Probability and Statistics for Engineers

First Edition

Benjamin Odoi, Abdulzeid Yen Anafo and

Seth Antanah

Contents

Preface IV

1 **Introduction to Statistics** 1

1.1 Learning objectives 1

1.2 Introduction to Statistics 1

1.3 Why Statistics ? 2

1.4 Branches of Statistics 3

1.4.1 Descriptive statistics 3

1.4.2 Inferential statistics 3

1.5 Variables 3

1.5.1 Qualitative vs. Quantitative Variables 3

1.5.2 Discrete vs. Continuous Variables 4

1.6 Univariate vs. Bivariate Data **4**

1.6.1 Univariate data 4

1.6.2 Bivariate data 4

1.6.3 Populations and Samples 5

1.6.4 Population vs.Sample 5

1.7 Summarizing data graphically **5**

1.8 Summary Statistics **6**

1.8.1 Measure of Location 6

1.8.2 Measure of Spread/ Variability 6

1.8.3 How to Describe Data Patterns in Statistics 6

1.9 Unusual Features **6**

1.9.1 Gaps 7

1.9.2 Outliers 7

1.9.3 How to Compare Data Sets 7

1.9.4 Four Ways to Describe Data Sets 7

1.10 Sampling Procedures **8**

1.10.1 Simple random sampling 8

1.10.2 Systematic random sampling 9

1.10.3 Stratified Sampling 10

1.10.4 Cluster sampling (also called block sampling) 10

1.11 Levels of Measurement (Types of Data) **13**

1.12 Frequency Distribution **14**

1.13 Measure of Location and Dispersion **15**

1.13.1 Measures of Location 16

1.13.2 Relation between Measure of Location and Types of Frequency Curves. 19
1.13.3 Measures of Dispersion, Skewness 19
1.13.4 Variance and Standard Deviation. 20
1.13.5 Coefficient of Variation. 21
1.13.6 Skewness 22
1.13.7 Kurtosis 22
1.13.8 Question 22
1.13.9 References 23

2 Introduction to Probability 25

2.1 Learning Objectives 25

2.2 Introduction 25

2.2.1 Determination of Probability of an Event 28
2.2.2 Probability of Compound Events 30
2.2.3 Multiplication Rule for $P(A \cap B)$ 34
2.2.4 Axioms of Probability 36
2.2.5 Some Rules of Probability 36
2.2.6 Application of Counting Techniques 37
2.2.7 Permutation of Objects 39

3 Random Variables and Distribution 43

3.1 Introduction 43

3.2 Random Variable 43

3.2.1 Types of Random Variables 45
3.2.2 Discrete Probability Distribution Variable 46

3.2.3 Continuous Probability Distribution Variable 49

3.2.4 Probability Density Function (PDF) 50

4 Special Distribution 53

4.1 Introduction 53

4.1.1 Discrete Probability Distribution 53

4.1.2 Continuous Probability Distribution 61

4.1.3 Mathematical Expectations 67

5 Estimations 69

5.0.1 Introduction 69

5.0.2 Properties of a Point Estimator 70

5.0.3 Interval Estimation 71

5.0.4 Confidence Interval For A Population Proportion 75

6 Hypothesis Testing 79

6.1 Tests of Hypotheses and Significance 79

6.1.1 Introduction 79

6.1.2 A Single Population Mean μ 82

6.1.3 Tests on the Mean of a Normal Distribution: Variance Unknown 84

6.1.4 Tests on a Population Proportion 85

6.1.5 7.5 The Difference Between two Population Means 87

7 Regression 95

7.1 Regression and Correlation Analysis 96

7.1.1 Introduction 96

7.1.2 The Regression Model 96

7.2 Method of Least Squares 100

7.3 Correlation Analysis 103

References 111

Authors 111

Preface

Even when flawless, the process of synthesis that all 'data' goes through before the communication step entails by its very nature reshaping and loss of information. As a statistician, I strongly believe my role to be:

- to help you learn how not to buy into single (data) stories
- to give you the tools not to tell them yourselves!

Well-intentioned or not (!!!), every data analysis task carries the risk to produce incomplete and misleading accounts!

Fundamentally, The book is designed to familiarise students with random variables and probability distributions of some special probability distribution with an introduction to R programming.

1. Introduction to Statistics

1.1 Learning objectives

Having worked through this chapter the student will be able to:

- Discuss the reasons for studying statistics as an engineer.
- Identify basic statistical concepts.
- Identify the levels of measurements
- discuss the sampling procedures
- understand data visualization using several graphical devices (Using R programming)

1.2 Introduction to Statistics

Statistics is a way to get information from data. Statistics is a discipline which is concerned with:

- summarizing information to aid understanding,
- drawing conclusions from data,
- estimating the present or predicting the future, and

- designing experiments and other data collection.

In making predictions, Statistics uses the concept of probability, which models chance mathematically and enables calculations of chance in complicated cases.

1.3 Why Statistics ?

The field of statistics deals with the collection, presentation, analysis, and use of data to make decisions, solve problems, and design products and processes. In simple terms, statistics is the science of data.

Because many aspects of engineering practice involve working with data, obviously knowledge of statistics is just as important to an engineer as are the other engineering sciences. Specifically, statistical techniques can be powerful aids in designing new products and systems, improving existing designs, and designing, developing, and improving production processes.

Statistical analysis provides objective ways of evaluating patterns of events or patterns in our data by computing the probability of observing such patterns by chance alone. Insisting on the use of statistical analyses on which to draw conclusions is an extension of the argument that objectivity is critical in science. Without the use of statistics, little can be learnt from most research studies.

Because of the increasing use of statistics in so many areas of our lives, it has become very desirable to understand and practice statistical thinking. This is important even if you do not use statistical methods directly.

1.4 Branches of Statistics

1.4.1 Descriptive statistics

This is the branch of statistics that involves the organization, summarization, and display of data. Two general techniques are used to accomplish this goal.

- Organize the entire set of scores into a table or a graph that allows researchers (and others) to see the whole set of scores. (summarizing data graphically)
- Compute one or two summary values (such as the average) that describe the entire group. (summarizing data numerically).

1.4.2 Inferential statistics

This is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

1.5 Variables

In statistics, a variable has two defining characteristics:

- A variable is an attribute that describes a person, place, thing, or idea.
- The value of the variable can "vary" from one entity to another.

For example, a person's hair color is a potential variable, which could have the value of "blond" for one person and "brunette" for another.

1.5.1 Qualitative vs. Quantitative Variables

Variables can be classified as qualitative (aka, categorical) or quantitative (aka, numeric).

1. Qualitative. Qualitative variables take on values that are names or labels. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, and terrier) would be examples of qualitative or categorical variables.

2. Quantitative. Quantitative variables are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable. In algebraic equations, quantitative variables are represented by symbols (e.g., x, y, or z).

1.5.2 Discrete vs. Continuous Variables

Quantitative variables can be further classified as discrete or continuous. If a variable can take on any value between its minimum value and its maximum value, it is called a continuous variable; otherwise, it is called a discrete variable.

1.6 Univariate vs. Bivariate Data

Statistical data are often classified according to the number of variables being studied.

1.6.1 Univariate data

When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

1.6.2 Bivariate data

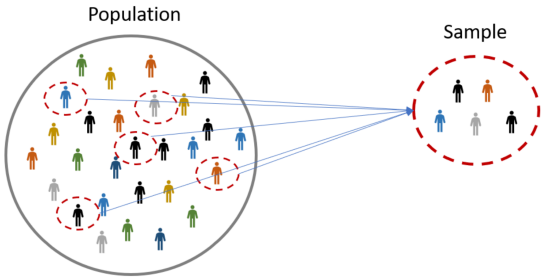
When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there was a relationship between the height and weight of high school students. Since we are working with two variables (height and weight), we would be working with bivariate data.

1.6.3 Populations and Samples

The study of statistics revolves around the study of data sets. This lesson describes two important types of data sets - populations and samples. Along the way, we introduce simple random sampling, the main method used in this tutorial to select samples.

1.6.4 Population vs.Sample

The main difference between a population and sample has to do with how observations are assigned to the data set. A population includes all of the elements from a set of data. A sample consists of one or more observations from the population. Depending on the sampling method, a sample can have fewer observations than the population, the same number of observations, or more observations. More than one sample can be derived from the same population.



A measurable characteristic of a population, such as a mean or standard deviation, is called a parameter; but a measurable characteristic of a sample is called a statistic. We will see in future lessons that the mean of a population is denoted by the symbol μ; but the mean of a sample is denoted by the symbol $\bar{X}$.

1.7 Summarizing data graphically

Selected graphs for qualitative data

- Pie chart
- Bar Chart (Also frequency distribution)

Selected graphs for Numerical data

- Box plot
- Dot plot
- Stem-and-leaf
- Histogram

1.8 Summary Statistics

1.8.1 Measure of Location

These provide an indication of the center of the distribution where most of the scores tend to cluster. There are three principal measures of central tendency: Mode, Median, and Mean.

1.8.2 Measure of Spread/ Variability

Variability is the measure of the spread in the data. The three common variability concepts are: Range, Variance and Standard deviation.

1.8.3 How to Describe Data Patterns in Statistics

Graphic displays are useful for seeing patterns in data. Patterns in data are commonly described in terms of: Center, Spread, Shape, Symmetry, Skewness and Kurtosis

1.9 Unusual Features

Sometimes, statisticians refer to unusual features in a set of data. The two most common unusual features are gaps and outliers.

1.9.1 Gaps

Gaps refer to areas of a distribution where there are no observations. The first figure below has a gap; there are no observations in the middle of the distribution.

1.9.2 Outliers

Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are called outliers. The second figure below illustrates a distribution with an outlier. Except for one lonely observation (the outlier on the extreme right), all of the observations fall between 0 and 4. As a "rule of thumb", an extreme value is often considered to be an outlier if it is at least 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3).

1.9.3 How to Compare Data Sets

Common graphical displays (e.g., dot plots, box plots, stem plots, bar charts) can be effective tools for comparing data from two or more data sets.

1.9.4 Four Ways to Describe Data Sets

When you compare two or more data sets, focus on four features:

- Center: Graphically, the center of a distribution is the point where about half of the observations are on either side.
- Spread: The spread of a distribution refers to the variability of the data. If the observations cover a wide range, the spread is larger. If the observations are clustered around a single value, the spread is smaller.
- Shape: The shape of a distribution is described by symmetry, skewness, number of peaks, etc.
- Unusual: features Unusual features refer to gaps (areas of the distribution where

there are no observations) and outliers.

1.10 Sampling Procedures

Statisticians employ different procedures in choosing the observations that will constitute their random samples of the population. The objective of these procedures is to select samples that will be representative of the population from where they originate. These samples, also known as random samples, will have the property that each sample has the same probability of being drawn from the population as another sample. There are two types of sam

1.10.1 Simple random sampling

Simple random sampling is used to make statistical inferences about a population. It helps ensure high internal validity: randomization is the best method to reduce the impact of potential confounding variables. However, simple random sampling can be challenging to implement in practice. To use this method, there are some prerequisites:

- You have a complete list of every member of the population.
- You can contact or access each member of the population if they are selected.
- You have the time and resources to collect data from the necessary sample size.

How is a simple random sampling perforned?

- Define the population
- Decide on the sample size
- Randomly select your sample
- Collect data from your sample

1.10.2 Systematic random sampling

Systematic sampling is a method that imitates many of the randomization benefits of simple random sampling, but is slightly easier to conduct.

You can use systematic sampling with a list of the entire population, like you would in simple random sampling. However, unlike with simple random sampling, you can also use this method when you're unable to access a list of your population in advance.

> *Example:* The (testable) Your population list alternates between men (on the even numbers) and women (on the odd numbers). You choose to sample every tenth individual, which will therefore result in only men being included in your sample. This would obviously be unrepresentative of the population.

> *Example 2:* You run a department store and are interested in how you can improve the store experience for your customers. To investigate this question, you ask an employee to stand by the store entrance and survey every 20th visitor who leaves, every day for a week. Although you do not necessarily have a list of all your customers ahead of time, this method should still provide you with a representative sample of your customers since their order of exit is essentially random.

How is a systematic random sampling performed?

- Define and list your population, ensuring that it is not ordered in a cyclical or periodic order.
- Decide on your sample size and calculate your interval,k, by dividing your population by your target sample size.
- Choose every k^{th} member of the population as your sample.

1.10.3 Stratified Sampling

In a stratified sample, researchers divide a population into homogeneous sub populations called strata (the plural of stratum) based on specific characteristics (e.g., race, gender identity, location, etc.). Every member of the population studied should be in exactly one stratum.

Each stratum is then sampled using another probability sampling method, such as cluster sampling or simple random sampling, allowing researchers to estimate statistical measures for each sub-population.

Researchers rely on stratified sampling when a population's characteristics are diverse and they want to ensure that every characteristic is properly represented in the sample. This helps with the generalizability and validity of the study, as well as avoiding research biases like undercoverage bias.

How do we perform stratified sampling?

- Define your population and subgroup
- Separate the population into strata
- Decide on the sample size for each stratum
- Randomly sample from each stratum

1.10.4 Cluster sampling (also called block sampling)

This is a sampling procedure that randomly selects clusters of observations from the population under study, and then chooses all, or a random selection, of the elements of these clusters, as the observations of the sample.

How is Cluster sampling performed?

STEP 1


Step 3: Randomly select clusters


Step 4: Collect data from clusters

1.11 Levels of Measurement (Types of Data)

Variables can be classified on the basis of their level of measurement. The way we classify variables greatly affects how we can use them in our analysis. Variables can be

- *Ordinal;* Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

  Ordinal data is qualitative data for which their values have some kind of relative position. These kinds of data can be considered "in-between" qualitative and quantitative data. The ordinal data only shows the sequences and cannot use for statistical analysis. Compared to nominal data, ordinal data have some kind of order that is not present in nominal data.

- *Norminal* Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

  The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

- *interval,* Measurements on a numerical scale in which the value of zero is arbitrary but the difference between values is important. Of all four levels of measurement, only the ratio scale is based on a numbering system in which zero is meaningful. Therefore, the arithmetic operations of multiplication and division also take on a rational interpretation. A ratio scale is used to measure

many types of data found in business and geoscientific analyses. Variables such as costs, profits, inventory levels and grades are expressed as ratio measures. The value of zero dollars to measure revenues, for example, can be logically interpreted to mean that no sales have occurred. Furthermore, a firm with a 40 percent market share has twice as much of the market as a firm with a 20 percent market share. Measurements such as weight, time, and distance are also measured on a ratio scale since zero is meaningful, and an item that weighs 100 pounds is one-half as heavy as an item weighing 200 pounds.

- *ratio.* Numerical measurements in which zero is a meaningful value and the difference between values is important. You may notice that the four levels of measurement increase in sophistication, progressing from the crude nominal scale to the more refined ratio scale. Each measurement offers more information about the variable than did the previous one. This distinction among the various degrees of refinement is important, since different statistical techniques require different levels of measurements. While most statistical tests require interval or ratio measurements, other tests, called nonparametric tests (which will be examined later in this text), are designed to use nominal or ordinal data.

1.12 Frequency Distribution

Graphical representation makes unwieldy data readily intelligible and brings to light the salient features of the data at a glance. It makes visual comparison of data easier. It facilitates the comparison of two frequency distributions.

Several graphical devices are often used to portray shapes of distributions. The following types of graphs are commonly used in representing frequency distributions.

- Stem-and –leaf display,
- Dot plot.
- Box-and-whiskers display (box plot)
- Histogram
- Frequency polygon and Frequency curve
- Cumulative frequency curve or the 'Ogive'
- Pareto chart
- Pie chart
- Bar chart

1.13 Measure of Location and Dispersion

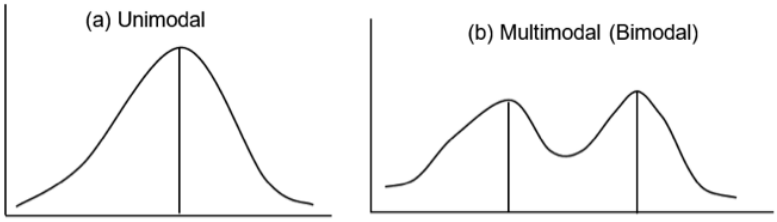
In addition to the histogram, the information in the frequency distribution can be further summarised by means of just two numbers. The first is the location of the data, and the various numbers that provide information about this are known as 'measures of location' or 'measures of central tendency'. 'Location of the data' refers to a value that is typical of all the sample observations.

The second important aspect of the data is the dispersion of the observations. This implies how the data are scattered (dispersed). This is also called 'measure of variation'.

1.13.1 Measures of Location

The Mode:

The mode is defined as the observation in the sample which occurs most frequently. If there is only one mode then it is unimodal otherwise it is multimodal



The Arithmetic Mean

It is the most commonly used measure of locations. Let the variable x take the values $x_1, x_2 \ldots x_n$. The arithmetic mean is defined as:

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

For large data it may be advantageous to classify the data. If these n observations have corresponding frequencies, the arithmetic mean is computed using the formula,

$$x = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{n} \tag{1.13.1}$$

can be rewritten as,

$$x = \frac{1}{n}\sum_{n}^{i=1} x_i f_i \tag{1.13.2}$$

The Geometric Mean

The geometric mean is the average of a set of products, the calculation of which is commonly used to determine the performance results of an investment or portfolio. It is technically defined as "the nth root product of n numbers."

$$G = antilog\left(\frac{1}{N}\sum f_i \log x_i\right) \tag{1.13.3}$$

where $x_i = x_1, x_2, \ldots, x_n$ and $f_i = f_1, f_2, \ldots f_n$ The geometric mean may be used to show percentage changes in a series of positive numbers. As such, it has wide application in business and economics, since we are often interested in determining the percentage change in sales, gross national product, or any other economic series. The geometric mean (GM) is found by taking the nth root of the product of n numbers. Thus:

$$GM = (X_1, X_2, \ldots, X_n)^{\frac{1}{n}} \tag{1.13.4}$$

GM is most often used to calculate the average growth rate over time of some given series.

Example: A farm labourer wishes to determine the average growth rate of his monthly income based on the figures in the table below. If the average annual growth rate of monthly salary is less than 10% he will resign. Using GM should he resign?

Year	**Revenue**	**Percentage of Previous Year**
1992	50 000	-
1993	55 000	55/50 =1.10
1994	66 000	66/55= 1.20
1995	70 000	70/66=1.06
1996	78 000	78/70 =1.11

Solution:

$$GM = (1.10 \times 1.20 \times 1.06 \times 1.11)^{(\frac{1}{4})} = 1.16 \tag{1.13.5}$$

Harmonic Mean

In statistics, harmonic mean is used to find the average rate.the harmonic mean is the reciprocal of the arithmetic mean of the reciprocals.

Note:

- Harmonic mean is important in problems in which variable-values are compared with a constant quantity of another variable , i.e. time, distance covered within a certain time, etc.

- Another word for average. Mean almost always refers to arithmetic mean. In certain contexts, however, it could refer to the geometric mean, harmonic mean, or root mean square

Example: Compute the Geometric and Harmonic means for the numbers 4 and 9.

Solution:

$$\text{Harmonic Mean} = \frac{2}{\frac{1}{4}+\frac{1}{9}} = 5.53 \tag{1.13.6}$$

Median

If the sample observations are arranged in order from smallest to largest, the median is defined as the middle observation if the number of observations is odd, and as the number halfway between the two middle observations if the number of observations is even. The genral formulae for the median is given as

$$MD = b_L + \frac{\frac{n}{2} - f_{m-1}}{f_m} \times c \tag{1.13.7}$$

where,

b_L = lower boundary of the median class

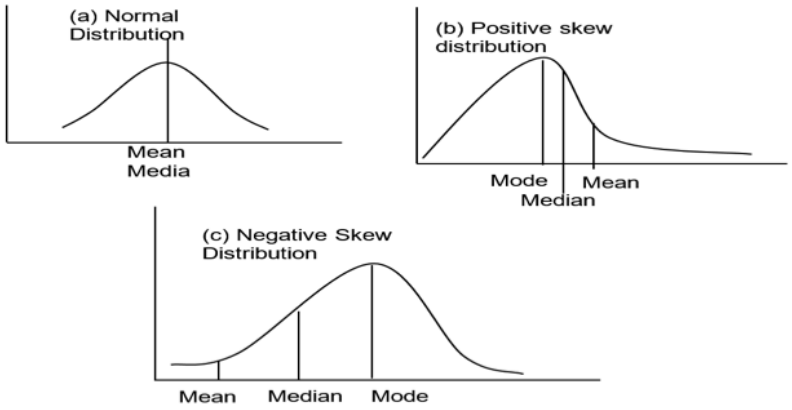
n = number of observations f_m = the number of observations in the median class

$f_m - 1$ = the cumulative frequency of the class preceding the median class.

c = class interval of the median class

Note: Because of the distorting effect of extreme observations on the mean, the median is often the preferred measure in such situations as salary negotiations.

1.13.2 Relation between Measure of Location and Types of Frequency Curves.



1.13.3 Measures of Dispersion, Skewness

It should be clear that a measure of central tendency by itself can exhibit only one of the important characteristics of a distribution and therefore while studying a distribution it is equally important to know how the variates are clustered around or away from the point of central tendency. The variation of the points about the mean is called dispersion. Spread or dispersion can be classified into three groups.

- Measures of the difference between representative variate values such as the range-the interquartile or the interdecile range.
- Measures obtained from the deviations of every variate value from some central value such as the mean deviation from the mean or the mean deviation from the median or the standard deviation.

- Measures obtained from the variations of all the variates among themselves, such as mean difference.

Range

It is the difference between the extreme values of the variate i.e. $(x_n - x_1)$ when the values are arranged in ascending order.

The Interquartile range

It is the difference between the 75% and 25% i.e. $(X_{75\%} - X_{25\%})$. The interdecile range is the difference between the ninth and first decile i.e. $X_{0.9} - X_{0.1}$. This combines eighty percent of the total frequency while the interquartile range contains fifty percent. They are only mainly used in descriptive statistics because of the mathematical difficulty in handling them in advanced statistics.

Average Deviation or Mean Deviation

The mean deviation is defined as:

$$MD = \frac{\sum |(x_i - \bar{x})|}{n} \tag{1.13.8}$$

For very large n, M.D may equal zero as some deviations may be negative and others positive but the individual deviations could be numerically large, thus giving a poor expression of the intrinsic dispersion. The mean absolute deviation (M.A.D). provides a better and more useful measure of dispersion.

1.13.4 Variance and Standard Deviation.

In order to minimise the inefficiency of the mean deviation outlined earlier a better option is the sum of the square deviations i.e. (known simply as the 'sum of squares'.

The mean of this sum of squares is the sample variance; denoted symbolically as:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n} \tag{1.13.9}$$

For theoretical reasons, the sum of squares is divided by (n˘1) rather than n because it represents a better estimate of the standard deviation.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{1.13.10}$$

For $n > 35$ there is practically no significant difference in the definitions. Sample standard deviation S is defined as: $S = \sqrt{S^2}$ or more appropriately:

$$s^2 = \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right) \tag{1.13.11}$$

For classified data, if the data have k classes

$$s^2 = \frac{1}{n-1}\left(\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}\right) \tag{1.13.12}$$

1.13.5 Coefficient of Variation.

Whilst the variance is very important in measuring dispersion it has certain limitations in its applications in comparing distributions that:

- Have significantly different means
- Are measured in different units

In such situations it is better to use the coefficient of variation which assesses the degree of dispersion of a data set relative to its mean.

$$CV = \frac{s}{\bar{x}} \times 100\% \tag{1.13.13}$$

1.13.6 Skewness

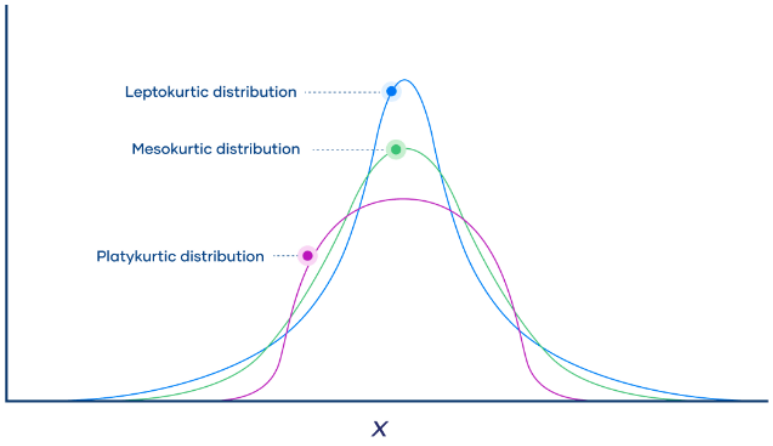
Measures describing the symmetry of distributions are called 'Coefficients of Skewness'. One such measure is given by:

$$\alpha_3 = \frac{\sum(x_i - \bar{x})^3}{S^3} \tag{1.13.14}$$

1.13.7 Kurtosis

Measures of the degree of peakedness of a distribution are called 'coefficients of kurtosis' or briefly 'kurtosis'. It is often measured as:

$$\alpha_4 = \frac{\sum(x_i - \bar{x})^4}{S^4} \tag{1.13.15}$$



1.13.8 Question

R A sample of size 40 produces the following arranged data. Note that the data has a missing value of x at the $x_{(39)}$ (the second largest number).This will NOT prevent you from answering the questions below.

14.1	46.0	49.3	53.0	54.2	54.7	54.7
54.7	54.8	55.4	57.6	58.2	58.3	58.7
58.9	60.8	60.9	61.0	61.1	63.0	64.3
65.6	66.3	66.6	67.0	67.9	70.1	70.3
72.1	72.4	72.9	73.5	74.2	75.3	75.4
75.9	76.5	77.0	x	88.9		

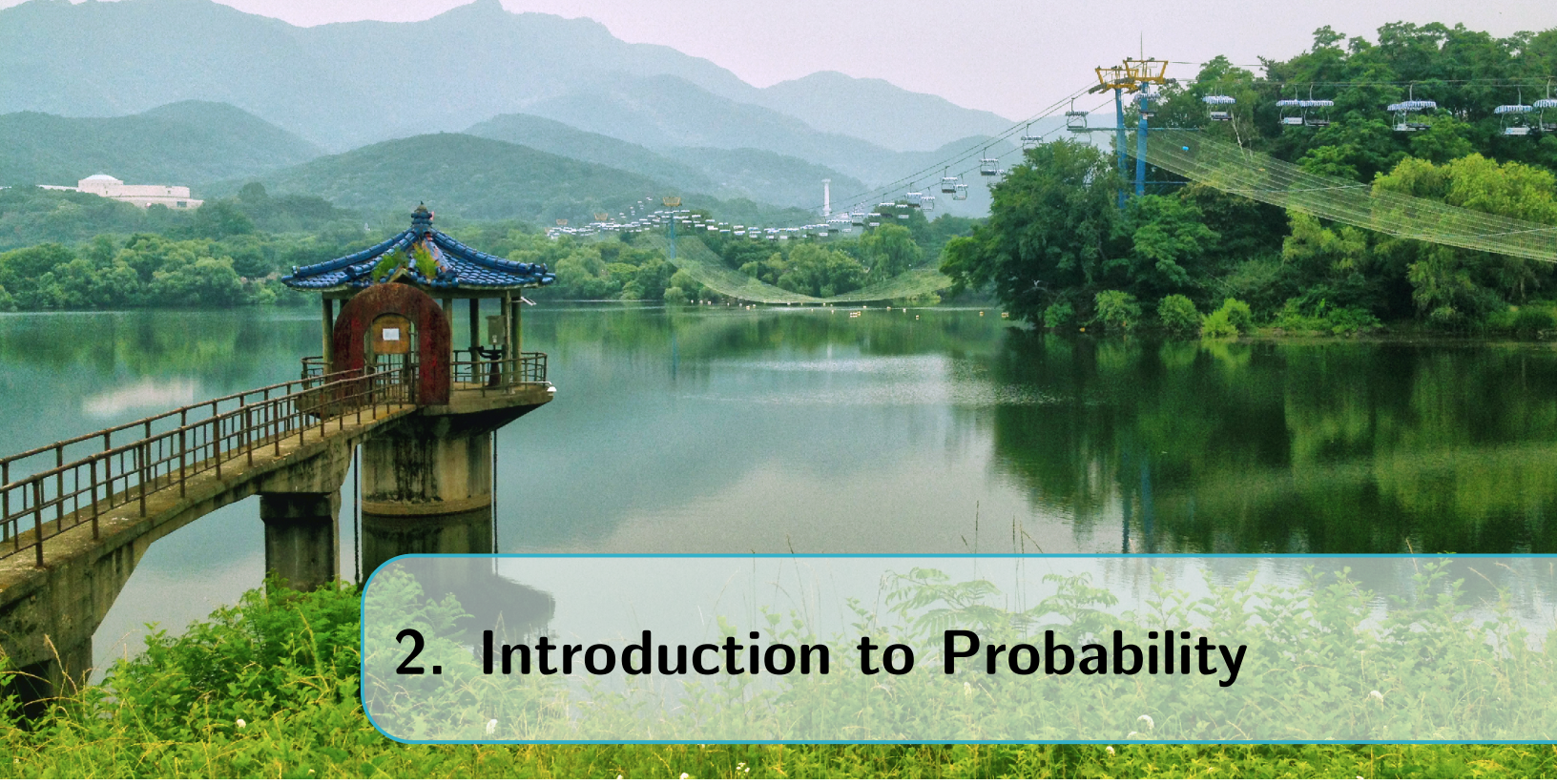
1. Calculate range,IQR,and median of these data.
2. Given that the mean of these datais 63.50(exactly) and the standard deviation is 12.33, what proportion of the data lie within one standard deviation of the mean?
3. Zeid decides to delete the smallest observation, 14.1, from these data.Thus,Zeid has a data set with $n = 39$. Calculate the range, IQR, and median of Zeid's new data set.
4. Refer to (3).Calculate the mean of Zeid's new data set.

Solution: In Class :-)

1.13.9 References

Since I found so much good information about pretty much everything I wanted to know about, I will just create a remark and let you know where you can find more specific information about, just like below.

R For more information about the cosmological principle, review Chapter 1: Why Learn Astronomy?, page 10, from **21st Century Astronomy**, *Hester* / *Smith* / *Blumenthal* / *Kay* / *Voss*, Third Edition, 2010.



2. Introduction to Probability

2.1 Learning Objectives

Having worked through this chapter the student will be able to:

- Interpret probabilities and use probabilities of outcomes to calculate probabilities of events in discrete sample spaces.
- Interpret and calculate conditional probabilities of events.
- Use Bayes' theorem to calculate conditional probabilities
- Discuss random variables.
- use counting techniques in calculating probabilities of events

2.2 Introduction

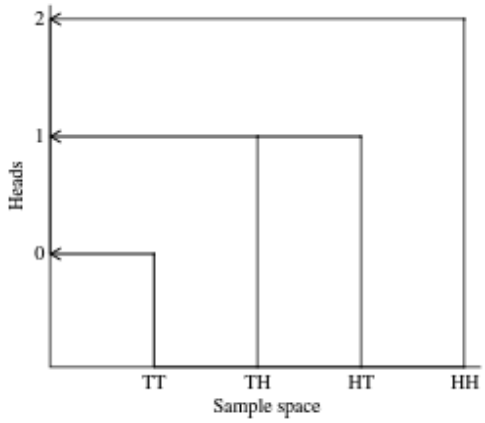
R *Probabilistic Experiment:* A probabilistic experiment is some occurrence such as the tossing of coins, rolling dice,or observation of rainfall on a particular day where a complex natural background leads to a chance outcome.

R *Trial:* Each repetition of an experiment is called a trial. That is, a trial is a single performance of an experiment.

R *Outcome:* The possible result of each trial of an experiment is called an outcome. When an outcome of an experiment has equal chance of occurring as the others the outcomes are said to be equally likely. For example, the toss of a coin and a die yield the possible outcomes in the sets, H, T and $1, 2, 3, 4, 5, 6$ and a play of a football match yields $win(W), loss(L), draw(D)$.

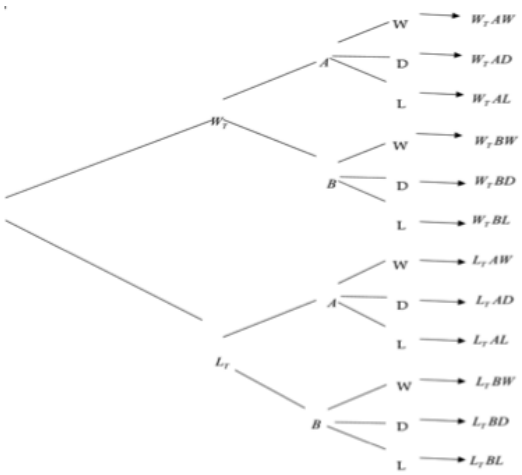
R *Random variable:* A random variable is a function that maps events defined on a sample space into a set of values. Several different random variables may be defined in relation to a given experiment. Thus, in the case of tossing two coins the number of heads observed is one random variable, the number of tails is another, and the number of double heads is another. The random variable "number of heads" associates the number 0 with the event TT, the number 1 with the events TH and HT, and the number 2 with the event HH. The Figure below illustrates this mapping.

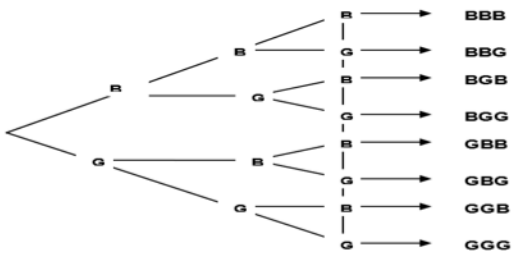
R *Sample space:* Sample space is the collection of all possible outcomes at a probability experiment. We use the notation S for sample space. Each element or outcome of the experiment is called a sample point. For example,



R *Tree Diagram:* The tree diagram represents pictorially the outcomes of random experiment. The probability of an outcome which is a sequence of trials, is represented by any path of the tree. For example,

1. Consider a couple planning to have three children, assuming each child born is equally likely to be a boy (B) or girl (G).
2. A soccer team on winning (WT) or losing (LT) a toss can defend either post A or B. It plays the match and either win (W), draw (D) or lose (L). We illustrate the experiment on a diagram as follows





2.2.1 Determination of Probability of an Event

The probability of an event A, denoted, $P(A)$, gives the numerical measure of the likelihood of the occurrence of event A which is such that $0 \leq P(A) \leq 1$. If $P(A) = 0$, the event A is said to be impossible to occur and if $P(A) = 1$, A is said to be certain. If A is the complement of the event A, then $P(A/) = 1 - P(A)$, called the probability that event A will not occur. There are three main schools of thought in defining and interpreting the probability of an event. These are the Classical Definition, Empirical Concept and the Subjective Approach. The first two are referred to as the Objective Approach.

The Classical Definition

This is based on the assumption that the outcomes of an experiment are equally likely. For example, if an experiment can lead to n mutually exclusive and equally likely outcomes, then the probability of the event A is defined by

$$P(A) = \frac{n(A)}{n(S)} = \frac{Number\ of\ succesful\ outcome}{Number\ of\ possible\ outcomes} \tag{2.2.1}$$

The classical definition of probability of event A is referred to as a prior probability because it is determined before any experiment is performed to observe the outcomes of event A.

The Empirical Concept

This concept uses the relative frequencies of past occurrences to develop probabilities for future. The probability of an event A happening in future is determined by observing what fraction of the time similar events happened in the past. That is,

$$P(A) = \frac{Number\ of\ times\ A\ occured\ in\ the\ past}{Total\ number\ of\ observations} \tag{2.2.2}$$

The relative frequency of the occurrence of the event A used to estimate $P(A)$ becomes more accurate if trials are largely repeated. The relative frequency approach of defining $P(A)$ is sometimes called posterior probability because $P(A)$ is determined only after event A is observed

The subjective Concept

The subjective concept of probability is based on the degree of belief through the evidence available. The probability of an event A may therefore be assessed through experience, intuitiveness, judgment or expertise. For example, determining the probability of getting a cure of a disease or going to rain today. This approach to probability has been developed relatively recently and is related to Bayesian Decision Analysis. Although the subjective view of probability has enjoyed increased attention over the years, it has not been fully accepted by statisticians who have traditional orientations.

Examples

R Consider the problem of a couple planning to have three children, assuming each child born is equally likely to be a boy (B) or a girl (G).

1. List the possible outcomes in this experiment
2. What is the probability of the couple having exactly two girls?

R Suppose a card is randomly selected from a packet of 52 playing cards.

1. What is the probability that it is a "Heart"?
2. What is the probability that the card bears the number 5 or a picture of a queen?

Solution: Let the sample space be the set, $S = playingcards$, $A = Heartcards$, $B = Cardsnumbered5$ and $Q = Cardswithapictureofqueen$. Then, $n(S) = 52$, $n(A) = 13$, $n(B) = 4$ and $n(Q) = 4$

R A die is tossed twice. List all the outcomes in each of the following events and compute the probability of each event.

1. The sum of the scores is less than 4
2. Each toss results in the same score
3. The sum of scores on both tosses is a prime number
4. The product of the scores is at least 20

2.2.2 Probability of Compound Events

Two or more events are combined to form a single event using the set operations, $\cap$ and $\cup$. The event

1. (A $\cap$ B) occurs if either A or B both occur(s).
2. (A $\cup$ B) occurs if both A and B occur.

Definitions:

1. **Mutually Exclusive Events:** Two or more events which have no common outcome(s) (i.e. never occur at the same time) are said to be mutually exclusive. If A and B are mutually exclusive events of an experiment, then $A \cap B = \varnothing$ and $P(A \cup B) = P(A) + P(B)$, since $P(A \cup B) = 0$.

2. **Independent Events:** Two or more events are said to be independent if the probability of occurrence of one is not influenced by the occurrence or non-occurrence of the other(s). Mathematically, the two events, A and B are said to be independent, if and only if $P(A \cap B) = P(A) \cdot P(B)$. However, if A and B are such that, $P(A \cap B) = P(A) \cdot P(A) \cdot P(B|A)$, they are said to be conditionally independent.
3. **Conditional Probability:** Let A and B be two events in the sample space, S with $P(B) > 0$. The probability that an event A occurs given that event B has already occurred, denoted $P(A/B)$, is called the conditional probability of A given B. The conditional probability of A given B is defined as.

$$P(A) = \frac{P(A \cap B)}{P(B)} \qquad P(B) > 0 \tag{2.2.3}$$

   In particular, if S is a finite equiprobable space, then

   (a) $P(A \cap B) = \frac{n(A \cap B)}{n(S)}$
   (b) $P(A) = \frac{n(S)}{n(S)}$
   (c) $P(A) = \frac{n(A \cap B)}{n(B)}$

4. **Exhaustive Events:** Two or more events defined on the same sample space are said to be exhaustive if their union is equal to the sample space (thus, if they partition the sample space mutually exclusively). Example: If $A_1, A_2, A_3 \in S$ and $A_1 \cup A_2 \cup A_3 = S$.
5. **partition of sample space:** The events form a partition of the same sample space if the following hold:

   - $A_i \neq \varnothing$ For all $i = 1, 2, 3, \ldots n$
   - $A_i \cap A_j = \varnothing$ For all $i \neq j$ $i = 1, 2, 3, \ldots n$ and $j = 1, 2, 3, \ldots n$
   - $\sum_{i=1}^{n} S$

Examples

R In a certain population of women, 40% have had breast cancer, 20% are smokers and 13% are smokers and have had breast cancer. If a woman is selected at random from the population, what is the probability that she had breast cancer, smokes or both?

Let A and B be event such that $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cup B) = 0.8$. Find

1. $P(A/B)$
2. Are A and B independent ?

Solution:

1. Let B be the event of women with breast cancer and W the event of women who smoke. Then, $P(B) = 0.4 \quad P(W) = 0.2 \quad P(B \cap W) = 0.13$

$$P(B \cup W) = P(B) + P(W) - P(B \cup W)$$
$$= 0.47$$

2. Given that, $P(A) = 0.6$, $P(B) = 0.5$ and $P(A \cup B) = 0.8$

   Applying,

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$= 0.3$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0 \tag{2.2.4}$$

   Hence, $P(A) = 0.6$

3. A and B are independent if $P(A) \cdot P(B) = P(A \cap B)$

   Hence, $P(A) \cdot P(B) = 0.6 \times 0.5 = 0.3 = P(A \cap B)$ Thus, A and B are independent.

Example on Conditional Probability: Complex components are assembled in a plant that uses two different assembly lines, A and $A/$. Line A uses older equipment than $A/$, so it is somewhat slower and less reliable. Suppose on a given day line A has assembled 8 components, of which 2 have been identified as defective (B) and 6 as non defective ($B/$), whereas $A/$ has produced 1 defective and 9 non defective components. This information is summarized in the accompanying table.

Condition		**Total**
B	B/	
2	6	8
1	9	10
3	15	18

Unaware of this information, the sales manager randomly selects 1 of these 18 components for a demonstration. Prior to the demonstration P(line A component selected)

$$P(A) = \frac{N(A)}{N} = \frac{8}{18} = 0.4 \tag{2.2.5}$$

However, if the chosen component turns out to be defective, then the event B has occurred, so the component must have been 1 of the 3 in the B column of the table. Since these 3 components are equally likely among themselves after B has occurred,

$$P(A) = \frac{P(A \cap B)}{P(B)} \tag{2.2.6}$$

$$= \frac{2/18}{3/18} \tag{2.2.7}$$

$$= \frac{2}{3} \tag{2.2.8}$$

R **Exercise**: Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and battery. Given that the selected individual purchased an extra battery, what is the probability that an optional card was also purchased?

2.2.3 Multiplication Rule for $P(A \cap B)$

The definition of conditional probability yields the following result, obtained by multiplying both sides of the conditional probability equation by $P(B)$.

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

- $P(A/B) \times P(B) = \frac{P(A \cap B)}{P(B)} \times P(B)$
- $P(A/B) \cdot P(B) = P(A \cap B)$

This rule is important because it is often the case that $P(A \cap B)$ is desired, whereas both $P(B)$ and $P(A/B)$ can be specified from the problem description.

The Law of Total Probability

Let $A_1, A_2, A_3, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = \sum_{i}^{k} P(B/A_i) \times P(A) \tag{2.2.9}$$

Bayes' Rule

The power of Bayes' rule is that in many situations where we want to compute $P(A|B)$ it turns out that it is difficult to do so directly, yet we might have direct information about $P(B|A)$. Bayes' rule enables us to compute $P(A|B)$ in terms of $P(B|A)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{2.2.10}$$

Bayes Theorem

Let A and A^c constitute a partition of the sample space S such that with $P(A) > 0$ and $P(A^c) > 0$, then for any event B in S such that $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \tag{2.2.11}$$

The denominator $P(B)$in the equation can be computed,

Example: A paint-store chain produces and sells latex and semigloss paint. Based on long-range sales, the probability that a customer will purchase latex paint is 0.75. Of those that purchase latex paint, 60% also purchase rollers. But only 30% of semi gloss pain buyers purchase rollers. A randomly selected buyer purchases a roller and a can of paint. What is the probability that the paint is latex?

Solution:

L =The customer purchases latex paint., $P(L) = 0.75$

S =The customer purchases semigloss paint., $P(S) = 0.25$

R =The customer purchases roller.

$P(R|L)$ =0.6

$P(R|S)$ =0.3

$$P(R) = P(R|L)P(L) + P(R|S) + P(S) = 0.6 \cdot 0.75 + 0.3 + 0.25 = 0.53$$

$$P(L|R) = \frac{P(L \cap R)}{P(R)} \tag{2.2.12}$$

$$= \frac{P(R|L)P(L)}{P(R)} \tag{2.2.13}$$

$$= \frac{0.6 \times 0.7}{0.6 \times 0.75 \times 0.3 \times 0.25} = 0.857 \tag{2.2.14}$$

2.2.4 Axioms of Probability

Given an experiment and a sample space, S , the objective of probability is to assign to each event A a number P(A), called the probability of the event A, which will give a precise measure of the chance that A will occur. To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments should satisfy the following axioms (basic properties) of probability.

1. For every event A, $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. If A and B are mutually exclusive events, i.e $A \cap B$ then $P(A \cup B) = P(A) + P(B)$
4. If $A_1, A_2, A_3, A_4, \ldots A_n$ is a sequence of n mutually exclusive events, then, $P(A_1 \cup A_2 \cup A_3 \cdots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_n)$

From the above axioms the following preposition are derived,

- If $\varnothing$ is a set then $P(\varnothing) = 0$
- If A^c is a complement of an event A, then $P(A^c) = 1 - P(A)$

2.2.5 Some Rules of Probability

Additive Rule

1. Let $A_1, A_2, A_3, \ldots, A_n$ be events of the sample space, S. Then,
   - $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
   - $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$

   If events $A_1, A_2, A_3, \ldots, A_n$ are mutually exclusive, then,

   (a) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$

   (b) $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$

   (c) $P(A_1 \cup A_2 \cup A_3 \ldots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_n)$

Multiplicative Rule

If events $A_1, A_2, A_3, \ldots, A_n$ are events of the same sample space S, then

(a) $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1)$

(b) $P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_2 \cup A_1)$

2.2.6 Application of Counting Techniques

The classical definition of probability of an event $A, P(A)$ requires the knowledge of the number of outcomes of A and the total possible outcomes of the experiment,S . To find these outcomes we list such outcomes explicitly, which may be impossible if they are too many. Counting Techniques may be useful to determine the number of outcomes and compute $P(A)$. We shall examine three basic counting techniques, namely the **Multiplication Principle**, **Permutation** and **Combination**.

The Multiplication Principle

The Multiplication Principle, also known as the Basic Counting Principle states that:

1. If an operation can be performed in ways, and a second operation can be performed in n_1 ways and so on for k_{th} operation which can be performed in n_k ways, then the combined experiment or operations can be performed in $n_1 \cdot n_2 \cdot n_3 \ldots n_K$ ways. **For example:** A homeowner doing some remodeling requires the services of both a plumbing contractor and an electrical contractor. If there are 12 plumbing contractors and 9 electrical contractors available in the area, in how many ways can the contractors be chosen? If we denote the plumbers by $P_1, \ldots, P_{12}$ and the electricians by $Q_1, ..., Q_9$, then we wish the number of pairs of the form (P_i, Q_j). With $n_1 = 12$ and $n_2 = 9$, the product rule yields $N = (12)(9) = 108$ possible ways of choosing the two types of contractors.

Examples

R Tossing a coin has two possible outcomes and tossing a die has six possible outcomes. Then the combined experiment, tossing the coin and die together will result in $(2*6=12)$ twelve possible outcomes provided below:

$$H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6$$

R Another example is the number of different ways for a man to get dressed if he has 8 different shirts and 6 different pairs of trousers. The combination of the 8 different shirts and the six different pairs of trousers results in $(8*6=64)$ possible ways.

R In a certain examination paper, students are required to answer 5 out of 10 questions from Section A another 3 out of 5 questions from Section B and 2 out of 5 questions from Section C. In how many ways can the students answer the examination paper? **Solution:**

1. The number of ways of answering the questions in Section A: $10*9*8*7*6=30\ 240$.
2. The number of ways of answering the questions in Section B: $5*4*3=60$.
3. The number of ways of answering the question in Section C: $5*4=20$.
4. Hence the students can answer the question in the three sections in : $30240*60*20=36\ 280\ 000$

Application of the multiplication principle results in the other two counting techniques: *Permutation* and *Combination*, used to find the number of possible ways when a fixed number of items are to be picked from a lot without replacement.

2.2.7 Permutation of Objects

An ordered arrangement of objects is called a permutation. For example, the possible permutations of the letters A, B and C are as follows: $ABC, ACB, BAC, BCA, CAB, CBA$

Definitions:

1. The number of permutations of $n!$ distinct objects, taken all together is:

$$n! = n(n-1)(n-2)(n-3)\dots \ or \ ({}^{n}P_{n})$$

2. The number of permutations of n distinct objects taken k at a time is:

$${}^{n}P_{k} = \frac{n!}{n-k!}$$

where $k < n$

3. The number of permutations of n objects consisting of groups of which n_1 of the first group are alike,n_2 of the second group are alike and so on for the k^{th} group with objects which are alike is:

$$\frac{n!}{n_1!n_2!n_3!,\dots n_k!}$$

4. Circular Permutations: Permutations that occur when objects are arranged in a circle are called circular permutations. The number of ways of arranging different objects in a circle is given by

$$\frac{n!}{n} = (n-1)!$$

Examples

R The number of permutations of 10 distinct digits taken two at a time is:

$$^{10}P_2 = \frac{10!}{(10-2)!} = 10*9 = 90$$

R A company codes its customers by giving each customer an eight character code. The first 3 characters are the letters A, B and C in any order and the remaining 5 are the digits $1,2,3,4$ and 5 also in any order. If each letter and digit can appear only once then the number of customers the company can code is obtained as follows:

1. The first 3 letters can be filled in 3!
2. The next 5 digits can be filled in 5!
3. Then the required number $3!*5! = 720$

R The number of permutations of the letters of the word, POSSIBILITY, which contains 3I's and 2S's is ?

R The number of arrangements of the letters of the word, ADDING, if the two letters D and D are together (ADDING)?

R In how many ways can 4 boys and 2 girls seat themselves in a row if

- the 2 girls are to sit next to each other?
- the 2 girls are not to sit next to each other?

Combination of Objects

A Combination is a selection of objects in which the order of selection does not matter. Definition: The number of ways in which objects can be selected from distinct objects, irrespective of their order is defined by:

$$^{n}C_{k} = \frac{n!}{(n-k)!k!}$$

where $k > n$.

Questions

R Find the number of ways in which a committee of 4 can be chosen from 6 boys and 5 girls if it must

1. consist of 2 boys and 2 girls
2. consist of at least 1 boy and 1 girl.

R A box contains 6 red, 3 white and 5 blue balls. If three balls are drawn at random, one after the other without replacement, find the probability that

1. all are red
2. 2 are red and 1 is white
3. at least 1 is red
4. 1 of each colour

R If the probability of achieving monthly production targets at Goldfields Ghana Limited, (A), and Ashanti (Obuasi), (B), are 0.8 and 0.9 respectively, what is $P(A \cap B)$?

R The Credit Manager at SSB collects data on 100 of her customers. Of the 60 men, 40 have credit cards (C). Of the 40 women, 30 have credit cards (C).Ten of the men with credit cards have balances (B), whilst 15 of the women have balances (B). The Credit Manager wants to determine the probability that a customer selected at random is:

1. A woman with credit card
2. A man with a balance.

R The probability that a mining company will make profit at an annual production rate of $5000t/yr$ is 0.7 if the gold price is $\$660/oz$. If the gold price goes below $660/oz$ the probability will fall to 0.40. The current world politics indicates that there is a 50% probability that the dollar will be strong and gold price will fall below $\$660/oz$. If:

A: Gold price falls below $\$660/oz$

B: The mine is profitable.

1. What is the probability that both A and B occur?
2. What is the probability that either A or B will occur?

R A coin is tossed twice. What is the probability that at least one head occurs?

R If a player picks 5 cards, find the probability of holding 2 aces and 3 jacks.

3. Random Variables and Distribution

3.1 Introduction

Having worked through this chapter the student will be able to:

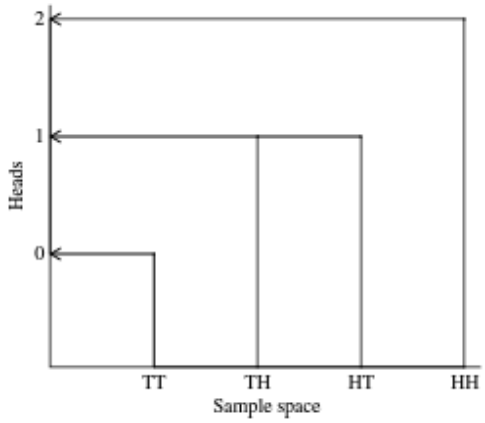
- Discuss random variables
- Determine probabilities from probability density functions
- Determine probabilities from cumulative distribution functions and cumulative distribution functions from probability density functions, and the reverse

3.2 Random Variable

R **Probabilistic Experiment:** A probabilistic experiment is some occurrence such as the tossing of coins, rolling dice,or observation of rainfall on a particular day where a complex natural background leads to a chance outcome.

R **Random variable:** A random variable is a function that maps events defined on a sample space into a set of values. Several different random variables may be defined in relation to a given experiment. Thus, in the case of tossing two

coins the number of heads observed is one random variable, the number of tails is another, and the number of double heads is another. The random variable "number of heads" associates the number 0 with the event TT, the number 1 with the events TH and HT, and the number 2 with the event HH. The Figure below illustrates this mapping.



R **Variate:** In the discussion of statistical distributions it is convenient to work in terms of variate. A variate is a generalization of the idea of a random variable and has similar probabilistic properties but is defined without reference to a particular type of probabilistic experiment. A variate is the set of all random variables that obey a given probabilistic law. The number of heads and the number of tails observed in independent coin tossing experiments are elements of the same variate since the probabilistic factors governing the numerical part of their outcome are identical. A multivariate is a vector or a set of elements,each of which is a variate.A matrix variate is a matrix or two-dimensional array of elements, each of which is a variate. In general, dependencies may exist between these elements.

R **Random number:**A random number associated with a given variate is a

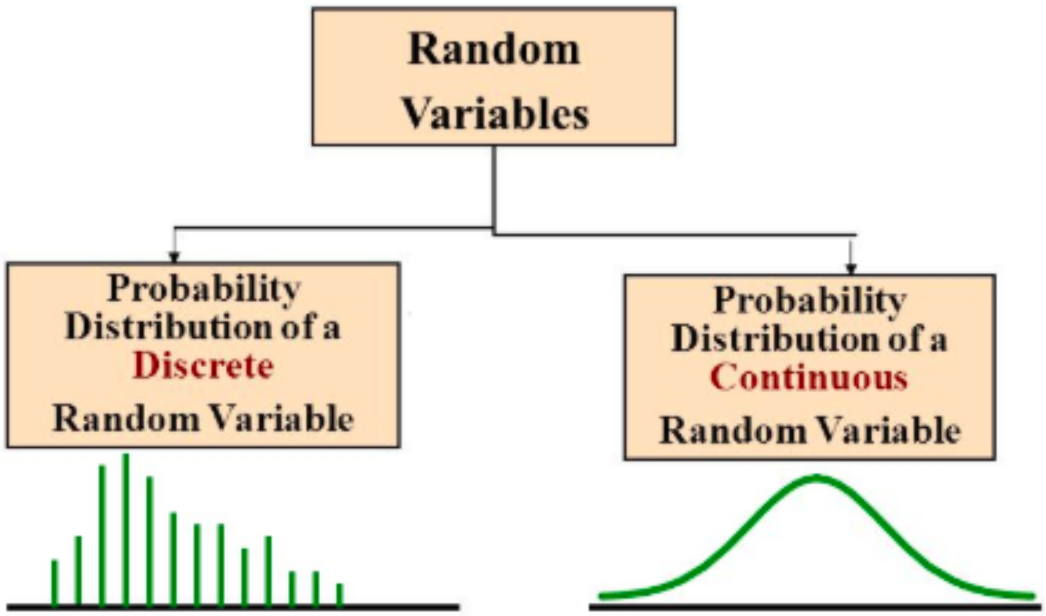
number generated at a realization of any random variable that is an element of that variate.

3.2.1 Types of Random Variables

There two types of random variables, The two random variables in the above examples are representatives of the two types of random variables that we will consider. These definitions are not quite precise, but more examples should make the idea clearer.

R **Discrete Random Variable:**A random variable X is discrete if the values it can take are separated by gaps. For example, X is discrete if it can take only finitely many values *(for example, the number of nuclear decays which take place in a second in a sample of radioactive material– the number is an integer but we can't easily put an upper limit on it.)*

R **Continuous Random Variable:** A random variable is continuous if there are no gaps between its possible values. In the first example, the height of a student could in principle be any real number between certain extreme limits. A random variable whose values range over an interval of real numbers, or even over all real numbers, is continuous. *In general, quantities such as pressure, height, mass, weight, density, volume, temperature, and distance are examples of continuous random variables*



We begin by considering discrete random variables.

3.2.2 Discrete Probability Distribution Variable

Cumulative Distribution Function (CDF)

Given a discrete random variable X, and its probability distribution function $P(X = x) = f(x)$, we define its cumulative distribution function, CDF, as:

$$F(x) = P(X \leq k)$$

where,

$$F(x) = P(X \leq x) = \sum_{t=x_{min}}^{x} P(X = t).$$

Properties of the CDF

The CDF has the following properties

- F(x) is non decreasing

- $\lim_{x\to-\infty} F(x) = 0$; $\lim_{x\to\infty} F(x) = 1$
- $F(x)$ is continuous from the right i.e ($\lim_{x\to 0^+} = F(x)$ for all x)

R *Question:* A discrete random variable X whose probability distribution function is:

$$P(X = x) = \frac{x}{15} \quad x \in 1,2,3,4,5 \tag{3.2.1}$$

Find $F(3)$, in other words: find $P(X \leq 3)$.

Solution

We use the cumulative distribution function and state:

$$P(X \leq 3) = \sum_{i=1}^{3} P(X = t)$$

That is:

$$P(X \leq 3) = P(X = 1) + P(X = 2) + (X = 3)$$

Using the fact that $P(X = x) = \frac{x}{15}$ we find:

$$P(X \leq 3) = (X = 1) + (X = 2) + (X = 3)$$

$$= \frac{1}{15} + \frac{2}{15} + \frac{3}{15}$$

$$P(X \leq 3) = \frac{6}{15}$$

Finally we can state $P(X \leq 3) = \frac{6}{15} = 0.4$.

R *Examples:*

1 A discrete random variable X has probability distribution function defined by:

$$P(X = x) = \frac{x^2}{30}$$

Where $x = \{1,2,3,4\}$.

Calculate the probability that $X \leq 2$.

2 A discrete random variable X has probability distribution function defined by:

$$f(x) = \frac{x}{15}$$

Where x = $\{1,2,3,4,5\}$.

Calculate the probability that $X < 4$.

Probability Mass Function (PMF), Discrete density function (DDF), Probabiltiy function

Let X be a discrete random variable, abd suppose that the possible values that it can assume are given by $x_1, x_2, \ldots x_n$ arranged in some order. Suppose alsp that these values are assumed with probabilities given by

$$P(X = x_k) = f(x_k) \quad k = 1,2,\ldots \tag{3.2.2}$$

Properties of the PMF

- $f(x)0$
- $\sum_x f(x) = 1$

Examples

R *Questions 1:* Show that the following can be probability mass function and explain your answers.

1. $f(x) = \frac{1}{5}$ where $x = 0, 1, 2, 3, 4, 5$
2. $f(x) = \frac{x^2}{30}$ where $x = 0, 1, 2, 3, 4$
3. $f(x) = \frac{x-2}{5}$ where $x = 1, 2, 3, 4, 5$

R *Question 2:* Suppose that a pair of fair coins is tossed and let the random variable X denote the number of heads minus the number of tails.

1. Obtain the probability distribution for X
2. Construct a graph for this distribution
3. Find $P(X = 1), f(-2), P(X \leq 2), P(-2 \leq X < 2), P(X < 0)$

R *Question 3:* A shipment of 8 similar microcomputers to a retail outlet contains 3 that are defective. If a school makes random purchase of 2 of these computers, find the probability distribution for the number of defectives

R *NOTE:* A probability distribution is a display of all possible outcomes of an experiment along with the probabilities of each outcome. In fact, it is a list of all possible outcomes of some experiment and the probability associated with each outcome

3.2.3 Continuous Probability Distribution Variable

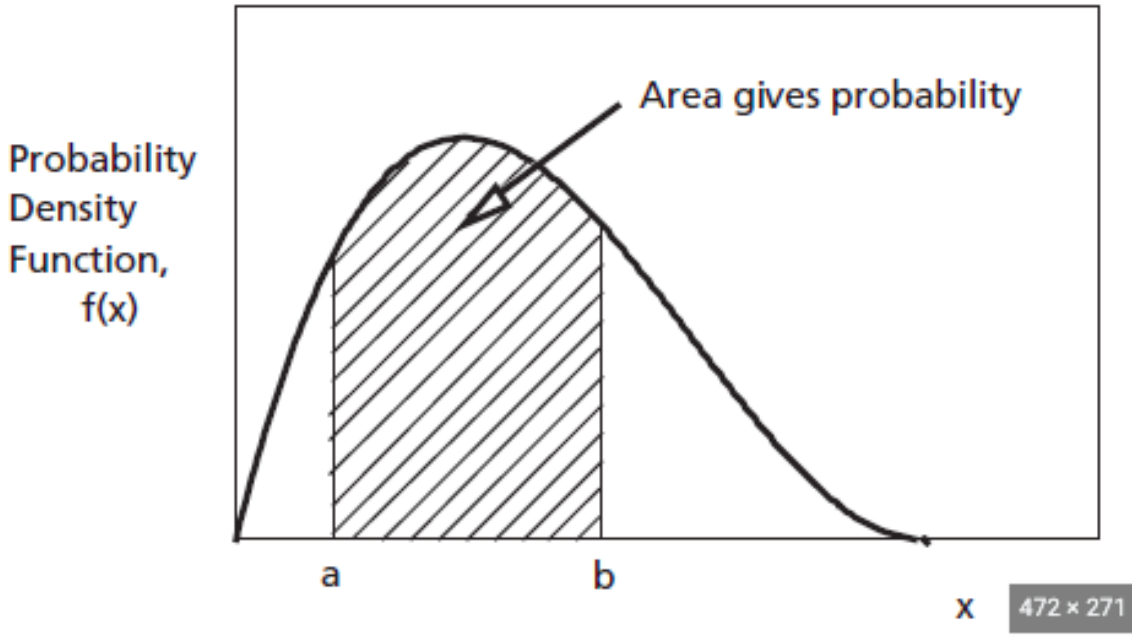
If X is a continuous random variable, the probability that takes on any one particular value is generally zero. Therefore, we cannot define a continuous random variable in

the same way as for a discrete random variable. In order to arrive at a probability distribution for a continuous random variable we note that the probability that lies between two different values is meaningful. Thus, a continuous random variable is the type whose spaces are not composed of a countable number of points but takes on values in some interval or a union of intervals of the real line.

3.2.4 Probability Density Function (PDF)

If the set of all possible values of a random variable X , takes on an uncountable infinite number of values or values in some interval or a union of intervals of the real line, it is called a continuous random variable if there exists a function f , called probability density function of X such that the following properties;

1. $f(x) \leq 0$ (Non negative)
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a \leq x \leq b) = \infty_{-\infty}^{\infty} f(x)dx$ where $-\infty \leq a \leq b \leq \infty$



Examples Suppose that the error in the reaction temperature, in $^o($ for a controlled laboratory experiment is a continuous random variable X having the probability density function:

$$f(x) = \begin{cases} \frac{x^2}{3} & \text{if } -1 < x2 \\ 0 & \text{if } xelsewhere \end{cases} \tag{3.2.3}$$

Verify,

- If $f(x)$ is a PDF
- $P(0 < x \geq 1)$

R Find the constant C such that the function below is a probability density function:

$$f(x) = \begin{cases} cx^2 & \text{if } 0 < x3 \\ 0 & \text{if } xelsewhere \end{cases} \tag{3.2.4}$$

Compute $P(1 < x < 2)$

R For each of the following functions, find the constant c so that $f(x)$ is a PDF of a random variable X.

1. $f(x) = 4xc,\ 0 \leq x \leq 1$
2. $f(x) = c\sqrt{4}, 0 \leq x \leq 4$

R The probability density function of a continuous random variable X is given by

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \leq x \leq 2 \\ 0 & x > 2 \end{cases}$$

Find the cumulative distribution function and sketch its graph.

4. Special Distribution

4.1 Introduction

Having worked through this chapter the student will be able to:

- Understand the assumptions for each of the discrete and continuous probability distributions presented.
- Select an appropriate discrete and continuous probability distribution to calculate probabilities in specific applications.
- Calculate probabilities, determine means and variances for each of the discrete and continuous probability distributions presented.

4.1.1 Discrete Probability Distribution

Bernoulli Distribution:

A single trial of an experiment may result in one of the two mutually exclusive outcomes such as defective and non-defective, dead or alive, yes or no, male or female, etc. Such a trial is called and a sequence of these trials form a process, satisfying the following conditions:

- Each trial results in one of the two mutually exclusive outcomes, success and failure.
- The probability of a success, p remains constant, from trial to trial. The probability of failure is denoted by $q = 1 - p$
- The trials are independent. That is, the outcome of any particular trial is not affected by the outcome of any other trial.

Definition

A random variable, is said to have a Bernoulli distribution if it assumes the values 0 and 1 for the two outcomes. The probability distribution for the success in the trial p is defined by

$$P(x) = p^x(1-p)^{1-x}, x = 0 \ or \ 1 \tag{4.1.1}$$

and $0 < p < 1$. where the mean and variance of the distribution are as follows:

- $\mu = E(x) = p$;
- $\sigma = Var(X) = p(1-p)$

An important distribution arising from counting the number of successes in a fixed number of independent Bernoulli trials is the Binomial distribution.

R Example 35: An urn contains 5 red and 15 green balls. Draw one ball at random from the urn. Let X=1 if the ball drawn is red, and X=0 if a green ball is drawn. Obtain;

- the p.d.f. of X,
- mean of X and
- variance of X.

The Binomial Distribution

The binomial distribution is a discrete probability distribution, where the experiment is repeated n times under identical conditions and each of the n trials is independent of each other which results in one of the two outcomes. Thus, in the event of independent trials (often called Bernoulli trials) let p be the probability that an event will happen (success) and $q = 1 - p$ the probability that the event will fail in any single trial. Such experiments are called Binomial experiments and the probability that the event will happen exactly times in trials is given by the probability function:

$$f\Big(x\Big) = Pr\Big(X = x\Big) = \left[\begin{array}{c} n \\ x \end{array}\right] p^x q^{n-x} = \frac{n!}{x!\Big(n-x\Big)!} p^x q^{n-x}$$

where the random variable X denotes the number of success in n trials and $x = 0, 1, 2, 3, 4, 5 \ldots$

The shape of the distribution depends on the two parameters n and p.

1. when $p < 0.5$ and n is small, the distribution will be skewed to the right.
2. when $p > 0.5$ and n is small, the distribution will be skewed to the left
3. when $p = 0.5$ the distribution will be symmetric.
4. In all cases, as n gets larger the distribution gets closer to being a symmetric, bell-shaped distribution.

Properties

1. Mean = np
2. Variance =npq
3. Standard Deviation = $\sqrt{npq}$

R If 20% of the bolts produced by a machine are bad. Determine the probability that out of 4 bolts chosen at random.

- one is defective

- none is defective
- at most 2 bolts will be defective.

R Suppose that it is known that 30% of a certain population is immune to some disease. If a random sample of 10 is selected from this population. What is the probability that it will contain exactly 4 immune persons?

R From the experiment "toss four coins and count the number of tails" what is the variance of X?

R Roll a fair 6 – sided die 20 times and count the number of times that 6 shows up. What is the standard development of your random variable?

R The following data are the number of seeds germinating out of 10 on damp filter paper for 80 sets of seeds. Fit a binomial distribution to these data.

x	0	1	2	3	4	5	6	7	8	9	10	Total
f	6.89	19.14	23.94	17.74	8.63	2.88	0.67	0.1	0.01	0.00	0.00	80

Negative Binomial Distribution (Pascal's Distribution)

Let us consider an experiment in which the properties are the same as those listed for a binomial experiment with the exception that the trials will be repeated until a fixed number of successes occur. Therefore, instead of finding the probability of x successes in n trials, where n is fixed, we are now interested in the probability that the. . . kth success occurs on the xth trial. Experiments of this kind are called 'negative binomial experiments'. (Walpole and Myres, 1993). The number X of trials to produce

k successes in a negative binomial experiment is called a "negative binomial random variable" and its probability distribution is called the "negative binomial distribution". Since its probabilities depend on the number of successes desired and the probability of success on a given trial, we shall denote them by the symbol $b^*(x;k,p)$. For the general formula $b^*(x;k,p)$, consider the probability of a success on the trial preceded by $k-1$ successes and $x-k$ failures in some specified order. The probability for the specified order ending in success is $p^{k-1}q^{x-k}p = p^kq^{x-k}$. The total number of sample points in the experiment ending in success, after the occurrence of $k-1$ successes and $x-k$ failures in any order is equal to the number of partitions of $x-1$ trials into two groups with $k-1$ successes corresponding to one group and $x-k$ failures corresponding to the other group. This number is given by the term $\binom{x-1}{k-1}$. each mutually exclusive and occurring with equal probability p^kq^{x-k}. We obtain the general formula by multiplying p^kq^{x-k} by $\binom{x-1}{k-1}$. In other words:

$$b^*\big(x;k,p\big) = \binom{x-1}{k-1}p^{k-1}q^{x-k}p = \binom{x-1}{k-1}p^kq^{x-k} \qquad x = k,k+1,.....$$

p = probability of success

q = (1-p) = probability of failure

x = total number of trials on which the k^{th} success occurs.

Areas of application of negative binomial distribution include many biological situations such as death of insects, number of insect bites per fruit (e.g. mango).

Examples

R Consider an exploration company that is determined to discover two new fields in a virgin basin it is prospecting, and will drill as many holes as required to achieve its goal. We can investigate the probability that it will require 2, 3,

4,. . . , n exploratory holes before two discoveries are made. The same conditions that govern the binomial distribution may be assumed, except that the number of trials is not fixed.

R Find the probability that a person tossing three coins will get either all heads or all tails for the second time in the fifth toss?

Geometric Distribution

The geometric distribution is a special case of the negative binomial distribution for which k = 1. This is the probability distribution for the number of trials required for a single success. Thus:

$$g(x;p) = pq^{x-1}$$

Examples

R In a certain theodolite manufacturing process, it is known that on the average, 1 in every 100 is defective. What is the probability that the fifth item inspected is the first defective theodolite found?

Poisson Distribution

Experiments yielding numerical values of a random variable (x), the number of successes occuring during a given time interval or in a specified region, are often called poisson experiments. The given time interval may be of any length, such as a minute, a day, a week, a month or even a year. Hence, a poisson experiment might generate observations for the random variable representing the number of telephone calls per hour received by an office, the number of days school is closed due to snow during the winter, or the number of postponed games due to rain during a basketball season. The specified region could be a line segment, an area, a volume or perhaps a material.

In this case, might represent the number of field mice per acre, the number of bacteria in a given culture, or the number of typing errors per page.

The Poisson process:

A Poisson experiment is derived from the Poisson process and possesses the following properties:

1. The number of successes occurring in one time interval or specified region are independent of those occurring in any other disjoint time interval or region of space.
2. The probability of a single success occurring during a very short time interval or in a small region is proportional to the length of the time interval or the size of the region and does not depend on the number of successes occuring outside this time interval or region.
3. The probability of more than one success occuring in such a short time interval or falling in such a small region is negligible.

The probability distribution of the Poisson random variable is called the Poisson distribution and is denoted by $P(x;\mu)$ since its values depend only on μ , the average number of successes occuring in the given time interval or specified region. This formula is given by the definition below:

Definition: The probability distribution of the Poisson random variable , representing the number of successes occurring in a given time interval or specified region is given by:

$$P(x;\mu) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 1,2,3,4,5,\ldots$$

where μ is the average number of successes occuring in the given time interval or specified region and $e = 2.7183$

Theorem: The mean and variance of the Poisson distribution both have the value μ.

Examples

R Suppose that an urn contains 100,000 marbles and 120 are red. If a random sample of 1000 is drawn what are the probabilities that 0, 1, 2, 3, and 4 respectively will be red.

R A hospital administrator, who has been studying daily emergency admissions over a period of several years, has come to the conclusion that they are distributed according to the Poisson law. Hospital records reveal that emergency admissions have averaged three per day during this period. If the administrator is correct in assuming a Poisson distribution. Find the probability that

1. exactly two emergency admissions will occur on a given day.
2. No emergency admissions will occur on a particular day.
3. Either 3 or 4 emergency cases will be admitted on a particular

R Fit a Poisson distribution to the following data which gives the number of yeast cells per square for 400 squares

No. of cells per square (x)	0	1	2	3	4	5	6	7	8	9	10	Total
No. of squares (f)	103	143	98	42	8	4	2	0	0	0	0	400

R In a manufacturing process in which glass is being produced, defects or bubbles occur, occasionally rendering the pieces undesirable for marketing. If it is known that on the average 1 in every 1000 of these items produced have one or more bubbles. What is the probability that a random sample of 8000 will yield fewer than 7 items possessing bubbles?

R 4. Suppose it is known that the probability of recovery from a certain disease is 0.4. If 15 people are stricken with the disease what is the probability that

1. or more will recover?
2. 4 or more will recover?
3. at least 5 will recover?
4. fewer than three recover?

4.1.2 Continuous Probability Distribution

Normal Distribution

The graph of the normal distribution which is a bell-shaped smooth curve approximately describes many phenomena that occur in nature, industry and research. In addition, errors in scientific measurements are extremely well approximated by a normal distribution. Thus, the normal distribution is one of the most widely used probability distributions for modelling random experiments. It provides a good model for continuous random variables involving measurements such as time, heights/weights of persons, marks scored in an examination, amount of rainfall, growth rate and many other scientific measurements.

Definition: The probability density function for the normal random variable X which is simply called normal distribution is defined by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{1-\mu}{\sigma})^2} \tag{4.1.2}$$

where $\sigma > 0$, $\mu > 0$ and $-\infty < x < \infty$.

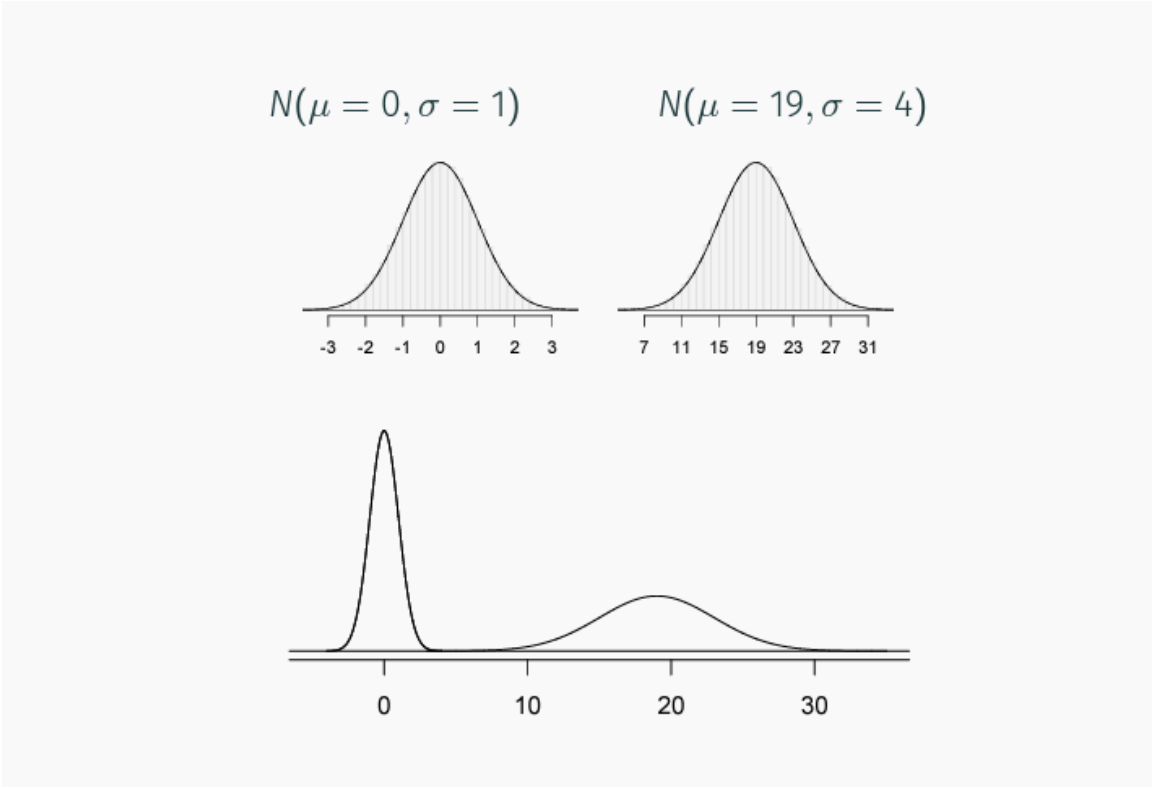
Properties

- Mean= $E(x) = \mu$
- Variance= σ^2

Reasons for importance

1. Many data sets well-modelled by normal distribution; for example heights and weights.
2. Many data sets can be transformed to near normality; for example $log(income)normal$.
3. Central limit theorem– sample means $\bar{x}$ normal for large sample sizes
4. Many distributions approach normality in some limit.

Standard normal distribution



A measure of the number of standard deviations the data falls above or below the mean.

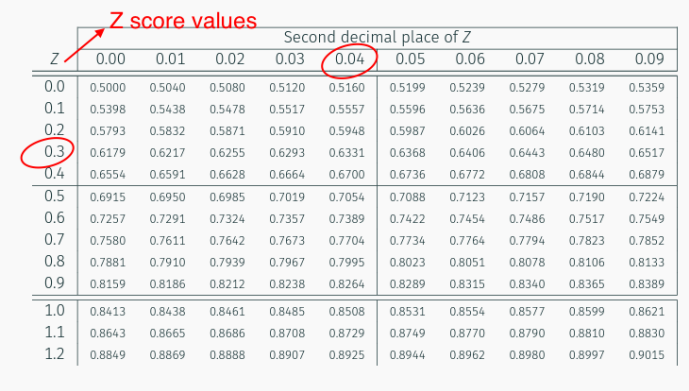
$$Z = \frac{observation - mean}{SD} \tag{4.1.3}$$

Wecan calculate Z scores for distributions of any shape, but with normal distributions we use Z scores to calculate probabilities. Observations that are more than 2 SD away from the mean are typically considered unusual. Another reason we use Z scores is if the distribution of X is nearly normal then the Z scores of X will have a Z distribution (unit normal). Note that the Z distribution is a special case of the normal distribution where $mean(\mu) = 0$ and $standarddeviation(\sigma) = 1$. Linear transformations of normally distributed random variable are also normally distributed. Hence, if

$$Z = \frac{X - \mu}{\sigma}$$

where X $N(\mu, \sigma)$. Here,

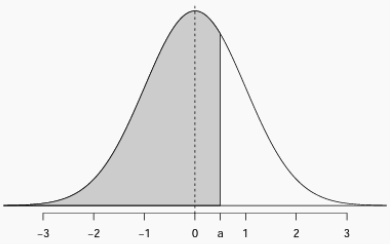
Calculating Probabilities - Z Table



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

The area under the unit normal curve from $-\infty$ to a is given by

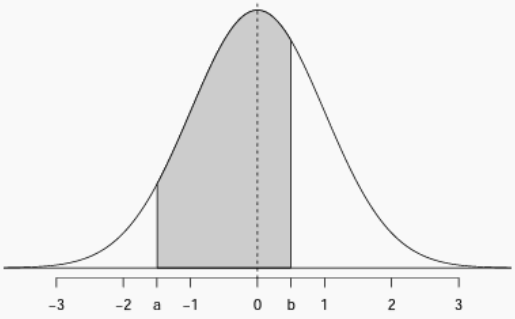
$$P(Z < a) = \omega(a)$$



*These left tail probabilities are sometimes called percentiles

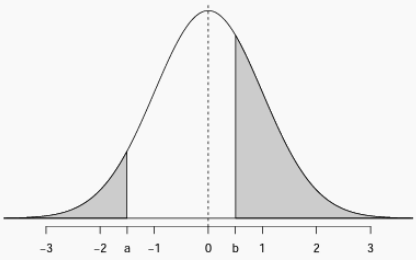
The area under the unit normal curve from a to b where $a \leq b$ is given by

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$



The area under the unit normal curve outside of a to b where $a \leq b$ is given by

$$P(a \geq Z \text{ or } Z \geq b) = \Phi(a) + (1 - \Phi(b)) = 1 - (\Phi(b) - \Phi(a))$$

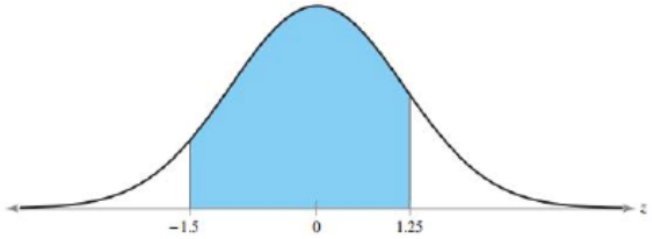


Examples

R Find the area under the standard normal curve between $z = -1.5$ and $z = 1.25$.

Solution

The area under the standard normal curve between $z = -1.5$ and $z = 1.25$ is shown



From the Standard Normal Table, the area to the left of $z = 1.25$ is 0.8944 and the area to the left of z = -1.5 is 0.0668. So, the area between z = -1.5 and z = 1.25 Area = 0.8944 - 0.0668 = 0.8276

Interpretation: So, 82.76% of the area under the curve falls between z = -1.5 and z = 1.25.

R A survey indicates that people use their cellular phones an average of 1.5 years before buying a new one. The standard deviation is 0.25 year. A cellular phone user is selected at random. Find the probability that the user will use their current phone for less than 1 year before buying a new one. Assume that the variable x is normally distributed.

Solution: The graph shows a normal curve with $\mu = 1.5$ and $\sigma = 0.25$ on a shaded area for x less than 1. The z-score that corresponds to 1 year is

$$z = \frac{1 - 1.15}{0.25} = -2$$

The Standard Normal Table shows that P(z < - 2) = 0.0288

Interpretation: The probability that the user will use their cellular phone for less than 1 year before buying a new one is 0.0228

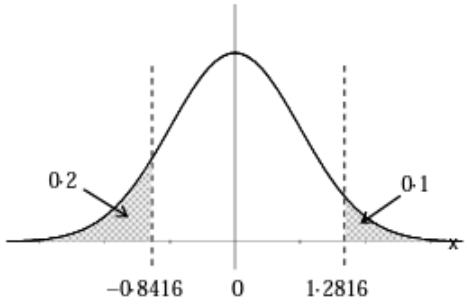
R The results of an examination were Normally distributed. 10% of the candidates had more than 70 marks and 20% had fewer than 35 marks. Find the mean and standard deviation of the marks.

Solution:

First we need the values from the tables

$\Rightarrow \quad \Phi(-0{\cdot}8416) = 0{\cdot}2,$

and $1 - \Phi(1{\cdot}2816) = 0{\cdot}1$



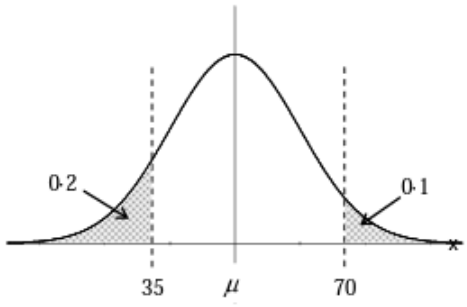
Using $Z = \dfrac{X-\mu}{\sigma}$ we have

$-0{\cdot}8416 = \dfrac{35-\mu}{\sigma}$

$\Rightarrow \quad \mu = 35 + 0{\cdot}8416\sigma$

and $1{\cdot}2816 = \dfrac{70-\mu}{\sigma}$

$\Rightarrow \quad \mu = 70 - 1{\cdot}2816\sigma$



$\Rightarrow \quad \sigma = 16{\cdot}5$ and $\mu = 48{\cdot}9$ to 3 S.F.

simultaneous equations

R The weights of chocolate bars are normally distributed with mean 205 g and standard deviation 26 g. The stated weight of each bar is 200 g.

1. Find the probability that a single bar is underweight
2. Four bars are chosen at random. Find the probability that fewer than two bars are underweight.

Solution: (a) Let W be the weight of a chocolate bar, W $N(205, 26^2)$ Then

$$Z = \frac{W-\mu}{\sigma} = \frac{200-205}{2.6} = -1923077$$

$P(W < 200) = P(Z < -192) = 1 - \Phi(192) = 1 - 09726$ **Interpretation: probability of an underweight bar is 00274.**

(b) We want the probability that 0 or 1 bars chosen from 4 are underweight.

Let U be underweight and C be correct weight

For $P(1 underweight) = P(CCCU) + P(CCUC) + P(CUCC) + P(UCCC)$

$= 4 \times 00274 \times 09726^3 = 01008354753$

$= 4 \times 00274 \times 09726^3 = 01008354753$

For $P(0\ underweight) = 0.9276^4 = 0.7404$

the probability that fewer than two bars are underweight is 0.841

4.1.3 Mathematical Expectations

A very important concept in probability and statistics is that of mathematical expectation, expected value or briefly expectation of a random variable. The expectation of X is very often called the mean of X and is denoted by µx or simply µ when a particular random variable is understood. This expected value of x gives a simple value, which acts as a representative, or average of the value of x and for this reason it is often called a measure of central tendency. Consider that the random variable x has the values $x_1, x_2, x_3, \ldots$. The mean or expected value of x is:

$$\mu = E(x) = \sum x f(x) \qquad for\ discrete\ case \tag{4.1.4}$$

$$\mu = E(x) = \int_{-\infty}^{\infty} f(x)dx \quad for\ continuous\ case \tag{4.1.5}$$

5. Estimations

5.0.1 Introduction

The basic reasons for the need to estimate population parameters from sample information is that it is ordinarily too expensive or simply infeasible to enumerate complete populations to obtain the required information. The cost of complete censuses may be prohibitive in finite populations while complete enumerations are impossible in the case of infinite populations. Hence, estimation procedures are useful in providing the means of obtaining estimates of population parameters with desired degree of precision. We now consider estimation, the first of the two general areas of statistical inference. The second general area is hypothesis testing which will be examined later. The subject of estimation is concerned with the methods by which population characteristics are measured from sample information. The objectives are to present:

1. properties for judging how well a given sample statistic estimates the parent population parameter.
2. several methods for estimating these parameters.

There are basically two types of estimation: point estimation and interval estimation. In point estimation, a single sample statistic, such as $\bar{X}$, s , or p is calculated from the sample to provide a best estimate of the true value of the corresponding population parameter such as μ, σ or p . Such a statistic is termed a point estimator. The function or rule that is used to estimate the value of a parameter is called an estimator. An estimate is a particular value calculated from a particular sample of observations. On the other hand, an interval estimate consists of two numerical values defining an interval which, with varying degrees of confidence, we feel includes the parameter being estimated.

5.0.2 Properties of a Point Estimator

Unbiasedness:

If the expected value or mean of all possible values of a statistic over all possible samples is equal to the population parameter being estimated, the sample statistic is said to be unbiased. That is, if the expected value of an estimator is equal to the corresponding population parameter, the estimator is unbiased

R The sample mean is an unbiased estimator of the population mean

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \mu \tag{5.0.1}$$

Efficiency:

The most efficient estimator among a group of unbiased estimators is the one with the smallest variance. This concept refers to the sampling variability of an estimator.

Consistency:

An estimator is consistent if as the sample size increases, the probability increases that the estimator will approach the true value of the population parameter. Alternatively,

an estimator is consistent if it satisfies the following conditions:

1. $Var(\bar{\theta}) \longrightarrow 0$ as $n \to \infty$
2. becomes unbiased as $n \to \infty$

5.0.3 Interval Estimation

For most practical purposes, it would not suffice to have merely a single value estimate of a population parameter. Any single point estimate will be either right or wrong. Therefore, instead of obtaining only a single estimate of a population parameter, it would certainly seem to extremely useful and perhaps necessary to obtain two estimators, say $\bar{X}_1 and \bar{X}_2$, and say with some confidence that the interval between $\bar{X}_1$ and $\bar{X}_2$ includes the true mean μ . Thus, an interval estimate of a population parameter θ is a statement of two values between which it is estimated that the parameter lies. We shall be discussing the construction of confidence intervals as a means of interval estimation. The confidence we have that a population parameter, θ , will fall within some confidence interval will equal (1− , where α is the probability that the interval does not contain θ (i.e. the probability , is an allowance for error). To construct a 95% confidence interval $\alpha = 0.05$. That is, the probability is 0.05 that the value θ will not lie within the interval.

Note that,

$$\alpha + (confidence\ interval) = 1$$

The larger the confidence interval, the smaller the probability of error α for the interval estimator

Confidence Interval For μ and (σ) unknown

A confidence interval is constructed on the basis of sample information. It also depends on the size of . Assume the population variance σ^2 is known and the population is

normal, then $100(1-\alpha)\%$ the percent C.I for μ is given by

$$\bar{X} - Z_{a/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_{a/2}\left(\frac{\sigma}{\sqrt{n}}\right) \tag{5.0.2}$$

Simply written as

$$\bar{X} - Z_{a/2}\left(\frac{\sigma}{\sqrt{n}}\right) \tag{5.0.3}$$

where $Z_{a/2}$ is the Z value representing an area $a/2$ to the right and left tails of the standard normal probability distribution.

R The yield of a chemical process is being studied. From previous experience yield is known to be normally distributed and . The past five days of plant operation have resulted in the following percent yields: 91.6, 88.75, 90.8, 89.95, and 91.3. Find a 95% two-sided confidence interval on the true mean yield.

Solution $n = 5$, $\sigma = 3$, $\bar{x} = 90.48$, $Z_{a/2} = Z_{0.025} = 1.96$

$$\begin{aligned} 90.48 - Z_{0.025}\left(\frac{3}{\sqrt{5}}\right) \leq \quad & \mu \quad \leq 90.48 + Z_{0.025}\left(\frac{3}{\sqrt{5}}\right) \\ 90.48 - 1.96(1.3416) \leq \quad & \mu \quad \leq 90.48 + 1.96(1.3416) \\ 87.8505 \leq \mu \leq 93.1095 & \end{aligned}$$

R A manufacturer produces piston rings for an automobile engine. It is known that ring diameter is normally distributed with millimeters. A random sample of 15 rings has a mean diameter of millimeters.

1. Construct a 99% two-sided confidence interval on the mean piston ring diameter.
2. Construct a 95% confidence interval on the mean piston ring diameter.

R ASTM Standard E23 defines standard test methods for notched bar impact testing of metallic materials. The Charpy V-notch (CVN) technique measures impact energy and is often used to determine whether or not a material experiences a ductile-to-brittle transition with decreasing temperature. Ten measurements of impact energy (J) on specimens of A238 steel cut at 60°C are as follows: 64.1, 64.7, 64.5, 64.6, 64.5, 64.3, 64.6, 64.8, 64.2, and 64.3. Assume that impact energy is normally distributed with . We want to find a 95% CI for , the mean impact energy. The resulting 95% CI?

Confidence Interval For $\mu(\sigma\ unknown/n \geq 30)$

In practice, the standard deviation σ of a population , is not likely to be known. When σ is unknown and n is 30 or more, we proceed as before and estimate σ with the sample standard deviation s. the resulting $1-\alpha$ large sample confidence interval for becomes

$$\bar{X} - Z_{a/2}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + Z_{a/2}\left(\frac{s}{\sqrt{n}}\right) \tag{5.0.4}$$

R 1. A sample of 40 ten-year-old girls gave a mean weight of 71.5 and standard deviation of 12 pounds respectively. Assuming normality, find the

1. 90% confidence interval for μ.
2. 95% confidence interval for μ.
3. 99% confidence interval for μ.

R A hospital administrator took a sample of 45 overdue accounts from which he computed a mean of \$250 and a standard deviation of \$75. Assuming that the amounts of all overdue accounts are normally distributed. Find the

1. 90% confidence interval for μ.
2. 95% confidence interval for μ.
3. The 99% confidence interval for μ.

Confidence Interval For $\mu(\sigma\ unknown\ /n < 30)$

When the is not known and the sample size is small, the procedure for interval estimation of population mean is based on a probability distribution known as the student t-distribution. When the population variance is unknown, and the sample size is small, the correct distribution for constructing a confidence interval for is the t-distribution. Here, an estimate s must be calculated from the sample to substitute for the unknown population standard deviation. The t-distribution is used such that

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

where

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{(n-1)}} \tag{5.0.5}$$

The t-distribution is based on the assumption that the population is normal. A $100(1-\alpha)\%$ CI for the population mean, with the population normal and unknown is given by

$$\bar{x} - t_{a/2,v}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{a/2,v}\left(\frac{s}{\sqrt{n}}\right) \tag{5.0.6}$$

where $v = n-1$. Notice that a requirement for the valid use of the t-distribution is that the sample must be drawn from a normal distribution.

R A sample of 25 ten-year-old boys yielded a mean weight and standard deviation of 73 and 10 pounds respectively. Assuming a normally distributed population, find 90, 95 and 99 percent confidence intervals for the mean of the population from which the sample came.

Solution: $n = 25$, $\bar{x} = 73$ and $s = 10$

$$\begin{aligned} \bar{x} &\pm t_{a/2}\left(\frac{10}{\sqrt{25}}\right) \\ \bar{x} &\pm t_{0.05}\,(2) \\ 75 &\pm 1.711\,(2) \\ &\quad (69.578, 76.422) \end{aligned}$$

Summary:

Confidence Interval for μ

Population Mean μ	Confidence Interval
Sample size	
Large	
$\rightarrow \sigma$ assumed known	$\bar{X} \pm Z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$
$\rightarrow \sigma$ unknown (*estimated by s*)	$\bar{X} \pm Z_{\alpha/2}\left(s/\sqrt{n}\right)$
Small	
$\rightarrow \sigma$ assumed known	$\bar{X} \pm Z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$
$\rightarrow \sigma$ unknown (*estimated by s*)	$\bar{X} \pm t_{\alpha/2}\left(s/\sqrt{n}\right)$

5.0.4 Confidence Interval For A Population Proportion

It is often necessary to construct confidence intervals on a population proportion. For example, suppose that a random sample of size n has been taken from a large (possibly infinite) population and that $X(\leq n)$ observations in this sample belong to a class of interest. Then $\bar{P} = X/n$ is a point estimator of the proportion of the population p that belongs to this class. Note that n and p are the parameters of a binomial distribution. Furthermore, we know that the sampling distribution of $\bar{P}$ is approximately normal with mean p and variance $p(1-p)/n$ if p is not too close to either 0 or 1 and if n is relatively large. Typically, to apply this approximation we

require np that and $n(1-p)$ be greater than or equal to 5. We will make use of the normal approximation in this regard.

Definition:

If n is large, the distribution of

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\bar{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \tag{5.0.7}$$

is approximately standard normal. The $100(1-\alpha)\%$ CI for p then given by;

$$\bar{P} - Z_{a/2}\sqrt{\frac{\bar{P}(1-\bar{P})}{n}} \leq p \leq \bar{P} + Z_{a/2}\sqrt{\frac{\bar{P}(1-\bar{P})}{n}} \tag{5.0.8}$$

This procedure depends on the adequacy of the normal approximation to the binomial. To be reasonably conservative, this requires that np and $n(1-p)$ be greater than or equal to 5. In situations where this approximation is inappropriate, particularly in cases where n is small, other methods must be used.

R A manufacturer of electronic calculators is interested in estimating the fraction of defective units produced. A random sample of 800 calculators contains 10 defectives. Compute a 99% confidence interval on the fraction defective.

Solution: $n = 800$, $x = 10$, $\bar{p} = x/n = 0.0125$, $n\bar{p} = 10$ and $n(1-p) = 790$

$$\begin{aligned} 0.0125 - Z_{0.05}\sqrt{\frac{0.0125(1-0.0125)}{800}} \leq \quad & p \quad \leq 0.0125 + Z_{a/2}\sqrt{\frac{0.0125(1-0.0125)}{800}} \\ 0.0125 \quad & \pm \quad 2.575(0.003928) \\ 0.0125 \quad & \pm \quad 0.0101 = (0.0025, 0.0226) \end{aligned}$$

R Of 1000 randomly selected cases of lung cancer, 823 resulted in death within 10

years. Construct a 95% confidence interval on the death rate from lung cancer.

Confidence Interval for the Difference Between Two Population Means (Variances Known)

There are instances where we are interested in estimating the difference between two population means. Here, from each of the populations a sample is drawn and from the data of each, the sample means x_1 and x_2 respectively, are computed. The estimator $x_1 - x_2$ yields an unbiased estimate of $\mu_1 - \mu_2$, the difference between the population means. The quantity

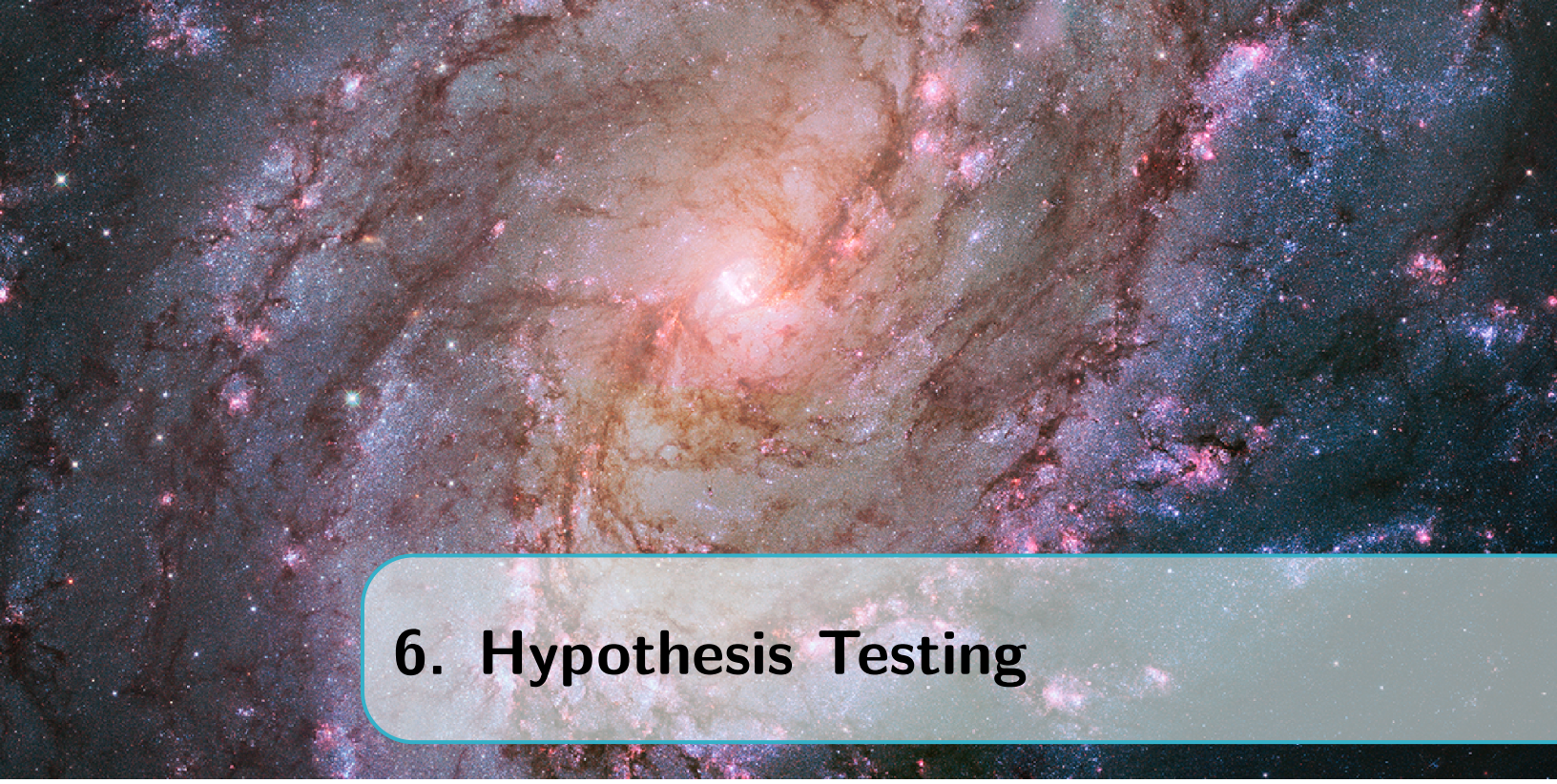
$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{5.0.9}$$

has $N(0,1)$ distribution.

The $100(1-\alpha)$ % confidence interval for $\mu_1 - \mu_2$ is given by,

$$\bar{X}_1 - \bar{X}_2 \pm Z_{a/2}\left(\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{5.0.10}$$

for large sample sizes n_1 and n_2 respectively.



6. Hypothesis Testing

Learning Objectives

Having worked through this chapter the student will be able to:

- Structure engineering decision-making problems as hypothesis tests.
- Test hypotheses on the mean of a normal distribution using either a Z-test or a t-test procedure.
- Test hypotheses on the variance or standard deviation of a normal distribution.
- Test hypotheses on a population proportion.

6.1 Tests of Hypotheses and Significance

6.1.1 Introduction

We now discuss the subject of hypothesis testing, which as earlier noted is one of the two basic classes of statistical inference. Testing of hypotheses involves using statistical inference to test the validity of postulated values for population parameters. If the hypothesis specifies the distribution completely it is called simple, otherwise it is called composite. For example, a demographer interested in the mean age of

residents in a certain local government area might pose a simple hypothesis such as $\mu = 24$ or he might specify a composite hypothesis as $\mu < 24$ or $\mu > 24$.

A statistical test is usually structured in terms of two mutually exclusive hypotheses referred to as the null hypothesis and the alternative hypothesis denoted by H_0 and H_1 respectively.

Two types of error occur in hypothesis testing; these are type I error and type II error. Type I error occurs if H_0 is rejected when it is true. The probability of a type I error is the conditional probability, P(reject $H_0|H_0$ is true) is denoted by α

Hence,

$$\alpha = P(reject \quad H_0|H_0 \quad is \quad true) \quad and$$

$$1-\alpha = P(accept \quad H_0|H_0 \quad is \quad true)$$

Type II error if H_0 is accepted when it is false. Its probability is denoted by the symbol, where β, where

$$\beta = P(accept \quad H_0|H_0 \quad is \quad false) \quad and$$

$$1-\beta = P(reject \quad H_0|H_0 \quad is \quad false) \quad called \quad power \quad of \quad the \quad test$$

Types I and II error can be explained as follows:

	H_0 is true	H_0 is false
Accept H_0	$1-\alpha$ (correct decision)	β (Type II errors)
Reject H_0	α (Type II errors)	$1-\beta$ (correct decision)

Standard format of hypothesis testing: this format involves five steps.

Step 1: State the null and alternative hypotheses.

Step 2: Determine the suitable test statistics.

This involves choosing the appropriate random variable to use in deciding to accept or reject the null hypothesis.

Unknown Parameter	H_0 Appropriate Test Statistic
'μ' σ known, population normal	$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
'μ' σ known, population normal	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ if n is 'large' usually $n \geq 30$
'μ' σ unknown, n small, population normal	$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$, with (n-1) df
'p' population normal, n large	$Z = \frac{\left(x/n\right) - p_0}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}$

Step 3: Determine the critical region using the cumulative distribution table for the test statistic. The set of values that lead to the rejection of the null hypothesis is called the critical region. A statistical test may be a one-tail or two-tail test. Whether one uses a one- or two- tail test of significance depends upon how the alternative hypothesis is formulated.

Types of Hypothesis	H_0	H_1	Decision Rule H_0 Rejected if
Two-tail	$\mu = \mu_0$	$\mu = \mu_0$	$Z < Z_{a/2}$ or $Z < -Z_{a/2}$
Right-tail	$\mu \leq \mu_0$	$\mu > \mu_0$	$Z > Z_{a/2}$
Left-tail	$\mu \geq \mu_0$	$\mu < \mu_0$	$Z < -Z_{a/2}$

Step 4: Compute the values of the test statistic based on the sample information,

e.g. Z_e, t_e, χ^2_e

Step 5: Make a statistical decision and interpretation. H_0 is rejected if the computed value of the test statistic falls in the critical region otherwise it is accepted.

Possible situation in testing a statistical hypothesis

	Hypothesis is correct	H_0 Hypothesis is incorrect
Hypothesis is Accepted	Correct decision	Type II error β
Hypothesis is Rejected	Type I error α	Correct decision

Type I error: We reject a hypothesis when it should be accepted. P(reject Ho|Ho true).

Type II error: We accept a hypothesis when it should be rejected. P(accepting Ho|Ho false).

6.1.2 A Single Population Mean μ

We shall consider testing of hypothesis about a population mean under three different conditions:

1 When sampling is from a normally distributed population with known variance.

2 When sampling is from a normally distributed population with unknown variance.

3 When sampling is from a population that is not normally distributed.

Sampling From Normally Distributed Populations: Population Variance Known

Examples

R A researcher is interested in the mean level of some enzyme in a certain population. The data available to the researcher are the enzyme determinations made on a sample of 10 individuals from the population of interest, and the sample mean is 22. If the sample came from a population that is normally distributed with a known variance, $\sigma^2 = 45$. Can the researcher conclude that the mean

enzyme level in this population is different from 25? Take $\alpha = 0.05$.

Solutions:

Step 1:

$H_0 : \mu_0 = 25$

$H_1 : \mu_0 \neq 25$

Step 2: $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$

since μ_0 and σ are known.

Step 3: $\sigma^2 = 45, n = 10, \bar{X} = 22$

Step 4: $Z_c = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{22-25}{\sqrt{45}/\sqrt{10}} = \frac{-3}{2.1213} = -1.41$

Step: We are unable to reject the null hypothesis, since -1.42 > -1.96.

R Aircrew escape systems are powered by a solid propellant. The burning rate of this propellant is an important product characteristic. Specifications require that the mean burning rate must be 50 centimeters per second. We know that the standard deviation of burning rate is $\sigma = 2$ centimeters per second. The experimenter decides to specify a type I error probability or significance level of $\alpha = 0.05$ and selects a random sample of and obtains a sample average burning rate of $\bar{X} = 51.3$ centimeters per second. What conclusions should be drawn?

Solution:

$H_0 : \mu_0 = 50$

$H_1 : \mu_0 \neq 50$

$Z_c = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} = \frac{51.3-50}{2/\sqrt{25}} = \frac{1.3}{0.4} = 3.25$

Conclusion: Since $z_c = 3.25 > 1.96$, we reject $H_0 : \mu_0 = 50$ at the 0.05 level of significance. We conclude that the mean burning rate differs from 50 centimeters per second, based on a sample of 25 measurements.

6.1.3 Tests on the Mean of a Normal Distribution: Variance Unknown

Test statistic is

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has a t distribution with n - *1 degrees of freedom*

R A study revealed that the upper limit of the Normal Body Temperature of males is 98.6. The body temperatures for 25 male subjects were taken and recorded as follows: 97.8, 97.2, 97.4, 97.6, 97.8, 97.9, 98.0, 98.0, 98.0, 98.1, 98.2, 98.3, 98.3, 98.4, 98.4, 98.4, 98.5, 98.6, 98.6, 98.7, 98.8, 98.8, 98.9, 98.9 and 99.0. Test the hypothesis $H_0 : \mu_0 = 98.6$ versus $H_1 : \mu_0 \neq 98.6$, using $\alpha = 0.05$

R Nine patients suffering from the same physical handicap, but otherwise comparable were asked to perform a certain task as part of an experiment. The average time required to perform the task was seven minutes with a standard deviation of two minutes. Assuming normality, can we conclude that the true mean time required to perform the task by this type of patient is at least ten minutes?

R The increased availability of light materials with high strength has revolutionized the design and manufacture of golf clubs, particularly drivers. Clubs with hollow heads and very thin faces can result in much longer tee shots, especially for players of modest skills. This is due partly to the "spring-like effect" that the thin face imparts to the ball. Firing a golf ball at the head of the club and measuring the ratio of the outgoing velocity of the ball to the incoming velocity can quantify this spring-like effect. The ratio of velocities is called the coefficient of restitution of the club. An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and

their coefficients of restitution measured. In the experiment, the golf balls were fired from an air cannon so that the incoming velocity and spin rate of the ball could be precisely controlled. Determine if there is evidence (with $\alpha = 0.05$) to support a claim that the mean coefficient of restitution exceeds 0.82. The observations are:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

The sample mean and sample standard deviation are $\bar{X} = 0.83725$ and $s = 0.02456$.

6.1.4 Tests on a Population Proportion

It is often necessary to test hypotheses on a population proportion. For example, suppose that a random sample of size n has been taken from a large population and that $X(\leq n)$observations in this sample belong to a class having a particular characteristic of interest. Then $\hat{P} = X/n$ is a point estimator of the proportion of the population p that belongs to this class. Note that n and p are the parameters of a binomial distribution. Recall that the sampling distribution of $\hat{p}$ is approximately normal with mean p and variance $p(1-p)/n$, if p is not too close to either 0 or 1 and if n is relatively large.

In many engineering problems, we are concerned with a random variable that follows the binomial distribution. For example, consider a production process that manufactures items that are classified as either acceptable or defective. It is usually reasonable to model the occurrence of defectives with the binomial distribution, where the binomial parameter p represents the proportion of defective items produced. Consequently, many engineering decision problems include hypothesis testing about p.

Considering testing

$$H_0 = p = p_0$$

$$H_1 = p \neq p_0$$

For large samples, the normal approximation to the binomial with the test statistic

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{X/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\bar{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

may be used.

This presents the test statistic in terms of the sample proportion instead of that number of items X in the sample that belongs to the class interest.

R In a study designed to assess the relationship between a certain drug and a certain anomaly in chick embryos, 50 fertilized eggs were injected with the drug on the fourth day of incubation. On the twentieth day of incubation the embryos were examined and in 12 the presence of the abnormality was observed. Test the null hypothesis that the drug causes abnormalities in not more than 20 percent of eggs into which it is introduced. Let $\alpha = 0.05$.

R A manufacturer of intraocular lenses is qualifying a new grinding machine and will qualify the machine if the percentage of polished lenses that contain surface defects does not exceed 2%. A random sample of 250 lenses contains six defective lenses. Formulate and test an appropriate set of hypotheses to determine if the machine can be qualified. Use $\alpha = 0.05$.

R A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step not exceed 0.05 and that the manufacturer demonstrate process capability at this level of quality using $\alpha = 0.05$. The semiconductor manufacturer takes a random sample of 200 devices and finds that four of them are defective. Test the null hypothesis that the process fallout does not exceed 0.05.

6.1.5 7.5 The Difference Between two Population Means

Hypothesis testing involving the difference between two population means is most frequently employed to determine whether or not it is reasonable to conclude that the two are unequal. In such cases, one or other of the following hypotheses is tested:

$$\begin{array}{ll} H_0 : \mu_1 = \mu_2 = 0 & H_1 : \mu_1 = \mu_2 \neq 0 \\ H_0 : \mu_1 = \mu_2 \geq 0 & H_1 : \mu_1 = \mu_2 < 0 \\ H_0 : \mu_1 = \mu_2 \leq 0 & H_1 : \mu_1 = \mu_2 > 0 \end{array}$$

Hypothesis Tests for a Difference in Means, Variances Known

Test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

R In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with mongolism yielded a mean serum uric acid value of $\bar{X}_1 = 4.4mg/100ml$. In a general hospital, a sample of 15 normal individuals of the same age and sex were found to have a mean value of $\bar{X}_2 = 43.4mg/100ml$. If it is reasonable to assume that the two populations of values are normally distributed with variances equal to 1. Do these data provide sufficient evidence

to indicate a difference in mean serum uric acid levels between normal individuals and individuals with mongolism? Let $\alpha = 0.05$.

Solution

$$n_1 = 12 \quad n_2 = 15$$

$$\sigma_1^2 = 1 \quad \sigma_2^2 = 1$$

$$\bar{x}_1 = 4.5 \quad \bar{x}_2 = 3.4$$

$$H_0 : \mu_1 = \mu_2 = 0$$

$$H_1 : \mu_1 = \mu_2 \neq 0$$

$$Z_c = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z_c = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(4.5 - 3.4) - 0}{\sqrt{\frac{1}{12} + \frac{1}{15}}}$$

$$= \frac{1.1}{\sqrt{0.15}} = \frac{1}{0.3873} = 2.84$$

Reject H_0 since $2.84 > 1.96$ on the basis of these data, there is an indication that the means are not equal.

R A product developer is interested in reducing the drying time of a primer paint. Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample average drying times are $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient, using $\alpha = 0.05$?

Solution:

$$n_1 = 10 \quad n_2 = 10$$

$$\sigma_1^2 = 64 \quad \sigma_2^2 = 64$$

$$\bar{x}_1 = 121 \quad \bar{x}_2 = 112$$

$$H_0 : \mu_1 \leq \mu_2 = 0$$

$$H_1 : \mu_1 > \mu_2$$

$$Z_c = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(121 - 112) - 0}{\sqrt{\frac{64}{10} + \frac{64}{10}}}$$

$$= \frac{9}{\sqrt{12.8}} = \frac{9}{3.5777} = 2.52$$

Conclusion: Reject H_0.

Exercises

i Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The fill volume can be assumed normal, with standard deviation $\alpha_1 = 0.020$ and $\alpha_2 = 0.025$ ounces. A member of the quality engineering staff suspects that both machines fill to the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from the output of each machine.

Machine 1		Machine 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.03
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

Do you think the engineer is correct? Use. $\alpha = 0.05$

ii Two different formulations of an oxygenated motor fuel are being tested to study their road octane numbers. The variance of road octane number for formulation 1 is σ_1^2, and for formulation 2 it is σ_2^2. Two random samples of size $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane numbers observed are $\bar{x}_1 = 89.6$ and $\bar{x}_2 = 92.5$. Assume normality, and if formulation 2 produces a higher road octane number than formulation 1, the manufacturer would like to detect it. Formulate and test an appropriate hypothesis, using $\alpha = 0.05$.

Hypothesis Tests for a Difference in Means, Variances unknown but Assumed Equal.

We now extend the results of the previous lecture to the difference in means of the two distributions when the variances of both distributions σ_1^2 and σ_2^2 are unknown. If

the sample sizes n_1 and n_2 exceed 30, the normal distribution procedures could be used. However, when small samples are taken, we will assume that the populations are normally distributed and base our hypothesis tests on the t distribution. This nicely parallels the case of inference on the mean of a single sample with unknown variance.

The normality assumption is required to develop the test procedure, but moderate departures from normality do not adversely affect the procedure. Two different situations must be treated. In the first case, we assume that the variances of the two normal distributions are unknown but equal; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. In the second, we assume that and are unknown and not necessarily equal. The test statistic is

$$t_c = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad with \quad n_1 + n_2 - 2df$$

The two sample variances are combined to form an estimator of σ^2. The pooled estimator of σ^2 is defined as follows.

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Examples

R The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are $\bar{x}_1 = 8.73$,$s_1^2 = 0.35$,$\bar{x}_2 = 8.68$, and $s_1^2 = 0.40$, respectively. Assume that $\sigma_1^2 = \sigma_2^2$ and that the data are drawn from a normal distribution. Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use $\alpha = 0.05$ in arriving at this conclusion.

R Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield. A test is run in the pilot plant and results in the data shown in the following table. Is there any difference between the mean yields? Use $\alpha = 0.05$, and assume equal variances.

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x} = 92.255$	$\bar{x} = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

R Serum amylase determinations were made on a sample of 15 apparently normal subjects. The sample yielded a mean of 96 units/100ml and a standard deviation 35 units/100ml. Serum amylase determinations were also made on 22 hospitalized subjects. The mean and standard deviation from this second group are 120 and 40 units/100ml., respectively. Would we be justified in calculating that the implied population means are different? Let $\alpha = 0.05$.

Hypothesis Tests for a Difference Between Two Population Proportions

Suppose that two independent random samples of sizes n_1 and n_2 are taken from two populations, and let X_1 and X_2 represent the number of observations that belong to the class of interest in samples 1 and 2, respectively. Furthermore, suppose that the normal :approximation to the binomial is applied to each population, so the estimators of the population proportions $\hat{P}_1 = X_1/n_1$ and $\hat{P}_2 = X_2/n_2$ have approximate normal distributions.

The test statistic is

$$Z = \frac{\bar{P}_1 - \bar{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n2}}}$$

is distributed approximately as standard normal and is the basis of a test for $H_0 : p_1 = p_2$. If the null hypothesis is true, using the fact that $p_1 = p_2 = p$, the random variable

$$Z = \frac{\bar{P}_1 - \bar{P}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is distributed approximately $N(0,1)$. An estimator of the common parameter p is

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

The test statistic for $H_0 : p_1 = p_2$ is then

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Examples

R A random sample of 500 adult residents of Maricopa County found that 385 were in favor of increasing the highway speed limit to 75 mph, while another sample of 400 adult residents of Pima County found that 267 were in favor of the increased speed limit. Do these data indicate that there is a difference in the support for increasing the speed limit between the residents of the two counties? Use $\alpha = 0.05$.

R Out of a sample of 150, selected from patients admitted over a two-year period to a large hospital, 129 had some type of hospitalization insurance. In a sample

of 160 similarly selected patients from a second hospital, 144 had some type of hospitalization insurance. Test the null hypothesis that $p_1 = p_2$. Let $\alpha = 0.05$.

7. Regression

Learning Objectives

Having worked through this chapter the student will be able to:

- Use simple linear regression for building empirical models of engineering and scientific data.
- Understand how the method of least squares is used to estimate the parameters in a linear regression model.
- Test statistical hypotheses and construct confidence intervals on regression model parameters.
- Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval on the future observation.
- Apply the correlation model.

7.1 Regression and Correlation Analysis

7.1.1 Introduction

Many problems in engineering and science involve exploring the relationships between two or more variables. *Regression analysis* is a statistical technique that is very useful for these types of problems. For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes. Other examples are, studying the relationship between blood pressure and age, the concentration of an injected drug and heart rate etc.

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the independent or explanatory variables with a view to estimating and predicting the (population) mean or average of the former (dependent) in terms of the known or fixed (in repeated sampling) values of the latter (independent).

Very often in practice, a relationship is found to exist between two (or more) variables and one wishes to express this relationship in mathematical form by determining an equation connecting the variables. Correlation analysis, on the other hand, is concerned with measuring the strength of the relationship between variables. When we compute measures of correlation from a set of data, we are interested in the degree of the correlation between variables.

7.1.2 The Regression Model

In the typical regression problem, the researcher has available for analysis a sample of observations from some real or hypothetical population. Based on the result of his

analysis of the sample data, he is interested in reaching decisions about the population from which the sample is presumed to have been drawn. It is important that the researcher understand the nature of the population in which he is interested.

In the simple linear regression model two variables X and Y, are of interest. The variable X is usually referred to as the independent variable, while the other variable, Y is called the dependent variable; and we speak of the regression of Y on X. The following are the assumptions underlying the simple linear regression model.

i Values of the independent variable X are fixed.

ii The variable X is measured without error.

iii Values Y are normally distributed.

iv The Y values are statistically independent.

v The variances of the subpopulations of Y are all equal.
$V(Y|x) = V(\alpha+\beta x+e) = V(\alpha+\beta x)+V(e) = 0+\sigma^2 = \sigma^2$

vi The means of the subpopulations of Y all lie on the same straight line, this is the assumption of linearity.

$$E\Big(Y|x\Big) = \mu_{Y|x} = \alpha+\beta x \tag{7.1.1}$$

These assumptions may be summarized by means of the following equation which is called the simple linear regression model because it has only one independent variable or *regressor*:

$$y = \alpha+\beta x+e \tag{7.1.2}$$

where α and β (slope and intercept) are called population regression coefficients, e is called the error term with *mean zero* and variance σ^2. The random errors corresponding to different observations are also assumed to be uncorrelated random variables.

The results of n observations of the set of random variables X and Y can be summarized by drawing a scatter diagram. A straight line passing closely to the points may be drawn. The main problem arises when the points do not all lie exactly on the straight line, but simply form a cloud of points around it. Thus, it may be possible by guess work to draw quite a number of lines each of which will appear to be able to explain the relationship between X and Y. We shall consider finding a best fit line. Such a line will then be used as a model relating the random variable Y with the random variable X.

Suppose that we have n pairs of observations $\big(x_1, y_1\big), \big(x_2, y_2\big), ... \big(x_n, y_n\big)$. The following figure shows a typical scatter plot of observed data and a candidate for the estimated regression line.

The estimates of α and β should result in a line that is (in some sense) a "best fit" to the data. The German scientist Karl Gauss proposed estimating the parametersandin Equation 1.1 to minimize the sum of the squares of the deviations in the diagram. We call this criterion for estimating the regression coefficients the (method of least squares.) Using Equation 7.1.2, we may express the n observations in the sample as

$$y_i = \alpha_i + \beta x_i + e_i \quad i = 1, 2, ..., n$$

and the sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} = \sum_{i=1}^{n} \big(y - \alpha - \beta x_i\big)^2 \tag{7.1.3}$$

The least squares estimators of α and β, say $\hat{\alpha}\$and\$$, must satisfy

$$\frac{\delta L}{\delta \alpha} = -2 \sum_{i=1}^{n} \big(y - \hat{\alpha} - \hat{\beta} x_i\big) = 0 \tag{7.1.4}$$

$$\frac{\delta L}{\delta \beta} = -2\sum_{i=1}^{n}\Big(y - \hat{\alpha} - \hat{\beta}x_i\Big)x_i = 0 \tag{7.1.5}$$

Simplifying these two equations yields

$$n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i y_i \tag{7.1.6}$$

Equations 1.4 are called the least squares normal equations. The solution to the normal equations results in the least squares estimators α and β.

The least squares estimates of the intercept and slope in the simple linear regression model are

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{7.1.7}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Equation 7.1.8 can also be written as

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

The estimated regression line is therefore $\hat{y} = \hat{\alpha} + \hat{\beta}x$ Alternatively

7.2 Method of Least Squares

We shall now find a and b, the estimates of and so that the sum of the squares of the residuals is a minimum. The residual sum of squares is often called the Sum of Squares of Errors (SSE) about the regression line. This minimisation procedure for estimating the parameter is called the "methods of least squares". Hence we shall find a and b so as to minimise

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \Big(y_i - \hat{y}_i\Big)^2 = \sum_{i=1}^{n} \Big(y_i - a - bx_i\Big)^2$$

Differentiating SSE with respect a and b, we have

$$\frac{\delta(SSE)}{\delta a} = -2\sum_{i=1}^{n}(Y_i - a - bx_i) \frac{\delta(SSE)}{\delta b} = -2\sum_{i=1}^{n}(Y_i - a - bx_i)x_i$$

Setting the partial derivative equal to zero and rearranging the terms, we obtain the equation (called the normal equations)

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad .....(1)$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad .....(2)$$

Solving for a and b from (1) and (2)

$$b = \frac{n\sum x_i y_i - \Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big)}{n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

Equations (1) and (2) can also be solved using matrices as:

$$\begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$\begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

Examples

R Assuming we have the following quantities $n = 8$, $\sum x = 140$, $= 382$, $\sum xy =$ 3870, and $\sum x2 =$ 3500.

Solution:

Using the above equation:

$$\begin{bmatrix} 8 & 140 \\ 140 & 3500 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 382 \\ 3870 \end{bmatrix}$$

Solving gives a = 94.67 and b = -2.68

Thus, the estimated regression line is given by:

= 94.67 - 2.68x

It may be noted that the least-squares line passes through the point (x, y) called the 'Centroid' or centre of gravity of the data. The slope b of the regression line is independent of the origin of coordinates. It is therefore said that b is invariable under the translation of axes. Besides assuming that the regression of y and x is a linear function having the form $E(Y|X) = \alpha + \beta x$ we have made three further assumptions which may be summarised as follows:

Patients' Scores on Standardized Test and New Test		
Patient Number	Score on New Test (X)	Score on Standardized Test (Y)
1	50	61
2	55	61
3	60	59
4	65	71
5	70	80
6	75	76
7	80	90
8	85	106
9	90	98
10	95	100
11	100	114

1 **Normality:** We have assumed that each variable y_i has a normal distribution.

2 **Independence:** We have assumed that the variables $y_i....y_n$ are independent.

3 **Homoscedasticity:** We have assumed that the variable $y_i....y_n$ have the same variance σ^2. This assumption is called the assumption of homoscedasticity. In general, it is said that random variables having the same variance are homoscedastic, and random variables having different variance are heteroscedastic.

Examples

R A team of professional mental health workers in a long-stay psychiatric hospital wished to measure the level of response of withdrawn patients to a program of remotivation therapy. A standard test was available for this purpose, but it was expensive and time-consuming to administer. To overcome this obstacle, the team developed a test that was much easier to administer. To test the usefulness of the new instrument for measuring the level of patient response, the team decided to examine the relationship between scores made on the new test and scores made on the standardized test. The objective was to use the new test if it could be shown that it was a good predictor of a patient's score on the standardized test. The results are shown in the table below: Obtain the estimates of the regression coefficients.

R A research on "Near Surface Characteristics of Concrete: Intrinsic Permeability", presented data on compressive strength x and intrinsic permeability y of various concrete mixes and cures. Summary quantities are $n14$, $\sum x_i = 43$, $\sum y_i = 572$, $\sum x_i^2 = 157.42$, $\sum y_i^2 = 23,530$, and $\sum x_i y_i = 1697.80$. Assume that the two variables are related according to the simple linear regression model.

(a) Calculate the least squares estimates of the slope and intercept.

(b) Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is $x = 4.3$.

(c) Give a point estimate of the mean permeability when compressive strength is $x = 3.7$

R The following data were obtained from a study investigating the relationship between noise exposure and hypertension.

Y	1	0	1	2	5	1	4	6	2	3	5	4
	6	8	4	5	7	9	7	6				
X	60	63	65	70	70	70	80	90	80	80	85	89
	90	90	90	90	94	100	100	100				

i Fit the simple linear regression model using least squares.

ii Find the predicted mean rise in blood pressure level associated with a sound pressure level of 85 decibels.

7.3 Correlation Analysis

Closely related but conceptually very much different from regression analysis is correlation analysis, where the primary objective is to measure the strength or degree of linear association between two variables. The *correlation coefficient* measures this strength of (linear) association. For example, we may be interested in finding the

correlation between smoking and lung cancer; between scores on mathematics and fluid mechanics examinations, between high school grades and college grades etc. In regression analysis, as already noted, we are not primarily interested in such a measure. Instead, we try to estimate the average value of one variable on the basis of the fixed values of another variable.

The population correlation coefficient between two random variables, X and Y is defined as

$$\rho = \frac{E[X - E(X)][Y - E(Y)]}{\sqrt{var(X)var(Y)}} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{xy}}{\sigma_X \sigma_Y}$$

where σ_{XY} is the covariance between variables X and Y, σ_X and σ_Y are the standard deviations of X and Y respectively. It is possible to draw inferences about the correlation coefficient ρ using its estimator, the sample correlation coefficient, r. "r" is the correlation coefficient between "n" pairs of observations whose values are (X_i, Y_i) and is given by

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}y}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n}\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Properties of r:

1 It is symmetrical in nature (the two variables are treated symmetrically). That is, there is no distinction between the dependent and independent variables.

2 Both variables are assumed to be random.

3 It can be positive or negative, the sign depending on the sign of the term in the numerator which measures the sample co variation of the two variables.

4 It lies between the limits of -1 and +1; that is, $-1 \leq r \leq +1$.

5 If X and Y are independent, the correlation coefficient between them is zero but

if r=0 it does not mean that the two variables are independent.

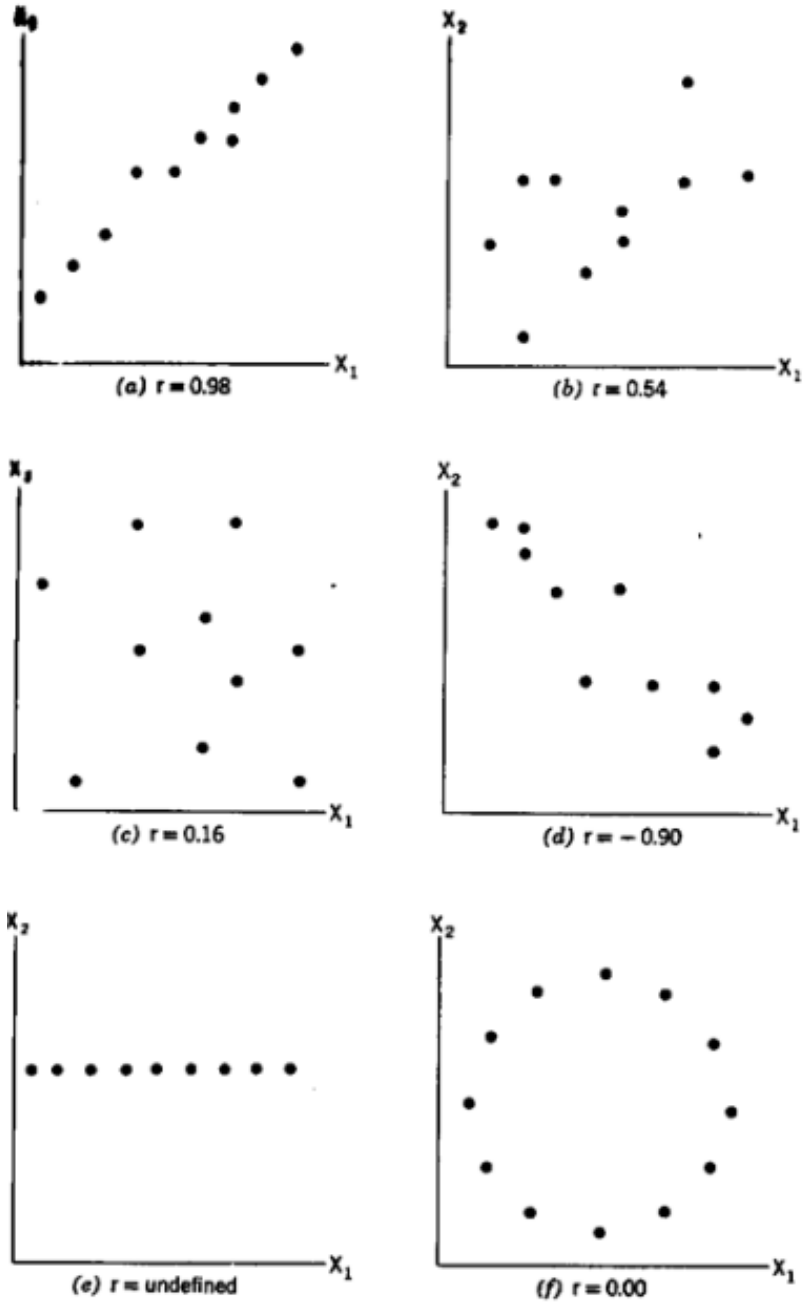


Figure 7.1: Scatter plots with various r values

Testing hypothesis about the correlation coefficient.

A test of the special hypothesis = 0 versus an appropriate alternative is equivalent to

testing $\beta = 0$ for the simple linear regression model. In doing this the t-distribution with n-2 degrees of freedom may be needed. $\frac{b}{\sqrt{SS_{xx}}}$ which can also be written as to test:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Examples

R Using the following data, test the hypothesis that there is no linear correlation among the variables that generated them; at 5% level of significance: $SSxx = 0.11273$ $SSyy = 11,807,324,786$ $SSxy = 34,42275972$

Solution:

$$r = \frac{SSxy}{\sqrt{SSxxSSyy}} = \frac{34422.75972}{\sqrt{(0.11273)(11807324786)}} = 0.9435$$

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

$$\alpha = 0.05 \qquad df = (n-2)$$

$$critical \quad region: \quad t < -2.052 \quad and \quad t > 2.052$$

$$t = \frac{0.9435(\sqrt{29-2})}{\sqrt{1-(0.9435)^2}} = 14.79$$

$$P < 0.0001$$

Decision: Since $t > t_{0.025}(27)$ reject the hypothesis of no linear correlation. More generally, if X and Y follow the bivariate normal distribution, it can be shown that quantity is a random variable that follows approximately the normal distribution with mean and variance equal to $1/(n-3)$. There the

procedure is to compute

$$z = \frac{\sqrt{n-3}}{2}\left[ln(\frac{1+r}{1-r}) - ln(\frac{1+\rho_0}{1-\rho_0})\right] = \frac{\sqrt{n-3}}{n} In\left[\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right]$$

Examples

R Consider the immediate preceding example data, test the null hypothesis that = 0.9 against the alternative that > 0.9 at 5% level of significance.

Solution:

$H_0 : \rho = 0.9$

$H_1 : \rho > 0.9$

Critical region : Z > 1.645

Decision: Since $Z < Z_{0.05}$ there is no evidence that the correlation coefficient is not equal to 0.9

In ordinary usage of this method, it is not necessary to use the formula for Z that corresponds to r values between 0.0 and 0.99. Tables contain fisher - Z values Z_f are available. In this case to test $H_0 : \rho = \rho_0$ vrs $H_1 : \rho \neq \rho_0$

Z = we have Z=

Critical region is $Z \leq -Z_2 and Z \geq Z_{\alpha/2}$ where Z_f and f are the fisher - Z values for r and ρ_0 respectively.

Examples

R The following data gave X = the water content of snow on April 1 and Y= the yield from April to July (in inches) on the Snake River watershed in Wyoming for 17 years.

x	y	X	y
23.1	10.5	37.9	22.8
32.8	16.7	30.5	14.1
31.8	18.2	25.1	12.9
32.0	17.0	12.4	8.8
30.4	16.3	35.1	17.4
24.0	10.5	31.5	14.9
39.5	23.1	21.1	10.5
24.2	12.4	27.6	16.1
52.5	24.9		

Estimate the correlation between X and Y

R Two methods of measuring cardiac output were compared in 10 experimental animals with the following results

Cardiac Output (1./min)	
Method I x	Method I Y
0.8	0.5
1.0	1.2
1.3	1.1
1.4	1.3
1.5	1.1
1.4	1.8
2.0	1.6
2.4	2.0
2.7	2.4
3.0	2.8

Compute the sample correlation coefficient.

R A group of eight athletes ran a 400 metres race twice. The times in seconds were recorded as follows for each athlete.

Runner	
1st Trial x	2nd Trial Y
48.4	48.0
51.2	54.3
48.6	49.4
49.5	48.4
51.6	54.0
49.3	47.2
50.8	51.8
49.7	50.3

Calculate the correlation coefficient between these two trials.

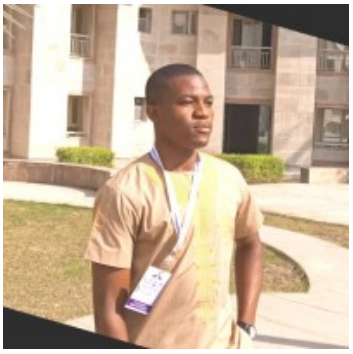
Author Profiles

Dr. Benjamin Odoi

Lecturer at the Department of Mathematical Sciences at the University of Mines and Technology, UMaT, Tarkwa. He is a contemporary statistician and big data analytics with diverse applications focusing on health, applied statistics, and climate change. As a statistician on many research projects, he has provided technical advice in areas ranging from sample size calculation, data collection design, data management procedures, data analysis, statistical and mathematical modeling, and many more. Also, as an inspiring teacher, he has taught many statistical and related courses at both the undergraduate and postgraduate levels. He has also guided the research work of many students in applied statistics, mathematical modeling, public health, finance, logistics, and engineering at the Ph.D., MPhil, MSc, and BSc levels. His computer programming skills include using SAS, R, Minitab, SPSS, Eviews, XL Miner, and Pasalade.

Abdulzeid Yen Anafo

Experienced Statistician and Machine Learning Engineer with a demonstrated history of working in the research industry. Skilled in Python and R programming, SPSS and Teamwork. Strong research professional with a Doctor of Philosophy - Ph.D. focused in Mathematical statistics from the University of Mines and Technology, Tarkwa.



Seth Antanah

Mathematician with a passion for learning. I have experience in programming languages like JavaScript, R, and Python as well as Relational Databases such as SQL and BigQuery. My personal goal is to become a highly productive Data Analyst.