

# Introductory Statistical Methods for Engineering

---

# *Introductory Statistical Methods for Engineering*

---

BASSIMC AND BRYANLEE

eCampus Ontario Open Access



Introductory Statistical Methods for Engineering Copyright © 2024 by bassimc and bryanlee is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-nc-sa/4.0/), except where otherwise noted.



# Contents

Introductory Statistical Methods for Engineering	xi
Title Page	1
About This Book	2
Learning Outcomes	4
Figurative Overview of Learning Modules	7
Python Installation and Review	10
GitHub Repository Link	11
1.0.1 Introduction to Exploring Data	12
1.0.2 Attributions Part 1	15
1.1.1 Statistical Methods in Engineering	16
1.1.2 Variability	20
1.1.3 Types of Statistical Studies and Statistical Methods	22
1.1.4 Sampling	27
1.1.5 Types of Data	30
1.1.6 Measurement: Its Importance and Difficulty	34
1.1.7 Mathematical Models, Reality, and Data Analysis	39
1.1.8 Taxonomy of Variables in a Model	43
1.1.9 Tutorial 1 - Exploring Data with Python	46
2.0.1 Introduction Summarize, Visualize, and Communicate with Data	47
2.0.2 Attributions Part 2	49
2.1.1 Quantitative Data and Quantiles Introduction	50
2.1.2 Dot Diagrams and Stem-and-Leaf Plots	51
2.1.3 Frequency Tables and Histograms	55
2.1.4 Scatterplots and Run Charts	60
2.1.5 Quantiles and Quantile Plots	64

2.1.6	Boxplots	69
2.1.7	Q-Q Plots and Comparing Distributional Shapes	74
2.2.1	Measures of Location	84
2.2.2	Measures of Spread	87
2.2.3	Statistics and Parameters	90
2.2.4	Plots of Summary Statistics with Time and Factors	92
2.2.5	Bar Charts and Plots for Qualitative and Count Data	98
2.2.6	Summary Statistics and Statistical Computing	103
2.2.7	Tutorial 2 - Data Cleaning, Summarization, and Plotting in Python	105
3.0.1	Introduction to Probability and Random Variables	106
3.0.2	Attributions Part 3	108
3.1.1	Probability of Random Events	109
3.1.2	Probability and Independence of Events	115
3.1.3	Random Variables and Probability Distributions	119
3.1.4	Cumulative Distribution Functions	122
3.1.5	Discrete Random Variables and Continuous Random Variables	123
3.1.6	Summary of Probability Models	125
3.2.0	Introduction to Discrete Probability Distributions	126
3.2.1	Probability Mass Function (PMF) for a Discrete Random Variable	129
3.2.2	Cumulative Distribution Function	133
3.2.3	Probability Expressed to Two Decimal Places	135
3.2.4	Mean or Expected Value and Standard Deviation of Discrete Probability Distributions	136
3.2.5	Binomial Distribution	140
3.2.6	Poisson Distribution	145
3.2.7	Working with Discrete Probability Distributions in Python	149
4.0.1	Introduction to Continuous Random Variables and Probability Distributions	150
4.0.2	Attributions Part 4	152
4.1.1	Probability Density Functions and Cumulative Probability Function	153

4.1.2 Means and Variances for Continuous Distributions	157
4.1.3 Normal Probability Distribution	159
4.1.4 Standard Normal Distribution	161
4.1.5 The Empirical Rule	168
4.1.6 Tutorial 3 - Normal Probability Distributions	169
4.2.0 Introduction Joint Distributions and Independence	170
4.2.1 Joint Distributions	171
4.2.2 Conditional Distributions and Independence	178
4.2.3 Means and Variances for Linear Combinations of Random Variables	190
4.2.4 The Central Limit Theorem	195
5.0.1 Introduction to Formal Statistical Inference	201
5.0.1 Attributions	202
5.1.1 Large-Sample Confidence Intervals for a Mean	203
5.1.2 Large-Sample Significance Tests for a Mean	212
5.1.3 A Five-Step Format for Summarizing Significance Tests	218
5.1.4 Generally Applicable Large-n Significance Tests for Means.	220
5.1.5 Significance Testing and Formal Statistical Decision Making	222
5.1.6 Statistical Significance, Estimation, and Practical Importance	228
5.2.0 Introduction One- and Two-Sample Inference for Means	231
5.2.1 Small-Sample Inference for a Single Mean	232
5.2.2 Large-Sample Comparisons of Two Means (Based on Independent Samples)	239
5.2.3 Small-Sample Comparisons of Two Means (Based on Independent Samples from Normal Distributions)	244
5.2.4 Two-Sample Inference for Variances	251
5.2.5 Inference for the Mean of Paired Differences	259
5.2.6 Tutorial 4A - Inferential Statistics & T-Tests	264
5.3.0 Introduction to Nonparametric Models	265
5.3.1 Nonparametric Methods	266
5.3.2 Choosing The Appropriate Statistical Test	268
5.3.3 Comparing Two Independent Conditions: The Mann-Whitney U Test	269

5.3.4 The Wilcoxon Test for Paired Samples	270
5.3.5 Differences Between Several Independent Groups: The Kruskal–Wallis Test	272
5.3.6 Tutorial 4 - Non-Parametric Tests	274
6.0.1 Introduction to the One-Way Normal Model	275
6.0.2 Attributions Part 6	276
6.1.1 Graphical Comparison of Several Samples of Measurement Data	277
6.1.2 The One-Way (Normal) Multisample Model, Fitted Values, and Residuals	282
6.1.3 A Pooled Estimate of Variance for Multisample Studies	291
6.2.0 Introduction Confidence Intervals Multisample Studies	295
6.2.1 Intervals for Means and for Comparing Means	296
6.2.2 Individual and Simultaneous Confidence Levels	299
6.2.3 Simultaneous Confidence Interval Methods	301
6.3.0 Introduction ANOVA	304
6.3.1 Significance Testing and Multisample Studies	305
6.3.2 The One-Way ANOVA F Test	307
6.3.3 The One-Way ANOVA Identity and Table	311
6.3.4 Computing ANOVA in Python	316
7.0.1 Introduction Least Squares and Simple Linear Regression Analysis	319
7.0.2 Attributions	320
7.1.0 Introduction to Least Squares: Describing the Relationship between Bivariate Quantitative Data	321
7.1.1: Applying the Least Squares Principle	322
7.1.2 The Sample Correlation and Coefficient of Determination	327
7.1.3 Computing and Using Residuals	330
7.1.4 Cautions When Using Least Squares Line Fitting	335
7.1.5 Using Statistical Computing	337
7.1.6 Tutorial 5 - Correlation and Covariance	339
7.2.0 Introduction to Simple Linear Regression Inference Methods Related to the Least Squares Fitting of a Line (Simple Linear Regression)	340



7.2.1 The Simple Linear Regression Model, Corresponding Variance Estimate, and Standardized Residuals	341
7.2.2 Inference for the Slope Parameter	348
7.2.3 Inference for the Mean System Response for a Particular Value of x	351
7.2.4 Prediction and Tolerance Intervals	357
7.2.5 Simple Linear Regression and ANOVA	361
7.2.6 Statistical Computing for Simple Linear Regression: Pressure and Density Example	365
7.2.7 Tutorial 6 & 7 - Simple Linear Regression	368
8.0.1 Introduction to Multiple and Logistic Regression	369
8.0.2 Attributions	370
8.1.0 Introduction to Multiple Linear Regression: Fitting Curves and Surfaces by Least Squares	371
8.1.1 Curve Fitting by Least Squares	372
8.1.2 Transformations	382
8.1.3 Surface Fitting by Least Squares	386
8.1.4 Common Residual Plots in Multiple Regression	394
8.1.5 Interactions	395
8.1.6 Some Additional Cautions: Extrapolation, Outliers, and Parsimony	400
8.1.7 Statistical Computing with Python	403
8.1.8 Tutorial 8 - Transformations	404
8.1.9 Transitioning from Simple to Multiple Linear Regression in Python	405
8.2.1 Categorical Variable Independent Variables and Dummy Variables	406
8.2.2 Matrix Algebra and Multiple Regression	410
9.0.2 Attribution	412
9.0.1 Introduction to Design of Experiments	413
9.1.1 Design of Experiments: Introduction	414
9.1.2 Design of Experiments: Analysis	419
9.1.3 Tutorial 9 - Design of Experiments	424
9.2.1 Design of Experiments: Full Factorial Designs	425
9.2.2 Design of Experiments: Disturbances and Blocking	431

9.2.3 Design of Experiments: Fractional Designs	433
9.3.1 Design of Experiments: Optimization and Response Surface Methods	437
9.3.2 Design of Experiments: The General Approach	445
9.4.1 Design of Experiments Project	447
Table A1.1 Table of Standard Normal Probabilities	449
Table A1.2. Upper Tail Standard Normal Probabilities	454
Table A1.3. t Distribution Quantiles Table	456
Table A1.4 Chi-Square Distribution Quantiles	458
Table A1.5 F Distribution Tables	460
Table A1.6 Critical values of the smallest rank sum for the Wilcoxon-Mann-Whitney test	467
Table A1.7 Critical Values of the Wilcoxon Signed Ranks Test	471
Table A1.8 Critical Values of the Mann-Whitney U	473

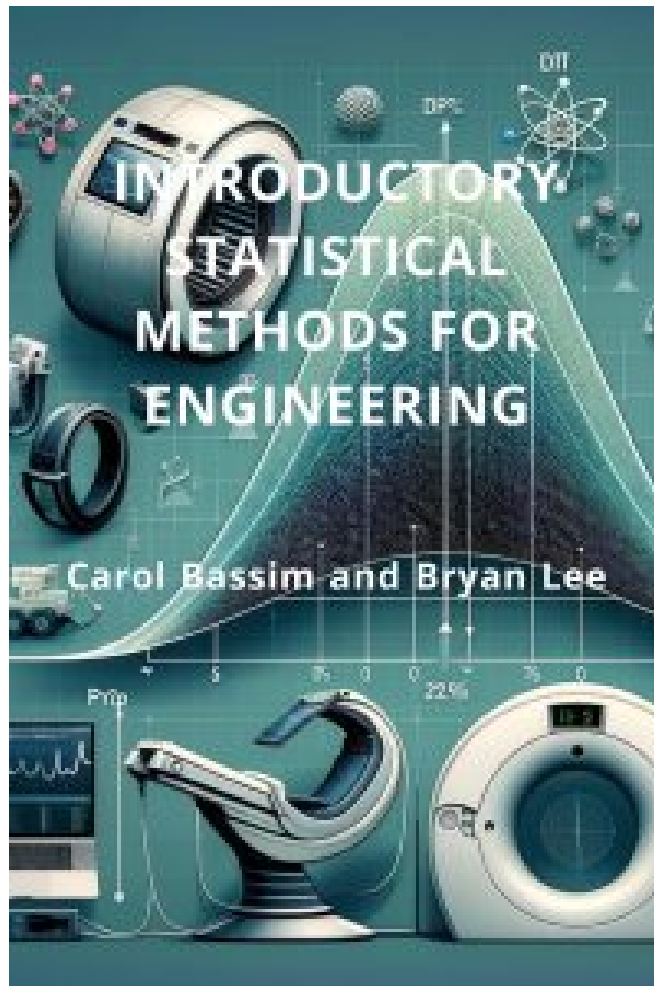
# *Introductory Statistical Methods for Engineering*

Funded by the Government of Ontario

The views expressed in this publication are the views of the author(s) and do not necessarily reflect those of the Government of Ontario or the Ontario Online Learning Consortium



## Title Page



Funded by the Government of Ontario

The views expressed in this publication are the views of the author(s) and do not necessarily reflect those of the Government of Ontario or the Ontario Online Learning Consortium



## About This Book

Welcome to the exciting and transformative world of Engineering Statistics, where mathematical theory and innovation converge to shape the future of engineering, technology, the environment, and healthcare. This open-access textbook is specially tailored for undergraduate students as an introductory or survey course, providing you with the foundational knowledge and practical skills necessary to thrive in the dynamic field of engineering and the specializations of the discipline.

### Why Statistics in Engineering?

Engineering is at the forefront of technological innovation and the lived experience of humanity.

### Exploring Diverse Domains

Throughout this textbook, you will embark on a journey through various domains within engineering and the need for statistical methods within these domains. Practical examples, case studies, and problem-solving exercises are woven into the fabric of this textbook and its associated resources, providing real-world context and hands-on experience. From theory to real-world applications, this text navigates through descriptive and analytical statistical tools and methodology, emphasizing their application in real-world engineering problems. You will learn not only the theory of statistics but also how to apply these concepts to design experiments, analyze data, and control processes in engineering contexts.

### Leveraging Open Access and Statistical Computing Resources for Exploration

We encourage you to make full use of the open access nature of this textbook, allowing you to comprehensively explore how statistics can be applied to engineering systems. Wherever your passion lies in engineering, this book will serve as an invaluable guide on your journey. Statistical computing support through tutorials residing at the associated GitHub repository offer practical examples and interaction with practical statistics and coding, as well as the ability to learn through simulation and exploration. The GitHub repository can be found here: [GitHub: Introductory Statistical Methods for Engineering](#).

### Attributions and What is New?

This first draft of the textbook is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

Iowa State University Professor Emeritus [Stephen Vardeman](#) and Miami University Professor Emeritus [J. Marcus Jobe](#) (ISU PhD, 1984) have made their book [Basic Engineering Data Collection and Analysis](#), originally published by Duxbury/Thompson Learning/Cengage, freely available for download under a (CC BY-NC-SA)

4.0 International license through the [Iowa State University Digital Press](#). The book is available [here](#) and has been assigned the following DOI: <https://doi.org/10.31274/isudp.2023.127>

The Basic Engineering Data Collection and Analysis book is essentially a revision/second edition of *Statistics for Engineering Problem Solving* by Vardeman that won the [American Society for Engineering Education](#) 1994 [Meriam/Wiley Distinguished Author Award](#).

This resource also has a reliance on a foundational statistics resources from “Process Improvement Using Data”. This is an invaluable legacy resource created and provided as an open educational resource by Kevin Dunn during his tenure at McMaster University between 2012 and 2016. Kevin’s resource was not just an invaluable for this text but for many educators globally, making engineering statistics and data science available, comprehensible, and applicable: [PID](#). This resource is [CC BY-SA 4.0](#).

It also draws on the very helpful “Introductory Statistics” from OpenStax by Barbara Illowsky and Susan Dean: [Introductory Statistics](#). [CC BY-NC-SA 4.0](#).

However, these resources, as well as many others, have benefited here from a synthesis approach for engineering and statistical computing support, and for a specificity for specializations in engineering. The use of Jupyter Notebooks and the coding language of Python are supported here as a practical experience and active learning experience, combining this text with the FAIR principles of open access resources, being Findable, Accessible, Interoperable, and Reusable.

### **A Journey of Impact**

As you embark on this educational adventure, remember that engineering is not just about engineering solutions; it’s about improving lives. Your work has the potential to significantly impact people and make a difference in the world. Together, we’ll embark on this transformative journey, where statistics and innovation go hand in hand.

**Let’s explore the exciting intersection of engineering, statistics, and technology, shaping the future together!**

# Learning Outcomes

## Learning Outcomes

Students will:

1. Master core principles of engineering statistics.
2. Implement data analytics tailored for engineering scenarios.
3. Develop hands-on Python skills through tutorials and simulations.
4. Apply statistical knowledge to real-world engineering challenges.

## Importance to the Field

These learning outcomes are essential for engineers because they provide a strong foundation in statistical analysis, data analytics, and practical programming skills in Python. By achieving these outcomes, students will be well-prepared to address complex engineering problems that require data-driven decision-making and statistical analysis.

## Parts, Modules, and Chapter

The following specific Parts and their associated learning outcomes, as taught in the Part modules and chapters, align with the broader goals outlined above.

### Part 1: Explore Data

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Understand the role of statistics in engineering.
- Apply statistical computing skills to data exploration.
- Clean data to prepare for statistical analysis applications.

### Part 2: Summarize, Visualize, and Communicate with Data

- Learn to plot and communicate effectively with data
- Display data graphically and interpret graphs.
- Recognize, describe, and calculate measures of data location and spread.

### Part 3: Probability and Discrete Random Variables

- Understand and use probability terminology.
- Calculate probabilities using Addition and Multiplication Rules.
- Construct and interpret Contingency Tables, Venn Diagrams, and Tree Diagrams.
- Recognize and understand discrete probability distribution functions.
- Calculate and interpret expected values.
- Apply various discrete probability distributions appropriately.

#### **Part 4: Continuous Random Variables and The Normal Probability Distribution**

- Recognize and understand continuous probability density functions.
- Apply continuous probability distributions appropriately.
- Recognize and apply the normal probability distribution.

#### **Part 5: Inferential Statistics and Hypothesis Testing with Samples**

- Apply and interpret the central limit theorem for means.
- Describe hypothesis testing and differentiate between types of hypothesis testing errors.
- Conduct and interpret hypothesis tests for population parameters.
- Conduct and interpret hypothesis tests for two population parameters.
- Understand and apply non-parametric methods for comparing distributions.
- Calculate and interpret confidence intervals for population parameters.
- Determine required sample sizes for confidence intervals.
- Understand and communicate about the p-value and statistical test conclusions.
- Confidently choose between statistical tests.

#### **Part 6: Inference for Unstructured Multisample Studies and ANOVA**

- Interpret the F probability distribution.
- Conduct and interpret one-way ANOVA and tests of variances.
- Conduct individual and simultaneous confidence interval methods for one-way ANOVA.

#### **Part 7: Least Squares and Simple Linear Regression Analysis**

- Discuss linear regression and correlation concepts.
- Create and analyze scatter plots, calculate correlation coefficients, and identify outliers.
- Make conclusions about simple linear regression models and confidently communicate conclusions.
- Fit established models and create new models from data.

#### **Part 8: Multiple Linear Regression Analysis**

- Apply multiple regression analysis.
- Learn model fitting and building for multiple linear regression.



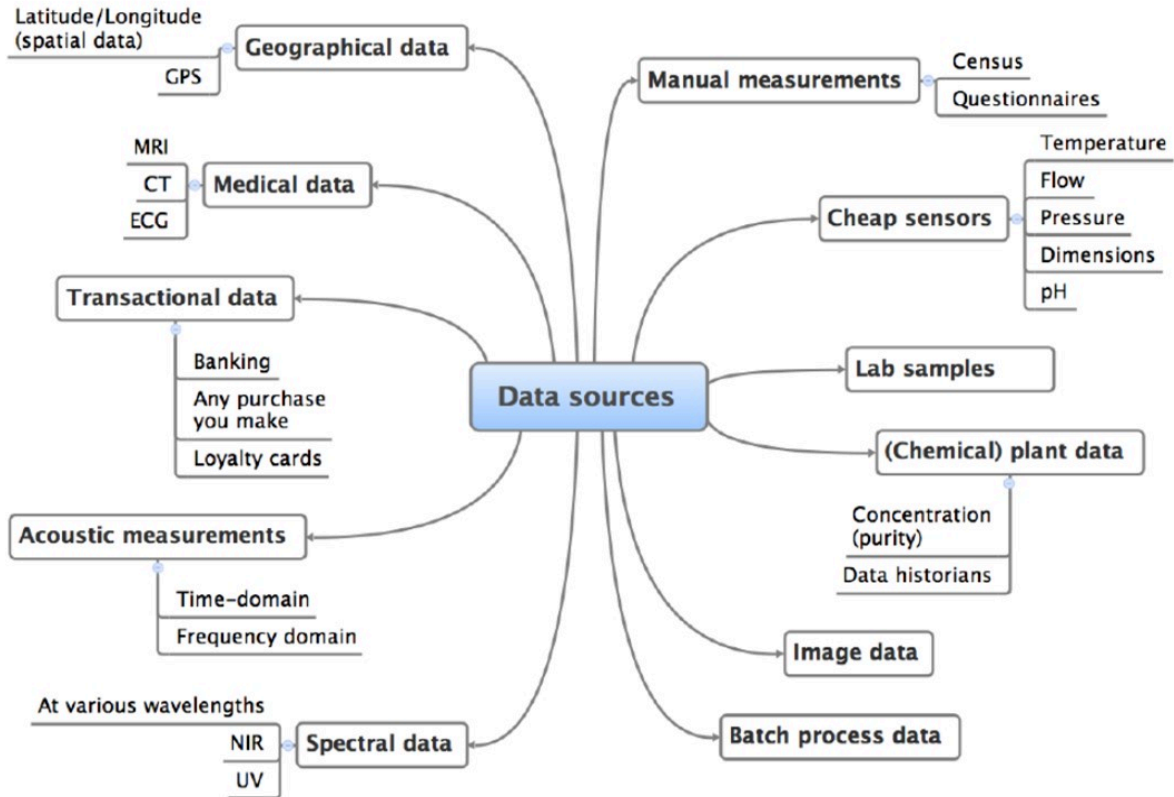
- Introduction to Full Factorial Design of Experiments.

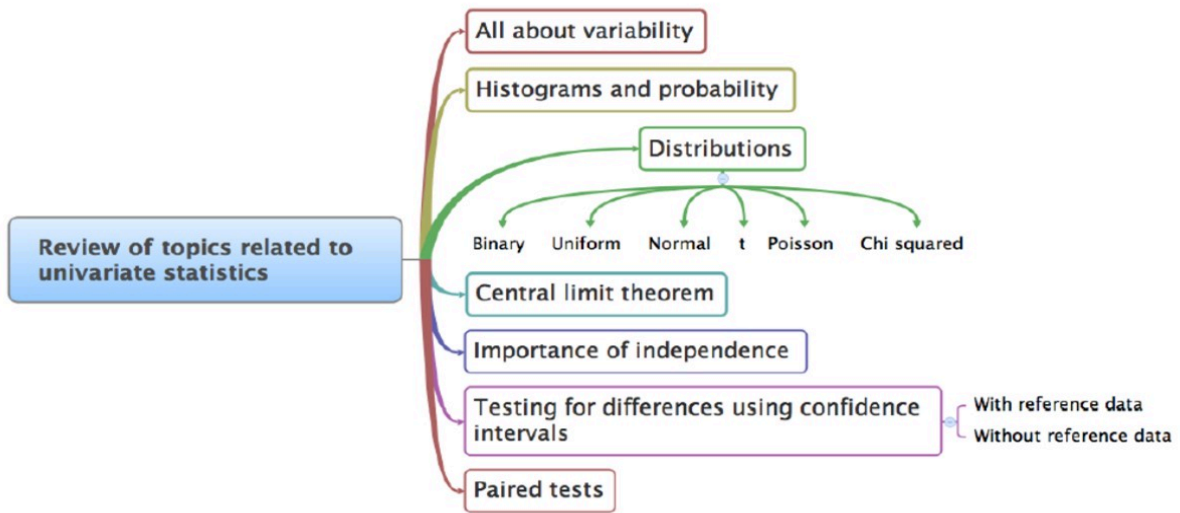
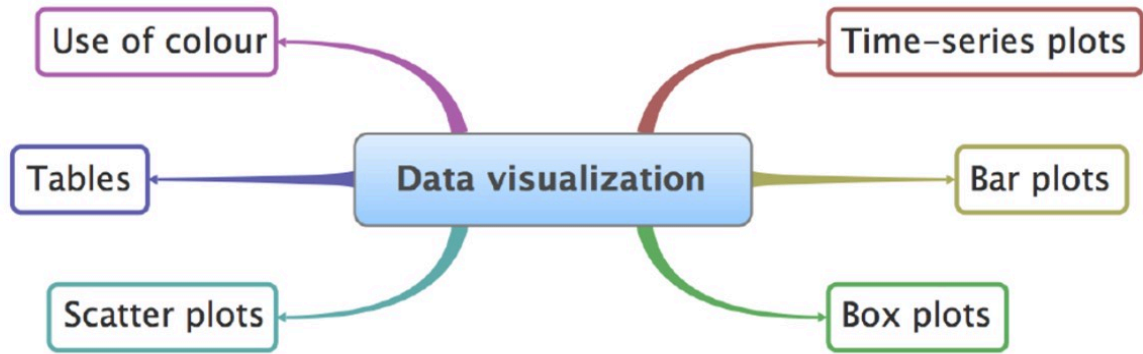
### **Part 9: Design of Experiments**

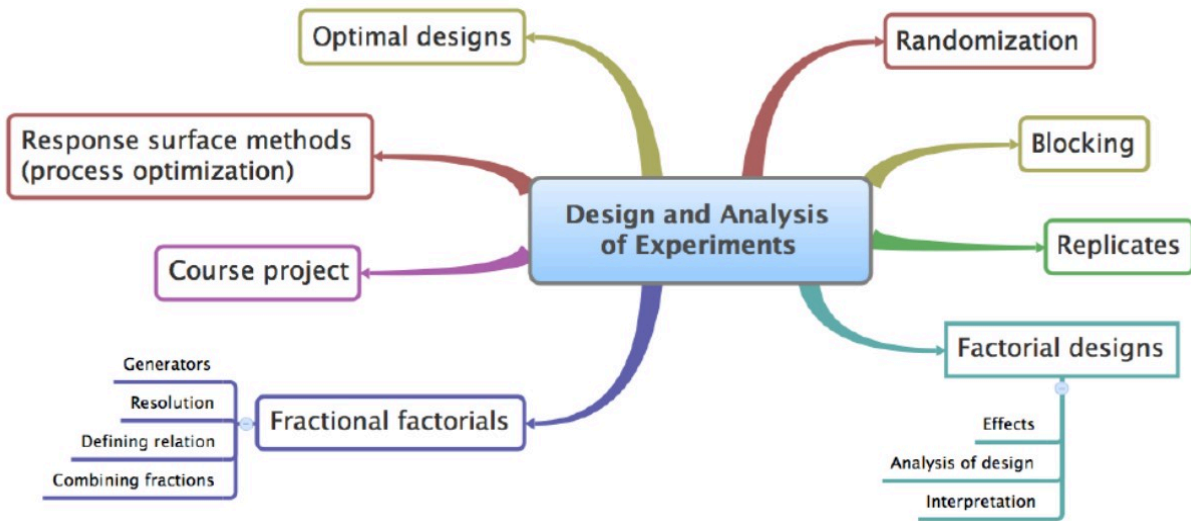
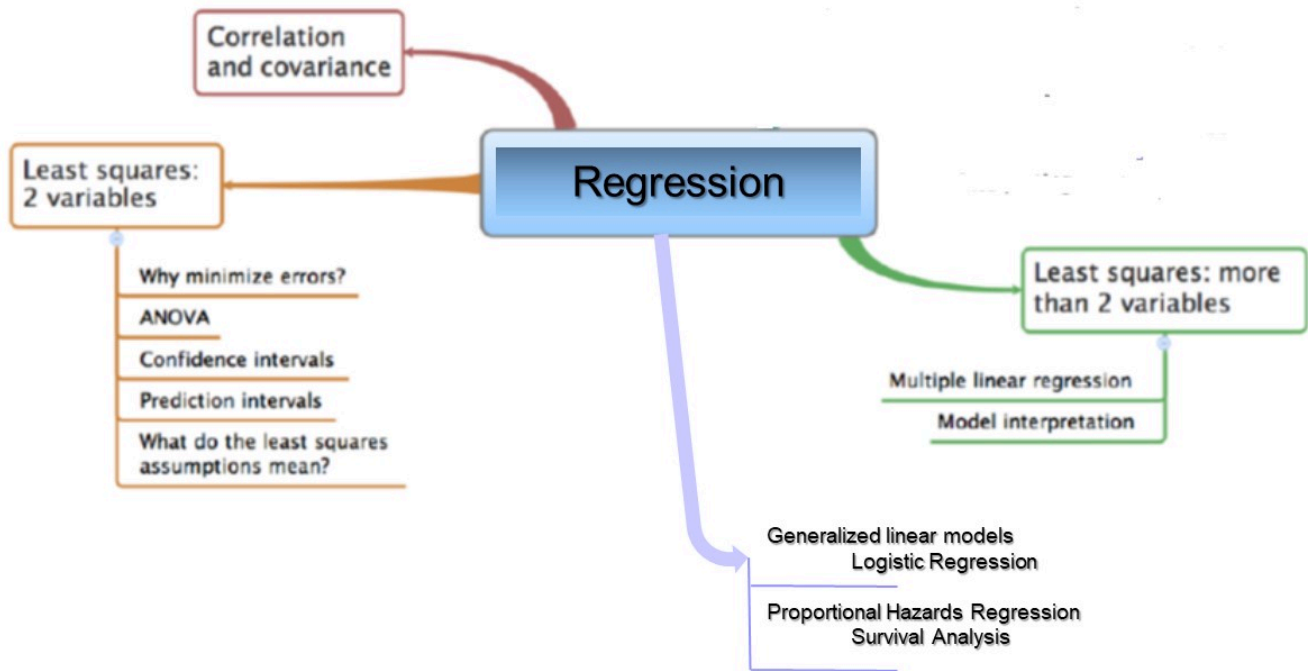
- Apply and implement a design of experiment.
- Apply full and fractional designs.
- Understand and utilize Surface Response Methods and Optimization Methods.

Overall, these modules and learning outcomes equip engineering students with the statistical knowledge and skills needed to excel in their field, enabling them to make data-driven decisions and tackle engineering challenges effectively.

## Figurative Overview of Learning Modules







Attribution: This Figurative Overview of Learning Modules is from "Process Improvement Using Data". by Kevin Dunn . This resource is available at [PID](#) and any material is copyrighted to him and shared by [CC BY-SA 4.0](#).

## Python Installation and Review

To take full advantage of this resource, it is strongly recommended that you utilize a statistical package that can read the Python code. We recommend using Jupyter Lab or Jupyter Notebook with the Anaconda package. See the instructions below for installation for different operating systems.

### Steps for Installation:

1. Navigate to the Anaconda Webpage and Download the appropriate setup file.
  - Link to download Anaconda: <https://www.anaconda.com/download#downloads>
2. Follow the appropriate instructions
  - 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. If you have trouble installing Python, you can use Google Colab: <https://colab.google/>
  - As long as you have a gmail account, you can log-in and use Jupyter Notebook through your internet browser.

**A Python Review is available at the [GitHub Site](#) of the course, at [Getting Started with Python](#).**

## GitHub Repository Link

### GitHub Repository

The GitHub repository can be found here: [GitHub: Introductory Statistical Methods for Engineering](#).

Statistical computing support through tutorials residing at the associated GitHub repository offer practical examples and interaction with practical statistics and coding, as well as the ability to learn through simulation and exploration. The GitHub repository can be found here: [GitHub: Introductory Statistical Methods for Engineering](#).

The repository holds the Python based Jupyter Notebook files for this course. **It is recommended that you download the specific files to your computer and run them locally.** However, you can also work through interactive Jupyter Notebooks associated with the course modules without using anything else, find the BinderHub badge on the ReadMe section of the repository and click on it.

These interaction links are also incorporated throughout the text of this resource to be able to work through examples in the text at the same time as you review the concepts in each module, through Special GitHub Site repositories.

## 1.0.1 Introduction to Exploring Data



1912 photograph of Karl Pearson (By Unknown author - Google Books - Nock, Albert Jay (1912-03). "A New Science And Its Findings". *The American Magazine* LXXIII: 579. The Phillips Publishing Co., Public Domain, <https://commons.wikimedia.org/w/index.php?curid=4578734>, and Google Books: Karl Pearson, *The Grammar of Science*, Adam and Charles Black, 1911 London: [https://www.google.com/books/edition/The\\_Grammar\\_of\\_Science/9mISAAAIAAJ?hl=en&gbpv=1&dq=grammar+of+science&printsec=frontcover](https://www.google.com/books/edition/The_Grammar_of_Science/9mISAAAIAAJ?hl=en&gbpv=1&dq=grammar+of+science&printsec=frontcover), Public Domain.

Karl Pearson, a pioneering and problematic English mathematician and biostatistician born in 1857, profoundly impacted the field of statistics. His book, "The Grammar of Science," first published in 1892, is a pivotal work in scientific philosophy, and can be seen as a link between statistics and the engineering in that it focuses on the importance of statistical methods in comprehending and articulating natural phenomena. This perspective is particularly resonant in engineering, where observation, measurement, description,

technical communication, and creative application— key aspects of the scientific method and heavily reliant on statistical reasoning— are fundamental.

Statistics and statistical methods are vital in engineering and biomedical engineering, playing a crucial role in the design, analysis, and interpretation of data. As these fields increasingly rely on technology and data, statistical literacy and being able to use “the grammar of science” becomes essential for biomedical engineers.

#### Key Takeaways

**This course will be about harnessing data and describing and communicating about its uncertainty using statistical methods.**

These methods are key in healthcare and necessary for creating, testing, and understanding the impact of new biomedical technologies, which produce vast data amounts. In real-world applications, unlike in pure mathematics, data always contain errors and variation. Statistics aid in making informed decisions amidst this inherent uncertainty, a critical skill in various fields including economics, health, business, and engineering.

Statistics involves two main areas- descriptive methods, which summarize sample data, and inferential methods, which draw conclusions about a larger population. Exploring and cleaning data and defining data type is crucial for choosing an appropriate statistical analysis. Understanding and communicating about data’s central tendency and variation is vital, involving measures like mean, median, mode, standard deviation, and interquartile range.

This Part of the course will focus on the core concepts of statistics and introduce the use of statistical computing and some fundamental concepts of data science to be able to apply statistical methods to data. Data science is the interdisciplinary field of statistics, scientific computing, and science and engineering used to extract and use knowledge from data. For this course, we will be using Python based JupyterLab Notebooks as a statistical computing tool to explore and practice the application of statistical concepts.

#### Learning Objectives

##### Learning Outcomes for Part 1:

- Differentiate between descriptive and inferential statistics and understand their applications in engineering contexts.
- Understand basic statistical samples and sampling techniques.
- Review and understand experimental design and designed experiments in engineering.
- Identify, classify, and use different types of statistical data and data types (categorical, ranked, discrete, continuous).
- Review the fundamentals of data cleaning and preparation for data exploration.

##### Learning Outcomes for Part 1- Jupyter Notebook Tutorials:

- Open and use a JupyterLab Notebook tutorial and read in a simple dataset.
- Use statistical computing to clean and prepare data.

This Course Part 1 lays a foundation for all that follows: It contains a road map for the study of engineering



statistics. The subject is defined, its importance is described, some basic terminology is introduced, and the important issue of measurement is discussed. Finally, the role of mathematical models in achieving the objectives of engineering statistics is investigated.

## 1.0.2 *Attributions Part 1*

This first draft of Part 1 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#). The chapter on Variability comes directly from this resource, and is the copyright of Kevin Dunn.

## 1.1.1 Statistical Methods in Engineering

In general terms, what a working engineer does is to design, build, operate, and/or improve physical systems and products. This work is guided by basic mathematical and physical theories learned in an undergraduate engineering curriculum. As the engineer's experience grows, these quantitative and scientific principles work along-side sound engineering judgment. But as technology advances and new systems and products are encountered, the working engineer is inevitably faced with questions for which theory and experience provide little help. When this happens, what is to be done?

On occasion, consultants can be called in, but most often an engineer must independently find out “what makes things tick.” It is necessary to **collect and interpret data** that will help in understanding how the new system or product works. Without specific training in data collection and analysis, the engineer's attempts can be haphazard and poorly conceived. Valuable time and resources are then wasted, and sometimes erroneous (or at least unnecessarily ambiguous) conclusions are reached. To avoid this, it is vital for a working engineer to have a toolkit that includes the best possible principles and methods for gathering and interpreting data. This toolkit is the **statistical methods for engineering**.

The goal of engineering statistics is to provide the concepts and methods needed by an engineer who faces a problem for which independent judgment is needed or new innovation is required. It supplies principles for how to efficiently acquire and process empirical information needed to understand and manipulate engineering systems.

### **DEFINITION 1.1.1.1. Engineering Statistics**

Engineering statistics is the study of how best to

1. collect engineering data,
2. summarize or describe engineering data, and
3. draw formal inferences and practical conclusions on the basis of engineering data, all the while recognizing the reality of variation.

To better understand the definition, it is helpful to consider how the elements of engineering statistics enter into a real problem.

### Example 1.1.1.1. Heat Treating Gears.

The article “Statistical Analysis: Mack Truck Gear Heat Treating Experiments” by P. Brezler (Heat Treating, November, 1986) describes a simple application of engineering statistics. A process engineer was faced with the question, “How should gears be loaded into a continuous carburizing furnace in order to minimize distortion during heat treating?” Various people had various semi-informed opinions about how it should be done—in particular, about whether the gears should be laid flat in stacks or hung on rods passing through the gear bores. But no one really knew the consequences of laying versus hanging.

#### Data Collection

In order to settle the question, the engineer decided to get the facts—to collect some data on “thrust face runout” (a measure of gear distortion) for gears laid and gears hung. Deciding exactly how this data collection should be done required careful thought. There were possible differences in gear raw material lots, machinists and machines that produced the gears, furnace conditions at different times and positions within the furnace, technicians and measurement devices that would produce the final runout measurements, etc. The engineer did not want these differences either to be mistaken for differences between the two loading techniques or to unnecessarily cloud the picture. Avoiding this required care.

In fact, the engineer conducted a well-thought-out and executed study. Table 1.1.1.1 shows the runout values obtained for 38 gears laid and 39 gears hung after heat treating. In raw form, the runout values are hardly understandable. They lack organization; it is not possible to simply look at Table 1.1.1.1 and tell what is going on. The data needed to be summarized.

#### Data Summarization

One thing that was done was to compute some numerical summaries of the data. For example, the process engineer found

$$\text{Mean laid runout} = 12.6$$

$$\text{Mean hung runout} = 17.9$$

#### Visualization

Figure 1.1.1.1

Further, a simple graphical summarization was made, as shown in

#### Variation

From these summaries of the runouts, several points are obvious. One is that there is variation in the runout values, even within a particular loading method. Variability is an omnipresent fact of life, and all statistical methodology explicitly recognizes this. In the case of the gears, it appears from Figure 1.1.1.1 that there is somewhat more variation in the hung values than in the laid values. But in spite of the variability that complicates comparison between the loading methods, Figure 1.1.1.1 and the two group means also carry the message that the laid runouts are on the whole smaller than the hung runouts. By how much? One answer is

$$\text{Mean hung runout} - \text{Mean laid runout} = 5.3$$

But how “precise” is this figure? Runout values are variable. So is there any assurance that the difference seen in the present means would reappear in further testing? Or is it possibly explainable as simply “stray background noise”? Laying gears is more expensive than hanging them. Can one know whether the extra expense is justified?

## Drawing Inferences from Data

These questions point to the need for methods of formal statistical inference from data and translation of those inferences into practical conclusions. Methods presented in this text can, for example, be

used to support the following statements about hanging and laying gears:

- One can be roughly 90% sure that the difference in long-run mean runouts produced under conditions like those of the engineer's study is in the range

3.2 to 7.4

- One can be roughly 95% sure that 95% of runouts for gears laid under conditions like those of the engineer's study would fall in the range

3.0 to 22.2

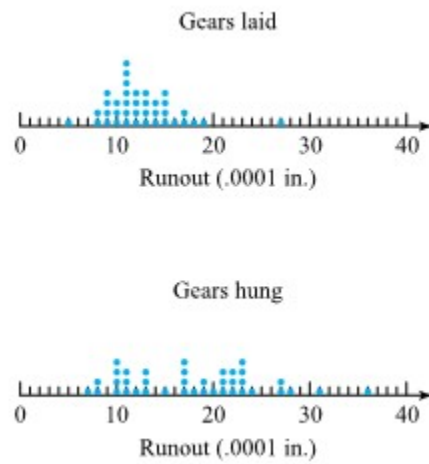
- One can be roughly 95% sure that 95% of runouts for gears hung under conditions like those of the engineer's study would fall in the range

.8 to 35.0

These are formal quantifications of what was learned from the study of laid and hung gears. To derive practical benefit from statements like these, the process engineer had to combine them with other information, such as the consequences of a given amount of runout and the costs for hanging and laying gears, and had to apply sound engineering judgment. Ultimately, the runout improvement was great enough to justify some extra expense, and the laying method was implemented.

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

Table 1.1.1.1. Thrust Face Runouts (.0001 in.)



*Figure 1.1.1.1. Dot diagrams of runouts*

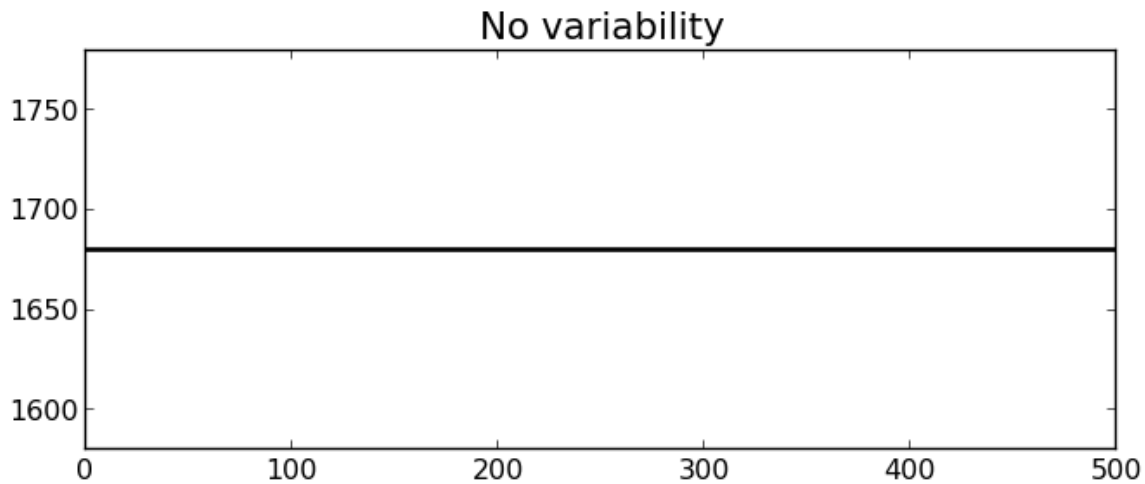
The example shows how the elements of statistics were helpful in solving an engineer's problem. Throughout this text, the intention is to emphasize that the topics discussed are not ends in themselves, but rather methods that engineers can use to help them do their jobs effectively.

---

## 1.1.2 Variability

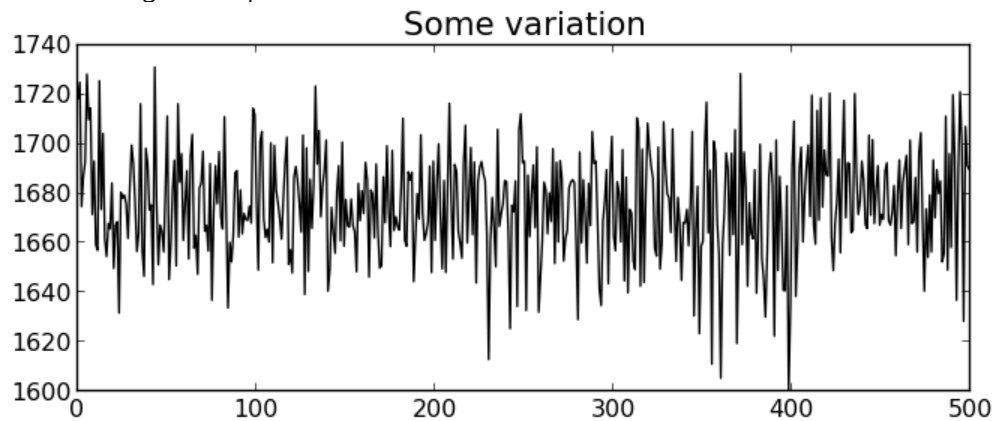
### WHAT IS VARIABILITY?

Life is pretty boring without variability, and this course, and almost all the field of statistics would be unnecessary if things did not naturally vary.

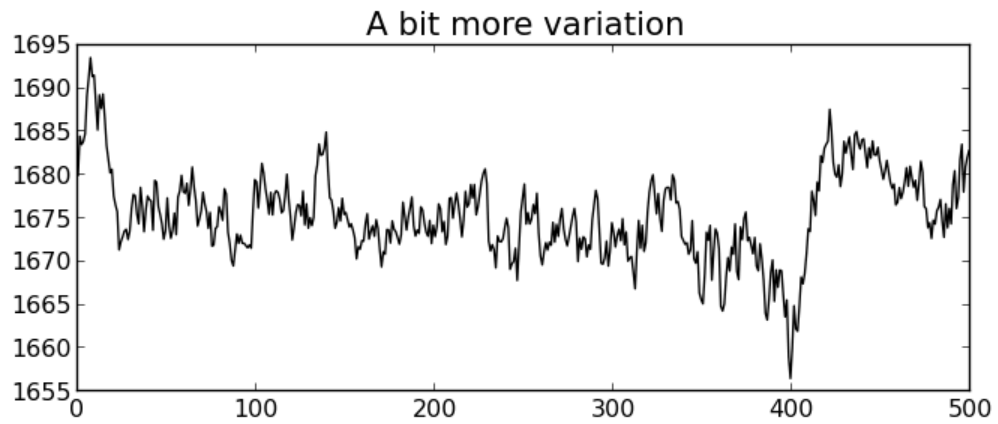


Fortunately, we have plenty of variability in the recorded data from our processes and systems:

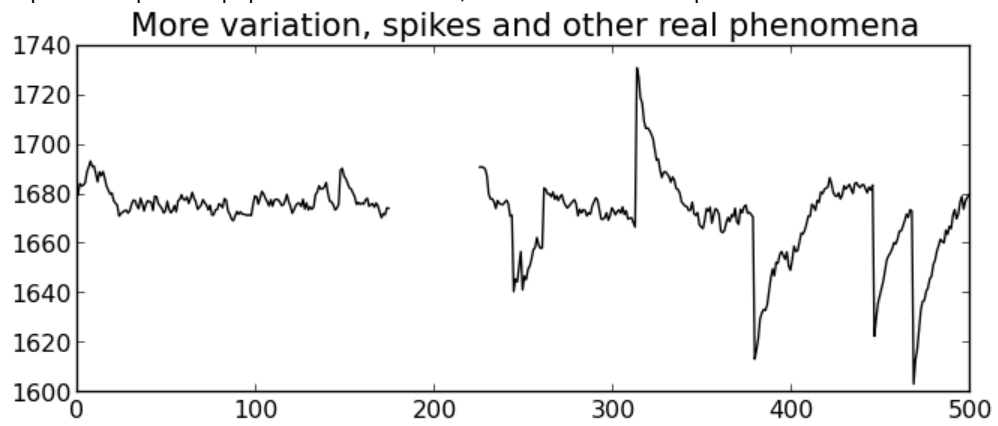
- Raw material and input properties are not constant.
- Unknown sources, often called “*error*” or “*noise*”. These errors are all sources of variation which our imperfect knowledge of the process cannot account for.



- Measurement and sampling variability: sensor drift, spikes, noise, recalibration shifts, errors in our sample analysis and laboratory equipment.



- Production disturbances:
  - external conditions change, such as ambient temperature, or humidity, and
  - pieces of plant equipment break down, wear out and are replaced.





## 1.1.3 Types of Statistical Studies and Statistical Methods

When an engineer sets about to gather data, he or she must decide how active to be. Will the engineer turn knobs and manipulate process variables or simply let things happen and try to record the salient features?

### **DEFINITION 1.2.3.1. Observational Study**

An observational study is one in which the investigator's role is basically passive. A process or phenomenon is watched and data are recorded, but there is no intervention on the part of the person conducting the study.

### **DEFINITION 1.2.3.2. Experimental Study**

An experimental study (or, more simply, an experiment) is one in which the investigator's role is active. Process variables are manipulated, and the study environment is regulated.

Most real statistical studies have both observational and experimental features, and these two definitions should be thought of as representing idealized opposite ends of a continuum. On this continuum, the experimental end usually provides the most efficient and reliable ways to collect engineering data. It is typically much quicker to manipulate process variables and watch how a system responds to the changes than to passively observe, hoping to notice something interesting or revealing.

### **Inferring causality**

In addition, it is far easier and safer to infer causality from an experiment than from an observational study. Real systems are complex. One may observe

several instances of good process performance and note that they were all surrounded by circumstances X without being safe in assuming that circumstances X cause good process performance. There may be important variables in the background that are changing and are the true reason for instances of favorable system behavior. These so-called lurking variables may govern both process performance and circumstances X. Or it may simply be that many variables change haphazardly without appreciable impact on the system and that by chance, during a limited period of observation, some of these happen to produce X at the same time that good performance occurs. In either case, an engineer's efforts to create X as a means of making things work well will be wasted effort.

On the other hand, in an experiment where the environment is largely regulated except for a few variables the engineer changes in a purposeful way, an inference of causality is much stronger. If circumstances created by the investigator are consistently accompanied by favorable results, one can be reasonably sure that they caused the favorable results.

#### Example 1.1.3.1. Pelletizing Hexamine Powder

Cyr, Ellson, and Rickard attacked the problem of reducing the fraction of non-conforming fuel pellets produced in the compression of a raw hexamine powder in a pelletizing machine. There were many factors potentially influencing the percentage of nonconforming pellets: among others, Machine Speed, Die Fill Level, Percent Paraffin added to the hexamine, Room Temperature, Humidity at manufacture, Moisture Content, "new" versus "reground" Composition of the mixture being pelletized, and the Roughness of the chute entered by the freshly stamped pellets. Correlating these many factors to process performance through passive observation was hopeless.

The students were, however, able to make significant progress by conducting an experiment. They chose three of the factors that seemed most likely to be important and purposely changed their levels while holding the levels of other factors as close to constant as possible. The important changes they observed in the percentage of acceptable fuel pellets were appropriately attributed to the influence of the system variables they had manipulated.

Besides the distinction between observational and experimental statistical studies, it is helpful to distinguish between studies on the basis of the intended breadth of application of the results. Two relevant terms, popularized by the late W. E. Deming, are defined next:

#### **DEFINITION 1.1.3.3. Enumerative study**

An enumerative study is one in which there is a particular, well-defined, finite group of objects under study. Data are collected on some or all of these objects, and conclusions are intended to apply only to these objects.

**DEFINITION 1.1.3.4. Analytical study**

An analytical study is one in which a process or phenomenon is investigated at one point in space and time with the hope that the data collected will be representative of system behavior at other places and times under similar conditions. In this kind of study, there is rarely, if ever, a particular well-defined group of objects to which conclusions are thought to be limited.

Most engineering studies tend to be of the second type, although some important engineering applications do involve enumerative work. One such example is the reliability testing of critical components—e.g., for use in a space shuttle. The interest is in the components actually in hand and how well they can be expected to perform rather than on any broader problem like “the behavior of all components of this type.” Acceptance sampling (where incoming lots are checked before taking formal receipt) is another important kind of enumerative study. But as indicated, most engineering studies are analytical in nature.

**Example 1.1.3.1. continued**

The students working on the pelletizing machine were not interested in any particular batch of pellets, but rather in the question of how to make the machine work effectively. They hoped (or tacitly assumed) that what they learned about making fuel pellets would remain valid at later times, at least under shop conditions like those they were facing. Their experimental study was analytical in nature.

Particularly when discussing enumerative studies, the next two definitions are needed.

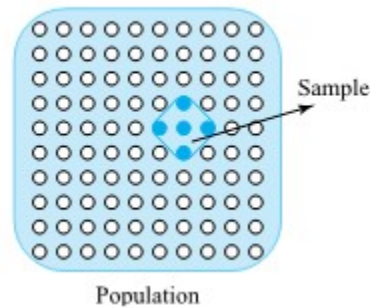
**DEFINITION 1.1.3.5. Population**

A population is the entire group of objects about which one wishes to gather information in a statistical study.

**DEFINITION 1.1.3.6. Sample**

A sample is the group of objects on which one actually gathers data. In the case of an enumerative investigation, the sample is a subset of the population (and can in some cases include the entire population).

Figure 1.1.3.1 shows the relationship between a population and a sample. If a crate of 100 machine parts is delivered to a loading dock and 5 are examined in order to verify the acceptability of the lot, the 100 parts constitute the population of interest, and the 5 parts make up a (single) sample of size 5 from the population. (Notice the word usage here: There is one sample, not five samples.)



*Figure 1.1.3.1. Population and sample*

There are several ways in which the meanings of the words population and sample are often extended. For one, it is common to use them to refer to not only objects under study but also data values associated with those objects. For example, if one thinks of Rockwell hardness values associated with 100 crated machine parts, the 100 hardness values might be called a population (of numbers). Five hardness values corresponding to the parts examined in acceptance sampling could be termed a sample from that population.

#### **Example 1.1.3.1. continued**

Cyr, Ellson, and Rickard identified eight different sets of experimental conditions under which to run the pelletizing machine. Several production runs of fuel pellets were made under each set of conditions, and each of these produced its own percentage of conforming pellets. These eight sets of percentages can be referred to as eight different samples (of numbers).

Also, although strictly speaking there is no concrete population being investigated in an analytical study, it is common to talk in terms of a conceptual population in such cases. Phrases like “the population consisting of all widgets that could be produced under these conditions” are sometimes used. This can sometimes be confusing. But it is a common usage, and it is supported by the fact that typically the same mathematics is used when drawing inferences in enumerative and analytical contexts.

## TYPES OF STATISTICAL METHODS

Two main statistical methods are used in data analysis: descriptive statistics and inferential statistics. Descriptive statistics summarize data from a sample, such as by using the mean and standard deviation of a sample, and will be the main consideration for Part 2 of this course. Inferential statistics draw conclusions from data drawn from a sample that are subject to random variation. Inferential statistics uses a probability model to describe the process from which the data were obtained, which we will learn about in Part 3 and Part 4. Data are then used to draw conclusions about the process by estimating parameters in the model and making predictions based on the model. We will first learn about formal inferential tests of statistics in Part 5 of this course. Figure 1.1.2.2 shows how descriptive and inferential statistics are related.

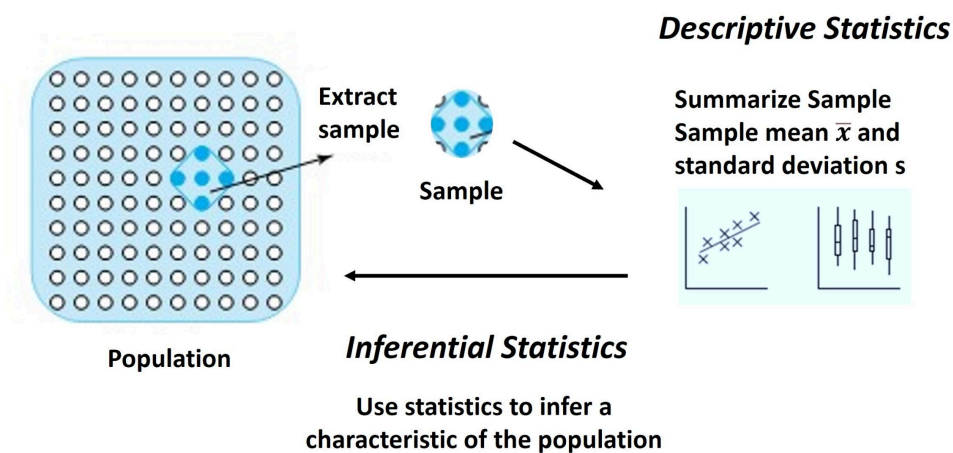


Figure 1.1.2.2. How descriptive and inferential statistics are related.

## 1.1.4 Sampling

### SAMPLING IN ENUMERATIVE STUDIES

---

An enumerative study has an identifiable, concrete population of items. This chapter discusses selecting a sample of the items to include in a statistical investigation.

Using a sample to represent a (typically much larger) population has obvious advantages. Measuring some characteristics of a sample of 30 electrical components from an incoming lot of 10,000 can often be feasible in cases where it would not be feasible to perform a census (a study that attempts to include every member of the population). Sometimes testing is destructive, and studying an item renders it unsuitable for subsequent use. Sometimes the timeliness and data quality of a sampling investigation far surpass anything that could be achieved in a census. Data collection technique can become lax or sloppy in a lengthy study. A moderate amount of data, collected under close supervision and put to immediate use, can be very valuable—often more valuable than data from a study that might appear more complete but in fact takes too long.

If a sample is to be used to stand for a population, how that sample is chosen becomes very important. The sample should somehow be representative of the population. The question addressed here is how to achieve this.

Systematic and judgment-based methods can in some circumstances yield samples that faithfully portray the important features of a population. If a lot of items is manufactured in a known order, it may be reasonable to select, say, every 20th one for inclusion in a statistical engineering study. Or it may be effective to force the sample to be balanced—in the sense that every operator, machine, and raw material lot (for example) appears in the sample. Or an old hand may be able to look at a physical population and fairly accurately pick out a representative sample.

But there are potential problems with such methods of sample selection. Humans are subject to conscious and subconscious preconceptions and biases. Accordingly, judgment-based samples can produce distorted pictures of populations. Systematic methods can fail badly when unexpected cyclical patterns are present. (For example, suppose one examines every 20th item in a lot according to the order in which the items come off a production line. Suppose further that the items are at one point processed on a machine having five similar heads, each performing the same operation on every fifth item. Examining every 20th item only gives a picture of how one of the heads is behaving. The other four heads could be terribly misadjusted, and there would be no way to find this out.)

Even beyond these problems with judgment-based and systematic methods of sampling, there is the additional difficulty that it is not possible to quantify their properties in any useful way. There is no good way to take information from samples drawn via these methods and make reliable statements of likely margins of error. The method introduced next avoids the deficiencies of systematic and judgment-based sampling.

**DEFINITION 1.1.4.1. Simple random sample**

A simple random sample of size  $n$  from a population is a sample selected in such a manner that every collection of  $n$  items in the population is a priori equally likely to compose the sample.

Probably the easiest way to think of simple random sampling is that it is conceptually equivalent to drawing  $n$  slips of paper out of a hat containing one for each member of the population.

**Example 1.1.4.1. Random Sampling Dorm Residents**

C. Black did a partially enumerative and partially experimental study comparing student reaction times under two different lighting conditions. He decided to recruit subjects from his coed dorm floor, selecting a simple random sample of 20 of these students to recruit. In fact, the selection method he used involved a table of so-called random digits. He could today use a random number generator using a statistical computing package. But he could have just as well written the names of all those living on his floor on standard-sized slips of paper, put them in a bowl, mixed thoroughly, closed his eyes, and selected 20 different slips from the bowl.

*Mechanical Methods, Random Digit Tables, and Simple Random Samples*

Methods for actually carrying out the selection of a simple random sample include mechanical methods and methods using “random digits.” Mechanical methods rely for their effectiveness on symmetry and/or thorough mixing in a physical randomizing device. So to speak, the slips of paper in the hat need to be of the same size and well scrambled before sample selection begins.

The first Vietnam-era U.S. draft lottery was a famous case in which adequate care was not taken to ensure appropriate operation of a mechanical randomizing device. Birthdays were supposed to be assigned priority numbers 1 through 366 in a “random” way. However, it was clear after the fact that balls representing birth dates were placed into a bin by months, and the bin was poorly mixed. When the balls were drawn out, birth dates near the end of the year received a disproportionately large share of the low draft numbers. In the present terminology, the first five dates out of the bin should not have been thought of as a simple random sample of size 5. Those who operate games of chance more routinely make it their business to know (via the collection of appropriate data) that their mechanical devices are operating in a more random manner.

Using random digits to do sampling implicitly relies for “randomness” on the appropriateness of the method used to generate those digits. Physical random processes like radioactive decay and pseudorandom number generators (complicated recursive numerical algorithms) are the most common sources of random digits. Until fairly recently, it was common to record such digits in printed tables.

*Statistical Software and Random Samples*

With the wide availability of personal computers, random digit tables have become largely obsolete. That is, random numbers can be generated “on the spot” using statistical or spreadsheet software.

### *Notes on Random Sampling*

---

Regardless of how Definition 1.1.4.1 is implemented, several comments about the method are in order. First, it must be admitted that simple random sampling meets the original objective of providing representative samples only in some average or long-run sense. It is possible for the method to produce particular realizations that are horribly unrepresentative of the corresponding population. A simple random sample of 20 out of 80 axles could turn out to consist of those with the smallest diameters. But this doesn't happen often. On the average, a simple random sample will faithfully portray the population. Definition 1.1.4.1 is a statement about a method, not a guarantee of success on a particular application of the method.

Second, it must also be admitted that there is no guarantee that it will be an easy task to make the physical selection of a simple random sample. Imagine the pain of retrieving 5 out of a production run of 1,000 microwave ovens stored in a warehouse. It would probably be a most unpleasant job to locate and gather 5 ovens corresponding to randomly chosen serial numbers to, for example, carry to a testing lab.

But the virtues of simple random sampling usually outweigh its drawbacks. For one thing, it is an objective method of sample selection. An engineer using it is protected from conscious and subconscious human bias. In addition, the method interjects probability into the selection process in what turns out to be a manageable fashion. As a result, the quality of information from a simple random sample can be quantified. Methods of formal statistical inference, with their resulting conclusions ("I am 95% sure that ..."), can be applied when simple random sampling is used.



## 1.1.5 Types of Data

Engineers encounter many types of data. One useful distinction concerns the degree to which engineering data are intrinsically numerical.

### **DEFINITION 1.1.5.1. Categorical Data**

Qualitative or categorical data are the values of basically nonnumerical characteristics associated with items in a sample. There can be an order associated with qualitative data, but aggregation and counting are required to produce any meaningful numerical values from such data.

Consider again 5 machine parts constituting a sample from 100 crated parts. If each part can be classified into one of the (ordered) categories (1) conforming, (2) rework, and (3) scrap, and one knows the classifications of the 5 parts, one has 5 qualitative data points. If one aggregates across the 5 and finds 3 conforming, 1 reworkable, and 1 scrap, then numerical summaries have been derived from the original categorical data by counting.

In contrast to categorical data are numerical data.

### **DEFINITION 1.1.5.2. Numerical Data**

Quantitative or numerical data are the values of numerical characteristics associated with items in a sample. These are typically either counts of the number of occurrences of a phenomenon of interest or measurements of some physical property of the items.

Returning to the crated machine parts, Rockwell hardness values for 5 selected parts would constitute a

set of quantitative measurement data. Counts of visible blemishes on a machined surface for each of the 5 selected parts would make up a set of quantitative count data.

It is sometimes convenient to act as if infinitely precise measurement were possible. From that perspective, measured variables are continuous in the sense that their sets of possible values are whole (continuous) intervals of numbers. For example, a convenient idealization might be that the Rockwell hardness of a machine part can lie anywhere in the interval  $(0, \infty)$ . But of course this is only an idealization. All real measurements are to the nearest unit (whatever that unit may be). This is becoming especially obvious as measurement instruments are increasingly equipped with digital displays. So in reality, when looked at under a strong enough magnifying glass, all numerical data (both measured and count alike) are discrete in the sense that they have isolated possible values rather than a continuum of available outcomes. Although  $(0, \infty)$  may be mathematically convenient and completely adequate for practical purposes, the real set of possible values for the measured Rockwell hardness of a machine part may be more like  $\{.1, .2, .3, \dots\}$  than like  $(0, \infty)$ .

Well-known conventional wisdom is that measurement data are preferable to categorical and count data. Statistical methods for measurements are simpler and more informative than methods for qualitative data and counts. Further, there is typically far more to be learned from appropriate measurements than from qualitative data taken on the same physical objects. However, this must sometimes be balanced against the fact that measurement can be more time-consuming (and thus expensive) than the gathering of qualitative data.

#### Example 1.1.5.1. Pellet Mass Measurements

As a preliminary to their experimental study on the pelletizing process (discussed in Example 1.1.3.1), Cyr, Ellson, and Rickard collected data on a number of aspects of machine behavior. Included was the mass of pellets produced under standard operating conditions. Because a nonconforming pellet is typically one from which some material has broken off during production, pellet mass is indicative of system performance. Informal requirements for (specifications on) pellet mass were from 6.2 to 7.0 grams.

Information on 200 pellets was collected. The students could have simply observed and recorded whether or not a given pellet had mass within the specifications, thereby producing qualitative data. Instead, they took the time necessary to actually measure pellet mass to the nearest .1 gram—thereby collecting measurement data. A graphical summary of their findings is shown in Figure 1.1.5.1

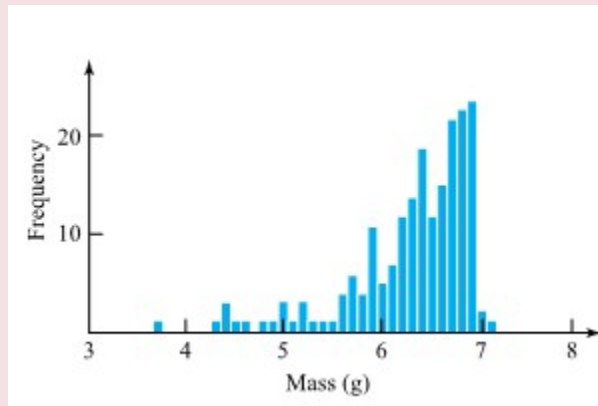


Figure 1.1.5.1 Pellet mass measurements

Notice that one can recover from the measurements the conformity/nonconformity information—about 28.5% (57 out of 200) of the pellets had masses that did not meet specifications. But there is much more in Figure 1.1.5.1 besides this. The shape of the display can

give insights into how the machine is operating and the likely consequences of simple modifications to the pelletizing process. For example, note the truncated or chopped-off appearance of the figure. Masses do not trail off on the high side as they do on the low side. The students reasoned that this feature of their data had its origin in the fact that after powder is dispensed into a die, it passes under a paddle that wipes off excess material before a cylinder compresses the powder in the die. The amount initially dispensed to a given die may have a fairly symmetric mound-shaped distribution, but the paddle probably introduces the truncated feature of the display.

Also, from the numerical data displayed in Figure 1.1.5.1, one can find a percentage of pellet masses in any interval of interest, not just the interval [6.2, 7.0]. And by mentally sliding the figure to the right, it is even possible to project the likely effects of increasing die size by various amounts.

It is typical in engineering studies to have several response variables of interest. The next definitions present some jargon that is useful in specifying how many variables are involved and how they are related.

**DEFINITION 1.1.5.3. Univariate**

Univariate data arise when only a single characteristic of each sampled item is observed.

**DEFINITION 1.1.5.4. Multivariate**

Multivariate data arise when observations are made on more than one characteristic of each sampled item. A special case of this involves two characteristics—**bivariate data**.

**DEFINITION 1.1.5.5. Repeated Measures**

When multivariate data consist of several determinations of basically the same characteristic (e.g., made with different instruments or at different times), the data are called repeated measures data. In the special case of bivariate responses, the term paired data is used.

It is important to recognize the multivariate character of data when it is present. Having Rockwell hardness values for 5 of 100 crated machine parts and determinations of the percentage of carbon for 5 other parts is not at all equivalent to having both hardness and carbon content values for a single sample of 5 parts. There are two samples of 5 univariate data points in the first case and a single sample of 5 bivariate data points in the second. The second situation is preferable to the first, because it allows analysis and exploitation of any relationships that might exist between the variables Hardness and Percent Carbon.

#### Example 1.1.5.2. Paired Distortion Measurements

In the furnace-loading scenario discussed in Example 1.1.1.1, radial runout measurements were actually made on all  $38 + 39 = 77$  gears both before and after heat treating. (Only after-treatment values were given in Table 1.1.) Therefore, the process engineer had two samples (of respective sizes 38 and 39) of paired data. Because of the pairing, the engineer was in the position of being able (if desired) to analyze how post-treatment distortion was correlated with pretreatment distortion.

## 1.1.6 Measurement: Its Importance and Difficulty

Success in statistical engineering studies requires the ability to measure. For some physical properties like length, mass, temperature, and so on, methods of measurement are commonplace and obvious. Often, the behavior of an engineering system can be adequately characterized in terms of such properties. But when it cannot, engineers must carefully define what it is about the system that needs observing and then apply ingenuity to create a suitable method of measurement.

### Example 1.1.6.1. Measuring Brittleness

A senior design capstone in metallurgical engineering took on the project of helping a manufacturer improve the performance of a spike-shaped metal part. In its intended application, this part needed to be strong but very brittle. When meeting an obstruction in its path, it had to break off rather than bend, because bending would in turn cause other damage to the machine in which the part functions. As the class planned a statistical study aimed at finding what variables of manufacture affect part performance, the students came to realize that the company didn't have a good way of assessing part performance. As a necessary step in their study, they developed a measuring device. It looked roughly as in Figure 1.1.7.1. A swinging arm with a large mass at its end was brought to a horizontal position, released, and allowed to swing through a test part firmly fixed in a vertical position at the bottom of its arc of motion. The number of degrees past vertical that the arm traversed after impact with the part provided an effective measure of brittleness.

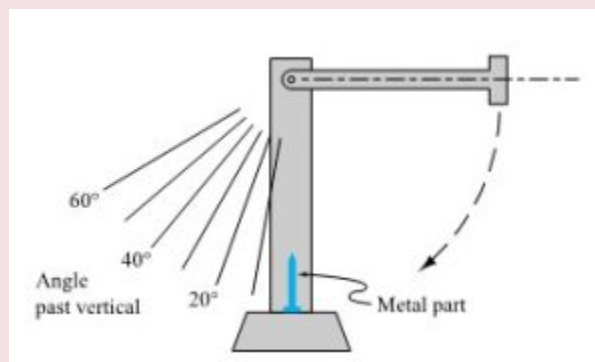


Figure 1.1.6.1. A device for measuring brittleness

### Example 1.1.6.2. Measuring Wood Joint Strength

Dimond and Dix wanted to conduct a study comparing joint strengths for combinations of three different woods and three glues. They

didn't have access to strength-testing equipment and so invented their own. To test a joint, they suspended a large container from one of the pieces of wood involved and poured water into it until the weight was sufficient to break the joint. Knowing the volume of water poured into the container and the density of water, they could determine the force required to break the joint.

Regardless of whether an engineer uses off-the-shelf technology or must fabricate a new device, a number of issues concerning measurement must be considered. These include validity, measurement variation/error, accuracy, and precision.

#### **DEFINITION 1.1.6.1. Validity**

A measurement or measuring method is called valid if it usefully or appropriately represents the feature of an object or system that is of engineering importance.

It is impossible to overstate the importance of facing the question of measurement validity before plunging ahead in a statistical engineering study. Collecting engineering data costs money. Expending substantial resources collecting data, only to later decide they don't really help address the problem at hand, is unfortunately all too common.

### **Measurement Error**

The point was made in Section 1.1.1.1 that when using data, one is quickly faced with the fact that variation is omnipresent. Some of that variation comes about because the objects studied are never exactly alike. But some of it is due to the fact that measurement processes also have their own inherent variability. Given a fine enough scale of measurement, no amount of care will produce exactly the same value over and over in repeated measurement of even a single object. And it is naive to attribute all variation in repeat measurements to bad technique or sloppiness. (Of course, bad technique and sloppiness can increase measurement variation beyond that which is unavoidable.)

An exercise suggested by W. J. Youden in his book *Experimentation and Measurement* is helpful in making clear the reality of measurement error. Consider measuring the thickness of the paper in this book. The technique to be used is as follows. The book is to be opened to a page somewhere near the beginning and one somewhere near the end. The stack between the two pages is to be grasped firmly between the thumb and index finger and stack thickness read to the nearest .1 mm using an ordinary ruler. Dividing the stack thickness by the number of sheets in the stack and recording the result to the nearest .0001 mm will then produce a thickness measurement.

#### **Example 1.1.6.3. Book Paper Thickness Measurements**

Presented below are ten measurements of the thickness of the paper in Box, Hunter, and Hunter's Statistics for Experimenters made one semester by engineering students Wendel and Gulliver.

Wendel: .0807,.0826,.0854,.0817,.0824,  
.0799,.0812,.0807,.0816,.0804

Gulliver: .0972,.0964,.0978,.0971,.0960,  
.0947,.1200,.0991,.0980,.1033

Figure 1.1.6.2 shows a graph of these data and clearly reveals that even repeated measurements by one person on one book will vary and also that the patterns of variation for two different individuals can be quite different. (Wendel's values are both smaller and more consistent than Gulliver's.)

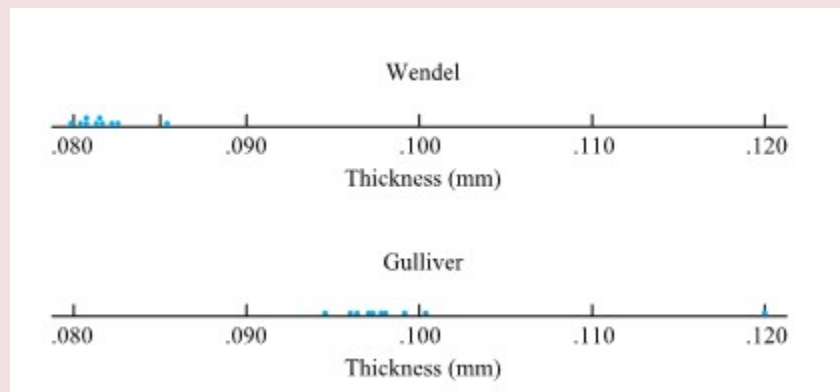


Figure 1.1.6.2. Dot diagrams of paper thickness measurements

The variability that is inevitable in measurement can be thought of as having both internal and external components.

#### Definition 1.1.7.2. Precision

A measurement system is called precise if it produces small variation in repeated measurement of the same object.

Precision is the internal consistency of a measurement system; typically, it can be improved only with basic changes in the configuration of the system.

#### Example 1.1.6.3. continued

Ignoring the possibility that some property of Gulliver's book was responsible for his values showing more spread than those of Wendel, it appears that Wendel's measuring technique was more precise than Gulliver's. The precision of both students' measurements could probably have been improved by giving each a binder clip and a micrometer. The binder clip would provide a relatively constant

pressure on the stacks of pages being measured, thereby eliminating the subjectivity and variation involved in grasping the stack firmly between thumb and index finger. For obtaining stack thickness, a micrometer is clearly a more precise instrument than a ruler.

Precision of measurement is important, but for many purposes it alone is not adequate.

### Definition 1.1.7.3 Accuracy

A measurement system is called accurate (or sometimes, unbiased) if on average it produces the true or correct value of a quantity being measured.

Accuracy is the agreement of a measuring system with some external standard. It is a property that can typically be changed without extensive physical change in a measurement method. Calibration of a system against a standard (bringing it in line with the standard) can be as simple as comparing system measurements to a standard, developing an appropriate conversion scheme, and thereafter using converted values in place of raw readings from the system.

### Example 1.1.6.3. continued

It is unknown what the industry-standard measuring methodology would have produced for paper thickness in Wendel's copy of the text. But for the sake of example, suppose that a value of .0850 mm/sheet was appropriate. The fact that Wendel's measurements averaged about .0817 mm/sheet suggests that her future accuracy might be improved by proceeding as before but then multiplying any figure obtained by the ratio of .0850 to .0817—i.e., multiplying by 1.04.

Maintaining Canada's reference sets for physical measurement is the business of Measurement Canada. In the USA it is the National Institute of Standards and Technology. It is important business. Poorly calibrated measuring devices may be sufficient for local purposes of comparing local conditions. But to establish the values of quantities in any absolute sense, or to expect local values to have meaning at other places and other times, it is essential to calibrate measurement systems against a constant standard. A millimeter must be the same today in Ontario as it was last week in British Columbia.

## Accuracy and statistical studies

The possibility of bias or inaccuracy in measuring systems has at least two important implications for planning statistical engineering studies. First, the fact that be monitored over time and that they be recalibrated as needed. The well-known phenomenon of instrument drift can ruin an otherwise flawless statistical study. Second, whenever possible, a single system should be used to do all measuring. If several measurement devices or technicians are used, it is hard to know whether the differences observed originate with the variables under study or from differences in devices or technician biases. If the use of several

The possibility of bias or inaccuracy in measuring systems has at least two important implications for planning statistical engineering studies. First, the fact that be monitored over time and that they be recalibrated as needed.



measurement systems is unavoidable, they must be calibrated against a standard (or at least against each other). The following example illustrates the role that human differences can play.

#### Example 1.1.6.4. Differences Between Technicians in Their Use of a Gauge

Cowan, Renk, Vander Leest, and Yakes worked with a company on the monitoring of a critical dimension of a high-precision metal part produced on a computer-controlled lathe. They encountered large, initially unexplainable variation in this dimension between different shifts at the plant. This variation was eventually traced not to any real shift-to-shift difference in the parts but to an instability in the company's measuring system. A single gauge was in use on all shifts, but different technicians used it quite differently when measuring the critical dimension. The company needed to train the technicians in a single, standardized method of using the gauge.

An analogy that is helpful in understanding the difference between precision and accuracy involves comparing measurement to target shooting. In target shooting, one can be on or off target (accurate or inaccurate) with a small or large cluster of shots (showing precision or imprecision). Figure 1.1.7.2 illustrates this analogy.

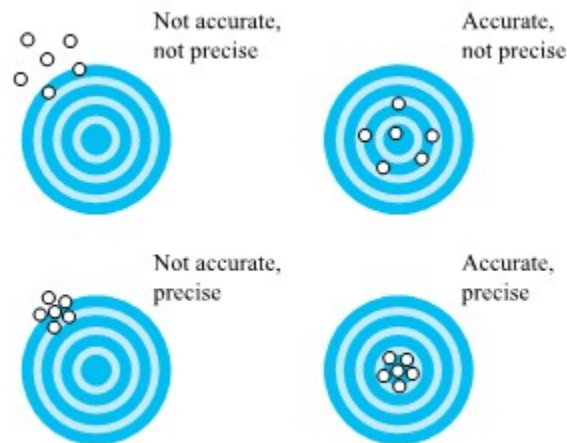


Figure 1.1.6.2. Measurement / Target shooting analogy.

Good measurement is hard work, but without it data collection is futile. To make progress, engineers must obtain valid measurements, taken by methods whose precision and accuracy are sufficient to let them see important changes in system behavior. Usually, this means that measurement inaccuracy and imprecision must be an order of magnitude smaller than the variation in measured response caused by those changes.

## 1.1.7 Mathematical Models, Reality, and Data Analysis

One can learn the basics of statistics and the statistical methods of engineering without an understanding of the underlying mathematics. Statistics contains a fair amount of mathematics that most engineering readers will find to be reasonably understandable—if unfamiliar and initially puzzling. But a learning context based in mathematics provides a much deeper and better path to being able to utilize the statistical methods of engineering. It is also a good application of the mathematical theory and application that students have learned in a practical application. Therefore, it seems wise to try to put the mathematical content of the book in perspective early. In this section, the relationships of mathematics to the physical world and to engineering statistics are discussed.

### Mathematical models and reality

Mathematics is a construct and a tool. While it is of interest to some people in its own right, engineers generally approach mathematics from the point of view that it can be useful in describing and predicting how physical systems behave. Indeed, mathematical theories are guides in every branch of modern engineering.

Throughout this text, we will frequently use the phrase mathematical model.

#### **DEFINITION 1.1.7.1. Mathematical model**

A mathematical model is a description or summarization of salient features of a real-world system or phenomenon in terms of symbols, equations, numbers, and the like.

Mathematical models are themselves not reality, but they can be extremely effective descriptions of reality. This effectiveness hinges on two somewhat opposing properties of a mathematical model: (1) its degree of simplicity and (2) its predictive ability. The most powerful mathematical models are those that simultaneously are simple and generate good predictions. A model's simplicity allows one to maneuver within its framework, deriving mathematical consequences of basic assumptions that translate into predictions of process behavior. When these are empirically correct, one has an effective engineering tool.

The elementary “laws” of mechanics are an outstanding example of effective mathematical modeling. For example, the simple mathematical statement that the acceleration due to gravity is constant,

$$a = g$$

yields, after one easy mathematical maneuver (an integration), the prediction that beginning with 0 velocity, after a time  $t$  in free fall an object will have velocity

$$v = gt$$

And a second integration gives the prediction that beginning with 0 velocity, a time  $t$  in free fall produces displacement

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

The beauty of this is that for most practical purposes, these easy predictions are quite adequate. They agree well with what is observed empirically and can be counted on as an engineer designs, builds, operates, and/or improves physical processes or products.

### Mathematical models in statistics

But then, how does the notion of mathematical modeling interact with the subject of engineering statistics? There are several ways. For one, data collection and analysis are essential in fitting or estimating parameters of mathematical models. To understand this point, consider again the example of a body in free fall. If one postulates that the acceleration due to gravity is constant, there remains the question of what numerical value that constant should have. The parameter  $g$  must be evaluated before the model can be used for practical purposes. One does this by gathering data and using them to estimate the parameter.

A standard first college physics lab has traditionally been to empirically evaluate  $g$ . The method often used is to release a steel bob down a vertical wire running through a hole in its center and allowing 60-cycle current to arc from the bob through a paper tape to another vertical wire, burning the tape slightly with every arc. A schematic diagram of the apparatus used is shown in Figure 1.1.7.1. The vertical positions of the burn marks are bob positions at intervals of  $\frac{1}{60}$  of a second. Table 1.1.7.1 gives measurements of such positions. (Dr.

Frank Peterson of the ISU Physics and Astronomy Department supplied the tape.) Plotting the bob positions in the table at equally spaced intervals produces the approximately quadratic plot shown in Figure 1.1.7.2. Picking a parabola to fit the plotted points involves identifying an appropriate value for  $g$ . A method of curve fitting called least squares produces a value for  $g$  of  $9.79m/sec^2$ , not far from the commonly quoted value of  $9.8m/sec^2$ .

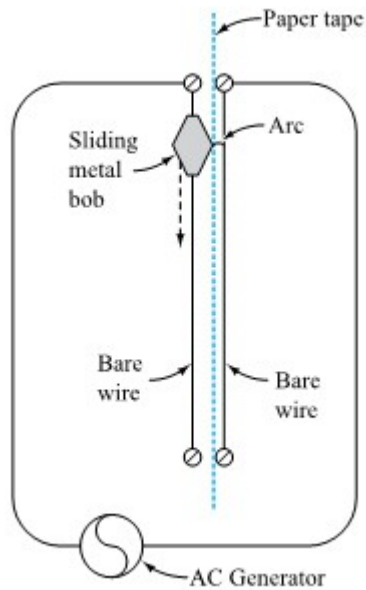


Figure 1.1.7.1. A device for measuring  $g$

Point Number	Displacement (mm)	Point Number	Displacement (mm)
1	.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		

Table 1.1.7.1. Measured Displacements of a Bob in Free Fall

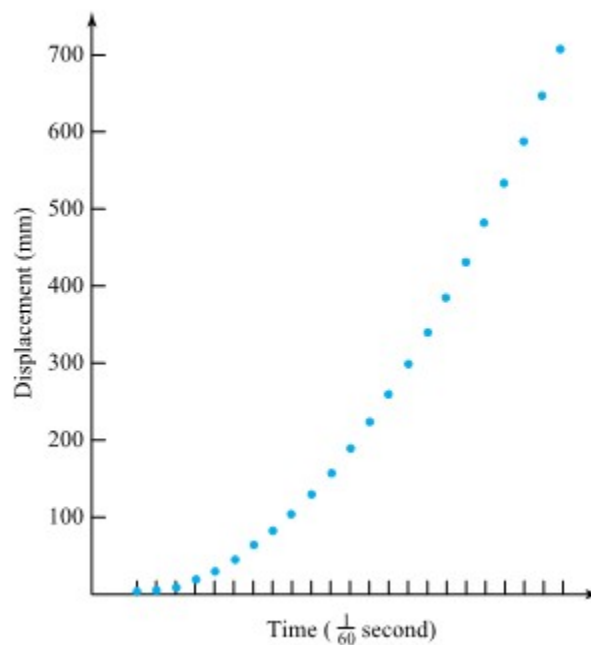


Figure 1.1.7.2. Bob positions in free fall

Notice that (at least before Newton) the data in Table 1.1.7.1 might also have been used in another way. The parabolic shape of the plot in Figure 1.1.7.2 could have suggested the form of an appropriate model for the motion of a body in free fall. That is, a careful observer viewing the plot of position versus time should conclude that there is an approximately quadratic relationship between position and time (and from that proceed via two differentiations to the conclusion that the acceleration due to gravity is roughly constant). This text is full of examples of how helpful it can be to use data both to identify potential forms for empirical models and to then estimate parameters of such models (preparing them for use in prediction).

This discussion has concentrated on the fact that statistics provides raw material for developing realistic mathematical models of real systems. But there is another important way in which statistics and mathematics interact. The mathematical theory of probability provides a framework for quantifying the uncertainty associated with inferences drawn from data.

**DEFINITION 1.1.7.2. Probability**

Probability is the mathematical theory intended to describe situations and phenomena that one would colloquially describe as involving chance.

If, for example, five students arrive at the five different laboratory values of  $g$ ,

$$9.78, 9.82, 9.81, 9.78, 9.79$$

questions naturally arise as to how to use them to state both a best value for  $g$  and some measure of precision for the value. The theory of probability provides guidance in addressing these issues. Material in Part 3 shows that probability considerations support using the class average of 9.796 to estimate  $g$  and attaching to it a precision on the order of plus or minus  $.02m/sec^2$ .

The mathematics of probability is a full subject on its own, so this text will only supply a minimal introduction to the subject. But do not lose sight of the fact that probability is not statistics—nor vice versa. Rather, probability is a branch of mathematics and a useful subject in its own right. It is met in a statistics course as a tool because the variation that one sees in real data is closely related conceptually to the notion of chance modeled by the theory of probability.

## 1.1.8 Taxonomy of Variables in a Model

One of the hard realities of statistical modelling and experiment planning is the multidimensional nature of the world. There are typically many characteristics of observed but non-experimental systems and system performance by experimentation that the engineer would like to understand and many variables that might influence them. Some terminology is needed to facilitate clear thinking and discussion in light of this complexity.

### **DEFINITION 1.1.8.1. Response Variable**

A response variable in an experiment is one that is monitored as characterizing system performance/behavior. It is the dependent variable in the system model.

### **DEFINITION 1.1.8.2. Input Variable**

For existing data that was not experimentally collected, a system input variable acts as the variable that influences the model, or the independent variable of interest in the system model.

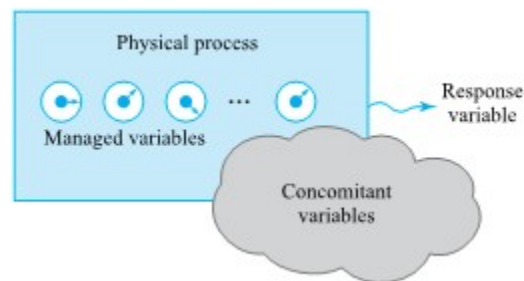
For experimental studies, the input variable is a supervised (or managed) variable in the experiment over which an investigator exercises power, choosing a setting or settings for use in the study. When a supervised variable is held constant (has only one setting), it is called a controlled variable. And when a supervised variable is given several different settings in a study, it is called an experimental variable.

Some of the variables that are neither primary responses nor managed in an experiment will nevertheless be observed.

### DEFINITION 1.1.8.3 Accompanying variable

An accompanying variable (or concomitant variable) in an experiment is one that is identified and included in an analysis but is neither a primary response variable nor an input variable. Such a variable can change in reaction to either input variables or unknown causes and may or may not itself have an impact on a response variable.

Figure 1.1.8.1 is an attempt to picture Definitions 1.1.8.1 through 1.1.8.3. In it, the blackbox physical process somehow produces values of a response in an experiment. “Knobs” on the process represent managed variables. Concomitant variables are floating about as part of the experimental environment without being its main focus.



*Figure 1.1.8.1. Variables in an experiment (“Basic Engineering Data Collection and Analysis” by Stephen B. Vardeman & J. Marcus Jobe which is licensed under CC BY-NC-SA 4.0.)*

Identification of variables that may affect system response requires expert knowledge of the process under study. Engineers who do not have hands-on experience with a system can sometimes contribute insights gained from experience with similar systems and from basic theory. But it is also wise (in most cases, essential) to include on a project team several people who have first-hand knowledge of the particular process and to talk extensively with those who work with the system on a regular basis.

Typically, the job of identifying factors of potential importance in a statistical engineering study is a group activity, carried out in brainstorming sessions. It is therefore helpful to have tools for lending order to what might otherwise be an inefficient and disorganized process. One tool that has proved effective is variously known as a cause-and-effect diagram, or fishbone diagram, or Ishikawa diagram. Figure 1.1.9.2 is a template of a fishbone diagram for a system. In root-cause analysis, the use of 5 (or 8) M’s, is one of the most common frameworks for root-cause analysis (Wikipedia contributors. (2023b, December 3). Ishikawa diagram. Wikipedia. [https://en.wikipedia.org/wiki/Ishikawa\\_diagram](https://en.wikipedia.org/wiki/Ishikawa_diagram)).

Without the time to think through these variables and some kind of organization, it is often difficult to develop anything like a complete list of important factors in a complex or real-world system.

## FISHBONE DIAGRAM

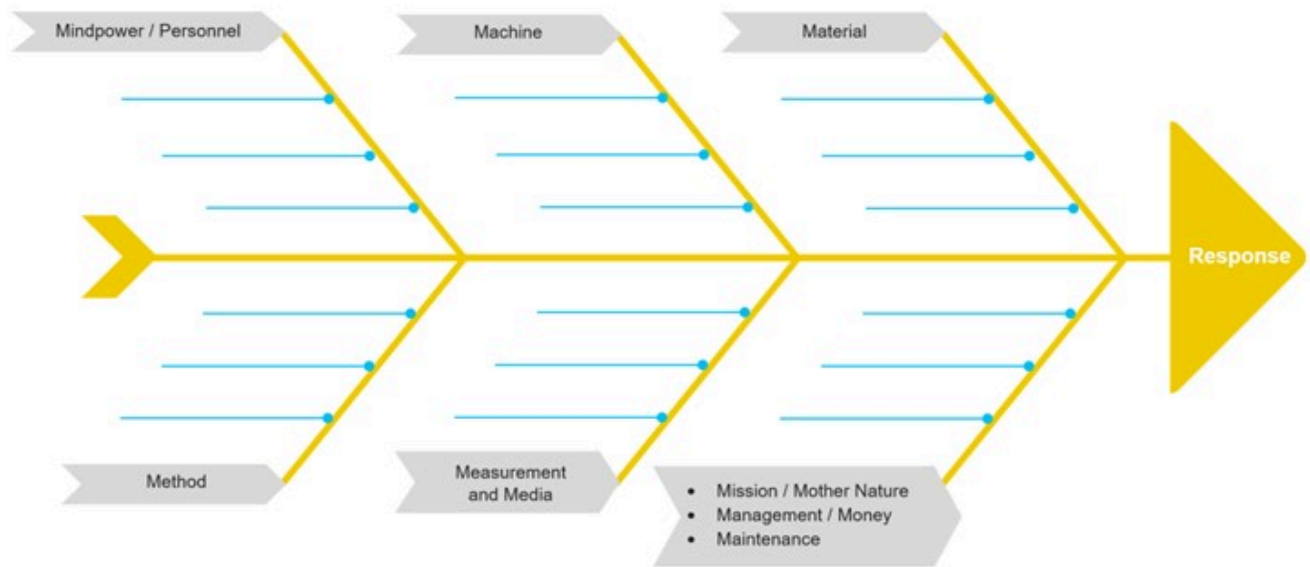


Figure 1.1.8.2. Fishbone diagram of a system.



## 1.1.9 Tutorial 1 - Exploring Data with Python

At this point, it is recommended that you work your way through the [Tutorial 1 exercise](#) found on the associated GitHub repository. This exercise will introduce you to importing data into Python and doing some basic manipulation.

**It is strongly recommended that you consult the [Reading Data into Python & Data Cleaning Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section.

## 2.0.1 Introduction Summarize, Visualize, and Communicate with Data



*NATURE* [MARCH 7, 1907]

The most exact estimate of the average weight of a particular thing or, made by the guesses of a number of persons.

Degree of accuracy of estimate in lbs.	Number of persons	* Results		Error of estimate in lbs.
		Observed	Normal	
100	100	100	100	0
90	100	100	100	0
80	100	100	100	0
70	100	100	100	0
60	100	100	100	0
50	100	100	100	0
40	100	100	100	0
30	100	100	100	0
20	100	100	100	0
10	100	100	100	0
0	100	100	100	0

The results are shown in the table. The observed results are given in the first column, and the normal results in the second. The error of estimate is given in the third column.



Figure 2.0.1.1. Sir Francis Galton, probably taken in the 1850s or early 1860s, image from Wikipedia [https://en.wikipedia.org/wiki/Francis\\_Galton#/media/File:Francis\\_Galton\\_1850s.jpg](https://en.wikipedia.org/wiki/Francis_Galton#/media/File:Francis_Galton_1850s.jpg) and Nature article Vox Populi 1907 image from [https://galton.org/cgi-bin/searchImages/search/essays/pages/galton-1907-vox-populi\\_1.htm](https://galton.org/cgi-bin/searchImages/search/essays/pages/galton-1907-vox-populi_1.htm).

Figure 2.0.1.2. William Playfair, images from Wikipedia [https://en.wikipedia.org/wiki/William\\_Playfair](https://en.wikipedia.org/wiki/William_Playfair).

Francis Galton was a British polymath (1822-1911), and was a pioneer in the use of summary statistics, Figure 2.0.1.1. He was fascinated with measurement and quantification and developed innovative (though deeply problematic) statistical concepts to deal with these. One interesting use of statistics including his insightful observation of the median through an oxen-weight estimation contest. At a livestock fair, Galton observed a competition where participants attempted to guess the weight of an ox. Intrigued by the diverse range of guesses, Galton analyzed the data and found that while individual estimates varied widely, the median of the guesses was surprisingly close to the actual weight of the ox. This discovery highlighted the effectiveness of the median as a measure of central tendency, especially in its robustness to outliers and skewed data, and was published in Nature in 1907.

William Playfair, born in 1786, is regarded as the founder of graphical methods of statistics, including the line, bar, area, and pie charts, Figure 2.0.1.2. He revolutionized the way data was presented and demonstrated that charts could communicate information more effectively than tables of data. After describing and summarizing data using descriptive statistics, data can be described and presented in many different graphical visualizations to present and underscore conclusions about data.

The need for and growth of visualizations of data emphasizes the critical role of statistical graphs as effective tools for understanding the distribution and shape of data. Unlike a mere collection of numbers,

graphs provide a visual representation that makes it easier to discern data clusters, trends, and outliers, a practice widely utilized in various media and industries for quick and efficient data comparison and for communication.

#### Key Takeaways

**Graphs provide a visual representation of data and allow for the communication and story-telling of descriptive statistics.**

We focus on fundamental graphical methods such as histograms, bar plots, box plots, time-series plots, and scatterplots. Practical applications of these concepts are demonstrated through exercises using Python based Jupyter Notebook tutorials. We conclude by emphasizing principles of graphical excellence and the importance of creating informative, truthful, and visually useful graphs.

Overall, this module provides a comprehensive blend of theoretical concepts, practical applications, and statistical computing tools, essential for mastering graphical communication of data in biomedical engineering statistics.

#### Learning Objectives

##### Learning Outcomes for Module 2:

- Learn the descriptive statistical summarizations based on central tendency and spread of data
- Learn to construct and interpret various types of graphs like histograms, bar plots, and box-plots.
- Understand how descriptive statistics summarize and describe the features of a dataset through visualizations.
- Create and interpret an appropriate visualization of data and understand how these graphical techniques are useful in uncovering and summarizing patterns and comparisons in data.
- Understand how to use simple time series plots to visualise the important features of time-directed data.
- Apply the principles of graphical excellence and effective data presentation.

##### Learning Outcomes for Module 2- Jupyter Notebook Tutorials:

- Utilize statistical software for data summarization, visualization, and interpretation.
- Learn to create basic plots using Python's plotting libraries.

## 2.0.2 *Attributions Part 2*

This first draft of Part 2 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

## 2.1.1 Quantitative Data and Quantiles Introduction

Engineering data are always variable. Given precise enough measurement, even supposedly constant process conditions produce differing responses. Therefore, it is not individual data values that demand an engineer's attention as much as the pattern or distribution of those responses. The task of summarizing data is to describe their important distributional characteristics. This chapter discusses simple methods that are helpful in this task.

### **ELEMENTARY GRAPHICAL AND TABULAR TREATMENT OF QUANTITATIVE DATA**

---

Almost always, the place to begin in data analysis is to make appropriate graphical and/or tabular displays. Indeed, where only a few samples are involved, a good picture or table can often tell most of the story about the data. The next few chapters discuss the usefulness of dot diagrams, stem-and-leaf plots, frequency tables, histograms, scatterplots, and run charts.

### **QUANTILES AND RELATED GRAPHICAL TOOLS**

---

After this review of some elementary graphical and tabular methods of data summarization, the concepts of quantiles of a distribution is then introduced and used to make other useful graphical displays.

## 2.1.2 Dot Diagrams and Stem-and-Leaf Plots

When an engineering study produces a small or moderate amount of univariate quantitative data, a dot diagram, easily made with pencil and paper, is often quite revealing. A dot diagram shows each observation as a dot placed at a position corresponding to its numerical value along a number line.

### Example 2.1.2.1. Portraying Thrust Face Runouts

[Module 1.1](#) considered a heat treating problem where distortion for gears laid and gears hung was studied. That figure has been reproduced here as Figure 2.1.2.1. It consists of two dot diagrams, one showing thrust face runout values for gears laid and the other the corresponding values for gears hung, and shows clearly that the laid values are both generally smaller and more consistent than the hung values.

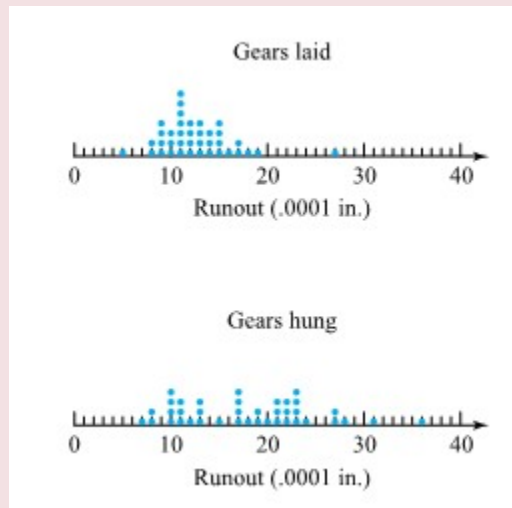


Figure 2.1.2.1. Dot diagrams of runouts.

### Example 2.1.2.2. Penetration of 200 grain bullets

Sale and Thom compared penetration depths for several types of .45 caliber bullets fired into oak wood from a distance of 15 feet. Table 2.1.2.1 gives the penetration depths (in **mm** from the target surface to the back of the bullets) for two bullet types. Figure 2.2.2.2 presents a corresponding pair of dot diagrams.

Bullet Penetration Depths (mm)	
230 Grain Jacketed Bullets	200 Grain Jacketed Bullets
40.50, 38.35, 56.00, 42.55,	63.80, 64.65, 59.50, 60.70,
38.35, 27.75, 49.85, 43.60,	61.30, 61.50, 59.80, 59.10,
38.75, 51.25, 47.90, 48.15,	62.95, 63.55, 58.65, 71.70,
42.90, 43.85, 37.35, 47.30,	63.30, 62.65, 67.75, 62.30,
41.15, 51.60, 39.75, 41.00	70.40, 64.05, 65.00, 58.00

Table 2.1.2.1. Bullet Penetration Depths

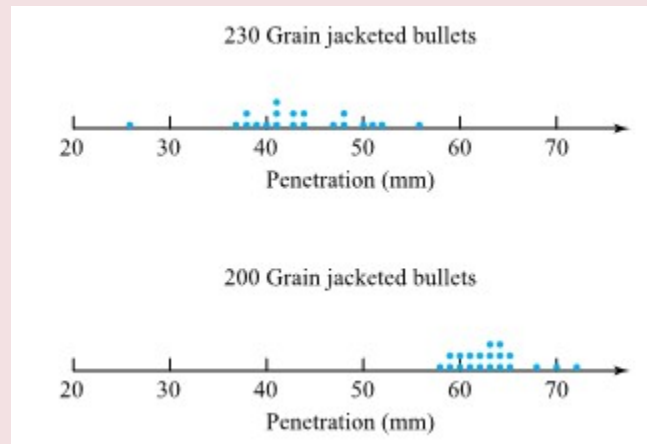


Figure 2.1.2.2. Dot diagrams of penetration depths

The dot diagrams show the penetrations of the 200 grain bullets to be both larger and more consistent than those of the 230 grain bullets. (The students had predicted larger penetrations for the lighter bullets on the basis of greater muzzle velocity and smaller surface area on which friction can act. The different consistencies of penetration were neither expected nor explained.)

Dot diagrams give the general feel of a data set but do not always allow the recovery of exactly the values used to make them. A stem-and-leaf plot carries much the same visual information as a dot diagram while preserving the original values exactly. A stem-and-leaf plot is made by using the last few digits of each data point to indicate where it falls.

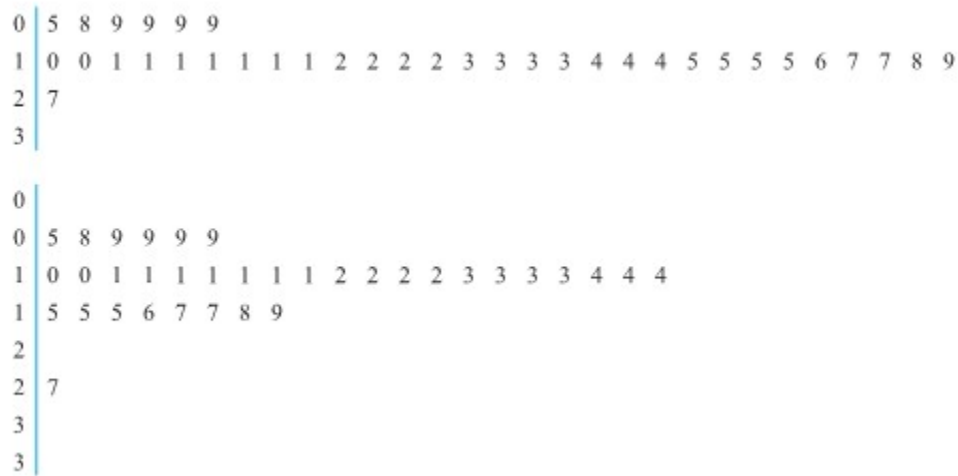


Figure 2.1.2.3. Stem-and-leaf plots of laid gear runouts

#### Example 2.1.2.1 Thrust face runouts of laid gears, continued

Figure 2.1.2.3 gives two possible stem-and-leaf plots for the thrust face runouts of laid gears. In both, the first digit of each observation is represented by the number to the left of the vertical line or “stem” of the diagram. The numbers to the right of the vertical line make up the “leaves” and give the second digits of the observed runouts. The second display shows somewhat more detail than the first by providing “0 — 4” and “5 — 9” leaf positions for each possible leading digit, instead of only a single “0 — 9” leaf for each leading digit.

#### Example 2.1.2.2 Penetration of 200 grain bullets, continued

Figure 2.1.2.4 gives two possible stem-and-leaf plots for the penetrations of 200 grain bullets in Table 2.1.2.1. On these, it was convenient to use two digits to the left of the decimal point to make the stem and the two following the decimal point to create the leaves. The first display was made by recording the leaf values directly from the table (from left to right and top to bottom). The second display is a better one, obtained by ordering the values that make up each leaf. Notice that both plots give essentially the same visual impression as the second dot diagram in Figure 2.2.1.2.





## 2.1.3 Frequency Tables and Histograms

Dot diagrams and stem-and-leaf plots are useful devices when mulling over a data set. But they are not commonly used in presentations and reports. In these more formal contexts, frequency tables and histograms are more often used.

A frequency table is made by first breaking an interval containing all the data into an appropriate number of smaller intervals of equal length. Then tally marks can be recorded to indicate the number of data points falling into each interval. Finally, frequencies, relative frequencies, and cumulative relative frequencies can be added.

### Example 2.1.3.1. Laid Gear Runouts, continued

Table 2.1.3.1 gives one possible frequency table for the laid gear runouts. The relative frequency values are obtained by dividing the entries in the frequency column by 38, the number of data points. The entries in the cumulative relative frequency column are the ratios of the totals in a given class and all preceding classes to the total number of data points. (Except for round-off, this is the sum of the relative frequencies on the same row and above a given cumulative relative frequency.) The tally column gives the same kind of information about distributional shape that is provided by a dot diagram or a stem-and-leaf plot.

Runout (.0001 in.)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
5–8		3	.079	.079
9–12		18	.474	.553
13–16		12	.316	.868
17–20		4	.105	.974
21–24		0	0	.974
25–28		1	.026	1.000
		38	1.000	

Table 2.1.3.1. Frequency Table for Laid Gear Thrust Face Runouts.

### Choosing intervals for a frequency table

The choice of intervals to use in making a frequency table is a matter of judgment. Two people will not necessarily choose the same set of intervals. However, there are a number of simple points to keep in mind when choosing them. First, in order to avoid visual distortion when using the tally column of the table to gain an impression of distributional shape, intervals of equal length should be employed. Also, for

aesthetic reasons, round numbers are preferable as interval endpoints. Since there is usually aggregation (and therefore some loss of information) involved in the reduction of raw data to tallies, the larger the number of intervals used, the more detailed the information portrayed by the table. On the other hand, if a frequency table is to have value as a summarization of data, it can't be cluttered with too many intervals.

After making a frequency table, it is common to use the organization provided by the table to create a histogram. A (frequency or relative frequency) histogram is a kind of bar chart used to portray the shape of a distribution of data points.

**Example 2.1.2.2. Penetration of 200 grain bullets, continued.**

Table 2.1.3.2 is a frequency table for the 200 grain bullet penetration depths, and Figure 2.1.3.1 is a translation of that table into the form of a histogram.

Penetration Depth (mm)	Tally	Frequency	Relative Frequency	Cumulative Relative Frequency
58.00–59.99		5	.25	.25
60.00–61.99		3	.15	.40
62.00–63.99		6	.30	.70
64.00–65.99		3	.15	.85
66.00–67.99		1	.05	.90
68.00–69.99		0	0	.90
70.00–71.99		2	.10	1.00
		20	1.00	

*Table 2.1.3.2. Frequency table for 200 grain penetration depths.*

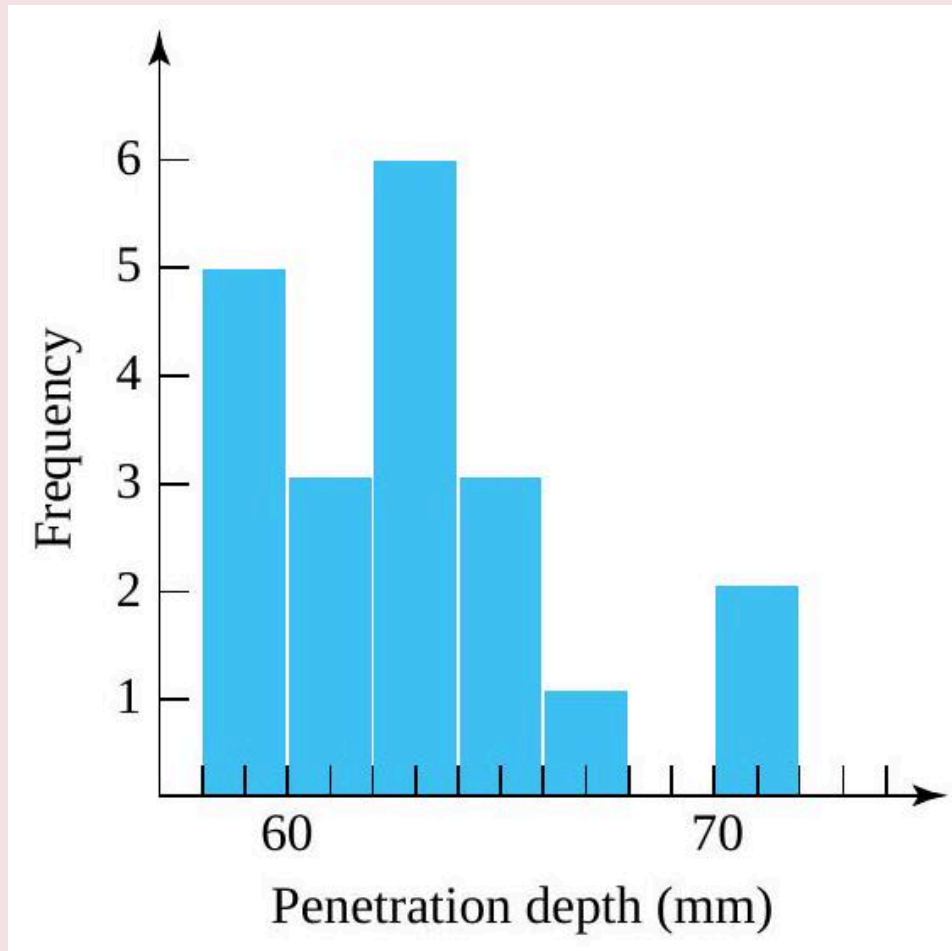


Figure 2.1.3.1. Histogram of the 200 grain penetration depths.

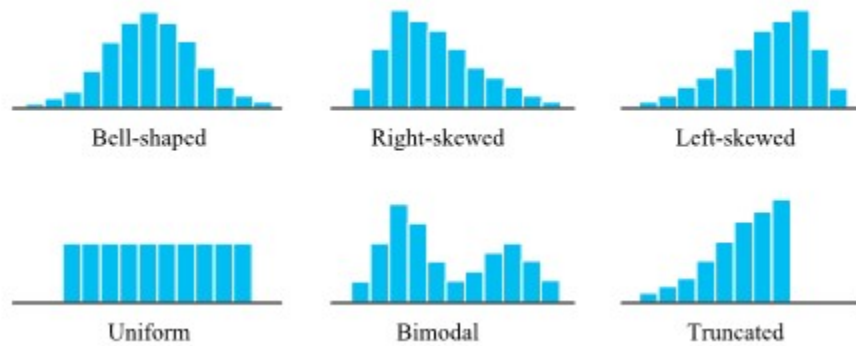
The vertical scale in Figure 2.1.3.1 is a frequency scale, and the histogram is a frequency histogram. By changing to relative frequency on the vertical scale, one can produce a relative frequency histogram.

## GUIDELINES FOR MAKING HISTOGRAMS

In making Figure 2.1.3.1, care was taken to:

1. (continue to) use intervals of equal length,
2. show the entire vertical axis beginning at zero,
3. avoid breaking either axis,
4. keep a uniform scale across a given axis, and
5. center bars of appropriate heights at the midpoints of the (penetration depth) intervals.

Following these guidelines results in a display in which equal enclosed areas correspond to equal numbers of data points. Further, data point positioning is clearly indicated by bar positioning on the horizontal axis. If these guidelines are not followed, the resulting bar chart will in one way or another fail to faithfully represent its data set. Figure 2.1.3.2 shows terminology for common distributional shapes encountered when making and using dot diagrams, stem-and-leaf plots, and histograms.



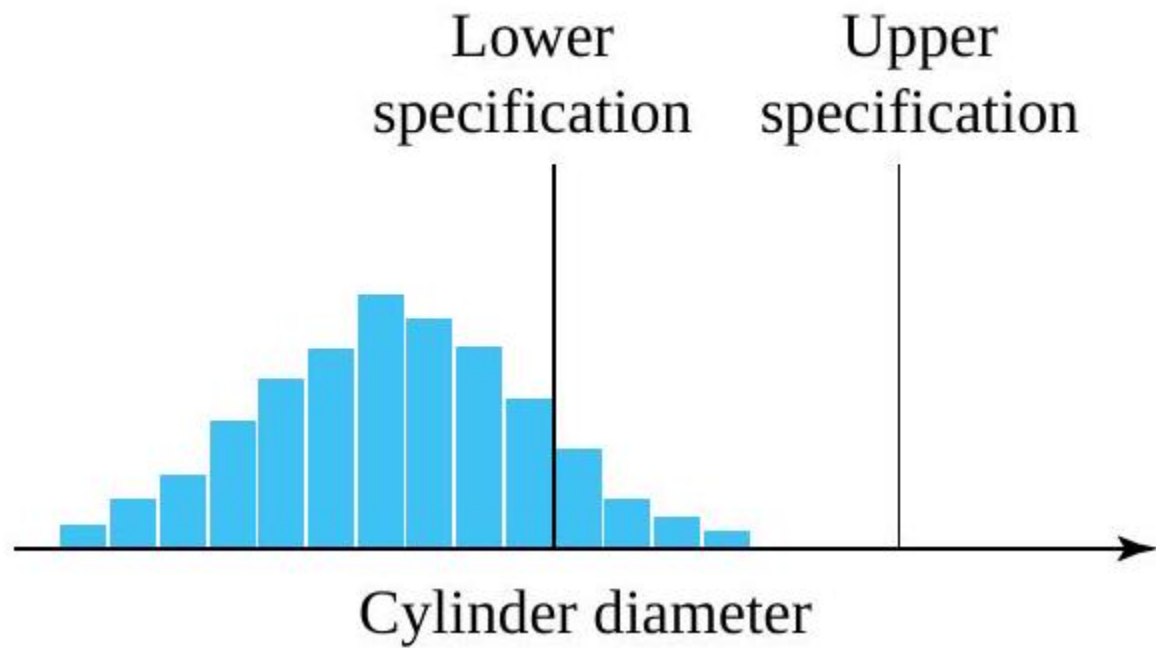
*Figure 2.1.3.2. Distributional Shapes.*

The graphical and tabular devices discussed to this point are deceptively simple methods. When routinely and intelligently used, they are powerful engineering tools. The information on location, spread, and shape that is portrayed so clearly on a histogram can give strong hints as to the functioning of the physical process that is generating the data. It can also help suggest physical mechanisms at work in the process.

#### *Examples of engineering interpretations of distribution shape*

---

For example, if data on the diameters of machined metal cylinders purchased from a vendor produce a histogram that is decidedly bimodal (or multimodal, having several clear humps), this suggests that the machining of the parts was done on more than one machine, or by more than one operator, or at more than one time. The practical consequence of such multichannel machining is a distribution of diameters that has more variation than is typical of a production run of cylinders from a single machine, operator, and setup. As another possibility, if the histogram is truncated, this might suggest that the lot of cylinders has been 100% inspected and sorted, removing all cylinders with excessive diameters. Or, upon marking engineering specifications (requirements) for cylinder diameter on the histogram, one may get a picture like that in Figure 2.1.3.3. It then becomes obvious that the lathe turning the cylinders needs adjustment in order to increase the typical diameter. But it also becomes clear that the basic process variation is so large that this adjustment will fail to bring essentially all diameters into specifications. Armed with this realization and a knowledge of the economic consequences of parts failing to meet specifications, an engineer can intelligently weigh alternative courses of action: sorting of all incoming parts, demanding that the vendor use more precise equipment, seeking a new vendor, etc.



*Figure 2.1.3.3. Histogram marked with engineering specifications.*

Investigating the shape of a data set is useful not only because it can lend insight into physical mechanisms but also because shape can be important when determining the appropriateness of methods of formal statistical inference like those discussed later in this book. A methodology appropriate for one distributional shape may not be appropriate for another.

## 2.1.4 Scatterplots and Run Charts

Dot diagrams, stem-and-leaf plots, frequency tables, and histograms are univariate tools. But engineering data are often multivariate and relationships between the variables are then usually of interest. The familiar device of making a two-dimensional scatterplot of data pairs is a simple and effective way of displaying potential relationships between two variables.

### Example 2.1.4.1. Bolt Torques on a Face Plate

Brenny, Christensen, and Schneider measured the torques required to loosen six distinguishable bolts holding the front plate on a type of heavy equipment component. Table 2.1.4.1 contains the torques (in **ftlb**) required for bolts number 3 and 4), respectively, on 34 different components. Figure 2.1.4.1 is a scatterplot of the bivariate data from Table 2.1.4.1. In this figure, where several points must be plotted at a single location, the number of points occupying the location has been plotted instead of a single dot.

Torques Required to Loosen Two Bolts on Face Plates (ft lb)					
Component	Bolt 3 Torque	Bolt 4 Torque	Component	Bolt 3 Torque	Bolt 4 Torque
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

Table 2.1.4.1. Torques Required to Loosen Two Bolts on Face Plates (ft lb)

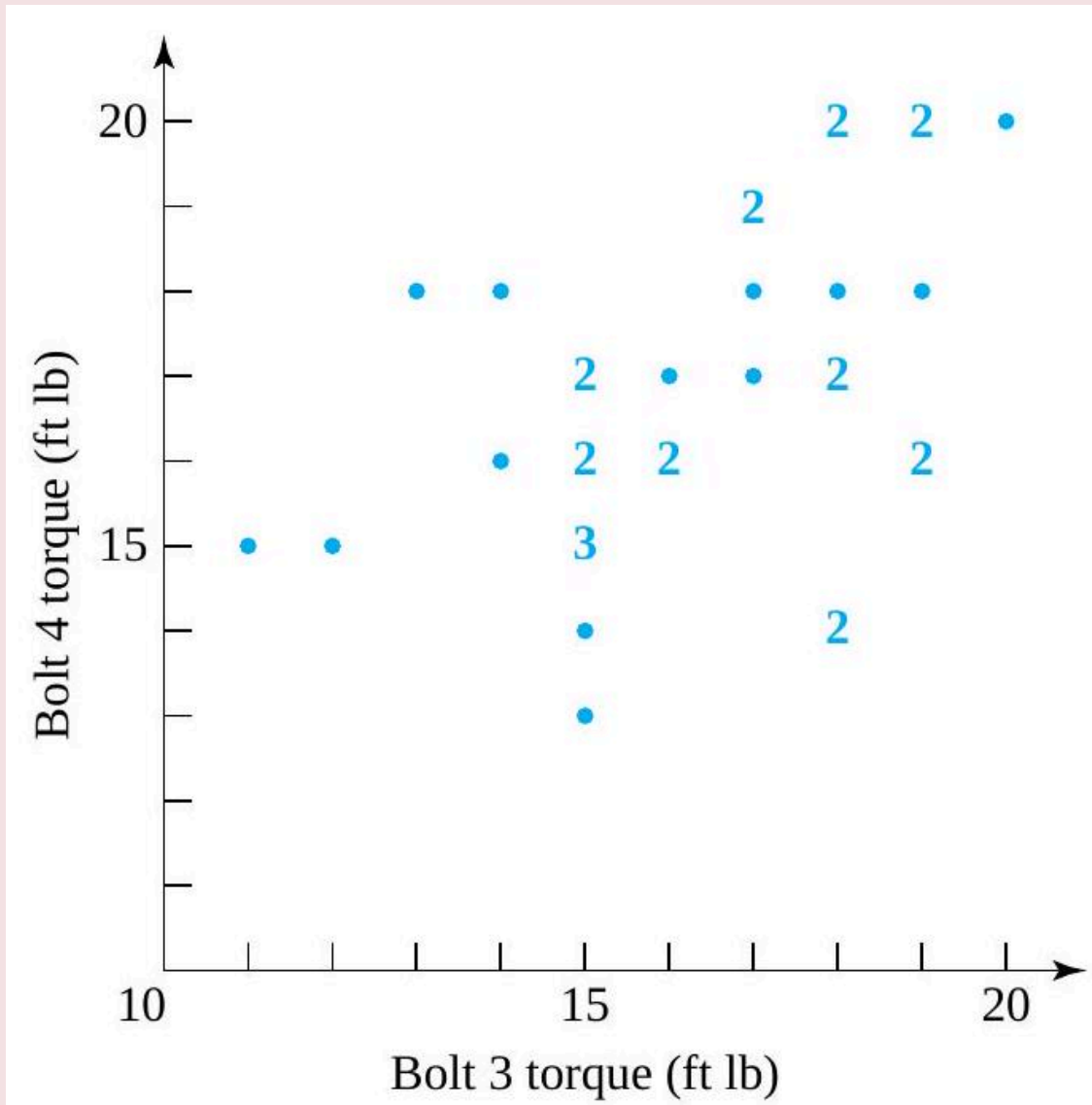


Figure 2.1.4.1. Scatterplot of bolt 3 and bolt 4 torques.

The plot gives at least a weak indication that large torques at position 3 are accompanied by large torques at position 4. In practical terms, this is comforting; otherwise, unwanted differential forces might act on the face plate. It is also quite reasonable that bolt 3 and bolt 4 torques be related, since the bolts were tightened by different heads of a single pneumatic wrench operating off a single source of compressed air. It stands to reason that variations in air pressure might affect the tightening of the bolts at the two positions similarly, producing the big-together, small-together pattern seen in Figure 2.1.4.1.

The previous example illustrates the point that relationships seen on scatterplots suggest a common physical cause for the behavior of variables and can help reveal that cause.

## RUN CHART

In the most common version of the scatterplot, the variable on the horizontal axis is a time variable. A scatterplot in which univariate data are plotted against time order of observation is called a run chart or



trend chart. Making run charts is one of the most helpful statistical habits an engineer can develop. Seeing patterns on a run chart leads to thinking about what process variables were changing in concert with the pattern. This can help develop a keener understanding of how process behavior is affected by those variables that change over time.

#### Example 2.1.4.2. Diameters of Consecutive Parts Turned on a Lathe

Williams and Markowski studied a process for rough turning of the outer diameter on the outer race of a constant velocity joint. Table 2.1.4.2 gives the diameters (in inches above nominal) for 30 consecutive joints turned on a particular automatic lathe. Figure 2.1.4.2 gives both a dot diagram and a run chart for the data in the table. In keeping with standard practice, consecutive points on the run chart have been connected with line segments.

Joint	Diameter (inches above nominal)	Joint	Diameter (inches above nominal)
1	-.005	16	.015
2	.000	17	.000
3	-.010	18	.000
4	-.030	19	-.015
5	-.010	20	-.015
6	-.025	21	-.005
7	-.030	22	-.015
8	-.035	23	-.015
9	-.025	24	-.010
10	-.025	25	-.015
11	-.025	26	-.035
12	-.035	27	-.025
13	-.040	28	-.020
14	-.035	29	-.025
15	-.035	30	-.015

Table 2.1.4.2. 30 Consecutive Outer Diameters Turned on a Lathe

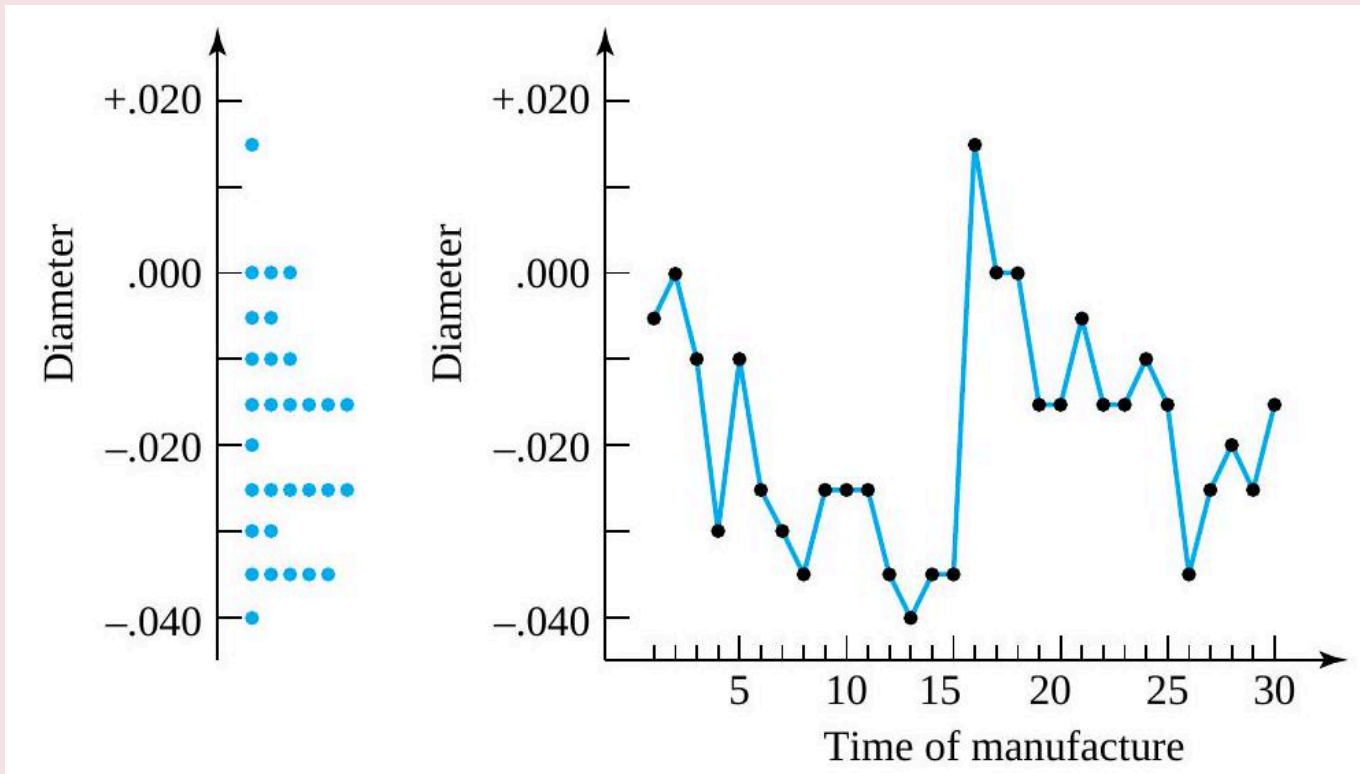


Figure 2.1.4.2. Dot diagram and run chart of consecutive outer diameters.

Here the dot diagram is not particularly suggestive of the physical mechanisms that generated the data. But the time information added in the run chart is revealing. Moving along in time, the outer diameters tend to get smaller until part 16, where there is a large jump, followed again by a pattern of diameter generally decreasing in time. In fact, upon checking production records, Williams and Markowski found that the lathe had been turned off and allowed to cool down between parts 15 and 16. The pattern seen on the run chart is likely related to the behavior of the lathe's hydraulics. When cold, the hydraulics probably don't do as good a job pushing the cutting tool into the part being turned as when they are warm. Hence, the turned parts become smaller as the lathe warms up. In order to get parts closer to nominal, the aimed-for diameter might be adjusted up by about .020 in. and parts run only after warming up the lathe.

## 2.1.5 Quantiles and Quantile Plots

Most readers will be familiar with the concept of a percentile. The notion is most famous in the context of reporting scores on educational achievement tests. For example, if a person has scored at the 80th percentile, roughly 80% of those taking the test had worse scores, and roughly 20% had better scores. This concept is also useful in the description of engineering data. However, because it is often more convenient to work in terms of fractions between 0 and 1 rather than in percentages between 0 and 100, slightly different terminology will be used here: “Quantiles,” rather than percentiles, will be discussed. After the quantiles of a data set are carefully defined, they are used to create a number of useful tools of descriptive statistics: quantile plots, boxplots,  $Q - Q$  plots, and normal plots (a type of theoretical  $Q - Q$  plot).

Roughly speaking, for a number  $p$  between 0 and 1, the  $p$  quantile of a distribution is a number such that a fraction  $p$  of the distribution lies to the left and a fraction  $1 - p$  of the distribution lies to the right. However, because of the discreteness of finite data sets, it is necessary to state exactly what will be meant by the terminology. Definition 1 gives the precise convention that will be used in this text.

### Definition 3.1.5.1 $p$ quantile

For a data set consisting of  $n$  values that when ordered are  $x_1 \leq x_2 \leq \dots \leq x_n$ ,

1. if  $p = \frac{i - .5}{n}$  for a positive integer  $i \leq n$ , the  $p$  quantile of the data set is

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

(The  $i$ th smallest data point will be called the  $\frac{i - .5}{n}$  quantile.)

2. for any number  $p$  between  $\frac{i - .5}{n}$  and  $\frac{i + .5}{n}$  that is not of the form  $\frac{i - .5}{n}$  for an integer  $i$ , the  $p$  quantile of the data set will be obtained by linear interpolation between the two values of  $Q\left(\frac{i - .5}{n}\right)$  with corresponding  $\frac{i - .5}{n}$  that bracket  $p$ .

In both cases, the notation  $Q(p)$  will be used to denote the  $p$  quantile.

Definition 2.1.5.1 identifies  $Q(p)$  for all  $p$  between  $.5/n$  and  $(n - .5)/n$ . To find  $Q(p)$  for such a value of  $p$ , one may solve the equation  $p = (i - .5)/n$  for  $i$ , yielding

$$\text{Index (i) of the ordered data point that is } Q(p) \\ i = np + .5$$

and locate the " $(np + .5)^{\text{th}}$  ordered data point."

#### Example 2.1.5.1. Quantiles for Dry Breaking Strengths of Paper Towel

Lee, Sebghati, and Straub did a study of the dry breaking strength of several brands of paper towel. Table 3.1.5.1 shows ten breaking strengths (in grams) reported by the students for a generic towel. By ordering the strength data and computing values of  $\frac{i - .5}{10}$ , one can easily find the .05, .15, .25, . . . , .85, and .95 quantiles of the breaking strength distribution, as shown in Table 31.5.2.

Test	Breaking Strength (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

Table 2.1.5.1.

Quantiles of the Paper Towel Breaking Strength Distribution

$i$	$\frac{i-5}{10}$	$i$ th Smallest Data Point, $x_i = Q\left(\frac{i-5}{10}\right)$
1	.05	7,583 = $Q(.05)$
2	.15	8,527 = $Q(.15)$
3	.25	8,572 = $Q(.25)$
4	.35	8,577 = $Q(.35)$
5	.45	9,011 = $Q(.45)$
6	.55	9,165 = $Q(.55)$
7	.65	9,471 = $Q(.65)$
8	.75	9,614 = $Q(.75)$
9	.85	9,614 = $Q(.85)$
10	.95	10,688 = $Q(.95)$

Table 2.1.5.2.

Since there are  $n = 10$  data points, each one accounts for  $10\%$  of the data set. Applying convention (1) in Definition 3.1.5.1 to find (for example) the .35 quantile, the smallest 3 data points and half of the fourth smallest are counted as lying to (continued) the left of the desired number, and the largest 6 data points and half of the seventh largest are counted as lying to the right. Thus, the fourth smallest data point must be the .35 quantile, as is shown in Table 2.1.5.2.

To illustrate convention (2) of Definition 1, consider finding the .5 and .93 quantiles of the strength distribution. Since .5 is  $\frac{.5 - .45}{.55 - .45} = .5$  of the way from .45 to .55, linear interpolation gives:

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

Then, observing that .93 is  $\frac{.93 - .85}{.95 - .85} = .8$  of the way from .85 to .95, linear interpolation gives:

$$Q(.93) = (1 - .8)Q(.85) + .8Q(.95) = .2(9,614) + .8(10,688) = 10,473.2 \text{ g}$$

Particular round values of  $p$  give quantiles  $Q(p)$  that are known by special names.

#### DEFINITION 2.1.5.2 Median

Definition 2  $Q(.5)$  is called the median of a distribution.

**DEFINITION 2.1.5.3 First (or lower) quartile and third (or upper) quartile**

Definition 3  $Q(.25)$  and  $Q(.75)$  are called the first (or lower) quartile and third (or upper) quartile of a distribution, respectively.

**Example 2.1.5.1 Dry Breaking Strengths of Paper Towel, continued**

Referring again to Table 2.1.5.2 and the value of  $Q(.5)$  previously computed, for the (continued) breaking strength distribution

$$\begin{aligned}\text{Median} &= Q(.5) = 9,088 \text{ g} \\ \text{1st quartile} &= Q(.25) = 8,572 \text{ g} \\ \text{3rd quartile} &= Q(.75) = 9,614 \text{ g}\end{aligned}$$

A way of representing the quantile idea graphically is to make a quantile plot.

**DEFINITION 2.1.5.4 Quantile Plot**

A quantile plot is a plot of  $Q(p)$  versus  $p$ . For an ordered data set of size  $n$  containing values  $x_1 \leq x_2 \leq \cdots \leq x_n$ , such a display is made by first plotting the points  $\left(\frac{i-.5}{n}, x_i\right)$  and then connecting consecutive plotted points with straight-line segments.

It is because convention (2) in Definition 2.1.5.1 calls for linear interpolation that straightline segments enter the picture in making a quantile plot.

**Example 2.1.5.1. Dry Breaking Strengths of Paper Towel, continued**

Referring again to Table 2.1.5.2 for the quantiles of the breaking strength distribution, it is clear that a quantile plot for these data will involve plotting and then connecting consecutive ones of the following ordered pairs.

(.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)		

Figure 2.1.5.1 gives such a plot.

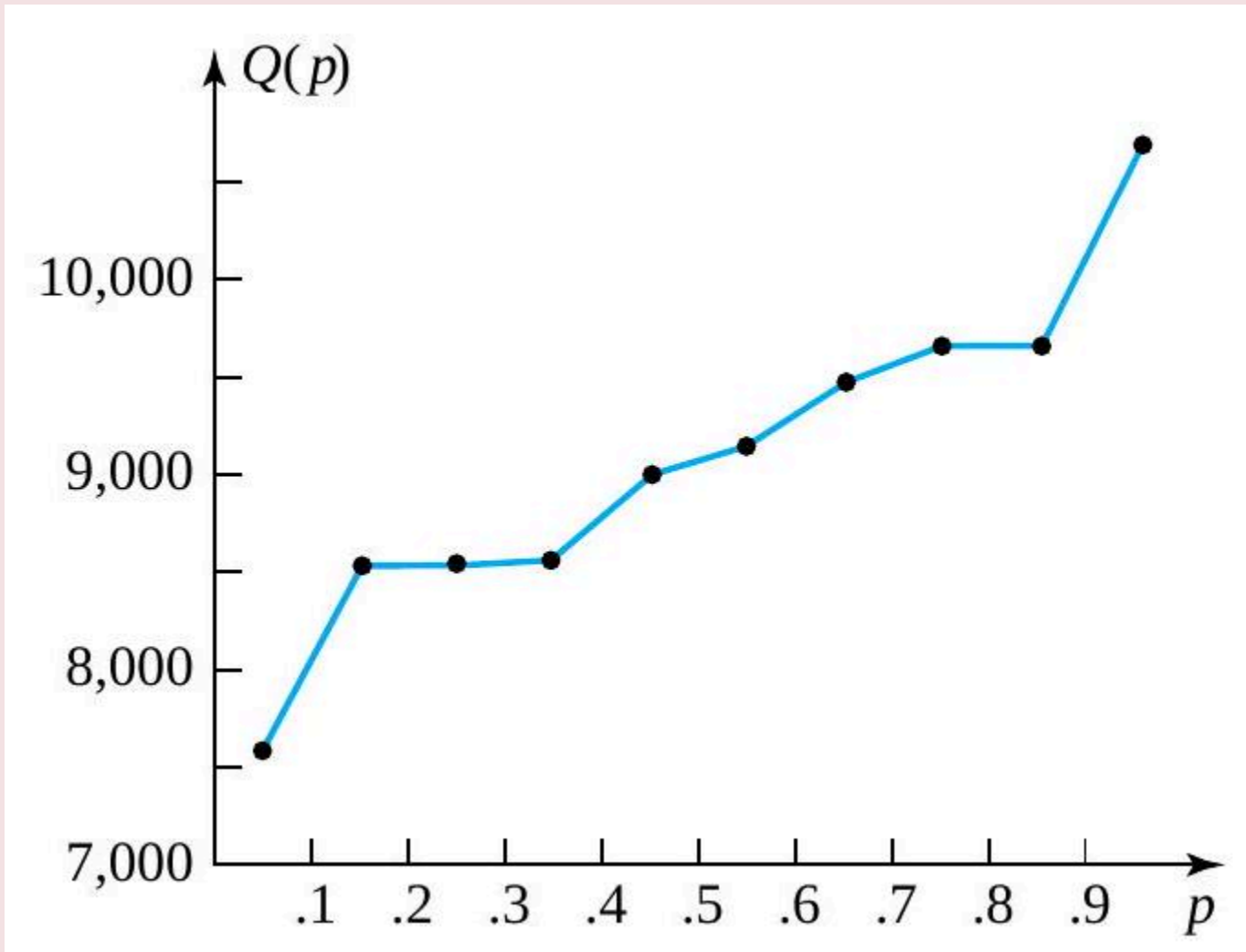


Figure 2.1.5.1. Quantile plot of papertowel strengths.

A quantile plot allows the user to do some informal visual smoothing of the plot to compensate for any jaggedness. (The tacit assumption is that the underlying datagenerating mechanism would itself produce smoother and smoother quantile plots for larger and larger samples.)

## 2.1.6 Boxplots

Familiarity with the quantile idea is the principal prerequisite for making boxplots, an alternative to dot diagrams or histograms. The boxplot carries somewhat less information, but it has the advantage that many can be placed side-by-side on a single page for comparison purposes.

There are several common conventions for making boxplots. The one that will be used here is illustrated in generic fashion in Figure 2.1.6.1. A box is made to extend from the first to the third quartiles and is divided by a line at the median. Then the interquartile range:

**DEFINITION 2.1.6.1. Interquartile Range: IQR**

$$IQR = Q(.75) - Q(.25)$$

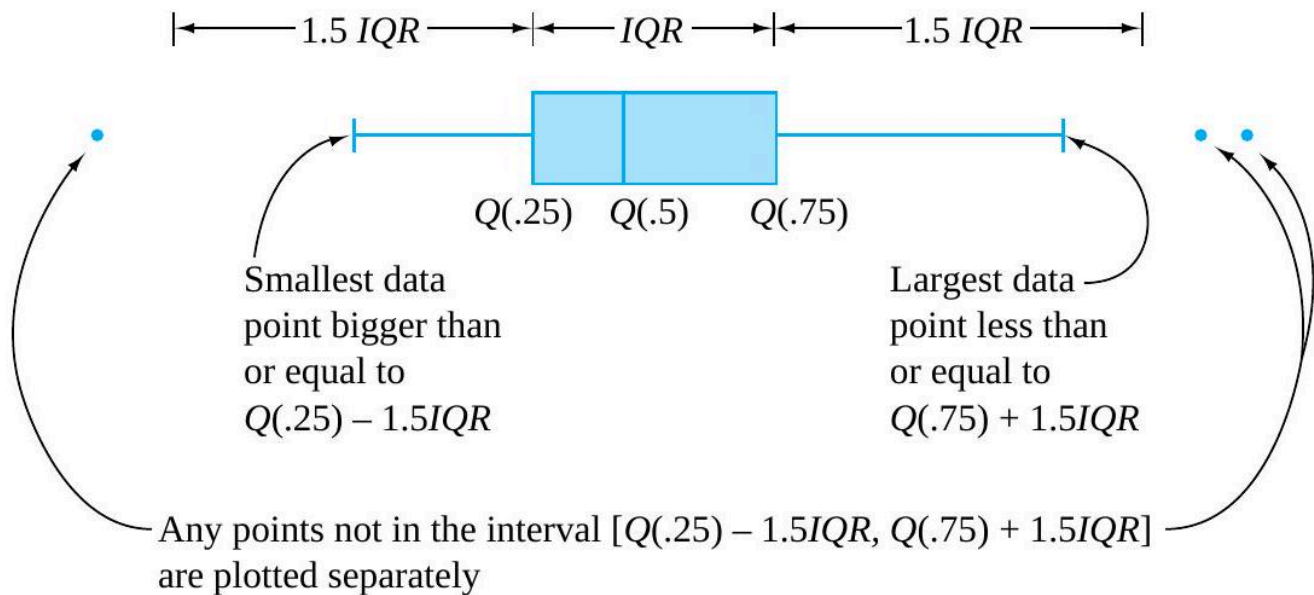


Figure 2.1.6.1. Generic Boxplot.

is calculated and the smallest data point within  $1.5IQR$  of  $Q(.25)$  and the largest data point within  $1.5IQR$  of  $Q(.75)$  are determined. Lines called whiskers are made to extend out from the box to these values.



Typically, most data points will be within the interval  $[Q(.25) - 1.5IQR, Q(.75) + 1.5IQR]$ . Any that are not then get plotted individually and are thereby identified as outlying or unusual.

#### Example 2.1.6.2. Dry Breaking Strengths of Paper Towel, continued

Consider making a boxplot for the paper towel breaking strength data. To begin,

$$Q(.25) = 8,572 \text{ g}$$

$$Q(.5) = 9,088 \text{ g}$$

$$Q(.75) = 9,614 \text{ g}$$

So

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

and

$$1.5IQR = 1,563 \text{ g}$$

Then

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

and

$$Q(.25) - 1.5IQR = 8,572 - 1,563 = 7,009 \text{ g}$$

Since all the data points lie in the range 7,009 g to 11,177 g, the boxplot is as shown in Figure 2.1.6.2.

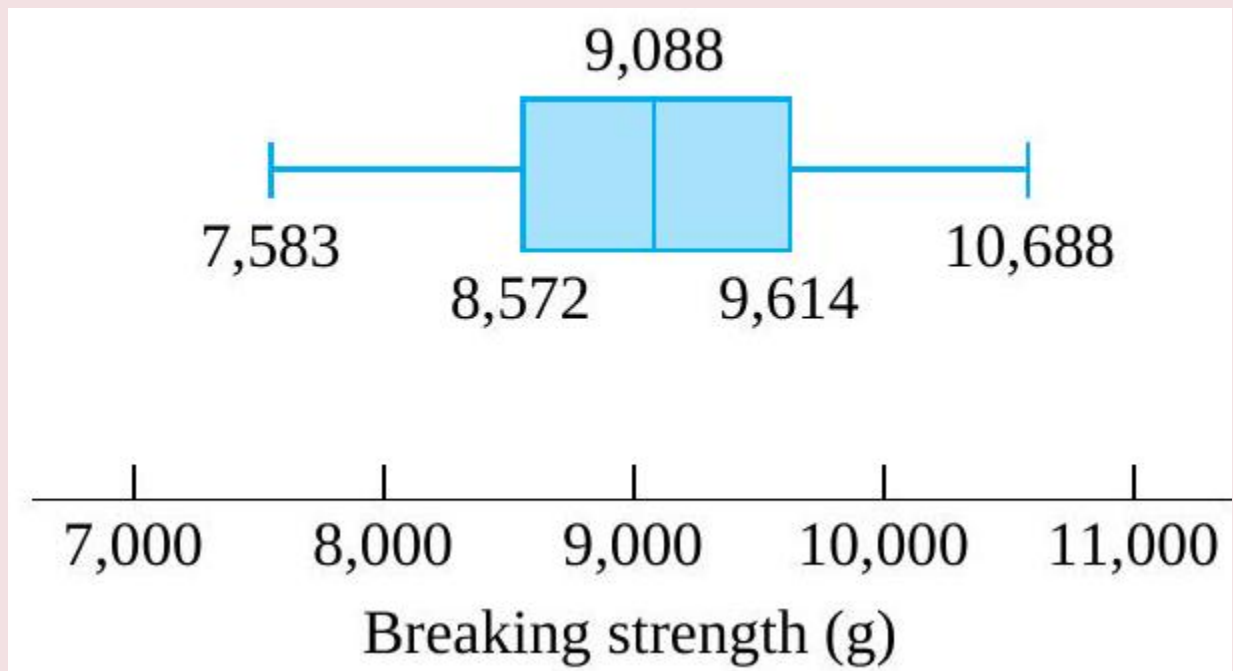


Figure 2.1.6.2. Boxplot of paper towel strengths.

A boxplot shows distributional location through the placement of the box and whiskers along a number line. It shows distributional spread through the extent of the box and the whiskers, with the box enclosing the middle 50% of the distribution. Some elements of distributional shape are indicated by the symmetry (or lack thereof) of the box and of the whiskers. And a gap between the end of a whisker and a separately plotted point serves as a reminder that no data values fall in that interval.

Two or more boxplots drawn to the same scale and side by side provide an effective way of comparing

samples.

### Example 2.1.6.3. Bullet penetraion depth, continued

Table 2.1.6.1 contains the raw information needed to find the  $\frac{i - .5}{20}$  quantiles for the two distributions of bullet penetration depth

introduced in the previous section. For the 230 grain bullet penetration depths, interpolation yields

$$Q(.25) = .5Q(.225) + .5Q(.275) = .5(38.75) + .5(39.75) = 39.25 \text{ mm}$$

$$Q(.5) = .5Q(.475) + .5Q(.525) = .5(42.55) + .5(42.90) = 42.725 \text{ mm}$$

$$Q(.75) = .5Q(.725) + .5Q(.775) = .5(47.90) + .5(48.15) = 48.025 \text{ mm}$$

So

$$IQR = 48.025 - 39.25 = 8.775 \text{ mm}$$

$$1.5IQR = 13.163 \text{ mm}$$

$$Q(.75) + 1.5IQR = 61.188 \text{ mm}$$

$$Q(.25) - 1.5IQR = 26.087 \text{ mm}$$

Similar calculations for the 200 grain bullet penetration depths yield

$$Q(.25) = 60.25 \text{ mm}$$

$$Q(.5) = 62.80 \text{ mm}$$

$$Q(.75) = 64.35 \text{ mm}$$

$$Q(.75) + 1.5IQR = 70.50 \text{ mm}$$

$$Q(.25) - 1.5IQR = 54.10 \text{ mm}$$

$i$	$\frac{i-5}{20}$	$i$ th Smallest 230 Grain Data Point = $Q(\frac{i-5}{20})$	$i$ th Smallest 200 Grain Data Point = $Q(\frac{i-5}{20})$
1	.025	27.75	58.00
2	.075	37.35	58.65
3	.125	38.35	59.10
4	.175	38.35	59.50
5	.225	38.75	59.80
6	.275	39.75	60.70
7	.325	40.50	61.30
8	.375	41.00	61.50
9	.425	41.15	62.30
10	.475	42.55	62.65
11	.525	42.90	62.95
12	.575	43.60	63.30
13	.625	43.85	63.55
14	.675	47.30	63.80
15	.725	47.90	64.05
16	.775	48.15	64.65
17	.825	49.85	65.00
18	.875	51.25	67.75
19	.925	51.60	70.40
20	.975	56.00	71.70

Table 2.1.6.1.

Figure 2.1.6.3 then shows boxplots placed side by side on the same scale. The plots show the larger and more consistent penetration depths of the 200 grain bullets. They also show the existence of one particularly extreme data point in the 200 grain data set. Further, the relative lengths of the whiskers hint at some skewness (recall the terminology introduced previously to discuss distributional shape) in the data. And all of this is done in a way that is quite uncluttered and compact. Many more of these boxes could be added to Figure 2.1.6.3 (to compare other bullet types) without visual overload.

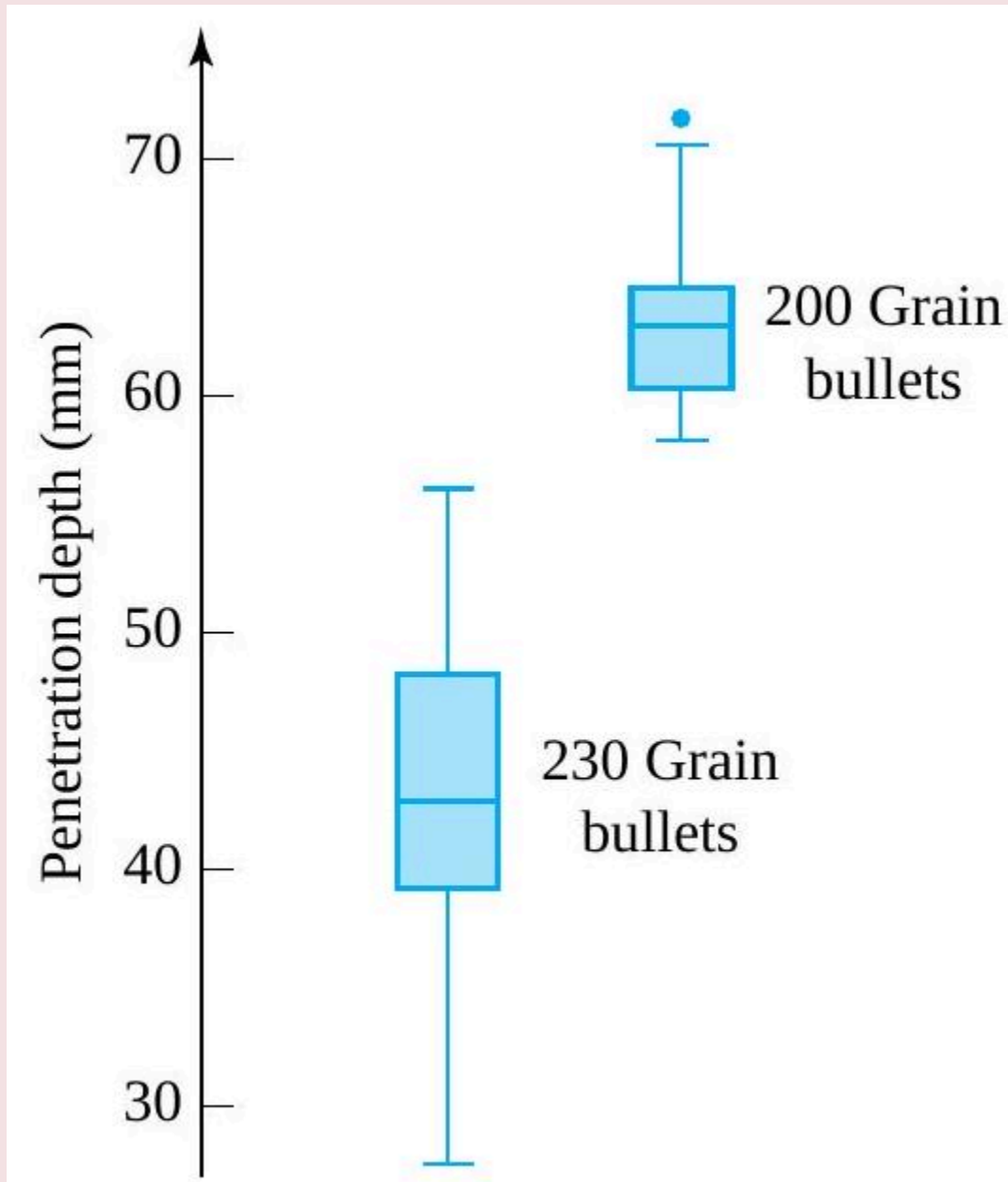


Figure 2.1.6.3. Side-by-side boxplots for the bullet penetration depths

## 2.1.7 Q-Q Plots and Comparing Distributional Shapes

It is often important to compare the shapes of two distributions. Comparing histograms is one rough way of doing this. A more sensitive way is to make a single plot based on the quantile functions for the two distributions and exploit the fact that “equal shape” is equivalent to “linearly related quantile functions.” Such a plot is called a **quantile-quantile plot** or, more briefly, a **Q – Q plot**.

Consider the two small artificial data sets given in Table 2.1.7.1. Dot diagrams of these two data sets are given in Figure 2.1.71.. The two data sets have the same shape. But why is this so? One way to look at the equality of the shapes is to note that

**2.1.7.1**

$$i \text{ th smallest value in data set 2} = 2(i \text{ th smallest value in data set 1}) + 1$$

Then, recognizing ordered data values as quantiles and letting  $Q_1$  and  $Q_2$  stand for the quantile functions of the two respective data sets, it is clear from display (2.1.7.1) that

**2.1.7.2**

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

Two Small Artificial Data Sets

Data Set 1	Data Set 2
3, 5, 4, 7, 3	15, 7, 9, 7, 11

Table 2.1.7.1.

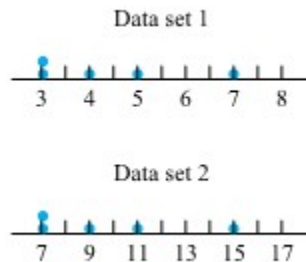


Figure 2.1.7.1 Dot diagrams for two small data sets.

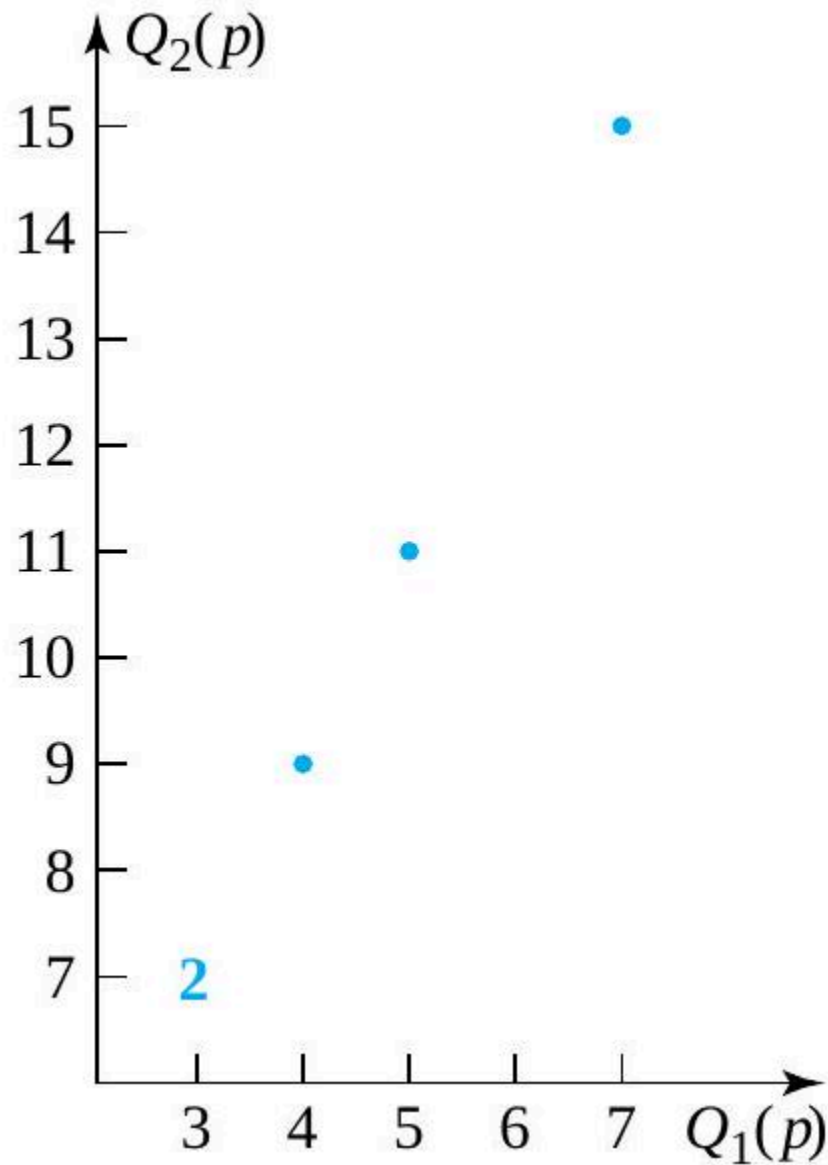


Figure 2.1.7.2. Q-Q plot for the data of Table 2.1.7.1.

That is, the two data sets have quantile functions that are linearly related. Looking at either display (2.1.7.1) or (2.1.7.2), it is obvious that a plot of the points:

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

(for  $i = 1, 2, 3, 4, 5$ ) should be exactly linear. Figure 3.16 illustrates this—in fact Figure 3.16 is a  $Q - Q$  plot for the data sets of Table 2.1.7.1.

**DEFINITION 2.1.7.1. A  $Q - Q$  plot.**

A  $Q - Q$  plot for two data sets with respective quantile functions  $Q_1$  and  $Q_2$  is a plot of ordered pairs  $(Q_1(p), Q_2(p))$  for appropriate values of  $p$ . When two data sets of size  $n$  are involved, the values of  $p$  used to make the plot will be  $\frac{i - .5}{n}$  for  $i = 1, 2, \dots, n$ . When two data sets of unequal sizes are involved, the values of  $p$  used to make the plot will be  $\frac{i - .5}{n}$  for  $i = 1, 2, \dots, n$ , where  $n$  is the size of the smaller set.

**STEPS IN MAKING A Q-Q PLOT**

To make a  $Q - Q$  plot for two data sets of the same size:

1. order each from the smallest observation to the largest,
2. pair off corresponding values in the two data sets, and
3. plot ordered pairs, with the horizontal coordinates coming from the first data set and the vertical ones from the second.

When data sets of unequal size are involved, the ordered values from the smaller data set must be paired with quantiles of the larger data set obtained by interpolation.

A  $Q - Q$  plot that is reasonably linear indicates the two distributions involved have similar shapes. When there are significant departures from linearity, the character of those departures reveals the ways in which the shapes differ.

**Example 2.1.7.1. Bullet penetration, continued.**

Returning again to the bullet penetration depths, the table previously gave the raw material for making a  $Q - Q$  plot. The depths on each row of that table need only be paired and plotted in order to make the plot given in Figure 2.1.7.3.

The scatterplot in Figure 2.1.7.3 is not terribly linear when looked at as a whole. However, the points corresponding to the 2nd through 13th smallest values in each data set do look fairly linear, indicating that (except for the extreme lower ends) the lower ends of the two distributions have similar shapes.

The horizontal jog the plot takes between the 13th and 14th plotted points indicates that the gap between **43.85 mm** and **47.30 mm** (for the 230 grain data) is out of proportion to the gap between 63.55 and **63.80 mm** (for the 200 grain data). This hints that there was some kind of basic physical difference in the mechanisms that produced the smaller and larger 230 grain penetration depths. Once this kind of indication is discovered, it is a task for ballistics experts or materials people to explain the phenomenon.

Because of the marked departure from linearity produced by the 1st plotted point (**27.75, 58.00**), there is also a drastic difference in the shapes of the extreme lower ends of the two distributions. In order to move that point back on line with the rest of the plotted points, it would need to be moved to the right or down (i.e., increase the smallest 230 grain observation or decrease the smallest 200 grain observation). That is, relative to the 200 grain distribution, the 230 grain distribution is long-tailed to the low side. (Or to put it differently, relative to the 230 grain distribution, the 200 grain distribution is short-tailed to the low side.) Note that the difference in shapes was already evident in the boxplot in Figure previously. Again, it would remain for a specialist to explain this difference in

distributional shapes.

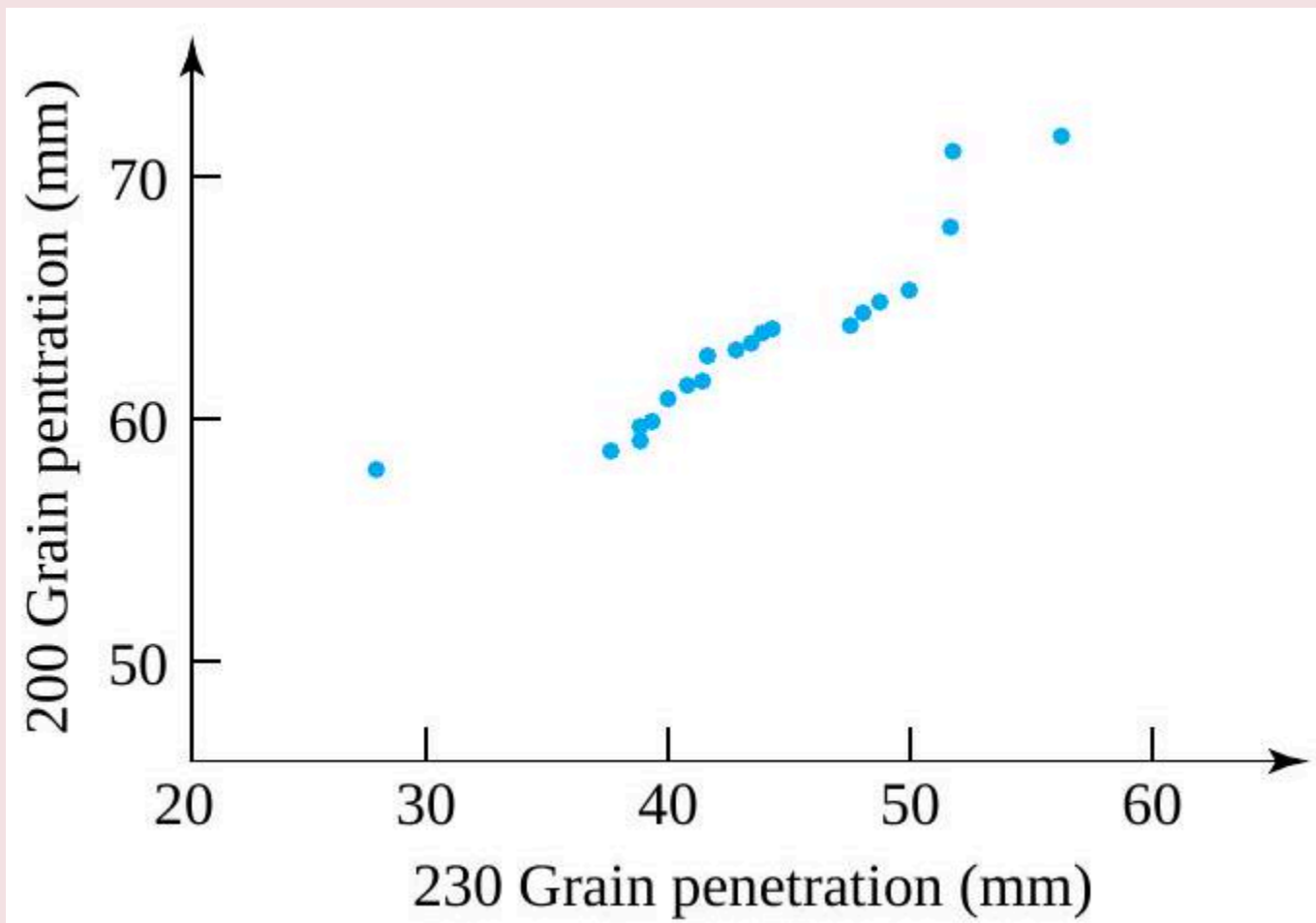


Figure 2.1.7.3. Q-Q plot for the bullet penetration depths

The Q-Q plotting idea is useful when applied to two data sets, and it is easiest to explain the notion in such an “empirical versus empirical” context. But its greatest usefulness is really when it is applied to one quantile function that represents a data set and a second that represents a *theoretical distribution*.

#### DEFINITION 2.1.7.2 A theoretical Q-Q plot

A theoretical Q-Q plot or probability plot for a data set of size  $n$  and a theoretical distribution, with respective quantile functions  $Q_1$  and  $Q_2$ , is a plot of ordered pairs  $(Q_1(p), Q_2(p))$  for appropriate values of  $p$ . In this text, the values of  $p$  of the form  $\frac{i - .5}{n}$  for  $i = 1, 2, \dots, n$  will be used.



Recognizing  $Q\left(\frac{i - .5}{n}\right)$  as the  $i$ th smallest data point, one sees that a theoretical

$Q$ - $Q$  plot is a plot of points with horizontal plotting positions equal to observed data and vertical plotting positions equal to quantiles of the theoretical distribution. That is, with ordered data  $x_1 \leq x_2 \leq \dots \leq x_n$ , the points

### 2.1.7.3 Ordered pairs making a probability plot

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

are plotted. Such a plot allows one to ask, "Does the data set have a shape similar to the theoretical distribution?"

## NORMAL PLOTTING

The most famous version of the theoretical  $Q - Q$  plot occurs when quantiles for the standard normal or Gaussian distribution are employed. This is the familiar bell-shaped distribution. Table 3.10 gives some quantiles of this distribution. In order to find  $Q(p)$  for  $p$  equal to one of the values .01, .02, . . . , .98, .99, locate the entry in the row labelled by the first digit after the decimal place and in the column labelled by the second digit after the decimal place. (For example,  $Q(.37) = -.33$ .) A simple numerical approximation to the values given in Table 3.10 adequate for most plotting purposes is

### 2.1.7.3 Approximate standard normal quantiles

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

The origin of Table 2.1.7.2 is not obvious at this point. It will be explained in Part 4, but for the time being consider the following crude argument to the effect that the quantiles in the table correspond to a bell-shaped distribution. Imagine that each entry in Table 2.1.7.2 corresponds to a data point in a set of size  $n = 99$ . A possible frequency table for those 99 data points is given as Table 2.1.7.3. The tally column in Table 2.1.7.3 shows clearly the bell shape.

The standard normal quantiles can be used to make a theoretical  $Q - Q$  plot as a way of assessing how bell-shaped a data set looks. The resulting plot is called a **normal (probability) plot**.

Standard Normal Quantiles

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0		-2.33	-2.05	-1.88	-1.75	-1.65	-1.55	-1.48	-1.41	-1.34
.1	-1.28	-1.23	-1.18	-1.13	-1.08	-1.04	-.99	-.95	-.92	-.88
.2	-.84	-.81	-.77	-.74	-.71	-.67	-.64	-.61	-.58	-.55
.3	-.52	-.50	-.47	-.44	-.41	-.39	-.36	-.33	-.31	-.28
.4	-.25	-.23	-.20	-.18	-.15	-.13	-.10	-.08	-.05	-.03
.5	0.00	.03	.05	.08	.10	.13	.15	.18	.20	.23
.6	.25	.28	.31	.33	.36	.39	.41	.44	.47	.50
.7	.52	.55	.58	.61	.64	.67	.71	.74	.77	.81
.8	.84	.88	.92	.95	.99	1.04	1.08	1.13	1.18	1.23
.9	1.28	1.34	1.41	1.48	1.55	1.65	1.75	1.88	2.05	2.33

Table 2.1.7.2. Standard Normal Quantiles

## A Frequency Table for the Standard Normal Quantiles

Value	Tally	Frequency
-2.80 to -2.30		1
-2.29 to -1.79		2
-1.78 to -1.28		7
-1.27 to -.77		12
-.76 to -.26		17
-.25 to .25		21
.26 to .76		17
.77 to 1.27		12
1.28 to 1.78		7
1.79 to 2.29		2
2.30 to 2.80		1

Table 2.1.7.3. A Frequency Table for the Standard Normal Quantiles

**Example 2.1.7.2. Paper towel strength, continued.**

Consider again the paper towel strength testing scenario and now the issue of how bell-shaped its data set is. Table 2.1.7.4 was made using the original table and Table 2.1.7.2; it gives the information needed to produce the theoretical  $Q - Q$  plot in Figure 2.1.7.4.

Considering the small size of the data set involved, the plot in Figure 2.1.4 is fairly linear, and so the data set is reasonably bell-shaped. As a practical consequence of this judgment, it is then possible to use the normal probability models discussed in Part 4 to describe breaking strength. These could be employed to make breaking strength predictions, and methods of formal statistical inference based on them could be used in the analysis of breaking strength data.

## Breaking Strength and Standard Normal Quantiles

$i$	$\frac{i-.5}{10}$	$\frac{i-.5}{10}$ Breaking Strength Quantile	$\frac{i-.5}{10}$ Standard Normal Quantile
1	.05	7,583	-1.65
2	.15	8,527	-1.04
3	.25	8,572	-.67
4	.35	8,577	-.39
5	.45	9,011	-.13
6	.55	9,165	.13
7	.65	9,471	.39
8	.75	9,614	.67
9	.85	9,614	1.04
10	.95	10,688	1.65

Table 2.1.7.4. Breaking Strength and Standard Normal Quantiles

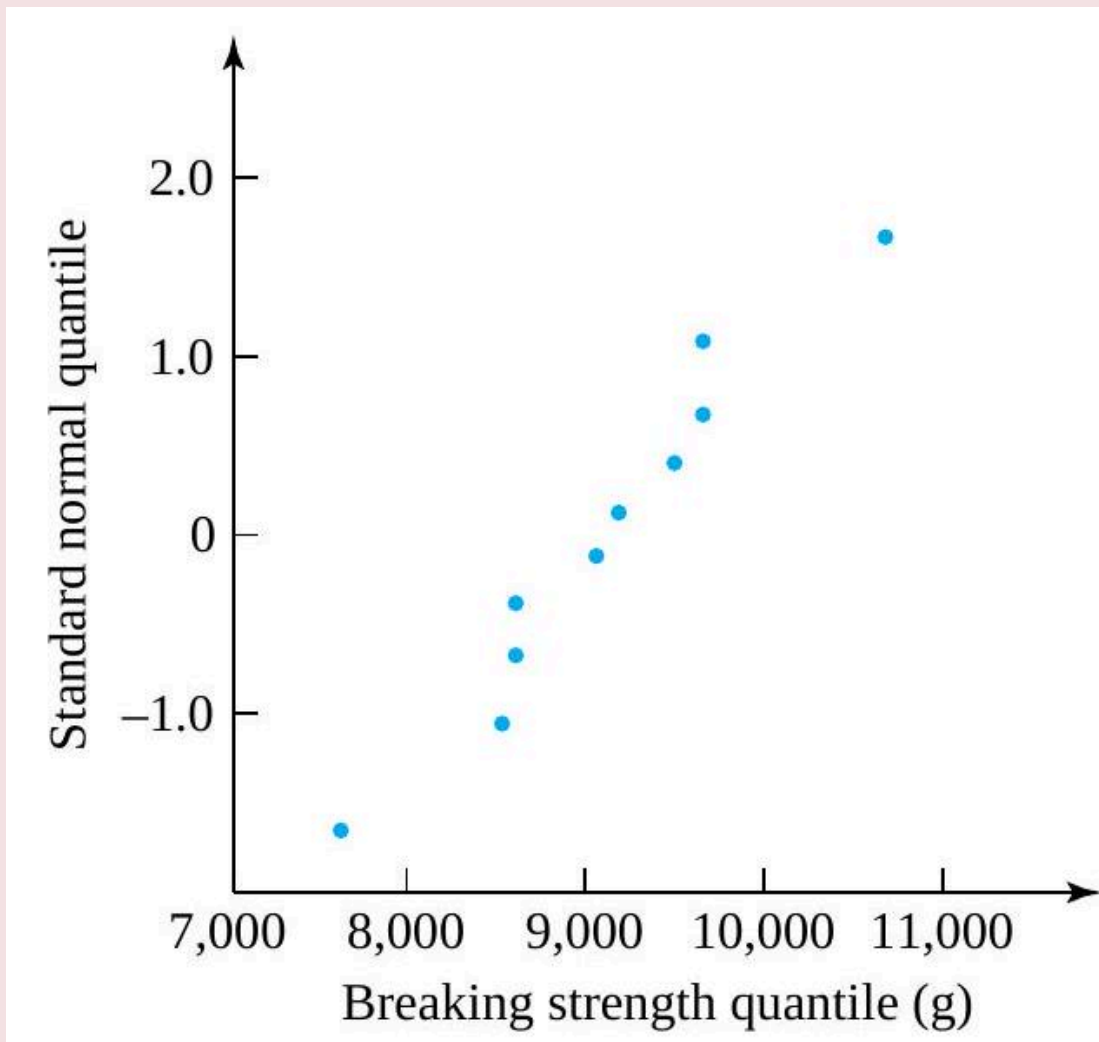
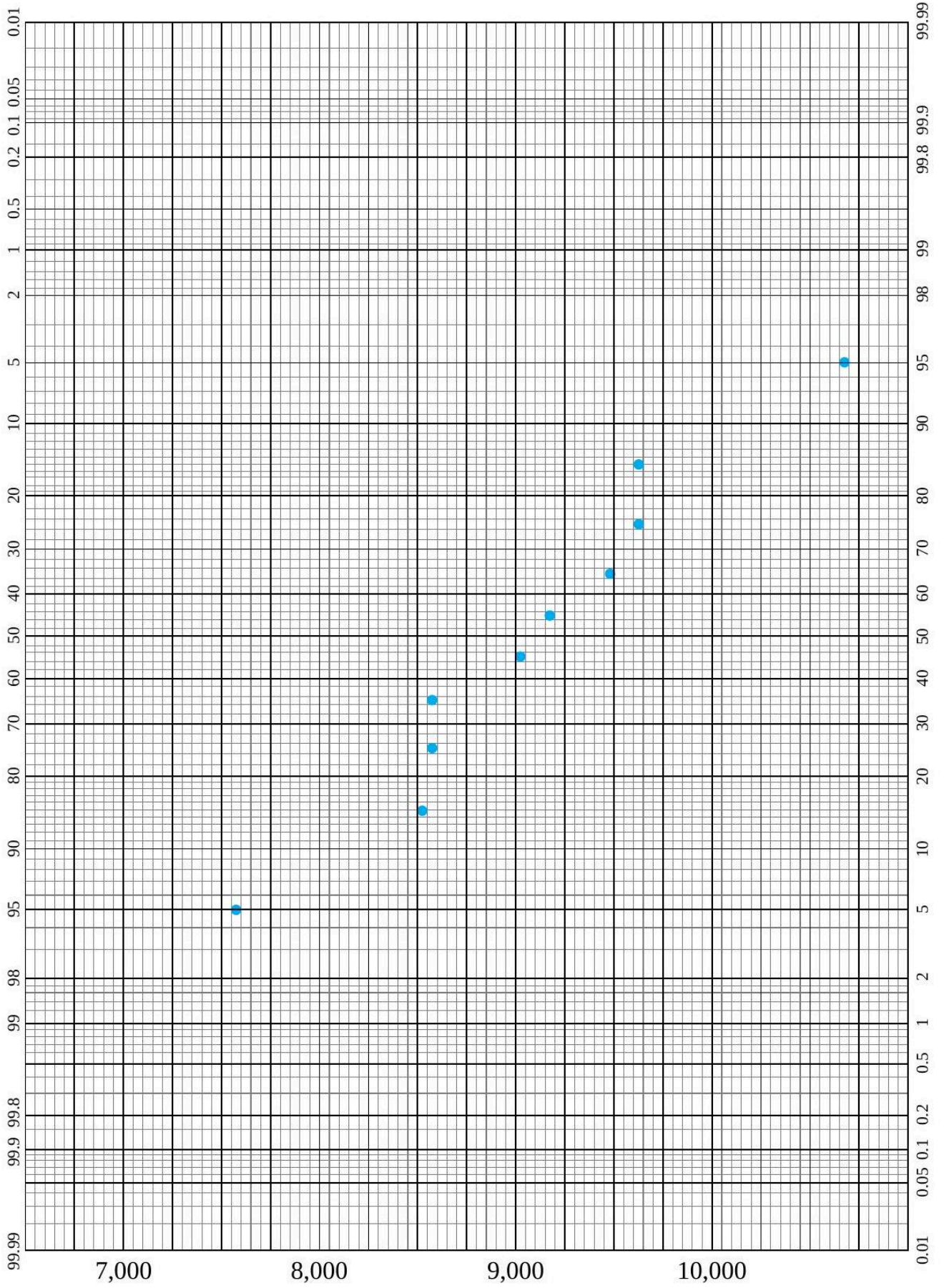


Figure 2.1.7.3. Theoretical Q-Q plot for the paper towel strength.

Special graph paper, called normal probability paper (or just probability paper), is available as an alternative way of making normal plots. Instead of plotting points on regular graph paper using vertical plotting positions taken from Table 2.1.7.2, points are plotted on probability paper using vertical plotting positions of the form  $\frac{i - .5}{n}$ . Figure 2.1.7.4 is a normal plot of the breaking strength data from Example 2.1.7.2 made on probability paper. Observe that this is virtually identical to the plot in Figure 2.1.7.2.



*Figure 2.1.7.4. Normal plot for the paper towel strengths (made on probability paper (image from Keuffel and Esser Company).*

Normal plots are not the only kind of theoretical  $Q - Q$  plots useful to engineers. Many other types of theoretical distributions are of engineering importance, and each can be used to make theoretical  $Q - Q$  plots. This point is discussed in more detail in other modules, but the introduction of theoretical  $Q - Q$  plotting here makes it possible to emphasize the relationship between probability plotting and (empirical)  $Q - Q$  plotting.

## 2.2.1 Measures of Location

Most people are familiar with the concept of an “average” as being representative of, or in the center of, a data set. Temperatures may vary between different locations in a blast furnace, but an average temperature tells something about a middle or representative temperature. Scores on an exam may vary, but one is relieved to score at least above average.

The word average, as used in colloquial speech, has several potential technical meanings. One is the median,  $Q(.5)$ , which was introduced in the last section. The median divides a data set in half. Roughly half of the area enclosed by the bars of a well-made histogram will lie to either side of the median. As a measure of center, it is completely insensitive to the effects of a few extreme or outlying observations. For example, the small set of data

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

has median 6, and this remains true even if the value 10 is replaced by 10,000,000 and/or the value 2 is replaced by  $-200,000$ .

The previous section used the median as a center value in the making of boxplots. But the median is not the technical meaning most often attached to the notion of average in statistical analyses. Instead, it is more common to employ the (arithmetic) mean.

### DEFINITION 2.2.1.1. Arithmetic mean.

The (arithmetic) mean of a sample of quantitative data, say,  $x_1, x_2, \dots, x_n$ , is

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

The mean is sometimes called the first moment or center of mass of a distribution, drawing on an analogy to mechanics. Think of placing a unit mass along the number line at the location of each value in a data set—the balance point of the mass distribution is at  $\bar{x}$ .

#### Example 2.2.1.1. Waste on Bulk Paper Rolls

Hall, Luethé, Pelszynski, and Ringhofer worked with a company that cuts paper from large rolls purchased in bulk from several suppliers.

The company was interested in determining the amount of waste (by weight) on rolls obtained from the various sources. Table 2.2.1.1 gives percent waste data, which the students obtained for six and eight rolls, respectively, of paper purchased from two different sources.

The medians and means for the two data sets are easily obtained. For the supplier 1 data,

$$Q(.5) = .5(.65) + .5(.92) = .785\% \text{ waste}$$

and

$$\bar{x} = \frac{1}{6}(.37 + .52 + .65 + .92 + 2.89 + 3.62) = 1.495\% \text{ waste}$$

For the supplier 2 data,

$$Q(.5) = .5(1.47) + .5(1.58) = 1.525\% \text{ waste}$$

and

$$\begin{aligned} \bar{x} &= \frac{1}{8}(.89 + .99 + 1.45 + 1.47 + 1.58 + 2.27 + 2.63 + 6.54) \\ &= 2.228\% \text{ waste} \end{aligned}$$

## Percent Waste by Weight on Bulk Paper Rolls

---

Supplier 1

Supplier 2

---

.37, .52, .65,  
.92, 2.89, 3.62

.89, .99, 1.45, 1.47,  
1.58, 2.27, 2.63, 6.54

---

Table 2.2.1.1.

Figure 2.2.1.1 shows dot diagrams with the medians and means marked. Notice that a comparison of either medians or means for the two suppliers shows the supplier 2 waste to be larger than the supplier 1 waste. But there is a substantial difference between the median and mean values for a given supplier. In both cases, the mean is quite a bit larger than the corresponding median. This reflects the right-skewed nature of both data sets. In both cases, the center of mass of the distribution is pulled strongly to the right by a few extremely large values.



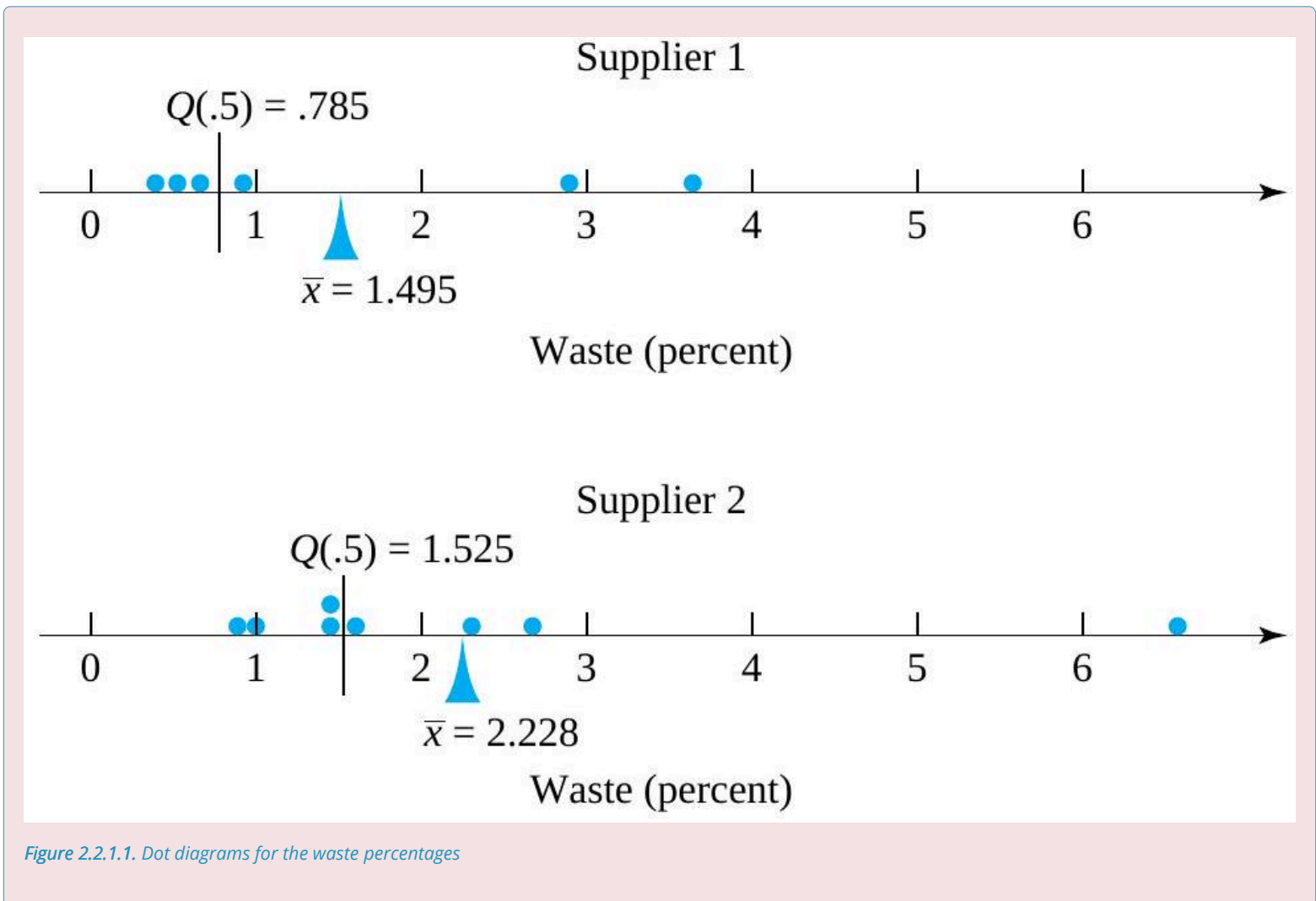


Figure 2.2.1.1. Dot diagrams for the waste percentages

Example 2.2.1.1 shows clearly that, in contrast to the median, the mean is a measure of center that can be strongly affected by a few extreme data values. People sometimes say that because of this, one or the other of the two measures is “better.” Such statements lack sense. Neither is better; they are simply measures with different properties. And the difference is one that intelligent consumers of statistical information do well to keep in mind. The “average” income of employees at a company paying nine workers each 10,000/year and a president 110,000/year can be described as 10,000/year or 20,000/year, depending upon whether the median or mean is being used.

## 2.2.2 Measures of Spread

Quantifying the variation in a data set can be as important as measuring its location. In manufacturing, for example, if a characteristic of parts coming off a particular machine is being measured and recorded, the spread of the resulting data gives information about the intrinsic precision or capability of the machine. The location of the resulting data is often a function of machine setup or settings of adjustment knobs. Setups can fairly easily be changed, but improvement of intrinsic machine precision usually requires a capital expenditure for a new piece of equipment or overhaul of an existing one.

Although the point wasn't stressed in Module 2.1, the interquartile range,  $IQR = Q(.75) - Q(.25)$ , is one possible measure of spread for a distribution. It measures the spread of the middle half of a distribution. Therefore, it is insensitive to the possibility of a few extreme values occurring in a data set. A related measure is the range, which indicates the spread of the entire distribution.

### DEFINITION 2.2.2.1. The range.

The range of a data set consisting of ordered values  $x_1 \leq x_2 \leq \dots \leq x_n$  is

$$R = x_n - x_1$$

Notice the word usage here. The word range could be used as a verb to say, "The data range from 3 to 21." But to use the word as a noun, one says, "The range is  $(21 - 3) = 18$ ." Since the range depends only on the values of the smallest and largest points in a data set, it is necessarily highly sensitive to extreme (or outlying) values. Because it is easily calculated, it has enjoyed long-standing popularity in industrial settings, particularly as a tool in statistical quality control.

However, most methods of formal statistical inference are based on another measure of distributional spread. A notion of "mean squared deviation" or "root mean squared deviation" is employed to produce measures that are called the **variance** and the **standard deviation**, respectively.

**DEFINITION 2.2.2.2. Sample variance and sample standard deviation**

The **sample variance** of a data set consisting of values  $x_1, x_2, \dots, x_n$  is

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

The **sample standard deviation**,  $s$ , is the nonnegative square root of the sample variance.

Apart from an exchange of  $n - 1$  for  $n$  in the divisor,  $s^2$  is an average squared distance of the data points from the central value  $\bar{x}$ . Thus,  $s^2$  is nonnegative and is 0 only when all data points are exactly alike. The units of  $s^2$  are the squares of the units in which the original data are expressed. Taking the square root of  $s^2$  to obtain  $s$  then produces a measure of spread expressed in the original units.

**Example 2.2.2.1. Waste on Bulk Paper Rolls, continued.**

The spreads in the two sets of percentage wastes recorded in Table 2.2.1.1 can be expressed in any of the preceding terms. For the supplier 1 data,

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

$$Q(.75) = 2.89$$

and so

$$IQR = 2.89 - .52 = 2.37\% \text{ waste}$$

Also,

$$R = 3.62 - .37 = 3.25\% \text{ waste}$$

Further,

$$\begin{aligned} s^2 &= \frac{1}{6-1} \left( (.37 - 1.495)^2 + (.52 - 1.495)^2 + (.65 - 1.495)^2 + (.92 - 1.495)^2 \right. \\ &\quad \left. + (2.89 - 1.495)^2 + (3.62 - 1.495)^2 \right) \\ &= 1.945(\% \text{ waste})^2 \end{aligned}$$

so that

$$s = \sqrt{1.945} = 1.394\% \text{ waste}$$

Similar calculations for the supplier 2 data yield the values

$$IQR = 1.23\% \text{ waste}$$

and

$$R = 6.54 - .89 = 5.65\% \text{ waste}$$

Further,

$$\begin{aligned}s^2 &= \frac{1}{8-1} \left( (.89 - 2.228)^2 + (.99 - 2.228)^2 + (1.45 - 2.228)^2 + (1.47 - 2.228)^2 \right. \\ &\quad \left. + (1.58 - 2.228)^2 + (2.27 - 2.228)^2 + (2.63 - 2.228)^2 + (6.54 - 2.228)^2 \right) \\ &= 3.383 (\% \text{ waste})^2\end{aligned}$$

so

$$s = 1.839\% \text{ waste}$$

Supplier 2 has the smaller IQR but the larger  $R$  and  $s$ . This is consistent with Figure 2.2.1.1 The central portion of the supplier 2 distribution is tightly packed. But the single extreme data point makes the overall variability larger for the second supplier than for the first.

The calculation of sample variances just illustrated is meant simply to reinforce the fact that  $s^2$  is a kind of mean squared deviation. Of course, the most sensible way to find sample variances in practice is by using either a handheld electronic calculator with a preprogrammed variance function or a statistical package on a personal computer.

## 2.2.3 Statistics and Parameters

At this point, it is important to introduce some more basic terminology. Jargon and notation for distributions of samples are somewhat different than for population distributions (and theoretical distributions).

### DEFINITION 2.2.3.1. Statistics and Parameters

Numerical summarizations of sample data are called (sample) **statistics**. Numerical summarizations of population and theoretical distributions are called (population or model) **parameters**. Typically, Roman letters are used as symbols for statistics, and Greek letters are used to stand for parameters.

As an example, consider the mean. Definition 2.2.1.1 refers specifically to a calculation for a sample. If a data set represents an entire population, then it is common to use the lowercase Greek letter mu ( $\mu$ ) to stand for the population mean and to write:

**Population mean 2.2.3.2.** 
$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

Comparing this expression to the one in Definition 2.2.1.1, not only is a different symbol used for the mean but also  $N$  is used in place of  $n$ . It is standard to denote a population size as  $N$  and a sample size as  $n$ . As another example of the usage suggested by Definition 2.2.3.1, consider the variance and standard deviation. Definition 2.2.1.2 refers specifically to the sample variance and standard deviation. If a data set represents an entire population, then it is common to use the lowercase Greek sigma squared ( $\sigma^2$ ) to stand for the population variance and to define:

**Population Variance 2.2.3.3.**

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

The non-negative square root of  $\sigma^2$  is then called the population standard deviation,  $\sigma$ .

On one point, this text will deviate from the Roman/Greek symbolism convention laid out in Definition 2.2.3.1: the notation for quantiles.  $Q(p)$  will stand for the  $p^{\text{th}}$  quantile of a distribution, whether it is from

a sample, a population, or a theoretical model.

## 2.2.4 Plots of Summary Statistics with Time and Factors

Plotting numerical summary measures in various ways is often helpful in the early analysis of engineering data. For example, plots of summary statistics against time are frequently revealing.

### Example 2.2.4.1 Monitoring a Critical Dimension of Machined Parts, continued.

Cowan, Renk, Vander Leest, and Yakes worked with a company that makes precision metal parts. A critical dimension of one such part was monitored by occasionally selecting and measuring five consecutive pieces and then plotting the sample mean and range. Table 2.2.4.1 gives the  $\bar{x}$  and  $R$  values for 25 consecutive samples of five parts. The values reported are in .0001 in.

Figure 2.2.4.1 is a plot of both the means and ranges against order of observation. Looking first at the plot of ranges, no strong trends are obvious, which suggests that the basic short-term variation measured in this critical dimension is stable. The combination of process and measurement precision is neither improving nor degrading with time. The plot of means, however, suggests some kind of physical change. The average dimensions from the second shift on October 27 (samples 9 through 15) are noticeably smaller than the rest of the means. It turned out to be the case that the parts produced on that shift were not really systematically any different from the others. Instead, the person making the measurements for samples 9 through 15 used the gauge in a fundamentally different way than other employees. The pattern in the  $\bar{x}$  values was caused by this change in measurement technique.

Means and Ranges for a Critical Dimension on Samples of $n = 5$ Parts											
Sample	Date	Time		$\bar{x}$	$R$	Sample	Date	Time	$\bar{x}$	$R$	
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						

Table 2.2.4.1





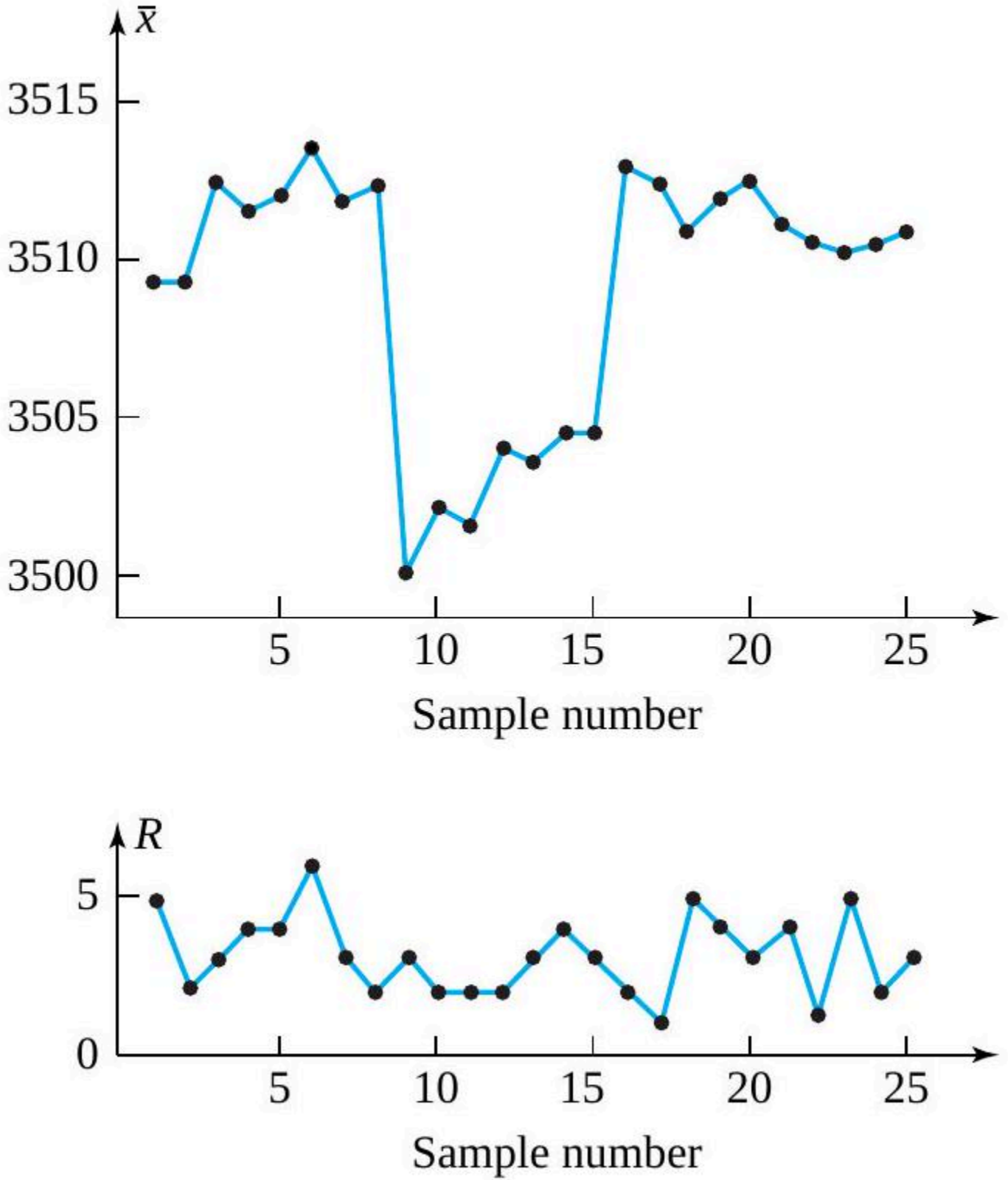


Figure 2.2.4.1 Plots of  $\bar{x}$  and  $R$  over time.

## TERMINOLOGY AND CAUSES FOR PATTERNS ON PLOTS AGAINST TIME

---

Patterns revealed in the plotting of sample statistics against time ought to alert an engineer to look for a physical cause and (typically) a cure. Systematic variations or cycles in a plot of means can often be related to process variables that come and go on a more or less regular basis. Examples include seasonal or daily variables like ambient temperature or those caused by rotation of gauges or fixtures. Instability or variation in excess of that related to basic equipment precision can sometimes be traced to mixed lots of raw material or overadjustment of equipment by operators. Changes in level of a process mean can originate in the introduction of new machinery, raw materials, or employee training and (for example) tool wear. Mixtures of several patterns of variation on a single plot of some summary statistic against time can sometimes (as in Example 2.2.4.1) be traced to changes in measurement calibration. They are also sometimes produced by consistent differences in machines or streams of raw material.

## PLOTS AGAINST PROCESS VARIABLES

---

Plots of summary statistics against time are not the only useful ones. Plots against process variables can also be quite informative.

### Example 2.2.4.2 Plotting the Mean Shear Strength of Wood Joints.

In their study of glued wood joint strength, Dimond and Dix obtained the values given in Table 2.2.4.2 as mean strengths (over three shear tests) for each combination of three woods and three glues. Figure 2.2.4.2 gives a revealing plot of these  $3 \times 3 = 9$  different  $\bar{x}$ 's.

From the plot, it is obvious that the gluing properties of pine and fir are quite similar, with pine joints averaging around 40–45 lb stronger. For these two soft woods, cascamate appears slightly better than carpenter's glue, both of which make much better joints than white glue. The gluing properties of oak (a hardwood) are quite different from those of pine and fir. In fact, the glues perform in exactly the opposite ordering for the strength of oak joints. All of this is displayed quite clearly by the simple plot in Figure 2.2.4.2.

## Mean Joint Strengths for Nine Wood/Glue Combinations

Wood	Glue	$\bar{x}$ Mean Joint Shear Strength (lb)
pine	white	131.7
pine	carpenter's	192.7
pine	cascamite	201.3
fir	white	92.0
fir	carpenter's	146.3
fir	cascamite	156.7
oak	white	257.7
oak	carpenter's	234.3
oak	cascamite	177.7

Table 2.2.4.2

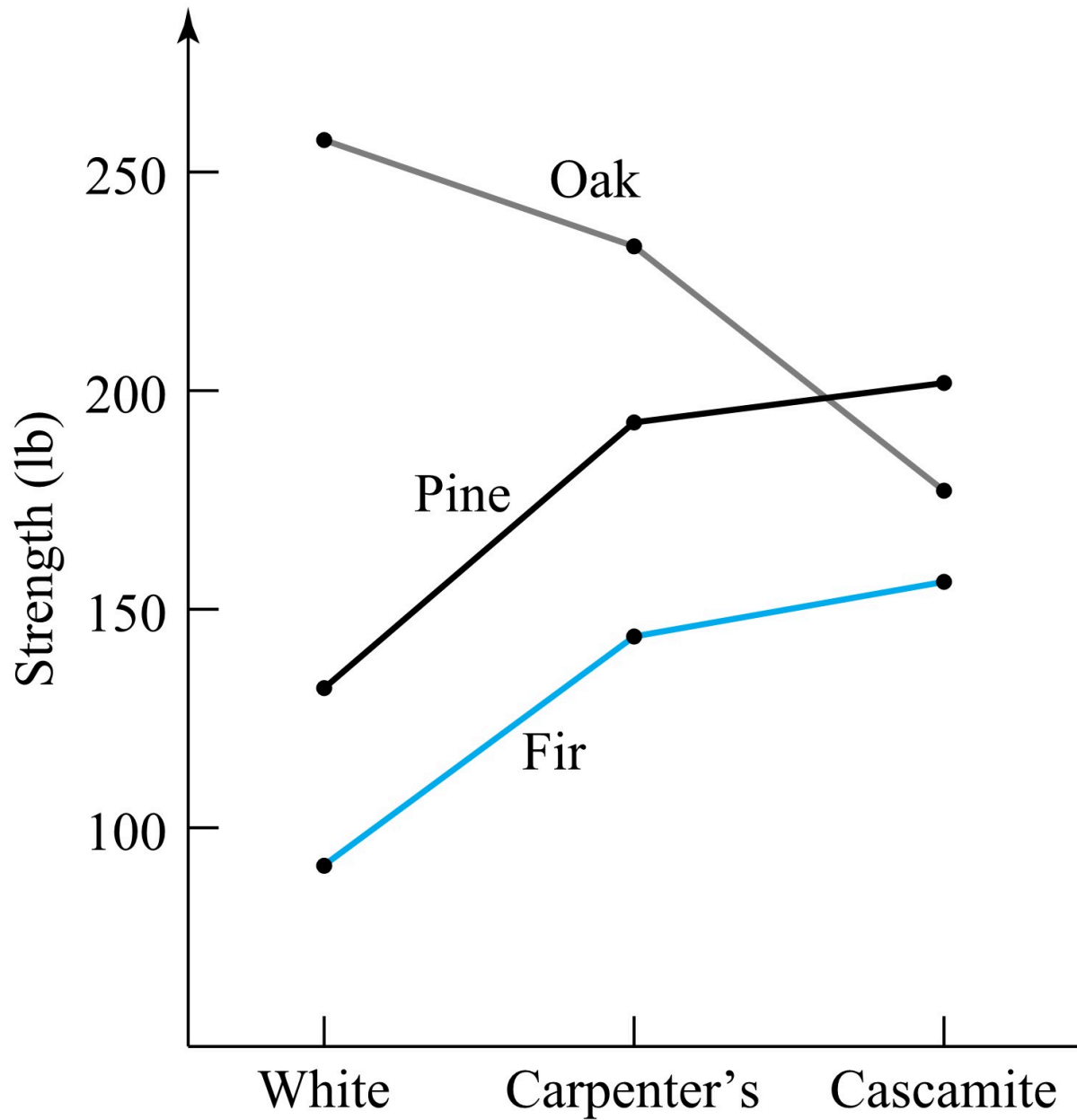


Figure 2.2.4.2 Plot of mean joint strength vs. glue type for three woods.

The two previous examples have illustrated the usefulness of plotting sample statistics against time and against levels of an experimental variable.

## 2.2.5 Bar Charts and Plots for Qualitative and Count Data

The techniques presented thus far in this chapter are primarily relevant to the analysis of measurement data. As noted in Part 1, conventional wisdom is that where they can be obtained, measurement data (or variables data) are generally preferable to count and qualitative data (or attributes data). Nevertheless, qualitative or count data will sometimes be the primary information available. It is therefore worthwhile to consider their summarization and visualization.

Often, a study will produce several values of  $\hat{p}$  or  $\hat{u}$  that need to be compared. Bar charts and simple bivariate plots can be a great aid in summarizing these results.

### Example 2.2.5.1. Defect Classifications of Cable Connectors.

Delva, Lynch, and Stephany worked with a manufacturer of cable connectors. Daily samples of 100 connectors of a certain design were taken over 30 production days, and each sampled connector was inspected according to a well-defined (operational) set of rules. Using the information from the inspections, each inspected connector could be classified as belonging to one of the following five mutually exclusive categories:

Category A: having “very serious” defects

Category B: having “serious” defects but no “very serious” defects

Category C: having “moderately serious” defects but no “serious” or “very serious” defects

Category D: having only “minor” defects

Category E: having no defects

Table 2.2.5.1 gives counts of sampled connectors falling into the first four categories (the four defect categories) over the 30-day period.

Then,	using	the	fact	that
				$\hat{p}_A = 3/3000 = .0010$
				$\hat{p}_B = 0/3000 = .0000$
				$\hat{p}_C = 11/3000 = .0037$
				$\hat{p}_D = 1/3000 = .0003$

Notice that here  $\hat{p}_E = 1 - (\hat{p}_A + \hat{p}_B + \hat{p}_C + \hat{p}_D)$ , because categories A through E represent a set of nonoverlapping and exhaustive classifications into which an individual connector must fall, so that the  $\hat{p}$  's must total to 1 .

## Counts of Connectors Classified into Four Defect Categories

Category	Number of Sampled Connectors
A	3
B	0
C	11
D	1

Table 2.2.5.1.

Figure 2.2.5.1 is a bar chart of the fractions of connectors in the categories A through D. It shows clearly that most connectors with defects fall into category C, having moderately serious defects but no serious or very serious defects. This bar chart is a presentation of the behavior of a single categorical variable.

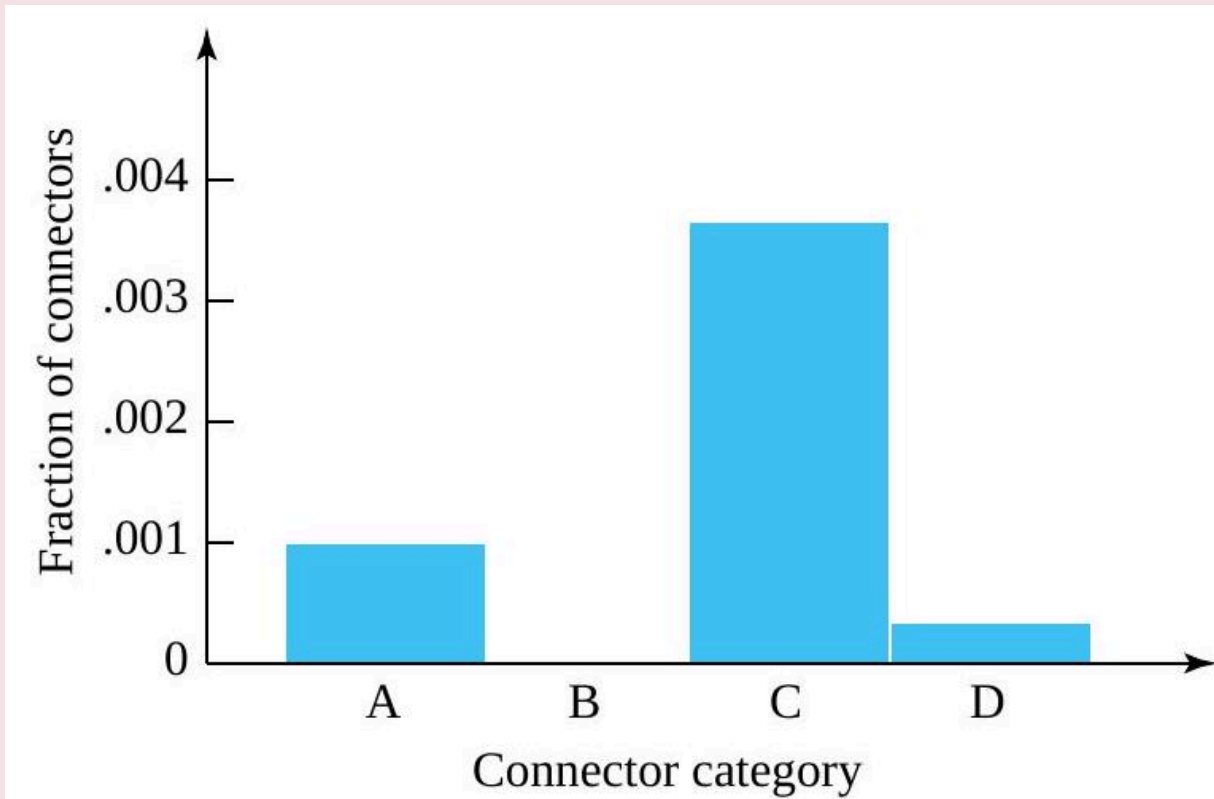


Figure 2.2.5.1. Bar Chart of connector defects.

#### Example 2.2.5.2. Pneumatic tool manufacture.

Kraber, Rucker, and Williams worked with a manufacturer of pneumatic tools. Each tool produced is thoroughly inspected before shipping. The students collected some data on several kinds of problems uncovered at final inspection. Table 2.2.5.2 gives counts of tools having these problems in a particular production run of 100 tools.

## Counts and Fractions of Tools with Various Problems

Problem	Number of Tools	$\hat{p}$
Type 1 leak	8	.08
Type 2 leak	4	.04
Type 3 leak	3	.03
Missing part 1	2	.02
Missing part 2	1	.01
Missing part 3	2	.02
Bad part 4	1	.01
Bad part 5	2	.02
Bad part 6	1	.01
Wrong part 7	2	.02
Wrong part 8	2	.02

Table 2.2.5.2.

This is a summarization of highly multivariate qualitative data. The categories listed in Table 2.2.5.2 are not mutually exclusive; a given



tool can fall into more than one of them. Instead of representing different possible values of a single categorical variable (as was the case with the connector categories in Example 2.2.5.1), the categories listed above each amount to 1 (present) of 2 (present and not present) possible values for a different categorical variable. For example, for type 1 leaks,  $\hat{p} = .08$ , so  $1 - \hat{p} = .92$  for the fraction of tools without type 1 leaks. The  $\hat{p}$  values do not necessarily total to the fraction of tools requiring rework at final inspection. A given faulty tool could be counted in several  $\hat{p}$  values.

Figure 2.2.5.2 is a bar chart of the information on tool problems in Table 2.2.5.1. It shows leaks to be the most frequently occurring problems on this production run.

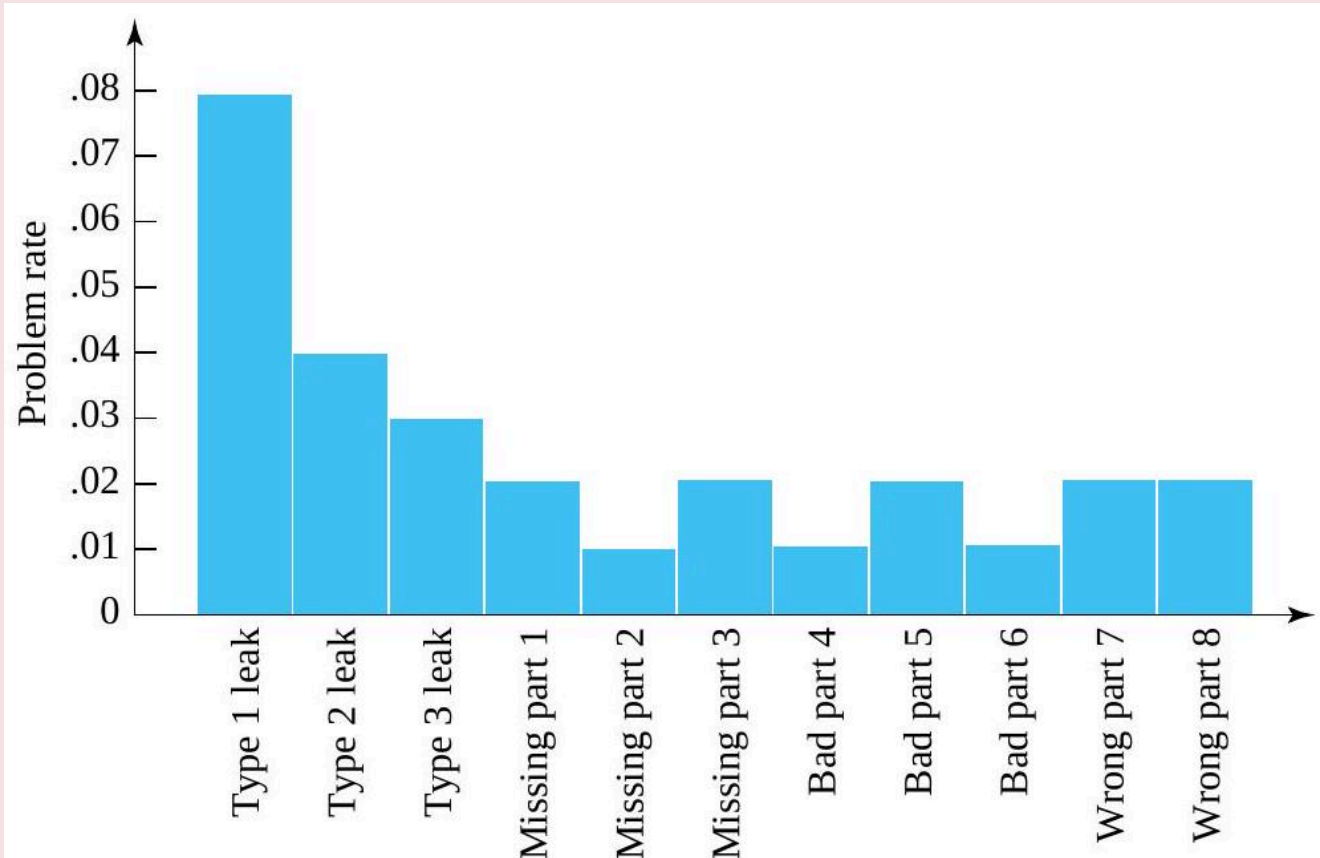


Figure 2.2.5.2. Bar chart for assembly problems.

Figures 2.2.5.1 and 2.2.5.2 are both bar charts, but they differ considerably. The first concerns the behavior of a single (ordered) categorical variable—namely, Connector Class. The second concerns the behavior of 11 different present-not present categorical variables, like Type 1 Leak, Missing Part 3, etc. There may be some significance to the shape of Figure 2.2.5.1, since categories A through D are arranged in decreasing order of defect severity, and this order was used in the making of the figure. But the shape of Figure 2.2.5.2 is essentially arbitrary, since the particular ordering of the tool problem categories used to make the figure is arbitrary. Other equally sensible orderings would give quite different shapes.

## 2.2.6 Summary Statistics and Statistical Computing

The numerical data summaries introduced in this chapter are relatively simple. For small data sets they can be computed quite easily using only a pocket calculator. However, for large data sets and in cases where subsequent additional calculations or plotting may occur, statistical software can be convenient.

This Jupyter Notebook using Python is available to look at and access for download at the course [GitHub Site](#) or at the [Special GitHub Site for Part 2](#).

Or you can open an interactive computing environment to work through the Jupyter Notebook using Python through a Binder Site using the Special GitHub Site for the Part 2 example. Click this [Binder Site](#) to go to the Binder Site for the Example (located at <https://mybinder.org/v2/gh/Statistical-Methods-for-Engineering/Special-GitHub-Site-Part-2-Example-Percent-Waste-by-Weight-on-Bulk-Paper-Rolls/HEAD>).

Printout 1 illustrates the use of the Python Jupyter Notebook statistical package to produce summary statistics for the percent waste data sets from this Part. The mean, median, and standard deviation values on the printout agree with those produced in the example. However, the first and third quartile figures on the printout do not match exactly those found earlier. Python's library for numpy and pandas uses slightly different conventions for those quantities than the ones introduced in Part 2.

Supply_1	
<b>count</b>	6.000000
<b>mean</b>	1.495000
<b>std</b>	1.394457
<b>min</b>	0.370000
<b>25%</b>	0.552500
<b>50%</b>	0.785000
<b>75%</b>	2.397500
<b>max</b>	3.620000

High-quality statistical packages like Python (or JMP, SAS, SPSS, SYSTAT, SPLUS, MINITAB, MATLAB, R etc.) are widely available. One of them should be on the electronic desktop of every working engineer. Unfortunately, this is not always the case, and engineers often assume that standard spreadsheet software (perhaps augmented with third party plug-ins) provides a workable substitute. Often this is true, but sometimes it is not. Statistical computing and some level of competence in Data Science are needed by the modern engineer.

Supply_2	
<b>count</b>	8.000000
<b>mean</b>	2.227500
<b>std</b>	1.839229
<b>min</b>	0.890000
<b>25%</b>	1.335000
<b>50%</b>	1.525000
<b>75%</b>	2.360000
<b>max</b>	6.540000

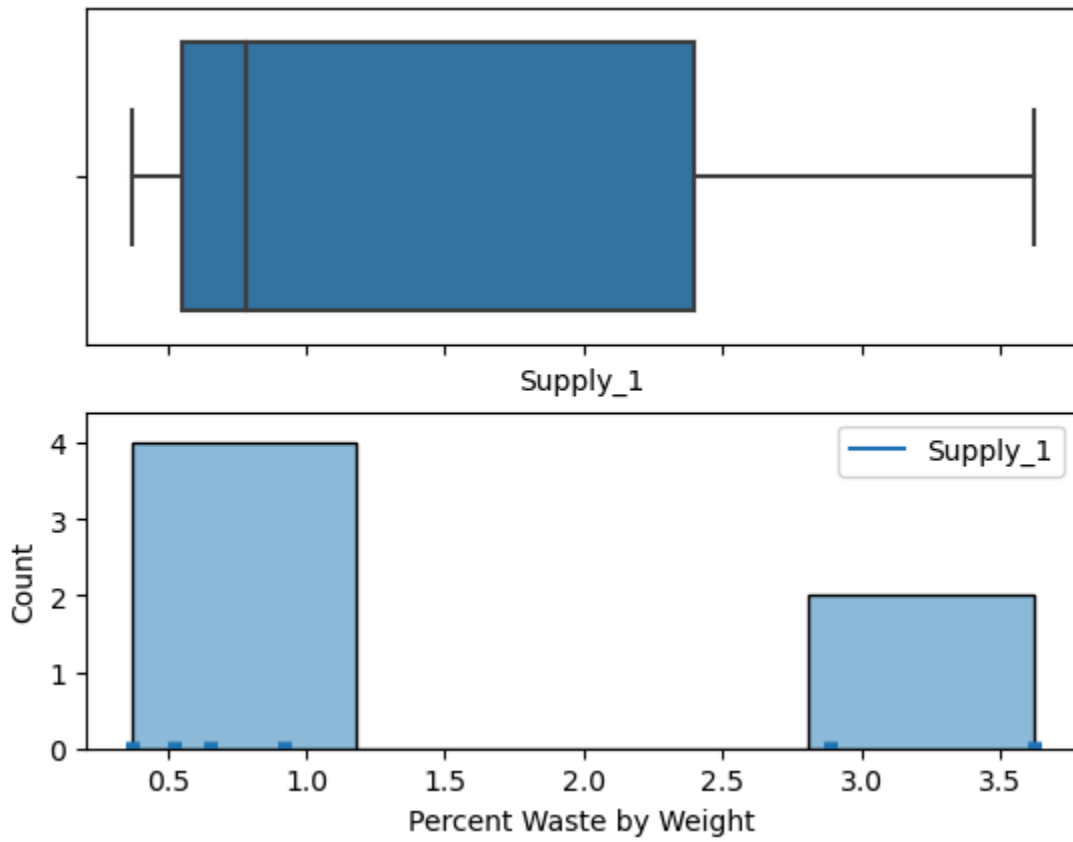


Figure 2.2.6.1.

Figure 2.2.6.1 is a Boxplot and Histogram of Supply 1 of the example. Look through the Jupyter Notebook and begin to summarize and visualize this data.

## 2.2.7 Tutorial 2 - Data Cleaning, Summarization, and Plotting in Python

At this point, it is recommended that you work your way through the [Tutorial 2 exercise](#) found on the associated GitHub repository. This exercise will introduce you to data cleaning and creation of simple plots in Python.

**It is strongly recommended that you consult the [Reading Data into Python & Data Cleaning Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section.

### 3.0.1 Introduction to Probability and Random Variables



Figure 3.1.0.1. Blaise Pascal, Palace of 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](https://commons.wikimedia.org/wiki/File:Pierre_de_Fermat.jpg)

The onset of probability as a practical and scientific discipline is primarily attributed to the joint efforts of Blaise Pascal (1623-1662) and Pierre de Fermat (1601-1665). Their collaboration began over a gambling problem posed by Chevalier de Mere in 1654. In their correspondence dating back to 1654, they delved into a gambling quandary famously known as The Problem of Points, also termed the problem of dividing the stakes. This problem essentially revolves around determining a fair method to distribute the pot when a game concludes prematurely, without a definitive winner. Through their correspondence, Pascal and Fermat not only addressed these specific problems but also laid the foundational groundwork for probability theory.

In the earlier modules, we explored how to use descriptive statistics and data visualization for data summarization. After data descriptions, it's often vital to infer about the originating process of the data,

especially when trying to predict a process's long-term performance from a limited data sample. This approach inherently involves some level of uncertainty due to the reliance on sample data.

Random variables serve as a fundamental tool for quantifying and managing the uncertainty inherent in various processes or experiments. These variables, which can be either discrete or continuous, assume numerical outcomes based on the randomness of the observed phenomena. A random variable describes the outcomes of a statistical observation or experiment, and the values of a random variable can vary with each repetition of an experiment.

#### Key Takeaways

- A discrete random variable is a random variable with a finite set of possible outcomes (interval data).
- A continuous random variable is a random variable with an interval of possible outcomes (continuous data).

Random variables are the outcome of an observation or experiment. Probability at its core is a best “guess” about the outcome of a random event in order to make a decision. Making a decision based on the most educated “guess” is what probability theory is based on. The necessity to make educated guesses about outcomes with inherent uncertainty is prevalent in various fields. For instance, politicians use polls to estimate their chances of winning an election, doctors select treatments based on expected outcomes, gamblers choose games based on perceived odds of winning, and career choices are often influenced by the perceived availability of job opportunities. Probability plays a fundamental role in the application of statistics within engineering, as it provides a framework for making sense of and interpreting statistical data. You are constantly calculating probabilities and then refining your best “guess”.

#### Key Takeaways

Statistical probability provides the framework for describing and analyzing random phenomena and uncertainty, providing us with a best “guess”.

#### Learning Objectives

##### Learning Outcomes for Module 3.1:

- Understand random variables in the context of a statistical observation or experiment.
- Demonstrate an understanding of long-term relative frequencies.
- Understand the properties and terminology of probability.
- Understand the concepts of mutually exclusive and independent events.
- Apply Addition and Multiplication Rules to calculate probabilities of multiple events.
- Recognize the role of inferential statistics within the wider field of statistics

## 3.0.2 *Attributions Part 3*

This first draft of Part 3 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

## 3.1.1 Probability of Random Events

### PROBABILITY

---

Probability is the mathematical framework concerning events from a particular activity and numerical descriptions of how likely they are to occur. Probability is a measure, assigned as a number between 0 and 1, inclusive, that is associated with how certain we are of outcomes of the particular activity.

First, we will review some probability terminology:

Terminology for Probability:

- An *experiment* is a process (a particular activity, an experience, a phenomenon, or a planned operation carried out under controlled conditions ) that produces an *observation*.
- An *outcome* is the *mutually exclusive* result of an experiment's possible observations.
- *Mutually exclusive* results means that only one of the possible outcomes can be observed.
- The set of all possible outcomes is called the *sample space*.
- An *event* is a subset of the sample space.
- A *trial* is a single running of an experiment.

### RANDOM EVENTS

---

Randomness and uncertainty exist in all experiments: in our daily lives and everywhere in the world, as well as in every discipline in medicine, science, and engineering. A random experiment is one where the outcome exists but is not predetermined or known. A random event is therefore the subset of the sample space from a random experiment. Flipping a fair coin is an example of a random experiment, since the outcome of being a heads or a tails is uncertain. Ways of representing sample space are to list the set, to visualize a Venn diagram, to draw a tree diagram, and to write out a contingency table. These methods may be useful when we begin to assign and calculate probabilities to multiple events.

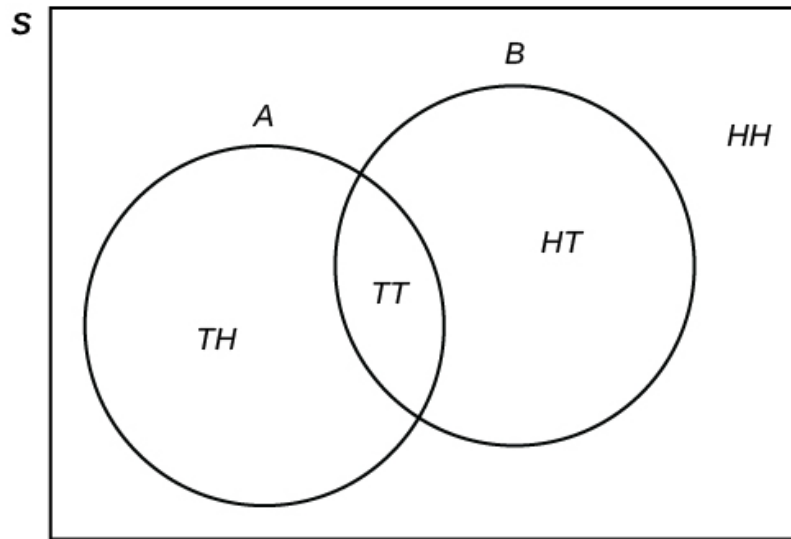
We will use capital letters to denote a set and will list all the outcomes in curly brackets. For example,



to define the sample space of the random experiment of flipping one fair coin:  $S = \{H, T\}$  where  $H$  = heads and  $T$  = tails are the outcomes. The sample space for flipping two fair coins once is shown:  $S = \{(HH), (HT), (TH), (TT)\}$ . We will also use capital letters to denote an event, like  $A$  and  $B$ . For example, we can define event  $A$  as realizing tails on the first coin and event  $B$  = tails on the second coin. This would be shown as  $A = \{TT, TH\}$  and  $B = \{TT, HT\}$ . Using diagrams is helpful in representing the operations of multiple events together.

### Venn Diagram

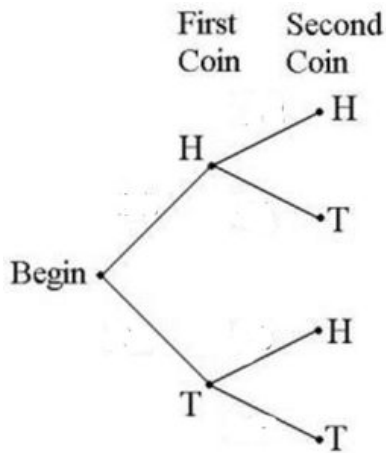
A Venn diagram is the visual representation of a sample space and events in the form of circles or ovals showing their intersections. For the example above of flipping a fair coin twice, we have event  $A$  and event  $B$ , and the outcome  $HH$  is in neither  $A$  nor  $B$ . The Venn diagram is as follows:



Flip two fair coins:  $A = \{TT, TH\}$  and  $B = \{TT, HT\}$ . Therefore,  $A \text{ AND } B = A \text{ INTERSECTION } B = A \cap B = \{TT\}$ .  $A \text{ OR } B = A \text{ UNION } B = A \cup B = \{TH, TT, HT\}$ .

### Tree Diagram

A tree diagram is a representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes.



### Contingency Table

Contingency tables classify outcomes and events. These tables contain rows and columns that display bivariate frequencies of categorical data. The joint events here are happening together in a cell, such as, for the above example, the joint events of  $A = \{TH, TH\}$  and  $B = \{TH, HT\} = TH$ . The marginal events are those shown on the margins of the table, and are those that occur for a single event with no regard for the other events in the table. For our example, we have marginal event  $A$  and the associated joint events,  $A = \{TH, TH\}$ .

		2 <sup>nd</sup> coin		
		Head	Tail	
1 <sup>st</sup> coin				
Head		HH	HT	HH,HT
Tail		TH	TT	TH,TT
Total		HH,TH	HT,TT	S

### SET THEORY

Since events of random experiments are sets, we will review some basic set theory:

$A$  and  $B$  are events in a sample space.

- If all the elements of  $A$  belong to  $B$ , it is shown as  $\subseteq$ .
- The empty set of no outcomes is shown as  $\emptyset$ .
- $A$  and  $B$  are disjoint, or mutually exclusive, if  $A \cap B = \emptyset$ .
- $A$  is a subset of  $B$  if every element of  $A$  is also in  $B$ , and shown as  $A \subset B$ .

For multiple events, we therefore state:

$A$  and  $B$  are events in a sample space.

- $A \cap B$  is the set of outcomes that are in both  $A$  and  $B$
- $A \cup B$  is the set of outcomes that are in either  $A$  or  $B$ , or both
- The complement of  $A$  is  $\overline{A} = S - A$ . Therefore  $\overline{A}$  is the set of outcomes that are not in  $A$ .

## PROBABILITY THEORY

---

The usefulness of probability is in assigning sensible likelihoods of occurrence to possible happenings for random experiments. Before we look at some practical ways to use probability, let's discuss ways that we can interpret probability theory for random experiments.

Probability can be interpreted as a quantification of our degree of subjective personal belief that an event will happen in the random experiment. The most common subjective approach is using Bayesian probability, but this is beyond the scope of this course. In a simple form, you can think of probability as the proportion of a favorable event that occurs over the number of total outcomes possible in an equally probable sample space. Another interpretation is based on quantifying the objective results of random experiments. This frequentist probability approach is the basis of most introduction to statistics courses and of much of statistical methods, and will be the framework we use for harnessing the randomness of random experiments.

Frequentist probability states that the probability of a random event is the relative frequency of the event when the experiment is repeated indefinitely. This interpretation is often stated as being the relative frequency of an experiment "in the long-run" or "in the long-term". Given an event  $A$  in a sample space,

the relative frequency of A is the ratio,  $\frac{m}{n}$ , with m being the number of outcomes in the the event A occurs and n being the total number of outcomes of the experiment. A claim of the frequentist approach is that, as the number of trials increases, the change in the relative frequency will diminish. Hence, one can view a probability as the *limiting value* of the corresponding relative frequencies. You can realize the relative frequency by either running real experiments and finding an empirical or estimated probability or by recognizing the theoretical model for the experiment and adopting a theoretical probability based on events from the sample space.

In the case of of a sample space where equally likely outcomes are stated, then

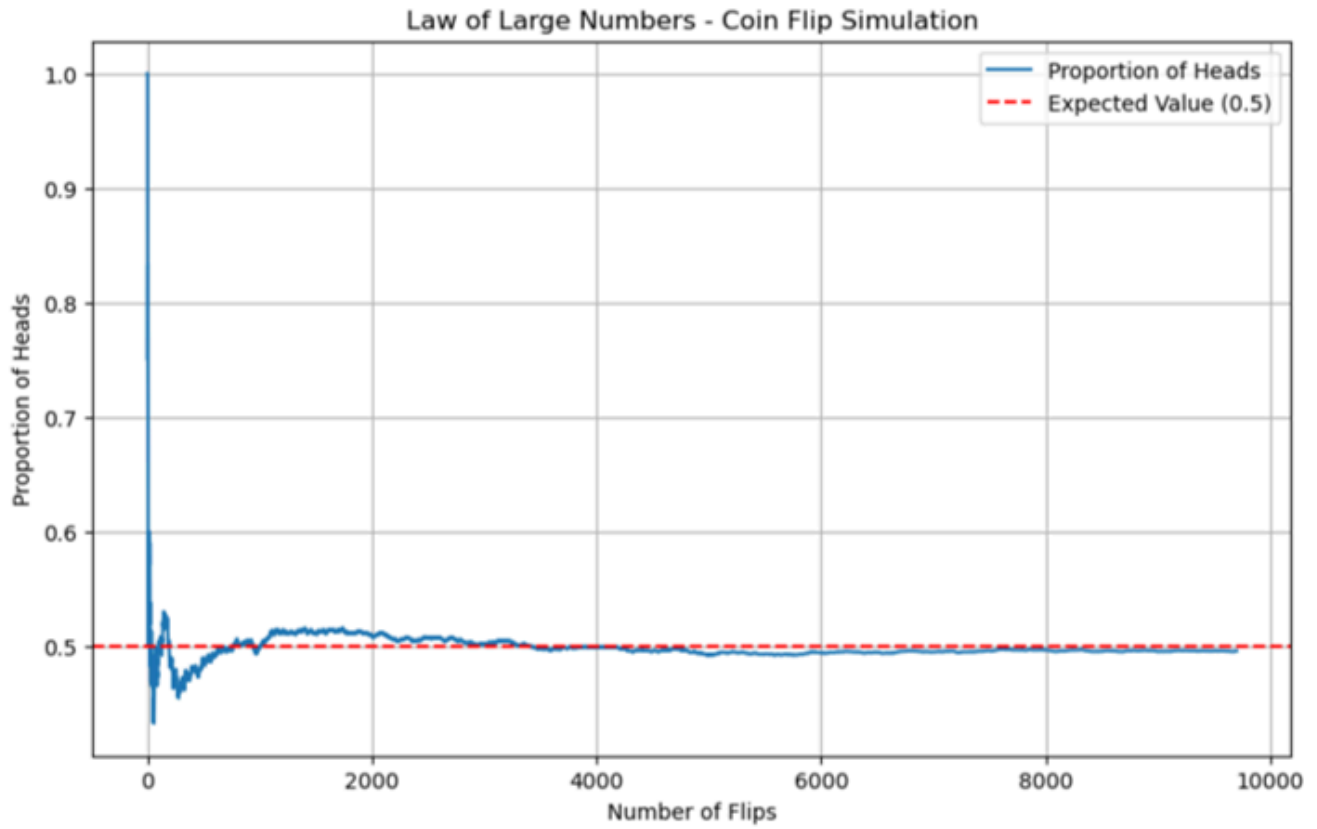
If the outcomes in a finite sample space S all have the same probability, then for any event A:

$$P[A] = \frac{\text{the number of outcomes in event A}}{\text{the number of outcomes in S}}$$

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a fair coin, a Head (*H*) and a Tail (*T*) are equally likely to occur, so you can count the number of outcomes for event A=getting one heads and divide by the total number of outcomes in the sample space. If you toss two fair coins, the sample space is {*HH, TH, HT, TT*}. There are two outcomes that meet this condition {*HT, TH*}, so  $P(A) = \frac{2}{4} = 0.5$ .

This text will use the notational convention that a capital P followed by an expression or phrase enclosed by brackets will be read “the probability” of that expression. So P(A) is the probability of the random variable of A.

Over the long-term, the relative frequency of tossing a fair coin will approach 0.5, the theoretical probability. Since there are only 2 possible outcomes to tossing a coin, this empirical probability of success as an experimental relative frequency will converge to the theoretical probability. The law of large numbers states that as the number of trials increases sample values tend to converge on the expected result. This can be interpreted here as the proportion of heads in a “large” number of coin flips “should be” roughly 0.5. In particular, the proportion of heads after *n* flips will converge to 0.5 as *n* approaches infinity. probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability.



See the Jupyter Notebook in the GitHub repository for a simulation for this demonstration of flipping a coin in the long-run: [CoinTossSimulation](#).

## 3.1.2 Probability and Independence of Events

### PROBABILITY OF RANDOM EVENTS

---

The goal of probability is to assign numbers between 0 and 1 as measures of the likelihood of random events. For example, if the experiment is to flip one fair coin, event  $A$  might be getting at most one head. The *probability of an event  $A$*  is written  $P(A)$ , is assigned a number between zero and one, inclusive, and describes the proportion of time we expect the event to occur over the long-term.  $P(A) = 0$  means the event  $A$  can never happen.  $P(A) = 1$  means the event  $A$  always happens.  $P(A) = 0.5$  means the event  $A$  is equally likely to occur or not to occur. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

We will review the axioms of probability to build up the rules of probability that we will use in this course:

A system of probabilities is an assignment of numbers (probabilities)  $P(A)$ , to events  $A$  in such a way that

- For each event  $A$ ,  $P(A)$  is a non-negative real number between 0 and 1 inclusive. This is:  $0 \leq P(A) \leq 1$ .
- The probability of the sample space  $S$  is 1 and the probability of the empty set is 0. This is:  $P(S) = 1$  and  $P(\emptyset) = 0$ .

- Probabilities are countably additive for disjoint events. This is: 
$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

### CONDITIONAL PROBABILITY AND THE INDEPENDENCE OF EVENTS

---

The idea of assigning probabilities for one event conditional on the value of another is essential to understand for statistics. For the conditional assignment of probabilities of events:

For event  $A$  and event  $B$ , provided event  $B$  has nonzero probability, the conditional probability of  $A$  given  $B$  is

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

We read  $P(A|B)$  as “the probability of  $A$  given  $B$ ”.

Often, event  $A$  and event  $B$  are dependent on each other. This means that conditional probabilities apply and the numerical values of  $P(A|B)$  and  $P(A)$  are different. The difference can be thought of as reflecting the change in one’s assessed likelihood of occurrence of  $A$  brought about by knowing that  $B$ ’s occurrence is certain. In cases where there is no difference, the terminology of independence is used.

Two events  $A$  and  $B$  are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two rolls of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. Two events are independent if one of the following are true:

If  $A$  and  $B$  are events with non-zero probability in the sample space  $S$ , and are independent, then the following are equivalent:

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

The probabilities of events obey rules that lead from the application of the axioms of probability and the application of independence, and can be shown as:

If  $A$  and  $B$  are events in sample space  $S$ :

- For any event  $A$ ,  $P(A) = 1 - P(\overline{A})$ .
- The additive rule states that, for any two events A and B:  $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- For *disjoint* events, the additive rule simplifies; for any two mutually exclusive events A and B:  $P(A \text{ or } B) = P(A) + P(B)$
- The multiplication rule states that, for  $P(B) > 0$ ,  $P(A \text{ and } B) = P(A | B) \cdot P(B)$ .
- For *independent* events, the multiplication rule simplifies; for any independent events A and B:  $P(A \text{ and } B) = P(A) \cdot P(B)$ .

We can now extend the definition of independence to mutual independence of multiple events. The independence of more than two events extends the understanding of independence: knowing something about some of these events gives no probabilistic information about the others. Mutual independence extends for all collections of events within the sample space. This ideas of mutually independence will become very important for assigning probabilities to the events of random experiements.

A collection of events  $A_1, A_2, \dots, A_n \subset S$  are mutually independent if for any sub-collection  $A_{i_1}, \dots, A_{i_k}$  there is:

$$\bullet P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1}) \cdot P(A_{i_2}) \cdot \dots \cdot P(A_{i_k})$$

## RANDOM SAMPLING AND INDEPENDENCE

---

Sampling may be done **with replacement** or **without replacement**.

- **With replacement:** If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.
- **Without replacement:** When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are



affected by the result of the first pick. The events are considered to be dependent or not independent.

### 3.1.2.1. Sampling from a well-shuffled deck

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

1. **Sampling with replacement:** Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.
2. **Sampling without replacement:** Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice.

### 3.1.3 Random Variables and Probability Distributions

#### RANDOMNESS AND VARIATION

---

We have discussed randomness as representing the fundamental element of chance, such as in flipping a coin, but it may also represent uncertainty, such as in measurement error. We have introduced the concept of random events and experiments in the previous chapter, and let us now also think of an experiment as taking an measurement from an engineering experiment as the numerical outcome. Data measures will generally have some chance involved, and will be subject to chance influences. In statistical sampling and frequency studies, chance is introduced by sampling techniques. Chance is also introduced through measurement error. Other sources of chance may be many small, unnameable causes that work to produce the measurement that is the observation taken from the chance phenomena. In analytical contexts, changes in system conditions work to make measured responses vary, and this is most often attributed to chance.

No matter how carefully an experiment is designed and conducted, variations often occur due to these chance phenomena. The goal is therefore to understand, quantify, and model variation, and to harness this variation into our analyses in order to make conclusions based on the data that is not invalidated by the variation.

#### RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

---

A **random variable** is a mathematical formalization, or function, of an event which is dependent on an underlying random experiment. It is a real-valued variable that assigns a numerical value to each possible outcome of the experiment.

In most cases, a random variable  $X$  is a function from the sample space (a probability measure space) to the real numbers (of a measurable space):

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

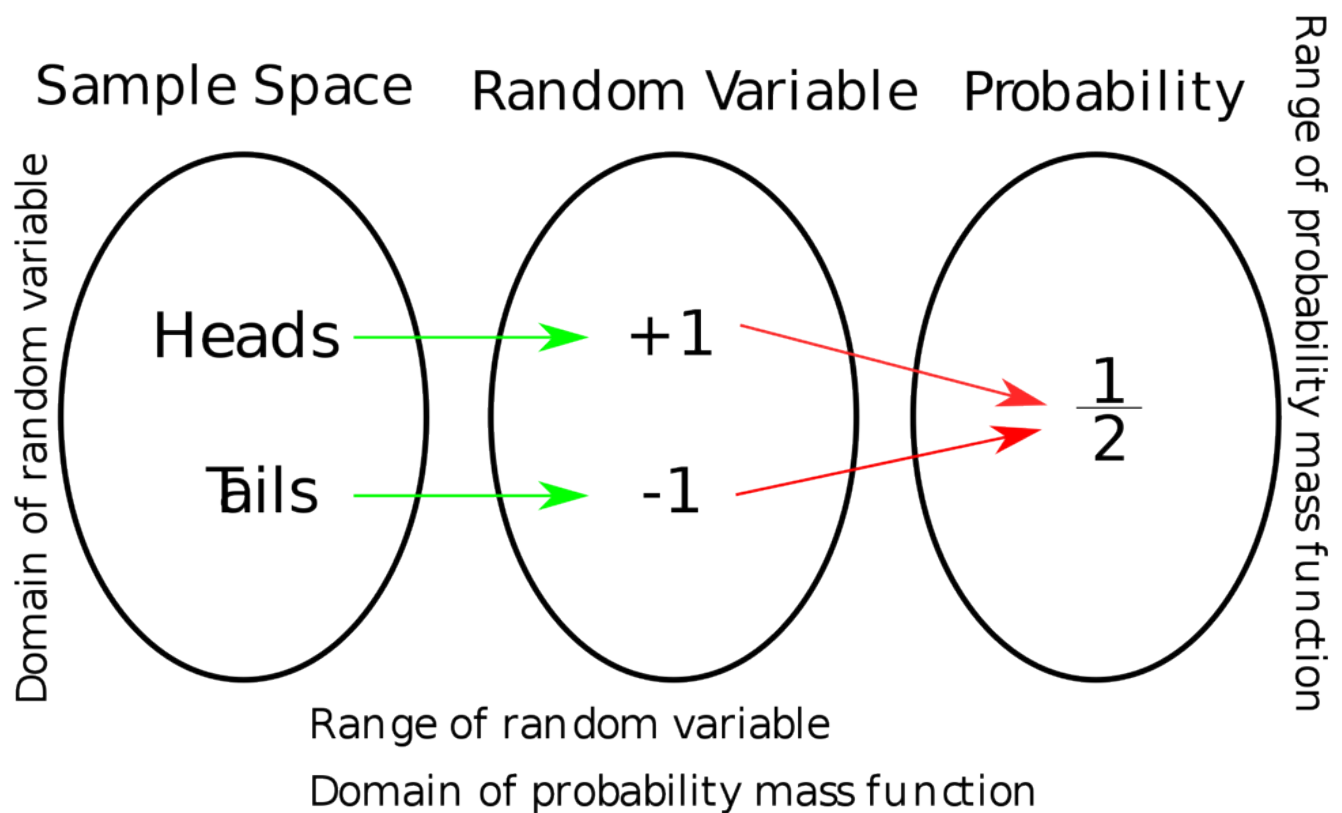
With this assignment from the sample space to real numbers, we can create a mathematical distribution of a random variable, based on our probability axioms and  $[0, 1] \subseteq \mathbb{R}$ , which provides the probability

measure on the set of all possible values of the random variable. Random variables are shown as Roman capital letters, often towards the end of the alphabet, such as  $X, Y, Z, T$ .

For the simple example that we have been using about flipping a fair coin, the function assigns values from the possible outcomes into a sample space set of  $S = \{H, T\}$  to a measurable space of  $\{-1, 1\}$ , where 1 is correspondent to H and -1 is correspondent to T, utilizing a random variable of  $X$  to represent the chance measurement of the experiment of flipping the coin.

Once we have defined the sample space of  $S$  by correspondent random variable  $X$ , we can now ask: "How likely is it that the value of  $X$  is equal to +1?". This is the probability of the event  $E = x = +1$ , written as  $P(X = 1)$ .

Recording all of the probabilities of the outputs of a random variable  $X$  will provide the probability distribution of  $X$ . A **probability distribution** is the mathematical function that defines the probabilities of occurrence of the event, or the defined subset of the sample space, and therefore defines the random experiment in terms of the event.



*Figure 3.1.3.1. A random variable is a function from all possible outcomes of a random experiment to real values. This figure shows how the outcome of flipping a coin is shown as a discrete random variable that is used for defining a probability mass function.*

For our coin example, if  $X$  is the random variable used to define the chance outcome of the experiment of

flipping the fair coin, then the probability distribution of  $X$  would take the value 0.5 (or 1/2) for  $X$ = Heads, and 0.5 for  $X$ = Tails.

---

#### Key Takeaways

Review of terms for random variables and probability distributions:

- Random variable: from values taken from a sample space, assigns probabilities based on how likely the experimental event is.
- Event: set of possible values (oucomes) of a random variable that occurs with a certain probability,
- Probability distribution: a function that provides the probability of occurrence of events for the experiement, or  $P(X \in E)$  for an event.

## 3.1.4 Cumulative Distribution Functions

### CUMULATIVE DISTRIBUTION FUNCTION

---

Probability distributions can be defined in different ways depending on how we will describe the random variable used, but can always be defined by a cumulative distribution function or CDF. This describes the probability that the random variable is no larger than a given value, or  $P(X \leq x)$ .

Every probability distribution supported on real numbers is defined by a right-continuous, non-decreasing function  $F: \mathbb{R} \rightarrow [0, 1]$ , where  $\lim_{x \rightarrow -\infty} F(x) = 0$  and  $\lim_{x \rightarrow \infty} F(x) = 1$ . Every function with these four properties is a CDF: for every such function, a random variable can be defined such that the function is the cumulative distribution function of that random variable.

#### Definition 3.1.4.1. Cumulative Distribution Function (CDF)

The cumulative probability function for a random variable  $X$  is a function  $F(x)$  that for each number  $x$  gives the probability that  $X$  takes that value or a smaller one. In symbols,

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

## 3.1.5 Discrete Random Variables and Continuous Random Variables

### DISCRETE RANDOM VARIABLES

---

We have already made a distinction between discrete and continuous data types when we explored data and descriptive statistics in Module 1. That terminology carries over to the present context and inspires two more definitions.

There are two types of random variables:

- A discrete random variable is one that has isolated or separated possible values (rather than a continuum of available outcomes).
- A continuous random variable is one that can be idealized as having an entire (continuous) interval of numbers as its set of possible values.

Random variables that are basically count variables clearly fall under the first definition and are discrete. It could be argued that all measurement variables are discrete—on the basis that all measurements are “to the nearest unit”, but for practical purposes we will continue with the definitions of data type and treat numerical values as continuous. We will learn about continuous probability distributions in the next module.

Remember that we use the notational convention that a capital P followed by an expression or phrase enclosed by parentheses or brackets will be read “the probability” of that expression. In these terms, a probability function for X the outcome of flipping a fair coin, which according to our definition is a discrete random variable, is a function f such that

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

That is, " $f(x)$  is the probability that (the random variable)  $X$  takes the value  $x$ " and is = 0.5 at the event of  $x$ =Heads or  $x$ =Tails.

## 3.1.6 Summary of Probability Models

### PROBABILITY MODELS

---

As we have learned previously, random variables serve as a fundamental tool for quantifying and managing the uncertainty inherent in various random processes or experiments. Probabilities for a random variable are usually determined from a model that describes the random experiment. Key to this understanding are the concepts of expected value, variance, and standard deviation, which respectively represent the average outcome, the variation, and the measure of dispersion of a random variable's potential values. This is the probability distribution of a random variable and is a description of the probabilities associated with the possible values of the random variable. These probability distributions are critical in engineering for modeling, predicting, and controlling system behaviors, enabling engineers to make informed decisions under conditions of uncertainty and risk.

#### Key Takeaways

The probability distribution of a random variable is a description of the probabilities associated with the possible values of the random variable.

A probability distribution is a mathematical description of the probabilities of events, subsets of the possible outcomes of the experiment. In simple terms, a probability distribution function is a theoretical model or pattern that you try to find so that you can use it to find your best "guess" or probability for.

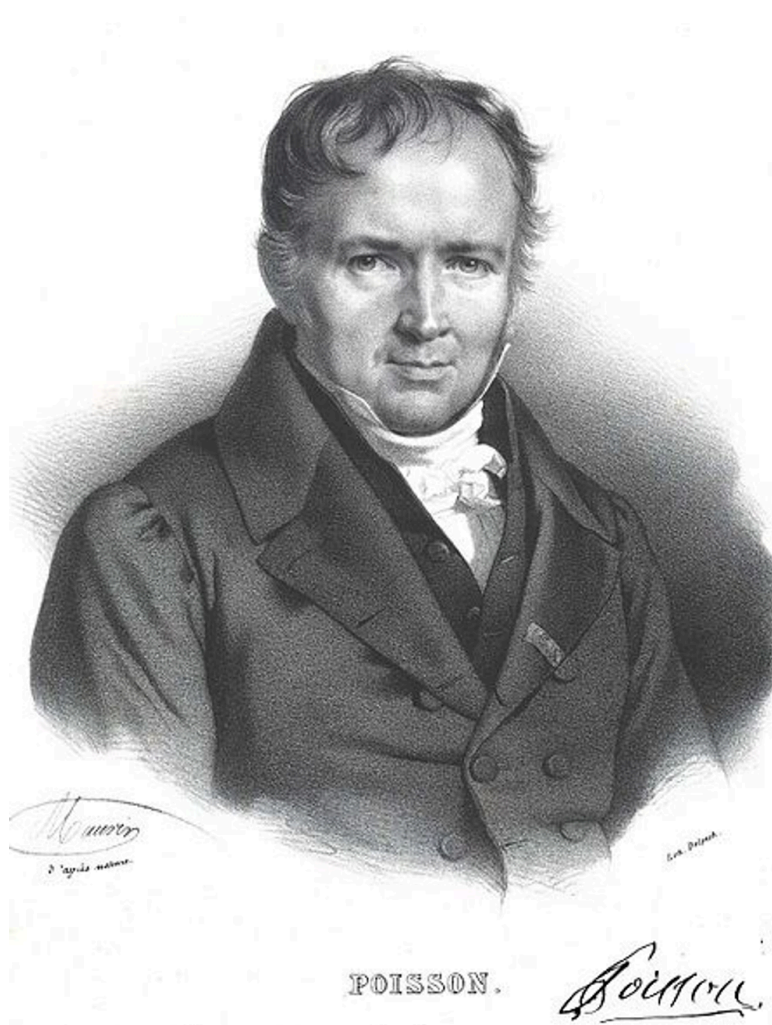
#### Key Takeaways

- Probability distributions are theoretical models or tools to make solving probability problems easier.

These probability distributions are theoretical models or tools to make solving probability problems easier. Each distribution has its own special assumptions, characteristics, and parameters. Learning these enables you to distinguish among the different distributions and choose the best model to use. By recognizing the probability distribution of an identified random variable, we are able to characterize and harness chance and variability in order to decide on probabilities, or on how likely an event is to occur. This provides us with the tools to enable the "best guess" of future, unknown experimental outcomes by choosing the most probable event. This "best guess" will lead us to be able to form predictions based on choosing a model and analyzing sample data.



## 3.2.0 Introduction to Discrete Probability Distributions



24

Zweites Kapitel. § 12.

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
I	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
II	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
III	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
IV	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
V	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
VI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
VII	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
VIII	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
IX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
X	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
XI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
XII	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
XIII	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
XIV	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
XV	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Figure 3.2.0.1. Siméon Poisson: François-Séraphin Delpech, Public domain, 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, Germany: B.G. Teubner, 1898). Bortkiewicz presents the Poisson distribution. On [pages 23–2](#)

The Poisson distribution, named after the French mathematician Siméon Denis Poisson, born in 1781, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, under the assumption that these events occur with a known constant mean

rate and independently of the time since the last event, Figure 3.2.0.1. A famous historical application of the Poisson distribution is its use in analyzing the incidence of deaths from horse kicks in the Prussian cavalry. This example is often cited to illustrate the power and utility of the Poisson distribution in modeling rare, random events in various domains.

See the GitHub Jupyter Notebook Activity for illustrating using the Poisson distribution to model horse kick mortality in the Prussian cavalry: [Poisson Distribution and the Prussian Cavalry](#).

## DISCRETE RANDOM VARIABLES

---

As we have learned, for a discrete random variable, it is sufficient to specify a probability mass function assigning a probability to each possible outcome or events. For the probability mass function, a cumulative distribution function evaluating the probability of the random variable taking on a value less than or equal to a value is defined.

### Key Takeaways

- For discrete random variables, a probability mass function defines the probability of an event from a random experiment.

These probability distributions are theoretical models or tools to make solving probability problems easier. Each distribution has its own special assumptions, characteristics, and parameters. Learning these enables you to distinguish among the different distributions and choose the best model to use. Some of the more common discrete probability functions are binomial, geometric, hypergeometric, and Poisson.

### Learning Objectives

#### Learning Objectives for Module 3.2:

- Recognize and apply discrete random variables to empirical and theoretical probabilities.
- Recognize and understand discrete probability distribution functions and their assumptions.
- Calculate and interpret expected values and distribution parameters of probability mass function.
- Understand the cumulative distribution function and apply it to calculations.
- Recognize the binomial probability distribution and apply it appropriately.
- Recognize the Poisson probability distribution and apply it appropriately.

- Recognize the geometric probability distribution and apply it appropriately.
- Recognize the hypergeometric probability distribution and apply it appropriately.

## 3.2.1 Probability Mass Function (PMF) for a Discrete Random Variable

### DISCRETE RANDOM VARIABLE

---

Let us review the definition of a **discrete random variable** that we learned about in the previous module:

A discrete random variable is one that has isolated or separated possible values (rather than a continuum of available outcomes).

Remember that a random variable is unpredictable and not known prior to a random experiment. Therefore, in describing or modeling it, the important thing is to specify its set of potential values and the likelihoods associated with those possible values.

#### **DEFINITION 3.2.1.1. Probability Distribution**

To specify a probability distribution for a random variable is to give the set of possible values and (in one way or another) consistently assign numbers between 0 and 1—called probabilities—as measures of the likelihood that the various numerical values will occur.

The tool most often used to describe a discrete probability distribution is the probability mass function.

**DEFINITION 3.2.1.2. Probability Mass Function**

A probability function for a discrete random variable  $X$ , having possible values  $x_1, x_2, \dots$ , is a non-negative function  $f(x)$ , with  $f(x_i)$  giving the probability that  $X$  takes the value  $x_i$ .

Remember that  $P(X)$  or  $P[X]$  is the probability of the expression or phrase  $X$ . Therefore the probability function (probability mass function) for  $X$  is the function  $f$  such that:

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

That is, “ $f(x)$  is the probability that (the random variable)  $X$  takes the value  $x$ .”

**Example 3.2.1.1. Revisiting bolt torques****A Torque Requirement Random Variable**

Consider again the example in Chapter 2, where Brenny, Christensen, and Schneider measured bolt torques on the face plates of a heavy equipment component. If we state that:

$Z$  = the next measured torque for bolt 3 (recorded to the nearest integer), and we will treat  $Z$  as a discrete random variable. Now we want to give a plausible probability function for it. The relative frequencies for the bolt 3 torque measurements recorded introduce the relative frequency distribution:

$z$ , Torque (ft lb)	Frequency	Relative Frequency
11	1	$1/34 \approx .02941$
12	1	$1/34 \approx .02941$
13	1	$1/34 \approx .02941$
14	2	$2/34 \approx .05882$
15	9	$9/34 \approx .26471$
16	3	$3/34 \approx .08824$
17	4	$4/34 \approx .11765$
18	7	$7/34 \approx .20588$
19	5	$5/34 \approx .14706$
20	1	$1/34 \approx .02941$
	34	1

Table 3.2.1.1.

This table shows, for example, that over the period the researchers were collecting data, about 15% of measured torques were 19 ft lb. If it is sensible to believe that the same system of causes that produced the data in this table will operate to produce the next bolt torque, then it also makes sense to base a probability function for  $Z$  on the relative frequencies in this table.

That is, the probability distribution specified in this next table might be used. (In going from the relative frequencies in the first table to proposed values for  $f(z)$  in the second table, there has been some slightly arbitrary rounding. This has been done so that probability values are expressed to two decimal places and now total to exactly 1.00.)

A Probability Function for  $Z$

Torque $z$	Probability $f(z)$
11	.03
12	.03
13	.03
14	.06
15	.26
16	.09
17	.12
18	.20
19	.15
20	.03

### *The probability mass distribution of a single value selected at random from a population*

The appropriateness of the probability function in the above table for describing  $Z$  depends essentially on the physical stability of the bolt-tightening process. But there is a second way in which relative frequencies can become obvious choices for probabilities. [Table 3.2.1.2.](#) For example, think of treating the 34 torques represented in the Table 3.2.1.1 as a population, from which  $n = 1$  item is to be sampled at random, and,  $Y$  = the torque value selected.

Then the probability function in the Table 3.2.1.2 is also approximately appropriate for  $Y$ . This point is not so important in this specific example as it is in general: Where one value is to be selected at random from a population, an appropriate probability distribution is one that is equivalent to the population relative frequency distribution.

#### Key Takeaways

The **probability distribution** for a random variable lists all the possible values of the random variable and the probability the random variable takes on each value. It describes how probabilities are distributed over the values of the random variable. If one value is to be selected at random from a population, an appropriate probability distribution is one that is equivalent to the population relative frequency distribution.

### *Properties of a mathematically valid probability function*

The probability function shown in Table 3.2.1.2 has two properties that are necessary for the mathematical consistency of a discrete probability distribution. The  $f(z)$  values are each in the interval  $[0, 1]$  and they total to 1. Negative probabilities or ones larger than 1 would make no practical sense. A probability of 1 is taken as indicating certainty of occurrence and a probability of 0 as indicating certainty of non-occurrence. Thus, according to the model specified in Table 3.2.1.2, since the values of  $f(z)$  sum to 1, the occurrence of one of the values 11, 12, 13, 14, 15, 16, 17, 18, 19, and 20 ft lb is certain.

A probability function  $f(x)$  gives probabilities of occurrence for individual values. Adding the appropriate values gives probabilities associated with the occurrence of one of a specified type of value for  $X$ .

#### Example 3.2.1.2. Revisiting bolt torques, continued

Consider using  $f(z)$  defined in Table 3B.1.2 to find:

$$P[Z > 17] = P[\text{the next torque exceeds 17}]$$

Adding the  $f(z)$  entries corresponding to possible values larger than 17 ft lb,

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

The likelihood of the next torque being more than 17 ft lb is about 38%.

If, for example, specifications for torques were 16 ft lb to 21 ft lb, then the likelihood that the next torque measured will be within specifications is:

$$\begin{aligned} P[16 \leq Z \leq 21] &= f(16) + f(17) + f(18) + f(19) + f(20) + f(21) \\ &= .09 + .12 + .20 + .15 + .03 + .00 \\ &= .59 \end{aligned}$$

In the torque measurement example, the probability function is given in tabular form. In other cases, it is possible to give a formula for  $f(x)$ .

### Example 3.2.1.3. A Random Tool Serial Number

The last step of the pneumatic tool assembly process studied by Kraber, Rucker, and Williams was to apply a serial number plate to the completed tool. Imagine going to the end of the assembly line at exactly 9:00 A.M. next Monday and observing the number plate first applied after 9:00.

Suppose that

$W$  = the last digit of the serial number observed

Suppose further that tool serial numbers begin with some code special to the tool model and end with consecutively assigned numbers reflecting how many tools of the particular model have been produced. The symmetry of this situation suggests that each possible value of  $W$  ( $w = 0, 1, \dots, 9$ ) is equally likely. That is, a plausible probability function for  $W$  is given by the formula

$$f(w) = \begin{cases} .1 & \text{for } w = 0, 1, 2, \dots, 9 \\ 0 & \text{otherwise} \end{cases}$$

## 3.2.2 Cumulative Distribution Function

### CUMULATIVE DISTRIBUTION FUNCTION

Another way of specifying a discrete probability distribution is sometimes used. That is to specify its cumulative distribution function (or cumulative probability function).

Remember the definition of a CDF.

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

Since (for discrete distributions) probabilities are calculated by summing values of  $f(x)$ , for a discrete distribution,

#### DEFINITION 3.2.2.1. Cumulative Distribution Function for a discrete variable X

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

The sum is over possible values less than or equal to  $x$ . In this discrete case, the graph of  $F(x)$  will be a stair-step graph with jumps located at possible values and equal in size to the probabilities associated with those possible values.

#### Example 3.2.2.1. Revisiting bolt torques

Torque Variable Example from [3.2.1](#) continued

Values of both the probability function and the cumulative probability function for the torque variable  $Z$  are given in Table 3.2.1.1. Values of  $F(z)$  for other  $z$  are also easily obtained. For example,

$$F(10.7) = P[Z \leq 10.7] = 0$$

$$F(16.3) = P[Z \leq 16.3] = P[Z \leq 16] = F(16) = .50$$

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



A graph of the cumulative probability function for  $Z$  is given in Figure 3.2.2.1. It shows the stair-step shape characteristic of cumulative probability functions for discrete distributions.

Values of the Probability Function and Cumulative Probability Function for $Z$		
$z$ , Torque	$f(z) = P[Z = z]$	$F(z) = P[Z \leq z]$
11	.03	.03
12	.03	.06
13	.03	.09
14	.06	.15
15	.26	.41
16	.09	.50
17	.12	.62
18	.20	.82
19	.15	.97
20	.03	1.00

Table 3.2.2.1. Values of the Probability Function and Cumulative Probability Function for  $Z$ .

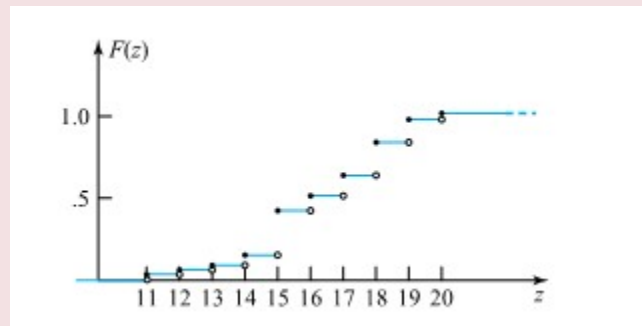


Figure 3.2.2.1. Graph of the cumulative probability function for  $Z$ .

The information about a discrete distribution carried by its cumulative probability function is equivalent to that carried by the corresponding probability function. The cumulative version is sometimes preferred for table making, because round-off problems are more severe when adding several  $f(x)$  terms than when taking the difference of two  $F(x)$  values to get a probability associated with a consecutive sequence of possible values, and because of ease of comprehension.

### 3.2.3 Probability Expressed to Two Decimal Places

#### EXPRESSING PROBABILITIES

---

We will usually express probabilities to two decimal places, such as shown in Table 3.2.1.2. Computations may be carried to several more decimal places, but final probabilities will typically be reported only to two places. This is because numbers expressed to more than two places tend to look too impressive and be taken too seriously by the uninitiated. Consider for example the statement “There is a .097328 probability of booster engine failure” at a certain missile launch. This may represent the results of some very careful mathematical manipulations and be correct to six decimal places in the context of the mathematical model used to obtain the value. But it is doubtful that the model used is a good enough description of physical reality to warrant that much apparent precision. Two-decimal precision is about what is warranted in most engineering applications of simple probability.

## 3.2.4 Mean or Expected Value and Standard Deviation of Discrete Probability Distributions

### SUMMARIZATION OF DISCRETE PROBABILITY DISTRIBUTIONS

Almost all of the devices for describing relative frequency (empirical) distributions in Modules 1 and 2 on exploring, summarizing, and visualizing data have versions that can describe (theoretical) probability distributions.

For a discrete random variable with equally spaced possible values, a probability histogram gives a picture of the shape of the variable's distribution. It is made by centering a bar of height  $f(x)$  over each possible value  $x$ . Probability histograms for the random variables  $Z$  and  $W$  in Examples 3.2.1 are given in Figure 3B.4.1. Interpreting such probability histograms is similar to interpreting relative frequency histograms, except that the areas on them represent (theoretical) probabilities instead of (empirical) fractions of data sets.

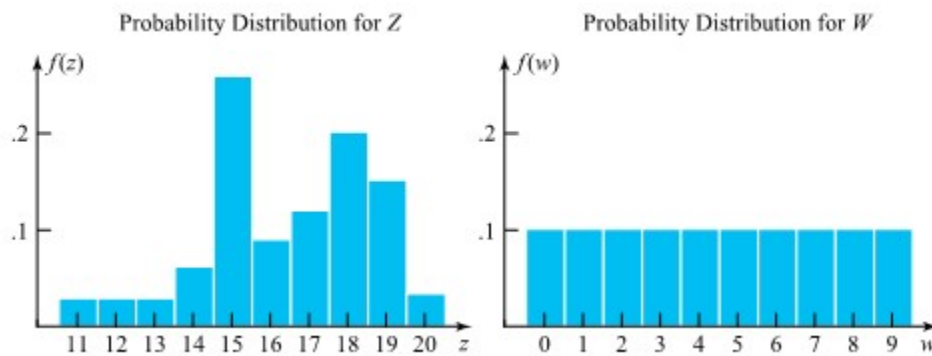


Figure 3.2.4.1. Probability histograms for  $Z$  and  $W$  (Examples 3.2.1.1 and 3.2.1.2)

It is useful to have a notion of mean value for a discrete random variable (or its probability distribution).

**DEFINITION 3.2.4.1. The mean of a discrete random variable**

The mean or expected value of a discrete random variable  $X$  (sometimes called the mean of its probability distribution) is

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

$EX$  is read as “the expected value of  $X$ ,” and sometimes the notation  $\mu$  is used in place of  $EX$ .

Remember that  $\mu$  stands for both the mean of a population and the mean of a probability distribution, as we discussed with empirical distributions.

#### Example 3.2.4.1. Bolt Torque Example, continued.

Returning to the bolt torque example, the expected (or theoretical mean) value of the next torque is

$$\begin{aligned} EZ &= \sum_z z f(z) \\ &= 11(.03) + 12(.03) + 13(.03) + 14(.06) + 15(.26) + 16(.09) + 17(.12) + 18(.20) + 19(.15) + 20(.03) \\ &= 16.35 \text{ ft lb} \end{aligned}$$

This value is essentially the arithmetic mean of the bolt 3 torques listed previously. This kind of agreement provides motivation for using the symbol  $\mu$ , first seen in Module 2, as an alternative to  $EZ$ .

The mean of a discrete probability distribution has a balance point interpretation, much like that associated with the arithmetic mean of a data set. Placing (point) masses of sizes  $f(x)$  at points  $x$  along a number line,  $EX$  is the center of mass of that distribution.

#### Example 3.2.4.2. Serial Number Example continued.

Considering again the serial number example, and the second part of Figure 3.2.4.1, if a balance point interpretation of expected value is to hold,  $EW$  had better turn out to be 4.5. And indeed

$$EW = 0(.1) + 1(.1) + 2(.1) + \cdots + 8(.1) + 9(.1) = 45(.1) = 4.5$$

It was convenient to measure the spread of a data set (or its relative frequency distribution) with the variance and standard deviation. It is similarly useful to have notions of spread for a discrete probability distribution.

**DEFINITION 3.2.4.2. Variance of discrete random variable X**

The variance of a discrete random variable X (or the variance of its distribution) is

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

The standard deviation of X is  $\sqrt{\text{Var } X}$ . Often the notation  $\sigma^2$  is used in place of Var X, and  $\sigma$  is used in place of  $\sqrt{\text{Var } X}$ .

The variance of a random variable is its expected (or mean) squared distance from the center of its probability distribution. The use of  $\sigma^2$  to stand for both the variance of a population and the variance of a probability distribution is motivated on the same grounds as the double use of  $\mu$ .

**Example 3.2.4.3. Bolt Torque Example, continued**

The calculations necessary to produce the bolt torque standard deviation are organized in Table 3.2.4.1. So

$$\sigma = \sqrt{\text{Var } Z} = \sqrt{4.6275} = 2.15 \text{ ft lb}$$

Except for a small difference due to round-off associated with the creation of Table 3.2.1.2, this standard deviation of the random variable Z is numerically the same as the population standard deviation associated with the bolt 3 torques in Table 2.X. (Again, this is consistent with the equivalence between the population relative frequency distribution and the probability distribution for Z.)

Calculations for Var Z			
$z$	$f(z)$	$(z - 16.35)^2$	$(z - 16.35)^2 f(z)$
11	.03	28.6225	.8587
12	.03	18.9225	.5677
13	.03	11.2225	.3367
14	.06	5.5225	.3314
15	.26	1.8225	.4739
16	.09	.1225	.0110
17	.12	.4225	.0507
18	.20	2.7225	.5445
19	.15	7.0225	1.0534
20	.03	13.3225	.3997
			Var Z = 4.6275

Table 3.2.4.1. Calculations for Var Z

**Example 3.2.4.4. Serial Number Example, continued.**

To illustrate the alternative for calculating a variance given in Definition 3.2.4.2, consider finding the variance and standard deviation of the serial number variable  $W$ . Table 3.2.4.2 shows the calculation of  $\sum w^2 f(w)$ .

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

Table 3.2.4.2.

Then

$$\text{Var } W = \sum w^2 f(w) - (EW)^2 = 28.5 - (4.5)^2 = 8.25$$

So that

$$\sqrt{\text{Var } W} = 2.87$$

Comparing the two probability histograms in the figure previously, notice that the distribution of  $W$  appears to be more spread out than that of  $Z$ . Happily, this is reflected in the fact that

$$\sqrt{\text{Var } W} = 2.87 > 2.15 = \sqrt{\text{Var } Z}$$

### 3.2.5 Binomial Distribution

Discrete probability distributions are sometimes developed from past experience with a particular physical phenomenon (as in Example 1). On the other hand, sometimes an easily manipulated set of mathematical assumptions having the potential to describe a variety of real situations can be put together. When those can be manipulated to derive generic distributions, those distributions can be used to model a number of different random phenomena. One such set of assumptions is that of **independent, identical success-failure trials**.

Many engineering situations involve repetitions of essentially the same “go-no go” (success-failure) scenario, where:

1. There is a constant chance of a go/success outcome on each repetition of the scenario (call this probability  $p$ ).
2. The repetitions are independent in the sense that knowing the outcome of any one of them does not change assessments of chance related to any others.

Examples of this kind include the testing of items manufactured consecutively, where each will be classified as either conforming or nonconforming; observing motorists as they pass a traffic checkpoint and noting whether each is traveling at a legal speed or speeding; and measuring the performance of workers in two different workspace configurations and noting whether the performance of each is better in configuration A or configuration B.

In this context, there are two generic kinds of random variables for which deriving appropriate probability distributions is straightforward. The first is the case of a count of the repetitions out of  $n$  that yield a go/success result. That is, consider a variable:

$X$  = the number of go/success results in  $n$   
independent identical  
success-failure trials

#### Binomial random variables

*DEFINITION 3.2.5.1. Binomial Definition*  
The binomial distribution is a discrete probability distribution with probability function

$$f(x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

for  $n$  a positive integer and  $0 < p < 1$ .

Equation (3.2.5.1) is completely plausible. In it there is one factor of  $p$  for each trial producing a go/success outcome and one factor of  $(1 - p)$  for each trial producing a no go/failure outcome. And the  $n!/x!(n - x)!$  term is a count of the number of patterns in which it would be possible to see  $x$  go/success outcomes in  $n$  trials. The name *binomial* distribution derives from the fact that the values  $f(0), f(1), f(2), \dots, f(n)$  are the terms in the expansion of

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

according to the binomial theorem.

We can take the time to plot probability histograms for several different binomial distributions. It turns out that for  $p < .5$ , the resulting histogram is right-skewed. For  $p > .5$ , the resulting histogram is left-skewed. The skewness increases as  $p$  moves away from  $.5$ , and it decreases as  $n$  is increased. Four binomial probability histograms are pictured in Figure 3.2.5.1.

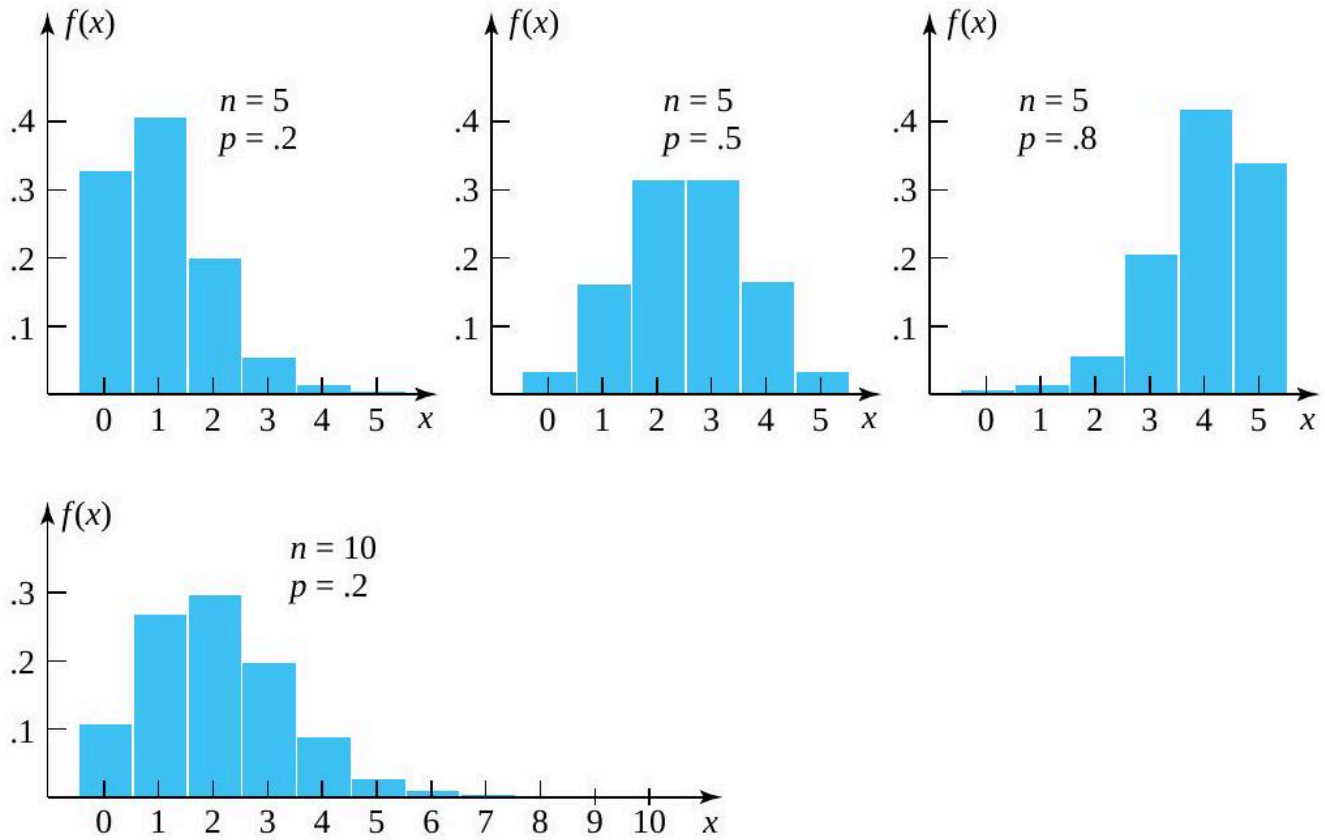


Figure 3.2.5.1. Four binomial probability histograms.

#### Example 3.2.5.1. The Binomial Distribution and Counts of Reworkable Shafts.

Consider a study of the performance of a process for turning steel shafts. Early in that study, around 20% of the shafts were typically classified as "reworkable." Suppose that  $p = .2$  is indeed a sensible figure for the chance that a given shaft will be reworkable. Suppose further that  $n = 10$  shafts will be inspected, and the probability that at least two are classified as reworkable is to be evaluated.

Adopting a model of independent, identical success-failure trials for shaft conditions,

$U =$  the number of reworkable shafts in the sample of 10

is a binomial random variable with  $n = 10$  and  $p = .2$ . So



$$\begin{aligned}
P[\text{at least two reworkable shafts}] &= P[U \geq 2] \\
&= f(2) + f(3) + \cdots + f(10) \\
&= 1 - (f(0) + f(1)) \\
&= 1 - \left( \frac{10!}{0!10!} (.2)^0 (.8)^{10} + \frac{10!}{1!9!} (.2)^1 (.8)^9 \right) \\
&= .62
\end{aligned}$$

(The trick employed here, to avoid plugging into the binomial probability function 9 times by recognizing that the  $f(u)$ 's have to sum up to 1, is a common and useful one.)

The .62 figure is only as good as the model assumptions that produced it. If an independent, identical success-failure trials description of shaft production fails to accurately portray physical reality, the .62 value is fine mathematics but possibly a poor description of what will actually happen. For instance, say that due to tool wear it is typical to see 40 shafts in specifications, then 10 reworkable shafts, a tool change, 40 shafts in specifications, and so on. In this case, the binomial distribution would be a very poor description of  $U$ , and the .62 figure largely irrelevant. (The independence-of-trials assumption would be inappropriate in this situation.)

### The binomial distribution and simple random sampling

There is one important circumstance where a model of independent, identical success-failure trials is not exactly appropriate, but a binomial distribution can still be adequate for practical purposes - that is, in describing the results of simple random sampling from a dichotomous population. Suppose a population of size  $N$  contains a fraction  $p$  of type A objects and a fraction  $(1 - p)$  of type B objects. If a simple random sample of  $n$  of these items is selected and

$X =$  the number of type A items in the sample

strictly speaking,  $x$  is not a binomial random variable. But if  $n$  is a small fraction of  $N$  (say, less than 10%), and  $p$  is not too extreme (i.e., is not close to either 0 or 1),  $X$  is approximately binomial  $(n, p)$ .

#### Examples 3.2.5.2. Simple Random Sampling from a Lot of Hexamine Pellets

In a pelletizing machine experiment, Greiner, Grimm, Larson, and Lukomski found a combination of machine settings that allowed them to produce 66 conforming pellets out of a batch of 100 pellets. Treat that batch of 100 pellets as a population of interest and consider selecting a simple random sample of size  $n = 2$  from it.

If one defines the random variable

$V =$  the number of conforming pellets in the sample of size 2

the most natural probability distribution for  $V$  is obtained as follows. Possible values for  $V$  are 0, 1, and 2.

$$f(0) = P[V = 0]$$

$= P[\text{first pellet selected is nonconforming and subsequently the second pellet is also nonconforming}]$

$$f(2) = P[V = 2]$$

$= P[\text{first pellet selected is conforming and subsequently the second pellet selected is conforming}]$

$$f(1) = 1 - (f(0) + f(2))$$

Then think, "In the long run, the first selection will yield a nonconforming pellet about 34 out of 100 times. Considering only cases where this occurs, in the long run the next selection will also yield a nonconforming pellet about 33 out of 99 times." That is, a sensible evaluation of  $f(0)$  is

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

Similarly,

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

and thus

$$f(1) = 1 - (.1133 + .4333) = 1 - .5467 = .4533$$

Now,  $V$  cannot be thought of as arising from exactly independent trials. For example, knowing that the first pellet selected was conforming would reduce most people's assessment of the chance that the second is also conforming from  $\frac{66}{100}$  to  $\frac{65}{99}$ . Nevertheless, for most practical purposes,  $V$  can be thought of as essentially binomial with  $n = 2$  and  $p = .66$ . To see this, note that

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

Here,  $n$  is a small fraction of  $N$ ,  $p$  is not too extreme, and a binomial distribution is a decent description of a variable arising from simple random sampling.

### Mean and variance of the binomial $(n, p)$ distribution

Calculation of the mean and variance for binomial random variables is greatly simplified by the fact that when the formulas from earlier in this module are used with the expression for binomial probabilities in equation (3.2.5.1), simple formulas result. For  $X$  a binomial  $(n, p)$  random variable,

#### DEFINITION 3.2.5.2. Mean of the binomial $(n, p)$ distribution

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

Further, it is the case that

**DEFINITION 3.2.5.3. Variance of the binomial (n,p) distribution**

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

**Example 3.2.5.3. Machining of steel shafts.**

Returning to the machining of steel shafts, suppose that a binomial distribution with  $n = 10$  and  $p = .2$  is appropriate as a model for

$U =$  the number of reworkable shafts in the sample of 10

Then, by formulas (3.2.5.2) and (3.2.5.3),

$$EU = (10)(.2) = 2 \text{ shafts}$$

$$\sqrt{\text{Var } U} = \sqrt{10(.2)(.8)} = 1.26 \text{ shafts}$$

### 3.2.6 Poisson Distribution

It is often important to keep track of the total number of occurrences of some relatively rare phenomenon, where the physical or time unit under observation has the potential to produce many such occurrences. A case of floor tiles has potentially many total blemishes. In a one-second interval, there are potentially a large number of messages that can arrive for routing through a switching center. And a 1 cc sample of glass potentially contains a large number of imperfections.

So probability distributions are needed to describe random counts of the number of occurrences of a relatively rare phenomenon across a specified interval of time or space. By far the most commonly used theoretical distributions in this context are the Poisson distributions.

#### DEFINITION 3.2.6.1. Poisson ( $\lambda$ ) distribution

The Poisson ( $\lambda$ ) distribution is a discrete probability distribution with probability function

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

for  $\lambda > 0$ .

The form of equation (3.2.6.1) may initially seem unappealing. But it is one that has sensible mathematical origins, is manageable, and has proved itself empirically useful in many different “rare events” circumstances. One way to arrive at equation (3.2.6.1) is to think of a very large number of independent trials (opportunities for occurrence), where the probability of success (occurrence) on any one is very small and the product of the number of trials and the success probability is  $\lambda$ . One is then led to the binomial

$\left(n, \frac{\lambda}{n}\right)$  distribution. In fact, for large  $n$ , the binomial  $\left(n, \frac{\lambda}{n}\right)$  probability function approximates the

one specified in equation (5.10). So one might think of the Poisson distribution for counts as arising through a mechanism that would present many tiny similar opportunities for independent occurrence or non-occurrence throughout an interval of time or space.

The Poisson distributions are right-skewed distributions over the values  $x = 0, 1, 2, \dots$ , whose probability histograms peak near their respective  $\lambda$ 's. Two different Poisson probability histograms are shown in Figure 3.2.6.1.

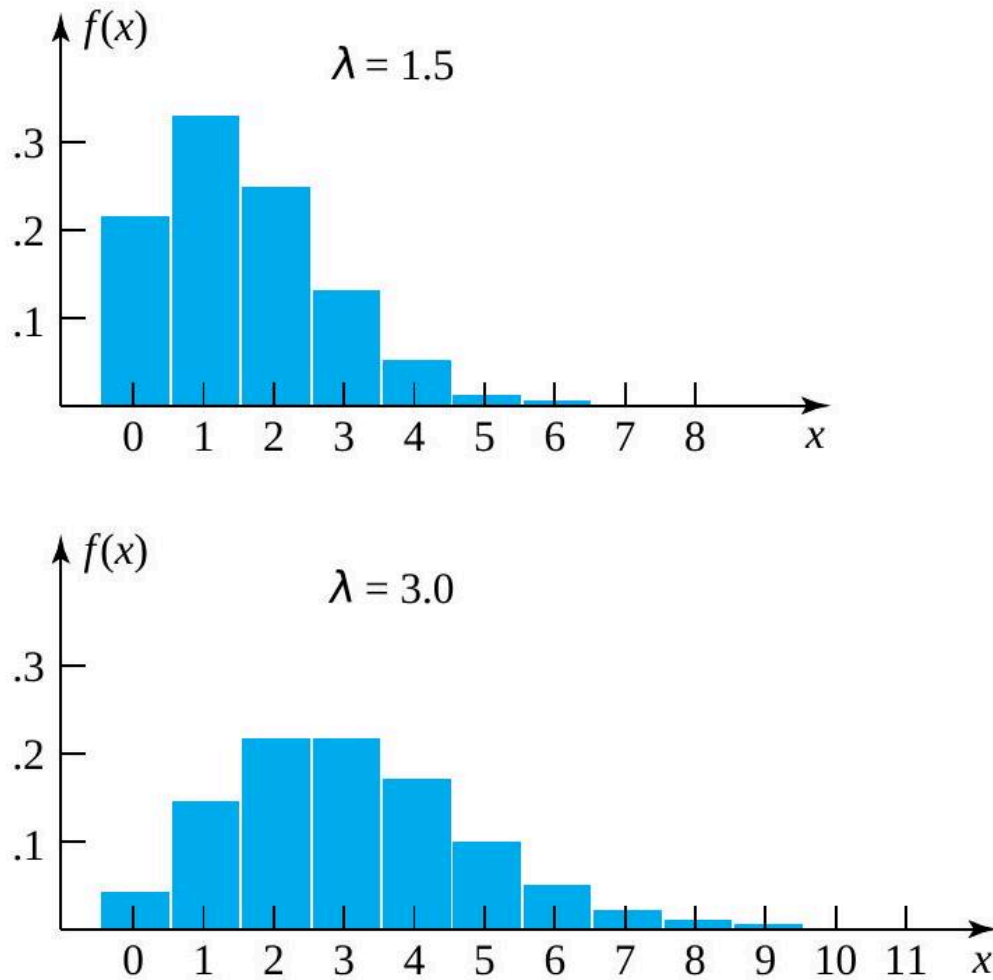


Figure 2.3.6.1. Two Poisson probability histograms.

$\lambda$  is both the mean and the variance for the Poisson ( $\lambda$ ) distribution. That is, if  $X$  has the Poisson ( $\lambda$ ) distribution, then

**DEFINITION 3.2.6.2. Mean of the Poisson ( $\lambda$ ) distribution**

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

and

**DEFINITION 3.2.6.3. Variance of the Poisson ( $\lambda$ ) distribution**

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

Fact (5.11) is helpful in picking out which Poisson distribution might be useful in describing a particular “rare events” situation.

**Example 3.5.6.1. The Poisson Distribution and Counts of  $\alpha$ -Particles**

A classical data set of Rutherford and Geiger, reported in Philosophical Magazine in 1910, concerns the numbers of  $\alpha$ -particles emitted from a small bar of polonium and colliding with a screen placed near the bar in 2,608 periods of 8 minutes each. The Rutherford and Geiger relative frequency distribution has mean 3.87 and a shape remarkably similar to that of the Poisson probability distribution with mean  $\lambda = 3.87$ .

In a duplication of the Rutherford/Geiger experiment, a reasonable probability function for describing

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

is then

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

Using such a model, one has (for example)

$$\begin{aligned} & P[\text{at least 4 particles are recorded}] \\ &= P[S \geq 4] \\ &= f(4) + f(5) + f(6) + \dots \\ &= 1 - (f(0) + f(1) + f(2) + f(3)) \\ &= 1 - \left( \frac{e^{-3.87} (3.87)^0}{0!} + \frac{e^{-3.87} (3.87)^1}{1!} + \frac{e^{-3.87} (3.87)^2}{2!} + \frac{e^{-3.87} (3.87)^3}{3!} \right) \\ &= .54 \end{aligned}$$

**Example 3.2.6.2. Arrivals at a University Library**

Stork, Wohlsdorf, and McArthur collected data on numbers of students entering the ISU library during various periods over a week's time. Their data indicate that between 12:00 and 12:10 P.M. on Monday through Wednesday, an average of around 125 students entered. Consider modeling

$M$  = the number of students entering the ISU library between 12:00 and 12:01 next Tuesday

Using a Poisson distribution to describe  $M$ , the reasonable choice of  $\lambda$  would seem to be

$$\lambda = \frac{125 \text{ students}}{10 \text{ minutes}} (1 \text{ minute}) = 12.5 \text{ students}$$

For this choice,

$$EM = \lambda = 12.5 \text{ students}$$

$$\sqrt{\text{Var } M} = \sqrt{\lambda} = \sqrt{12.5} = 3.54 \text{ students}$$

and, for example, the probability that between 10 and 15 students (inclusive) arrive at the library between 12:00 and 12:01 would be evaluated as

$$\begin{aligned} P[10 \leq M \leq 15] &= f(10) + f(11) + f(12) + f(13) + f(14) + f(15) \\ &= \frac{e^{-12.5} (12.5)^{10}}{10!} + \frac{e^{-12.5} (12.5)^{11}}{11!} + \frac{e^{-12.5} (12.5)^{12}}{12!} \\ &\quad + \frac{e^{-12.5} (12.5)^{13}}{13!} + \frac{e^{-12.5} (12.5)^{14}}{14!} + \frac{e^{-12.5} (12.5)^{15}}{15!} \\ &= .60 \end{aligned}$$

## 3.2.7 Working with Discrete Probability Distributions in Python

If you were interested in working with discrete probability distributions in python it is strongly recommended that you consult the [Normal Probability & Confidence Intervals](#) Jupyter Notebook Files. These can be found in the “How do I do X in Python?” section. Specifically the file on “Discrete Probability Distributions” will be particularly useful.



## 4.0.1 Introduction to Continuous Random Variables and Probability Distributions

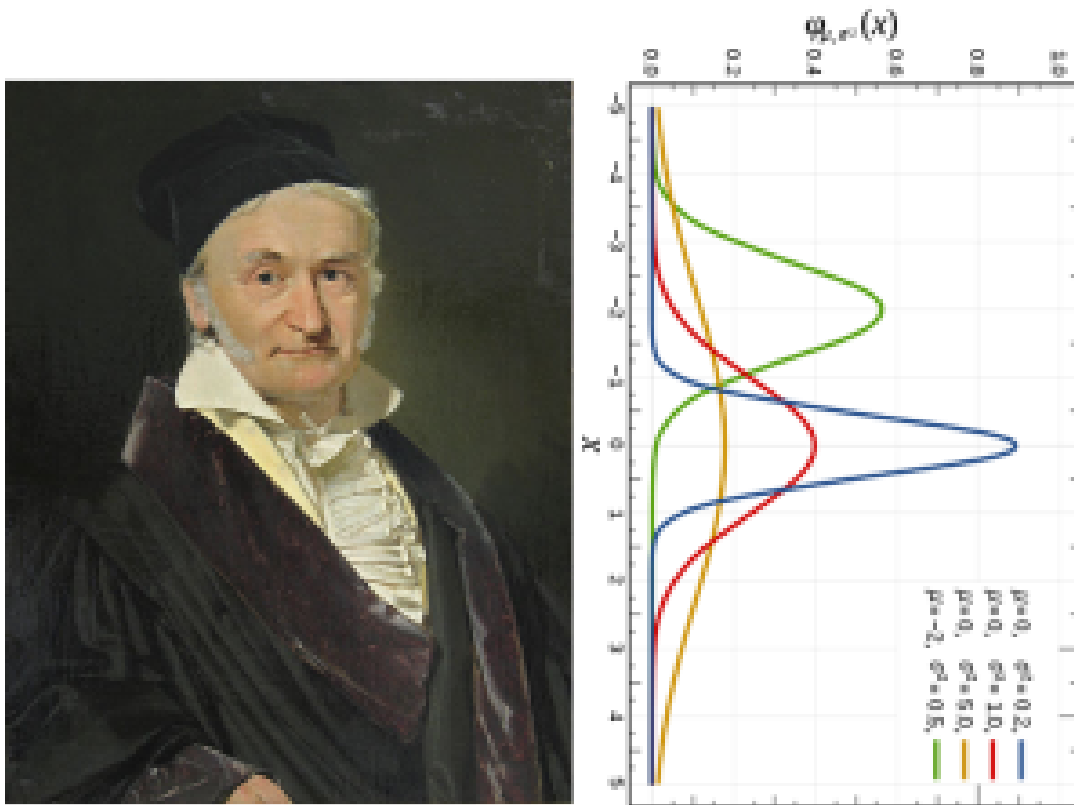


Figure 4.1.0.1. Friedrich Gauss: [https://en.wikipedia.org/wiki/Carl\\_Friedrich\\_Gauss](https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss)

Recognized as a “Prince of Mathematicians”, Carl Friedrich Gauss (born in Germany 1777-1855) holds a paramount place in the history of statistics and mathematics, Figure 4.1.0.1. Gauss made prodigious contributions across various fields, but his work in statistics and the theory of probability are notable. He is best known for developing the method of least squares and the normal distribution, also known as the Gaussian distribution or the bell curve, vital for statistical analysis in various fields, from social sciences to natural sciences to engineering. The normal distribution is a symmetric probability distribution that describes the way that a continuous random variable can be distributed. Its distinctive bell-shaped curve emerges when a dataset has a high frequency of values near the mean, with frequencies gradually decreasing as values move further away from the mean. It is ubiquitous because it naturally models many real-world phenomena and because many random processes and experiments tend to produce averaged data that follow a normal distribution. Its importance lies in its ability to provide a simple, yet powerful, framework for understanding and interpreting datasets, making it a cornerstone of statistical analysis.

## CONTINUOUS RANDOM VARIABLES

---

It is often convenient to think of a random variable as not discrete but rather continuous in the sense of having a whole (continuous) interval for its set of possible values. The devices used to describe continuous probability distributions differ from the tools studied in the last section. So the first tasks here are to introduce the notion of a probability density function, to show its relationship to the cumulative probability function for a continuous random variable, and to show how it is used to find the mean and variance for a continuous distribution. Then a couple of useful distributions will be reviewed: the uniform, the exponential, and the Weibull distributions. After this, the most important and standard continuous distribution useful in engineering applications of probability theory will be discussed: the normal distribution.

## 4.0.2 *Attributions Part 4*

This first draft of Part 4 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

## 4.1.1 Probability Density Functions and Cumulative Probability Function

### PROBABILITY DENSITY FUNCTION

The methods used to specify and describe probability distributions have parallels in mechanics. When considering continuous probability distributions, the analogy to mechanics becomes especially helpful. In mechanics, the properties of a continuous mass distribution are related to the possibly varying density of the mass across its region of location. Amounts of mass in particular regions are obtained from the density by integration.

The concept in probability theory corresponding to mass density in mechanics is probability density. To specify a continuous probability distribution, one needs to describe “how thick” the probability is in the various parts of the set of possible values. The formal definition is:

#### DEFINITION 4.1.1.1. Probability Density Function (PDF)

##### EXPRESSION 4.1.1.1.

A probability density function for a continuous random variable  $X$  is a nonnegative function  $f(x)$  with:

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

and such that for all  $a \leq b$ , one is willing to assign  $P[a \leq X \leq b]$  according to:

##### EXPRESSION 4.1.1.2.

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

A generic probability density function (PDF) is pictured in Figure 4.1.1.1. As can be seen, the graph of a continuous probability distribution is a curve. In keeping with equations for the definition of the PDF, the plot of  $f(x)$  does not dip below the  $x$ -axis, the total area under the curve  $y=f(x)$  is 1, and areas under the curve above particular intervals give probabilities corresponding to those intervals. We define the function  $f(x)$  so that the area between it and the  $x$ -axis is equal to a probability. Since the maximum probability is one, the maximum area is also one.

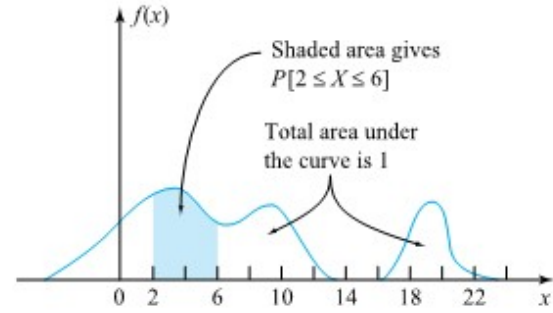


Figure 4.1.1.1. A generic probability density function.

The curve is the pdf. We use the symbol  $f(x)$  to represent the curve.  $f(x)$  is the function that corresponds to the graph; we use the density function  $f(x)$  to draw the graph of the probability distribution. The area under the curve represents the probability.

#### Continued mechanics analogy for probability density

In direct analogy to what is done in mechanics, if  $f(x)$  is indeed the “density of probability” around  $x$ , then the probability in an interval of small length  $dx$  around  $x$  is approximately  $f(x) dx$ . (In mechanics, if  $f(x)$  is mass density around  $x$ , then the mass in an interval of small length  $dx$  around  $x$  is approximately  $f(x) dx$ .) Then to get a probability between  $a$  and  $b$ , one needs to sum up such  $f(x) dx$  values.  $\int_a^b f(x) dx$  is exactly the

limit of  $\sum f(x) dx$  values as  $dx$  gets small. (In mechanics,  $\int_a^b f(x) dx$  is the mass between  $a$  and  $b$ .) So the expression in the definition for the PDF and expression 4.1.1.2 is reasonable.

#### For $X$ a continuous random variable, $P(X = a) = 0$

One point about continuous probability distributions that may at first seem counterintuitive concerns the probability associated with a continuous random variable assuming a particular prespecified value (say,  $a$ ). Just as the mass that a continuous mass distribution places at a single point is 0, so also is  $P(X = a) = 0$  for a continuous random variable  $X$ . This follows from expression 4.1.1.2, because:

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

One consequence of this mathematical curiosity is that when working with continuous random variables, you don't need to worry about whether or not inequality signs you write are strict inequality signs. That is, if  $X$  is continuous:

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

## CUMULATIVE DISTRIBUTION FUNCTION

---

Previously we gave a perfectly general definition of the cumulative distribution function for a random variable and this was specialized in the case of a discrete variable. Now, equation 4.1.1.2 can be used to express the cumulative distribution function for a continuous random variable in terms of an integral of its probability density. That is, for  $X$  continuous with probability density  $f(x)$ :

### DEFINITION 4.1.1.3. Cumulative Distribution Function (CDF) for a Continuous Variable

#### EXPRESSION 4.1.1.3

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

$F(x)$  is obtained from  $f(x)$  by integration, and applying the fundamental theorem of calculus to equation (4.1.1.3):

### Another relationship between $F(x)$ and $f(x)$

#### EXPRESSION 4.1.1.4

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

That is,  $f(x)$  is obtained from  $F(x)$  by differentiation.

The area under the curve of the pdf is given by this different function of the cdf. The cumulative distribution function is used to evaluate probability, and can be found by using geometry, by formulas, by statistical technology, or probability tables.

### *Continuous Probability Distributions*

---

There are many continuous probability distributions. When using a continuous probability distribution to model probability, the distribution used is selected to model and fit the particular situation in the best way. In this module, we will study the uniform distribution, the exponential distribution, and the Weibull distribution, and then focus on the most important distribution for introductory statistics: the normal distribution.

### Property Review of Continuous Distributions

The probability density function (pdf) is used to describe probabilities for continuous random variables. The area under the density curve between two points corresponds to the probability that the variable falls between those two values. In other words, the area under the density curve between points  $a$  and  $b$  is equal to  $P(a < x < b)$ . The cumulative distribution function (cdf) gives the probability as an area. If  $X$  is a continuous random variable, the probability density function (pdf),  $f(x)$ , is used to draw the graph of the probability distribution. The total area under the graph of  $f(x)$  is one. The area under the graph of  $f(x)$  and between values  $a$  and  $b$  gives the probability  $P(a < x < b)$ . This is shown in Figure 4.1.1.2.

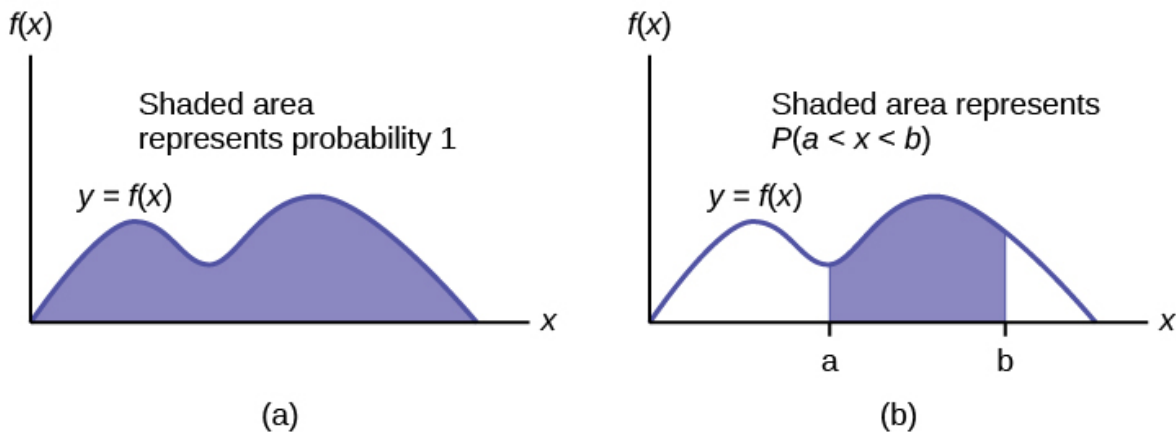


Figure 4.1.1.2. The graph on the left shows a general density curve,  $y = f(x)$ . The region under the curve and above the  $x$ -axis is shaded. The area of the shaded region is equal to 1. This shows that all possible outcomes are represented by the curve. The graph on the right shows the same density curve. Vertical lines  $x = a$  and  $x = b$  extend from the axis to the curve, and the area between the lines is shaded. The area of the shaded region represents the probability that a value  $x$  falls between  $a$  and  $b$ .

The cumulative distribution function (cdf) of  $X$  is defined by  $P(X \leq x)$ . It is a function of  $x$  that gives the probability that the random variable is less than or equal to  $x$ .

- The outcomes are measured, not counted.
- The entire area under the curve and above the  $x$ -axis is equal to one.
- Probability is found for intervals of  $x$  values rather than for individual  $x$  values.
- $P(c < x < d)$  is the probability that the random variable  $X$  is in the interval between the values  $c$  and  $d$ .  $P(c < x < d)$  is the area under the curve, above the  $x$ -axis, to the right of  $c$  and the left of  $d$ .
- $P(x = c) = 0$ . The probability that  $x$  takes on any single individual value is zero. The area below the curve, above the  $x$ -axis, and between  $x = c$  and  $x = c$  has no width, and therefore no area (area = 0). Since the probability is equal to the area, the probability is also zero.
- $P(c < x < d)$  is the same as  $P(c \leq x \leq d)$  because probability is equal to area.

## 4.1.2 Means and Variances for Continuous Distributions

### MEANS AND VARIANCES FOR CONTINUOUS DISTRIBUTIONS

A plot of the probability density  $f(x)$  is a kind of idealized histogram. It has the same kind of visual interpretations that have already been applied to relative frequency histograms and probability histograms. Further, it is possible to define a mean and variance for a continuous probability distribution. These numerical summaries are used in the same way that means and variances are used to describe data sets and discrete probability distributions.

#### DEFINITION 4.1.2.1. Mean of Continuous Random Variable $X$

##### EXPRESSION 4.1.2.1.

The mean or expected value of a continuous random variable  $X$  (sometimes called the mean of its probability distribution) is:

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

As for discrete random variables, the notation  $\mu$  is sometimes used in place of  $EX$ .

Formula 4.1.2.1 is perfectly plausible from at least two perspectives. First, the probability in a small interval around  $x$  of length  $dx$  is approximately  $f(x) dx$ . So multiplying this by  $x$  and summing, one has  $\sum xf(x) dx$ , and formula 4.1.2.1 is exactly the limit of such sums as  $dx$  gets small. And second, in mechanics the center of mass of a continuous mass distribution is of the form given in equation 4.1.2.1 except for division by a total mass, which for a probability distribution is 1.

“Continuization” of the formula for the variance of a discrete random variable produces a definition of the variance of a continuous random variable.



**DEFINITION 4.1.2.2. Variance of Continuous Random Variable X****EXPRESSION 4.1.2.2.**

The variance of a continuous random variable X (sometimes called the variance of its probability distribution) is:

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

The standard deviation of X is  $\sqrt{\text{Var } X}$ . Often the notation  $\sigma^2$  is used in place of  $\text{Var } X$ , and  $\sigma$  is used in place of  $\sqrt{\text{Var } X}$ .

## 4.1.3 Normal Probability Distribution

### NORMAL PROBABILITY DISTRIBUTION

---

Though there are a number of continuous distributions commonly applied to engineering problems, the normal distribution is of unique importance. Formally, the normal distribution is:

#### DEFINITION 4.1.3.1. The Normal Distribution

#### EXPRESSION 4.1.3.1

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

for all  $x$  and for  $\sigma > 0$ .

It is not necessarily obvious, but formula (4.1.3.1) does yield a legitimate probability density, in that the total area under the curve  $y = f(x)$  is 1. Further, it is also the case that:

Normal Distribution Mean and Variance

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

and

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

### Parameters of the Normal Distribution

The normal distribution has two parameters (two numerical descriptive measures of the theoretical distribution), the mean  $\mu$  and the variance  $\sigma^2$  (remember that the standard deviation =  $\sqrt{\sigma^2} = \sigma$ ). Figure 4.1.3.1 shows the notation for the standard normal distribution, and that the distribution shape depends on these parameters. Since the area under the curve must equal one, a change in the standard deviation,  $\sigma$ , causes a change in the shape of the curve; the curve becomes fatter or skinnier depending on  $\sigma$ . A change in  $\mu$  causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions.

The parameters  $\mu$  and  $\sigma^2$  used in Definition (4.1.3.1) are, respectively, the mean and variance (as defined in Definitions 4.1.2.1 and 4.1.2.2) of the distribution. Figure 4.1.3.2 is a graph of the probability density specified by formula (4.1.3.1). The bell-shaped curve shown there is symmetric about  $x = \mu$  and has inflection points at  $\mu - \sigma$  and  $\mu + \sigma$ .

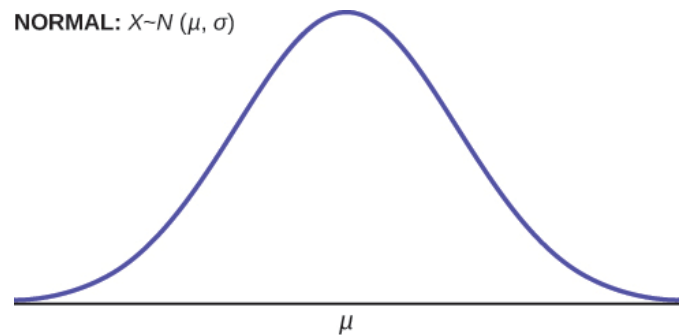


Figure 4.1.3.1. Notation for the Standard Normal Distribution.

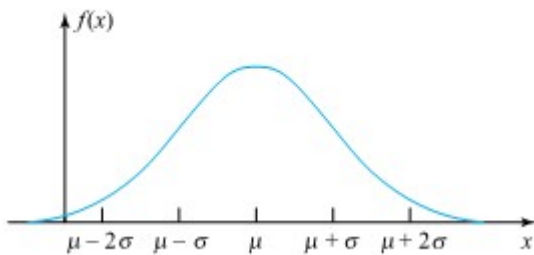


Figure 4.1.3.2. Graph of a normal probability density function

The exact form of formula (4.1.3.1) has a number of theoretical origins. It is also a form that turns out to be empirically useful in a great variety of applications. In theory, probabilities for the normal distributions can be found directly by integration using formula (4.1.3.1). Indeed, readers with pocket calculators that are preprogrammed to do numerical integration may find it instructive to check some of the calculations in the examples that follow, by straightforward use of formulas (4.1.1.2) and (4.1.3.1). We will also use statistical computing to find these by the use of formula. But the freshman calculus methods of evaluating integrals via antidifferentiation will fail when it comes to the normal densities. They do not have antiderivatives that are expressible in terms of elementary functions. Instead, normal probability tables are typically used based on a specialized form of normal distribution: the standard normal distribution.

## 4.1.4 Standard Normal Distribution

### STANDARD NORMAL DISTRIBUTION

The use of tables for evaluating normal probabilities depends on the following relationship. If  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ ,

#### EXPRESSION 4.1.4.1.

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

where the second inequality follows from the change of variable or substitution of:

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

This z-score is a standardized value measured in units of the standard deviation. For example, if the mean of a normal distribution is five and the standard deviation is two, the value 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:  $x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$ , and the z-score is three:  $z = (11-5)/2 = 3$ . The z-score tells you how many standard deviations the value  $x$  is above (to the right of) or below (to the left of) the mean,  $\mu$ . Values of  $x$  that are larger than the mean have positive z-scores, and values of  $x$  that are smaller than the mean have negative z-scores. If  $x$  equals the mean, then  $x$  has a z-score of zero.

Equation (4.1.4.1) involves an integral of the normal density with  $\mu = 0$  and  $\sigma = 1$ . The transformation with  $z = \frac{x - \mu}{\sigma}$  produces the distribution  $Z \sim N(0,1)$ . This states that the value  $x$  in the given equation comes from a normal distribution with mean of 0 and standard deviation of 1. It says that evaluation of all normal probabilities can be reduced to the evaluation of normal probabilities for that special case. So, the standard normal distribution is a normal distribution of standardized values using z-scores.

**DEFINITION 4.1.4.2. THE STANDARD NORMAL DISTRIBUTION****EXPRESSION 4.1.4.2.**

The normal distribution with  $\mu = 0$  and  $\sigma = 1$  is called the standard normal distribution.

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

*Z-value for a value x of a normal ( $\mu, \sigma^2$ ) random variable*

---

Relationship (4.1.4.2) shows how to use the standard normal cumulative probability function to find general normal probabilities. For  $X$  normal  $(\mu, \sigma^2)$  and a value  $x$  associated with  $X$ , one converts to units of standard deviations above the mean via:

**EXPRESSION 4.1.4.3.**

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

and then consults the standard normal table using  $z$  instead of  $x$ .

*Relation between normal ( $\mu, \sigma^2$ ) probabilities and standard normal probabilities: the standard normal cumulative probability*

---

The relationship between normal ( $\mu, \sigma^2$ ) and standard normal probabilities is illustrated in Figure 4.1.4.1

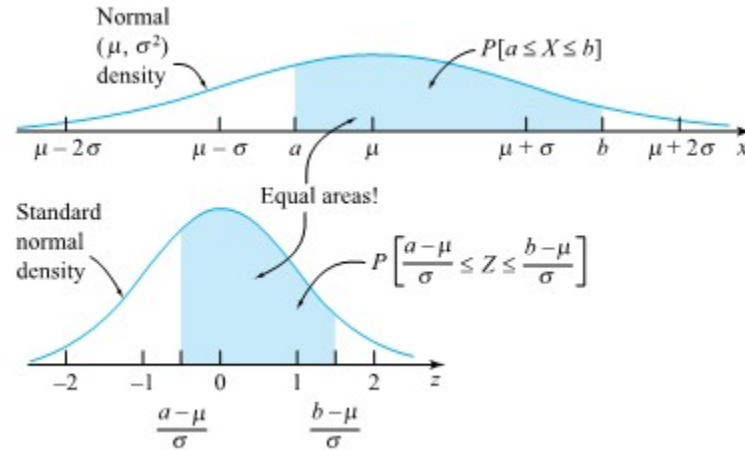


Figure 4.1.4.1. Illustration of the relationship between normal  $(\mu, \sigma^2)$  and standard normal probabilities.

Once one realizes that probabilities for all normal distributions can be had by tabulating probabilities for only the standard normal distribution, it is a relatively simple matter to use techniques of numerical integration to produce a standard normal table. The one that will be used in this text (other forms are possible) is given in the Table A1.1. Table of Standard Normal Probabilities in the Tables Appendix 1. It is a table of the standard normal cumulative probability function. That is, for values  $z$  located on the table's margins, the entries in the table body are:

**EXPRESSION 4.1.4.4**

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

where  $\Phi(z)$  is used to stand for the standard normal cumulative probability function, instead of the more generic  $F$ .

*Relationship between the standard normal cumulative probability function and the standard normal quantile function.*

In mathematical symbols, for  $\Phi(z)$ , the standard normal cumulative probability function, and  $Q_z(p)$ , the standard normal quantile function,

**EXPRESSION 4.1.4.5.**

$$\left. \begin{aligned} \Phi(Q_z(p)) &= p \\ Q_z(\Phi(z)) &= z \end{aligned} \right\}$$

Relationships (4.7.5) mean that  $Q_z(p)$  and  $\Phi(z)$  are inverse functions. (In fact, the relationship  $Q = F^{-1}$  is not just a standard normal phenomenon but is true in general for continuous distributions.)

**EXAMPLES****Example 4.1.4.1. Standard Normal Probabilities**

Suppose that  $Z$  is a standard normal random variable. We will find some probabilities for  $Z$  using Table 1. Table of Standard Normal Probabilities in the Tables Appendix. By a straight table look-up,

Cumulative probability of a value of  $Z$

$$P[Z < 1.76] = \Phi(1.76) = 0.96$$

(The tabled value is .9608, but in keeping with the earlier promise to state final probabilities to only two decimal places, the tabled value was rounded to get 0.96.)

After two table look-ups and a subtraction,

$$\text{Probability between two values of } Z \text{ } P[.57 < Z < 1.32] = P[Z < 1.32] - P[Z \leq .57]$$

$$= \Phi(1.32) - \Phi(0.57)$$

$$= 0.9066 - 0.7157$$

$$= 0.19$$

And a single table look-up and a subtraction yield a right-tail probability, such as,

Right-tailed probability of a  $Z$  value

$$P[Z > -0.89] = 1 - P[Z \leq -0.89] = 1 - 0.1867 = 0.81$$

As the table was used in these examples, probabilities for values  $z$  located on the table's margins were found in the table's body. The process can be run in reverse. Probabilities located in the table's body can be used to specify values  $z$  on the margins. For example, consider locating a value  $z$  such that,

$$P[-z < Z < z] = 0.95$$

$z$  will then put probability  $\frac{1 - 0.95}{2} = .025$  in the right tail of the standard normal distribution—i.e., be such that  $\Phi(z) = .975 = .975$ . Locating .975 in the table body, one sees that  $z = 1.96$ .

This amounts to finding the .975 quantile for the standard normal distribution and allows us to understand and describe standard normal quantiles.

Figure 4.1.4.2 illustrates all of the calculations for this example.

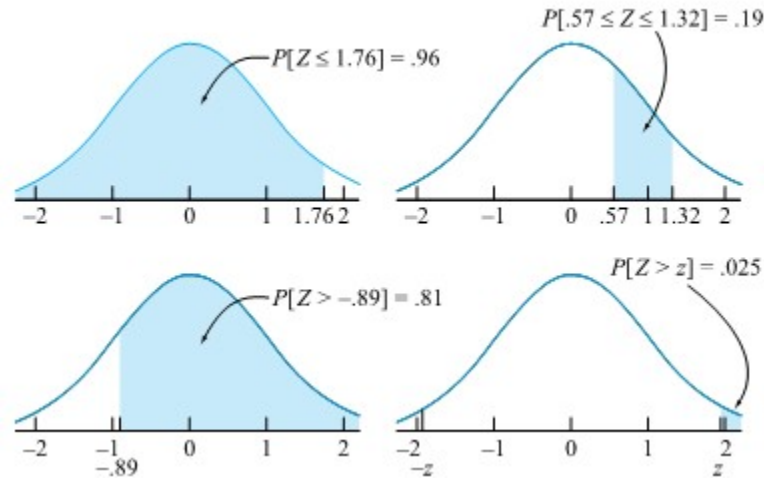


Figure 4.1.4.2. Standard normal probabilities for Example 4.7.1.

#### Example 4.1.4.2. Net Weights of Jars of Baby Food

J. Fisher, in his article “Computer Assisted Net Weight Control” (Quality Progress, June 1983), discusses the filling of food containers by weight. In the article, there is a reasonably bell-shaped histogram of individual net weights of jars of strained plums with tapioca. The mean of the values portrayed is about 137.2 g, and the standard deviation is about 1.6 g. The declared (or label) weight on jars of this product is 135.0 g.

Suppose that it is adequate to model

$W$  = the next strained plums and tapioca fill weight

with a normal distribution with  $\mu = 137.2$  and  $\sigma = 1.6$ . And further suppose the probability that the next jar filled is below declared weight (i.e.,  $P[W < 135.0]$ ) is of interest. Using formula (4.7.3),  $w = 135.0$  is converted to units of standard deviations above  $\mu$  (converted to a  $z$ -value) as

$$z = \frac{135.0 - 137.2}{1.6} = -1.38$$

Then, using Table 1. Table of Standard Normal Probabilities in the Tables Appendix,

$$P[W < 135.0] = \Phi(-1.38) = .08$$

This model puts the chance of obtaining a below-nominal fill level at about 8%.



As a second example, consider the probability that  $W$  is within 1 gram of nominal (i.e.,  $P[134.0 < W < 136.0]$ ). Using formula (4.7.3), both  $w_1 = 134.0$  and  $w_2 = 136.0$  are converted to  $z$ -values (or units of standard deviations above the mean) as

$$z_1 = \frac{134.0 - 137.2}{1.6} = -2.00$$

$$z_2 = \frac{136.0 - 137.2}{1.6} = -.75$$

So, then

$$P[134.0 < W < 136.0] = \Phi(-.75) - \Phi(-2.00) = .2266 - .0228 = .20$$

The preceding two probabilities and their standard normal counterparts are shown in Figure 4.1.4.3.

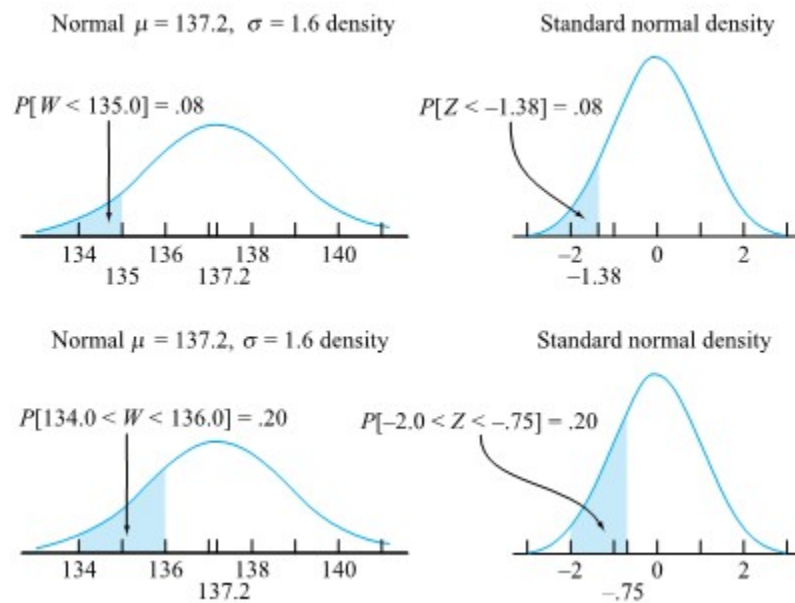


Figure 4.1.4.3. Normal probabilities for Example 4.1.4.2.

#### Example 4.1.4.3. Net Weights of Jars of Baby Food continued.

The calculations for this example have consisted of starting with all of the quantities on the right of formula (4.1.4.3) and going from the margin of Table A1.1. Table of Standard Normal Probabilities in the Tables Appendix to its body to find probabilities for  $W$ . An important variant on this process is to instead go from the body of the table to its margins to obtain  $z$ , and then—given only two of the three quantities on the right of formula (4.1.4.3)—to solve for the third.

For example, suppose that it is easy to adjust the aim of the filling process (i.e., the mean  $\mu$  of  $W$ ) and one wants to decrease the probability that the next jar is below the declared weight of 135.0 to .01 by increasing  $\mu$ . What is the minimum  $\mu$  that will achieve this (assuming that  $\sigma$  remains at 1.6 g)?

Figure 4.1.4.4 shows what to do.  $\mu$  must be chosen in such a way that  $w = 135.0$  becomes the .01 quantile of the normal distribution

with mean  $\mu$  and standard deviation  $\sigma = 1.6$ . Consulting Table A1.1, it is easy to determine that the .01 quantile of the standard normal distribution is

$$z = Q_z(.01) = -2.33$$

So in light of equation (4.1.4.3) one wants

$$-2.33 = \frac{135.0 - \mu}{1.6}$$

that is:  $\mu = 138.7\text{g}$

An increase of about  $138.7 - 137.2 = 1.5\text{ g}$  in fill level aim is required.

In practical terms, the reduction in  $P[W < 135.0]$  is bought at the price of increasing the average give-away cost associated with filling jars so that on average they contain much more than the nominal contents. In some applications, this type of cost will be prohibitive. There is another approach open to a process engineer. That is to reduce the variation in fill level through acquiring more precise filling equipment. In terms of equation (4.1.4.3), instead of increasing  $\mu$  one might consider paying the cost associated with reducing  $\sigma$ . The interested engineer is encouraged to verify that a reduction in  $\sigma$  to about .94 g would also produce  $P[W < 135.0] = .01$  without any change in  $\mu$ .

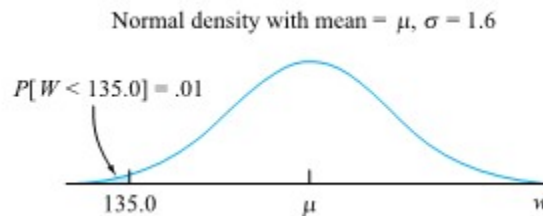


Figure 4.1.4.4. Normal distribution and  $P[W < 135.0] = .01$

As these examples illustrate, equation (4.1.4.3) is the fundamental relationship used in problems involving normal distributions. One way or another, three of the four entries in the equation are specified, and the fourth must be obtained.

## 4.1.5 The Empirical Rule

### THE EMPIRICAL RULE

---

If  $X$  is a random variable and has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then the Empirical Rule states the following:

- About 68% of the  $x$  values lie between  $-1\sigma$  and  $+1\sigma$  of the mean  $\mu$  (within one standard deviation of the mean).
- About 95% of the  $x$  values lie between  $-2\sigma$  and  $+2\sigma$  of the mean  $\mu$  (within two standard deviations of the mean).
- About 99.7% of the  $x$  values lie between  $-3\sigma$  and  $+3\sigma$  of the mean  $\mu$  (within three standard deviations of the mean). Notice that almost all the  $x$  values lie within three standard deviations of the mean.
- The z-scores for  $+1\sigma$  and  $-1\sigma$  are  $+1$  and  $-1$ , respectively.
- The z-scores for  $+2\sigma$  and  $-2\sigma$  are  $+2$  and  $-2$ , respectively.
- The z-scores for  $+3\sigma$  and  $-3\sigma$  are  $+3$  and  $-3$  respectively.

The empirical rule is also known as the 68-95-99.7 rule, and is shown in Figure 4.1.5.1.

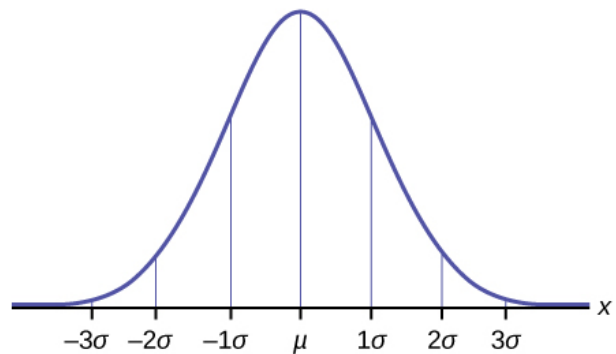


Figure 4.1.5.1. The Empirical Rule.

## 4.1.6 Tutorial 3 - Normal Probability Distributions

At this point, it is recommended that you work your way through the [Tutorial 3 exercise](#) found on the associated GitHub repository. This exercise will introduce you to the calculation of probabilities using the standard normal distribution in Python.

**It is strongly recommended that you consult the [Normal Probability & Confidence Intervals Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “Standard Normal Distribution in Python” will be particularly useful.

## 4.2.0 Introduction Joint Distributions and Independence

Most applications of probability to engineering statistics involve not one but several random variables. In some cases, the application is intrinsically multivariate. It then makes sense to think of more than one process variable as subject to random influences and to evaluate probabilities associated with them in combination. Take, for example, the assembly of a ring bearing with nominal inside diameter 1.00 in. on a rod with nominal diameter .99 in. If:

X = the ring bearing inside diameter

Y = the rod diameter

one might be interested in

$$P [ X < Y ] = P [\text{there is an interference in assembly}]$$

which involves both variables.

But even when a situation is univariate, samples larger than size 1 are essentially always used in engineering applications. The  $n$  data values in a sample are usually thought of as subject to chance causes and their simultaneous behavior must then be modeled. The methods so far discussed are capable of dealing with only a single random variable at a time. They must be generalized to create methods for describing several random variables simultaneously.

Entire books are written on various aspects of the simultaneous modeling of many random variables. This section can give only a brief introduction to the topic. We will start by considering first the comparatively simple case of jointly discrete random variables, the topics of joint and marginal probability functions, conditional distributions, and independence are discussed primarily through reference to simple bivariate examples.

The concepts for joint and marginal probability density functions, conditional distributions, and independence for jointly continuous random variables are not reviewed in this course, but are analogous to those discussed.

## 4.2.1 Joint Distributions

### DESCRIBING JOINTLY DISCRETE RANDOM VARIABLES

For several discrete variables the device typically used to specify probabilities is a **joint probability function**. The two-variable version of this is defined next.

#### DEFINITION 4.2.1.1. Joint probability function

##### EXPRESSION 4.2.1.1

A joint probability function for discrete random variables  $X$  and  $Y$  is a nonnegative function  $f(x, y)$ , giving the probability that (simultaneously)  $X$  takes the value  $x$  and  $Y$  takes the value  $y$ . That is,

$$f(x, y) = P[X = x \text{ and } Y = y]$$

#### Example 4.2.1.1. The Joint Probability Distribution of Two Bolt Torques

Return again to the situation of Brenny, Christensen, and Schneider and the measuring of bolt torques on the face plates of a heavy equipment component to the nearest integer. With

$X$  = the next torque recorded for bolt 3

$Y$  = the next torque recorded for bolt 4

The data displayed in the previous table and figure suggest, for example, that a sensible value for  $P[X = 18 \text{ and } Y = 18]$  might be  $\frac{1}{34}$ , the relative frequency of this pair in the data set. Similarly, the assignments

$$P[X = 18 \text{ and } Y = 17] = \frac{2}{34}$$

$$P[X = 14 \text{ and } Y = 9] = 0$$

also correspond to observed relative frequencies.

If one is willing to accept the whole set of relative frequencies defined by the students' data as defining probabilities for  $X$  and  $Y$ , these can be collected conveniently in a two-dimensional table specifying a joint probability function for  $X$  and  $Y$ . This is illustrated in Table 4.2.1.1. (To avoid clutter, 0 entries in the table have been left blank.)

$f(x, y)$  for the Bolt Torque Problem

$y \backslash x$	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					

Table 4.2.1.1.

## PROPERTIES OF A JOINT PROBABILITY FUNCTION FOR $X$ AND $Y$

The probability function given in tabular form in Table 4.2.1.1 has two properties that are necessary for mathematical consistency. These are that the  $f(x, y)$  values are each in the interval  $[0, 1]$  and that they total to 1. By summing up just some of the  $f(x, y)$  values, probabilities associated with  $X$  and  $Y$  being configured in patterns of interest are obtained.

### Example 4.2.1.2 Bolt Torques example, continued.

Consider using the joint distribution given in Table 4.2.1.1 to evaluate

$$P[X \geq Y],$$

$$P[|X - Y| \leq 1],$$

and  $P[X = 17]$

Take first  $P[X \geq Y]$ , the probability that the measured bolt 3 torque is at least as big as the measured bolt 4 torque. Figure 4.2.1.1 indicates with asterisks which possible combinations of  $x$  and  $y$  lead to bolt 3 torque at least as large as the bolt 4 torque. Referring to Table 4.2.1.1 and adding up those entries corresponding to the cells that contain asterisks,

$$\begin{aligned}
 P[X \geq Y] &= f(15, 13) + f(15, 14) + f(15, 15) + f(16, 16) \\
 &\quad + f(17, 17) + f(18, 14) + f(18, 17) + f(18, 18) \\
 &\quad + f(19, 16) + f(19, 18) + f(20, 20) \\
 &= \frac{1}{34} + \frac{1}{34} + \frac{3}{34} + \frac{2}{34} + \cdots + \frac{1}{34} = \frac{17}{34}
 \end{aligned}$$

Similar reasoning allows evaluation of  $P[|X - Y| \leq 1]$ -the probability that the bolt 3 and 4 torques are within 1ftlb of each other. Figure 4.2.1.2 shows combinations of  $x$  and  $y$  with an absolute difference of 0 or 1. Then, adding probabilities corresponding to these combinations,

$$\begin{aligned}
 P[|X - Y| \leq 1] &= f(15, 14) + f(15, 15) + f(15, 16) + f(16, 16) \\
 &\quad + f(16, 17) + f(17, 17) + f(17, 18) + f(18, 17) \\
 &\quad + f(18, 18) + f(19, 18) + f(19, 20) + f(20, 20) = \frac{18}{34}
 \end{aligned}$$

$y \backslash x$	11	12	13	14	15	16	17	18	19	20
20										*
19									*	*
18								*	*	*
17							*	*	*	*
16						*	*	*	*	*
15					*	*	*	*	*	*
14				*	*	*	*	*	*	*
13			*	*	*	*	*	*	*	*

Figure 4.2.1.1. Combinations of bolt 3 and bolt 4 torques with  $x \geq y$



$y \backslash x$	11	12	13	14	15	16	17	18	19	20
20									*	*
19								*	*	*
18							*	*	*	
17						*	*	*		
16					*	*	*			
15				*	*	*				
14			*	*	*					
13		*	*	*						

Figure 4.2.1.2. Combinations of bolt 3 and bolt 4 torques with  $|x-y| \leq 1$ .

Finally,  $P[X = 17]$ , the probability that the measured bolt 3 torque is 17ftlb, is obtained by adding down the  $x = 17$  column in Table 4.2.1.1. That is,

$$\begin{aligned}
 P[X = 17] &= f(17, 17) + f(17, 18) + f(17, 19) \\
 &= \frac{1}{34} + \frac{1}{34} + \frac{2}{34} \\
 &= \frac{4}{34}
 \end{aligned}$$

## FINDING MARGINAL PROBABILITY FUNCTIONS USING A BIVARIATE JOINT PROBABILITY FUNCTION

In bivariate problems like the present one, one can add down columns in a twoway table giving  $f(x, y)$  to get values for the probability function of  $X$ ,  $f_X(x)$ . And one can add across rows in the same table to get values for the probability function of  $Y$ ,  $f_Y(y)$ . One can then write these sums in the margins of the two-way table. So it should not be surprising that probability distributions for individual random variables obtained from their joint distribution are called marginal distributions. A formal statement of this terminology in the case of two discrete variables is next.

**DEFINITION 4.2.1.2. Marginal probability function****EXPRESSION 4.2.1.2**

The individual probability functions for discrete random variables  $X$  and  $Y$  with joint probability function  $f(x, y)$  are called **marginal probability functions**. They are obtained by summing  $f(x, y)$  values over all possible values of the other variable. In symbols, the marginal probability function for  $X$  is

$$f_X(x) = \sum_y f(x, y)$$

and the marginal probability function for  $Y$  is

$$f_Y(y) = \sum_x f(x, y)$$

**Example 4.2.1.3. continued.**

Table 4.2.1.2 is a copy of Table 4.2.1.1, augmented by the addition of marginal probabilities for  $X$  and  $Y$ . Separating off the margins from the two-way table produces tables of marginal probabilities in the familiar format of earlier. For example, the marginal probability function of  $Y$  is given separately in Table 4.2.1.3.

$y \setminus x$	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	

Table 4.2.1.2.

# Marginal Probability Function for $Y$

---

$y$	$f_Y(y)$
-----	----------

---

13	1/34
----	------

14	3/34
----	------

15	5/34
----	------

16	7/34
----	------

17	6/34
----	------

Table 4.2.1.3.

Getting marginal probability functions from joint probability functions raises the natural question whether the process can be reversed. That is, if  $f_X(x)$  and  $f_Y(y)$  are known, is there then exactly one choice for  $f(x, y)$ ? The answer to this question is "No." Figure 5.29 shows two quite different bivariate joint distributions that nonetheless possess the same marginal distributions. The marked difference between the distributions in Figure 4.2.1.3 has to do with the joint, rather than individual, behavior of  $X$  and  $Y$ .

Distribution 1					Distribution 2				
$y \backslash x$	1	2	3		$y \backslash x$	1	2	3	
3	.4	0	0	.4	3	.16	.16	.08	.4
2	0	.4	0	.4	2	.16	.16	.08	.4
1	0	0	.2	.2	1	.08	.08	.04	.2
	.4	.4	.2			.4	.4	.2	

Figure 4.2.1.3. Two different joint distributions with the same marginal distributions.

## 4.2.2 Conditional Distributions and Independence

### CONDITIONAL DISTRIBUTIONS AND INDEPENDENCE FOR DISCRETE RANDOM VARIABLES

---

When working with several random variables, it is often useful to think about what is expected of one of the variables, given the values assumed by all others. For example, in the bolt ( $X$ ) torque situation, a technician who has just loosened bolt 3 and measured the torque as **15ftlb** ought to have expectations for bolt 4 torque ( $Y$ ) somewhat different from those described by the marginal distribution in Table 4.2.1.3. After all, returning to the data in that led to Table 4.2.1.1, the relative frequency distribution of bolt 4 torques for those components with bolt 3 torque of **15ftlb** is as in Table 4.2.2.1. Somehow, knowing that  $X = 15$  ought to make a probability distribution for  $Y$  like the relative frequency distribution in Table 4.2.2.1 more relevant than the marginal distribution given in Table 4.1.1.3.

## Relative Frequency Distribution for Bolt 4 Torques When Bolt 3 Torque Is 15 ft lb

$y$ , Torque (ft lb)	Relative Frequency
13	1/9
14	1/9
15	3/9
16	2/9
17	2/9

Table 4.2.2.1.

The theory of probability makes allowance for this notion of “distribution of one variable knowing the values of others” through the concept of conditional distributions. The two-variable version of this is defined next.

**DEFINITION 4.2.2.1. Conditional probability function of  $X$  given  $Y=y$**

**EXPRESSION 4.2.2.1**

For discrete random variables  $X$  and  $Y$  with joint probability function  $f(x, y)$ , the conditional probability function of  $X$  given  $Y = y$  is the function of  $x$

$$f_{X|Y}(x | y) = \frac{f(x, y)}{\sum_x f(x, y)}$$

The conditional probability function of  $Y$  given  $X = x$  is the function of  $y$

$$f_{Y|X}(y | x) = \frac{f(x, y)}{\sum_y f(x, y)}$$

Comparing Definitions 4.2.1.1 and 4.2.2.1

**The conditional probability function for  $X$  given  $Y=y$**  4.2.2.2

$$f_{X|Y}(x | y) = \frac{f(x, y)}{f_Y(y)}$$

and

**The conditional probability function for  $Y$  given  $X=x$**  4.2.2.3

$$f_{Y|X}(y | x) = \frac{f(x, y)}{f_X(x)}$$

### Finding conditional distributions from a joint probability function

And formulas (4.2.2.2) and (4.2.2.3) are perfectly sensible. Equation (4.2.2.2) says that starting from  $f(x, y)$  given in a two-way table and looking only at the row specified by  $Y = y$ , the appropriate (conditional) distribution for  $X$  is given by the probabilities in that row (the  $f(x, y)$  values) divided by their sum ( $f_Y(y) = \sum_x f(x, y)$ ), so that they are renormalized to total to 1. Similarly, equation (4.2.2.3) says that looking only at the column specified by  $X = x$ , the appropriate conditional distribution for  $Y$  is given by the probabilities in that column divided by their sum.

#### Example 4.2.2.1. Bolt Torques continued.

To illustrate the use of equations (4.2.2.2) and (4.2.2.3), consider several of the conditional distributions associated with the joint distribution for the bolt 3 and bolt 4 torques, beginning with the conditional distribution for  $Y$  given that  $X = 15$ .

From equation (4.2.2.3),

$$f_{Y|X}(y | 15) = \frac{f(15, y)}{f_X(15)}$$

Referring to Table 4.2.1.2, the marginal probability associated with  $X = 15$  is  $\frac{9}{34}$ . So dividing values in the  $X = 15$  column of that

table by  $\frac{9}{34}$ , leads to the conditional distribution for  $Y$  given in Table 4.2.2.2. Comparing this to Table 4.2.1.4, indeed formula (4.2.2.3) produces a conditional distribution that agrees with intuition.

## The Conditional Probability Function for $Y$ Given $X = 15$

$y$	$f_{Y X}(y   15)$
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	$\left(\frac{2}{34}\right) \div \left(\frac{9}{34}\right) = \frac{2}{9}$

Table 4.2.2.2.

Next consider  $f_{Y|X}(y | 18)$  specified by



$$f_{Y|X}(y | 18) = \frac{f(18, y)}{f_X(18)}$$

Consulting Table 4.2.1.2 again leads to the conditional distribution for  $Y$  given that  $X = 18$ , shown in Table 4.2.2.3. Tables 4.2.2.2 and 4.2.4.3 confirm that the conditional distributions of  $Y$  given  $X = 15$  and given  $X = 18$  are quite different. For example, knowing that  $X = 18$  would on the whole make one expect  $Y$  to be larger than when  $X = 15$ .

## The Conditional Probability Function for $Y$ Given $X = 18$

$y$	$f_{Y X}(y   18)$
14	2/7
17	2/7
18	1/7
20	2/7

Table 4.2.2.3.

To make sure that the meaning of equation (4.2.2.2) is also clear, consider the conditional distribution of the bolt 3 torque ( $X$ ) given that the bolt 4 torque is 20 ( $Y = 20$ ). In this situation, equation (4.2.2.2) gives

$$f_{X|Y}(x | 20) = \frac{f(x, 20)}{f_Y(20)}$$

(Conditional probabilities for  $X$  are the values in the  $Y = 20$  row of Table 4.2.1..2 divided by the marginal  $Y = 20$  value.) Thus,  $f_{X|Y}(x | 20)$  is given in Table 4.2.2.4.

The Conditional Probability Function for $X$ Given $Y = 20$	
$x$	$f_{X Y}(x   20)$
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	$\left(\frac{1}{34}\right) \div \left(\frac{5}{34}\right) = \frac{1}{5}$

Table 4.2.2.4.

The bolt torque example has the feature that the conditional distributions for  $Y$  given various possible values for  $X$  differ. Further, these are not generally the same as the marginal distribution for  $Y$ .  $X$  provides some information about  $Y$ , in that depending upon its value there are differing probability assessments for  $Y$ . Contrast this with the following example.

#### Example 4.2.2.2. Random Sampling Two Bolt 4 Torques

Suppose that the 34 bolt 4 torques obtained by Brenny, Christensen, and Schneider and given in Table 4.2.2.5 are written on slips of paper and placed in a hat. Suppose further that the slips are mixed, one is selected, the corresponding torque is noted, and the slip is replaced. Then the slips are again mixed, another is selected, and the second torque is noted. Define the two random variables

$U =$  the value of the first torque selected

and

$V =$  the value of the second torque selected

Component	Bolt 3 Torque	Bolt 4 Torque	Component	Bolt 3 Torque	Bolt 4 Torque
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

Table 4.2.2.5.

Intuition dictates that (in contrast to the situation of  $X$  and  $Y$  in Example 4.2.2.1) the variables  $U$  and  $V$  don't furnish any information about each other. Regardless of what value  $U$  takes, the relative frequency distribution of bolt 4 torques in the hat is appropriate as the (conditional) probability distribution for  $V$ , and vice versa. That is, not only do  $U$  and  $V$  share the common marginal distribution given in Table 4.2.2.6 but it is also the case that for all  $u$  and  $v$ , both

$$4.2.2.4 \quad f_{U|V}(u | v) = f_U(u)$$

and

$$4.2.2.5 \quad f_{V|U}(v | u) = f_V(v)$$

Equations (4.2.2.4) and (4.2.2.5) say that the marginal probabilities in Table 4.2.2.6 also serve as conditional probabilities. They also specify how joint probabilities for  $U$  and  $V$  must be structured. That is, rewriting the left-hand side of equation (4.2.2.4) using expression (4.2.2.2),

$$\frac{f(u, v)}{f_V(v)} = f_U(u)$$

That is,

$$4.2.2.6 \quad f(u, v) = f_U(u)f_V(v)$$

(The same logic applied to equation (4.2.2.5) also leads to equation (4.2.2.6).) Expression (4.2.2.6) says that joint probability values for  $U$  and  $V$  are obtained by multiplying corresponding marginal probabilities. Table 4.2.2.7 gives the joint probability function for  $U$  and  $V$ .

## The Common Marginal Probability Function for $U$ and $V$

---

$u$ or $v$	$f_U(u)$ or $f_V(v)$
------------	----------------------

---

13	1/34
----	------

14	3/34
----	------

15	5/34
----	------

16	7/34
----	------

17	6/34
----	------

18	5/34
----	------

19	2/34
----	------

20	5/35
----	------

---

Table 4.2.2.6.

Joint Probabilities for $U$ and $V$									
$v \setminus u$	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	$\frac{1}{(34)^2}$	$\frac{3}{(34)^2}$	$\frac{5}{(34)^2}$	$\frac{7}{(34)^2}$	$\frac{6}{(34)^2}$	$\frac{5}{(34)^2}$	$\frac{2}{(34)^2}$	$\frac{5}{(34)^2}$	$1/34$
$f_U(u)$	$1/34$	$3/34$	$5/34$	$7/34$	$6/34$	$5/34$	$2/34$	$5/34$	

Table 4.2.2.7.

## INDEPENDENCE OF OBSERVATIONS IN STATISTICAL STUDIES

Example 18 suggests that the intuitive notion that several random variables are unrelated might be formalized in terms of all conditional distributions being equal to their corresponding marginal distributions. Equivalently, it might be phrased in terms of joint probabilities being the products of corresponding marginal probabilities. The formal mathematical terminology is that of **independence** of the random variables. The definition for the two-variable case is next.

### DEFINITION 4.2.2.7. Independence of random variables

#### EXPRESSION 4.2.2.7

Discrete random variables  $X$  and  $Y$  are called independent if their joint probability function  $f(x, y)$  is the product of their respective marginal probability functions. That is, independence means that

$$f(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y$$

If formula (4.2.2.7) does not hold, the variables  $X$  and  $Y$  are called dependent.

(Formula (4.2.2.7) does imply that conditional distributions are all equal to their corresponding marginals, so that the definition does fit its “unrelatedness” motivation.)

$U$  and  $V$  in Example 2.4.2.2 are independent, whereas  $X$  and  $Y$  in Example 2.4.2.1 are dependent.

Further, the two joint distributions depicted in Figure 2.4.1.3 give an example of a highly dependent joint distribution (the first) and one of independence (the second) that have the same marginals.

The notion of independence is a fundamental one. When it is sensible to model random variables as independent, *great mathematical simplicity results*. Where engineering data are being collected in an analytical context, and care is taken to make sure that all obvious physical causes of carryover effects that might influence successive observations are minimal, an assumption of independence between observations is often appropriate. And in enumerative contexts, relatively small (compared to the population size) simple random samples yield observations that can typically be considered as at least approximately independent.

#### Example 4.2.2.3. Bolt Torques example, continued

Again consider putting bolt torques on slips of paper in a hat. The method of torque selection described earlier for producing  $U$  and  $V$  is not simple random sampling. Simple random sampling as defined in Part 1 is without-replacement sampling, not the with-replacement sampling method used to produce  $U$  and  $V$ . Indeed, if the first slip is not replaced before the second is selected, the probabilities in Table 4.2.2.7 are not appropriate for describing  $U$  and  $V$ . For example, if no replacement is done, since only one slip is labeled **13ftlb**, one clearly wants

$$f(13, 13) = P[U = 13 \text{ and } V = 13] = 0$$

not the value

$$f(13, 13) = \frac{1}{(34)^2}$$

indicated in Table 4.2.2.7. Put differently, if no replacement is done, one clearly wants to use

$$f_{V|U}(13 | 13) = 0$$

rather than the value

$$f_{V|U}(13 | 13) = f_V(13) = \frac{1}{34}$$

which would be appropriate if sampling is done with replacement. Simple random sampling doesn't lead to exactly independent observations.

But suppose that instead of containing 34 slips labeled with torques, the hat contained  $100 \times 34$  slips labeled with torques with relative frequencies as in Table 4.2.2.6. Then even if sampling is done without replacement, the probabilities developed earlier for  $U$  and  $V$  (and placed in Table 4.2.2.7) remain at least approximately valid. For example, with 3,400 slips and using without-replacement sampling,

$$f_{V|U}(13 | 13) = \frac{99}{3,399}$$

is appropriate. Then, using the fact that

$$f_{V|U}(v | u) = \frac{f(u, v)}{f_U(u)}$$

so that

$$f(u, v) = f_{V|U}(v | u)f_U(u)$$

without replacement, the assignment

$$f(13, 13) = \frac{99}{3,399} \cdot \frac{1}{34}$$

is appropriate. But the point is that

$$\frac{99}{3,399} \approx \frac{1}{34}$$

and so

$$f(13, 13) \approx \frac{1}{34} \cdot \frac{1}{34}$$

For this hypothetical situation where the population size  $N = 3,400$  is much larger than the sample size  $n = 2$ , independence is a suitable approximate description of observations obtained using simple random sampling.

Where several variables are both independent and have the same marginal distributions, some additional jargon is used.

## INDEPENDENT AND IDENTICALLY DISTRIBUTED RANDOM VARIABLES

### DEFINITION 4.2.2.8. Independent and identically distributed.

If random variables  $X_1, X_2, \dots, X_n$  all have the same marginal distribution and are independent, they are termed iid or independent and identically distributed.

For example, the joint distribution of  $U$  and  $V$  given in Table 4.2.2.7 shows  $U$  and  $V$  to be iid random variables.

*When observations can be modeled as iid.*

The standard statistical examples of iid random variables are successive measurements taken from a stable process and the results of random sampling with

When can observations be modeled as iid? replacement from a single population. The question of whether an iid model is appropriate in a statistical application thus depends on whether or not the datagenerating

mechanism being studied can be thought of as conceptually equivalent to these.



## 4.2.3 Means and Variances for Linear Combinations of Random Variables

The last section introduced the mathematics used to simultaneously model several random variables. An important engineering use of that material is in the analysis of system outputs that are functions of random inputs. This section studies how the variation seen in an output random variable depends upon that of the variables used to produce it. We will focus on when using linear combinations of random variables.

### *The Distribution of a Function of Random Variables*

The problem considered in this section is this. Given a joint distribution for the random variables  $X, Y, \dots, Z$  and a function  $g(x, y, \dots, z)$ , the object is to predict the behavior of the random variable

$$4.2.3.1 \quad U = g(X, Y, \dots, Z)$$

In some special simple cases, it is possible to figure out exactly what distribution  $U$  inherits from  $X, Y, \dots, Z$

#### **Example 4.2.3.1 The Distribution of the Clearance Between Two Mating Parts with Randomly Determined Dimensions**

Suppose that a steel plate with nominal thickness .15 in. is to rest in a groove of nominal width .155 in., machined on the surface of a steel block. A lot of plates has been made and thicknesses measured, producing the relative frequency distribution in Table 4.2.3.1; a relative frequency distribution for the slot widths measured on a lot of machined blocks is given in Table 4.2.3.2.

If a plate is randomly selected and a block is separately randomly selected, a natural joint distribution for the random variables

$X$  = the plate thickness

$Y$  = the slot width

is one of independence, where the marginal distribution of  $X$  is given in Table 4.2.3.1 and the marginal distribution of  $Y$  is given in Table 4.2.3.2. That is, Table 4.2.3.3 gives a plausible joint probability function for  $X$  and  $Y$ .

Plate Thickness (in.)	Relative Frequency
.148	.4
.149	.3
.150	.3

Table 4.2.3.1

Slot Width (in.)	Relative Frequency
.153	.2
.154	.2
.155	.4
.156	.2

Table 4.2.3.2

A variable derived from X and Y that is of substantial potential interest is the clearance involved in the plate/block assembly,

$$U=Y-X$$

Notice that taking the extremes represented in Tables 4.2.3.1 and 4.2.3.2, U is guaranteed to be at least  $.153 - .150 = .003$  in. but no more than  $.156 - .148 = .008$  in. In fact, much more than this can be said. Looking at Table 4.2.3.3, one can see that the diagonals of entries (lower left to upper right) all correspond to the same value of  $Y - X$ . Adding probabilities on those diagonals produces the distribution of U given in Table 4.2.3.4.

$y \setminus x$	.148	.149	.150	$f_Y(y)$
.156	.08	.06	.06	.2
.155	.16	.12	.12	.4
.154	.08	.06	.06	.2
.153	.08	.06	.06	.2
$f_X(x)$	.4	.3	.3	

Table 4.2.3.3

The Probability Function for the Clearance  $U = Y - X$

$u$	$f(u)$
.003	.06
.004	.12 = .06 + .06
.005	.26 = .08 + .06 + .12
.006	.26 = .08 + .12 + .06
.007	.22 = .16 + .06
.008	.08

Table 4.2.3.4

Example 4.2.3.1 involves a very simple discrete joint distribution and a very simple function  $g$ —namely,  $g(x, y) = y - x$ . In general, exact complete solution of the problem of finding the distribution of  $U = g(X, Y, \dots, Z)$  is not practically possible. Happily, for many engineering applications of probability, approximate and/or partial solutions suffice to answer the questions of practical interest.

## MEANS AND VARIANCES FOR LINEAR COMBINATIONS OF RANDOM VARIABLES

For engineering purposes, it often suffices to know the mean and variance for  $U$  given in formula (4.2.3.1) (as opposed to knowing the whole distribution of  $U$ ). When this is the case and  $g$  is linear, there are explicit formulas for these.

### PROPOSITION 4.2.3.2

If  $X, Y, \dots, Z$  are  $n$  independent random variables and  $a_0, a_1, a_2, \dots, a_n$  are  $n + 1$  constants, then the random variable  $U = a_0 + a_1X + a_2Y + \dots + a_nZ$  has mean

$$4.2.3.3 \quad EU = a_0 + a_1EX + a_2EY + \dots + a_nEZ$$

and variance

$$4.2.3.4 \quad \text{Var}U = a_1^2 \text{Var}X + a_2^2 \text{Var}Y + \cdots + a_n^2 \text{Var}Z$$

Formula (4.2.3.3) actually holds regardless of whether or not the variables  $X, Y, \dots, Z$  are independent, and although formula (4.2.3.4) does depend upon independence, there is a generalization of it that can be used even if the variables are dependent. However, the form of Proposition 1 given here is adequate for present purposes.

One type of application in which Proposition 1 is immediately useful is that of geometrical tolerancing problems, where it is applied with  $a_0 = 0$  and the other  $a_i$ 's equal to plus and minus 1's.

#### Example 4.2.3.2 Clearance steel plate.

Consider a situation of the clearance involved in placing a steel plate in a machined slot on a steel block. With  $X, Y$ , and  $U$  being (respectively) the plate thickness, slot width, and clearance, means and variances for these variables can be calculated. The reader is encouraged to verify that

$$\begin{aligned} EX &= .1489 & \text{and} & & \text{Var}X &= 6.9 \times 10^{-7} \\ EY &= .1546 & \text{and} & & \text{Var}Y &= 1.04 \times 10^{-6} \end{aligned}$$

Now, since

$$U = Y - X = (-1)X + 1Y$$

Proposition 1 can be applied to conclude that

$$\begin{aligned} EU &= -1EX + 1EY = -.1489 + .1546 = .0057 \text{ in.} \\ \text{Var}U &= (-1)^2 6.9 \times 10^{-7} + (1)^2 1.04 \times 10^{-6} = 1.73 \times 10^{-6} \end{aligned}$$

so that

$$\sqrt{\text{Var}U} = .0013 \text{ in.}$$

It is worth the effort to verify that the mean and standard deviation of the clearance produced using Proposition 1 agree with those obtained using the distribution of  $U$  given in Table 4.2.3.4 and the formulas for the mean and variance given in Part 3. The advantage of using Proposition 1 is that if all that is needed are  $EU$  and  $\text{Var}U$ , there is no need to go through

the intermediate step of deriving the distribution of  $U$ . The calculations via Proposition 1 use only characteristics of the marginal distributions.

## WHEN RANDOM VARIABLES $X_1, X_2, \dots, X_n$ ARE RANDOM SELECTIONS (WITH REPLACEMENT) FROM A SINGLE NUMERICAL POPULATION

Another particularly important use of Proposition 1 concerns  $n$  iid random variables where each  $a_i$  is  $\frac{1}{n}$ .

That is, in cases where random variables  $X_1, X_2, \dots, X_n$  are conceptually equivalent to random selections (with replacement) from a single numerical population, Proposition 1 tells how the mean and variance of the random variable

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$$

are related to the population parameters  $\mu$  and  $\sigma^2$ . For independent variables  $X_1, X_2, \dots, X_n$  with common mean  $\mu$  and variance  $\sigma^2$ , Proposition 1 shows that

### 4.2.3.5 The mean of an average of $n$ iid random variables

$$E\bar{X} = \frac{1}{n}EX_1 + \frac{1}{n}EX_2 + \dots + \frac{1}{n}EX_n = n \left( \frac{1}{n}\mu \right) = \mu$$

and

### 4.2.3.6 The variance of an average of $n$ iid random variables

$$\begin{aligned} \text{Var } \bar{X} &= \left(\frac{1}{n}\right)^2 \text{Var } X_1 + \left(\frac{1}{n}\right)^2 \text{Var } X_2 + \dots + \left(\frac{1}{n}\right)^2 \text{Var } X_n \\ &= n \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

is decreasing in  $n$ , equations (4.2.3.5) and (4.2.3.6) give the reassuring picture of  $\bar{X}$  having a probability distribution centered at the population mean  $\mu$ , with spread that decreases as the sample size increases.

Relationships (4.2.3.5) and (4.2.3.6), which perfectly describe the random behavior of  $\bar{X}$  under random sampling with replacement, are also approximated descriptions of the behavior of  $\bar{X}$  under simple random sampling in enumerative contexts. (Recall the discussion about the approximate independence of observations resulting from simple random sampling of large populations.)

## 4.2.4 The Central Limit Theorem

### CENTRAL LIMIT EFFECT

One of the most frequently used statistics in engineering applications is the sample mean. Formulas related to the mean and variance of the probability distribution of the sample mean to those of a single observation when an iid model is appropriate have been discussed. One of the most useful facts of applied probability is that if the sample size is reasonably large, it is also possible to approximate the shape of the probability distribution of  $\bar{X}$ , independent of the shape of the underlying distribution of individual observations. That is, there is the following fact:

.

#### Proposition 4.2.2.1 The Central Limit Theorem

If  $X_1, X_2, \dots, X_n$  are iid random variables (with mean  $\mu$  and variance  $\sigma^2$ ), then for large  $n$ , the variable  $\bar{X}$  is approximately normally distributed. (That is, approximate probabilities for  $\bar{X}$  can be calculated using the normal distribution with mean  $\mu$  and variance  $\sigma^2/n$ .)

.

A proof of Proposition 4.2.2.1 is outside the purposes of this text. But intuition about the effect is fairly easy to develop through an example.

#### Example 4.2.2.1 The Central Limit Effect and the Sample Mean of Tool Serial Numbers, continued.

Consider again the example from Section 3.2.1.2 involving the last digit of essentially randomly selected serial numbers of pneumatic tools. Suppose now that

$W_1$  = the last digit of the serial number observed next Monday at 9 A.M.

$W_2$  = the last digit of the serial number observed the following Monday at 9 A.M.

A plausible model for the pair of random variables  $W_1, W_2$  is that they are independent, each with the marginal probability function

$$4.2.2.1 \quad f(w) = \begin{cases} .1 & \text{if } w = 0, 1, 2, \dots, 9 \\ 0 & \text{otherwise} \end{cases}$$

that is pictured in Figure 4.2.2.1

Using such a model, it is a straightforward exercise to reason that  $\bar{W} = \frac{1}{2}(W_1 + W_2)$  has the probability function given in Table 4.2.2.1 and pictured in Figure 4.2.2.2

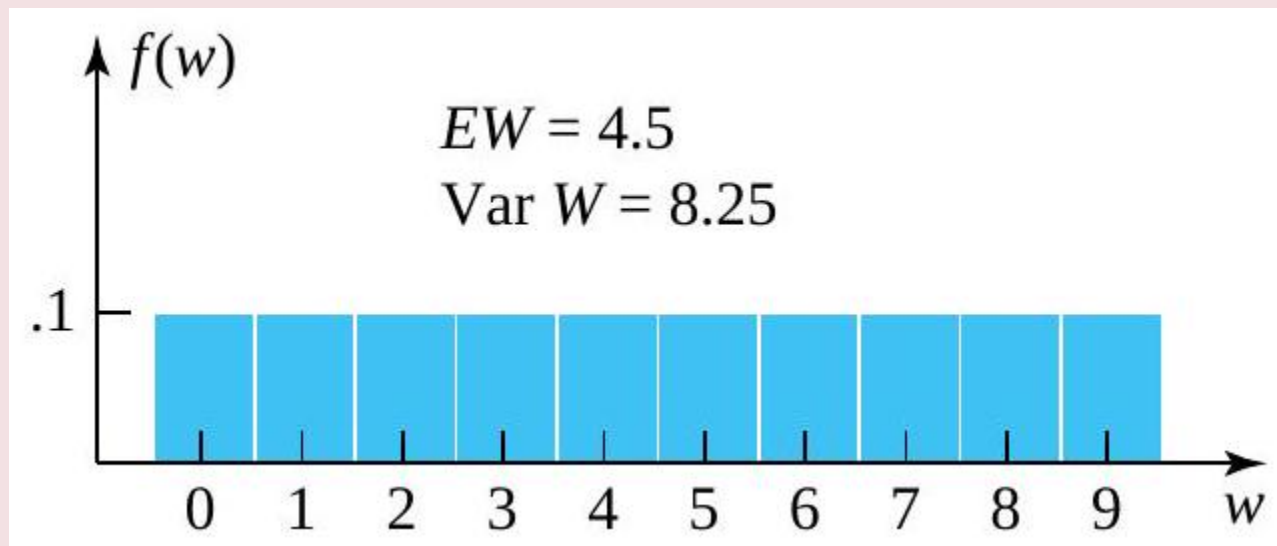


Figure 4.2.2.1 Probability histogram for  $W$ .

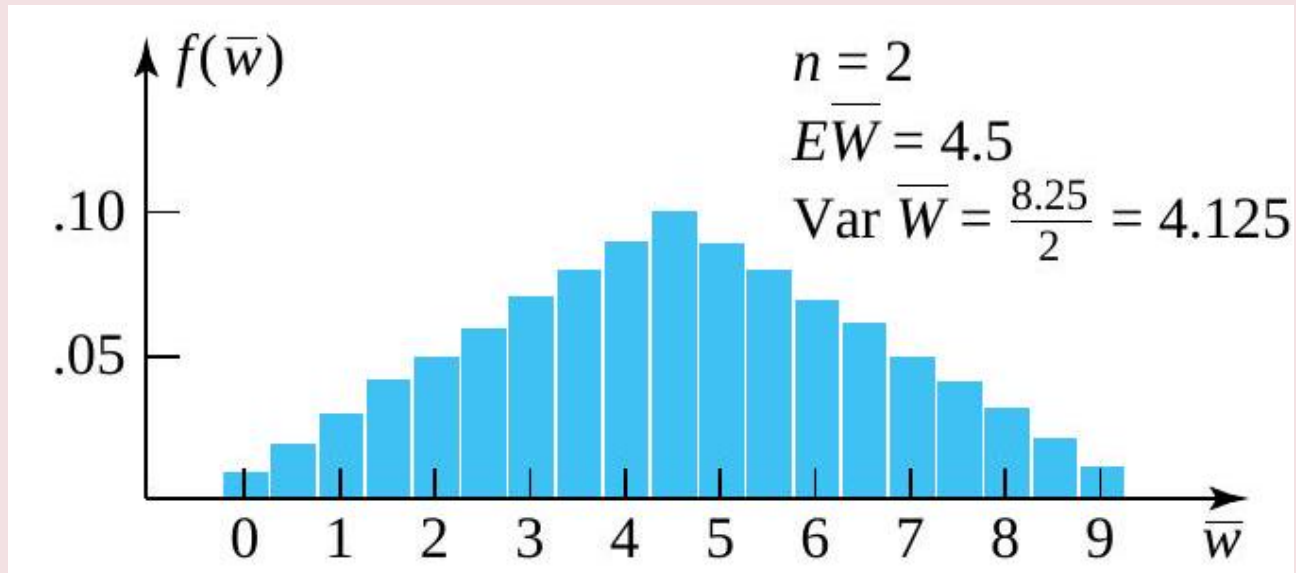


Figure 4.2.2.2 Probability histogram for  $\bar{W}$  based on  $n=2$ .

The Probability Function for  $\bar{W}$  for  $n = 2$

$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$	$\bar{w}$	$f(\bar{w})$
0.0	.01	2.0	.05	4.0	.09	6.0	.07	8.0	.03
0.5	.02	2.5	.06	4.5	.10	6.5	.06	8.5	.02
1.0	.03	3.0	.07	5.0	.09	7.0	.05	9.0	.01
1.5	.04	3.5	.08	5.5	.08	7.5	.04		

Table 4.2.2.1

Comparing Figures 4.2.2.1 and 4.2.2.2, it is clear that even for a completely flat/uniform underlying distribution of  $W$  and the small sample size of  $n = 2$ , the probability distribution of  $\bar{W}$  looks far more bell-shaped than the underlying distribution. It is clear why this is so. As you move away from the mean or central value of  $\bar{W}$ , there are relatively fewer and fewer combinations of  $w_1$  and  $w_2$  that can produce a given value of  $\bar{w}$ . For example, to observe  $\bar{W} = 0$ , you must have  $W_1 = 0$  and  $W_2 = 0$ —that is, you must observe not one but two extreme values. On the other hand, there are ten different combinations of  $w_1$  and  $w_2$  that lead to  $\bar{W} = 4.5$ .

It is possible to use the same kind of logic leading to Table 4.2.2.1 to produce exact probability distributions for  $\bar{W}$  based on larger sample sizes  $n$ . But such work is tedious, and for the purpose of indicating roughly how the central limit effect takes over as  $n$  gets larger, it is sufficient to approximate the distribution of  $\bar{W}$  via simulation for a larger sample size. To this end, 1,000 sets of values for iid variables  $W_1, W_2, \dots, W_8$  (with marginal distribution were simulated and each set averaged to produce 1,000 simulated values of  $\bar{W}$  based on  $n = 8$ . Figure 4.2.2.3 is a histogram of these 1,000 values. Notice the bell-shaped character of the plot. (The simulated mean of  $\bar{W}$  was  $4.508 \approx 4.5 = E\bar{W} = EW$ , while the variance of  $\bar{W}$  was  $1.025 \approx 1.013 = \text{Var } \bar{W} = 8.25/8$ , in close agreement with formulas.)



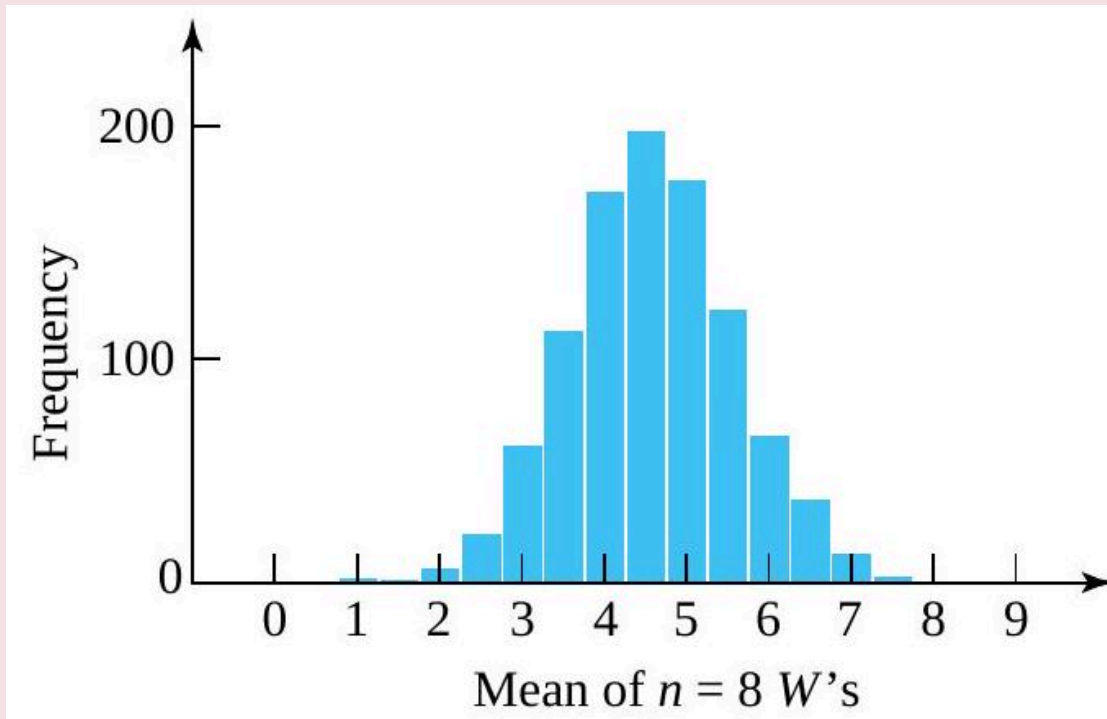


Figure 4.2.2.3 Histogram of 1,000 simulated values of  $\bar{W}$  based on  $n=8$ .

### Sample size and the central limit effect

What constitutes “large  $n$ ” in Proposition 4.2.2.1 isn’t obvious. The truth of the matter is that what sample size is required before  $\bar{X}$  can be treated as essentially normal depends on the shape of the underlying distribution of a single observation. Underlying distributions with decidedly nonnormal shapes require somewhat bigger values of  $n$ . But for most engineering purposes,  $n \geq 25$  or so is adequate to make  $\bar{X}$  essentially normal for the majority of data-generating mechanisms met in practice. (The exceptions are those subject to the occasional production of wildly outlying values.) Indeed, as Example 4.2.2.2 suggests, in many cases  $\bar{X}$  is essentially normal for sample sizes much smaller than 25.

The practical usefulness of Proposition 4.2.2.1 is that in many circumstances, only a normal table is needed to evaluate probabilities for sample averages.

#### Example 4.2.2.2 Stamp sale time requirement.

There is a stamp sale time requirements and we need to consider observing and averaging the next  $n = 100$  excess service times, to produce

$\bar{S}$  = the sample mean time (over a 7.5sec threshold) required to complete the next 100 stamp sales

And consider approximating  $P[\bar{S} > 17]$ .

We will assume that an iid model with marginal exponential  $\alpha = 16.5$  distribution is plausible for the individual excess service times,  $S$ . Then

$$E\bar{S} = \alpha = 16.5\text{sec}$$

and

$$\sqrt{\text{Var } \bar{S}} = \sqrt{\frac{\alpha^2}{100}} = 1.65\text{sec}$$

are appropriate for  $\bar{S}$ , via formulas. Further, in view of the fact that  $n = 100$  is large, the normal probability table may be used to find approximate probabilities for  $\bar{S}$ . Figure 4.2.2.4 shows an approximate distribution for  $\bar{S}$  and the area corresponding to  $P[\bar{S} > 17]$ .

The approximate probability distribution of  $\bar{S}$  is normal with mean 16.5 and standard deviation 1.65

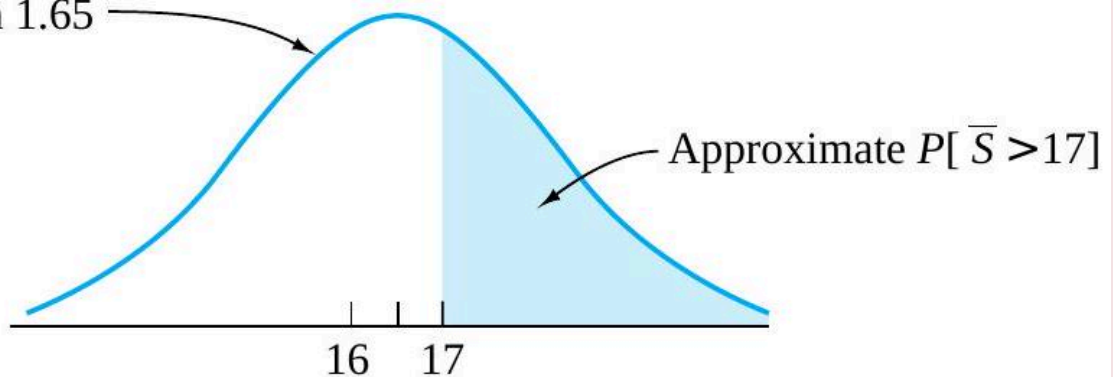


Figure 4.2.2.4 Approximate probability distribution for  $\bar{S}$  and  $P[\bar{S} > 17]$ .

As always, one must convert to  $z$ -values before consulting the standard normal table. In this case, the mean and standard deviation to be used are (respectively) 16.5sec and 1.65sec. That is, a  $z$ -value is calculated as

$$z = \frac{17 - 16.5}{1.65} = .30$$

So

$$P[\bar{S} > 17] \approx P[Z > .30] = 1 - \Phi(.30) = .38$$

## Z-VALUE FOR A SAMPLE MEAN

---

The  $z$ -value calculated in the example is an application of the general form

### 4.2.2.1 $z$ -value calculated for a sample mean

$$z = \frac{\bar{x} - E\bar{X}}{\sqrt{\text{Var } \bar{X}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

appropriate when using the central limit theorem to find approximate probabilities for a sample mean. Formula (4.2.2.1) is relevant because by Proposition 4.2.2.1,  $\bar{X}$  is approximately normal for large  $n$  and the formulas give its mean and standard deviation.

## *5.0.1 Introduction to Formal Statistical Inference*

Formal statistical inference uses probability theory to quantify the reliability of data-based conclusions. This chapter introduces the logic involved in several general types of formal statistical inference. Then the most common specific methods for one- and two-sample statistical studies are discussed.

The chapter begins with an introduction to confidence interval estimation, using the important case of large-sample inference for a mean. Then the topic of significance testing is considered, again using the case of large-sample inference for a mean. With the general notions in hand, successive sections treat the standard one- and two-sample confidence interval and significance-testing methods for means, then variances, and then proportions. Finally, the important topics of tolerance and prediction intervals are introduced.

## 5.0.1 Attributions

Part 5 of this open educational resource is composed of text mostly as an adaptation of “[Basic Engineering Data Collection and Analysis](#)” by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material from Chapters 6 of this text. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made.

Iowa State University Professor Emeritus [Stephen Vardeman](#) and Miami University Professor Emeritus [J. Marcus Jobe](#) (ISU PhD, 1984) have made their book [Basic Engineering Data Collection and Analysis](#), originally published by Duxbury/Thompson Learning/Cengage, freely available for download under a (CC BY-NC-SA) 4.0 International license through the [Iowa State University Digital Press](#). The book is available at

<https://www.iastatedigitalpress.com/plugins/books/127/>

and has been assigned the DOI

<https://doi.org/10.31274/isudp.2023.127>

The Basic Engineering Data Collection and Analysis book is essentially a revision/second edition of *Statistics for Engineering Problem Solving* by Vardeman that won the [American Society for Engineering Education](#) 1994 [Meriam/Wiley Distinguished Author Award](#).

---

Module 5.3 is text from “Statistics for Research Students”, by [Erich C. Fein; John Gilmour; Tanya Machin; and Liam Hendry](#) <https://usq.pressbooks.pub/statisticsforresearchstudents/>

Chapter 9: Nonparametric Statistics

[Statistics for Research Students](#) Copyright © 2022 by University of Southern Queensland is licensed under a [Creative Commons Attribution 4.0 International License](#), except where otherwise noted.

## 5.1.1 Large-Sample Confidence Intervals for a Mean

### LARGE-SAMPLE CONFIDENCE INTERVALS FOR A MEAN

---

Many important engineering applications of statistics fit the following standard mold. Values for parameters of a data-generating process are unknown. Based on data, the object is

1. identify an interval of values likely to contain an unknown parameter (or a function of one or more parameters) and
2. quantify “how likely” the interval is to cover the correct value.

For example, a piece of equipment that dispenses baby food into jars might produce an unknown mean fill level,  $\mu$ . Determining a data-based interval likely to contain  $\mu$  and an evaluation of the reliability of the interval might be important. Or a machine that puts threads on U-bolts might have an inherent variation in thread lengths, describable in terms of a standard deviation,  $\sigma$ . The point of data collection might then be to produce an interval of likely values for  $\sigma$ , together with a statement of how reliable the interval is. Or two different methods of running a pelletizing machine might have different unknown propensities to produce defective pellets, (say,  $p_1$  and  $p_2$ ). A data-based interval for  $p_1 - p_2$ , together with an associated statement of reliability, might be needed

#### **DEFINITION 5.1.1.1 Confidence Interval**

A confidence interval for a parameter (or function of one or more parameters) is a data-based interval of numbers thought likely to contain the parameter (or function of one or more parameters) possessing a stated probability-based confidence or reliability.

This section discusses how basic probability facts lead to simple large-sample formulas for confidence intervals for a mean,  $\mu$ . The unusual case where the standard deviation  $\sigma$  is known is treated first. Then

parallel reasoning produces a formula for the much more common situation where  $\sigma$  is not known. The section closes with discussions of three practical issues in the application of confidence intervals.

## A LARGE-N CONFIDENCE INTERVAL FOR $\mu$ INVOLVING $\sigma$

The final example in Chapter 4.2.2.4 involved a physically stable filling process known to have a net weight standard deviation of  $\sigma = 1.6$  g. Since, for large  $n$ , the sample mean of iid random variables is approximately normal, the final example of Chapter 4.2.4 argued that for  $n = 47$  and

$\bar{x}$  = the sample mean net fill weight of 47 jars filled by the process (g)

here is an approximately 80% chance that  $\bar{x}$  is within .3 gram of  $\mu$ . This fact is pictured again in Figure 5.1.1.1.

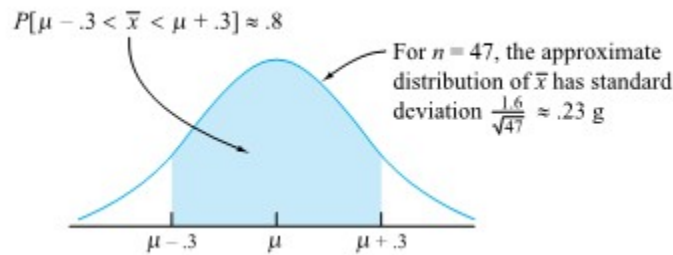


Figure 5.1.1.1 Approximate probability distribution for  $\bar{x}$  based on  $n = 47$

### Notation Conventions

We need to interrupt for a moment to discuss notation. In Part 4, capital letters were carefully used as symbols for random variables and corresponding lowercase letters for their possible or observed values. But here a lowercase symbol,  $\bar{x}$  has been used for the sample mean random variable. This is fairly standard statistical usage, and it is in keeping with the kind of convention used in earlier Parts. We are thus going to now abandon strict adherence to the capitalization convention introduced in Chapter 4. Random variables will often be symbolized using lowercase letters and the same symbols used for their observed values. The Chapter 4 capitalization convention is especially helpful in learning the basics of probability. But once those basics are mastered, it is common to abuse notation and to determine from context whether a random variable or its observed value is being discussed.

The most common way of thinking about a graphic like Figure 5.1.1.1 is to think of the possibility that

$$5.1.1.1 \quad \mu - .3 < \bar{x} < \mu + .3$$

in terms of whether or not  $\bar{x}$  falls in an interval of length  $2(.3) = .6$  centered at  $\mu$ . But the equivalent is to consider whether or not an interval of length .6 centered at  $\bar{x}$  falls on top of  $\mu$ . Algebraically, inequality (5.1.1.1) is equivalent to

$$5.1.1.2 \quad \bar{x} - .3 < \mu < \bar{x} + .3$$

which shifts attention to this second way of thinking. The fact that expression (5.1.1.2) has about an 80% chance of holding true anytime a sample of 47 fill weights is taken suggests that the random interval

$$5.1.1.3 \quad (\bar{x} - .3, \bar{x} + .3)$$

can be used as a confidence interval for  $\mu$ , with 80% associated reliability or confidence.

#### Example 5.1.1.1 A Confidence Interval for a Process Mean Fill Weight

Suppose a sample of  $n = 47$  jars produces  $\bar{x} = 138.2$  g. Then expression (5.1.1.3) suggests that the interval with endpoints

$$138.2 \text{ g} \pm .3 \text{ g}$$

(i.e., the interval from 137.9 g to 138.5 g) be used as an 80% confidence interval for the process mean fill weight.

It is not hard to generalize the logic that led to expression (5.1.1.3). Anytime an iid model is appropriate for the elements of a large sample, the central limit theorem implies that the sample mean  $\bar{x}$  is approximately normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ . Then, if for  $p > .5$ ,  $z$  is the  $p$  quantile of the standard normal distribution, the probability that

$$5.1.1.4 \quad \mu - z \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z \frac{\sigma}{\sqrt{n}}$$

is approximately  $1 - 2(1 - p)$ . But inequality (5.1.1.4) can be rewritten as

$$5.1.1.5 \quad \bar{x} - z \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z \frac{\sigma}{\sqrt{n}}$$

and thought of as the eventuality that the random interval with endpoints

#### EXPRESSION 5.1.1.6 Large-Sample Known $\sigma$ Confidence Limits for $\mu$

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

brackets the unknown  $\mu$ . So an interval with endpoints (5.1.1.6) is an approximate confidence interval for  $\mu$  (with confidence level  $1 - 2(1 - p)$ ).

In an application,  $z$  in equation (5.1.1.6) is chosen so that the standard normal probability between  $-z$  and  $z$  corresponds to a desired confidence level. Appendix Table A1.1 (of standard normal cumulative probabilities) can be used to verify the appropriateness of the entries in Table 5.1.1.1. (This table gives values of  $z$  for use in expression (5.1.1.6) for some common confidence levels.)



z's for Use in Two-sided  
Large-*n* Intervals for  $\mu$

Desired Confidence	<i>z</i>
80%	1.28
90%	1.645
95%	1.96
98%	2.33
99%	2.58

Table 5.1.1.1

### Example 5.1.1.2 Confidence Interval for the Mean Deviation from Nominal in a Grinding Operation

Dib, Smith, and Thompson studied a grinding process used in the rebuilding of automobile engines. The natural short-term variability associated with the diameters of rod journals on engine crankshafts ground using the process was on the order of  $\sigma = .7 \times 10^{-4}$  in. Suppose that the rod journal grinding process can be thought of as physically stable over runs of, say, 50 journals or less. Then if 32 consecutive rod journal diameters have mean deviation from nominal of  $\bar{x} = -.16 \times 10^{-4}$  in., it is possible to apply expression (5.1.1.6) to make a confidence interval for the current process mean deviation from nominal. Consider a 95% confidence level. Consulting Table 5.1.1.1 (or otherwise, realizing that 1.96 is the  $p = .975$  quantile of the standard normal distribution),  $z = 1.96$  is called for in formula (5.1.1.6) (since  $.95 = 1 - 2(1 - .975)$ ). Thus, a 95% confidence interval for the current process mean deviation from nominal journal diameter has endpoints

$$-.16 \times 10^{-4} \pm (1.96) \frac{.7 \times 10^{-4}}{\sqrt{32}}$$

that is, endpoints

$$-.40 \times 10^{-4} \text{ in. and } .08 \times 10^{-4} \text{ in.}$$

An interval like this one could be of engineering importance in determining the advisability of making an adjustment to the process aim. The interval includes both positive and negative values. So although  $\bar{x} < 0$ , the information in hand doesn't provide enough precision to tell with any certainty in which direction the grinding process should be adjusted. This, coupled with the fact that potential machine adjustments are probably much coarser than the best-guess misadjustment of  $\bar{x} = -.16 \times 10^{-4}$  in., speaks strongly against making a change in the process aim based on the current data.

## A GENERALLY APPLICABLE LARGE-N CONFIDENCE INTERVAL FOR $\mu$

Although expression (5.1.1.6) provides a mathematically correct confidence interval, the appearance of  $\sigma$  in the formula severely limits its practical usefulness. It is unusual to have to estimate a mean  $\mu$  when the corresponding  $\sigma$  is known (and can therefore be plugged into a formula). These situations occur primarily in manufacturing situations like those of Examples 5.1.1.1 and 2. Considerable past experience can sometimes give a sensible value for  $\sigma$ , while physical process drifts over time can put the current value of  $\mu$  in question.

Happily, modification of the line of reasoning that led to expression (5.1.1.1) produces a confidence interval formula for  $\mu$  that depends only on the characteristics of a sample. The argument leading to formula

(5.1.1.6) depends on the fact that for large  $n$ ,  $\bar{x}$  is approximately normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n}$ —i.e., that

$$5.1.1.7 \quad Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

is approximately standard normal. The appearance of  $\sigma$  in expression (5.1.1.7) is what leads to its appearance in the confidence interval formula (5.1.1.6). But a slight generalization of the central limit theorem guarantees that for large  $n$ ,

$$5.1.1.8 \quad Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is also approximately standard normal. And the variable (5.1.1.8) doesn't involve  $\sigma$ .

Beginning with the fact that (when an iid model for observations is appropriate and  $n$  is large) the variable (5.1.1.8) is approximately standard normal, the reasoning is much as before. For a positive  $z$ ,

$$-z < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < z$$

is equivalent to

$$\mu - z \frac{S}{\sqrt{n}} < \bar{x} < \mu + z \frac{S}{\sqrt{n}}$$

which in turn is equivalent to

$$\bar{x} - z \frac{S}{\sqrt{n}} < \mu < \bar{x} + z \frac{S}{\sqrt{n}}$$

Thus, the interval with random center  $\bar{x}$  and random length  $2zs/\sqrt{n}$ —i.e., with random endpoints

#### EXPRESSION 5.1.1.9 Large-Sample Confidence Levels for $\mu$

$$\bar{x} \pm z \frac{s}{\sqrt{n}}$$

can be used as an approximate confidence interval for  $\mu$ . For a desired confidence,  $z$  should be chosen such that the standard normal probability between  $-z$  and  $z$  corresponds to that confidence level.

### Example 5.1.1.3 Breakaway Torques and Hard Disk Failures

F. Willett, in the article “The Case of the Derailed Disk Drives” (Mechanical Engineering, 1988), discusses a study done to isolate the cause of “blink code A failure” in a model of Winchester hard disk drive. Included in that article are the data given in Figure 5.1.1.2. These are breakaway torques (units are inch ounces) required to loosen the drive’s interrupter flag on the stepper motor shaft for 26 disk drives returned to the manufacturer for blink code A failure. For these data,  $\bar{x} = 11.5$  in. oz and  $s = 5.1$  in. oz.

If the disk drives that produced the data in Figure 5.1.1.2 are thought of as representing the population of drives subject to blink code A failure, it seems reasonable to use an iid model and formula (5.1.1.9) to estimate the population mean breakaway torque. Choosing to make a 90% confidence interval for  $\mu$ ,  $z = 1.645$  is indicated in Table 5.1.1.1. And using formula (5.1.1.9), endpoints

$$11.5 \pm 1.645 \frac{5.1}{\sqrt{26}}$$

(i.e., endpoints 9.9 in. oz and 13.1 in. oz) are indicated.

The interval shows that the mean breakaway torque for drives with blink code A failure was substantially below the factory’s 33.5 in. oz target value. Recognizing this turned out to be key in finding and eliminating a design flaw in the drives.

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 Torques required to loosen 26 interrupter flags

## C: SOME COMMENTS CONCERNING CONFIDENCE INTERVALS

Formulas (5.1.1.6) and (5.1.1.9) have been used to make confidence statements of the type “ $\mu$  is between a and b.” But often a statement like “ $\mu$  is at least c” or “ $\mu$  is no more than d” would be of more practical value. For example, an automotive engineer might wish to state, “The mean NO emission for this engine is at most 5 ppm.” Or a civil engineer might want to make a statement like “the mean compressive strength for specimens of this type of concrete is at least 4188 psi.” That is, practical engineering problems are sometimes best addressed using one-sided confidence intervals.

### Making one-sided confidence intervals

There is no real problem in coming up with formulas for one-sided confidence intervals. If you have a workable two-sided formula, all that must be done is to

1. replace the lower limit with  $-\infty$  or the upper limit with  $+\infty$  and
2. adjust the stated confidence level appropriately upward (this usually means dividing the “unconfidence level” by 2).

This prescription works not only with formulas (5.1.1.6) and (5.1.1.9) but also with the rest of the two-sided confidence intervals introduced in this chapter.

#### Example 5.1.1.4 continued

For the mean breakaway torque for defective disk drives, consider making a one-sided 90% confidence interval for  $\mu$  of the form  $(-\infty, \#)$ , for  $\#$  an appropriate number. Put slightly differently, consider finding a 90% upper confidence bound for  $\mu$ , (say,  $\#$ ).'

Beginning with a two-sided 80% confidence interval for  $\mu$ , the lower limit can be replaced with  $-\infty$  and a one-sided 90% confidence interval determined. That is, using formula (6.9), a 90% upper confidence bound for the mean breakaway torque is

$$\bar{x} + 1.28 \frac{s}{\sqrt{n}} = 11.5 + 1.28 \frac{5.1}{\sqrt{26}} = 12.8 \text{ in. oz}$$

Equivalently, a 90% one-sided confidence interval for  $\mu$  is  $(-\infty, 12.8)$ .

The 12.8 in. oz figure here is less than (and closer to the sample mean than) the 13.1 in. oz upper limit from the 90% two-sided interval found earlier. In the one-sided case,  $-\infty$  is declared as a lower limit so there is no risk of producing an interval containing only numbers larger than the unknown  $\mu$ . Thus an upper limit smaller than that for a corresponding two-sided interval can be used.

### Interpreting a confidence interval

A second issue in the application of confidence intervals is a correct understanding of the technical meaning of the term confidence. Unfortunately, there

are many possible misunderstandings. So it is important to carefully lay out what confidence does and doesn't mean.

Prior to selecting a sample and plugging into a formula like (5.1.1.6) or (5.1.1.9), the meaning of a confidence level is obvious. Choosing a (two-sided) 90% confidence level and thus  $z = 1.645$  for use in formula (5.1.1.9), before the fact of sample selection and calculation, “there is about a 90% chance of winding up with an interval that brackets  $\mu$ .” In symbols, this might be expressed as

$$P \left[ \bar{x} - 1.645 \frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.645 \frac{s}{\sqrt{n}} \right] \approx .90$$

But how to think about a confidence level after sample selection? This is an entirely different matter. Once numbers have been plugged into a formula like (5.1.1.6) or (5.1.1.9), the die has already been cast, and the numerical interval is either right or wrong. The practical difficulty is that while which is the case can't be determined, it no longer makes logical sense to attach a probability to the correctness of the interval. For example, it would make no sense to look again at the two-sided interval found in Example 5.1.1.3 and try to say something like “there is a 90% probability that  $\mu$  is between 9.9 in. oz and 13.1 in. oz.”  $\mu$  is not a random variable. It is a fixed (although unknown) quantity that either is or is not between 9.9 and 13.1. There is no probability left in the situation to be discussed.

So what does it mean that (9.9, 13.1) is a 90% confidence interval for  $\mu$ ? Like it or not, the phrase “90%

confidence” refers more to the method used to obtain the interval (9.9, 13.1) than to the interval itself. In coming up with the interval, methodology has been used that would produce numerical intervals bracketing  $\mu$  in about 90% of repeated applications. But the effectiveness of the particular interval in this application is unknown, and it is not quantifiable in terms of a probability. A person who (in the course of a lifetime) makes many 90% confidence intervals can expect to have a “lifetime success rate” of about 90%. But the effectiveness of any particular application will typically be unknown.

A short statement summarizing this discussion as “the interpretation of confidence” will be useful.

#### **DEFINITION 5.1.1.2 Interpretation of a Confidence Interval**

To say that a numerical interval (a, b) is (for example) a 90% confidence interval for a parameter is to say that in obtaining it, one has applied methods of data collection and calculation that would produce intervals bracketing the parameter in about 90% of repeated applications. Whether or not the particular interval (a, b) brackets the parameter is unknown and not describable in terms of a probability.

The reader may feel that the statement in Definition 5.1.1.2 is a rather weak meaning for the reliability figure associated with a confidence interval. Nevertheless, the statement in Definition 5.1.1.2 is the correct interpretation and is all that can be rationally expected. And despite the fact that the correct interpretation may initially seem somewhat unappealing, confidence interval methods have proved themselves to be of great practical use.

### **D: SAMPLE SIZES FOR ESTIMATING $\mu$**

As a final consideration in this introduction to confidence intervals, note that formulas like (5.1.1.6) and (5.1.1.9) can give some crude quantitative answers to the question, “How big must n be?” Using formula (5.1.1.9), for example, if you have in mind (1) a desired confidence level, (2) a worst-case expectation for the sample standard deviation, and (3) a desired precision of estimation for  $\mu$ , it is a simple matter to solve for a corresponding sample size. That is, suppose that the desired confidence level dictates the use of the value  $z$  in formula (5.1.1.9),  $s$  is some likely worst-case value for the sample standard deviation, and you want to have confidence limits (or a limit) of the form  $\bar{x} \pm \Delta$ . Setting

$$\Delta = z \frac{S}{\sqrt{n}}$$

and solving for  $n$  produces the requirement

$$n = \left( \frac{zS}{\Delta} \right)^2$$

Suppose that in the disk drive problem, engineers plan to follow up the analysis of the data in Figure 5.1.1.2 with the testing of a number of new drives. This will be done after subjecting them to accelerated (high) temperature conditions, in an effort to understand the mechanism behind the creation of low breakaway torques. Further suppose that the mean breakaway torque for temperature-stressed drives is to be estimated with a two-sided 95% confidence interval and that the torque variability expected in the new temperature-stressed drives is no worse than the  $s = 5.1$  in. oz figure obtained from the returned drives. A  $\pm 1$  in. oz precision of estimation is desired. Then using the plus-or-minus part of formula (5.1.1.9) and remembering Table 5.1.1.1, the requirement is

$$1 = 1.96 \frac{5.1}{\sqrt{n}}$$

which, when solved for  $n$ , gives

$$n = \left( \frac{(1.96)(5.1)}{1} \right)^2 \approx 100$$

A study involving in the neighborhood of  $n = 100$  temperature-stressed new disk drives is indicated. If this figure is impractical, the calculations at least indicate that dropping below this sample size will (unless the variability associated with the stressed new drives is less than that of the returned drives) force a reduction in either the confidence or the precision associated with the final interval.

For two reasons, the kind of calculations in the previous example give somewhat less than an ironclad answer to the question of sample size. The first is that they are only as good as the prediction of the sample standard deviation,  $s$ . If  $s$  is underpredicted, an  $n$  that is not really large enough will result. (By the same token, if one is excessively conservative and overpredicts  $s$ , an unnecessarily large sample size will result.) The second issue is that expression (5.1.1.9) remains a large-sample formula. If calculations like the preceding ones produce  $n$  smaller than, say, 25 or 30, the value should be increased enough to guarantee that formula (5.1.1.9) can be applied.

## 5.1.2 Large-Sample Significance Tests for a Mean

### THE GOAL OF SIGNIFICANCE TESTING

---

The last chapter illustrated how probability can enable confidence interval estimation. This chapter makes a parallel introduction of significance testing.

Significance testing amounts to using data to quantitatively assess the plausibility of a trial value of a parameter (or function of one or more parameters). This trial value typically embodies a status quo/“pre-data” view. For example, a process engineer might employ significance testing to assess the plausibility of an ideal value of 138 g as the current process mean fill level of baby food jars. Or two different methods of running a pelletizing machine might have unknown propensities to produce defective pellets, (say,  $p_1$  and  $p_2$ ), and significance testing could be used to assess the plausibility of  $p_1 - p_2 = 0$  — i.e., that the two methods are equally effective.

This section describes how basic probability facts lead to simple large-sample significance tests for a mean,  $\mu$ . It introduces significance testing terminology in the case where the standard deviation  $\sigma$  is known. Next, a five-step format for summarizing significance testing is presented. Then the more common situation of significance testing for  $\mu$  where  $\sigma$  is not known is considered. The section closes with two discussions about practical issues in the application of significance-testing logic.

### LARGE-N SIGNIFICANCE TESTS FOR $\mu$ INVOLVING $\sigma$

---

Recall once more the final example in Chapter 4.2.4, where a physically stable filling process is known to have  $\sigma = 1.6$  g for net weight. Suppose further that with a declared (label) weight of 135 g, process engineers have set a target mean net fill weight at  $135 + 3\sigma = 139.8$  g. Finally, suppose that in a routine check of filling-process performance, intended to detect any change of the process mean from its target value, a sample of  $n = 25$  jars produces  $\bar{x} = 139.0$  g. What does this value have to say about the plausibility of the current process mean actually being at the target of 139.8 g?

The central limit theorem can be called on here. If indeed the current process mean is at 139.8 g,  $\bar{x}$  has an approximately normal distribution with mean 139.8 g and standard deviation  $\sigma/\sqrt{n} = 1.6/\sqrt{25} = .32$  g, as pictured in Figure 5.1.2.1 along with the observed value of  $\bar{x} = 139.0$  g.

Figure 5.1.2.2 shows the standard normal picture that corresponds to Figure 5.1.2.1. It is based on the fact that if the current process mean is on target at 139.8 g, then the fact that  $\bar{x}$  is approximately normal with mean  $\mu$  and standard deviation  $\sigma/\sqrt{n} = .32$  g implies that

5.1.2.1

$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 139.8}{.32}$$

is approximately standard normal. The observed  $\bar{x} = 139.0$  g in Figure 5.1.2.1 has corresponding observed  $z = -2.5$  in Figure 5.1.2.2.

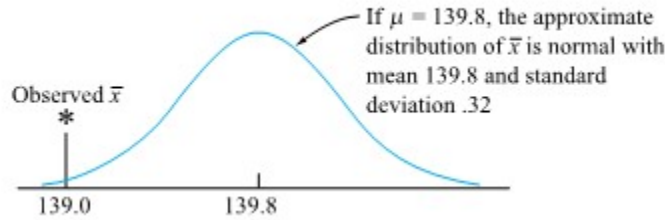


Figure 5.1.2.1 Approximate probability distribution for  $\bar{x}$  if  $\mu = 139.8$ , and the observed value of  $\bar{x} = 139.0$

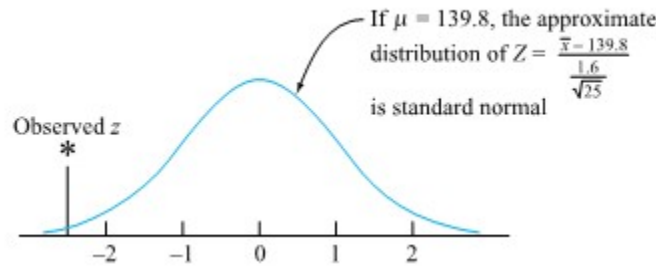


Figure 5.1.2.2 The standard normal picture corresponding to Figure 5.1.2.1

It is obvious from either Figure 5.1.2.1 or Figure 5.1.2.2 that if the process mean is on target at 139.8 g (and thus the figures are correct), a fairly extreme/rare  $\bar{x}$ , or equivalently  $z$ , has been observed. Of course, extreme/rare things occasionally happen. But the nature of the observed  $\bar{x}$  (or  $z$ ) might instead be considered as making the possibility that the process is on target implausible.

The figures even suggest a way of quantifying their own implausibility—through calculating a probability associated with values of  $\bar{x}$  (or  $Z$ ) at least as extreme as the one actually observed. Now “at least as extreme” must be defined in relation to the original purpose of data collection—to detect either a decrease of  $\mu$  below target or an increase above target. Not only are values  $\bar{x} \leq 139.0$  g ( $z \leq -2.5$ ) as extreme as that observed but so also are values  $\bar{x} \geq 140.6$  g ( $z \geq 2.5$ ). (The first kind of  $\bar{x}$  suggests a decrease in  $\mu$ , and the second suggests an increase.) That is, the implausibility of being on target might be quantified by noting that if this were so, only a fraction

$$\Phi(-2.5) + (1 - \Phi(2.5)) = .01$$

of all samples would produce a value of  $\bar{x}$  (or  $Z$ ) as extreme as the one actually observed. Put in those terms, the data seem to speak rather convincingly against the process being on target.



The argument that has just been made is an application of typical significance- testing logic. In order to make the pattern of thought obvious, it is useful to isolate some elements of it in definition form. This is done next, beginning with a formal restatement of the overall purpose.

#### **DEFINITION 5.1.2.1 Statistical Significance Testing**

Statistical significance testing is the use of data in the quantitative assessment of the plausibility of some trial value for a parameter (or function of one or more parameters).

Logically, significance testing begins with the specification of the trial or hypothesized value. Special jargon and notation exist for the statement of this value.

#### **DEFINITION 5.1.2.2 Null Hypothesis**

A null hypothesis is a statement of the form

$$\text{Parameter} = \#$$

or

$$\text{Function of parameters} = \#$$

(for some number, #) that forms the basis of investigation in a significance test. A null hypothesis is usually formed to embody a status quo/"pre-data" view of the parameter (or function of the parameter(s)). It is typically denoted as  $H_0$ .

The notion of a null hypothesis is so central to significance testing that it is common to use the term hypothesis testing in place of significance testing. The "null" part of the phrase "null hypothesis" refers to the fact that null hypotheses are statements of no difference, or equality. For example, in the context of the filling operation, standard usage would be to write

**5.1.2.2**       $H_0 : \mu = 139.8$

meaning that there is no difference between  $\mu$  and the target value of 139.8 g.

After formulating a null hypothesis, what kinds of departures from it are of interest must be specified.

**DEFINITION 5.1.2.3 Alternative Hypothesis**

An alternative hypothesis is a statement that stands in opposition to the null hypothesis. It specifies what forms of departure from the null hypothesis are of concern. An alternative hypothesis is typically denoted as  $H_a$ . It is of the same form as the corresponding null hypothesis, except that the equality sign is replaced by =, >, or <.

Often, the alternative hypothesis is based on an investigator's suspicions and/or hopes about the true state of affairs, amounting to a kind of research hypothesis that the investigator hopes to establish. For example, if an engineer tests what is intended to be a device for improving automotive gas mileage, a null hypothesis expressing "no mileage change" and an alternative hypothesis expressing "mileage improvement" would be appropriate.

Definitions 5.1.2.2 and 5.1.2.3 together imply that for the case of testing about a single mean, the three possible pairs of null and alternative hypotheses are

$$\begin{array}{lll} H_0 : \mu = \# & H_0 : \mu = \# & H_0 : \mu = \# \\ H_a : \mu > \# & H_a : \mu < \# & H_a : \mu \neq \# \end{array}$$

In the example of the filling operation, there is a need to detect both the possibility of consistently underfilled ( $\mu < 139.8$  g) and the possibility of consistently overfilled ( $\mu > 139.8$  g) jars. Thus, an appropriate alternative hypothesis is

**5.1.2.3**       $H_a : \mu \neq 139.8$

Once null and alternative hypotheses have been established, it is necessary to lay out carefully how the data will be used to evaluate the plausibility of the null hypothesis. This involves specifying a statistic to be calculated, a probability distribution appropriate for it if the null hypothesis is true, and what kinds of observed values will make the null hypothesis seem implausible.

**DEFINITION 5.1.2.4 Test Statistic**

A test statistic is the particular form of numerical data summarization used in a significance test. The formula for the test statistic typically involves the number appearing in the null hypothesis.

**DEFINITION 5.1.2.5 Null Distribution**

A reference (or null) distribution for a test statistic is the probability distribution describing the test statistic, provided the null hypothesis is in fact true.

The values of the test statistic considered to cast doubt on the validity of the null hypothesis are specified after looking at the form of the alternative hypothesis. Roughly speaking, values are identified that are more likely to occur if the alternative hypothesis is true than if the null hypothesis holds.

The discussion of the filling process scenario has vacillated between using  $\bar{x}$  and its standardized version  $Z$  given in equation (5.1.2.1) for a test statistic. Equation (5.1.2.1) is a specialized form of the general (large- $n$ , known  $\sigma$ ) test statistic for  $\mu$ ,

$$5.1.2.4 \quad Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

for the present scenario, where the hypothesized value of  $\mu$  is 139.8,  $n = 25$ , and  $\sigma = 1.6$ . It is most convenient to think of the test statistic for this kind of problem in the standardized form shown in equation (5.1.2.4) rather than as *bar x* itself. Using form (5.1.2.4), the reference distribution will always be the same—namely, standard normal.

Continuing with the filling example, note that if instead of the null hypothesis (5.1.2.2), the alternative hypothesis (5.1.2.3) is operating, observed  $\bar{x}$ 's much larger or much smaller than 139.8 will tend to result. Such  $\bar{x}$ 's will then, via equation (5.1.2.4), translate respectively to large or small (that is, large negative numbers in this case) observed values of  $Z$ —i.e., large values  $|z|$ . Such observed values render the null hypothesis implausible.

Having specified how data will be used to judge the plausibility of the null hypothesis, it remains to collect them, plug them into the formula for the test statistic, and (using the calculated value and the reference distribution) arrive at a quantitative assessment of the plausibility of  $H_0$ . There is jargon for the form this will take.

**DEFINITION 5.1.2.6 P-value**

The observed level of significance or p-value in a significance test is the probability that the reference distribution assigns to the set of possible values of the test statistic that are at least as extreme as the one actually observed (in terms of casting doubt on the null hypothesis).

**Small p-values are evidence against  $H_0$** 

The smaller the observed level of significance, the stronger the evidence against the validity of the null hypothesis. In the context of the filling operation, with an observed value of the test statistic of

$$z = -2.5$$

the p-value or observed level of significance is

$$\Phi(-2.5) + (1 - \Phi(2.5)) = .01$$

which gives fairly strong evidence against the possibility that the process mean is on target.

## 5.1.3 A Five-Step Format for Summarizing Significance Tests

### FIVE STEP SIGNIFICANCE TESTING FORMAT

It is helpful to lay down a step-by-step format for organizing write-ups of significance tests. The one that will be used in this text includes the following five steps:

**Step 1** State the null hypothesis.

**Step 2** State the alternative hypothesis.

**Step 3** State the test criteria. That is, give the formula for the test statistic (plugging in only a hypothesized value from the null hypothesis, but not any sample information) and the reference distribution. Then state in general terms what observed values of the test statistic will constitute evidence against the null hypothesis.

**Step 4** Show the sample-based calculations.

**Step 5** Report an observed level of significance and (to the extent possible) state its implications in the context of the real engineering problem.

#### Example 5.1.3.1 A Significance Test Regarding a Process Mean Fill Level

The five-step significance-testing format can be used to write up the preceding discussion of the filling process.

1.  $H_0: \mu = 139.8$ .
2.  $H_a: \mu \neq 139.8$ .
3. The test statistic is

$$Z = \frac{\bar{x} - 139.8}{\frac{\sigma}{\sqrt{n}}}$$

The reference distribution is standard normal, and large observed values  $|z|$  will constitute evidence against  $H_0$ .

4. The sample gives

$$z = \frac{139.0 - 139.8}{\frac{1.6}{\sqrt{100}}} = -2.5$$

5. The observed level of significance is

$$\begin{aligned} &P[\text{a standard normal variable} \leq -2.5] \\ &+ P[\text{a standard normal variable} \geq 2.5] \\ &= P[|\text{a standard normal variable}| \geq 2.5] \\ &= .01 \end{aligned}$$

This is reasonably weak evidence supporting the null hypothesis. This is reasonably strong evidence that the process mean fill level is not on target.

## 5.1.4 Generally Applicable Large- $n$ Significance Tests for Means.

The significance-testing method used to carry the discussion thus far is easy to discuss and understand but of limited practical use. The problem with it is that statistic (5.1.2.4) involves the parameter  $\sigma$ . As remarked in Chapter 5.1.1, there are few engineering contexts where one needs to make inferences regarding  $\mu$  but knows the corresponding  $\sigma$ . Happily, because of the same probability fact that made it possible to produce a large-sample confidence interval formula for  $\mu$  free of  $\sigma$ , it is also possible to do large- $n$  significance testing for  $\mu$  without having to supply  $\sigma$ .

For observations that are describable as essentially equivalent to random selections with replacement from a single population with mean  $\mu$  and variance  $\sigma^2$ , if  $n$  is large,

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is approximately standard normal. This means that for large  $n$ , to test

$$H_0 : \mu = \#$$

a widely applicable method will simply be to use the logic already introduced but with the statistic

**EXPRESSION 5.1.4.1 Large-sample test statistic for  $\mu$**

$$Z = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

in place of statistic (5.1.2.4).

### Example 5.1.4.1. Significance Testing and Hard Disk Failures continued.

Consider the problem of disk drive blink code A failure. Breakaway torques set at the factory on the interrupter flag connection to the

stepper motor shaft averaged 33.5 in. oz, and there was suspicion that blink code A failure was associated with reduced breakaway torque. Recall that a sample of  $n = 26$  failed drives had breakaway torques (given in Figure 5.1.2.2) with  $\bar{x} = 11.5$  in. oz and  $s = 5.1$  in. oz.

Consider the situation of an engineer wishing to judge the extent to which the data in hand debunk the possibility that drives experiencing blink code A failure have mean breakaway torque equal to the factory-set mean value of 33.5 in. oz. The five-step significance-testing format can be used.

1.  $H_0: \mu = 33.5$ .

2.  $H_a: \mu < 33.5$ .

(Here the alternative hypothesis is directional, amounting to a research hypothesis based on the engineer's suspicions about the relationship between drive failure and breakaway torque.)

3. The test statistic is

$$Z = \frac{\bar{x} - 33.5}{\frac{s}{\sqrt{n}}}$$

The reference distribution is standard normal, and small observed values  $z$  will constitute evidence against the validity of  $H_0$ . (Means less than 33.5 will tend to produce  $\bar{x}$ 's of the same nature and therefore small—i.e., large negative— $z$ 's.)

4. The sample gives

$$z = \frac{11.5 - 33.5}{\frac{5.1}{\sqrt{26}}} = -22.0$$

5. The observed level of significance is

$$P[\text{a standard normal variable} < -22.0] \approx 0$$

The sample provides overwhelming evidence that failed drives have a mean breakaway torque below the factory-set level.

It is important not to make too much of a logical jump here to an incorrect conclusion that this work constitutes the complete solution to the real engineering problem. Drives returned for blink code A failure have substandard breakaway torques. But in the absence of evidence to the contrary, it is possible that they are no different in that respect from nonfailing drives currently in the field. And even if reduced breakaway torque is at fault, a real-world fix of the drive failure problem requires the identification and prevention of the physical mechanism producing it. This is not to say the significance test lacks importance, but rather to remind the reader that it is but one of many tools an engineer uses to do a job.



## 5.1.5 Significance Testing and Formal Statistical Decision Making

The basic logic introduced in this section is sometimes applied in a decision-making context, where data are being counted on to provide guidance in choosing between two rival courses of action. In such cases, a decision-making framework is often built into the formal statistical analysis in an explicit way, and some additional terminology and patterns of thought are standard.

In some decision-making contexts, it is possible to conceive of two different possible decisions or courses of action as being related to a null and an alternative hypothesis. For example, in the filling-process scenario,  $H_0 : \mu = 139.8$  might correspond to the course of action “leave the process alone,” and  $H_a : \mu = 139.8$  could correspond to the course of action “adjust the process.” When such a correspondence holds, two different errors are possible in the decision-making process.

### DEFINITION 5.1.5.1 Type I Error

When significance testing is used in a decision-making context, deciding in favor of  $H_a$  when in fact  $H_0$  is true is called a type I error.

### DEFINITION 5.1.5.2 Type II Error

When significance testing is used in a decision-making context, deciding in favor of  $H_0$  when in fact  $H_a$  is true is called a type II error.

The content of these two definitions is represented in the  $2 \times 2$  table pictured in Figure 5.1.5.1. In the filling-process problem, a type I error would be adjusting an on-target process. A type II error would be failing to adjust an off-target process.

		$H_0$	$H_a$
The true state of affairs is described by:	$H_0$		Type I error
	$H_a$	Type II error	

Figure 5.1.5.1. Four potential outcomes in a decision problem

Significance testing is harnessed and used to come to a decision by choosing a critical value and, if the observed level of significance is smaller than the critical value (thus making the null hypothesis correspondingly implausible), deciding in favor of  $H_a$ . Otherwise, the course of action corresponding to  $H_0$  is followed. The critical value for the observed level of significance ends up being the a priori probability the decision maker runs of deciding in favor of  $H_a$ , calculated supposing  $H_0$  to be true. There is special terminology for this concept.

#### DEFINITION 5.1.5.3 Significance Level

When significance testing is used in a decision-making context, a critical value separating those large observed levels of significance for which  $H_0$  will be accepted from those small observed levels of significance for which  $H_0$  will be rejected in favor of  $H_a$  is called the type I error probability or the significance level. The symbol  $\alpha$  is usually used to stand for the type I error probability.

It is standard practice to use small numbers, like .1, .05, or even .01, for  $\alpha$ . This puts some inertia in favor of  $H_0$  into the decision-making process. (Such a practice guarantees that type I errors won't be made very often. But at the same time, it creates an asymmetry in the treatment of  $H_0$  and  $H_a$  that is not always justified.)

Definition 5.1.5.2 and Figure 5.1.5.1 make it clear that type I errors are not the only undesirable possibility. The possibility of type II errors must also be considered.

#### DEFINITION 5.1.5.4 Type II Errors

When significance testing is used in a decision-making context, the probability—calculated supposing a particular parameter value described by  $H_a$  holds—that the observed level of significance is bigger than  $\alpha$  (i.e.,  $H_0$  is not

rejected) is called a type II error probability. The symbol  $\beta$  is usually used to stand for a type II error probability.  $1-\beta$  is called the power of the significance test.

For most of the testing methods studied in this book, calculation of  $\beta$ 's is more than the limited introduction to probability given in Part 4 will support. But the job can be handled for the simple known- $\sigma$  situation that was used to introduce the topic of significance testing. And making a few such calculations will provide some intuition consistent with what, qualitatively at least, holds in general.

#### Example 5.1.5.1 continued

Again consider the filling process and testing  $H_0 : \mu = 139.8$  vs.  $H_a : \mu \neq 139.8$ . This time suppose that significance testing based on  $n = 25$  will be used tomorrow to decide whether or not to adjust the process. Type II error probabilities, calculated supposing  $\mu = 139.5$  and  $\mu = 139.2$  for tests using  $\alpha = .05$  and  $\alpha = .2$ , will be compared.

First consider  $\alpha = .05$ . The decision will be made in favor of  $H_0$  if the p-value exceeds .05. That is, the decision will be in favor of the null hypothesis if the observed value of Z given in the equation is such that

$$|z| < 1.96$$

ie, if

$$139.8 - 1.96(.32) < \bar{x} < 139.8 + 1.96(.32)$$

ie. if

$$139.2 < \bar{x} < 140.4$$

Now if  $\mu$  described by  $H_a$  given in  $H_a : \mu \neq 139.8$  is the true process mean,  $\bar{x}$  is not approximately normal with mean 139.8 and standard deviation .32, but rather approximately normal with mean  $\mu$  and standard deviation .32. So for such a  $\mu$ , expression (5.1.5.1) and Definition 5.1.5.4 show that the corresponding  $\beta$  will be the probability the corresponding normal distribution assigns to the possibility that  $139.2 < \bar{x} < 140.4$ . This is pictured in Figure 5.1.5.2 for the two means  $\mu = 139.5$  and  $\mu = 139.2$ .

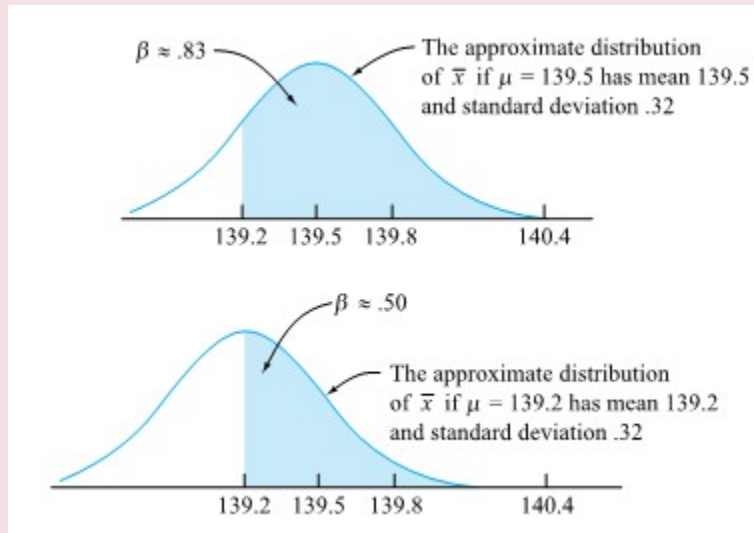


Figure 5.1.5.2 Approximate probability distributions for  $\bar{x}$  for two different values of  $\mu$  described by  $H_a$  and the corresponding  $\beta$ 's, when  $\alpha = .05$

It is an easy matter to calculate z-values corresponding to  $\bar{x} = 139.2$  and  $\bar{x} = 140.4$  using means of 139.5 and 139.2 and a standard deviation of .32 and to consult a standard normal table in order to verify the correctness of the two  $\beta$ 's marked in Figure 5.1.5.2.

Parallel reasoning for the situation with  $\alpha = .2$  is as follows. The decision will be in favor of  $H_0$  if the p-value exceeds .2. That is, the decision will be in favor of  $H_0$  if  $|z| < 1.28$ —i.e., if

$$139.4 < \bar{x} < 140.2$$

If  $\mu$  described by  $H_a$  is the true process mean,  $\bar{x}$  is approximately normal with mean  $\mu$  and standard deviation .32. So the corresponding  $\beta$  will be the probability this normal distribution assigns to the possibility that  $139.4 < \bar{x} < 140.2$ . This is pictured in Figure 5.1.5.3 for the two means  $\mu = 139.5$  and  $\mu = 139.2$ , having corresponding type II error probabilities  $\beta = .61$  and  $\beta = .27$ .

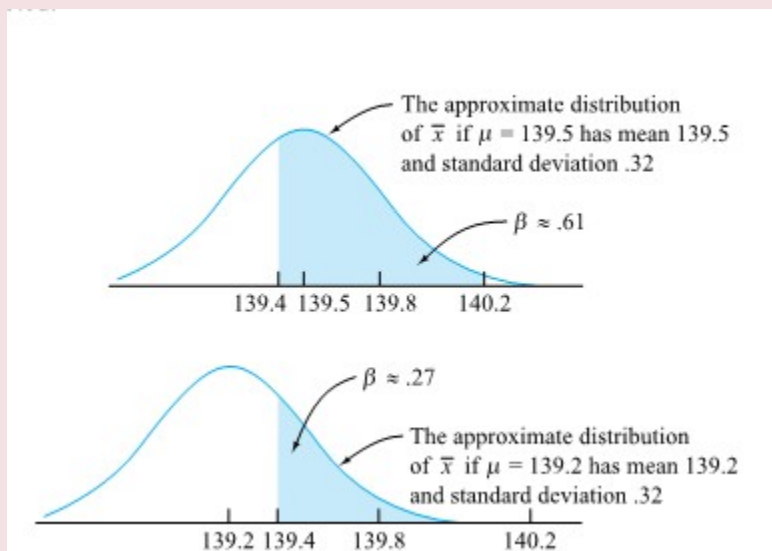


Figure 5.1.5.3. Approximate probability distributions for  $\bar{x}$  for two different values of  $\mu$  described by  $H_a$  and the corresponding  $\beta$ 's, when  $\alpha = .2$

The calculations represented by the two figures are collected in Table 5.1.5.1. Notice two features of the table. First, the  $\beta$  values for  $\alpha = .05$  are larger than those for  $\alpha = .2$ . If one wants to run only a 5% chance of (incorrectly) deciding to adjust an on-target process, the price to be paid is a larger probability of failure to recognize an off-target condition. Secondly, the  $\beta$  values for  $\mu = 139.2$  are smaller than the  $\beta$  values for  $\mu = 139.5$ . The further the filling process is from being on target, the less likely it is that the off-target condition will fail to be detected.

		$\mu$	
		139.2	139.5
$\alpha$	.05	.50	.83
	.2	.27	.61

Table 5.1.5.1  $\beta$  values.

The story told by Table 5.1.5.1 applies in qualitative terms to all uses of significance testing in decision-making contexts. The further  $H_0$  is from being true, the smaller the corresponding  $\beta$ . And small  $\alpha$ 's imply large  $\beta$ 's and vice versa.

### The effect of sample size on $\beta$ s

There is one other element of this general picture that plays an important role in the determination of error probabilities. That is the matter of sample size. If a sample size can be increased, for a given  $\alpha$ , the corresponding  $\beta$ 's can be reduced. Redo the calculations of the previous example, this time supposing that  $n = 100$  rather than 25. Table 5.1.5.2 shows the type II error probabilities that should result, and comparison with Table 5.1.5.1 serves to indicate the sample-size effect in the filling-process example.

		$\mu$	
		139.2	139.5
$\alpha$	.05	.04	.53
	.2	.01	.28

Table 5.1.5.2  $\beta$  values.

### Analogy between testing and a criminal trial

opposing hypotheses, namely

An analogy helpful in understanding the standard logic applied when significance testing is employed in decision-making involves thinking of the process of coming to a decision as a sort of legal proceeding, like a criminal trial. In a criminal trial, there are two

$H_0$  : The defendant is innocent

$H_a$  : The defendant is guilty

Evidence, playing a role similar to the data used in testing, is gathered and used to decide between the two hypotheses. Two types of potential error exist in a criminal trial: the possibility of convicting an innocent person (parallel to the type I error) and the possibility of acquitting a guilty person (similar to the type II error). A criminal trial is a situation where the two types of error are definitely thought of as having differing consequences, and the two hypotheses are treated asymmetrically. The a priori presumption in a criminal trial is in favor of  $H_0$ , the defendant's innocence. In order to keep the chance of a false conviction small (i.e., keep  $\alpha$  small), overwhelming evidence is required for conviction, in much the same way that if small  $\alpha$  is used in testing, extreme values of the test statistic are needed in order to indicate rejection of  $H_0$ . One consequence of this method of operation in criminal trials is that there is a substantial chance that a guilty individual will be acquitted, in the same way that small  $\alpha$ 's produce big  $\beta$ 's in testing contexts.

This significance testing/criminal trial parallel is useful, but do not make more of it than is justified. Not all significance-testing applications are properly thought of in this light. And few engineering scenarios are simple enough to reduce to a "decide between  $H_0$  and  $H_a$ " choice. Sensible applications of significance testing are often only steps of "evidence evaluation" in a many-faceted, data-based job necessary to solve an engineering problem. And even when a real problem can be reduced to a simple "decide between  $H_0$  and  $H_a$ " framework, it need not be the case that the "choose a small  $\alpha$ " logic is appropriate. In some engineering contexts, the practical consequences of a type II error are such that rational decision-making strikes a balance between the opposing goals of small  $\alpha$  and small  $\beta$ 's.

## 5.1.6 Statistical Significance, Estimation, and Practical Importance

### SOME COMMENTS CONCERNING SIGNIFICANCE TESTING AND ESTIMATION

Confidence interval estimation and significance testing are the two most commonly used forms of formal statistical inference. These having been introduced, it is appropriate to offer some comparative comments about their practical usefulness and, in the process, admit to an *estimation orientation* that will be reflected in much of the rest of this book's treatment of formal inference.

More often than not, engineers need to know "What is the value of the parameter?" rather than "Is the parameter equal to some hypothesized value?" And it is confidence interval estimation, not significance testing, that is designed to answer the first question. A confidence interval for a mean breakaway torque of from 9.9 in. oz to 13.1 in. oz says what values of  $\mu$  seem plausible. A tiny observed level of significance in testing  $H_0: \mu = 33.5$  says only that the data speak clearly against the possibility that  $\mu = 33.5$ , but it doesn't give any clue to the likely value of  $\mu$ .

#### "Statistical Significance" and Practical Importance

The fact that significance testing doesn't produce any useful indication of what parameter values are plausible is sometimes obscured by careless interpretation of semistandard jargon. For example, it is common in some fields to term p-values less than

.05 "statistically significant" and ones less than .01 "highly significant." The danger in this kind of usage is that "significant" can be incorrectly heard to mean "of great practical consequence" and the p-value incorrectly interpreted as a measure of how much a parameter differs from a value stated in a null hypothesis. One reason this interpretation doesn't follow is that the observed level of significance in a test depends not only on how far  $H_0$  appears to be from being correct but on the sample size as well. Given a large enough sample size, any departure from  $H_0$ , whether of practical importance or not, can be shown to be "highly significant."

#### Example 5.1.6.1 Statistical Significance and Practical Importance in a Regulatory Agency Test

A good example of the previous points involves the newspaper article in Figure 5.1.6.1 Apparently the Pass Master manufacturer did enough physical mileage testing (used a large enough  $n$ ) to produce a p-value less than .05 for testing a null hypothesis of no mileage improvement. That is, a "statistically significant" result was obtained.

But the size of the actual mileage improvement reported is only "small but real," amounting to about .8 mpg. Whether or not this improvement is of practical importance is a matter largely separate from the significance-testing result. And an engineer equipped with a confidence interval for the mean mileage

improvement is in a better position to judge this than is one who knows only that the p-value was less than .05.

WASHINGTON (AP)—A gadget that cuts off a car's air conditioner when the vehicle accelerates has become the first product aimed at cutting gasoline consumption to win government endorsement.

The device, marketed under the name "Pass Master," can provide a "small but real fuel economy benefit," the Environmental Protection Agency said Wednesday.

Motorists could realize up to 4 percent fuel reduction while using their air conditioners on cars equipped with the device, the agency said. That would translate into .8-miles-per-gallon improvement for a car that normally gets 20 miles to the gallon with the air conditioner on.

The agency cautioned that the 4 percent figure was a maximum amount and could be less depending on a motorist's driving habits, the type of car and the type of air conditioner.

But still the Pass Master, which sells for less than \$15, is the first of 40 products to pass the EPA's tests as making any "statistically significant" improvement in a car's mileage.

*Figure 5.1.6.1 Article from The Lafayette Journal and Courier, Page D-3, August 28, 1980. Reprinted by permission of the Associated Press. © 1980 the Associated Press to Stephen B. Vardeman and J. Marcus Jobe. Basic Engineering Data Collection and Analysis (Figure 6.8 of Chapter 6).*

#### Example 5.1.6.2 continued

To illustrate the effect that sample size has on observed level of significance, return to the breakaway torque problem and consider two hypothetical samples, one based on  $n = 25$  and the other on  $n = 100$  but both giving  $\bar{x} = 32.5$  in. oz and  $s = 5.1$  in. oz.

For testing  $H_0: \mu = 33.5$  with  $H_a: \mu < 33.5$ , the first hypothetical sample gives

$$z = \frac{32.5 - 33.5}{\frac{5.1}{\sqrt{25}}} = -.98$$

with associated observed level of significance

$$\Phi(-.98) = .16$$

The second hypothetical sample gives

$$z = \frac{32.5 - 33.5}{\frac{5.1}{\sqrt{100}}} = -1.96$$

with corresponding p-value



$$\Phi(-1.96) = .02$$

Because the second sample size is larger, the second sample gives stronger evidence that the mean breakaway torque is below 33.5 in. oz. But the best data-based guess at the difference between  $\mu$  and 33.5 is  $\bar{x} - 33.5 = -1.0$  in. oz in both cases. And it is the size of the difference between  $\mu$  and 33.5 that is of primary engineering importance.

It is further useful to realize that in addition to doing its primary job of providing an interval of plausible values for a parameter, a confidence interval itself also provides some significance-testing information. For example, a 95% confidence interval for a parameter contains all those values of the parameter for which significance tests using the data in hand would produce p-values bigger than 5%. (Those values not covered by the interval would have associated p-values smaller than 5%.)

#### Example 6.1.6.3 continued

Recall from Chapter 5.1.1 that a 90% one-sided confidence interval for the mean breakaway torque for failed drives is  $(-\infty, 12.8)$ . This means that for any value, #, larger than 12.8 in. oz, a significance test of  $H_0: \mu = \#$  with  $H_a: \mu < \#$  would produce a p-value less than .1. So clearly, the observed level of significance corresponding to the null hypothesis  $H_0: \mu = 33.5$  is less than .1. (Infact, as was seen earlier in this section, the p-value is 0 to two decimal places.) Put more loosely, the interval  $(-\infty, 12.8)$  is a long way from containing 33.5 in. oz and therefore makes such a value of  $\mu$  quite implausible.

The discussion here could well raise the question “What practical role remains for significance testing?” Some legitimate answers to this question are

1. In an almost negative way, p-values can help an engineer gauge the extent to which data in hand are inconclusive. When observed levels of significance are large, more information is needed in order to arrive at any definitive judgment.
2. Sometimes legal requirements force the use of significance testing in a compliance or effectiveness demonstration. (This was the case in Example 5.1.6.2, where before the Pass Master could be marketed, some mileage improvement had to be legally demonstrated.)
3. There are cases where the use of significance testing in a decision-making framework is necessary and appropriate. (An example is acceptance sampling: Based on information from a sample of items from a large lot, one must determine whether or not to receive shipment of the lot.)
4. As additional evidence and reinforcements of reports or scientific journal results.

So, properly understood and handled, significance testing does have its place in engineering practice. Thus, although the rest of this book features estimation over significance testing, methods of significance testing will not be completely ignored.

## *5.2.0 Introduction One- and Two-Sample Inference for Means*

Part 5 introduced the basic concepts of confidence interval estimation and significance testing. There are thousands of specific methods of these two types. This book can only discuss a small fraction that are particularly well known and useful to engineers. The next sections consider the most elementary of these—some of those that are applicable to one- and two-sample studies—beginning in this section with methods of formal inference for means.

Inferences for a single mean, based not on the large samples of Part 5 but instead on small samples, are considered first. In the process, it is necessary to introduce the so-called (Student)  $t$  probability distributions. Presented next are methods of formal inference for paired data. The section concludes with discussions of both large- and small- $n$  methods for data-based comparison of two means based on independent samples.

## 5.2.1 Small-Sample Inference for a Single Mean

The most important practical limitation on the use of the methods of the previous two sections is the requirement that  $n$  must be large. That restriction comes from the fact that without it, there is no way to conclude that

$$5.2.1.1 \quad \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is approximately standard normal. So if, for example, one mechanically uses the large- $n$  confidence interval formula

$$5.2.1.2 \quad \bar{x} \pm z \frac{S}{\sqrt{n}}$$

with a small sample, there is no way of assessing what actual level of confidence should be declared. That is, for small  $n$ , using  $z = 1.96$  in formula (5.2.1.2) generally doesn't produce 95% confidence intervals. And without a further condition, there is neither any way to tell what confidence might be associated with  $z = 1.96$  nor any way to tell how to choose  $z$  in order to produce a 95% confidence level.

There is one important special circumstance in which it is possible to reason in a way parallel to the work in Part 5 and arrive at inference methods for means based on small sample sizes. That is the situation where it is sensible to model the observations as iid normal random variables. The normal observations case is convenient because although the variable (5.2.1.1) is not standard normal, it does have a recognized, tabled distribution. This is the Student  $t$  distribution.

### DEFINITION 5.2.1.1 The (Student) $t$ distribution

The (Student)  $t$  distribution with degrees of freedom parameter  $\nu$  is a continuous probability distribution with probability density

#### EXPRESSION 5.2.1.3

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right) \sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}$$

for all  $t$ .

If a random variable has the probability density given by formula (5.2.1.3), it is said to have a  $t_v$  distribution.

The word Student in Definition 5.2.1.1 was the pen name of the statistician who first came upon formula (5.2.1.3). Expression (5.2.1.3) is rather formidable looking. No direct computations with it will actually be required in this book. But, it is useful to have expression (5.2.1.3) available in order to sketch several  $t$  probability densities, to get a feel for their shape. Figure 5.2.1.1 pictures the  $t$  densities for degrees of freedom  $v = 1, 2, 5,$  and  $11$ , along with the standard normal density.

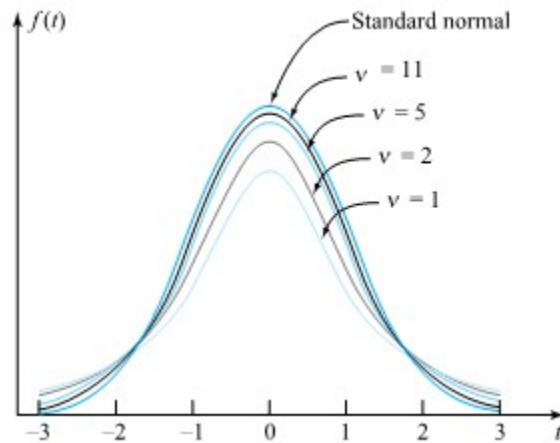


Figure 5.2.1.1  $t$  Probability densities for  $v = 1, 2, 5,$  and  $11$  and the standard normal density

### **$t$ distributions and the standard normal distribution'**

The message carried by Figure 5.2.1.1 is that the  $t$  probability densities are bell shaped and symmetric about 0. They are flatter than the standard normal density but are increasingly like it as  $v$  gets larger. In fact, for most practical purposes, for  $v$  larger than about 30, the  $t$  distribution with  $v$  degrees of freedom and the standard normal distribution are indistinguishable.

Probabilities for the  $t$  distributions are not typically found using the density in expression (5.2.1.3), as no

simple antiderivative for  $f(t)$  exists. Instead, it is common to use tables (or statistical software) to evaluate common t distribution quantiles and to get at least crude bounds on the types of probabilities needed in significance testing. Table A1.2 in the Appendix 1 of statistical tables is a typical table of t quantiles. Across the top of the table are several cumulative probabilities. Down the left side are values of the degrees of freedom parameter,  $\nu$ . In the body of the table are corresponding quantiles. Notice also that the last line of the table is a " $\nu = \infty$ " (i.e., standard normal) line.

#### Example 5.2.1.1 Use of a Table of t Distribution Quantiles

Suppose that  $T$  is a random variable having a t distribution with  $\nu = 5$  degrees of freedom. Consider first finding the .95 quantile of  $T$ 's distribution, then seeing what Table A1.2 reveals about  $P[T < -1.9]$  and then about  $P[|T| > 2.3]$ .

First, looking at the  $\nu = 5$  row of Table A1.2 under the cumulative probability .95, 2.015 is found in the body of the table. That is,  $Q(.95) = 2.015$  or (equivalently)  $P[T \leq 2.015] = .95$ .

Then note that by symmetry,

$$P[T < -1.9] = P[T > 1.9] = 1 - P[T \leq 1.9]$$

Looking at the  $\nu = 5$  row of Table A1.2, 1.9 is between the .90 and .95 quantiles of the  $t_5$  distribution. That is,

$$.90 < P[T \leq 1.9] \leq .95$$

so, finally

$$.05 < P[T < -1.9] < .10$$

Then, from the  $\nu = 5$  row of Table A1.2, 2.3 is seen to be between the .95 and .975 quantiles of the  $t_5$  distribution. That is,

$$.95 < P[T \leq 2.3] < .975$$

so

$$.05 < P[|T| > 2.3] < .10$$

The three calculations of this example are pictured in Figure 5.2.1.2

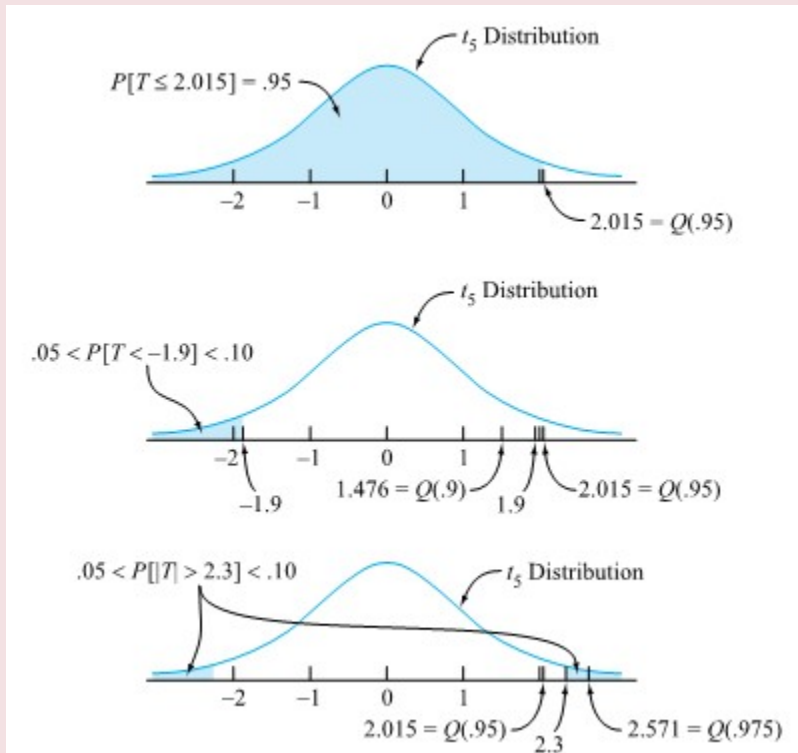


Figure 5.2.1.2 Three  $t_5$  probability calculations for Example 5.2.1.1.

The connection between expressions (5.2.1.3) and (5.2.1.1) that allows the development of small- $n$  inference methods for normal observations is that if an iid normal model is appropriate,

$$5.2.1.4 \quad T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has the  $t$  distribution with  $\nu = n - 1$  degrees of freedom. (This is consistent with the basic fact used in the previous two sections. That is, for large  $n$ ,  $\nu$  is large, so the  $t_\nu$  distribution is approximately standard normal; and for large  $n$ , the variable (5.2.1.4) has already been treated as approximately standard normal.)

Since the variable (5.2.1.4) can under appropriate circumstances be treated as a  $t_{n-1}$  random variable, we are in a position to work in exact analogy to what was done in Part 5 to find methods for confidence interval estimation and significance testing. That is, if a data-generating mechanism can be thought of as essentially equivalent to drawing independent observations from a single normal distribution, a two-sided confidence interval for  $\mu$  has endpoints

#### EXPRESSION 5.2.1.5 Normal distribution confidence limits for $\mu$

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

where  $t$  is chosen such that the  $t_{n-1}$  distribution assigns probability corresponding to the desired confidence level to the interval between  $-t$  and  $t$ . Further, the null hypothesis

$$H_0 : \mu = \#$$

can be tested using the statistic

**EXPRESSION 5.2.1.6 Normal distribution test statistic for  $\mu$**

$$T = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

and a  $t_{n-1}$  reference distribution.

*Operationally*, the only difference between the inference methods indicated here and the large-sample methods of the previous two sections is the exchange of standard normal quantiles and probabilities for ones corresponding to the  $t_{n-1}$  distribution. *Conceptually*, however, the nominal confidence and significance properties here are practically relevant only under the extra condition of a reasonably normal underlying distribution. Before applying either expression (5.2.1.5) or (5.2.1.6) in practice, it is advisable to investigate the appropriateness of a normal model assumption.

**Example 5.2.1.2 Small-Sample Confidence Limits for a Mean Spring Lifetime**

Part of a data set of W. Armstrong (appearing in Analysis of Survival Data by Cox and Oakes) gives numbers of cycles to failure of ten springs of a particular type under a stress of 950 N/mm<sup>2</sup>. These spring-life observations are given in Table 5.2.1.1 in units of 1,000 cycles.

Cycles to Failure of Ten  
Springs under 950 N/mm<sup>2</sup>  
Stress (10<sup>3</sup> cycles)

Spring Lifetimes

225, 171, 198, 189, 189

135, 162, 135, 117, 162

Table 5.2.1.1 An important question here might be “What is the average spring lifetime under conditions of 950 N/mm<sup>2</sup> stress?” Since only  $n = 10$  observations are available, the large-sample method of Part 5.1 is not applicable. Instead, only the method indicated by expression (5.2.1.5) is a possible option. For it to be appropriate, lifetimes must be normally distributed.

Without a relevant base of experience in materials engineering, it is difficult to speculate a priori about the appropriateness of a normal lifetime model in this context. But at least it is possible to examine the data in Table 5.2.1.1 themselves for evidence of strong departure from normality. Figure 5.2.1.3 is a normal plot for the data. It shows that in fact no such evidence exists.

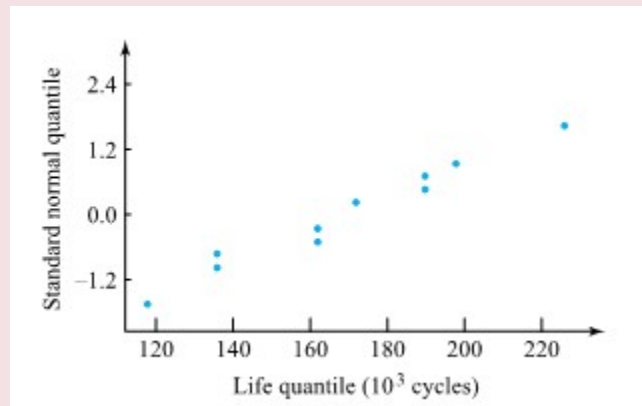


Figure 5.2.1.3 Normal plot of spring lifetimes

For the ten lifetimes,  $\bar{x} = 168.3$  ( $\times 10^3$  cycles) and  $s = 33.1$  ( $\times 10^3$  cycles). So to estimate the mean spring lifetime, these values may be used in expression (5.2.1.5), along with an appropriately chosen value of  $t$ . Using, for example, a 90% confidence level and a two-sided interval,  $t$  should be chosen as the .95 quantile of the  $t$  distribution with  $\nu = n - 1 = 9$  degrees of freedom. That is, one uses the  $t_9$  distribution and chooses  $t > 0$  such that

$$P\left[-t < \text{random variable} < t\right] = 0.90$$

Consulting Table A1.2 the choice  $t = 1.833$  is in order. So a two-sided 90% confidence interval for  $\mu$  has endpoints

$$168.3 \pm 1.833 \frac{33.1}{\sqrt{10}}$$

that is

$$168.3 \pm 19.2$$

ie

$$149.1 \times 10^3 \text{ cycles and } 187.5 \times 10^3 \text{ cycles}$$



## CHECKING NORMAL PLOTS

---

As illustrated in Example 5.2.1.2, normal-plotting the data as a rough check on the plausibility of an underlying normal distribution is a sound practice, and one that is used repeatedly in this text. However, it is important not to expect more than is justified from the method. It is certainly preferable to use it rather than making an unexamined leap to a possibly inappropriate normal assumption. But it is also true that when used with small samples, the method doesn't often provide definitive indications as to whether a normal model can be used. Small samples from normal distributions will often have only marginally linear-looking normal plots. At the same time, small samples from even quite nonnormal distributions can often have reasonably linear normal plots. In short, because of sampling variability, small samples don't carry much information about underlying distributional shape. About all that can be counted on from a small-sample preliminary normal plot, like that in Example 5.2.1.2, is a warning in case of gross departure from normality associated with an underlying distributional shape that is much heavier in the tails than a normal distribution (i.e., producing more extreme values than a normal shape would).

## SMALL SAMPLE TESTS FOR $\mu$

---

Example 5.2.1.2 shows the use of the confidence interval formula (5.2.1.5) but not the significance testing method (5.2.1.6). Since the small-sample method is exactly analogous to the large-sample method of Section 5.1 (except for the substitution of the  $t$  distribution for the standard normal distribution), and the source from which the data were taken doesn't indicate any particular value of  $\mu$  belonging naturally in a null hypothesis, the use of the method indicated in expression (5.2.1.6) by itself will not be illustrated at this point.

## 5.2.2 Large-Sample Comparisons of Two Means (Based on Independent Samples)

Methods that can be used to compare two means where two different “unrelated” samples form the basis of inference are studied next, beginning with large-sample methods.

### Example 5.2.2.1 Comparing the Packing Properties of Molded and Crushed Pieces of a Solid

A company research effort involved finding a workable geometry for molded pieces of a solid. One comparison made was between the weight of molded pieces of a particular geometry, that could be poured into a standard container, and the weight of irregularly shaped pieces (obtained through crushing), that could be poured into the same container. A series of 24 attempts to pack both molded and crushed pieces of the solid produced the data (in grams) that are given in Figure 5.2.2.1 in the form of back-to-back stem-and-leaf diagrams.

Notice that although the same number of molded and crushed weights are represented in the figure, there are two distinctly different samples represented. This is in no way comparable to a paired-difference situation treated in another Chapter, and a different method of statistical inference is appropriate.

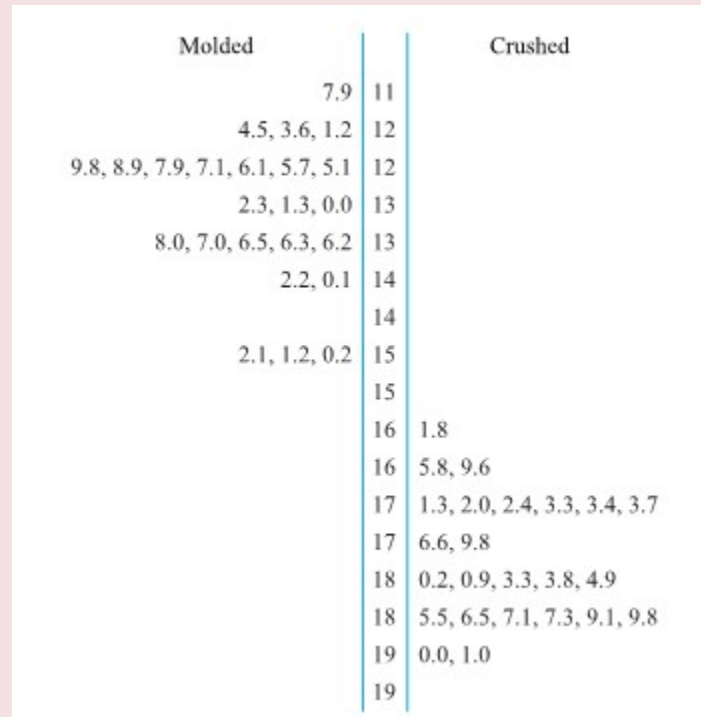


Figure 5.2.2.1 Back-to-back stem-and-leaf plots of packing weights for molded and crushed pieces.

In situations like Example 5.2.2.1, it is useful to adopt subscript notation for both the parameters and the statistics—for example, letting  $\mu_1$  and  $\mu_2$  stand for underlying distributional means corresponding to the first and second conditions and  $\bar{x}_1$  and  $\bar{x}_2$  stand for corresponding sample means. Now if the two data-generating mechanisms are conceptually essentially equivalent to sampling with replacement from two distributions, Part 4 says that  $\bar{x}_1$  has mean  $\mu_1$  and variance  $\sigma_1^2/n_1$  and  $\bar{x}_2$  has mean  $\mu_2$  and variance  $\sigma_2^2/n_2$ . The difference in sample means  $\bar{x}_1 - \bar{x}_2$  is a natural statistic to use in comparing  $\mu_1$  and  $\mu_2$ . Part 4 implies that if it is reasonable to think of the two samples as separately chosen/independent, the random variable has

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

and

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

if, in addition,  $n_1$  and  $n_2$  are large (so that  $\bar{x}_1$  and  $\bar{x}_2$  are each approximately normal),  $\bar{x}_1 - \bar{x}_2$  is approximately normal—i.e.,

**EXPRESSION 5.2.2.1**

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has an approximately standard normal probability distribution.

It is possible to begin with the fact that the variable (5.2.2.1) is approximately standard normal and end up with confidence interval and significance-testing methods for  $\mu_1 - \mu_2$  by using logic exactly parallel to that in the “known- $\sigma$ ” parts of Sections 5.1. But practically, it is far more useful to begin instead with an expression that is free of the parameters  $\sigma_1$  and  $\sigma_2$ . Happily, for large  $n_1$  and  $n_2$ , not only is the variable (5.2.2.1) approximately standard normal but so is

**EXPRESSION 5.2.2.2**

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Then the standard logic of Section 5.1 shows that a two-sided large-sample confidence interval for the difference  $\mu_1 - \mu_2$  based on two independent samples has endpoints

**EXPRESSION 5.2.2.3 Large-sample confidence limits for  $\mu_1 - \mu_2$** 

$$\bar{x}_1 - \bar{x}_2 \pm z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where  $z$  is chosen such that the probability that the standard normal distribution assigns to the interval between  $-z$  and  $z$  corresponds to the desired confidence. And the logic of Section 5.2 shows that under the same conditions,

$$H_0 : \mu_1 - \mu_2 = \#$$

can be tested using the statistic

**EXPRESSION 5.2.2.4 Large-sample test statistic for  $\mu_1 - \mu_2$**

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and a standard normal reference distribution.

**Example 5.2.2 continued.**

In the molding problem, the crushed pieces were a priori expected to pack better than the molded pieces (that for other purposes are more convenient). Consider testing the statistical significance of the difference in mean weights and also making a 95% one-sided confidence interval for the difference (declaring that the crushed mean weight minus the molded mean weight is at least some number).

The sample sizes here ( $n_1 = n_2 = 24$ ) are borderline for being called large. It would be preferable to have a few more observations of each type. Lacking them, we will go ahead and use the methods of expressions (5.2.2.3) and (5.2.2.4) but remain properly cautious of the results should they in any way produce a “close call” in engineering or business terms.

Arbitrarily labeling “crushed” condition 1 and “molded” condition 2 and calculating from the data in Figure 5.2.2.2 that  $\bar{x}_1 = 179.55$  g,  $s_1 = 8.34$  g,  $\bar{x}_2 = 132.97$  g, and  $s_2 = 9.31$  g, the five-step testing format produces the following summary:

1.  $H_0 : \mu_1 - \mu_2 = 0$

2.  $H_a : \mu_1 - \mu_2 > 0$

(The research hypothesis here is that the crushed mean exceeds the molded mean so that the difference, taken in this order, is positive.)

3. The test statistic is

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The reference distribution is standard normal, and large observed values  $z$  will constitute evidence against  $H_0$  and in favor of  $H_a$ .

4. The samples give

$$z = \frac{179.55 - 132.97 - 0}{\sqrt{\frac{(8.34)^2}{24} + \frac{(9.31)^2}{24}}} = 18.3$$

5. The observed level of significance is  $P[\text{a standard normal variable} \geq 18.3] \approx 0$ . The data present overwhelming evidence that  $\mu_1 - \mu_2 > 0$ . i.e., that the mean packed weight of crushed pieces exceeds that of the molded pieces.

Then turning to a one-sided confidence interval for  $\mu_1 - \mu_2$ , note that only the lower endpoint given in display (5.2.2.3) will be used. So  $z = 1.645$  will be appropriate. That is, with 95% confidence, we conclude that the difference in means (crushed minus molded) exceeds

$$(179.55 - 132.97) - 1.645 \sqrt{\frac{(8.34)^2}{24} + \frac{(9.31)^2}{24}}$$

i.e., exceeds

$$46.58 - 4.20 = 42.38 \text{ g}$$

Or differently put, a 95% one-sided confidence interval for  $\mu_1 - \mu_2$  is

$$(42.38, \infty)$$

Students are sometimes uneasy about the arbitrary choice involved in labeling the two conditions in a two-sample study. The fact is that either one can be used. As long as a given choice is followed through consistently, the real-world conclusions reached will be completely unaffected by the choice. In Example 5.5.2.2, if the molded condition is labeled number 1 and the crushed condition number 2, an appropriate one-sided confidence for the molded mean minus the crushed mean is

$$(-\infty, -42.38)$$

This has the same meaning in practical terms as the interval in the example.

Remember that the present methods apply where single measurements are made on each element of two different samples. This stands in contrast to problems of paired data (where there are bivariate observations on a single sample) and which we will study later.

## 5.2.3 Small-Sample Comparisons of Two Means (Based on Independent Samples from Normal Distributions)

The last inference methods presented in this section are those for the difference in two means in cases where at least one of  $n_1$  and  $n_2$  is small. All of the discussion for this problem is limited to cases where observations are normal. And in fact, the most straightforward methods are for cases where, in addition, the two underlying standard deviations are comparable. The discussion begins with these.

### GRAPHICAL CHECK ON THE PLAUSIBILITY OF THE MODEL

A way of making at least a rough check on the plausibility of “normal distributions with a common variance” model assumptions in an application is to normal-plot two samples on the same set of axes, checking not only for approximate linearity but also for approximate equality of slope.

#### Example 5.2.3.1 continued

The data of W. Armstrong on spring lifetimes (appearing in the book by Cox and Oakes) not only concern spring longevity at a 950 N/mm<sup>2</sup> stress level but also longevity at a 900 N/mm<sup>2</sup> stress level. Table 5.2.3.1 repeats the 950 N/mm<sup>2</sup> data from before and gives the lifetimes of ten springs at the 900 N/mm<sup>2</sup> stress level as well.

Spring Lifetimes under Two Different Levels of Stress (10 <sup>3</sup> cycles)	
950 N/mm <sup>2</sup> Stress	900 N/mm <sup>2</sup> Stress
225, 171, 198, 189, 189	216, 162, 153, 216, 225
135, 162, 135, 117, 162	216, 306, 225, 243, 189

Table 5.2.3.1

Figure 5.2.3.1 consists of normal plots for the two samples made on a single set of axes. In light of the kind of variation in linearity and slope exhibited by the normal plots for samples of this size ( $n = 10$ ) from a single normal distribution, there is certainly no strong evidence in Figure 5.2.3.1 against the appropriateness of an “equal variances, normal distributions” model for spring lifetimes.

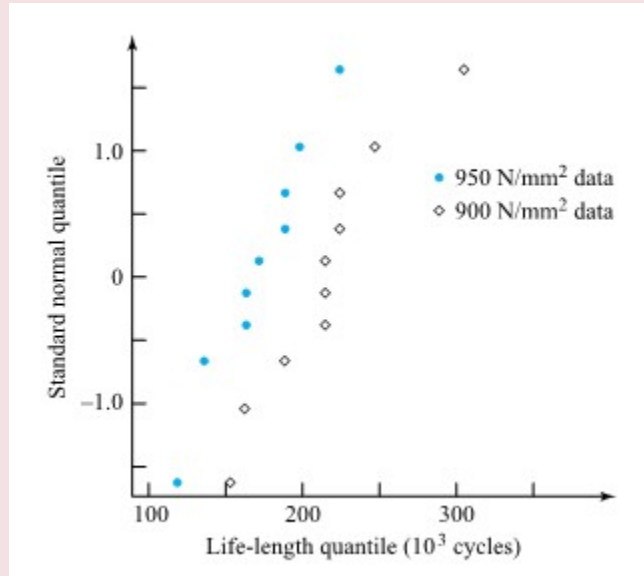


Figure 5.2.3.1 Normal plots of spring lifetimes under two different levels of stress

## POOLED SAMPLE VARIANCE

If the assumption that  $\sigma_1 = \sigma_2$  is used, then the common value is called  $\sigma$ , and it makes sense that both  $s_1$  and  $s_2$  will approximate  $\sigma$ . That suggests that they should somehow be combined into a single estimate of the basic, baseline variation. As it turns out, mathematical convenience dictates a particular method of combining or *pooling* the individual  $s$ 's to arrive at a single estimate of  $\sigma$ .

**DEFINITION pooled sample variance  $sp^2$**

**EXPRESSION 5.2.3.1**

If two numerical samples of respective sizes  $n_1$  and  $n_2$  produce respective sample variances  $s_1^2$  and  $s_2^2$ , the pooled sample variance,  $sp^2$ , is the weighted average of  $s_1^2$  and  $s_2^2$  where the weights are the sample sizes minus 1. That is,

$$s_P^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

The pooled sample standard deviation,  $s_P$ , is the square root of  $sp^2$ .



$s_P$  is a kind of average of  $s_1$  and  $s_2$  that is guaranteed to fall between the two values  $s_1$  and  $s_2$ . Its exact form is dictated more by considerations of mathematical convenience than by obvious intuition.

#### Example 5.2.3.2 continued

In the spring-life case, making the arbitrary choice to call the  $900 \text{ N/mm}^2$  stress level condition 1 and the  $950 \text{ N/mm}^2$  stress level condition 2,  $s_1 = 42.9(10^3 \text{ cycles})$  and  $s_2 = 33.1(10^3 \text{ cycles})$ . So pooling the two sample variances via formula (5.2.3.1) produces

$$s_P^2 = \frac{(10 - 1)(42.9)^2 + (10 - 1)(33.1)^2}{(10 - 1) + (10 - 1)} = 1,468(10^3 \text{ cycles})^2$$

Then, taking the square root,

$$s_P = \sqrt{1,468} = 38.3(10^3 \text{ cycles})$$

In the argument leading to large-sample inference methods for  $\mu_1 - \mu_2$ , the quantity given in the expression

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

was briefly considered. In the  $\sigma_1 = \sigma_2 = \sigma$  context, this can be rewritten as

$$5.2.3.3 \quad Z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

One could use the fact that expression (5.2.3.3) is standard normal to produce methods for confidence interval estimation and significance testing. But for use, these would require the input of the parameter  $\sigma$ . So instead of beginning with expression (5.2.3.3), it is standard to replace  $\sigma$  in expression (5.2.3.3) with  $s_P$  and begin with the quantity

$$5.2.3.4 \quad T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Expression (5.2.3.4) is crafted exactly so that under the present model assumptions, the variable (5.2.3.4) has a well-known, tabled probability distribution: the t distribution with  $v = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$  degrees of freedom. (Notice that the  $n_1 - 1$  degrees of freedom associated with the first sample add together with the  $n_2 - 1$  degrees of freedom associated with the second to produce  $n_1 + n_2 - 2$  overall.) This probability fact, again via the kind of reasoning developed in Sections 5.1 and 5.2, produces inference methods for  $\mu_1 - \mu_2$ . That is, a two-sided confidence interval for the difference  $\mu_1 - \mu_2$ , based on independent samples from normal distributions with a common variance, has endpoints

**EXPRESSION 5.2.3.5 Normal distributions ( $\sigma_1 = \sigma_2$ ) confidence limits for  $\mu_1 - \mu_2$** 

$$\bar{x}_1 - \bar{x}_2 \pm t_{SP} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where  $t$  is chosen such that the probability that the  $t_{n_1+n_2-2}$  distribution assigns to the interval between  $-t$  and  $t$  corresponds to the desired confidence. And under the same conditions,

$$H_0 : \mu_1 - \mu_2 = \#$$

can be tested using the statistic

**EXPRESSION 5.2.3.6 Normal distributions ( $\sigma_1 = \sigma_2$ ) test statistic for  $\mu_1 - \mu_2$** 

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \#}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and a  $t_{n_1+n_2-2}$  reference distribution.

**Example 5.2.3.3 continued**

We return to the spring-life case to illustrate small-sample inference for two means. First consider testing the hypothesis of equal mean lifetimes with an alternative of increased lifetime accompanying a reduction in stress level. Then consider making a two-sided 95% confidence interval for the difference in mean lifetimes.

Continuing to call the 900 N/mm<sup>2</sup> stress level condition 1 and the 950 N/mm<sup>2</sup> stress level condition 2, from Table 5.3.3.1  $\bar{x}_1 = 215.1$  and  $\bar{x}_2 = 168.3$ , while (from before)  $s_P = 38.3$ . The five-step significance-testing format then gives the following:

1.  $H_0 : \mu_1 - \mu_2 = 0$
2.  $H_a : \mu_1 - \mu_2 > 0$ .

(The engineering expectation is that condition 1 produces the larger life-times.)

3. The test statistic is

$$T = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

The reference distribution is  $t$  with  $10 + 10 - 2 = 18$  degrees of freedom, and large observed  $t$  will count as evidence against  $H_0$ .

4. The samples give

$$t = \frac{215.1 - 168.3 - 0}{38.3 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 2.7$$

5. The observed level of significance is  $P$  [a  $t_{18}$  random variable  $\geq 2.7$ ], which (according to Table A1.2) is between .01 and .005. This is strong evidence that the lower stress level is associated with larger mean spring lifetimes.

Then, if the expression (5.5.3.5) is used to produce a two-sided 95% confidence interval, the choice of  $t$  as the .975 quantile of the  $t_{18}$  distribution is in order. Endpoints of the confidence interval for  $\mu_1 - \mu_2$  are

$$(215.1 - 168.3) \pm 2.101(38.3) \sqrt{\frac{1}{10} + \frac{1}{10}}$$

that is

$$46.8 \pm 36.0$$

that is

$$10.8 \times 10^3 \text{ cycles and } 82.8 \times 10^3 \text{ cycles}$$

The data in Table 5.2.3.1 provide enough information to establish convincingly that increased stress is associated with reduced mean spring life. But although the apparent size of that reduction when moving from the  $900 \text{ N/mm}^2$  level (condition 1) to the  $950 \text{ N/mm}^2$  level (condition 2) is  $46.8 \times 10^3$  cycles, the variability present in the data is large enough (and the sample sizes small enough) that only a precision of  $\pm 36.0 \times 10^3$  cycles can be attached to the figure  $46.8 \times 10^3$  cycles.

## SMALL-SAMPLE INFERENCE FOR $\mu_1 - \mu_2$ WITHOUT THE $\sigma_1 = \sigma_2$ ASSUMPTION

There is no completely satisfactory answer to the question of how to do inference for  $\mu_1 - \mu_2$  when it is not sensible to assume that  $\sigma_1 = \sigma_2$ . The most widely accepted (but approximate) method for the problem is one due to Satterthwaite that is related to the large-sample formula (from 5.2.1). That is, while endpoints (from 5.2.1) are not appropriate when  $n_1$  or  $n_2$  is small (they don't produce actual confidence levels near the nominal one), a modification of them is appropriate. Let

**EXPRESSION 5.3.3.7 Satterthwaite's "estimated degrees of freedom"**

$$\hat{\nu} = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{(n_1-1)n_1^2} + \frac{s_2^4}{(n_2-1)n_2^2}}$$

and for a desired confidence level, suppose that  $\hat{t}$  is such that the  $t$ -distribution with  $\hat{\nu}$  degrees of freedom assigns that probability to the interval between  $-\hat{t}$  and  $\hat{t}$ . Then the two endpoints

**EXPRESSION 5.2.3.8 Satterthwaite (approximate) normal distribution confidence limits for  $\mu_1 - \mu_2$** 

$$\bar{x}_1 - \bar{x}_2 \pm \hat{t} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

can serve as confidence limits for  $\mu_1 - \mu_2$  with a confidence level approximating the desired one. (One of the two limits (5.2.3.8) may be used as a single confidence bound with the two-sided unconfidence level halved.)

**Example 5.2.3.4 continued**

Armstrong collected spring lifetime data at stress levels besides the 900 and 950 N/mm<sup>2</sup> levels used thus far in this example. Ten springs tested at 850 N/mm<sup>2</sup> had lifetimes with  $\bar{x} = 348.1$  and  $s = 57.9$  (both in 103 cycles) and a reasonably linear normal plot. But taking the 850, 900, and 950 N/mm<sup>2</sup> data together, there is a clear trend to smaller and more consistent lifetimes as stress is increased. In light of this fact, should mean lifetimes at the 850 and 950 N/mm<sup>2</sup> stress levels be compared, use of a constant variance assumption seems questionable.

Consider then what the Satterthwaite method (5.2.3.8) gives for two-sided approximate 95% confidence limits for the difference in 850 and 950 N/mm<sup>2</sup> mean lifetimes. Equation (5.2.2.7) gives

$$\hat{v} = \frac{\left( \frac{(57.9)^2}{10} + \frac{(33.1)^2}{10} \right)^2}{\frac{(57.9)^4}{9(100)} + \frac{(33.1)^4}{9(100)}} = 14.3$$

and (rounding “degrees of freedom” down) the .975 quantile of the  $t_{14}$  distribution is 2.145. So the 95% limits (5.3.3.8) for the (850 N/mm<sup>2</sup> minus 950 N/mm<sup>2</sup>) difference in mean lifetimes ( $\mu_8 50 - \mu_9 50$ ) are

$$348.1 - 168.3 \pm 2.145 \sqrt{\frac{(57.9)^2}{10} + \frac{(33.1)^2}{10}}$$

that is

$$179.8 \pm 45.2$$

that is

$$134.6 \times 10^3 \text{ cycles and } 225.0 \times 10^3 \text{ cycles}$$

## COMMENTS ON SMALL-SAMPLE METHODS

The inference methods represented in this chapter are the last of the standard one- and two-sample methods for means. We will now look at a parallel methods for variances. But before leaving this section to consider this method, a final comment is appropriate about the small-sample methods.

This discussion has emphasized that, strictly speaking, the nominal properties (in terms of coverage probabilities for confidence intervals and relevant p-value declarations for significance tests) of the small-sample methods depend on the appropriateness of exactly normal underlying distributions and (in the cases of the methods (5.2.3.5) and (5.2.3.6)) exactly equal variances. On the other hand, when actually applying the methods, rather crude probability-plotting checks have been used for verifying (only) that the models are roughly plausible. According to conventional statistical wisdom, the small-sample methods presented here are remarkably robust to all but gross departures from the model assumptions. That is, as long as the model assumptions are at least roughly a description of reality, the nominal confidence levels and p-values will not be ridiculously incorrect. (For example, a nominally 90% confidence interval method might in reality be only an 80% method, but it will not be only a 20% confidence interval method.) So the kind of plotting that has been illustrated here is often taken as adequate precaution against unjustified application of the small-sample inference methods for means.

## 5.2.4 Two-Sample Inference for Variances

### INFERENCE FOR THE RATIO OF TWO VARIANCES (BASED ON INDEPENDENT SAMPLES FROM NORMAL DISTRIBUTIONS)

To move from inference for a single variance to inference for comparing two variances requires the introduction of yet another new family of probability distributions: (Snedecor's) F distributions.

#### DEFINITION 5.2.4.1 F Distribution

#### EXPRESSION 5.2.4.1

The (Snedecor)  $F$  distribution with numerator and denominator degrees of freedom parameters  $\nu_1$  and  $\nu_2$  is a continuous probability distribution with probability density

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1/2)-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{(\nu_1+\nu_2)/2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

If a random variable has the probability density given by formula (5.2.4.1), it is said to have the  $F_{\nu_1, \nu_2}$  distribution.

As Figure 5.2.4.1 reveals, the F distributions are strongly right-skewed distributions, whose densities achieve their maximum values at arguments somewhat less than 1. Roughly speaking, the smaller the values  $\nu_1$  and  $\nu_2$ , the more asymmetric and spread out is the corresponding F distribution.

#### Using the F distribution tables, Table A3

Direct use of formula (5.2.4.1) to find probabilities for the  $F$  distributions requires numerical integration methods. For purposes of applying the  $F$  distributions in statistical inference, the typical path is to instead make use of either statistical software or some fairly abbreviated tables of  $F$  distribution quantiles. Table Appendix Tables A3 are tables of  $F$  quantiles. The body of a particular one of these tables, for a single  $p$ , gives the  $F$  distribution  $p$  quantiles for various combinations of  $\nu_1$  (the numerator degrees of freedom) and  $\nu_2$  (the denominator degrees of

freedom). The values of  $\nu_1$  are given across the top margin of the table and the values of  $\nu_2$  down the left margin.

Tables A3 give only  $p$  quantiles for  $p$  larger than .5. Often  $F$  distribution quantiles for  $p$  smaller than .5 are needed as well. Rather than making up tables of such values, it is standard practice to instead make use of a computational trick. By using a relationship between  $F_{\nu_1, \nu_2}$  and  $F_{\nu_2, \nu_1}$  quantiles, quantiles for small  $p$  can be determined. If one lets  $Q_{\nu_1, \nu_2}$  stand for the  $F_{\nu_1, \nu_2}$  quantile function and  $Q_{\nu_2, \nu_1}$  stand for the quantile function for the  $F_{\nu_2, \nu_1}$  distribution,

**EXPRESSION 5.2.4.2 Relationship between  $F_{\nu_1, \nu_2}$  and  $F_{\nu_2, \nu_1}$  quantiles**

$$Q_{\nu_1, \nu_2}(p) = \frac{1}{Q_{\nu_2, \nu_1}(1-p)}$$

Fact (5.2.4.2) means that a small lower percentage point of an distribution may be obtained by taking the reciprocal of a corresponding small upper percentage point of the distribution with degrees of freedom reversed.

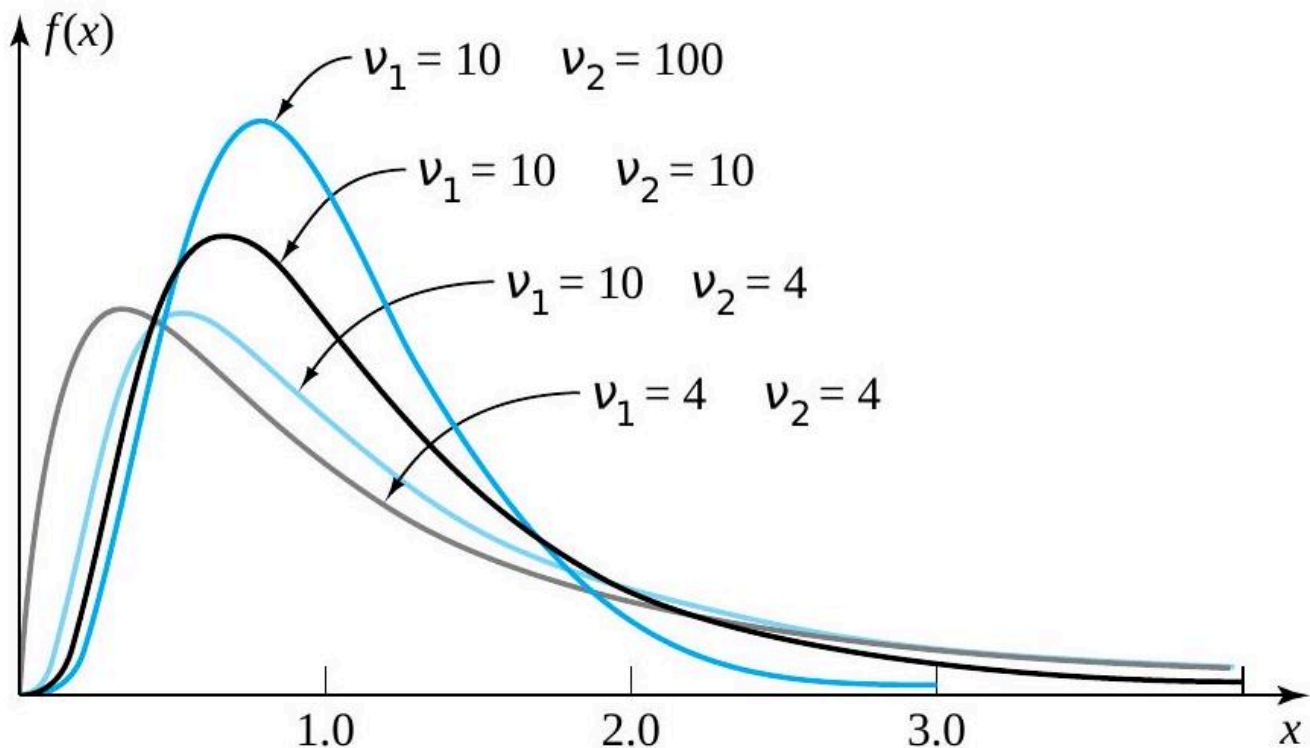


Figure 5.2.4.1 Four different  $F$  probability densities

**Example 5.2.4.1** Use of Tables of Distribution Quantiles

Suppose that  $V$  is an  $F_{3,5}$  random variable. Consider finding the .95 and .01 quantiles of  $V$ 's distribution and then seeing what Tables A3 reveal about  $P[V > 4.0]$  and  $P[V < .3]$ .

First, a direct look-up in the  $p = .95$  table of quantiles, in the  $v_1 = 3$  column and  $v_2 = 5$  row, produces the number 5.41. That is,  $Q(.95) = 5.41$ , or (equivalently)  $P[V < 5.41] = .95$ .

To find the  $p = .01$  quantile of the  $F_{3,5}$  distribution, expression (5.2.4.2) must be used. That is,

$$Q_{3,5}(.01) = \frac{1}{Q_{5,3}(.99)}$$

so that using the  $v_1 = 5$  column and  $v_2 = 3$  row of the table of  $F_{.99}$  quantiles, one has

$$Q_{3,5}(.01) = \frac{1}{28.24} = .04$$

Next, considering  $P[V > 4.0]$ , one finds (using the  $v_1 = 3$  columns and  $v_2 = 5$  rows of Tables A3) that 4.0 lies between the .90 and .95 quantiles of the  $F_{3,5}$  distribution. That is,

$$.90 < P[V \leq 4.0] < .95$$

so that

$$.05 < P[V > 4.0] < .10$$

Finally, considering  $P[V < .3]$ , note that none of the entries in Tables A3 is less than 1.00. So to place the value 3 in the  $F_{3,5}$  distribution, one must locate its reciprocal,  $3.33 (= 1/.3)$ , in the  $F_{5,3}$  distribution and then make use of expression (5.2.4.2). Using the  $v_1 = 5$  columns and  $v_2 = 3$  rows of Tables A3, one finds that 3.33 is between the .75 and .90 quantiles of the  $F_{5,3}$  distribution. So by expression (5.2.4.2), .3 is between the .1 and .25 quantiles of the  $F_{3,5}$  distribution, and

$$.10 < P[V < 0.3] < 0.25$$

The extra effort required to find small F distribution quantiles is an artifact of standard table-making practice, rather than being any intrinsic extra difficulty associated with the F distributions. One way to eliminate the difficulty entirely is to use standard statistical software or a statistical calculator to find F quantiles.

The F distributions are of use here because a probability fact ties the behavior of ratios of independent sample variances based on samples from normal distributions to the variances  $\sigma_1^2$  and  $\sigma_2^2$  of those underlying distributions. That is, when  $s_1^2$  and  $s_2^2$  come from independent samples from normal distributions, the variable  $F = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2}$  has  $n_1 - 1$  associated degrees of freedom and is in the numerator of

has an  $F_{n_1-1, n_2-1}$  distribution. ( $s_1^2$  has  $n_1 - 1$  associated degrees of freedom and is in the numerator of



this expression, while  $s_2^2$  has  $n_2 - 1$  associated degrees of freedom and is in the denominator, providing motivation for the language introduced in Definition 5.2.4.1)

This fact is exactly what is needed to produce formal inference methods for the ratio  $\sigma_1^2/\sigma_2^2$ . For example, it is possible to pick appropriate F quantiles L and U such that the probability that the variable (5.2.4.3) falls between L and U corresponds to a desired confidence level. (Typically, L and U are chosen to "split the 'unconfidence' " between the upper and lower  $F_{n_1-1, n_2-1}$  tails.) But

$$L < \frac{s_1^2}{\sigma_1^2} \cdot \frac{\sigma_2^2}{s_2^2} < U$$

is algebraically equivalent to

$$\frac{1}{U} \cdot \frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{L} \cdot \frac{s_1^2}{s_2^2}$$

That is, when a data-generating mechanism can be thought of as essentially equivalent to independent random sampling from two normal distributions, a two-sided confidence interval for  $\sigma_1^2/\sigma_2^2$  has endpoints

#### 5.2.4.4 Normal distributions confidence limits for $\sigma_1^2/\sigma_2^2$

$$\frac{s_1^2}{U \cdot s_2^2} \quad \text{and} \quad \frac{s_1^2}{L \cdot s_2^2}$$

where L and U are ( $F_{n_1-1, n_2-1}$  quantiles) such that the  $F_{n_1-1, n_2-1}$  probability assigned to the interval (L, U) corresponds to the desired confidence.

In addition, there is an obvious significance-testing method for  $\sigma_1^2/\sigma_2^2$ . That is, subject to the same modeling limitations as needed to support the confidence interval method,

#### 5.2.4.5

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \#$$

can be tested using the statistic

#### 5.2.4.6 Normal distributions test statistic for $\sigma_1^2/\sigma_2^2$

$$F = \frac{s_1^2/s_2^2}{\#}$$

and an  $F_{n_1-1, n_2-1}$  reference distribution. (The choice of  $\# = 1$  in displays (5.2.4.5) and (5.2.4.6), so that the null hypothesis is one of equality of variances, is the only one commonly used in practice.)

*P-values for testing*  $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = \#$

*p-values for the one-sided alternative hypotheses*  $H_a : \sigma_1^2/\sigma_2^2 < \#$  and  $H_a : \sigma_1^2/\sigma_2^2 > \#$  are

(respectively) the left and right  $F_{n_1-1, n_2-1}$  tail areas beyond the observed values of the test statistic. For the two-sided alternative hypothesis  $H_a : \sigma_1^2 / \sigma_2^2 \neq \#$ , the standard convention is to report twice the  $F_{n_1-1, n_2-1}$  probability to the right of the observed  $f$  if  $f > 1$  and to report twice the  $F_{n_1-1, n_2-1}$  probability to the left of the observed  $f$  if  $f < 1$ .

#### Example 5.2.4.2 Comparing Uniformity of Hardness Test Results for Two Types of Steel

Condon, Smith, and Woodford did some hardness testing on specimens of 4% carbon steel. Part of their data are given in Table 5.2.4.1, where Rockwell hardness measurements for ten specimens from a lot of heat-treated steel specimens and five specimens from a lot of cold-rolled steel specimens are represented.

Consider comparing measured hardness uniformity for these two steel types (rather than mean hardness, as might have been done in Chapter 5.2.3). Figure 5.2.4.2 shows side-by-side dot diagrams for the two samples and suggests that there is a larger variability associated with the heat-treated specimens than with the cold-rolled specimens. The two normal plots in Figure 5.2.4.3 indicate no obvious problems with a model assumption of normal underlying distributions.

Heat-Treated	Cold-Rolled
32.8, 44.9, 34.4, 37.0, 23.6,	21.0, 24.5, 19.9, 14.8, 18.8
29.1, 39.5, 30.1, 29.2, 19.2	

Table 5.2.4.1 Rockwell Hardness Measurements for Steel Specimens of Two Types

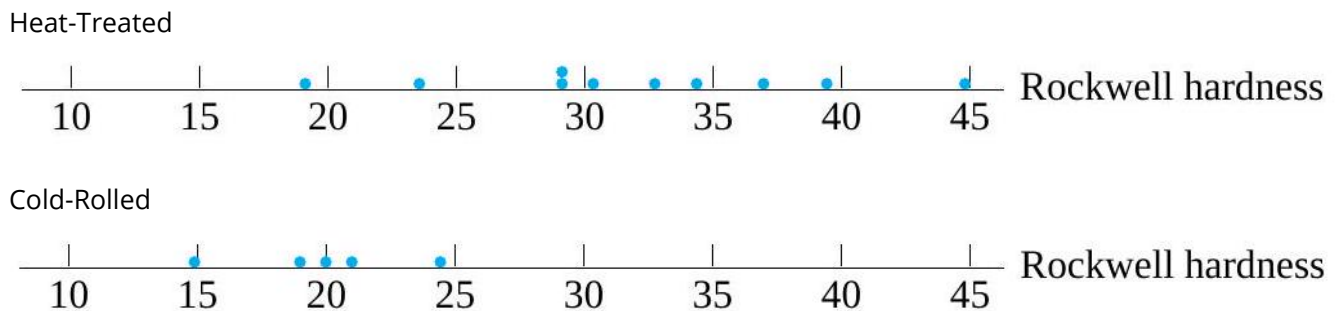


Figure 5.2.4.2 Dot diagrams of hardness for heat-treated and cold-rolled steels

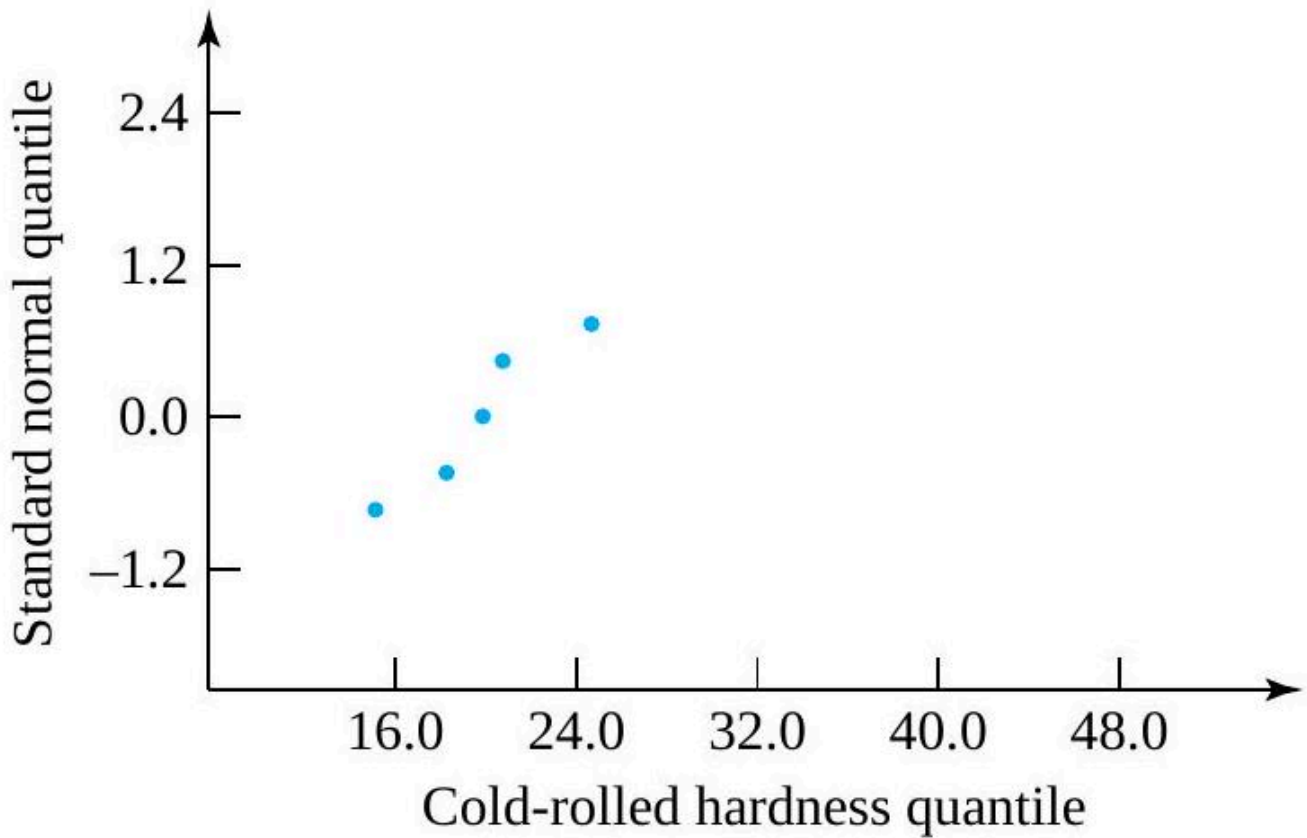
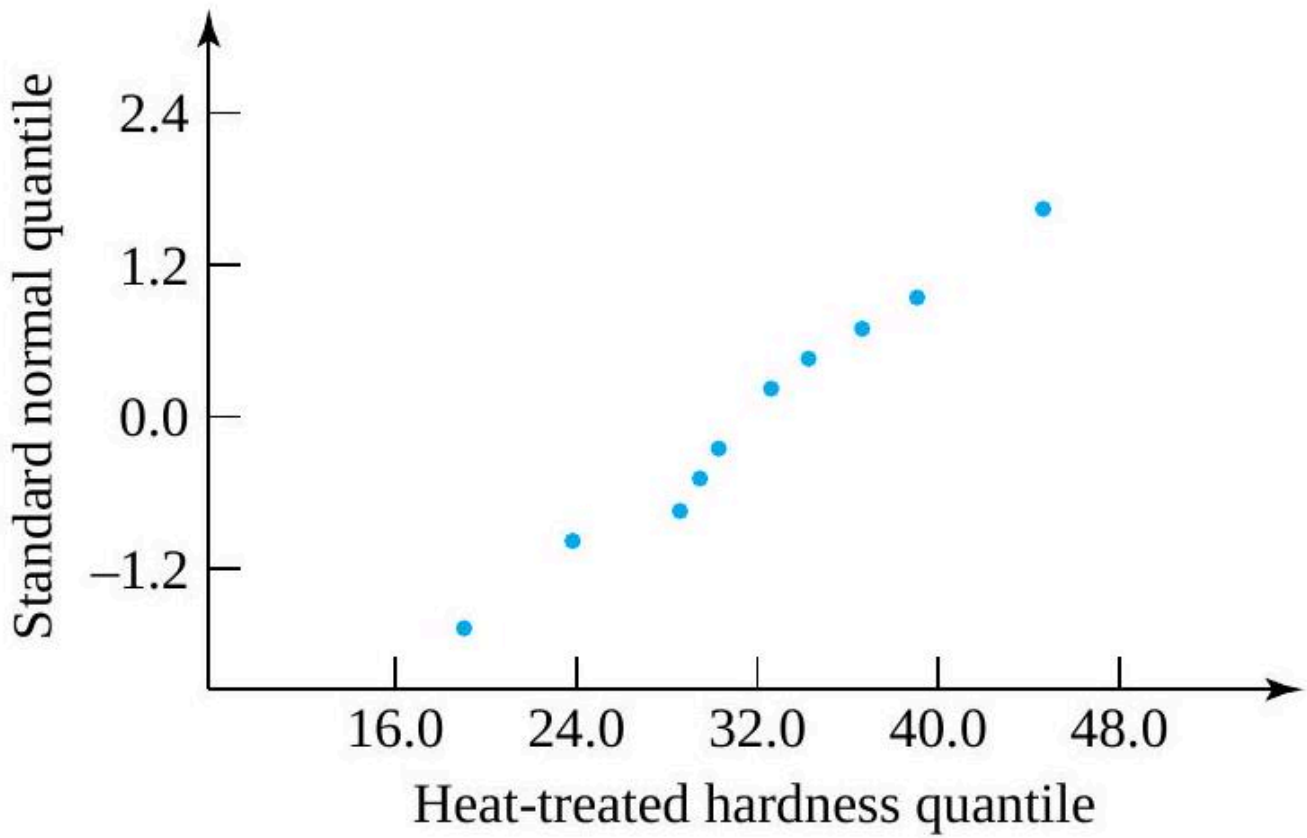


Figure 5.2..4.3 Normal plots of hardness for heat-treated and cold-rolled steels

Then, arbitrarily choosing to call the heat-treated condition number 1 and the cold-rolled condition 2,  $s_1 = 7.52$  and  $s_2 = 3.52$ , and a five-step significance test of equality of variances based on the variable (5.2.4.6) proceeds as follows:

$$1. ]H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$2. H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

(If there is any materials-related reason to pick a one-sided alternative hypothesis here, the authors don't know it.)

3. The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

The reference distribution is the  $F_{9,4}$  distribution, and both large observed  $f$  and small observed  $f$  will constitute evidence against  $H_0$ .

4. The samples give

$$f = \frac{(7.52)^2}{(3.52)^2} = 4.6$$

5. Since the observed  $f$  is larger than 1, for the two-sided alternative, the  $p$ -value is

$$2P [ \text{an } F_{9,4} \text{ random variable } \geq 4.6 ]$$

From Tables A3, 4.6 is between the  $F_{9,4}$  distribution .9 and .95 quantiles, so the observed level of significance is between .1 and .2. This makes it moderately (but not completely) implausible that the heat-treated and cold-rolled variabilities are the same.

In an effort to pin down the relative sizes of the heat-treated and cold-rolled hardness variabilities, the square roots of the expressions in display (4.2.4.6) may be used to give a 90 % two-sided confidence interval for  $\sigma_1/\sigma_2$ . Now the .95 quantile of the  $F_{9,4}$  distribution is 6.0, while the .95 quantile of the  $F_{4,9}$  distribution is 3.63, implying that the .05 quantile of the  $F_{9,4}$  distribution is  $\frac{1}{3.63}$ . Thus, a 90 % confidence interval for the ratio of standard deviations  $\sigma_1/\sigma_2$  has endpoints

$$\sqrt{\frac{(7.52)^2}{6.0(3.52)^2}} \text{ and } \sqrt{\frac{(7.52)^2}{(1/3.63)(3.52)^2}}$$

That is

0.87 and 4.07

The fact that the interval (.87, 4.07) covers values both smaller and larger than 1 indicates that the data in hand do not provide definitive evidence even as to which of the two variabilities in material hardness is larger.

One of the most important engineering applications of the inference methods represented by these expressions are in the comparison of inherent precisions for different pieces of equipment and for different methods of operating a single piece of equipment.

#### Example 4.2.4.3 Comparing Uniformities of Operation of Two Ream Cutters

Abassi, Afinson, Shezad, and Yeo worked with a company that cuts rolls of paper into sheets. The uniformity of the sheet lengths is important, because the better the uniformity, the closer the average sheet length can be set to the nominal value without producing undersized sheets, thereby reducing the company's giveaway costs. The students compared the uniformity of sheets cut on a ream cutter having a manual brake to the uniformity of sheets cut on a ream cutter that had an automatic brake. The basis of that comparison was estimated standard deviations of sheet lengths cut by the two machines—just the kind of information used to frame formal inferences in this section. The students estimated  $\sigma_{\text{manual}} / \sigma_{\text{automatic}}$  to be on the order of 1.5 and predicted a period of two years or less for the recovery of the capital improvement cost of equipping all the company's ream cutters with automatic brakes.

## CAVEATS ABOUT INFERENCES FOR VARIANCE

The methods of this section are, strictly speaking, normal distribution methods. It is worthwhile to ask, "How essential is this normal distribution restriction to the predictable behavior of these inference methods for one and two variances?" There is a remark at the end of Module 5.2.3 to the effect that the methods presented there for means are fairly robust to moderate violation of the section's model assumptions. Unfortunately, such is not the case for the methods for variances presented here.

These are methods whose nominal confidence levels and p-values can be fairly badly misleading unless the normal models are good ones. This makes the kind of careful data scrutiny that has been implemented in the examples (in the form of normal-plotting) essential to the responsible use of the methods of this section. And it suggests that since normal-plotting itself isn't typically terribly revealing unless the sample size involved is moderate to large, formal inferences for variances will be most safely made on the basis of moderate to large normal-looking samples.

The importance of the "normal distribution(s)" restriction to the predictable operation of the methods of this section is not the only reason to prefer large sample sizes for inferences on variances. A little experience with the formulas in this section will convince the reader that (even granting the appropriateness of normal models) small samples often do not prove adequate to answer practical questions about variances. Confidence intervals for variances and variance ratios based on small samples can be so big as to be of little practical value, and the engineer will typically be driven to large sample sizes in order to solve variance-related real-world problems. This is not in any way a failing of the present methods. It is simply a warning and quantification of the fact that learning about variances requires more data than (for example) learning about means.

## 5.2.5 Inference for the Mean of Paired Differences

An important type of application of the methods of confidence interval estimation and significance testing is to paired data. In many engineering problems, it is natural to make two measurements of essentially the same kind, but differing in timing or physical location, on a single sample of physical objects. The goal in such situations is often to investigate the possibility of consistent differences between the two measurements.

### Example 5.2.5.1 Comparing Leading-Edge and Trailing-Edge Measurements on a Shaped Wood Product

Drake, Hones, and Mulholland worked with a company on the monitoring of the operation of an end-cut router in the manufacture of a wood product. They measured a critical dimension of a number of pieces of a particular type as they came off the router. Both a leading-edge and a trailing-edge measurement were made on each piece. The design for the piece in question specified that both leading-edge and trailing-edge values were to have a target value of .172 in. Table 5.2.5.1 gives leading- and trailing-edge measurements taken by the students on five consecutive pieces.

In this situation, the correspondence between leading- and trailing-edge dimensions was at least as critical to proper fit in a later assembly operation as was the conformance of the individual dimensions to the nominal value of .172 in. This was thus a paired-data situation, where one issue of concern was the possibility of a consistent difference between leading- and trailing-edge dimensions that might be traced to a machine misadjustment or unwise method of router operation.

Piece	Leading-Edge Measurement (in.)	Trailing-Edge Measurement (in.)
1	.168	.169
2	.170	.168
3	.165	.168
4	.165	.168
5	.170	.169

Table 5.2.5.1 Leading-Edge and Trailing-Edge Dimensions for Five Workpieces

In situations like Example 5.2.5.1, one simple method of investigating the possibility of a consistent difference between paired data is to first reduce the two measurements on each physical object to a single difference between them. Then the methods of confidence interval estimation and significance testing studied thus far may be applied to the differences. That is, after reducing paired data to differences  $d_1, d_2, \dots, d_n$ , if  $n$  (the number of data pairs) is large, endpoints of a confidence interval for the underlying mean difference,  $\mu_d$ , are

#### 5.2.5.1 Large-sample confidence limits for $\mu_d$

$$\bar{d} \pm z \frac{s_d}{\sqrt{n}}$$

where  $s_d$  is the sample standard deviation of  $d_1, d_2, \dots, d_n$ . Similarly, the null hypothesis

$$5.2.5.2 \quad H_0 : \mu_d = \#$$

can be tested using the test statistic

### 5.2.5.3 Large-sample test statistic for $\mu_d$

$$Z = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}}$$

and a standard normal reference distribution.

If  $n$  is small, in order to come up with methods of formal inference, an underlying normal distribution of differences must be plausible. If that is the case, a confidence interval for  $\mu_d$  has endpoints

### 5.2.5.4 Normal distribution confidence limits for $\mu_d$

$$\bar{d} \pm t \frac{s_d}{\sqrt{n}}$$

and the null hypothesis (5.2.5.2) can be tested using the test statistic

### 5.2.5.5 Normal distribution test statistic for $\mu_d$

$$T = \frac{\bar{d} - \#}{\frac{s_d}{\sqrt{n}}}$$

and a  $t_{n-1}$  reference distribution.

#### Example 5.2.5.2 continued

To illustrate this method of paired differences, consider testing the null hypothesis  $H_0: \mu_d = 0$  and making a 95% confidence interval for any consistent difference between leading- and trailing-edge dimensions,  $\mu_d$ , based on the data in Table 5.2.5.1

Begin by reducing the  $n = 5$  paired observations in Table 5.2.5.1 to differences

$$d = \text{leading-edge dimension} - \text{trailing-edge dimension}$$

appearing in Table 5.2.5.2. Figure 5.2.5.1 is a normal plot of the  $n = 5$  differences in Table 5.2.5.2. A little experimenting with normal plots

of simulated samples of size  $n = 5$  from a normal distribution will convince you that the lack of linearity in Figure 5.2.5.1 would in no way be atypical of normal data. This, together with the fact that normal distributions are very often appropriate for describing machined dimensions of mass-produced parts, suggests the conclusion that the methods represented by expressions 5.2.5.4 and 5.2.5.5 are in order in this example.

The differences in Table 6.6 have  $\hat{d} = -.0008$  in. and  $s_d = .0023$  in. So, first investigating the plausibility of a “no consistent difference” hypothesis in a five-step significance testing format, gives the following:

1.  $H_0: \mu_d = 0$ .

2.  $H_a: \mu_d \neq 0$ .

(There is a priori no reason to adopt a one-sided alternative hypothesis.)

3. The test statistic will be

$$T = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

The reference distribution will be the t distribution with  $\nu = n - 1 = 4$  degrees of freedom. Large observed  $|t|$  will count as evidence against  $H_0$  and in favor of  $H_a$ .

4. The sample gives

$$t = \frac{-.0008}{\frac{.0023}{\sqrt{5}}} = -.78$$

5. The observed level of significance is  $P[|a t_4 \text{ random variable}| \geq .78]$ , which can be seen from Table A1.2 to be larger than  $2(.10) = .2$ . The data in hand are not convincing in favor of a systematic difference between leading- and trailing-edge measurements.

Consulting Table A1.2 for the .975 quantile of the  $t_4$  distribution,  $t = 2.776$  is the appropriate multiplier for use in the expression for 95% confidence. That is, a two-sided 95% confidence interval for the mean difference between the leading- and trailing-edge dimensions has endpoints

$$-.0008 \pm 2.776 \frac{.0023}{\sqrt{5}}$$

that is

$$-.0008 \text{ in.} \pm .0029 \text{ in.}$$

that is

$$-.0037 \text{ in. and } .0021 \text{ in.}$$

This confidence interval for  $\mu_d$  implicitly says (since 0 is in the calculated interval) that the observed level of significance for testing  $H_0: \mu_d = 0$  is more than .05 ( $= 1 - .95$ ). Put slightly differently, it is clear from display the calculated CI above that the imprecision



represented by the plus-or-minus part of the expression is large enough to make it believable that the perceived difference,  $\bar{d} = -.0008$ , is just a result of sampling variability.

Piece	$d = \text{Difference in Dimensions (in.)}$	
1	-.001	(= .168 - .169)
2	.002	(= .170 - .168)
3	-.003	(= .165 - .168)
4	-.003	(= .165 - .168)
5	.001	(= .170 - .169)

Table 5.2.5.2 Five Differences in Leading- and Trailing-Edge Measurements

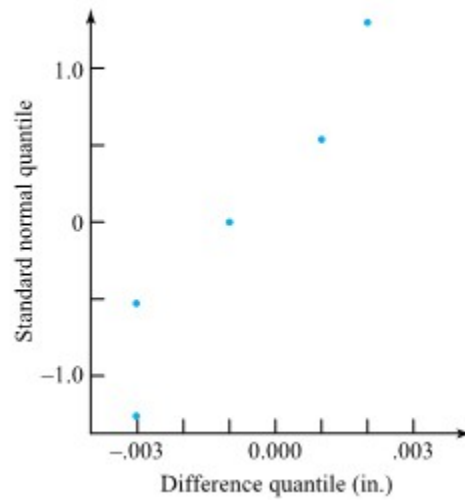


Figure 5.2.5.1 Normal plot of  $n = 5$  differences

### Large-sample inference for $\mu_d$

Example 5.2.5.2 treats a small-sample problem. No example for large  $n$  is included here, because after the taking of differences just illustrated, such an example would reduce to a rehash of things already learned. In fact, since for large  $n$  the  $t$  distribution with  $\nu = n - 1$  degrees of freedom becomes essentially standard normal, one could even imitate Example 5.2.5.2 for large  $n$  and get into no logical problems. So at this point, it makes sense to move on from consideration of the paired-difference method.

### Paired or Unpaired data

This problem of paired data (where there are bivariate observations on a single sample) stands in contrast to the previous problem where methods apply to a single measurements that are made on each element of two different samples. In the woodworking case of Example 5.2.5.2, the data are paired because both

leading-edge and trailing-edge measurements were made on each piece. If leading-edge measurements were taken from one group of items and trailing-edge measurements from another, a two-sample (not a paired difference) analysis would be in order.

## 5.2.6 Tutorial 4A - Inferential Statistics & T-Tests

At this point, it is recommended that you work your way through the [Tutorial 4A exercise](#) found on the associated GitHub repository. This exercise will teach you how to conduct t-tests using Python syntax.

**It is strongly recommended that you consult the [Hypothesis Testing Jupyter Notebook Files](#).** These can be found in the "How do I do X in Python?" section. Specifically the files on "T-tests" and "P-Values" will be particularly useful. The "Confidence Intervals - Difference of Means" file will be useful if looking to compute intervals when comparing multiple groups. Additionally, if you are looking to compute sample size or power calculations, the "Sample Size & Power Calculations" file will be useful.

### 5.3.0 Introduction to Nonparametric Models

The aim of this Module is to discuss the idea of nonparametric statistics. Nonparametric statistics are types of test statistics with related formulas that can be used to estimate associations between two or more variables without basing these associations on changes from the mean. The arithmetic mean can be seriously influenced by extreme values and values that are dispersed in non-normal ways. Essentially if collections of data are not arranged according to the *normal distribution*, and when researchers can be reasonably sure that the actual distribution of variable values in a population is *not* normal, nonparametric statistics can then be used to better estimate associations between variables.

## 5.3.1 Nonparametric Methods

### NON-PARAMETRIC METHODS

---

What can be done when the assumptions we have discussed in past lessons (t-tests, correlation etc.) are not maintained? There are tests used when a number of assumptions are not maintained for regular tests like t-tests or correlations (e.g. nonnormal distribution or small sample sizes). These tests – called non-parametric tests – use the same type of comparisons but with different assumptions.

### PARAMETRIC ASSUMPTIONS

---

Parametric statistics is a branch of statistics that assumes that sample data comes from a population that follows parameters and assumptions that hold true in most, in not all, cases. Most well-known elementary statistical methods are parametric, many of which we have discussed , and which can be found discussed on the Parametric Statistics Wikipedia [webpage](#).

### PARAMETRIC ASSUMPTIONS AND THE NORMAL DISTRIBUTION

---

Normal distribution is a common assumption for many tests, including t-tests, ANOVAs and regression. Recall that parametric tests we have discussed here met the following assumptions of the normal distribution: minimal or no skewness and kurtosis of variables and error terms are independent across variables.

These assumptions allow us to infer a normal distribution in the population.

### NON-PARAMETRIC METHODS

---

Statistical methods which do not require us to make distributional assumptions about the data are called non-parametric methods. Non-parametric, as a term, actually does not apply to the data, but to the method used to analyse the data. These tests use rankings to analyse differences. Non-parametric methods can be used for different types of comparisons or models

### NONPARAMETRIC ASSUMPTIONS

---

1. Nonparametric tests make assumptions about sampling (that it is generally random).
2. There are assumptions about the independence or dependence of samples, depending on which nonparametric test is used, there are no assumptions about the population distribution of scores.

## NONPARAMETRIC TESTS AND LEVEL OF MEASUREMENT

---

Variables at particular categorical levels of measurement may require Nonparametric Tests

Consider variables like autonomy, skill, income. Would such variables always follow a normal distribution? It is possible that when looking at income, you would expect the data to be skewed, as there are a small minority of the population who earn extremely high salaries.

## MEAN VS MEDIAN

---

When a distribution is highly skewed, the mean is affected by the high number of relative outliers. For example, when measuring something like income, where there are few high-income earners but many middle and low-income earners, the center of the distribution is quite skewed. This means that the median (i.e., the middle amount with 50% above and below this amount) is best used.

## SAMPLE SIZE

---

Sample size is another consideration when deciding if one should use a parametric or nonparametric test. Often, researchers will want to run a certain type of parametric test, but might not have the recommended minimum number of participants. Additionally, if the sample is very small, tests of normality often cannot be run. This is due to the lack of power needed to provide an interpretable result. When this is coupled with non-normal distributions of data, researchers might decide to use nonparametric tests.

## OUTLIERS

---

As discussed in previous chapters, parametric tests can only use continuous data for the dependant variable. This data should be normally distributed and not have any spurious outliers. However, some nonparametric tests can use data that is ordinal, or ranked for the dependant variable. These tests may also not be impacted severely by non-normal data or outliers. Each parametric test has its own requirements, so it is advisable to check the assumptions for each test.

## 5.3.2 *Choosing The Appropriate Statistical Test*

### **CHOOSING APPROPRIATE STATISTICAL TESTS**

---

#### **MULTIPLE CONSIDERATIONS REQUIRED**

---

When deciding to use nonparametric statistics, an examination of whether the mean or the median is the best representation of the center of the data distribution is needed. If it is found that the median is the best representation of the data's center, then nonparametric tests are most likely to be appropriate, even with a larger sample of participants. If you have a small sample, then nonparametric statistics may be appropriate either way.

#### **DIFFERENT TESTS**

---

Each parametric test of difference we have discussed previously has a nonparametric equivalent, which can be used in cases where there is nonnormal data or a small sample size.

### 5.3.3 Comparing Two Independent Conditions: The Mann–Whitney U Test

When examining differences between two groups, Mann-Whitney U Test is best. This test examines the differences in median scores, as well as the size of the differences. Example: Is there a difference in the median number of Facebook Friends for male and female internet users? If a researcher wanted to compare Two Related Conditions, the test to use would be the Wilcoxon Signed-Rank Test.

Ranks			
	<i>Gender</i>	N	Mean Rank
<i>FacebookFriends</i>	Male	82	159.46
	Female	285	191.06
	Total	367	

Test Statistics	
	<i>FacebookFriends</i>
<i>Chi-Square</i>	5.65
<i>df</i>	1
<i>Asymp. Sig.</i>	.017

#### INTERPRETATION FOR THE MANN-WHITNEY U TEST

As can be seen in the blue, there is a statistically significant difference, note the p value. The chi-squared value, and degrees of freedom are also needed for reporting. The median ranks indicate that female internet users have more Facebook Friends than male users.

#### WRITE-UP

The results of the Mann-Whitney U Test indicate that female internet users reported having a statistically significantly higher number of Facebook Friends (Median = 191.06) than male users (Median = 159.46; U = 5.65, p = .017).



### 5.3.4 The Wilcoxon Test for Paired Samples

When examining within groups differences, Wilcoxon Signed Ranks Test is best. This test examines the differences in scores, as well as the size of the differences.

Example: The levels of perceived social support a group of Australians reported before engaging with a social skills building program and after completing the program.

Ranks		N	Mean Rank	Sum of Ranks
SocialSupportPre - SocialSupportPost	Negative Ranks	259	184.30	47732.50
	Positive Ranks	68	86.70	5895.50
	Ties	40		
	Total	367		

Test Statistics		SocialSupportPre - SocialSupportPost
Z		-12.24
Asymp. Sig. (2-tailed)		.000

#### INTERPRETATION OF THE WILCOXON TEST

Using the same example from the t-test module, the levels of perceived social support a group of Australians reported before engaging with a social skills building program and after completing the program. As can be seen in red, the Z score, and in green the p value. These indicate that there is a difference in median pre- vs post-test rank score. The scores appear to improve from time 1 to time 2, which we can infer by the negative Z score, and the number of positive ranks in time 2.

**WRITE-UP**

---

An example write up: A Wilcoxon Sign-Rank Test indicated that median post-test ranks for social support were statistically significantly higher than the pre-test ranks ( $Z = -12.24, p < .001$ ).

### 5.3.5 Differences Between Several Independent Groups: The Kruskal–Wallis Test

#### THE KRUSKAL-WALLIS H TEST FOR THREE OR MORE INDEPENDENT SAMPLES

When examining the differences between three or more groups, Kruskal-Wallis H Test is best. This test examines the differences in median scores, as well as the size of the differences. This test examines the main effect of your variable, similar to an ANOVA. Example: Is there a difference in the median reported levels of mental distress for full-time, part-time, and casual employees? If one wanted to compare differences between several related groups, the test to use would be Friedman's ANOVA.

Ranks			
	<i>Are you employed?</i>	N	Mean Rank
<i>MentalDistress</i>	Full-time	161	157.01
	Part-time	83	185.11
	Casual	123	218.59
	Total	367	

Test Statistics	
	<i>MentalDistress</i>
<i>Chi-Square</i>	23.53
<i>df</i>	2
<i>Asymp. Sig.</i>	.000

#### INTERPRETATION OF THE KRUSKAL-WALLIS H TEST

As can be seen in the blue, there is a statistically significant difference, note the p value. The chi-squared value, and degrees of freedom are also needed for reporting. The median ranks indicate that casual employees have the highest scores of mental distress. It is important to note that follow-up tests are required for individual group differences (like Mann-Whitney U Tests), similar to posthoc tests in ANOVA.

**WRITE-UP**

---

A Kruskal-Wallis H test showed that there was a statistically significant difference in levels of mental distress,  $\chi^2(2) = 23.53$ ,  $p < .001$ , for full-time (Median = 157.01), part-time (Median = 185.11), and casual employees (Median = 218.58).

## 5.3.6 Tutorial 4 - Non-Parametric Tests

At this point, it is recommended that you work your way through the [Tutorial 4 exercise](#) found on the associated GitHub repository. This exercise will teach you how to conduct a non-parametric test using Python syntax.

**It is strongly recommended that you consult the [Hypothesis Testing Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “Non-Parametric Tests” will be particularly useful.

## *6.0.1 Introduction to the One-Way Normal Model*

Statistical engineering studies often produce samples taken under not one or two, but rather many different sets of conditions. So although the inference methods of Part 5 are a start, they are not a complete statistical toolkit for engineering problem solving. Methods of formal inference appropriate to multisample studies are also needed.

This section begins to provide such methods. First the reader is reminded of the usefulness of some of the simple graphical tools of Part 2 for making informal comparisons in multisample studies. Next the “equal variances, normal distributions” model is introduced. The role of residuals in evaluating the reasonableness of that model in an application is explained and emphasized. The section then proceeds to introduce the notion of combining several sample variances to produce a single pooled estimate of baseline variation. Finally, there is a discussion of how standardized residuals can be helpful when sample sizes vary considerably.

## 6.0.2 Attributions Part 6

This first draft of Part 6 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

## 6.1.1 Graphical Comparison of Several Samples of Measurement Data

Any thoughtful analysis of several samples of engineering measurement data should begin with the making of graphical representations of those data. Where samples are small, side-by-side dot diagrams are the most natural graphical tool. Where sample sizes are moderate to large (say, at least six or so data points per sample), side-by-side boxplots are effective.

### Example 1 Comparing Compressive Strengths for Eight Different Concrete Formulas

Armstrong, Babb, and Campen did compressive strength testing on 16 different concrete formulas. Part of their data are given in Table 7.1, where eight different formulas are represented. (The only differences between formulas 1 through 8 are their water/cement ratios. Formula 1 had the lowest water/cement ratio, and the ratio increased with formula number in the progression .40, .44, .49, .53, .58, .62, .66, .71. Of course, knowing these water/cement ratios suggests that a curve-fitting analysis might be useful with these data, but for the time being this possibility will be ignored.)

Making side-by-side dot diagrams for these eight samples of sizes  $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = n_7 = n_8 = 3$  amounts to making a scatterplot of compressive strength versus formula number. Such a plot is shown in Figure 6.1.1.1. The general message conveyed by Figure 6.1.1.1 is that there are clear differences in mean compressive strengths between the formulas but that the variabilities in compressive strengths are roughly comparable for the eight different formulas.



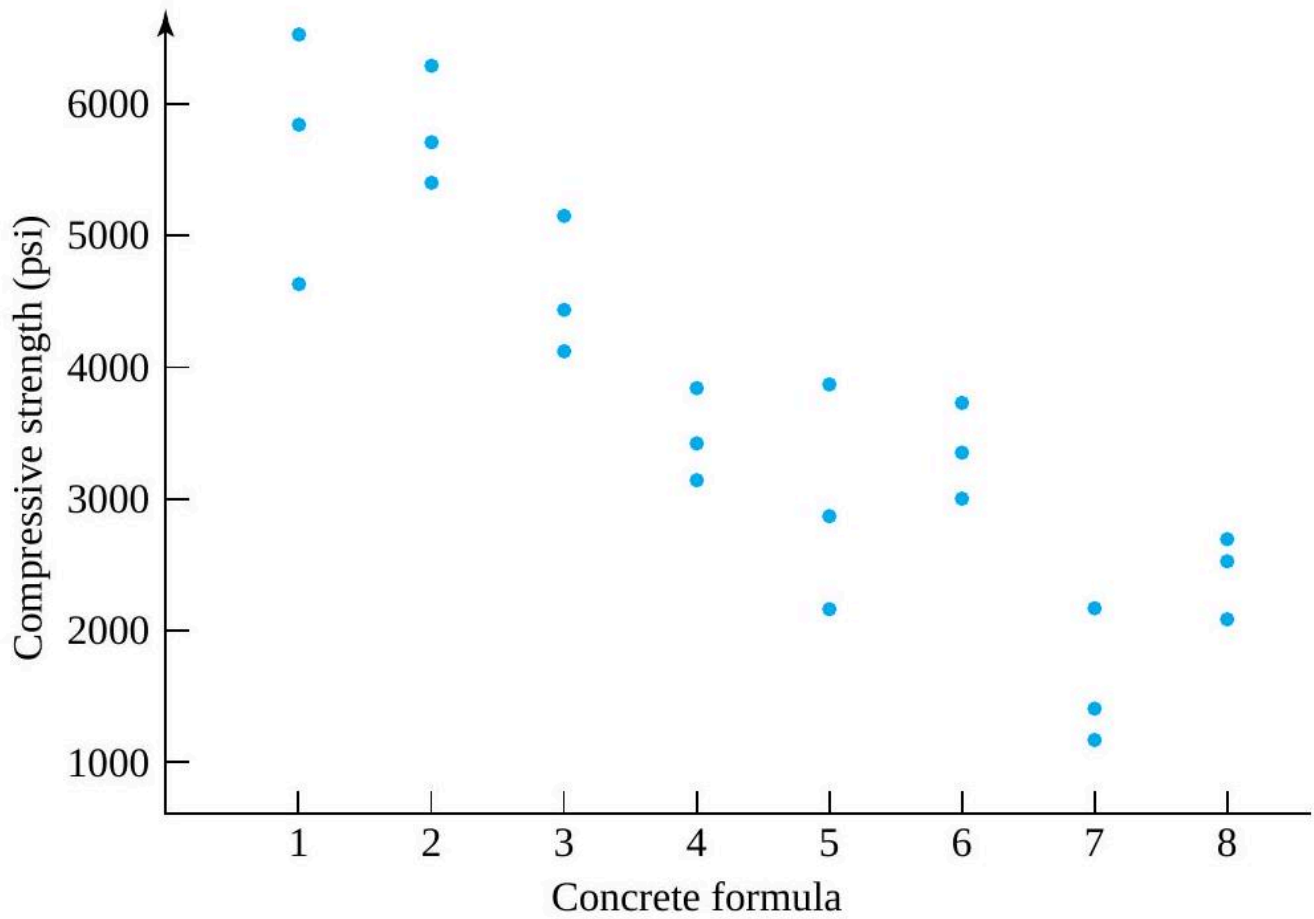


Figure 6.1.1.1 Side-by-side dot diagrams for eight samples of compressive strengths

Compressive Strengths for 24 Concrete Specimens

Specimen	Concrete Formula	28-Day Compressive Strength (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 Compressive Strengths for 24 Concrete Specimens

## Example 6.1.1.2 Comparing Empirical Spring Constants for Three Different Types of Springs

Hunwardsen, Springer, and Wattonville did some testing of three different types of steel springs. They made experimental determinations of spring constants for  $n_1 = 7$  springs of type 1 (a 4 in. design with a theoretical spring constant of 1.86),  $n_2 = 6$  springs of type 2 (a 6 in. design with a theoretical spring constant of 2.63), and  $n_3 = 6$  springs of type 3 (a 4 in. design with a theoretical spring constant of 2.12), using an 8.8lb load. The students' experimental values are given in Table 6.1.1.2

These samples are just barely large enough to produce meaningful boxplots. Figure 6.6.1.2 gives a side-by-side boxplot representation of these data. The primary qualitative message carried by Figure 6.6.1.2 is that there is a substantial difference in empirical spring constants between the 6 in. spring type and the two 4 in. spring types but that no such difference between the two 4 in. spring types is obvious. Of course, the information in Table 6.1.1.2 could also be presented in side-by-side dot diagram form, as in Figure 6.1.1.3.

Empirical Spring Constants

Type 1 Springs	Type 2 Springs	Type 3 Springs
1.99, 2.06, 1.99	2.85, 2.74, 2.74	2.10, 2.01, 1.93
1.94, 2.05, 1.88	2.63, 2.74, 2.80	2.02, 2.10, 2.05
2.30		

Table 6.1.1.2 Empirical Spring Constants

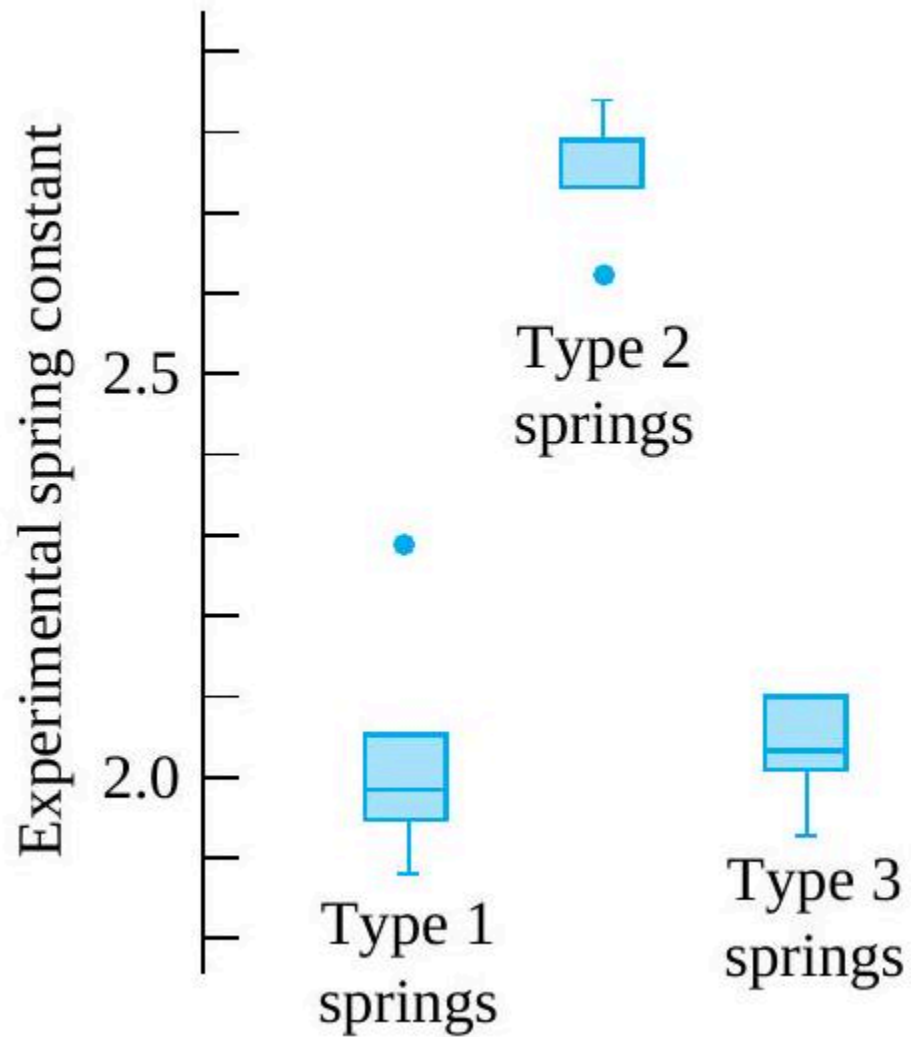


Figure 6.1.1.2 Side-by-side boxplots of empirical spring constants for springs of three types

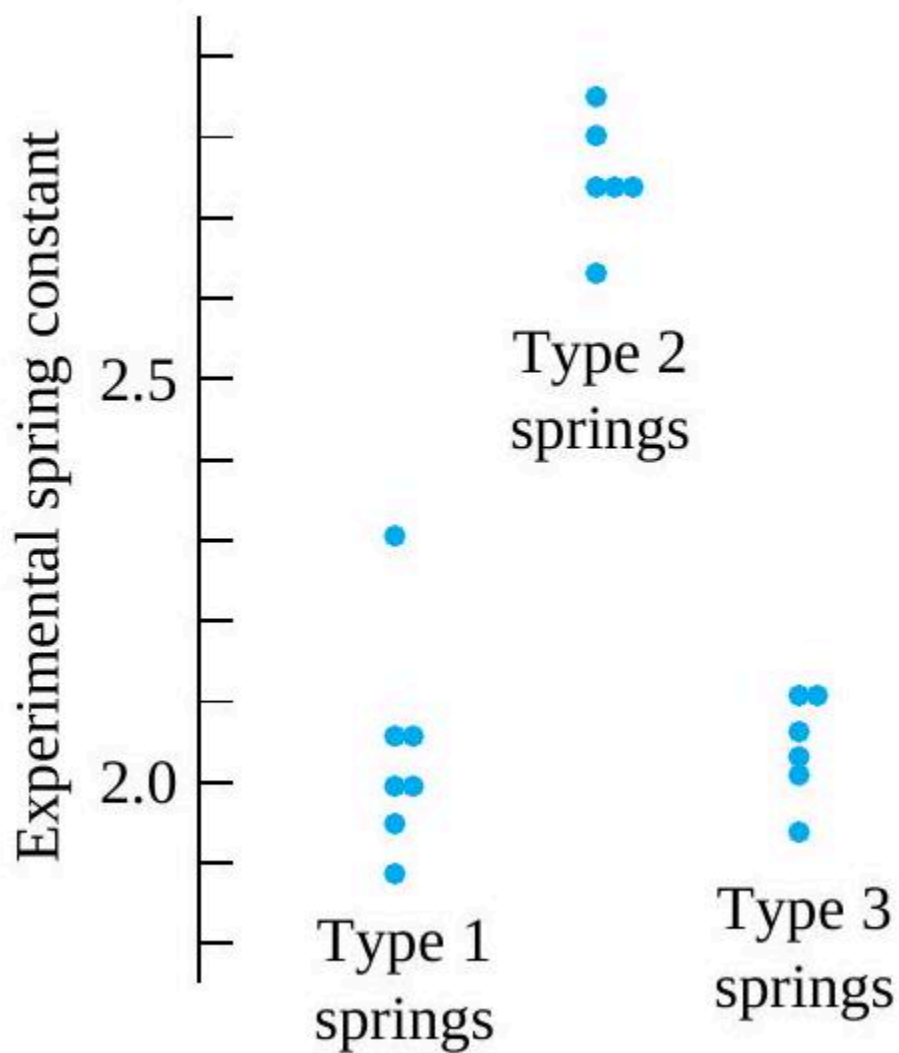


Figure 6.2.1.3 Side-by-side dot diagrams for three samples of empirical spring constants

Methods of formal statistical inference are meant to sharpen and quantify the impressions that one gets when making a descriptive analysis of data. But an intelligent graphical look at data and a correct application of formal inference methods rarely tell completely different stories. Indeed, the methods of formal inference offered here for simple, unstructured multisample studies are confirmatory-in cases like Examples 1 and 2, they should confirm what is clear from a descriptive or exploratory look at the data.

## 6.1.2 The One-Way (Normal) Multisample Model, Fitted Values, and Residuals

### ONE-WAY NORMAL MODEL ASSUMPTIONS

Part 5 emphasized repeatedly that to make one- and two-sample inferences, one must adopt a model for data generation that is both manageable and plausible. The present situation is no different, and standard inference methods for unstructured multisample studies are based on a natural extension of the model used in Section 5.3 to support small-sample comparison of two means. The present discussion will be carried out under the assumption that  $r$  samples of respective sizes  $n_1, n_2, \dots, n_r$  are independent samples from normal underlying distributions with a common variance—say,  $\sigma^2$ . Just as in Section 5.3 the  $r = 2$  version of this one-way (as opposed, for example, to several-way factorial) model led to useful inference methods for  $\mu_1 - \mu_2$ , this general version will support a variety of useful inference methods for  $r$ -sample studies. Figure 6.1.2.1 shows a number of different normal distributions with a common standard deviation. It represents essentially what must be generating measured responses if the methods of this chapter are to be applied.

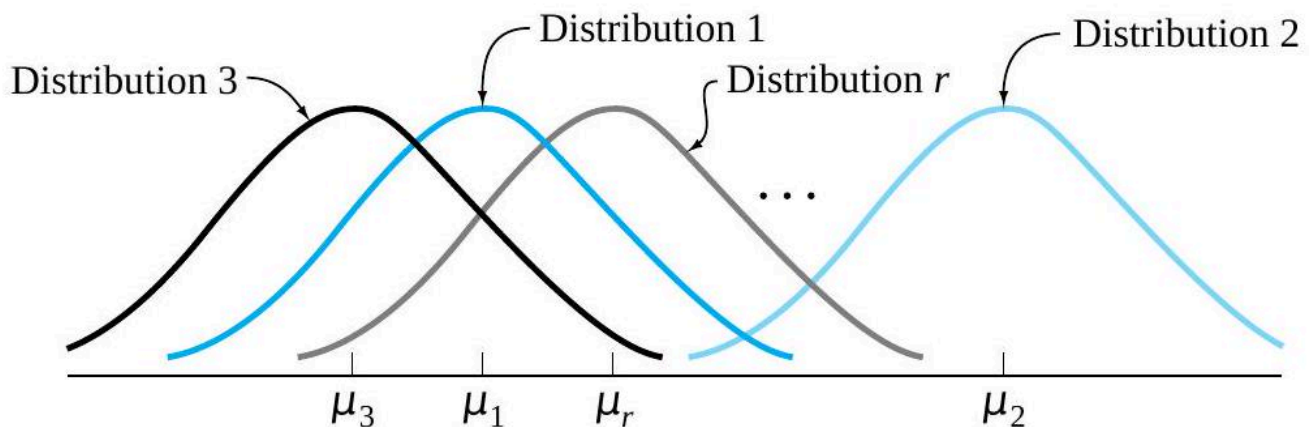


Figure 6.1.2.1 normal distributions with a common standard deviation

In addition to a description of the one-way model in words and the pictorial representation given in Figure 6.1.2.1, it is helpful to have a description of the model in symbols. This and the next three sections will employ the notation

$$y_{ij} = \text{the } j \text{ th observation in sample } i$$

The model equation used to specify the one-way model is then

### 6.1.2.1 One-way model statement in symbols

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where  $\mu_i$  is the  $i$ th underlying mean and the quantities  $\epsilon_{11}, \epsilon_{12}, \dots, \epsilon_{1n_1}, \epsilon_{21}, \epsilon_{22}, \dots, \epsilon_{2n_2}, \dots, \epsilon_{r1}, \epsilon_{r2}, \dots, \epsilon_{rn_r}$  are independent normal random variables with mean 0 and variance  $\sigma^2$ . (In this statement, the means  $\mu_1, \mu_2, \dots, \mu_r$  and the variance  $\sigma^2$  are typically unknown parameters.)

Equation (6.1.2.1) says exactly what is conveyed by Figure 6.1.2.1 and the statement of the one-way assumptions in words. This equation (6.1.2.1) says that an observation in sample  $i$  is made up of the corresponding underlying mean plus some random noise, namely

$$\epsilon_{ij} = y_{ij} - \mu_i$$

This is a theoretical counterpart of an empirical notion that we will see later in fitting a line using least squares. There, it will be useful to decompose data into fitted values and the corresponding residuals.

In the present situation, since any structure relating the  $r$  different samples is specifically being ignored, it may not be obvious how to apply the notions of fitted values and residuals. But a plausible meaning for

$$\hat{y}_{ij} = \text{the fitted value corresponding to } y_{ij}$$

in the present context is the  $i$ th sample mean

**$i$ th sample mean**

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

That is,

### 6.1.2.2 Fitted values for the one-way model

$$\hat{y}_{ij} = \bar{y}_i$$

Taking equation (6.1.2.2) to specify fitted values for an  $r$ -sample study, the pattern established then says that residuals are differences between observed values and sample means. That is, with

$$e_{ij} = \text{the residual corresponding to } y_{ij}$$

one has

### 6.1.2.3 Residuals for the one-way model

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$$

Rearranging display (6.1.2.3) gives the relationship

$$6.1.2.4 \quad y_{ij} = \hat{y}_{ij} + e_{ij} = \bar{y}_i + e_{ij}$$

which is an empirical counterpart of the theoretical statement (6.1.2.1). In fact, combining equations (6.1.2.1) and (6.1.2.4) into a single statement gives

$$6.1.2.5 \quad y_{ij} = \mu_i + \epsilon_{ij} = \bar{y}_i + e_{ij}$$

This is a specific instance of a pattern of thinking that runs through all of the common normal-distribution-based methods of analysis for multisample studies. In words, equation (6.1.2.5) says

$$6.1.2.6 \quad \text{Observation} = \text{deterministic response} + \text{noise} = \text{fitted value} + \text{residual}$$

and display (6.1.2.6) is a paradigm that provides a unified way of approaching the majority of the analysis methods presented in the rest of this book.

The decompositions (6.1.2.5) and (6.1.2.6) suggest that

1. the fitted values ( $\hat{y}_{ij} = \bar{y}_i$ ) are meant to approximate the deterministic part of a system response ( $\mu_i$ ), and
2. the residuals ( $e_{ij}$ ) are therefore meant to approximate the corresponding noise in the response ( $\epsilon_{ij}$ ).

The fact that the  $\epsilon_{ij}$  in equation (6.1.2.1) are assumed to be iid normal  $(0, \sigma^2)$  random variables then suggests that the  $e_{ij}$  ought to look at least approximately like a random sample from a normal distribution.

So the normal-plotting of an entire set of residuals is a way of checking on the reasonableness of the one-way model. The plotting of residuals against (1) fitted values, (2) time order of observation, or (3) any other potentially relevant variable-hoping to see only random scatter-are other ways of investigating the appropriateness of the model assumptions.

These kinds of plotting, which combine residuals from all  $r$  samples, are often especially useful in practice. When  $r$  is large at all, budget constraints on total data collection costs often force the individual sample sizes  $n_1, n_2, \dots, n_r$  to be fairly small. This makes it fruitless to investigate "single variance, normal distributions" model assumptions using (for example) sample-by-sample normal plots. (Of course, where all of  $n_1, n_2, \dots, n_r$  are of a decent size, a sample-by-sample approach can be effective.)

**Example 6.1.2.1 continued**

Returning again to the concrete strength study, consider investigating the reasonableness of model (6.1.2.1) for this case. Figure 6.1.1.1 is a first step in this investigation. As remarked earlier, it conveys the visual impression that at least the “equal variances” part of the one-way model assumptions is plausible. Next, it makes sense to compute some summary statistics and examine them, particularly the sample standard deviations. Table 6.1.2.1 gives sample sizes, sample means, and sample standard deviations for the data in Table 6.1.1.1.

At first glance, it might seem worrisome that in this table  $s_1$  is more than three times the size of  $s_8$ . But the sample sizes here are so small that a largest ratio of sample standard deviations on the order of 3.2 is hardly unusual (for  $r = 8$  samples of size 3 from a normal distribution). Note from the  $F$  tables (Tables A3) that for samples of size 3, even if only 2 (rather than 8) sample standard deviations were involved, a ratio of sample variances of  $(965.6/302.5)^2 \approx 10.2$  would yield a  $p$ -value between .10 and .20 for testing the null hypothesis of equal variances with a two-sided alternative. The sample standard deviations in Table 6.1.2.1 really carry no strong indication that the one-way model is inappropriate.

Since the individual sample sizes are so small, trying to see anything useful in eight separate normal plots of the samples is hopeless. But some insight can be gained by calculating and plotting all  $8 \times 3 = 24$  residuals. Some of the calculations necessary to compute residuals for the data in Table 6.1.1.1 (using the fitted values appearing as sample means in Table 6.1.2.1) are shown in Table 6.1.2.2. Figures 6.1.2.2 and 6.1.2.3 are, respectively, a plot of residuals versus fitted  $y$  ( $e_{ij}$  versus  $\bar{y}_{ij}$ ) and a normal plot of all 24 residuals.

Figure 6.1.2.2 gives no indication of any kind of strong dependence of  $\sigma$  on  $\mu$  (which would violate the “constant variance” restriction). And the plot in Figure 6.1.2.3 is reasonably linear, thus identifying no obvious difficulty with the assumption of normal distributions. In all, it seems from examination of both the raw data and the residuals that analysis of the data in Table 6.1.1.1 on the basis of model (6.1.2.1) is perfectly sensible.

Summary Statistics for the Concrete Strength Study

$i$ , Concrete Formula	$n_i$ , Sample Size	$\bar{y}_i$ , Sample Mean (psi)	$s_i$ , Sample Standard Deviation (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	$n_8 = 3$	$\bar{y}_8 = 2,390.7$	$s_8 = 302.5$

Table 6.1.2.1



Example Computations of Residuals for the Concrete Strength Study

Specimen	$i$ , Concrete Formula	$y_{ij}$ , Compressive Strength (psi)	$\hat{y}_{ij} = \bar{y}_i$ , Fitted Value	$e_{ij}$ , Residual
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

Table 6.1.2.2

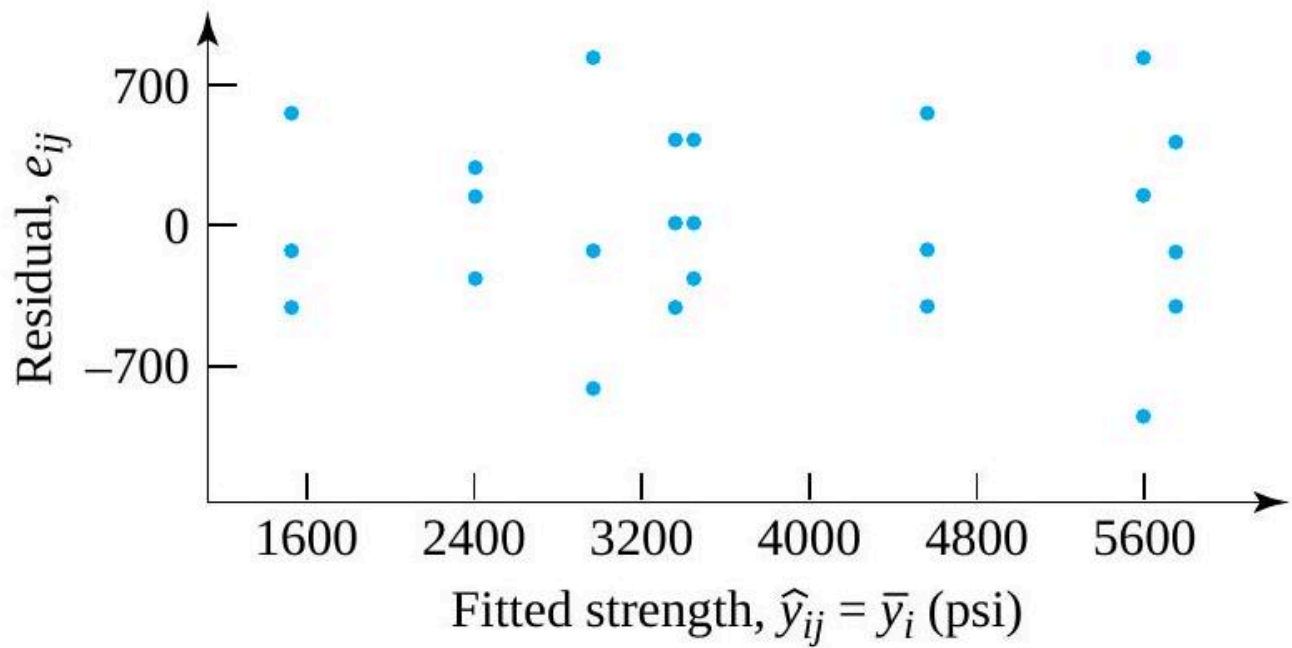


Figure 6.1.2.2 Plot of residuals versus fitted responses for the compressive strengths

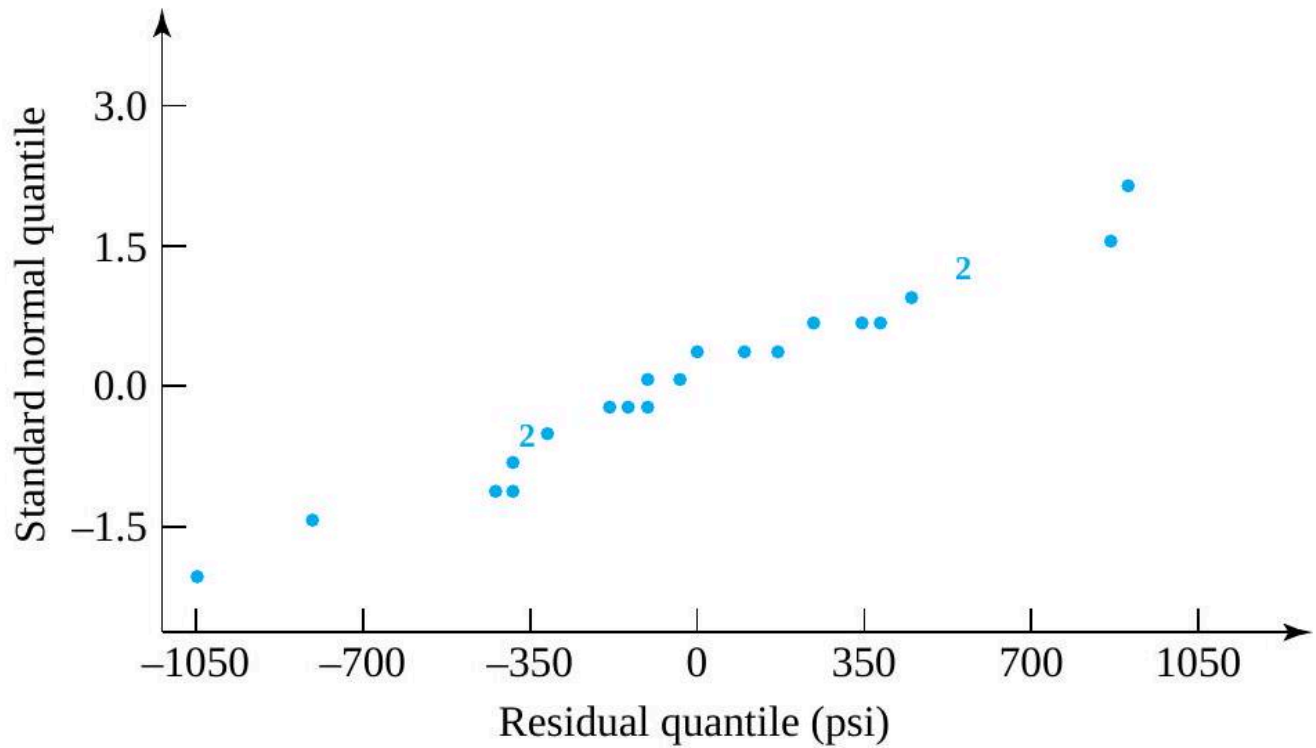


Figure 6.1.2.3 Normal plot of the compressive strength residuals

#### Example 6.1.2.2 Spring Testing continued

The spring testing data can also be examined with the potential use of the one-way normal model (6.1.1.1) in mind. Figures 6.1.1.2 and 6.1.1.3 indicate reasonably comparable variabilities of experimental spring constants for the  $r = 3$  different spring types. The single very large value (for spring type 1) causes some doubt both in terms of this judgment and also (by virtue of its position on its boxplot as an outlying value) regarding a “normal distribution” description of type 1 experimental constants. Summary statistics for these samples are given in Table 6.1.2.3.

Without the single extreme value of 2.30, the first sample standard deviation would be .068, completely in line with those of the second and third samples. But even the observed ratio of largest to smallest sample variance (namely  $(.134/.064)^2 = 4.38$ ) is not a compelling reason to abandon a one-way model description of the spring constants. (A look at the  $F$  tables with  $v_1 = 6$  and  $v_2 = 5$  shows that 4.38 is between the  $F_{6,5}$  distribution .9 and .95 quantiles. So even if there were only two rather than three samples involved, a variance ratio of 4.38 would yield a  $p$ -value between .1 and .2 for (two-sided) testing of equality of variances.) Before letting the single type 1 empirical spring constant of 2.30 force abandonment of the highly tractable model (6.1.2.1) some additional investigation is warranted.

Sample sizes  $n_1 = 7$  and  $n_2 = n_3 = 6$  are large enough that it makes sense to look at sample-by-sample normal plots of the spring constant data. Such plots, drawn on the same set of axes, are shown in Figure 6.1.2.4. Further, use of the fitted values  $(\bar{y}_i)$  listed in Table 6.1.2.3 with the original data given in Table 6.1.1.2 produces 19 residuals, as partially illustrated in Table 6.1.2.4. Then Figures

6.1.2.5 and 6.1.2.6, respectively, show a plot of residuals versus fitted responses and a normal plot of all 19 residuals.

But Figures 6.1.2.5 and 6.1.2.6 again draw attention to the largest type 1 empirical spring constant. Compared to the other measured values, 2.30 is simply too large (and thus produces a residual that is too large compared to all the rest) to permit serious use of model (6.1.2.1) with the spring constant data. Barring the possibility that checking of original data sheets would show the 2.30 value to be an arithmetic blunder or gross error of measurement (which could be corrected or legitimately force elimination of the 2.30 value from consideration), it appears that the use of model (6.1.2.1) with the  $r = 3$  spring types could produce inferences with true (and unknown) properties quite different from their nominal properties.

One might, of course, limit attention to spring types 2 and 3. There is nothing in the second or third samples to render the “equal variances, normal distributions” model untenable for those two spring types. But the pattern of variation for springs of type 1 appears to be detectably different from that for springs of types 2 and 3, and the one-way model is not appropriate when all three types are considered.

Summary Statistics for the Empirical Spring Constants

$i$ , Spring Type	$n_i$	$\bar{y}_i$	$s_i$
1	7	2.030	.134
2	6	2.750	.074
3	6	2.035	.064

*6.1.2.3 Table Summary Statistics for the Empirical Spring Constants*

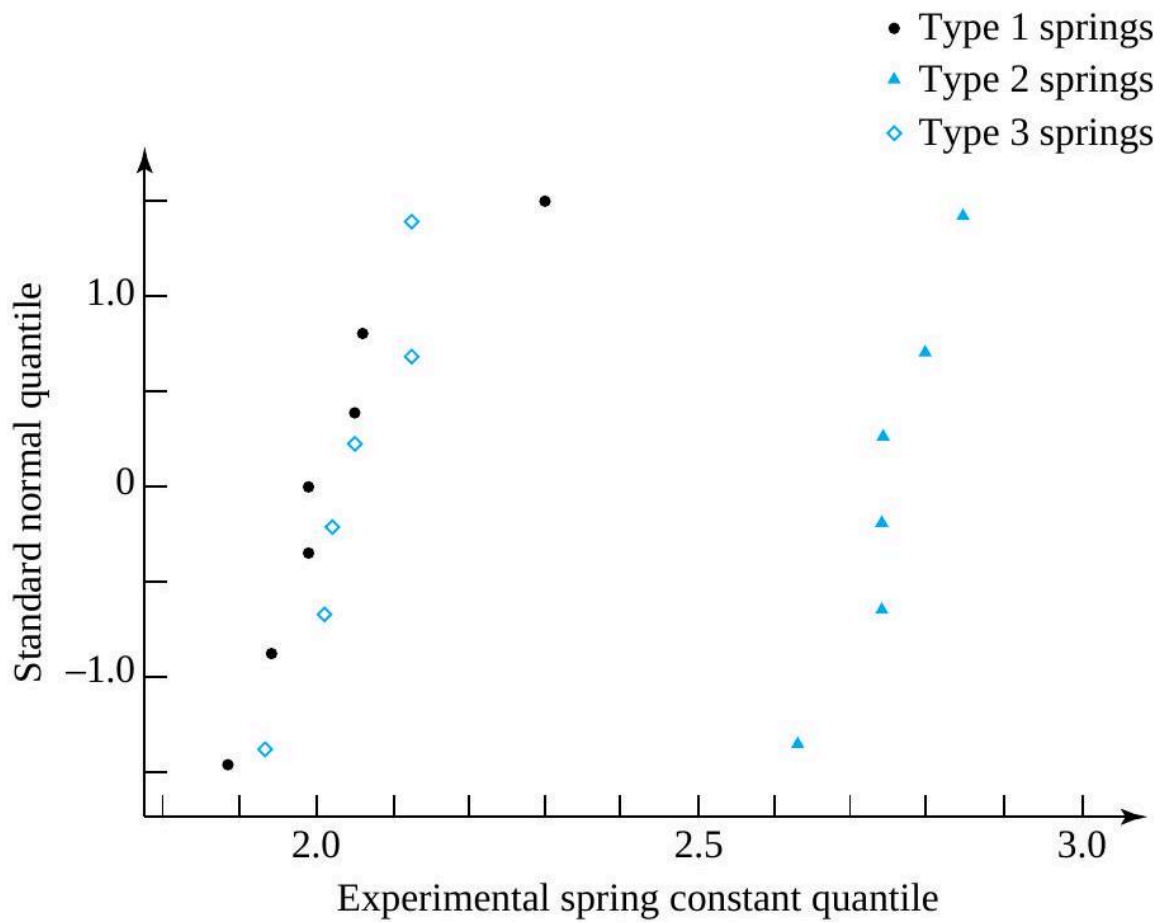


Figure 6.1.2.4 Normal plots of empirical spring constants for springs of three types

Example Computations of Residuals for the Spring Constant Study

$i,$ Spring Type	$j,$ Observation Number	$y_{ij},$ Spring Constant	$\hat{y}_{ij} = \bar{y}_i,$ Sample Mean	$e_{ij},$ Residual
1	1	1.99	2.030	-.040
⋮	⋮	⋮	⋮	⋮
1	7	2.30	2.030	.270
2	1	2.85	2.750	.100
⋮	⋮	⋮	⋮	⋮
2	6	2.80	2.750	.050
3	1	2.10	2.035	.065
⋮	⋮	⋮	⋮	⋮
3	6	2.05	2.035	.015

Table 6.1.2.4 Example Computations of Residuals for the Spring Constant Study

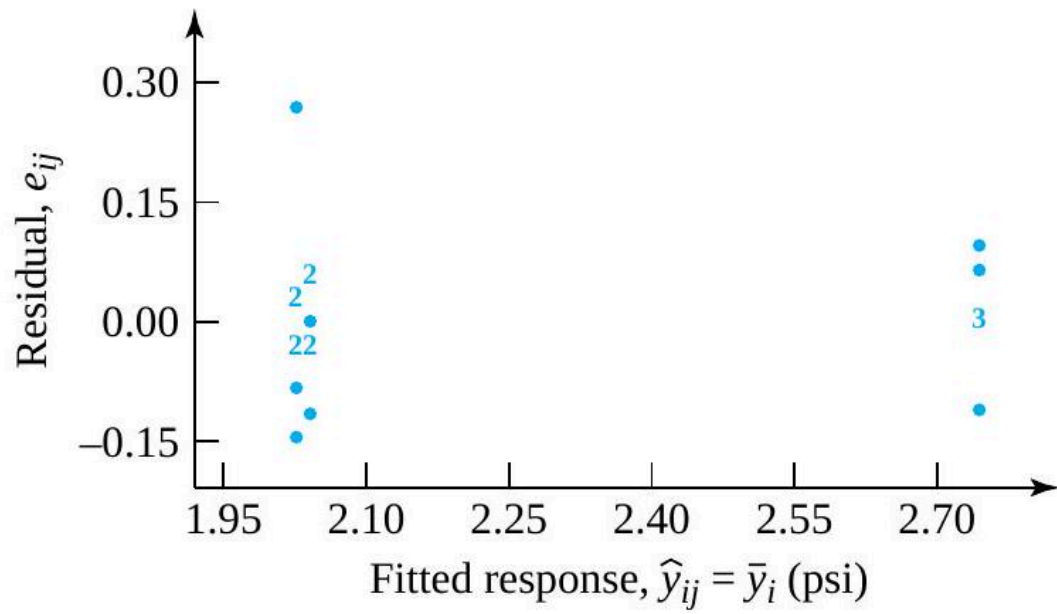


Figure 6.1.2.5 Plot of residuals versus fitted responses for the empirical spring constants

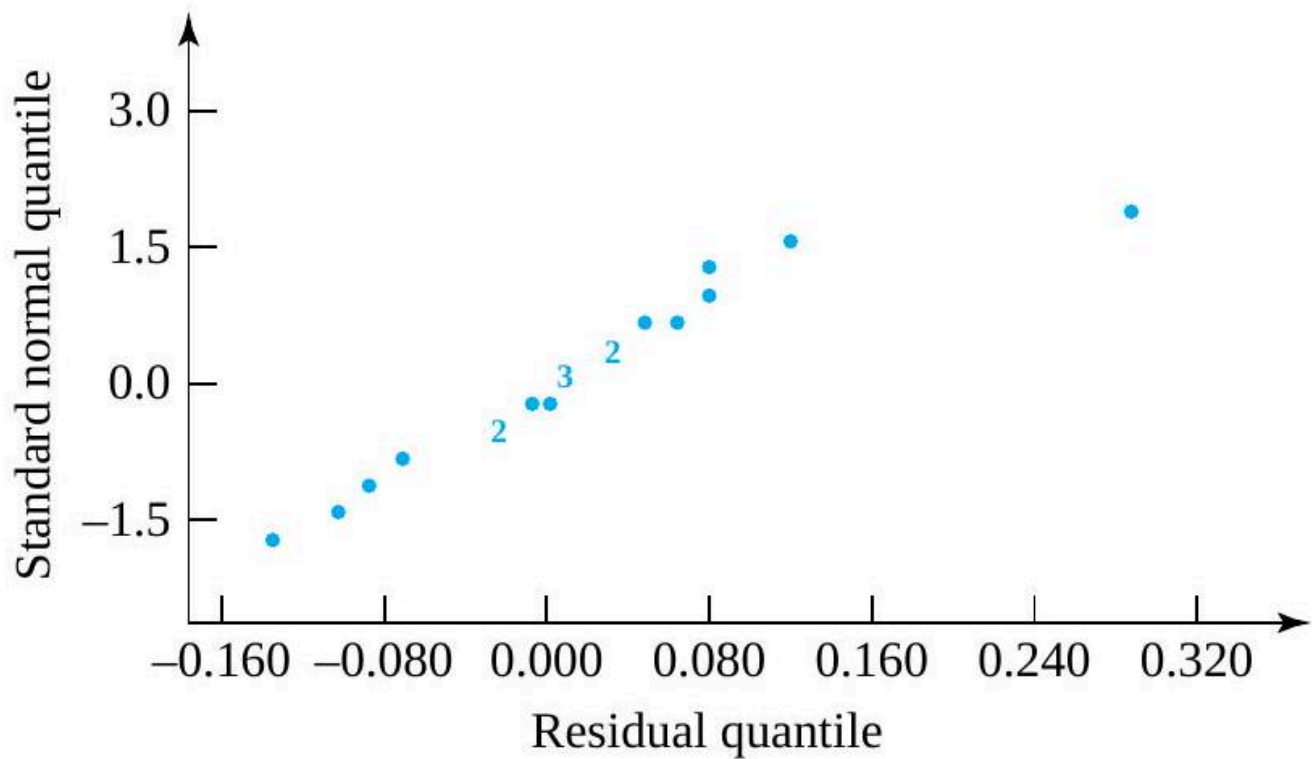


Figure 6.1.2.6 Normal plot of the spring constant residuals

## 6.1.3 A Pooled Estimate of Variance for Multisample Studies

The “equal variances, normal distributions” model (6.1.2.1) has as a fundamental parameter,  $\sigma$ , the standard deviation associated with responses from any of conditions **1**, **2**, **3**, . . . ,  $r$ . Similar to what was done in the  $r = 2$  situation of Part 5, it is typical in multisample studies to pool the  $r$  sample variances to arrive at a single estimate of  $\sigma$  derived from all  $r$  samples.

### DEFINITION Pooled Standard Deviation

#### EXPRESSION 6.1.3.1

If  $r$  numerical samples of respective sizes  $n_1, n_2, \dots, n_r$  produce sample variances  $s_1^2, s_2^2, \dots, s_r^2$ , the **pooled sample variance**,  $s_P^2$ , is the weighted average of the sample variances, where the weights are the sample sizes minus 1. That is,

$$s_P^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2 + \dots + (n_r - 1) s_r^2}{(n_1 - 1) + (n_2 - 1) + \dots + (n_r - 1)}$$

The pooled sample standard deviation,  $s_P$ , is the square root of  $s_P^2$ .

Definition 6.1.3.1 is just redefining that in Part 5 restated for the case of more than two samples. As was the case for  $s_p$  based on two samples,  $s_P$  is guaranteed to lie between the largest and smallest of the  $s_i$  and is a mathematically convenient form of compromise value.

Equation (6.1.3.1) can be rewritten in a number of equivalent forms. For one thing, letting

■ The total number of observations in an  $r$ -sample study

$n = \sum_{i=1}^r n_i =$  the total number of observations in all  $r$  samples

it is common to rewrite the denominator on the right of equation (6.1.3.1) as

$$\sum_{i=1}^r (n_i - 1) = \sum_{i=1}^r n_i - \sum_{i=1}^r 1 = n - r$$

And noting that the  $i$  th sample variance is

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

the numerator on the right of equation (6.1.3.1) is

6.1.3.2 and 6.1.3.3

$$\begin{aligned} \sum_{i=1}^r (n_i - 1) \left( \frac{1}{(n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \right) &= \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \\ &= \sum_{i=1}^r \sum_{j=1}^{n_i} e_{ij}^2 \end{aligned}$$

Alternative formulas for  $s_P^2$

So one can define  $s_P^2$  in terms of the right-hand side of equation (6.1.3.2) or (6.1.3.3) divided by  $n - r$ .

#### Example 6.1.3.1 Compression Strength continued.

For the compressive strength data, each of  $n_1, n_2, \dots, n_8$  are 3, and  $s_1$  through  $s_8$  are given in Table 6.1.2.1. So using equation (6.1.3.1),

$$\begin{aligned} s_{\mathrm{P}}^2 &= \frac{(3-1)(965.6)^2 + (3-1)(432.3)^2 + \cdots + (3-1)(302.5)^2}{(3-1) + (3-1) + \cdots + (3-1)} \\ &= \frac{2 \left[ (965.6)^2 + (432.3)^2 + \cdots + (302.5)^2 \right]}{16} \quad \&= \frac{2,705,705}{8} \quad \&= 338,213 (\mathrm{psi})^2 \end{aligned}$$

and thus

$$s_P = \sqrt{338,213} = 581.6 \text{ psi}$$

One estimates that if a large number of specimens of any one of formulas 1 through 8 were tested, a standard deviation of compressive strengths on the order of 582 psi would be obtained.

### The meaning of $s_P$

$s_P$  is an estimate of the intrinsic or baseline variation present in a response variable at a fixed set of conditions, calculated supposing that the baseline variation is constant across the conditions under which the samples were collected. When that supposition is reasonable, the pooling idea allows a number of individually unreliable small-sample estimates to be combined into a single, relatively more reliable combined estimate. It is a fundamental measure that figures prominently in a variety of useful methods of formal inference.

### Confidence limits for the one-way model variance

On occasion, it is helpful to have not only a single number as a data-based best guess at  $\sigma^2$  but a confidence interval as well. Under model restrictions (6.2.1.1), the variable

$$\frac{(n - r) s_P^2}{\sigma^2}$$

has a  $\chi_{n-r}^2$  distribution. Thus, in a manner exactly parallel to the derivation in Part 5, a two-sided confidence interval for

$\sigma^2$  has endpoints

$\frac{(n-r) s_P^2}{U}$  and  $\frac{(n-r) s_P^2}{L}$

where  $L$  and  $U$  are such that the  $\chi_{n-r}^2$  probability assigned to the interval  $(L, U)$  is the desired confidence level. And, of course, a one-sided interval is available by using only one of the endpoints (6.1.3.4) and choosing  $U$  or  $L$  such that the  $\chi_{n-r}^2$  probability assigned to the interval  $(0, U)$  or  $(L, \infty)$  is the desired confidence.

#### Example 6.1.3.2 continued

In the concrete compressive strength case, consider the use of display (6.1.3.4) in making a two-sided 90% confidence interval for  $\sigma$ . Since  $n - r = 16$  degrees of freedom are associated with  $s_P^2$ , one consults Table A1.4 for the .05 and .95 quantiles of the  $\chi_{16}^2$  distribution. These are 7.962 and 26.296, respectively. Thus a confidence interval for  $\sigma^2$  has endpoints

$$\frac{16(581.6)^2}{26.296} \text{ and } \frac{16(581.6)^2}{7.962}$$

So a two-sided 90% confidence interval for  $\sigma$  has endpoints



$$\sqrt{\frac{16(581.6)^2}{26.296}} \text{ and } \sqrt{\frac{16(581.6)^2}{7.962}}$$

.  
that is,  
.

453.7psi and 824.5psi

## 6.2.0 Introduction Confidence Intervals Multisample Studies

Part 5 illustrates how useful confidence intervals for means and differences in means can be in one- and two-sample studies. Estimating an individual mean and comparing a pair of means are every bit as important when there are  $r$  samples as they are when there are only one or two. The methods of Part 5 can be applied in  $r$ -sample studies by simply limiting attention to one or two of the samples at a time.

But since individual sample sizes in multisample studies are often small, such a strategy of inference often turns out to be relatively uninformative. Under the one-way model assumptions discussed in the previous section, it is possible to base inference methods on the pooled standard deviation,  $s_p$ . Those tend to be relatively more informative than the direct application of the formulas from Part 5 in the present context. This section first considers the confidence interval estimation of a single mean and of the difference between two means under the “equal variances, normal distributions” model. Finally, the section closes with some comments concerning the notions of individual and simultaneous confidence levels.

## 6.2.1 Intervals for Means and for Comparing Means

The primary drawback to applying the formulas from Part 5 in a multisample context is that typical small sample sizes lead to small degrees of freedom, large  $t$  multipliers in the plus-or-minus parts of the interval formulas, and thus long intervals. But based on the one-way model assumptions, confidence interval formulas can be developed that tend to produce shorter intervals.

That is, in a development parallel to that in Part 5, under the one-way normal model,

$$T = \frac{\bar{y}_i - \mu_i}{\frac{s_P}{\sqrt{n_i}}}$$

has a  $t_{n-r}$  distribution. Hence, a two-sided confidence interval for the  $i$ th mean,  $\mu_i$ , has endpoints

### 6.2.1.1 Confidence limits for $\mu_i$ based on the one-way model

$$\bar{y}_i \pm t \frac{s_P}{\sqrt{n_i}}$$

where the associated confidence is the probability assigned to the interval from  $-t$  to  $t$  by the  $t_{n-r}$  distribution. This is exactly formula the formulas from Part 5, except that  $s_P$  has replaced  $s_i$  and the degrees of freedom have been adjusted from  $n_i - 1$  to  $n - r$ .

In the same way, for conditions  $i$  and  $i'$ , the variable

$$T = \frac{\bar{y}_i - \bar{y}_{i'} - (\mu_i - \mu_{i'})}{s_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}}$$

has a  $t_{n-r}$  distribution. Hence, a two-sided confidence interval for  $\mu_i - \mu_{i'}$  has endpoints

### 6.2.1.2 Confidence limits for $\mu_i - \mu_{i'}$ based on the one-way model

$$\bar{y}_i - \bar{y}_{i'} \pm t s_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

where the associated confidence is the probability assigned to the interval from  $-t$  to  $t$  by the  $t_{n-r}$  distribution. Display (6.2.1.2) is essentially a formula from Part 5, except that  $s_P$  is calculated based on  $r$  samples instead of two and the degrees of freedom are  $n - r$  instead of  $n_i + n_{i'} - 2$ .

Of course, use of only one endpoint from formula (6.2.1.1) or (6.2.1.2) produces a one-sided confidence interval with associated confidence corresponding to the  $t_{n-r}$  probability assigned to the interval  $(-\infty, t)$  (for  $t > 0$ ). The virtues of formulas (6.2.1.1) and (6.2.1.2) (in comparison to the corresponding formulas from Part 5) are that (when appropriate) for a given confidence, they will tend to produce shorter intervals than their Part 5 counterparts.

#### Example 6.2.1.1 Confidence Intervals for Individual, and Differences of Mean Concrete Compressive Strengths, continued

Return to the concrete strength study of Armstrong, Babb, and Campen. Consider making first a 90% two-sided confidence interval for the mean compressive strength of an individual concrete formula and then a 90% two-sided confidence interval for the difference in mean compressive strengths for two different formulas. Since  $n = 24$  and  $r = 8$ , there are  $n - r = 16$  degrees of freedom associated with  $s_P = 581.6$ . So the .95 quantile of the  $t_{16}$  distribution, namely 1.746, is appropriate for use in both formulas (6.2.1.1) and (6.2.1.2).

Turning first to the estimation of a single mean compressive strength, since each  $n_i$  is 3, the plus-or-minus part of formula (6.2.1.1) gives

$$t \frac{s_P}{\sqrt{n_i}} = 1.746 \frac{581.6}{\sqrt{3}} = 586.3 \text{ psi}$$

So  $\pm 586.3$  psi precision could be attached to any one of the sample means in Table 6.2.1.1 as an estimate of the corresponding formula's mean strength. For example, since  $\bar{y}_3 = 4,527.3$  psi, a 90% two-sided confidence interval for  $\mu_3$  has endpoints

$$4,527.3 \pm 586.3$$

that is,

$$3,941.0 \text{ psi and } 5,113.6 \text{ psi}$$

In parallel fashion, consider estimation of the difference in two mean compressive strengths with 90% confidence. Again, since each  $n_i$  is 3, the plus-or-minus part of formula (6.2.1.2) gives

$$t s_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} = 1.746(581.6) \sqrt{\frac{1}{3} + \frac{1}{3}} = 829.1 \text{ psi}$$

Thus,  $\pm 829.1$  psi precision could be attached to any difference between sample means in Table 6.2.1.1 as an estimate of the corresponding difference in formula mean strengths. For instance, since  $\bar{y}_3 = 4,527.3$  psi and  $\bar{y}_7 = 1,551.3$  psi, a 90%

two-sided confidence interval for  $\mu_3 - \mu_7$  has endpoints

$$(4,527.3 - 1,551.3) \pm 829.1$$

That is,

$$2,146.9\text{psi} \quad \text{and} \quad 3,805.1\text{psi}$$

Concrete Formula Sample Mean Strengths

Concrete Formula	Sample Mean Strength (psi)
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

Table 6.2.1.1 Concrete Formula Sample Mean Strengths

The use of  $n - r = 16$  degrees of freedom in Example 6.2.1.1 instead of  $n_i - 1 = 2$  and  $n_i + n_{i'} - 2 = 4$  reflects the reduction in uncertainty associated with  $s_P$  as an estimate of  $\sigma$  as compared to that of  $s_i$  and of  $s_P$  based on only two samples. That reduction is, of course, bought at the price of restriction to problems where the “equal variances” model is tenable.

## 6.2.2 Individual and Simultaneous Confidence Levels

This section has introduced a variety of confidence intervals for multisample studies. In a particular application, several of these might be used, perhaps several times each. For example, even in the relatively simple context of Example 6.2.1.1. (the paper towel absorbency study), it would be reasonable to desire confidence intervals for each of

$$\mu_1, \mu_2, \mu_3, \mu_1 - \mu_2, \mu_1 - \mu_3, \mu_2 - \mu_3, \text{ and } \mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$$

Since many confidence statements are often made in multisample studies, it is important to reflect on the meaning of a confidence level and realize that it is attached to one interval at a time. If many 90% confidence intervals are made, the 90% figure applies to the intervals individually. One is "90% sure" of the first interval, separately "90% sure" of the second, separately "90% sure" of the third, and so on. It is not at all clear how to arrive at a reliability figure for the intervals jointly or simultaneously (i.e., an a priori probability that all the intervals are effective). But it is fairly obvious that it must be less than 90%. That is, the simultaneous or joint confidence (the overall reliability figure) to be associated with a group of intervals is generally not easy to determine, but it is typically less (and sometimes much less) than the individual confidence level(s) associated with the intervals one at a time.

There are at least three different approaches to be taken once the difference between simultaneous and individual confidence levels is recognized. The most obvious option is to make individual confidence intervals and be careful to interpret them as such (being careful to recognize that as the number of intervals one makes increases, so does the likelihood that among them are one or more intervals that fail to cover the quantities they are meant to locate).

A second way of handling the issue of simultaneous versus individual confidence is to use very large individual confidence levels for the separate intervals and then employ a somewhat crude inequality to find at least a minimum value for the simultaneous confidence associated with an entire group of intervals. That is, if  $k$  confidence intervals have associated confidences  $\gamma_1, \gamma_2, \dots, \gamma_k$ , the Bonferroni inequality says that the simultaneous or joint confidence that all  $k$  intervals are effective (say,  $\gamma$ ) satisfies

### 6.2.2.1 The Bonferroni inequality

$$\gamma \geq 1 - ((1 - \gamma_1) + (1 - \gamma_2) + \dots + (1 - \gamma_k))$$

(Basically, this statement says that the joint "unconfidence" associated with  $k$  intervals ( $1 - \gamma$ ) is no larger than the sum of the  $k$  individual unconfidences. For example, five intervals with individual 99%

confidence levels have a joint or simultaneous confidence level of at least 95%.)

The third way of approaching the issue of simultaneous confidence is to develop and employ methods that for some specific, useful set of unknown quantities provide intervals with a known level of simultaneous confidence. There are whole books full of such simultaneous inference methods. In the next section, one of the better known and simplest of these are discussed.

## 6.2.3 Simultaneous Confidence Interval Methods

As Section 6.2.2 illustrated, there are several kinds of confidence intervals for means and linear combinations of means that could be made in a multisample study. The issue of individual versus simultaneous confidence was also raised, but only the use of the Bonferroni inequality was given as a means of controlling a simultaneous confidence level.

This section presents a method for making a number of confidence intervals and in the process maintaining a desired simultaneous confidence. This is Tukey's method for the simultaneous confidence interval estimation of all differences in pairs of underlying means.

### TUKEY'S METHOD

A second set of quantities often of interest in an  $r$ -sample study consists of the differences in all  $\frac{r(r-1)}{2}$  pairs of mean responses  $\mu_i$  and  $\mu_{i'}$ . Section 6.2 argued that a single difference in mean responses,  $\mu_i - \mu_{i'}$ , can be estimated using an interval with endpoints

$$6.2.3.1 \quad \bar{y}_i - \bar{y}_{i'} \pm t_{SP} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$$

where the associated confidence level is an individual one. But if, for example,  $r = 8$ , there are 28 different two-at-a-time comparisons of underlying means to be considered ( $\mu_1$  versus  $\mu_2$ ,  $\mu_1$  versus  $\mu_3, \dots, \mu_1$  versus  $\mu_8$ ,  $\mu_2$  versus  $\mu_3, \dots$ , and  $\mu_7$  versus  $\mu_8$ ). If one wishes to guarantee a reasonable simultaneous confidence level for all these comparisons via the crude Bonferroni idea, a huge individual confidence level is required for the intervals (6.2.3.1). For example, the Bonferroni inequality requires 99.82% individual confidence for 28 intervals in order to guarantee simultaneous 95% confidence.

A better approach to the setting of simultaneous confidence limits on all of the differences  $\mu_i - \mu_{i'}$  is to replace  $t$  in formula (6.2.3.1) with a multiplier derived specifically for the purpose of providing exact, stated, simultaneous confidence in the estimation of all such differences. J. Tukey first pointed out that it is possible to provide such multipliers using quantiles of the Studentized range distributions. Tables A5A and A5B give values of constants  $q^*$  such that the set of two-sided intervals with endpoints

#### 6.2.3.2 Tukey's twosided simultaneous confidence limits for all differences in $r$ means}



$$\bar{y}_i - \bar{y}_{i'} \pm \frac{q^*}{\sqrt{2}} S_P \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}$$

has simultaneous confidence at least 95% or 99% (depending on whether  $Q(.95)$  is read from Table A5A or  $Q(.99)$  is read from Table A5B) in the estimation of all differences  $\mu_i - \mu_{i'}$ . If all the sample sizes  $n_1, n_2, \dots, n_r$  are equal, the 95% or 99% nominal simultaneous confidence figure is exact, while if the sample sizes are not all equal, the true value is at least as big as the nominal value.

In order to apply Tukey's method, one must find (using interpolation as needed) the column in Tables A5, corresponding to the number of samples/means to be compared and the row corresponding to the degrees of freedom associated with  $s_P$ , (namely,  $v = n - r$ ).

#### Example 6.2.3.1 Compressive Strengths continued

Figure 6.2.3.1 shows a plot of eight sample mean compressive strengths, enhanced with error bars derived from simultaneous confidence limit.

Consider the making of confidence intervals for differences in formula mean compressive strengths. If a 95% two-sided individual confidence interval is desired for a specific difference  $\mu_i - \mu_{i'}$ , formula (6.2.3.1) shows that appropriate endpoints are

$$\bar{y}_i - \bar{y}_{i'} \pm 2.120(581.6) \sqrt{\frac{1}{3} + \frac{1}{3}}$$

that is,

$$\bar{y}_i - \bar{y}_{i'} \pm 1,006.7 \text{psi}$$

On the other hand, if one plans to estimate all differences in mean compressive strengths with simultaneous 95% confidence, by formula (6.2.3.2) Tukey two-sided intervals with endpoints

$$\bar{y}_i - \bar{y}_{i'} \pm \frac{4.90}{\sqrt{2}} (581.6) \sqrt{\frac{1}{3} + \frac{1}{3}}$$

that is,

$$\bar{y}_i - \bar{y}_{i'} \pm 1,645.4 \text{psi}$$

are in order (4.90 is the value in the  $r = 8$  column and  $v = 16$  row of Table A5A.)

In keeping with the fact that the confidence level associated with the second intervals is a simultaneous one, the Tukey intervals are wider than those indicated in the first formula.

The plus-or-minus part of the final display is not as big as twice the plus-or-minus part of expression previously. Thus, when looking at Figure 6.3.2.1, it is not necessary that the error bars around two means fail to overlap before it is safe to judge the corresponding underlying means to be detectably different. Rather, it is only necessary that the two sample means differ by the plus-or-minus part of formula (6.2.3.2)-1,645.4 psi in the present situation.

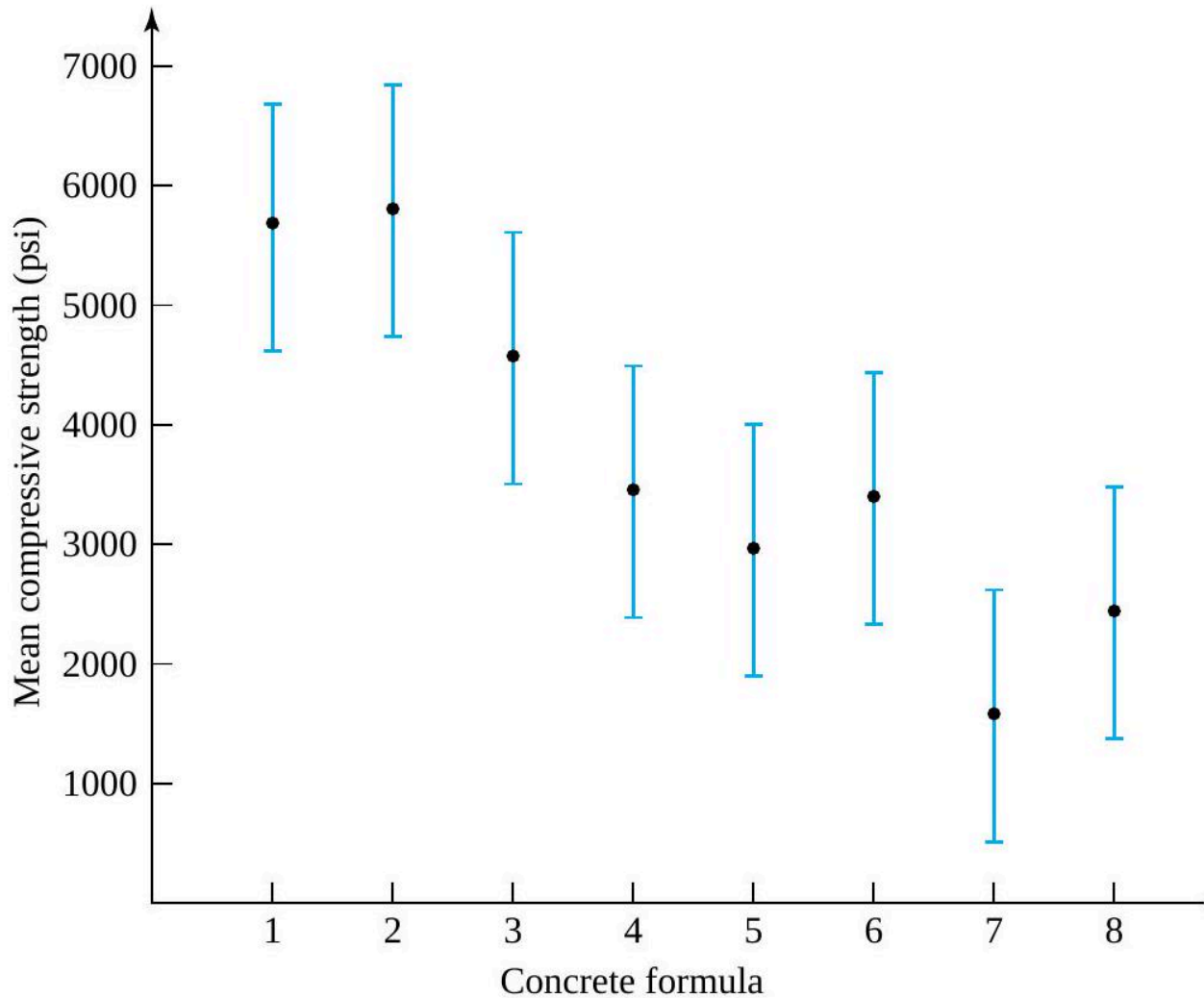


Figure 6.2.3.1 Plot of eight sample mean compressive strengths, enhanced with error bars derived from simultaneous confidence limits

## 6.3.0 Introduction ANOVA

This course's approach to inference in multisample studies has to this point been completely "interval-oriented." But there are also significance-testing methods that are appropriate to the multiple-sample context. This section considers some of these and the issues raised by their introduction. It begins with some general comments regarding significance testing in  $r$ -sample studies. Then the one-way analysis of variance (ANOVA) test for the equality of  $r$  means is discussed. Next, the oneway ANOVA table and the organization and intuition that it provides are presented.

### 6.3.1 Significance Testing and Multisample Studies

Just as there are many quantities one might want to estimate in a multisample study, there are potentially many issues of statistical significance to be judged. For instance, one might desire  $p$ -values for hypotheses like

6.3.1.1  $H_0 : \mu_3 = 7$

6.3.1.2  $H_0 : \mu_3 - \mu_7 = 0$

6.3.1.3  $H_0 : \mu_1 - \frac{1}{2}(\mu_2 + \mu_3) = 0$

The confidence interval methods discussed in Section 6.2 have their significance testing analogs for treating hypotheses that, like all three of these, involve linear combinations of the means  $\mu_1, \mu_2, \dots, \mu_r$ .

In general (under the standard one-way model), if

$$L = c_1\mu_1 + c_2\mu_2 + \dots + c_r\mu_r$$

the hypothesis

6.3.1.4  $H_0 : L = \#$

can be tested using the test statistic

6.3.1.5

$$T = \frac{\hat{L} - \#}{s_P \sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \dots + \frac{c_r^2}{n_r}}}$$

and a  $t_{n-r}$  reference distribution. This fact specializes to cover hypotheses of types (6.3.1.1) to (6.3.1.3) by appropriate choice of the  $c_i$  and  $\#$ .

But the significance-testing method most often associated with the one-way normal model is not for hypotheses of the type (6.3.1.4). Instead, the most common method concerns the hypothesis that all  $r$

underlying means have the same value. In symbols, this is

$$\mathbf{6.3.1.6} \quad H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

Given that one is working under the assumptions of the one-way model to begin with, hypothesis (6.3.1.6) amounts to a statement that all  $r$  underlying distributions are essentially the same – or “There are no differences between treatments.”

Hypothesis (6.3.1.6) can be thought of in terms of the simultaneous equality of  $\frac{r(r-1)}{2}$  pairs of means

– that is, as equivalent to the statement that simultaneously

$$\begin{aligned} \mu_1 - \mu_2 = 0, \quad \mu_1 - \mu_3 = 0, \quad \dots, \quad \mu_1 - \mu_r = 0 \\ \mu_2 - \mu_3 = 0, \quad \dots, \quad \text{and} \quad \mu_{r-1} - \mu_r = 0 \end{aligned}$$

And this fact should remind the reader of the ideas about simultaneous confidence intervals from the previous section (specifically, Tukey’s method). In fact, one way of judging the statistical significance of an  $r$ -sample data set in reference to hypothesis (6.3.1.6) is to apply Tukey’s method of simultaneous interval estimation and note whether or not all the intervals for differences in means include 0. If they all do, the associated  $p$ -value is larger than 1 minus the simultaneous confidence level. If not all of the intervals include 0, the associated  $p$ -value is smaller than 1 minus the simultaneous confidence level. (If simultaneous 95% intervals all include 0, no differences between means are definitively established, and the corresponding  $p$ -value exceeds .05.)

The authors admit a bias toward estimation over testing per se. A consequence of this bias is a fondness for deriving a rough idea of a  $p$ -value for hypothesis (6.3.1.6) as a byproduct of Tukey’s method. But a most famous significance-testing method for hypothesis (6.3.1.6) also deserves discussion: the one-way analysis of variance test.

At this point it may seem strange that a test about means has a name apparently emphasizing variance. The motivation for this jargon is that the test is associated with a very helpful way of thinking about partitioning the overall variability that is encountered in a response variable. This is the one-way ANOVA F Test.

## 6.3.2 The One-Way ANOVA $F$ Test

The standard method of testing the hypothesis (6.3.2.6)

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

of no differences among  $r$  means against

$$H_a : \text{not } H_0$$

is based essentially on a comparison of a measure of variability among the sample means to the pooled sample variance,  $s_p^2$ . In order to fully describe this method some additional notational conventions are needed.

Repeatedly in the balance of this book, it will be convenient to have symbols for the summary measures of Part 2 (sample means and variances) applied to the data from multisample studies, ignoring the fact that there are  $r$  different samples involved. Already the unsubscripted letter  $n$  has been used to stand for  $n_1 + n_2 + \cdots + n_r$ , the number of observations in hand ignoring the fact that  $r$  samples are involved. This kind of convention will now be formally extended to include statistics calculated from the  $n$  responses. For emphasis, this will be stated in definition form.

### DEFINITION 6.3.2.1 A Notational Convention for Multisample Studies

In multisample studies, symbols for sample sizes and sample statistics appearing without subscript indices or dots will be understood to be calculated from all responses in hand, obtained by combining all samples.

So  $n$  will stand for the total number of data points (even in an  $r$ -sample study),  $\bar{y}$  for the grand sample average of response  $y$ , and  $s^2$  for a grand sample variance calculated completely ignoring sample boundaries.

For present purposes (of writing down a test statistic for testing hypothesis (6.3.1.6)), one needs to make use of  $\bar{y}$ , the grand sample average. It is important to recognize that  $\bar{y}$  and

### 6.3.2.1 The (unweighted) average of $r$ sample means

$$\bar{y}_{\cdot} = \frac{1}{r} \sum_{i=1}^r \bar{y}_i$$

are not necessarily the same unless all sample sizes are equal. That is, when sample sizes vary,  $\bar{y}$  is the (unweighted) arithmetic average of the raw data values  $y_{ij}$  but is a weighted average of the sample means  $\bar{y}_i$ . On the other hand,  $\bar{y}_i$  is the (unweighted) arithmetic mean of the sample means  $\bar{y}_i$  but is a weighted average of the raw data values  $y_{ij}$ . For example, in the simple case that  $r = 2$ ,  $n_1 = 2$ , and  $n_2 = 3$ ,

$$\bar{y} = \frac{1}{5}(y_{11} + y_{12} + y_{21} + y_{22} + y_{23}) = \frac{2}{5}\bar{y}_1 + \frac{3}{5}\bar{y}_2$$

while

$$\bar{y}_i = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{6}y_{21} + \frac{1}{6}y_{22} + \frac{1}{6}y_{23}$$

and, in general,  $\bar{y}_{\cdot}$  and  $\bar{y}_i$  will not be the same.

Now, under the hypothesis (6.3.1.6), that  $\mu_1 = \mu_2 = \dots = \mu_r$ ,  $\bar{y}$  is a natural estimate of the common mean. (All underlying distributions are the same, so the data in hand are reasonably thought of not as  $r$  different samples, but rather as a single sample of size  $n$ .) Then the differences  $\bar{y}_i - \bar{y}$  are indicators of possible differences among the  $\mu_i$ . It is convenient to summarize the size of these differences  $\bar{y}_i - \bar{y}$  in terms of a kind of total of their squares—namely,

$$6.3.2.2 \quad \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2$$

One can think of statistic (6.3.2.2) either as a weighted sum of the quantities  $(\bar{y}_i - \bar{y})^2$  or as an unweighted sum, where there is a term in the sum for each raw data point and therefore  $n_i$  of the type  $(\bar{y}_i - \bar{y})^2$ . The quantity (6.3.2.2) is a measure of the between-sample variation in the data. For a given set of sample sizes, the larger it is, the more variation there is between the sample means  $\bar{y}_i$ .

In order to produce a test statistic for hypothesis (6.3.1.6), one simply divides the measure (6.3.2.2) by  $(r - 1)s_p^2$ , giving

### 6.3.2.3 One-way ANOVA test statistic for equality of $r$ means

$$F = \frac{\frac{1}{r-1} \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2}{s_P^2}$$

The fact is that if  $\mathbf{H}_0 : \mu_1 = \mu_2 = \dots = \mu_r$  is true, the one-way model assumptions imply that this statistic has an  $F_{r-1, n-r}$  distribution. So the hypothesis of equality of  $r$  means can be tested using the statistic in equation (6.3.2.3) with an  $F_{r-1, n-r}$  reference distribution, where large observed values of  $F$  are taken as evidence against  $\mathbf{H}_0$  in favor of  $\mathbf{H}_a : \text{not } \mathbf{H}_0$ .

#### Example 6.3.2.1 Concrete Compression Study continued.

Returning again to the concrete compressive strength study of Armstrong, Babb, and Campen,  $\bar{y} = 3,693.6$  and the 8 sample means  $\bar{y}_i$  have differences from this value given in Table 6.3.2.1.

Then since each  $n_i = 3$ , in this situation,

$$\begin{aligned} \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2 &= 3(1,941.7)^2 + 3(2,059.7)^2 + \dots \\ &\quad + 3(-2,142.3)^2 + 3(-1,302.9)^2 \\ &= 47,360,780(\text{psi})^2 \end{aligned}$$

In order to use this figure to judge statistical significance, one standardizes via equation (6.3.2.3) to arrive at the observed value of the test statistic

$$f = \frac{\frac{1}{8-1}(47,360,780)}{(581.6)^2} = 20.0$$

It is easy to verify from Tables A3 (The F Tables) that 20.0 is larger than the .999 quantile of the  $F_{7,16}$  distribution. So

$$p\text{-value} = P[\text{an } F_{7,16} \text{ random variable} \geq 20.0] < .001$$

That is, the data provide overwhelming evidence that  $\mu_1, \mu_2, \dots, \mu_8$  are not all equal.



Sample Means and Their  
Deviations from  $\bar{y}$  in the Concrete  
Strength Study

$i$ , Formula	$\bar{y}_i$	$\bar{y}_i - \bar{y}$
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

*Table 6.3.2.1 Sample Means and Their  
Deviations from  $\bar{y}$  in the  
Concrete Strength Study*

For pedagogical reasons, the one-way ANOVA test has been presented after discussing interval-oriented methods of inference for  $r$ -sample studies. But if it is to be used in applications, the testing method typically belongs chronologically before estimation. That is, the ANOVA test can serve as a screening device to determine whether the data in hand are adequate to differentiate conclusively between the means, or whether more data are needed.

### 6.3.3 The One-Way ANOVA Identity and Table

Associated with the ANOVA test statistic is some strong intuition related to the partitioning of observed variability. This is related to an algebraic identity that is stated here in the form of a proposition.

**Proposition 6.3.3.1**

**One-Way ANOVA Identity**

For any  $n$  numbers  $y_{ij}$

$$6.3.3.1 \quad (n - 1)s^2 = \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2 + (n - r)s_P^2$$

or in other symbols,

A second statement of the one-way ANOVA identity

$$6.3.3.2 \quad \sum_{i,j} (y_{ij} - \bar{y})^2 = \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Proposition 6.3.3.1 should begin to shed some light on the phrase “analysis of variance.” It says that an overall measure of variability in the response  $y$ , namely,

$$(n - 1)s^2 = \sum_{i,j} (y_{ij} - \bar{y})^2$$

can be partitioned or decomposed algebraically into two parts. One,

$$\sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2$$

can be thought of as measuring variation between the samples or “treatments,” and the other,

$$(n - r)s_P^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

measures variation within the samples (and in fact consists of the sum of the squared residuals). The  $F$  statistic (6.3.2.3), developed for testing  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ , has a numerator related to the first of these and a denominator related to the second. So using the ANOVA  $F$  statistic amounts to a kind of analyzing of the raw variability in  $y$ .

In recognition of their prominence in the calculation of the one-way ANOVA  $F$  statistic and their usefulness as descriptive statistics in their own right, the three sums (of squares) appearing in formulas (6.3.3.1) and (6.3.3.2) are usually given special names and shorthand. These are stated here in definition form.

**DEFINITION 6.3.3.1 Total Sum of Squares SSTot**

In a multisample study,  $(n - 1)s^2$ , the sum of squared differences between the raw data values and the grand sample mean, will be called the total sum of squares and denoted as SSTot.

**DEFINITION 6.3.3.2 Treatment Sum of Squares SSTR**

In an unstructured multisample study,  $\sum n_i (\bar{y}_i - \bar{y})^2$  will be called the treatment sum of squares and denoted as SSTR.

**DEFINITION 6.3.3.3 Error Sum of Squares SSE**

In a multisample study, the sum of squared residuals,  $\sum (y - \hat{y})^2$  (which is  $(n - r)s_p^2$  in the unstructured situation) will be called the error sum of squares and denoted as SSE.

In the new notation introduced in these definitions, Proposition 1 states that in an unstructured multisample context,

**6.3.3.4 A third statement of the one-way ANOVA identity**

$$SSTot = SSTr + SSE$$

Partially as a means of organizing calculation of the  $F$  statistic given in formula (6.3.2.3) and partially because it reinforces and extends the variance partitioning insight provided by formulas (6.3.3.1), (6.3.3.2), and (6.3.3.3), it is useful to make an ANOVA table. There are many forms of ANOVA tables corresponding to various multisample analyses. The form most relevant to the present situation is given in symbolic form as Table 6.3.3.1.

The column headings in Table 6.3.3.1 are Source (of variation), Sum of Squares (corresponding to the source), degrees of freedom (corresponding to the source), Mean Square (corresponding to the source), and  $\underline{F}$  (for testing the significance of the source in contributing to the overall observed variability). The entries in the Source column of the table are shown here as being Treatments, Error, and Total. But the name Treatments is sometimes replaced by Between (Samples), and the name Error is sometimes replaced by Within (Samples) or Residual. The first two entries in the SS column must sum to the third, as indicated in equation (6.3.3.3). Similarly, the Treatments and Error degrees of freedom add to the Total degrees of freedom,  $(n - 1)$ . Notice that the entries in the  $df$  column are those attached to the numerator and denominator, respectively, of the test statistic in equation (6.3.2.3). The ratios of sums of squares to degrees of freedom are called mean squares, here the mean square for treatments (MSTr) and the mean square for error (MSE). Verify that in the present context,  $MSE = s_p^2$  and  $MSTr$  is the numerator of the  $F$  statistic given in equation (6.3.2.3). So the single ratio appearing in the  $F$  column is the observed value of  $F$  for testing  $H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$ .

General Form of the One-Way ANOVA Table

ANOVA Table (for testing $H_0: \mu_1 = \mu_2 = \dots = \mu_r$ )				
Source	SS	df	MS	F
Treatments	$SSTr$	$r - 1$	$SSTr/(r - 1)$	$MSTr/MSE$
Error	$SSE$	$n - r$	$SSE/(n - r)$	
Total	$SSTot$	$n - 1$		

Table 6.3.3.1 General Form of the One-Way ANOVA Table

**Example 6.3.3.1 Concrete Strength Study continued.**

Consider once more the concrete strength study. It is possible to return to the raw data given in Table 6.1.1.1 and find that  $\bar{y} = 3,693.6$ , so

$$\begin{aligned} SSTot &= (n - 1)s^2 \\ &= (5,800 - 3,693.6)^2 + (4,598 - 3,693.6)^2 + (6,508 - 3,693.6)^2 \\ &\quad + \dots + (2,631 - 3,693.6)^2 + (2,490 - 3,693.6)^2 \\ &= 52,772,190(\text{psi})^2 \end{aligned}$$

Further, as in Section 6.1.1.1,  $s_p^2 = 338,213.1(\text{psi})^2$  and  $n - r = 16$ , so

$$SSE = (n - r)s_p^2 = 5,411,410(\text{psi})^2$$

And from earlier in this section,

$$SSTr = \sum_{i=1}^r n_i (\bar{y}_i - \bar{y})^2 = 47,360,780$$

Then, plugging these and appropriate degrees of freedom values into the general form of the one-way ANOVA table produces the table for the concrete compressive strength study, presented here as Table 6.3.3.2.

Notice that, as promised by the one-way ANOVA identity, the sum of the treatment and error sums of squares is the total sum of squares. Also, Table 6.3.3.2 serves as a helpful summary of the testing process, showing at a glance the observed value of  $F$ , the appropriate degrees of freedom, and  $s_p^2 = MSE$ .

One-Way ANOVA Table for the Concrete Strength Study

<b>ANOVA Table (for testing <math>H_0: \mu_1 = \mu_2 = \dots = \mu_8</math>)</b>				
<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Treatments	47,360,780	7	6,765,826	20.0
Error	5,411,410	16	338,213	
Total	52,772,190	23		

Table 6.3.3.2 One-Way ANOVA Table for the Concrete Strength Study

## 6.3.4 Computing ANOVA in Python

The computations here in Part 6 are by no means impossible to do “by hand.” But the most sensible way to handle them is to employ a statistical package.

Using Python and Jupyter Notebooks, here we show the figures, ANOVA Table, and output for the a One-Way Analysis of the Concrete Strength Data, illustrating much of the content from this Part 6. **It is strongly recommended that you consult the [Hypothesis Testing Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “ANOVA” will be particularly useful.

For an interactive and step by step discussion of this example and output, [Binder Site for Special GitHub Part 6 ANOVA Example](#).

Or go to special GitHub site: <https://github.com/Statistical-Methods-for-Engineering/Special-GitHub-Site-Part-6-ANOVA-and-Compression-Strength-Example>, the Special GitHub Site Part 6: ANOVA and Compression Strength Example, and click on the BinderLink to enable the interactive tutorial.

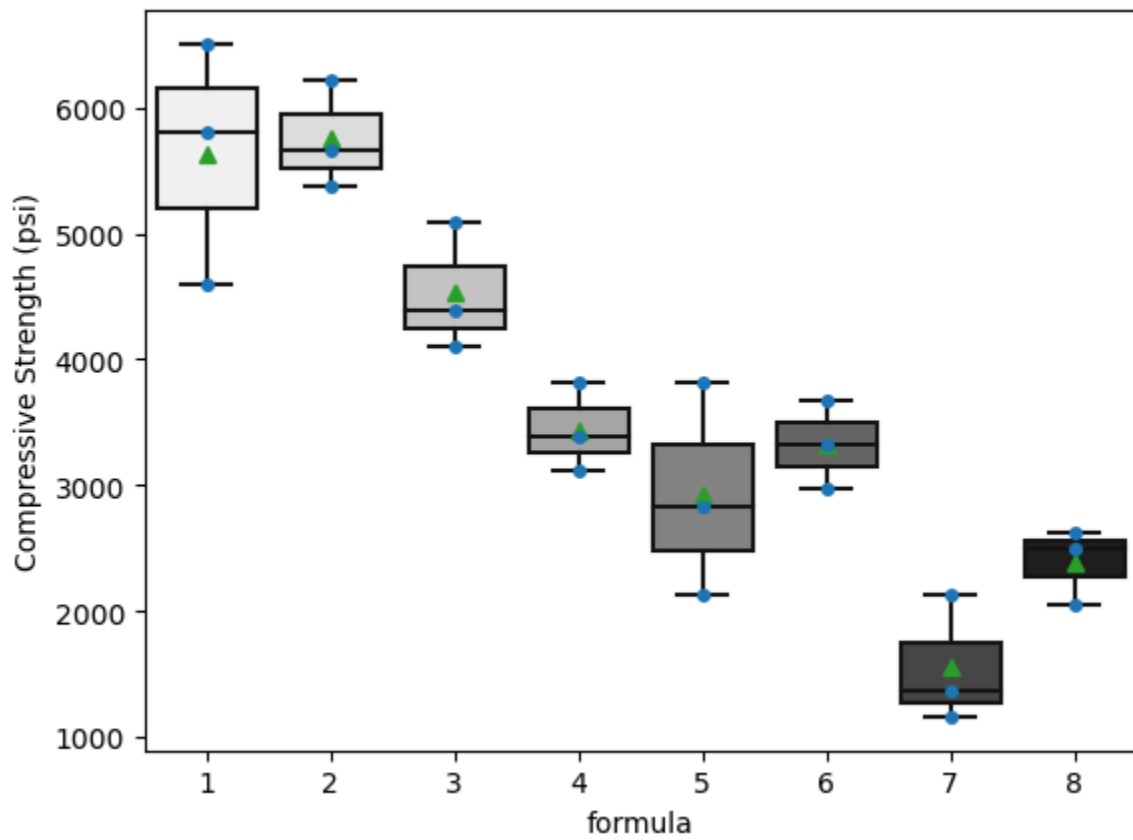


Figure 6.3.4.1 Boxplot of the eight formulas showing the compression strength.

	df	sum_sq	mean_sq	F	PR(>F)
<b>formula</b>	7.0	4.736078e+07	6.765826e+06	20.004625	8.550775e-07
<b>Residual</b>	16.0	5.411409e+06	3.382131e+05	NaN	NaN

Table 6.3.4.1 ANOVA Table for compression strength example.



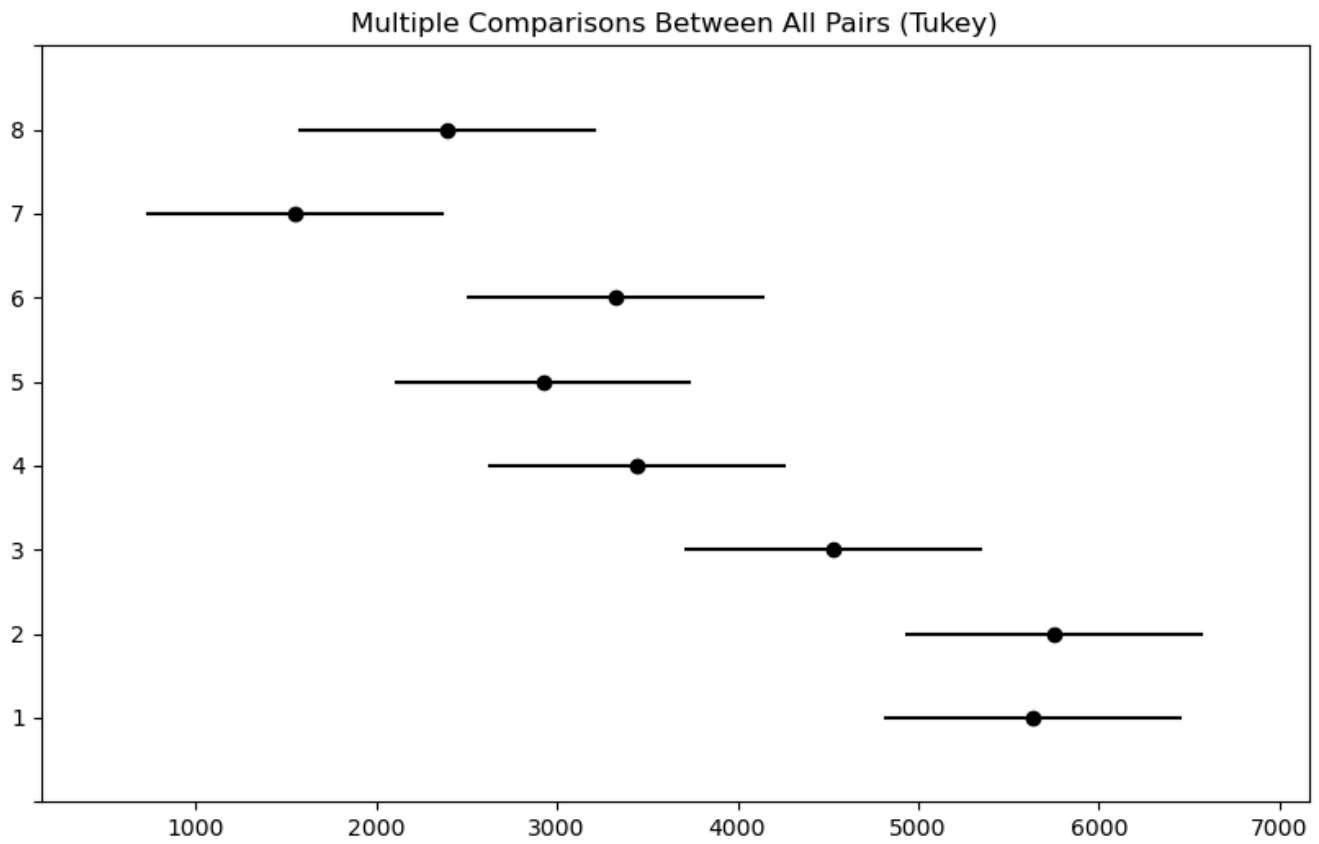


Figure 6.3.4.2 Multiple Comparison between pairs,, (Tukey Method) for simultaneous comparisons.

Or as,always, this Jupyter Notebook can be found on the course GitHub site with Tutorials and other examples under Part 6 ANOVA: [IntroEngStatsMethods\\_GitHub Site](#).

## 7.0.1 Introduction Least Squares and Simple Linear Regression Analysis

This Part begins a new idea: we start considering more than one variable at a time. However, you will see the tools of confidence intervals and visualization from the previous sections coming into play so that we can interpret our least squares models both analytically and visually.

The following sections, on design and analysis of experiments will build on the least squares model we learn about here.

The material in this section is used whenever you need to interpret and quantify the relationship between two or more variables. Examples of this kind quantification that can be explored include:

- *Colleague*: How is the yield from our lactic acid batch fermentation related to the purity of the sucrose substrate?
- *You*: The yield can be predicted from sucrose purity with an error of plus/minus 8%
- *Colleague*: And how about the relationship between yield and glucose purity?
- *You*: Over the range of our historical data, there is no discernible relationship.
  
- *Engineer 1*: The theoretical equation for the melt index is non-linearly related to the viscosity
- *Engineer 2*: The linear model does not show any evidence of that, but the model's prediction ability does improve slightly when we use a non-linear transformation in the least squares model.
  
- *HR manager*: We use a least squares regression model to graduate personnel through our pay grades. The model is a function of education level and number of years of experience. What do the model coefficients mean?

## 7.0.2 Attributions

This first draft of Part 7 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

## 7.1.0 Introduction to Least Squares: Describing the Relationship between Bivariate Quantitative Data

Bivariate data often arise because a quantitative experimental variable  $x$  has been varied between several different settings, producing a number of samples of a response variable  $y$ . For purposes of summarization, interpolation, limited extrapolation, and/or process optimization/adjustment, it is extremely helpful to have an equation relating  $y$  to  $x$ . A linear (or straight line) equation:

### EXPRESSION 7.1.0.1

$$y \approx \beta_0 + \beta_1 x$$

relating  $y$  to  $x$  is about the simplest potentially useful equation to consider after making a simple  $(x, y)$  scatterplot.

In this section, the principle of least squares is used to fit a line to  $(x, y)$  data. The appropriateness of that fit is assessed using the sample correlation and the coefficient of determination. Plotting of residuals is introduced as an important method for further investigation of possible problems with the fitted equation. A discussion of some practical cautions and the use of statistical software in fitting equations to data follows.

## 7.1.1: Applying the Least Squares Principle

### Example 7.1.1.1: Pressing Pressures and Specimen Densities for a Ceramic Compound

Benson, Locher, and Watkins studied the effects of varying pressing pressures on the density of cylindrical specimens made by dry pressing a ceramic compound. A mixture of  $\text{Al}_2\text{O}_3$ , polyvinyl alcohol, and water was prepared, dried overnight, crushed, and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures from 2,000 psi to 10,000 psi, and cylinder densities were calculated. Table 7.1.1.1 gives the data that were obtained, and a simple scatterplot of these data is given in Figure 7.1.1.1.

$x$ , Pressure (psi)	$y$ , Density (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

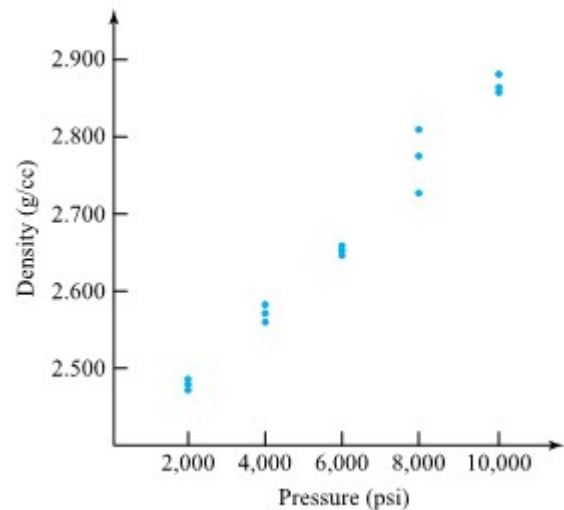


Figure 7.1.1.1: Scatterplot of density vs. pressing pressure

Table 7.1.1.1: Pressing Pressures and Resultant Specimen Densities

It is very easy to imagine sketching a straight line through the plotted points in Figure 7.1.1.1. Such a line could then be used to summarize how density depends upon pressing pressure. The principle of least squares provides a method of choosing a “best” line to describe the data.

**DEFINITION Principle of Least Squares****EXPRESSION 7.1.1.1**

To apply the principle of least squares in the fitting of an equation for  $y$  to an  $n$ -point data set, values of the equation parameters are chosen to minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where  $y_1, y_2, \dots, y_n$  are the observed responses and  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$  are corresponding responses predicted or fitted by the equation.

In the context of fitting a line to  $(x, y)$  data, the prescription offered by Definition 7.1.1.1 amounts to choosing a slope and intercept so as to minimize the sum of squared vertical distances from  $(x, y)$  data points to the line in question. This notion is shown in generic fashion in Figure 7.1.1.2 for a fictitious five-point data set. (It is the squares of the five indicated differences that must be added and minimized.)

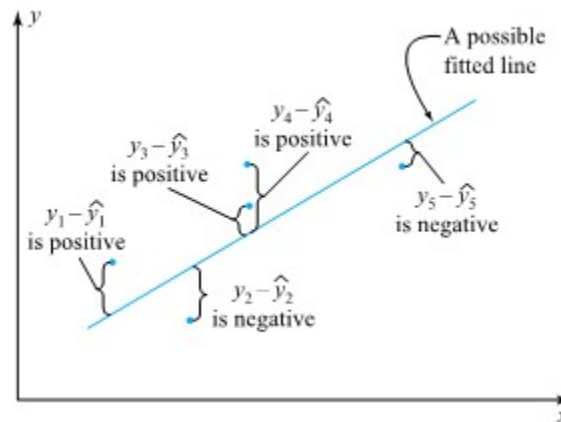


Figure 7.1.1.2 Five data points  $(x, y)$  and a possible fitted line.

Looking at the form of display (7.1.0.1), for the fitting of a line,

$$\hat{y} = \beta_0 + \beta_1 x$$

Therefore, the expression to be minimized by choice of slope ( $\beta_1$ ) and intercept ( $\beta_0$ ) is

$$7.1.1.2 \quad S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

The minimization of the function of two variables  $S(\beta_0, \beta_1)$  is an exercise in calculus. The partial derivatives of  $S$  with respect to  $\beta_0$  and  $\beta_1$  may be set equal to zero, and the two resulting equations may be solved simultaneously for  $\beta_0$  and  $\beta_1$ . The equations produced in this way are

$$7.1.1.3 \quad n\beta_0 + \left( \sum_{i=1}^n x_i \right) \beta_1 = \sum_{i=1}^n y_i$$

and

$$7.1.1.4 \quad \left( \sum_{i=1}^n x_i \right) \beta_0 + \left( \sum_{i=1}^n x_i^2 \right) \beta_1 = \sum_{i=1}^n x_i y_i$$

For reasons that are not obvious, equations (7.1.1.3) and (7.1.1.4) are sometimes called the normal (as in perpendicular) equations for fitting a line. They are two linear equations in two unknowns and can be fairly easily solved for  $\beta_0$  and  $\beta_1$  (provided there are at least two different  $x_i$ 's in the data set). Simultaneous solution of equations (7.1.1.3) and (7.1.1.4) produces values of  $\beta_0$  and  $\beta_1$  given by

Slope of the least squares line,  $\beta_1$  7.1.1.5

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and

Intercept of the least squares line,  $\beta_0$  7.1.1.6

$$b_0 = \bar{y} - b_1 \bar{x}$$

Notice the notational convention here. The particular numerical slope and intercept minimizing  $S(\beta_0, \beta_1)$  are denoted (not as  $\beta$ 's but) as  $b_1$  and  $b_0$ .

A note about expression (7.1.1.5) and the somewhat standard practice that has been followed (and the summation notation abused) by not indicating the variable or range of summation ( $i$ , from 1 to  $n$ ).

Example 7.1.1.2 continued

It is possible to verify that the data in Table 7.1.1.1 yield the following summary statistics:

$$\sum x_i = 2,000 + 2,000 + \cdots + 10,000 = 90,000,$$

$$\text{so } \bar{x} = \frac{90,000}{15} = 6,000$$

$$\sum (x_i - \bar{x})^2 = (2,000 - 6,000)^2 + (2,000 - 6,000)^2 + \cdots +$$

$$\sum y_i = 2.486 + 2.479 + \cdots + 2.858 = 40.005,$$

$$\text{so } \bar{y} = \frac{40.005}{15} = 2.667$$

$$\sum (y_i - \bar{y})^2 = (2.486 - 2.667)^2 + (2.479 - 2.667)^2 + \cdots +$$

$$(2.858 - 2.667)^2 = .289366$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (2,000 - 6,000)(2.486 - 2.667) + \cdots +$$

$$(10,000 - 6,000)(2.858 - 2.667) = 5,840$$

Then the least squares slope and intercept,  $b_1$  and  $b_0$ , are given via equations (7.1.1.5) and (7.1.1.6) as

$$b_1 = \frac{5,840}{120,000,000} = .000048\bar{6} \text{ (g/cc)/psi}$$

and

$$b_0 = 2.667 - (.000048\bar{6})(6,000) = 2.375 \text{ g/cc}$$

Figure 7.1.1.3 shows the least squares line

$$\hat{y} = 2.375 + .0000487x$$

sketched on a scatterplot of the  $(x, y)$  points from Table 7.1.1.1.

### Interpretation of the slope of the least squares line

Note that the slope on this plot,  $b_1 \approx 0.0000487 \text{ (g/cc)/psi}$ , has physical meaning as the (approximate) increase in  $y$  (density) that accompanies a unit (1 psi) increase in  $x$  (pressure).

### Interpretation of y-intercept and careful for extrapolation

The intercept on the plot,  $b_0 = 2.375 \text{ g/cc}$ , positions the line vertically and is the value at which the line cuts the  $y$  axis. But it should probably not be interpreted as the density that would accompany a pressing pressure of  $x = 0$  psi. The point is that the reasonably linear-looking relation that the investigators found for pressures between 2,000 psi and 10,000 psi could well break down at larger or smaller pressures. Thinking of  $b_0$  as a 0 pressure density amounts to an extrapolation outside the range of data used to fit the equation, something that ought always to be approached with extreme caution.



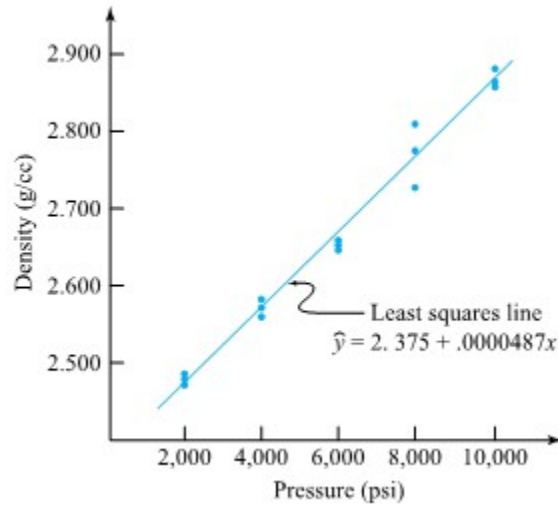


Figure 7.1.1.3 Scatterplot of the pressure/density data and the least squares line.

As indicated in Definition 7.1.1.1, the value of  $y$  on the least squares line corresponding to a given  $x$  can be termed a fitted or predicted value. It can be used to represent likely  $y$  behavior at that  $x$ .

Example. 7.1.1.3 continued.

Consider the problem of determining a typical density corresponding to a pressure of 4,000 psi and one corresponding to 5,000 psi. First, looking at  $x = 4,000$ , a simple way of representing a typical  $y$  is to note that for the three data points having  $x = 4,000$ ,

$$\bar{y} = \frac{1}{3}(2.558 + 2.570 + 2.580) = 2.5693 \text{ g/cc}$$

and so to use this as a representative value. But assuming that  $y$  is indeed approximately linearly related to  $x$ , the fitted value

$$\hat{y} = 2.375 + .000048\bar{6}(4,000) = 2.5697 \text{ g/cc}$$

might be even better for representing average density for 4,000 psi pressure.

### Interpolation

fitted value

Looking then at the situation for  $x = 5,000$  psi, there are no data with this  $x$  value. The only thing one can do to represent density at that pressure is to ask

whether interpolation is sensible from a physical viewpoint. If so, the

$$\hat{y} = 2.375 + .000048\bar{6}(5,000) = 2.6183 \text{ g/cc}$$

an be used to represent density for 5,000 psi pressure.

## 7.1.2 The Sample Correlation and Coefficient of Determination

### CORRELATION

Visually, the least squares line in Figure 7.1.1.3 seems to do a good job of fitting the plotted points. However, it would be helpful to have methods of quantifying the quality of that fit. One such measure is the sample correlation.

#### DEFINITION Sample (linear) correlation

#### EXPRESSION 7.1.2.1

The sample (linear) correlation between  $x$  and  $y$  in a sample of  $n$  data pairs  $(x_i, y_i)$  is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

#### Interpreting the sample correlation

The sample correlation always lies in the interval from  $-1$  to  $1$ . Further, it is  $-1$  or  $1$  only when all  $(x, y)$  data points fall on a single straight line. Comparison of formulas (7.1.1.5) and (7.1.2.1) shows that  $r = b_1 \left( \sum (x_i - \bar{x})^2 / \sum (y_i - \bar{y})^2 \right)^{1/2}$  so that  $b_1$  and  $r$  have the same sign. So a sample correlation of  $-1$  means that  $y$  decreases linearly in increasing  $x$ , while a sample correlation of  $+1$  means that  $y$  increases linearly in increasing  $x$ .

Real data sets do not often exhibit perfect ( $+1$  or  $-1$ ) correlation. Instead  $r$  is typically between  $-1$  and  $1$ . But drawing on the facts about how it behaves, people take  $r$  as a measure of the strength of an apparent linear relationship:  $r$  near  $+1$  or  $-1$  is interpreted as indicating a relatively strong linear relationship;  $r$  near  $0$  is taken as indicating a lack of linear relationship. The sign of  $r$  is thought of as indicating whether  $y$  tends to increase or decrease with increased  $x$ .

Example 7.1.2.2 continued

For the pressure/density data, the summary statistics in the example produces

$$r = \frac{5,840}{\sqrt{(120,000,000)(.289366)}} = .9911$$

This value of  $r$  is near +1 and indicates clearly the strong positive linear relationship evident in Figures 7.1.1.1 and 7.1.1.3

## COEFFICIENT OF DETERMINATION

### DEFINITION Coefficient of Determination

#### EXPRESSION 7.1.2.2

The coefficient of determination for an equation fitted to an  $n$ -point data set via least squares and producing fitted  $y$  values

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

### Interpretation of $R^2$

$R^2$  may be interpreted as *the fraction of the raw variation in  $y$  accounted for using the fitted equation*. That is, provided the fitted equation includes a constant term,  $\sum (y_i - \bar{y})^2 \geq \sum (y_i - \hat{y}_i)^2$ . Further,  $\sum (y_i - \bar{y})^2$  is a measure of raw variability in  $y$ , while  $\sum (y_i - \hat{y}_i)^2$  is a measure of variation in  $y$  remaining after fitting the equation. So the nonnegative difference  $\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$  is a measure of the variability in  $y$  accounted for in the equation-fitting process.  $R^2$  then expresses this difference as a fraction (of the total raw variation).

Example 7.1.2.2 continued.

Using the fitted line, one can find  $\hat{y}$  values for all  $n = 15$  data points in the original data set. These are given in Table 7.1.2.1

$x$ , Pressure	$\hat{y}$ , Fitted Density
2,000	2.4723
4,000	2.5697
6,000	2.6670
8,000	2.7643
10,000	2.8617

Table 7.1.2.1 Fitted Density Values

Then, referring again to Table 7.1.1.1,

$$\begin{aligned} \sum (y_i - \hat{y}_i)^2 &= (2.486 - 2.4723)^2 + (2.479 - 2.4723)^2 + (2.472 - 2.4723)^2 \\ &\quad + (2.558 - 2.5697)^2 + \dots + (2.879 - 2.8617)^2 \\ &\quad + (2.858 - 2.8617)^2 \\ &= .005153 \end{aligned}$$

Further, since  $\sum (y_i - \bar{y})^2 = .289366$  from equation 7.1.2.2

$$R^2 = \frac{.289366 - .005153}{.289366} = .9822$$

and the fitted line accounts for over 98% of the raw variability in density, reducing the “unexplained” variation from .289366 to .005153.

### $R^2$ as a squared correlation

The coefficient of determination has a second useful interpretation. For equations that are linear in the parameters (which are the only ones considered here and which will be discussed in detail later),  $R^2$  turns out to be a squared correlation. It is the squared correlation between the observed values  $y_i$  and the fitted values  $\hat{y}_i$ . (Since in the present situation of fitting a line, the  $\hat{y}_i$  values are perfectly correlated with the  $x_i$  values,  $R^2$  also turns out to be the squared correlation between the  $y_i$  and  $x_i$  values.)

Example 7.1.2.3 continued.

For the pressure/density data, the correlation between  $x$  and  $y$  is

$$r = .9911$$

Since  $\hat{y}$  is perfectly correlated with  $x$ , this is also the correlation between  $\hat{y}$  and  $y$ . But notice as well that

$$r^2 = (.9911)^2 = .9822 = R^2$$

so  $R^2$  is indeed the squared sample correlation between  $y$  and  $\hat{y}$ .

## 7.1.3 Computing and Using Residuals

When fitting an equation to a set of data, the hope is that the equation extracts the main message of the data, leaving behind (unpredicted by the fitted equation) only the variation in  $y$  that is uninterpretable. That is, one hopes that the  $y_i$ 's will look like the  $\hat{y}_i$ 's except for small fluctuations explainable only as random variation. A way of assessing whether this view is sensible is through the computation and plotting of **residuals**.

### DEFINITION Residuals

#### EXPRESSION 7.1.3.1

If the fitting of an equation or model to a data set with responses  $y_1, y_2, \dots, y_n$  produces fitted values  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$  then the corresponding residuals are the values

$$e_i = y_i - \hat{y}_i$$

If a fitted equation is telling the whole story contained in a data set, then its residuals ought to be patternless. So when they're plotted against time order of observation, values of experimental variables, fitted values, or any other sensible quantities, the plots should look randomly scattered. When they don't, the patterns can themselves suggest what has gone unaccounted for in the fitting and/or how the data summary might be improved.

#### Example 7.1.3.1 Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phosphate Additive

As an exaggerated example of the previous point, consider the naive fitting of a line to some data of B. Roth. Roth studied the compressive strength of concrete-like fly ash cylinders. These were made using varying amounts of ammonium phosphate as an additive. Part of Roth's data are given in Table 7.1.3.1. The ammonium phosphate values are expressed as a percentage by weight of the amount of fly ash used.

x, Ammonium Phosphate (%)	y, Compressive Strength (psi)	x, Ammonium Phosphate (%)	y, Compressive Strength (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

Table 7.1.3.1. Additive Concentrations and Compressive Strengths for Fly Ash Cylinders

Using formulas (7.1.1.5) and (7.1.1.6), it is possible to show that the least squares line through the (x, y) data in Table 7.1.3.1 is

7.1.3.2

$$\hat{y} = 1498.4 - .6381x$$

Then straightforward substitution into equation (7.1.3.2) produces fitted values  $\hat{y}_i$  and residuals  $e_i = y_i - \hat{y}_i$ , as given in Table 7.1.3.2. The residuals for this straight-line fit are plotted against x in Figure 7.1.3.1.

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

Table 7.1.3.2 Residuals from a Straight-Line Fit to the Fly Ash Data

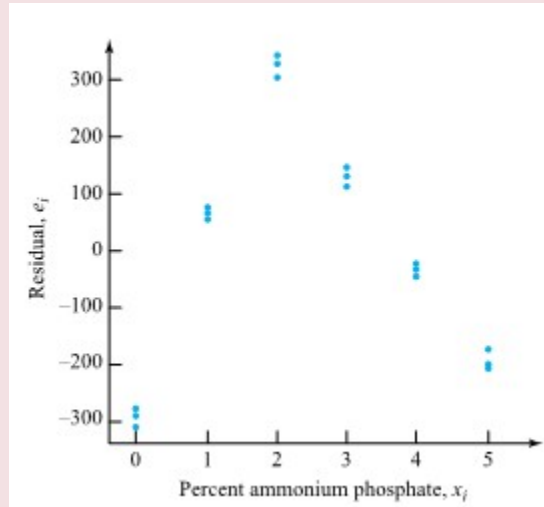


Figure 7.1.3.1 Plot of residuals vs.  $x$  for a linear fit to the fly ash data.

The distinctly “up-then-back-down-again” curvilinear pattern of the plot in Figure 7.1.3.1 is not typical of random scatter. Something has been missed in the fitting of a line to Roth’s data. Figure 7.1.3.2 is a simple scatterplot of Roth’s data (which in practice should be made before fitting any curve to such data). It is obvious from the scatterplot that the relationship between the amount of ammonium phosphate and compressive strength is decidedly nonlinear. In fact, a quadratic function would come much closer to fitting the data in Table 7.1.3.1.

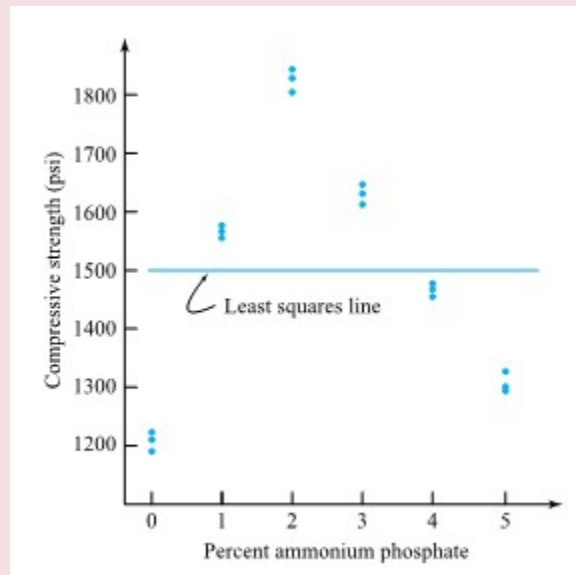


Figure 7.1.3.2 Scatterplot of the fly ash data.

## Interpreting patterns on residual plots

Figure 7.1.3.3 shows several patterns that can occur in plots of residuals against various variables. Plot 1 of Figure 7.1.3.3 shows a trend on a plot of residuals versus time order of observation. The pattern

suggests that some variable changing in time is acting on  $y$  and has not been accounted for in fitting  $\hat{y}$  values. For example, instrument drift (where an instrument reads higher late in a study than it did early on) could produce a pattern like that in Plot 1. Plot 2 shows a fan-shaped pattern on a plot of residuals versus fitted values. Such a pattern indicates that large responses are fitted (and quite possibly produced and/or measured) less consistently than small responses. Plot 3 shows residuals corresponding to observations made by Technician 1 that are on the whole smaller than those made by Technician 2. The suggestion is that Technician 1's work is more precise than that of Technician 2.

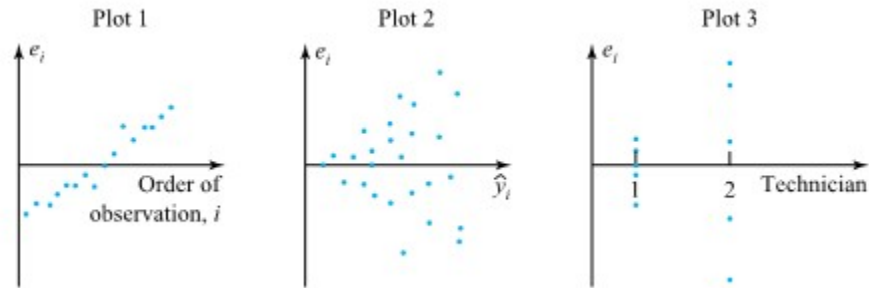


Figure 7.1.3.3 Patterns in residual plots.

## Normal-plotting residuals

Another useful way of plotting residuals is to normal-plot them. The idea is that the normal distribution shape is typical of random variation and that normal-plotting of residuals is a way to investigate whether such a distributional shape applies to what is left in the data after fitting an equation or model.

Another useful way of plotting residuals is to normal-plot them. The idea is that the normal distribution shape is typical of random variation and that normal-

Example 7.1.3.2 continued.

Table 7.1.3.3 gives residuals for the fitting of a line to the pressure/density data. The residuals  $e_i$  were treated as a sample of 15 numbers and normal-plotted (using the methods we have introduced previously) to produce Figure 7.1.3.4.

The central portion of the plot in Figure 7.1.3.4 is fairly linear, indicating a generally bell-shaped distribution of residuals. But the plotted point corresponding to the largest residual, and probably the one corresponding to the smallest residual, fail to conform to the linear pattern established by the others. Those residuals seem big in absolute value compared to the others.

From Table 7.1.3.3 and the scatterplot in Figure 7.1.1.3, one sees that these large residuals both arise from the 8,000 psi condition. And the spread for the three densities at that pressure value does indeed look considerably larger than those at the other pressure values. The normal plot suggests that the pattern of variation at 8,000 psi is genuinely different from those at other pressures. It may be that a different physical compaction mechanism was acting at 8,000 psi than at the other pressures. But it is more likely that there was a problem with laboratory technique, or recording, or the test equipment when the 8,000 psi tests were made.

In any case, the normal plot of residuals helps draw attention to an idiosyncrasy in the data of Table 7.1.1.1 that merits further investigation, and perhaps some further data collection.



$x$ , Pressure	$y$ , Density	$\hat{y}$	$e = y - \hat{y}$
2,000	2.486	2.4723	.0137
2,000	2.479	2.4723	.0067
2,000	2.472	2.4723	-.0003
4,000	2.558	2.5697	-.0117
4,000	2.570	2.5697	.0003
4,000	2.580	2.5697	.0103
6,000	2.646	2.6670	-.0210
6,000	2.657	2.6670	-.0100
6,000	2.653	2.6670	-.0140
8,000	2.724	2.7643	-.0403
8,000	2.774	2.7643	.0097
8,000	2.808	2.7643	.0437
10,000	2.861	2.8617	-.0007
10,000	2.879	2.8617	.0173
10,000	2.858	2.8617	-.0037

Table 7.3.3.3 Residuals from the Linear Fit to the Pressure/Density Data.

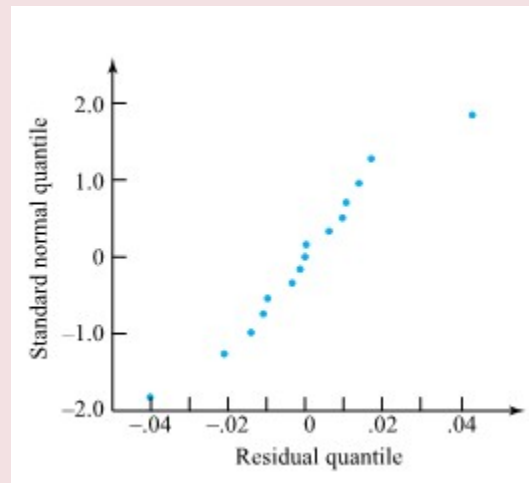


Figure 7.1.3.4 Normal plot of residuals from a linear fit to the pressure/density data.

## 7.1.4 Cautions When Using Least Squares Line Fitting

The methods of this section are extremely useful engineering tools when thoughtfully applied. But a few additional comments are in order, warning against some errors in logic that often accompany their use.

### **r Measures only linear association**

y and yet have a value of  $r$  near 0. In fact,

our second example is an excellent example of this. Compressive strength is strongly related to the ammonium phosphate content. But  $r = -.005$ , very nearly 0, for the data set in Table 7.1.3.1.

The first warning regards the correlation. It must be remembered that  $r$  measures only the linear relationship between  $x$  and  $y$ . It is perfectly possible to have a strong nonlinear relationship between  $x$  and

### **Correlation and causation**

y or vice versa. It may be the case that another variable (say,  $z$ ) drives the system under study and causes simultaneous changes in both  $x$  and  $y$ .

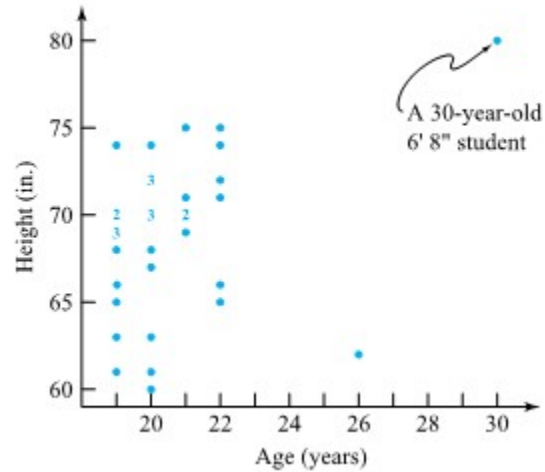
The second warning is essentially a restatement of one implicit in the early part of this discussion: Correlation is not necessarily causation. One may observe a large correlation between  $x$  and  $y$  in an observational study without it being true that  $x$  drives

### **The influence of extreme observations**

course plotted in Figure 4.8. By the time people reach college age, there is little useful relationship between age and height, but the correlation between ages and heights is .73. This fairly large value is produced by essentially a single data point. If the data point corresponding to the 30-year-old student who happened to be 6 feet 8 inches tall is removed from the data set, the correlation drops to .03.

The last warning is that  $r$ ,  $R^2$ , and least squares fitting can be drastically affected by a few unusual data points. As an example of this, consider the ages and

heights of 36 students from an elementary statistics



*Figure 7.1.4.1 Scatterplot of ages and heights of 36 students.*

An engineer's primary insurance against being misled by this kind of phenomenon is the habit of plotting data in as many different ways as are necessary to get a feel for how they are structured. Even a simple boxplot of the age data or height data alone would have identified the 30-year-old student in Figure 7.1.4.1 as unusual. That would have raised the possibility of that data point strongly influencing both  $r$  and any curve that might be fitted via least squares.

## 7.1.5 Using Statistical Computing

The examples in this section have no doubt left the impression that computations were done “by hand.” In practice, such computations are almost always done with a statistical analysis package. The fitting of a line by least squares is generally done using a regression program. Such programs usually also compute  $R^2$  and have an option that allows the computing and plotting of residuals.

This course uses Python coding and Jupyter Notebooks as the statistical computing platform, but there are many others available. Annotated printouts are often included to show how Python formats and shows its outputs.

Printout 7.1.5.1 is such a printout from our GitHub site for an analysis of the pressure/density data in the example from Module 7.1.1, paralleling the discussion in this Part. This can be found to look at or download (as usual) at [Intro Statistical Methods for Engineering](#) under Part 7 or at the [Special GitHub Site for Part 7](#).

Or you can open an interactive computing environment to work through the Jupyter Notebook using Python through a Binder Site using the Special GitHub Site for the Part 7 example. Click [HERE](#) to go to the Binder Site (located at ).

The Statsmodels library from Python that we are using gives its user much more in the way of analysis for least squares curve fitting than has been discussed to this point, so your understanding of the Printout will be incomplete. But it should be possible to locate values of the major summary statistics discussed here.

The regression equation is  
 $\text{density} = 2.375 + 4.867e-05 * \text{pressure}$

```

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 Tutorial 5 - Correlation and Covariance

At this point, it is recommended that you work your way through the [Tutorial 5 exercise](#) found on the associated GitHub repository. This exercise will teach you how to compute covariance and correlation using Python syntax.

**It is strongly recommended that you consult the [Simple Linear Regression Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “Covariance and Correlation” will be particularly useful.

## *7.2.0 Introduction to Simple Linear Regression Inference Methods Related to the Least Squares Fitting of a Line (Simple Linear Regression)*

We have begun a study of inference methods for multisample studies by considering first those which make no explicit use of structure relating several samples and we will end the course discussing some directed at the analysis of factorial structure. The discussion in this module will primarily consider inference methods for multisample studies where factors involved are inherently quantitative and it is reasonable to believe that some approximate functional relationship holds between the values of the system/input/independent variables and observed system responses. That is, this chapter introduces and applies inference methods for the line-fitting contexts discussed in Module 7.1.

This Module begins with a discussion of the simplest situation of this type— namely, where a response variable  $y$  is approximately linearly related to a single quantitative input variable  $x$ . In this specific context, it is possible to give explicit formulas and illustrate in concrete terms what is possible in the way of inference methods for regression analyses. We will then move on to multiple regression (curve- and surface-fitting) analysis in our next module.

This Module considers inference methods that are applicable where a response  $y$  is approximately linearly related to an input/system variable  $x$ . It begins by introducing the (normal) simple linear regression model and discussing how to estimate response variance in this context. Next there is a look at standardized residuals. Then inference for the rate of change ( $\Delta y / \Delta x$ ) is considered, along with inference for the average response at a given  $x$ . There follows a discussion of prediction and tolerance intervals for responses at a given setting of  $x$ . Next is an exposition of ANOVA ideas in the present situation. The section then closes with an illustration of how statistical software expedites the calculations introduced in the section.

## 7.2.1 The Simple Linear Regression Model, Corresponding Variance Estimate, and Standardized Residuals

Part 6 introduced the one-way (equal variances, normal distributions) model as the most common probability basis of inference methods for multisample studies. It was represented in symbols as

$$7.2.1.1 \quad y_{ij} = \mu_i + \epsilon_{ij}$$

where the means  $\mu_1, \mu_2, \dots, \mu_r$  were treated as  $r$  unrestricted parameters. Turning now to the matter of inference based on data pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  exhibiting an approximately linear scatterplot, one once again proceeds by imposing a restriction on the one-way model (7.2.1.1). In words, the model assumptions will be that there are underlying normal distributions for the response  $y$  with a common variance  $\sigma^2$  but means  $\mu_{y|x}$  that change linearly in  $x$ . In symbols, it is typical to write that for  $i = 1, 2, \dots, n$ ,

### The (normal) simple linear regression model 7.2.1.2

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the  $\epsilon_i$  are (unobservable) iid normal  $(0, \sigma^2)$  random variables, the  $x_i$  are known constants, and  $\beta_0$ ,  $\beta_1$ , and  $\sigma^2$  are unknown model parameters (fixed constants). Model (7.2.1.2) is commonly known as the (normal) simple linear regression model.

If one thinks of the different values of  $x$  in an  $(x, y)$  data set as separating it into various samples of  $y$ 's, expression (7.2.1.2) is the specialization of model (7.2.1.1) where the (previously unrestricted) means of  $y$  satisfy the linear relationship  $\mu_{y|x} = \beta_0 + \beta_1 x$ . Figure 7.2.1.1 is a pictorial representation of the "constant variance, normal, linear (in  $x$ ) mean" model.



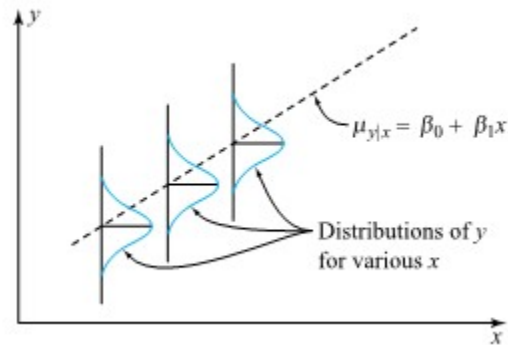


Figure 7.2.1.1 Graphical representation of the simple linear regression model

Inferences about quantities involving those  $x$  values represented in the data (like the mean response at a single  $x$  or the difference between mean responses at two different values of  $x$ ) will typically be sharper when methods based on model (7.2.1.2) can be used in place of the general methods of Part 6 an dANOVA. And to the extent that model (7.2.1.2) describes system behavior for values of  $x$  not included in the data, a model like (7.2.1.2) provides for inferences involving limited interpolation and extrapolation on  $x$ .

Module 7.1 contains an extensive discussion of the use of least squares in the fitting of the approximately linear relation

7.2.1.3 
$$y \approx \beta_0 + \beta_1 x$$

to a set of  $(x, y)$  data. Now we can observe that Module 7.1 can be thought of as an exposition of fitting and the use of residuals in model checking for the simple linear regression model (7.2.1.2). In particular, associated with the simple linear regression model are the estimates of  $\beta_1$  and  $\beta_0$  which we will show again here:

**Slope of the least squares line,  $b_1$**   
7.2.1.3

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and

**Intercept of the least squares line,  $b_0$**   
7.2.1.4

$$b_0 = \bar{y} - b_1 \bar{x}$$

and the corresponding fitted values

**Fitted values for simple linear regression**  
7.2.1.5

$$\hat{y}_i = b_0 + b_1 x_i$$

and residuals

## Residuals for simple linear regression 7.2.1.6

$$e_i = y_i - \hat{y}_i$$

Further, the residuals (or errors) (7.2.1.6) can be used to make up an estimate of  $\sigma^2$ . As always, a sum of squared residuals is divided by an appropriate number of degrees of freedom. That is, there is the following definition of a simple linear regression or line-fitting sample variance, which we will call the mean squared error of the line-fitting ( $MSE_{LF}$ ).

## MEAN SQUARED ERROR OF THE LINE-FITTING SIMPLE LINEAR REGRESSION MODEL

**DEFINITION** Mean squared error of the line-fitting simple linear regression model ( $MSE_{LF}$ )

**EXPRESSION 7.2.1.7**

$$MSE_{LF} = s_{LF}^2 = \frac{1}{n-2} \sum (y - \hat{y})^2 = \frac{1}{n-2} \sum e^2$$

will be called the mean squared error of the line-fitting ( $MSE_{LF}$ ). This is the line-fitting (by simple linear regression) of the sample error variance ( $s_{LF}^2$ ).

Associated with it are  $\nu = n - 2$  degrees of freedom and the standard error of the line-fitting model ( $\text{sqrt}MSE_{LF}$ , an estimated standard deviation of the response variable ( $\sqrt{s_{LF}^2}$ ).

**DEFINITION** Standard error of the line-fitting simple linear regression model ( $\sqrt{MSE_{LF}}$ )

**EXPRESSION 7.2.1.8**

$$\sqrt{MSE_{LF}} = \sqrt{s_{LF}^2} = s_{LF}$$

$s_{LF}$  estimates the level of basic background variation,  $\sigma^2$ , whenever the model (7.2.1.2, the simple linear regression model) is an adequate description of the system under study.

When it is not,  $s_{LF}$  will tend to overestimate  $\sigma$ . So comparing  $s_{LF}$  to  $s_p$  (the pooled sample standard

deviation) is another way of investigating the appropriateness of model 7.2.1.2. A  $s_{LF}$  much larger than  $s_P$  suggests the linear regression model is a poor one.

Example 7.2.1.1 Inference in the Ceramic Powder Pressing Study (continued from 7.1)

The main example in this section will be the pressure/density study of Benson, Locher, and Watkins (used extensively in Module 7.1 to illustrate the descriptive analysis of  $(x, y)$  data). Table 7.2.1.1 lists again those  $n = 15$  data pairs  $(x, y)$  (first presented in Table 7.1.1.1) representing

$x$  = the pressure setting used (psi)  
 $y$  = the density obtained (g/cc)

in the dry pressing of a ceramic compound into cylinders, and Figure 7.2.1.1 is a scatterplot of the data.

Recall further from the calculation of  $R^2$  that the data of Table 7.2.1.1 produce fitted values in Table 7.1.1.2 and then

$$\sum (y - \hat{y})^2 = .005153$$

So for the pressure/density data, one has (via formula (7.2.1.7)) that

$$s_{LF}^2 = \frac{1}{15 - 2} (.005153) = .000396 (\text{g/cc})^2$$

so

$$s_{LF} = \sqrt{.000396} = .0199 \text{ g/cc}$$

If one accepts the appropriateness of model (7.2.1.2) in this powder pressing example, for any fixed pressure the standard deviation of densities associated with many cylinders made at that pressure would be approximately .02 g/cc.

The original data in this example can be thought of as organized into  $r = 5$  separate samples of size  $m = 3$ , one for each of the pressures 2,000 psi, 4,000 psi, 6,000 psi, 8,000 psi, and 10,000 psi. It is instructive to consider what this thinking leads to for an alternative estimate of  $\sigma$ —namely,  $s_P$ . Table 7.2.1.2 gives  $\hat{y}$  and  $s$  values for the five samples.

The sample standard deviations in Table 7.2.1.2 can be employed in the usual way to calculate  $s_P$ . That is (from the expression from Part 5),

$$\begin{aligned} s_P^2 &= \frac{(3 - 1)(.0070)^2 + (3 - 1)(.0110)^2 + \cdots + (3 - 1)(.0114)^2}{(3 - 1) + (3 - 1) + \cdots + (3 - 1)} \\ &= .000424 (\text{g/cc})^2 \end{aligned}$$

from which

$$s_P = \sqrt{s_P^2} = .0206 \text{ g/cc}$$

Comparing  $s_{LF}$  and  $s_P$ , there is no indication of poor fit carried by these values.

$x_i$ Pressure (psi)	$y_i$ Density (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

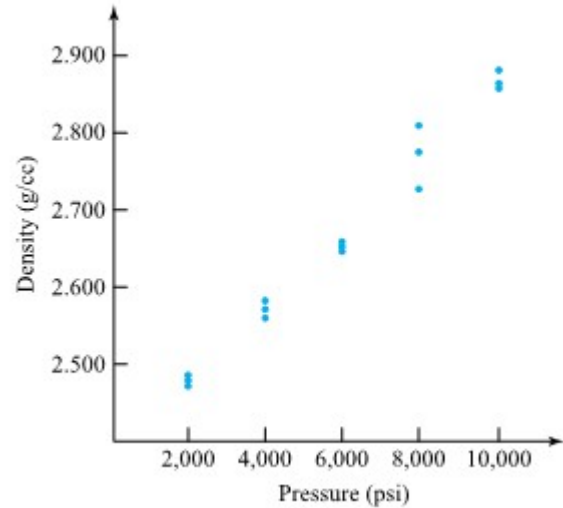


Figure 7.2.1.2: Scatterplot of density vs. pressing pressure

Table 7.2.1.1: Pressing Pressures and Resultant Specimen Densities

$x_i$ Pressure (psi)	$\bar{y}_i$ Sample Mean	$s_i$ Sample Standard Deviation
2,000	2.479	.0070
4,000	2.569	.0110
6,000	2.652	.0056
8,000	2.769	.0423
10,000	2.866	.0114

Table 7.2.1.2 Sample Means and Standard Deviations of Densities for Five Different Pressing Pressures.

Module 7.1 includes some plotting of the residuals (Expression 7.2.1.6) for the pressure/density data (in particular, a normal plot that appears as Figure 7.1.3.4). Although the (raw) residuals (7.2.1.6) are most easily calculated, most commercially available regression programs provide standardized residuals as well as, or even in preference to, the raw residuals.

## STANDARDIZED RESIDUALS

In curve- and surface-fitting analyses, the variances of the residuals depend on the corresponding  $x$ 's. Standardizing before plotting is a way to prevent mistaking a pattern on a residual plot that is explainable on the basis of these different variances for one that is indicative of problems with the basic model. Under model (7.2.1.2), for a given  $x$  with corresponding response  $y$ ,

$$7.2.1.7 \quad \text{Var}(y - \hat{y}) = \sigma^2 \left( 1 - \frac{1}{n} - \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right)$$

So using formula (7.2.1.7) and standardization discussions, corresponding to the data pair  $(x_i, y_i)$  is the standardized residual for simple linear regression

### Standardized residuals for simple linear regression 7.2.1.8

$$e_i^* = \frac{e_i}{s_{LF} \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2}}}$$

The more sophisticated method of examining residuals under model (7.2.1.2) is thus to make plots of the values (7.2.1.8) instead of plotting the raw residuals (7.2.1.6).

Example 7.2.1.2 continued.

Consider how the standardized residuals for the pressure/density data set are related to the raw residuals. Recalling that

$$\sum (x - \bar{x})^2 = 120,000,000$$

and that the  $x_i$  values in the original data included only the pressures 2,000 psi, 4,000 psi, 6,000 psi, 8,000 psi, and 10,000 psi, it is easy to obtain the necessary values of the radical in the denominator of expression (7.2.1.8). These are collected in Table 7.2.1.3.

$x$	$\sqrt{1 - \frac{1}{15} - \frac{(x - 6,000)^2}{120,000,000}}$
2,000	.894
4,000	.949
6,000	.966
8,000	.949
10,000	.894

Table 7.2.1.3 Calculations for Standardized Residuals in the Pressure/Density Study

The entries in Table 7.2.1.3 show, for example, that one should expect residuals corresponding to  $x = 6,000$  psi to be (on average) about  $.966/.894 = 1.08$  times as large as residuals corresponding to  $x = 10,000$  psi. Division of raw residuals by  $s_{FL}$  times the appropriate entry of the second column of Table 7.2.1.3 then puts them all on equal footing, so to speak. Table 7.2.1.4 shows both the raw residuals (taken from Module 7.1) and their standardized counterparts.

$x$	$e$			Standardized Residual
2,000	.0137	.0067	-.0003	.77, .38, -.02
4,000	-.0117	.0003	.0103	-.62, .02, .55
6,000	-.0210	-.0100	-.0140	-1.09, -.52, -.73
8,000	-.0403	.0097	.0437	-2.13, .51, 2.31
10,000	-.0007	.0173	-.0037	-.04, .97, -.21

Table 7.2.1.4 Residuals and Standardized Residuals for the Pressure/Density Study

In the present case, since the values .894, .949, and .966 are roughly comparable, standardization via formula (9.12) doesn't materially affect conclusions about model adequacy. For example, Figures 7.2.1.3 and 7.2.1.4 are normal plots of (respectively) raw residuals and standardized residuals. For all intents and purposes, they are identical. So any conclusions (like those made in Module 7.1) about model adequacy supported by Figure 7.2.1.3 are equally supported by Figure 7.2.1.4, and vice versa.

In other situations, however (especially those where a data set contains a few very extreme  $x$  values), standardization can involve more widely varying denominators for formula (7.2.1.8) than those implied by Table 7.2.1.3 and thereby affect the results of a residual analysis.

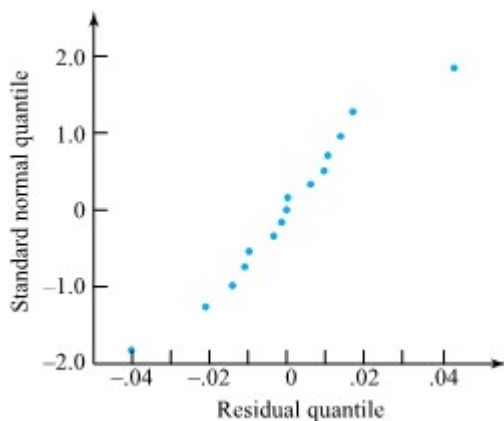


Figure 7.2.1.3 Normal plot of residuals from a linear fit to the pressure/density data.

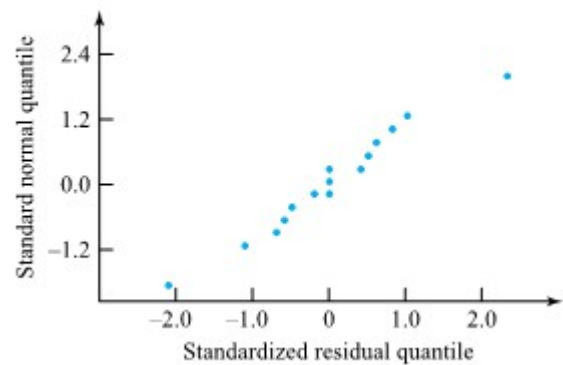


Figure 7.2.1.4 Normal plot of standardized residuals for a linear fit to the pressure/density data

## 7.2.2 Inference for the Slope Parameter

Especially in applications of the simple linear regression model (7.2.1.1) where  $x$  represents a variable that can be physically manipulated by the engineer, the slope parameter  $\beta_1$  is of fundamental interest. It is the rate of change of average response with respect to  $x$ , and it governs the impact of a change in  $x$  on the system output. Inference for  $\beta_1$  is fairly simple, because of the distributional properties that  $b_1$  (the slope of the least squares line) inherits from the model. That is, under model (7.2.1.1),  $b_1$  has a normal distribution with

$$Eb_1 = \beta_1$$

and

$$7.2.2.1 \quad \text{Var } b_1 = \frac{\sigma^2}{\sum(x - \bar{x})^2}$$

which in turn imply that

$$Z = \frac{b_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum(x - \bar{x})^2}}}$$

is standard normal. In a manner similar to many of the arguments in Parts 5 and 6, this motivates the fact that the quantity

$$7.2.2.2 \quad T = \frac{b_1 - \beta_1}{\frac{s_{\text{LF}}}{\sqrt{\sum(x - \bar{x})^2}}}$$

has a  $t_{t-2}$  distribution. The standard arguments of Part 5 applied to expression 7.2.2.2 then show that

$$7.2.2.3 \quad H_0 : \beta_1 = \#$$

can be tested using the test statistic

### 7.2.2.4 Test statistic for $H_0 : \beta_1 = \#$

$$T = \frac{b_1 - \#}{\frac{s_{LF}}{\sqrt{\sum(x-\bar{x})^2}}}$$

and a  $t_{n-2}$  reference distribution. More importantly, under the simple linear regression model (7.2.1.2), a two-sided confidence interval for  $\beta_1$  can be made using endpoints

### 7.2.2.5 Confidence limits for the slope $\beta_1$

$$b_1 \pm t \frac{s_{LF}}{\sqrt{\sum(x - \bar{x})^2}}$$

where the associated confidence is the probability assigned to the interval between  $-t$  and  $t$  by the  $t_{n-2}$  distribution. A one-sided interval is made in the usual way, based on one endpoint from formula (7.2.2.5).

#### Example 7.2.2.1 Powder Pressing Study continued.

In the context of the powder pressing study, Module 7.1 showed that the slope of the least squares line through the pressure/density data is

$$b_1 = .000048\bar{6} \text{ (g/cc)/psi}$$

Then, for example, a 95% two-sided confidence interval for  $\beta_1$  can be made using the .975 quantile of the  $t_1 3$  distribution in formula (7.2.2.5). That is, one can use endpoints

$$.000048\bar{6} \pm 2.160 \frac{.0199}{\sqrt{120,000,000}}$$

that is,

$$.000048\bar{6} \pm .0000039$$

or

$$.0000448 \text{ (g/cc)/psi and } .0000526 \text{ (g/cc)/psi}$$

A confidence interval like this one for  $\beta_1$  can be translated into a confidence interval for a difference in mean responses for two different values of  $x$ . According to model (7.2.1.2), two different values of  $x$  differing by  $\Delta x$  have mean responses differing by  $\beta_1 \Delta x$ . One then simply multiplies endpoints of a confidence interval for  $\beta_1$  by  $\Delta x$  to obtain a confidence interval for the difference in mean responses. For example, since  $8,000 - 6,000 = 2,000$ , the difference between mean densities at 8,000 psi and 6,000 psi levels has a 95% confidence interval with endpoints



2,000(.0000448)g/cc and 2,000(.0000526)g/cc

that is

.0896 g/cc and .1052 g/cc

## CONSIDERATIONS IN THE SELECTION OF X VALUES

Formula (7.2.2.5) allows a kind of precision to be attached to the slope of the least squares line. It is useful to consider how that precision is related to study characteristics that are potentially under an investigator's control. Notice that both formulas (7.2.2.1) and (7.2.2.5) indicate that the larger  $\sum (x - \bar{x})^2$  is (i.e., the more spread out the  $x_i$  values are), the more precision  $b_1$  offers as an estimator of the underlying slope  $\beta_1$ . Thus, as far as the estimation of  $\beta_1$  is concerned, in studies where  $x$  represents the value of a system variable under the control of an experimenter, they should choose settings of  $x$  with the largest possible sample variance. (In fact, if one has  $n$  observations to spend and can choose values of  $x$  anywhere in some interval  $[a, b]$ , taking  $\frac{n}{2}$  of them at  $x = a$  and  $\frac{n}{2}$  at  $x = b$  produces the best possible precision for estimating the slope  $\beta_1$ .)

However, this advice (to spread the  $x_i$ 's out) must be taken with a grain of salt. The approximately linear relationship (7.2.1.2) may hold over only a limited range of possible  $x$  values. Choosing experimental values of  $x$  beyond the limits where it is reasonable to expect formula (7.2.1.2) to hold, hoping thereby to obtain a good estimate of slope, is of course nonsensical. And it is also important to recognize that precise estimation of  $\beta_1$  under the assumptions of model (7.2.1.2) is not the only consideration when planning data collection. It is usually also important to be in a position to tell when the linear form of (7.2.1.2) is inappropriate. That dictates that data be collected at a number of different settings of  $x$ , not simply at the smallest and largest values possible.

## 7.2.3 Inference for the Mean System Response for a Particular Value of $x$

Chapter 6 considered the problem of estimating the mean of  $y$  with levels of the factor (or factors) of interest. In the present context, the analog is the problem of estimating the mean response for a fixed value of the system variable  $x$ ,

**7.2.3.1**

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

The natural data-based approximation of the mean in formula (7.2.3.1) is the corresponding  $y$  value taken from the least squares line. The notation

**7.2.3.2 Estimator of  $\mu_{y|x} = \beta_0 + \beta_1 x$**

$$\hat{y} = b_0 + b_1 x$$

will be used for this value on the least squares lines. (This is in spite of the fact that the value in formula (7.2.3.2) may not be a fitted value in the sense that the phrase has most often been used to this point.  $x$  need not be equal to any of  $x_1, x_2, \dots, x_n$  for both expressions (7.2.3.1) and (7.2.3.2) to make sense.) The simple linear regression model (7.2.1.2) leads to simple distributional properties for  $\hat{y}$  that then produce inference methods for  $\mu_{y|x}$ .

Under model (7.2.1.2),  $\hat{y}$  has a normal distribution with

$$E\hat{y} = \mu_{y|x} = \beta_0 + \beta_1 x$$

and

**7.2.3.3**

$$\text{Var } \hat{y} = \sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2} \right)$$

(In expression (7.2.3.3), notation is being abused somewhat. The  $i$  subscripts and indices of summation in  $\sum (x - \bar{x})^2$  have been suppressed. This summation runs over the  $n$  values  $x_i$  included in the original data set. On the other hand, in the  $(x - \bar{x})^2$  term appearing as a numerator in expression (7.2.3.3), the  $x$  involved is not necessarily equal to any of  $x_1, x_2, \dots, x_n$ . Rather, it is simply the value of the system variable at which the mean response is to be estimated.) Then

$$Z = \frac{\hat{y} - \mu_{y|x}}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}}$$

has a standard normal distribution. This in turn motivates the fact that

**7.2.3.4**

$$T = \frac{\hat{y} - \mu_{y|x}}{s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}}$$

has a  $t_{n-2}$  distribution. The standard arguments of Part 5 applied to expression 7.2.3.4 then show that

**7.2.3.5**

$$H_0 : \mu_{y|x} = \#$$

can be tested using the test statistic

### 7.2.3.6 Test statistic for

$$H_0 : \mu_{y|x} = \#$$

$$T = \frac{\hat{y} - \#}{s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}}$$

and a  $t_{n-2}$  reference distribution. Further, under the simple linear regression model (7.2.1.2), a two-sided individual confidence interval for  $\mu_{y|x}$  can be made using endpoints

### 7.2.3.7 Confidence limits for the mean response, $\mu_{y|x} = \beta_0 + \beta_1 x$

$$\hat{y} \pm t s_{\text{LF}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

where the associated confidence is the probability assigned to the interval between  $-t$  and  $t$  by the  $t_{(n-2)}$  distribution. A one-sided interval is made in the usual way based on one endpoint from formula (7.2.3.7).

#### Example 7.2.3.1. continued

Returning again to the pressure/density study, consider making individual 95% confidence intervals for the mean densities of cylinders produced first at 4,000 psi and then at 5,000 psi.

Treating first the 4,000 psi condition, the corresponding estimate of mean density is

$$\hat{y} = 2.375 + .000048\bar{6}(4,000) = 2.5697 \text{ g/cc}$$

Further, from formula (7.2.3.7) and the fact that the .975 quantile of the  $t_{13}$  distribution is 2.160, a precision of plus-or-minus

$$2.160(.0199) \sqrt{\frac{1}{15} + \frac{(4,000 - 6,000)^2}{120,000,000}} = .0136 \text{ g/cc}$$

can be attached to the 2.5697 g/cc figure. That is, endpoints of a two-sided 95% confidence interval for the mean density under the 4,000 psi condition are

$$2.5561 \text{ g/cc and } 2.5833 \text{ g/cc}$$

Under the  $x = 5,000$  psi condition, the corresponding estimate of mean density is

$$\hat{y} = 2.375 + .000048\bar{6}(5,000) = 2.6183 \text{ g/cc}$$

Using formula (7.2.3.7), a precision of plus-or-minus

$$2.160(.0199) \sqrt{\frac{1}{15} + \frac{(5,000 - 6,000)^2}{120,000,000}} = .0118 \text{ g/cc}$$

can be attached to the 2.6183 g/cc figure. That is, endpoints of a two-sided 95% confidence interval for the mean density under the 5,000 psi condition are

$$2.6065 \text{ g/cc and } 2.6301 \text{ g/cc}$$

The reader should compare the plus-or-minus parts of the two confidence intervals found here. The interval for  $x = 5,000$  psi is shorter and therefore more informative than the interval for  $x = 4,000$  psi. The origin of this discrepancy should be clear, at least upon scrutiny of formula (7.2.3.7). For the researchers' data,

$\bar{x} = 6,000$  psi.  $x = 5,000$  psi is closer to  $\bar{x}$  than is  $x = 4,000$  psi, so the  $(x - \bar{x})^2$  term (and thus the interval length) is smaller for  $x = 5,000$  psi than for  $x = 4,000$  psi.

The phenomenon noted in the preceding example—that the length of a confidence interval for  $\mu_{y|x}$  increases as one moves away from  $\bar{x}$ —is an important one. And it has an intuitively plausible implication for the planning of experiments where an approximately linear relationship between  $y$  and  $x$  is expected, and  $x$  is under the investigators' control. If there is an interval of values of  $x$  over which one wants good precision in estimating mean responses, it is only sensible to center one's data collection efforts in that interval.

### *Inference for the intercept $\beta_0$*

Proper use of displays (7.2.3.5), (7.2.3.6) and (7.2.3.7) give inference methods for the parameter  $\beta_0$  in model (7.2.1.2).  $\beta_0$  is the  $y$  intercept of the linear relationship (7.2.3.1). So by setting  $x = 0$  in displays (9.22), (9.23), and (9.24), tests and confidence intervals for  $\beta_0$  are obtained. However, unless  $x = 0$  is a feasible value for the input variable

and the region where the linear relationship (7.2.3.1) is a sensible description of physical reality includes  $x = 0$ , inference for  $\beta_0$  alone is rarely of practical interest.

## SIMULTANEOUS TWO-SIDED CONFIDENCE LIMITS FOR ALL MEANS, $\mu_{y|x}$

### 7.2.3.8 95% Confidence Interval of the Mean Response

$$(b_0 + b_1 x) \pm \sqrt{2f} s_{LF} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

where for positive  $f$ , the associated simultaneous confidence is the  $F_{2,n-2}$  probability assigned to the interval  $(0, f)$ .

Of course, the practical meaning of the phrase “for all means  $\mu_{y|x}$ ” is more like “for all mean responses in an interval where the simple linear regression model (7.2.1.2) is a workable description of the relationship between  $x$  and  $y$ .” As is always the case in curve- and surface-fitting situations, extrapolation outside of the range of  $x$  values where one has data (and even to some extent interpolation inside that range) is risky business. When it is done, it should be supported by subject-matter expertise to the effect that it is justifiable.

It may be somewhat difficult to grasp the meaning of a simultaneous confidence figure applicable to all possible intervals of the form (7.2.3.8). To this point, the confidence levels considered have been for finite sets of intervals. Probably the best way to understand the theoretically infinite set of intervals given by formula (7.2.3.8) is as defining a region in the  $(x, y)$ -plane thought likely to contain the line  $\mu_{y|x} = \beta_0 + \beta_1 x$ . Figure 7.2.3.1 is a sketch of a typical confidence region represented by formula (7.2.3.8). There is a region indicated about the least squares line whose vertical extent increases with distance from  $\bar{x}$  and which has the stated confidence in covering the line describing the relationship between  $x$  and  $\mu_{y|x}$ .

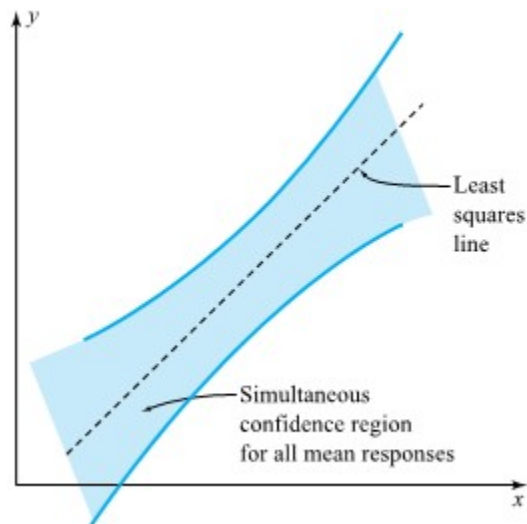


Figure 7.2.3.1 Region in the  $(x, y)$ -plane defined by simultaneous confidence intervals for all values of  $\mu_{y|x}$ .

## Example 3.2.3.2 continued

Using formula (7.2.3.8), find the simultaneous 95% confidence intervals for mean cylinder densities produced under the five conditions actually used by the researchers in their study.

Since  $v_1 = 2$  and  $v_2 = 13$  degrees of freedom are involved in the use of formula (7.2.3.8), simultaneous limits of the form

$$\hat{y} \pm \sqrt{2(3.81)} s_{LF} \sqrt{\frac{1}{15} + \frac{(x - 6,000)^2}{120,000,000}}$$

are indicated.

We can also compare this to the use of P-R method from Part 6 for simultaneous 95% CI calculation.

First, formula (from Module shows that with  $n - r = 15 - 5 = 10$  degrees of freedom for  $s_P$  and  $r = 5$  conditions under study, 95% simultaneous two-sided confidence limits for all five mean densities are of the form

$$\bar{y}_i \pm 3.103 \frac{s_P}{\sqrt{n_i}}$$

which in the example is

$$\bar{y}_i \pm 3.103 \frac{.0206}{\sqrt{3}}$$

that is,

$$\bar{y}_i \pm .0369 \text{ g/cc}$$

Table 3.2.3.1 shows the five intervals that result from the use of the two simultaneous confidence method, together with individual intervals (7.2.3.7).

Two points are evident from Table 3.2.3.1. First, the intervals that result from formula (7.3.3.8) are somewhat wider than the corresponding individual intervals given by formula (7.3.3.7). But it is also clear that the use of the simple linear regression model assumptions in preference to the more general one-way assumptions of Part 6 can lead to shorter simultaneous confidence intervals and correspondingly sharper real-world engineering inferences.

$x$ , Pressure	$\mu_{y x}$ (P-R Method) Mean Density	$\mu_{y x}$ (from formula (9.25)) Mean Density	$\mu_{y x}$ (from formula (9.24)) Mean Density
2,000 psi	2.4790 ± .0369 g/cc	2.4723 ± .0246 g/cc	2.4723 ± .0136 g/cc
4,000 psi	2.5693 ± .0369 g/cc	2.5697 ± .0174 g/cc	2.5697 ± .0118 g/cc
6,000 psi	2.6520 ± .0369 g/cc	2.6670 ± .0142 g/cc	2.6670 ± .0111 g/cc
8,000 psi	2.7687 ± .0369 g/cc	2.7643 ± .0174 g/cc	2.7643 ± .0118 g/cc
10,000 psi	2.8660 ± .0369 g/cc	2.8617 ± .0246 g/cc	2.8617 ± .0136 g/cc

Table 7. 2.3.1 Simultaneous (and Individual) 95% Confidence Intervals for Mean Cylinder Densities



## 7.2.4 Prediction and Tolerance Intervals

Inference for  $\mu_{y|x}$  is one kind of answer to the qualitative question, “If I hold the input variable  $x$  at some particular level, what can I expect in terms of a system response?” It is an answer in terms of mean or long-run average response. Sometimes an answer in terms of individual responses is of more practical use. And in such cases it is helpful to know that the simple linear regression model assumptions (7.2.1.2) lead to their own specialized formulas for prediction and tolerance intervals.

The basic fact that makes possible prediction intervals under assumptions (7.2.1.2) is that if  $y_{n+1}$  is one additional observation, coming from the distribution of responses corresponding to a particular  $x$ , and  $\hat{y}$  is the corresponding fitted value at that  $x$  (based on the original  $n$  data pairs), then

$$T = \frac{y_{n+1} - \hat{y}}{s_{\text{LF}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}}$$

has a  $t_{n-2}$  distribution. This fact leads in the usual way to the conclusion that under model (7.2.1.2) the two-sided interval with endpoints

### 7.2.4.1 Simple Linear Regression prediction limits for an additional $y$ at a given $x$

$$\hat{y} \pm t s_{\text{LF}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

can be used as a prediction interval for an additional observation  $y$  at a particular value of the input variable  $x$ . The associated prediction confidence is the probability  $t_{n-2}$  distribution assigns to the interval between  $-t$  and  $t$ . One-sided intervals are made in the usual way, by employing only one of the endpoints (7.2.4.1) and adjusting the confidence level appropriately.

It is possible not only to derive prediction interval formulas from the simple linear regression model assumptions but also to develop relatively simple formulas for approximate one-sided tolerance bounds. That is, the intervals



### 7.2.4.2 A one-sided tolerance interval for the y distribution at x

$$(\hat{y} - \tau s_{LF}, \infty)$$

and

### 7.2.4.3 Another one-sided tolerance interval for the y distribution at x

$$(-\infty, \hat{y} + \tau s_{LF})$$

can be used as one-sided tolerance intervals for a fraction  $p$  of the underlying distribution of responses corresponding to a particular value of the system variable  $x$ , provided  $\tau$  is appropriately chosen (depending upon the data,  $p$ ,  $x$ , and the desired confidence level).

### 7.2.4.4 The ratio of $\sqrt{\text{Var } \hat{y}}$ to $\sigma$ for simple linear regression

$$A = \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

will be adopted for the multiplier that is used (e.g., in previous formula to go from an estimate of  $\sigma$  to an estimate of the standard deviation of  $\hat{y}$ ). Then, for approximate  $\gamma$  level confidence in locating a fraction  $p$  of the responses  $y$  at the  $x$  of interest,  $\tau$  appropriate for use in interval (7.2.4.2) or (7.2.4.3) is

### 7.2.4.5 Multiplier to use in tolerance bounds

$$\tau = \frac{Q_z(p) + A Q_z(\gamma) \sqrt{1 + \frac{1}{2(n-2)} \left( \frac{Q_z^2(p)}{A^2} - Q_z^2(\gamma) \right)}}{1 - \frac{Q_z^2(\gamma)}{2(n-2)}}$$

#### Example 7.2.4.1 continued

To illustrate the use of prediction and tolerance interval formulas in the simple linear regression context, consider a 90% lower prediction bound for a single additional density in powder pressing, if a pressure of 4,000 psi is employed. Then, additionally consider finding a 95% lower tolerance bound for 90% of many additional cylinder densities if that pressure is used.

Treating first the prediction problem, formula (7.2.4.1) shows that an appropriate prediction bound is

$$2.5697 - 1.350(.0199) \sqrt{1 + \frac{1}{15} + \frac{(4,000 - 6,000)^2}{120,000,000}} = 2.5796 - .0282$$

that is

$$2.5514 \text{ g/cc}$$

If, rather than predicting a single additional density for  $x = 4,000$  psi, it is of interest to locate 90% of additional densities corresponding to a 4,000 psi pressure, a tolerance bound is in order. First use formula (7.2.4.4) and find that

$$A = \sqrt{\frac{1}{15} + \frac{(4,000 - 6,000)^2}{120,000,000}} = .3162$$

Next, for 95% confidence, applying formula (7.4.4.5),

$$\tau = \frac{1.282 + (.3162)(1.645) \sqrt{1 + \frac{1}{2(15-2)} \left( \frac{(1.282)^2}{(.3162)^2} - (1.645)^2 \right)}}{1 - \frac{(1.645)^2}{2(15-2)}} = 2.149$$

So finally, an approximately 95% lower tolerance bound for 90% of densities produced using a pressure of 4,000 psi is (via formula (7.2.4.2))

$$2.5697 - 2.149(.0199) = 2.5697 - .0428$$

that is

$$2.5269 \text{ g/cc}$$

## CAUTIONS ABOUT PREDICTION AND TOLERANCE INTERVALS IN REGRESSION

The fact that curve-fitting facilitates interpolation and extrapolation makes it imperative that care be taken in the interpretation of prediction and tolerance intervals. All of the warnings regarding the interpretation of prediction and tolerance intervals raised in Part 5 apply equally to the present situation. But the new element here (that formally, the intervals can be made for values of  $x$  where one has absolutely no data) requires additional caution. If one is to use formulas (7.2.4.1), (7.2.4.2), and (7.2.4.3) at a value of  $x$  not represented among  $x_1, x_2, \dots, x_n$ , it must be plausible that model (7.2.1.2) not only describes system behavior at those  $x$  values where one has data, but at the additional value of  $x$  as well. And even when this is "plausible" the application of formulas (7.2.4.1), (7.2.4.2), and (7.2.4.3) to new values of  $x$  should be treated with a good dose of care. Should one's (unverified) judgment prove wrong, the nominal confidence level has unknown practical relevance.



## 7.2.5 Simple Linear Regression and ANOVA

Part 6 illustrates how, for unstructured studies, partition of the total sum of squares into interpretable pieces provides both (1) intuition and quantification regarding the origin of observed variation and also (2) the basis for an F test of “no differences between mean responses.” It turns out that something similar is possible in simple linear regression contexts.

In the unstructured context of Part 6, it was useful to name the difference between SSTot (Sum of Squares Total) and SSE (Sum of Squares Error). The corresponding convention for curve- and surface-fitting situations is stated next in definition form.

### DEFINITION REGRESSION SUM OF SQUARES (SSR)

#### EXPRESSION 7.2.5.1

In curve- and surface-fitting analyses of multisample studies, the difference

$$SSR = SSTot - SSE$$

will be called the regression sum of squares (SSReg or SSR).

It is not obvious, but the difference referred to in Definition (7.2.5.1) in general has the form of a sum of squares of appropriate quantities. In the present context of fitting a line by least squares,

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Without using the particular terminology of Definition (7.2.5.1), this text has already made fairly extensive use of  $SSR = SSTot - SSE$ . A review of Definition (7.1.2.2) (the coefficient of determination  $R^2$ ) in Part 7.1 and Definitions in Part 6 will show that in curve- and surface-fitting contexts,

### 7.2.5.1 The coefficient of determination for simple linear regression in sum of squares notation

$$R^2 = \frac{SSR}{SSTot}$$

That is, SSR is the numerator of the coefficient of determination defined first in Definition (7.1.2.2) (Part 7.1). It is commonly thought of as the part of the raw variability in  $y$  that is accounted for in the curve- or surface-fitting process.

SSR and SSE not only provide an appealing partition of SSTot but also form the raw material for an F test of

7.2.5.2  $H_0 : \beta_1 = 0$

versus

7.2.5.3  $H_a : \beta_1 \neq 0$

Under model (7.2.1.2), hypothesis (7.2.5.2) can be tested using the statistic

### 7.2.5.4 An F statistic for testing

$$H_0 : \beta_1 = 0$$

$$F = \frac{SSR/1}{s_{LF}^2} = \frac{SSR/1}{SSE/(n-2)}$$

and an  $F_{1,n-2}$  reference distribution, where large observed values of the test statistic constitute evidence against  $H_0$ .

Earlier in this section, the general null hypothesis  $H_0 : \beta_1 = \#$  was tested using the t statistic. It is thus reasonable to consider the relationship of the F test indicated in displays (7.2.5.2), (7.2.5.3), and (7.2.5.4) to the earlier t test. The null hypothesis  $H_0 : \beta_1 = 0$  is a special form of the hypothesis,  $H_0 : \beta_1 = \#$ . It is the most frequently tested version of the hypothesis because it can (within limits) be interpreted as the null hypothesis that mean response doesn't depend on  $x$ . This is because when hypothesis (7.2.5.2) is true within the simple linear regression model (7.2.1.2),  $\mu_{y|x} = \beta_0 + 0 \cdot x = \beta_0$ , which doesn't depend on  $x$ . (Actually, a better interpretation of a test of hypothesis (7.2.5.2) is as a test of whether a linear term in  $x$  adds significantly to one's ability to model the response  $y$  after accounting for an overall mean response.)

If one then considers testing hypotheses (7.2.5.2) and (7.2.5.3), it might appear that the  $\# = 0$  version of formulas from Module 7 represent two different testing methods. But they are equivalent. The statistic (7.2.5.4) turns out to be the square of the  $\# = 0$  version of the statistic, and (two-sided) observed significance levels based on the statistic and the  $t_{n-2}$  distribution turn out to be the same as observed significance levels based on statistic (7.2.5.2) and the  $F_{1,n-2}$ . So, from one point of view, the F test specified here is redundant, given the earlier discussion. But it is introduced here because of its relationship to the ANOVA ideas of Part 6, and because it has an important natural generalization to more complex curve- and surface-fitting contexts. (This generalization is discussed in Part 8 and cannot be made equivalent to a t test.)

The partition of  $SSTot$  into its parts,  $SSR$  and  $SSE$ , and the calculation of the statistic (7.2.5.4) can be organized in ANOVA table format. Table 7.2.5.1 shows the general format that this book will use in the simple linear regression context.

ANOVA Table (for testing $H_0 : \beta_1 = 0$ )				
Source	$SS$	$df$	$MS$	$F$
Regression	$SSR$	1	$SSR/1$	$MSR/MSE$
Error	$SSE$	$n - 2$	$SSE/(n - 2)$	
Total	$SSTot$	$n - 1$		

Table 7.2.5.1 General Form of the ANOVA Table for Simple Linear Regression

#### Example 7.2.5.1 continued

Recall again from the discussion of the pressure/density example in Module 7.1.1 that

$$SSTot = \sum (y - \bar{y})^2 = .289366$$

and that

$$SSE = \sum (y - \hat{y})^2 = .005153$$

Thus,

$$SSR = SSTot - SSE = .289366 - .005153 = .284213$$

and the specific version of Table 7.2.5.1 for the present example is given as Table 7.2.5.2.

Then the observed level of significance for testing  $[latex]H_0 : \beta_1 = 0[/latex]$  is

$$P [ \text{an } F_{1,13} \text{ random variable} > 717 ] < .001$$

and one has very strong evidence against the possibility that  $\beta_1 = 0$ . A linear term in Pressure is an important contributor to one's ability to describe the behavior of Cylinder Density. This is, of course, completely consistent with the earlier interval-oriented analysis that produced 95% confidence limits for  $\beta_1$  of

$$.0000448(\text{ g/cc})/\text{psi} \text{ and } .0000526(\text{ g/cc})/\text{psi}$$

that do not bracket 0.

The value of  $R^2 = .9822$  (found first in Module 7) can also be easily derived, using the entries of Table 7.2.5.2 and the relationship (7.2.5.1).

<b>ANOVA Table (for testing <math>H_0 : \beta_1 = 0</math>)</b>				
Source	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Regression	<i>SSR</i>	1	<i>SSR/1</i>	<i>MSR/MSE</i>
Error	<i>SSE</i>	$n - 2$	$SSE/(n - 2)$	
Total	<i>SSTot</i>	$n - 1$		

Table 7.2.5.2 ANOVA Table for the Pressure/Density Data

## 7.2.6 Statistical Computing for Simple Linear Regression: Pressure and Density Example

Many of the calculations needed for the methods of this section are made easier by statistical software packages. None of the methods of this section are so computationally intensive that they absolutely require the use of such software, but it is worthwhile to consider its use in the simple linear regression context. Learning where on a typical printout to find the various summary statistics corresponding to calculations made in this section helps in locating important summary statistics for the more complicated curve- and surface-fitting analyses of the next Part.

Printout 7.2.6.1 is from a Python JupyterLab Notebook analysis of the pressure/density data for the Pressure/Density Data Example. This Notebook is located on our GitHub site at: [Intro Statistal Methods for Engineering GitHub Site](#) and is located under Part 7A.

It is also available to look at and access for download at he [Special GitHub Site for Part 7](#).

Or you can open an interactive computing environment to work thourgh the Jupyter Notebook using Python thourgh a Binder Site using the Special GitHub Site for the Part 2 example. Click [here](#) to go to the Binder Site (located at <https://mybinder.org/v2/gh/Statistical-Methods-for-Engineering/Special-GitHub-Site-Part-2-Example-Percent-Waste-by-Weight-on-Bulk-Paper-Rolls/HEAD>).

This is typical of summaries of regression analyses printed by available statistical packages. The most basic piece of information on the printout is, of course, the fitted equation. Then we show a summary output of a table giving the estimated coefficients ( $b_0$  and  $b_1$ ), their estimated standard deviations, and the t ratios (appropriate for testing whether coefficients  $\beta$  are 0). The printout includes the values of Scale =  $MSE_{LF}$  =  $s_{LF}^2$  and  $R^2$ . We also show an ANOVA table printout. For the several observed values of test statistics printed out in these printouts, observed levels of significance are shown. The ANOVA table is followed by a table of values of y, fitted y, standard deviation of the fitted y, and residual, and standardized residual corresponding to the n data points. Statsmodels in Python's regression program has an option that allows one to request fitted values, confidence intervals for  $\mu_{y|x}$ , and prediction intervals for x values of interest. This overview of the printouts finishes with this information for the value  $x = 5,000$ .

The reader is encouraged to compare the information on this Printout 7.2.6.1 with the various results obtained in Example from this Part 7 of the course and verify that these pieces of the output are familiar. We will continue to learn about the remaining pieces in Part 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_low    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 Tutorial 6 & 7 - Simple Linear Regression

At this point, it is recommended that you work your way through the [Tutorial 6 exercise](#) and the [Tutorial 7 exercise](#) found on the associated GitHub repository. Tutorial 6 will teach you how to interpret the various outputs that you receive when computing an OLS model in Python as well as how to compute it by-hand. Tutorial 7 will teach you how to compute an OLS model using Python syntax.

**It is strongly recommended that you consult the [Simple Linear Regression Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the files on “Ordinary Least Squares Regression” and “Goodness of Fit” will be particularly useful.

## 8.0.1 Introduction to Multiple and Logistic Regression

The principles of simple linear regression lay the foundation for more sophisticated regression methods used in a wide range of challenging settings. In this section, we explore multiple regression, which introduces the possibility of more than one predictor,. The basic ideas introduced in Part 7 on Simple Linear Regression generalize to produce a powerful engineering tool: multiple linear regression, which is introduced in this section.

Multiple regression extends simple two-variable regression to the case that still has one response but many predictors (denoted  $x_1, x_2, x_3, \dots$ ). The method is motivated by scenarios where many variables may be simultaneously connected to an output.

## 8.0.2 Attributions

This first draft of Part 7 is mostly a direct adoption of the text of of [“Basic Engineering Data Collection and Analysis”](#) by [Stephen B. Vardeman & J. Marcus Jobe](#) which is licensed under [CC BY-NC-SA 4.0](#).

Changes include rewriting some of the passages and adding some minor original material. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. Python based Jupyter Notebooks have been adapted from the text examples and linked throughout.

This resource also draws on Kevin Dunns “Process Improvement Using Data” at [PID](#). Portions of this work are the copyright of Kevin Dunn, and shared through [CC BY-SA 4.0](#).

Material for Chapters 8.2.1.1 and 8.2.2.2 come from Quantitative Research Methods for Political Science, Public Policy and Public Administration: 4th Edition With Applications in R, by *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, and Josie Davis*, located at <https://bookdown.org/ripberjt/qrmbook/>. This work is shared through the licensed under a [Creative Commons Attribution 4.0 International License](#) (CC BY 4.0).

## *8.1.0 Introduction to Multiple Linear Regression: Fitting Curves and Surfaces by Least Squares*

This Part 8.1 first covers fitting curves defined by polynomials and other functions that are linear in their parameters to  $(x, y)$  data. Next comes the fitting of surfaces to data where a response  $y$  depends upon the values of several variables  $x_1, x_2, \dots, x_k$ . In both cases, the discussion will stress how useful  $R^2$  and residual plotting are and will consider the question of choosing between possible fitted equations. Lastly, we include some additional practical cautions.

## 8.1.1 Curve Fitting by Least Squares

In Part 7.1, a straight line did a reasonable job of describing the pressure/density data. But in the fly ash study, the ammonium phosphate/compressive strength data were very poorly described by a straight line. This section first investigates the possibility of fitting curves more complicated than a straight line to  $(x, y)$  data. As an example, an attempt will be made to find a better equation for describing the fly ash data.

A natural generalization of the linear equation

$$8.1.1.1 \quad y \approx \beta_0 + \beta_1 x$$

is the **polynomial equation**

$$8.1.1.2 \quad y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

The least squares fitting of equation (8.1.1.2) to a set of  $n$  pairs  $(x_i, y_i)$  is conceptually only slightly more difficult than the task of fitting equation (8.1.1.1). The function of  $k + 1$  variables

$$\begin{aligned} y &\approx \frac{1}{2}g \left( t_0 + \frac{1}{60}(x - 1) \right)^2 \\ &= \frac{g}{2} \left( \frac{x}{60} \right)^2 + g \left( t_0 - \frac{1}{60} \right) \left( \frac{x}{60} \right) + \frac{g}{2} \left( t_0 - \frac{1}{60} \right)^2 \\ &= \frac{g}{7200} x^2 + \frac{g}{60} \left( t_0 - \frac{1}{60} \right) x + \frac{g}{2} \left( t_0 - \frac{1}{60} \right)^2 \end{aligned}$$

must be minimized. Upon setting the partial derivatives of  $S(\beta_0, \beta_1, \dots, \beta_k)$  equal to 0, the set of **normal equations** is obtained for this least squares problem, generalizing the pair of equations from Part 7.1. There are  $k + 1$  linear equations in the  $k + 1$  unknowns  $\beta_0, \beta_1, \dots, \beta_k$ . And typically, they can be solved simultaneously for a single set of values,  $b_0, b_1, \dots, b_k$ , minimizing  $S(\beta_0, \beta_1, \dots, \beta_k)$ .

**Example 8.1.1.1 More on the Fly Ash Data**

Return to the fly ash study of B. Roth and the Table 7.1.3.1. A quadratic equation might fit the data better than the linear one. So consider fitting the  $k = 2$  version of equation (8.1.1.2)

$$8.1.1.3 \quad y \approx \beta_0 + \beta_1 x + \beta_2 x^2$$

to the data of Table 7.1.3.1. Printouts 8.1.1.1 and 8.1.1.2 show the Python Jupyter Notebook Output for this regression model. (After entering  $x$  and  $y$  values from Table 8.1.1.2 into two columns of the dataframe, an additional column was created by squaring the  $x$  values, creating the  $x\_sqr$  variable). This Python based Jupyter Notebook is available through the course [GitHub Site](#).

This Notebook can also be viewed through an interactive [Binder Site](#) for the Special GitHub Site for the Fly\_Ash Data Example.

The regression equation is  
 $y = 1.243e+03 + 382.7 x + -76.66 x\_sqr$

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	
No. 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.	Std.Err.	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			Condition No.:	38	
=====						

Printout 8.1.1.1 Quadratic Fit to the Fly Ash Data



	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

Printout 8.1.1.2 ANOVA table for Quadratic Fit to Fly Ash Data.

The fitted quadratic equation is

$$\hat{y} = 1242.9 + 382.7x - 76.7x^2$$

Figure 8.1.1.1 shows the fitted curve sketched on a scatterplot of the  $(x, y)$  data. Although the quadratic curve is not an altogether satisfactory summary of Roth's data, it does a much better job of following the trend of the data than the line sketched previously.

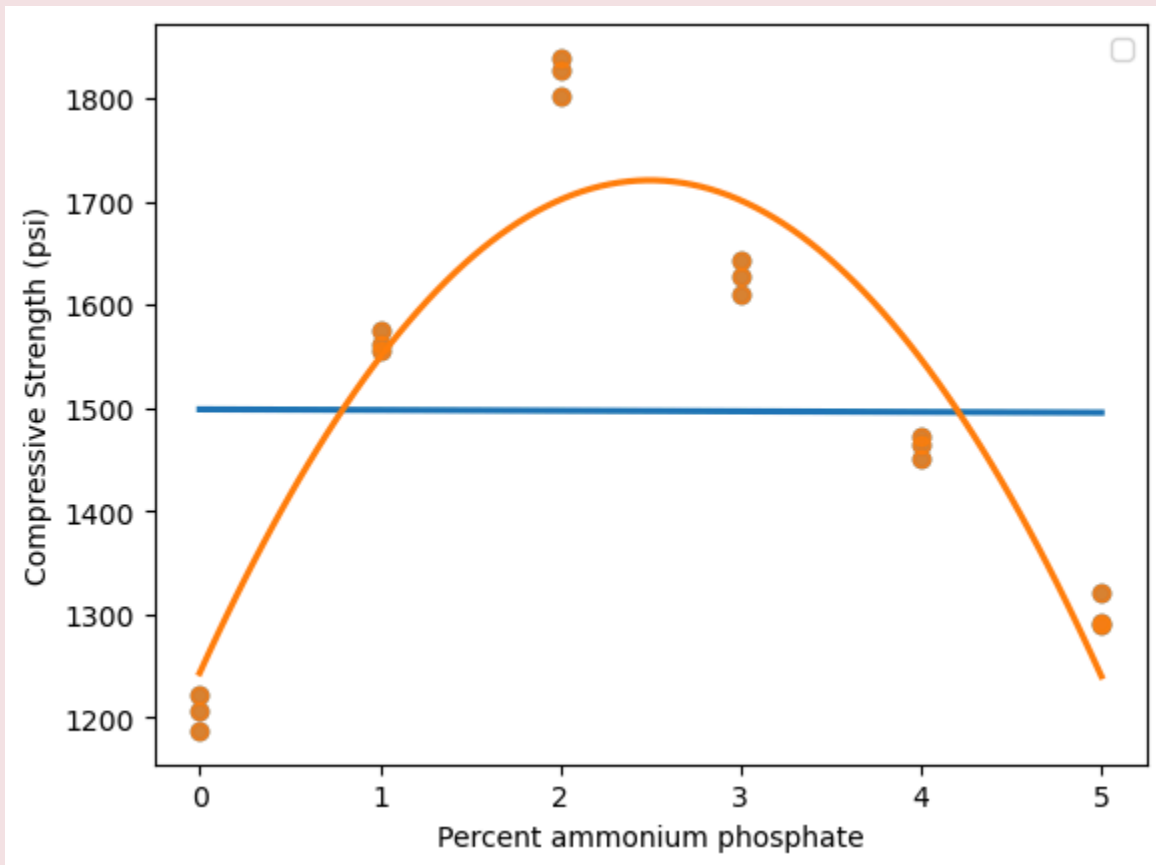


Figure 8.1.1.1 Scatterplot and fitted simple linear fit (as the blue line) and the fitted quadratic for the fly ash example data.

The previous Part showed that when fitting a line to  $(x, y)$  data, it is helpful to quantify the goodness of that fit using  $R^2$ . The coefficient of determination can also be used when fitting a polynomial of form (8.1.1.2). Recall once more from Definition 3 that

**8.1.1.3 DEFINITION and Expression for the Coefficient of Determination**

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

is the fraction of the raw variability in  $y$  accounted for by the fitted equation. Calculation by hand from formula (8.1.1.3) is possible, but of course the easiest way to obtain  $R^2$  is to use a statistical computing.

**Example 8.1.1.2 continued.**

Consulting the Printouts above, it can be seen that the equation  $\hat{y} = 1242.9 + 382.7x - 76.7x^2$  produces  $R^2 = .867$ . So 86.7% of the raw variability in compressive strength is accounted for using the fitted quadratic. The sample correlation between the observed strengths  $y_i$  and fitted strengths  $\hat{y}_i$  is  $+\sqrt{.867} = .93$ .

Comparing what has been done in the present section to what was done in Part 7.1, it is interesting that for the fitting of a line to the fly ash data,  $R^2$  obtained there was only .000 (to three decimal places). The present quadratic is a remarkable improvement over a linear equation for summarizing these data.

A natural question to raise is "What about a cubic version of equation (8.1.1.2)?" Printout 8.1.1.3 and 8.1.1.4 shows some results of a run made to investigate this possibility, and Figure 8.1.1.2 shows a scatterplot of the data and a plot of the fitted cubic equation.  $x$  values were squared and cubed to provide  $x$ ,  $x^2$ , and  $x^3$  for each  $y$  value to use in the fitting.

```

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
No. 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.   Std.Err.   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             Condition No.:      324
=====

```

*Printout 8.1.1.3 Cubic fit to fly ash data.*

	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

*Printout 8.1.1.4 ANOVA table for the cubic fit of the fly ash data.*

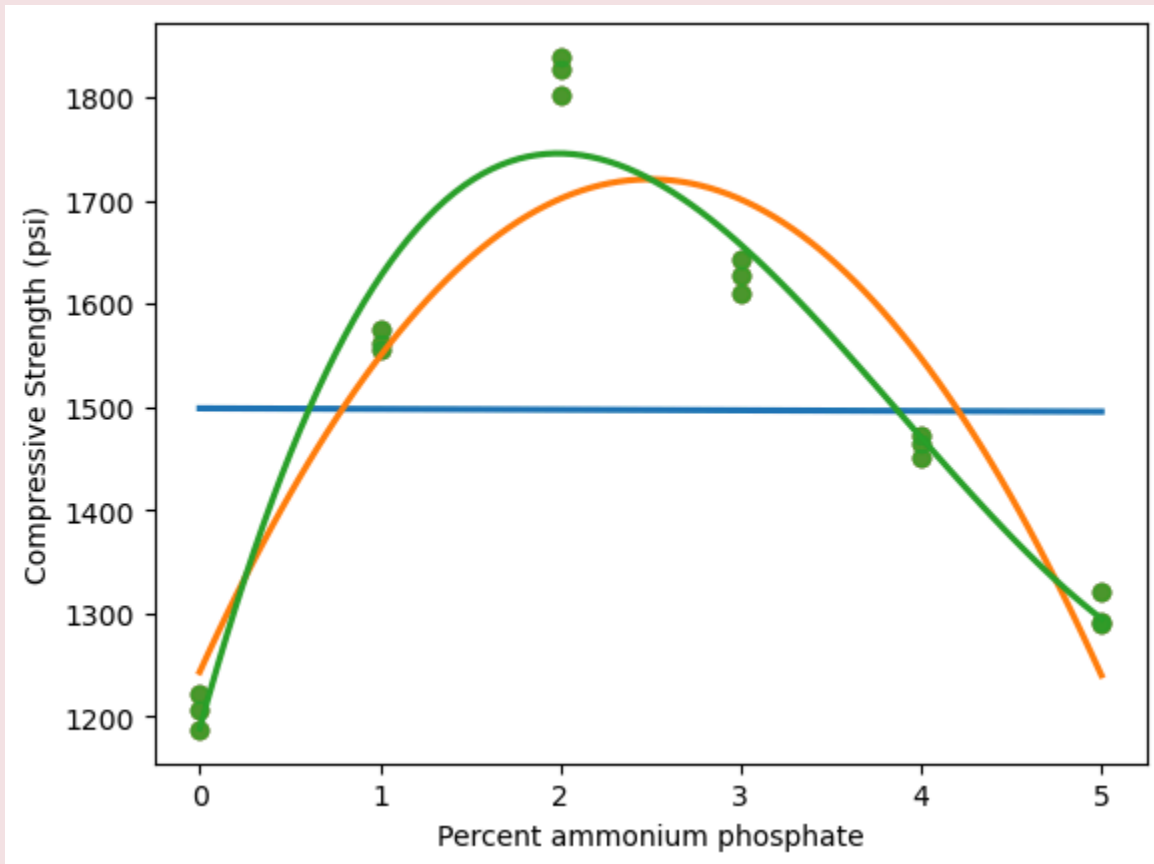


Figure 8.1.1.2 Scatterplot and fitted cubic for the fly ash data (least squares cubic shown in green).

$R^2$  for the cubic equation is .952, somewhat larger than for the quadratic. But it is fairly clear from Figure 8.1.1.2 that even a cubic polynomial is not totally satisfactory as a summary of these data. In particular, both the fitted quadratic in Figure 8.1.1.2 and the fitted cubic in Figure 8.1.1.1 fail to fit the data adequately near an ammonium phosphate level of 2%. Unfortunately, this is where compressive strength is greatest—precisely the area of greatest practical interest.

The example illustrates that  $R^2$  is not the only consideration when it comes to judging the appropriateness of a fitted polynomial. The examination of plots is also important. Not only scatterplots of  $y$  versus  $x$  with superimposed fitted curves but plots of residuals can be helpful. This can be illustrated on a data set where  $y$  is expected to be nearly perfectly quadratic in  $x$ .

#### Example 8.1.1.3 Analysis of the Bob Drop Data

Consider again the experimental determination of the acceleration due to gravity (through the dropping of the steel bob) data given in Part 1 and reproduced here in the first two columns of Table 8.1.1.1. Recall that the positions  $y$  were recorded at  $\frac{1}{60}$  sec intervals beginning at some unknown time  $t_0$  (less than  $\frac{1}{60}$  sec) after the bob was released. Since Newtonian mechanics predicts the bob displacement to be

$$\text{displacement} = \frac{gt^2}{2}$$

one expects

#### 8.1.1.4

$$\begin{aligned} y &\approx \frac{1}{2}g\left(t_0 + \frac{1}{60}(x-1)\right)^2 \\ &= \frac{g}{2}\left(\frac{x}{60}\right)^2 + g\left(t_0 - \frac{1}{60}\right)\left(\frac{x}{60}\right) + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \\ &= \frac{g}{7200}x^2 + \frac{g}{60}\left(t_0 - \frac{1}{60}\right)x + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \end{aligned}$$

That is,  $y$  is expected to be approximately quadratic in  $x$  and, indeed, the plot of  $(x, y)$  points in Figure for Part 1 appears to have that character.

As a slight digression, note that this expression shows that if a quadratic is fitted to the data in Table 8.1.1.1 via least squares,

#### 8.1.1.5

$$\hat{y} = b_0 + b_1x + b_2x^2$$

is obtained and an experimentally determined value of  $g$  (in  $\text{mm}/\text{sec}^2$ ) will be  $7200b_2$ . This is in fact how the value  $9.79 \text{ m}/\text{sec}^2$ , quoted in Section 1.4, was obtained.

$$\hat{y} = .0645 - .4716x + 1.3597x^2$$

(from which  $g \approx 9790 \text{ mm}/\text{sec}^2$ ) with  $R^2$  that is 1.0 to 6 decimal places. Residuals for this fit can be calculated using Definition 8.1.1.3 and are also given in Table 8.1.1.1. Figure 8.1.1.3 is a normal plot of the residuals. It is reasonably linear and thus not remarkable (except for some small suggestion that the largest residual or two may not be as extreme as might be expected, a circumstance that suggests no obvious physical explanation).

Data, Fitted Values, and Residuals for a Quadratic Fit to the Bob Displacement

$x$ , Point Number	$y$ , Displacement	$\hat{y}$ , Fitted Displacement	$e$ , Residual
1	.8	.95	-.15
2	4.8	4.56	.24
3	10.8	10.89	-.09
4	20.1	19.93	.17
5	31.9	31.70	.20
6	45.9	46.19	-.29
7	63.3	63.39	-.09
8	83.1	83.31	-.21
9	105.8	105.96	-.16
10	131.3	131.32	-.02
11	159.5	159.40	.10
12	190.5	190.21	.29
13	223.8	223.73	.07
14	260.0	259.97	.03
15	299.2	298.93	.27
16	340.5	340.61	-.11
17	385.0	385.01	-.01
18	432.2	432.13	.07
19	481.8	481.97	-.17
20	534.2	534.53	-.33
21	589.8	589.80	.00
22	647.7	647.80	-.10
23	708.8	708.52	.28

Table 8.1.1.1 Data, Fitted Values, and Residuals for a Quadratic Fit to the Bob Displacement.

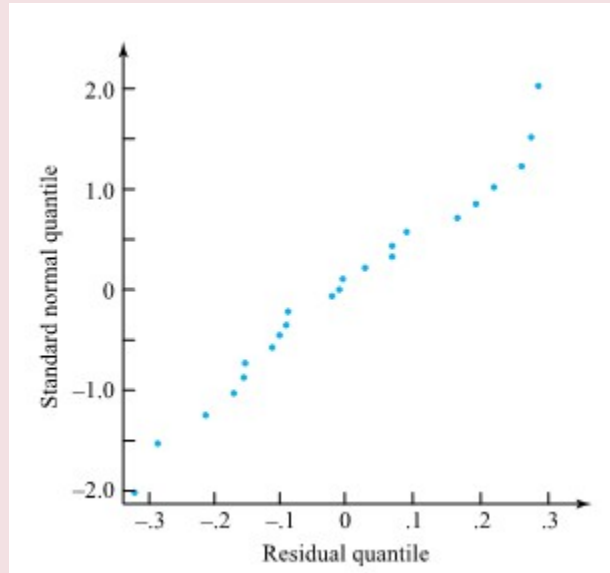


Figure 8.1.1.3 Normal plot of the residuals from a quadratic fit to the bob drop data

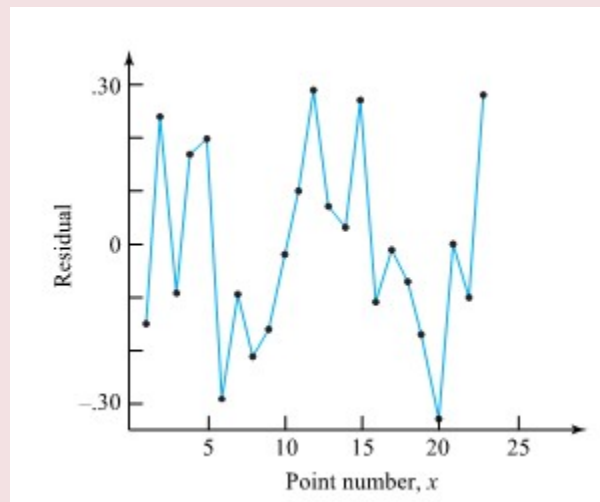


Figure 8.1.1.4 Plot of the residuals from the bob drop quadratic fit vs.  $x$

However, a plot of residuals versus  $x$  (the time variable) is interesting. Figure 8.1.1.4 is such a plot, where successive plotted points have been connected with line segments. There is at least a hint in Figure 8.1.1.4 of a cyclical pattern in the residuals. Observed displacements are alternately too big, too small, too big, etc. It would be a good idea to look at several more tapes, to see if a cyclical pattern appears consistently, before seriously thinking about its origin. But should the pattern suggested by Figure 8.1.1.4 reappear consistently, it would indicate that something in the mechanism generating the 60 cycle current may cause cycles to be alternately slightly shorter then slightly longer than  $\frac{1}{60}$  sec. The practical implication of this would be that if a better determination of  $g$  were desired, the regularity of the AC current waveform is one matter to be addressed.

## WHAT IF A POLYNOMIAL DOESN'T FIT $(x, y)$ DATA?

---

Examples 8.1.1.2 and 8.1.1.3 (respectively) illustrate only partial success and then great success in describing an  $(x, y)$  data set by means of a polynomial equation. Situations like Example 8.1.1.3 obviously do sometimes occur, and it is reasonable to wonder what to do when they happen. There are two simple things to keep in mind.

For one, although a polynomial may be unsatisfactory as a global description of a relationship between  $x$  and  $y$ , it may be quite adequate locally-i.e., for a relatively restricted range of  $x$  values. For example, in the fly ash study, the quadratic representation of compressive strength as a function of percent ammonium phosphate is not appropriate over the range 0 to 5%. But having identified the region around 2% as being of practical interest, it would make good sense to conduct a follow-up study concentrating on (say) 1.5 to 2.5% ammonium phosphate. It is quite possible that a quadratic fit only to data with  $1.5 \leq x \leq 2.5$  would be both adequate and helpful as a summarization of the follow-up data.

The second observation is that the terms  $x, x^2, x^3, \dots, x^k$  in equation (8.1.1.2) can be replaced by any (known) functions of  $x$  and what we have said here will remain essentially unchanged. This can lead us to considering transforming a term to find a better fit.



## 8.1.2 Transformations

### TRANSFORMATIONS FOR LINE FITTING

---

The second observation discussed in the previous Chapter 8.1.1 for when a model does not seem to fit is that the terms  $x, x^2, x^3, \dots, x^k$  in equation (8.1.12) can be replaced by any (known) functions of  $x$  and what we have said here will remain essentially unchanged. The normal equations will still be  $k + 1$  linear equations in  $\beta_0, \beta_1, \dots, \beta_k$ , and a multiple linear regression program will still produce least squares values  $b_0, b_1, \dots, b_k$ . This can be quite useful when there are theoretical reasons to expect a particular (nonlinear but) simple functional relationship between  $x$  and  $y$ . For example, Taylor's equation for tool life is of the form

$$y \approx \alpha x^\beta$$

for  $y$  tool life (e.g., in minutes) and  $x$  the cutting speed used (e.g., in sfpm). Taking logarithms,

$$\ln(y) \approx \ln(\alpha) + \beta \ln(x)$$

This is an equation for  $\ln(y)$  that is linear in the parameters  $\ln(\alpha)$  and  $\beta$  involving the variable  $\ln(x)$ . So, presented with a set of  $(x, y)$  data, empirical values for  $\alpha$  and  $\beta$  could be determined by

1. taking logs of both  $x$ 's and  $y$ 's,
2. fitting the linear version of (4.12), and
3. identifying  $\ln(\alpha)$  with  $\beta_0$  (and thus  $\alpha$  with  $\exp(\beta_0)$ ) and  $\beta$  with  $\beta_1$ .

### TRANSFORMATIONS OF VARIABLES IN MODELING

---

This course is an introduction to one of the main themes of engineering statistical analysis: the discovery and use of simple structure in complicated situations. Sometimes this can be done by reexpressing variables on some other (nonlinear) scales of measurement besides the ones that first come to mind. That is, sometimes simple structure may not be obvious on initial scales of measurement, but may emerge after some or all variables have been transformed. This section presents several examples where transformations are helpful. In the process, some comments about commonly used types of transformations, and more specific reasons for using them, are offered.

### Transformations and Single Samples

In disucced in Part 3 and Part 4, there are a number of standard theoretical distributions. When one of these standard models can be used to describe a response  $y$ , all that is known about the model can be brought to bear in making predictions and inferences regarding  $y$ . However, when no standard distributional shape can be found to describe  $y$ , it may nevertheless be possible to so describe  $g(y)$  for some function  $g(\cdot)$ .

#### Example 8.1.2.1 Discovery time.

Elliot, Kibby, and Meyer studied operations at an auto repair shop. They collected some data on what they called the “discovery time” associated with diagnosing what repairs the mechanics were going to recommend to the car owners. Thirty such discovery times (in minutes) are given in Figure 8.1.2.1, in the form of a stem-and-leaf plot.

The stem-and-leaf plot shows these data to be somewhat skewed to the right. Many of the most common methods of statistical inference are based on an assumption that a data-generating mechanism will in the long run produce not skewed, but rather symmetrical and bell-shaped data. Therefore, using these methods to draw inferences and make predictions about discovery times at this shop is highly questionable. However, suppose that some transformation could be applied to produce a bell-shaped distribution of transformed discovery times. The standard methods could be used to draw inferences about transformed discovery times, which could then be translated (by undoing the transformation) to inferences about raw discovery times.

One common transformation that has the effect of shortening the right tail of a distribution is the logarithmic transformation,  $g(y) = \ln(y)$ . To illustrate its use in the present context, normal plots of both discovery times and log discovery times are given in Figure 8.1.2.2. These plots indicate that Elliot, Kibby, and Meyer could not have reasonably applied standard methods of inference to the discovery times, but they could have used the methods with log discovery times. The second normal plot is far more linear than the first.



Figure 8.1.2.1 Stem-and-leaf plot of discovery times.

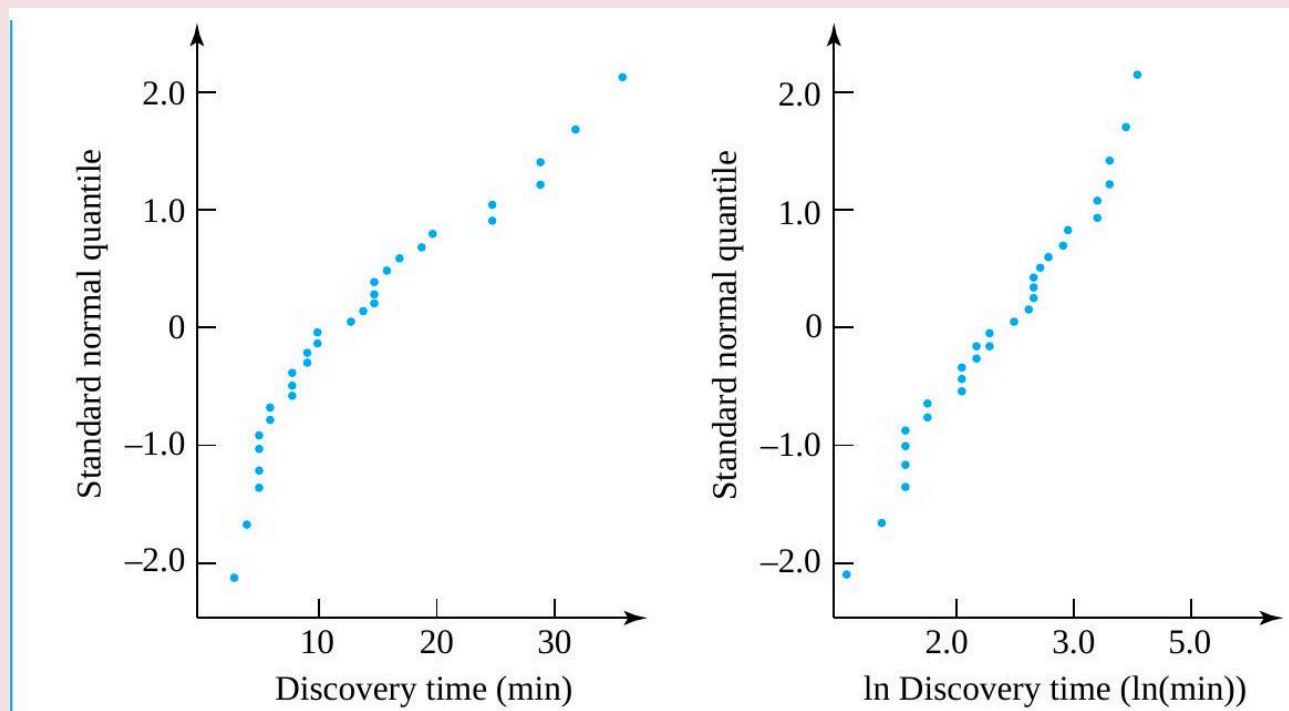


Figure 8.1.2.2 Normal plots for discovery times and log discovery times.

The logarithmic transformation was useful in the preceding example in reducing the skewness of a response distribution. Some other transformations commonly employed to change the shape of a response distribution in statistical engineering studies are the power transformations,

#### 8.1.2.1 Power Transformations

$$g(y) = (y - \gamma)^\alpha$$

In transformation (8.1.2.1), the number  $\gamma$  is often taken as a threshold value, corresponding to a minimum possible response. The number  $\alpha$  governs the basic shape of a plot of  $g(y)$  versus  $y$ . For  $\alpha > 1$ , transformation (8.1.2.1) tends to lengthen the right tail of a distribution for  $y$ . For  $0 < \alpha < 1$ , the transformation tends to shorten the right tail of a distribution for  $y$ , the shortening becoming more drastic as  $\alpha$  approaches 0 but not as pronounced as that caused by the logarithmic transformation

#### 8.1.2.2 Logarithmic Transformation

$$g(y) = \ln(y - \gamma)$$

### *Transformations and Multiple Samples*

---

Comparing several sets of process conditions is one of the fundamental problems of statistical engineering analysis. It is advantageous to do the comparison on a scale where the samples have comparable variabilities, for at least two reasons. The first is the obvious fact that comparisons then reduce simply to comparisons between response means. Second, standard methods of statistical inference often have wellunderstood properties only when response variability is comparable for the different sets of conditions.

When response variability is not comparable under different sets of conditions, a transformation can sometimes be applied to all observations to remedy this. This possibility of **transforming to stabilize variance** exists when response variance is roughly a function of response mean. Some theoretical calculations suggest the following guidelines as a place to begin looking for an appropriate variance-stabilizing transformation:

1. If response standard deviation is approximately proportional to response mean, try a logarithmic transformation.
2. If response standard deviation is approximately proportional to the  $\delta$  power of the response mean, try transformation (4.34) with  $\alpha = 1 - \delta$ .

Where several samples (and corresponding  $\bar{y}$  and  $s$  values) are involved, an empirical way of investigating whether (1) or (2) above might be useful is to plot  $\ln(s)$  versus  $\ln(\bar{y})$  and see if there is approximate linearity. If so, a slope of roughly 1 makes (1) appropriate, while a slope of  $\delta \neq 1$  signals what version of (2) might be helpful.

## 8.1.3 Surface Fitting by Least Squares

It is a small step from the idea of fitting a line or a polynomial curve to realizing that essentially the same methods can be used to summarize the effects of several different quantitative variables  $x_1, x_2, \dots, x_k$  on some response  $y$ . Geometrically the problem is fitting a surface described by an equation

$$8.1.3.1 \quad y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

to the data using the least squares principle. This is pictured for a  $k = 2$  case in Figure 8.1.3.1, where six  $(x_1, x_2, y)$  data points are pictured in three dimensions, along with a possible fitted surface of the form (8.1.3.1). To fit a surface defined by equation (8.1.3.1) to a set of  $n$  data points  $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$  via least squares, the function of  $k + 1$  variables

$$S(\beta_0, \beta_1, \beta_2, \dots, \beta_k) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2$$

must be minimized by choice of the coefficients  $\beta_0, \beta_1, \dots, \beta_k$ . Setting partial derivatives with respect to the  $\beta$ 's equal to 0 gives normal equations generalizing equations for linear regression. The solution of these  $k + 1$  linear equations in the  $k + 1$  unknowns  $\beta_0, \beta_1, \dots, \beta_k$  is the first task of a multiple linear regression program. The fitted coefficients  $b_0, b_1, \dots, b_k$  that it produces minimize  $S(\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ .

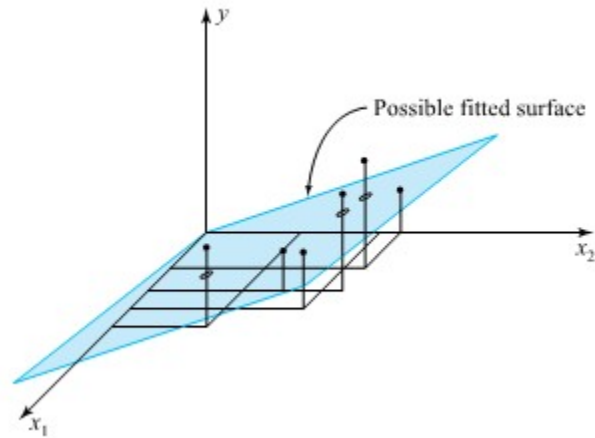


Figure 8.1.3.1 Six data points  $(x_1, x_2, y)$  and a possible fitted plane.

#### Example 8.1.3.1 Surface Fitting and Brownlee's Stack Loss Data

Table 8.1.3.1 contains part of a set of data on the operation of a plant for the oxidation of ammonia to nitric acid that appeared first in Brownlee's *Statistical Theory and Methodology in Science and Engineering*. In plant operation, the nitric oxides produced are absorbed in a countercurrent absorption tower.

The air flow variable,  $x_1$ , represents the rate of operation of the plant. The acid concentration variable,  $x_3$ , is the percent circulating minus 50 times 10. The response variable,  $y$ , is ten times the percentage of ingoing ammonia that escapes from the absorption column unabsorbed (i.e., an inverse measure of overall plant efficiency). For purposes of understanding, predicting, and possibly ultimately optimizing plant performance, it would be useful to have an equation describing how  $y$  depends on  $x_1$ ,  $x_2$ , and  $x_3$ . Surface fitting via least squares is a method of developing such an empirical equation.

Printout 8.1.3.1 shows results from a Python Jupyter Notebook run made to obtain a fitted equation of the form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Brownlee's Stack Loss Data				
$i,$ Observation Number	$x_{1i},$ Air Flow	$x_{2i},$ Cooling Water Inlet Temperature	$x_{3i},$ Acid Concentration	$y_i,$ Stack Loss
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

Tab1 8.1.3.1 Brownlee's Stack Loss Data.

The equation produced by the program is

$$8.1.3.2 \quad \hat{y} = -37.65 + .80x_1 + .58x_2 - .07x_3$$

with  $R^2 = .975$ . The coefficients in this equation can be thought of as rates of change of stack loss with respect to the individual variables  $x_1$ ,  $x_2$ , and  $x_3$ , holding the others fixed. For example,  $b_1 = .80$  can be interpreted as the increase in stack loss  $y$  that accompanies a one-unit increase in air flow  $x_1$  if inlet temperature  $x_2$  and acid concentration  $x_3$  are held fixed. The signs on the coefficients indicate whether  $y$  tends to increase or decrease with increases in the corresponding  $x$ . For example, the fact that  $b_1$  is positive indicates that the higher the rate at which the plant is run, the larger  $y$  tends to be (i.e., the less efficiently the plant operates). The large value of  $R^2$  is a preliminary indicator that the equation (8.1.3.2) is an effective summarization of the data.

The regression equation is

stack = -37.65 + 0.80 air + 0.58 water + -0.07 acid.

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	
No. 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.	Std.Err.	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
water	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			Condition No.:	1644	
=====						

Printout 8.1.3.1 Multiple Regression for the Stack Loss Data.

	df	sum_sq	mean_sq	F	PR(>F)
air	1.0	775.482188	775.482188	494.160440	9.969916e-12
water	1.0	18.492672	18.492672	11.784083	4.452801e-03
acid	1.0	1.859634	1.859634	1.185015	2.961071e-01
Residual	13.0	20.400800	1.569292	NaN	NaN

Printout 8.1.3.2 ANOVA table for multiple regression stack loss data.

### The goal of multiple regression

Although the mechanics of fitting equations of the form (8.1.3.1) to multivariate data are relatively straightforward, the choice and interpretation of appropriate equations are not so clear-cut. Where many  $x$  variables are involved, the number of potential equations of form (8.1.3.1) is huge. To make matters worse, there is no completely satisfactory way to plot multivariate  $(x_1, x_2, \dots, x_k, y)$  data to “see” how an equation is fitting. About all that we can do at this point is to (1) offer the broad advice that what is wanted is the simplest equation that adequately fits the data and then (2) provide examples of how  $R^2$  and residual plotting can be helpful tools in clearing up the difficulties that arise.



In the context of the nitrogen plant, it is sensible to ask whether all three variables,  $x_1$ ,  $x_2$ , and  $x_3$ , are required to adequately account for the observed variation in  $y$ . For example, the behavior of stack loss might be adequately explained using only one or two of the three  $x$  variables. There would be several consequences of practical engineering importance if this were so. For one, in such a case, a simple or **parsimonious** version of equation (8.1.3.1) could be used in describing the oxidation process. And if a variable is not needed to predict  $y$ , then it is possible that the expense of measuring it might be saved. Or, if a variable doesn't seem to have much impact on  $y$  (because it doesn't seem to be essential to include it when writing an equation for  $y$ ), it may be possible to choose its level on purely economic grounds, without fear of degrading process performance.

As a means of investigating whether indeed some subset of  $x_1$ ,  $x_2$ , and  $x_3$  is adequate to explain stack loss behavior,  $R^2$  values for equations based on all possible subsets of  $x_1$ ,  $x_2$ , and  $x_3$  were obtained and placed in Table 8.1.3.2. This shows, for example, that 95% of the raw variability in  $y$  can be accounted for using a linear equation in only the air flow variable  $x_1$ . Use of both  $x_1$  and the water temperature variable  $x_2$  can account for 97.3% of the raw variability in stack loss. Inclusion of  $x_3$ , the acid concentration variable, in an equation already involving  $x_1$  and  $x_2$ , increases  $R^2$  only from .973 to .975.

Equation Fit	$R^2$
$y \approx \beta_0 + \beta_1 x_1$	.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

Table 8.1.3.2

If identifying a simple equation for stack loss that seems to fit the data well is the goal, the message in Table 8.1.3.2 would seem to be "Consider an  $x_1$  term first, and then possibly an  $x_2$  term." On the basis of  $R^2$ , including an  $x_3$  term in an equation for  $y$  seems unnecessary. And in retrospect, this is entirely consistent with the character of the fitted equation (8.1.3.1):  $x_3$  varies from 72 to 93 in the original data set, and this means that  $\hat{y}$  changes only a total amount

$$.07(93 - 72) \approx 1.5$$

based on changes in  $x_3$ . (Remember that  $.07 = b_3 =$  the fitted rate of change in  $y$  with respect to  $x_3$ .) 1.5 is relatively small in comparison to the range in the observed  $y$  values.

Once  $R^2$  values have been used to identify potential simplifications of the equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

these can and should go through thorough residual analyses before they are adopted as data summaries. As an example, consider a fitted equation involving  $x_1$  and  $x_2$ . A multiple linear regression program can be used to produce the fitted equation

## 8.1.3.3

$$\hat{y} = -42.00 - .78x_1 + .57x_2$$

(Notice that  $b_0$ ,  $b_1$ , and  $b_2$  in equation (8.1.3.3) differ somewhat from the corresponding values in equation (8.1.3.2). That is, equation (8.1.3.3) was not obtained from equation

Dropping variables from a fitted equation typically changes coefficients

(8.1.3.2) by simply dropping the last term in the equation. In general, the values of the coefficients  $b$  will change depending on which  $x$  variables are and are not included in the fitting.)

Residuals for equation (8.1.3.3) can be computed and plotted in any number of potentially useful ways. Figure 8.1.3.2 shows a normal plot of the residuals and three other plots of the residuals against, respectively,  $x_1$ ,  $x_2$ , and  $\hat{y}$ . There are no really strong messages carried by the plots in Figure 8.1.3.2 except that the data set contains one unusually large  $x_1$  value and one unusually large  $\hat{y}$  (which corresponds to the large  $x_1$ ). But there is enough of a curvilinear “up-then-down-then-back-up-again” pattern in the plot of residuals against  $x_1$  to suggest the possibility of adding an  $x_1^2$  term to the fitted equation (8.1.3.3).

You might want to verify that fitting the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$$

to the data of Table 8.1.3.1 yields approximately

## 8.1.3.4

$$\hat{y} = -15.409 - .069x_1 + .528x_2 + .007x_1^2$$

with corresponding  $R^2 = .980$  and residuals that show even less of a pattern than those for the fitted equation (8.1.3.3). In particular, the hint of curvature on the plot of residuals versus  $x_1$  for equation (8.1.3.3) is not present in the corresponding plot for equation (8.1.3.4). Interestingly, looking back over this example, one sees that fitted equation (8.1.3.4) has a better  $R^2$  value than even fitted equation (8.1.3.2), in spite of the fact that equation (8.1.3.2) involves the process variable

's and also eliminates the slight pattern seen on the plot of residuals for equation (8.1.3.3) versus  $x_1$ , it seems an attractive choice for summarizing the stack loss data. A 3D scatterplot of  $x_1$  and  $x_2$  on the fitted line from equation 8.1.3.4 is shown in Figure 8.1.3.3 A two-dimensional representation of the fitted surface defined by equation (8.1.3.4) is given in Figure 8.1.3.4. The slight curvature on the plotted curves is a result of the  $x_1^2$  term appearing in equation (8.1.3.4). Since most of the data have  $x_1$  from 50 to 62 and  $x_2$  from 17 to 24, the curves carry the message that over these ranges, changes in  $x_1$  seem to produce larger changes in stack loss than do changes in  $x_2$ . This conclusion is consistent with the discussion centered around Table 8.1.3.2.

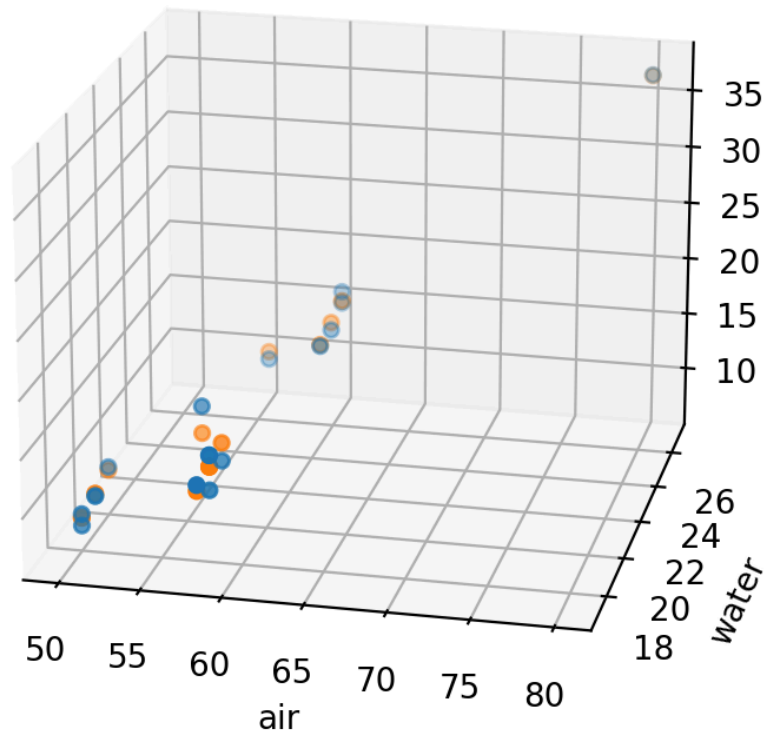


Figure 8.1.3.3 3D scatterplot of fitted values from 8.1.3.4.

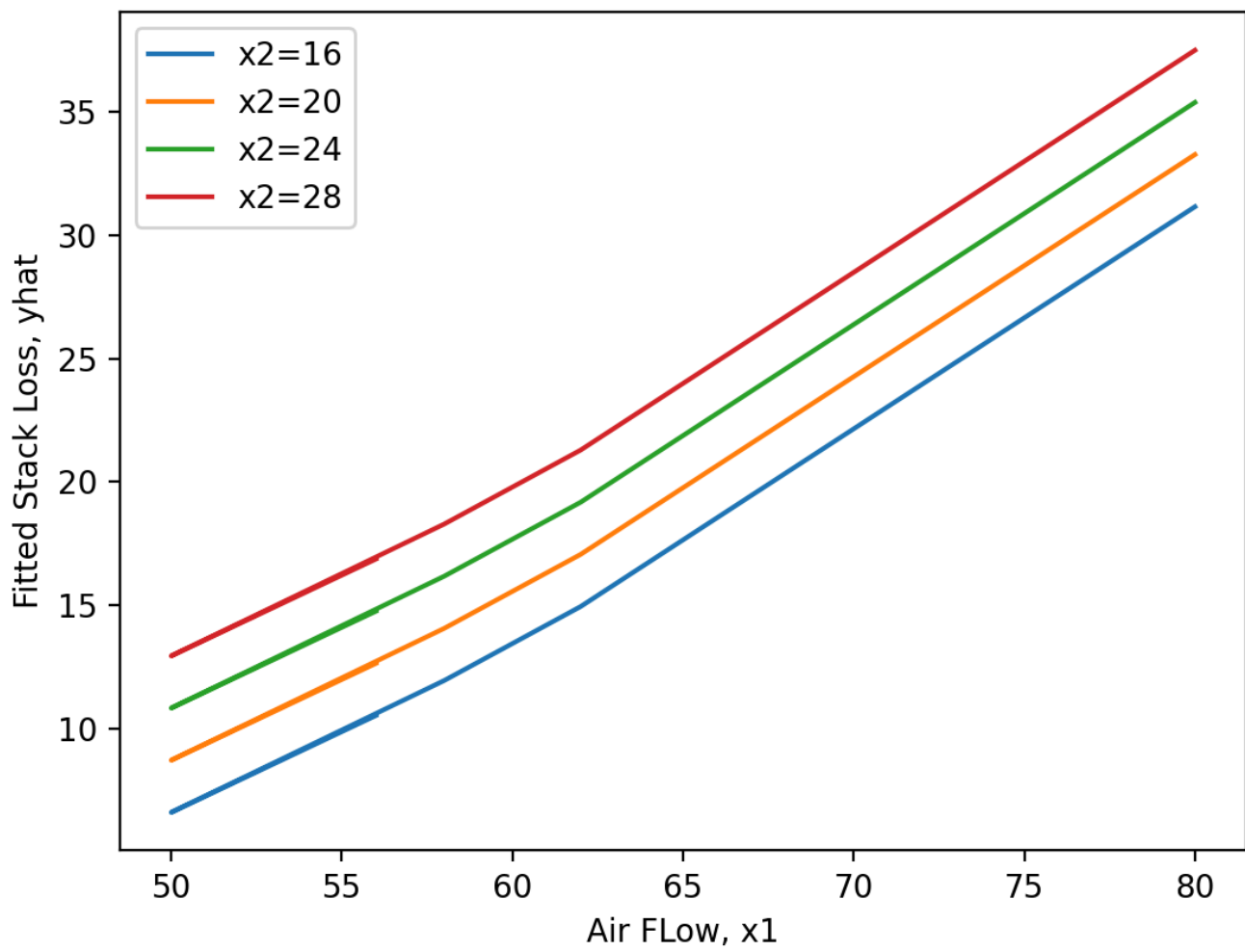


Figure 8.1.3.4 Plots of fitted stack loss from equation 8.3.1.4.

## 8.1.4 Common Residual Plots in Multiple Regression

The plots of residuals used in Example 8.1.3 are typical. They are

1. normal plots of residuals,
2. plots of residuals against all  $x$  variables,
3. plots of residuals against  $\hat{y}$ ,
4. plots of residuals against time order of observation, and
5. plots of residuals against variables (like machine number or operator) not used in the fitted equation but potentially of importance.

All of these can be used to help assess the appropriateness of surfaces fit to multivariate data, and they all have the potential to tell an engineer something not previously discovered about a set of data and the process that generated them.

## 8.1.5 Interactions

Earlier in this section, there was a discussion of the fact that an " $x$  term" in the equations fitted via least squares can be a known function (e.g., a logarithm) of a basic process variable. In fact, it is frequently helpful to allow an " $x$  term" in to be a known function of several basic process variables. The next example illustrates this point.

### Example 8.1.5.1 Lift/Drag Ratio for a Three-Surface Configuration

P. Burriss studied the effects of the positions relative to the wing of a canard (a forward lifting surface) and tail on the lift/drag ratio for a three-surface configuration. Part of his data are given in Table 8.1.5.1, where

$x_1$  = canard placement in inches above the plane defined by the main wing

$x_2$  = tail placement in inches above the plane defined by the main wing

(The front-to-rear positions of the three surfaces were constant throughout the study.)

A straightforward least squares fitting of the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

to these data produces  $R^2$  of only .394. Even the addition of squared terms in both  $x_1$  and  $x_2$ , i.e., the fitting of

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

produces an increase in  $R^2$  to only .513. However, Printout 8.1.5.1 shows that fitting the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

yields  $R^2 = .641$  and the fitted relationship

$$\hat{y} = 3.4284 + .5361x_1 + .3201x_2 - .5042x_1x_2$$

Lift/Drag Ratios for 9 Canard/Tail Position Combinations

$x_1$ , Canard Position	$x_2$ , Tail Position	$y$ , Lift/Drag Ratio
-1.2	-1.2	.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

Table 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.	Std.Err.	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			Condition No.:	1	
=====						

Printout 8.1.5.1 Multiple Regression for Lift/Drag Data

	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

*Printout 8.1.5.2 ANOVA table for multiple regression for the Lift/Drag Ratio Data*

The regression equation is

$$y = 3.43 + 0.536x_1 + 0.320x_2 - 0.504x_1 * x_2$$

After reading  $x_1$ ,  $x_2$ , and  $y$  values from Table 8.1.5.1 into columns,  $x_1 x_2$  products were created and  $y$  fitted to the three predictor variables  $x_1$ ,  $x_2$ , and  $x_1 x_2$  in order to create this printout.)

Figure 8.1.5.1 shows the nature of the fitted surface (8.1.5.1). Raising the canard (increasing  $x_1$ ) has noticeably different predicted impacts on  $y$ , depending on the value of  $x_2$  (the tail position). (It appears that the canard and tail should not be lined up-i.e.,  $x_1$  should not be near  $x_2$ . For large predicted response, one wants small  $x_1$  for large  $x_2$  and large  $x_1$  for small  $x_2$ .) It is the cross-product term  $x_1 x_2$  in relationship (8.1.5.1) that allows the response curves to have different characters for different  $x_2$  values. Without it, the slices of the fitted  $(x_1, x_2, \hat{y})$  surface would be parallel for various  $x_2$ , much like the situation in Module 8.1.4.



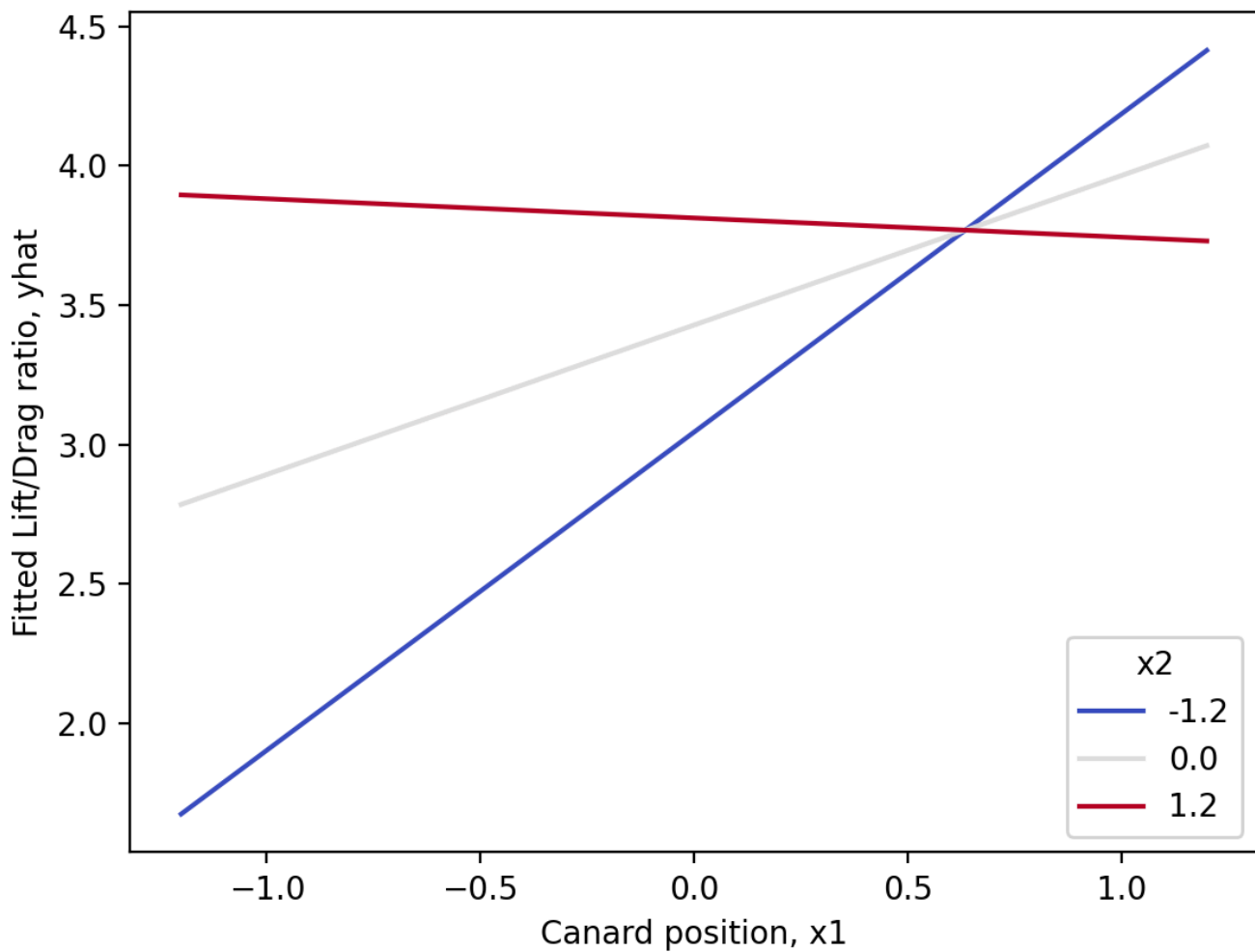


Figure 8.1.5.1 Plots of fitted Lift/Drag from Equation 8.1.5.1

Although the main new point of this example has by now been made, it probably should be mentioned that equation (8.1.5.1) is not the last word for fitting the data of Table 8.1.5.1. Figure 8.1.5.2 gives a plot of the residuals for relationship (8.1.5.1) versus canard position  $x_1$ , and it shows a strong curvilinear pattern. In fact, the fitted equation

$$8.1.5.2 \quad \hat{y} = 3.9833 + .5361x_1 + .3201x_2 - .4843x_1^2 - .5042x_1x_2$$

provides  $R^2 = .754$  and generally random-looking residuals. It can be verified by plotting  $\hat{y}$  versus  $x_1$  curves for several  $x_2$  values that the fitted relationship (8.1.5.2) yields nonparallel parabolic slices of the fitted  $(x_1, x_2, \hat{y})$  surface, instead of the nonparallel linear slices seen in Figure 8.1.5.1.

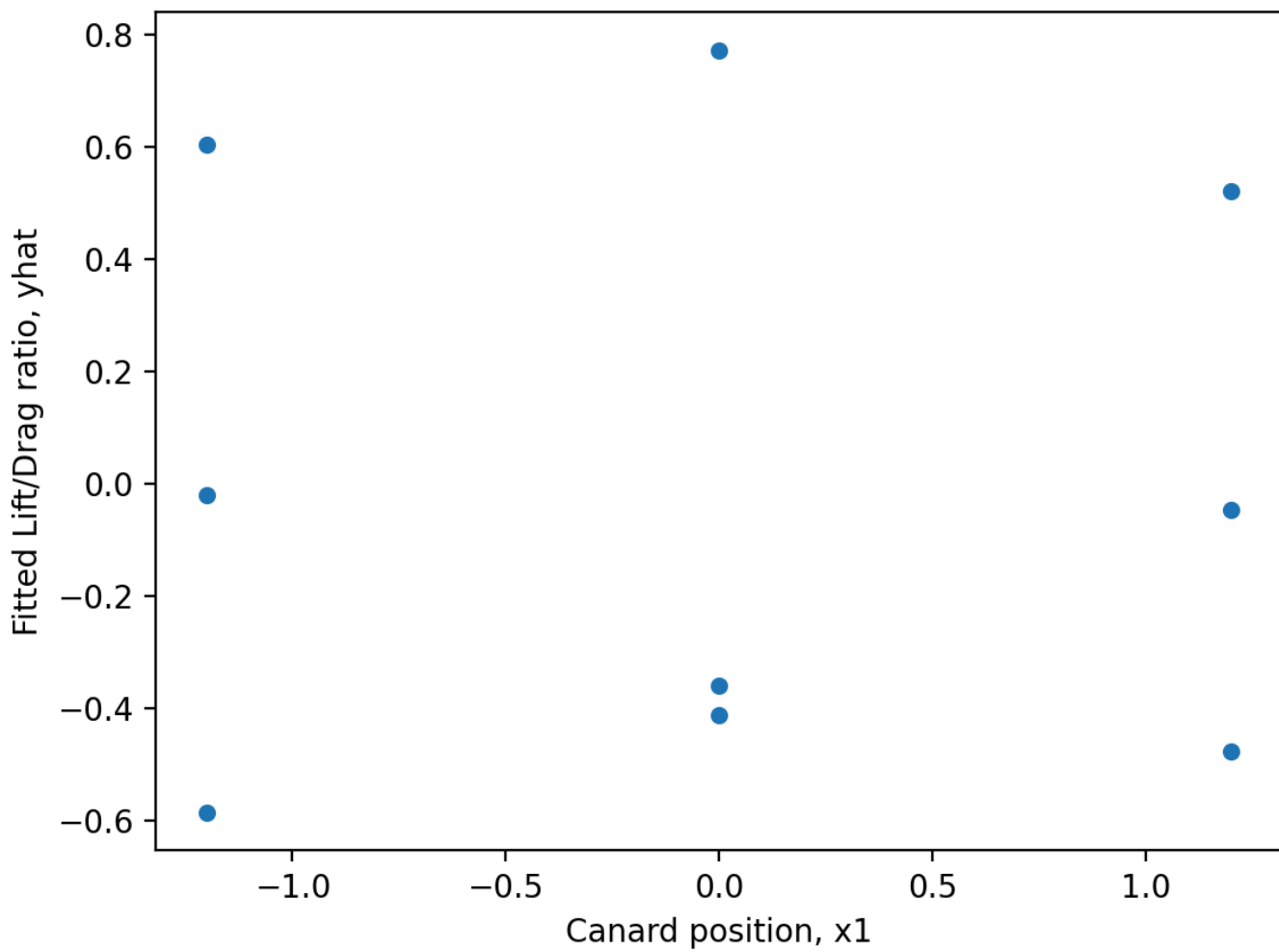


Figure 8.1.5.2 Plot of residuals from equation 8.1.5.1 versus  $x_1$ .

This example is available through Python Jupyter Notebook on the course [GitHub Site](#).

Or use this Binder link for an interactive environment to review this example (special GitHub Site for Example 8.1.5). [Binder Site for example 8.1.5](#).

## 8.1.6 Some Additional Cautions: Extrapolation, Outliers, and Parsimony

Least squares fitting of curves and surfaces is of substantial engineering importance-but it must be handled with care and thought. Before leaving the subject until the next Module 8.2, which explains methods of formal inference associated with it, a few more warnings must be given.

### EXTRAPOLATION

First, it is necessary to warn of the dangers of extrapolation substantially outside the “range” of the  $(x_1, x_2, \dots, x_k, y)$  data. It is sensible to count on a fitted equation to describe the relation of  $y$  to a particular set of inputs  $x_1, x_2, \dots, x_k$  only if they are like the sets used to create the equation. The challenge surface fitting affords is

that when several different  $x$  variables are involved, it is difficult to tell whether a particular  $(x_1, x_2, \dots, x_k, y)$  vector is a large extrapolation. About all one can do is check to see that it comes close to matching some single data point in the set on each coordinate  $(x_1, x_2, \dots, x_k, y)$ . It is not sufficient that there be some point with  $x_1$  value near the one of interest, another point with  $x_2$  value near the one of interest, etc. For example, having data with  $1 \leq x_1 \leq 5$  and  $10 \leq x_2 \leq 20$  doesn't mean that the  $(x_1, x_2)$  pair  $(3, 15)$  is necessarily like any of the pairs in the data set. This fact is illustrated in Figure 8.1.6.1 for a fictitious set of  $(x_1, x_2)$  values.

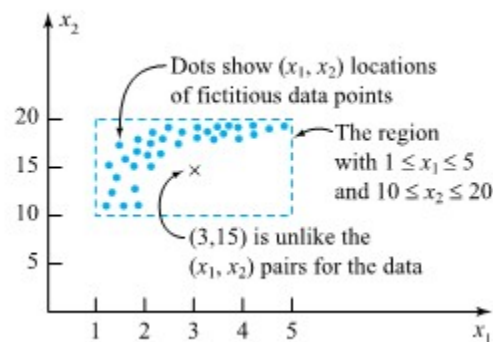


Figure 8.1.6.1 Hypothetical plot of  $(x_1, x_2)$  pairs.

## THE INFLUENCE OF OUTLYING DATA VECTORS

Another potential pitfall is that the fitting of curves and surfaces via least squares can be strongly affected by a few outlying or extreme data points. One can try to identify such points by examining plots and comparing fits made with and without the suspicious point(s).

### Example 8.1.6.1 Stack Loss Data continued

Figure 8.1.3.2 earlier called attention to the fact that the nitrogen plant data set contains one point with an extreme  $x_1$  value. Figure 8.1.6.2 is a scatterplot of  $(x_1, x_2)$  pairs for the data in Table 8.1.3. It shows that by most qualitative standards, observation 1 in Table 8.1.3. is unusual or outlying.

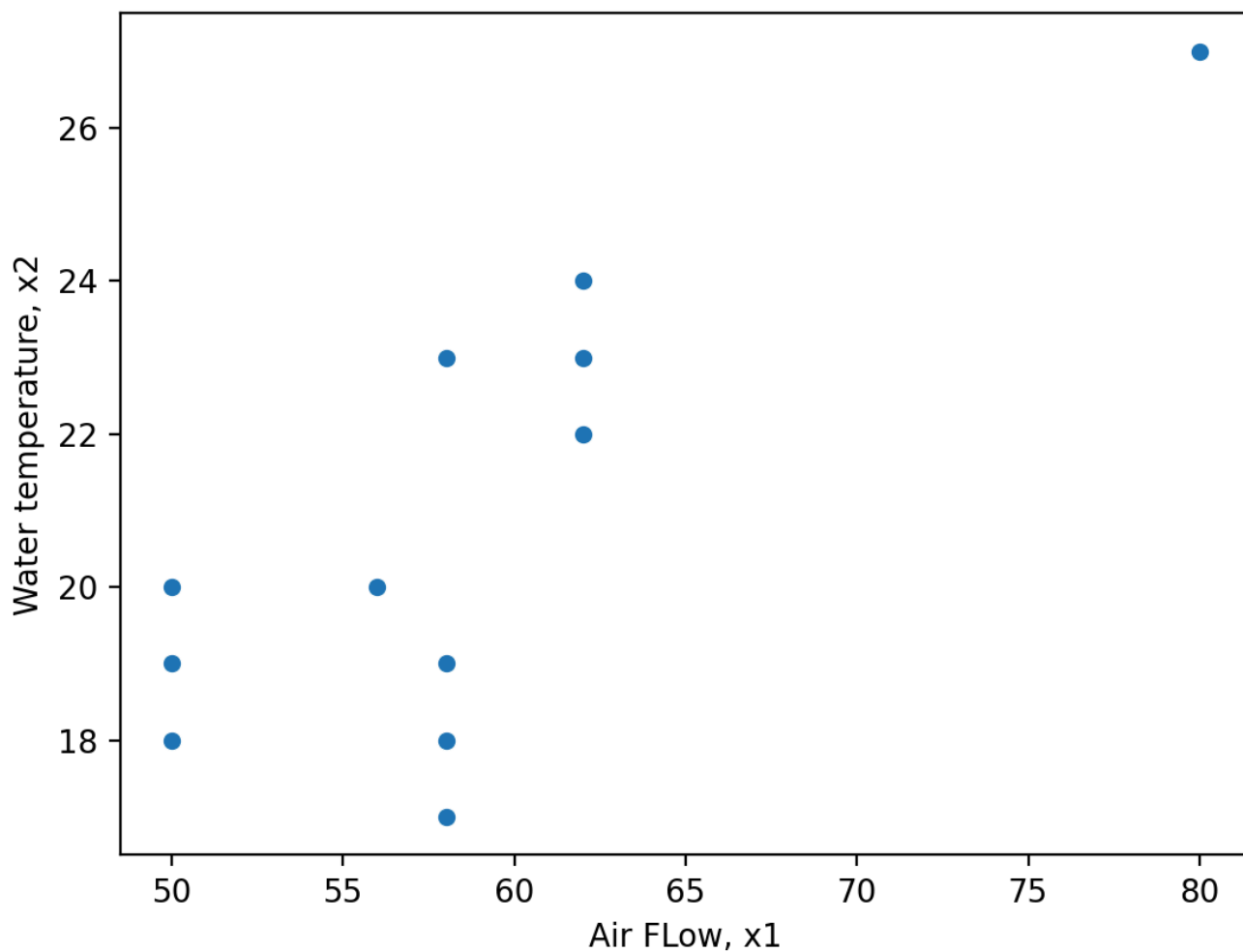


Figure 8.1.6.1

If the fitting of the equation is redone using only the last 16 data points in Table 8.1.3, the equation

$$8.1.6.1 \quad \hat{y} = -56.797 + 1.404x_1 + .601x_2 - .007x_1^2$$

and  $R^2 = .942$  are obtained. Using equation (8.1.61) as a description of stack loss and limiting attention to  $x_1$  in the range 50 to 62 could be considered. But it is possible to verify that though some of the coefficients (the  $b$ 's) in equations (8.1.3.) and (8.1.6.1) differ substantially, the two equations produce comparable  $\hat{y}$  values for the 16 data points with  $x_1$  between 50 and 62. In fact, the largest difference in fitted values is about .4. So, since point 1 in Table 4.8 doesn't radically change predictions made using the fitted equation, it makes sense to leave it in consideration, adopt equation (8.1.3.), and use it to describe stack loss for  $(x_1, x_2)$  pairs interior to the pattern of scatter in Figure 8.1.6.2.

## THE POSSIBILITY OF OVERFITTING

---

Another caution is that the notion of equation simplicity (parsimony) is important for reasons in addition to simplicity of interpretation and reduced expense involved in using the equation. It is also important from the point of view of typically giving smooth interpolation and not **overfitting** a data set. As a hypothetical example, consider the artificial, generally linear  $(x, y)$  data plotted in Figure 8.1.6.3. It would be possible to run a (wiggly)  $k = 10$  version of a polynomial through each of these points. But in most physical problems, such a curve would do a much worse job of predicting  $y$  at values of  $x$  not represented by a data point than would a simple fitted line. A tenth-order polynomial would overfit the data in hand.

## EMPIRICAL MODELS AND ENGINEERING

---

As a final point in this section, consider how the methods discussed here fit into the broad picture of using models for attacking engineering problems. It must be said that physical theories of physics, chemistry, materials, etc. rarely produce equations of the simple forms presented here. Sometimes pertinent equations from those theories can be rewritten in such forms, as was possible with Taylor's equation for tool life earlier in this section. But the majority of engineering applications of the methods in this section are to the large number of problems where no commonly known and simple physical theory is available, and a simple empirical description of the situation would be helpful. In such cases, the tool of least squares fitting of curves and surfaces can function as a kind of "guess" or "template", allowing an engineer to develop approximate empirical descriptions of how a response  $y$  is related to system inputs  $x_1, x_2, \dots, x_k$ .

## 8.1.7 Statistical Computing with Python

Several of the Jupyter Notebook using Python that have been used in this Part on MLR is available to look at and access for download at the course [GitHub Site](#) or at the Special GitHub Sites for Part 8:

[Special GitHub Site for 8.1.1 Fly Ash Data](#) or go to the coding Binder environment at the [Module 8.1.1. Fly Ash Binder Site](#).

[Special GitHub Site for 8.1.3 Stack Loss Data](#) or go to the coding Binder environment at the [Module 8.1.3 Stack Loss Binder Site](#).

[Special GitHub Site 8.1.5 Drag Lift Data](#) or go to the coding Binder environment at the [Module 8.1.5 Drag Lift Binder Site](#).

## 8.1.8 Tutorial 8 - Transformations

At this point, it is recommended that you work your way through the [Tutorial 8 exercise](#) found on the associated GitHub repository. This exercise will teach you how to transform non-linear data so that it can be used with linear models using Python syntax.

**It is strongly recommended that you consult the [Simple Linear Regression Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “Transformations” will be particularly useful.

## 8.1.9 *Transitioning from Simple to Multiple Linear Regression in Python*

Multiple linear regression builds on simple linear regression conceptually but the generation and interpretation of results within Python differs somewhat.

**To facilitate this, it is strongly recommended that you consult the [Multiple Linear Regression Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the files on “Transitioning from Simple to Multiple Linear Regression” and “Multiple Linear Regression” will be particularly useful.



## 8.2.1 Categorical Variable Independent Variables and Dummy Variables

Thus far, we have considered Ordinary Least Squares (OLS) models that include variables measured on interval level scales (or, in a pinch and with caution, ordinal scales). That is fine when we have variables for which we can develop valid and reliable interval (or ordinal) measures. But in engineering, we often want to include in our analysis concepts that do not readily admit to interval measure – including many cases in which a variable has an “on – off”, or “present – absent” quality. In other cases we want to include a concept that is essentially nominal in nature, such that an observation can be categorized as a subset but not measured on a “high-low” or “more-less” type of scale. In these instances we can utilize what is generally known as a dummy variable, but are also referred to as indicator variables, Boolean variables, or categorical variables.

What are “Dummy Variables”?

- – A dichotomous variable, with values of 0 and 1 ;
- – A value of 1 represents the presence of some quality, a zero its absence;
- – The 1 s are compared to the 0s, who are known as the “reference group”;
- – Dummy variables are often thought of as a proxy for a qualitative variable.

Dummy variables allow for tests of the differences in overall value of the  $\bar{Y}$  for different nominal groups in the data. They are akin to a difference of means test for the groups identified by the dummy variable. Dummy variables allow for comparisons between an included (the 1s) and an omitted (the 0s) group. Therefore, it is important to be clear about which group is omitted and serving as the “comparison category.”

It is often the case that there are more than two groups represented by a set of nominal categories. In that case, the variable will consist of two or more dummy variables, with 0/1 codes for each category except the referent group (which is omitted). Several examples of categorical variables that can be represented in multiple regression with dummy variables include:

- Experimental treatment and control groups (treatment = 1, control = 0)
- Gender ( male = 1, female = 0 or vice versa)
- Race and ethnicity (a dummy for each group, with one omitted referent group)
- Product lot (dummy for each product lot with one omitted reference lot)
- Machine setting (dummy for each type with omitted reference type)

The value of the dummy coefficient represents the estimated difference in  $\bar{Y}$  between the dummy group

and the reference group. Because the estimated difference is the average over all of the  $Y$  observations, the dummy is best understood as a change in the value of the intercept ( $A$ ) for the “dummied” group. This is illustrated in Figure 8.2.1.1. In this illustration, the value of  $Y$  is a function of  $X_1$  (a continuous variable) and  $X_2$  (a dummy variable). When  $X_2$  is equal to 0 (the referent case) the top regression line applies. When  $X_2 = 1$ , the value of  $Y$  is reduced to the bottom line. In short,  $X_2$  has a negative estimated partial regression coefficient represented by the difference in height between the two regression lines.

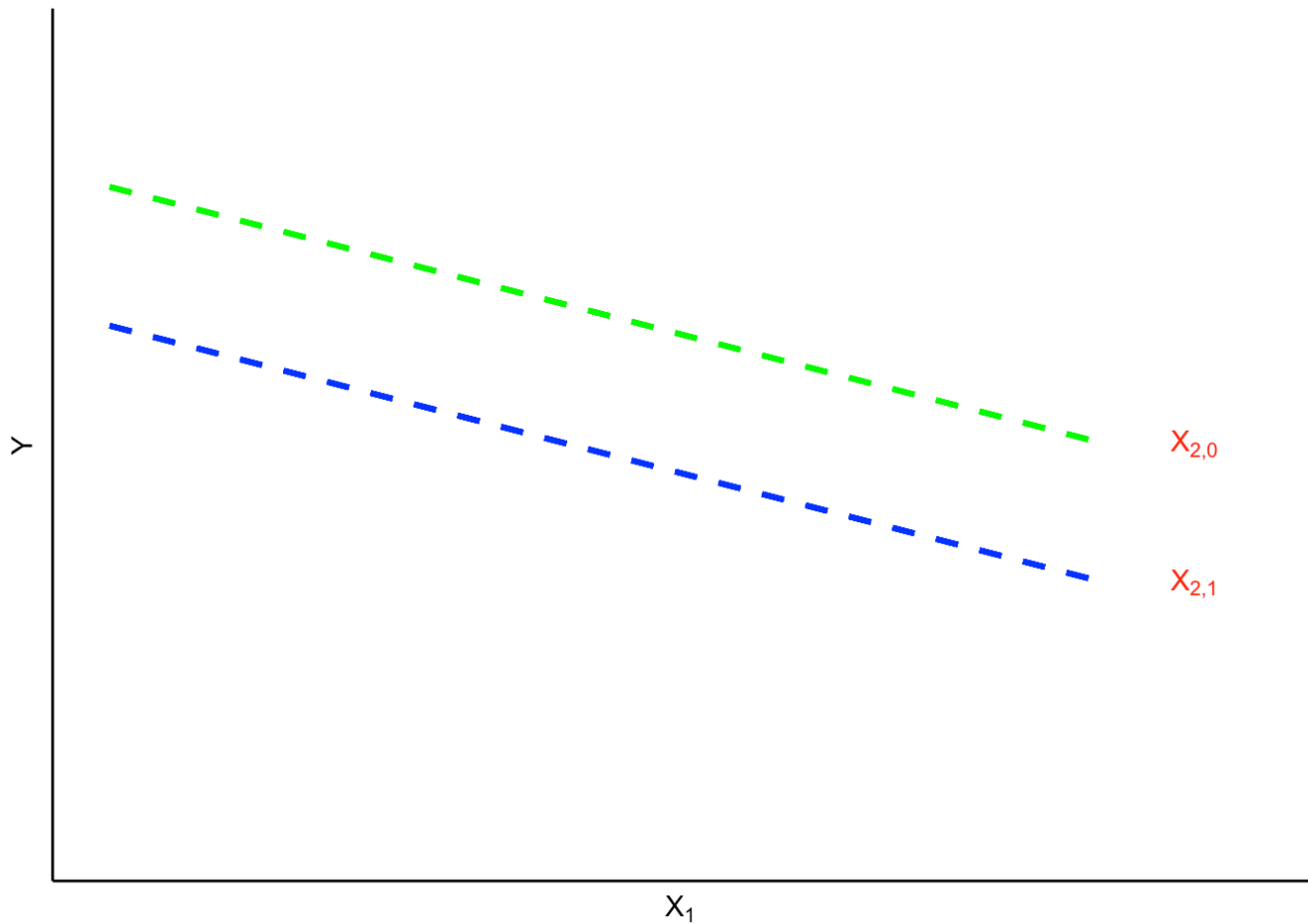


Figure 8.2.1.1 Dummy Intercept Variables

For a case with multiple nominal categories (e.g., region) the procedure is as follows: (a) determine which category will be assigned as the referent group; (b) create a dummy variable for each of the other categories. For example, if you are coding a dummy for four regions (North, South, East and West), you could designate the South as the referent group. Then you would create dummies for the other three regions. Then, all observations from the North would get a value of 1 in the North dummy, and zeros in all others. Similarly, East and West observations would receive a 1 in their respective dummy category and zeros elsewhere. The observations from the South region would be given values of zero in all three categories. The interpretation of the partial regression coefficients for each of the three dummies would then be the estimated difference in  $Y$  between observations from the North, East and West and those from the South.

## INTERACTION EFFECTS WITH DUMMY VARIABLES

---

Dummy variables can also be used to estimate the ways in which the effect of a variable differs across subsets of cases. These kinds of effects are generally called “interactions.” When an interaction occurs, the effect of one  $X$  is dependent on the value of another. Typically, an OLS model is additive, where the  $B$ ’s are added together to predict  $Y$ ;

$$Y_i = ABX_1 + BX_2 + BX_3 + BX_4 + E_i.$$

However, an interaction model has a multiplicative effect where two of the IVs are multiplied;

$$Y_i = ABX_1 + BX_2 + BX_3 * BX_4 + E_i.$$

A “slope dummy” is a special kind of interaction in which a dummy variable is interacted with (multiplied by) a scale (ordinal or higher) variable. Suppose, for example, that you hypothesized that the effects of political ideology on perceived risks of climate change were different for men and women. Perhaps men are more likely than women to consistently integrate ideology into climate change risk perceptions. In such a case, a dummy variable (0 = women, 1 = men) could be interacted with ideology (1 = strong liberal, 7 = strong conservative) to predict levels of perceived risk of climate change (0 = no risk, 10 = extreme risk). If your hypothesized interaction was correct, you would observe the kind of pattern as shown in Figure 8.2.1.2.

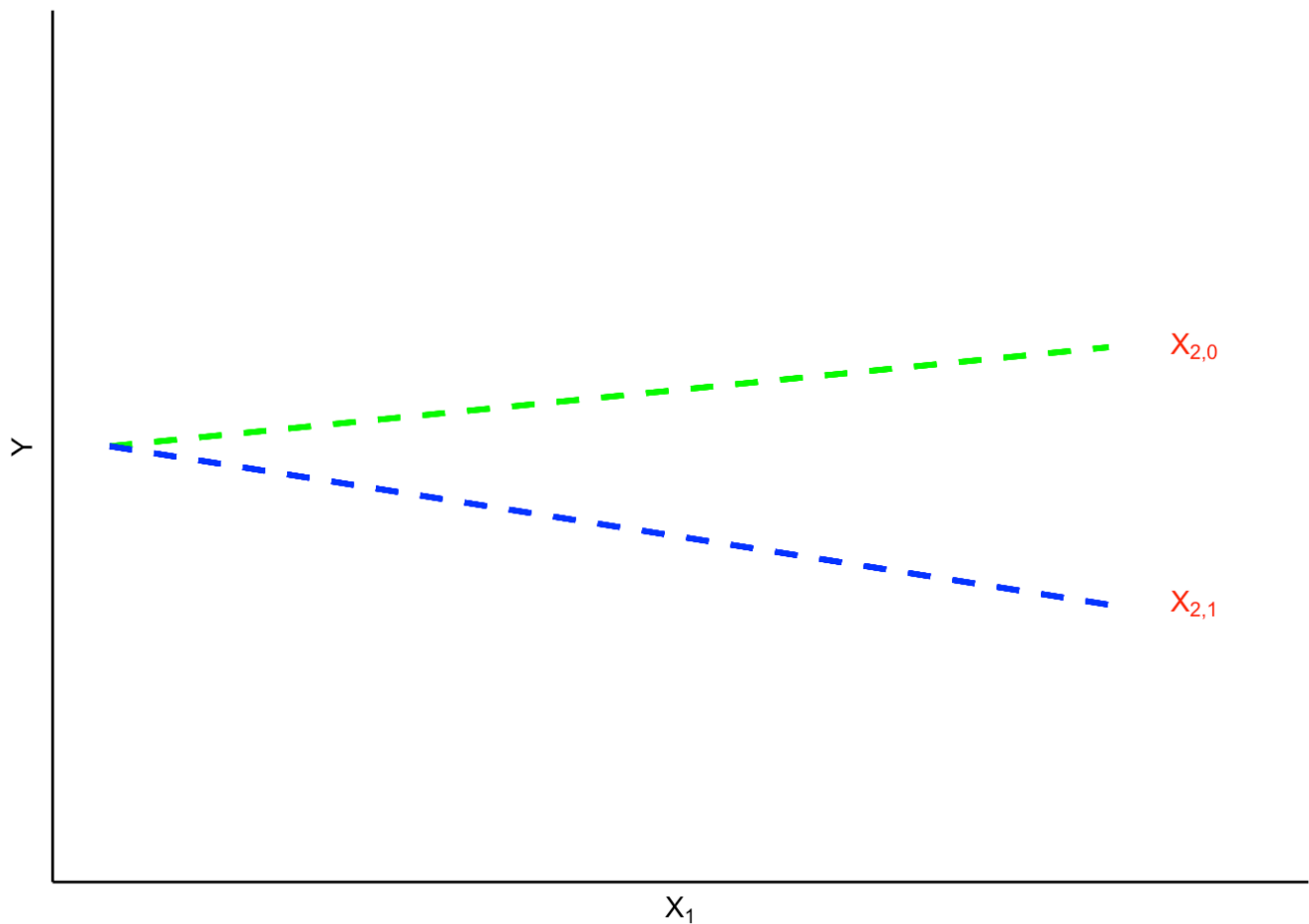


Figure 8.2.1.2 Illustration of Slope Interaction

In sum, dummy variables add greatly to the flexibility of OLS model specification. They permit the inclusion of categorical variables, and they allow for testing hypotheses about interactions of groups with other independent variables within the model. This kind of flexibility is one reason that OLS models are widely used by social scientists and policy analysts.

#### Attribution

---

Material for Chapters 8.2.1.1 and 8.2.2.2 come from Quantitative Research Methods for Political Science, Public Policy and Public Administration: 4th Edition With Applications in R, by *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, and Josie Davis*, located at <https://bookdown.org/ripberjt/qrmbook/>. This work is shared through the licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0).

## 8.2.2 Matrix Algebra and Multiple Regression

Matrix algebra is widely used for the derivation of multiple regression because it permits a compact, intuitive depiction of regression analysis. For example, an estimated multiple regression model in scalar notation is expressed as:  $Y = A + BX_1 + BX_2 + BX_3 + E$ . Using matrix notation, the same equation can be expressed in a more compact and (believe it or not!) intuitive form:  $y = Xb + e$ .

In addition, matrix notation is flexible in that it can handle any number of independent variables. Operations performed on the model matrix  $X$ , are performed on all independent variables simultaneously. Lastly, you will see that matrix expression is widely used in statistical presentations of the results of OLS analysis. For all these reasons, then, we begin with the development of multiple regression in matrix form.

### THE BASICS OF MATRIX ALGEBRA

A matrix is a rectangular array of numbers with rows and columns. As noted, operations performed on matrices are performed on all elements of a matrix simultaneously. In this section we provide the basic understanding of matrix algebra that is necessary to make sense of the expression of multiple regression in matrix form.

#### Matrix Basics

The individual numbers in a matrix are referred to as “elements”. The elements of a matrix can be identified by their location in a row and column, denoted as  $A_{r,c}$ . In the following example,  $m$  will refer to the matrix row and  $n$  will refer to the column.

$$A_{m,n} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

Therefore, in the following matrix;

$$A = \begin{bmatrix} 10 & 5 & 8 \\ -12 & 1 & 0 \end{bmatrix}$$

element  $a_{2,3} = 0$  and  $a_{1,2} = 5$ .

### Vectors

---

A vector is a matrix with single column or row. Here are some examples:

$$A = \begin{bmatrix} 6 \\ -1 \\ 8 \\ 11 \end{bmatrix}$$

or

$$A = [1 \quad 2 \quad 8 \quad 7]$$

### Matrix Operations

---

There are several “operations” that can be performed with and on matrices. Most of these can be computed with python, so we will use this example as we go along.

Go to the interactive [Binder site for the Special GitHub Repository](#) for a tutorial on multiple linear regression using a Wire dataset that will walk you through the concepts of multiple linear regression and using matrix operations to fit the model.

As always, this repository can be found for download at the [course GitHub Site](#).

### Attribution

---

Text for Chapters 8.2.1.1 and 8.2.2.2 come from Quantitative Research Methods for Political Science, Public Policy and Public Administration: 4th Edition With Applications in R, by *Hank Jenkins-Smith, Joseph Ripberger, Gary Copeland, Matthew Nowlin, Tyler Hughes, Aaron Fister, Wesley Wehde, and Josie Davis*, located at <https://bookdown.org/ripberjt/qrmbook/>. This work is shared through the licensed under a [Creative Commons Attribution 4.0 International License](#) (CC BY 4.0).

## 9.0.2 Attribution

Part 9 of this open educational resource is composed of text and is an adaptation of "[Process Improvement Using Data](#)" by [Kevin Dunn](#) is licensed under [CC BY-SA 4.0](#).

While the majority of the text has been rewritten with new examples, strong inspiration was taken from from Chapter 5: Design and Analysis of Experiments of the text and some portions were adapted from that chapter. Formatting for Pressbooks and adaptation of the chapter numbering and nesting have been made. This work and portions of this work are the copyright of Kevin Dunn.

## 9.0.1 Introduction to Design of Experiments

So far, the majority of this resource has been focused on identifying correlations. In this chapter we will investigate how to identify *cause and effect*. We have to disturb and change a system in order to be certain of *cause and effect* between factors and a measurable outcome. It should be emphasized, that despite the name “Design of Experiments”, these principles do not just apply to laboratory work or applied research. Principles in this module will be wide reaching and can be applied to systems as simple as baking cookies to advanced scenarios such as process improvement in a hospital or production facility.



## 9.1.1 Design of Experiments: Introduction

### 9.1.1 DESIGN AND ANALYSIS OF EXPERIMENTS IN CONTEXT

This module will go over how we can purposely disturb the system to learn more about it. Principles presented in earlier modules, in particular those focused on Hypothesis testing and Linear Regression, will be applied here.

### 9.1.2. TERMINOLOGY

To ensure that everyone is on the same page, here is some common terminology that will be used in this section (Table 9.1.2.1) when discussing Design of Experiments (DoE).

**Table 9.1.2.1 Terminology for Design of Experiments**

Term	Definition
Experiment	Changing a system and using the resulting information to improve it
Objective	An objective to improve
Outcome	The measurable result of your experiment
Factor	Things you can actively change to influence the outcome
Levels	Scale to your factors

#### *Objectives & Outcomes*

Let's say that your objective is to improve the yield of a single batch of cookies in your recipe. One such objective could be to increase the number of cookies and thus the measured outcome would be the number of cookies. Alternatively, your objective could be to improve the aesthetics of your cookies and thus your outcome could be the colour of the cookies (eg. white, brown, golden brown). Some more examples are given in the table below (9.1.2.2) :

**Table 9.1.2.2. Example Objectives and Outcomes for Baking Cookies**

Objective	Measured Outcome	Quantitative or Qualitative Outcome
Increase the number of cookies	Number of cookies	Quantitative
Improve cookie aesthetics	Cookie colour	Qualitative
Reduce baking time	Baking time	Quantitative
Improve taste	Taste tester ratings	Qualitative

Each experiment typically has an objective which combines an outcome and the need to adjust that outcome. This objective can be to increase, decrease, or to keep something the same. Outcomes should

always be measurable but can be quantitative or qualitative. Without any outcomes you can not conduct any analysis!

### Factors

---

Factors are the central aspect of DoE as they are the variables that you will change to influence the outcome. In order to perform an experiment, at least one factor should be changed. As with all types of data, you can have numeric or categorical factors and most experiments will have both.

Using the cookie baking example, here are some potential factors:

1. The amount of sugar used in the recipe → *numeric factor*
2. The type of milk used (oat or almond milk) → *categorical factor*
3. The time spent mixing → *numeric factor*
4. Using a stand-mixer or mixing by hand → *categorical factor*

Numeric factors are quantified by measuring and, as such, there is some implied ordering. Using the amount of sugar as an example, 2 cups is greater than 1 cup. Conversely, categorical factors take on a limited number of values. The choice of oat or almond milk has no implicit ordering. However, it should be noted though that many categorical factors could be converted into numeric, continuous variables. For example, the calcium content in oat and almond milk might differ and could be converted to 300 and 400 mg of calcium/cup respectively.

### Levels

---

In the simplest form of Design of Experiments each factor will only have 2 levels, as in the previous example. The above examples all represent factors with two levels: 2 cups or 1 cup of sugar, stand-mixer or hand mixing, 300 or 400 mg calcium/cup. The choice of levels for an experiment is an important decision for the designer and this typically relies on some expertise and/or knowledge of the system. In more complex experiments, factors can have 3, 4 or even more levels. This module will focus on designs with 2 levels per factor since designs with 2 or 3 levels per factor are the most common.

The choice of levels is important. Here are some good practices for choosing the range of levels:

- The level range should be sufficient to show a difference in outcome (too wide though and it may not fit a linear model)
- Do not use extreme values to start
- You want to perturb the system but you do not want to be too granular
- Without prior knowledge, a range of 25% of the normal operating range is a good starting point

When we perform an experiment, we call it a run. If we perform eight experiments, we can say “there are eight runs” in the set of experiments.

### 9.1.3. EXAMPLE OF DESIGN OF EXPERIMENTS

---

Let's say that we are running a bakery and are looking to increase profits. We propose to run an experiment

to determine what the optimal solution is. In this case, we have simplified it to just 2 factors. In later chapters we will discuss methods to narrow down the number of factors for an experiment. We can summarize this problem as follows:

### Example 9.1.3.1. Example of Design of Experiments

**Objective:** Increase profit

**Outcome:** Profit made in a day while selling cookies

**Factors:** Amount of light in the store & Price of Product (see Table 9.1.3.1 for Levels)

9.1.3.1. Levels for Design of Experiments Cookie Example

Factor	Low Level	High Level
Light	Low light (50%)	High Light (75%)
Price	\$7.79	\$8.49

In order to run an experiment, it is essential to consider all possible factor combinations. This is typically displayed in a table known as a **standard order table**. Standard tables are typically given with discrete/coded values of 0, -1, 1 etc.

Table 9.1.3.2. Example Standard Order Table

Experiment	Light	Price
1	-1 (Low)	-1 (Low)
2	1 (High)	-1 (Low)
3	-1 (Low)	1 (High)
4	1 (High)	1 (High)

As shown in Table 9.1.3.2., this order helps us identify all of the possible combinations of factors that you could have in the experiment. Some statistical software packages are also designed around receiving data prepared in this manner. If we were to run these experiments, the table would turn into Table 9.1.3.3 where profit is our measured outcome. Note the column that says "Run". It is imperative that experiments are run in a random order to avoid the impact of disturbances (see 10.5).

Table 9.1.3.3. Experimental Runs for Cookie Design of Experiments

Experiment	Run	Light Level	Price Level	Profit
1	2	Low light (50%)	Low (\$7.79)	\$490
2	1	High light (75%)	Low (\$7.79)	\$570
3	4	Low light (50%)	High (\$8.49)	\$370
4	3	High light (75%)	High (\$8.49)	\$450

Figure 9.1.3.1 visualized this table and from this, certain results can be extracted:

Moving from low to high lighting increases profit, on average, by \$80.

- The difference in profit at low price but changing from low to high lighting gives:  $(\$570 - \$490) = \$80$

- The difference in profit at high price but changing from low to high lightning gives:  $(\$450 - \$370) = \$80$

Increasing the price from \$7.79 to \$8.49 decreases profit, on average, by \$120.

- The difference in profit at low lightning but changing price from \$7.79 to \$8.49 gives:  $(\$370 - \$490) = -\$120$
- The difference in profit at high lightning but changing price from \$7.79 to \$8.49 gives:  $(\$450 - \$570) = -\$120$

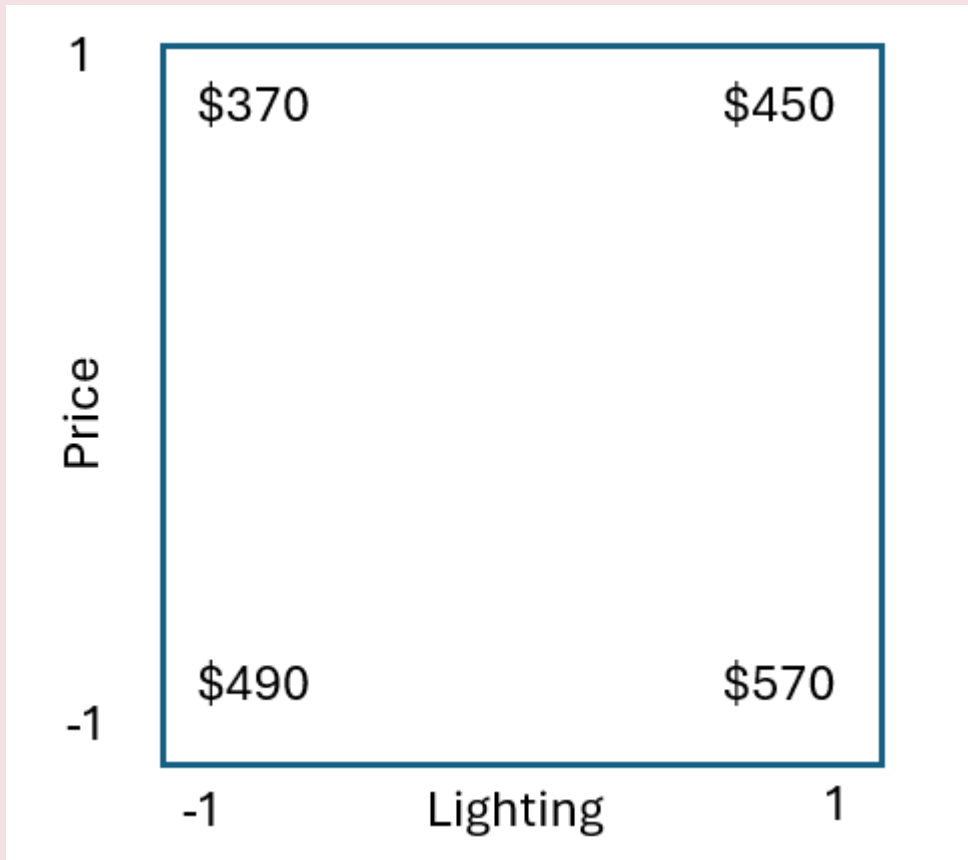


Figure 9.1.3.1. Plot visualizing the standard order table. Profit is shown for the different combinations of lighting and price.

The use of design of experiments allows us to examine interactions between these factors. More specifically, you could then plot contour lines between the various data points, which would allow us to get the “center” (all the potential datapoints inside the square) and not just the perimeter. Additionally, this process can be expanded to multiple factors.

#### 9.1.4. WHY USE DESIGN OF EXPERIMENTS?

A common question that arises when design of experiments is considered is why bother at all? For many systems, there is lots of historical data that exists, why can that not be used? Existing data = historical data = potential happenstance data. Unless there are detailed records, you can not assume that the data was properly disturbed and thus we can only be certain of any identified correlations within the data. Designed experiments are the only way that we can be sure that correlated events are causal! Additionally, without design of experiments, experiments are typically conducted using trial-and-error methods which means changing one-factor at a time. Design of experiments methods reach the optimal solution quicker, are

more efficient and more structured compared to trial-and-error methods. This will be explained in detail in subsequent chapters!

## 9.1.2 Design of Experiments: Analysis

As with any experiment, analysis is necessary before we can to decide what to make of the results. This chapter will introduce you to methods to analyze your Design of Experiments making use of knowledge learned in the [regression modules](#).

### 9.1.5 ANALYSIS OF DESIGN OF EXPERIMENTS

Let's say that we are biomaterials engineers looking to improve upon the design of a dental implant. We are considering the impact of surface roughness and water contact angle on the viability of a potential biomaterial for this application. For the implant to be useful, we want to encourage large amounts of bone cells (osteoblasts) to grow on its surface. Similar to the example in [9.1.3](#), tables are shown below for: the levels (Table 9.1.5.1), standard order (Table 9.1.5.2) and experimental results (Table 9.1.5.3) are shown below. We can summarize this example as follows:

#### Example 9.1.5.1. Analysis of Design of Experiments

**Objective:** Increase viability of dental implant

**Outcome:** Cell viability on the surface of the prospective material

**Factors:** Surface Roughness & Water Contact Angle (see Table 9.1.5.1 for Levels)

Table 9.1.5.1. Levels for Design of Experiments for Dental Implant

Factor	Low Level	High Level
Surface Roughness	300 $\mu\text{m}$	350 $\mu\text{m}$
Water Contact Angle	50°	100°

Table 9.1.5.2. Dental Implant Standard Order Table

Experiment	Surface Roughness	Water Contact Angle
1	-1	-1
2	1	-1
3	-1	1
4	1	1

Table 9.1.5.3. Experimental Runs for Dental Implant Design of Experiments

Experiment	Run	Surface Roughness	Water Contact Angle	Cell Viability (a.u.)
1	4	Low (300 $\mu\text{m}$ )	Low (50°)	31
2	1	High (350 $\mu\text{m}$ )	Low (50°)	70
3	2	Low (300 $\mu\text{m}$ )	High (100°)	56
4	3	High (350 $\mu\text{m}$ )	High (100°)	82

From these four runs we also have a midpoint, the mean, which is 59.75 a.u.. From this we can identify the main effects of Roughness and Water Contact Angle by hand (see Figure 9.1.5.1).

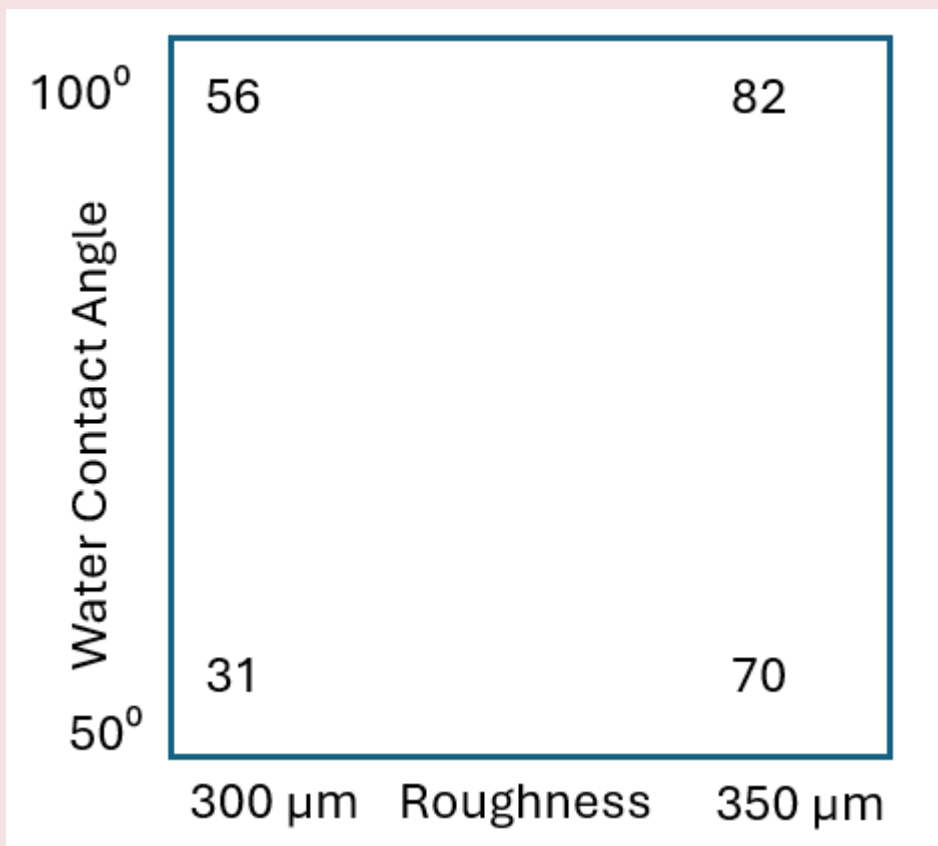


Figure 9.1.5.1. Plot visualizing the standard order table. Cell viability is shown for the different combinations of roughness and water contact angle.

#### Surface Roughness:

Moving from 300 to 350  $\mu\text{m}$  of roughness increases cell viability, on average, by 32.5 a.u. per 50  $\mu\text{m}$ .

- The difference in cell viability at a water contact angle of 50° but changing from 300 to 350  $\mu\text{m}$  of roughness gives:  $(70-31) = 39$  a.u.

- The difference in cell viability at a water contact angle of 100° but changing from 300 to 350  $\mu\text{m}$  of roughness gives:  $(82-56) = 26$  a.u.

#### Water Contact Angle:

Increasing water contact angle from 50 to 100° decreases cell viability, on average, by 18.5 a.u. per 50°.

- The difference in cell viability at 300  $\mu\text{m}$  roughness but changing water contact angle from 50 to 100° gives:  $(56-31) = 25$  a.u.
- The difference in cell viability at 350  $\mu\text{m}$  roughness but changing water contact angle from 50 to 100° gives:  $(82-70) = 12$  a.u.

In most statistical software, these effects are considered to be half of what we just calculated above. This is because we have coded the levels as if we are going from -1 to 1 but these levels are viewed mathematically as being between 0 and 1. As such, our reported half-effects are:

*Surface Roughness increases cell viability, on average, by 16.25 a.u. per 25  $\mu\text{m}$ .*

*Water contact angle increases cell viability, on average, by 9.25 a.u. per 25°.*

**Using ordinary least squares, it can be determined that the OLS model for this system is:**

$$y = 59.75 + 16.25x_1 + 9.25x_2$$

**Where  $y$  is cell viability,  $x_1$  is surface roughness and  $x_2$  is water contact angle.**

### 9.1.6. INTERACTIONS

---

As with linear regression, interactions should also be considered with Design of Experiments. Recall, interactions are when the effect of one factor depends on the level of another factor.

Using the dental implant example from 9.1.5, interaction plots can be generated for roughness and water contact angle (Figure 9.1.6.1). As the two lines are not parallel, this is an overt signal that there is an interaction between roughness and water contact angle. (Any biomaterials engineer would know this to be true!) In fact, any interaction must be symmetrical: if roughness interacts with water contact angle, water contact angle interacts with roughness to the same magnitude.



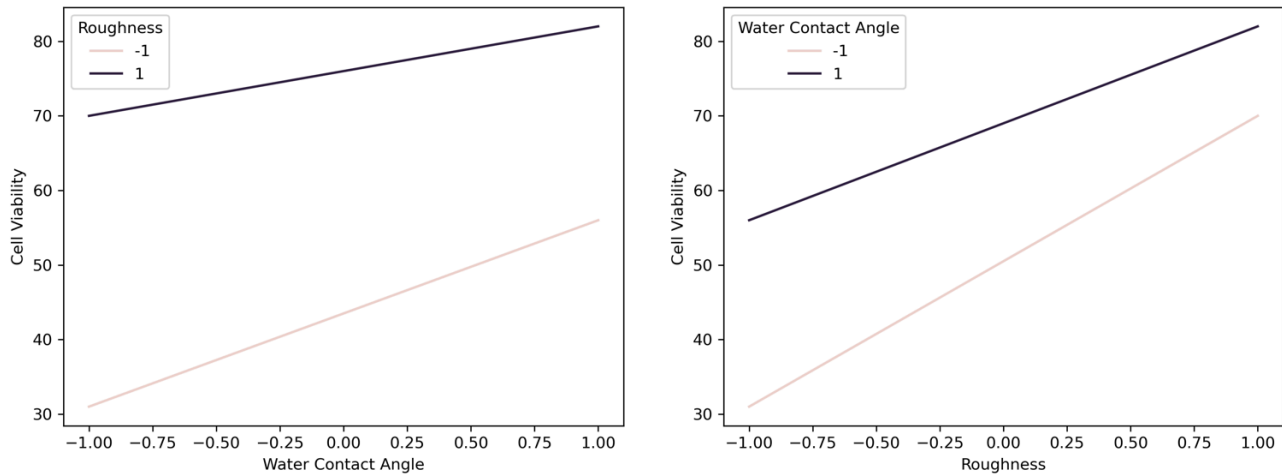


Figure 9.1.6.1. Interaction plots of surface roughness and water contact angle for dental implant design of experiments example

If we wished to calculate the interaction terms by hand it would look like this:

Surface Roughness:

- At high water contact angle:  $82 - 56 = 26$  a.u.
- At low water contact angle:  $70 - 31 = 39$  a.u.
- $(26 - 39)/2 = -6.5$

Water Contact Angle:

- At high roughness:  $82 - 70 = 12$  a.u.
- At low roughness:  $56 - 31 = 25$  a.u.
- $(26 - 39)/2 = -6.5$

Average interaction term =  $-6.5/2 = -3.25$  a.u.

Recall that we divide by two again because we have coded the levels as if we are going from -1 to 1 but these levels are viewed mathematically as being between 0 and 1.

**With the interaction term included, we can create the following OLS model:**

$$y = 59.75 + 16.25x_1 + 9.25x_2 - 3.25x_3$$

Where  $y$  is cell viability,  $x_1$  is surface roughness,  $x_2$  is water contact angle, and  $x_3$  is the interaction term.

### 9.1.7. WHERE DO WE GO NEXT?

Your experiments are just the “first guess” to help you understand your system. If you want to truly optimize your system, subsequent experiments will be necessary. However, this leads us to the question of “where do we go next?”.

To determine this, we have to move the levels of our factors in a direction that optimizes our objective. In the case of the example used in 9.1.5 and 9.1.6, this would be altering our levels of surface roughness and water contact angle in a direction that we believe will lead to improved cell viability. This is best visualized through the use of a contour plot (see Figure 9.1.7.1). Based on this plot, our next experiments for this dental implant would be in the *top right* portion of the plot (ie. higher contact angle and higher surface roughness). The use of contour plots is useful for 2 or 3 factors systems but with increased complexity we can not visualize it anymore. Instead, a vector can be calculated to determine the direction to pursue to increase the measured outcome.

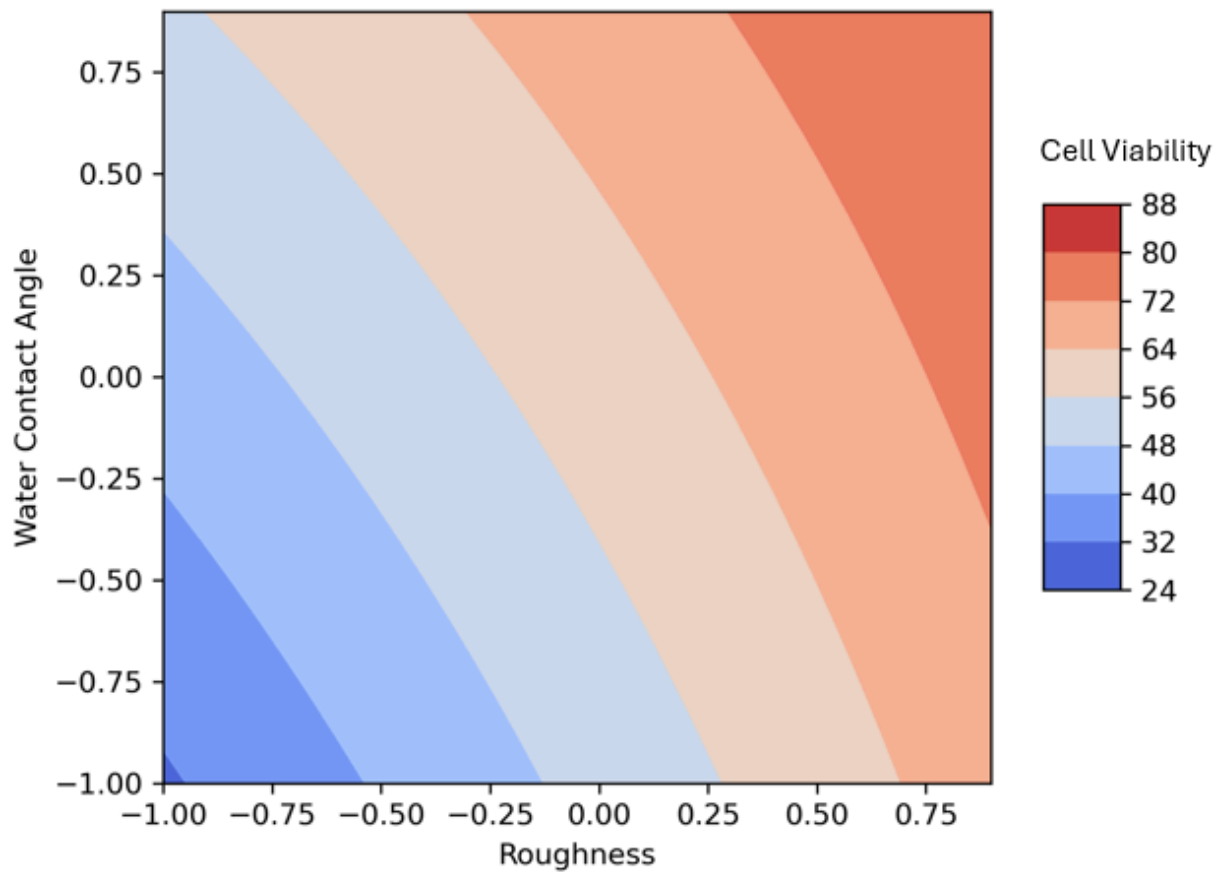


Figure 9.1.7.1. Contour plot showing the interactions between surface roughness and water contact angle on cell viability.

## 9.1.3 Tutorial 9 - Design of Experiments

At this point, it is recommended that you work your way through the [Tutorial 9 exercise](#) found on the associated GitHub repository. This exercise will teach you how to properly import data from a Standard Order Table so that you can compute an OLS model using Python syntax.

**It is strongly recommended that you consult the [Design of Experiments Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the file on “Full Factorial Example” will be particularly useful.

## 9.2.1 Design of Experiments: Full Factorial Designs

### 9.2.1. FULL FACTORIAL DESIGNS

---

As we have demonstrated in [9.1.3](#) and [9.1.5](#), we can use a design of experiments to investigate the effects of several factors simultaneously. This is a more efficient approach for gathering information on our system.

Ultimately, we need to determine how many experiments are required. Based on the number of factors ( $k$ ), and their corresponding number of levels ( $X$ ), the number of experiments in a factorial design is given by:  $X^k$ .

For the cookie example in [9.1.3](#) we had 2 factors (light & price) and each factor had two levels. Therefore, the number of experiments was  $2^2$  experiments or 4 experiments. This was a factorial design. Naturally this can be scaled up to 3, 4 or 5 factors (or even higher) giving us 8, 16 and 32 experiments respectively assuming each factor has 2 levels. These are known as *Full Factorial Designs*.

### 9.2.2 APPLYING LINEAR REGRESSION TO FACTORIAL DESIGNS

---

Now, suppose we apply a linear regression model to a factorial design where we have four parameters to estimate and four data points. This means that we have *no degrees of freedom* afterwards and thus we will have no residual errors. This means that we can not compute any hypothesis tests on the parameters or generate confidence intervals. In section 9.2.3 we will address how we can adjust the design so that we have residual errors and can compute desired hypothesis tests.

#### Example 9.2.2.1 Applying Linear Regression to Factorial Designs

For now, using the example from [9.1.5](#) (see Table 9.2.2.1), we can generate the following least squares regression model for the sample as:

$$y_i = b_0 + b_R x_R + b_W x_W + b_{RW} x_{RW} + e$$

Table 9.2.2.1. Experimental Runs for Dental Implant Design of Experiments

Experiment	Run	Surface Roughness	Water Contact Angle	Cell Viability (a.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

We can conceptualize this set of experiments using matrices as per below (where  $\mathbf{x}_R$  = Surface Roughness and  $\mathbf{x}_W$  = Water Contact Angle):

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & R- & W- & (R - W-) \\ 1 & R+ & W- & (R + W-) \\ 1 & R- & W+ & (R - W+) \\ 1 & R+ & W+ & (R + W+) \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

$$\begin{bmatrix} 31 \\ 70 \\ 56 \\ 82 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

We can solve this system using our knowledge of linear regression. Since our system is orthogonal, the Matrix  $(\mathbf{X}^T \mathbf{X})$  has only non-zero values on the diagonal. Therefore:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y})$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 239 \\ 65 \\ 37 \\ -13 \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{y}) = \begin{bmatrix} (1/4) & 0 & 0 & 0 \\ 0 & (1/4) & 0 & 0 \\ 0 & 0 & (1/4) & 0 \\ 0 & 0 & 0 & (1/4) \end{bmatrix} \begin{bmatrix} 239 \\ 65 \\ 37 \\ -13 \end{bmatrix} = \begin{bmatrix} 59.75 \\ 16.25 \\ 9.25 \\ -3.25 \end{bmatrix}$$

The resulting equation,  $\mathbf{y} = 59.75 + 16.25\mathbf{x}_R + 9.25\mathbf{x}_W - 3.25\mathbf{x}_{RW}$ , can be interpreted the same manner as before. For example, a 1 unit increase in roughness corresponds to a 16.25 a.u. increase in cell viability. This method also explains why

we had to divide by 2 a second time earlier since this coefficient represents the effect of changing surface roughness from 0 to 1 or from 325 to 350  $\mu\text{m}$ . The same is true for water contact angle as well. Finally, the interaction term decreases cell viability by 3.25 units if both surface roughness and water contact angle are at the same level (both high or both low).

### 9.2.3. DETERMINING STATISTICAL SIGNIFICANCE

---

As mentioned in the previous section, with no available degrees of freedom, no hypothesis tests or confidence intervals can be generated for the main effects or interaction terms.

With a Full Factorial Design there are a couple of choices:

1. Run a full set of replicates
2. Add center points
3. Remove factors that have low magnitude or are not of interest
4. Utilize a confounding pattern or fractional design

#### 1) *Run a full set of replicates*

With infinite resources and time, this would be the simplest method as you would have more experiments than parameters. This would give you the required degrees of freedom to calculate the standard error of all the model coefficients. However, this is usually an inefficient solution and will utilize a significant amount of resources. There are better choices available but it is always an option. Once you have degrees of freedom, you can identify which coefficients are significant or not and then removing coefficients will give you additional degrees of freedom.

#### 2) *Add center points*

Center points are halfway parameters between the levels of a given factor. Using the biomaterials example, a center point at 325  $\mu\text{m}$  surface roughness and 75° water contact angle could be run. This may be performed as many times as desired since adding these does not change the orthogonality of X and adds degrees of freedom to facilitate calculation of the standard error. As it does not require as many runs as the full set of replicates, adding center points is always a viable option. Similar to the full replicate situation, once you have degrees of freedom you can identify which coefficients are significant or not and then removing coefficients will give you additional degrees of freedom. In matrix notation it would look like the following if we did three replicates:

**Example 9.2.2.2 Demonstrating how adding replicates provides degrees of freedom**

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & R- & W- & R-W- \\ 1 & R+ & W- & R+W- \\ 1 & R- & W+ & R-W+ \\ 1 & R+ & W+ & R+W+ \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_{RW} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_{RW} \end{bmatrix}$$

### 3) Remove factors that have low magnitude or interest

With a full factorial design, you also have the choice of removing a coefficient, even if you do not have confidence intervals to support your choice. If you coefficient yields a magnitude of 0.00001 but you are working in a practical setting where changes in the system work with values in the order of 100s or 1000s, it might not be practical to keep the coefficient (even if it was statistically significant) because it would have little practical and/or clinical relevance. Removing coefficients this way should be done with caution as context and knowledge of the system is needed to do this properly. As with the previous two examples, doing so will give you available degrees of freedom to calculate the standard error.

Pareto Plots (see Figure 9.2.3.1) are a way to help you visualize this concept. By sorting the coefficients from lowest to highest magnitude (excluding the intercept) and then plotting them in a bar plot, one can quickly establish which coefficients have larger impacts on the system.

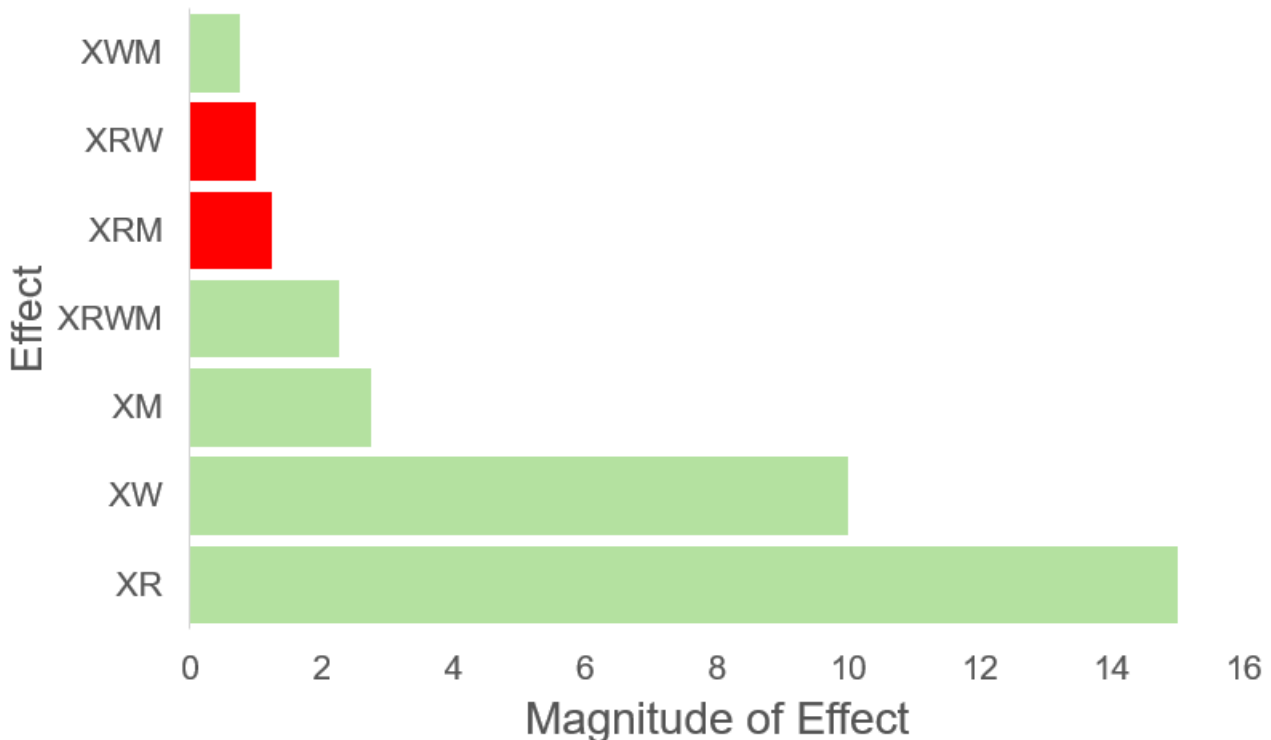


Figure 9.2.3.1 Pareto plot for dental implant design of experiments

From Figure 9.2.3.1 we can quickly identify which coefficients have larger impacts on the outcome. Use of

colour here can also allow us to quickly identify which coefficients have a positive (green) impact on the outcome compared to the coefficients that have a negative (red) impact on the outcome.

#### 4) Utilize a confounding pattern or fractional design

See chapters [9.2.6](#) and [9.2.7](#) as this is an essential concept to fractional designs.

## 9.2.4. INCREASING THE NUMBER OF FACTORS

All the examples used so far have been for situations where we have two factors. Increasing the number of factors does add complexity but the underlying methods and math remains the same.

### Example 9.2.4.1. Increasing the Number of Factors

Let's say we take the biomaterials example and now we consider a 3rd factor, the material, as dental implants at your company have been designed to use either titanium or stainless steel. Tables for: the levels (Table 9.2.4.1), standard order (Table 9.2.4.2) and experimental results (Table 9.2.4.3) are shown below.

**Table 9.2.4.1. Levels for Design of Experiments for Dental Implant**

Factor	Low Level	High Level
Surface Roughness	300 $\mu\text{m}$	350 $\mu\text{m}$
Water Contact Angle	50°	100°
Material	Titanium	Stainless Steel

**Table 9.2.4.2. Dental Implant Standard Order Table**

Experiment	Surface Roughness	Water Contact Angle	Material
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



Table 9.2.4.3. Experimental Runs for Dental Implant Design of Experiments

Experiment	Run	Surface Roughness	Water Contact Angle	Material
1	8	-1 (300 $\mu\text{m}$ )	-1 (50°)	-1 (Titanium)
2	5	+1 (350 $\mu\text{m}$ )	-1 (50°)	-1 (Titanium)
3	2	-1 (300 $\mu\text{m}$ )	+1 (100°)	-1 (Titanium)
4	6	+1 (350 $\mu\text{m}$ )	+1 (100°)	-1 (Titanium)
5	1	-1 (300 $\mu\text{m}$ )	-1 (50°)	1 (Stainless Steel)
6	7	+1 (350 $\mu\text{m}$ )	-1 (50°)	1 (Stainless Steel)
7	3	-1 (300 $\mu\text{m}$ )	+1 (100°)	1 (Stainless Steel)
8	4	+1 (350 $\mu\text{m}$ )	+1 (100°)	1 (Stainless Steel)

The corresponding matrix model is (where  $\mathbf{x}_M$  = Material Coefficient):

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{bmatrix} 31 \\ 70 \\ 56 \\ 82 \\ 42 \\ 67 \\ 61 \\ 91 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_R \\ b_W \\ b_M \\ b_{RW} \\ b_{RM} \\ b_{WM} \\ b_{RWM} \end{bmatrix} + \begin{bmatrix} e_0 \\ e_R \\ e_W \\ e_M \\ e_{RW} \\ e_{RM} \\ e_{WM} \\ e_{RWM} \end{bmatrix}$$

The resulting equation is then:

$$\mathbf{y} = 62.5 + 15\mathbf{x}_R + 10\mathbf{x}_W + 2.75\mathbf{x}_M - 1\mathbf{x}_{RW} - 1.25\mathbf{x}_{RM} + 0.75\mathbf{x}_{WM} + 2.25\mathbf{x}_{RWM}$$

## 9.2.2 Design of Experiments: Disturbances and Blocking

### 9.2.5. UNDERSTANDING DISTURBANCES

Every experiment will have external elements that can or will impact the outcomes. We call these disturbances. As scientists or engineers, it is our job to design our experiments to reduce the impact of disturbances where possible.

Generally, we can classify disturbances as:

- Known vs. Unknown
- Controllable vs. Uncontrollable
- Measurable vs. Unmeasurable

In an ideal situation all disturbances would be known, controllable and measurable but this is almost never the case. Whether its the ambient temperature, an unexpected change in the stock market, or the choice of an individual operator, much of this can not be controlled or even planned for and this is why randomization is so critical. Randomization will ensure that disturbances cannot systematically affect the outcome.

A common method to handle disturbances is to design the experiment to account for them. If the disturbance is controlled and is held constant for all experiments, it is no longer a disturbance since its effect cancels out. Pairing can also cancel out the effect of disturbances by using the same subject/specimens for the same reasons as stated in [Module 5](#). We can classify factors depending on their capacity to be controlled and/or measured (Table 9.2.5.1). Covariates are parameters that are capable of altering the outcome but are not of interest to you. An example is something like ambient temperature. For many experiments it is not of major interest but it could influence the outcome. Blocking will be discussed in 9.2.6.

		Measurable	
		Yes	No
Controllable	Yes	Factors	Blocking
	No	Covariates	Disturbances

Table 9.2.5.1. Table demonstrating how to classify factors depending on whether they are measurable and/or controllable.

## 9.2.6. BLOCKING (AND CONFOUNDING)

Through clever design, blocking allows us to minimize the impact of a disturbance on our interpretation of the system. Blocking is used when we have disturbances that we are aware of but we do not have the means to control them. The solution is to purposely confound the effect of the disturbance with another effect in the system that is anticipated to be small (or insignificant).

Let's say we have a system with 3 factors: A, B and C. In factorial designs, the highest order interaction terms tend to have very small impacts on the outcome so this makes them an appealing coefficients to confound with a disturbance. Effectively, we will not be able to tell the difference between the interaction effect of ABC and the disturbance. You could also state that the corresponding coefficient is:  $b_{ABC} = \text{ABC interaction effect} + \text{disturbance}$ .

This concept can be combined with experimental runs to use a process called blocking. Normally, with 3 factors we would have  $2^3$  experiments but with blocking we split the runs in half so that half the runs are at ABC+ and half are at ABC-.

### Example 9.2.6.1. Blocking (and Confounding) Example

For example, let's say that we are experimenting with marketing for a cell phone app with the measured outcome of in-app purchases 60-days after marketing. Our three factors are the promotion (A), the message sent (B), and the price (C). However, we quickly realize that some people in our study will have iPhones, while others will have Androids. The type of phone that our users have fits the criteria of a factor that we can measure but not control. See Table 9.2.6.1 to see how this is conceptualized.

Table 9.2.6.1. Standard Order Table for Cell Phone App Experiment

Experiment	A (Promotion)	B (Message)	C (Price)	AB	AC	BC	ABC (Confound)
1	-	-	-	+	+	+	-(iPhone)
2	+	-	-	-	-	+	+(Android)
3	-	+	-	-	+	-	+(Android)
4	+	+	-	+	-	-	-(iPhone)
5	-	-	+	+	-	-	+(Android)
6	+	-	+	-	+	-	-(iPhone)
7	-	+	+	-	-	+	-(iPhone)
8	+	+	+	+	+	+	+(Android)

There is inevitably some confusion present now as the effect of the ABC interaction term and the type of phone can not be separated. However, this trade-off is beneficial to us as our main effects and two-factor interactions can be interpreted without bias assuming that the disturbance was held constant.

## 9.2.3 Design of Experiments: Fractional Designs

### 9.2.7. FRACTIONAL DESIGNS

With  $2^k$  runs, it should become quite apparent that as we increase the number of factors ( $k$ ), the amount of resources required will quickly inflate. As such, there is a necessary discussion that should be had around methods of reducing the amount of work we need to do and conserve resources spent. This is most applicable for scenarios when you are screening or evaluating a new system. This could be lab-scale exploration, making a new product or even troubleshooting a problem.

This concept relies on us using the concept of confounding, previously introduced in [9.2.6](#). By confounding factors with one another, we can reduce the number of required runs to effectively halve the amount of work needed. A  $2^k$  experiment can become a  $2^{k-1}$  experiment through this principle, this is known as fractional design. This works because we typically care more about the main effects and interactions tend to have limited practical significance (especially 3-factor and above).

#### Example 9.2.7.1 Fractional Design Example

As shown in Table 9.2.7.1, we can take a  $2^3$  experiment, which has 8 runs, and halve it to 4 runs by confounding one factor with the interaction of the other two factors. We write the first two factors as normal but the third factor is written as a product of the first two factors.

Table 9.2.7.1. Experimental runs for  $2^{3-1}$  system where factor C is confounded with the interaction of AB

Run	A	B	C=AB
1	-	-	+
2	+	-	-
3	-	+	-
4	+	+	+

Now the important question is what is the consequence of doing this?

1) We only have to do half the work! This can not be understated. We reduced the amount of used resources and the efficiency of the process has increased. Especially when we consider that initial experiments will not find our optimal parameters and we will have to conduct serve experiments to determine these (see [9.3](#)).

2) We now have several confounded factors. Each of the main effects (**ABC**) will now be confounded with an interaction term.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$y_i = b_0 + b_A x_a + b_B x_B + b_C x_C + b_{AB} x_{AB} + b_{AC} x_{AC} + b_{BC} x_{BC} + b_{ABC} x_{ABC} + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_A \\ b_B \\ b_C \\ b_{AB} \\ b_{AC} \\ b_{BC} \\ b_{ABC} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

It should be apparent that this system is now *underdetermined* as we have 8 unknowns but only 4 equations as a result of only doing 4 runs. We also note that the X matrix is no longer orthogonal. The solution to this is to exactly what was stated above – confound the main effects with interaction terms. This is shown below:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} b_0 + b_{ABC} \\ b_A + b_{BC} \\ b_B + b_{AC} \\ b_C + b_{AB} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

For instance, we would now state that the main effect of **A** is confounded with the interaction of **BC**, since the model coefficient is the sum of these two effects. From this we can state that **A** is an alias for **BC**, that **B** is an alias for **AC**, that **C** is an alias for **AB** and that the intercept is aliased with the 3-factor interaction **ABC**. This can be expressed by the series of equations below.

$$b_0 + b_{ABC} \rightarrow \mathbf{I} + \mathbf{ABC}$$

$$b_A + b_{BC} \rightarrow \mathbf{A} + \mathbf{BC}$$

$$b_B + b_{AC} \rightarrow \mathbf{B} + \mathbf{AC}$$

$$b_C + b_{AB} \rightarrow \mathbf{C} + \mathbf{AB}$$

## 10.8. GENERATORS

For a 3 factor system, it is fairly simple to determine the confounding patterns. However, with larger numbers of factors this becomes much more complicated. We can use generators to simplify this process for us.

For a 4 factor system,  $2^4$ , we would have factors **A**, **B**, **C** and **D**. Factors **A** through **C** would be considered as normal but Factor **D** would be written as **D = ABC**. This is called the *generating relation*.

To work with a *generating relation*, one needs to be aware of some rules:

1. The intercept **I** is a column of ones.
2. When a factor is multiplied by itself it is the identity (or intercept):  $\mathbf{AxA} = \mathbf{I}$
3. A factor multiplied by the identity (or intercept or a column of ones) is equal to itself:  $\mathbf{AxI} = \mathbf{A}$
4. Through some algebra we can also establish the defining relation of  $\mathbf{I} = \mathbf{ABCD}$ . Take the

generating relation and multiply both sides by  $\mathbf{D}$ . Through some algebra we can also establish the defining relation of  $\mathbf{D}$ .

By multiplying a main effect by the defining relationship, we can quickly determine the factor that it is aliased with. For example, for the  $2^{4-1}$  half fraction we can see that  $\mathbf{A}$  is aliased with  $\mathbf{BCD}$  by the following equation:

$$\mathbf{A} \times \mathbf{I} = \mathbf{A} \times \{\mathbf{ABCD}\} = \{\mathbf{AA}\} \times \{\mathbf{BCD}\} = \mathbf{I} \times \{\mathbf{BCD}\} = \{\mathbf{BCD}\}$$

We know that for the  $2^{3-1}$  half fraction that the generating relation is  $\mathbf{I} = \mathbf{ABC}$  which tells us that  $\mathbf{B}$  is aliased with  $\mathbf{AC}$  by:  $\mathbf{B} \times \mathbf{I} = \mathbf{B} \times \mathbf{ABC} = \mathbf{AC}$

### 9.2.9. RESOLUTION

A result of confounding and/or using fractional designs is a trade-off with regards to resolution. Resolution is the degree to which an estimated main effect(s) is aliased (or confounded) with estimated 2-level, 3-level, or higher interactions. The resolution is considered to be one more than the smallest order interaction that some main effect is confounded with. This can be best visualized through the trade-off table shown below (Figure 9.2.9.1).


		Number of factors, $k$						
		3	4	5	6	7	8	9
increasing cost  Number of runs	4	$2^{3-1}_{III}$						
	8	$2^3$ $\pm C=AB$ <i>full</i>	$2^{4-1}_{IV}$ $\pm D=ABC$	$2^{5-2}_{III}$ $\pm D=AB$ $\pm E=AC$	$2^{6-3}_{III}$ $\pm D=AB$ $\pm E=AC$ $\pm F=BC$	$2^{7-4}_{III}$ $\pm D=AB$ $\pm E=AC$ $\pm F=BC$ $\pm G=ABC$		
	16	$2^3$ <i>twice</i>	$2^4$ <i>full</i>	$2^{5-1}_{V}$ $\pm E=ABCD$	$2^{6-2}_{IV}$ $\pm E=ABC$ $\pm F=ABD$	$2^{7-3}_{IV}$ $\pm E=ABC$ $\pm F=ABD$ $\pm G=ACD$	$2^{8-4}_{IV}$ $\pm E=ABC$ $\pm F=ABD$ $\pm G=ACD$ $\pm H=BCD$	$2^{9-5}_{III}$
	32	$2^3$ <i>4 times</i>	$2^4$ <i>twice</i>	$2^5$ <i>full</i>	$2^{6-1}_{VI}$ $\pm F=ABCDE$	$2^{7-2}_{IV}$ $\pm F=ABC$ $\pm G=ABDE$	$2^{8-3}_{IV}$ $\pm F=ABC$ $\pm G=ABD$ $\pm H=ACDE$	$2^{9-4}_{IV}$
	64	$2^3$ <i>8 times</i>	$2^4$ <i>4 times</i>	$2^5$ <i>twice</i>	$2^6$ <i>full</i>	$2^{7-1}_{VII}$ $\pm G=ABCDEF$	$2^{8-2}_{V}$ $\pm G=ABCD$ $\pm H=ABEF$	$2^{9-3}_{IV}$
		increasing information about additional factors				lower resolution greater aliasing		

Figure 9.2.9.1. Trade-off Table for Design of Experiments demonstrating how resolution and aliasing are related.

Consider the example of:  $2_{\text{IV}}^{4-1}$ . Here the roman numerals **IV** indicate the level of resolution for the design. This number is equivalent to the number of factors present in the defining relation. Since **I = ABCD** for a  $2_{\text{IV}}^{4-1}$  experiment, we say that this is a resolution **IV** design.

As a general practice:

- Resolution III designs are good for screening
- Resolution IV designs are good for characterizing
- Resolution V designs are good for optimizing

Note that none of these designs have any confounding between the main effects.

*Unique Features of Resolution III, IV & V Designs are given as follows:*

---

Resolution III Designs:

- Main effects confounded with two-factor interactions

Resolution IV Designs:

- Main effects are aliased with three-factor interactions
- Two-factor interactions are aliased with each other

Resolution V Designs:

- Have no aliasing between main effects or two-factor interactions
- Two-factor interactions are aliased with three-factor interactions

## 9.3.1 Design of Experiments: Optimization and Response Surface Methods

### 9.3.1. OPTIMIZATION

Ultimately, we are experimenting with the goal of optimizing a system. Factorial or fractional designs are good for initial trials when we have limited information. After this we can proceed with a sequence of experiments to ensure that we slowly replace factorial experiments with designs that are closer to the optimal conditions. This procedure is called response surface methods (RSM).

#### RSM for a Single Variable

First let's consider the effect of a single factor,  $x_1$  as it relates to our response,  $y$ . This is to illustrate the general response surface process.

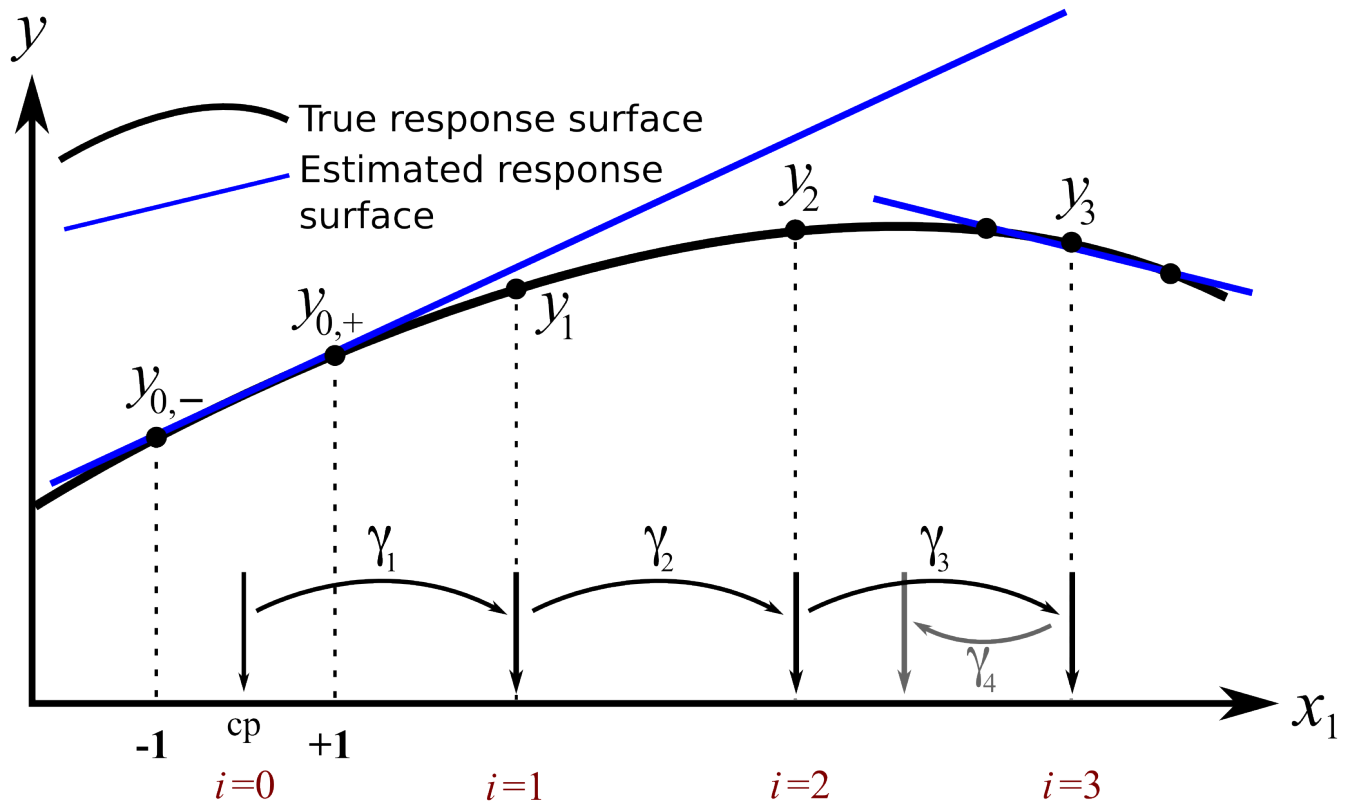


Figure 9.3.1.1 Plot demonstrating the response surface methods with a single factor.



We start at the point marked  $\mathbf{i} = \mathbf{0}$  as our initial baseline (cp=center point). Then, we run a 2-level experiment, above and below this baseline at -1 and 1 and obtain the corresponding response values of  $\mathbf{y}_{0,-}$  and  $\mathbf{y}_{0,+}$ . From this we can estimate a line of best fit and move in the direction that increases  $\mathbf{y}$ . Note that the sloping tangential line is also called the path of steepest ascent. Make a move of step-size =  $\gamma_1$  units along  $\mathbf{x}_1$  and measure the response,  $\mathbf{y}_1$ . Since the response variable increased, we keep going in this direction.

Make another step-size, this time of  $\gamma_2$  units in the direction that increases  $\mathbf{y}$ . We measure the response,  $\mathbf{y}_2$ , and are still increasing. Encouraged by this, we take another step of size  $\gamma_3$ . The step-sizes,  $\gamma_i$  should be of a size that is big enough to cause a change in the response in a reasonable number of experiments, but not so big as to miss an optimum.

Our next value of  $\mathbf{y}_3$  is about the same size as  $\mathbf{y}_2$ , indicating that we have plateaued. At this point we can take some exploratory steps and refit the tangential line (which now has a slope in the opposite direction). Or we can use the accumulated datapoints to fit a non-linear curve. Either way, we can then estimate a different step-size  $\gamma_4$  that will bring us closer to the optimum.

This approach works well when there is only a single factor that affects the response. However, in most systems there are multiple factors that affect the response, we need to adapt this method to find optimums for those systems.

### 9.3.2. OPTIMIZATION OF A 2-VARIABLE SYSTEM

Let's say we are looking to optimize a bioreactor where two factors, temperature  $\mathbf{T}$ , and substrate concentration  $\mathbf{S}$  are known to affect the yield. However, our outcome of interest is actually total profit which takes into account energy costs, raw materials costs and other relevant factors. Figure 9.3.2.1 shows (hypothetical) contours of profit in light grey, but in practice these are often unknown. We currently operate at these baseline conditions:

- $\mathbf{T} = 325 \text{ K}$
- $\mathbf{S} = 0.75 \text{ g/L}$
- **Profit** = \$407 per day

We create a full factorial around this baseline by choosing  $\Delta_{\mathbf{T}} = 10\text{K}$ , and  $\Delta_{\mathbf{S}} = 0.5\text{g/L}$  based on our knowledge that these are sufficiently large changes to show an actual difference in the response value (see Table 9.3.2.1), but not too large so as to move to a totally different form of operation in the bioreactor.

Table 11.2.1 Bioreactor Experiment Design of Experiments

Experiment	T (actual)	S (actual)	T (coded)	S (coded)	Profit
Baseline	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

It is evident that we can maximize profit by operating at higher temperatures and higher substrate

concentrations. The only way, however, to know how much higher is to build a linear model of the system from the factorial data:

$$\hat{y} = b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S$$

$$\hat{y} = 389.8 + 55x_T + 134x_S - 3.50x_T x_S$$

where  $x_T = \frac{x_{T,\text{actual}} - \text{center}_T}{\Delta_T/2}$

$$= \frac{x_{T,\text{actual}} - 325}{5}$$

and similarly,  $x_S = \frac{x_{S,\text{actual}} - 0.75}{0.25}$ .

The model shows that we can expect an increase of \$55/day of profit for a unit increase in T. In real-world units that would require increasing temperature by  $\Delta x_{T,\text{actual}} = 1 \times \Delta_T/2 = 5K$  to achieve that goal. This scaling factor comes from the coding we used:

$$\frac{x_{T,\text{actual}} - \text{center}_T}{\Delta_T/2}$$

Similarly, we can increase (S) by  $\Delta x_{S,\text{actual}} = 1 \times \Delta_S/2 = 0.25g/L$  to achieve a \$134 per day profit increase.

The interaction term is small, indicating that the response surface is mostly linear in this region. Figure 9.3.2.1 shows the model's contours (straight, green lines). Notice that the model contours are a good approximation to the actual contours (dotted, light grey), which are unknown in practice.

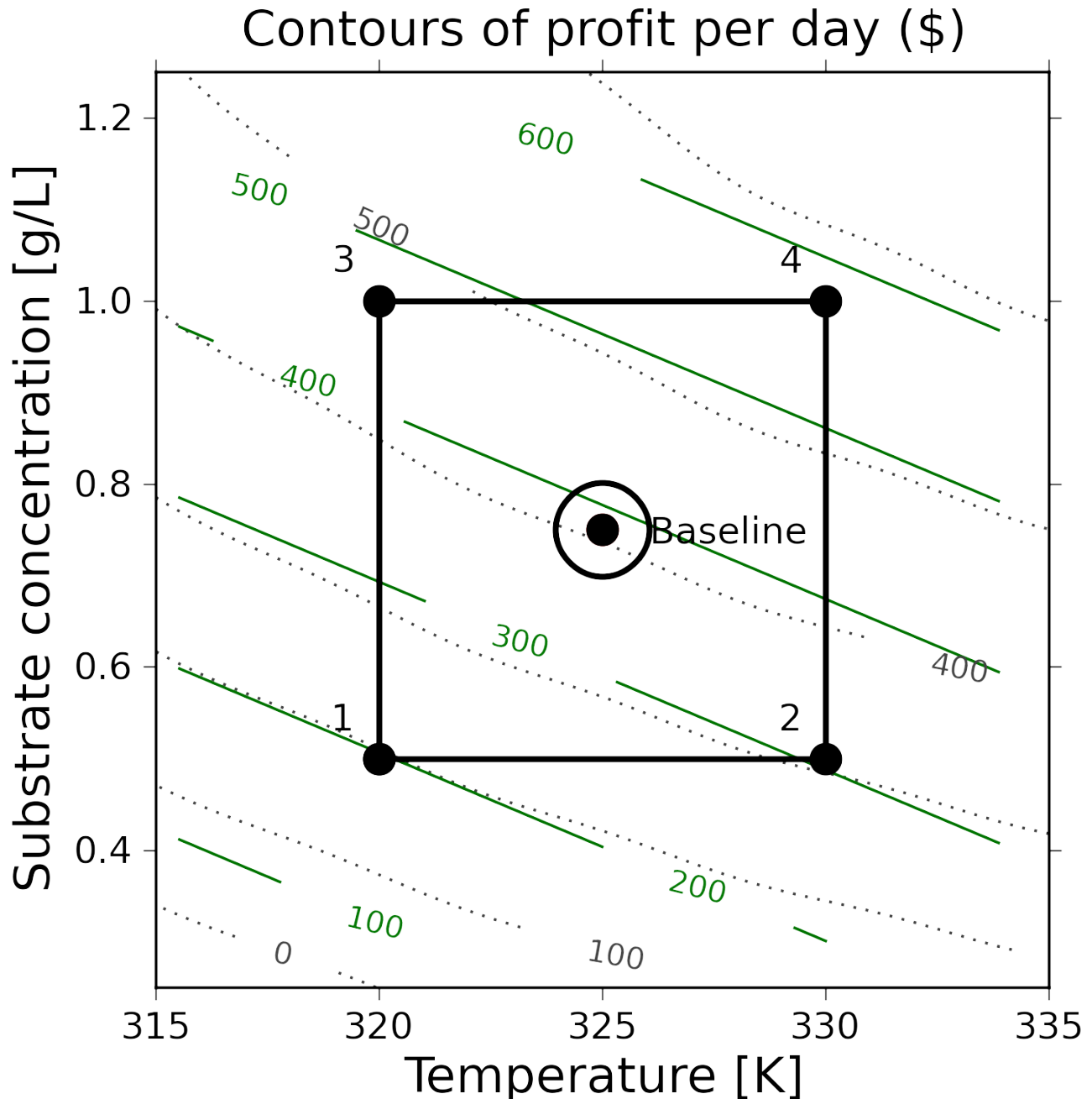


Figure 9.3.2.1. First factorial experiment for bioreactor example.

To improve our profit in the optimal way we move along our estimated model's surface, in the direction of steepest ascent. This direction is found by taking partial derivatives of the model function, but ignoring the interaction term since it is so small.

$$\left[ \frac{\partial \hat{y}}{\partial x_T} = b_T = 55 \quad \quad \quad \frac{\partial \hat{y}}{\partial x_S} = b_S = 134 \right]$$

This means that for every  $b_T = 55$  coded units that we move by in  $x_T$  we should also move  $x_S$  by  $b_S = 134$  coded units. Mathematically:

$$\left[ \frac{\Delta x_S}{\Delta x_T} = \frac{134}{55} \right]$$

The simplest way to do this is just to pick a movement size for one of the variables, then change the movement size of the other appropriately.



<https://ecampusontario.pressbooks.pub/app/uploads/sites/4023/2024/02/RSM-base-case-combined.png> alt="Figure 11.2.2. Illustration of Response Surface Methods of Bioreactor Example. From the original factorial data, a path of steepest ascent is established. Once an optimum is reached, an additional factorial is done to determine the next path of steepest ascent." width="3300" height="2400">

**Figure 9.3.2.2.** Illustration of Response Surface Methods of bioreactor example. From the original factorial data, a path of steepest ascent is established. Once an optimum is reached, an additional factorial is done to determine the next path of steepest ascent.

A least squares model from the 4 factorial points (experiments 8, 9, 10, and 11 run in random order), seem to show that the promising direction now is to increase temperature but decrease the substrate concentration.

$$\hat{y} = b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S$$

$$= 673.8 + 13.25 x_T - 39.25 x_S - 2.25 x_T x_S$$

As before we take a step in the direction of steepest ascent by  $b_T$  units along the  $x_T$  direction and  $b_S$  units along the  $x_S$  direction. Again we choose  $\Delta x_T = 1$  unit, though we must emphasize that we could use a smaller or larger amount, if desired. Therefore:

$$\frac{\Delta x_S}{\Delta x_T} = \frac{-39}{13}$$

$$\Delta x_S = \frac{-39}{13} \times 1$$

$$\Delta x_{S, \text{actual}} = \frac{-39}{13} \times 1 \times 0.4 / 2 = -0.6 \text{ g/L}$$

$$\Delta x_{T, \text{actual}} = 4 \text{ K}$$

This gives us the following conditions for run 12:

- $T_{12} = T_6 + \Delta x_{T, \text{actual}} = 335 + 4 = 339 \text{ K}$
- $S_{12} = S_6 + \Delta x_{S, \text{actual}} = 1.97 - 0.6 = 1.37 \text{ g/L}$

We determine that at run 12 the profit is \$ 716. But our previous factorial had a profit value of \$725 on one of the corners. Now it could be that we have a noisy system; after all, the difference between \$716 and \$725 is not too much, but there is a relatively large difference in profit between the other points in the factorial.

Some considerations when you are approaching an optimum:

- The response variable will start to plateau (remember that the first derivative is zero at an optimum)
- If the response variable remains roughly constant for two consecutive jumps (you may have bypassed the optimum)
- The response variable can decrease, sometimes very rapidly, if you overshoot the optimum
- The presence of curvature can also be inferred when interaction terms are similar or larger in magnitude than the main effect terms

This means that an optimum will exhibit some form of curvature. Thus, a model that only has linear terms will be unable to find the direction of steepest ascent along the *true response surface*. We must add terms that account for this curvature.

### 9.3.3. CHECKING FOR CURVATURE

When the measured center point is quite different from the predicted center point in your linear model, that

is a signal that there is curvature present. This can be accommodated for by adding polynomial terms to the model.

The factorial's center point can be predicted from  $(x_T, x_S) = (0, 0)$ , and is just the intercept term. In the last factorial, the predicted center point was  $\hat{y}_{cp} = \$670$ . Yet the actual center point from run 6 showed a profit of \$688. This difference of \$18 is substantial, especially when compared to the main effects' coefficients.

### 9.3.4. CENTRAL COMPOSITE DESIGNS

It is beyond the scope of this pressbook to go into detail about central composite designs. However, this section will show you what they look like for the case of 2 and 3 variables, taking an existing orthogonal factorial and augmenting it with axial points. Conveniently, these points can be added later as well to account for nonlinearity.

The axial points are placed  $4^{0.25} = 1.4$  coded units away from the center for a 2 factor system, and  $8^{0.25} = 1.7$  units away for a three factor system.

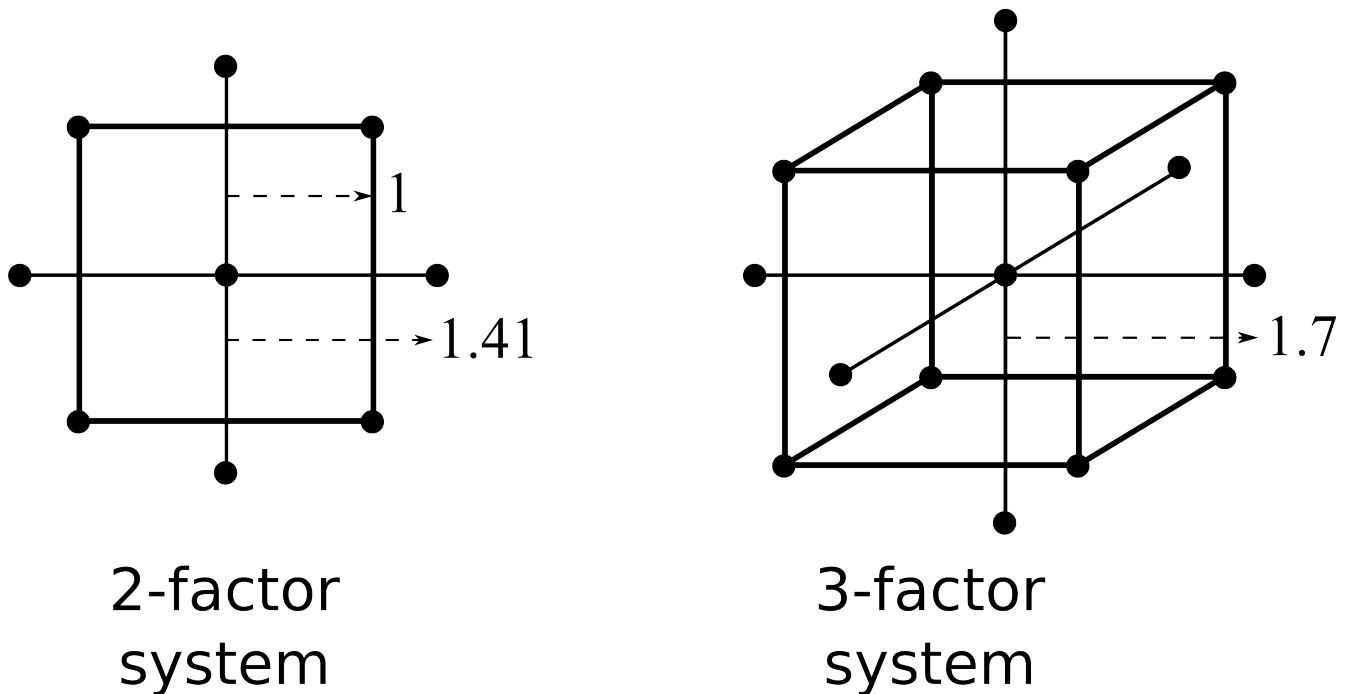


Figure 9.3.4.1. Illustration of central composite design for 2 and 3 factor systems. Axial points are placed at 1.4 and 1.7 units away from the center of 2 and 3 factor systems respectively.

A central composite design layout was added to the factorial in the above example and the experiments run, randomly, at the 4 axial points.

The four response values were  $y_{13} = 720$ ,  $y_{14} = 699$ ,  $y_{15} = 610$ , and  $y_{16} = 663$ . This allows us to estimate a model with quadratic terms in it:  $y = b_0 + b_T x_T + b_S x_S + b_{TS} x_T x_S + b_{TT} x_T^2 + b_{SS} x_S^2$ . The parameters in this model are found in the usual way, using a least-squares model:

$$y = 688 + 13x_T - 39x_S - 2.4x_Tx_S - 4.2x_T^2 - 12.2x_S^2$$

Notice how the linear terms estimated previously are the same! The quadratic effects are quite significant compared to the other effects, which was what prevented us from successfully using a linear model to project out to point 12 previously.

The final step in the response surface methodology is to plot this model's contour plot and predict where to run the next few experiments. As the solid contour lines in the illustration show, we should run our next experiments roughly at  $T = 343\text{K}$  and  $S = 1.60\text{ g/L}$  where the expected profit is around \$736. You can determine this approximately with your eyes or analytically. This is not exactly where the true process optimum is, but it is pretty close to it.

This example has demonstrated how powerful response surface methods are. A minimal number of experiments quickly converged towards the true, unknown process optimum. We achieved this by building successive least squares models that approximate the underlying surface. Those least squares models are built using the tools of fractional and full factorials, as well as basic optimization theory, to climb the hill of steepest ascent.

## 9.3.2 Design of Experiments: The General Approach

### 9.3.4 .THE GENERAL APPROACH FOR RESPONSE SURFACE MODELLING

---

1. Start at your baseline conditions and identify the main factors based on the process, expert opinion input, and intuition. Perform factorial experiments (full or fractional factorials), completely randomized. Use the results from the experiment to estimate a linear model of the system:  

$$\hat{y} = b_0 + b_A x_A + b_B x_B + b_C x_C \dots + b_{AB} x_A x_B + b_{AC} x_A x_C + \dots$$
2. The main effects are usually significantly larger than the two-factor interactions, so these higher interaction terms can be safely ignored. Any main effects that are not significant may be dropped for future iterations. Consider what was discussed in 10.3.

3. Use the model to estimate the path of steepest ascent (or descent if minimizing):  

$$\frac{\partial \hat{y}}{\partial x_1} = b_1 \quad \frac{\partial \hat{y}}{\partial x_2} = b_2 \quad \dots$$

The path of steepest ascent is climbed. Move any one of the main effects, e.g.  $(b_A)$  by a certain amount,  $\Delta x_A$ . Then move the other effects:  $\Delta x_i = \frac{b_i}{b_A} \Delta x_A$ . For example,  $\Delta x_C$  is moved by

$$\frac{b_C}{b_A} \Delta x_A.$$

If any of the  $\Delta x_i$  values are too large to safely implement, take a smaller proportional step in all factors. Recall that these are coded units, so unscale them to obtain the move amount in real-world units.

4. One can make several sequential steps until the response starts to level off, or if you become certain you have entered a different operating mode of the process.
5. At this point you repeat the factorial experiment from step 1, making the last best response value your new baseline. This is also a good point to reintroduce factors that you may have omitted earlier. Also, if you have a binary factor, investigate the effect of alternating its sign at this point. These additional factorial experiments should also include center points.
6. Repeat steps 1 through 5 until the linear model estimate starts to show evidence of curvature, or that the interaction terms start to dominate the main effects. This indicates that you are reaching an optimum.
  - Curvature can be assessed by comparing the predicted center point, i.e. the model's intercept =  $b_0$ , against the actual center point response(s). A large difference in the prediction, when compared to the model's effects, indicates the response surface is curved.



7. If there is curvature, add axial points to expand the factorial into a central composite design. Now estimate a quadratic model of the form:  
$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{12}x_1x_2 + \dots + b_{11}x_1^2 + b_{22}x_2^2 + \dots$$
8. Draw contour plots of this estimated response surface and determine where to place your sequential experiments. You can also find the model's optimum analytically by taking derivatives of the model function.

## SUMMARY

---

1. In the previous sections we used factorials and fractional factorials for screening the important factors. When we move to process optimization, we are assuming that we have already identified the important variables. In fact, we might find that variables that were previously important, appear unimportant as we approach the optimum. Conversely, variables that might have been dropped out earlier, become important at the optimum.
2. Response surface methods generally work best when the variables we adjust are numerically continuous. Categorical variables (yes/no, catalyst A or B) are handled by fixing them at one value, or the other, and then performing the optimization conditional on those selected values. It is always worth investigating the alternative values once the optimum has been reached.

## 9.4.1 Design of Experiments Project

At this point you should feel comfortable attempting your own design of experiments project! This will draw on everything that this pressbook has covered, from hypothesis testing to regression to design of experiments.

**It is strongly recommended that you consult the [Design of Experiments Jupyter Notebook Files](#).** These can be found in the “How do I do X in Python?” section. Specifically the files on “Full Factorial Example” and “Standard Error & Replicates” will be particularly useful.

### Design of Experiments Project

This DOE (design of experiments) mini project gives you an opportunity to learn about designed experiments in a more hands-on manner.

The project is *not long* and should *not be elaborate*. You only have a few weeks to plan your experiments, perform them and then analyze the data. Some examples are given below, but you are free to choose any topic like optimizing a favourite recipe or dessert, a hobby or sport.

The intention is that you discover for yourself how important the following topics are in DOE. Once you have decided on a system to investigate you will be faced with questions such as:

- Which variables should we use?
- What range should these variables cover?
- How do we measure these variables (especially the response/y variable)?
- What other variability is in the system, is it measurable, and is it controllable?
- Choosing the type of experimental design (full factorial, fractional factorial), confounding pattern, and handling constraints.
- How many experiments should be run, are replicates and/or center points possible, and how to randomize the runs.

These are issues that are not easily reproduced or understood from assignment questions and exams.

### Project Topic

You might be passionate about a hobby, or cooking, or sports, or a research area, etc., so coming up with a system to investigate shouldn't be a problem. However, some systems are too complex for the short time you have available, and you might have to cut back to something simpler. Below are some ideas that you can think about (and modify), but please work on anything you are interested in, or anything you have ever wondered about. Don't pick a project only because it “looks easy”, pick one that you have good ideas for a strong experimental setup.

Example topics:

- Yield of stovetop/microwave popcorn
- Rise height of bread
- Fuel efficiency/gas milage of a car
- Flight time of a paper plane
- Plant growth
- Bounce height of a ball
- Distance you can kick a ball
- Towel absorbency
- Burst time of soap bubbles

Regardless of which topic you choose there are some general guidelines you should follow:

- The experiment should be reproducible/repeatable
- Avoid time-based effects – e.g., learning a language using different methods; can't "unlearn" what you previous have learned
- Objective should be quantifiable – avoid subjective outcomes such as 'taste'
- Ideally it should include both numeric and categorical factors, but this depends on the experiment

*Table A1.1 Table of Standard Normal Probabilities*

## STANDARD NORMAL CUMULATIVE PROBABILITIES

---

## Standard Normal Cumulative Probabilities

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt$$

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641







*Table A1.2. Upper Tail Standard Normal Probabilities*



*Table A1.3. t Distribution Quantiles Table*

*t* Distribution Quantiles

$\nu$	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
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
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.088	3.281

*Table A1.4 Chi-Square Distribution Quantiles*

Chi-Square Distribution Quantiles

$\nu$	$Q(.005)$	$Q(.01)$	$Q(.025)$	$Q(.05)$	$Q(.1)$	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$
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

## *Table A1.5 F Distribution Tables*

F Distribution .75 Quantiles

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (Numerator Degrees of Freedom)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
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	1.89	1.88	1.88	1.88	1.88	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
$\infty$	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



## F Distribution .90 Quantiles

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (Numerator Degrees of Freedom)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
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
$\infty$	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

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (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
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

F Distribution of .95 Quantiles (continued)

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (Numerator Degrees of Freedom)								
	12	15	20	24	30	40	60	120	$\infty$
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
$\infty$	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

F Distribution .99 Quantiles

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (Numerator Degrees of Freedom)									
	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
$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

F Distribution of .99 Quantiles (continued)

$\nu_2$ (Denominator Degrees of Freedom)	$\nu_1$ (Numerator Degrees of Freedom)								
	12	15	20	24	30	40	60	120	$\infty$
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
$\infty$	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

*Table A1.6 Critical values of the smallest rank sum for the Wilcoxon-Mann-Whitney test*

**Table Critical values of the smallest rank sum for the Wilcoxon-Mann-Whitney test** $n_1$  = number of elements in the largest sample; $n_2$  = number of elements in the smallest sample.

		Level of significance $\alpha$						Level of significance $\alpha$					
		Two-sided	0.20	0.10	0.05	0.01			Two-sided	0.20	0.10	0.05	0.01
		One-sided	0.10	0.05	0.025	0.005			One-sided	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		
10	2	6	4	3	-	13	13	149	142	136	125		
10	3	12	10	9	6								
10	4	20	17	15	12								
10	5	28	26	23	19								

Level of significance $\alpha$						Level of significance $\alpha$					
Two-sided		0.20	0.10	0.05	0.01	Two-sided		0.20	0.10	0.05	0.01
One-sided		0.10	0.05	0.025	0.005	One-sided		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	103	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						





*Table A1.7 Critical Values of the Wilcoxon Signed Ranks Test*

## Critical Values of the Wilcoxon Signed Ranks Test

n	Two-Tailed Test		One-Tailed Test	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
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
27	107	83	119	92
28	116	91	130	101
29	126	100	140	110
30	137	109	151	120

*Table A1.8 Critical Values of the Mann-Whitney U*

**Critical Values of the Mann-Whitney U**  
(Two-Tailed Testing)

n <sub>2</sub>	α	n <sub>1</sub>																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

**Critical Values of the Mann-Whitney U  
(One-Tailed Testing)**

n <sub>2</sub>	α	n <sub>1</sub>																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	0	0	1	2	2	3	4	4	5	5	6	7	7	8	9	9	10	11
	.01	--	0	0	0	0	0	1	1	1	2	2	2	3	3	4	4	4	5
4	.05	0	1	2	3	4	5	6	7	8	9	10	11	12	14	15	16	17	18
	.01	--	--	0	1	1	2	3	3	4	5	5	6	7	7	8	9	9	10
5	.05	1	2	4	5	6	8	9	11	12	13	15	16	18	19	20	22	23	25
	.01	--	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
6	.05	2	3	5	7	8	10	12	14	16	17	19	21	23	25	26	28	30	32
	.01	--	1	2	3	4	6	7	8	9	11	12	13	15	16	18	19	20	22
7	.05	2	4	6	8	11	13	15	17	19	21	24	26	28	30	33	35	37	39
	.01	0	1	3	4	6	7	9	11	12	14	16	17	19	21	23	24	26	28
8	.05	3	5	8	10	13	15	18	20	23	26	28	31	33	36	39	41	44	47
	.01	0	2	4	6	7	9	11	13	15	17	20	22	24	26	28	30	32	34
9	.05	4	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54
	.01	1	3	5	7	9	11	14	16	18	21	23	26	28	31	33	36	38	40
10	.05	4	7	11	14	17	20	24	27	31	34	37	41	44	48	51	55	58	62
	.01	1	3	6	8	11	13	16	19	22	24	27	30	33	36	38	41	44	47
11	.05	5	8	12	16	19	23	27	31	34	38	42	46	50	54	57	61	65	69
	.01	1	4	7	9	12	15	18	22	25	28	31	34	37	41	44	47	50	53
12	.05	5	9	13	17	21	26	30	34	38	42	47	51	55	60	64	68	72	77
	.01	2	5	8	11	14	17	21	24	28	31	35	38	42	46	49	53	56	60
13	.05	6	10	15	19	24	28	33	37	42	47	51	56	61	65	70	75	80	84
	.01	2	5	9	12	16	20	23	27	31	35	39	43	47	51	55	59	63	67
14	.05	7	11	16	21	26	31	36	41	46	51	56	61	66	71	77	82	87	92
	.01	2	6	10	13	17	22	26	30	34	38	43	47	51	56	60	65	69	73
15	.05	7	12	18	23	28	33	39	44	50	55	61	66	72	77	83	88	94	100
	.01	3	7	11	15	19	24	28	33	37	42	47	51	56	61	66	70	75	80
16	.05	8	14	19	25	30	36	42	48	54	60	65	71	77	83	89	95	101	107
	.01	3	7	12	16	21	26	31	36	41	46	51	56	61	66	71	76	82	87
17	.05	9	15	20	26	33	39	45	51	57	64	70	77	83	89	96	102	109	115
	.01	4	8	13	18	23	28	33	38	44	49	55	60	66	71	77	82	88	93
18	.05	9	16	22	28	35	41	48	55	61	68	75	82	88	95	102	109	116	123
	.01	4	9	14	19	24	30	36	41	47	53	59	65	70	76	82	88	94	100
19	.05	10	17	23	30	37	44	51	58	65	72	80	87	94	101	109	116	123	130
	.01	4	9	15	20	26	32	38	44	50	56	63	69	75	82	88	94	101	107
20	.05	11	18	25	32	39	47	54	62	69	77	84	92	100	107	115	123	130	138
	.01	5	10	16	22	28	34	40	47	53	60	67	73	80	87	93	100	107	114

