



Core Concepts in

Mathematics: Statistics

Diego Fernandez

2ND EDITION

 **BIBLIOTEX**
Digital Library

CORE CONCEPTS IN MATHEMATICS: **STATISTICS**

Diego Fernandez

2nd Edition



www.bibliotex.com

CORE CONCEPTS IN MATHEMATICS: STATISTICS

2nd Edition

Diego Fernandez



www.bibliotex.com

email: info@bibliotex.com

e-book Edition 2024

ISBN: 978-1-98469-082-1 (e-book)

This book contains information obtained from highly regarded resources. Reprinted material sources are indicated. Copyright for individual articles remains with the authors as indicated and published under Creative Commons License. A Wide variety of references are listed. Reasonable efforts have been made to publish reliable data and views articulated in the chapters are those of the individual contributors, and not necessarily those of the editors or publishers. Editors or publishers are not responsible for the accuracy of the information in the published chapters or consequences of their use. The publisher assumes no responsibility for any damage or grievance to the persons or property arising out of the use of any materials, instructions, methods or thoughts in the book. The editors and the publisher have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission has not been obtained. If any copyright holder has not been acknowledged, please write to us so we may rectify.

Notice: Registered trademark of products or corporate names are used only for explanation and identification without intent of infringement.

© 2024 3G E-learning LLC

In Collaboration with 3G E-Learning LLC. Originally Published in printed book format by 3G E-Learning LLC with ISBN 978-1-98468-865-1

EDITORIAL BOARD



Aleksandar Mratinković was born on May 5, 1988. in Arandjelovac, Serbia. He has graduated on Economic high school (2007), The College of Tourism in Belgrade (2013), and also has a master degree of Psychology (Faculty of Philosophy, University of Novi Sad). He has been engaged in different fields of psychology (Developmental Psychology, Clinical Psychology, Educational Psychology and Industrial Psychology) and has published several scientific works



Dan Piestun (PhD) is currently a startup entrepreneur in Israel working on the interface of Agriculture and Biomedical Sciences and was formerly president-CEO of the National Institute of Agricultural Research (INIA) in Uruguay. Dan is a widely published scientist who has received many honours during his career including being a two-time recipient of the Amit Golda Meir Prize from the Hebrew University of Jerusalem, his areas of expertise includes stem cell molecular biology, plant and animal genetics and bioinformatics. Dan's passion for applied science and technological solutions did not stop him from pursuing a deep connection to the farmer, his family and nature. Among some of his interest and practices counts enjoying working as a beekeeper and onboard fishing.



Felecia Killings is the Founder and CEO of LiyahAmore Publishing, a publishing company committed to providing technical and educational services and products to Christian Authors. She operates as the Senior Editor and Writer, the Senior Writing Coach, the Content Marketing Specialist, Editor-in-Chief to the company's quarterly magazine, the Executive and Host of an international virtual network, and the Executive Director of the company's online school for Authors. She is a former high-school English instructor and professional development professor. She possesses a Master of Arts degree in Education and a Bachelor's degree in English and African American studies.



Dr. Sandra El Hajj, Ph.D. in Health Sciences from Nova Southeastern University, Florida, USA is a health professional specialized in Preventive and Global Health. With her 12 years of education obtained from one of the most prominent universities in Beirut, in addition to two leading universities in the State of Florida (USA), Dr. Sandra made sure to incorporate interdisciplinary and multicultural approaches in her work. Her long years of studies helped her create her own miniature world of knowledge linking together the healthcare field with Medical Research, Statistics, Food Technology, Environmental & Occupational Health, Preventive Health and most noteworthy her precious last degree of Global Health. Till today, she is the first and only doctor specialized in Global Health in the Middle East area.



Hazem Shawky Fouda has a PhD. in Agriculture Sciences, obtained his PhD. From the Faculty of Agriculture, Alexandria University in 2008, He is working in Cotton Arbitration & Testing General Organization (CATGO).



Fozia Parveen has a Dphil in Sustainable Water Engineering from the University of Oxford. Prior to this she has received MS in Environmental Sciences from National University of Science and Technology (NUST), Islamabad Pakistan and BS in Environmental Sciences from Fatima Jinnah Women University (FJWU), Rawalpindi.



Igor Kronic 2003-2007 in the School of Economics. After graduating in 2007, he went on to study at The College of Tourism, at the University of Belgrade where he got his bachelor degree in 2010. He was active as a third-year student representative in the student parliament. Then he went on the Faculty of science, at the University of Novi Sad where he successfully defended his master's thesis in 2013. The crown of his study was the work titled "Opportunities for development of cultural tourism in Cacak". Later on, he became part of a multinational company where he got promoted to a deputy director of logistic. Nowadays he is a consultant and writer of academic subjects in the field of tourism.



Dr. Jovan Pehceviski obtained his PhD in Computer Science from RMIT University in Melbourne, Australia in 2007. His research interests include big data, business intelligence and predictive analytics, data and information science, information retrieval, XML, web services and service-oriented architectures, and relational and NoSQL database systems. He has published over 30 journal and conference papers and he also serves as a journal and conference reviewer. He is currently working as a Dean and Associate Professor at European University in Skopje, Macedonia.



Dr. Tanjina Nur finished her PhD in Civil and Environmental Engineering in 2014 from University of Technology Sydney (UTS). Now she is working as Post-Doctoral Researcher in the Centre for Technology in Water and Wastewater (CTWW) and published about eight International journal papers with 80 citations. Her research interest is wastewater treatment technology using adsorption process.



Stephen obtained his PhD from the University of North Carolina at Charlotte in 2013 where his graduate research focused on cancer immunology and the tumor microenvironment. He received postdoctoral training in regenerative and translational medicine, specifically gastrointestinal tissue engineering, at the Wake Forest Institute of Regenerative Medicine. Currently, Stephen is an instructor for anatomy and physiology and biology at Forsyth Technical Community College.



Michelle holds a Masters of Business Administration from the University of Phoenix, with a concentration in Human Resources Management. She is a professional author and has had numerous articles published in the Henry County Times and has written and revised several employee handbooks for various YMCA organizations throughout the United States.

Table of Contents

Preface

ix

1 Introduction to Statistics 1

Introduction.....	2
Core Concept: Descriptive Statistics:The First Step Towards Statistical Analysis	3
Core Concept: Exploratory Data Analysis	26
Summary	38
Multiple Choice Questions	39
Review Questions	40
References.....	41

2 Statistics and Correlation Formulae 43

Introduction.....	44
Core Concept: Statistics and its Formulae	44
Core Concept: Fundamental of Correlation.....	60
Core Concept: Simple Linear Regression	67
Summary	77
Multiple Choice Questions	78
Review Questions	79
References.....	80

3 Statistical Inference 81

Introduction.....	82
Core Concept: Point Estimation.....	82
Core Concept: Statistical Inference: One and Two Sample Problems.....	86
Core Concept: Kolmogorov-Smirnov Test	93
Core Concept: Discrete And Continuous Data.....	97

Core Concept: Frequency Distributions	103
Summary	110
Multiple Choice Questions	111
Review Questions	112
References.....	113

4 **Multivariate Analysis** **115**

Introduction.....	116
Core Concept: Statistics of Multivariate Analysis	116
Core Concept: Multivariate Normal Distribution.....	121
Core Concept: Relationship Between Variables, Multiple Regression	126
Core Concept: Discriminant Analysis.....	133
Core Concept: Mahalanobis Distance.....	146
Summary	155
Multiple Choice Questions	156
Review Questions	157
References.....	158

5 **Tests of Hypothesis** **159**

Introduction.....	160
Core Concept: Nature of Hypothesis	160
Core Concept: Tests of Hypotheses on the Population Mean	163
Core Concept: Types of Hypotheses	172
Core Concept: Hypothesis Testing and Sampling Distributions.....	184
Summary	191
Multiple Choice Questions	192
Review Questions	193
References.....	194

6 **Parameter Estimation** **195**

Introduction.....	196
Core Concept: Samples and Statistics	196

Core Concept: Quality Criteria for Estimates.....	201
Core Concept: Methods of Estimation	213
Summary	223
Multiple Choice Questions	224
Review Questions	225
References.....	226

7

Simple Linear Regression and Correlation

227

Introduction.....	228
Core Concept: Simple Linear Regression	229
Core Concept: The Model Behind Linear Regression.....	237
Core Concept: Interpreting Regression Coefficients	241
Core Concept: Robustness of Simple Linear Regression	243
Core Concept: Correlation.....	249
Summary	259
Multiple Choice Questions	260
Review Questions	261
References.....	262

Index

263

Preface

Statistics plays a vital role in every field of human activity due to its enormous importance in daily life. With applications in a vast number of different fields, statistics is a very broad subject. In general, it can be said that statistics is the methodology for collecting, analyzing, interpreting, and drawing conclusions from information. Every day, we encounter a variety of data-based evaluations and claims. Determining the existing position of per capita income, unemployment rates, population growth rates, housing availability, schooling facilities, medical resources - all these aspects are made possible by the knowledge of statistics and exploratory data analysis. These tools enable us to evaluate the truth value behind such claims and protect ourselves from manipulative or deceiving tricks. By leveraging statistical tools and methods, we can make sense of all the data and related information that we come across on a regular basis. Statistics serves as a crucial process not only in how we make discoveries in science but also how we make decisions based on data-driven insights and predictions.

The objective of the book *Core Concepts in Statistics* (2nd Edition) is to provide students with a comprehensive and accessible introduction to statistical concepts. Through clear explanations and examples, the book aims to enhance students' understanding of statistics and how it relates to real-world applications. It seeks to equip readers with essential skills for interpreting data, analyzing trends, and making informed decisions based on empirical evidence. The book strives to promote critical thinking by encouraging readers to evaluate statistical claims critically and differentiate between valid data analysis techniques from misleading ones. The book further strives to familiarize students with basic principles of study design, sampling methods, probability theory, hypothesis testing, confidence intervals, correlation analysis, regression modeling, and experimental design. With its user-friendly approach and practical exercises that reinforce key concepts, *Core Concepts in Statistics* assists learners in developing statistical literacy necessary for success in various fields where statistics play a crucial role.

INTRODUCTION TO STATISTICS



Statistics is the study of data: how to collect, organize, analyze, interpret, and present information in order to make informed decisions and draw meaningful conclusions. It plays a crucial role in various fields, including science, business, social sciences, economics, and more. Statistics enables us to understand patterns, relationships, and trends in data, which can help us make predictions and inform decisions.

After studying this chapter, you will understand the core concepts of:

- Descriptive Statistics: The First Step Towards Statistical Analysis
- Exploratory Data Analysis

INTRODUCTION

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory.

People who do statistics are referred to as statisticians. They're particularly concerned with determining how to draw reliable conclusions about large groups and general events from the behavior and other observable characteristics of small samples. These small samples represent a portion of the large group or a limited number of instances of a general phenomenon.

Statistics are used in virtually all scientific disciplines, such as the physical and social sciences as well as in business, the humanities, government, and manufacturing. Statistics is fundamentally a branch of applied mathematics that developed from the application of mathematical tools, including calculus and linear algebra to probability theory.

Statistics is the idea that we can learn about the properties of large sets of objects or events (a population) by studying the characteristics of a smaller number of similar objects or events (a sample). Gathering comprehensive data about an entire population is too costly, difficult, or impossible in many cases so statistics start with a sample that can be conveniently or affordably observed.

Statisticians measure and gather data about the individuals or elements of a sample, then they analyze this data to generate descriptive statistics. They can then use these observed characteristics of the sample data, which are properly called 'statistics, to make inferences or educated guesses about the unmeasured (or unmeasured) characteristics of the broader population, known as the parameters.





CORE CONCEPT: DESCRIPTIVE STATISTICS: THE FIRST STEP TOWARDS STATISTICAL ANALYSIS

Descriptive statistics is distinguished from inferential statistics, in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Statistics is a set of procedures for gathering, measuring, classifying, computing, describing, synthesizing, analyzing, and interpreting systematically acquired quantitative data. Statistics has major two components: the Descriptive Statistics and the Inferential Statistics. Descriptive Statistics gives numerical and graphic procedures to summarize a collection of data in a clear and understandable way whereas Inferential Statistics provides procedures to draw inferences about a **population** from a sample.

Population is the number of living things that live together in the same place.



Descriptive statistics are numbers that are used to summarize and describe data. The word “data” refers to the information that has been collected from an experiment, a survey, a historical record, etc. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand. These are ways of summarizing large sets of quantitative (numerical) information. If you have a large number of measurements, the best thing you can do is to make a graph with all the possible scores along the bottom (x axis), and the number of

```
graph LR; A[Descriptive Statistics] --> B[Measures of Central Tendency]; A --> C[Measures of Dispersion]; B --> D[Mean]; B --> E[Median]; B --> F[Mode]; C --> G[Range]; C --> H[InterQuartile Range]; C --> I[Standard Deviation]; C --> J[Variance];
```

Descriptive Statistics

- Measures of Central Tendency
 - Mean
 - Median
 - Mode
- Measures of Dispersion
 - Range
 - InterQuartile Range
 - Standard Deviation
 - Variance

The plots contain detailed information about the distribution. Graphical methods are better suited than numerical methods for identifying patterns in the data. Numerical approaches are more precise and objective. Since the numerical and

graphical approaches complement each other, it is wise to use both. There are three major characteristics of a single variable that we tend to look at:

- Distribution
- Central Tendency
- Dispersion

Distribution

The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of times each value occurs. One of the most common ways to describe a single variable is with a frequency distribution. Frequency distributions can be depicted in two ways, as a table or as a graph. Distributions may also be displayed using percentages. **Frequency distribution** organizes raw data or observations that have been collected into ungrouped data and grouped data. The Ungrouped Data provide listing of all possible scores that occur in a distribution and then indicating how often each score occurs. Grouped Data combines all possible scores into classes and then indicating how often each score occurs within each class. It is easier to see patterns in the data, but the information about individual scores is lost. Graphs make it easier to see certain characteristics and trends in a set of data. Graphs for quantitative data include Histogram, Frequency Polygon etc. and graphs for qualitative data include **Bar Chart**, Pie Chart etc.

Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.

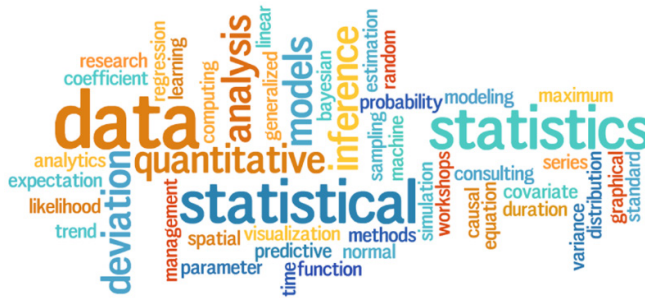


Bar chart is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent.

Types of Data

A variate or random variable is a quantity or attribute whose value may vary from one unit of investigation to another. For example, the units might be headache sufferers and the variate

might be the time between taking an aspirin and the headache ceasing. An observation or response is the value taken by a variate for some given unit.



There are various types of variate.

- Qualitative or nominal; described by a word or phrase (e.g. blood group, color).
- Quantitative; described by a number (e.g. time till cure, number of calls arriving at a telephone exchange in 5 seconds).
- Ordinal; this is an “in-between” case. Observations are not numbers but they can be ordered (e.g. much improved, improved, same, worse, much worse).

Averages etc. can sensibly be evaluated for quantitative data, but not for the other two. Qualitative data can be analyzed by considering the frequencies of different categories. Ordinal data can be analyzed like qualitative data, but really requires special techniques called nonparametric methods.

Quantitative data can be:

- Discrete: the variate can only take one of a finite or countable number of values (e.g. a count)
- Continuous: the variate is a measurement which can take any value in an interval of the real line (e.g. a weight)

Displaying Data

It is nearly always useful to use graphical methods to illustrate your data. We shall describe in this section just a few of the methods available.



Frequency table and bar chart

Suppose that you have collected some discrete data. It will be difficult to get a “feel” for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.

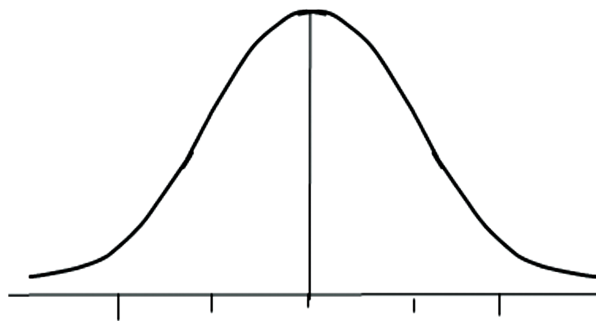
- The frequency of a value is the number of observations taking that value.
- A frequency table is a list of possible values and their frequencies.
- A bar chart consists of bars corresponding to each of the possible values, whose heights are equal to the frequencies.

Continuous data: histograms

When the variate is continuous, we do not look at the frequency of each value, but group the values into intervals. The plot of frequency against interval is called a histogram. Be careful to define the interval boundaries unambiguously.

Shape of the Distribution

An important aspect of the “description” of a variable is the shape of its distribution, which tells the frequency of values from different ranges of the variable. Typically, a researcher is interested in how well the distribution can be approximated by the normal distribution. Simple descriptive statistics can provide some information relevant to this issue. For example, if the skewness (which measures the deviation of the distribution from symmetry) is clearly different from 0, then that distribution is asymmetrical, while normal distributions are perfectly symmetrical. If the kurtosis (which measures the peakedness of the distribution) is clearly different from 0, then the distribution is either flatter or more peaked than normal; the kurtosis of the normal distribution is 0.



Normal Distribution

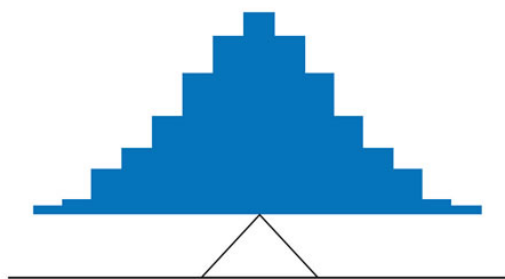
More precise information can be obtained by performing one of the tests of normality to determine the probability that the sample came from a normally distributed population of observations (e.g., the so-called Kolmogorov-Smirnov test, or the Shapiro-Wilks’ W test). However, none of these tests can entirely substitute for a visual examination of the data using a histogram (i.e., a graph that shows the frequency distribution of a variable).

Central Tendency

The central tendency of a distribution is an estimate of the “center” of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The Mean or average is probably the most commonly used method of describing central tendency. It is the most common measure of central tendency. To compute the mean, all the values are added up and divided by the number of values.



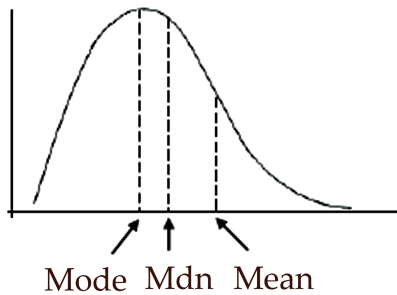
The Median is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score number 250 would be the median. Let the 8 scores be ordered as 15, 15, 15, 20, 20, 21, 25, and 36. Score number 4 and number 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

The Mode is the most frequently occurring value in the set of scores. To determine the mode, order the scores as shown above, and then count each one. The most frequently occurring value is the mode. It is used for either numerical or categorical data. In the above example, the value 15 occurs three times and is the mode. In some distributions, there may be more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently. Further, there may not be a mode. Mode is not affected by extreme value. If the yield of paddy from different fields are 6.0, 4.9, 6.0, 5.8, 6.2, 6.0, 6.3, 4.8, 6.0, 5.7 and 6.0 tonnes per hectare, the modal value is 6.0 tonnes per hectare.

For the same set of 8 scores, three different values, 20.875, 20, and 15 for the mean, median and mode respectively have been obtained. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other. While the mean

Statistics informally dates back centuries. An early record of correspondence between French mathematicians Pierre de Fermat and Blaise Pascal in 1654 is often cited as an early example of statistical probability analysis.

is the most frequently used measure of central tendency, it does suffer from one major drawback. Unlike other measures of central tendency, the mean can be influenced profoundly by one extreme data point (referred to as an “outlier”). The median and mode clearly do not suffer from this problem. There are certainly occasions where the mode or median might be appropriate. For qualitative and categorical data, the mode makes sense, but the mean and median do not. For example, when we are interested in knowing the typical soil type in a locality or the typical cropping pattern in a region we can use mode. On the other hand, if the data is quantitative one, we can use any one of the averages.



If the data is quantitative, then we have to consider the nature of the frequency distribution. When the frequency distribution is skewed (not symmetrical), the median or mode will be proper average. In case of raw data in which extreme values, either small or large, are present, the median or mode is the proper average. In case of a symmetrical distribution either mean or median or mode can be used. However, as seen already, the mean is preferred over the other two. When dealing with rates, speed and prices, use harmonic mean. If interest is in relative change, as in the case of bacterial growth, cell division etc., geometric mean is the most appropriate average. The mean, median, and mode can be related (approximately) to the histogram: the mode is the highest bump, the median is where half the area is to the right and half is to the left, and the mean is where the **histogram** would balance.

Measures of Central Tendency

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution.

Histogram

is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson.

Measures of Central Tendency

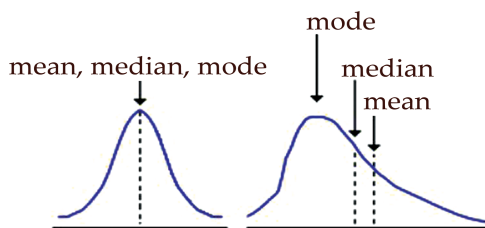


Figure 1: Measures of central tendency.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

Mean (Arithmetic)

The mean (or average) is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values $x_1, x_2, \dots, x_n$, the sample mean, usually denoted by $\bar{x}$ (pronounced x bar), is:

$$\bar{x} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capital letter, Σ , pronounced “sigma”, which means “sum of...”:

$$\bar{x} = \frac{\Sigma x}{n}$$

You may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the [sample mean](#), we use the Greek lower case letter “mu”, denoted as μ :

Sample Mean
refers to the mean value of a sample of data calculated from within a large population of data.

$$\mu = \frac{\sum x}{n}$$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set.

However, one of its important properties is that it minimizes error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set. An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

When not to use the mean

The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is \$30.7k. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the \$12k to 18k range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation.

Another time when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed). If we consider the normal distribution - as this is the most frequently assessed in statistics - when the data is perfectly normal, the mean, median and mode are identical. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values.

Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92
----	----	----	----	----	----	----	----	----	----	----

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89	92
----	----	----	----	----	-----------	----	----	----	----	----

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45
----	----	----	----	----	----	----	----	----	----

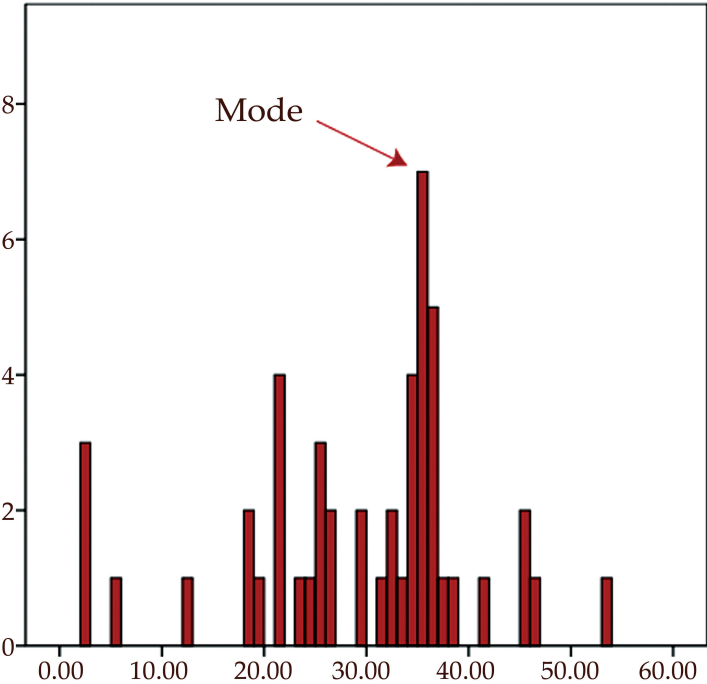
We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	56	56	65	87	89
----	----	----	----	-----------	-----------	----	----	----	----

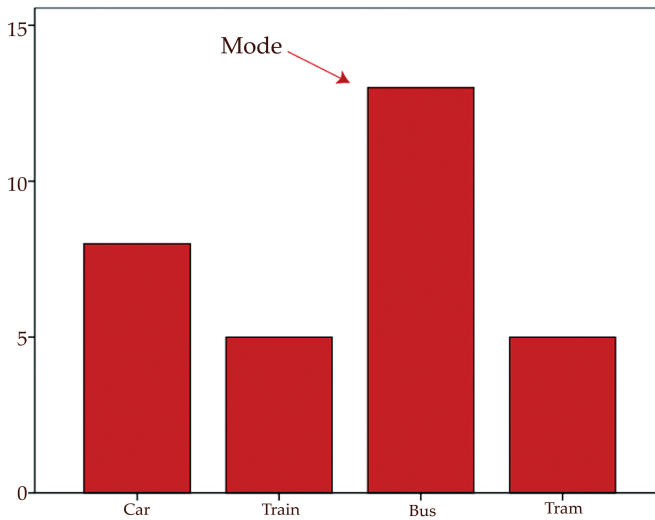
Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Mode

The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below:

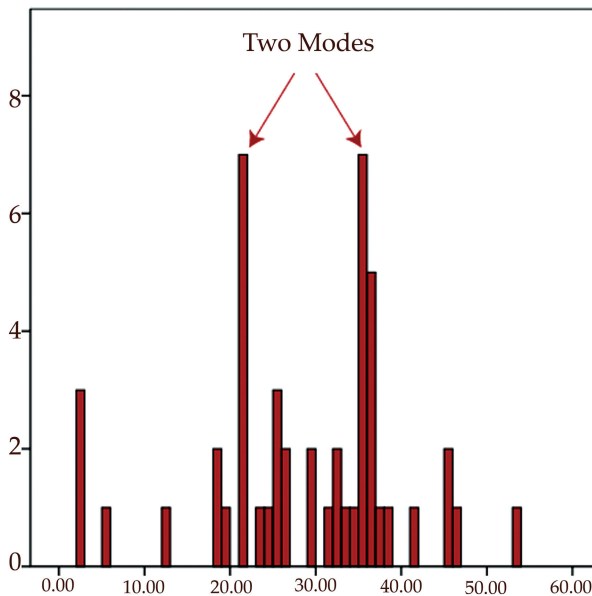


Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:



We can see above that the most common form of transport, in this particular data set, is the bus. However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest **frequency**, such as below:

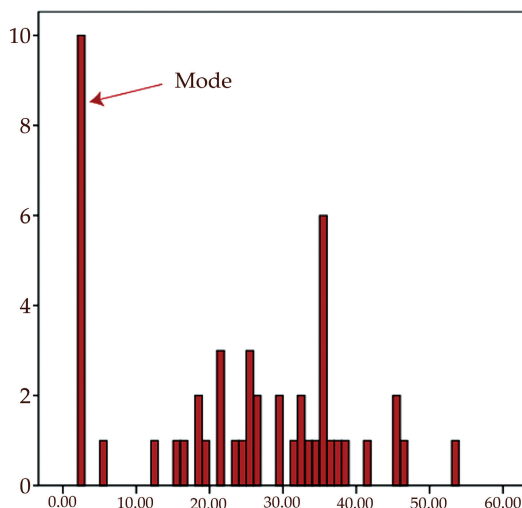
Frequency is the number of occurrences of a repeating event per unit of time.



We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with exactly the same weight (e.g., 67.4 kg)?

The answer, is probably very unlikely - many people might

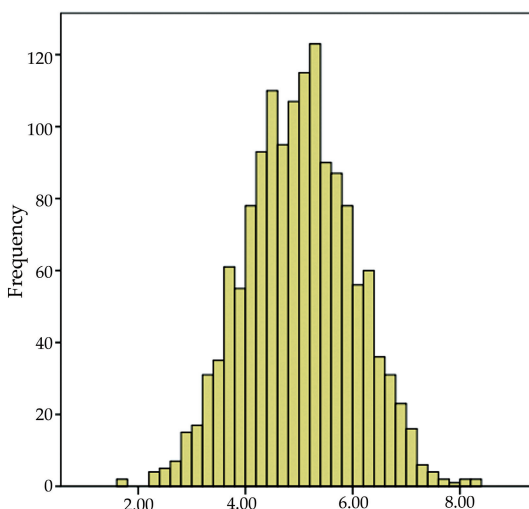
be close, but with such a small sample (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight; that is, to the nearest 0.1 kg. This is why the mode is very rarely used with continuous data. Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the diagram below:



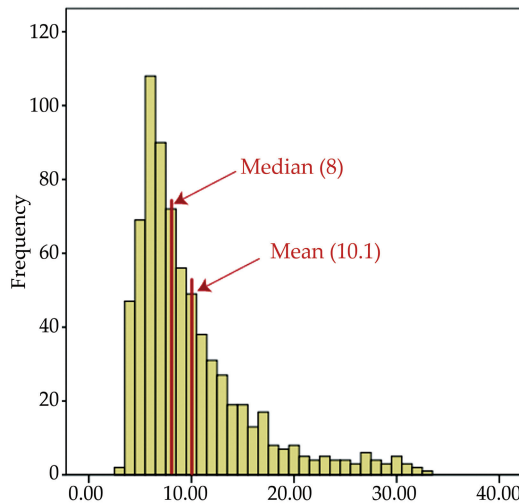
In the above diagram the mode has a value of 2. We can clearly see, however, that the mode is not representative of the data, which is mostly concentrated around the 20 to 30 value range. To use the mode to describe the central tendency of this data set would be misleading.

Skewed Distributions and the Mean and Median

We often test whether our data is normally distributed because this is a common assumption underlying many statistical tests. An example of a normally distributed set of data is presented below:



When you have a normally distributed sample you can legitimately use both the mean and the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of central tendency because it is the measure that includes all the values in the data set for its calculation, and any change in any of the scores will affect the value of the mean. This is not the case with the median or mode. However, when our data is skewed, for example, as with the right-skewed data set below:



We find that the mean is being dragged in the direction of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean. A classic example of the above right-skewed distribution is income (salary), where higher-earners provide a false representation of the typical income if expressed as a mean and not a median. If dealing with a normal distribution, and tests of normality show that the data is non-normal, it is customary to use the median instead of the mean. However, this is more a rule of thumb than a strict guideline.

If the data set depends on a sample of a larger population, then the analyst can develop interpretations about the population primarily based on the statistical outcomes from the sample.

Dispersion

Averages are representatives of a frequency distribution but they fail to give a complete picture of the distribution. They do not tell anything about the scatterness of observations within the distribution.

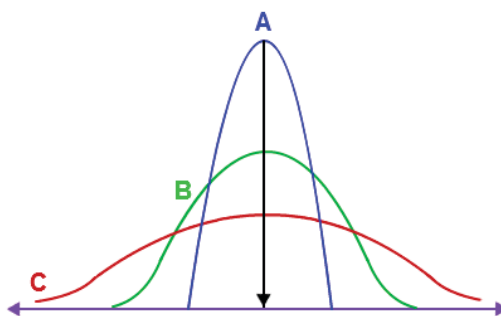


Figure 2: Dispersion.

Suppose that we have the distribution of the yields (kg per plot) of two paddy varieties from 5 plots each. The distribution may be as follows:

- Variety I: 45 42 42 41 40
- Variety II: 54 48 42 33 30

It can be seen that the mean yield for both varieties is 42 kg. But we cannot say that the performance of the two varieties are same. There is greater uniformity of yields in the first variety whereas there is more variability in the yields of the second variety. The first variety may be preferred since it is more consistent in yield performance. From the above example, it is obvious that a measure of central tendency alone is not sufficient to describe a frequency distribution. In addition to it we should have a measure of scatterness of observations. The scatterness or variation of observations from their average is called the dispersion. There are different measures of dispersion like the range, the quartile deviation, the mean deviation and the standard deviation.

The Range is simply the highest value minus the lowest value. The Standard Deviation (S.D) is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range. The Standard Deviation shows the relation that set of scores has to the mean of the sample. The standard deviation is the square root of the sum of the squared deviations from the mean divided by the number of scores minus one. Mean deviation is the mean of the deviations of individual values from their average. The average may be either mean or median. For raw data the mean deviation from the median is the least. Although we can calculate these **univariate statistics** by hand, it gets quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you.

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable.

Measures of Dispersion

The mean is the value usually used to indicate the center of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.

- **Coefficient of dispersion:** A “catch-all” term for a variety of formulas, including distance between quartiles.
- **Standard deviation:** probably the most common measure. It tells you how spread out numbers are from the mean,
- **Index of Dispersion:** a measure of dispersion commonly used with nominal variables.
- **Interquartile range (IQR):** describes where the bulk of the data lies (the “middle fifty” percent).
- **Interdecile range:** the difference between the first decile (10%) and the last decile (90%).
- **Range:** the difference between the smallest and largest number in a set of data.
- **Mean difference or difference in means:** measures the absolute difference between the mean values in two different groups in clinical trials.
- **Median absolute deviation (MAD):** the median of the absolute deviations from a data set’s median.
- **Quartiles:** Numbers that split the data into four quarters (first, second, third, and fourth quartiles).

In some processes, like manufacturing or measurement, low dispersion is associated with high precision. High dispersion is associated with low precision. We will consider several measures of dispersion and discuss the merits and pitfalls of each.

The Range

One very simple measure of **dispersion** is the range. Let’s consider the two distributions given in Figures 3 and 4. They represent the marks of a group of thirty students on two tests.

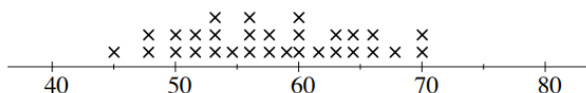


Figure 3: Marks on test A.

Dispersion is the state of getting dispersed or spread. Statistical dispersion means the extent to which numerical data is likely to vary about an average value.

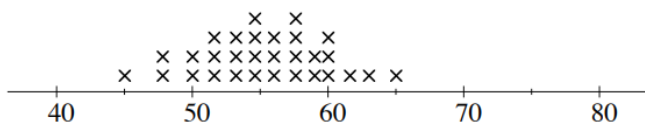


Figure 4: Marks on test B.

Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

- On test A, the range of marks is $70 - 45 = 25$.
- On test B, the range of marks is $65 - 45 = 20$.

Here the range gives us an accurate picture of the dispersion of the two distributions. However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers. This is illustrated very well if the students' marks are distributed as in Figures 5 and 6.

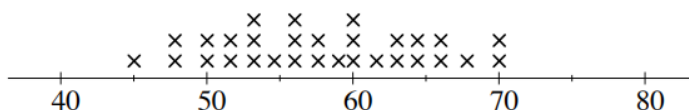


Figure 5: Marks on test A.

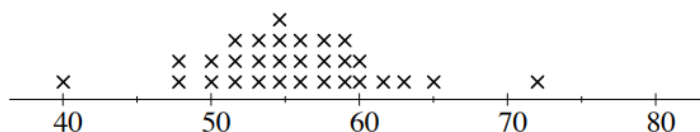


Figure 6: Marks on test B.

On test A, the range is still $70 - 45 = 25$.

On test B, the range is now $72 - 40 = 32$, but apart from the outliers, the distribution of marks on test B is clearly less spread out than that of A. We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion most commonly used, the **standard deviation**.

Standard deviation

is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

Standard Deviation

Suppose we have a set of data where there is no variability in the observed values. Each observation would have the same

value, say 3, 3, 3, 3 and the mean would be that same value, 3. Each observation would not be different or deviate from the mean. Now suppose we have a set of observations where there is variability. The observed values would deviate from the mean by varying amounts. The standard deviation is a kind of average of these deviations from the mean. This is best explained by considering the following example.

Take, for example, the following grades of 6 students:

56 48 63 60 51 52.

Mean = 55.

To find how much our observed values deviate from the mean, we subtract the mean from each.

Observed values	56	48	63	60	51	52
Deviations from Mean	+1	-7	+8	+5	-4	-3

We cannot, at this stage, simply take the average of the deviations as their sum is zero.

$$(+1) + (-7) + (+8) + (+5) + (-4) + (-3) = 0$$

We get around this difficulty by taking the square of the deviations. This gets rid of the minus signs. (Remember $(-7) \times (-7) = 49$.)

Deviations	+1	-7	+8	+5	-4	-3
Squared deviations	1	49	64	25	16	9

We can now take the mean of these squared deviations. This is called the variance.

$$\text{Variance} = \frac{1 + 49 + 64 + 25 + 16 + 9}{6} = 27.33.$$

The variance is a very useful measure of dispersion for statistical inference, but for our purposes it has a major disadvantage. Because we squared the deviations, we now have a quantity in square units. So to get the measure of dispersion back into the same units as the observed values, we define standard deviation as the square root of the variance.

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{27.33} = 5.228.$$

The standard deviation may be thought of as the ‘give or take’ number. That is, on average, the student’s grade will be 55, give or take 5 marks. The standard deviation is a very good measure of dispersion and is the one to use when the mean is used as the measure of central tendency.

Example: Calculate the mean and standard deviation of the following set of data.
Birthweight of ten babies (in kilograms)

2.977 3.155 3.920 3.412 4.236 2.593 3.270 3.813 4.042 3.387

Solution:

Birthweight in kilograms	Deviations from Mean score – mean	Squared Deviations (score – mean) ²
2.977	–0.5035	0.2535
3.155	–0.3255	0.1060
3.920	0.4395	0.1932
3.412	–0.0685	0.0047
4.236	0.7555	0.5708
2.593	–0.8875	0.7877
3.270	–0.2105	0.0443
3.813	0.3325	0.1106
4.042	0.5615	0.3153
3.387	–0.0935	0.0087
Sum = 34.805	Sum = 0	Sum = 2.3948

$$\text{Mean} = \frac{\text{sum of observations}}{\text{number of observations}} = \frac{34.805}{10} = 3.4805 = \mu.$$

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of observations}} = \frac{2.3948}{10} = 0.2395 = \sigma^2.$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{0.2395} = 0.4894 = \sigma.$$

The Interquartile Range

The interquartile range is another useful measure of dispersion or spread. It is used when the median is used as the measure of central tendency. It gives the range in which the middle 50% of the distribution lies. In order to describe this in detail, we first need to discuss what we mean by quartiles.

1. Quartiles

Suppose we start with a large set of data, say the heights of all adult males. We can represent these data in a graph, which if smoothed out a bit, may look like Figure 7.

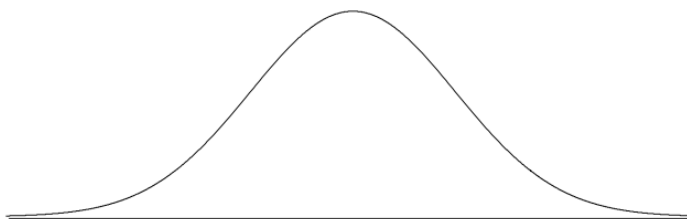


Figure 7: Graph representing heights of adult males.

As the name ‘quartile’ suggests, we want to divide the data into four equal parts. In the above example, we want to divide the area under our curve into four equal areas.

a. The second quartile or median

It is easy to see how to divide the area in Figure 7 into two equal parts, since the graph is symmetric. The point which gives us 50% of the area to the left of it and 50% to the right of it is called the second quartile or median. This is illustrated in Figure 8.

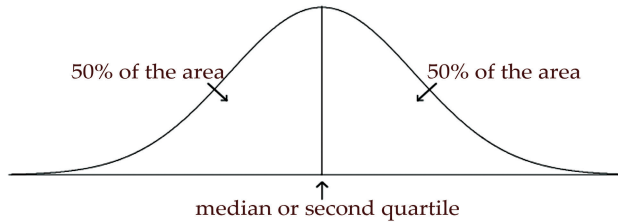


Figure 8: Graph showing the median or second quartile.

This exactly corresponds to our before idea of median as the middle value.

b. The first quartile

The first quartile is the point which gives us 25% of the area to the left of it and 75% to the right of it. This means that 25% of the observations are less than or equal to the first quartile and 75% of the observations greater than or equal to the first quartile. The first quartile is also called the 25th percentile. This is illustrated in Figure 9.

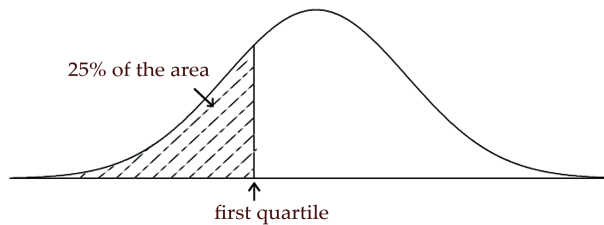


Figure 9: Graph showing the first quartile.

c. The third quartile

The third quartile is the point which gives us 75% of the area to the left of it and 25% of the area to the right of it. This means that 75% of the observations are less than or equal to the third quartile and 25% of the observation are greater than or equal to the third quartile. The third quartile is also called the 75th percentile. This is illustrated in Figure 10.

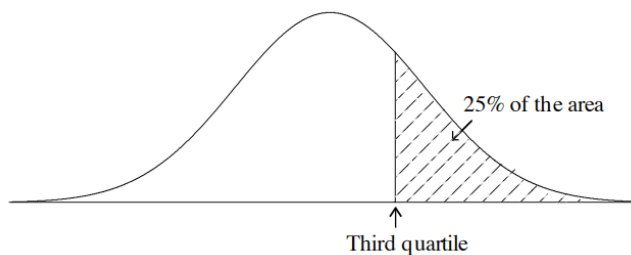


Figure 10: Graph showing the third quartile.

The first (Q_1), second (Q_2) and third (Q_3) quartiles divide the distribution into four equal parts. This is illustrated in Figure 11.

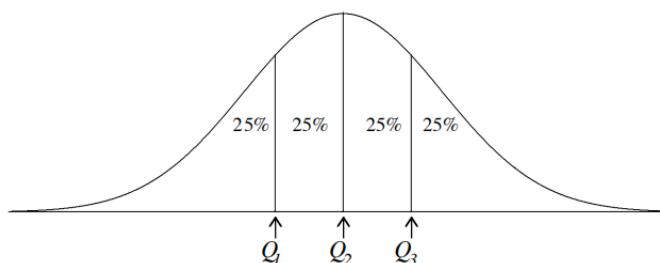


Figure 11: Graph showing all quartiles.

2. Quartiles for small data sets

Suppose now we have a small data set of twelve observations which we can write in ascending order as follows. (A data set, where the number of observations is a multiple of four, has been chosen to avoid some technical difficulties.)

15 18 19 20 20 20 21 23 23 24 24 25

In this case, we want to divide the data into four equal sets, so that there are 25% of the observations in each. First, we find the median

15	18	19	20	20	20	↑	21	23	23	24	24	25
median												

The median is 20.5 (half way between the 6th and 7th observations), and divides the data into two equal sets with exactly 50% of the observations in each: the 1st to the 6th observations in the first set and the 7th to 12th observations in the other. To find the first quartile we consider the observations less than the median.

15 18 19 20 20 20

The first quartile is the median of these data. In this case, the first quartile is half way between the 3rd and 4th observations and is equal to 19.5. Now, we consider the observations which are greater than the median.

21 23 23 24 24 25

The third quartile is the median of these data and is equal to 23.5. So, for our small data set of 12 observations, the quartiles divide the set into four subsets.

$$\begin{array}{cccccccccccccccc} 15 & 18 & 19 & \uparrow & 20 & 20 & 20 & \uparrow & 21 & 23 & 23 & \uparrow & 24 & 24 & 25 \\ & & & & Q_1 & & & & Q_2 & & & & Q_3 & & \end{array}$$

We will now use the quartiles to define a measure of spread called interquartile range.

3. The interquartile range

The interquartile range quantifies the difference between the third and first quartiles. If we were to remove the median (Q_2) from Figure 11 we would have a graph like that in Figure 12.

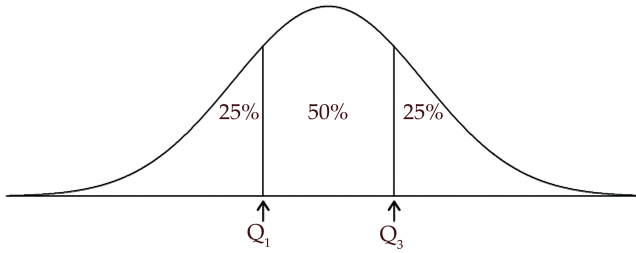


Figure 12: Graph showing the first and third quartiles.

From Figure 12, we see that 50% of the area is between the first and third quartiles. This means that 50% of the observations lie between the first and third quartiles. We define the interquartile range as:

The interquartile range = Third quartile – First quartile.

For our small data set, the first quartile was 19.5 and our third quartile was 23.5. So, the interquartile range is $23.5 - 19.5 = 4$. We will use the interquartile range later to draw a box-plot. For now we are interested in it as a measure of spread. The interquartile range is particularly useful to describe data sets where there are a few extreme values. Unlike the range, and to a lesser extent the standard deviation, it is not sensitive to extreme values as it relies on the spread of the middle 50% of the distribution. So, if there are data sets which have extreme values, it can be more appropriate to use the median to describe central tendency and the **interquartile range** to describe the spread.

Difference between Descriptive and Inferential Statistics

In today's fast-paced world, statistics is playing a major role in the field of research; that helps in the collection, analysis and presentation of data in a measurable form. It is quite hard to identify, whether the research relies on descriptive statistics or inferential statistics, as people usually, lacks knowledge about these two branches of statistics. As the name suggests, descriptive statistics is one which describes the population.

On the other end, inferential statistics is used to make the generalization about the population based on the samples. So, there is a big difference between descriptive and inferential statistics, i.e. what you do with your data. Let's take a glance at this section to get some more details on the two topics.

Interquartile range

is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts.



Comparison Chart

BASIS FOR COMPARISON	DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Function	It explains the data, which is already known, to summarize sample.	It attempts to reach the conclusion to learn about the population that extends beyond the data available.

Definition of Descriptive Statistics

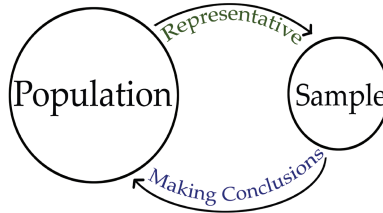
Descriptive Statistics refers to a discipline that quantitatively describes the important characteristics of the dataset. For the purpose of describing properties, it uses measures of central tendency, i.e. mean, median, mode and the measures of dispersion i.e. range, standard deviation, quartile deviation and variance, etc. The data is summarized by the researcher, in a useful way, with the help of numerical and graphical tools such as charts, tables, and graphs, to represent data in an accurate way. Moreover, the text is presented in support of the diagrams, to explain what they represent.

Definition of Inferential Statistics

Inferential Statistics is all about generalizing from the sample to the population, i.e. the results of analysis of the sample can be deduced to the larger population, from which the sample is taken. It is a convenient way to draw conclusions about the population when it is not possible to query each and every member of the universe.

The sample chosen is a representative of the entire population; therefore, it should contain important features of the population.

Inferential Statistics



Inferential Statistics is used to determine the probability of properties of the population on the basis of the properties of the sample, by employing probability theory. The major inferential statistics are based on the statistical models such as Analysis of Variance, chi-square test, student's t distribution, regression analysis, etc. Methods of inferential statistics:

- Estimation of parameters
- Testing of hypothesis

Key Differences between Descriptive and Inferential Statistics

The difference between descriptive and inferential statistics can be drawn clearly on the following grounds:

- Descriptive Statistics is a discipline which is concerned with describing the population under study. Inferential Statistics is a type of statistics; that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
- Descriptive Statistics collects, organizes, analyzes and presents data in a meaningful way. On the contrary, Inferential Statistics, compares data, test hypothesis and make predictions of the future outcomes.
- There is a diagrammatic or tabular representation of final result in descriptive statistics whereas the final result is displayed in the form of probability.
- Descriptive statistics describes a situation while inferential statistics explains the likelihood of the occurrence of an event.
- Descriptive statistics explains the data, which is already known, to summarize sample. Conversely, inferential statistics attempts to reach the end to learn about the population; that extends beyond the data available.

The use of descriptive and summary statistics has an extensive history and, indeed, the simple tabulation of populations and of economic data was the first way the topic of statistics appeared.



CORE CONCEPT: EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis refers to a set of techniques originally developed by John Tukey to display data in such a way that interesting features will become apparent. Unlike classical methods which usually begin with an assumed model for the data, EDA techniques are used to encourage the data to suggest models that might be appropriate.

Exploratory Data Analysis (EDA) is an approach to analyzing data. It is where the researcher takes a bird's eye view of the data and tries to make some sense of it. It is often the first step in data analysis, implemented before any formal statistical techniques are applied. Although specific statistical techniques can be used, like creating histograms or box plots. EDA is considered by some to be more of an art form than a science.



Exploratory data analysis is a complement to inferential statistics, which tends to be fairly rigid with rules and formulas. EDA involves the analyst trying to get a “feel” for the data set, often using their own judgment to determine what the most important elements in the data set are.

The purpose of exploratory data analysis is to:

- Check for missing data and other mistakes.
- Gain maximum insight into the data set and its underlying structure.
- Uncover a parsimonious model, one which explains the data with a minimum number of predictor variables.
- Check assumptions associated with any model fitting or hypothesis test.
- Create a list of outliers or other anomalies.
- Find parameter estimates and their associated confidence intervals or margins of error.

- Identify the most influential variables.

Other, specific knowledge can be obtained through EDA such as creating a ranked list of relevant factors. You may not necessarily include all of the above items in your data analysis, although it is likely you will want to include at least a few. They should be viewed as guidelines, rather than rigid rules.

Types of Exploratory Data Analysis

EDA falls into four main areas:

- Univariate non-graphical — looking at one variable of interest, like age, height, income level etc.
- Univariate graphical.
- Multivariate non-graphical — analysis of multiple variables at the same time.
- Multivariate graphical.

Exploratory Data Analysis Techniques

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of techniques to

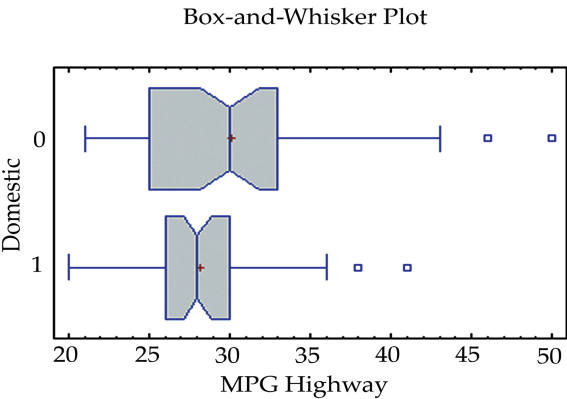
- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models; and
- Determine optimal factor settings.

Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to open-mindedly explore, and graphics gives the analysts unparalleled power to do so, enticing the data to reveal its structural secrets, and being always ready to gain some new, often unsuspected, insight into the data. In combination with the natural pattern recognition capabilities that we all possess, graphics provides, of course, unparalleled power to carry this out. The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

- Plotting the raw data (such as data traces, histograms, bihistograms, stem and leaf display, probability plots, lag plots, block plots, Youden plots, scatter plots, character plots, residual plots.
- Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
- Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

Box-and-Whisker Plots

Box-and-whisker plots are graphical displays based upon Tukey’s 5-number summary of a data sample. In his original plot, a box is drawn covering the center 50% of the sample. A vertical line is drawn at the median, and whiskers are drawn from the central box to the smallest and largest data values. If some points are far from the box, these “outside points” may be shown as separate point symbols. Later analysts have added notches showing approximate confidence intervals for the median, and plus signs at the sample mean.

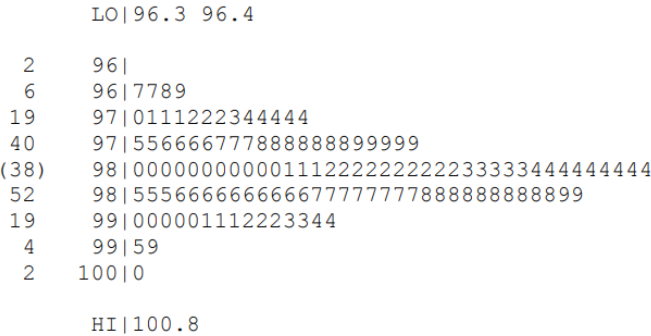


Stem-and-Leaf Display

Stem-and-leaf displays take each data value and divide it into a stem and a leaf. For example, the temperature of the first subject in the data sample to the left had a body temperature of 98.4 degrees. The first two digits (“98”) are called the stem and plotted at the left, while the third digit (“4”) is called the leaf. Although similar to a histogram turned on its side, Tukey thought that the stem-and-leaf plot was preferable to a bar chart since the **data** values could be recovered from the display.

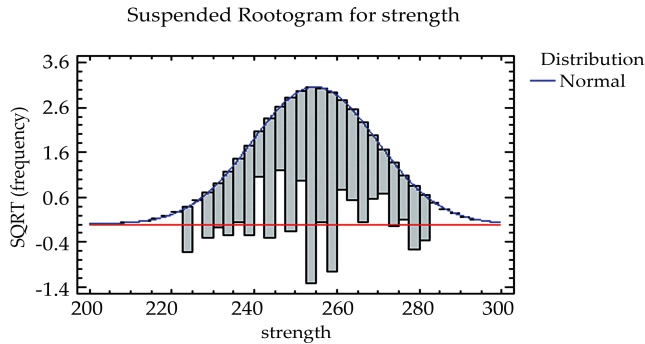
Data is a set of values of qualitative or quantitative variables.

Stem-and-Leaf Display for Temperature: unit = 0.1 1|2 represents 1.2



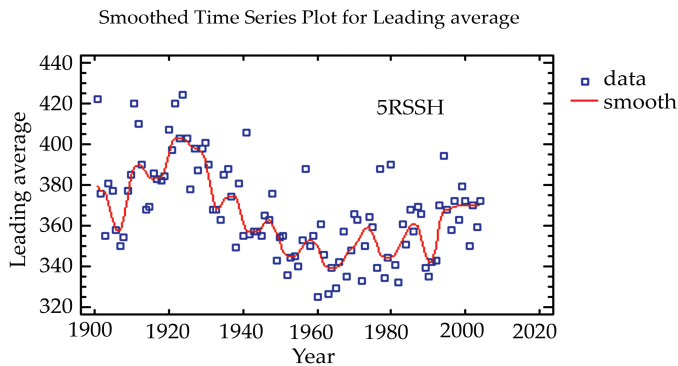
Rootogram

A rootogram is similar to a histogram, except that it plots the square roots of the number of observations observed in different ranges of a quantitative variable. It is usually plotted together with a fitted distribution. The idea of using square roots is to equalize the variance of the deviations between the bars and the curve, which otherwise would increase with increasing frequency. Sometimes, the bars are suspended from the fitted distribution, which allows for easier visual comparison with the horizontal line drawn at 0, since visual comparison with a curved line may be deceiving.



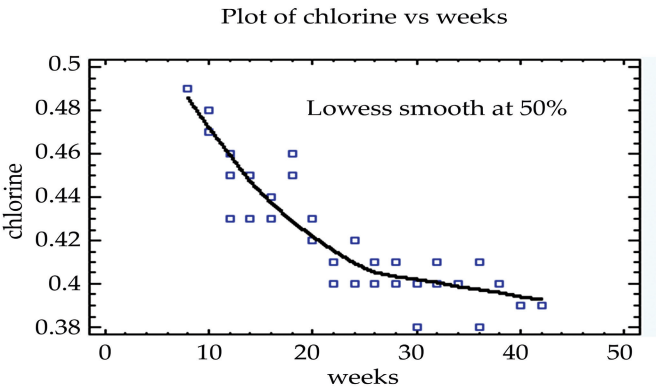
Resistant Time Series Smoothing

Tukey invented a number of nonlinear smoothers, used to smooth sequential time series data that are very good at ignoring outliers and are often applied as a first step to reduce the influence of potential outliers before a moving average is applied. These include 3RSS, 3RSSH, 5RSS, 5RSSH, and 3RSR smoothers. Each symbol in the name of the smoother indicates an operation that is applied to the data.



Scatterplot Smoothing

X-Y scatterplots may be smoothed using any of several methods: running means, running lines, LOWESS (locally weighted scatterplot smoothing), and resistant LOWESS. Smoothers are useful for suggesting the type of regression model that might be appropriate to describe the relationship between two variables.



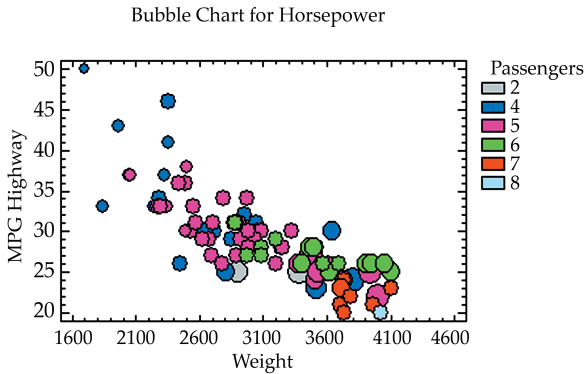
Median Polish

The Median Polish procedure constructs a model for data contained in a two-way table. The model represents the contents of each cell in terms of a common value, a row effect, a column effect, and a residual. Although the model used is similar to that estimated using a two-way analysis of variance, the terms in the model are estimated using medians rather than means. This makes the estimates more resistant to the possible presence of outliers.

Cause	None	Grams 1 14	Grams 15 24	Grams 25	Row effect
Lung cancer	-0.5	-0.2025	0.2	0.86	0.1175
Upper resp. cancer	0.0	0.0275	0.0	-0.02	-0.4525
Stomach cancer	0.24	0.0875	-0.16	-0.09	-0.2825
Colon cancer	0.0025	0.0	-0.1575	0.0725	-0.015
Prostrate caner	0.405	0.0125	-0.015	-0.035	-0.3075
Other cancer	-0.015	-0.0375	0.015	0.135	0.2025
TB	-0.06	-0.0025	0.03	0.0	-0.3925
Bronchitis	-0.125	-0.0575	0.055	0.245	-0.2075
Other respiratory	0.24	-0.0025	0.0	-0.28	-0.0025
Thrombosis	-0.305	0.0125	-0.015	1.235	4.0725
Cardiovascular	0.0925	-0.09	0.2425	-0.1175	1.685
Hemorrhage	0.0875	-0.085	-0.1525	0.1775	1.47
Ulcer	-0.0175	0.02	0.0525	-0.0275	-0.435
Violence	-0.125	0.1725	-0.185	0.125	0.0925
Other	0.035	0.2925	-0.035	-0.075	0.9625
Column effect	-0.09375	0.00875	-0.00375	0.13625	0.54625

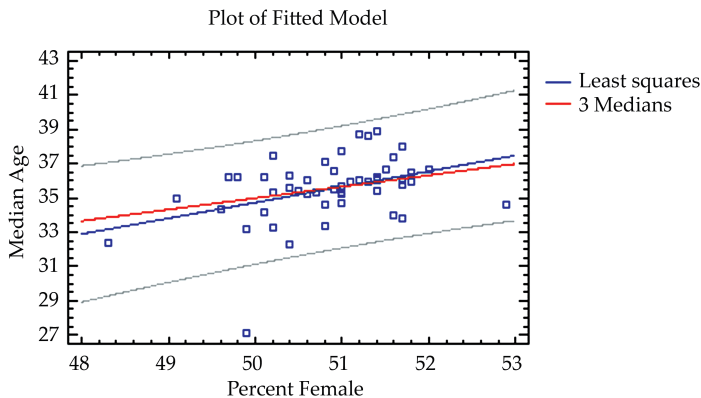
Bubble Chart

The Bubble Chart is an X-Y scatterplot on which the value of a third and possibly fourth variable is shown by changing the size and/or color of the point symbols. It is one way to plot multivariate data in 2 dimensions.



Resistant Curve Fitting

Tukey proposed a method for fitting lines and other curves that is less influenced by any outliers that might be present. Called the method of 3 medians, the data are first divided into 3 groups according to the value of X. Medians are then computed within each group, and the curve is determined from the 3 medians.



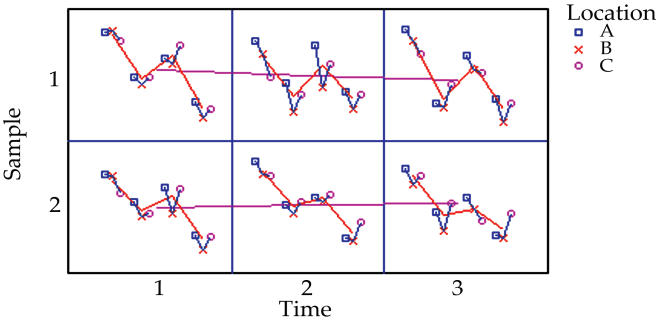
Multi-Vari Chart

A Multi-Vari Chart is a chart designed to display multiple sources of variability in a way that enables the analyst to identify easily which factors are the most important. This exploratory data analysis technique is commonly used to display EDA data from a designed experiment prior to performing a formal **statistical analysis**.

Statistical analysis

is a component of data analytics. In the context of business intelligence (BI), statistical analysis involves collecting and scrutinizing every data sample in a set of items from which samples can be drawn.

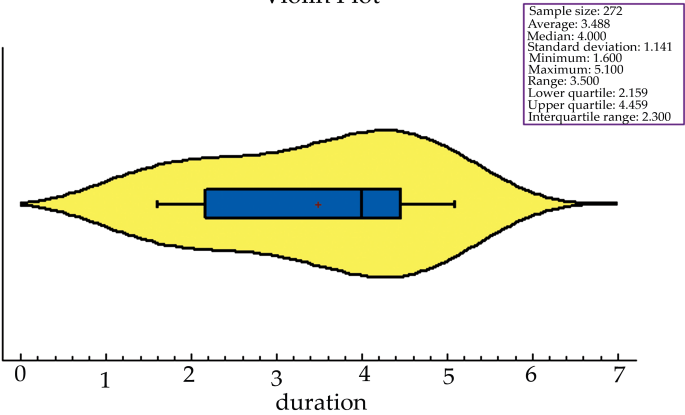
Multi-Vari Chart for Diameter
Cavity=1,2,3,4



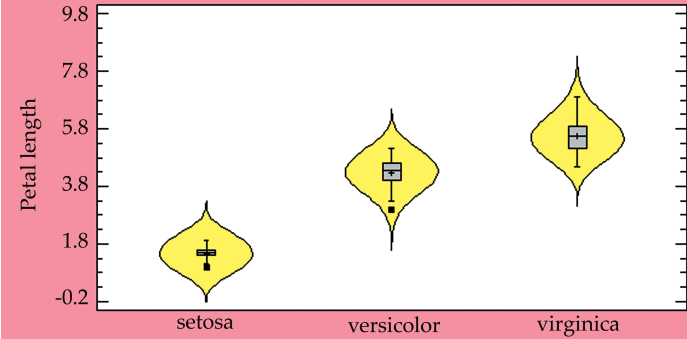
Violin Plot

The Violin Plot Statlet displays data for a single quantitative sample using a combination of a box-and-whisker plot and a nonparametric density estimator. It is very useful for visualizing the shape of the probability density function for the population from which the data came. A separate procedure is available for creating violin plots for multiple samples.

Violin Plot

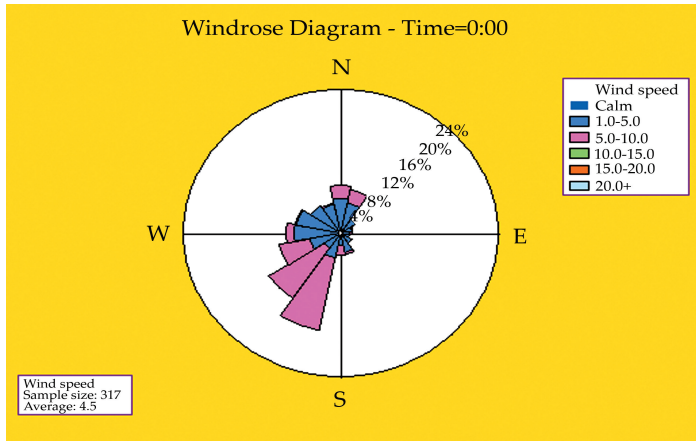


Violin Plot by Species



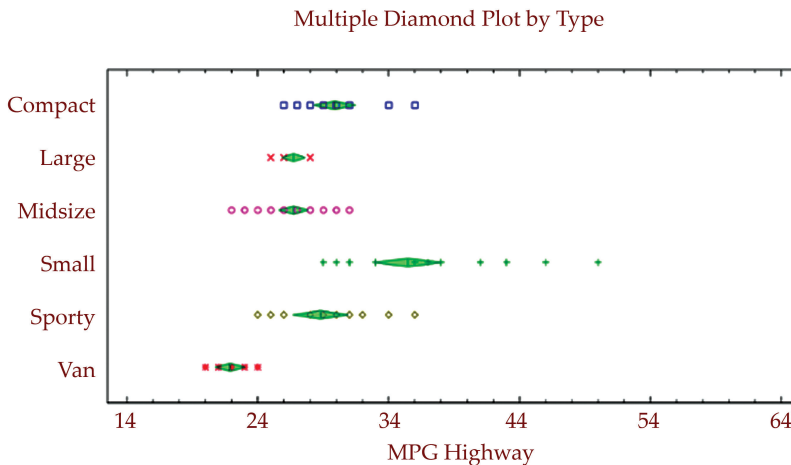
Wind Rose

The Wind Rose Statlet displays data on a circular plot, depicting the frequency distribution of variables such as wind speed and direction. It may be used to display the distribution at a single point in time, or it may show changes over time in a dynamic manner.



Diamond Plot

The Diamond Plot procedure creates a plot for a single quantitative variable showing the n sample observations together with a confidence interval for the population mean. A separate procedure is available for creating diamond plots for multiple samples.

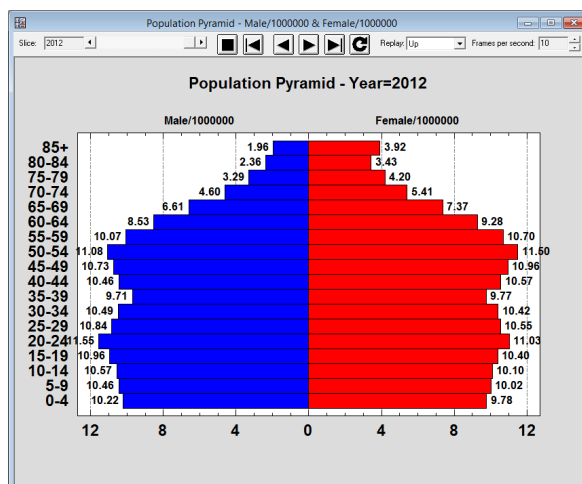


Heat Map

The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent values of the quantitative variable.

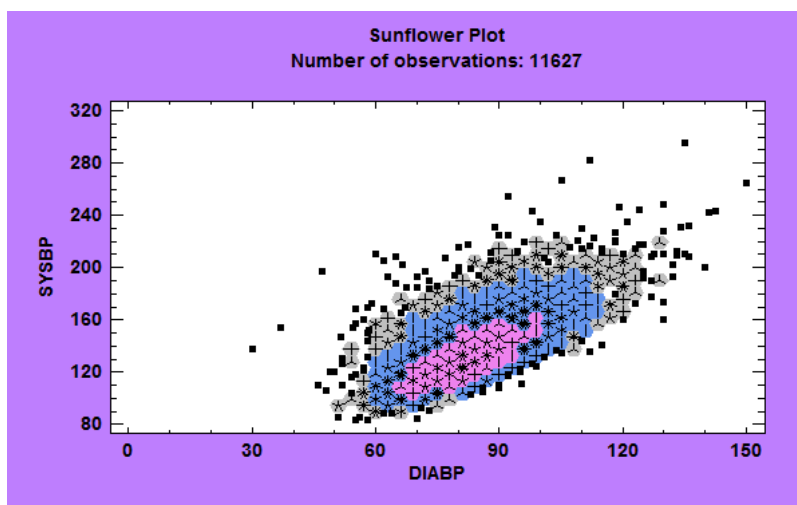
Population Pyramid

The Population Pyramid Statlet is designed to compare the distribution of population counts (or similar values) between 2 groups. It may be used to display that distribution at a single point in time, or it may show changes over time in a dynamic manner. In the latter case, various options are offered for smoothing the data and for dealing with missing values.



Sunflower Plot

The *Sunflower Plot Statlet* is used to display an X-Y scatterplot when the number of observations is large. To avoid the problem of over plotting point symbols with large amounts of data, glyphs in the shape of sunflowers are used to display the number of observations in small regions of the X-Y space.



Measurement Scales

Variables differ in “how well” they can be measured, i.e., in how much measurable information their measurement scale can provide. There is obviously some

measurement error involved in every measurement, which determines the “amount of information” that we can obtain. Another factor that determines the amount of information that can be provided by a variable is its “type of measurement scale.” Specifically variables are classified as (a) nominal, (b) ordinal, (c) interval or (d) ratio.

Remember, statistics can be a powerful tool for making informed decisions and drawing conclusions, but it's important to be aware of potential biases, limitations, and pitfalls in the analysis process.

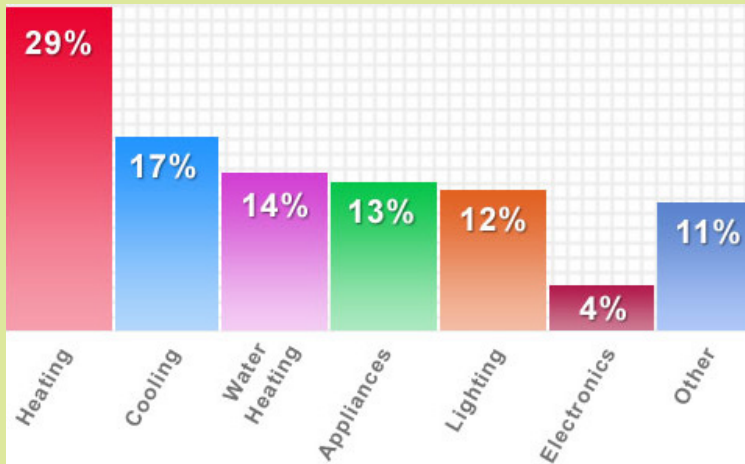
- Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. For example, all we can say is that 2 individuals are different in terms of variable A (e.g., they are of different race), but we cannot say which one “has more” of the quality represented by the variable. Typical examples of nominal variables are soil type, variety, gender, race, color, city, etc.
- Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say “how much more.” A typical example of an ordinal variable is the socioeconomic status of families. For example, we know that upper-middle is higher than middle but we cannot say that it is, for example, 18% higher. Also this very distinction between nominal, ordinal, and interval scales itself represents a good example of an ordinal variable. For example, we can say that nominal measurement provides less information than ordinal measurement, but we cannot say “how much less” or how this difference compares to the difference between ordinal and interval scales.
- Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Fahrenheit or Celsius, constitutes an **interval scale**. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.
- Ratio variables are very similar to interval variables;

Interval Scale is a numeric scale in which the numbers are assigned to the objects in such a way that numerically equal distances on the scale represent the equal distances between the characteristics of the objects being measured.

in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y . Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, we can correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

A CONCEPT OF BAR CHART

A bar chart (also known as a bar graph) is a graphical representation of data that uses rectangular bars to display values. It is particularly useful for showing and comparing the frequency, distribution, or relationship between different categories or groups. Bar charts can be created for both categorical and numerical data.



Here's a breakdown of the key components and how to interpret a bar chart:

- **Categories or Groups:** The x-axis of the bar chart represents the different categories or groups being compared. These categories can be anything from types of fruits to age groups, countries, or any other distinct classifications.
- **Values or Frequencies:** The y-axis of the bar chart represents the values or frequencies associated with each category. The height of each bar corresponds to the value or frequency of the respective category.
- **Bars:** The bars in a bar chart are rectangular columns that are drawn vertically or horizontally, depending on the orientation of the chart. Each category is represented by a separate bar, and the length or height of the bar corresponds to the value it represents.
- **Spacing:** There is typically a small gap between adjacent bars to visually distinguish them from each other.

SUMMARY

- Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data.
- Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.
- The central tendency of a distribution is an estimate of the “center” of a distribution of values.
- The interquartile range is another useful measure of dispersion or spread. It is used when the median is used as the measure of central tendency.
- Exploratory Data Analysis (EDA) is an approach to analyzing data. It is where the researcher takes a bird’s eye view of the data and tries to make some sense of it
- A rootogram is similar to a histogram, except that it plots the square roots of the number of observations observed in different ranges of a quantitative variable.
- The Bubble Chart is an X-Y scatterplot on which the value of a third and possibly fourth variable is shown by changing the size and/or color of the point symbols.

MULTIPLE CHOICE QUESTIONS

1. **What is the primary goal of statistics?**
 - a. To make data look more appealing
 - b. To manipulate data for personal gain
 - c. To draw meaningful conclusions from data
 - d. To create fictional data
2. **Which of the following is a measure of central tendency?**
 - a. Variance
 - b. Range
 - c. Standard Deviation
 - d. Median
3. **The average of a set of numbers is known as:**
 - a. Median
 - b. Mode
 - c. Mean
 - d. Range
4. **What does inferential statistics involve?**
 - a. Describing data patterns
 - b. Summarizing data
 - c. Drawing conclusions about a population based on a sample
 - d. Organizing data
5. **What is a random sample?**
 - a. A sample chosen without any specific criteria
 - b. A sample with the highest variability
 - c. A sample chosen based on convenience
 - d. A sample with the lowest variability

Answers

1. (c) 2. (d) 3. (c) 4. (c) 5. (a)

REVIEW QUESTIONS

1. What do you understand by descriptive statistics?
2. Differentiate between descriptive and inferential statistics.
3. Write the different types of data.
4. Discuss about central tendency.
5. The following values are the number of customers a restaurant served for lunch on ten consecutive days.

46 50 51 60 62 64 72 41 53 55

Find the mean and standard deviation of these data.

6. The ages of seven geography students who went on a field trip were

20 19 19 25 20 18 19

and the age of the instructor who accompanied them was 52. Find the mean and standard deviation of ages of these eight people.

7. The raw scores that eight students got on a history test were:

69 84 93 61 79 88 57 67

Find the mean and standard deviation of these scores.

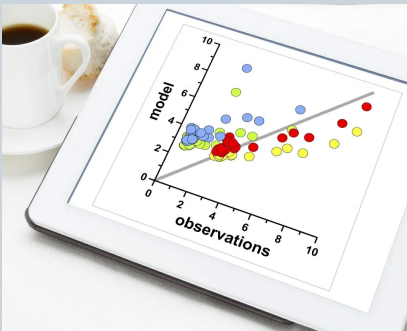
8. Create the stem and leaf display and box plot.

REFERENCES

1. Agresti, A., & Finlay, B. (2018). *Statistical Methods for the Social Sciences*. Pearson.
2. Agresti, A., & Franklin, C. A. (2018). *Statistics: The Art and Science of Learning from Data*. Pearson.
3. Bluman, A. G. (2019). *Elementary Statistics: A Step by Step Approach*. McGraw-Hill Education.
4. De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2020). *Intro Stats*. Pearson.
5. Diez, D. M., Barr, C. D., & Cetinkaya-Rundel, M. (2019). *OpenIntro Statistics*. OpenIntro, Inc.
6. Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W. W. Norton & Company.
7. Moore, D. S., Notz, W. I., & Fligner, M. A. (2018). *The Basic Practice of Statistics*. W. H. Freeman.
8. Spiegel, M. R., Stephens, L. J., & Schiller, J. L. (2017). *Schaum's Outline of Statistics*. McGraw-Hill Education.
9. Triola, M. F. (2019). *Elementary Statistics*. Pearson.

CHAPTER 2

STATISTICS AND CORRELATION FORMULAE



Statistics is the scientific discipline of collecting, analyzing, interpreting, presenting, and organizing data. It provides tools and methods to understand patterns, make informed decisions, and draw meaningful conclusions from data. A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. Using a scatterplot, we can generally

assess the relationship between the variables and determine whether they are correlated or not.

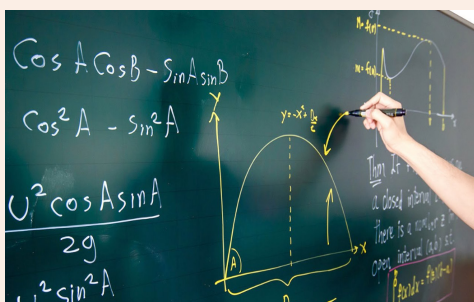
After studying this chapter, you will understand the core concepts of:

- Statistics and Its Formulae
- Fundamental of Correlation
- Simple Linear Regression

INTRODUCTION

The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a perfect negative, or inverse, correlation, with values in one series rising as those in the other decline, and vice versa. A coefficient of 1 shows a perfect positive correlation, or a direct relationship. A correlation coefficient of 0 means there is no linear relationship.

Correlation coefficients are used in science and in finance to assess the degree of association between two variables, factors, or data sets. For example, since high oil prices are favorable for crude producers, one might assume the correlation between oil prices and forward returns on oil stocks is strongly positive. Calculating the correlation coefficient for these variables based on market data reveals a moderate and inconsistent correlation over lengthy periods.



CORE CONCEPT: STATISTICS AND ITS FORMULAE

Statistical formulas are used to calculate values related to statistical concepts or analyses. Here we will discuss common formulas and what they stand for.

Population Mean

The term population mean, which is the average score of the population on a given variable, is represented by:

$$\mu = (\sum X_i) / N$$

Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances. The science of statistics deals with the collection, analysis, interpretation, and presentation of data. We see and use data in our everyday lives. To be able to use data correctly is essential to many professions and is in your own best self-interest.

Statistics is a branch of mathematics which deals with numbers and data analysis. Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. Statistical theory defines a statistic as a function of a sample where the function

Correlation coefficient formulas are used to find how strong a relationship is between data.

itself is independent of the sample's distribution. In short, Statistics is associated with collecting, classifying, arranging and presenting numerical data. It allows us to interpret various results from it and forecast many possibilities.

Mean	$\bar{x} = \frac{\sum x}{n}$	x = Items given n = Total number of items
Median	If n is odd, then $M = \left(\frac{n+1}{2}\right)^{th} \text{ term}$ If n is even, then $M = \frac{\left(\frac{n}{2}\right)^{th} \text{ term} + \left(\frac{n}{2} + 1\right)^{th} \text{ term}}{2}$	n = Total number of items
Mode	The value which occurs most frequently	
Variance	$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$	x = Items given $\bar{x}$ = Mean n = Total number of items
Standard Deviation	$S = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$	x = Items given $\bar{x}$ = Mean n = Total number of items

The Mean and Median: Measures of Central Tendency

The mean and the median are summary measures used to describe the most “typical” value in a set of values. Statisticians refer to the mean and median as measures of central tendency. The “mean” is the “average” you’re used to, where you add up all the numbers and then divide by the number of numbers. The “median” is the “middle” value in the list of numbers. To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median.

The Mean and the Median

The difference between the mean and median can be illustrated with an example. Suppose we draw a sample of five women and measure their weights. They weigh 100 pounds, 100 pounds, 130 pounds, 140 pounds, and 150 pounds.

To find the median, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values. Thus, in the sample of five women, the median value would be 130 pounds; since 130 pounds is the middle weight.

The mean of a sample or a population is computed by adding all of the observations and dividing by the number of observations. Returning to the example of the five women, the mean weight would equal $(100 + 100 + 130 + 140 + 150)/5 = 620/5 = 124$ pounds. In the general case, the mean can be calculated, using one of the following equations:

$$\text{Population mean} = \mu = \sum X / N \quad \text{OR} \quad \text{Sample mean} =$$

$$\bar{x} = \sum x / n$$

where $\sum X$ is the sum of all the population observations, N is the number of population observations, $\sum x$ is the sum of all the sample observations, and n is the number of sample observations.

When statisticians talk about the mean of a population, they use the Greek letter μ to refer to the mean score. When they talk about the mean of a sample, statisticians use the symbol $\bar{x}$ to refer to the mean score.

The Mean vs. the Median

As measures of central tendency, the mean and the median each have advantages and disadvantages. Some pros and cons of each measure are summarized below.

The median may be a better indicator of the most typical value if a set of scores has an outlier. An outlier is an extreme value that differs greatly from other values. However, when the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency.

To illustrate these points, consider the following example. Suppose we examine a sample of 10 households to estimate the typical family income. Nine of the households have incomes between \$20,000 and \$100,000; but the tenth household has an annual income of \$1,000,000,000. That tenth household is an outlier. If we choose a measure to estimate the income of a typical household, the mean will greatly over-estimate the income of a typical family (because of the outlier); while the median will not.

Median is the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half. For a data set, it may be thought of as the “middle” value.

Effect of Changing Units

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of central tendency are affected when we change units.

If you add a constant to every value, the mean and median increase by the same constant. For example, suppose you have a set of scores with a mean equal to 5 and a median equal to 6. If you add 10 to every score, the new mean will be $5 + 10 = 15$; and the new median will be $6 + 10 = 16$.

Suppose you multiply every value by a constant. Then, the mean and the median will also be multiplied by that constant. For example, assume that a set of scores has a mean of 5 and

a median of 6. If you multiply each of these scores by 10, the new mean will be $5 * 10 = 50$; and the new **median** will be $6 * 10 = 60$.

Problem. Four friends take an IQ test. Their scores are 96, 100, 106, 114. Which of the following statements is true?

- I. The mean is 103.
- II. The mean is 104.
- III. The median is 100.
- IV. The median is 106.

- (A) I only
- (B) II only
- (C) III only
- (D) IV only
- (E) None is true

Solution

The correct answer is (B). The mean score is computed from the equation:

$$\text{Mean score} = \Sigma x / n = (96 + 100 + 106 + 114) / 4 = 104$$

Since there are an even number of scores (4 scores), the median is the average of the two middle scores. Thus, the median is $(100 + 106) / 2 = 103$.

Measure Variability in Statistics

Statisticians use summary measures to describe the amount of variability or spread in a set of data. The most common measures of **variability** are the range, the interquartile range (IQR), variance, and standard deviation.

The Range

The range is the difference between the largest and smallest values in a set of values.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. For this set of numbers, the range would be $11 - 1$ or 10.

Variability is the extent to which data points in a statistical distribution or data set diverge—vary—from the average value, as well as the extent to which these data points differ from each other.

The Interquartile Range (IQR)

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

- Q1 is the “middle” value in the first half of the rank-ordered data set.
- Q2 is the median value in the set.
- Q3 is the “middle” value in the second half of the rank-ordered data set.

The interquartile range is equal to Q3 minus Q1.

For example, consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. Q1 is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, $Q1 = (3 + 4)/2$ or $Q1 = 3.5$. Q3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, $Q3 = (6 + 7)/2$ or $Q3 = 6.5$. The interquartile range is Q3 minus Q1, so $IQR = 6.5 - 3.5 = 3$.

The Variance

In a population, variance is the average squared deviation from the population mean, as defined by the following formula:

$$\sigma^2 = \sum (X_i - \mu)^2 / N$$

where σ^2 is the population variance, μ is the population mean, X_i is the i th element from the population, and N is the number of elements in the population.

Observations from a simple random sample can be used to estimate the variance of a population. For this purpose, sample variance is defined by slightly different formula, and uses a slightly different notation:

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

where s^2 is the sample variance, $\bar{x}$ is the sample mean, x_i is the i th element from the sample, and n is the number of elements in the sample. Using this formula, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate an unknown population variance, based on data from a simple random sample, this is the formula to use.

The Standard Deviation

The standard deviation is the square root of the variance. Thus, the standard deviation of a population is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

where σ is the population standard deviation, μ is the population mean, X_i is the i th element from the population, and N is the number of elements in the population.

Statisticians often use simple random samples to estimate the standard deviation of a population, based on sample data. Given a simple random sample, the best estimate of the standard deviation of a population is:

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

where s is the sample standard deviation, $\bar{x}$ is the sample mean, x_i is the i th element from the sample, and n is the number of elements in the sample.

Effect of Changing Units

Sometimes, researchers change units (minutes to hours, feet to meters, etc.). Here is how measures of variability are affected when we change units.

- If you add a constant to every value, the distance between values does not change. As a result, all of the measures of variability (range, interquartile range, standard deviation, and variance) remain the same.
- On the other hand, suppose you multiply every value by a constant. This has the effect of multiplying the range, **interquartile range (IQR)**, and standard deviation by that constant. It has an even greater effect on the variance. It multiplies the variance by the square of the constant.

Problem: A population consists of four observations: {1, 3, 5, 7}. What is the variance?

- (A) 2
- (B) 4
- (C) 5
- (D) 6
- (E) None of the above

Solution

Interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

The correct answer is (C). First, we need to compute the population mean.

$$\mu = (1 + 3 + 5 + 7) / 4 = 4$$

Then we plug all of the known values into formula for the variance of a population, as shown below:

$$\sigma^2 = \sum (X_i - \mu)^2 / N$$

$$\sigma^2 = [(1 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (7 - 4)^2] / 4$$

$$\sigma^2 = [(-3)^2 + (-1)^2 + (1)^2 + (3)^2] / 4$$

$$\sigma^2 = [9 + 1 + 1 + 9] / 4 = 20 / 4 = 5$$

Sometimes, students are unsure about whether the denominator in the formula for the variance should be N or $(n - 1)$. We use N to compute the variance of a population, based on population data; and we use $(n - 1)$ to estimate the variance of a population, based on sample data. In this problem, we are computing the variance of a population based on population data, so this solution uses N in the denominator.

Measure Variability in a Data Set

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. A data set is organized into some type of data structure. Here three measures that are used to quantify the amount of variation in a data set - the range, the variance, and the standard deviation.

For example, consider a population of elements $\{5, 5, 5, 5\}$. Here, each of the values in the data set are equal, so there is no variation. The set $\{3, 5, 5, 7\}$, on the other hand, has some variation since some elements in the data set have different values.

Notation

The following notation is helpful, when we talk about variability.

- σ^2 : The variance of the population.
- σ : The standard deviation of the population.
- s^2 : The variance of the sample.
- s : The standard deviation of the sample.
- μ : The population mean.
- x : The sample mean.
- N : Number of observations in the population.
- n : Number of observations in the sample.
- P : The proportion of elements in the population that has a particular attribute.
- p : The proportion of elements in the sample that has a particular attribute.
- Q : The proportion of elements in the population that does not have a specified attribute. Note that $Q = 1 - P$.

- q : The proportion of elements in the sample that does not have a specified attribute. Note that $q = 1 - p$.

Note that capital letters refer to population parameters, and lower-case letters refer to sample statistics.

The Range

The range is the simplest measure of variation. It is difference between the biggest and smallest random variable.

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Therefore, the range of the four random variables $\{3, 5, 5, 7\}$ would be 7 minus 3 or 4.

Variance of the Mean

It is important to distinguish between the variance of a population mean and the variance of a sample mean. They have different notation, and they are computed differently. The variance of a population mean is denoted by σ^2 ; and the variance of a sample mean, by s^2 .

The variance of a population mean is the average squared deviation from the population mean, as defined by the following formula:

$$\sigma^2 = \sum (X_i - \mu)^2 / N$$

where σ^2 is the population variance, μ is the population mean, X_i is the i th element from the population, and N is the number of elements in the population.

The variance of a sample mean is defined by slightly different formula:

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

where s^2 is the sample variance, $\bar{x}$ is the sample mean, x_i is the i th element from the sample, and n is the number of elements in the sample. If you are working with a simple random sample, the sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you want to estimate the unknown population variance, based on known data from a simple random sample, use this formula.

Example. A population consists of four observations: $\{1, 3, 5, 7\}$. What is the variance?

Solution: First, we need to compute the population mean.

$$\mu = (1 + 3 + 5 + 7) / 4 = 4$$

Then we plug all of the known values in to formula for the variance of a population, as shown below:

$$\sigma^2 = \sum (X_i - \mu)^2 / N$$

$$\sigma^2 = [(1 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (7 - 4)^2] / 4$$

$$\sigma^2 = [(-3)^2 + (-1)^2 + (1)^2 + (3)^2] / 4$$

$$\sigma^2 = [9 + 1 + 1 + 9] / 4 = 20 / 4 = 5$$

Example. A simple random sample consists of four observations: {1, 3, 5, 7}. What is the best estimate of the population variance?

Solution: This problem is handled exactly like the previous problem, except that we use the formula for calculating sample variance, rather than the formula for calculating population variance.

$$s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$$

$$s^2 = [(1 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 + (7 - 4)^2] / (4 - 1)$$

$$s^2 = [(-3)^2 + (-1)^2 + (1)^2 + (3)^2] / 3$$

$$s^2 = [9 + 1 + 1 + 9] / 3 = 20 / 3 = 6.667$$

Variance of a Proportion

Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of (random) numbers are spread out from their average value. However, for proportions the formulas can be expressed in a form that is easier to compute. When all of the elements of the population are known, the variance of a population proportion is defined by the following formula:

$$\sigma^2 = PQ$$

where P is the population proportion and Q equals 1 - P.

When the population proportion is estimated from sample data, the variance of the sample proportion is estimated by slightly different formula:

$$s^2 = pq$$

where p is the sample estimate of the true proportion, and q is equal to 1 - p. Given a simple random sample, this sample variance can be considered an unbiased estimate of the true population variance. Therefore, if you need to estimate the unknown population variance, based on known data from a simple random sample, this is the formula to use.

Standard Deviation of a Proportion

The standard deviation of a proportion is the square root of the variance of the proportion. Thus, the standard deviation of a population proportion is:

$$\sigma = \sqrt{\sigma^2} = \sqrt{PQ}$$

where P is the population proportion and Q equals 1 - P.

And, using sample data, the standard deviation of a population proportion can be estimated from the following formula:

$$s = \sqrt{s^2} = \sqrt{pq}$$

where p is the sample proportion and q equals $1 - p$.

Hypothesis Test: Difference between Proportions

A hypothesis test to determine whether the difference between two proportions is significant. A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

- Null hypothesis. The null hypothesis, denoted by H_0 , is usually the hypothesis that sample observations result purely from chance.
- Alternative hypothesis. The alternative hypothesis, denoted by H_1 or H_a , is the hypothesis that sample observations are influenced by some non-random cause.

The test procedure, called the two-proportion z-test, is appropriate when the following conditions are met:

- The sampling method for each population is simple random sampling.
- The samples are independent.
- Each sample includes at least 10 successes and 10 failures.
- Each population is at least 20 times as big as its sample.

This approach consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis. The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of “no effect” or “no difference”. The alternative hypothesis is the statement you want to be able to conclude is true.

Based on the sample data, the test determines whether to reject the null hypothesis. You use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point that you define, then you can reject the null hypothesis.

State the Hypotheses

Every hypothesis test requires the analyst to state a null hypothesis and an alternative hypothesis. The table below shows three sets of hypotheses. Each makes a statement about the difference d between two population proportions, P_1 and P_2 . (In the table, the symbol $\neq$ means “not equal to”.)

Set	Null hypothesis	Alternative hypothesis	Number of tails
1	$P_1 - P_2 = 0$	$P_1 - P_2 \neq 0$	2
2	$P_1 - P_2 \geq 0$	$P_1 - P_2 < 0$	1
3	$P_1 - P_2 \leq 0$	$P_1 - P_2 > 0$	1

The first set of hypotheses (Set 1) is an example of a two-tailed test, since an extreme value on either side of the sampling distribution would cause a researcher to reject the null hypothesis. The other two sets of hypotheses (Sets 2 and 3) are one-tailed tests, since an extreme value on only one side of the sampling distribution would cause a researcher to reject the null hypothesis.

When the null hypothesis states that there is no difference between the two population proportions (i.e., $d = 0$), the null and alternative hypothesis for a two-tailed test are often stated in the following form.

$H_0: P_1 = P_2$

$H_a: P_1 \neq P_2$

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. It should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the two-proportion z-test to determine whether the hypothesized difference between population proportions differs significantly from the observed sample difference.

Analyze Sample Data

Using sample data, complete the following computations to find the test statistic and its associated P-Value.

- Pooled sample proportion. Since the null hypothesis states that $P_1 = P_2$, we use a pooled sample proportion (p) to compute the standard error of the sampling distribution.

$p = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2)$

where p_1 is the sample proportion from population 1, p_2 is the sample proportion from population 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- Standard error. Compute the standard error (SE) of the sampling distribution difference between two proportions.

$$SE = \sqrt{p * (1 - p) * [(1/n_1) + (1/n_2)] }$$

where p is the pooled sample proportion, n_1 is the size of sample 1, and n_2 is the size of sample 2.

- Test statistic. The test statistic is a z-score (z) defined by the following equation.

$$z = (p_1 - p_2) / SE$$

where p_1 is the proportion from sample 1, p_2 is the proportion from sample 2, and SE is the standard error of the sampling distribution.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a z-score, use the Normal Distribution Calculator to assess the probability associated with the z-score.

The analysis described above is a two-proportion z-test.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Test Your Understanding

In this section, two sample problems illustrate how to conduct a hypothesis test for the difference between two proportions. The first problem involves an a two-tailed test; the second problem, a one-tailed test.

Problem: Two-Tailed Test: Suppose the Acme Drug Company develops a new drug, designed to prevent colds. The company states that the drug is equally effective for men and women. To test this claim, they choose a simple random sample of 100 women and 200 men from a population of 100,000 volunteers.

At the end of the study, 38% of the women caught a cold; and 51% of the men caught a cold. Based on these findings, can we reject the company's claim that the drug is equally effective for men and women? Use a 0.05 level of significance.

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- State the hypotheses. The first step is to state the null hypothesis and an alternative hypothesis.

Null hypothesis: $P_1 = P_2$

Alternative hypothesis: $P_1 \neq P_2$

Note that these hypotheses constitute a two-tailed test. The null hypothesis will be rejected if the proportion from population 1 is too big or if it is too small.

- Formulate an analysis plan. For this analysis, the significance level is 0.05. The test method is a two-proportion z-test.
- Analyze sample data. Using sample data, we calculate the pooled sample proportion (p) and the standard error (SE). Using those measures, we compute the z-score test statistic (z).

$$p = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2) = [(0.38 * 100) + (0.51 * 200)] / (100 + 200) = 140/300 = 0.467$$

$$SE = \sqrt{p * (1 - p) * [(1/n_1) + (1/n_2)]}$$

$$SE = \sqrt{0.467 * 0.533 * (1/100 + 1/200)} = \sqrt{0.003733} = 0.061$$

$$z = (p_1 - p_2) / SE = (0.38 - 0.51)/0.061 = -2.13$$

where p_1 is the sample proportion in sample 1, where p_2 is the sample proportion in sample 2, n_1 is the size of sample 1, and n_2 is the size of sample 2.

Since we have a two-tailed test, the P-value is the probability that the z-score is less than -2.13 or greater than 2.13.

We use the Normal Distribution Calculator to find $P(z < -2.13) = 0.017$, and $P(z > 2.13) = 0.017$. Thus, the P-value = $0.017 + 0.017 = 0.034$.

- Interpret results. Since the P-value (0.034) is less than the significance level (0.05), we cannot accept the null hypothesis.

Confidence Interval: Difference between Means

Statisticians use a confidence interval to express the degree of uncertainty associated with a sample statistic. A confidence interval is an interval estimate combined with a probability statement.

For example, suppose a statistician conducted a survey and computed an interval estimate, based on survey data. The statistician might use a confidence level to describe uncertainty associated with the interval estimate. He/she might describe the interval estimate as a “95% confidence interval”. This means that if we used the same sampling method to select different samples and computed an interval estimate for each sample, we would expect the true population parameter to fall within the interval estimates 95% of the time.

Estimation Requirements

- Both samples are simple random samples.
- The samples are independent.
- Each population is at least 20 times larger than its respective sample.

A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

- The sampling distribution of the difference between means is approximately normally distributed.

Generally, the sampling distribution will be approximately normally distributed when the sample size is greater than or equal to 30.

The Variability of the Difference between Sample Means

To construct a confidence interval, we need to know the variability of the difference between sample means. This means we need to know how to compute the standard deviation of the sampling distribution of the difference.

- If the population standard deviations are known, the standard deviation of the sampling distribution is:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \text{sqrt} [\sigma_1^2 / n_1 + \sigma_2^2 / n_2]$$

where σ_1 is the standard deviation of the population 1, σ_2 is the standard deviation of the population 2, and n_1 is the size of sample 1, and n_2 is the size of sample 2.

- When the standard deviation of either population is unknown and the sample sizes (n_1 and n_2) are large, the standard deviation of the sampling distribution can be estimated by the standard error, using the equation below.

$$SE_{\bar{x}_1 - \bar{x}_2} = \text{sqrt} [s_1^2 / n_1 + s_2^2 / n_2]$$

where SE is the standard error, s_1 is the standard deviation of the sample 1, s_2 is the standard deviation of the sample 2, and n_1 is the size of sample 1, and n_2 is the size of sample 2.

Find the Confidence Interval for the Difference between Means

For convenience, we repeat the key steps below.

- Identify a sample statistic. Use the difference between sample means to estimate the difference between population means.
 - Select a confidence level. The confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
 - Find the margin of error. The margin of error expresses the amount of the random variation underlying a survey's results. This can be thought of as a measure of the variation one would see in reported percentages if the same poll were taken multiple times. The larger the margin of error, the less confidence one has that the poll's reported percentages are close to the "true" percentages, that is the percentages in the whole population. Previously, we showed how to compute the margin of error, based on the critical value and standard deviation.

When the sample size is large, you can use a t statistic or a z score for the critical value. Since it does not require computing degrees of freedom, the z score is a little

Degrees of freedom

is the number of values in the final calculation of a statistic that are free to vary.

easier. When the sample sizes are small (less than 40), use a t score for the critical value.

If you use a t statistic, you will need to compute **degrees of freedom (DF)**. Here's how.

- The following formula is appropriate whenever a t statistic is used to analyze the difference between means.

$$DF = (s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)] \}$$

- If you are working with a pooled standard deviation (see above), $DF = n_1 + n_2 - 2$.

The next section presents sample problems that illustrate how to use z scores and t statistics as critical values.

- Specify the confidence interval. The range of the confidence interval is defined by the sample statistic $\pm$ margin of error. And the uncertainty is denoted by the confidence level.

Problem: Small Samples

Suppose that simple random samples of college freshman are selected from two universities - 15 students from school A and 20 students from school B. On a standardized test, the sample from school A has an average score of 1000 with a standard deviation of 100. The sample from school B has an average score of 950 with a standard deviation of 90.

What is the 90% confidence interval for the difference in test scores at the two schools, assuming that test scores came from normal distributions in both schools? (Hint: Since the sample sizes are small, use a t score as the critical value.)

- (A) 50 ± 1.70
- (B) 50 ± 28.49
- (C) 50 ± 32.74
- (D) 50 ± 55.66
- (E) None of the above

Solution

The correct answer is (D). The approach that we used to solve this problem is valid when the following conditions are met.

- The sampling method must be simple random sampling. This condition is satisfied; the problem statement says that we used simple random sampling.

- The samples must be independent. Since responses from one sample did not affect responses from the other sample, the samples are independent.
- The sampling distribution should be approximately normally distributed. The problem states that test scores in each population are normally distributed, so the difference between test scores will also be normally distributed.

Since the above requirements are satisfied, we can use the following four-step approach to construct a confidence interval.

- Identify a sample statistic. Since we are trying to estimate the difference between population means, we choose the difference between sample means as the sample statistic. Thus, $x_1 - x_2 = 1000 - 950 = 50$.
 - Select a confidence level. In this analysis, the confidence level is defined for us in the problem. We are working with a 90% confidence level.
 - Find the margin of error. Elsewhere on this site, we show how to compute the margin of error when the sampling distribution is approximately normal. The key steps are shown below.
- o Find standard error. Using the sample standard deviations, we compute the standard error (SE), which is an estimate of the standard deviation of the difference between sample means.

$$SE = \sqrt{s_1^2 / n_1 + s_2^2 / n_2}$$

$$SE = \sqrt{(100)^2 / 15 + (90)^2 / 20}$$

$$SE = \sqrt{10,000/15 + 8100/20} = \sqrt{666.67 + 405} = 32.74$$

- o Find critical value. The critical value is a factor used to compute the margin of error. Because the sample sizes are small, we express the critical value as a t score rather than a z score. To find the critical value, we take these steps.

□ Compute alpha (α): $\alpha = 1 - (\text{confidence level} / 100) = 1 - 90/100 = 0.10$

□ Find the critical probability (p^*): $p^* = 1 - \alpha/2 = 1 - 0.10/2 = 0.95$

□ Find the degrees of freedom (df):

$$DF = (s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2/n_1)^2 / (n_1 - 1)] + [(s_2^2/n_2)^2 / (n_2 - 1)] \}$$

$$DF = (100^2/15 + 90^2/20)^2 / \{ [(100^2/15)^2 / 14] + [(90^2/20)^2 / 19] \}$$

$$DF = (666.67 + 405)^2 / (31746.03 + 8632.89) = 1150614.5 / 40378.92 = 28.495$$

Rounding off to the nearest whole number, we conclude that there are 28 degrees of freedom.

- The critical value is the t statistic having 28 degrees of freedom and a cumulative probability equal to 0.95. From the t Distribution Calculator, we

Margin of error is a statistic expressing the amount of random sampling error in a survey's results.

find that the critical value is 1.7.

- o Compute margin of error (ME): $ME = \text{critical value} * \text{standard error} = 1.7 * 32.74 = 55.66$
- Specify the confidence interval. The range of the confidence interval is defined by the sample statistic $\pm$ **margin of error**. And the uncertainty is denoted by the confidence level.

Therefore, the 90% confidence interval is 50 ± 55.66 ; that is, -5.66 to 105.66. Here's how to interpret this confidence interval. Suppose we repeated this study with different random samples for school A and school B. Based on the confidence interval, we would expect the observed difference in sample means to be between -5.66 and 105.66 90% of the time.



CORE CONCEPT: FUNDAMENTAL OF CORRELATION

Correlations are useful because if you can find out what relationship variables have, you can make predictions about future behavior. Knowing what the future holds is very important in the

social sciences like government and healthcare. Businesses also use these statistics for budgets and business plans.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5" is less than the average weight of people 5'6", and their average weight is less than that of people 5'7", etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

Although this correlation is fairly obvious your data may contain unsuspected correlations. You may also suspect there are correlations, but don't know which are the strongest. An intelligent correlation analysis can lead to a greater understanding of your data.

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which

those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. When the fluctuation of one variable reliably predicts a similar fluctuation in another variable, there's often a tendency to think that means that the change in one causes the change in the other. However, correlation does not imply causation. There may be, for example, an unknown factor that influences both variables similarly.

Here's one example: A number of studies report a positive correlation between the amount of television children watch and the likelihood that they will become bullies. Media coverage often cites such studies to suggest that watching a lot of television causes children to become bullies. However, the studies only report a correlation, not causation. It is likely that some other factor – such as a lack of parental supervision – may be the influential factor.

The least-squares regression analysis method best suits prediction models and trend analysis. One may best use it in economics, finance, and stock markets, wherein the value of any future variable is predicted with the help of existing variables and the relationship between them.

Techniques in Determining Correlation

There are several different correlation techniques. The Survey System's optional Statistics Module includes the most common type, called the Pearson or product-moment correlation. The module also includes a variation on this type called **partial correlation**. The latter is useful when you want to look at the relationship between two variables while removing the effect of one or two other variables. Like all statistical techniques, correlation is only appropriate for certain kinds of data.

Partial correlation

is a method used to describe the relationship between two variables whilst taking away the effects of another variable, or several other variables, on this relationship.

Rating Scales

Rating scales are a controversial middle case. The numbers in rating scales have meaning, but that meaning isn't very precise. They are not like quantities. With a quantity (such as dollars), the difference between 1 and 2 is exactly the same as between 2 and 3. With a rating scale, that isn't really the case. You can be sure that your respondents think a rating of 2 is between a rating of 1 and a rating of 3, but you cannot be sure they think it is exactly halfway between. This is especially true if you labeled the mid-points of your scale (you cannot assume

“good” is exactly half way between “excellent” and “fair”). Most statisticians say you cannot use correlations with rating scales, because the mathematics of the technique assume the differences between numbers are exactly equal. Nevertheless, many survey researchers do use correlations with rating scales, because the results usually reflect the real world. Our own position is that you can use correlations with rating scales, but you should do so with care. When working with quantities, correlations provide precise measurements. When working with rating scales, correlations provide general indications.

Correlation Coefficient

The main result of a correlation is called the correlation coefficient (or r). It ranges from -1.0 to $+1.0$. The closer r is to $+1$ or -1 , the more closely the two variables are related.

If r is close to 0 , it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an “inverse” correlation).

While correlation coefficients are normally reported as $r =$ (a value between -1 and $+1$), squaring them makes then easier to understand. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r , ignore the decimal point. An r of $.5$ means 25% of the variation is related ($.5$ squared $= .25$). An r value of $.7$ means 49% of the variance is related ($.7$ squared $= .49$).

A correlation report can also show a second result of each test - statistical significance. In this case, the significance level will tell you how likely it is that the correlations reported may be due to chance in the form of random sampling error. If you are working with small sample sizes, choose a report format that includes the significance level. This format also reports the sample size.

A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable causes a change in another. Sales of personal computers and athletic shoes have both risen strongly in the last several years and there is a high correlation between them, but you cannot assume that buying computers causes people to buy athletic shoes (or vice versa).

The second caveat is that the Pearson correlation technique works best with linear relationships: as one variable gets larger, the other gets larger (or smaller) in direct proportion. It does not work well with curvilinear relationships (in which the relationship does not follow a straight line). An example of a curvilinear relationship is age and health care. They are related, but the relationship doesn't follow a straight line. Young children and older people both tend to use much more health care than teenagers or young adults. Multiple regression (also included in the Statistics Module) can be used to examine curvilinear relationships.

Interpret a Correlation Coefficient

Correlation coefficients measure the strength of association between two variables. The most common correlation coefficient, called the Pearson product-moment correlation coefficient, measures the strength of the linear association between variables. Generally, the correlation coefficient of a sample is denoted by r , and the correlation coefficient of a population is denoted by ρ or R .

The sign and the absolute value of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.

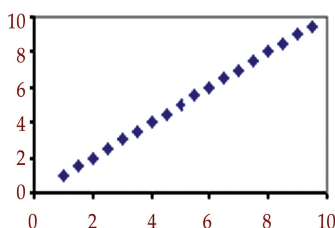
- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of a correlation coefficient, the stronger the linear relationship.
- The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.
- The weakest **linear relationship** is indicated by a correlation coefficient equal to 0.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

Linear relationship is a statistical term used to describe a straight-line relationship between two variables.

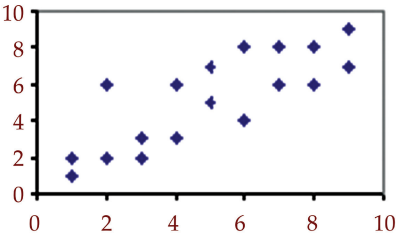
Keep in mind that the Pearson product-moment correlation coefficient only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero linear relationship. (It is possible for two variables to have zero linear relationship and a strong curvilinear relationship at the same time.)

Scatterplots and Correlation Coefficients

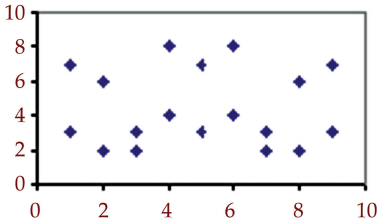
The scatterplots below show how different patterns of data produce different degrees of correlation.



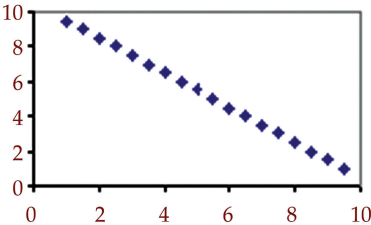
Maximum positive correlation ($r = 1.0$)



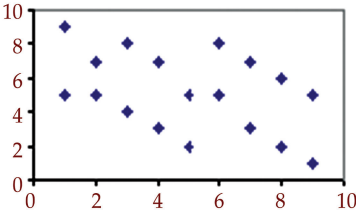
Strong positive correlation ($r = 0.80$)



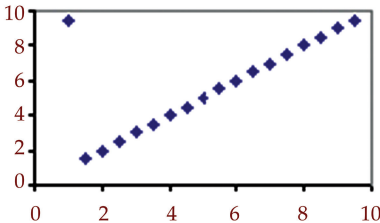
Zero correlation ($r = 0$)



Maximum negative correlation ($r = -1.0$)



Moderate negative correlation ($r = -0.43$)



Strong correlation & outlier ($r = 0.71$)

Several points are evident from the scatterplots.

- When the slope of the line in the plot is negative, the correlation is negative; and vice versa.
- The strongest correlations ($r = 1.0$ and $r = -1.0$) occur when data points fall exactly on a straight line.
- The correlation becomes weaker as the data points become more scattered.
- If the data points fall in a random pattern, the correlation is equal to zero.
- Correlation is affected by outliers. Compare the first scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 1.00 to 0.71).

Calculate a Correlation Coefficient

The most common formula for computing a product-moment correlation coefficient (r) is given below:

Product-moment correlation coefficient. The correlation r between two variables is:

$$r = \Sigma(xy) / \sqrt{[\Sigma x^2] * [\Sigma y^2]}$$

where Σ is the summation symbol, $x = x_i - \bar{x}$, x_i is the x value for observation i , $\bar{x}$ is the mean x value, $y = y_i - \bar{y}$, y_i is the y value for observation i , and $\bar{y}$ is the mean y value.

The formula below uses population means and population standard deviations to compute a population correlation coefficient (ρ) from population data.

Population correlation coefficient. The correlation ρ between two variables is:

$$\rho = [1 / N] * \Sigma \{[(X_i - \mu_x) / \sigma_x] * [(Y_i - \mu_y) / \sigma_y]\}$$

where N is the number of observations in the population, Σ is the summation symbol, X_i is the X value for observation i , μ_x is the population mean for variable X , Y_i is the Y value for observation i , μ_y is the population mean for variable Y , σ_x is the population standard deviation of X , and σ_y is the population standard deviation of Y .

The formula below uses sample means and sample standard deviations to compute a correlation coefficient (r) from sample data.

Sample correlation coefficient. The correlation r between two variables is:

$$r = [1 / (n - 1)] * \Sigma \{[(x_i - \bar{x}) / s_x] * [(y_i - \bar{y}) / s_y]\}$$

where n is the number of observations in the sample, Σ is the summation symbol, x_i is the x value for observation i , $\bar{x}$ is the sample mean of x , y_i is the y value for observation i , $\bar{y}$ is the sample mean of y , s_x is the sample standard deviation of x , and s_y is the sample standard deviation of y .

Each of the latter two formulas can be derived from the first formula. Use the first or second formula when you have data from the entire population. Use the third formula when you only have sample data, but want to estimate the correlation in the population. When in doubt, use the first formula.

Fortunately, you will rarely have to compute a correlation coefficient by hand. Many software packages (e.g., Excel) and most graphing calculators have a correlation function that will do the job for you.

Problem: A national consumer magazine reported the following correlations.

- The correlation between car weight and car reliability is -0.30 .
- The correlation between car weight and annual maintenance cost is 0.20 .

Which of the following statements are true?

- I. Heavier cars tend to be less reliable.
- II. Heavier cars tend to cost more to maintain.
- III. Car weight is related more strongly to reliability than to maintenance cost.

(A) I only

(B) II only

(C) III only

(D) I and II only

(E) I, II, and III

Solution: The correct answer is (E). The correlation between car weight and reliability is negative. This means that reliability tends to decrease as car weight increases. The correlation between car weight and maintenance cost is positive. This means that maintenance costs tend to increase as car weight increases.

The strength of a relationship between two variables is indicated by the absolute value of the correlation coefficient. The correlation between car weight and reliability has an absolute value of 0.30 . The correlation between car weight and maintenance cost has an absolute value of 0.20 . Therefore, the relationship between car weight and reliability is stronger than the relationship between car weight and maintenance cost.

For a statistical test to be legitimate, sampling and data collection must be done in a way that is meant to test your hypothesis. You cannot draw statistical conclusions about the population you are interested in if your data is not representative.

$$Y = mX + c$$

$$m = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{(\overline{x})^2 - \overline{x^2}}$$

$$c = \overline{y} - m \cdot \overline{x}$$

CORE CONCEPT: SIMPLE LINEAR REGRESSION

Simple linear regression is a method that enables you to determine the relationship between a continuous process output (Y) and one factor (X). The relationship is typically expressed in terms of a

mathematical equation such as $Y = b + mX$

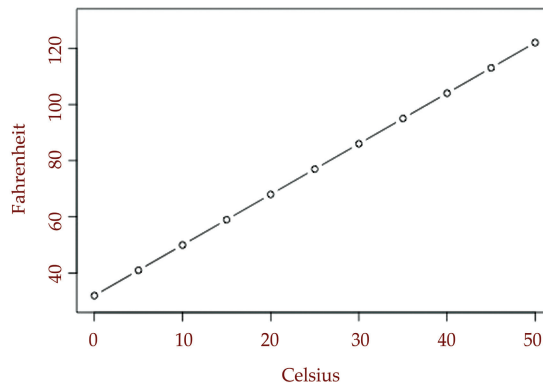
Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x , is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y , is regarded as the response, outcome, or dependent variable.

Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course. The other terms are mentioned only to make you aware of them should you encounter them in other arenas. Simple linear regression gets its adjective "simple," because it concerns the study of only one predictor variable. In contrast, multiple linear regression, which we study later in this course, gets its adjective "multiple," because it concerns the study of two or more predictor variables.

Types of relationships

Before proceeding, we must clarify what types of relationships we won't study in this course, namely, deterministic (or functional) relationships. Here is an example of a deterministic relationship.



Note that the observed (x , y) data points fall directly on a line. As you may remember, the relationship between degrees Fahrenheit and degrees Celsius is known to be:

$$\text{Fahr} = \frac{9}{5} \text{Cels} + 32$$

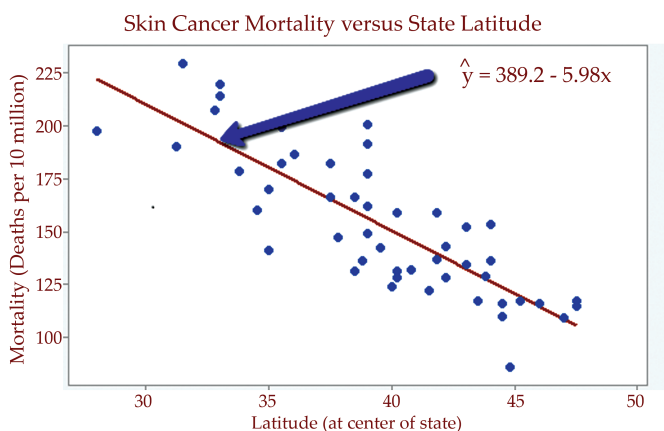
That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit exactly.

Here are some examples of other deterministic relationships:

- Circumference = $\pi \times$ diameter
- Hooke's Law: $Y = \alpha + \beta X$, where Y = amount of stretch in a spring, and X = applied weight.
- Ohm's Law: $I = V/r$, where V = voltage applied, r = resistance, and I = current.
- Boyle's Law: For a constant temperature, $P = \alpha/V$, where P = pressure, α = constant for each gas, and V = volume of gas.

For each of these deterministic relationships, the equation exactly describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in statistical relationships, in which the relationship between the variables is not perfect.

Here is an example of a statistical relationship. The response variable y is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable x is the latitude (degrees North) at the center of each of 49 states in the U.S. (skincancer.txt) (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states. And, Washington, D.C. is included in the data set even though it is not technically a state.)



You might anticipate that if you lived in the higher latitudes of the northern U.S., the less exposed you'd be to the harmful rays of the sun, and therefore, the less risk you'd have of death due to skin cancer. The scatter plot supports such a hypothesis.

There appears to be a negative linear relationship between latitude and mortality due to skin cancer, but the relationship is not perfect. Indeed, the plot exhibits some “trend,” but it also exhibits some “scatter.” Therefore, it is a statistical relationship, not a deterministic one.

Some other examples of statistical relationships might include:

- Height and weight — as height increases, you’d expect weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content — as alcohol consumption increases, you’d expect one’s blood alcohol content to increase, but not perfectly.
- Vital lung capacity and pack-years of smoking — as amount of smoking increases (as quantified by the number of pack-years of smoking), you’d expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
- Driving speed and gas mileage — as driving speed increases, you’d expect gas mileage to decrease, but not perfectly.

Variable of Linear Regression

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares **linear regression** is a method for predicting the value of a dependent variable Y , based on the value of an independent variable X .

Prerequisites for Regression

Simple linear regression is appropriate when the following conditions are satisfied.

- The dependent variable Y has a linear relationship to the independent variable X . To check this, make sure that the XY scatterplot is linear and that the residual plot shows a random pattern.
- For each value of X , the probability distribution of Y has the same standard deviation σ . When this condition is satisfied, the variability of the residuals will be relatively constant across all values of X , which is easily checked in a residual plot.
- For any given value of X ,

Linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X .

- The Y values are independent, as indicated by a random pattern on the residual plot.
- The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large. A histogram or a dot plot will show the shape of the distribution.

The Least Squares Regression Line

Linear regression finds the straight line, called the least squares regression line or LSRL, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is:

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the regression coefficient, X is the value of the independent variable, and Y is the value of the dependent variable.

Given a random sample of observations, the population regression line is estimated by:

$$\hat{y} = b_0 + b_1x$$

where b_0 is a constant, b_1 is the regression coefficient, x is the value of the independent variable, and $\hat{y}$ is the predicted value of the dependent variable.

How to Define a Regression Line

Normally, you will use a computational tool - a software package (e.g., Excel) or a graphing calculator- to find b_0 and b_1 . You enter the X and Y values into your program or calculator, and the tool solves for each parameter.

In the unlikely event that you find yourself on a desert island without a computer or a graphing calculator, you can solve for b_0 and b_1 «by hand». Here are the equations.

$$b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$$

$$b_1 = r * (s_y / s_x)$$

$$b_0 = \bar{y} - b_1 * \bar{x}$$

where b_0 is the constant in the regression equation, b_1 is the regression coefficient, r is the correlation between x and y, x_i is the X value of observation i, y_i is the Y value of observation i, $\bar{x}$ is the mean of X, $\bar{y}$ is the mean of Y, s_x is the standard deviation of X, and s_y is the standard deviation of Y

Properties of the Regression Line

When the regression parameters (b_0 and b_1) are defined as described above, the regression line has the following properties.

- The line minimizes the sum of squared differences between observed values (the y values) and predicted values (the $\hat{y}$ values computed from the regression equation).

- The regression line passes through the mean of the X values ($\bar{x}$) and through the mean of the Y values ($\bar{y}$).
- The regression constant (b_0) is equal to the y intercept of the regression line.
- The regression coefficient (b_1) is the average change in the dependent variable (Y) for a 1-unit change in the independent variable (X). It is the slope of the regression line.

The **least squares** regression line is the only straight line that has all of these properties.

Least squares

is a standard approach in regression analysis to the approximate solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns.

The Coefficient of Determination

The coefficient of determination (denoted by R^2) is a key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10% of the variance in Y is predictable from X; an R^2 of 0.20 means that 20% is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

Coefficient of determination. The coefficient of determination (R^2) for a linear regression model with one independent variable is:

$$R^2 = \{ (1/N) * \sum [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, $\bar{x}$ is the mean x value, y_i is the y value for observation i, $\bar{y}$ is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y.

If you know the linear correlation (r) between two variables, then the coefficient of determination (R^2) is easily computed using the following formula: $R^2 = r^2$.

Standard Error

The standard error about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts. The higher the coefficient of determination, the lower the standard error; and the more accurate predictions are likely to be.

Problem: A researcher uses a regression equation to predict home heating bills (dollar cost), based on home size (square feet). The correlation between predicted bills and home size is 0.70. What is the correct interpretation of this finding?

- 70% of the variability in home heating bills can be explained by home size.
- 49% of the variability in home heating bills can be explained by home size.
- For each added square foot of home size, heating bills increased by 70 cents.
- For each added square foot of home size, heating bills increased by 49 cents.
- None of the above.

Solution

The correct answer is (B). The coefficient of determination measures the proportion of variation in the dependent variable that is predictable from the independent variable. The coefficient of determination is equal to R^2 ; in this case, $(0.70)^2$ or 0.49. Therefore, 49% of the variability in heating bills can be explained by home size.

Hypothesis Test for Regression Slope

A hypothesis test to determine whether there is a significant linear relationship between an independent variable X and a dependent variable Y . The test focuses on the slope of the regression line

$$Y = B_0 + B_1X$$

where B_0 is a constant, B_1 is the slope (also called the regression coefficient), X is the value of the independent variable, and Y is the value of the dependent variable. If we find that the slope of the regression line is significantly different from zero, we will conclude that there is a significant relationship between the independent and dependent variables.

Test Requirements

The approach described the standard requirements for simple linear regression are met.

- The dependent variable Y has a linear relationship to the independent variable X .

- For each value of X , the probability distribution of Y has the same standard deviation σ .
- For any given value of X ,
 - The Y values are independent.
 - The Y values are roughly normally distributed (i.e., symmetric and unimodal). A little skewness is ok if the sample size is large.

The test procedure consists of four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

If there is a significant linear relationship between the independent variable X and the dependent variable Y , the slope will not equal zero.

$$H_0: B_1 = 0$$

$$H_a: B_1 \neq 0$$

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.

Analyze Sample Data

Using sample data, find the standard error of the slope, the slope of the regression line, the degrees of freedom, the test statistic, and the P-value associated with the test statistic.

- Standard error. Many statistical software packages and some graphing calculators provide the standard error of the slope as a regression analysis output. The table below shows hypothetical output for the following regression equation: $y = 76 + 35x$.

Predictor	Coef	SE Coef	T	P
Constant	76	30	2.53	0.01
X	35	20	1.75	0.04

In the output above, the standard error of the slope (shaded in gray) is equal to 20. In this example, the standard error is referred to as “SE Coeff”. However, other software packages might use a different label for the standard error. It might be “StDev”, “SE”, “Std Dev”, or something else. If you need to calculate the standard error of the slope (SE) by hand, use the following formula:

$$SE = s_{b_1} = \text{sqrt} [\Sigma(y_i - \hat{y}_i)^2 / (n - 2)] / \text{sqrt} [\Sigma(x_i - \bar{x})^2]$$

where y_i is the value of the dependent variable for observation i , $\hat{y}_i$ is estimated value of the dependent variable for observation i , x_i is the observed value of the independent variable for observation i , $\bar{x}$ is the mean of the independent variable, and n is the number of observations.

- Slope. Like the standard error, the slope of the regression line will be provided by most statistics software packages. In the hypothetical output above, the slope is equal to 35.
- Degrees of freedom. For simple linear regression (one independent and one dependent variable), the degrees of freedom (DF) is equal to:

$$DF = n - 2$$

where n is the number of observations in the sample.

- Test statistic. The test statistic is a t statistic (t) defined by the following equation.

$$t = b_1 / SE$$

where b_1 is the slope of the sample regression line, and SE is the standard error of the slope.

- P-value. The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a t statistic, use the t Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

Interpret Results

If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Problem: The local utility company surveys 101 randomly selected customers. For each survey participant, the company collects the following: annual electric bill (in dollars) and home size (in square feet). Output from a regression analysis appears below.

Regression equation: Annual bill = 0.55 * Home size + 15				
Predictor	Coef	SE Coef	T	P
Constant	15	3	5.0	0.00
Home size	0.55	0.24	2.29	0.01

Is there a significant linear relationship between annual bill and home size? Use a 0.05 level of significance.

Solution: The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.

H_0 : The slope of the regression line is equal to zero.

H_a : The slope of the regression line is not equal to zero.

If the relationship between home size and electric bill is significant, the slope will not equal zero.

- **Formulate an analysis plan.** For this analysis, the significance level is 0.05. Using sample data, we will conduct a linear regression t-test to determine whether the slope of the regression line differs significantly from zero.
- **Analyze sample data.** To apply the linear regression t-test to sample data, we require the standard error of the slope, the slope of the regression line, the degrees of freedom, the t statistic test statistic, and the P-value of the test statistic.

We get the slope (b_1) and the standard error (SE) from the regression output.

$$b_1 = 0.55 \quad SE = 0.24$$

We compute the degrees of freedom and the t statistic test statistic, using the following equations.

$$DF = n - 2 = 101 - 2 = 99$$

$$t = b_1 / SE = 0.55 / 0.24 = 2.29$$

where DF is the degrees of freedom, n is the number of observations in the sample, b_1 is the slope of the regression line, and SE is the standard error of the slope.

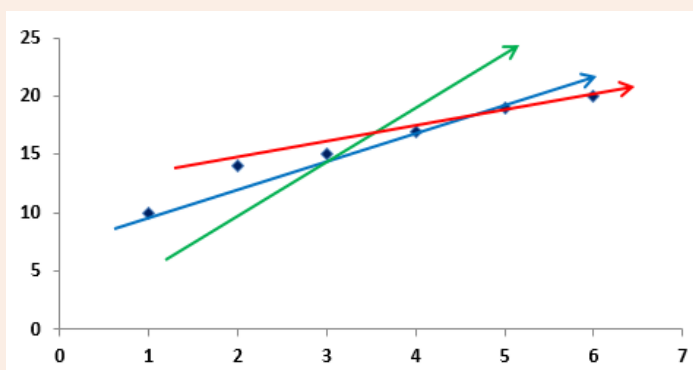
Based on the t statistic test statistic and the degrees of freedom, we determine the P-value. The P-value is the probability that a t statistic having 99 degrees of freedom is more extreme than 2.29. Since this is a two-tailed test, “more extreme” means greater than 2.29 or less than -2.29. We use the t Distribution Calculator to find $P(t > 2.29) = 0.0121$ and $P(t < -2.29) = 0.0121$. Therefore, the P-value is $0.0121 + 0.0121$ or 0.0242.

- **Interpret results.** Since the P-value (0.0242) is less than the significance level (0.05), we cannot accept the null hypothesis.

A CONCEPT OF LINE OF BEST FIT IN THE LEAST SQUARE REGRESSION

The line of best fit is a straight line drawn through a scatter of data points that best represents the relationship between them.

Let us consider the following graph wherein a data set plot along the x and y-axis. These data points represent using the blue dots. Three lines are drawn through these points – a green, a red, and a blue line. The green line passes through a single point, and the red line passes through three data points. However, the blue line passes through four data points, and the distance between the residual points and the blue line is minimal compared to the other two lines.



In the above graph, the blue line represents the line of best fit as it lies closest to all the values and the distance between the points outside the line to the line is minimal (the distance between the residuals to the line of best fit – also referred to as the sums of squares of residuals). However, in the other two lines, the orange and the green, the distance between the residuals and the lines is greater than the blue line.

The least-squares method provides the closest relationship between the dependent and independent variables by minimizing the distance between the residuals, and the line of best fit, i.e., the sum of squares of residuals is minimal under this approach. Hence, the term “least squares.”

SUMMARY

- The correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its values can range from -1 to 1.
- The “mean” is the “average” you’re used to, where you add up all the numbers and then divide by the number of numbers.
- Statisticians use summary measures to describe the amount of variability or spread in a set of data.
- A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity.
- A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true.
- Statisticians use a confidence interval to express the degree of uncertainty associated with a sample statistic. A confidence interval is an interval estimate combined with a probability statement.
- A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another.
- Rating scales are a controversial middle case. The numbers in rating scales have meaning, but that meaning isn’t very precise. They are not like quantities.
- Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.
- The standard error about the regression line (often denoted by SE) is a measure of the average amount that the regression equation over- or under-predicts.

MULTIPLE CHOICE QUESTIONS

1. **The Pearson correlation coefficient (r) ranges from:**
 - a. -1 to 1
 - b. 0 to 1
 - c. $-\infty$ to ∞
 - d. 1 to 100
2. **Spearman's rank correlation coefficient is used to:**
 - a. Measure linear relationships
 - b. Measure non-linear relationships
 - c. Measure monotonic relationships
 - d. Measure causal relationships
3. **What does a negative correlation coefficient (r) indicate?**
 - a. No correlation
 - b. Perfect positive correlation
 - c. Perfect negative correlation
 - d. As one variable increases, the other tends to decrease
4. **Kendall's Tau is primarily used to:**
 - a. Measure linear relationships
 - b. Measure non-linear relationships
 - c. Measure rank-based relationships
 - d. Measure categorical relationships
5. **A strong correlation coefficient (r) close to -1 indicates:**
 - a. Strong positive correlation
 - b. Weak positive correlation
 - c. No correlation
 - d. Strong negative correlation

Answer

1. (a) 2. (c) 3. (d) 4. (c) 5. (d)

REVIEW QUESTIONS

1. Define the central tendency of mean and median.
2. Explain how to measure variability in statistics.
3. What is hypothesis test and how it differ from proportions?
4. Identify useful techniques for determining the correlation.
5. Explain the variable of linear regression.
6. Illustrate the hypothesis test for regression slope.

REFERENCES

1. Agresti, A., & Franklin, C. A. (2018). *Statistics: The Art and Science of Learning from Data*. Pearson.
2. Bluman, A. G. (2019). *Elementary Statistics: A Step by Step Approach*. McGraw-Hill Education.
3. Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W. W. Norton & Company.
4. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
5. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
6. Moore, D. S., Notz, W. I., & Fligner, M. A. (2018). *The Basic Practice of Statistics*. W. H. Freeman.
7. Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press.
8. Triola, M. F. (2019). *Elementary Statistics*. Pearson.
9. Weisberg, S. (2005). *Applied Linear Regression*. Wiley.

A collage of various statistical charts and graphs, including bar charts, line graphs, and pie charts, with a magnifying glass focusing on one of the bar charts. The charts are labeled with figures like Fig. 1, Fig. 5, and Fig. 8. A clipboard is also visible in the top left.

CHAPTER 3

STATISTICAL INFERENCE



Statistical Inference is the process of drawing conclusions about a population based on sample data. It's a fundamental aspect of statistics that enables us to make predictions, estimate parameters, and test hypotheses using information from a subset of the larger population. Statistical inference involves making informed decisions and generalizations while accounting for uncertainty and variability in the data.

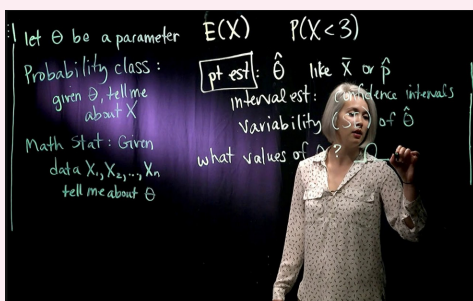
After studying this chapter, you will understand the core concepts of:

- Point Estimation
- Statistical Inference: One and two Sample Problems
- Kolmogorov-Simonov test
- Discrete and Continuous Data
- Frequency Distributions

INTRODUCTION

Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics. Hypothesis testing and confidence intervals are the applications of the statistical inference. Statistical inference is a method of making decisions about the parameters of a population, based on random sampling. It helps to assess the relationship between the dependent and independent variables. The purpose of statistical inference to estimate the uncertainty or sample to sample variation. It allows us to provide a probable range of values for the true values of something in the population. The components used for making statistical inference are:

- Sample Size
- Variability in the sample
- Size of the observed differences



CORE CONCEPT: POINT ESTIMATION

Point estimation, in statistics, the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population. The accuracy of any particular

approximation is not known precisely, though probabilistic statements concerning the accuracy of such numbers as found over many experiments can be constructed. For a contrasting estimation method, see interval estimation.

Recall that a parameter is a numerical characteristic of the population and a statistic is a numerical characteristic of random sample from the population of interest. A statistic is any function of **random variables**.

Estimator: Any statistic (function of random variables) whose values are used to estimate θ

Estimate: A value of estimator calculated from particular sample

Example: Sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, where $X_1, X_2, \dots, X_n$ are random variables is an estimator of a population mean μ and a $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ calculated from particular inge $x_1, x_2, \dots, x_n$ is an estimate of the population mean μ .

A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ . A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data.

In probability and statistics, a **random variable**, random quantity, aleatory variable, or stochastic variable is a variable whose possible values are outcomes of a random phenomenon.

- Point estimate results in a number that we consider as a plausible value of θ .
- The symbol $\hat{\theta}$ (“theta hat”) is used to denote both the estimator of θ and the point estimate resulting from a given sample.
- Example 1: Consider the following 20 observations on dielectric breakdown voltage for pieces of epoxy resin

24.46 25.61 26.25 26.42 26.66 27.15 27.31 27.54
27.74 27.94
27.98 28.04 28.28 28.49 28.50 28.87 29.11 29.13
29.50 30.88



Suppose that the distribution of dielectric breakdown voltage for pieces of epoxy resin follows normal distribution with unknown mean μ and unknown variance σ^2 .

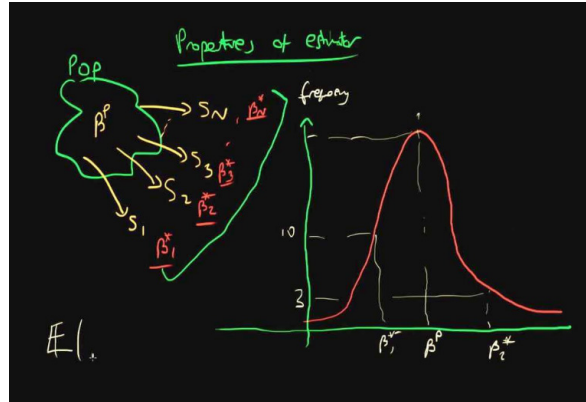
A point estimator of the unknown mean μ is $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and the point estimate calculated from the above observations using this point estimator is $\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{20} x_i = (24.46 + 25.61 + 26.25 + \dots + 29.50 + 30.88)/20 = 27.93$.

A point estimator of the unknown variance σ^2 is $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and the point estimate calculated from the above observations using this point estimator is $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{20} (x_i - \bar{x})^2 = \frac{1}{20-1} \{(24.46 - 27.93)^2 + (25.61 - 27.93)^2 + (26.25 - 27.93)^2 + \dots + (29.50 - 27.93)^2 +$

$$(30.88 - 27.93)^2\} = 2.14.$$

Properties of Point Estimator

There might be several criteria for finding a “best” estimator among many possible estimators. “Unbiased” and “minimum variance” properties are the two main criteria for a best estimator.



Unbiased estimator

A point estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ .

If $\hat{\theta}$ is not unbiased estimator, we call it biased estimator. That is, the difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.

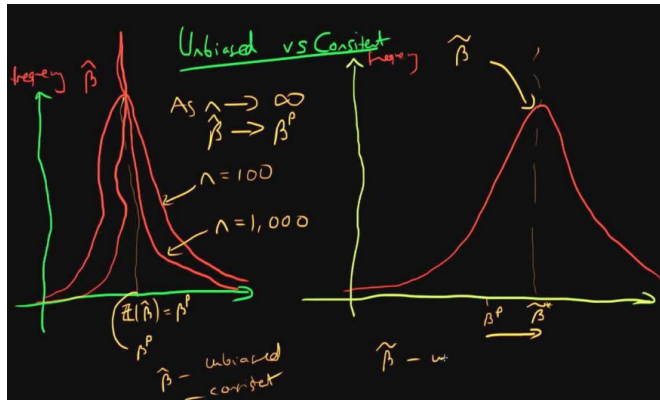
In statistics, the term **population** is used to describe the subjects of a particular study—everything or everyone who is the subject of a statistical observation.

- (Example 2) Let $X_1, X_2, \dots, X_n$ be a random sample from a **population** with mean μ and variance σ^2 . Then, the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is an “unbiased” estimator of the population mean μ and the sample variance $\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an “unbiased” estimator of the population variance σ^2 .

(sol) First we show that $E(\bar{X}) = \mu$.

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n E(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu, \text{ since } E(X_i) = \mu \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{n} \cdot n\mu \\
 &= \mu
 \end{aligned}$$



For the unbiasedness of S^2 ,

$$\begin{aligned}
 E(S^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\
 &= \frac{1}{n-1} E\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\
 &= \frac{1}{n-1} [E\left(\sum_{i=1}^n X_i^2\right) - nE(\bar{X}^2)] \\
 &= \frac{1}{n-1} [\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)] \\
 &= \frac{1}{n-1} [\sum_{i=1}^n (\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)] \\
 &= \frac{1}{n-1} [n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2] \\
 &= \frac{1}{n-1} [n\sigma^2 - \sigma^2] \\
 &= \frac{1}{n-1} [(n-1) \cdot \sigma^2] \\
 &= \sigma^2
 \end{aligned}$$

Therefore, $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2 , respectively

Minimum Variance Unbiased Estimation Among all estimators of θ that are unbiased, the one $\hat{\theta}$ that has minimum variance is called the minimum variance unbiased estimator (MVUE) of θ .

- (Example 3) Let $X_1, X_2, \dots, X_n$ be a random sample from a normal distribution with parameters μ and σ . Then the estimator $\hat{\mu} = \bar{X}$ is the MVUE for μ .

Standard error is the standard deviation of a statistic. When we report the value of a point estimate from sample, some indication of its precision should be given. The usual measure of precision is the standard error of the estimator used.

- (Example 4) In example 3, the standard error of $\bar{X}$ is $\sqrt{V(\bar{X})} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.



CORE CONCEPT: STATISTICAL INFERENCE: ONE AND TWO SAMPLE PROBLEMS

A one-sample t-test helps determine whether μ (the population mean) is equal to a hypothesized value (the test mean). The test uses the standard deviation of the sample to estimate σ (the population standard deviation). A two-sample t-test helps determine whether two population means are different. The test uses the sample standard deviations to estimate σ for each population.

Inference, in statistics, the process of drawing conclusions about a parameter one is seeking to measure or estimate.

Statistical inference means drawing conclusions based on data.

There are a many contexts in which **inference** is desirable:

- Exploratory data analysis (EDA).
- Statistical inference
- Nonparametric methods.

Exploratory data analysis (EDA)

Motivation: the classical methods of statistical inference depend heavily on the assumption that the data is outlier-free and nearly normal, and that the data is serially uncorrelated. Exploratory data analysis (EDA) uses graphical displays to help you obtain an understanding of whether or not such assumptions hold. The basic graphical exploratory data analysis (EDA) include

- histogram and the boxplot: used to get the shape of the distribution and to detect the presence of the outlier.

- normal qq-plot: used to check if the normal assumption is sensible and also used to detect the outlier.
- time series plot and autocorrelation function (ACF) plot: used to check for the serial correlation.

Now, we introduce how the qq-plot can be obtained via a simple example.

Example:

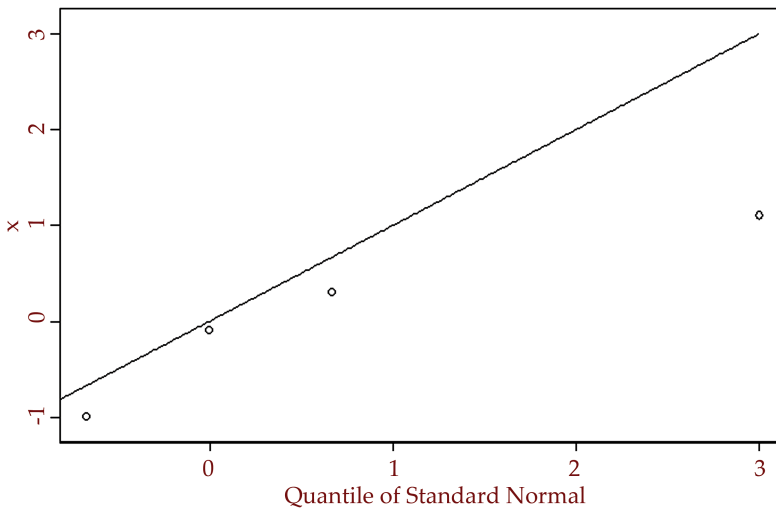
$n=4$

x	-1.0	-0.1	0.3	1.1
i	1	2	3	4
Empirical CDF (i/n)	0.25	0.5	0.75	1

Therefore,

Percent	25	50	75	100
Percentile of x	-1.0	-0.1	0.3	1.1
Percentile of $N(0,1)$	-0.67	0	0.67	3

Intuitively, if **x** is nearly standard normal, the percentile of **x** should be very close to the corresponding percentile of the standard normal random variable. Then, if we plot these percentiles of the standard normal variable in the above table versus the corresponding percentile of **x**, the points $(-0.67, -1)$, $(0, -0.1)$, $(0.67, 0.3)$, and $(3, 1.1)$ should fall around the line $y=x$.



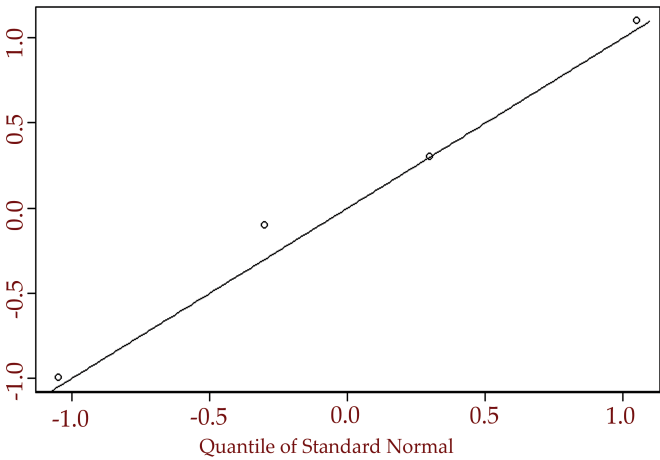
In S-plus, the above process has been modified. As $n \leq 10$, the empirical CDF formula used in S-plus is $\frac{(i-3/8)}{(n+1/4)}$. As $n > 10$, the empirical CDF is $\frac{(i-1/2)}{n}$. Therefore, in S-plus,

x	-1.0	-0.1	0.3	1.1
i	1	2	3	4
Empirical CDF	0.147	0.382	0.617	0.852

Thus,

Percent	14.7	38.2	61.7	85.2
Percentile of x	-1.0	-0.1	0.3	1.1
Percentile of N(0,1)	-1.04	-0.29	0.29	1.04

The plot of the percentile of $N(0,1)$ versus the percentile of **x** is



In S-plus, the `qqnorm(x)` will generate the same plot as the one of the percentile of $N(0,1)$ versus the percentile of **x** as the plot. The following S-plus commands can compare the plot generated by `qqnorm(x)` directly and the plot using the above process.

Example (Splus):

```
>par(mfrow=c(2,1))
>x<-c(-1.0,-0.1,0.3,1.1)
>qqnorm(x) # the qq-normal plot
>lines(seq(-1.1,1.1,by=0.01),seq(-1.1,1.1,by=0.01)) # the line y=x
>px<-ppoints(x) # px: the empirical CDF for x
>qn<-qnorm(px) # the percentile for the standard normal
>plot(qn,x) # distribution. The plot is the same as the one
> lines(seq(-1.1,1.1,by=0.01),seq(-1.1,1.1,by=0.01)) # generated by qqnorm(x)
```

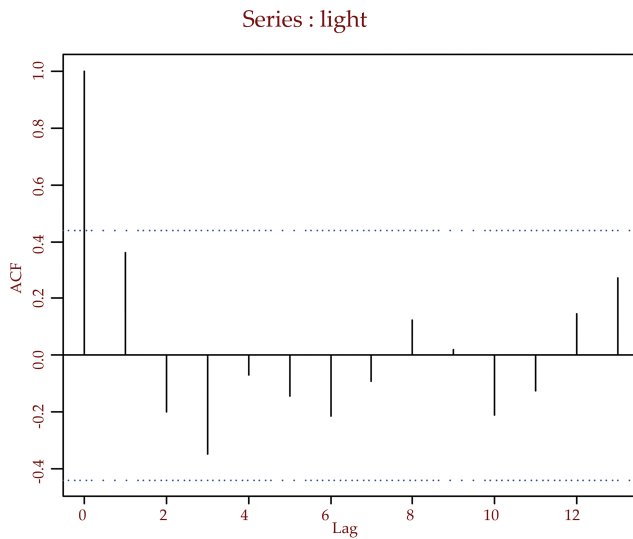
We now introduce the autocorrelation function (ACF) which is important tool for describing the serial (or temporal) dependence structure. Suppose we have the data $x^1, x^2, \dots, x^n$ generated by the random variables $X^1, X^2, \dots, X^n$. The autocovariance function at lag k is define as $r(k)=Cov(X^1,X^{k+1})$. The estimated

autocovariance function at lag k is $\tilde{r}(k) = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{n}$ where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$. Note that $\tilde{r}(0) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ is the biased estimate of the population variance $r(0)$. The autocorrelation function at lag k

is then defined as $\rho(k) = \frac{\text{Cov}(X_1, X_{k+1})}{\sqrt{\text{Var}(X_1)\text{Var}(X_{k+1})}} = \frac{r(k)}{r(0)}$. Therefore, a natural

estimate of $\rho(k)$ is $\tilde{\rho}(k) = \frac{\tilde{r}(k)}{\tilde{r}(0)} = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. Intuitively, if there is

no serial correlation, then, except $\tilde{\rho}(0) = 1$, the other estimated autocorrelation functions, $\tilde{\rho}(1), \tilde{\rho}(2), \dots$, should be very close to 0. Thus, the estimated **autocorrelation function** plot should look like



Autocorrelation function

of a signal is defined as the measure of similarity or coherence between a signal and its time delayed version.

On the other hand, if some $\tilde{\rho}(k)$ stands out, this might imply there is serial correlation at lag k .

We now use EDA method for the famous experiment data, measurements of the speed of light, conducted by the first American Nobel Prize winner A. A. Michelson.

Example (Splus):

```
>light<-c(850,740,900,1070,930,850,950,980,      # the ture
light=light
+980,880,1000,980,930,650,760,810,1000,1000,960,960)  #
+299000
>par(mfrow=c(2,2))
```

Distribution means to spread the product throughout the marketplace such that a large number of people can buy it.

```
>plot(light)
>hist(light,nclass=5)
>boxplot(light)
>qqnorm(light)
>qqline(light)
```

Note: the above plots seem to imply the distribution might not be normal. The **distribution** is skewed toward the left, but rather normal in the middle region by the histogram. 650 seems to be an outlier according to the box plot. qq-plot also indicates the data might not be normally distributed.

Statistical Inference (Parametric Test under Normal Assumption)

Two problems will be considered:

- a. One sample problem
- b. Two sample problem

(a) One Sample Problem:

$X^1, X^2, \dots, X^n$ i.i.d. normal random variables with mean μ and variance σ^2 . For testing $H_0: \mu = \mu_0$, the statistic, $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, can be used, where $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ and $s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ is the sample variance and n is the sample size. Theoretically, as H_0 is true, t is distributed as a t distribution with degrees of freedom $n-1$. Example (Splus):

```
>t.test(light,mu=990)           # test  $H_0: \mu = 990$  vs  $H_1: \mu \neq 990$ 

>t.test(light,conf.level=0.9,mu=990)   # 90% confidence interval can be obtained

>t.test(light,alternative="greater",mu=990) # test  $H_0: \mu = 990$  vs  $H_1: \mu > 990$ 

>help(t.test)
>qt(0.975,19)
```

(b) Two Sample Problem:

Suppose $X^1, X^2, \dots, X^n$ are i.i.d. normal random variables with mean μ_1 and variance σ_1^2 and $Y^1, Y^2, \dots, Y^n$ are i.i.d. random variables with mean μ_2 and variance σ_2^2 . X 's and Y 's are independent. To test $H^0: \mu_1 = \mu_2$, we first need to check if $H^0: \sigma_1^2 = \sigma_2^2$.

(i) Variance test:

For testing $H^0: \sigma_1^2 = \sigma_2^2$ the statistic $F = \frac{s_1^2}{s_2^2}$ can be used, where $s_1^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ and $s_2^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1}$ are the estimates of σ_1^2 and σ_2^2 , respectively, and where respectively, and $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ and $\bar{Y} = \frac{\sum_{i=1}^m Y_i}{m}$. Intuitively, if H^0 is true, F statistic should be close to 1. On the other hand, if F statistic has much larger value than 1, then this might imply H^0 is not true. Theoretically, as H^0 is true, F is distributed as F distribution with degrees of freedom $n-1$ and $m-1$.

Example (Splus):

Protein(H)	134	146	104	119	124	161	107	83	113	129	97	123
Protein(L)	70	118	101	85	107	132	94					

19 rats were divided into two groups, one group with 12 rats and the other with 7 rats. The larger group was given high protein food while the smaller group is given low protein food. The data of the weight gains for these rats under the two diets are given in above table. We now demonstrate the variance test using this data in S-Plus.

```
>rat1<-c(134,146,104,119,124,161,107,83,113,129,97,1123)
```

```
>rat2<-c(70,118,101,85,107,132,94)
```

```
>var.test(rat1,rat2) # H0:  $\sigma_1^2 = \sigma_2^2$  .vs H1:  $\sigma_1^2 \neq \sigma_2^2$ 
```

```
>var.test(rat1,rat2,conf.level=0.9,alt="g") # H0:  $\sigma_1^2 = \sigma_2^2$  .vs H1:  $\sigma_1^2 > \sigma_2^2$ 
```

```
>qf(0.95,11,6) #  $f_{11,7,0.95}$ 
```

(ii) T test:

As the variance test indicates $\sigma_1^2 = \sigma_2^2$, then for testing $H^0: \mu_1 = \mu_2$ the following

statistic can be used $t = \frac{\bar{X} - \bar{Y}}{\sqrt{(\frac{1}{n} + \frac{1}{m})s_p^2}}$, where $s_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n+m-2}$ is the pooled estimate of $\sigma_1^2 = \sigma_2^2$

Nonparametric Methods

Both the variance test and t-test used in depend heavily on the common normal assumption of the observations. In addition, these tests are very sensitive to some

unusual observations, such as outliers. However, the conclusions drawn from the nonparametric methods are not much affected if the normal assumption holds or some outliers exist.

(i) *One sample: problem:*

$X^1, X^2, \dots, X^n$ are random variables with common distribution (not necessarily normal distribution). Suppose M is the median of the distribution. We want to test if $H_0: M = M_0$, where M_0 is some number. We now introduce the Wilcoxon sign rank test.

Wilcoxon sign rank test:

Let $Z_i = |X_i - M_0|$ and R_i is the rank of Z_i . Let $T^+ = \sum_{(X_i - M_0) > 0} R_i$ be the sum of all the ranks corresponding to $X_i - M_0 > 0$ while $T^- = \sum_{(X_i - M_0) < 0} R_i$ is the sum of all the ranks corresponding to $X_i - M_0 < 0$.

Example:

$H_0: M = M_0 = 8$

X_i	7.1	13.5	5.2	4.2	6.4	10.4	5.7	5.4	6.3	7.5
Z_i	0.9	5.5	2.8	3.8	1.6	2.4	2.3	2.6	1.7	0.5
R_i	2	10	8	9	3	6	5	7	4	1
Sign of $X_i - 8$	-	+	-	-	-	+	-	-	-	-

Then, $T^+ = 10 + 6 = 16, T^- = 2 + 8 + 9 + 3 + 5 + 7 + 4 + 1 = 39$.

Statisticians and data analysts often establish a confidence interval between 90% and 99%, as plausibility must be less than 100%. Typically, a 95% confidence interval is desirable when evaluating inferential data in statistics.



CORE CONCEPT: KOLMOGOROV-SMIRNOV TEST

In statistics, the Kolmogorov–Smirnov test (K–S test or KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test).

Let $X_1, X_2, \dots, X_n$ be a random sample from a population X . We hypothesized that the distribution of X is $F(x)$. Further, we wish to test our hypothesis. Thus our null **hypothesis** is

$$H_o : X \sim F(x)$$

We would like to design a test of this null hypothesis against the alternative

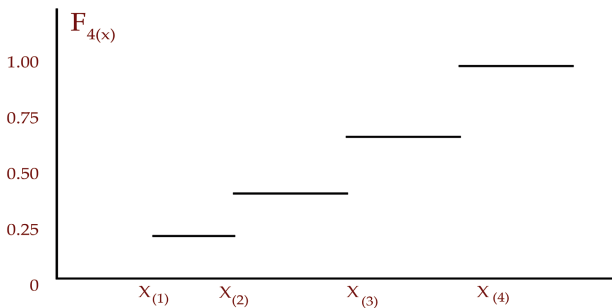
$$H_a : X \not\sim F(x)$$

A **hypothesis** (plural hypotheses) is a proposed explanation for a phenomenon.

In order to design a test, first of all we need a statistic which will un-biasedly estimate the unknown distribution $F(x)$ of the population X using the random sample $X_1, X_2, \dots, X_n$. Let $x_{(1)} < x_{(2)} < \dots < x_{(n)}$ be the observed values of the ordered statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. The empirical distribution of the random sample is defined as

$$F_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ \frac{k}{n} & \text{if } x_{(k)} \leq x < x_{(k+1)}, \quad \text{for } k = 1, 2, \dots, n-1 \\ 1 & \text{if } x_{(n)} \leq x. \end{cases}$$

The graph of the empirical distribution function $F_4(x)$ is shown follows.



Empirical Distribution Function

Figure 1: Empirical Distribution Function.

For a fixed value of x , the empirical distribution function can be considered as a random variable that takes on the values

$$0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, \frac{n}{n}.$$

First we show that $F_n(x)$ is an unbiased estimator of the population distribution $F(x)$. That is,

$$E(F_n(x)) = F(x)$$

for a fixed value of x . We need the probability density function of the random variable $F_n(x)$. From the definition of the empirical distribution we see that if exactly k observations are less than or equal to x , then

$$F_n(x) = \frac{k}{n}$$

which is

$$n F_n(x) = k.$$

The probability that an observation is less than or equal to x is given by $F(x)$.

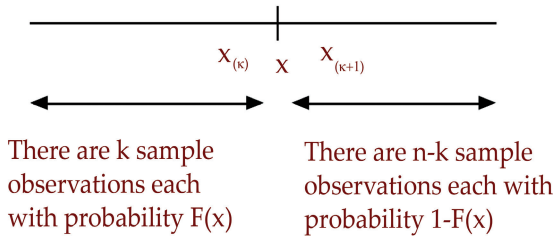


Figure 2: Distribution of the empirical distribution function.

Hence

$$\begin{aligned} P(n F_n(x) = k) &= P\left(F_n(x) = \frac{k}{n}\right) \\ &= \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} \end{aligned}$$

for $k = 0, 1, \dots, n$. Thus

$$n F_n(x) \sim \text{BIN}(n, F(x)).$$

Thus the expected value of the random variable $n F_n(x)$ is given by

$$E(n F_n(x)) = n F(x)$$

$$n E(F_n(x)) = n F(x)$$

$$E(F_n(x)) = F(x).$$

This shows that, for a fixed x , $F_n(x)$, on an **average**, equals to the population distribution function $F(x)$. Hence the empirical distribution function $F_n(x)$ is an unbiased estimator of $F(x)$.

The statistical hypothesis is called the null hypothesis and is typically stated as no effect or no difference, this is often opposite to the research hypothesis that motivated the study.

Since

$$n F_n(x) \sim BIN(n, F(x))$$

the variance of $n F_n(x)$ is given by

$$Var(n F_n(x)) = n F(x) [1 - F(x)].$$

Hence the variance of $F_n(x)$ is

$$Var(F_n(x)) = \frac{F(x) [1 - F(x)]}{n}$$

It is easy to see that $Var(F_n(x)) \rightarrow 0$ as $n \rightarrow \infty$ for all values of x . Thus the empirical distribution function $F_n(x)$ and $F(x)$ tend to be closer to each other with large n . As a matter of fact, Glivenko, a Russian mathematician, proved that $F_n(x)$ converges to $F(x)$ uniformly in x as $n \rightarrow \infty$ with probability one.

Because of the convergence of the empirical distribution function to the theoretical distribution function, it makes sense to construct a goodness of fit test based on the closeness of $F_n(x)$ and hypothesized distribution $F(x)$.

Let

$$D_n = \max_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

That is D_n is the maximum of all point wise differences $|F_n(x) - F(x)|$. The distribution of the Kolmogorov-Smirnov statistic, D_n can be derived. However, we shall not do that here as the derivation is quite involved. Instead, we give a closed form formula for $P(D_n \leq d)$. If $X_1, X_2, \dots, X_n$ is a sample from a population with continuous distribution function $F(x)$, then

$$P(D_n \leq d) = \begin{cases} 0 & \text{if } d \leq \frac{1}{2n} \\ n! \prod_{i=1}^n \int_{2^{i-d}}^{2^{i-\frac{1}{n}+d}} du & \text{if } \frac{1}{2n} < d < 1 \\ 1 & \text{if } d \geq 1 \end{cases}$$

where $d_u = d_{u_1} d_{u_2} \cdots d_{u_n}$ with $0 < u_1 < u_2 < \cdots < u_n < 1$. Further,

$$\lim_{n \rightarrow \infty} P(\sqrt{n} D_n \leq d) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2 k^2 d^2}$$

These formulas show that the distribution of the Kolmogorov-Smirnov statistic D_n is distribution free, that is, it does not depend on the distribution F of the population.

For most situations, it is sufficient to use the following approximations due to Kolmogorov:

In statistics, an **average** is defined as the number that measures the central tendency of a given set of numbers.

$$P(\sqrt{n} D_n \leq d) \approx 1 - 2e^{-2nd^2} \quad \text{for } d > \frac{1}{\sqrt{n}}.$$

If the null hypothesis $H_0: X \sim F(x)$ is true, the statistic D_n is small. It is therefore reasonable to reject H_0 if and only if the observed value of D_n is larger than some constant d_n .

If the level of significance is given to be α , then the constant d_n can be found from

$$\alpha = P(D_n > d_n / H_0 \text{ is true}) \approx 2e^{-2nd_n^2}.$$

This yields the following hypothesis test: Reject H_0 if $D_n \geq d_n$ where

$$d_n = \sqrt{-\frac{1}{2n} \ln \left(\frac{\alpha}{2} \right)}$$

The normal distribution can be used as an **approximation** to the binomial distribution, under certain circumstances, namely: If $X \sim B(n, p)$ and if n is large and/or p is close to $1/2$, then X is approximately $N(np, npq)$.

is obtained from the above Kolmogorov's **approximation**. Note that the approximate value of d_{12} obtained by the above formula is equal to 0.3533 when $\alpha = 0.1$, however more accurate value of d_{12} is 0.34.

Next we address the issue of the computation of the statistics D_n . Let us define

$$D_n^+ = \max_{x \in \mathbb{R}} \{F_n(x) - F(x)\}$$

and

$$D_n^- = \max_{x \in \mathbb{R}} \{F(x) - F_n(x)\}$$

Then it is easy to see that

$$D_n = \max\{D_n^+, D_n^-\}.$$

Further, since $F_n(x_{(i)}) = \frac{i}{n}$, it can be shown that

$$D_n^+ = \max \left\{ \max_{1 \leq i \leq n} \left[\frac{i}{n} - F(x_{(i)}) \right], 0 \right\}$$

and

$$D_n^- = \max \left\{ \max_{1 \leq i \leq n} \left[F(x_{(i)}) - \frac{i-1}{n} \right], 0 \right\}$$

Therefore it can also be shown that

$$D_n = \max_{1 \leq i \leq n} \left\{ \max \left[\frac{i}{n} - F(x_{(i)}), F(x_{(i)}) - \frac{i-1}{n} \right] \right\}$$

The following figure illustrates the Kolmogorov-Smirnov statistics D_n when $n = 4$.

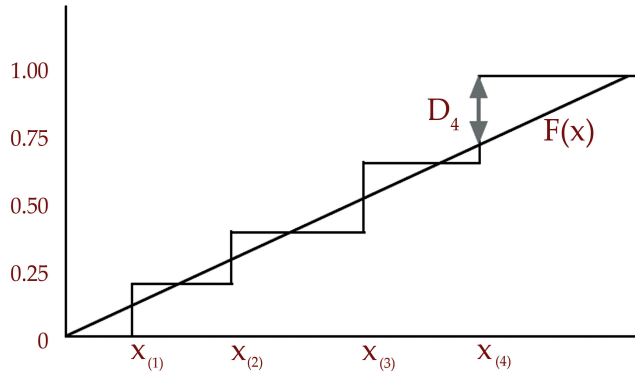
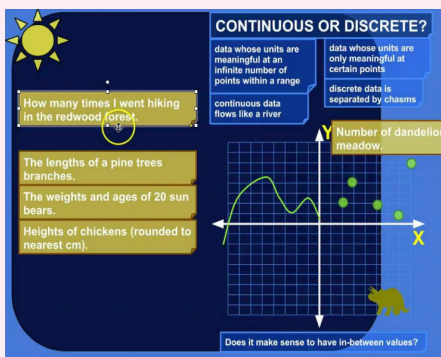


Figure 3: Kolmogorov-Smirnov Statistic.



CORE CONCEPT: DISCRETE AND CONTINUOUS DATA

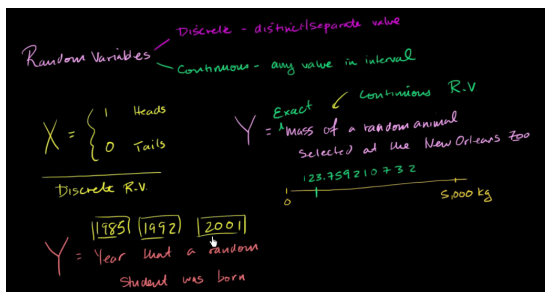
Discrete data take particular values, while continuous data are not restricted to separate values. Discrete data can take on only integer values whereas continuous data can take on any value.

Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be discrete or continuous, depending on the area of application.

For example, a chemical engineer may be interested in conducting an experiment that will lead to conditions where yield is maximized. Here, of course, the yield may be in percent or grams/pound, measured on a continuum. On the other hand, a toxicologist conducting a combination drug experiment may encounter data that are binary in nature (i.e., the patient either responds or does not).

Great distinctions are made between discrete and continuous data in the **probability theory** that allow us to draw statistical inferences. Often applications of statistical inference are found when the data are count data.

Probability theory is the branch of mathematics concerned with probability, the analysis of random phenomena.



For example, an engineer may be interested in studying the number of radioactive particles passing through a counter in, say, 1 millisecond. Personnel responsible for the efficiency of a port facility may be interested in the properties of the number of oil tankers arriving each day at a certain port city.

The probability distribution of a statistic is actually the sampling distribution.

Special attention even at this early stage of the textbook should be paid to some details associated with binary data. Applications requiring statistical analysis of binary data are voluminous. Often the measure that is used in the analysis is the sample proportion. Obviously the binary situation involves two categories. If there are n units involved in the data and x is defined as the number that fall into category 1, then $n - x$ fall into category 2. Thus, x/n is the sample proportion in category 1, and $1 - x/n$ is the sample proportion in category 2. In the biomedical application, 50 patients may represent the sample units, and if 20 out of 50 experienced an improvement in a stomach ailment (common to all 50) after all were given the drug, then $20/50 = 0.4$ is the sample proportion for which the drug was a success and $1 - 0.4 = 0.6$ is the sample proportion for which the drug was not successful. Actually the basic numerical measurement for binary data is generally denoted by either 0 or 1.

For example, in our medical example, a successful result is denoted by a 1 and a nonsuccess a 0. As a result, the sample proportion is actually a sample mean of the ones and zeros. For the successful category,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

Different Kinds of Problems are solved in Binary Data Situations

The kinds of problems facing scientists and engineers dealing in binary data are not a great deal unlike those seen where continuous measurements are of interest. However, different techniques are used since the statistical properties of sample proportions are quite different from those of the sample means that result from averages taken from continuous populations. The statistical problem underlying this illustration focuses on whether an intervention, say, an increase in curing temperature, will alter the population mean tensile strength associated with the silicone rubber process. On the

other hand, in a quality control area, suppose an automobile tire manufacturer reports that a shipment of 5000 tires selected randomly from the process results in 100 of them showing blemishes. Here the sample proportion is $100/5000 = 0.02$.

Following a change in the process designed to reduce blemishes, a second sample of 5000 is taken and 90 tires are blemished. The sample proportion has been reduced to $90/5000 = 0.018$. The question arises, “Is the decrease in the sample proportion from 0.02 to 0.018 substantial enough to suggest a real improvement in the population proportion?” Both of these illustrations require the use of the statistical properties of sample averages one from samples from a continuous population, and the other from samples from a discrete (binary) population. In both cases, the sample mean is an estimate of a population parameter, a population mean in the first illustration (i.e., mean tensile strength), and a population proportion in the second case (i.e., proportion of blemished tires in the population). So here we have sample estimates used to draw scientific conclusions regarding population parameters.



Statistical Modeling, Scientific Inspection, and Graphical Diagnostics

Often the end result of a statistical analysis is the estimation of parameters of a postulated model. This is natural for scientists and engineers since they often deal in modeling. A statistical model is not deterministic but, rather, must entail some probabilistic aspects. A model form is often the foundation of assumptions that are made by the analyst.

For example, in example 2 the scientist may wish to draw some level of distinction between the nitrogen and no-nitrogen populations through the sample information. The analysis may require a certain model for the data, for example, that the two samples come from normal or **Gaussian distributions**.

The **Gaussian distribution** is a continuous function which approximates the exact binomial distribution of events.

Obviously, the user of statistical methods cannot generate sufficient information or experimental data to characterize the population totally. But sets of data are often used to learn about certain properties of the population. Scientists and engineers are accustomed to dealing with data sets. The importance of characterizing or summarizing the nature of collections of data should be obvious. Often a summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken. The role of sampling and the display of data for enhancement of statistical inference is explored in detail. We merely introduce some simple but often effective displays that complement the study of statistical populations.

Scatter Plot

At times the model postulated may take on a somewhat complicated form. Consider, for example, a textile manufacturer who designs an experiment where cloth specimens that contain various percentages of cotton are produced.

Consider the data in following table:

Cotton Percentage	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Five cloth specimens are manufactured for each of the four cotton percentages. Both the model for the experiment and the type of analysis used should take into account the goal of the experiment and important input from the textile scientist. Some simple graphics can shed important light on the clear distinction between the samples. The sample means and variability are depicted nicely in the scatter plot. One possible goal of this experiment is simply to determine which cotton percentages are truly distinct from the others.

In other words, as in the case of the nitrogen/no-nitrogen data, for which cotton percentages are there clear distinctions between the populations or, more specifically, between the populations means? Perhaps a reasonable model is that each sample comes from a normal distribution. Here the goal is very much like that of the nitrogen/no-nitrogen data except that more samples are involved.

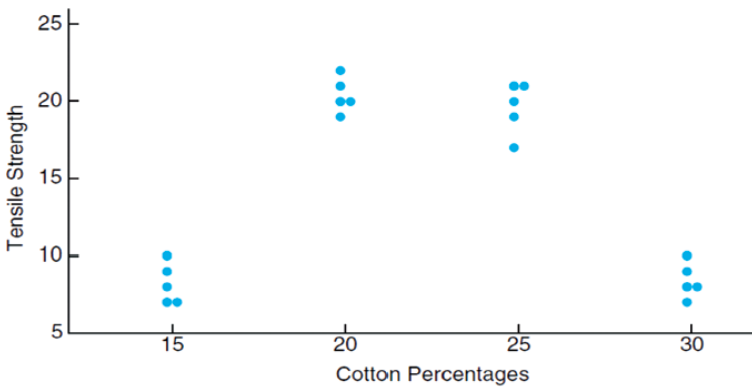
Incidentally, this formality is perhaps not necessary in light of the diagnostic plot. But does this describe the real goal of the experiment and hence the proper approach to data analysis? It is likely that the scientist anticipates the existence of a maximum population mean tensile strength in the range of cotton concentration in the experiment. Here the analysis of the data should revolve around a different type of model, one that postulates a type of structure relating the population means tensile strength to the cotton concentration. In other words, a model may be written:

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2$$

where $\mu_{t,c}$ is the population mean tensile strength, which varies with the amount of cotton in the product C. The implication of this model is that for a fixed cotton level, there is a population of tensile strength measurements and the population mean is $\mu_{t,c}$. This type of model, called a regression model. The functional form is chosen by the scientist.

At times the data analysis may suggest that the model be changed. Then the data analyst “entertains” a model that may be altered after some analysis is done. The use of an empirical model is accompanied by [estimation theory](#), where β_0 , β_1 , and β_2 are estimated by the data. Further, statistical inference can then be used to determine model adequacy.

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.



Two points become evident from the two data illustrations here:

- The type of model used to describe the data often depends on the goal of the experiment
- The structure of the model should take advantage of non-statistical scientific input.

A selection of a model represents a fundamental assumption upon which the resulting statistical inference is based. Often, plots can illustrate information that allows the results of the formal statistical inference to be better communicated to the scientist or engineer. At times, plots or exploratory data analysis can teach the analyst something not retrieved from the formal analysis. Almost any formal analysis requires assumptions that evolve from the model of the data. Graphics can nicely highlight violation of assumptions that would otherwise go unnoticed.

Stem-and-Leaf Plot

Statistical data, generated in large masses, can be very useful

for studying the behavior of the distribution if presented in a combined tabular and graphic display called a stem-and-leaf plot.

To illustrate the construction of a stem-and-leaf plot, consider the data of following table, which specifies the “life” of 40 similar car batteries recorded to the nearest tenth of a year. The batteries are guaranteed to last 3 years. First, split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf. The four stems 1, 2, 3, and 4 for our data are listed vertically on the left side in following table; the leaves are recorded on the right side opposite the appropriate stem value. Thus, the leaf 6 of the number 1.6 is recorded opposite the stem 1; the leaf 5 of the number 2.5 is recorded opposite the stem 2; and so forth. The number of leaves recorded opposite each stem is summarized under the frequency column.

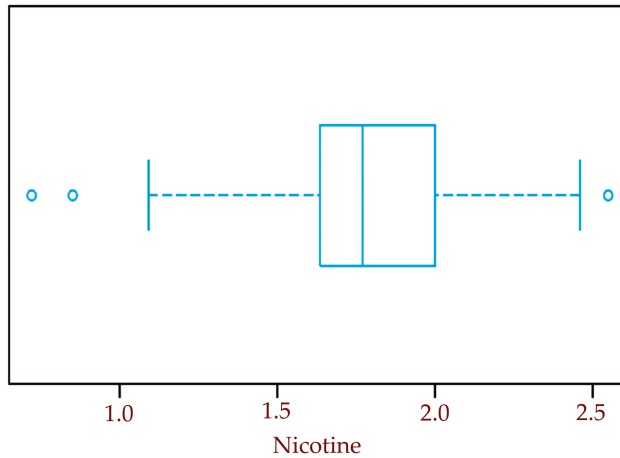
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

The stem-and-leaf plot of table contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time.

Box-and-Whisker Plot or Box Plot

Another display that is helpful for reflecting properties of a sample is the box and-whisker plot. This plot encloses the interquartile range of the data in a box that has the median displayed within. The interquartile range has as its extremes the 75th percentile (upper quartile) and the 25th percentile (lower quartile). In addition to the box, “whiskers” extend, showing extreme observations in the sample. For reasonably large samples, the display shows center of location, variability, and the degree of asymmetry. In addition, a variation called a box plot can provide the viewer with information regarding which observations may be outliers. Outliers are observations that are considered to be unusually far from the bulk of the data. There are many statistical tests that are designed to detect outliers. Technically, one may view an outlier as being an observation that represents a “rare event” (there is a small probability of obtaining a value that far from the bulk of the data).



The visual information in the box-and-whisker plot or box plot is not intended to be a formal test for outliers. Rather, it is viewed as a diagnostic tool. While the determination of which observations are outliers varies with the type of software that is used, one common procedure is to use a multiple of the interquartile range.

Other Distinguishing Features of a Sample

There are features of the distribution or sample other than measures of center of location and variability that further define its nature.



CORE CONCEPT: FREQUENCY DISTRIBUTIONS

Frequency distribution is a representation, either in a graphical or tabular format that displays the number of observations within a given interval.

The intervals must be mutually exclusive and exhaustive, and the interval size depends on the data being analyzed and the goals of the analyst.

Frequency distribution table can be presented in various ways, depending upon the usage and utilization of the data. Frequency distribution is divided into several kinds also due to nature of raw data. Much useful information can be inferred from the frequency distribution table; therefore, frequency distribution table can be presented in proper and useful manner.

Following are the various types.

Frequency Distribution for Discrete Data

The class limits in discrete data are the true class limits and there will be no class boundaries because discrete data are not in fractions. For example; following figures represents number of children born to 50 women in a certain locality up to the age of 40 years.

1	5	1	0	2	5
9	2	6	3	5	7
8	4	6	8	9	10
9	3	5	7	9	9
4	5	4	5	5	7
3	4	2	3	4	6
3	4	2	5	6	4
0	5	6	8	5	4
7	6				

The following Table shows the frequency distribution table for discrete data, taking the class interval size of 1.

Number of Children	Tally Marks	Number of Women
0	//	2
1	//	2
2	////	4
3	////	5
4	//// ///	8
5	//// ////	10
6	//// /	6
7	////	4
8	///	3
9	////	5
10	/	1
Total		50

Cumulative Frequency Distribution

Cumulative frequency is defined as a running total of frequencies.

Cumulative frequency distribution represents the sum of all succeeding up to certain class. The table showing the cumulative frequency is called cumulative frequency distribution or cumulative frequency distribution table or simply **cumulative frequency**. For example, referring Table, the cumulative frequency for class 120-129 is $1 + 4 = 5$. Similarly, the cumulative frequency of the class 130-139 is $1 + 4 + 17 = 22$. It

will be interpreted as there are 22 children who have weights less than 139.5 pounds.

The cumulative frequency is shown in the following Table.

Weights (lb)	Cumulative Frequency
Less than 109.5	0
Less than 119.5	$0 + 1 = 1$
Less than 129.5	$1 + 4 = 5$
Less than 139.5	$5 + 17 = 22$
Less than 149.5	$22 + 28 = 50$
Less than 159.5	$50 + 25 = 75$
Less than 169.5	$75 + 18 = 93$
Less than 179.5	$93 + 13 = 106$
Less than 189.5	$106 + 6 = 112$
Less than 199.5	$112 + 5 = 117$
Less than 209.5	$117 + 2 = 119$
Less than 219.5	$119 + 1 = 120$

Relative Frequency Distribution

The frequency of a class divided by the total frequency is called the relative frequency of that particular class. The frequency distribution table showing the relative frequencies is called relative frequency distribution or relative frequency or percentage table. Relative frequencies are generally expressed as a percentage. The sum of the relative frequencies of all the classes is 1 or 100%. For example; referring Table, the relative frequency of the class 160-169 is $18/120 \times 100 = 15\%$. The following Table gives the relative frequency distribution for the weight distribution of Table.

Weight (lb)	Relative Frequency
110-119	$1/120 = 0.0083$ or 0.83%
120-129	$4/120 = 0.0333$ or 3.33%
130-139	$17/120 = 0.1417$ or 14.17%
140-149	$28/120 = 0.2333$ or 23.33%
150-159	$25/120 = 0.2084$ or 20.84%
160-169	$18/120 = 0.15$ or 15%
170-179	$13/120 = 0.1083$ or 10.83%
180-189	$6/120 = 0.05$ or 5%
190-199	$5/120 = 0.0417$ or 4.17%
200-209	$2/120 = 0.0167$ or 1.67%
210-219	$1/120 = 0.0083$ or 0.83%

Relative Cumulative Frequency Distribution

The cumulative frequency of a class divided by the total frequency is called relative cumulative frequency. It is also called percentage cumulative frequency since it is expressed in percentage. The table showing relative cumulative frequencies is called the relative cumulative frequency distribution or percentage cumulative frequency distribution. For example, referring Table, the relative cumulative frequency of weight less than 159.5 is $75/120 \times 100 = 62.5\%$. it means that 62.5% of the students have weight less than 159.5 pounds.

Weight (lb)	Relative Cumulative Frequency
Less than 109.5	0%
Less than 119.5	$1/120 = 0.0083$ or 0.83%
Less than 129.5	$5/120 = 0.0417$ or 4.17%
Less than 139.5	$22/120 = 0.1833$ or 18.33%
Less than 149.5	$50/120 = 0.4167$ or 41.67%
Less than 159.5	$75/120 = 0.6250$ or 62.5%
Less than 169.5	$93/120 = 0.7750$ or 77.5%
Less than 179.5	$106/120 = 0.8833$ or 88.33%
Less than 189.5	$112/120 = 0.9333$ or 93.33%
Less than 199.5	$117/120 = 0.9750$ or 97.5%
Less than 209.5	$119/120 = 0.9917$ or 99.17%
Less than 219.5	$120/120 = 1$ or 100%

Bivariate Frequency Distribution

So far we have considered frequency distributions which involved only one variable. Such frequency distributions are called uni-variable frequency distribution because they involve only one variable. We can also construct a distribution taking two variables at a time. The frequency distribution involving two variables is called bivariate frequency distribution or bivariate frequency table or simply bivariate distribution or bivariate table. For example; in the data provided below we have the heights in inches and weighs in pounds of 50 students at a certain college.

Height (inches)	60	62	61	70	64	60	65	65	73	71
Weight (lb)	100	105	104	115	110	102	110	108	119	118
Height (inches)	61	60	63	64	67	68	69	64	66	62
Weight (lb)	109	108	107	112	115	117	117	111	113	104
Height (inches)	63	67	71	70	68	68	71	64	63	68
Weight (lb)	108	108	116	110	114	116	119	107	108	105
Height (inches)	73	69	64	67	67	64	62	67	62	64

Weight (lb)	119	107	115	111	114	108	105	117	105	107
Height (inches)	65	66	67	68	61	64	65	67	66	69
Weight (lb)	108	116	118	115	104	108	109	113	113	115

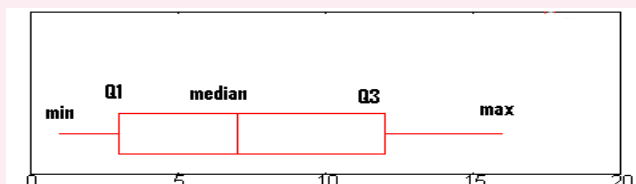
From the above data we will prepare frequency distribution, taking the class interval of size 3 for heights and a class interval of size 5 pounds for weights. We will arrange the class limits for heights in columns and those of weights in rows as provided in Table below. The classification of data will be done by taking pair of values of two variables and a tally mark will be marked in a cell lying at the intersection of appropriate class of the two variables. For example; the tally mark for the height 60 inches and weight of 100 pounds will be marked at the intersection of the classes 60-62 for heights and 100-104 for weights.

Weight (lb)	Height (inches)				
	60-62	63-65	66-68	69-71	72-74
100-104	###				
105-109	###	### ###	//	/	
110-114		////	### /	/	
115-119		/	### //	### /	//

Weight (lb)	Height (inches)					
	60-62	63-65	66-68	69-71	72-74	Total
100-104	5					5
105-109	5	10	2	1		18
110-114		4	6	1		11
115-119		1	7	6	2	16
Total	10	15	15	8	2	50

A CONCEPT OF BOXPLOT

A boxplot, also called a box and whisker plot, is a way to show the spread and centers of a data set. Measures of spread include the interquartile range and the mean of the data set. Measures of center include the mean or average and median (the middle of a data set).



The box and whiskers chart shows you how your data is spread out. Five pieces of information (the “five number summary”) are generally included in the chart:

- The minimum (the smallest number in the data set). The minimum is shown at the far left of the chart, at the end of the left “whisker.”
- First quartile, Q1, is the far left of the box (or the far right of the left whisker).
- The median is shown as a line in the center of the box.
- Third quartile, Q3, shown at the far right of the box (at the far left of the right whisker).
- The maximum (the largest number in the data set), shown at the far right of the box.

How to Read a Box Plot

A boxplot is a way to show a five number summary in a chart. The main part of the chart (the “box”) shows where the middle portion of the data is: the interquartile range. At the ends of the box, you” find the first quartile (the 25% mark) and the third quartile (the 75% mark). The far left of the chart (at the end of the left “whisker”) is the minimum (the smallest number in the set) and the far right is the maximum (the largest number in the set). Finally, the median is represented by a vertical bar in the center of the box.

Box plots aren’t used that much in real life. However, they can be a useful tool for getting a quick summary of data.

Step 1: Find the minimum.

The minimum is the far left hand side of the graph, at the tip of the left whisker. For this graph, the left whisker end is at approximately 0.75.

Step 2: Find Q1, the first quartile.

Q1 is represented by the far left hand side of the box. In this case, about 2.5.

Step 3: Find the median.

The median is represented by the vertical bar. In this boxplot, it can be found at about 6.5.

Step 4: Find Q3, the third quartile.

Q3 is the far right hand edge of the box, at about 12 in this graph.

Step 5: Find the maximum.

The maximum is the end of the “whiskers”: in this graph, at approximately 16.

SUMMARY

- Statistical inference is the process of analysing the result and making conclusions from data subject to random variation. It is also called inferential statistics.
- A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ . A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data.
- Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be discrete or continuous, depending on the area of application.
- The statistical problem underlying this illustration focuses on whether an intervention, say, an increase in curing temperature, will alter the population mean tensile strength associated with the silicone rubber process.
- A statistical model is not deterministic but, rather, must entail some probabilistic aspects. A model form is often the foundation of assumptions that are made by the analyst.
- Cumulative frequency distribution represents the sum of all succeeding up to certain class. The table showing the cumulative frequency is called cumulative frequency distribution or cumulative frequency distribution table or simply cumulative frequency.
- The cumulative frequency of a class divided by the total frequency is called relative cumulative frequency. It is also called percentage cumulative frequency since it is expressed in percentage.

MULTIPLE CHOICE QUESTIONS

1. **What is the primary goal of statistical inference?**
 - a. Describing data patterns
 - b. Drawing conclusions about a population based on sample data
 - c. Collecting raw data
 - d. Creating graphs
2. **What does point estimation involve?**
 - a. Providing a range of values within which a parameter is likely to fall
 - b. Calculating a test statistic
 - c. Providing a single value as an estimate of a parameter
 - d. Comparing sample data to a population distribution
3. **A confidence interval is used for:**
 - a. Hypothesis testing
 - b. Estimating population parameters
 - c. Comparing two sample means
 - d. Ranking data
4. **What is the central limit theorem?**
 - a. The process of calculating a test statistic
 - b. The distribution of sample data
 - c. The distribution of a statistic across all possible samples from a population
 - d. The distribution of data in a population
5. **Type I error in hypothesis testing occurs when:**
 - a. A true null hypothesis is rejected
 - b. A false null hypothesis is rejected
 - c. A true alternative hypothesis is rejected
 - d. A false alternative hypothesis is rejected

Answers

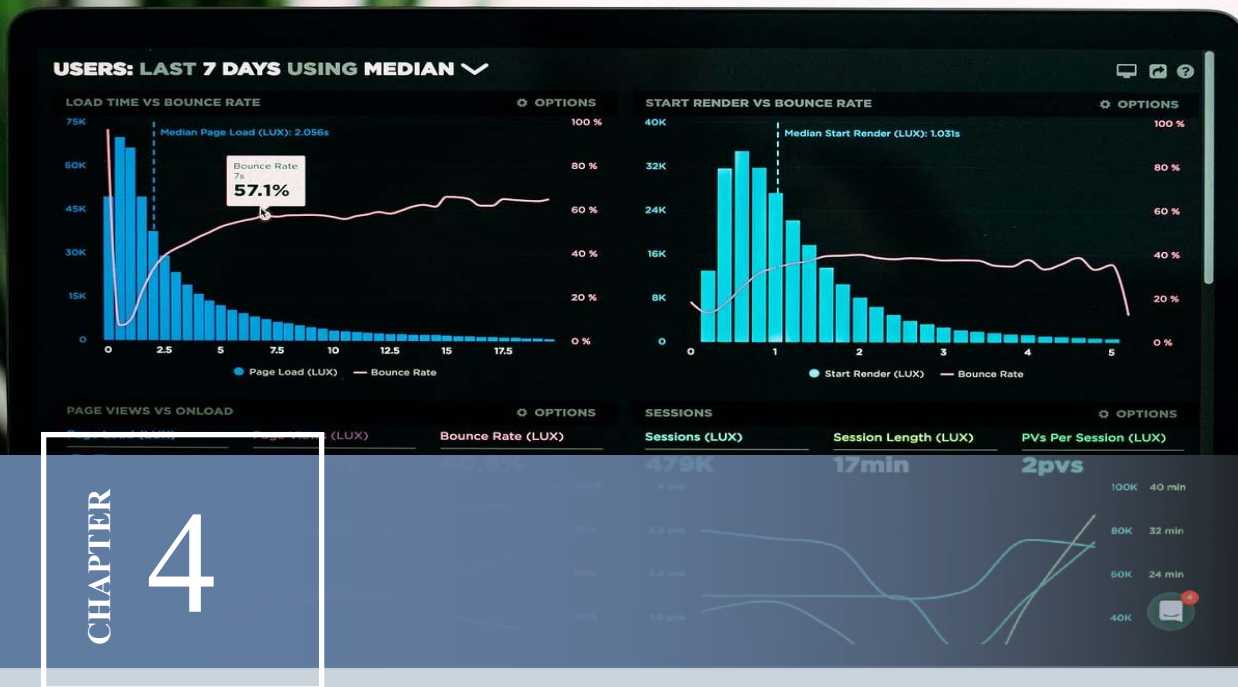
1. (b) 2. (c) 3. (b) 4. (c) 5. (a)

REVIEW QUESTIONS

1. What is statistical inference?
2. Describe the properties of point estimator.
3. Discuss about data collection and descriptive statistics.
4. Differentiate between inferential statistics and probability models.
5. Focus on the populations and samples in data analysis.
6. What is the variability in scientific data?
7. Explain simple random sampling.
8. Discuss about exploratory data analysis (EDA).
9. Explain the nonparametric methods.

REFERENCES

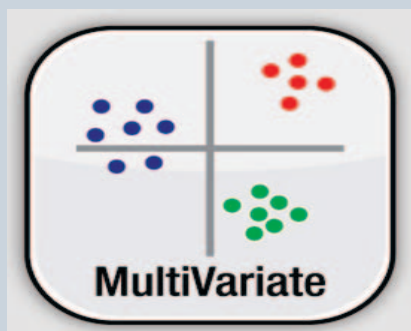
1. Berger, J. O. (2005). *Statistical Decision Theory and Bayesian Analysis*. Springer.
2. Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Brooks/Cole.
3. Efron, B., & Hastie, T. (2016). *Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.
4. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.
5. Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical Statistics*. Pearson.
6. Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.
7. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
8. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.



CHAPTER

4

MULTIVARIATE ANALYSIS



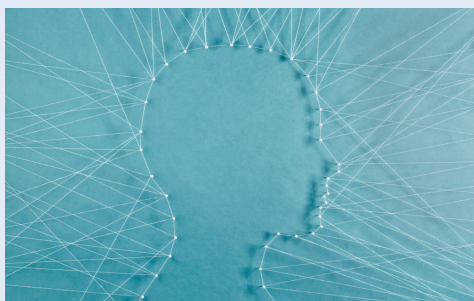
Multivariate Analysis is a branch of statistics that deals with the analysis of data involving multiple variables simultaneously. In many real-world scenarios, data are not limited to just one variable; they often involve multiple variables that interact and influence each other. Multivariate analysis provides techniques to understand the relationships, patterns, and structures among these variables.

After studying this chapter, you will understand the core concepts of:

- Statistics of Multivariate Analysis
- Multivariate Normal Distribution
- Relationship between Variables, Multiple Regression
- Discriminant Analysis
- Mahalanobis Distance

INTRODUCTION

Multivariate Analysis is defined as a process involving multiple dependent variables resulting in one outcome. This explains that the majority of the problems in the real world are Multivariate. For example, we cannot predict the weather of any year based on the season. There are multiple factors like pollution, humidity, precipitation, etc. Here, we will introduce you to multivariate analysis, its history, and its application in different fields.



CORE CONCEPT: STATISTICS OF MULTIVARIATE ANALYSIS

Multivariate analysis (MVA) is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time. In design and analysis, the technique is used to perform trade studies across multiple dimensions while taking into account the effects of all variables on the responses of interest.

The subject of multivariate analysis deals with the statistical analysis of the data collected on more than one (response) variable. These variables may be correlated with each other, and their statistical dependence is often taken into account when analyzing such data. In fact, this consideration of statistical dependence makes multivariate analysis somewhat different in approach and considerably more complex than the corresponding univariate analysis, when there is only one response variable under consideration.

Response variables under consideration are often described as random variables and since their dependence is one of the things to be accounted for in the analyses, these response variables are often described by their joint probability distribution. This consideration makes the modeling issue relatively manageable and provides a convenient framework for scientific analysis of the data. Multivariate normal distribution is one of the most frequently made distributional assumptions for the analysis of multivariate data. However, if possible, any such consideration should ideally be dictated by the particular context. Also, in many cases, such as when the data are collected on a nominal or ordinal scales, multivariate normality may not be an appropriate or even viable assumption.

In the real world, most data collection schemes or designed experiments will result in multivariate data. A few examples of such situations are given below.

- During a survey of households, several measurements on each household are taken. These measurements, being taken on the same household, will be dependent. For example, the education level of the head of the household and the annual income of the family are related.
- During a production process, a number of different measurements such as the tensile strength, brittleness, diameter, etc. are taken on the same unit. Collectively such data are viewed as multivariate data.
- On a sample of 100 cars, various measurements such as the average gas mileage, number of major repairs, noise level, etc. are taken. Also each car is followed for the first 50,000 miles and these measurements are taken after every 10,000 miles. Measurements taken on the same car at the same mileage and those taken at different mileage are going to be correlated. In fact, these data represent a very complex multivariate analysis problem.
- An engineer wishes to set up a control chart to identify the instances when the production process may have gone out of control. Since an out of control process may produce an excessively large number of out of specification items, detection at an early stage is important. In order to do so, she may wish to monitor several process characteristics on the same units. However, since these characteristics are functions of process parameters (conditions), they are likely to be correlated leading to a set of multivariate data. Thus many times, it is appropriate to set up a single (or only a few) multivariate control chart(s) to detect the occurrence of any out of control conditions. On the other hand, if several univariate control charts are separately set up and individually monitored, one may witness too many false alarms, which is clearly an undesirable situation.
- A new drug is to be compared with a control for its effectiveness. Two different groups of patients are assigned to each of the two treatments and they are observed weekly for next two months. The periodic measurements on the same patient will exhibit dependence and thus the basic problem is multivariate in nature. Additionally, if the measurements on various possible side-effects of the drugs are also considered, the subsequent analysis will have to be done under several carefully chosen models.
- In a designed experiment conducted in a research and development center, various factors are set up at desired levels and a number of response variables are measured for each of these treatment combinations. The problem is to find a combination of the levels of these factors where all the responses are at their 'optimum'. Since a treatment combination which optimizes one response variable may not result in the optimum for the other response variable, one has a problem of conflicting objectives especially when the problem is treated as collection of several univariate optimization problems. Due to dependence among responses, it may be more meaningful to analyze response variables simultaneously.
- In many situations, it is more economical to collect a large number of

measurements on the same unit but such measurements are made only on a few units. Such a situation is quite common in many remote sensing data collection plans. Obviously, it is practically impossible to collectively interpret hundreds of univariate analyses to come up with some definite conclusions. A better approach may be that of data reduction by using some meaningful approach. One may eliminate some of the variables which are deemed redundant in the presence of others. Better yet, one may eliminate some of the linear combinations of all variables which contain little or no information and then concentrate only on a few important ones. Which linear combinations of the variables should be retained can be decided using certain multivariate methods such as principal component analysis.

Most of the problems stated above require (at least for the convenience of modeling and for performing statistical tests) the assumption of multivariate normality. There are however, several other aspects of multivariate analysis such as factor analysis, cluster analysis, etc. which are largely distribution free in nature.

Population versus Sample

As we pointed out, the rows in the n by p data matrix $\mathbf{X}$ are viewed as multivariate observations on n units. If the set of these n units constitutes the entire (finite) set of all possible units, then we have data available on the entire reference population. An example of such a situation is the data collected on all cities in the United States that have a population of 1,000,000 or more, and on three variables, namely, cost-of-living, average annual salary, and the quality of health care facilities. Since each U.S. city that qualifies for the definition is included, any summary of these data will be the *true* summary of the population.

However, more often than not, the data are obtained through a survey in which, on each of the units, all p characteristics are measured. Such a situation represents a multivariate sample. A sample (adequately or poorly) represents the underlying population from which it is taken. As the population is now represented through only a few units taken from it, any summary derived from it merely represents the *true* population summary in the sense that we hope that, generally, it will be close to the true summary, although no assurance about an exact match between the two can be given.

How can we measure and ensure that the summary from a sample is a good representative of the population summary? To quantify it, some kinds of indexes based on probabilistic ideas seem appropriate. That requires one to build some kind of probabilistic structure over these units. This is done by artificially and intentionally introducing the probabilistic structure into the sampling scheme. Of course, since we want to ensure that the sample is a good representative of the population, the probabilistic structure should be such that it treats all the population units in an equally fair way. Thus, we require that the sampling is done in such a way that each unit of (finite or infinite) population has an equal chance of being included in the sample. This requirement can be met by a simple random sampling with or without

replacement. It may be pointed out that in the case of a finite population and sampling without replacement, observations are *not* independent, although the strength of dependence diminishes as the sample size increases.

Although a probabilistic structure is introduced over different units through random sampling, the same cannot be done for the p different measurements, as there is neither a reference population nor do all p measurements (such as weight, height, etc.) necessarily represent the same thing. However, there is possibly some inherent dependence between these measurements, and this dependence is often assumed and modeled as some joint probability distribution. Thus, we view each row of $\mathbf{X}$ as a multivariate observation from some p -dimensional population that is represented by some p -dimensional multivariate distribution. Thus, the rows of $\mathbf{X}$ often represent a random sample from a p -dimensional population. In much multivariate analysis work, this population is assumed to be infinite and quite frequently it is assumed to have a multivariate normal distribution.

Data Reduction, Description, and Estimation

We presented some of the basic population summary quantities and their sample counterparts commonly referred to as **descriptive statistics**. The basic idea was to summarize the population or sample data through smaller sized matrices or simply numbers. All the quantities (except correlation) defined there were straightforward generalizations of their univariate counterparts. However, the multivariate data do have some of their own unique features and needs, which do not exist in the univariate situation.

Based on a number of measurements such as average housing prices, cost of living, health care facilities, crime rate, etc., we would like to describe which cities in the country are most livable and also try to observe any unique similarities or differences among cities. There are several variables to be measured, and it is unlikely that attempts to order cities with respect to any one variable will result in the same ordering if another variable were used. For example, a city with a low crime rate (a desirable feature) may have a high cost of living (an undesirable feature), and thus these variables often tend to offset each other. How do we decide which cities are the best to live in? The problem here is that of data reduction. However,

Descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.

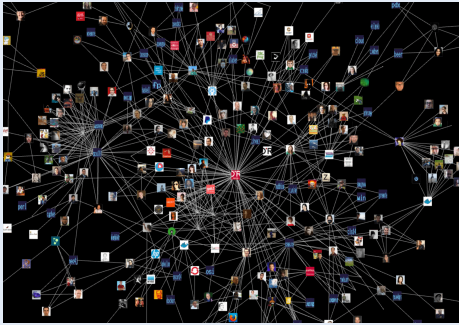
this problem can neither be described as that of variable selection (there is no dependent variable and no model) nor can it be viewed as a prediction problem. It is more a problem of attempting to detect and understand the unique features that the data set may contain and then to interpret them. This requires some meaningful approach for data description.

As another example, suppose we have a set of independent variables which in turn have effects on a large number of dependent variables. Such a situation is quite common in the chemical industry and in economic data, where the two sets can be clearly defined as those containing input and output variables. We are not interested in individual variables, but we want to come up with a few new variables in each group. These may themselves be functions of all variables in the respective groups, so that each new variable from one group can be paired with another new variable in the other group in some meaningful sense, with the hope that these newly defined variables can be appropriately interpreted in the context. We must emphasize that analysis is not being done with any specific purpose of proving or disproving some claims. It is only an attempt to understand the data. As the information is presented in terms of new variables, which are fewer in number, it is easier to observe any striking features or associations in this latter situation.

An automobile company wants to know what determines the customer's preference for various cars. A sample of 100 randomly selected individuals were asked to give a score between 1 (low) and 10 (high) on six variables, namely, price, reliability, status symbol related to car, gas mileage, safety in an accident, and average miles driven per week. What kind of analysis can be made for these data? With the assumptions that there are some underlying hypothetical and unobservable variables on which the scores of these six observable variables depend, a natural inquiry would be to identify these hypothetical variables. Intuitively, safety consciousness and economic status of the individual may be two (perhaps of several others) traits that may influence the scores on some of these six observable variables. Thus, some or all of the observed variables can be written as a function of, say, these two unobservable traits.). Note, however, that the analysis provides only the functions and, their interpretations as some meaningful unobservable trait, is left to the analyst. Nonetheless, it is again a problem of **data reduction** and description in that many measurements are reduced to only a few traits with the objective of providing an appropriate description of the data.

Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume.

As is clear from these examples, many multivariate problems involve data reduction, description and, in the process of doing so, estimation. As a general theme, most of the situations either require some matrix decomposition and transformations or use a distance-based approach. Distributional assumptions such as multivariate normality are also helpful (usually but not always, in assessing the quality of estimation) but not crucial.



CORE CONCEPT: MULTIVARIATE NORMAL DISTRIBUTION

A multivariate normal distribution is a vector in multiple normally distributed variables, such that any linear combination of the variables is also normally distributed. It is mostly useful in

extending the central limit theorem to multiple variables, but also has applications to Bayesian inference and thus machine learning, where the multivariate normal distribution is used to approximate the features of some characteristics; for instance, in detecting faces in pictures.

A probability distribution that plays a pivotal role in multivariate analysis is multivariate normal distribution. We say that $\mathbf{y}$ has a multivariate normal distribution (with a mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$) if its density is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \cdot \exp \left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right).$$

In notation, we state this fact as $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Observe that the above density is a straightforward extension of the univariate normal density to which it will reduce when $p = 1$.

Important properties of the multivariate normal distribution include some of the following:

- Let $\mathbf{A}_{r \times p}$ be a fixed matrix, then $\mathbf{A}\mathbf{y} \sim N_r(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}') (r \leq p)$. It may be added that $\mathbf{A}\mathbf{y}$ will admit the density if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ is nonsingular, which will happen if and only if all rows of $\mathbf{A}$ are linearly independent. Further, in principle, r can also be greater than p . However, in that case, the matrix $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ will not be nonsingular. Consequently, the vector $\mathbf{A}\mathbf{y}$ will not admit a density function.
- Let $\mathbf{G}$ be such that $\boldsymbol{\Sigma}^{-1} = \mathbf{G}\mathbf{G}'$, then $\mathbf{G}'\mathbf{y} \sim N_p(\mathbf{G}'\boldsymbol{\mu}, \mathbf{I})$ and $\mathbf{G}'(\mathbf{y} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I})$.
- Any fixed linear combination of $y_1, \dots, y_p$, say $\mathbf{c}'\mathbf{y}$, $\mathbf{c}_{p \times 1} \neq \mathbf{0}$ is also normally

Variance-covariance matrix is a square matrix that contains the variances and covariances associated with several variables.

distributed. Specifically, $c'y \sim N_1(c'\mu, c'\Sigma c)$

- The subsectors y_1 and y_2 are also normally distributed, specifically $y_1 \sim N_{p_1}(\mu_1, \Sigma_{11})$ and $y_2 \sim N_{p-p_1}(\mu_2, \Sigma_{22})$.
- Individual components $y_1, \dots, y_p$ are all normally distributed. That is, $y_i \sim N_1(\mu_i, \sigma_{ii})$, $i = 1, \dots, p$.
- The conditional distribution of y_1 given y_2 , written as $y_1|y_2$, is also normal. Specifically,

$$y_1|y_2 \sim N_{p_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

- Let $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2) = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}y_2 = B_0 + B_1y_2$, and $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.
- The conditional expectation of y_1 for given values of y_2 or the regression function of y_1 on y_2 is $B_0 + B_1y_2$, which is linear in y_2 . This is a key fact for multivariate multiple linear regression modeling. The matrix $\Sigma_{11.2}$ is usually represented by the **variance-covariance matrix** of error components in these models. An analogous result (and the interpretation) can be stated for the conditional distribution of y_2 given y_1 .
- Let δ be a fixed $p \times 1$ vector, then

$$y + \delta \sim N_p(\mu + \delta, \Sigma).$$

- The random components $y_1, \dots, y_p$ are all independent if and only if Σ is a diagonal matrix; that is, when all the covariances (or correlations) are zero.
- Let u_1 and u_2 be respectively distributed as $N_p(\mu_{u_1}, \Sigma_{u_1})$ and $N_p(\mu_{u_2}, \Sigma_{u_2})$, then $u_1 \pm u_2 \sim N_p(\mu_{u_1} \pm \mu_{u_2}, \Sigma_{u_1} + \Sigma_{u_2} \pm (\text{cov}(u_1, u_2) + \text{cov}(u_2, u_1)))$.

Note that if u_1 and u_2 were independent, the last two covariance terms would drop out.

Univariate Normal Distribution

Just to refresh memory, in the case of a single random variable, the probability distribution or density function of that variable x is represented by Equation (2.1):

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (1)$$

Bivariate Normal Distribution

The bivariate distribution represents the joint distribution of two random variables. The two random variables x_1 and x_2 are related to each other in the sense that they are not independent of each other. This dependence is reflected by the correlation ρ between the two variables x_1 and x_2 . The density function for the two variables jointly is

$$\Phi(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \right\} \quad (2.2) \quad (2)$$

This function can be represented graphically as in Figure 1:

The Isodensity contour is defined as the set of points for which the values of x_1 and x_2 give the same value for the density function Φ . This contour is given by Equation (3) for a fixed value of C , which defines a constant probability:

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} = C \quad (3)$$

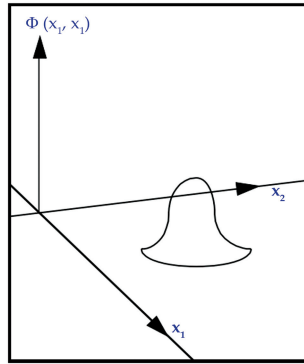


Figure 1: The bivariate normal distribution.

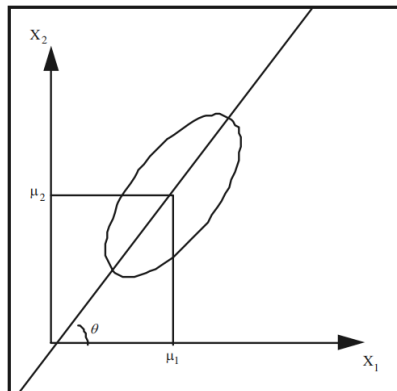


Figure 2: The locus of points of the bivariate normal distribution at a given density level.

Equation (3) defines an ellipse with centroid (μ_1, μ_2) . This ellipse is the locus of points representing the combinations of the values of x_1 and x_2 with the same probability, as defined by the constant C (Figure 2).

For various values of C , we get a family of concentric ellipses (at a different cut, i.e., cross section of the density surface with planes at various elevations) (Figure 3).

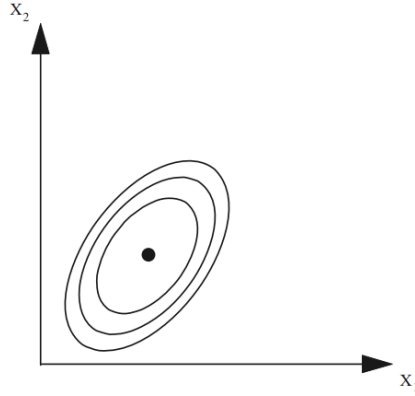


Figure 3: Concentric ellipses at various density levels.

The angle θ depends only on the values of σ_1 , σ_2 , and ρ but is independent of C . The higher the correlation between x_1 and x_2 , the steeper the line going through the origin with angle θ , i.e., the bigger the angle.

Generalization to Multivariate Case

Random variable is a variable whose possible values are outcomes of a random phenomenon.

Let us represent the bivariate distribution in matrix algebra notation in order to derive the generalized format for more than two **random variables**. The covariance matrix of (x_1, x_2) can be written as

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (4)$$

The determinant of the matrix Σ is

$$|\Sigma| = \sigma_1^2\sigma_2^2(1 - \rho^2) \quad (5)$$

Equation (3) can now be re-written as

$$C = [x_1 - \mu_1, x_2 - \mu_2] \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \quad (6)$$

where

$$\Sigma^{-1} = 1/\left[\sigma_1^2\sigma_2^2(1-\rho^2)\right] \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \quad (7)$$

Note that $\Sigma^{-1} = |\Sigma|^{-1} \times \text{matrix of cofactors}$.

Let

$$\mathbf{X} = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

then $\mathbf{X}^{-1}\mathbf{X} = \chi^2$, which is a quadratic form of the variables $\mathbf{x}$ and is, therefore, a chi-square variate.

Also, because $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$, $|\Sigma|^{1/2} = \sigma_1\sigma_2\sqrt{(1-\rho^2)}$, and consequently,

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} = (2\pi)^{-1} |\Sigma|^{-1/2} \quad (8)$$

the bivariate distribution function can be now expressed in matrix notation as

$$\Phi(x_1, x_2) = (2\pi)^{-1} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{X}'\Sigma^{-1}\mathbf{X}} \quad (9)$$

Now, more generally with p random variables ($x^1, x^2, \dots, x^p$), let

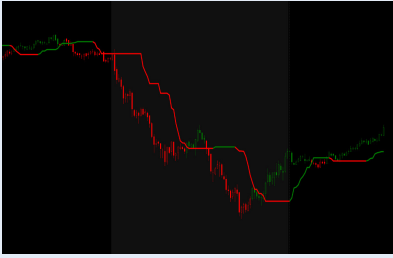
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

The density function is

$$\Phi(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)}$$

For a fixed value of the density Φ , an ellipsoid is described. Let $\mathbf{X} = \mathbf{x} - \mu$. The inequality $\mathbf{X}'\Sigma^{-1}\mathbf{X} \leq \chi^2$ defines any point within the ellipsoid.

The statistical study of data where multiple measurements are made on each experimental unit and where the relationships among multivariate measurements and their structure are important.



CORE CONCEPT: RELATIONSHIP BETWEEN VARIABLES, MULTIPLE REGRESSION

Multiple regression is an extension of simple linear regression. It is used when anyone want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how “pretty” the house is (subjective rating). You may also detect “outliers,” that is, houses that should really sell for more, given their location and characteristics.

Regression equation

can be defined as a statistical model, used to determine the specific relationship between the predictor variable and the outcome variable.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. You can determine a number of factors or dimensions such as “amount of responsibility” (*Resp*) or “number of people to supervise” (*No_Super*) that you believe to contribute to the value of a job. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a **regression equation** of the form:

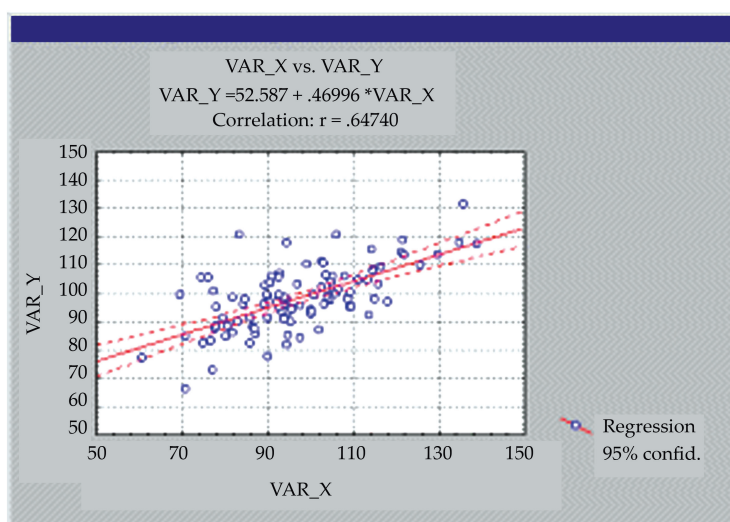
$$\text{Salary} = .5 * \text{Resp} + .8 * \text{No_Super}$$

Once this so-called regression line has been determined, the analyst can now easily construct a graph of the expected

(predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably. In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question “what is the best predictor of ...”. For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

Computational Approach

The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line to a number of points.



In the simplest case - one dependent and one independent variable - you can visualize this in a scatterplot.

- Least Squares
- The Regression Equation
- Unique Prediction and Partial Correlation
- Predicted and Residual Scores
- Residual Variance and R-square
- Interpreting the Correlation Coefficient R

Least Squares

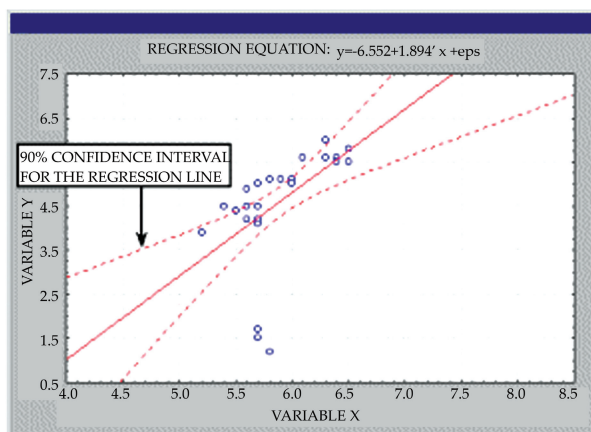
In the scatterplot, we have an independent or X variable, and a dependent or Y variable. These variables may, for example, represent IQ (intelligence as measured by a test) and school achievement (grade point average; GPA), respectively. Each

point in the plot represents one student, that is, the respective student's IQ and GPA. The goal of linear regression procedures is to fit a line through the points. Specifically, the program will compute a line so that the squared deviations of the observed points from that line are minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

The Regression Equation

A line in a two dimensional or two-variable space is defined by the equation $Y=a+b*X$; in full text: the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant is also referred to as the *intercept*, and the slope as the *regression coefficient* or *B coefficient*. For example, GPA may best be predicted as $1+.02*IQ$. Thus, knowing that a student has an *IQ* of 130 would lead us to predict that her GPA would be 3.6 (since, $1+.02*130=3.6$).

For example, the animation below shows a two dimensional regression equation plotted with three different confidence intervals (90%, 95% and 99%).



In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily. For example, if in addition to IQ we had additional predictors of achievement (e.g., Motivation, Self-discipline) we could construct a linear equation containing all those variables. In general then, multiple regression procedures will estimate a linear equation of the form:

$$Y = a + b_1 * X_1 + b_2 * X_2 + \dots + b_p * X_p$$

Unique Prediction and Partial Correlation

Note that in this equation, the regression coefficients (or B coefficients) represent the independent contributions of each independent variable to the prediction of the dependent variable. Another way to express this fact is to say that, for example,

Regression models predict a value of the Y variable given known values of the X variables. Prediction within the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction outside this range of the data is known as extrapolation.

variable X_1 is correlated with the Y variable, after controlling for all other independent variables. This type of correlation is also referred to as a partial correlation (this term was first used by Yule, 1907). Perhaps the following example will clarify this issue. You would probably find a significant negative correlation between hair length and height in the population (i.e., short people have longer hair). At first this may seem odd; however, if we were to add the variable Gender into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have longer hair than men; they also are shorter on the average than men. Thus, after we remove this gender difference by entering Gender into the equation, the relationship between hair length and height disappears because hair length does not make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable Gender. Put another way, after controlling for the variable Gender, the partial correlation between hair length and height is zero.

Predicted and Residual Scores

The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points around the fitted regression line (as in the scatterplot shown earlier). The deviation of a particular point from the regression line (its predicted value) is called the **residual value**.

Residual value refers to the estimated worth of an asset after the asset has fully depreciated.

Residual Variance and R-Square

R-Square, also known as the Coefficient of determination is a commonly used statistic to evaluate model fit. R-square is 1 minus the ratio of residual variability. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. For example, if there is no relationship between the X and Y variables, then the ratio of the residual variability of the Y variable to the original variance is equal to 1.0. Then R-square would be 0. If X and Y are perfectly related then there is no residual variance and the ratio of variance would be 0.0, making R-square = 1. In most cases, the ratio and R-square will fall somewhere between these extremes, that is, between 0.0 and 1.0. This ratio value is immediately interpretable in the following manner. If we have an R-square

of 0.4 then we know that the variability of the Y values around the regression line is 1-0.4 times the original variance; in other words we have explained 40% of the original variability, and are left with 60% residual variability. Ideally, we would like to explain most if not all of the original variability. The R -square value is an indicator of how well the model fits the data (e.g., an R -square close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model).

Interpreting The Correlation Coefficient R

Customarily, the degree to which two or more predictors (independent or X variables) are related to the dependent (Y) variable is expressed in the correlation coefficient R , which is the square root of R -square. In multiple regression, R can assume values between 0 and 1. To interpret the direction of the relationship between variables, look at the signs (plus or minus) of the regression or B coefficients. If a B coefficient is positive, then the relationship of this variable with the dependent variable is positive (e.g., the greater the IQ the better the grade point average); if the B coefficient is negative then the relationship is negative (e.g., the lower the class size the better the average test scores). Of course, if the B coefficient is equal to 0 then there is no relationship between the variables.

Assumptions, Limitations, Practical Considerations

- Assumption of Linearity
- Normality Assumption
- Limitations
- Choice of the number of variables
- Multicollinearity and matrix ill-conditioning
- The importance of residual analysis

Assumption Of Linearity

First of all, as is evident in the name multiple *linear* regression, it is assumed that the relationship between variables is linear. In practice this assumption can virtually never be confirmed; fortunately, multiple regression procedures are not greatly affected by minor deviations from this assumption. However, as a rule it is prudent to *always* look at bivariate scatterplot of the variables of interest. If curvature in the relationships is evident, you may consider either transforming the variables, or explicitly allowing for nonlinear components.

Normality Assumption

It is assumed in multiple regression that the residuals (predicted minus observed values) are distributed normally (i.e., follow the normal distribution). Again, even though most tests (specifically the F -test) are quite robust with regard to violations of

this assumption, it is *always* a good idea, before drawing final conclusions, to review the distributions of the major variables of interest. You can produce histograms for the residuals as well as normal probability plots, in order to inspect the distribution of the residual values.

Limitations

The major conceptual limitation of all regression techniques is that you can only ascertain *relationships*, but never be sure about underlying *causal* mechanism. For example, you would find a strong positive relationship (correlation) between the damage that a fire does and the number of firemen involved in fighting the blaze. Do we conclude that the firemen cause the damage? Of course, the most likely explanation of this correlation is that the size of the fire (an external variable that we forgot to include in our study) caused the damage as well as the involvement of a certain number of firemen (i.e., the bigger the fire, the more firemen are called to fight the blaze). Even though this example is fairly obvious, in real correlation research, alternative causal explanations are often not considered.

Choice of The Number of Variables

Multiple regression is a seductive technique: “plug in” as many predictor variables as you can think of and usually at least a few of them will come out significant. This is because you are capitalizing on chance when simply including as many variables as you can think of as predictors of some other variable of interest. This problem is compounded when, in addition, the number of observations is relatively low. Intuitively, it is clear that you can hardly draw conclusions from an analysis of 100 questionnaire items based on 10 respondents. Most authors recommend that you should have at least 10 to 20 times as many observations (cases, respondents) as you have variables; otherwise the estimates of the regression line are probably very unstable and unlikely to replicate if you were to conduct the study again.

Multicollinearity and Matrix Ill-Conditioning

This is a common problem in many correlation analyses. Imagine that you have two predictors (X variables) of a person's height:

1. weight in pounds and
2. weight in ounces.

Obviously, our two predictors are completely redundant; weight is one and the same variable, regardless of whether it is measured in pounds or ounces. Trying to decide which one of the two measures is a better predictor of height would be rather silly; however, this is exactly what you would try to do if you were to perform a multiple regression analysis with height as the dependent (Y) variable and the two measures of weight as the independent (X) variables. When there are very many variables involved, it is often not immediately apparent that this problem exists, and

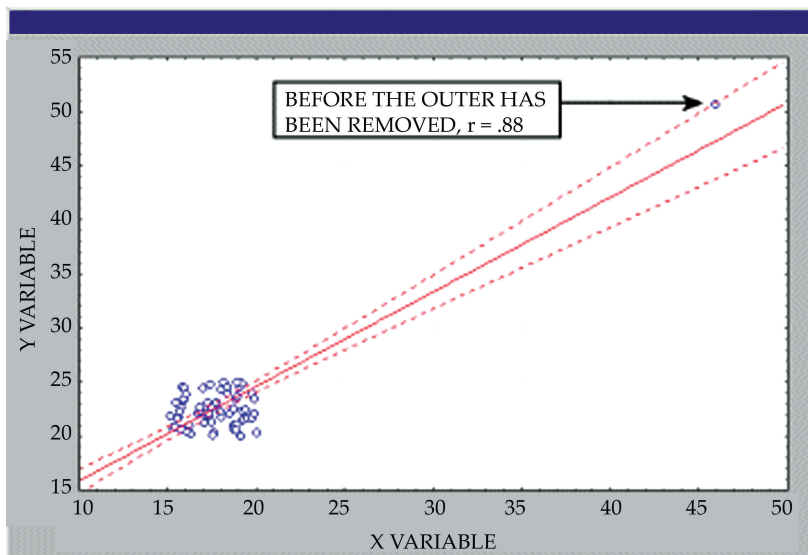
it may only manifest itself after several variables have already been entered into the regression equation. Nevertheless, when this problem occurs it means that at least one of the predictor variables is (practically) completely redundant with other predictors. There are many statistical indicators of this type of redundancy (tolerances, semi-partial R , etc., as well as some remedies (e.g., Ridge regression).

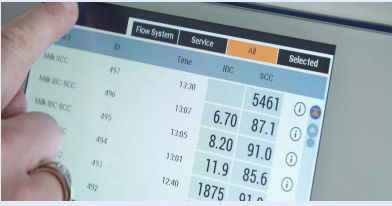
Fitting Centered Polynomial Models

The fitting of higher-order polynomials of an independent variable with a mean not equal to zero can create difficult multicollinearity problems. Specifically, the polynomials will be highly correlated due to the mean of the primary independent variable. With large numbers (e.g., Julian dates), this problem is very serious, and if proper protections are not put in place, can cause wrong results. The solution is to “center” the independent variable (sometimes, this procedure is referred to as “centered polynomials”), i.e., to subtract the mean, and then to compute the polynomials.

The Importance of Residual Analysis

Even though most assumptions of multiple regression cannot be tested explicitly, gross violations can be detected and should be dealt with appropriately. In particular outliers (i.e., extreme cases) can seriously bias the results by “pulling” or “pushing” the regression line in a particular direction, thereby leading to biased regression coefficients. Often, excluding just a single extreme case can yield a completely different set of results.





ID	Time	INC	WCC	Selected
497	13:30	6.70	87.1	5461
496	13:07	8.20	91.0	
495	13:05	11.9	85.6	
494	13:01	1875	01.8	
493	12:40			
492				

CORE CONCEPT: DISCRIMINANT ANALYSIS

Discriminant Analysis finds a set of prediction equations based on independent variables that are used to classify individuals into groups. There are two possible objectives in a discriminant analysis: finding a predictive equation for classifying new individuals or interpreting the predictive equation to better understand the relationships that may exist among the variables.

Discriminant analysis is a statistical method that is used by researchers to help them understand the relationship between a “dependent variable” and one or more “independent variables.” A dependent variable is the variable that a researcher is trying to explain or predict from the values of the independent variables. Discriminant analysis is similar to regression analysis and analysis of variance (ANOVA). The principal difference between discriminant analysis and the other two methods is with regard to the nature of the dependent variable.

Discriminant analysis requires the researcher to have measures of the dependent variable and all of the independent variables for a large number of cases. In regression analysis and ANOVA, the dependent variable must be a “continuous variable.” A numeric variable indicates the degree to which a subject possesses some characteristic, so that the higher the value of the variable, the greater the level of the characteristic. A good example of a continuous variable is a person’s income.

In discriminant analysis, the dependent variable must be a “categorical variable.” The values of a categorical variable serve only to name groups and do not necessarily indicate the degree to which some characteristic is present. An example of a categorical variable is a measure indicating to which one of several different market segments a customer belongs; another example is a measure indicating whether or not a particular employee is a “high potential” worker. The categories must be mutually exclusive; that is, a subject can belong to one and only one of the groups indicated by the categorical variable. While a categorical variable must have at least two values (as in the “high potential” case), it may have numerous values (as in the case of the market segmentation measure). As the mathematical methods used in discriminant analysis are complex, they are described here only in general terms. We will do this by providing an example of a simple case in which the dependent variable has only two categories.

Discriminant analysis is most often used to help a researcher predict the group or category to which a subject belongs. For example, when individuals are interviewed for a job, managers will not know for sure how job candidates will perform on the job if hired. Suppose, however, that a human resource manager has a list of current employees who have been classified into two groups: “high performers” and “low performers.”

Discriminant analysis (DA) is a multivariate technique used to separate two or more groups of observations (individuals) based on k variables measured on each experimental unit (sample) and find the contribution of each variable in separating the groups.

These individuals have been working for the company for some time, have been evaluated by their supervisors, and are known to fall into one of these two mutually exclusive categories. The manager also has information on the employees' backgrounds: educational attainment, prior work experience, participation in **training** programs, work attitude measures, personality characteristics, and so forth. This information was known at the time these employees were hired. The manager wants to be able to predict, with some confidence, which future job candidates are high performers and which are not. A researcher or consultant can use discriminant analysis, along with existing data, to help in this task.

There are two basic steps in **discriminant analysis**. The first involves estimating coefficients, or weighting factors, that can be applied to the known characteristics of job candidates (i.e., the independent variables) to calculate some measure of their tendency or propensity to become high performers. This measure is called a “discriminant function.” Second, this information can then be used to develop a decision rule that specifies some cut-off value for predicting which job candidates are likely to become high performers.

The tendency of an individual to become a high performer can be written as a linear equation. The values of the various predictors of high performer status (i.e., independent variables) are multiplied by “discriminant function coefficients” and these products are added together to obtain a predicted discriminant function score. This score is used in the second step to predict the job candidates likelihood of becoming a high performer. Suppose that you were to use three different independent variables in the discriminant analysis. Then the discriminant function has the following form:

where D = discriminant function score,

B_i = discriminant function coefficient relating independent variable i to the discriminant function score,

X_i = value of independent variable i .

The equation is quite similar to a regression equation. Conventional regression analysis should not be used in place of discriminant analysis. The dependent variable would have only two values (high performer and low performer) and would thus violate important assumptions of the regression model. Discriminant analysis does not have these limitations with respect to the dependent variable.

Estimation of the discriminant function coefficients requires a set of cases in which values of the independent variables and the dependent variables are known. In the case described above, the company has this information for a current group of employees. There are several different ways that can be used to estimate discriminant function coefficients, but all work on the same general principle: the values of the coefficients are selected so that differences between the groups defined by the dependent variable are maximized with regard to some objective function. One commonly used objective function is the F-ratio, which is defined as it is in ANOVA and regression problems. The coefficients are chosen to maximize the F-ratio when analysis of variance is performed on the resulting discriminant function, using the dependent variable (i.e., job performance) as the grouping variable. Most general statistical programs, such as the Statistical Package for the Social Sciences, contain discriminant analysis modules.

There are various tests of significance that can be used in discriminant analysis. One widely used test statistic is based on Wilks lambda, which provides an assessment of the discriminating power of the function derived from the analysis. If this value is found to be statistically significant, then the set of independent variables can be assumed to differentiate between the groups of the categorical variable. This test, which is analogous to the F-ratio test in ANOVA and regression, is useful in evaluating the overall adequacy of the analysis.

Unfortunately, discriminant analysis does not generate estimates of the standard errors of the individual coefficients, as in regression, so it is not quite so simple to assess the statistical significance of each coefficient. For example, most discriminant analysis programs have a stepwise option. Independent variables are entered into the equation one at a time. Again, Wilks lambda can be used to assess the potential contribution of each variable to the explanatory power of the model. Variables from the set of independent variables are added to the equation until a point is reached for which additional items provide no statistically significant increment in explanatory power.

Once the analysis is completed, the discriminant function coefficients can be used to assess the contributions of the various independent variables to the tendency of an employee to be a high performer. The discriminant function coefficients are analogous regression coefficients and they range between values of -1.0 and 1.0. The second box provides the within-group averages for the discriminant function for the two categories of the dependent variable. Note that the high performers have an average score of 1.45 on the discriminant function, while the low performers have an average score of -.89. The discriminant function is treated as a standardized variable, so it has a mean of zero and a standard deviation of one. The average values of the discriminant function scores are meaningful only in that they help us interpret the coefficients. Since the high performers are at the upper end of the scale, all of the positive coefficients indicate that the greater the value of those variables, the greater the likelihood of a worker being a high performer (e.g., education, motivation). The magnitudes of the coefficients also tell us something about the relative contributions of the independent variables. The closer the value of a coefficient is to zero, the weaker

it is as a predictor of the dependent variable. On the other hand, the closer the value of a coefficient is to either 1.0 or -1.0, the stronger it is as a predictor of the dependent variable. In this example, then, years of education and ability to handle stress both have positive coefficients, though the latter is quite weak. Finally, individuals who place high importance on family life are less likely to be high performers than those who do not.

The second step in discriminant analysis involves predicting to which group in the dependent variable a particular case belongs. A subject's discriminant score can be translated into a probability of being in a particular group by means of Bayes Rule. Separate probabilities are computed for each group and the subject is assigned to the group with the highest probability.

Another test of the adequacy of a model is the degree to which known cases are correctly classified. As in other statistical procedures, it is generally preferable to test the model on a set of cases that were not used to estimate the model's parameters. This provides a more conservative test of the model. Thus, a set of cases should, if possible, be saved for this purpose. Having completed the analysis, the results can be used to predict the work potential of job candidates and hopefully serve to improve the selection process.

General Purpose

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (1) to go to college, (2) to attend a trade or professional school, or (3) to seek no further training or education. For that purpose, the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. *Discriminant Analysis* could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

A medical researcher may record different variables relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely to recover completely (group 1), partially (group 2), or not at all (group 3). A biologist could record different characteristics of similar types (groups) of flowers, and then perform a discriminant function analysis to determine the set of characteristics that allows for the best discrimination between the types.

Computational Approach

Computationally, discriminant function analysis is very similar to analysis of variance (ANOVA). Let us consider a simple example. Suppose we measure height in a random sample of 50 males and 50 females. Females are, on the average, not as tall as males, and this difference will be reflected in the difference in means (for the variable *Height*). Therefore, variable height allows us to discriminate between males

and females with a better than chance probability: if a person is tall, then he is likely to be a male, if a person is short, then she is likely to be a female. We can generalize this reasoning to groups and variables that are less “trivial.” For example, suppose we have two groups of high school graduates: Those who choose to attend college after graduation and those who do not.

We could have measured students’ stated intention to continue on to college one year prior to graduation. If the means for the two groups (those who actually went to college and those who did not) are different, then we can say that intention to attend college as stated one year prior to graduation allows us to discriminate between those who are and are not college bound (and this information may be used by career counselors to provide the appropriate guidance to the respective students).

Analysis of Variance

Stated in this manner, the discriminant function problem can be rephrased as a one-way analysis of variance (ANOVA) problem. Specifically, one can ask whether or not two or more groups are *significantly different* from each other with respect to the mean of a particular variable. To learn more about how one can test for the statistical significance of differences between means in different groups you may want to read the Overview to ANOVA/MANOVA. However, it should be clear that, if the means for a variable are significantly different in different groups, then we can say that this variable discriminates between the groups.

In the case of a single variable, the final significance test of whether or not a variable discriminates between groups is the F test. As described in Elementary Concepts and ANOVA /MANOVA, F is essentially computed as the ratio of the between-groups variance in the data over the pooled (average) within-group variance. If the between-group variance is significantly larger then there must be significant differences between means.

Multiple Variables

Usually, one includes several variables in a study in order to see which one(s) contribute to the discrimination between groups. In that case, we have a matrix of total variances and covariance’s; likewise, we have a matrix of pooled within-group variances and covariance’s.

We can compare those two matrices via multivariate F tests in order to determine whether or not there are any significant differences (with regard to all variables) between groups. This procedure is identical to multivariate analysis of variance or MANOVA. As in MANOVA, one could first perform the multivariate test, and, if statistically significant, proceed to see which of the variables have significantly different means across the groups. Thus, even though the computations with multiple variables are more complex, the principal reasoning still applies, namely, that we are looking for variables that discriminate between groups, as evident in observed mean differences.

Stepwise Discriminant Analysis

Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups. For example, an educational researcher interested in predicting high school graduates' choices for further education would probably include as many measures of personality, achievement motivation, academic performance, etc. as possible in order to learn which one(s) offer the best prediction.

Model

Put another way, we want to build a “model” of how we can best predict to which group a case belongs. In the following discussion we will use the term “in the model” in order to refer to variables that are included in the prediction of group membership, and we will refer to variables as being “not in the model” if they are not included.

Forward Stepwise Analysis

In stepwise discriminant function analysis, a model of discrimination is built step-by-step. Specifically, at each step all variables are reviewed and evaluated to determine which one will contribute most to the discrimination between groups. That variable will then be included in the model, and the process starts again.

Backward Stepwise Analysis

One can also step backwards; in that case all variables are included in the model and then, at each step, the variable that contributes least to the prediction of group membership is eliminated. Thus, as the result of a successful discriminant function analysis, one would only keep the “important” variables in the model, that is, those variables that contribute the most to the discrimination between groups.

F to enter, F to remove

The stepwise procedure is “guided” by the respective F to enter and F to remove values. The F value for a variable indicates its statistical significance in the discrimination between groups, that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership. If you are familiar with stepwise multiple regression procedures, then you may interpret the F to enter/remove values in the same way as in stepwise regression.

Capitalizing on Chance

A common misinterpretation of the results of stepwise discriminant analysis is to take statistical significance levels at face value. By nature, the stepwise procedures will capitalize on chance because they “pick and choose” the variables to be included in the model so as to yield maximum discrimination. Thus, when using the stepwise approach, the researcher should be aware that the significance levels do not reflect

the true *alpha* error rate, that is, the probability of erroneously rejecting H_0 (the null hypothesis that there is no discrimination between groups).

Interpreting a Two-Group Discriminant Function

In the two-group case, discriminant function analysis can also be thought of as (and is analogous to) multiple regression. If we code the two groups in the analysis as 1 and 2, and use that variable as the dependent variable in a multiple regression analysis, then we would get results that are analogous to those we would obtain via *Discriminant Analysis*. In general, in the two-group case we fit a linear equation of the type:

$$\text{Group} = a + b_1 * x_1 + b_2 * x_2 + \dots + b_m * x_m$$

where a is a constant and b_1 through b_m are regression coefficients. The interpretation of the results of a two-group problem is straightforward and closely follows the logic of multiple regression: Those variables with the largest (standardized) regression coefficients are the ones that contribute most to the prediction of group membership.

Discriminant Functions for Multiple Groups

When there are more than two groups, then we can estimate more than one discriminant function like the one presented above. For example, when there are three groups, we could estimate (1) a function for discriminating between group 1 and groups 2 and 3 combined, and (2) another function for discriminating between group 2 and group 3. For example, we could have one function that discriminates between those high school graduates that go to college and those who do not, and a second function to discriminate between those graduates that go to a professional or trade school versus those who get a job. The b coefficients in those discriminant functions could then be interpreted as before.

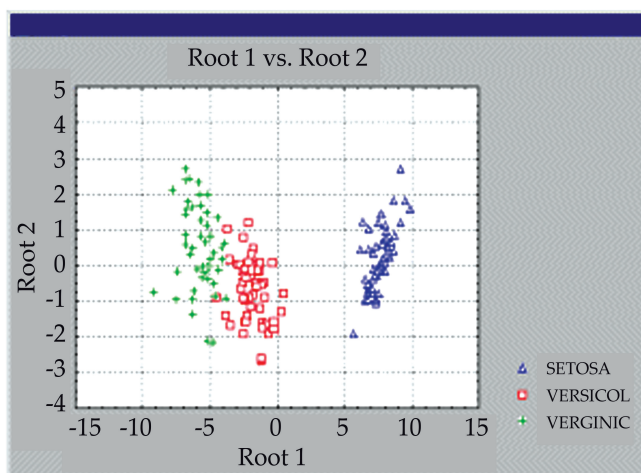
Canonical Analysis

When actually performing a multiple group discriminant analysis, we do not have to specify how to combine groups so as to form different discriminant functions. Rather, you can automatically determine some optimal combination of variables so that the first function provides the most overall discrimination between groups, the second provides second most, and so on. Moreover, the functions will be independent or *orthogonal*, that is, their contributions to the discrimination between groups will not overlap. Computationally, you will perform a canonical correlation analysis that will determine the successive functions and canonical *roots* (the term root refers to the eigenvalues that are associated with the respective canonical function). The maximum number of functions will be equal to the number of groups minus one, or the number of variables in the analysis, whichever is smaller.

Interpreting the Discriminant Functions

As before, we will get b (and standardized *beta*) coefficients for each variable in

each discriminant (now also called *canonical*) function, and they can be interpreted as usual: the larger the standardized coefficient, the greater is the contribution of the respective variable to the discrimination between groups. (Note that we could also interpret the structure coefficients; see below.) However, these coefficients do not tell us between which of the groups the respective functions discriminate. We can identify the nature of the discrimination for each discriminant (canonical) function by looking at the means for the functions across groups. We can also visualize how the two functions discriminate between groups by plotting the individual scores for the two discriminant functions (see the example graph below).



In this example, Root (function) 1 seems to discriminate mostly between groups Setosa, and Virginic and Versicolcombined. In the vertical direction (Root 2), a slight trend of Versicol points to fall below the center line (0) is apparent.

Factor Structure Matrix

Another way to determine which variables “mark” or define a particular discriminant function is to look at the factor structure. The factor structure coefficients are the correlations between the variables in the model and the discriminant functions; if you are familiar with factor analysis you may think of these correlations as factor *loadings* of the variables on each discriminant function.

Some authors have argued that these structure coefficients should be used when interpreting the substantive “meaning” of discriminant functions. The reasons given by those authors are that (1) supposedly the structure coefficients are more stable, and (2) they allow for the interpretation of factors (discriminant functions) in the manner that is analogous to factor analysis. However, subsequent Monte Carlo research has shown that the discriminant function coefficients and the structure coefficients are about equally unstable, unless the n is fairly large (e.g., if there are 20 times more cases than there are variables). The most important thing to remember is that the discriminant function coefficients denote the unique (partial) contribution of each variable to the discriminant function(s), while the structure coefficients denote the simple correlations between the variables and the function(s). If one wants to

assign substantive “meaningful” labels to the discriminant functions (akin to the interpretation of factors in factor analysis), then the structure coefficients should be used (interpreted); if one wants to learn what is each variable’s unique contribution to the discriminant function, use the discriminant function coefficients (weights).

Significance of Discriminant Functions

One can test the number of roots that add *significantly* to the discrimination between group. Only those found to be statistically significant should be used for interpretation; non-significant functions (roots) should be ignored.

To summarize, when interpreting multiple discriminant functions, which arise from analyses with more than two groups and more than one variable, one would first test the different functions for statistical significance, and only consider the significant functions for further examination. Next, we would look at the standardized *b* coefficients for each variable for each significant function. The larger the standardized *b* coefficient, the larger is the respective variable’s unique contribution to the discrimination specified by the respective discriminant function. In order to derive substantive “meaningful” labels for the discriminant functions, one can also examine the factor structure matrix with the correlations between the variables and the discriminant functions. Finally, we would look at the means for the significant discriminant functions in order to determine between which groups the respective functions seem to discriminate.

Assumptions

As mentioned earlier, discriminant function analysis is computationally very similar to MANOVA, and all assumptions for *MANOVA* mentioned in ANOVA/MANOVA apply. In fact, you may use the wide range of diagnostics and statistical tests of assumption that are available to examine your data for the discriminant analysis.

Normal Distribution

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution. You can examine whether or not variables are normally distributed with histograms of frequency distributions. However, note that violations of the normality assumption are usually not “fatal,” meaning, that the resultant significance tests etc. are still “trustworthy.” You may use specific tests for normality in addition to graphs.

Homogeneity of Variances/Covariances

It is assumed that the variance/covariance matrices of variables are homogeneous across groups. Again, minor deviations are not that important; however, before accepting final conclusions for an important study it is probably a good idea to review the within-groups variances and correlation matrices. In particular, a [scatterplot matrix](#) can be produced and can be very useful for this purpose. When in doubt, try

Scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables.

re-running the analyses excluding one or two groups that are of less interest. If the overall results (interpretations) hold up, you probably do not have a problem. You may also use the numerous tests available to examine whether or not this assumption is violated in your data. However, as mentioned in *ANOVA/MANOVA*, the multivariate Box *M* test for homogeneity of variances/covariances is particularly sensitive to deviations from multivariate normality, and should not be taken too “seriously.”

Correlations between Means and Variances

The major “real” threat to the validity of significance tests occurs when the means for variables across groups are correlated with the variances (or standard deviations). Intuitively, if there is large variability in a group with particularly high means on some variables, then those high means are not reliable. However, the overall significance tests are based on pooled variances, that is, the average variance across all groups. Thus, the significance tests of the relatively larger means (with the large variances) would be based on the relatively smaller pooled variances, resulting erroneously in statistical significance. In practice, this pattern may occur if one group in the study contains a few extreme outliers, who have a large impact on the means, and also increase the variability. To guard against this problem, inspect the descriptive statistics, that is, the means and standard deviations or variances for such a correlation.

The Matrix Ill-conditioning Problem

Another assumption of discriminant function analysis is that the variables that are used to discriminate between groups are not completely redundant. As part of the computations involved in discriminant analysis, you will invert the variance/covariance matrix of the variables in the model. If any one of the variables is completely redundant with the other variables then the matrix is said to be *ill-conditioned*, and it cannot be inverted. For example, if a variable is the sum of three other variables that are also in the model, then the matrix is ill-conditioned.

Tolerance Values

In order to guard against matrix ill-conditioning, constantly check the so-called tolerance value for each variable. This tolerance value is computed as *1 minus R-square* of the respective variable with all other variables included in the

current model. Thus, it is the proportion of variance that is unique to the respective variable. You may also refer to Multiple Regression to learn more about multiple regression and the interpretation of the tolerance value. In general, when a variable is almost completely redundant (and, therefore, the matrix ill-conditioning problem is likely to occur), the tolerance value for that variable will approach 0.

Classification

Another major purpose to which discriminant analysis is applied is the issue of predictive classification of cases. Once a model has been finalized and the discriminant functions have been derived, how well can we *predict* to which group a particular case belongs?

A priori and post hoc predictions

Before going into the details of different estimation procedures, we would like to make sure that this difference is clear. Obviously, if we estimate, based on some data set, the discriminant functions that best discriminate between groups, and then use the *same* data to evaluate how accurate our prediction is, then we are very much capitalizing on chance. In general, one will *always* get a worse classification when predicting cases that were not used for the estimation of the discriminant function. Put another way, *post hoc* predictions are always better than *a priori* predictions. (The trouble with predicting the future *a priori* is that one does not know what will happen; it is much easier to find ways to predict what we already know has happened.) Therefore, one should never base one's confidence regarding the correct classification of future observations on the same data set from which the discriminant functions were derived; rather, if one wants to classify cases predictively, it is necessary to collect new data to "try out" (cross-validate) the utility of the discriminant functions.

Classification Functions

These are not to be confused with the discriminant functions. The classification functions can be used to determine to which group each case most likely belongs. There are as many classification functions as there are groups. Each function allows us to compute *classification scores* for each case for each group, by applying the formula:

$$S_i = c_i + w_{i1} * x_1 + w_{i2} * x_2 + \dots + w_{im} * x_m$$

In this formula, the subscript i denotes the respective group; the subscripts $1, 2, \dots, m$ denote the m variables; c_i is a constant for the i 'th group, w_{ij} is the weight for the j 'th variable in the computation of the classification score for the i 'th group; x_j is the observed value for the respective case for the j 'th variable. S_i is the resultant classification score.

We can use the classification functions to directly compute classification scores for some new observations.

Classification of Cases

Once we have computed the classification scores for a case, it is easy to decide how to classify the case: in general we classify the case as belonging to the group for which it has the highest classification score. Thus, if we were to study high school students' post-graduation career/educational choices based on several variables assessed one year prior to graduation, we could use the classification functions to predict what each student is most likely to do after graduation. However, we would also like to know the *probability* that the student will make the predicted choice. Those probabilities are called *posterior* probabilities, and can also be computed. However, to understand how those probabilities are derived, let us first consider the so-called *Mahalanobis* distances.

Mahalanobis Distances

You may have read about these distances in other parts of the manual. In general, the Mahalanobis distance is a measure of distance between two points in the space defined by two or more correlated variables. For example, if there are two variables that are uncorrelated, then we could plot points (cases) in a standard two-dimensional scatterplot; the Mahalanobis distances between the points would then be identical to the Euclidean distance; that is, the distance as, for example, measured by a ruler. If there are three uncorrelated variables, we could also simply use a ruler (in a 3-D plot) to determine the distances between points. If there are more than 3 variables, we cannot represent the distances in a plot any more. Also, when the variables are correlated, then the axes in the plots can be thought of as being *non-orthogonal*; that is, they would not be positioned in right angles to each other. In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations.

Probability is the measure of the likelihood that an event will occur.

Mahalanobis Distances and Classification

For each group in our sample, we can determine the location of the point that represents the means for all variables in the multivariate space defined by the variables in the model. These points are called group *centroids*. For each case we can then compute the Mahalanobis distances (of the respective case) from each of the group centroids. Again, we would classify the case as belonging to the group to which it is closest, that is, where the Mahalanobis distance is smallest.

Posterior Classification Probabilities

Using the Mahalanobis distances to do the classification, we can now derive probabilities. The **probability** that a case belongs to a particular group is basically proportional to the Mahalanobis distance from that group centroid (it is not exactly proportional because we assume a multivariate normal distribution around each centroid). Because we compute the location of each case from our prior knowledge of the values for that case on the variables in the model, these probabilities are called posterior probabilities. In summary, the posterior probability is the probability, based on our knowledge of the values of other variables, that the respective case belongs to a particular group. Some software packages will automatically compute those probabilities for all cases.

A priori classification probabilities. There is one additional factor that needs to be considered when classifying cases. Sometimes, we know ahead of time that there are more observations in one group than in any other; thus, the a priori probability that a case belongs to that group is higher. For example, if we know ahead of time that 60% of the graduates from our high school usually go to college, then we should adjust our prediction accordingly: a priori, and all other things being equal, it is more likely that a student will attend college than choose either of the other two options. You can specify different a priori probabilities, which will then be used to adjust the classification of cases (and the computation of posterior probabilities) accordingly.

In practice, the researcher needs to ask him or herself whether the unequal number of cases in different groups in the sample is a reflection of the true distribution in the population, or whether it is only the (random) result of the sampling procedure. In the former case, we would set the a priori probabilities to be proportional to the sizes of the groups in our sample, in the latter case we would specify the a priori probabilities as being equal in each group. The specification of different a priori probabilities can greatly affect the accuracy of the prediction.



CORE CONCEPT: MAHALANOBIS DISTANCE

The Mahalanobis distance (MD) is the distance between two points in multivariate space. In a regular Euclidean space, variables (e.g. x , y , z) are represented by axes drawn at right angles to each other; The distance between any two points can be measured with a ruler. For uncorrelated variables, the Euclidean distance equals the MD.

Many machine learning techniques make use of distance calculations as a measure of similarity between two points. For example, in k-means clustering, we assign data points to clusters by calculating and comparing the distances to each of the cluster centers. Similarly, Radial Basis Function (RBF) Networks, such as the RBF SVM, also make use of the distance between the input vector and stored prototypes to perform classification.

What happens, though, when the components have different variances, or there are correlations between components? In this post, I'll be looking at why these data statistics are important, and describing the Mahalanobis distance, which takes these into account.

First, a note on terminology. "Covariance" and "correlation" are similar concepts; the correlation between two variables is equal to their covariance divided by their variances, as explained here. For our discussion, they're essentially interchangeable, and you'll see me using both terms below.

Differences in Component Variances

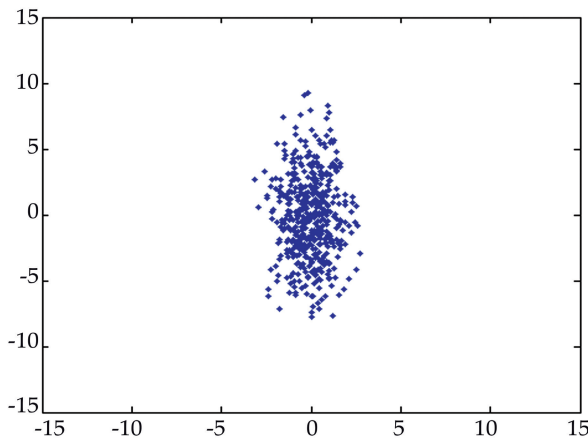
Let's start by looking at the effect of different variances, since this is the simplest to understand.

Consider the following cluster, which has a multivariate distribution. This cluster was generated from a normal distribution with a horizontal variance of 1 and a vertical variance of 10, and no covariance. In order to assign a point to this cluster, we know intuitively that the distance in the horizontal dimension should be given a different weight than the distance in the vertical direction.

We can account for the differences in variance by simply dividing the component differences by their variances.

First, here is the component-wise equation for the Euclidean distance (also called the "L2" distance) between two vectors, x and y :

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$$



Let's modify this to account for the different variances. Using our above cluster example, we are going to calculate the adjusted distance between a point 'x' and the centre of this cluster 'c'.

σ_1^2 - Horizontal variance

σ_2^2 - Vertical variance

$$d(x, c) = \sqrt{\frac{(x_1 - c_1)^2}{\sigma_1^2} + \frac{(x_2 - c_2)^2}{\sigma_2^2}}$$

The Mahalanobis Distance

The equation above is equivalent to the Mahalanobis distance for a two dimensional vector with no **covariance**. The general equation for the Mahalanobis distance uses the full covariance matrix, which includes the covariances between the vector components.

Covariance is a measure of the joint variability of two random variables.

Before looking at the Mahalanobis distance equation, it's helpful to point out that the Euclidean distance can be re-written as a dot-product operation:

$$d_{L2}(x, y) = \sqrt{(x - y)^T (x - y)}$$

With that in mind, below is the general equation for the Mahalanobis distance between two vectors, x and y, where S is the covariance matrix.

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Note: As you might expect, the probability density function for a multivariate Gaussian distribution uses the Mahalanobis distance instead of the Euclidean.

The Mahalanobis Distance with Zero Covariance

Before we move on to looking at the role of correlated components, let's first walk through how the Mahalanobis distance equation reduces to the simple two dimensional example from early in the post when there is no correlation.

Assuming no correlation, our covariance matrix is:

$$S = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

The inverse of a 2x2 matrix can be found using the following:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Applying this to get the inverse of the covariance matrix:

$$S^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2} \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$$

Now we can work through the Mahalanobis equation to see how we arrive at our earlier variance-normalized distance equation.

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

$$\begin{aligned} &= \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}} \\ &= \sqrt{\begin{bmatrix} \frac{x_1 - y_1}{\sigma_1^2} & \frac{x_2 - y_2}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}} \\ &= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2}} \end{aligned}$$

Correlation

So far we've just focused on the effect of variance on the distance calculation. But when happens when the components are correlated in some way?

Correlation is computed as part of the covariance matrix, S. For a dataset of m samples, where the i th sample is denoted as $x^{(i)}$, the **covariance matrix** S is computed as:

$$S = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

Note that the placement of the transpose operator creates a matrix here, not a single value.

For two dimensional data (as we've been working with so far), here are the equations for each individual cell of the 2x2 covariance matrix, so that you can get more of a feel for what each element represents.

$$\begin{array}{c|c} \frac{1}{m} \sum_{i=1}^m (x_1 - \mu_1)^2 & \frac{1}{m} \sum_{i=1}^m (x_1 - \mu_1)(x_2 - \mu_2) \\ \hline \frac{1}{m} \sum_{i=1}^m (x_2 - \mu_2)(x_1 - \mu_1) & \frac{1}{m} \sum_{i=1}^m (x_2 - \mu_2)^2 \end{array}$$

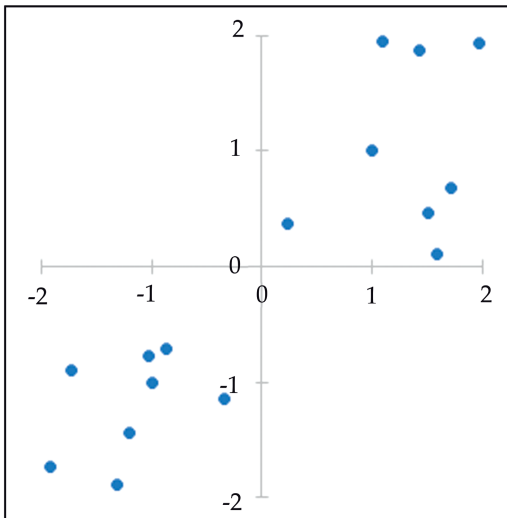
If you subtract the means from the dataset ahead of time, then you can drop the “minus mu” terms from these equations.

Subtracting the means causes the dataset to be centered around (0, 0). It's critical to appreciate the effect of this mean-subtraction on the signs of the values. Your original dataset could be all positive values, but after moving the mean to (0, 0), roughly half the component values should now be negative.

The top-left corner of the covariance matrix is just the variance—a measure of how much the data varies along the horizontal dimension. Similarly, the bottom-right corner is the variance in the vertical dimension.

The bottom-left and top-right corners are identical. These indicate the correlation between x_1 and x_2 .

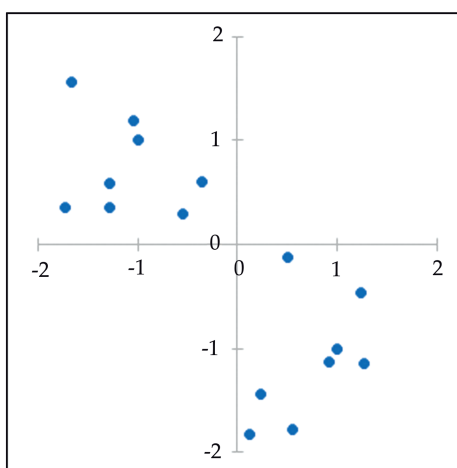
If the data is mainly in quadrants one and three, then all of the $x_1 * x_2$ products are going to be positive, so there's a *positive* correlation between x_1 and x_2 .



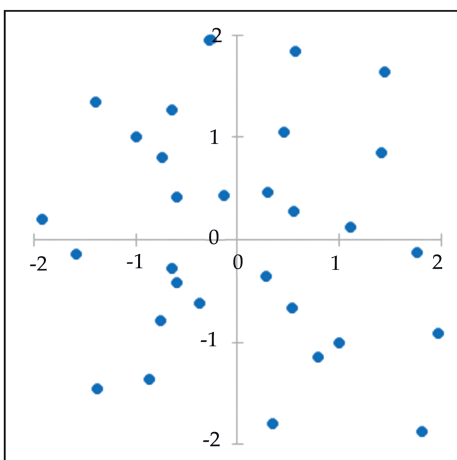
Covariance

matrix is a type of matrix that is used to represent the covariance values between pairs of elements given in a random vector.

If the data is all in quadrants two and four, then the all of the products will be negative, so there's a *negative* correlation between x_1 and x_2 .



If the data is evenly dispersed in all four quadrants, then the positive and negative products will cancel out, and the covariance will be roughly zero. This indicates that there is no correlation.

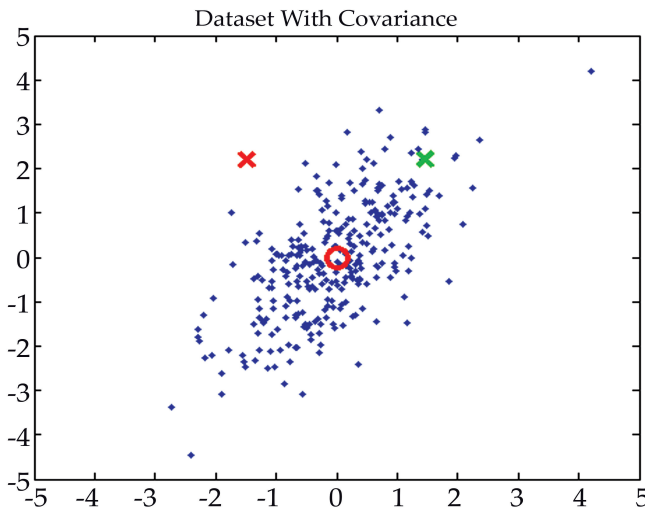


As another example, imagine two pixels taken from different places in a black and white image. If the pixels tend to have the same value, then there is a positive correlation between them.

If the pixels tend to have opposite brightnesses (e.g., when one is black the other is white, and vice versa), then there is a negative correlation between them. If the pixel values are entirely independent, then there is no correlation.

Effect of Correlation on Distance

To understand how correlation confuses the distance calculation, let's look at the following two-dimensional example. The cluster of blue points exhibits positive correlation.



I've marked two points with X's and the mean $(0, 0)$ with a red circle. Looking at this plot, we know intuitively the red X is less likely to belong to the cluster than the green X. However, I selected these two points so that they are equidistant from the center $(0, 0)$. Even taking the horizontal and vertical variance into account, these points are still nearly equidistant from the center.

It's clear, then, that we need to take the correlation into account in our distance calculation.

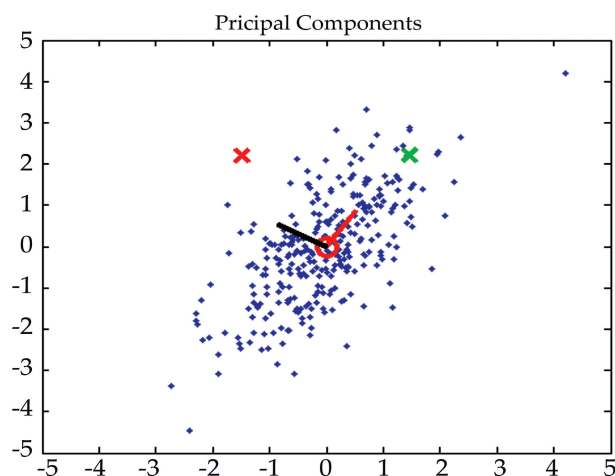
The Mahalanobis distance takes correlation into account; the covariance matrix contains this information. However, it's difficult to look at the Mahalanobis equation and gain an intuitive understanding as to how it actually does this.

We can gain some insight into it, though, by taking a different approach. Instead of accounting for the covariance using Mahalanobis, we're going to transform the data to remove the correlation and variance.

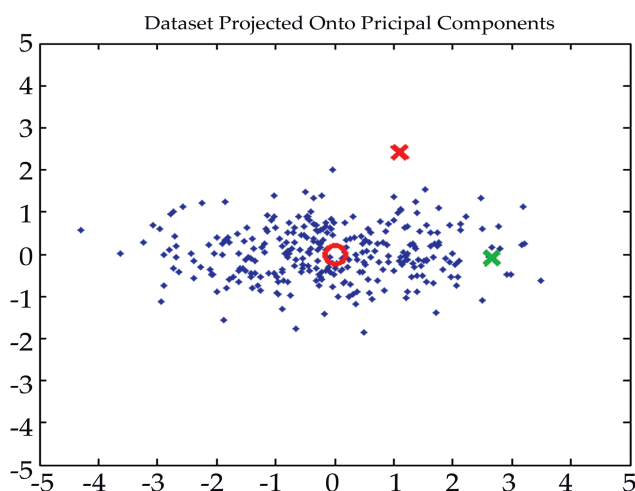
We'll remove the correlation using a technique called **Principal Component Analysis (PCA)**. To perform PCA, you calculate the eigenvectors of the data's covariance matrix. The two eigenvectors are the principal components. I've overlaid the eigenvectors on the plot. You can see that the first principal component, drawn in red, points in the direction of the highest variance in the data. The second principal component, drawn in black, points in the direction with the second highest variation.

Principal component analysis (PCA)

is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.



Using these vectors, we can rotate the data so that the highest direction of variance is aligned with the x-axis, and the second direction is aligned with the y-axis. This rotation is done by projecting the data onto the two principal components.



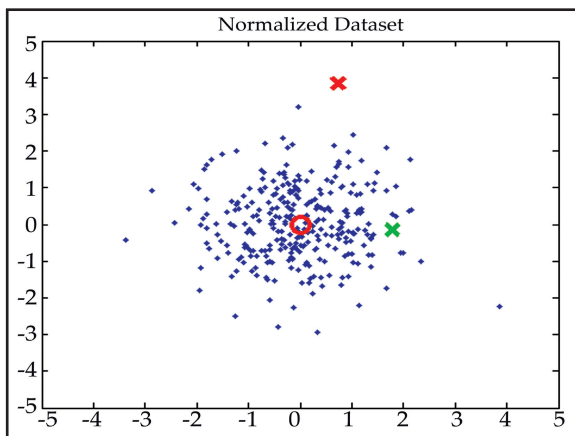
If we calculate the covariance matrix for this rotated data, we can see that the data now has zero covariance:

2.2396 0.0000

0.0000 0.3955

What does it mean that there's no correlation? Just that the data is evenly distributed among the four quadrants around $(0, 0)$. We've rotated the data such that the slope of the trend line is now zero.

You'll notice, though, that we haven't really accomplished anything yet in terms of normalizing the data. The two points are still equidistant from the mean. However, the principal directions of variation are now aligned with our axes, so we can normalize the data to have unit variance (we do this by dividing the components by the square root of their variance). This turns the data cluster into a sphere.



And now, finally, we see that our green point is closer to the mean than the red. Hurray!

Given that removing the correlation alone didn't accomplish anything, here's another way to interpret correlation: Correlation implies that there is some variance in the data *which is not aligned with the axes*. It's still variance that's the issue, it's just that we have to take into account the direction of the variance in order to normalize it properly.

A CONCEPT OF PARTIAL CORRELATION

Partial Correlation is a statistical technique used to measure the strength and direction of a linear relationship between two variables while controlling for the effects of one or more additional variables. It allows you to examine the unique relationship between two variables by removing the influence of other variables that might be confounding the analysis.

In simpler terms, partial correlation helps you understand the direct relationship between two variables after accounting for the effects of other variables that might be affecting both of them.

Key Points about Partial Correlation:

- **Controlling for Confounding Variables:** When you observe a correlation between two variables, there might be a third variable (a confounding variable) that affects both of them. Partial correlation helps isolate the relationship between the two variables of interest while neutralizing the influence of the third variable.
- **Formula:** The formula for calculating the partial correlation coefficient ($r_{xy \cdot z}$) between two variables X and Y , while controlling for a third variable Z , is:

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

Here, r_{xy} is the correlation between X and Y , r_{xz} is the correlation between X and Z , and r_{yz} is the correlation between Y and Z .

- **Interpretation:** The resulting partial correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative partial correlation, 1 indicates a perfect positive partial correlation, and 0 indicates no partial correlation.
- **Example:** Suppose you are studying the relationship between sleep quality (X) and academic performance (Y) among students. You suspect that the amount of coffee consumed (Z) might influence both variables. By calculating the partial correlation between X and Y while controlling for Z , you can determine if the relationship between sleep quality and academic performance remains significant when accounting for coffee consumption.

Partial correlation is a valuable tool in research, especially when studying complex relationships among multiple variables. It helps researchers disentangle the direct associations between variables from the potential confounding effects of other variables, leading to more accurate and meaningful insights.

SUMMARY

- Multivariate Analysis is defined as a process involving multiple dependent variables resulting in one outcome.
- The basic idea was to summarize the population or sample data through smaller sized matrices or simply numbers. All the quantities (except correlation) defined there were straightforward generalizations of their univariate counterparts.
- A probability distribution that plays a pivotal role in multivariate analysis is multivariate normal distribution.
- The bivariate distribution represents the joint distribution of two random variables.
- The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.
- The regression line expresses the best prediction of the dependent variable (Y), given the independent variables (X).
- Multiple regression is a seductive technique: “plug in” as many predictor variables as you can think of and usually at least a few of them will come out significant.
- Discriminant analysis is a statistical method that is used by researchers to help them understand the relationship between a “dependent variable” and one or more “independent variables.”
- Probably the most common application of discriminant function analysis is to include many measures in the study, in order to determine the ones that discriminate between groups.

MULTIPLE CHOICE QUESTIONS

1. **What does multivariate analysis involve?**
 - a. Analyzing data with only one variable
 - b. Analyzing data with multiple variables
 - c. Analyzing data with a single categorical variable
 - d. Analyzing data with a single continuous variable
2. **Which of the following is a goal of multivariate analysis?**
 - a. Reducing data to a single variable
 - b. Ignoring relationships between variables
 - c. Exploring patterns among multiple variables
 - d. Analyzing data from a single source
3. **Principal Component Analysis (PCA) is used for:**
 - a. Creating additional variables
 - b. Expanding data dimensions
 - c. Data visualization and dimensionality reduction
 - d. Hypothesis testing
4. **Canonical Correlation Analysis (CCA) is used to study:**
 - a. The relationship between a dependent variable and an independent variable
 - b. The relationship between two sets of variables
 - c. The variance within a single variable
 - d. The distribution of a single variable
5. **Cluster analysis is used for:**
 - a. Studying linear relationships between variables
 - b. Grouping similar observations based on multiple variables
 - c. Analyzing a single variable's distribution
 - d. Studying the central tendency of data

Answers

1. (b) 2. (c) 3. (c) 4. (b) 5. (b)

REVIEW QUESTIONS

1. Discuss about multivariate statistical analysis.
2. What is the difference between bivariate and multivariate analysis?
3. What is the multivariate normal distribution?
4. Differentiate between unique prediction and partial correlation.
5. Explain the importance of residual analysis.
6. Write the effect of correlation on distance.

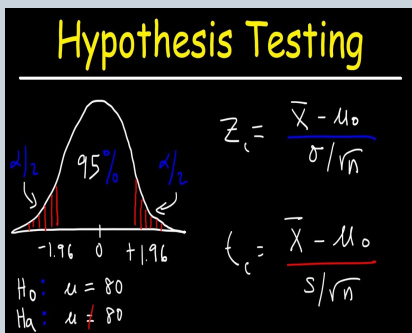
REFERENCES

1. De Leeuw, J., & Meulman, J. (2003). Introduction to Multidimensional Scaling. In P. M. Green, P. E. Sørensen, & R. H. Ysseldyke (Eds.), *Handbook of Statistics, Volume 21: Behavioral Statistics in Action* (pp. 363-397). Elsevier.
2. Everitt, B. S., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer.
3. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis*. Pearson.
4. Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.
5. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
6. Jolliffe, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
7. Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis*. Wiley.
8. Shalizi, C. R., & Shalizi, K. L. (2013). *Advanced Data Analysis from an Elementary Point of View*. Vol. 171. CRC Press.
9. Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson.
10. Tatsuoka, M. M. (2002). *Data Analysis Methods for Multivariate Social Science Research*. Plenum.



CHAPTER 5

TESTS OF HYPOTHESIS



Tests of Hypothesis, also known as statistical hypothesis tests or significance tests, are fundamental tools in statistics that help us make decisions about population parameters based on sample data. These tests allow us to determine whether observed differences or relationships in the sample data are statistically significant or if they could have occurred due to random chance.

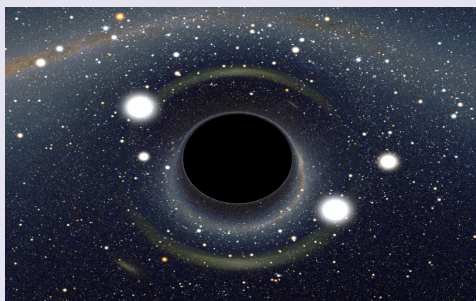
After studying this chapter, you will understand the core concepts of:

- Nature of Hypothesis
- Tests of Hypotheses on the Population Mean
- Types of Hypotheses
- Hypothesis Testing and Sampling Distributions

INTRODUCTION

Hypothesis testing is a statistical method used to evaluate a hypothesis or claim about a population parameter based on sample data. In hypothesis testing, we start with a null hypothesis (H_0), which is a statement about the value of a population parameter that we want to test. We also have an alternative hypothesis (H_a), which is a statement that contradicts the null hypothesis and represents the possibility of an effect or difference.

The hypothesis test is aimed to test if the null hypothesis should be rejected in favor of the alternative hypothesis. The basic logic of a hypothesis test is to compare two statistical data sets. One data set is obtained by sampling and the other data set originates from an idealized model. If the sample data is consistent with the idealized model, the null hypothesis is not rejected. If the sample data is inconsistent with the idealized model and thus supports an alternative hypothesis, the null hypothesis is rejected in favor of the alternative hypothesis.



CORE CONCEPT: NATURE OF HYPOTHESIS

For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. The adjective hypothetical, meaning “having the nature of a hypothesis”, or “being assumed to exist as an immediate consequence

of a hypothesis”, can refer to any of these meanings of the term “hypothesis”.

The various discussions of the hypothesis which have appeared in works on inductive logic and in writings on scientific method, its structure and function have received considerable attention, while its origin has been comparatively neglected.

The hypothesis has generally been treated as that part of scientific procedure which marks the stage where a definite plan or method is proposed for dealing with new or unexplained facts.

It is regarded as an invention for the purpose of explaining the given, as a definite conjecture which is to be tested by an appeal to experience to see whether deductions made in accordance with it will be found true in fact.

The function of the hypothesis is to unify, to furnish a method of dealing with things, and its structure must be suitable to this end. It must be so formed that it will be likely to prove valid, and writers have formulated various rules to be followed

in the formation of hypotheses. These rules state the main requirements of a good hypothesis, and are intended to aid in a general way by pointing out certain limits within which it must fall.

In respect to the origin of the hypothesis, writers have usually contented themselves with pointing out the kind of situations in which hypotheses are likely to appear. But after this has been done, after favorable external conditions have been given, the rest must be left to “genius,” for hypotheses arise as “happy guesses,” for which no rule or law can be given. In fact, the genius differs from the ordinary plodding mortal in just this ability to form fruitful.

Hypotheses in the midst of the same facts which to other less gifted individuals remain only so many disconnected experiences. Hypothesis is to determine its nature a little more precisely through an investigation of its rather obscure origin, and to call attention to certain features of its function which have not generally been accorded their due significance.

Testing of Hypothesis

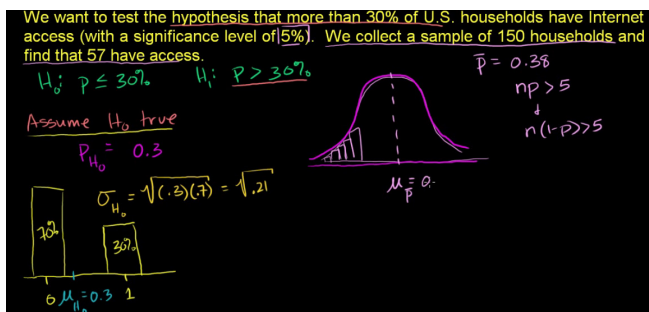
Hypothesis testing refers to the process of using statistical analysis to determine if the observed differences between two or more samples are due to random chance or to true differences in the samples.

Hypothesis testing is the process of using a variety of statistical tools to analyze data and, ultimately, to fail to reject or reject the null hypothesis. From a practical point of view, finding statistical evidence that the null hypothesis is false allows you to reject the null hypothesis and accept the alternate hypothesis. Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.

The usual process of hypothesis testing consists of four steps:

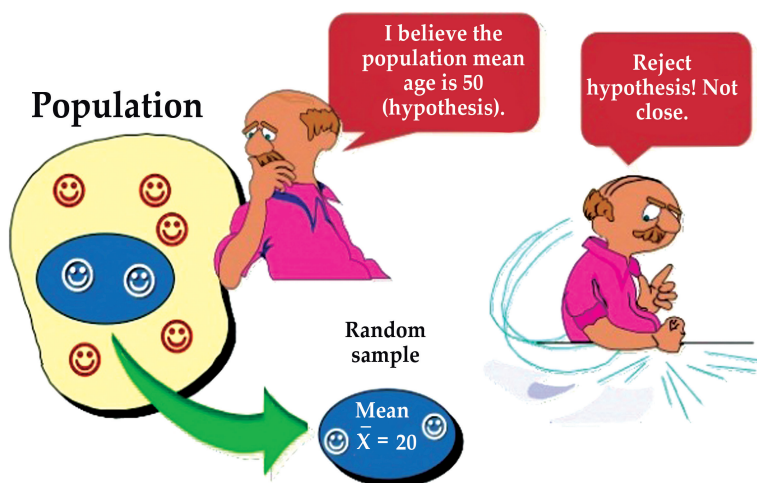
- Formulate the null hypothesis (commonly, that the observations are the result of pure chance) and the alternative hypothesis (commonly, that the observations show a real effect combined with a component of chance variation).
- Identify a test statistic that can be used to assess the truth of the null hypothesis.
- Compute the P-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true.
- Compare the -value to an acceptable significance value (sometimes called an alpha value). If, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid

We always do hypothesis testing for population parameters. We never hypothesized sample statistics.

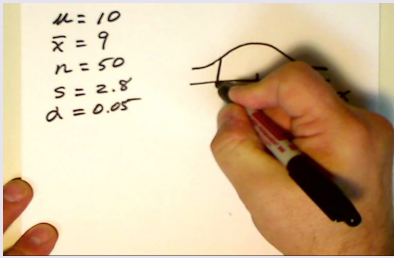


A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population. A hypothesis test examines two opposing hypotheses about a population: the null hypothesis and the alternative hypothesis. The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of “no effect” or “no difference”. The alternative hypothesis is the statement you want to be able to conclude is true.

HYPOTHESIS TESTING



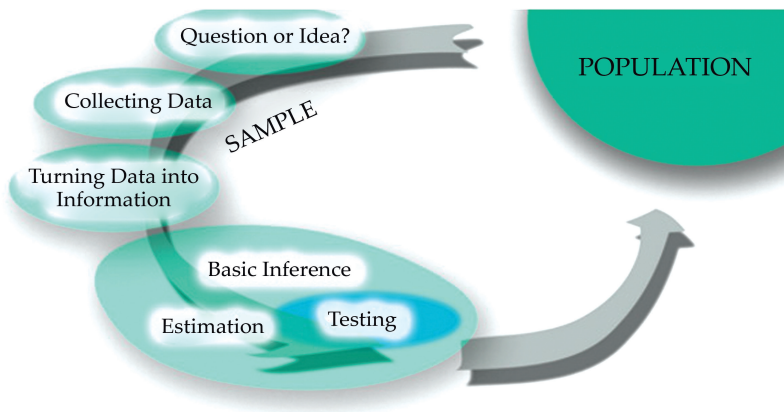
Based on the sample data, the test determines whether to reject the null hypothesis. You use a p-value, to make the determination. If the p-value is less than or equal to the level of significance, which is a cut-off point that you define, then you can reject the null hypothesis. A common misconception is that statistical hypothesis tests are designed to select the more likely of two hypotheses. Instead, a test will remain with the null hypothesis until there is enough evidence (data) to support the alternative hypothesis.



CORE CONCEPT: TESTS OF HYPOTHESES ON THE POPULATION MEAN

Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

A statistical hypothesis is an assumption about a population parameter. This assumption may or may not be true.



Statistical Hypotheses

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

There are two types of statistical hypotheses.

- *Null hypothesis:* The null hypothesis, denoted by H_0 , is usually the hypothesis that sample observations result purely from chance.
- *Alternative hypothesis:* The alternative hypothesis, denoted by H_1 or H_a , is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as

$$H_0: P = 0.5$$

$$H_a: P \neq 0.5$$

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

The Testing Process

In the statistics literature, statistical hypothesis testing plays a fundamental role. Statisticians follow a formal process to determine whether to reject a null hypothesis, based on sample data.

This process, called hypothesis testing, consists of four steps:

- *State the hypotheses:* This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.
- *Formulate an analysis plan:* The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
- *Analyze sample data:* Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.
- *Interpret results:* Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.



Hypothesis is a proposed explanation for a phenomenon.

Let us now consider the actual testing. First of all, there seem to be various degrees of ‘testability’. The strongest requirement for an empirical hypothesis is that it should be falsifiable: it should be possible to prove the **hypothesis** wrong. For instance, a descriptive hypothesis that all ravens are black would be falsified by the occurrence of a single non-black

raven. In the human sciences, most hypotheses are weaker: to falsify the claim that most people have two legs I cannot simply find a single one-legged person, but I have to make a statistical survey to see whether, in a representative sample (typical, non-biased, e.g. not made in a hospital amputation clinic), there are indeed more people with two legs than with some other number. Hypotheses formulated as tendencies are weaker still, for we can always argue about what we mean by a tendency. What percentage of instances (increasing at what rate over what period of time) would constitute a tendency? This is where we need inferential statistics.

If a hypothesis is not, strictly speaking, falsifiable, a weaker requirement is that it should nevertheless be testable. A claim that cannot be tested at all is not worth making, from an empirical point of view: it would be mere speculation. It may nevertheless be the case that if a hypothesis cannot be tested directly, it still has testable consequences. This is an important point when we test interpretive hypotheses. Because these are hypotheses about how best to interpret (understand) something, they cannot actually be falsified: they are not claims about the distribution or causation of features of empirical reality, but about the usefulness of particular ways of making sense of these features. Interpretive hypotheses are thus tested in use, by seeing what benefits they bring in conceptualizing the object of study. If they do not turn out to be beneficial, for ex-ample in facilitating fruitful analysis or generating new empirical hypotheses, they will simply fade out of use.

For instance, suppose you claim that it would be a good idea to teach translator trainees to think in such-and-such a way about translation: to think of translation in terms of, say, creative performance. This claim (an interpretive hypothesis about how best to understand translation) has testable consequences if we assume that if translators think in this new way, then their translations will be somehow different, perhaps better. That is, their translations will have different profile features and we can then go ahead and test for the presence of these features, in comparison with translations produced by trainees who had not been expressly taught to think about translation in this way.

Phenomenon is anything which manifests itself.

Hypotheses can be tested on four criteria: these are the ACID tests.....

- A for Added value in general: new understanding
- C for Comparative value, in comparison with other hypotheses

- I for Internal value: logic, clarity, elegance, economy
- D against Data, empirical evidence.

Testing means checking these things:

- Ascertain that the hypothesis does indeed add to our understanding of the **phenomenon**, it brings something new, it is not trivial, it is genuinely interesting
- Check it against other competing hypotheses: in what respects is it better than others?
- Is it logical, elegant, and parsimonious (economical), with no unnecessary concepts or assumptions? Is it plausible?
- Does it accurately represent the empirical evidence? Does it account for the facts? Does it cover a wider variety of data, is it more general than competing hypotheses?

Whatever the hypothesis, it is worth bearing in mind that, strictly speaking, a hypothesis can never be proved true, or confirmed to be true. Science does not proceed by piling up truths, but by developing better and better hypotheses, which may well approximate closer and closer to being accurate descriptions or explanations of reality. An empirical test may support a hypothesis, or corroborate it; or it may not support it; or it may falsify it. In Translation Studies the results of a single test are seldom conclusive, one way or the other.

Hypothesis-testing often pertains to the scope of a claim. Unrestricted claims are maximally general, such as those proposing translation universals. But most hypotheses specify a narrower scope of application. For instance, someone might claim that the retranslation hypothesis only holds for fiction, not non-fiction. Tests might then suggest that this scope is still too wide, and that the hypothesis should only apply to certain kinds of fiction. Other tests might suggest the opposite conclusion: that the hypothesis actually has a wider scope than fiction alone. Tests are thus carried out, among other things, in order to determine the scope within which a given hypothesis seems to be valid.

If a hypothesis turns out to be supported by empirical evidence, this of course might simply be due to chance, so a replication of the test might be needed. Replications of hypothesis-testing procedures are a standard part of empirical science, but so far they have been relatively rare in *Translation Studies*. For a test to be replicated, the methodology must be described explicitly, in enough detail. Different tests might also be warranted, to check the validity of the hypothesis. A well-corroborated hypothesis can then lead to further generalizations, so that understanding grows.

To perform hypothesis testing, we first define the null and alternative hypotheses based on the research question and the sample data. Then, we choose a suitable statistical test and calculate a test statistic and its corresponding p-value.

If a hypothesis is not supported, this is usually an interesting result in itself and may be valuable especially if the hypothesis seemed to be well justified in the first place. Such a result raises new questions. Was the empirical test perhaps inappropriate or not sensitive enough? Was the material badly chosen, not typical, not valid? Were the calculations wrong, not reliable? If you come to suspect the test itself rather than the hypothesis, the next stage is to test again, or to replicate the test on other material. Or maybe the hypothesis itself needs to be refined, or even rejected?

Decision Errors

Two types of errors can result from a hypothesis test.

- *Type I error:* A Type I error occurs when, it rejects a null hypothesis when it is true. The probability of committing a Type I error is called the significance level. This probability is also called alpha, and is often denoted by α .
- *Type II error:* A Type II error occurs when, it fails to reject a null hypothesis that is false. The probability of committing a Type II error is called Beta, and is often denoted by β . The probability of *not* committing a Type II error is called the Power of the test.

Decision Rules

The analysis plan includes decision rules for rejecting the null hypothesis. In practice, statisticians describe these decision rules in two ways - with reference to a P-value or with reference to a region of acceptance.

- *P-value.* The strength of evidence in support of a null hypothesis is measured by the P-value. Suppose the test statistic is equal to S . The P-value is the probability of observing a test statistic as extreme as S , assuming the null hypothesis is true. If the P-value is less than the significance level, we reject the null hypothesis.
- *Region of acceptance.* The region of acceptance is a range of values. If the test statistic falls within the region of acceptance, the null hypothesis is not rejected. The region of acceptance is defined so that the chance of making a Type I error is equal to the significance level.

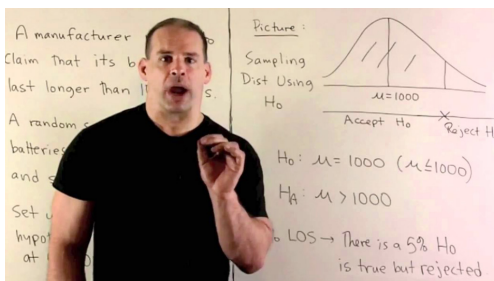
Sampling distribution is a concept used in statistics. It is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

The set of values outside the region of acceptance is called the region of rejection. If the test statistic falls within the region of rejection, the null hypothesis is rejected. In such cases, we say that the hypothesis has been rejected at the α level of significance.

One-Tailed and Two-Tailed Tests

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test. *For example*, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of **sampling distribution**; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test. For example, suppose the null hypothesis states that the mean is equal to 10. The alternative hypothesis would be that the mean is less than 10 or greater than 10. The region of rejection would consist of a range of numbers located on both sides of sampling distribution; that is, the region of rejection would consist partly of numbers that were less than 10 and partly of numbers that were greater than 10.



Population Proportion

Parameter is any summary number, like an average or percentage that describes the entire population.

A population proportion, generally denoted by P and in some textbooks by π , is a parameter that describes a percentage value associated with a population. *For example*, the 2010 United States Census showed that 83.7% of the American Population was identified as not being Hispanic or Latino. The value of .837 is a population proportion. In general, the population proportion and other population parameters are unknown. A census can be conducted in order to determine the actual value of a population **parameter**, but often a census is not practical due to its costs and time consumption.

$$\begin{aligned} & \hat{p} \pm \text{margin of error} \\ & \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ & \left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \end{aligned}$$

A population proportion is usually estimated through an unbiased sample statistic obtained from an observational study or experiment. For example, the National Technological Literacy Conference conducted a national survey of 2,000 adults to determine the percentage of adults who are economically illiterate. The study showed that 72% of the 2,000 adults sampled did not understand what a gross domestic product is. The value of 72% is a sample proportion.

P-value Approach

The *P*-value approach involves determining “likely” or “unlikely” by determining the probability — assuming the null hypothesis were true — of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed. If the *P*-value is small, say less than (or equal to) α , then it is “unlikely.” And, if the *P*-value is large, say more than α , then it is “likely.”

If the *P*-value is less than (or equal to) α , then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the *P*-value is greater than α , then the null hypothesis is not rejected.

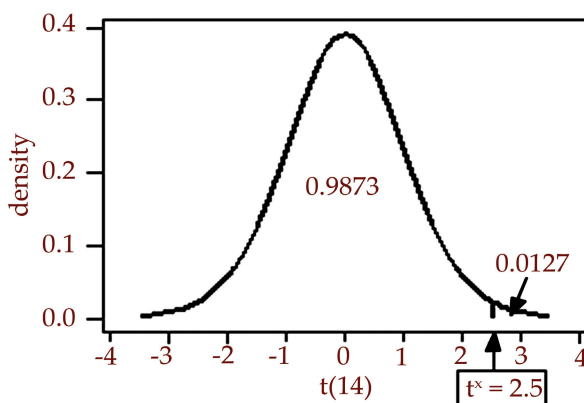
Specifically, the four steps involved in using the *P*-value approach to conducting any hypothesis test are:

- Specify the null and alternative hypotheses.
- Using the sample data and assuming the null hypothesis is true, calculate the value of the test statistic. Again, to conduct the hypothesis test for the population mean μ , we use the *t*-statistic $t^* = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ which follows a *t*-distribution with $n - 1$ degrees of freedom.
- Using the known distribution of the test statistic, calculate the ***P***-value: “If the null hypothesis is true, what is the probability that we’d observe a more extreme test statistic in the direction of the alternative hypothesis than we did?” (Note how this question is equivalent to the question answered in criminal trials: “If the defendant is innocent, what is the chance that we’d observe such extreme criminal evidence?”)
- Set the significance level, α , the probability of making a Type I error to be small — 0.01, 0.05, or 0.10. Compare the *P*-value to α . If the *P*-value is less than (or equal to) α , reject the null hypothesis in favor of the alternative hypothesis. If the *P*-value is greater than α , do not reject the null hypothesis.

In our example concerning the mean grade point average, suppose that our random

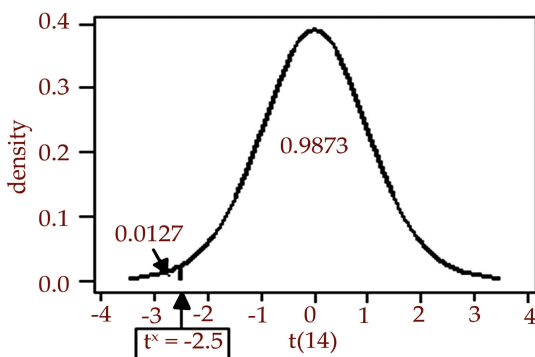
sample of $n = 15$ students majoring in mathematics yields a test statistic t^* equaling 2.5. Since $n = 15$, our test statistic t^* has $n - 1 = 14$ degrees of freedom. Also, suppose we set our significance level α at 0.05, so that we have only a 5% chance of making a Type I error.

The P -value for conducting the right-tailed test $H_0: \mu = 3$ versus $H_A: \mu > 3$ is the probability that we would observe a test statistic greater than $t^* = 2.5$ if the population mean μ really were 3. Recall that probability equals the area under the probability curve. The P -value is therefore the area under a $t_{n-1} = t_{14}$ curve and to the *right* of the test statistic $t^* = 2.5$. It can be shown using statistical software that the P -value is 0.0127:



The P -value, 0.0127, tells us it is “unlikely” that we would observe such an extreme test statistic t^* in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the P -value, 0.0127, is less than $\alpha = 0.05$, we reject the null hypothesis $H_0: \mu = 3$ in favor of the alternative hypothesis $H_A: \mu > 3$.

Note that we would not reject $H_0: \mu = 3$ in favor of $H_A: \mu > 3$ if we lowered our willingness to make a Type I error to $\alpha = 0.01$ instead, as the P -value, 0.0127, is then greater than $\alpha = 0.01$.



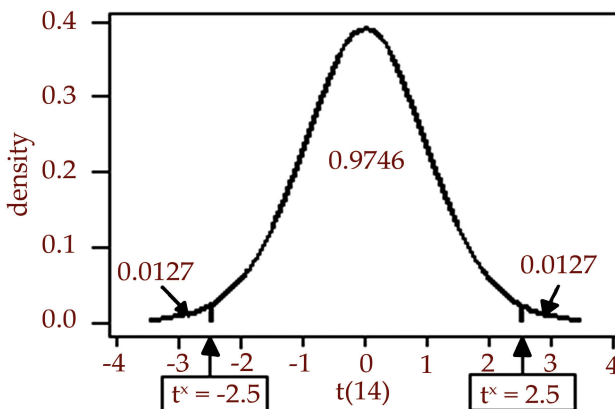
In our example concerning the mean grade point average, suppose that our random sample of $n = 15$ students majoring in mathematics yields a test statistic t^* instead equaling -2.5. The P -value for conducting the left-tailed test $H_0: \mu = 3$ versus $H_A: \mu < 3$

is the probability that we would observe a test statistic less than $t^* = -2.5$ if the population mean μ really were 3. The P -value is therefore the area under a $t_{n-1} = t_{14}$ curve and to the *left* of the test statistic $t^* = -2.5$. It can be shown using statistical software that the P -value is 0.0127:

The P -value, 0.0127, tells us it is “unlikely” that we would observe such an extreme test statistic t^* in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the P -value, 0.0127, is less than $\alpha = 0.05$, we reject the null hypothesis $H_0: \mu = 3$ in favor of the alternative hypothesis $H_A: \mu < 3$.

Note that we would not reject $H_0: \mu = 3$ in favor of $H_A: \mu < 3$ if we lowered our willingness to make a Type I error to $\alpha = 0.01$ instead, as the P -value, 0.0127, is then greater than $\alpha = 0.01$.

In our example concerning the mean grade point average, suppose again that our random sample of $n = 15$ students majoring in mathematics yields a test statistic t^* instead equaling -2.5. The P -value for conducting the two-tailed test $H_0: \mu = 3$ versus $H_A: \mu \neq 3$ is the probability that we would observe a test statistic less than -2.5 or greater than 2.5 if the population mean μ really were 3. That is, the two-tailed test requires taking into account the possibility that the test statistic could fall into either tail (and hence the name “two-tailed” test). The P -value is therefore the area under a $t_{n-1} = t_{14}$ curve to the *left* of -2.5 and to the *right* of the 2.5. It can be shown using statistical software that the P -value is $0.0127 + 0.0127$, or 0.0254:



Note that the P -value for a two-tailed test is always two times the P -value for either of the one-tailed tests. The P -value, 0.0254, tells us it is “unlikely” that we would observe such

Test statistic is a number calculated from a statistical test of a hypothesis. It shows how closely your observed data match the distribution expected under the null hypothesis of that statistical test.

an extreme test statistic t^* in the direction of H_A if the null hypothesis were true. Therefore, our initial assumption that the null hypothesis is true must be incorrect. That is, since the P -value, 0.0254, is less than $\alpha = 0.05$, we reject the null hypothesis $H_0: \mu = 3$ in favor of the alternative hypothesis $H_A: \mu \neq 3$.

Note that we would not reject $H_0: \mu = 3$ in favor of $H_A: \mu \neq 3$ if we lowered our willingness to make a Type I error to $\alpha = 0.01$ instead, as the P -value, 0.0254, is then greater than $\alpha = 0.01$.

Now that we have reviewed the critical value and P -value approach procedures for each of three possible hypotheses, let's look at three new examples — one of a right-tailed test, one of a left-tailed test, and one of a two-tailed test.



CORE CONCEPT: TYPES OF HYPOTHESES

Hypotheses in empirical disciplines (e.g. physics) are propositions proposed to predict or explain regular phenomena. Using hypotheses to predict or explain regular phenomena is often called “the hypothetico-

deductive method” in science.

The different types of hypotheses are:

Null Hypothesis

A hypothesis is a speculation or theory based on insufficient evidence that lends itself to further testing and experimentation. With further testing, a hypothesis can usually be proven true or false. Let's look at an example. Little Susie speculates, or hypothesizes, that the flowers she waters with club soda will grow faster than flowers she waters with plain water. She waters each plant daily for a month (experiment) and proves her hypothesis true!

A null hypothesis is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that trying to disprove. In the example, Susie's null hypothesis would be something like this: There is no statistically significant relationship between the type of water I feed the flowers and growth of the flowers. It is challenged by the null hypothesis and usually wants to disprove it, to demonstrate that there is a statistically-significant relationship between the two variables in the hypothesis.

This is the **conventional approach** to making a prediction. It involves a statement that says there is no relationship between two groups that the compares on a certain

variable. The hypothesis also may state that there is no significant difference when different groups are compared with respect to a particular variable. For example, “There is no difference in the academic performance of high school students who participate in extracurricular activities and those who do not participate in such activities” is a null hypothesis. In many cases, the purpose of a null hypothesis is to allow the experimental results to contradict the hypothesis and prove the point that there is a definite relationship.

Conventional approach means a product development approach ensuring reproducibility with the use of specified values and operation ranges of parameters.

Alternative Hypothesis

An alternative hypothesis simply is the inverse, or opposite, of the null hypothesis. So, if we continue with the above example, the alternative hypothesis would be that there IS indeed a statistically-significant relationship between what type of water the flower plant is fed and growth. More specifically, here would be the null and alternative hypotheses for Susie’s study:

- *Null:* If one plant is fed club soda for one month and another plant is fed plain water, there will be no difference in growth between the two plants.
- *Alternative:* If one plant is fed club soda for one month and another plant is fed plain water, the plant that is fed club soda will grow better than the plant that is fed plain water.

Logical Hypothesis

It is that type in which hypothesis is verified logically. J.S. Mill has given four canons of these hypothesis e.g. agreement, disagreement, difference and residue.

Statistical Hypothesis

A hypothesis which can be verified statistically called statistical hypothesis. The statement would be logical or illogical but if statistic verifies it, it will be statistical hypothesis.

Simple Hypothesis

Simple hypothesis is that one in which there exists relationship between two variables one is called independent variable or cause and other is dependent variable or effect.

For example

- Smoking leads to Cancer
- The higher ratio of unemployment leads to crimes.

Complex Hypothesis

Complex hypothesis is that one in which as relationship among variables exists. I recommend you should read characteristics of a good research hypothesis. In this type dependent as well as independent variables are more than two.

For example

- Smoking and other drugs leads to cancer, tension chest infections etc.
- The higher ration of unemployment poverty, illiteracy leads to crimes like dacoit, Robbery, Rape, prostitution and killing etc.

Empirical Hypothesis

Working hypothesis is that one which is applied to a field. During the formulation it is an assumption only but when it is pat to a test become an empirical or working hypothesis.

Type I and II Errors and Significance Levels

The confidence at which a null hypothesis is accepted or rejected is called level of significance. The level of significance is denoted by α .

Type I Error

Rejecting the null hypothesis when it is in fact true is called a Type I error. Many people decide, before doing a hypothesis test, on a maximum p-value for which they will reject the null hypothesis. This value is often denoted α (alpha) and is also called the significance level.

When a hypothesis test results in a p-value that is less than the significance level, the result of the hypothesis test is called statistically significant.

	Null Hypothesis True	Null Hypothesis False
Reject Null Hypothesis	Type I Error	Correct
Fail to Reject Null Hypothesis	Correct	Type II Error

Common mistake

Confusing statistical significance and practical significance. *For Example*, A large clinical trial is carried out to compare a new medical treatment with a standard one. The statistical analysis shows a statistically significant difference in lifespan when using the new treatment compared to the old one. But the increase in lifespan is at most three days, with average increase less than 24 hours, and with poor quality of life during the period of extended life. Most people would not consider the improvement practically significant.

Connection between Type I error and significance level:

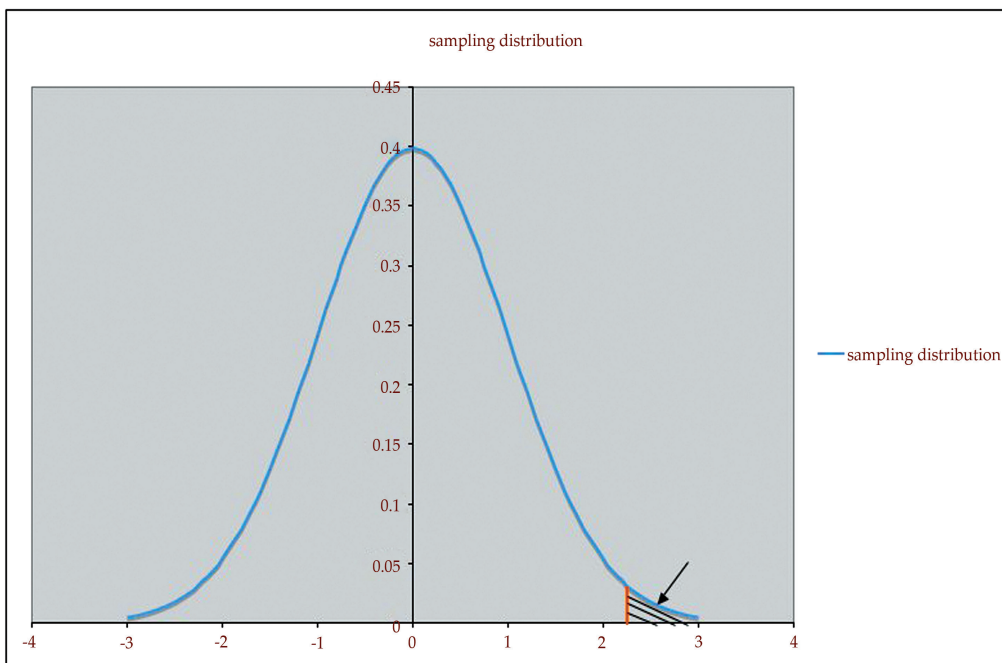
A significance level α corresponds to a certain value of the test statistic, say t_{α} , represented by the orange line in the picture of a sampling distribution below (the picture illustrates a hypothesis test with alternate hypothesis " $\mu > 0$ ")

Since the shaded area indicated by the arrow is the p-value corresponding to t_{α} , that p-value (shaded area) is α .

- To have p-value less than α , a t-value for this test must be to the right of t_{α} .
- So the probability of rejecting the null hypothesis when it is true is the probability that $t > t_{\alpha}$, which we saw above is α .
- In other words, the probability of Type I error is α .

Rephrasing using the definition of Type I error:

The significance level α is the probability of making the wrong decision when the null hypothesis is true.



Pros and Cons of Setting a Significance Level:

- Setting a significance level (*before* doing inference) has the *advantage* that the analyst is not tempted to chose a cut-off on the basis of what he or she hopes is true.
- It has the *disadvantage* that it neglects that some p-values might best be considered borderline. *This is one reason² why it is important to report p-values when reporting results of hypothesis tests. It is also good practice to include confidence intervals corresponding to the hypothesis test.* (For example, if a hypothesis test for the difference of two means is performed, also give a confidence interval for the difference of those means. If the significance level for the hypothesis test is .05, then use confidence level 95% for the confidence interval.)

Type II Error

Not rejecting the null hypothesis when in fact the alternate hypothesis is true is called a Type II error. (The second example below provides a situation where the concept of Type II error is important.)

Note: “The alternate hypothesis” in the definition of Type II error may refer to the alternate hypothesis in a hypothesis test, or it may refer to a “specific” alternate hypothesis.

Example: In a t-test for a sample mean μ , with null hypothesis “ $\mu = 0$ ” and alternate hypothesis “ $\mu > 0$ ”, we may talk about the Type II error relative to the *general alternate hypothesis* “ $\mu > 0$ ”, or may talk about the Type II error relative to the *specific alternate hypothesis* “ $\mu > 1$ ”. Note that the specific alternate hypothesis is a special case of the general alternate hypothesis.

In practice, people often work with Type II error relative to a *specific* alternate hypothesis. In this situation, the probability of Type II error relative to the specific alternate hypothesis is often called β . In other words, β is the probability of making the wrong decision when the *specific alternate* hypothesis is true.

Considering both types of error Together

The following table 1 summarizes Type I and Type II errors:

Table 1.

		Truth (for population studied)	
		Null Hypothesis True	Null Hypothesis False
Decision (based on sample)	Reject Null Hypothesis	Type I Error	Correct Decision
	Fail to reject Null Hypothesis	Correct Decision	Type II Error

An analogy that some people find helpful (but others don't) in understanding the two types of error is to consider a defendant in a trial. The null hypothesis is "defendant is not guilty;" the alternate is "defendant is guilty." A Type I error would correspond to convicting an innocent person; a Type II error would correspond to setting a guilty person free.

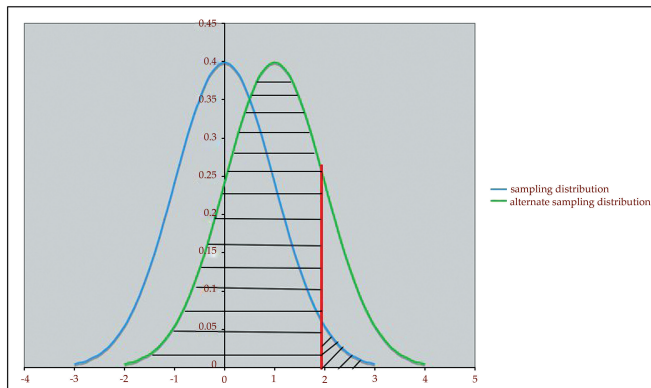
The analogous table 2 would be:

Table 2.

		Truth	
		Not Guilty	Guilty
Verdict	Guilty	Type I Error -- Innocent person goes to jail (and maybe guilty person goes free)	Correct Decision
	Not Guilty	Correct Decision	Type II Error -- Guilty person goes free

The following diagram illustrates the Type I error and the Type II error against the specific alternate hypothesis " $\mu = 1$ " in a hypothesis test for a population mean μ , with null hypothesis " $\mu = 0$," alternate hypothesis " $\mu > 0$," and significance level $\alpha = 0.05$.

- The blue (leftmost) curve is the sampling distribution assuming the null hypothesis " $\mu = 0$."
- The green (rightmost) curve is the sampling distribution assuming the specific alternate hypothesis " $\mu = 1$ ".
- The vertical red line shows the cut-off for rejection of the null hypothesis: the null hypothesis is rejected for values of the test statistic to the *right* of the red line (and *not* rejected for values to the *left* of the red line)>
- The area of the diagonally hatched region to the *right* of the red line and under the *blue* curve is the probability of type I error (α)
- The area of the horizontally hatched region to the *left* of the red line and under the *green* curve is the probability of Type II error (β)



Deciding what significance level to use:

This should be done *before* analyzing the data -- preferably before gathering the data. The choice of significance level should be based on the consequences of Type I and Type II errors.

- If the consequences of a type I error are serious or expensive, then a very small significance level is appropriate.

Example 1: Two drugs are being compared for effectiveness in treating the same condition. Drug 1 is very affordable, but Drug 2 is extremely expensive. The null hypothesis is “both drugs are equally effective,” and the alternate is “Drug 2 is more effective than Drug 1.” In this situation, a Type I error would be deciding that Drug 2 is more effective, when in fact it is no better than Drug 1, but would cost the patient much more money. That would be undesirable from the patient’s perspective, so a small significance level is warranted.

- If the consequences of a Type I error are not very serious (and especially if a Type II error has serious consequences), then a larger significance level is appropriate.

Example 2: Two drugs are known to be equally effective for a certain condition. They are also each equally affordable. However, there is some suspicion that Drug 2 causes a serious side-effect in some patients, whereas Drug 1 has been used for decades with no reports of the side effect. The null hypothesis is “the incidence of the side effect in both drugs is the same”, and the alternate is “the incidence of the side effect in Drug 2 is greater than that in Drug 1.” Falsely rejecting the null hypothesis when it is in fact true (Type I error) would have no great consequences for the consumer, but a Type II error (i.e., failing to reject the null hypothesis when in fact the alternate is true, which would result in deciding that Drug 2 is no more harmful than Drug 1 when it is in fact more harmful) could have serious consequences from a public health standpoint. So setting a large significance level is appropriate.

Common Mistake

Neglecting to think adequately about possible consequences of Type I and Type II errors (and deciding acceptable levels of Type I and II errors based on these consequences) before conducting a study and analyzing data.

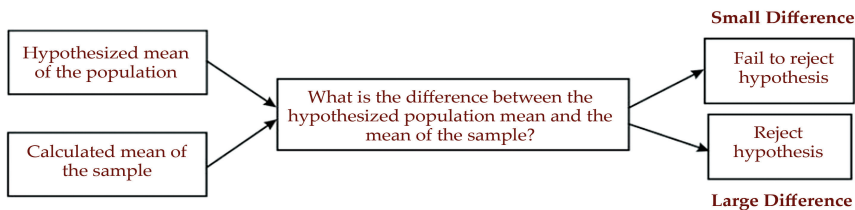
- Sometimes there may be serious consequences of each alternative, so some compromises or weighing priorities may be necessary. The trial analogy illustrates this well: Which is better or worse, imprisoning an innocent person or letting a guilty person go free? *This is a value judgment; value judgments are often involved in deciding on significance levels. Trying to avoid the issue by always choosing the same significance level is itself a value judgment.*
- Sometimes different stakeholders have different interests that compete (e.g., in the second example above, the developers of Drug 2 might prefer to have a smaller significance level.)

Common mistake: *Claiming that an alternate hypothesis has been “proved” because it has been rejected in a hypothesis test.*

- This is an instance of the common mistake of expecting too much certainty.
- There is always a possibility of a Type I error; the sample in the study might have been one of the small percentage of samples giving an unusually extreme test statistic.
- This is why *replicating experiments* (i.e., repeating the experiment with another sample) is important. The more experiments that give the same result, the stronger the evidence.
- There is also the possibility that the sample is biased or the method of analysis was inappropriate; either of these could lead to a misleading result.

Deciding Whether to Reject the Null Hypothesis: One and Two-Tailed Hypothesis Tests

The alternative hypothesis can be supported only by rejecting the null hypothesis. To reject the null hypothesis means to find a large enough difference between your sample mean and the hypothesized (null) mean that it raises real doubt that the true population mean is 20. If the difference between the hypothesized mean and the sample mean is very large, we reject the null hypothesis. If the difference is very small, we do not. In each hypothesis test, we have to decide in advance what the magnitude of that difference must be to allow us to reject the null hypothesis. Below is an overview of this process. Notice that if we fail to find a large enough difference to reject, we fail to reject the null hypothesis. Those are your only two alternatives.



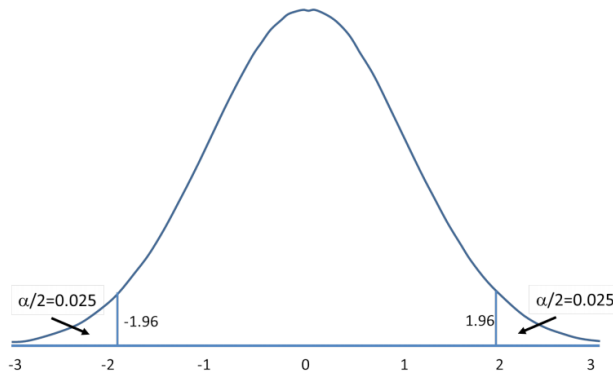
When a hypothesis is tested, a statistician must decide on how much of a difference between means is necessary in order to reject the null hypothesis.

Statisticians first choose a level of significance or alpha (α) level for their hypothesis test. Similar to the significance level you used in constructing confidence intervals, this alpha level tells us how improbable a sample mean must be for it to be deemed “significantly different” from the hypothesized mean. The most frequently used levels of significance are 0.05 and 0.01. An alpha level of 0.05 means that we will consider our sample mean to be significantly different from the hypothesized mean if the chances of observing that sample mean are less than 5%. Similarly, an alpha level of 0.01 means that we will consider our sample mean to be significantly different from the hypothesized mean if the chances of observing that sample mean are less than 1%.

Two-tailed Hypothesis

Tests A hypothesis test can be one-tailed or two-tailed. The examples above are all two-tailed hypothesis tests. We indicate that the average study time is either 20 hours per week, or it is not. Computer use averages 3.2 hours per week, or it does not. We do not specify whether we believe the true mean to be higher or lower than the hypothesized mean. We just believe it must be different. In a two-tailed test, you will reject the null hypothesis if your sample mean falls in either tail of the distribution. For this reason, the alpha level (let's assume .05) is split across the two tails. The curve below shows the critical regions for a two-tailed test. These are the regions under the normal curve that, together, sum to a probability of 0.05. Each tail has a probability of 0.025. The z-scores that designate the start of the critical region are called the critical values

If the sample mean taken from the population falls within these critical regions, or “rejection regions,” we would conclude that there was too much of a difference and we would reject the null hypothesis. However, if the mean from the sample falls in the middle of the distribution (in between the critical regions) we would fail to reject the null hypothesis.



One-Tailed Hypothesis Test

We would use a single-tail hypothesis test when the direction of the results is anticipated or we are only interested in one direction of the results.

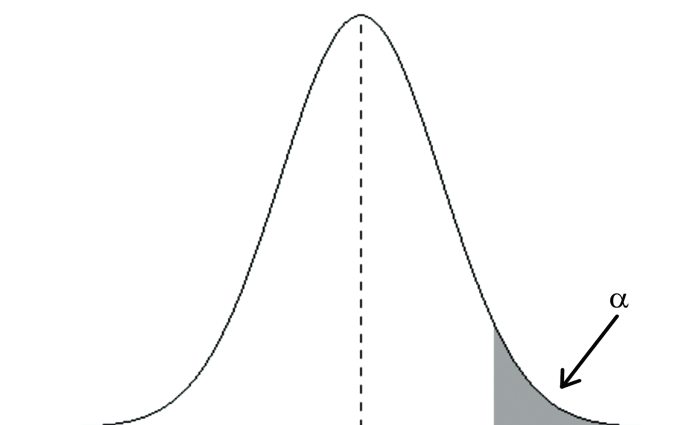
When performing a single-tail hypothesis test, our alternative hypothesis looks a bit different. We use the symbols of greater than or less than. *For example*, let's say we were claiming that the average SAT score of graduating seniors was GREATER than 1,110. Remember, our own personal hypothesis is the alternative hypothesis. Then our null and alternative hypothesis could look something like:

$$H_0 : \mu \leq 1100$$

$$H_a : \mu > 1100$$

In this scenario, our null hypothesis states that the mean SAT scores would be less than or equal to 1,100 while the alternate hypothesis states that the SAT scores

would be greater than 1,100. A single-tail hypothesis test also means that we have only one critical region because we put the entire critical region into just one side of the distribution. When the alternative hypothesis is that the sample mean is greater, the critical region is on the right side of the distribution (see below). When the alternative hypothesis is that the sample is smaller, the critical region is on the left side of the distribution.



Visual Representation of Critical Region:
One-tailed

To calculate the critical regions, we must first find the critical values or the cut-offs where the critical regions start

Identifies the Appropriate form of the Test-statistic when

Identify the appropriate test **statistic** and interpret the results for a hypothesis test concerning, the variance of a normally distributed population, and the equality of the variances of two normally distributed populations based on two independent random samples.

Statistic is a single measure of some attribute of a sample (e.g., its arithmetic mean value).

When Population Variance is known

Let's start by acknowledging that it is completely unrealistic to think that we'd find ourselves in the situation of knowing the population variance, but not the population mean. Therefore, the hypothesis testing method that we learn on this page has limited practical use. We study it only because you will use it later to learn about the "power" of a hypothesis test (by learning how to calculate Type II error rates). As usual, let's start with an example.

Population is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired.

When Population Variance is Unknown

Now that, for purely pedagogical reasons, we have the unrealistic situation (of a known population variance) behind us, let's turn our attention to the realistic situation in which both the population mean and **population** variance are unknown.

The Central Limit Theorem is to be used

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement text annotation indicator, then the distribution of the sample means will be approximately normally distributed. This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually $n > 30$). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that $\min(np, n(1-p)) > 5$, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

For the random samples we take from the population, we can compute the mean of the sample means:

$$\mu_{\bar{x}} = \mu$$

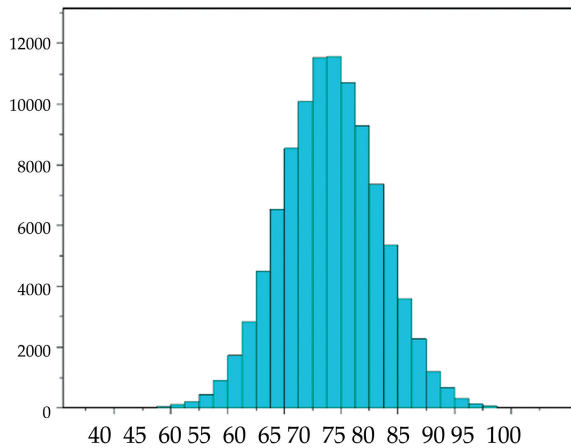
and the standard deviation of the sample means:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

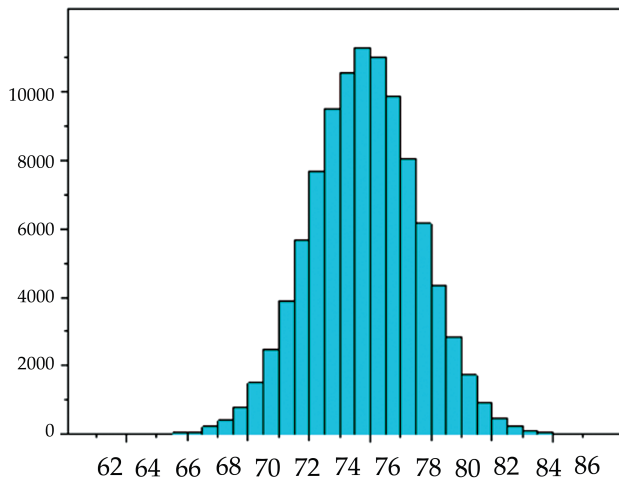
Before illustrating the use of the Central Limit Theorem (CLT) we will first illustrate the result. In order for the result of the CLT to hold, the sample must be sufficiently large ($n \geq 30$). Again, there are two exceptions to this. If the population is normal, then the result holds for samples of any size (i.e. the sampling distribution of the sample means will be approximately normal even for samples of size less than 30).

Central Limit Theorem with a Normal Population

The figure below illustrates a normally distributed characteristic, X , in a population in which the population mean is 75 with a standard deviation of 8.



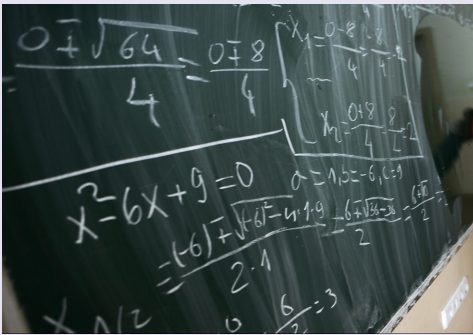
If we take simple random samples (with replacement) of size $n=10$ from the population and compute the mean for each of the samples, the distribution of sample means should be approximately normal according to the Central Limit Theorem. Note that the sample size ($n=10$) is less than 30, but the source population is normally distributed, so this is not a problem. The distribution of the sample means is illustrated below. Note that the horizontal axis is different from the previous illustration, and that the range is narrower.



The mean of the sample means is 75 and the standard deviation of the sample means is 2.5, with the standard deviation of the sample means computed as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{10}} = 2.5$$

If we were to take samples of $n=5$ instead of $n=10$, we would get a similar distribution, but the variation among the sample means would be larger. In fact, when we did this we got a sample mean = 75 and a sample standard deviation = 3.6.



CORE CONCEPT: HYPOTHESIS TESTING AND SAMPLING DISTRIBUTIONS

The underlying distribution of a statistic is called its sampling distribution. The sampling distribution of a statistic gives all the values that the statistic can take, and the probability of each value occurring by chance alone. The sampling distribution assumes that the null hypothesis is true.

The logic of hypothesis testing is rooted in an understanding of the sampling distribution of the mean

- The sample mean is an unbiased estimator of the population mean. On average, a randomly selected sample will have a mean equal to that in the population. In hypothesis testing, we begin by stating the null hypothesis. We expect that, if the null hypothesis is true, then a random sample selected from a given population will have a sample mean equal to the value stated in the null hypothesis.
- Regardless of the distribution in the population, the sampling distribution of the sample mean is normally distributed. Hence, the probabilities of all other possible sample means we could select are normally distributed. Using this distribution, we can therefore state an alternative hypothesis to locate the probability of obtaining sample means with less than a 5% chance of being selected if the value stated in the null hypothesis is true.

Table 3: A review of the notation used for the mean, variance, and standard deviation in population, sample, and sampling distributions

Characteristic	Population	Sample	Sampling Distribution
Mean	μ	M or $\bar{X}$	$\mu_M = \mu$
Variance	σ^2	s^2 or SD^2	$\sigma_M^2 = \frac{\sigma^2}{n}$
Standard deviation	σ	s or SD	$\sigma_M = \frac{\sigma}{\sqrt{n}}$

Table 4: A review of the key differences between population, sample, and sampling distributions

	Population Distribution	Sample Distribution	Distribution of Sample Means
What is it?	Scores of all persons in a population	Scores of a select portion of persons from the population	All possible sample means that can be drawn, given a certain sample size
Is it accessible?	Typically, no	Yes	Yes
What is the shape?	Could be any shape	Could be any shape	Normally distributed

Making a Decision: Types of Error

We decide whether to retain or reject the null hypothesis. Because we are observing a sample and not an entire population, it is possible that a conclusion may be wrong. Table 5 shows that there are four decision alternatives regarding the truth and falsity of the decision we make about a null hypothesis:

- The decision to retain the null hypothesis could be correct.
- The decision to retain the null hypothesis could be incorrect.
- The decision to reject the null hypothesis could be correct.
- The decision to reject the null hypothesis could be incorrect.

Table 5: Four outcomes for making a decision. The decision can be either correct (correctly reject or retain null) or wrong (incorrectly reject or retain null)

		Decision	
		Retain the null	Reject the null
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

We investigate each decision alternative in this section. Since we will observe a sample, and not a population, it is impossible to know for sure the truth in the population. So for the sake of illustration, we will assume we know this. This assumption is labeled as truth in the population in Table 5.

Decision: Retain the Null Hypothesis

When we decide to retain the null hypothesis, we can be correct or incorrect. The correct decision is to retain a true null hypothesis. This decision is called a null result or null finding. This is usually an uninteresting decision because the decision is to retain what we already assumed: that the value stated in the null hypothesis is correct. For this reason, null results alone are rarely published in behavioral research. The incorrect decision is to retain a false null hypothesis. This decision is an example of

a Type II error, or b error. With each test we make, there is always some probability that the decision could be a Type II error. In this decision, we decide to retain previous notions of truth that are in fact false. While it's an error, we still did nothing; we retained the null hypothesis. We can always go back and conduct more studies.

Examples Using The Z Test

The test statistic we use depends largely on what we know about the population. When we know the mean and standard deviation in a single population, we can use the one-independent sample z test, which we will use in this section to illustrate the four steps of hypothesis testing.

The one-independent sample z test is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance.

Directional, Lower-Tail Critical Hypothesis Tests

We will use the z test for a directional, or one-tailed test, where the alternative hypothesis is stated as less than ($<$) the null hypothesis. For a lower-tail critical test, or a less than statement, we place the level of significance or critical value in the lower tail of the sampling distribution. So we are interested in any alternative less than the value stated in the null hypothesis. This test is appropriate when it is not possible or highly unlikely that a sample mean will fall above the population mean stated in the null hypothesis.

Step 1: State the hypotheses. The population mean is 558, and we are testing whether the alternative is less than ($<$) this value:

$$H_0: \mu = 558$$

Mean test scores are equal to 558 in the population at this school.

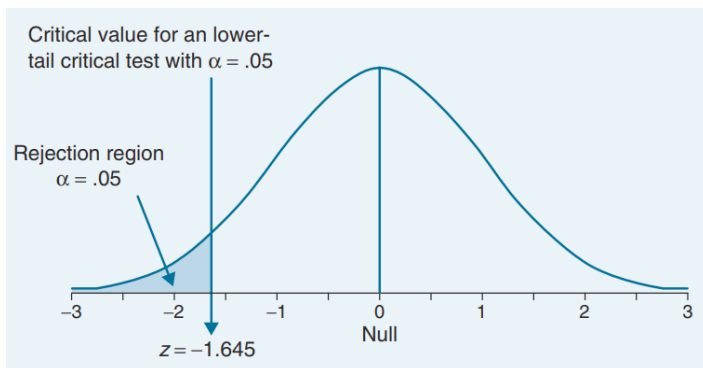
$$H_1: \mu < 558$$

Mean test scores are less than 558 in the population at this school.

Set the criteria for a decision. The level of significance is .05, which makes the alpha level $\alpha = .05$. To determine the critical value for a lower-tail critical test, we locate the probability .0500 toward the tail in column C in the unit normal table. The z-score associated with this probability is again $z = 1.645$. Since this test is a lower-tail critical test, we place the critical value the same distance below the mean: The critical value for this test is $z = -1.645$. All of the alpha level is placed in the lower tail of the distribution beyond the critical value. Figure shows the standard normal distribution, with the rejection region beyond the critical value.

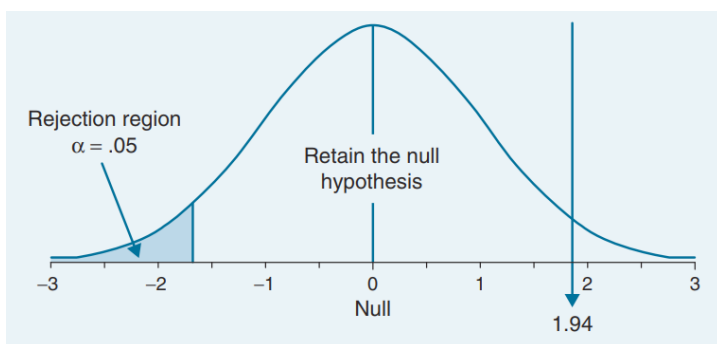
The probability is less than 5% that we will obtain a sample mean that is at least 1.645 standard deviations below the value of the population mean stated in the null hypothesis.

We are testing the same population, and we measured the same value of the sample mean. We changed



Make a decision. To make a decision, we compare the obtained value to the critical value. We reject the null hypothesis if the obtained value exceeds the critical value.

The test statistic does not reach the rejection region; retain the null hypothesis. This is actually a Type III error—this result would have been significant if the rejection region were placed in the upper tail.



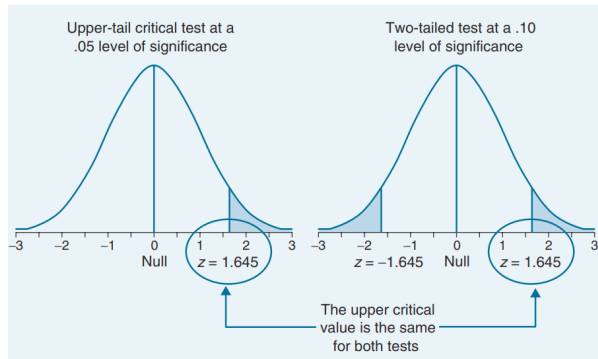
We would have decided to reject the null hypothesis. We anticipated that scores would be worse, and instead, they were better than the value stated in the null hypothesis.

A Type III error occurs with one-tailed tests, where decides to retain the null hypothesis because the rejection region was located in the wrong tail. The “wrong tail” refers to the opposite tail from where a difference was observed and would have otherwise been significant.

Directional versus No directional Tests

These results to a group of participants, they found that participants were more persuaded by a significant result when it was described as a one-tailed test, $p < .05$, than when it was described as a two-tailed test, $p < .10$. This was interesting because the two results were identical—both tests were associated with the same critical value in the upper tail.

Most editors of peer-reviewed journals in behavioral research will not publish the results of a study where the level of significance is greater than .05.



When $\alpha = .05$, all of that value is placed in the upper tail for an upper-tail critical test. The two tailed equivalent would require a test with $\alpha = .10$, such that .05 is placed in each tail.

The two-tailed test is more conservative; it makes it more difficult to reject the null hypothesis. It also eliminates the possibility of committing a Type III error. The one-tailed test, though, is associated with greater power. If the value stated in the null hypothesis is false, then a one-tailed test will make it easier to detect this (i.e., lead to a decision to reject the null hypothesis). Because the one-tailed test makes it easier to reject the null hypothesis, it is important that we justify that an outcome can occur in only one direction.

Effect size is most meaningfully reported with significant effects when the decision was to reject the null hypothesis. If an effect is not significant, as in instances when we retain the null hypothesis, then we are concluding that an effect does not exist in a population. It makes little sense to compute the size of an effect that we just concluded doesn't exist. In this section, we describe how far scores shifted in the population using a measure of effect size called Cohen's d .

Cohen's d measures the number of standard deviations an effect shifted above or below the population mean stated by the null hypothesis. The formula for Cohen's d replaces the standard error in the denominator of the test statistic with the population standard deviation.

$$\text{Cohen's } d = \frac{M - \mu}{\sigma}.$$

The sign of d indicates the direction of the shift. When values of d are positive, an effect shifted above the population mean; when values of d are negative, an effect shifted below the population mean.

Table 6: Cohen's effect size conventions

Description of Effect	Effect Size (<i>d</i>)
Small	$d < 0.2$
Medium	$0.2 < d < 0.8$
Large	$d > 0.8$

Cohen's *d* is a measure of effect size in terms of the number of standard deviations that mean scores shifted above or below the population mean stated by the null hypothesis. The larger the value of *d*, the larger the effect in the population.

Cohen's effect size conventions are standard rules for identifying small, medium, and large effects based on typical findings in behavioral research.

The Relationship between Sample Size and Power

To overcome low effect size, we can increase the sample size. Increasing sample size decreases standard error, thereby increasing power. To illustrate, let's compute the test statistic for the one-tailed significance test for Class 1, which had a small effect size.

$$z_{obt} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40 - 38}{\frac{10}{\sqrt{30}}} = 1.10.$$

For a one-tailed test that is upper-tail critical, the critical value is 1.645. The value of the test statistic (+1.10) does not exceed the critical value (+1.645), so we retain the null hypothesis.

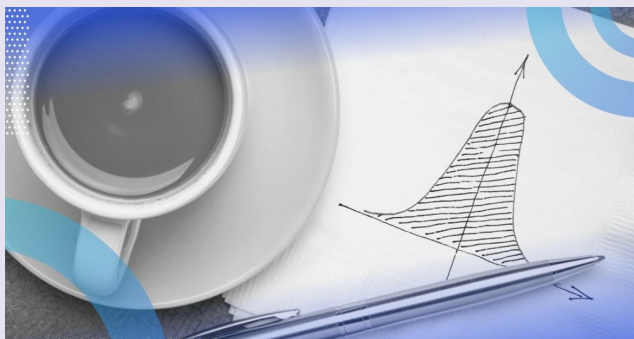
Increase the sample size to $n = 100$.

$$z_{obt} = \frac{M - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{40 - 38}{\frac{10}{\sqrt{100}}} = 2.00.$$

The critical value is still 1.645. The value of the test statistic (+2.00) now exceeds the critical value (+1.645), so we reject the null hypothesis.

A CONCEPT OF Z-TEST

Z-test is a statistical tool that is used in hypothesis testing. It is the go-to method when the sample size is large. The test derives the difference between the two large population samples, provided the variance is known. Z-tests are similar to t-tests; the only difference is that t-tests are conducted for small sample sizes or when the variance is unknown.



Again, t-tests are not conducted for large datasets; on the other hand, z-tests do not work if the sample size is too small. There is a minimum limit of 30; if the sample size is above 30. Thus, experiments that feature less than 30 subjects are referred to as a small sample size.

Before venturing further into the test, let us quickly examine hypothesis testing. Hypothesis testing ascertains whether a particular assumption is true for the whole population. It is a statistical application. It determines the validity of inference by evaluating sample data from the overall population.

The concept of hypothesis works on the probability of an event's occurrence. It confirms whether the primary hypothesis results are correct or not. In research, it is very important to eliminate randomness. The data should not have been caused by chance or a random factor. Hypothesis testing eliminates such uncertainties.

The z-test definition stresses an important assumption—the sample data is a normal distribution. That is, a given sample is normally distributed; there is no influence of an external factor.

Z-trials are classified into two:

- One sample test compares a single sample average with the population means.
- Two-sample tests compare the average mean of two samples.

SUMMARY

- Hypothesis testing is a statistical method used to evaluate a hypothesis or claim about a population parameter based on sample data.
- Hypothesis testing is the process of using a variety of statistical tools to analyze data and, ultimately, to fail to reject or reject the null hypothesis.
- Hypothesis-testing often pertains to the scope of a claim. Unrestricted claims are maximally general, such as those proposing translation universals.
- A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test.
- A null hypothesis is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. It is the hypothesis that trying to disprove.
- A hypothesis which can be verified statistically called statistical hypothesis. The statement would be logical or illogical but if statistic verifies it, it will be statistical hypothesis.
- The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement text annotation indicator, then the distribution of the sample means will be approximately normally distributed.

MULTIPLE CHOICE QUESTIONS

1. **What is the primary purpose of a test of hypothesis in statistics?**
 - a. To prove a hypothesis
 - b. To verify the population parameter
 - c. To make informed decisions based on sample data
 - d. To generate random data
2. **In a hypothesis test, the null hypothesis (H_0) usually represents:**
 - a. The researcher's desired outcome
 - b. The alternative hypothesis
 - c. No effect or difference in the population
 - d. A predetermined conclusion
3. **The p-value in a hypothesis test represents:**
 - a. The probability that the null hypothesis is true
 - b. The probability that the alternative hypothesis is true
 - c. The probability of obtaining observed results as extreme as or more extreme than the sample data, assuming the null hypothesis is true
 - d. The probability of obtaining sample data that perfectly match the null hypothesis
4. **A significance level (α) of 0.05 means:**
 - a. There's a 5% chance of making a Type I error
 - b. The null hypothesis is always true
 - c. The alternative hypothesis is true
 - d. The sample size is 5
5. **What does a p-value less than the significance level (α) indicate in hypothesis testing?**
 - a. Strong evidence to reject the null hypothesis
 - b. Strong evidence to accept the null hypothesis
 - c. Weak evidence to accept the null hypothesis
 - d. No evidence to accept or reject the null hypothesis

Answers

1. (c) 2. (c) 3. (c) 4. (a) 5. (a)

REVIEW QUESTIONS

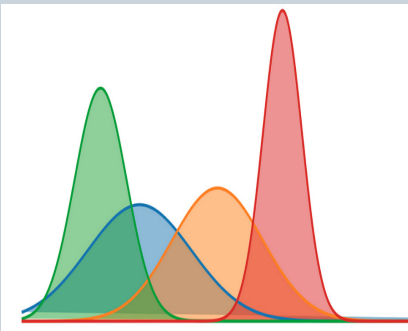
1. Discuss about the nature of hypothesis.
2. Explain hypothesis testing process.
3. What do you understand by population proportion?
4. Focus on P-value approach.
5. Explain the null and alternative hypothesis.
6. Describe the type I and II errors and significance levels.
7. Discuss about relationship between sample size and power

REFERENCES

1. Agresti, A., & Franklin, C. A. (2018). *Statistics: The Art and Science of Learning from Data*. Pearson.
2. Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Brooks/Cole.
3. DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.
4. Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. W. W. Norton & Company.
5. Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical Statistics*. Pearson.
6. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson.
7. Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer.
8. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.
9. Rosner, B. (2010). *Fundamentals of Biostatistics*. Brooks/Cole.
10. Triola, M. F. (2019). *Elementary Statistics*. Pearson.
11. Zar, J. H. (2010). *Biostatistical Analysis*. Pearson.

CHAPTER 6

PARAMETER ESTIMATION



Parameter Estimation is a fundamental concept in statistics that involves estimating unknown population parameters using sample data. In real-world situations, we often have access to only a subset of the entire population, and we use statistical methods to infer information about the population as a whole. Parameter estimation helps us make informed decisions and draw conclusions based on the available data.

After studying this chapter, you will understand the core concepts of:

- Samples and Statistics
- Quality Criteria for Estimates
- Methods of Estimation

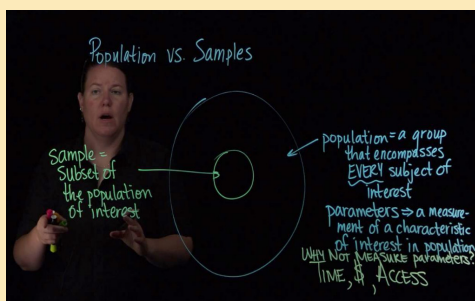
INTRODUCTION

Statistics are used as estimators of population quantities with an estimate denoting a given realisation of an estimator. We explore key properties that we wish estimators to have such as unbiasedness, efficiency and consistency. We study the properties of the sample mean and sample variance as estimators of the population mean and variance, respectively.

Parameters are descriptive measures of an entire population. However, their values are usually unknown because it is infeasible to measure an entire population. Because of this, you can take a random sample from the population to obtain parameter estimates. One goal of statistical analyses is to obtain estimates of the population parameters along with the amount of error associated with these estimates. These estimates are also known as sample statistics.

There are several types of parameter estimates:

- Point estimates are the single, most likely value of a parameter. For example, the point estimate of population mean (the parameter) is the sample mean (the parameter estimate).
- Confidence intervals are a range of values likely to contain the population parameter.



CORE CONCEPT: SAMPLES AND STATISTICS

In statistics and quantitative research methodology, a data sample is a set of data collected and/or selected from a statistical population by a defined procedure. The

elements of a sample are known as sample points, sampling units or observations. The sample usually represents a subset of manageable size.

Given an independent data set $x_1, x_2, \dots, x_n$, let

$$\hat{\theta} = h(x_1, x_2, \dots, x_n) \quad (1)$$

be an estimate of parameter θ . In order to ascertain its general properties, it is recognized that, if the experiment that yielded the data set were to be repeated, we would obtain different values for $x_1, x_2, \dots, x_n$. The function $h(x_1, x_2, \dots, x_n)$ when applied to the new data set would yield a different value for $\hat{\theta}$. We thus see that

estimate is $\hat{\theta}$ itself a random variable possessing a probability distribution, which depends both on the functional form defined by h and on the distribution of the underlying random variable X . The appropriate representation of $\hat{\theta}$ is thus

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n), \quad (2)$$

where $X_1, X_2, \dots, X_n$ are random variables, representing a sample from random variable X , which is referred to in this context as the population. In practically all applications, we shall assume that sample $X_1, X_2, \dots, X_n$ possesses the following properties:

- Property 1: $X_1, X_2, \dots, X_n$ are independent.
- Property 2: $f_{X_j}(x) = f_X(x)$ for all $x, j = 1, 2, \dots, n$.

The random variables $X_1, \dots, X_n$ satisfying these conditions are called a random sample of size n . The word ‘random’ in this definition is usually omitted for the sake of brevity. If X is a random variable of the discrete type with **probability** mass function (pmf) $p_X(x)$, then P_{X_j} for each j .

A specific set of observed values $(x_1, x_2, \dots, x_n)$ is a set of sample values assumed by the sample. The problem of parameter estimation is one class in the broader topic of **statistical inference** in which our object is to make inferences about various aspects of the underlying population distribution on the basis of observed sample values. For the purpose of clarification, the interrelationships among X , $(X_1, X_2, \dots, X_n)$, and $(x_1, x_2, \dots, x_n)$ are schematically shown in Figure 1.

Let us note that the properties of a sample as given above imply that certain conditions are imposed on the manner in which observed data are obtained. Each datum point must be observed from the population independently and under identical conditions. In sampling a population of percentage yield, for example, one would avoid taking adjacent batches if correlation between them is to be expected.

A statistic is any function of a given sample $X_1, X_2, \dots, X_n$ that does not depend on the unknown parameter. The function $h(X_1, X_2, \dots, X_n)$ in Equation (2) is thus a statistic for which the value can be determined once the sample values have been observed. It is important to note that a statistic, being a function of random variables, is a random variable. When used to estimate a distribution parameter, its statistical properties, such as mean, variance, and distribution, give information concerning the quality of this particular estimation procedure.

Probability is the measure of the likeliness that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data.

Certain statistics play an important role in statistical estimation theory; these include sample mean, sample variance, order statistics, and other sample moments. Some properties of these important statistics are discussed below.

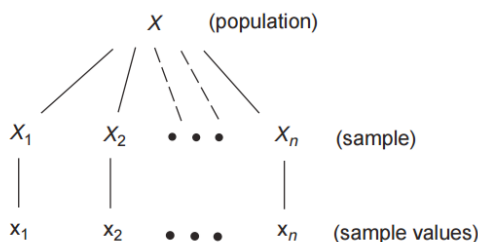


Figure 1: Population, sample, and sample values.

Sample Mean

The statistic

$$\boxed{\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i} \quad (3)$$

is called the sample mean of population X . Let the population mean and variance be, respectively,

$$\left. \begin{aligned} E\{X\} &= m, \\ \text{var}\{X\} &= \sigma^2. \end{aligned} \right\} \quad (4)$$

The mean and variance of $\bar{X}$, the sample mean, are easily found to be

$$E\{\bar{X}\} = \frac{1}{n} \sum_{i=1}^n E\{X_i\} = \frac{1}{n} (nm) = m, \quad (5)$$

and, owing to independence,

$$\begin{aligned} \text{var}\{\bar{X}\} &= E\{(\bar{X} - m)^2\} = E\left\{\left[\frac{1}{n} \sum_{i=1}^n (X_i - m)\right]^2\right\} \\ &= \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}, \end{aligned} \quad (6)$$

which is inversely proportional to sample size n . As n increases, the variance of $\bar{X}$ decreases and the distribution of $\bar{X}$ becomes sharply peaked at $E\{\bar{X}\} = m$. Hence, it is intuitively clear that statistic $\bar{X}$ provides a good procedure for estimating population mean m . This is another statement of the law of large numbers that was discussed in Examples.

Since $\bar{X}$ is a sum of independent **random variables**, its distribution can also be determined either by the use of techniques developed in this section or by means of the method of characteristic functions. We further observe that, on the basis of the central limit theorem, sample mean $\bar{X}$ approaches a normal distribution as $n \rightarrow \infty$. More precisely, random variable

$$(\bar{X} - m) \left(\frac{\sigma}{n^{1/2}} \right)^{-1}$$

approaches $N(0, 1)$ as $n \rightarrow \infty$.

Sample Variance

The statistic

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (7)$$

is called the sample variance of population X . The mean of S^2 can be found by expanding the squares in the sum and taking termwise expectations. We first write Equation (7) as

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n [(X_i - m) - (\bar{X} - m)]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left[(X_i - m) - \frac{1}{n} \sum_{j=1}^n (X_j - m) \right]^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - m)^2 - \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^n (X_i - m)(X_j - m). \end{aligned}$$

Taking termwise expectations and noting mutual independence, we have

$$E\{S^2\} = \sigma^2, \quad (8)$$

where m and σ^2 are defined in Equations (4). We remark at this point that the reason for using $1/(n-1)$ rather than $1/n$ in Equation (7) is to make the mean of S^2 equal to σ^2 . As we shall see in the next section, this is a desirable property for S^2 if it is to be used to estimate σ^2 , the true variance of X .

The variance of S^2 is found from

$$\text{var}\{S^2\} = E\{(S^2 - \sigma^2)^2\}. \quad (9)$$

Upon expanding the right-hand side and carrying out expectations term by term, we find that

$$\text{var}\{S^2\} = \frac{1}{n} \left[\mu_4 - \frac{n-3}{n-1} \sigma^4 \right], \quad (10)$$

where μ_4 is the fourth central moment of X ; that is,

$$\mu_4 = E\{(X - m)^4\}. \quad (11)$$

Equation (6) shows again that the variance of S^2 is an inverse function of n .

In principle, the distribution of S^2 can be derived with use of

Random variable

is usually written X , is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.

advanced techniques. It is, however, a tedious process because of the complex nature of the expression for S^2 as defined by Equation (7). For the case in which population X is distributed according to $N(m, \sigma^2)$, we have the following result (Theorem 1).

Theorem 1: Let S^2 be the sample variance of size n from normal population $N(m, \sigma^2)$, then $(n-1)S^2/\sigma^2$ has a chi-squared (χ^2) distribution with $(n-1)$ degrees of freedom.

Proof of Theorem 1: the chi-squared distribution is given, in order to sketch a proof for this theorem, let us note from this section that random variable Y ,

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - m)^2, \quad (12)$$

has a chi-squared distribution of n degrees of freedom since each term in the sum is a squared normal random variable and is independent of other random variables in the sum. Now, we can show that the difference between Y and $(n-1)S^2/\sigma^2$ is

$$Y - \frac{(n-1)S^2}{\sigma^2} = \left[(\bar{X} - m) \left(\frac{\sigma}{n^{1/2}} \right)^{-1} \right]^2. \quad (13)$$

Since the right-hand side of Equation (6) is a random variable having a chisquared distribution with one degree of freedom, Equation (13) leads to the result that $(n-1)S^2/\sigma^2$ is chi-squared distributed with $(n-1)$ degrees of freedom provided that independence exists between $(n-1)S^2/\sigma^2$ and

$$\left[(\bar{X} - m) \left(\frac{\sigma}{n^{1/2}} \right)^{-1} \right]^2$$

The proof of this independence is not given here but can be found in more advanced texts.

Sample Moments

The k th sample moment is

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (14)$$

Following similar procedures as given above, we can show that

$$\left. \begin{aligned} E\{M_k\} &= \alpha_k, \\ \text{var}\{M_k\} &= \frac{1}{n}(\alpha_{2k} - \alpha_k^2), \end{aligned} \right\} \quad (15)$$

where α_k is the k th moment of population X .

Parameter estimation plays a critical role in accurately describing system behavior through mathematical models such as statistical probability distribution functions, parametric dynamic models, and data-based Simulink® models.

Order Statistics

A sample $X_1, X_2, \dots, X_n$ can be ranked in order of increasing **numerical magnitude**. Let $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ be such a rearranged sample, where $X_{(1)}$ is the smallest and $X_{(n)}$ the largest. Then $X_{(k)}$ is called the k th-order statistic. Extreme values $X_{(1)}$ and $X_{(n)}$ are of particular importance in applications, and their properties have been discussed this Section.

In terms of the probability distribution function (PDF) of population X , $F_X(x)$, it follows from Equations that the PDFs of $X_{(1)}$ and $X_{(n)}$ are

$$F_{X_{(1)}}(x) = 1 - [1 - F_X(x)]^n, \quad (16)$$

$$F_{X_{(n)}}(x) = F_X^n(x). \quad (17)$$

If X is continuous, the pdfs of $X_{(1)}$ and $X_{(n)}$ are of the form

$$f_{X_{(1)}}(x) = n[1 - F_X(x)]^{n-1}f_X(x), \quad (18)$$

$$f_{X_{(n)}}(x) = nF_X^{n-1}(x)f_X(x). \quad (19)$$

The means and variances of order statistics can be obtained through integration, but they are not expressible as simple functions of the moments of population X .

Numerical magnitude imply that one can relatively estimate and comprehend an amount if the number is large.



CORE CONCEPT: QUALITY CRITERIA FOR ESTIMATES

An estimator is the rule or the method of estimation and an estimate is the value the estimator yields in a particular case. There may be several estimators but an appropriate estimator must have the following

properties.

- Unbiasedness
- Consistency
- Minimum Variability
- Efficiency
- Sufficiency

We are now in a position to propose a number of criteria under which the quality of an estimate can be evaluated. These criteria define generally desirable properties for an estimate to have as well as provide a guide by which the quality of one estimate can

be compared with that of another.

Before proceeding, a remark is in order regarding the notation to be used. As seen in Equation (2), our objective in parameter estimation is to determine a statistic

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n), \quad (20)$$

which gives a good estimate of parameter θ . This statistic will be called an estimator for θ , for which properties, such as mean, variance, or distribution, provide a measure of quality of this estimator. Once we have observed sample values $x_1, x_2, \dots, x_n$, the observed estimator,

$$\hat{\theta} = h(x_1, x_2, \dots, x_n), \quad (21)$$

has a numerical value and will be called an estimate of parameter θ .

Unbiasedness

An estimator $\hat{\Theta}$ is said to be an unbiased estimator for θ if

$$E\{\hat{\Theta}\} = \theta, \quad (22)$$

for all θ . This is clearly a desirable property for $\hat{\Theta}$, which states that, on average, we expect $\hat{\Theta}$ to be close to true parameter value θ . Let us note here that the requirement of unbiasedness may lead to other undesirable consequences. Hence, the overall quality of an estimator does not rest on any single criterion but on a set of criteria.

We have studied two statistics, $\bar{X}$ and S^2 . It is seen from Equations (5) and (8) that, if $\bar{X}$ and S^2 are used as estimators for the population mean m and population variance σ^2 , respectively, they are unbiased estimators. This nice property for S^2 suggests that the sample variance defined by Equation (7) is preferred over the more natural choice obtained by replacing $1/(n-1)$ by $1/n$ in Equation (7). Indeed, if we let

$$S^{2*} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2, \quad (23)$$

its mean is

$$E\{S^{2*}\} = \frac{n-1}{n} \sigma^2,$$

and estimator S^{2*} has a bias indicated by the coefficient $(n-1)/n$.

Minimum Variance

It seems natural that, if $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$ is to qualify as a good estimator for θ , not only its mean should be close to true value θ but also there should be a good probability that any of its observed values $\hat{\theta}$ will be close to θ . This can be achieved

by selecting a statistic in such a way that not only is $\hat{\Theta}$ unbiased but also its **variance** is as small as possible. Hence, the second desirable property is one of minimum variance.

Definition 1: let $\hat{\Theta}$ be an unbiased estimator for θ . It is an unbiased minimum-variance estimator for θ if, for all other unbiased estimators Θ^* of θ from the same sample,

$$\boxed{\text{var}\{\hat{\Theta}\} \leq \text{var}\{\Theta^*\}}, \quad (24)$$

for all θ .

Given two unbiased estimators for a given parameter, the one with smaller variance is preferred because smaller variance implies that observed values of the estimator tend to be closer to its mean, the true parameter value.

Example 1: Problem: we have seen that $\bar{X}$ obtained from a sample of size n is an unbiased estimator for population mean m . Does the quality of $\bar{X}$ improve as n increases?

Answer: we easily see from Equation (5) that the mean of $\bar{X}$ is independent of the sample size; it thus remains unbiased as n increases. Its variance, on the other hand, as given by Equation (6) is

$$\text{var}\{\bar{X}\} = \frac{\sigma^2}{n}, \quad (25)$$

which decreases as n increases. Thus, based on the minimum variance criterion, the quality of $\bar{X}$ as an estimator for m improves as n increases.

Example 2: Part 1. Problem: based on a fixed sample size n , is $\bar{X}$ the best estimator for m in terms of unbiasedness and minimum variance?

Approach: in order to answer this question, it is necessary to show that the variance of $\bar{X}$ as given by Equation (6) is the smallest among all unbiased estimators that can be constructed from the sample. This is certainly difficult to do. However, a powerful theorem shows that it is possible to determine the minimum achievable variance of any unbiased estimator obtained from a given sample. This lower bound on the variance thus permits us to answer questions such as the one just posed.

Theorem 2: the Cramér–Rao inequality. Let $X_1, X_2, \dots$

Variance in probability theory and statistics, variance measures how far a set of numbers is spread out. A variance of zero indicates that all the values are identical.

, X_n denote a sample of size n from a population X with pdf $f(x; \theta)$, where θ is the unknown parameter, and let $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$ be an unbiased estimator for θ . Then, the variance of $\hat{\Theta}$ satisfies the inequality

$$\text{var}\{\hat{\Theta}\} \geq \left\{ n E \left\{ \left[\frac{\partial \ln f(X; \theta)}{\partial \theta} \right]^2 \right\} \right\}^{-1}, \quad (26)$$

if the indicated expectation and differentiation exist. An analogous result with $p(X; \theta)$ replacing $f(X; \theta)$ is obtained when X is discrete

Proof of Theorem 2: the joint probability density function (jpdf) of $X_1, X_2, \dots, X_n$ is, because of their mutual independence, $f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)$.

The mean of statistic $\hat{\Theta}$ is $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$,
 $E\{\hat{\Theta}\} = E\{h(X_1, X_2, \dots, X_n)\}$,

and, since $\hat{\Theta}$ is unbiased, it gives

$$\theta = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n. \quad (27)$$

Another relation we need is the identity:

$$1 = \int_{-\infty}^{\infty} f(x_i; \theta) dx_i, \quad i = 1, 2, \dots, n. \quad (28)$$

Upon differentiating both sides of each of Equations (27) and (28) with respect to θ , we have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) \left[\sum_{j=1}^n \frac{1}{f(x_j; \theta)} \frac{\partial f(x_j; \theta)}{\partial \theta} \right] f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) \left[\sum_{j=1}^n \frac{\partial \ln f(x_j; \theta)}{\partial \theta} \right] f(x_1; \theta) \cdots f(x_n; \theta) dx_1 \cdots dx_n, \end{aligned} \quad (29)$$

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \frac{\partial f(x_i; \theta)}{\partial \theta} dx_i \\ &= \int_{-\infty}^{\infty} \frac{\partial \ln f(x_i; \theta)}{\partial \theta} f(x_i; \theta) dx_i, \quad i = 1, 2, \dots, n. \end{aligned} \quad (30)$$

Let us define a new random variable Y by

$$Y = \sum_{j=1}^n \frac{\partial \ln f(X_j; \theta)}{\partial \theta}. \quad (31)$$

Equation (30) shows that

$$E\{Y\} = 0.$$

Moreover, since Y is a sum of n independent random variables, each with mean zero and variance $E\{[\partial \ln f(X; \theta) / \partial \theta]^2\}$, the variance of Y is the sum of the n variances and has the form

$$\sigma_Y^2 = nE\left\{\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2\right\}. \quad (32)$$

Now, it follows from Equation (29) that

$$1 = E\{\hat{\Theta}Y\}. \quad (33)$$

Recall that

$$E\{\hat{\Theta}Y\} = E\{\hat{\Theta}\}E\{Y\} + \rho_{\hat{\Theta}Y}\sigma_{\hat{\Theta}}\sigma_Y,$$

Or

$$1 = \theta(0) + \rho_{\hat{\Theta}Y}\sigma_{\hat{\Theta}}\sigma_Y. \quad (34)$$

As a consequence of property $\rho^2 \leq 1$, we finally have

$$\frac{1}{\sigma_{\hat{\Theta}}^2 \sigma_Y^2} \leq 1,$$

or, using Equation (32),

$$\sigma_{\hat{\Theta}}^2 \geq \frac{1}{\sigma_Y^2} = \left\{ nE\left\{\left[\frac{\partial \ln f(X; \theta)}{\partial \theta}\right]^2\right\} \right\}^{-1}. \quad (35)$$

The proof is now complete.

In the above, we have assumed that differentiation with respect to θ under an integral or sum sign are permissible. Equation (26) gives a lower bound on the variance of any unbiased estimator and it expresses a fundamental limitation on the accuracy with which a parameter can be estimated. We also note that this lower bound is, in general, a function of θ , the true parameter value.

Several remarks in connection with the Cramér–Rao lower bound (CRLB) are now in order.

- *Remark 1:* the expectation in Equation (26) is equivalent to $-E\{\partial^2 \ln f(X; \theta) / \partial \theta^2\}$, or

$$\sigma_{\hat{\Theta}}^2 \geq -\left\{ nE\left[\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2}\right] \right\}^{-1}. \quad (36)$$

This alternate expression offers computational advantages in some cases.

- *Remark 2:* the result given by Equation (26) can be extended easily to multiple parameter cases. Let $\theta_1, \theta_2, \dots$, and $\theta_m = (m \leq n)$ be the unknown parameters in $f(x; \theta_1, \dots, \theta_m)$ which are to be estimated on the basis of a sample of size n . In vector notation, we can write

$$\boldsymbol{\theta}^T = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_m], \quad (37)$$

with corresponding vector unbiased estimator

$$\hat{\boldsymbol{\theta}}^T = [\hat{\theta}_1 \quad \hat{\theta}_2 \quad \cdots \quad \hat{\theta}_m]. \quad (38)$$

Following similar steps in the derivation of Equation (26), we can show that the Cram r–Rao inequality for multiple parameters is of the form

$$\text{cov}\{\hat{\boldsymbol{\theta}}\} \geq \frac{\Lambda^{-1}}{n}, \quad (39)$$

where Λ^{-1} is the inverse of matrix Λ for which the elements are

$$\Lambda_{ij} = E \left\{ \frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(X; \boldsymbol{\theta})}{\partial \theta_j} \right\}, \quad i, j = 1, 2, \dots, m. \quad (40)$$

Equation (39) implies that

$$\text{var}\{\hat{\theta}_j\} \geq \frac{(\Lambda^{-1})_{jj}}{n} \geq \frac{1}{n\Lambda_{jj}}, \quad j = 1, 2, \dots, m, \quad (41)$$

where $(\Lambda^{-1})_{jj}$ is the jj th element of Λ^{-1} .

- *Remark 3:* the CRLB can be transformed easily under a transformation of the parameter. Suppose that, instead of $\boldsymbol{\theta}$, parameter $\boldsymbol{\phi} = \mathbf{g}(\boldsymbol{\theta})$ is of interest, which is a one-to-one transformation and differentiable with respect to $\boldsymbol{\theta}$; then,

$$\text{CRLB for var}\{\hat{\boldsymbol{\Phi}}\} = \left[\frac{d\mathbf{g}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} \right]^2 [\text{CRLB for var}(\hat{\boldsymbol{\theta}})], \quad (42)$$

where $\hat{\boldsymbol{\Phi}}$ is an unbiased estimator for $\boldsymbol{\phi}$.

- *Remark 4:* given an unbiased estimator $\hat{\boldsymbol{\theta}}$ for parameter $\boldsymbol{\theta}$, the ratio of its CRLB to its variance is called the efficiency of $\hat{\boldsymbol{\theta}}$. The efficiency of any unbiased estimator is thus always less than or equal to 1. An unbiased estimator with efficiency equal to 1 is said to be efficient. We must point out, however, that efficient estimators exist only under certain conditions.

We are finally in the position to answer the question posed in Example 2.

Example 2: part 2. Answer: first, we note that, in order to apply the CRLB, pdf $f(x; \boldsymbol{\theta})$ of population X must be known. Suppose that $f(x; m)$ for this example is $N(m, \sigma^2)$. We have

$$\begin{aligned} \ln f(X; m) &= \ln \left\{ \frac{1}{(2\pi)^{1/2}\sigma} \exp \left[-\frac{(X-m)^2}{2\sigma^2} \right] \right\} \\ &= \ln \left[\frac{1}{(2\pi)^{1/2}\sigma} \right] - \frac{(X-m)^2}{2\sigma^2}, \end{aligned}$$

and

$$\frac{\partial \ln f(X; m)}{\partial m} = \frac{X - m}{\sigma^2}.$$

Thus,

$$E \left\{ \left[\frac{\partial \ln f(X; m)}{\partial m} \right]^2 \right\} = \frac{1}{\sigma^4} E \{ (X - m)^2 \} = \frac{1}{\sigma^2}.$$

Equation (26) then shows that the CRLB for the variance of any unbiased estimator for m is σ^2/n . Since the variance of $\bar{X}$ is σ^2/n , it has the minimum variance among all unbiased estimators for m when population X is distributed normally.

Example 3: Problem: consider a population X having a normal distribution $N(0, \sigma^2)$ where σ^2 is an unknown parameter to be estimated from a sample of size $n > 1$.

(a) Determine the CRLB for the variance of any unbiased estimator for σ^2 .

(b) Is sample variance S^2 an efficient estimator for σ^2 ?

Answer: let us denote σ^2 by θ . Then,

$$f(X; \theta) = \frac{1}{(2\pi\theta)^{1/2}} \exp\left(-\frac{X^2}{2\theta}\right),$$

and

$$\ln f(X; \theta) = -\frac{X^2}{2\theta} - \frac{1}{2} \ln 2\pi\theta,$$

$$\frac{\partial \ln f(X; \theta)}{\partial \theta} = \frac{X^2}{2\theta^2} - \frac{1}{2\theta},$$

$$\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2} = -\frac{X^2}{\theta^3} + \frac{1}{2\theta^2},$$

$$E \left\{ \frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2} \right\} = -\frac{\theta}{\theta^3} + \frac{1}{2\theta^2} = -\frac{1}{2\theta^2}.$$

Hence, according to Equation (36), the CRLB for the variance of any unbiased estimator for θ is $2\theta^2/n$.

For S^2 , it has been shown that it is an unbiased estimator for θ and that its variance is [see Equation (10)]

$$\begin{aligned} \text{var}\{S^2\} &= \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right) \\ &= \frac{1}{n} \left(3\sigma^4 - \frac{n-3}{n-1} \sigma^4 \right) \\ &= \frac{2\sigma^4}{n-1} = \frac{2\theta^2}{n-1}, \end{aligned}$$

since $\mu_4 = 3\sigma^4$ when X is normally distributed. The efficiency of S^2 , denoted by $e(S^2)$, is thus

$$e(S^2) = \frac{\text{CRLB}}{\text{var}(S^2)} = \frac{n-1}{n}.$$

We see that the sample variance is not an efficient estimator for θ in this case. It is, however, asymptotically efficient in the sense that $e(S^2) \rightarrow 1$ as $n \rightarrow \infty$.

Example 4: Problem: determine the CRLB for the variance of any unbiased estimator for θ in the lognormal distribution

$$f(x; \theta) = \begin{cases} \frac{1}{x(2\pi\theta)^{1/2}} \exp\left(-\frac{1}{2\theta} \ln^2 x\right), & \text{for } x \geq 0, \text{ and } \theta > 0; \\ 0, & \text{elsewhere.} \end{cases}$$

Answer: we have

$$\frac{\partial \ln f(X; \theta)}{\partial \theta} = -\frac{1}{2\theta} + \frac{\ln^2 X}{2\theta^2},$$

$$\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2} = \frac{1}{2\theta^2} - \frac{\ln^2 X}{\theta^3},$$

$$E\left\{\frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2}\right\} = \frac{1}{2\theta^2} - \frac{\theta}{\theta^3} = -\frac{1}{2\theta^2}.$$

It thus follows from Equation (36) that the CRLB is $2\theta^2/n$.

Before going to the next criterion, it is worth mentioning again that, although unbiasedness as well as small variance is desirable it does not mean that we should discard all biased estimators as inferior. Consider two estimators for a parameter θ , $\hat{\theta}_1$ and $\hat{\theta}_2$, the pdfs of which are depicted in Figure 2(a). Although $\hat{\theta}_2$ is biased, because of its smaller variance, the probability of an observed value of $\hat{\theta}_2$ being closer to the true value θ can well be higher than that associated with an observed value of $\hat{\theta}_1$. Hence, one can argue convincingly that $\hat{\theta}_2$ is the better estimator of the two. A more dramatic situation is shown in Figure 2(b). Clearly, based on a particular sample of size n , an observed value of $\hat{\theta}_2$ will likely be closer to the true value θ than that of $\hat{\theta}_1$ even though $\hat{\theta}_1$ is again unbiased. It is worthwhile for us to reiterate our remark advanced that the quality of an estimator does not rest on any single criterion but on a combination of criteria.

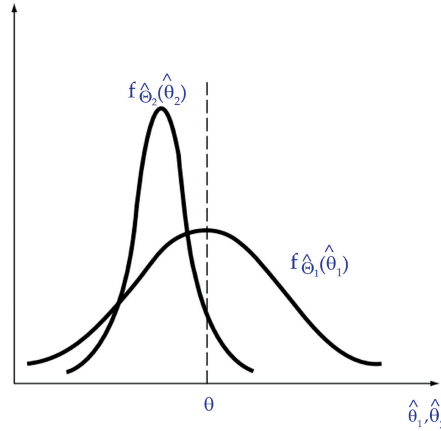
Example 5: To illustrate the point that unbiasedness can be outweighed by other considerations, consider the problem of estimating parameter θ in the binomial distribution

If the value an estimator estimates for the parameter, θ' , always converges to the actual parameter value θ as the quantity of data used for parameter estimation increases, we say an estimator is consistent.

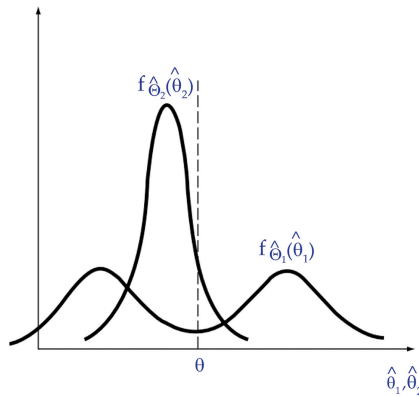
$$p_X(k) = \theta^k (1 - \theta)^{1-k}, \quad k = 0, 1. \quad (43)$$

Let us propose two estimators, $\hat{\Theta}_1$ and $\hat{\Theta}_2$, for θ given by

$$\left. \begin{aligned} \hat{\Theta}_1 &= \bar{X}, \\ \hat{\Theta}_2 &= \frac{n\bar{X} + 1}{n + 2}, \end{aligned} \right\} \quad (44)$$



(a)



(b)

Figure 2: Probability density functions of $\hat{\Theta}_1$ and $\hat{\Theta}_2$.

where $\bar{X}$ is the sample mean based on a sample of size n . The choice of $\hat{\Theta}_1$ is intuitively obvious since $E\{\bar{X}\} = \theta$, and the choice of $\hat{\Theta}_2$ is based on a prior probability argument that is not our concern at this point.

Since

$$E\{\bar{X}\} = \theta,$$

and

$$\sigma_{\bar{X}}^2 = \frac{\theta(1 - \theta)}{n}$$

we have

$$\left. \begin{aligned} E\{\hat{\Theta}_1\} &= \theta, \\ E\{\hat{\Theta}_2\} &= \frac{n\theta + 1}{n + 2}, \end{aligned} \right\} \quad (45)$$

and

$$\left. \begin{aligned} \sigma_{\hat{\Theta}_1}^2 &= \frac{\theta(1 - \theta)}{n}, \\ \sigma_{\hat{\Theta}_2}^2 &= \frac{n^2}{(n + 2)^2} \sigma_{\bar{X}}^2 = \frac{n\theta(1 - \theta)}{(n + 2)^2}. \end{aligned} \right\} \quad (46)$$

We see from the above that, although $\hat{\Theta}_2$ is a biased estimator, its variance is smaller than that of $\hat{\Theta}_1$, particularly when n is of a moderate value. This is a valid reason for choosing $\hat{\Theta}_2$ as a better estimator, compared with $\hat{\Theta}_1$, for θ , in certain cases.

Consistency

An estimator $\hat{\Theta}$ is said to be a consistent estimator for θ if, as sample size n increases,

$$\lim_{n \rightarrow \infty} P[|\hat{\Theta} - \theta| \geq \varepsilon] = 0,$$

(47)

for all $\varepsilon > 0$. The consistency condition states that estimator $\hat{\Theta}$ converges in the sense above to the true value θ as sample size increases. It is thus a large-sample concept and is a good quality for an estimator to have.

Example 6: Problem: show that estimator S^2 in Example 3 is a consistent estimator for σ^2 .

Answer: using the Chebyshev inequality, we can write

$$P\{|S^2 - \sigma^2| \geq \varepsilon\} \leq \frac{1}{\varepsilon^2} E\{(S^2 - \sigma^2)^2\}.$$

We have shown that $E\{S^2\} = \sigma^2$, and $\text{var}\{S^2\} = 2\sigma^2 / (n - 1)$. Hence,

Statistics and Machine Learning Toolbox™, which supports these and similar parameter estimation tasks with more than 40 different probability distributions, including Normal, Weibull, Gamma, Generalized Pareto, and Poisson. The toolbox also supports linear and nonlinear regression.

$$\lim_{n \rightarrow \infty} P\{|S^2 - \sigma^2| \geq \varepsilon\} \leq \lim_{n \rightarrow \infty} \frac{1}{\varepsilon^2} \left(\frac{2\theta^2}{n-1} \right) = 0.$$

Thus S^2 is a consistent estimator for σ^2 .

Example 6 gives an expedient procedure for checking whether an estimator is consistent. We shall state this procedure as a theorem below (Theorem 3). It is important to note that this theorem gives a sufficient, but not necessary, condition for consistency.

Theorem 3: Let $\hat{\Theta}$ be an estimator for θ based on a sample of size n . Then, if

$$\lim_{n \rightarrow \infty} E\{\hat{\Theta}\} = \theta, \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{var}\{\hat{\Theta}\} = 0, \quad (48)$$

estimator $\hat{\Theta}$ is a consistent estimator for θ .

The proof of Theorem 3 is essentially given in Example 6 and will not be repeated here.

Sufficiency

Let $X_1, X_2, \dots, X_n$ be a sample of a population X the distribution of which depends on unknown parameter θ . If $Y = h(X_1, X_2, \dots, X_n)$ is a statistic such that, for any other statistic

$$Z = g(X_1, X_2, \dots, X_n),$$

the conditional distribution of Z , given that $Y = y$ does not depend on θ , then Y is called a sufficient statistic for θ . If also $E\{Y\} = \theta$, then Y is said to be a sufficient estimator for θ .

In words, the definition for sufficiency states that, if Y is a sufficient statistic for θ , all sample information concerning θ is contained in Y . A sufficient statistic is thus of interest in that if it can be found for a parameter then an estimator based on this statistic is able to make use of all the information that the sample contains regarding the value of the unknown parameter. Moreover, an important property of a sufficient estimator is that, starting with any unbiased estimator of a parameter θ that is not a function of the sufficient estimator, it is possible to find an unbiased estimator based on the sufficient statistic that has a variance smaller than that of the initial estimator. Sufficient estimators thus have variances that are smaller than any other unbiased estimators that do not depend on sufficient statistics.

If a sufficient statistic for a parameter θ exists, Theorem 4, stated here without proof, provides an easy way of finding it.

Theorem 4: Fisher–Neyman factorization criterion. Let

$$Y = h(X_1, X_2, \dots, X_n)$$

be a statistic based on a sample of size n . Then Y is a sufficient statistic for θ if and only if the joint probability density function of $X_1, X_2, \dots, X_n$, $f_X(x_1; \theta) \dots f_X(x_n; \theta)$ can be factorized in the form

$$\prod_{j=1}^n f_X(x_j; \theta) = g_1[h(x_1, \dots, x_n), \theta] g_2(x_1, \dots, x_n). \quad (49)$$

If X is discrete, we have

$$\prod_{j=1}^n p_X(x_j; \theta) = g_1[h(x_1, \dots, x_n), \theta] g_2(x_1, \dots, x_n). \quad (50)$$

The foregoing results can be extended to the multiple parameter case. Let $\theta^T = [\theta_1 \dots \theta_m]$, $m \leq n$ be the parameter vector. Then $Y_1 = h_1(X_1, \dots, X_n), \dots, Y_r = h_r(X_1, \dots, X_n)$, $r \leq m$, is a set of sufficient statistics for θ if and only if

$$\prod_{j=1}^n f_X(x_j; \theta) = g_1[h(x_1, \dots, x_n), \theta] g_2(x_1, \dots, x_n), \quad (51)$$

where $h^T = [h_1 \dots h_r]$. A similar expression holds when X is discrete.

Example 7: Let us show that statistic $\bar{X}$ is a sufficient statistic for θ in Example 5. In this case,

$$\begin{aligned} \prod_{j=1}^n p_X(x_j; \theta) &= \prod_{j=1}^n \theta^{x_j} (1 - \theta)^{1-x_j} \\ &= \theta^{\sum x_j} (1 - \theta)^{n - \sum x_j}. \end{aligned} \quad (52)$$

We see that the joint probability mass function (jpmf) is a function of $\sum x_j$ and θ . If we let

$$Y = \sum_{j=1}^n X_j,$$

the jpmf of $X_1, \dots, X_n$ takes the form given by Equation (50), with

$$g_1 = \theta^{\sum x_j} (1 - \theta)^{n - \sum x_j},$$

And

$$g_2 = 1.$$

In this example,

$$\sum_{j=1}^n X_j$$

is thus a sufficient statistic for θ . We have seen in Example 5 that both $\hat{\theta}_1$ and $\hat{\theta}_2$, where $\hat{\theta}_1 = \bar{X}$, and $\hat{\theta}_2 = (n\bar{X} + 1)/(n+2)$ are based on this sufficient statistic. Furthermore

$\hat{\theta}_1$, being unbiased, is a sufficient estimator for θ .

Example 8: Suppose $X_1, X_2, \dots$, and X_n are a sample taken from a Poisson distribution; that is,

$$p_X(k; \theta) = \frac{\theta^k e^{-\theta}}{k!}, \quad k = 0, 1, 2, \dots, \quad (53)$$

where θ is the unknown parameter. We have

$$\prod_{j=1}^n p_X(x_j; \theta) = \frac{\theta^{\sum x_j} e^{-n\theta}}{\prod x_j!}, \quad (54)$$

which can be factorized in the form of Equation (50) by letting

$$g_1 = \theta^{\sum x_j} e^{-n\theta},$$

and

$$g_2 = \frac{1}{\prod x_j!}$$

It is seen that

$$Y = \sum_{j=1}^n X_j$$

is a sufficient statistic for θ .



CORE CONCEPT: METHODS OF ESTIMATION

The Method of Moments is a simple technique based on the idea that the sample moments are “natural” estimators of population moments. In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Based on the estimation criteria defined, some estimation techniques that yield ‘good’, and sometimes ‘best’, estimates of distribution parameters are now developed. Two approaches to the parameter estimation problem are discussed in what follows: point estimation and interval estimation. In point estimation, we use certain prescribed methods to arrive at a value for $\hat{\theta}$ as a function of the observed data that we accept as a ‘good’ estimate of θ – good in terms of unbiasedness, minimum variance, etc.,

as defined by the estimation criteria. In many scientific studies it is more useful to obtain information about a parameter beyond a single number as its estimate. Interval estimation is a procedure by which bounds on the parameter value are obtained that not only give information on the numerical value of the parameter but also give an indication of the level of confidence one can place on the possible numerical value of the parameter on the basis of a sample. Point estimation will be discussed first, followed by the development of methods of interval estimation.

Point Estimation

A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ . A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the point estimator of θ .

Given a parameter of interest, such as a population mean μ or population proportion p , the objective of point estimation is to use a sample to compute a number that represents in some sense a good guess for the true value of the parameter. The resulting number is called a point estimate. We present some general concepts of point estimation. We describe and illustrate two important methods for obtaining point estimates: the method of moments and the method of maximum likelihood.

We now proceed to present two general methods of finding point estimators for distribution parameters on the basis of a sample from a population.



Some General Concepts of Point Estimation

Statistical inference is almost always directed toward drawing some type of conclusion about one or more parameters (population characteristics). To do so requires that an investigator obtain sample data from each of the populations under study. Conclusions can then be based on the computed values of various sample quantities.

When discussing general concepts and methods of inference, it is convenient to have a generic symbol for the parameter of interest. We will use the Greek letter θ for this purpose. The objective of point estimation is to select a single number, based on sample data, that represents a sensible value for θ .

Suppose we want to estimate a parameter of a single population (e.g., μ or σ) based on a random sample of size n . Recall from the before data is available, the sample observations must be considered random variables (rv's) $X_1, X_2, \dots, X_n$. It follows that any function of the X_i 's—that is, any statistic—such as the sample mean $\bar{X}$ or sample standard deviation S is also a random variable. The same is true if available data consists of more than one sample. For example, we can represent tensile strengths of m type 1 specimens and n type 2 specimens by $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$, respectively. The difference between the two sample mean strengths is $\bar{X} - \bar{Y}$, the natural statistic for making inferences about $\mu_1 - \mu_2$, the difference between the population mean strengths.

PERT Three Point Estimate Technique

The PERT Three Point Estimate technique is a type of three point estimate. The only difference is that it applies weighting so that the most-likely estimate is weighted 4 times more than the other two estimates (optimistic and pessimistic). This formula is most valuable in estimating time or cost of activities for projects that are especially unique, such as in research and development where there are many unknowns. For projects that are similar to previous projects and there is good historical data and expert experience, the formula is less useful because you could use other techniques like analogous estimating (based on previous experience and projects).

Pert Estimate

$$E = (o + 4m + p) / 6$$

Standard Deviation

$$SD = (p - o) / 6$$

where SD is **Standard Deviation**; p = pessimistic estimate; o = optimistic estimate

To produce a PERT three point estimate:

- Decompose the project into a list of estimable tasks, i.e. identify tasks for each Work Breakdown Structure (WBS) work package
- Estimate the E value and SD for each task.
- Calculate the E value for the total project work as $E(\text{Project Work}) = \sum E(\text{Task})$
- Calculate the SD value for the total project work as

In statistics, the **standard deviation** (SD, also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

$$SD (\text{Project Work}) = \sqrt{\sum SD (\text{Task})^2}$$

The E and SD values are then used to convert the project estimates to confidence levels as follows:

- Confidence level in E value $\pm$ SD is approximately 68%
- Confidence level in E value $\pm 1.645 \times SD$ is approximately 90%
- Confidence level in E value $\pm 2 \times SD$ is approximately 95%
- Confidence level in E value $\pm 3 \times SD$ is approximately 99.7%
- Information Systems typically use the 95% confidence level, i.e. E Value + $1.645 \times SD$, for all project and task estimates.[2]

PERT Three Point Estimate Example:

For Activity A:

- $o = 4$ hours
- $p = 16$ hours
- $m = 8$ hours

Using the estimates above for Activity A, calculate the Estimate:

- $E = (4 + 4(8) + 16) / 6$
- $E = 52 / 6$
- $E = 6.5$ hours

Using the estimates above for Activity A, calculate the Standard Deviation:

- $SD = (16 - 4) / 6$
- $SD = 12 / 6$
- $SD = 2$ hours

PERT Three Point Estimate Results for Activity A:

- 4.5h – 8.5h hours: Confidence level in E value $\pm$ SD is approximately 68.2%
- 5.3h – 7.7h: Confidence level in E value $\pm 1.645 \times SD$ is approximately 90%
- 2.5h – 10.5h: Confidence level in E value $\pm 2 \times SD$ is approximately 95%
- .5h – 12.5h: Confidence level in E value $\pm 3 \times SD$ is approximately 99.7%
- Information Systems typically use the 90% confidence level, i.e. E Value + $1.645 \times SD$, for all project and task estimates.

Interval Estimation

We now examine another approach to the problem of parameter estimation. The interval estimation provides, on the basis of a sample from a population, not only information on the parameter values to be estimated, but also an indication of the level of confidence that can be placed on possible numerical values of the parameters. Before developing the theory of interval estimation, an example will be used to

demonstrate that a method that appears to be almost intuitively obvious could lead to conceptual difficulties.

This is the technique used when we use a point estimator (sample estimate) to construct an interval where we expect to find the population parameter with a certain probability. To do this we need to have a sample estimate and we need to specify what degree of certainty we want to know that the actual population parameter lies in the interval.

Statistical Inference

Using sample data and estimates to draw a conclusion about the population.

Interval Estimate

This is when we construct an interval by adding or subtracting a margin of error to a point estimate.

Mathematically: Point Estimate $\pm$ margin of error

Note: $\bar{x}$ & $\bar{p}$ are examples of point estimates commonly used.

Interval Estimation about a Population Mean (Large Sample; $n \geq 30$)

a. *sampling error* – tells us how close the population parameter is to the sample estimate.

Mathematically: $|\bar{x} - \mu|$

-note that since most likely the population parameter is unknown we can't say for sure how far away we are from the actual mean, but we can construct a sampling distribution and make probability statements about it assuming that we have a normal distribution.

b. *precision statement* – tells us with what precision the sampling error will take.

So the sampling error ($|\bar{x} - \mu|$) takes on some error (it could be less) with a certain probability.

Mathematically: $\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$

Note: our margin of error here is the $z_{\alpha/2} \sigma_{\bar{x}}$; so we are within some standard deviations away from the mean.

c. Confidence Coefficient & Precision Terms

- *confidence level* – the precision which you make the particular intervals $\rightarrow$ i.e. 95% or some percentage
- *confidence coefficient* – the actual percent level used
- *level of significance* – α – tells use the chance that we are not in the

confidence interval. We can take 1- confidence level = α .

ex: so if we use the 95% level we get $1 - 0.95 = \alpha = 0.05$.

Rule: If we have σ known we can use the population **variance** to construct the sampling distributions variance as $\sigma / \sqrt{n}$, but many times we don't know the population variance. In this case we can estimate the population variance with s , or the estimate of the population variance. This is calculated in the same manner as we discussed earlier. The interval is the exact same as before, but instead of the population variance we use $s \rightarrow \bar{x} \pm z_{\alpha/2} s_{\bar{x}}$

d. Example: Suppose we have the following information. Construct a 95% confidence interval.

$$\bar{x} = 80$$

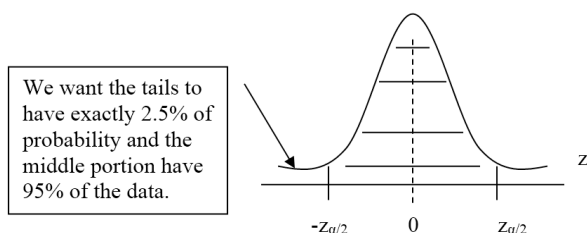
$$\sigma = 3$$

$$n = 10$$

$$\alpha = 0.05$$

Step 1: Find the appropriate z-score. So we want an interval that looks like the following.

Graph 1



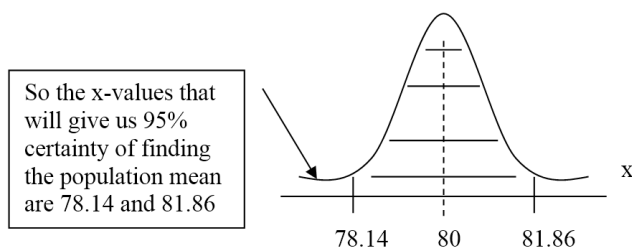
So we go to the table and find the Z that would give use this and it is $z = 1.96$.

Step 2: Find the sampling distributions variance $\rightarrow \sigma_{\bar{x}} = \sigma / \sqrt{n} \rightarrow 3 / \sqrt{10} = 0.949$

Step 3: Construct the confidence interval:

$$\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}} \rightarrow 80 \pm 1.96 (0.949) \rightarrow 80 \pm 1.86 \rightarrow 78.14 \text{ to } 81.86$$

Graph-2:



Interval Estimation about a Population Mean (small sample; $n < 30$)

In a small sample case we have to worry about the central limit theorem. Recall that as we have larger samples we can be assured that the data takes on a normal distribution. In small samples we cannot be assured this. So we must use another distribution (t-distribution) in the case where the population variance is unknown.

a. Case 1: σ is known

In this case we still use the normal distribution because we can conclude that the population is normal and therefore the sampling distribution is also normal. This is not any different from the calculations that we did above.

Mathematically: $\bar{x} \pm z_{\alpha/2} \sigma \bar{x}$

Where $1-\alpha$ = confidence coefficient and $z_{\alpha/2}$ gives the z-scores that gives us $\alpha/2$ probability in the upper and lower tails.

b. Case 2: σ is unknown

In this case we cannot assume that our sampling distribution follows a completely normal distribution. Now we must use the t-distribution (sometimes called the students t-distribution) to make probability statements. The technique is still the same as before, but now we go to another chart to get the critical values to construct our confidence interval.

– t-distribution

It is a family of distributions that is based on degrees of freedom. Each distribution with its degrees of freedom has its own features. As the degrees of freedom (d.o.f) go up the distributions get closer and closer to the standard normal. This is b/c as the d.o.f. go up the variability is reduced.

– We read the chart the same way as the standard normal with the only difference coming from the fact that we now have d.o.f which is $(n-1)$

- $Confidence\ Interval = \bar{x} \pm t_{\alpha/2} s \bar{x} = \bar{x} \pm t_{\alpha/2} s / \sqrt{n}$

recall that $s^2 = \sum (x_i - \bar{x})^2 / n-1$; this is just sample variance. To get s we simply take the square root of s^2 .

so the only time you use the t-distribution is in the case of a small sample and when the population variance is unknown. The CI is constructed in the same fashion.

Example: Consider a case where you want a 95% CI for $\bar{x}=10$, $n=25$, and $s=3$

We use a t-distribution in this case since the population standard deviation is unknown and we have a small sample.

$$\text{So } \bar{x} \pm t_{\alpha/2} s_{\bar{x}} = \bar{x} \pm t_{\alpha/2} s / \sqrt{n} = 10 \pm 2.064(3/\sqrt{3}) = 10 \pm 1.24 \rightarrow 8.76 \text{ to } 11.24$$

This implies we are 95% confident the true population mean lies somewhere between 8.76 and 11.24

Determining Sample Size

We can use this technique to find sample size required when we want to know a sample confidence interval contain the population parameter with a certain probability.

If we let E = margin of error (could also be M) = $z_{\alpha/2} \sigma_{\bar{x}}$; this is because this is how far we allow the sample stat to vary (see sampling error above)

$$\text{So we have } E = z_{\alpha/2} \sigma / \sqrt{n} \rightarrow \sqrt{n} = z_{\alpha/2} \sigma / E \rightarrow n = [(z_{\alpha/2} \sigma) / E]^2$$

Example: Suppose we are given the following information and want to know how large a sample we need to be sure that we are 90% confident we will find our population parameter in our sample.

$$E = 2 \text{ units}$$

$$\sigma = 3$$

$$\alpha = 0.10$$

Step 1: Find the appropriate $z \rightarrow z_{\alpha/2} = 1.645$; this is the z value that gives us 90% of the data between 1.645 and -1.645.

Step 2: Plug in values and find $n \rightarrow n = [(z_{\alpha/2} \sigma) / E]^2 \rightarrow [(1.645 * 3) / 2]^2 = 6.08$ or 7

****Note: generally we round up to ensure that we fall within our confidence limits.

A **population proportion**, generally denoted by p and in some textbooks by π , is a parameter that describes a percentage value associated with a population.

Constructing a CI for p

a. σ known

If we want to construct a CI it is exactly the same method as before but now we use our to construct a confidence interval for the true **population proportion**.

$$\pm z_{\alpha/2}$$

The standard deviation of the sampling distribution is equal to

-so we can reduce our formula above to:

$$\pm z_{\alpha/2}$$

Example: If we have a sample of 49 children from a school and there are 9 of them with a cold. Find a 99% confidence interval for the true proportion of children at the school who have a cold.

$$= 0.183$$

$$z_{\alpha/2} = 2.576$$

$$= 0.0552$$

So the 99% CI is $0.183 \pm (2.576)(0.0552)$ or 0.0407 to 0.325

Other Notes for P

a. Finding Sample Size: Once again we can find the appropriate sample size for a desired margin of error as follows:

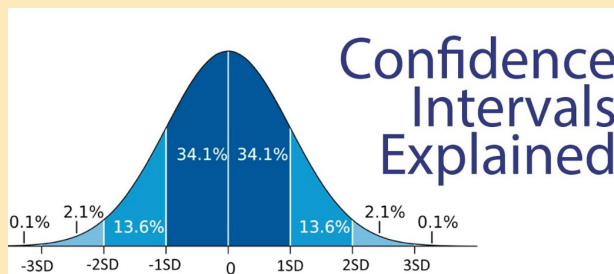
$$ME = z_{\alpha/2} \sqrt{p(1-p)} \quad \text{So if we solve for } n \text{ we get } n = \frac{z_{\alpha/2}^2 p(1-p)}{ME^2}$$

So if we knew that $p = 0.6$ and you wanted a margin of error of say 4% with a 95% confidence then the appropriate sample would be:

$$n = \frac{z_{\alpha/2}^2 p(1-p)}{ME^2} = \frac{(1.96)^2 (0.6)(1-0.6)}{(0.04)^2} = 576.24 \text{ or about } 577 \text{ observations.}$$

A CONCEPT OF CONFIDENCE INTERVAL

In Statistics, a confidence interval is a kind of interval calculation, obtained from the observed data that holds the actual value of the unknown parameter. It is associated with the confidence level that quantifies the confidence level in which the interval estimates the deterministic parameter. Also, we can say, it is based on Standard Normal Distribution, where Z value is the z-score. Here, let us look at the definition, formula, table, and the calculation of the confidence level in detail.



Confidence Interval Definition

The confidence level represents the proportion (frequency) of acceptable confidence intervals that contain the true value of the unknown parameter. In other terms, the confidence intervals are evaluated using the given confidence level from an endless number of independent samples. So that the proportion of the range contains the true value of the parameter that will be equal to the confidence level.

Mostly, the confidence level is selected before examining the data. The commonly used confidence level is 95% confidence level. However, other confidence levels are also used, such as 90% and 99% confidence levels.

Confidence Interval Formula

The confidence interval is based on the mean and standard deviation. Thus, the formula to find CI is

$$\bar{X} \pm Z_{\alpha/2} \times \left[\sigma / \sqrt{n} \right]$$

Where

$\bar{X}$ = Mean

Z = Confidence coefficient

α = Confidence level

σ = Standard deviation

N = sample space

The value after the $\pm$ symbol is known as the margin of error.

Note: This interval is only accurate when the population distribution is normal. But, in the case of large samples from other population distributions, the interval is almost accurate by the Central Limit Theorem.

SUMMARY

- Statistics are used as estimators of population quantities with an estimate denoting a given realisation of an estimator.
- A sufficient statistic is thus of interest in that if it can be found for a parameter then an estimator based on this statistic is able to make use of all the information that the sample contains regarding the value of the unknown parameter.
- A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ . A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data.
- The PERT Three Point Estimate technique is a type of three point estimate. The only difference is that it applies weighting so that the mostlikely estimate is weighted 4 times more than the other two estimates (optimistic and pessimistic).
- The interval estimation provides, on the basis of a sample from a population, not only information on the parameter values to be estimated, but also an indication of the level of confidence that can be placed on possible numerical values of the parameters.

MULTIPLE CHOICE QUESTIONS

1. **What is parameter estimation in statistics?**
 - a. Finding sample statistics
 - b. Estimating population parameters from sample data
 - c. Counting the number of variables in a dataset
 - d. Calculating p-values
2. **Point estimation involves:**
 - a. Providing a range of values within which a parameter is likely to fall
 - b. Calculating a test statistic
 - c. Providing a single value estimate for a parameter
 - d. Comparing observed data to a population distribution
3. **What is the main goal of interval estimation (confidence intervals)?**
 - a. Providing a single point estimate for a parameter
 - b. Finding the population parameter
 - c. Reducing variability in the data
 - d. Providing a range of plausible values for a parameter
4. **Bias in an estimator refers to:**
 - a. The variability of the estimator's values
 - b. The average difference between the estimator's values and the true parameter value
 - c. The number of observations in the sample
 - d. The efficiency of the estimator
5. **Maximum Likelihood Estimation (MLE) seeks parameter values that:**
 - a. Minimize the likelihood of observing the sample data
 - b. Minimize the variance of the sample data
 - c. Maximize the likelihood of observing the sample data
 - d. Maximize the mean of the sample data

Answers

1. (b) 2. (c) 3. (d) 4. (b) 5. (c)

REVIEW QUESTIONS

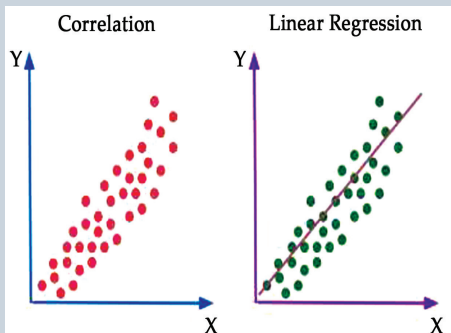
1. What do you understand by sample mean?
2. Describe the point estimation.
3. What is the sample variance?
4. Write the order of sample statistics.
5. Explain the minimum variance of quality criteria.
6. What do you understand by unbiasedness?

REFERENCES

1. Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Brooks/Cole.
2. DeGroot, M. H., & Schervish, M. J. (2002). *Probability and Statistics*. Pearson.
3. Ghosh, M., Delampady, M., & Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer.
4. Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical Statistics*. Pearson.
5. Mendenhall, W., & Sincich, T. (2011). *Statistics for Engineering and the Sciences*. Pearson.
6. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

CHAPTER 7

SIMPLE LINEAR REGRESSION AND CORRELATION



Simple Linear Regression is a method used to model the relationship between a dependent variable (also known as the response variable or outcome) and an independent variable (also known as the predictor variable or input). The goal is to find a linear equation that best describes the relationship between the two variables. Correlation measures the strength and direction of the linear relationship between two continuous variables. It quantifies how closely the data points resemble a

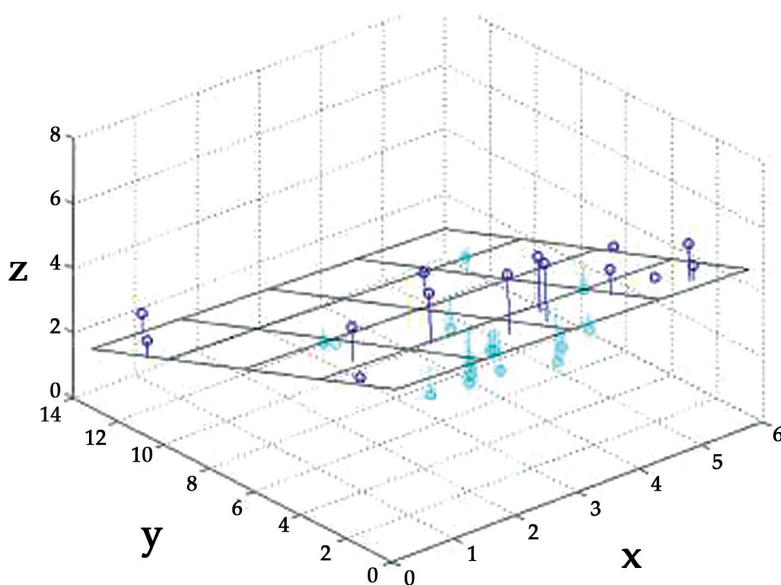
straight line. Correlation is typically denoted by the symbol r , and it ranges between -1 and 1.

After studying this chapter, you will understand the core concepts of:

- Understand the simple linear regression
- Define the model behind linear regression
- Describe the interpreting regression coefficients
- Illustrate the robustness of simple linear regression
- Focus on correlation

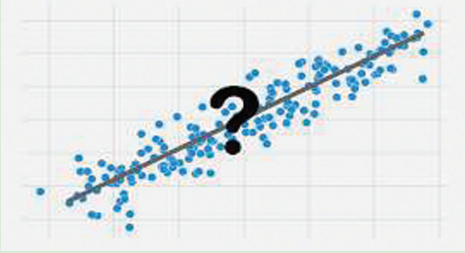
INTRODUCTION

Simple linear regression refers to fitting a response y as a linear function of a single predictor x . Although calibration of assay methods sometimes requires more than a single predictor x , simple linear regression is often adequate. In fact, assay developers often aim for a simple linear relation between a quick-to-measure response and a single predictor, such as radioactive source strength. Regression fitting is used during calibration and then to apply the assay method, the quick-to-measure response is used to predict the source strength (such as grams of nuclear material) of new test items. The simple linear regression, residual diagnostics, and variance propagation in a form suitable for implementation, intended both for practitioners who calibrate instruments and also as a case study of good applied statistical practice.



The word “regression” has a much different meaning outside the realm of statistics than it does within it. The literal definition of the word is to revert back (i.e., “regress”) to a state or form.

Simple Linear Regression



CORE CONCEPT: SIMPLE LINEAR REGRESSION

Simple Linear Regression is a statistical technique used to model and analyze the relationship between two continuous variables: a dependent variable (also called the response variable) and an independent variable (also

known as the predictor variable). It assumes that there is a linear relationship between the two variables, meaning that changes in the independent variable are associated with proportional changes in the dependent variable.

In regression there is always a single “dependent” variable, and one or more “independent” variables.

For example, we might think of the number of nonconforming units produced within a particular company each month as being dependent on the amount of time devoted to training operators. The amount of time (in minutes, say) would be the single independent variable. This should not be construed to mean, however, that there is a cause-and-effect relationship. There might be such a relationship, but that is not assumed in regression analysis. Furthermore, a cause-and-effect relationship can be detected only through the use of designed experiments. The word simple is used to denote the fact that a single independent variable is being used. Linear does not have quite the meaning that one would expect. Specifically, it does not necessarily mean that the relationship between the dependent variable and the independent variable is a straight-line relationship.

$$Y = \beta_0 + \beta_1 X + C \quad (1)$$

The equation is simply one example of a linear regression model, although this is the standard model. Notice how this closely resembles the equation for a straight line in algebra, with β_1 denoting the slope and β_0 is representing the Y-intercept. The symbol $\hat{I}$ represents the error term. This does not mean that a mistake is being made; it is simply a symbol used to indicate the absence of an exact relationship between the independent variable X and the **dependent variable** Y , as exact relationships do not exist in regression analysis. Other examples of linear regression models with a single X include

Dependent variable is a variable in an expression that depends on the value of another variable.

$$Y = \beta'_0 + \beta'_1 X + \beta'_{11} X^2 + \epsilon$$

and

$$Y = \beta'_1 X + \epsilon \quad (2)$$

A regression model is linear if it is linear in the parameters (the betas), and both of these last equations satisfy that condition.

“A linear regression model is a model that is linear in the parameters, regardless of the form of the predictors in the model”. We will note at the outset that the model given in equation (2) is rarely used, even though in some applications that might seem counterintuitive.

For example, assume that X represents an input variable such as grams of a product and Y is an output variable that is related directly to the input and depends entirely on the input. If the input is zero, then the output will also be zero, so the model given in equation (2) would seem to be appropriate. But we generally are not interested in using a regression model when X is approximately zero, however, and to do so would generally require extrapolating beyond the range of X -values in the sample. Therefore, the model in equation (1) can be expected to almost always provide a better fit to the data for the range of X that is of interest.

Regression analysis is not used for variables that have an exact linear relationship, as there would be no need for it.

For example, the equation expresses temperature in Fahrenheit as a function of temperature measured on the Celsius scale.

$$F = \frac{9}{5}C + 32 \quad (3)$$

Thus, F can be determined exactly for any given value of C . This is not the case for Y in equation (1), however, as β_0 and β_1 are generally unknown and therefore must be estimated. Even if they were known [as is the case with the constants $9/5$ and 32 in equation (3)], the presence of $\hat{\epsilon}$ in equation (1) would still prevent Y from being determined exactly for a given value of X .

It is important to recognize that a model such as equation (1) is simply an approximation to reality. Certainly, the world is nonlinear, so there is the need to determine if the linear regression model is an adequate approximation to reality for a particular application.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Regression Equation

The parameters β_0 and β_1 must be estimated. Various methods have been proposed

for doing so, and the most frequently used method for doing so is given. Once the parameters have been estimated, the equation is called the regression equation.

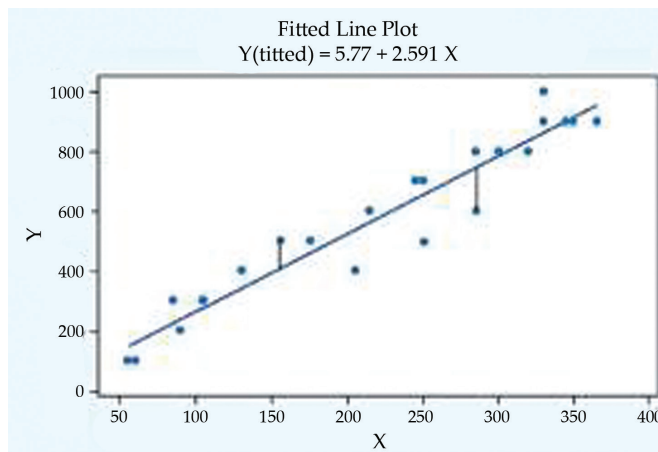
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (4)$$

When it is to be used to predict future values of Y , it could be called the prediction equation, with the values of $\hat{Y}$ the predicted values. (When the regression equation is used to compute the values of $\hat{Y}$ to perhaps compare with the observed values of Y in the dataset that is used to estimate the parameters, the values of $\hat{Y}$ are termed the *fitted values*.)

Estimating β_0 and β_1

Point estimates of β_0 and β_1 are needed to obtain the prediction equation given in general form in equation (4). A crude approach would be to draw a line through the center of the points and then use the Y -intercept and slope of the line as the estimates of β_0 and β_1 , respectively.

Before one could even attempt to do so, however, it would be necessary to define what is meant by “center.” We could attempt to minimize the sum of the slant distances (with each distance measured from the point to the line), but since we will be using X to predict Y , it would make more sense to try to minimize the sum of the vertical distances. If we use the signs of those distances we have a problem, however, because then the line is not uniquely defined.



The standard approach is to minimize the sum of the squares of the vertical distances, and this is accomplished by using the method of least squares. Figure illustrates the general idea; with the line that is shown being the least squares fitted line, which minimizes the sum of the squared vertical distances of each point to the line.

For the purpose of illustration, we must assume that equation (1) is the correct model, although as stated, the correct model will generally be unknown. The starting point is to write the model as

$$\epsilon = Y - (\beta_0 + \beta_1 X) \quad (5)$$

Since $\hat{I}$ represents the vertical distance from the observed value of Y to the line represented by $Y = \beta_0 + \beta_1 X$ that we would have if β_0 and β_1 were known, we want to minimize $\sum \epsilon^2$.

Recall that it was shown that this method minimizes the sum of the squares of the differences between the observations and their mean. The mean of $\hat{I}$ is assumed to be zero (i.e., the sum of the differences of the values on the right-hand side would be zero if the parameters were known). With this assumption, it follows that the mean of Y is $(\beta_0 + \beta_1 \bar{X})$, or properly stated as $\mu_{y|x}$, the conditional mean of Y given X since the mean depends on X . Thus,

$$\sum \epsilon^2 = \sum (Y - \mu_{y/x})^2$$

For convenience we define L as

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (6)$$

with n denoting the number of data points in a sample that has been obtained. To minimize L , we take the **partial derivative** of L with respect to each of the two parameters that we are estimating and set the resulting expressions equal to zero. Thus,

$$\frac{\partial L}{\partial \beta_0} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-1) = 0 \quad (7a)$$

and

$$\frac{\partial L}{\partial \beta_1} = 2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0 \quad (7b)$$

Dropping the 2 and the -1 from equations (7a) and (7b), the solutions for β_0 and β_1 would be obtained by solving the equations (which are generally called the normal equations)

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) \quad \text{and} \quad \sum_{i=1}^n (X_i Y_i - X_i \beta_0 - \beta_1 X_i^2) = 0$$

Which become

$$n\beta_0 + \beta_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \quad \text{and} \quad \beta_0 \sum_{i=1}^n X_i + \beta_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

The solution of these two equations (for the estimators of β_0 and β_1) produces the least squares estimators

Partial derivative is a way to find the slope in either the x or y direction, at the point indicated.

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - (\sum X_i)(\sum Y_i) / n}{\sum X_i^2 - (\sum X_i)^2 / n} \quad (8)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (9)$$

For notational convenience we will denote the numerator of equation (8) by S_{xy} and the denominator by S_{xx} . Thus $\hat{\beta}_1 = S_{xy} / S_{xx}$. The fitted line that results from the use of these parameter estimates in equation (4), as depicted in Figure, minimizes the sum of the squared distances of the points to the line.

Like, the astute reader will recognize that L in equation (6) is not automatically minimized just by setting the first derivatives equal to zero and solving the resultant equations. It can be shown, however, that the determinant of the matrix of second-order partial derivatives of L is positive, thus ensuring that a minimum, rather than a maximum, has been attained. Consider the following simple example.

X	2	3	6	5	4
Y	6	4	5	3	1

Applying equation (8), this would have to be used first

$$\hat{\beta}_1 = \frac{83 - (20)(19)/5}{90 - (20)^2/5} = 0.7$$

Then, $\hat{\beta}_0 = 4.8 - (0.7)4 = 2$. The fitted equation is then $\hat{Y} = 0.7 + 2X$.

Assumptions

If we simply wanted to obtain the prediction equation and compute some measure of the worth of the model, we would need to assume only that the model is a good proxy for the true, unknown model. We almost always want to go beyond that point, however, and make various types of inferences (confidence intervals, etc.). In order to do so, we have to make a distributional assumption on the error term, $\hat{\epsilon}$, and the usual assumption is that $\hat{\epsilon}$ has a normal distribution. Specifically, we assume that $\epsilon_i \sim N \text{ ID } (0, \sigma^2_{\epsilon})$. That is, the errors are assumed to be normally distributed (“N”), independent (“ID”), with a mean of zero and a variance that is constant for each error term and also does not depend on any variable. It is important that the assumptions of normality, independence, and constant error variance be checked, as violation of these assumptions can cause problems, especially violation of the independence assumption.

Sequence of Steps

The first step in any regression analysis with a single predictor is to plot the data. This would be done to see if there is evidence of a linear relationship between X and

Y as well as for other purposes such as checking for outlying observations (outliers). The importance of plotting the data cannot be overemphasized. Anscombe drove that point home with four sample sets he provided in his paper all of which produced virtually the same regression equation but had completely different graphs.

Example

Objective

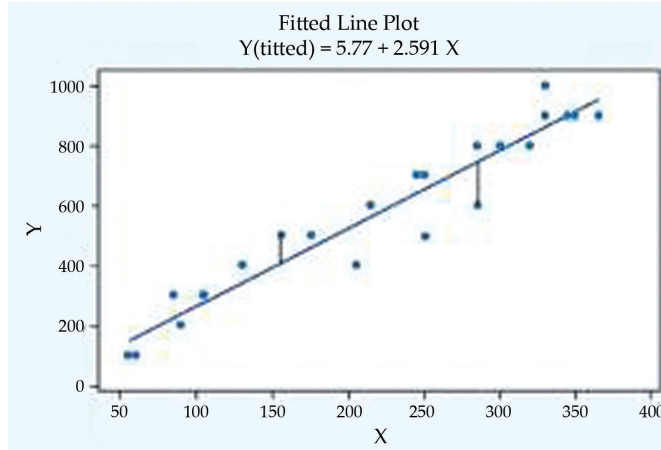
We will use a dataset given by T. Raz and M. Barad in their paper, “In- Process Control of Design Inspection Effectiveness” to emphasize the importance of first plotting the data. The objective was to model the defect escape probability (PROB) as part of a new methodology for the in-process control of design inspection. The probability was the dependent variable and size of the work product inspected was the independent variable. Clusters of the data were formed initially, with 2 of the 12 clusters discarded because the clusters contained fewer than 5 observations.

Data and Preliminary Analysis

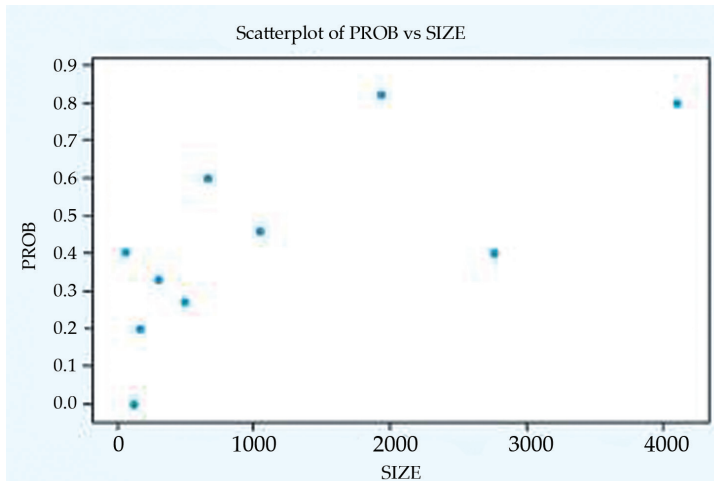
In the first modeling phase, the researchers looked at both SIZE and LOG (SIZE) as the independent variable, and if we plot the data given, we can see why we should use LOG (SIZE). (The units for SIZE are thousands of lines of code.)

Linear relationship
is a statistical term used to describe a straight-line relationship between two variables.

Prob	Size
0.82	1936.39
0.40	2766.31
0.80	4101.67
0.42	51.15
0.00	107.72
0.20	160.32
0.27	486.20
0.46	1039.34
0.33	295.68
0.60	657.47



Clearly, there is not a straight-line relationship; rather, there is evidence of a logarithmic relationship. Plotting PROB versus LOG (SIZE) in Figure shows some evidence of a **linear relationship**, although a couple of points do not conform to that relationship.



Method of Ordinary Least Square (OLS)

Ordinary least squares (OLS) regression is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line. In this entry, OLS regression will be discussed in the context of a bivariate model, that is, a model in which there is only one independent variable (X) predicting a dependent variable (Y). However, the logic of OLS regression is easily extended to the multivariate model in which there are two or more independent variables.

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between

one or more independent quantitative variables and a dependent variable (simple or multiple linear regression). Least squares stands for the minimum squares error (SSE). Maximum likelihood and Generalized method of moments estimator are alternative approaches to OLS.

In practice, you can use linear regression in many fields:

- meteorology, if you need to predict temperature or rainfall based on external factors.
- biology, if you need to predict the number of remaining individuals in a species depending on the number of predators or life resources.
- economy, if you need to predict a company's turnover based on the amount of sale.

Equations for the Ordinary Least Squares Regression

In the case of a model with p explanatory variables, the OLS regression model writes:

$$Y = \beta_0 + \sum_{j=1..p} \beta_j X_j + \varepsilon$$

where Y is the dependent variable, β_0 is the intercept of the model, X_j corresponds to the j^{th} explanatory variable of the model ($j= 1$ to p), and ε is the random error with expectation 0 and variance σ^2 .

In the case where there are n observations, the estimation of the predicted value of the dependent variable Y for the i^{th} observation is given by:

$$y_i = \beta_0 + \sum_{j=1..p} \beta_j X_{ij}$$

Example: We want to predict the height of plants depending on the number of days they have spent in the sun. Before getting exposure, they are 30 cm. A plant grows 1 mm (0.1 cm) after being exposed to the sun for a day.

- Y is the height of the plants
- X is the number of days spent in the sun
- β_0 is 30 because it is the value of Y when X is 0.
- β_1 is 0.1 because it is the coefficient multiplied by the number of days.

A plant being exposed 5 days to the sun has therefore an estimated height of $Y = 30 + 0.1 \cdot 5 = 30.5$ cm.

Of course, it is not always exact, which is why we must take into account the random error ε .

Moreover, before predicting, our method has to find the β coefficients: we just start out by inputting a table containing heights of several plants along with the number of days they have spent in the sun.

How does the method work?

The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values.

That way, the vector β of the coefficients can be estimated by the following formula

$$\beta = (X'DX)^{-1} X'Dy$$

with X the matrix of the explanatory variables preceded by a vector of 1s, D is a matrix with the w_i weights on its diagonal, and y the vector of the n observed values of the dependent variable

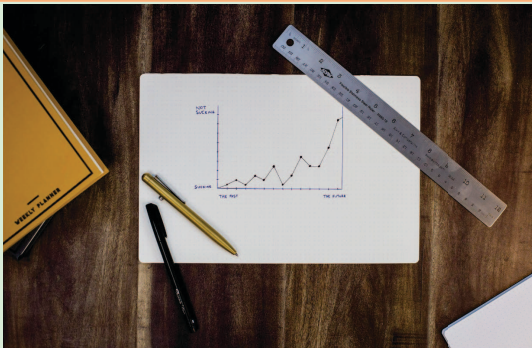
The vector of the predicted values can be written as follows:

$$y^* = X\beta = X(X'DX)^{-1} X'Dy$$

We can even the variance σ^2 of the random error ε by the following formula:

$$\sigma^2 = 1/(W - p^*) \sum_{i=1..n} w_i(y_i - y^*_i)$$

where p^* is the number of explanatory variables to which we add 1 if the intercept is not fixed, w_i is the weight of the i th observation, W is the sum of the w_i weights, y the $*$ vector of the observed values and y^* the vector of predicted values.



CORE CONCEPT: THE MODEL BEHIND LINEAR REGRESSION

The model behind linear regression is a mathematical representation that captures the relationship between the independent variable(s) and the dependent variable. In the case of simple linear regression, the

model assumes a linear relationship between a single independent variable (x) and a single dependent variable (y).

When we are examining the relationship between a quantitative outcome and a single quantitative explanatory variable, simple linear regression is the most commonly considered analysis method. (The “simple” part tells us we are only considering a single explanatory variable.) In linear regression we usually have many different values of the explanatory variable, and we usually assume that values between the observed values of the explanatory variables are also possible values of the explanatory variables. We postulate a linear relationship between the population mean of the outcome and the value of the explanatory variable. If we let Y be some outcome, and x be some **explanatory variable**, then we can express the structural model using the equation

Explanatory variable is what you manipulate or observe changes in, while a response variable is what changes as a result.

$$E(Y|x) = \beta_0 + \beta_1 x$$

where $E()$, which is read “expected value of”, indicates a population mean; $Y|x$, which is read “Y given x”, indicates that we are looking at the possible values of Y when x is restricted to some single value; β_0 , read “beta zero”, is the intercept parameter; and β_1 , read “beta one”, is the slope parameter. A common term for any parameter or parameter estimate used in an equation for predicting Y from x is coefficient. Often the “1” subscript in β_1 is replaced by the name of the explanatory variable or some abbreviation of it.

The structural model underlying a linear regression analysis is that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is $\beta_0 + \beta_1 x$.

So the structural model says that for each value of x the population mean of Y (over all of the subjects who have that particular value “x” for their explanatory variable) can be calculated using the simple linear expression $\beta_0 + \beta_1 x$. Of course we cannot make the calculation exactly, in practice, because the two parameters are unknown “secrets of nature”. In practice, we make estimates of the parameters and substitute the estimates into the equation.

In real life we know that although the equation makes a prediction of the true mean of the outcome for any fixed value of the explanatory variable, it would be unwise to use extrapolation to make predictions outside of the range of x values that we have available for study. On the other hand it is reasonable to interpolate, i.e., to make predictions for unobserved x values in between the observed x values. The structural model is essentially the assumption of “linearity”, at least within the range of the observed explanatory data.

It is important to realize that the “linear” in “linear regression” does not imply that only linear relationships can be studied. Technically it only says that the beta’s must not be in a transformed form. It is OK to transform x or Y, and that allows many non-linear relationships to be represented on a new scale that makes the relationship linear.

The error model that we use is that for each particular x, if we have or could collect many subjects with that x value, their distribution around the population mean is Gaussian with a spread, say σ^2 , that is the same value for each value of x (and corresponding population mean of y). Of course, the value of σ^2 is an unknown parameter, and we can make an estimate of it from the data. The error model described so far includes not only the assumptions of “Normality” and “equal variance”, but also the assumption of “fixed-x”. The “fixed-x” assumption is that the explanatory variable is measured without error. Sometimes this is possible, e.g., if it is a count, such as the number of legs on an insect, but usually there is some error in the measurement of the explanatory variable. In practice, we need to be sure that the size of the error in measuring x is small compared to the variability of Y at any given x value. In addition to the three error model assumptions just discussed, we also assume “independent errors”. This assumption comes down to the idea that the error (deviation of the true

outcome value from the population mean of the outcome for a given x value) for one observational unit (usually a subject) is not predictable from knowledge of the error for another observational unit. For example, in predicting time to complete a task from the dose of a drug suspected to affect that time, knowing that the first subject took 3 seconds longer than the mean of all possible subjects with the same dose should not tell us anything about how far the next subject's time should be above or below the mean for their dose. This assumption can be trivially violated if we happen to have a set of identical twins in the study, in which case it seems likely that if one twin has an outcome that is below the mean for their assigned dose, then the other twin will also have an outcome that is below the mean for their assigned dose (whether the doses are the same or different).

A more interesting cause of correlated errors is when subjects are trained in groups, and the different trainers have important individual differences that affect the trainees performance. Then knowing that a particular subject does better than average gives us reason to believe that most of the other subjects in the same group will probably perform better than average because the trainer was probably better than average.

Another important example of non-independent errors is serial correlation in which the errors of adjacent observations are similar. This includes adjacency in both time and space. For example, if we are studying the effects of fertilizer on plant growth, then similar soil, water, and lighting conditions would tend to make the errors of adjacent plants more similar. In many task-oriented experiments, if we allow each subject to observe the previous subject perform the task which is measured as the outcome, this is likely to induce **serial correlation**. And worst of all, if you use the same subject for every observation, just changing the explanatory variable each time, serial correlation is extremely likely. Breaking the assumption of independent errors does not indicate that no analysis is possible, only that linear regression is an inappropriate analysis. Other methods such as time series methods or mixed models are appropriate when errors are correlated.

Serial correlation is the relationship between a given variable and a lagged version of itself over various time intervals.

Before going into the details of linear regression, it is worth thinking about the variable types for the explanatory and outcome variables and the relationship of ANOVA to linear regression. For both ANOVA and linear regression we assume a Normal distribution of the outcome for each value

of the explanatory variable. (It is equivalent to say that all of the errors are Normally distributed.) Implicitly this indicates that the outcome should be a continuous quantitative variable. Practically speaking, real measurements are rounded and therefore some of their continuous nature is not available to us. If we round too much, the variable is essentially discrete and, with too much rounding, can no longer be approximated by the smooth Gaussian curve. Fortunately regression and ANOVA are both quite robust to deviations from the Normality assumption, and it is OK to use discrete or continuous outcomes that have at least a moderate number of different values, e.g., 10 or more. It can even be reasonable in some circumstances to use regression or ANOVA when the outcome is ordinal with a fairly small number of levels.

The explanatory variable in ANOVA is categorical and nominal. Imagine we are studying the effects of a drug on some outcome and we first do an experiment comparing control (no drug) vs. drug (at a particular concentration). Regression and ANOVA would give equivalent conclusions about the effect of drug on the outcome, but regression seems inappropriate. Two related reasons are that there is no way to check the appropriateness of the linearity assumption, and that after a regression analysis it is appropriate to interpolate between the x (dose) values, and that is inappropriate here.

Now consider another experiment with 0, 50 and 100 mg of drug. Now ANOVA and regression give different answers because ANOVA makes no assumptions about the relationships of the three population means, but regression assumes a linear relationship. If the truth is linearity, the regression will have a bit more power than ANOVA.

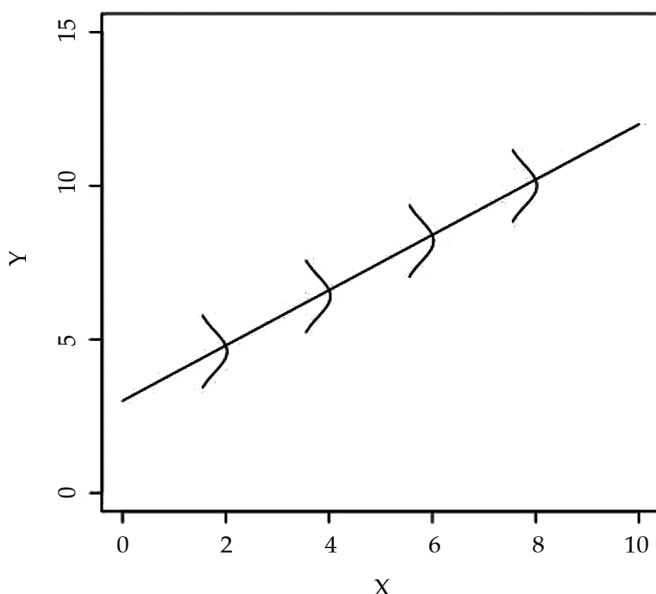
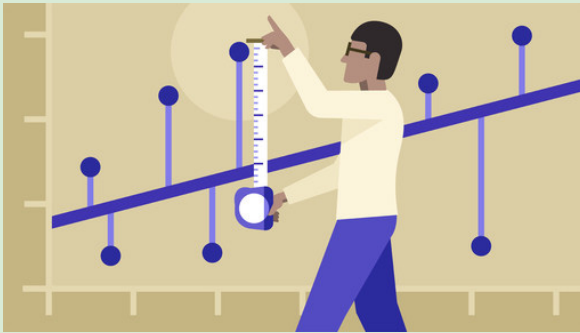


Figure 1. Mnemonic for the simple regression model.

If the truth is non-linearity, regression will make inappropriate predictions, but at least regression will have a chance to detect the non-linearity. ANOVA also loses

some power because it incorrectly treats the doses as nominal when they are at least ordinal. As the number of doses increases, it is more and more appropriate to use regression instead of ANOVA, and we will be able to better detect any non-linearity and correct for it, e.g., with a data transformation. Figure 1 shows a way to think about and remember most of the regression model assumptions. The four little Normal curves represent the Normally distributed outcomes (Y values) at each of four fixed x values. The fact that the four Normal curves have the same spreads represents the equal variance assumption. And the fact that the four means of the Normal curves fall along a straight line represents the linearity assumption. Only the fifth assumption of independent errors is not shown on this mnemonic plot.

The worst case of breaking the independent errors assumption in regression is when the observations are repeated measurement on the same experimental unit (subject).



CORE CONCEPT: INTERPRETING REGRESSION COEFFICIENTS

Interpreting regression coefficients is a crucial aspect of understanding the relationship between variables in a

linear regression model. In the context of simple linear regression ($y = \beta_0 + \beta_1 x + \epsilon$), the two coefficients of interest are β_0 (the intercept) and β_1 (the slope).

It is very important that you learn to correctly and completely interpret the coefficient estimates. From $E(Y|x) = \beta_0 + \beta_1 x$ we can see that b_0 represents our estimate of the mean outcome when $x = 0$. Before making an interpretation of b_0 , first check the range of x values covered by the experimental data. If there is no x data near zero, then the intercept is still needed for calculating $\hat{y}$ and residual values, but it should not be interpreted because it is an extrapolated value.

If there are x values near zero, then to interpret the intercept you must express it in terms of the actual meanings of the outcome and explanatory variables. We would say that b_0 is the estimated corn plant weight (in grams) when no nitrogen is added to the pots (which is the meaning of $x = 0$). This point estimate is of limited value, because it does not express the degree of uncertainty associated with it. So often it is better to use the CI for b_0 . In this case we say that we are 95% confident that the mean

weight for corn plants with no added nitrogen is between 47 and 122 gm, which is quite a wide range. (It would be quite misleading to report the mean no-nitrogen plant weight as 94.821 gm because it gives a false impression of high precision.)

After interpreting the estimate of b_0 and its CI, you should consider whether the null hypothesis, $\beta_0 = 0$ makes scientific sense. For the corn example, the null hypothesis is that the mean plant weight equals zero when no nitrogen is added. Because it is unreasonable for plants to weigh nothing, we should stop here and not

interpret the p-value for the intercept. For another example, consider a regression of weight gain in rats over a 6 week period as it relates to dose of an anabolic steroid. Because we might be unsure whether the rats were initially at a stable weight, it might make sense to test $H_0: \beta_0 = 0$. If the null hypothesis is rejected then we conclude that it is not true that the weight gain is zero when the dose is zero (control group), so the initial weight was not a stable baseline weight.

For interpretation of a slope coefficient, this section will assume that the setting is a randomized experiment, and conclusions will be expressed in terms of causation. Be sure to substitute association if you are looking at an observational study. The general meaning of a slope coefficient is the change in Y caused by a one-unit increase in x . It is very important to know in what units x are measured, so that the meaning of a one-unit increase can be clearly expressed. For the corn experiment, the slope is the change in mean corn plant weight (in grams) caused by a one mg increase in nitrogen added per pot. If a one-unit change is not substantively meaningful, the effect of a larger change should be used in the interpretation. For the corn example we could say the a 10 mg increase in nitrogen added causes a 52.7 gram increase in plant weight on average. We can also interpret the CI for β_1 in the corn

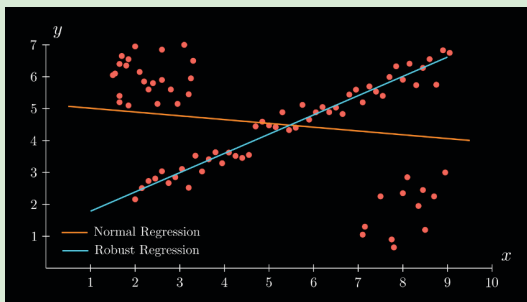
Interpret the estimate, b_0 , only if there are data near zero and setting the explanatory variable to zero makes scientific sense. The meaning of b_0 is the estimate of the mean outcome when $x = 0$, and should always be stated in terms of the actual variables of the study. The pvalue for the intercept should be interpreted (with respect to retaining or rejecting $H_0: \beta_0 = 0$) only if both the equality and the inequality of the mean outcome to zero when the explanatory variable is zero are scientifically plausible.

Absolute value means to remove any negative sign in front of a number, and to think of all numbers as positive (or zero).

experiment by saying that we are 95% confident that the change in mean plant weight caused by a 10 mg increase in nitrogen is 46.8 to 58.9 gm.

Be sure to pay attention to the sign of b_1 . If it is positive then b_1 represents the increase in outcome caused by each one-unit increase in the explanatory variable. If b_1 is negative, then each one-unit increase in the explanatory variable is associated with a fall in outcome of magnitude equal to the **absolute value** of b_1 .

A significant p-value indicates that we should reject the null hypothesis that $\beta_1 = 0$. We can express this as evidence that plant weight is affected by changes in nitrogen added. If the null hypothesis is retained, we should express this as having no good evidence that nitrogen added affects plant weight. Particularly in the case of when we retain the null hypothesis, the interpretation of the CI for β_1 is better than simply relying on the general meaning of retain.



CORE CONCEPT: ROBUSTNESS OF SIMPLE LINEAR REGRESSION

The robustness of a statistical method, including simple linear regression, refers to its ability to provide reliable and accurate results even in the presence of violations of its underlying assumptions or the influence of outliers. Robust methods are less affected by extreme data points that could otherwise have a disproportionate impact on the results.

Robust regression techniques aim to provide more stable and accurate parameter estimates even when assumptions are violated or outliers are present. However, it's important to note that while robust methods can mitigate some issues, they might not be a one-size-fits-all solution and have their limitations as well. The choice of method depends on the specific characteristics of the data and the goals of the analysis.

No model perfectly represents the real world. It is worth learning how far we can “bend” the assumptions without breaking the value of a regression analysis. If the linearity assumption is violated more than a fairly small amount, the regression loses its meaning. The most obvious way this happens is in the interpretation of b_1 . We interpret b_1 as the change in the mean of Y for a one-unit increase in x . If the relationship between x and Y is curved, then the change in Y for a one-unit increase in x varies at different parts of the curve, invalidating the interpretation. Luckily it is fairly easy to detect non-linearity through EDA (scatterplots) and/or residual analysis.

If non-linearity is detected, you should try to fix it by transforming the x and/or y variables. Common transformations are log and square root. Alternatively it is common to add additional new explanatory variables in the form of a square, cube, etc. of the original x variable one at a time until the residual vs. fit plot shows linearity of the residuals. For data that can only lie between 0 and 1, it is worth knowing (but not memorizing) that the square root of the arcsine of y is often a good transformation.

Regression analysis is not very robust to violations of the linearity, fixed- x , and independent errors assumptions. It is somewhat robust to violation of equal variance, and moderately robust to violation of the Normality assumption.

You should not feel that transformations are “cheating”. The original way the data is measured usually has some degree of arbitrariness. Also, common measurements like pH for acidity, decibels for sound, and the Richter earthquake scale are all log scales. Often transformed values are transformed back to the original scale when results are reported.

Regression is reasonably robust to the equal variance assumption. Moderate degrees of violation, e.g., the band with the widest variation is up to twice as wide as the band with the smallest variation, tend to cause minimal problems. For more severe violations, the p -values are incorrect in the sense that their null hypotheses tend to be rejected more than $100\alpha\%$ of the time when the null hypothesis is true. The confidence intervals (and the SE's they are based on) are also incorrect. For worrisome violations of the equal variance assumption, try transformations of the y variable (because the assumption applies at each x value, transformation of x will be ineffective).

Regression is quite robust to the Normality assumption. You only need to worry about severe violations. For markedly skewed or kurtotic residual distributions, we need to worry that the p -values and confidence intervals are incorrect. In that case try transforming the y variable. Also, in the case of data with less than a handful of different y values or with severe truncation of the data (values piling up at the ends of a limited width scale), regression may be inappropriate due to non-Normality.

The fixed- x assumption is actually quite important for regression. If the variability of the x measurement is of similar or larger magnitude to the variability of the y measurement, then regression is inappropriate. Regression will tend to give smaller than correct slopes under these conditions, and the null hypothesis on the slope will be retained far too often. Alternate techniques are required if the fixed- x assumption is broken, including so-called Type 2 regression or “errors in variables regression”.

The independent errors assumption is also critically important to regression. A slight violation, such as a few twins in the study doesn't matter, but other mild to moderate violations destroy the validity of the p -value and confidence intervals. In that case, use alternate techniques such as the paired t -test, repeated measures analysis, mixed models, or time series analysis, all of which model correlated errors rather than assume zero correlation.

Fixed Versus Random Regressors

A regressor might be “fixed,” as when we fix the temperature in a lab experiment, or it may be random such as when a regression model is used to predict college grade point average and one of the regressors is high school GPA. Obviously, the latter is random, not fixed, and probably most of the applications of regression that we can think of would have one or more random regressors. The random regressor case has been somewhat controversial over the years, with various methods of analysis proposed. The usual message to practitioners, however, is that in the random X case we can proceed the same as in the fixed X case provided that the conditional distribution of Y given X is (approximately) normal for each value of X . (Conditional distributions are not covered but here the meaning of the term should be apparent.)

Transformations

Transformations can be very useful in regression. There are two general types of transformations:

(1). Transforming the model and (2) transforming Y and/or X .

Transforming the Model

Assume that we “know” that the true model is $Y = \beta_0 \beta_1^{x_1} \epsilon$. This is not a linear regression model because it is not linear in the parameters. If we take the natural logarithm of each side of the model, we then obtain a linear regression model:

$$Y' = \beta'_0 + \beta'_1 X_1 + \epsilon$$

with

$$Y' = \log(Y), \beta'_0 = \log(\beta_0), \beta'_1 = \log(\beta_1), \text{ and } \epsilon' = \log(\epsilon)$$

Although this is straightforward mathematically, we need to know the error structure of the model and to know that it is as specified here in order for this approach to work. If the error term had been additive instead of multiplicative, then taking logs would not work. In that case we would have a nonlinear regression model. Such models are not covered. Oftentimes knowledge about the form of the error term will not be available or will vary from application to application even when the same nonlinear model is used.

Transforming Y and/or X

We never want to transform Y to try to improve the fit because if the assumptions were approximately met before the transformation, they will not be met after the transformation. Of course, we might want to transform Y in an effort to satisfy the assumptions. If we transform Y and want to see how good the fit is, we need to transform the values of $\hat{Y}$ back to the original scale before computing R^2 . If we have transformed Y in order to meet the model assumptions, the fit of the model may

Scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

have seriously been degraded. In order to regain the quality of the fit, it is necessary to either transform the entire right-hand side of the model using the same transformation as was employed for Y , or transform X in that manner. That is, if Y^λ is the transformation, with $\log(Y)$ used for $\lambda = 0$ [this family of transformations is due to Box and Cox], the choices for the right-hand side of the model are $(\beta_0 + \beta_1 X)^\lambda + \hat{1}$ or $\beta_0 + \beta_1 X^\lambda + \hat{1}$. The former, which is called the transform both-side's method, is preferable as that preserves the functional relationship. The other form has the advantage of simplicity, however, since it is a simple linear regression model (with a nonlinear term in X), whereas the other form is a nonlinear regression model. Nonlinear regression is not covered, but as mentioned there are various nonlinear regression texts, including Bates and Watts.

Frequently, a nonlinear term in X will be needed, either in place of the linear term or in addition to it. *For example*, the scatter plot may indicate the need for a quadratic term. Such a term would generally be used in addition to the linear term, rather than as a replacement for it. So if we are fitting polynomial terms, we usually want to include the lower-order terms, as justifying their exclusion could be difficult. If the scatter plot suggests a logarithmic or reciprocal relationship, however, such a term should be used without the linear term. A plot of the standardized residuals against each predictor is one of the plots used in multiple regression (i.e., more than one predictor) to check for the need to transform a predictor, but in simple linear regression all of the information needed to make that determination is contained in the **scatter plot**.

Prediction Intervals and Confidence Intervals

Since one of the uses of regression is for prediction, it would be logical to construct an interval about a predicted value, and we naturally call the interval a prediction interval. A prediction interval in regression will be wider than the prediction interval given by equation because in regression there is the added variability of $\hat{\beta}_1$, which is of course not involved in constructing a prediction interval that is not constructed using a regression model.

A confidence interval for a parameter θ is of the general form $\hat{\theta} \pm ts \hat{\sigma}$ whenever the use of t is appropriate. A prediction interval for Y is not an interval for a parameter, however, and will thus not be of that form. We will lead into the interval by

first going through a somewhat similar interval, as that will make the transition easier. *For example*, assume that we wish to predict college grade point average using high school grade point average as the sole predictor. Thus, Y = college GPA and X = high school GPA. Would a confidence interval for the average GPA of all students who have entered a certain college with a particular value of X and subsequently graduated be of any value? Clearly, it is of no value if we are trying to reach a decision regarding the admittance of a particular student. What we need in that case is an interval for the student's GPA if the student were admitted and attended for four years or at least until possibly encountering insurmountable academic problems.

The most intuitive way to obtain the expression for the prediction interval is to first obtain the expression for the confidence interval.

For the confidence interval, we want an interval for $\mu_{Y|X_0}$. In words, this is the mean of Y when $X = X_0$, which would also be called the conditional mean. Since this mean is our parameter θ , the next step is to determine $\hat{\theta}$. Because the mean of $\hat{I}$ is zero, it follows that the conditional mean is equal to $\beta_0 + \beta_1 X_0$. This of course would be estimated by $\hat{Y}_0$, the predicted value of Y when $X = X_0$, as clearly $E(\hat{Y}_0) = \mu_{Y|X_0}$ since the least squares estimators are unbiased. We then need $S_{\hat{Y}_0}$. Obtaining the latter requires a moderate amount of algebra, which will not be given here. The result is

$$S_{\hat{Y}_0} = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \text{ with } s^2 = \hat{\sigma}_e^2 = \sum \frac{(Y - \hat{Y})^2}{(n-2)}$$

Thus, the $100(1-\alpha)\%$ confidence interval is given by

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

It is relatively straightforward to go from the confidence interval expression to the expression for the prediction interval. The latter, of course, is to be constructed about Y_0 since this is the predicted value, so this is common to both expressions. That should not be surprising because the only difference between the model expression for $Y_0 = \beta_0 + \beta_1 X_0 + \hat{I}$ and $\mu_{Y|X_0}$ is a random error term that has a mean of zero and thus is not a term that can contribute to either estimation or prediction.

Since we are trying to predict a random variable, Y_0 , rather than estimate a parameter, $\mu_{Y|X_0}$, we must account for the variability of the random variable in constructing the interval. Since $\text{Var}(\hat{I}) = \text{Var}(Y_0)$, it follows that we must add σ_e^2 to $\sigma_{\hat{Y}_0}^2$. That is, we want $\text{Var}(Y_0 - \hat{Y}_0) = \text{Var}(Y_0) + \text{Var}(\hat{Y}_0)$, with the variances being added because a new observation is independent of the data used in producing the regression equation, and thus independent of $\hat{Y}_0$. The estimator of $\text{Var}(\hat{Y}_0)$ is of course the same as is used for the confidence interval and the estimator of $\text{Var}(\hat{Y}_0)$ is s^2 . Therefore, using these estimators and taking the square root, we obtain the prediction interval:

$$\hat{Y}_0 \pm t_{\alpha/2, n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Although the expressions are similar mathematically, conceptually they are quite different since a prediction interval is for the future and a confidence interval is for the present. It is useful to know what the expression is for the prediction interval, but we do not want to obtain a prediction interval using hand calculation, at least not more than once. We can easily obtain a prediction interval in MINITAB, *for example*, and if we obtain a 95% prediction interval for the SAT example with $X_0 = 1200$, which must be for a school not included in the sample, the output is as follows.

(Recall that X_0 denotes the 25th percentile SAT score.)

Predicted Value for New Observation

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	1395.33	2.75	(1389.79, 1400.87)	(1358.54, 1432.12)

The first interval is the confidence interval and the second is the prediction interval. Of course, the latter is always wider and we observe that here. Note that the observed value of 1450 for one of the sample data points lies outside the prediction interval. Why is that? This result should not be surprising because we picked a point that did not fit well with the other points. Therefore, the fitted value, in this case 1395.33, is not close to the actual value, so the amount that is added to (or subtracted from) the fitted value will not necessarily cover the actual value, especially when s is small. The prediction interval might still work for new data with $X_0 = 1200$, however, and of course a prediction interval is used for new X values. This illustrates the importance of using regression diagnostics, as they are called, rather than just taking a perfunctory approach.

To give a “better,” but less profound, example, observation is (1300, 1480), so when we produce the prediction interval using this X -value, we obtain predicted value for new observation.

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	1480.09	3.12	(1473.82, 1486.36)	(1443.18, 1517.00)

Thus, the actual value is virtually in the middle of the prediction interval because the predicted value is virtually the same as the observed value. There are many areas of application in which prediction intervals are extremely important. Consider again the scenario with Y = college GPA and X = high school GPA and consider two possible prediction intervals given by (1.8, 2.6) and (2.1, 2.3). The value of $\hat{Y}$ is 2.2 in each case (since $\hat{Y}$ lies in the center of the interval), but an admissions officer would undoubtedly look more favorably upon the second interval than the first interval. In other areas of application, such as fisheries management, having a well-estimated prediction interval is thought to be probably as important as having good estimates of the regression parameters.

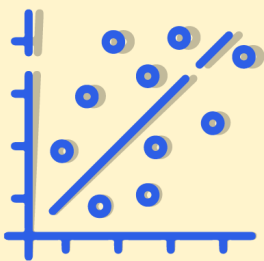
Another interval that is potentially useful is a confidence interval on β_1 . The signal we receive from a confidence interval agrees with the signal that we receive from the corresponding hypothesis test. Therefore, we know that the confidence interval will not come close to containing zero because the p-value for the hypothesis test is zero to three decimal places. Since $\hat{\beta}_1 = 0.84761$ and $S_{\hat{\beta}_1} = 0.02699$, a 95% confidence interval for $\hat{\beta}_1$ would

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} s_{\hat{\beta}_1} = 0.84761 \pm 2.0117(0.02699)$$

which gives us the interval (0.7933, 0.9019). Thus, the interval does not come close to containing zero, and as stated we knew that would happen before we constructed the interval. There is an important point to be made before we leave this example. Although it is quite reasonable to construct a confidence interval for β_1 when there is only a single regressor, we should construct confidence intervals for the β_i in the case of more than one regressor only when the X's are fixed.

Model Validation

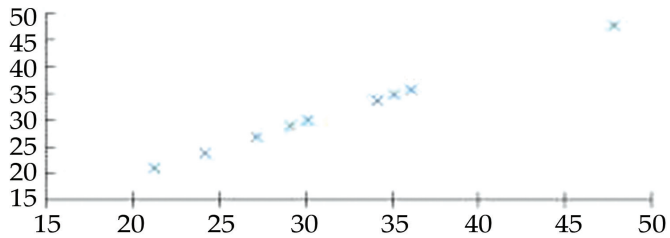
Bearing in mind the sequential, repetitive nature of experimentation. A regression model should be “validated” before it is used. That is, its performance should be assessed on data other than the data used to develop the model and to estimate the parameters. So how is model validation accomplished? One approach that has been used is to split the data in half, making the two halves approximately equal in terms of a selected similarity measure, and use the second half for validation. An argument against such practice has effectively been made in the literature, however: the loss of efficiency caused by using only half the data for parameter estimation militates against the use of data splitting. The other alternative is to obtain new data, if possible and practical, and test model performance on the new data.



CORE CONCEPT: CORRELATION

Correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It provides insight into how changes in one variable are associated with changes in another variable. Correlation is widely used to describe and understand relationships between variables in various fields such as statistics, economics, social sciences, and more.

The relationships graphed in Figure can first be summarized numerically by measuring the correlation coefficient between any pair of variables. We illustrate this by calculating the correlation coefficient between the birth rate (B) and growth (G), though we also present the results for the other cases. Just as the mean is a number that summarizes information about a single variable, so the correlation coefficient is a number which summarizes the relationship between two variables.



The different types of possible relationship between any two variables X and Y may be summarized as follows:

- High values of X tend to be associated with low values of Y and vice versa. This is termed negative correlation, and appears to be the case for B and G.
- High (low) values of X tend to be associated with high (low) values of Y. This is positive correlation and reflects (rather weakly) the relationship between B and the income ratio (IR).
- No relationship between X and Y exists. High (low) values of X are associated about equally with high and low values of Y. This is zero, or the absence of, correlation. There appears to be little correlation between the birth rate and per capita GNP.

It should be noted that positive correlation does not mean that high values of X are always associated with high values of Y, but usually they are.

It is also the case that correlation only represents a linear relationship between the two variables. As a counter-example, consider the backward-bending labor supply curve, as suggested by economic theory (higher wages initially encourage extra work effort, but above a certain point the benefit of higher wage rates is taken in the form of more leisure).

Correlation refers to a process for establishing the relationships between two variables.

The relationship is non-linear and the measured degree of **correlation** between wages and hours of work is likely to be low, even though the former obviously influences the latter. The sample correlation coefficient, r , is a numerical statistic

which distinguishes between the types of cases. It has the following properties:

- It always lies between -1 and $+1$.
- A positive value of r indicates positive correlation, a higher value indicating a stronger correlation between X and Y (i.e. the observations lie closer to a straight line). $r = 1$ indicates perfect positive correlation and means that all the observations lie precisely on a straight line with positive slope.
- A negative value of r indicates negative correlation. Similar to the above, a larger negative value indicates stronger negative correlation and $r = -1$ signifies perfect negative correlation.
- A value of $r = 0$ (or close to it) indicates a lack of correlation between X and Y .
- The relationship is symmetric, i.e. the correlation between X and Y is the same as between Y and X . It does not matter which variable is labeled Y and which is labeled X .

Country	Birth Rate Y	GNP Growth X	y^2	x^2	XY
Brazil	30	5.1	900	26.01	153.0
Colombia	29	3.2	841	10.24	92.8
Costa Rica	30	3.0	900	9.00	90.0
India	35	1.4	1225	1.96	49.0
Mexico	36	3.8	1296	14.44	136.8
Peru	36	1.0	1296	1.00	36.0
Philippines	34	2.8	1,156	7.84	95.2
Senegal	48	-0.3	2304	0.09	-14.4
South Korea	24	6.9	576	47.61	165.6
Sri Lanka	27	2.5	729	6.25	67.5
Taiwan	21	6.2	441	38.44	130.2
Thailand	30	4.6	900	21.16	138.0
Totals	380	40.2	12564	184.04	1139.7

The formula for calculating the correlation coefficient is given in equation:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

The calculation of r for the relationship between birth rate (Y) and growth (X) is shown in Table and equation. From the totals in Table we calculate

$$r = \frac{12 \times 1139.7 - 40.2 \times 380}{\sqrt{(12 \times 184.04 - 40.2^2)(12 \times 12\,564 - 380^2)}} = -0.824$$

This result indicates a fairly strong negative correlation between the birth rate and growth. Countries which have higher economic growth rates also tend to have lower birth rates. The result of calculating the correlation coefficient for the case of the birth rate and the income ratio is $r = 0.35$, which is positive as expected. Greater inequality (higher IR) is associated with a higher birth rate, though the degree of correlation is not particularly strong and less than the correlation with the growth rate. Between the birth rate and GNP per capita the value of r is only 0.26 indicating only a modest degree of correlation. All of this begins to cast doubt upon Todaro's interpretation of the data.

Correlation and Uses of Regression

A regression analysis can be performed when X is either fixed or random, but the correlation between two variables is defined only when they are both random. This convention is not strictly adhered to; however, as the term "correlation" is often used in conjunction with designed experiments in which the regressors are fixed. A correlation coefficient, such as r_{xy} , measures the extent to which the two random variables are related. The sample (Pearson) correlation coefficient is computed as

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (17)$$

with S_{yy} defined analogous to the way that S_{xx} was defined, and the other components of equation (17) are as defined. Formally, the population correlation coefficient, r_{xy} , is defined as the covariance between X and Y divided by the standard deviation of X times the standard deviation of Y . When those parameters are estimated by the sample statistics, equation (17) results.

Miscellaneous Uses of Regression

There are some important specialized uses of regression, one of which occurs extensively in engineering. Assume that we have two measuring instruments; one is accurate to the intended level of accuracy but is both expensive to use and slow, and the other is fast and less expensive to use but is inaccurate. If the measurements obtained from the two devices were highly correlated, the measurement that would have been made using the expensive measuring device could be predicted fairly well from the measurement that is actually obtained using the less expensive device.

Calibration

There are two approaches to calibration: inverse regression and the classical theory of calibration. If the less expensive device has an almost constant bias and we define X = measurement from the accurate instrument and Y = measurement from the inaccurate

device, and then regress X on Y to obtain $\hat{X}^* = \hat{\beta}_0^* + \hat{\beta}_1^* Y$ we would expect $\hat{\beta}_1^*$ to be close to 1.0 and $\hat{\beta}_0^*$ to be approximately equal to the bias.

Measurement Error

Measurement error was involved in the calibration example discussed, in which the objective was to work around the error using calibration and thus eliminate its effect. Measurement error is generally not incorporated into regression models, however, and only within the past ten years or so have statisticians and scientists given much attention to the effect of measurement errors. If measurement error in Y is random, normally distributed, and has a mean of zero, it adds to the model error (if the measurement error and model error are independent), leaves more variability in Y unexplained by the model, and thus reduces the value of R^2 . If the measurement error is large relative to the random errors and is markedly non-normal, an alternative to least squares may have to be used.

The effect of measurement error in X will depend on the nature of the error and also on whether or not X is a random variable. In general, it is desirable to have the spread in X large relative to measurement error in X whenever the latter is anticipated.

Regression for Control

An important but infrequently discussed application of regression is to attempt to control Y at a desired level through the manipulation of X . Doing so may be difficult, however, for a number of reasons. First, since a cause-and effect relationship is being inferred, the prediction equation must have been produced from preselected X values, and there must not be any other independent variables that are related to Y . The latter is not likely to be true when only a single X is being used, however, and the consequences when it is not true may be great.

Box gave a good example of this in regard to a hypothetical chemical process in which Y = process yield, X_1 = pressure, and X_2 = an unsuspected impurity. The scenario is that undesirable frothing can be reduced by increasing pressure, but a high value of X_2 is what actually produced frothing and also lowers yield, with the latter unrelated to pressure. Thus, although regression for control is potentially useful, there is also a high risk.

Correlation Coefficient

Correlation Coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment. The calculated value of the correlation coefficient explains the exactness between the predicted and actual values.

Correlation Coefficient value always lies between -1 to +1. If correlation coefficient value is positive, then there is a similar and identical relation between the two variables. Else it indicates the dissimilarity between the two variables.

The covariance of two variables divided by the product of their standard deviations gives Pearson's correlation coefficient. It is usually represented by ρ (rho).

$$\rho(X, Y) = \text{cov}(X, Y) / \sigma_X \cdot \sigma_Y.$$

Here cov is the covariance. σ_X is the standard deviation of X and σ_Y is the standard deviation of Y. The given equation for correlation coefficient can be expressed in terms of means and expectations.

$$\rho(X, Y) = E \frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x \cdot \sigma_y}$$

μ_x and μ_y are mean of x and mean of y respectively. E is the expectation.

Assumptions of Karl Pearson's Correlation Coefficient

The assumptions and requirements for calculating Pearson's correlation coefficient are as follows:

- The data set which is to be correlated should approximate to the normal distribution. If the data is normally distributed, then the data points tend to lie closer to the mean.
- The word homoscedastic is a greek originated meaning 'able to disperse'. Homoscedasticity means 'equal variances'. For all the values of the independent variable, the error term is the same. Suppose the error term is smaller for a certain set of values of independent variable and larger for another set of values, then homoscedasticity is violated. It can be checked visually through a scatter plot. The data is said to be homoscedastic if the points lie equally on both sides of the line of best fit.
- When the data follows a linear relationship, it is said to be linearity. If the data points are in the form of a straight line on the scatter plot, then the data satisfies the condition of linearity.
- The variables which can take any value in an interval are continuous variables. The data set must contain continuous variables to compute the Pearson correlation coefficient. If one of the data sets is ordinal, then Spearman's rank correlation is an appropriate measure.
- The data points must be in pairs which are termed as paired observations. There exists a dependent variable for every observation of the independent variable.
- There must be no outliers in the data. If the outliers are present, then they can skew the correlation coefficient and make it inappropriate. A point is considered to be an outlier if it is beyond +3.29 or -3.29 standard deviations away. They can be easily determined visually from a scatter plot.

Correlation Coefficient Properties

Correlation coefficient is all about establishing relationships between two variables. Some properties of correlation coefficient are as follows:

- Correlation coefficient remains in the same measurement as in which the two variables are.
- The sign which correlations of coefficient have will always be the same as the variance.
- The numerical value of correlation of coefficient will be in between -1 to +1. It is known as real number value.
- The negative value of coefficient suggests that the correlation is strong and negative. And if 'r' goes on approaching toward -1 then it means that the relationship is going towards the negative side.

When 'r' approaches to the side of +1 then it means the relationship is strong and positive. By this we can say that if +1 is the result of the correlation then the relationship is in a positive state.

- The weak correlation is signaled when the coefficient of correlation approaches to zero. When 'r' is near about zero then we can deduce that the relationship is weak.
- Correlation coefficient can be very dicey because we cannot say that the participants are truthful or not.

The coefficient of correlation is not affected when we interchange the two variables.

- Coefficient of correlation is a pure number without effect of any units on it. It also not get affected when we add the same number to all the values of one variable. We can multiply all the variables by the same positive number. It does not affect the correlation coefficient. As we discussed, 'r' is not affected by any unit because 'r' is a scale invariant.
- We use correlation for measuring the association but that does not mean we are talking about causation. By this, we simply mean that when we are correlating the two variables then it might be the possibility that the third variable may be influencing them.

Spearman Rank Correlation Coefficient

The Spearman's rank coefficient of correlation is a nonparametric measure of rank correlation (statistical dependence of ranking between two variables).

Named after Charles Spearman, it is often denoted by the Greek letter ' ρ ' (rho) and is primarily used for data analysis.

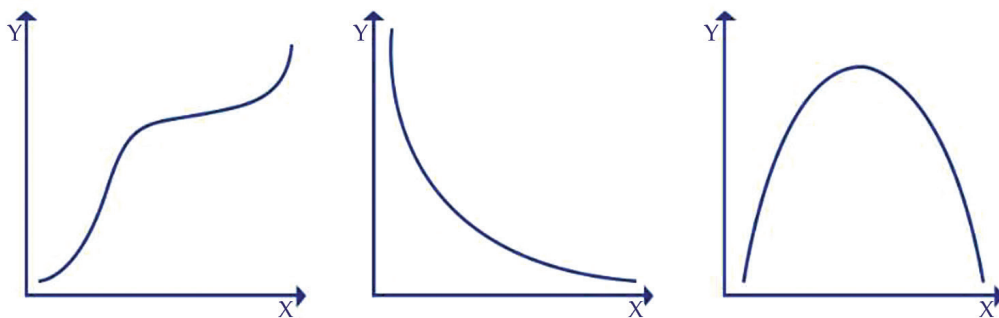
It measures the strength and direction of the association between two ranked variables. But before we talk about the Spearman correlation coefficient, it is important to understand Pearson's correlation first. A Pearson correlation is a statistical measure of the strength of a linear relationship between paired data.

For the calculation and significance testing of the ranking variable, it requires the following data assumption to hold true:

- Interval or ratio level

- Linearly related
- Bivariant distributed

If your data doesn't meet the above assumptions, then you would need Spearman's Coefficient. It is necessary to know what monotonic function is to understand Spearman correlation coefficient. A monotonic function is one that either never decreases or never increases as it is an independent variable increase. A monotonic function can be explained using the figure below:



Three concepts in monotonic function:

- Monotonically increasing: When the 'x' variable increases and the 'y' variable never decreases.
- Monotonically decreasing: When the 'x' variable increases but the 'y' variable never increases
- Not monotonic: When the 'x' variable increases and the 'y' variable sometimes increases and sometimes decreases.

Monotonic relation is less restrictive when compared to a linear relationship that is used in Pearson's coefficient. Although monotonicity is not the ultimate requirement for Spearman correlation coefficient, it will not be meaningful to pursue Spearman's correlation without actually determining the strength and direction of a monotonic relationship if it was already known that the relationship between the variable is non-monotonic.

Spearman correlation coefficient: Formula and Calculation with Example

$$rR = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

Here,

n= number of data points of the two variables

di= difference in ranks of the "ith" element

The Spearman Coefficient, ρ , can take a value between +1 to -1 where,

A ρ value of +1 means a perfect association of rank

A ρ value of 0 means no association of ranks

A ρ value of -1 means a perfect negative association between ranks.

Closer the ρ value to 0, weaker is the association between the two ranks.

We must be able to rank the data before proceeding with the Spearman's Rank Coefficient of Correlation. It is important to observe if increasing one variable, the other variable follows a monotonic relation.

At every level, you will need to compare the values of the two variables. Here is how the calculations work:

Step 1- Create a table of the data obtained.

Step 2- Start by ranking the two data sets. Data ranking can be achieved by assigning the ranking "1" to the biggest number in the column, "2" to the second biggest number and so forth. The smallest value will usually get the lowest ranking. This should be done for both sets of measurements.

Step 3- Add a third column d to your data set, d here denotes the difference between ranks. For example, if the first student's physics rank is 3 and the math rank is 5 then the difference in the rank is 3. In the fourth column, square your d values.

History	Rank	Geography	Rank	d	d square
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0
					12

Step 4- Add up all your d square values, which is 12 ($\sum d$ square)

Step 5- Insert these values in the formula

$$rR = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$= 1 - (6 * 12) / (9(81 - 1))$$

$$= 1 - 72 / 720$$

$$= 1 - 0.1$$

$$= 0.9$$

The Spearman's Rank Correlation for this data is 0.9 and as mentioned above if the ρ value is nearing +1 then they have a perfect association of rank.

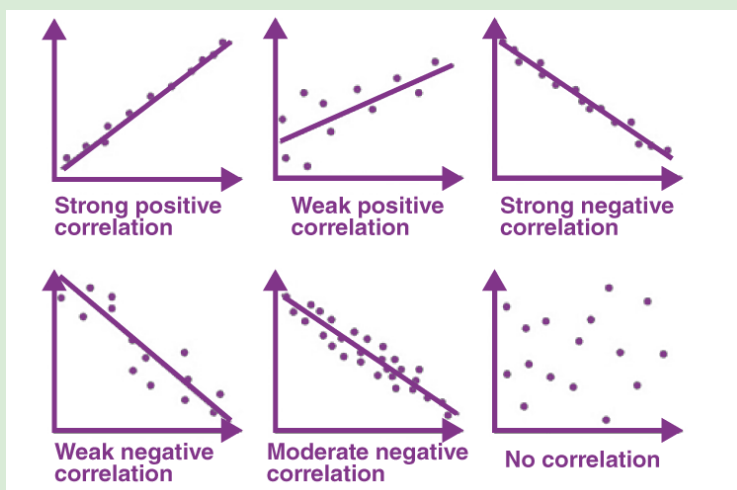
CONCEPTS OF CORRELATION IN STATISTICS

This section shows how to calculate and interpret correlation coefficients for ordinal and interval level scales. Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r , and it ranges from -1 to $+1$.

A correlation coefficient quite close to 0, but either positive or negative, implies little or no relationship between the two variables. A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.

A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable. A correlation coefficient can be produced for ordinal, interval or ratio level variables, but has little meaning for variables which are measured on a scale which is no more than nominal.

For ordinal scales, the correlation coefficient can be calculated by using Spearman's rho. For interval or ratio level scales, the most commonly used correlation coefficient is Pearson's r , ordinarily referred to as simply the correlation coefficient.



SUMMARY

- Simple linear regression refers to fitting a response y as a linear function of a single predictor x . Although calibration of assay methods sometimes requires more than a single predictor x , simple linear regression is often adequate.
- Regression analysis is not used for variables that have an exact linear relationship, as there would be no need for it.
- Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).
- Robust regression techniques aim to provide more stable and accurate parameter estimates even when assumptions are violated or outliers are present.
- A positive value of r indicates positive correlation, a higher value indicating a stronger correlation between X and Y (i.e. the observations lie closer to a straight line). $r = 1$ indicates perfect positive correlation and means that all the observations lie precisely on a straight line with positive slope.
- Correlation Coefficient is a statistical concept, which helps in establishing a relation between predicted and actual values obtained in a statistical experiment.

MULTIPLE CHOICE QUESTIONS

1. Which of the following are types of correlation?
 - a. Positive and Negative
 - b. Simple, Partial and Multiple
 - c. Linear and Nonlinear
 - d. All of the above
2. Which of the following is true for the coefficient of correlation?
 - a. The coefficient of correlation is not dependent on the change of scale
 - b. The coefficient of correlation is not dependent on the change of origin
 - c. The coefficient of correlation is not dependent on both the change of scale and change of origin
 - d. None of the above
3. In practice, Line of best fit or regression line is found when _____
 - a. Sum of residuals ($\sum(Y - h(X))$) is minimum
 - b. Sum of the absolute value of residuals ($\sum|Y-h(X)|$) is maximum
 - c. Sum of the square of residuals ($\sum (Y-h(X))^2$) is minimum
 - d. Sum of the square of residuals ($\sum (Y-h(X))^2$) is maximum
4. If Linear regression model perfectly first i.e., train error is zero, then _____
 - a. Test error is also always zero
 - b. Test error is non zero
 - c. Couldn't comment on Test error
 - d. Test error is equal to Train error
5. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?
 - a. 1
 - b. 2
 - c. 3
 - d. 4

Answers

1. (d) 2. (c) 3. (c) 4. (c) 5. (b)

REVIEW QUESTIONS

1. What is a linear regression?
2. How do you interpret a Q-Q plot in a linear regression model?
3. Discuss about method of ordinary least square (OLS).
4. Define the fixed versus random regressors.
5. Write the correlation coefficient.
6. Explain the assumptions of Karl Pearson's correlation coefficient.
7. Describe the Spearman rank correlation coefficient.

REFERENCES

1. Allen, D. M. "Mean square error of prediction as a criterion for selecting variables," *Technometrics*, 13, 469–475.
2. Allen, D. M. "The relationship between variable selection and data augmentation and a method for prediction," *Technometrics*, 16, 125–127.
3. Andrews, D. F. "Significance tests based on residuals," *Biometrika*, 58, 139–148.
4. Andrews, D. F. "The robustness of residual displays," in R. L. Launer and G. N. Wilkinson (Eds.), *Robustness in Statistics*, Academic Press, New York, pp. 19–32.
5. Andrews, D. F. "A robust method for multiple linear regression," *Technometrics*, 16, 523–531.
6. Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey [1972], *Robust Estimates of Location*, Princeton University Press, Princeton, N.J.
7. Anscombe, F. J. "Examination of residuals," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California, Berkeley, pp. 1–36.
8. *Applying Regression and Correlation: A Guide for Students and Researchers*, by Jeremy Miles, Mark Shevlin
9. *Correlation and Regression Analysis: A Historian's Guide*, by Thomas J. Archdeacon
10. Dodge, Yadolah (2010). *The Concise Encyclopedia of Statistics*. Springer-Verlag New York. p. 502.
11. Lehman, Ann (2005). *Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. Cary, NC: SAS Press. p. 123.
12. *Methods of Correlation and Regression Analysis, Linear and Curvilinear*, by Mordecai Ezekiel, Karl August Fox
13. Myers, Jerome L.; Well, Arnold D. (2003). *Research Design and Statistical Analysis* (2nd ed.). Lawrence Erlbaum. pp. 508.
14. Poston, Dudley L., Jr. 2002. The Statistical Modeling of the Fertility of Chinese Women. *Journal of Modern Applied Statistical Methods* 1 (2): 387–396.
15. Poston, Dudley L., Jr., and Sherry L. McKibben. 2003. Zero-inflated Count Regression Models to Estimate the Fertility of U.S. Women. *Journal of Modern Applied Statistical Methods* 2 (2): 371–379.
16. Vittinghoff, Eric, David V. Glidden, Stephen C. Shiboski, and Charles E. McCulloch. 2005. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. New York: Springer.
17. Yule, G. U.; Kendall, M. G.. *An Introduction to the Theory of Statistics* (14th ed.). Charles Griffin & Co. p. 268.

Index

A

Autocorrelation Function (ACF) 87, 88

B

Bar Chart 5

Box-plot 23

C

Calibration 252

Central Limit Theorem (CLT) 182

Central tendency 8, 9, 10, 11, 13, 14, 15,
16, 19, 20, 23, 24, 40, 45, 46, 79

Complex hypothesis 174

Correlation 43, 60, 61, 62, 63, 65, 87, 89

Correlation coefficient 250, 251, 252, 253,
254, 255, 256, 261

Crude approach 231

D

Data 5, 6, 26, 27, 28

Data analysis 26, 27, 31

Descriptive Statistics 3, 24, 25

Distribution 83, 85, 86, 88, 90, 91, 92, 93,
94, 95, 96, 99, 100, 102, 103, 104, 105,
106, 107, 165, 168, 169, 175, 177, 180,
181, 182, 183, 184, 186

Distribution parameter 197

E

Empirical distribution function 93, 94, 95

Error variance 233

Exploratory Data Analysis (EDA) 26, 27

F

Function 32

H

Histogram 7, 9, 12, 28, 29

Hypothesis 25, 26, 93, 96, 160, 161, 162,
163, 164, 165, 166, 167, 168, 169, 170,
171, 172, 173, 174, 175, 176, 177, 179,
180, 181, 184, 185, 186, 187, 188, 189,
193

Hypothesis testing 161, 163

I

Income ratio (IR) 250

Inferential statistics 3, 4, 23, 25, 26, 40

Interquartile Range (IQR) 48

Interval scale 35

K

Kolmogorov-Smirnov statistic 95

Kolmogorov-Smirnov test 93

L

Linear function 228, 259
 Linear regression model 229
 Linear relationship 63, 69, 72, 73, 75

M

Mahalanobis distance 144, 145, 146,
 147, 148, 151
 Measurement error 253
 Median 8, 9, 10, 11, 12, 15, 16, 17, 20,
 21, 22, 23, 24, 28
 Median absolute deviation (MAD) 17
 Minimum Variance 84, 85
 Monotonic relation 256
 Multiple Regression 143
 Multivariate graphical 27

N

Natural choice 202
 Normal distribution 233
 Null hypothesis 161, 162, 163, 164, 167,
 168, 169, 170, 171, 172, 173, 174,
 175, 176, 177, 178

P

Parameter 82, 83, 86, 99
 Parameter estimation 197, 202, 213, 216
 Parameter value 202, 203, 205, 214
 Plural hypotheses 93
 Point Estimation 82
 Population 3, 4, 7, 10, 15, 23, 24, 25, 32,
 33, 34, 82, 84, 86, 89, 93, 94, 95, 98,
 99, 100, 101, 159, 162, 163, 168,
 169, 170, 171, 177, 179, 180, 181,
 182, 183, 184, 185, 186, 187, 188,
 189, 193
 Population mean 159, 169, 170, 171,
 177, 179, 181, 182, 184, 186, 188,
 189
 Population parameter 163, 168
 Population proportion 168, 169
 Population variance 48, 51, 52
 Posterior 145
 Precision 17
 Prediction 172
 Prediction equation 231, 233, 253
 Probability 7, 25, 27, 32, 83, 93, 94, 95,

97, 102, 112, 161, 167, 169, 170, 171,
 175, 176, 177, 180, 182, 184, 186

Probability distribution 197, 201
 Probability theory 97

Q

Qualitative data 5, 6
 Quantitative 3, 4, 5, 6, 9, 27, 29, 32, 33
 Quartile 16, 20, 21, 22, 23, 24

R

Radial Basis Function (RBF) 146
 Random variables 82, 88, 90, 91, 92
 Regression analysis 25, 229, 230, 233, 238,
 240, 243, 252, 259
 Regression equation 231, 234, 247
 Regression fitting 228
 Russian mathematician 95

S

Sample mean 10, 28
 Sample variance 48, 51, 52
 Sampling distribution 54, 55, 57, 59
 Simple linear regression 228, 246, 259
 Simple random sample 48, 49, 51, 52, 55
 Spearman's rank coefficient 255
 Standard Deviation (S.D) 16
 Statistic 82, 83, 86, 90, 91, 93, 96
 Statistical analysis 4, 31, 98, 99
 Statistical data analysis 36
 Statistical hypothesis 162, 163, 164, 168,
 173
 Statistical inference 86, 97
 Statistical Observation 84
 Statistical populations 100
 Statistical theory 44
 Statistics 3, 4, 7, 10, 11, 16, 23, 24, 25, 27,
 40
 Straight-line relationship 229, 235

T

Tabular representation 25

U

Univariate graphical 27

Core Concepts in Mathematics: Statistics

2ND EDITION

Statistics plays a vital role in every field of human activity due to its enormous importance in daily life. With applications in a vast number of different fields, statistics is a very broad subject. In general, it can be said that statistics is the methodology for collecting, analyzing, interpreting, and drawing conclusions from information. Every day, we encounter a variety of data-based evaluations and claims. Determining the existing position of per capita income, unemployment rates, population growth rates, housing availability, schooling facilities, medical resources - all these aspects are made possible by the knowledge of statistics and exploratory data analysis. These tools enable us to evaluate the truth value behind such claims and protect ourselves from manipulative or deceiving tricks. By leveraging statistical tools and methods, we can make sense of all the data and related information that we come across on a regular basis. Statistics serves as a crucial process not only in how we make discoveries in science but also how we make decisions based on data-driven insights and predictions.

The objective of the book *Core Concepts in Statistics (2nd Edition)* is to provide students with a comprehensive and accessible introduction to statistical concepts. Through clear explanations and examples, the book aims to enhance students' understanding of statistics and how it relates to real-world applications. It seeks to equip readers with essential skills for interpreting data, analyzing trends, and making informed decisions based on empirical evidence. The book strives to promote critical thinking by encouraging readers to evaluate statistical claims critically and differentiate between valid data analysis techniques from misleading ones. The book further strives to familiarize students with basic principles of study design, sampling methods, probability theory, hypothesis testing, confidence intervals, correlation analysis, regression modeling, and experimental design. With its user-friendly approach and practical exercises that reinforce key concepts, *Core Concepts in Statistics* assists learners in developing statistical literacy necessary for success in various fields where statistics play a crucial role.

Diego Fernandez is Associate Professor in the faculty of Mathematics. His teaching portfolio encompasses courses on Bayesian Statistics, Biostatistics and Bioinformatics, Data Science, Dependent Data, and Foundations of Probability and Statistics. Regarding research pursuits, Dr. Fernandez shows a keen interest in Bayesian methodology as well as Environmental and Ecological Statistics, Machine Learning, Official Statistics, Spatial Statistics, Spatio-temporal models, Survey Methodology. Spearheading Statistical Theory & Methodologies are Spatial Point Process analysis techniques such as Bayesian Nonparametric methods alongside Big Data Inference modeling strategies for Causal Inference and Survival Analysis investigations.

ISBN 978-1-98469-082-1

