratique, ces courbes sont souvent inconnues. Le système fonctionne actuellement dans les conditions suivantes (conditions de référence) :

- **T** = 325 K
- **S** = 0,75 g/L
- **Profit** = 407 $ par jour

On crée une analyse factorielle complète à partir de ces conditions en choisissant $\mathbf{\Delta_T = 10K}$, et $\mathbf{\Delta_S = 0{,}5g/L}$, sachant que ces pas sont suffisamment grands pour révéler une différence dans la valeur de la réponse (voir le tableau 9.3.2.1), sans toutefois être si grands qu'ils exploreraient un tout autre régime du bioréacteur.

Tableau 11.2.1 Expérience du bioréacteur – Plan des expériences

Expérience	T (réelle)	S (réelle)	T (codée)	S (codée)	Profit
Conditions de référence	325 K	0,75 g/L	0	0	407
1	320 K	0,50 g/L	–	–	193
2	330 K	0,50 g/L	+	–	310
3	320 K	1,0 g/L	–	+	468
4	330 K	1,0 g/L	+	+	571

Il apparaît clairement qu'on peut maximiser les profits en opérant à des températures plus élevées et en utilisant de plus fortes concentrations de substrat. Néanmoins, la seule manière de mesurer ce taux d'augmentation est d'établir un modèle linéaire du système à partir des données factorielles :
<img src= »https://ecampusontario.pressbooks.pub/app/uploads/sites/4171/2024/03/974e987bb8c34da437e33d1d3d54f1a5.png »
alt= »\begin{split}\hat{y} &= b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S \\ \hat{y} &= 389,8 + 55 x_T + 134 x_S – 3,50 x_T x_S\end{split}

où Formula does not parse x_T = \dfrac{x_{T,\text{réelle}} - \text{centre}_T}{\Delta_T/2}" title="\begin{split}\hat{y} &= b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S \\ \hat{y} &= 389,8 + 55 x_T + 134 x_S - 3,50 x_T x_S\end{split}
où Formula does not parse x_T = \dfrac{x_{T,\text{réelle}} - \text{centre}_T}{\Delta_T/2}" class="latex mathjax"> $= \frac{x_{T,\text{réelle}} - 325}{5}$

et $x_S = \frac{x_{S,\text{réelle}} - 0{,}75}{0{,}25}$.

Le modèle démontre que l'on peut s'attendre à une hausse de profit de 55 $ par jour pour une augmentation d'une unité de T. En unités réelles, il faudrait augmenter la température de $\Delta x_{T,\text{ réelle}} = 1 \times \Delta_T/2 = 5K$ pour atteindre cet objectif. Ce facteur d'échelle provient du codage que nous avons utilisé :
\[\begin{split}x_T &= \displaystyle \frac{x_{T,\text{réelle}} - \text{centre}_T}{\Delta_T / 2} \\
\Delta x_T &= \displaystyle \frac{\Delta x_{T,\text{réelle}}}}{\Delta_T / 2}\end{split}\]
Au même titre, on peut augmenter \(S\) par $\Delta x_{S,\text{réelle}} = 1 \times \Delta_S/2 = 0{,}25 g/L$ pour obtenir une hausse de profit de 134 $ par jour.

Le terme d'interaction est faible, ce qui suggère que la surface de réponse est plutôt linéaire dans cette région. La figure 9.3.2.1 illustre les contours du modèle (lignes droites vertes). Observez que les contours du modèle représentent une bonne approximation des contours réels (en pointillé, gris clair), lesquels restent inconnus en pratique.

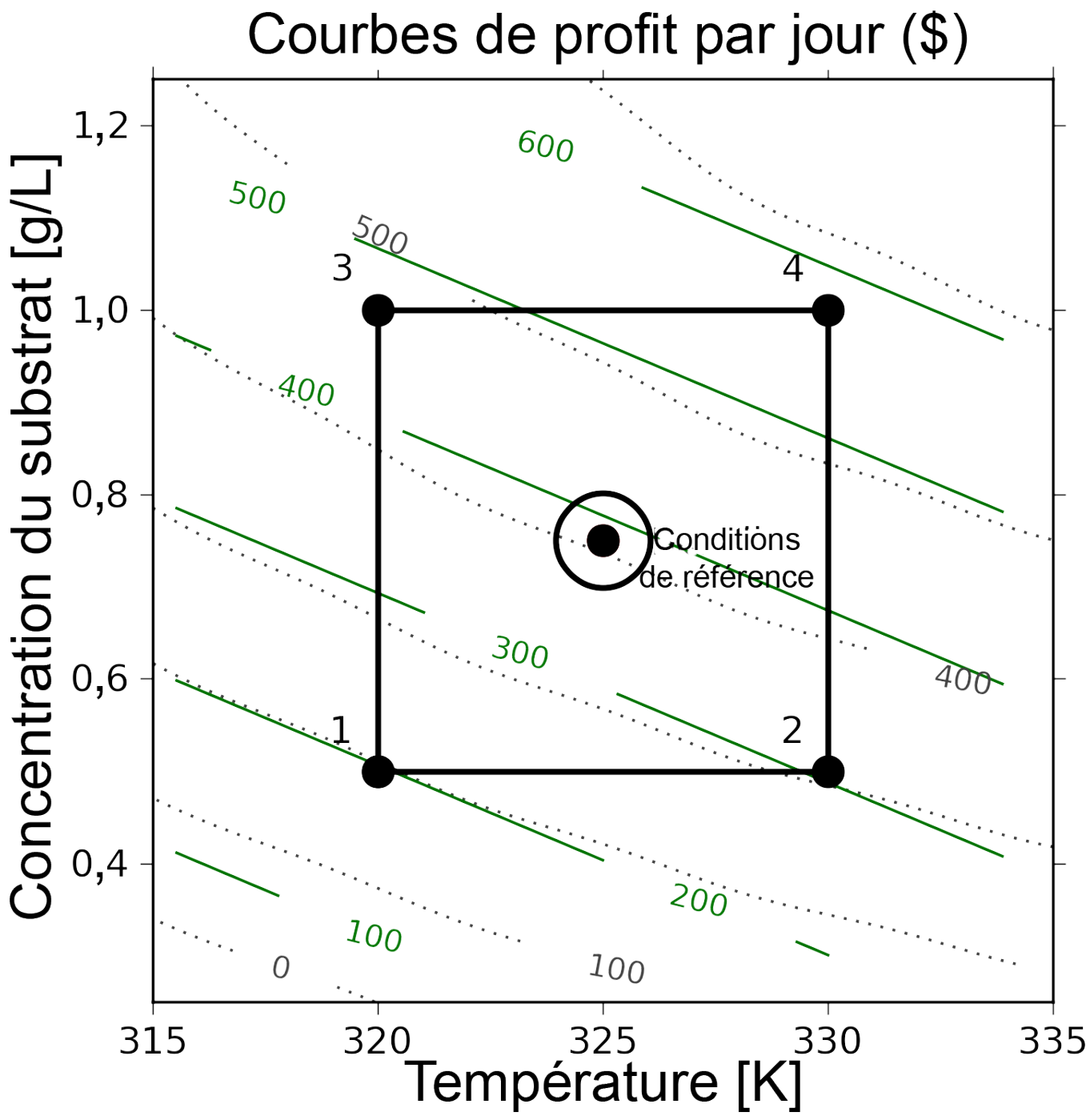


Figure 9.3.2.1. Première expérience factorielle pour l'exemple du bioréacteur.

Pour accroître le profit de manière optimale, on se déplace le long de la surface du modèle estimé, dans la direction de la pente maximale. Pour obtenir cette direction, il suffit de prendre les dérivées partielles de la fonction du modèle en ignorant le terme d'interaction (qui est négligeable).

\[\dfrac{\partial \hat{y}}{\partial x_T} = b_T = 55 \qquad \qquad \dfrac{\partial \hat{y}}{\partial x_S} = b_S = 134\]

Cela signifie que pour chaque déplacement $b_T = 55$ unités codées selon x_T, il faut également faire un déplacement selon x_S de $b_S = 134$ unités codées. Mathématiquement :

\[\frac{\Delta x_S}{\Delta x_T} = \frac{134}{55}\]

Le plus simple, c'est de choisir le pas de l'une des variables, puis d'ajuster l'autre en conséquence.

Ainsi, on choisit d'augmenter de $\Delta x_T = 1$ unité codée, ce qui signifie :

\[\begin{split} \Delta x_T &= 1 \\
\Delta x_{T,\text{réelle}} &= 5 \,\text{K} \\
\Delta x_S &= \frac{b_S}{b_T} \Delta x_T = \frac{134}{55} \Delta x_T \\
\text{mais comme}\qquad\qquad \Delta x_S &= \frac{x_{S,\text{réelle}}}{\Delta_S / 2} \\
\Delta x_{S,\text{réelle}} &= \frac{134}{55} \times 1 \times \Delta_S / 2\,\, \text{en comparant
les deux lignes précédentes} \\
\Delta x_{S,\text{réelle}} &= \frac{134}{55} \times 1 \times 0,5 / 2 = \bf{0,61}\,\,\text{g/L}\\\end{split}\]

Ce qui donne les conditions suivantes pour la cinquième expérience :

- $T_5 = T_{\text{T\_référence}} + \Delta x_{T,\text{réelle}} = 325 + 5 = 330K$
- $S_5 = S_{\text{référence}} + \Delta x_{S,\text{réelle}} = 0,75 + 0,6 = 1,36g/L$

On mène donc l'expérience suivante selon ces conditions, et le profit quotidien vaut y_5 = 669 $ y_5 = 669 $. Il s'agit d'une amélioration substantielle par rapport aux conditions de référence.

On décide d'effectuer un autre déplacement, dans la même direction de pente maximale, c'est-à-dire le long du vecteur qui pointe dans la direction $\frac{134}{55}$. On augmente la température de 5K (mais le pas aurait pu être plus grand ou plus petit), et on obtient les conditions suivantes pour l'expérience 6 :

- $T_6 = T_5 + \Delta x_{T,\text{réelle}} = 330 + 5 = 335K$
- $S_6 = S_5 + \Delta x_{S,\text{réelle}} = 1,36 + 0,61 = 1,97g/L$

Cette fois, le profit est de $y_6 = \$688$. La croissance se poursuit, mais dans une proportion moindre. Elle commence peut-être à se stabiliser. On décide toutefois d'encore monter la température de 5 K et d'augmenter la concentration de substrat en conséquence. On obtient ainsi les conditions suivantes pour l'expérience 7 :

- $T_7 = T_6 + \Delta x_{T,\text{réelle}} = 335 + 5 = 340K$
- $S_7 = S_6 + \Delta x_{S,\text{réelle}} = 1,97 + 0,61 = 2,58g/L$

Cette fois, le profit est de $y_7 = \$463$. Nous sommes allés trop loin, puisque les profits ont chuté. Il faut donc revenir au *meilleur point précédent*, parce que la surface a manifestement changé, et réajuster le modèle avec une nouvelle analyse factorielle dans ce voisinage :

Tableau 9.3.2.2 Exécutions séquentielles de l'expérience du bioréacteur

Expérience	T (réelle)	S (réelle)	T (codée)	S (codée)	Profit
6	335 K	1,97 g/L	0	0	688 $
8	331 K	1,77 g/L	-	-	694 $
9	339 K	1,77 g/L	+	-	725 $
10	331 K	2,17 g/L	-	+	620 $
11	339 K	2,17 g/L	+	+	642 $

Pour se déplacer plus lentement le long de la surface, on opte pour de plus petites étendues dans la factorielle : $\text{range}_T = 8 = (339 - 331)K$ et $\text{étendue\_s}_S = 0,4 = (2,17 - 1,77)g/L$.

<img src="https://ecampusontario.pressbooks.pub/app/uploads/sites/4023/2024/02/RSM-base-case-

combined.png" alt="**Figure 11.2.2.** Illustration de la méthodologie des surfaces de réponse dans l'exemple du bioréacteur."> À partir des données factorielles d'origine, on détermine la trajectoire de pente maximale. Une fois l'optimum atteint, on effectue une nouvelle analyse factorielle pour déterminer la prochaine trajectoire de pente maximale. <"width="3300" height="2400"> **Figure 9.3.2.2.** Illustration de la méthodologie des surfaces de réponse dans l'exemple du bioréacteur. À partir des données factorielles d'origine, on détermine la trajectoire de pente maximale. Une fois l'optimum atteint, on effectue une nouvelle analyse factorielle pour déterminer la prochaine trajectoire de pente maximale.

Un modèle des moindres carrés basé sur les quatre points factoriels (les expériences 8, 9, 10 et 11, exécutées dans un ordre aléatoire), suggère que la tendance la plus favorable serait d'augmenter la température tout en réduisant la concentration de substrat.

\[\begin{split}\hat{y} &= b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S \\
\hat{y} &= 673,8 + 13,25 x_T - 39,25 x_S - 2,25 x_T x_S\end{split}\]

Comme auparavant, on avance dans la direction la pente maximale en faisant un pas de b_T unités le long de la directionx_T et de b_S unités le long de la direction x_S. On choisi à nouveau $\Delta x_T = 1$ unité. (Rappelons qu'on pourrait opter pour un pas plus petit ou plus grand, si nécessaire.) Par conséquent :

\[\begin{split}\frac{\Delta x_S}{\Delta x_T} &= \frac{-39}{13} \\
\Delta x_S &= \frac{-39}{13} \times 1\\
\Delta x_{S, \text{réelle}} &= \frac{-39}{13} \times 1 \times 0,4 /2 = -0,6\,\text{g/L}\\
\Delta x_{T, \text{réelle}} &= 4\,\text{K}\\\end{split}\]

On obtient ainsi les conditions suivantes pour l'expérience 12 :

- $T_{12} = T_6 + \Delta x_{T,\text{réelle}} = 335 + 4 = 339K$
- $S_{12} = S_6 + \Delta x_{S,\text{réelle}} = 1,97 - 0,6 = 1,37g/L$

On détermine que le profit s'élève alors à 716 $. Or, l'analyse factorielle précédente présentait une valeur de profit de 725 $ à l'un des coins. Il se pourrait qu'il y ait du bruit dans le systèmes. En effet, la différence entre 716 $ et 725 $ ne représente pas un montant très élevé; en revanche, on observe une différence de profit relativement importante entre les autres points de l'analyse factorielle.

Quelques considérations à prendre en compte lorsqu'on s'approche d'un optimum :

- La variable de réponse atteindra un plateau (rappelez-vous qu'à un optimum, la première dérivée première vaut zéro).
- Si la variable de réponse reste à peu près constante pendant deux sauts consécutifs, vous avez peut-être dépassé l'optimum.
- La variable de réponse peut diminuer, parfois très rapidement, si vous dépassez l'optimum.
- On peut aussi déduire que la surface est courbe si les termes d'interaction sont du même ordre de grandeur (ou plus grands) que les termes d'effets principaux.

En d'autres termes, les optimums présentent une certaine courbure. Ainsi, un modèle qui ne comporte que des termes linéaires ne pourra pas vous indiquer la direction de pente maximale le long de la *surface de réponse réelle*. Il faut ajouter des termes qui tiennent compte de cette courbure.

9.3.3. VÉRIFICATION DE LA COURBURE

Lorsque le point de centrage mesuré diffère sensiblement du point de centrage prédit par le modèle linéaire, c'est que la surface de réponse est courbe, ce qu'il faut représenter par l'ajout de termes polynomiaux.

Le point de centrage de l'analyse factorielle peut être prédit à partir de $(x_T, x_S) = (0, 0)$ – c'est simplement le terme d'intersection. Dans la dernière analyse factorielle, le point de centrage prédit correspondait à $\hat{y}_{\text{cp}} = \$670$. Or, le point de centrage réel de l'expérience 6 affichait un profit de 688 $. Cette différence de 18 $ s'avère substantielle, surtout si on la compare aux coefficients des effets principaux.

9.3.4. PLANS COMPOSITES CENTRÉS

L'analyse détaillée des plans composites centrés ne sera pas abordée dans le présent manuel. Toutefois, cette partie montre quelques exemples à deux et trois variables, à partir d'une factorielle orthogonale existante à laquelle on ajoute des points axiaux. Ces points pourront être aisément ajoutés ultérieurement pour tenir compte de la non-linéarité.

Les points axiaux sont placés à $4^{0.25} = 1.4$ unité codée du centre pour un système à deux facteurs, et à $8^{0,25} = 1,7$ unité codée pour un système à trois facteurs.

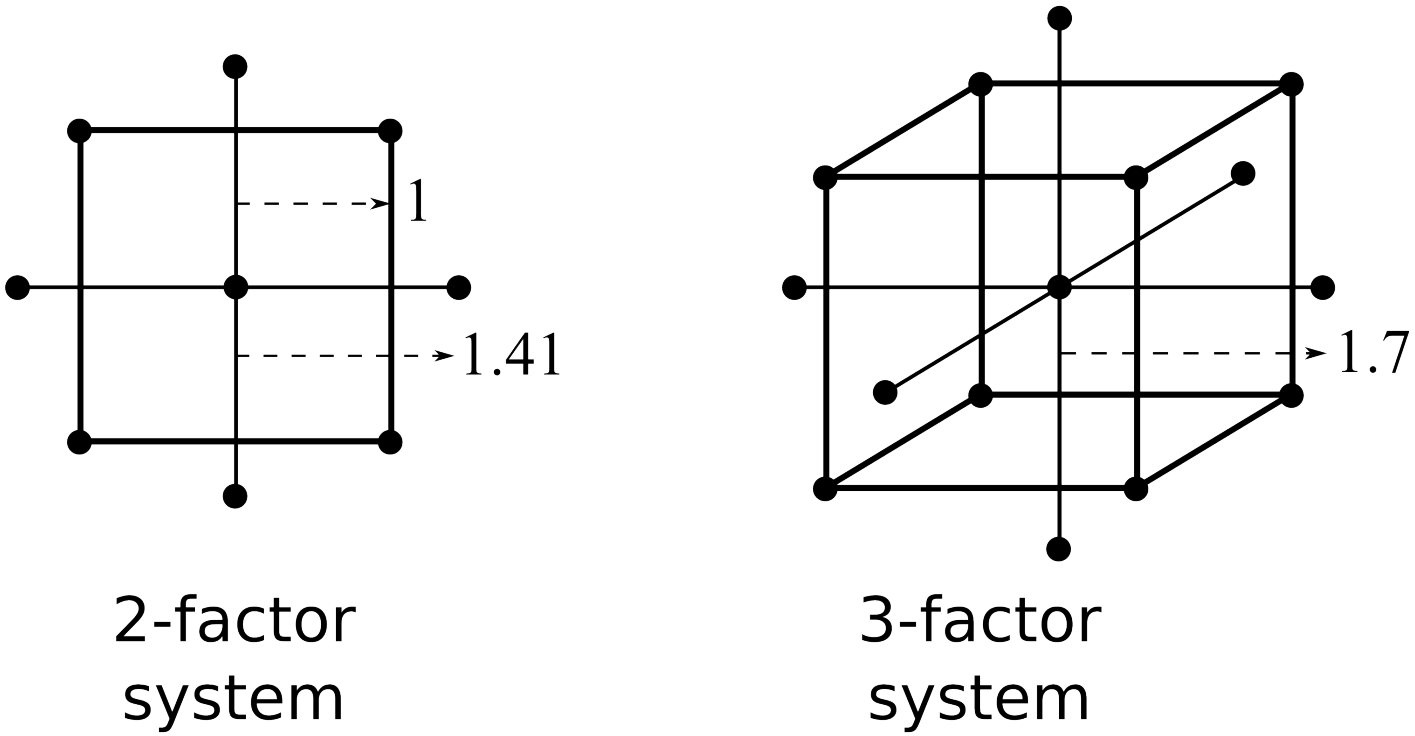


Figure 9.3.4.1. Illustration du plan composite centré pour les systèmes à deux et à trois facteurs. Les points axiaux sont respectivement placés à 1,4 et 1,7 unité du centre des systèmes à deux et à trois facteurs.

Un plan composite central a été ajouté à l'analyse factorielle dans l'exemple ci-dessus, puis les expériences ont été exécutées, de manière aléatoire, aux quatre points axiaux.

Les quatre valeurs de réponse étaient $y_{13} = 720$, $y_{14} = 699$, $y_{15} = 610$ et $y_{16} = 663$. Cela nous permet d'estimer un modèle contenant des termes quadratiques : $y = b_0 + b_T x_T + b_S x_S + bTS x_T x_S + b_{TT} x_T^2 + b_{SS} x_S^2$. Les paramètres de ce modèle s'obtiennent selon la procédure habituelle, en utilisant un modèle des moindres carrés :

$$y = 688 + 13x_T - 39x_S - 2{,}4x_Tx_S - 4{,}2x_T^2 - 12{,}2x_S^2$$

Remarquez que les termes linéaires sont identiques à ce qu'on avait auparavant! Les effets quadratiques se révèlent assez significatifs par rapport aux autres effets; en effet, c'est ce qui a empêché le modèle linéaire de prédire le résultat de l'expérience 12.

La dernière étape de la méthodologie des surfaces de réponse consiste à tracer le contour de ce modèle et à prédire où exécuter les prochaines expériences. Comme le démontrent les lignes continues de contour dans l'illustration, il faudrait exécuter les prochaines expériences approximativement à T = 343K et à S = 1,60 g/L, où le profit escompté avoisine 736 $. (Ces valeurs peuvent être obtenues de manière approximative par un examen visuel, ou de manière analytique.) Le véritable optimum du processus ne se situe pas exactement là, mais il s'en rapproche beaucoup.

Cet exemple a démontré toute la puissance de la méthode des surfaces de réponse. Un nombre minimal d'expériences a rapidement convergé vers le véritable optimum (qui était inconnu) du processus. Pour y parvenir, nous avons établi une succession de modèles des moindres carrés permettant d'obtenir une approximation de la surface sous-jacente. Ces modèles des moindres carrés se construisent avec les outils des analyse factorielles fractionnaires et complètes, ainsi qu'avec les fondements de l'optimisation. Ainsi, sommes en mesure de gravir la pente maximale.

9.3.2 Plan d'expériences : L'approche générale

9.3.4. L'APPROCHE GÉNÉRALE DE LA MODÉLISATION DES SURFACES DE RÉPONSE

1. À partir des conditions de référence, on identifie les facteurs principaux en se référant au processus, à des opinions d'experts ou à son intuition. On réalise des expériences factorielles (complètes ou fractionnaires) complètement randomisées. On utilise les résultats de ces expériences pour estimer un modèle linéaire du système :
   \[\hat{y} = b_0 + b_Ax_A + b_B x_B + b_C x_C \ldots + b_{AB}x_Ax_B + b_{AC} x_A x_C + \ldots\]
2. Les effets principaux sont généralement beaucoup plus importants que les interactions à deux facteurs, de sorte qu'on peut négliger ces termes d'ordre supérieur. Tout effet principal non significatif peut se voir écarté lors d'itérations ultérieures. (Pensez à ce qui a été abordé à la section précédente.)
3. On utilise le modèle pour estimer la trajectoire de pente maximale :
   \[\dfrac{\partial \hat{y}}{\partial x_1} = b_1 \qquad\
   qquad \dfrac{\partial \hat{y}}{\partial x_2} = b_2 \qquad \ldots\]
   Pour gravir (ou descendre, dans un problème de minimisation) la trajectoire de pente maximale, on déplace n'importe lequel des effets principaux, p. ex. \(b_A\) d'une certaine quantité Δx_A, puis on déplace les autres effets d'un pas $\Delta x_i = \frac{b_i}{b_A}\Delta x_A$. Par exemple, le pas Δx_C vaut $\frac{b_C}{b_A}\Delta x_A$.

   Si l'une des valeurs Δx_i se révèle trop élevée pour être mise en application sans risque, on rapetisse le pas et on applique le même facteur de proportionnalité à tous les autres pas. Rappelez-vous qu'il s'agit d'unités codées; il faut donc appliquer le facteur d'échelle pour obtenir la valeur du déplacement en unités réelles.
4. On répète ce processus jusqu'à ce que la réponse commence à se stabiliser ou que l'on ait la certitude d'être passé à un autre régime du processus.
5. À ce stade, on répète l'expérience factorielle à partir de l'étape 1, en utilisant la dernière meilleure valeur de réponse comme nouveau point de référence. C'est aussi le bon moment de réintroduire les termes auparavant négligés. Aussi, dans le cas d'un facteur binaire, on examine l'effet de l'alternance de son signe à cette même étape. Ces expériences factorielles supplémentaires devraient également inclure des points de centrage.
6. On répète les étapes 1 à 5 jusqu'à ce que l'estimation du modèle linéaire affiche des signes de courbure ou que les termes d'interaction commencent à dominer les effets principaux. Ceci indique qu'on s'approche d'un optimum.

   - On peut évaluer la courbure en comparant le point de centrage prédit, c'est-à-dire le point d'intersection du modèle = b_0, avec la ou les réponses réelles du point de centrage. Une grande différence entre la prédiction et les effets du modèle indique que la surface de réponse est courbe.

7. En cas de courbure, on ajoute des points axiaux pour transformer le plan factoriel en plan composite centré. On estime ensuite un modèle quadratique de la forme :
   \[y = b_0 + b_1x_1 + b_2 x_2 + \ldots + b_{12}x_1x_2 + \ldots + b_{11}x_1^2 + b_{22}x_2^2 + \ldots\]

8. On dessine les tracés de contour de cette surface de réponse estimée et on détermine où positionner les expériences séquentielles. On peut également obtenir l'optimum du modèle de manière analytique en calculant les dérivées de la fonction du modèle.

RÉSUMÉ

1. Dans les sections précédentes, on a utilisé des analyses factorielles complètes ou fractionnaires pour sélectionner les facteurs importants. Au moment de procéder à l'optimisation des processus, on présume que les variables importantes ont déjà été identifiées. En réalité, il arrive que des variables auparavant considérées comme importantes deviennent secondaires à mesure que l'on s'approche de l'optimum. Inversement, des variables qui auraient pu être écartées au départ deviennent importantes à l'optimum.
2. En règle générale, la méthode des surfaces de réponse fonctionne mieux lorsque les variables prises en compte sont continues. Les variables de catégorie (oui ou non, catalyseur A ou B) se traitent en les associant à une valeur ou à l'autre, puis en effectuant l'optimisation en fonction de ces valeurs sélectionnées. Il s'avère toujours utile d'examiner les autres valeurs une fois l'optimum atteint.

9.4.1 Projet de plan d'expériences

À ce stade, vous devriez vous sentir suffisamment en confiance pour entreprendre votre propre projet de plan d'expériences! Cet exercice s'appuiera sur tout ce qui a été exposé dans ce guide, des tests d'hypothèses jusqu'à la régression, en passant par le plan d'expériences.

Il est fortement recommandé de consulter les fichiers du Jupyter Notebook sur le plan d'expériences. On les retrouve dans le chapitre « How do I do X in Python? ». Les fichiers « Full Factorial Example » et « Standard Error & Replicates » s'avéreront particulièrement utiles.

Projet de plan d'expériences

Ce mini projet de plan d'expériences permet de se familiariser avec la planification d'expériences de manière concrète.

Ce projet est *relativement court* et ne nécessite *pas une élaboration poussée*. Vous ne disposez que de quelques semaines pour planifier vos expériences, les réaliser et analyser les données. Nous proposons quelques suggestions ci-dessous, mais vous pouvez choisir ce que vous voulez : l'optimisation d'une recette ou d'un dessert favori, un hobby, un sport, etc.

L'objectif visé consiste à découvrir l'importance des thèmes suivants dans le projet de plan d'expérience. Une fois le système à l'étude choisi, certaines questions se poseront :

- Quelles variables devront être utilisées?
- Quelle étendue ces variables doivent-elles couvrir?
- Comment mesure-t-on ces variables (en particulier la variable de réponse y)?
- Quelle autre variabilité trouve-t-on dans le système? Peut-on la mesurer et la contrôler?
- Quels types de plan expérimental (factoriel complet, factoriel fractionné) faut-il choisir? Quels sont les facteurs de confusion et les contraintes de manipulation?
- Combien d'expériences faut-il exécuter? Peut-on les répéter ou trouver des points de centrage? Comment randomiser les exécutions?

Voilà des questions difficiles à aborder dans un devoir ou un examen!

Thème du projet

Vous vous passionnez sans doute pour un passe-temps, la cuisine, le sport, un quelconque domaine de recherche, etc. Par conséquent trouver un système à étudier ne devrait pas poser de problème. Toutefois, certains systèmes sont trop complexes pour le temps alloué, et il vous faudra choisir un système plus simple. Vous trouverez ci-dessous quelques pistes de réflexion, mais n'hésitez pas à vous inspirer de tout ce qui vous intéresse ou de tout autre sujet qui suscite votre curiosité. Ne sélectionnez pas un projet uniquement pour sa simplicité; optez plutôt pour celui qui présente de solides perspectives en matière d'expérimentation.

Exemples de thèmes :

- rendement du maïs soufflé cuisiné sur la cuisinière ou au four micro-ondes
- hauteur de levée du pain
- efficacité énergétique et consommation de carburant d'une voiture
- durée de vol d'un avion en papier
- croissance des plantes

- hauteur du rebond d'une balle
- distance à laquelle on peut botter un ballon
- capacité d'absorption d'une serviette
- délai d'éclatement des bulles de savon

Quel que soit le thème choisi, il existe des lignes directrices générales à respecter :

- L'expérience à mener doit être reproductible et répétable.
- Évitez les effets liés au temps, comme l'apprentissage d'une langue à l'aide de différentes méthodes – ce qui a été appris ne peut pas être « désappris ».
- Les objectifs doivent être quantifiables – il faut éviter les résultats subjectifs comme le « goût ».
- Il est recommandé d'inclure à la fois des facteurs quantitatifs et qualitatifs, en fonction de l'expérience menée.

Tableau A1.1 Table de probabilités de la loi normale centrée réduite

PROBABILITÉS CUMULATIVES DE LA LOI NORMALE CENTRÉE RÉDUITE

Probabilités cumulatives de la loi normale centrée réduite

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
—3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
—3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
—3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
—3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
—3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
—2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
—2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
—2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
—2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
—2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
—2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
—2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
—2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
—2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
—2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
—1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
—1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
—1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
—1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
—1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
—1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
—1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
—1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
—1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
—1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
—0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
—0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
—0,7	0,2420	0,2389	0,2358	0,2327	0,2297	0,2266	0,2236	0,2206	0,2177	0,2148
—0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
—0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
—0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
—0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
—0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
—0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
—0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Probabilités cumulatives de la loi normale centrée réduite (*suite*)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9773	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Tableau A1.2. Table de probabilités de la loi normale centrée réduite – Moitié supérieure

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002

Tableau A1.3. Table de distribution des quantiles t

Quantiles de distribution t

ν	$Q(0,9)$	$Q(0,95)$	$Q(0,975)$	$Q(0,99)$	$Q(0,995)$	$Q(0,999)$	$Q(0,9995)$
1	3,078	6,314	12,706	31,821	63,657	318,317	636,607
2	1,886	2,920	4,303	6,965	9,925	22,327	31,598
3	1,638	2,353	3,182	4,541	5,841	10,215	12,924
4	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	1,476	2,015	2,571	3,365	4,032	5,893	6,869
6	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	1,415	1,895	2,365	2,998	3,499	4,785	5,408
8	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	1,330	1,734	2,101	2,552	2,878	3,610	3,922
19	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	1,325	1,725	2,086	2,528	2,845	3,552	3,849
21	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	1,319	1,714	2,069	2,500	2,807	3,485	3,768
24	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	1,310	1,697	2,042	2,457	2,750	3,385	3,646

Tableau A1.4 Table de distribution des quantiles chi2

Distribution des quantiles du chi²

v	$Q(0{,}005)$	$Q(0{,}01)$	$Q(0{,}025)$	$Q(0{,}05)$	$Q(0{,}1)$	$Q(0{,}9)$	$Q(0{,}95)$	$Q(0{,}975)$	$Q(0{,}99)$	$Q(0{,}995)$
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,143	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,290	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,653	40,647	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,994
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
31	14,458	15,655	17,539	19,281	21,434	41,422	44,985	48,232	52,192	55,003
32	15,134	16,362	18,291	20,072	22,271	42,585	46,194	49,480	53,486	56,328
33	15,815	17,074	19,047	20,867	23,110	43,745	47,400	50,725	54,775	57,648
34	16,501	17,789	19,806	21,664	23,952	44,903	48,602	51,966	56,061	58,964
35	17,192	18,509	20,569	22,465	24,797	46,059	49,802	53,204	57,342	60,275
36	17,887	19,233	21,336	23,269	25,643	47,212	50,998	54,437	58,619	61,581
37	18,586	19,960	22,106	24,075	26,492	48,364	52,192	55,668	59,893	62,885
38	19,289	20,691	22,878	24,884	27,343	49,513	53,384	56,896	61,163	64,183
39	19,996	21,426	23,654	25,695	28,196	50,660	54,572	58,120	62,429	65,477
40	20,707	22,164	24,433	26,509	29,051	51,805	55,759	59,342	63,691	66,767

Tableau A1.5. Tables de distribution F

Distribution F – Quantiles 0,75

ν_2 (Degrés de liberté du dénominateur)	ν_1 (degrés de liberté du numérateur)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	5,83	7,50	8,20	8,58	8,82	8,98	9,10	9,19	9,26	9,32	9,41	9,49	9,58	9,63	9,67	9,71	9,76	9,80	9,85
2	2,57	3,00	3,15	3,23	3,28	3,31	3,34	3,35	3,37	3,38	3,39	3,41	3,43	3,43	3,44	3,45	3,46	3,47	3,48
3	2,02	2,28	2,36	2,39	2,41	2,42	2,43	2,44	2,44	2,44	2,45	2,46	2,46	2,46	2,47	2,47	2,47	2,47	2,47
4	1,81	2,00	2,05	2,06	2,07	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08	2,08
5	1,69	1,85	1,88	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,89	1,88	1,88	1,88	1,88	1,87	1,87	1,87
6	1,62	1,76	1,78	1,79	1,79	1,78	1,78	1,78	1,77	1,77	1,77	1,76	1,76	1,75	1,75	1,75	1,74	1,74	1,74
7	1,57	1,70	1,72	1,72	1,71	1,71	1,70	1,70	1,69	1,69	1,68	1,68	1,67	1,67	1,66	1,66	1,65	1,65	1,65
8	1,54	1,66	1,67	1,66	1,66	1,65	1,64	1,64	1,64	1,63	1,62	1,62	1,61	1,60	1,60	1,59	1,59	1,58	1,58
9	1,51	1,62	1,63	1,63	1,62	1,61	1,60	1,60	1,59	1,59	1,58	1,57	1,56	1,56	1,55	1,54	1,54	1,53	1,53
10	1,49	1,60	1,60	1,59	1,59	1,58	1,57	1,56	1,56	1,55	1,54	1,53	1,52	1,52	1,51	1,51	1,50	1,49	1,48
11	1,47	1,58	1,58	1,57	1,56	1,55	1,54	1,53	1,53	1,52	1,51	1,50	1,49	1,49	1,48	1,47	1,47	1,46	1,45
12	1,46	1,56	1,56	1,55	1,54	1,53	1,52	1,51	1,51	1,50	1,49	1,48	1,47	1,46	1,45	1,45	1,44	1,43	1,42
13	1,45	1,55	1,55	1,53	1,52	1,51	1,50	1,49	1,49	1,48	1,47	1,46	1,45	1,44	1,43	1,42	1,42	1,41	1,40
14	1,44	1,53	1,53	1,52	1,51	1,50	1,49	1,48	1,47	1,46	1,45	1,44	1,43	1,42	1,41	1,41	1,40	1,39	1,38
15	1,43	1,52	1,52	1,51	1,49	1,48	1,47	1,46	1,46	1,45	1,44	1,43	1,41	1,41	1,40	1,39	1,38	1,37	1,36
16	1,42	1,51	1,51	1,50	1,48	1,47	1,46	1,45	1,44	1,44	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,35	1,34
17	1,42	1,51	1,50	1,49	1,47	1,46	1,45	1,44	1,43	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,35	1,34	1,33
18	1,41	1,50	1,49	1,48	1,46	1,45	1,44	1,43	1,42	1,42	1,40	1,39	1,38	1,37	1,36	1,35	1,34	1,33	1,32
19	1,41	1,49	1,49	1,47	1,46	1,44	1,43	1,42	1,41	1,41	1,40	1,38	1,37	1,36	1,35	1,34	1,33	1,32	1,30
20	1,40	1,49	1,48	1,47	1,45	1,44	1,43	1,42	1,41	1,40	1,39	1,37	1,36	1,35	1,34	1,33	1,32	1,31	1,29
21	1,40	1,48	1,48	1,46	1,44	1,43	1,42	1,41	1,40	1,39	1,38	1,37	1,35	1,34	1,33	1,32	1,31	1,30	1,28
22	1,40	1,48	1,47	1,45	1,44	1,42	1,41	1,40	1,39	1,39	1,37	1,36	1,34	1,33	1,32	1,31	1,30	1,29	1,28
23	1,39	1,47	1,47	1,45	1,43	1,42	1,41	1,40	1,39	1,38	1,37	1,35	1,34	1,33	1,32	1,31	1,30	1,28	1,27
24	1,39	1,47	1,46	1,44	1,43	1,41	1,40	1,39	1,38	1,38	1,36	1,35	1,33	1,32	1,31	1,30	1,29	1,28	1,26
25	1,39	1,47	1,46	1,44	1,42	1,41	1,40	1,39	1,38	1,37	1,36	1,34	1,33	1,32	1,31	1,29	1,28	1,27	1,25
26	1,38	1,46	1,45	1,44	1,42	1,41	1,39	1,38	1,37	1,37	1,35	1,34	1,32	1,31	1,30	1,29	1,28	1,26	1,25
27	1,38	1,46	1,45	1,43	1,42	1,40	1,39	1,38	1,37	1,36	1,35	1,33	1,32	1,31	1,30	1,28	1,27	1,26	1,24
28	1,38	1,46	1,45	1,43	1,41	1,40	1,39	1,38	1,37	1,36	1,34	1,33	1,31	1,30	1,29	1,28	1,27	1,25	1,24
29	1,38	1,45	1,45	1,43	1,41	1,40	1,38	1,37	1,36	1,35	1,34	1,32	1,31	1,30	1,29	1,27	1,26	1,25	1,23
30	1,38	1,45	1,44	1,42	1,41	1,39	1,38	1,37	1,36	1,35	1,34	1,32	1,30	1,29	1,28	1,27	1,26	1,24	1,23
40	1,36	1,44	1,42	1,40	1,39	1,37	1,36	1,35	1,34	1,33	1,31	1,30	1,28	1,26	1,25	1,24	1,22	1,21	1,19
60	1,35	1,42	1,41	1,38	1,37	1,35	1,33	1,32	1,31	1,30	1,29	1,27	1,25	1,24	1,22	1,21	1,19	1,17	1,15
120	1,34	1,40	1,39	1,37	1,35	1,33	1,31	1,30	1,29	1,28	1,26	1,24	1,22	1,21	1,19	1,18	1,16	1,13	1,10
∞	1,32	1,39	1,37	1,35	1,33	1,31	1,29	1,28	1,27	1,25	1,24	1,22	1,19	1,18	1,16	1,14	1,12	1,08	1,00

Distribution F – Quantiles 0,90

ν_2 (Degrés de liberté du dénominateur)	ν_1 (degrés de liberté du numérateur)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39,86	49,50	53,59	55,84	57,24	58,20	58,90	59,44	59,85	60,20	60,70	61,22	61,74	62,00	62,27	62,53	62,79	63,05	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,28	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

F Distribution .95 Quantiles

ν_2 (Denominator Degrees of Freedom)	ν_1 (Numerator Degrees of Freedom)									
	1	2	3	4	5	6	7	8	9	10
1	161.44	199.50	215.69	224.57	230.16	233.98	236.78	238.89	240.55	241.89
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Distribution F – Quantiles 0,99 (*suite*)

ν_2 (Degrés de liberté du dénominateur)	ν_1 (degrés de liberté du numérateur)								
	12	15	20	24	30	40	60	120	∞
1	243,91	245,97	248,02	249,04	250,07	251,13	252,18	253,27	254,31
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Distribution F – Quantiles 0,99

ν_2 (Degrés de liberté du dénominateur)	ν_1 (degrés de liberté du numérateur)									
	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5403	5625	5764	5859	5929	5981	6023	6055
2	98,51	99,00	99,17	99,25	99,30	99,33	99,35	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,19	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Distribution F – Quantiles 0,99

ν_2 (Degrés de liberté du dénominateur)	ν_1 (degrés de liberté du numérateur)								
	12	15	20	24	30	40	60	120	∞
1	6107	6157	6209	6235	6260	6287	6312	6339	6366
2	99,41	99,43	99,44	99,45	99,47	99,47	99,48	99,49	99,50
3	27,05	26,87	26,69	26,60	26,51	26,41	26,32	26,22	26,13
4	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Tableau A1.6 Tableau des valeurs critiques de la plus petite somme des rangs du test de Wilcoxon-Mann-Whitney

Valeurs critiques de la plus petite somme des rangs du test de Wilcoxon-Mann-Whitney

n_1 = nombre d'éléments dans le grand échantillon;

n_2 = nombre d'éléments dans le petit échantillon.

Niveau de signification α						Niveau de signification α					
Bilatéral		**0,20**	**0,10**	**0,05**	**0,01**	**Bilatéral**		**0,20**	**0,10**	**0,05**	**0,01**
Unilatéral		**0,10**	**0,05**	**0,025**	**0,005**	**Unilatéral**		**0,10**	**0,05**	**0,025**	**0,005**
n_1	n_2					n_1	n_2				
3	2	3	-	-	-	10	6	38	35	32	27
3	3	7	6	-	-	10	7	49	45	42	37
4	2	3	-	-	-	10	8	60	56	53	47
4	3	7	6	-	-	10	9	73	69	65	58
4	4	13	11	10	-	10	10	87	82	78	71
5	2	4	3	-	-	11	1	1	-	-	-
5	3	8	7	6	-	11	2	6	4	3	-
5	4	14	12	11	-	11	3	13	11	9	6
5	5	20	19	17	15	11	4	21	18	16	12
						11	5	30	27	24	20
6	2	4	3	-	-	11	6	40	37	34	28
6	3	9	8	7	-	11	7	51	47	44	38
6	4	15	13	12	10	11	8	63	59	55	49
6	5	22	20	18	16	11	9	76	72	68	61
6	6	30	28	26	13	11	10	91	86	81	73
						11	11	106	100	96	87
7	2	4	3	-	-						
7	3	10	8	7	-	12	1	1	-	-	-
7	4	16	14	13	10	12	2	7	5	4	-
7	5	23	21	20	16	12	3	14	11	10	7
7	6	32	29	27	24	12	4	22	19	17	13
7	7	41	39	36	32	12	5	32	28	26	21
						12	6	42	38	35	30
8	2	5	4	3	-	12	7	54	49	46	40
8	3	11	9	8	-	12	8	66	62	58	51
8	4	17	15	14	11	12	9	80	75	71	63
8	5	25	23	21	17	12	10	94	89	84	76
8	6	34	31	29	25	12	11	110	104	99	90
8	7	44	41	38	34	12	12	127	120	115	105
8	8	55	51	49	43						
						13	1	-	-	-	-
9	1	1	-	-	-	13	2	7	5	4	-
9	2	5	4	3	-	13	3	15	12	10	7
9	3	11	9	8	6	13	4	23	20	18	14
9	4	19	16	14	11	13	5	33	30	27	22
9	5	27	24	22	18	13	6	44	40	37	31
9	6	36	33	31	26	13	7	56	52	48	44
9	7	46	43	40	35	13	8	69	64	60	53
9	8	58	54	51	45	13	9	83	78	73	64
9	9	70	66	62	56	13	10	98	92	88	79
						13	11	114	108	103	93
10	1	1	-	-	-	13	12	131	125	119	109

Niveau de signification α						Niveau de signification α					
Bilatéral		0,20	0,10	0,05	0,01	Bilatéral		0,20	0,10	0,05	0,01
Unilatéral		0,10	0,05	0,025	0,005	Unilatéral		0,10	0,05	0,025	0,005
n_1	n_2					n_1	n_2				
14	1	1	-	-	-	17	4	28	25	21	16
14	2	7	5	4	-	17	5	40	35	32	25
14	3	16	13	11	7	17	6	52	47	43	36
14	4	25	21	19	14	17	7	66	61	56	47
14	5	35	31	28	22	17	8	81	75	70	60
14	6	46	42	38	32	17	9	97	90	84	74
14	7	59	54	50	43	17	10	113	106	100	89
14	8	72	67	62	54	17	11	131	123	117	105
14	9	86	81	76	67	17	12	150	142	135	122
14	10	102	96	91	81	17	13	170	161	154	140
14	11	118	112	106	96	17	14	190	182	174	159
14	12	136	129	123	112	17	15	212	203	195	180
14	13	154	147	141	129	17	16	235	225	217	201
14	14	174	166	160	147	17	17	259	249	240	223
15	1	1	-	-	-	18	1	1	-	-	-
15	2	8	6	4	-	18	2	9	7	5	-
15	3	16	13	11	8	18	3	19	15	13	8
15	4	26	22	20	15	18	4	30	26	22	16
15	5	37	33	29	23	18	5	42	37	33	26
15	6	48	44	40	33	18	6	55	49	45	37
15	7	61	56	52	44	18	7	69	63	58	49
15	8	75	69	65	56	18	8	84	77	72	62
15	9	90	84	79	69	18	9	100	93	87	76
15	10	106	99	94	84	18	10	117	110	13	92
15	11	123	116	110	99	18	11	135	127	121	108
15	12	141	133	127	115	18	12	155	146	139	125
15	13	159	152	145	133	18	13	175	166	158	144
15	14	179	171	164	151	18	14	196	187	179	163
15	15	200	192	184	171	18	15	218	208	200	184
						18	16	242	231	222	206
16	1	1	-	-	-	18	17	266	255	246	228
16	2	8	6	4	-	18	18	291	280	270	252
16	3	17	14	12	8						
16	4	27	24	21	15	19	1	2	1	-	-
16	5	38	34	30	24	19	2	10	7	5	3
16	6	50	46	42	34	19	3	20	16	13	9
16	7	64	58	54	46	19	4	31	27	23	17
16	8	78	72	67	58	19	5	43	38	34	27
16	9	93	87	82	72	19	6	57	51	46	38
16	10	109	103	97	86	19	7	71	65	60	50
16	11	127	120	113	102	19	8	87	80	74	64
16	12	145	138	131	119	19	9	103	96	90	78
16	13	165	156	150	130	19	10	121	113	107	94
16	14	185	176	169	155	19	11	139	131	124	111
16	15	206	197	190	175	19	12	159	150	143	129
16	16	229	219	211	196	19	13	180	171	163	147
						19	14	202	192	182	168
17	1	1	-	-	-	19	15	224	214	205	189
17	2	9	6	5	-	19	16	248	237	228	210
17	3	18	15	12	8						

Tableau A1.7 Tableau des valeurs critiques du test des rangs signés de Wilcoxon

Valeurs critiques du test des rangs signés de Wilcoxon

n	Test bilatéral		Test unilatéral	
	$\alpha = 0{,}05$	$\alpha = 0{,}01$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
5	--	--	0	--
6	0	--	2	--
7	2	--	3	0
8	3	0	5	1
9	5	1	8	3
10	8	3	10	5
11	10	5	13	7
12	13	7	17	9
13	17	9	21	12
14	21	12	25	15
15	25	15	30	19
16	29	19	35	23
17	34	23	41	27
18	40	27	47	32
19	46	32	53	37
20	52	37	60	43
21	58	42	67	49
22	65	48	75	55
23	73	54	83	62
24	81	61	91	69
25	89	68	100	76
26	98	75	110	84

Tableau A1.8 Tableau des valeurs critiques du test U de Mann-Whitney

Valeurs critiques du test U de Mann-Whitney
(test bilatéral)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	0,05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	0,01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	0,05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	0,01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	0,05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	0,01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	0,05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	0,01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	0,05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	0,01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	0,05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	0,01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	0,05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	0,01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	0,05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	0,01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	0,05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	0,01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	0,05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	0,01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	0,05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	0,01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	0,05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	0,01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	0,05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	0,01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	0,05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	0,01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	0,05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	0,01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	0,05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	0,01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	0,05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	0,01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	0,05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	0,01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

Valeurs critiques du test U de Mann-Whitney
(test unilatéral)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	0,05	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
	0,01	--	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	0,05	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	0,01	--	--	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	0,05	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	0,01	--	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6	0,05	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	0,01	--	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7	0,05	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	0,01	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8	0,05	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	0,01	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9	0,05	4	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	0,01	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10	0,05	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	0,01	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11	0,05	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	0,01	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12	0,05	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	0,01	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13	0,05	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	0,01	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14	0,05	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	0,01	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15	0,05	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	0,01	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16	0,05	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	0,01	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17	0,05	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	0,01	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18	0,05	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	0,01	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19	0,05	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	0,01	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20	0,05	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	0,01	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114