CAMPION COLLEGE
AT THE UNIVERSITY OF REGINA

STAT 100

Elementary Statistics for Applications

by Robert G. Petry and Blake Friesen

published by Campion College

2^{nd} Edition Copyright © 2012 Robert G. Petry, Blake Friesen, Mike DeCorby.

1^{st} Edition Copyright © 2011 Robert G. Petry, Blake Friesen, Mike DeCorby.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

Permission is granted to retain (if desired) the original title of this document on modified copies.

All numerical data in this document should be considered fabricated unless a source is cited directly. Comments in the transparent copy of the document may contain references indicating inspiration behind some of the data.

History

- 1^{st} Edition produced in 2011 entitled "Elementary Statistics for Applications" written by principal authors Robert G. Petry and Blake Friesen with contributions from Mike Decorby. Published by Campion College at the University of Regina.
- 2^{nd} Edition produced in 2012 entitled "Elementary Statistics for Applications" written by principal authors Robert G. Petry and Blake Friesen with contributions from Mike Decorby. Published by Campion College at the University of Regina.

A transparent copy of this document is available via
http://www.campioncollege.ca/about-us/faculty-listing/dr-robert-petry

About the cover: The cover image is the 3D probability density of an electron in a hydrogen atom in an excited state with principal quantum numbers n=5, l=3, m=1. The slices give the probability density found on each of the respective planes. Dark regions indicate areas of greater probability. The image was generated in part using the open source Visualization Toolkit software (VTK).

Contents

*Sections denoted with an asterisk are optional and may be omitted in the course.

Unit 1: Descriptive Statistics

1.1 Introduction to Statistics

What is statistics?

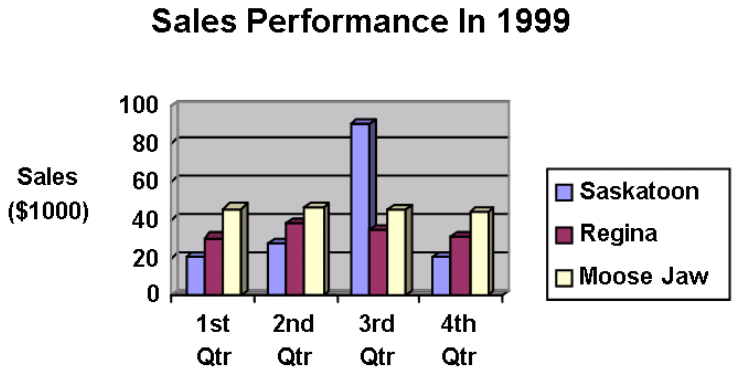
Depending upon one's background, the word statistics has many usages within our language.

1. Some people use the word statistics to mean *numerical facts* presented in forms such as in tables or graphs. Here are some examples of some economic statistics which fit this usage:

Table 1.1: Mineral Production in Saskatchewan

	Production ('000 Tonnes)	
Mineral	1996	1997
Coal	10854	11648
Salt	813	875
Sodium Sulphate	326	292
Potash	6969	8346
Uranium[1]	13.35	14.173
Gold[1]	3.432	4.366

[1] Based on reported sales.
Source: "Mineral Production", Monthly Statistical Review, Volume 24, no. 3, Government of Saskatchewan, Bureau of Statistics, p 12.



2. To some people the word statistics means a *calculation on a collection of numerical values.* General crime rates per 100,000 people for each province and territory for 2006 are given in the following table. (Source: Statistics Canada)

Locale	Rate	Locale	Rate
Canada	8,269	Man.	12,325
N.L.	6,571	Sask.	15,276
P.E.I.	7,486	Alta.	10,336
N.S.	8,698	B.C.	12,564
N.B.	6,781	Y.T.	22,197
Que.	6,626	N.W.T.	44,074
Ont.	6,251	Nvt.	32,831

   Why is it appropriate to report the rate per 100,000 people?

3. To some people statistics means a *methodology for arranging data in a format useful for decision making.* The following statistics on the growth of income disparity in Canada might be useful for informing government social or taxation policy. (Source: Statistics Canada)

	Median Earnings (2005 $; full-time, full-year earners)				Percentage change	
Population Segment	1980	1990	2000	2005	1980 to 2005	2000 to 2005
Bottom 20 percent	19,367	16,345	15,861	15,375	-20.6	-3.1
Middle 20 percent	41,348	40,778	40,433	41,101	0.1	2.4
Top 20 percent	74,084	76,616	81,224	86,253	16.4	6.2

4. Statistical procedures can be very analytical and use *theoretical information from probability functions to make decisions where randomness plays a part in observed outcomes.* The following description of a statistical study illustrates this.

   A large hotel derives income from its room rental service and its dining lounge. Management is interested in determining the contribution to sales at the dining lounge that are associated with room rentals. Data are selected from a random sample of 100 days of business at the hotel. A statistical plot of sales at the dining lounge against rooms occupied is made. It is determined that the plot closely follows a linear trend given by the equation $y = 38x + 650$. This is loosely interpreted as follows:
   Based on these observations, on average for each room rented, \$38 of new revenue per day are produced in the dining lounge and \$650 in revenue per day come from clients who do not rent rooms. The \$38 marginal contribution to daily revenue from room rentals is precise to $\pm$ \$3 in 95 cases out of 100.

   Similar examples of statistics can be found in trade journals, professional magazines, daily newspapers, periodicals, business reports, statistical journals, etc. Look for statistical examples in your daily reading material. It will make the examples and calculations done in the lectures more relevant to you.

   No matter what the usage, the sole need for statistics in data presentation is to handle fluctuations or variation in some quantity which is being analyzed for the purpose of decision making. Statistical analysis has no application in situations where quantities under consideration never vary.

Where do we use statistics?

Situations which give rise to analysis of quantities which vary, and as a result involve statistical methods, are activities such as:

- experimental analysis in the traditional sciences and social sciences
- quality control
- forecasting for the purpose of planning (business, government, etc.)
- statistical reports of business activities
- estimating
- testing
- any procedure which relies on sampling
- simulation and experimentation

In particular some sectors which rely on statistical procedures are:

1. Risk taking organizations like the insurance business analyze risks and found their business on the basis of these risks. People called actuaries analyze and provide strategies to manage the risks involved in the insurance business. Some other businesses which are founded on risk taking are gambling, investment management, and health plan carriers.
2. Retail businesses rely on statistics to conduct marketing analysis. Market areas can be analyzed to target potential sales, assess marketing strategies, evaluate customer preferences, etc.
3. Accountants use statistical methods to evaluate the accuracy of journal entries using sampling techniques.
4. Human resource workers use statistics in various ways. Personnel administrators use statistical procedures to interpret achievement and aptitude tests. Labour negotiators use published statistical information to discuss labour variables such as hours of work, pay rates, salary benefits.
5. Large organizations use statistics to determine efficient means of traffic control of inventory between branch locations.
6. Statistical methods are employed by companies to display the activities of the firm to the shareholders through annual reports.
7. In the manufacturing industry where one requires quality control of the production, experimentation and testing of their products.
8. Scientists and engineers apply statistical analysis to study vast amounts of data from their numerous experiments and tests they must conduct.
9. Activists and politicians use statistics when seeking to advance complex arguments based on reason and facts.

As a student, how will statistics be useful?

1. It will be useful for you to communicate your information through statistical data presentation techniques are useful in reports and articles.
2. It will be useful in order to understand the techniques based on sampling which are used by decision-makers in your field of study, your workplace, and the world around you and to apply them yourselves.
3. To be technically literate in a complex technical world, a person should understand the meaning of the statistical measures on which decisions are based.

1.2 Some Introductory Statistical Definitions

Most professionals develop a collection of specialized terms and definitions which are necessary in order to discuss the concepts that are peculiar to their field of study. This is also the case in statistics. Before statistical concepts can be applied to problem solving it is necessary to have an operational knowledge about their meaning. Here are some terms and definitions with which a person using statistical analysis must be able to apply correctly. The emphasis in this course will be on the application rather than the memorization of these terms.

Statistics (Recall from Section 1.1 the term has many uses.)

1. A collection of numbers or a number. (e.g. an accident statistic)
2. A calculated value from a collection. (e.g. an average grade)
3. The science of data analysis for the purpose of decision making. (e.g. statistical analysis of an opinion survey.)

Statistics, as an area of study, is a body of knowledge concerned with collection, summarizing, analyzing and interpreting of data for the purpose of making accurate decisions in the face of uncertainty.

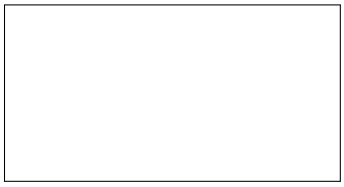
Population (It doesn't necessarily mean people.)

A collection of all possible individuals, objects, or measurements of interest.

Examples:

- The population of claims receipts at an insurance claims centre.
- The population of items produced on an assembly line process.
- The population of drivers over 65 in Saskatchewan.

We will use many diagrams for illustration purposes in this course for the purpose of helping us to visualize abstract concepts. The diagram used to illustrate a population in statistics is a rectangle.



Note that statisticians do not use the word population to mean the size of a population, as in "What is the population of Regina?"

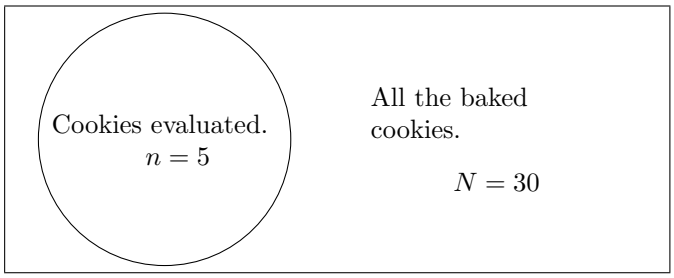
Sample (The everyday definition meaning *to take a portion of* is similar to the statistical one.)

A selection of some of the objects from the population. There are many different ways of making the selection of which one is random selection.

Example:

- To ensure quality, a child takes a sample of five cookies from a recently baked batch of thirty chocolate chip cookies.

For illustrative purposes, a sample is designated by a circle drawn within the population rectangle.



We will use n to refer to the size of a sample and N to refer to the size of a population.

Statistical Variable (The quantity designated by the symbol x)

The quantity under study whose value fluctuates with chance. Because it fluctuates with chance it is often called a **random variable**. It is because this fluctuation exists that statistical analysis is useful.

Examples:

- In the above example the child could test quality by observing the number of chocolate chips found in each cookie (x=number of chocolate chips).
- The value of a used car on a sales lot in Moose Jaw is a statistical variable because upon inspection it would be observed that the price varies from car to car on a sales lot.

x= price of a car

N = # of used cars on lots in Moose Jaw

Data (It most often means a collection of numbers.)

Whereas numbers are often referred to as statistics, a statistician would call these numbers data. To a statistician, a data set is the collection of observed values of the statistical variable.

Example:

- An inspection (sample) of 8 cars picked randomly revealed the following observations: (\$)

  7800 4500 5600 560 780 2300 999 1200

  Note: Usually the data are given as a listing as above, called an **array**, or it is rank ordered into a **ranked array**: (\$)

  560 780 999 1200 2300 4500 5600 7800

  The variable symbol x designates the price of a car. Individual observations within the array are identified by means of a subscripted variable. x_5 means the fifth observation in the array, \$2300. The individual observations in the above array can be referred to by symbol as x_1, x_2, x_3, ..., x_8. This is helpful for calculations with statistical formulae.

Returning to our definition of population given above one could consider the population in this example to be all the objects of interest, in this case the set of all used cars on sales lots in Moose Jaw. Alternatively if one considers the population to be all the measurements, then this would refer to the set of all of their actual prices, that is the data set. A similar ambiguity exists in the use of the word sample, either as the eight used cars or their eight prices. To clearly specify the objects upon which the measurements are taken (the former definitions) they may be called the **experimental units** by an experimenter while a statistician doing a survey would call them the **elements of the sample**. To refer clearly to the measurements themselves one may say **population data** or **sample data**. When the words population and sample are used on their own the statistical meaning intended is often apparent from the context.

Qualitative (or Categorical) Data (Usually non-numerical labels or categories called attributes)

Data[1] are referred to as being qualitative when the observations made are arrived at by classifying according to category or description.

Example:

- The religious denomination of community members is to be recorded and analyzed. (Example Data: "Lutheran", "Sikh", "Muslim", ...)
- The ranks of military personnel at an armed forces base are recorded. (Example Data: "Corporal", "Sergeant", "Corporal", ...)

Quantitative (or Numerical) Data (Numerical observations)

Data are called quantitative when the observations are arrived at by either measuring or counting.

Example:

- The volume of fuel purchased by customers at a self serve gas bar was to be analyzed. The volume was measured and displayed by a device on the the gas pump. (Example Data: 53.5 litres, 75.2 litres, 33.9 litres, ...)
- The number of phone calls received per month per household for charity solicitation in a certain neighbourhood was analyzed. A random sample of households were asked to record a log of their charity phone calls. (Example Data: 8, 5, 0, 5, ...)

Discrete Data (Counts)

Data are called discrete when the possible values of the variable are countable. It often results from a counting process. In this case the variable's value is known exactly. Assuming the counting is done accurately, a count should have no error attached to it.

Example:

- The number of children per household.
- The number of magazines subscribed to by a household.
- The number of birds visiting a bird feeder in an hour.

Continuous Data (Measurements)

Data are called continuous when the variable can assume any real number. Such data often result from a measuring process. Because of the limitations of the measuring process, the value of an observation can only be determined to the precision of the device used to record the measurement. All measurements contain some error.

[1] We will also refer to the variable itself which produces such data as being qualitative. Similarly variables are referred to as quantitative, continuous, discrete, etc. based on the type of data they produce.

Example:

- A store keeps a shelf of a freezer stocked with bricks of butter. The bricks have a labeled weight of 454 g. The weights of the bricks are to be analyzed. Government regulatory agencies allow for some deviation from the stated weight as long as it is not significantly different.
- A vehicle used for deliveries is refueled at the end of each day. The volume of gas filled is entered into a log book.

The terms associated with statistical data can be remembered easily with the following diagram:

$$\text{Statistical Data} \begin{cases} \text{Qualitative Data (categories)} \\ \text{Quantitative Data (numbers)} \begin{cases} \text{Discrete (counts)} \\ \text{Continuous (measurements)} \end{cases} \end{cases}$$

Levels of Measurement (The precision of the data)

Once observations have been taken for a particular statistical variable what can be done with the data depends upon its level of measurement (or precision). A lower level of measurement restricts the type of calculations that can be done with the observations. Four levels of measurement (starting with the simplest) are identified:

Nominal level is the most crude form of data for analysis. In this form, data can only be categorized such as by religious denomination, political affiliation, etc. Outside of counting the number of observations in each category, there are very few arithmetic calculations that can be done on this data.

Ordinal level is a level above nominal level with the added feature that the data can be rank ordered as well as counted. An example is rating form information with the options of good, average, poor. Note: The observations cannot be placed on a number line or added.

Interval level is a level above ordinal level in that the data can be quantified to the extent that it can be placed on a number scale. The number scale has the limitation of not having a meaningful zero point for comparison purposes. A meaningful zero indicates the absence of a quantity. The Celsius or Fahrenheit temperature scale is a an example of this. Zero degrees on these temperature scales does not represent an absence of temperature. A rating scale for things like consumer preference is another example. Note: Data values cannot be compared as ratios when there is no meaningful zero.

Ratio level is the most precise level of measurement. Data in this form can be placed on a number line with a meaningful zero point. The weight of the net contents of a packaged consumer product is an example. There is no limitation to the arithmetic that can be done with this data. All the legitimate operations of arithmetic can be done on data measured at this level.

Example:
An eccentric census[2] form requires you to provide the following information. Classify the data produced by the following variables as qualitative/quantitative and for a quantitative variable decide on whether it is discrete or continuous. Identify the level of measurement (nominal/ordinal/interval/ratio) of each variable. Hint: It may help to think about how you would answer the question to imagine the type of data produced.

1. Your nationality.
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .
2. Your social insurance number.
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .
3. The longitude of your community.
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .
4. The hottest temperature recorded in your community last summer in degrees Kelvin.
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .
5. Your yearly household income.
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .
6. The softness of your pillow (firm, medium, oh so soft).
   Circle: qualitative/quantitative(discrete/continuous), nominal/ordinal/interval/ratio .

Descriptive Statistics (Describing a data collection)

A branch of statistics concerned with describing collections of data for the purpose of placing some order among a collection of unorganized observations. The data collections that are being described can be either samples or populations.

Example:

- Categorizing the collection in a table or constructing a pictorial chart.
- Calculating the average or the amount of spread in the data values.

Inferential Statistics (Drawing conclusions from a sample.)

A branch of statistics concerned with making general conclusions about a population on the basis of a data sample. Probability concepts are involved.

Example:

- A random sample of 1200 Canadian adults was polled regarding their opinion about the quality of health care received by hospital patients in Canada. On the basis of the sample results, it is projected that 75% of all Canadian adults believe that patients receive good care while in hospital in Canada. The poll has an error of $\pm$ 3% in 95 samples out of 100.

The breakdown of statistics into its types can be pictured more easily by the following diagram.

$$\text{Types of Statistical Analysis} \begin{cases} \text{Descriptive} \begin{cases} \text{Describes what was actually} \\ \text{observed.} \end{cases} \\ \text{Inferential} \begin{cases} \text{Infers to a population based on} \\ \text{sample values. (i.e. Statements} \\ \text{are made based on incomplete} \\ \text{information.)} \end{cases} \end{cases}$$

[2] A **census** is a data set involving every individual in the population. They usually don't come cheap!

Statistical Categorization Of Observations

One of the methods of bringing order to a set of observations is by constructing a table of categories into which the data fit. These categories should be **mutually exclusive** and **exhaustive**.

Categories are mutually exclusive when each observation can go into one and only one category. This means the categories should not overlap. A table of categories is exhaustive if all of the raw data observations fit into the categories chosen.

Example:
A table giving political preference that only listed the three political parties Conservative, Liberal, and N.D.P. and the Undecided category would be mutually exclusive since someone who preferred one of the parties the most could also not simultaneously prefer a different party as much nor be undecided. However the table would not be exhaustive because someone could be decided but not prefer one of the three parties listed. For instance a respondent who supported the Green Party would not fit into any category on the table.

Example:
In table 1.1 on page 2 are the mineral categories mutually exclusive? (Yes/No)
Is the list exhaustive for mineral production for Saskatchewan? (Yes/No)

Assignment: For this unit the exercises will focus upon several case studies to be found at the end of the unit starting on page 67 which will be analyzed with increasing sophistication. To begin with, read the case studies. For each case study answer the following questions (which are repeated at the beginning of each study):

1. Identify the population (i.e. experimental unit) under consideration.
2. Identify whether the data encompasses the whole population (population data) or is for a sample of it (sample data).
3. Identify the statistical variable and label the column with the appropriate symbol.
4. Determine whether the variable is qualitative or quantitative.
5. For a quantitative variable identify if it is continuous or discrete.
6. Identify the level of measurement (precision) of each variable.

1.3 Organizing Numerical Data for Analysis

The objective of statistical data presentation is to bring order to the array of observations of the statistical variable. The most common method of doing this is by table construction.

Here are some examples of the methods by which quantitative data are presented. Data presented in these three ways will be the starting point for our calculations of descriptive statistical measures.

1.3.1 Raw Data

This is data listed in an array often in the order it was observed. This format is used when there are a very few observations to consider in the presentation.

Example:
A courier takes 10 trips from office A to office B. The time taken was observed to be: (min.)

15, 23, 30, 18, 26, 27, 19, 25, 31, 23

Note: Before this data is of much use to anyone, it must be rank ordered or perhaps treated mathematically by computing an average. The listing of the data values only show that there is variation in the times taken.

If x equals the time taken for a trip, here is a ranked array of the $(n = 10)$ observations of variable x.

Trip Time (min)
15
18
19
23
23
25
26
27
30
31

1.3.2 Ungrouped Frequency Distribution

This method is used to present data when the raw data array contains many observations but only a few distinct data values are observed.

Example:
The manager of a store examines the number of items purchased per customer passing through the express checkout lane. Here is what was observed:

1	5	7	4	3
2	2	3	2	4
3	2	3	1	1
4	3	2	3	1
1	2	3	7	3

This quantitative data may be presented by means of a table called a **frequency distribution**. The word **frequency**, in statistics, means the number of times a data value is observed to occur. The idea is to replace all of the same values in a raw data array with a single identification of the value and the number of times it occurs in the raw data, f, its frequency.

Example:
The previous information could be summarized by keeping track of the number of times each value occurred in the raw data first through a tally (left) and then converting the latter to the corresponding ungrouped frequency table (right):

# of items	Tally
1	𝍸
2	𝍸 |
3	𝍸 |||
4	|||
5	|
7	||

$\Rightarrow$

# of items	f
1	5
2	6
3	8
4	3
5	1
7	2
	$\sum f = 25$

Here the value 1 occurred 5 times, in the original data, etc.

Note:

1. The statistical variable x= *# of items* appears in the column heading.
2. There are 6 distinct data values but there are 25 observations in total.
3. The sum of the frequency column always equals the number of observations made.
4. The frequency column is always totaled to show that the table is exhaustive.
5. The mathematical symbol for the sum of the frequency column is $\sum f$ where

   $$\sum f = f_1 + f_2 + f_3 + \ldots + f_m \,.$$

   Here m is the number of distinct values taken by the variable x and f_i is the frequency of the i^{th} distinct value.
6. Only the distinct values observed are listed. For example, there is no data value 6 observed so none is listed. If one were listed, it would have a frequency of 0.
7. The table puts the observations in rank order.
8. The x column is never totaled because that is meaningless information for a frequency distribution.
9. In this example x is a discrete variable. Ungrouped frequency distributions are often used to organize discrete data.

1.3.3 Grouped Frequency Distribution

This method is used to summarize data when the raw data array of observations is large and many of the data values listed are distinct.

Example:
Suppose an insurance company examines its claims record over the past 5 years. It discovers upon observation of the records that it has had many claims of various values. The claims' values were summarized as follows:

Value ($)	f
0.00 - 1000.00	900
1000.00 - 2000.00	700
2000.00 - 3000.00	313
3000.00 - 4000.00	240
4000.00 - 5000.00	127
5000.00 - 6000.00	20
	$\sum f = 2300$

Note: There are 2300 claim values summarized in this table. There are so many distinct data values observed that a listing of the raw data array of observations would be difficult to interpret for analysis purposes.

Unlike an ungrouped frequency distribution, the individual observations are lost when the data is grouped. The specific values observed cannot be determined from the table. This type of summary is useful to establish trends among data values such as the range, areas of concentration, averages, dispersion, and to draw graphs.

Each grouping is called a **class**. In the first class there are 900 claims. The values \$0.00 and \$1000.00 are the **lower class limit** and **upper class limit** respectively of the first class. When class limits overlap only the lower class limit belongs to the class. A value of \$1000.00 would belong to the second class.

The **class width** Δx is the difference between the upper and lower class limits. Here

$$\Delta x = \$1000.00 - \$0.00 = \$1000.00$$

for the first class. All other classes in this example may be verified to have the same class width which is typical and preferable.

As in the previous example, x here is also technically a discrete variable since a claim cannot be in fractions of a penny. Still the diversity in the claims requires a grouped frequency distribution for its organization. Continuous random variables also typically appear in grouped frequency distributions because most of their data values will be unique. (e.g. 1.31 cm, 1.35 cm, 2.29 cm, etc. for a measurement.)

Assignment: For each case study:

1. Identify the type of data presentation (raw data, ungrouped frequency distribution or grouped frequency distribution).
2. For frequency distributions, identify the frequency column with the appropriate symbol.
3. Find the number of data elements in each case by direct counting (raw data) or by summing the frequency column (frequency distributions). Use appropriate symbols.

1.4 Summing Data Values

When calculating statistical measures (such as an average value), one of the quantities which will have to be computed is the **sum of all observations**. The method used to find this sum depends on which of the three methods is used to present the data.

1.4.1 Raw Data

The sum of all data values is given by:[3]

$$\sum x = x_1 + x_2 + x_3 + ... + x_n \, .$$

When this is done by paper and pencil, the calculation should be done tabularly (i.e. in a table). List the data values in a column, preferably rank ordered, and place the symbol x with appropriate units at the top and the symbol $\sum x$ at the bottom.

Example:
Three lengths are measured (in mm):

x(mm)
52
65
73
$\sum x = 190$

1.4.2 Ungrouped Frequency Distribution

In this case the sum of all observations is:

$$\sum xf = x_1 f_1 + x_2 f_2 + x_3 f_3 + ... + x_m f_m \, .$$

Here we must multiply each of the m distinct data values x_i by its frequency f_i because the equivalent raw data would have had the value x_i occurring f_i times.

Example:
Tickets are sold in three denominations \$5, \$10, and \$15:

x(\$)	f	xf(\$)
5	10	50
10	20	200
15	20	300
	$\sum f = 50$	$\sum xf = 550$

There are 50 observations (tickets sold) and the sum of these 50 observations is \$550 (the net sales).

Note the xf column will have the same units as the x column since the variable is multiplied by the (dimensionless) frequency.

[3] A more precise way of writing the sum is $\sum_{i=1}^{n} x_i$. Such indexed notation will largely be avoided in this course.

1.4.3 Grouped Frequency Distribution

It is impossible to find the exact total of all observations when the data is presented like this. An approximation can be found if it is assumed that the observations in a class are concentrated at the midpoint of the class. The sum of the observations is approximately $\sum xf$ where x is the midpoint of the class.

Example:
An exam has the following grade results:

Score (%)	f	x(%)	xf(%)
35 - 45	5	40	200
45 - 55	10	50	500
55 - 65	20	60	1200
65 - 75	10	70	700
75 - 85	5	80	400
	$\sum f = 50$		$\sum xf = 3000$

Note: There are 50 scores summarized in the table and the approximate sum of all scores is 3000%. We added the midpoint column x to facilitate the calculation.

Using the Calculator

On modern calculators with statistical functions it is possible to enter data and have the calculator derive values of interest such as the sum of observations. Entering data is typically done by placing the calculator in a statistical mode.[4] One then keys in each raw data value followed by a data entry key. When all data is entered one may select to calculate the sum of observations ($\sum x$), or any other value of interest, and the result will be presented for the stored data.

The calculator can also calculate the result for frequency distributions. When entering data values, one enters each distinct value. Following each value one usually presses a comma or semi-colon key followed by the frequency (number of times) the value occurred in the data. Then the data entry key is pressed and further data may be entered. To retrieve the sum of observations in this case one still would select the ($\sum x$) key. For a grouped frequency distribution, enter the midpoint of each class for x.[5] The student should verify the above sums on her calculator.

Assignment: For each case study find the sum of observations using a tabular method. In which cases are the sums only approximate? Verify the sums using the statistical data keys of your calculator.

[4] If you have lost the manual for your calculator try Googling the model number to find a pdf version of it for the necessary statistical instructions.

[5] On a calculator without statistical functions one may calculate such sums by entering the first observation into the memory M with the X→M key, and the rest of the list with the M+ key. The sum is retrieved by the RM key. For frequency distributions the sum can be obtained on the calculator using the same keys as for raw data. Before summing with the M+ key, however, multiply the data value by its frequency of occurrence.

1.5 Relative Frequency

Frequency distribution tables can be extended with additional columns to provide further information about the data.

1.5.1 Relative Frequency (P)

The **relative frequency** column lists the proportion, as a decimal fraction, of observations found in a given class or at a specific value. We will use the symbol P for the column caption. In symbols

$$\boxed{P = \frac{f}{\sum f}}$$

Example:
The number of occupants in apartments was surveyed with the following results:

# of Occupants	f	P
1	23	0.383
2	25	0.417
3	5	0.083
4	5	0.083
5	1	0.017
6	1	0.017
	$\sum f = 60$	$\sum P = 1.000$

In theory the sum of the relative frequency column will always be 1 because the sum of all proportional parts of the total equals 1. Round these values to two or more places after the decimal for this column. The total may not sum precisely to 1 in practice because of round off.

Relative frequency distributions are useful for comparing distributions where the number of observations in each array is different. In those cases, frequencies cannot be compared directly but proportions can be.

Example:
The following age distributions were found for two towns:

Age Distribution in Smallville		
Age (yr)	f	P
0 - 20	200	0.125
20 - 40	600	0.375
40 - 60	400	0.250
60 - 80	300	0.188
80 - 100	100	0.063
	$\sum f = 1600$	$\sum P = 1.001$

Age Distribution in Bigtown		
Age (yr)	f	P
0 - 20	1500	0.025
20 - 40	12000	0.200
40 - 60	22000	0.367
60 - 80	16000	0.267
80 - 100	8500	0.142
	$\sum f = 60000$	$\sum P = 1.001$

While the frequency f indicates that Bigtown has more young people, this is only due to its overall larger population. Smallville has a proportionately younger population as indicated by the relative frequency P.

Relative frequency columns are also useful in the presentation of data because they are largely insensitive to the number of observations that happen to have been taken.

Example:
Rather than taking a census of Smallville and collecting the ages of all N=1600 inhabitants, a statistician randomly sampled half the people ($n = 800$) and compiled the following frequency distribution:

Age Distribution in Smallville		
Age (yr)	f	P
0 - 20	103	0.129
20 - 40	296	0.370
40 - 60	200	0.250
60 - 80	152	0.190
80 - 100	49	0.061
	$\sum f = 800$	$\sum P = 1.000$

Compared to the results for the complete population, the frequency f of each class drops roughly in half. The relative frequency P, however, is approximately the same as the population results.

Many people are more familiar with interpreting percentages than decimal fractions. In this course we will use decimal and percent notation interchangeably and one must remember to shift the decimal over two places as required (e.g. $35\% = 0.35$). When used in calculations the decimal form is usually necessary. Finally fractions themselves are also convenient and may be converted to decimal through division on the calculator as usual (e.g. $\frac{7}{20} = 0.35$).

1.5.2 Relative Frequency Density (p)

While the relative frequency P is largely insensitive to the number of observations of a statistical variable it has a shortcoming for *grouped* frequency distributions because it still depends on the class size, Δx.

Example:
Two statisticians are given the **same** raw data for the heights of a sample of male high school basketball players in western Canada. They produce the following grouped frequency distributions, the first with class size $\Delta x = 2.0$ cm, and the second with $\Delta x = 4.0$ cm:

$\Delta x = 2.0$ cm

Height(cm)	x(cm)	f	P
169.0-171.0	170.0	4	0.020
171.0-173.0	172.0	10	0.050
173.0-175.0	174.0	14	0.070
175.0-177.0	176.0	73	0.365
177.0-179.0	178.0	70	0.350
179.0-181.0	180.0	17	0.085
181.0-183.0	182.0	10	0.050
183.0-185.0	184.0	2	0.010
		$\sum f = 200$	$\sum P = 1.000$

$\Delta x = 4.0$ cm

Height(cm)	x(cm)	f	P
170.0-174.0	172.0	20	0.100
174.0-178.0	176.0	128	0.640
178.0-182.0	180.0	46	0.230
182.0-186.0	184.0	6	0.030
		$\sum f = 200$	$\sum P = 1.000$

The left frequency distribution has four classes whose centres (boxed) line up with those of the right distribution. Despite this fact their relative frequencies differ considerably. This is because the classes on the right distribution are twice as wide as those on the left and hence the corresponding class will typically have about twice as much data as the ones on the left distribution.

To remove the arbitrariness of the choice of class width, one defines the **relative frequency density**[6] (symbol p) to be the relative frequency divided by the class size. In symbols,

$$\boxed{p = \frac{P}{\Delta x}} = \frac{f}{(\Delta x \cdot \sum f)} \quad .$$

Example:
Using a p column in the previous example yields:

$\Delta x = 2.0$ cm

Height(cm)	x(cm)	f	p(cm^{-1})
169.0-171.0	170.0	4	0.010
171.0-173.0	172.0	10	0.025
173.0-175.0	174.0	14	0.035
175.0-177.0	176.0	73	0.183
177.0-179.0	178.0	70	0.175
179.0-181.0	180.0	17	0.043
181.0-183.0	182.0	10	0.025
183.0-185.0	184.0	2	0.005
		$\sum f = 200$	

$\Delta x = 4.0$ cm

Height(cm)	x(cm)	f	p(cm^{-1})
170.0-174.0	172.0	20	0.025
174.0-178.0	176.0	128	0.160
178.0-182.0	180.0	46	0.058
182.0-186.0	184.0	6	0.008
		$\sum f = 200$	

Even with the coarse class widths, one sees that the relative frequency density p at the same values of x are now roughly equal. Note that summing the relative frequency density p column is meaningless.

The odd units for p (in the last example reciprocal centimetres, cm^{-1}), reminds us that to return to a proportion we must multiply by some interval of the variable x, namely Δx.

Example:
What proportion of basketball players had a height of 178 cm to the nearest cm? In this case the heights would have to lie between 177.5 cm and 178.5 cm, so $\Delta x = 178.5$ cm $- 177.5$ cm $= 1.0$ cm centred on $x = 178.0$ cm. At $x = 178.0$ cm, p is 0.175 cm^{-1} so, solving our density formula for P, gives

$$P = p \cdot \Delta x = (0.175 \text{ cm}^{-1})(1.0 \text{ cm}) = 0.175$$

Approximately 17.5% were 178 cm high to the nearest cm.

In this example our approximation is only valid because Δx was narrow enough that we could expect p to be relatively constant over the interval. Determining the proportion over wide intervals can be done through the introduction of cumulative frequency.

The usefulness of these new columns will become clearer when we interpret their meaning graphically.

Assignment: For each case study that is a frequency distribution append a column for the relative frequency, P. For grouped frequency distributions also append the relative frequency density p.

[6] The term *density* is appropriate here because, just like a mass density where we divide mass by volume to get mass per unit volume, here we are interested in getting proportion per unit of x.

*1.6 Cumulative Frequency

Further columns, useful for understanding how far into the data a particular variable value is, can also be added to frequency distribution tables.

1.6.1 Cumulative Frequency (*<Cf*)

The **(less than) cumulative frequency column** lists the number of observations below the upper limit of the class in a grouped frequency distribution or the number of observations at or below a specific value in an ungrouped one. The caption at the top of the column is $<Cf$. In symbols:

$$<Cf_i = f_1 + f_2 + \ldots + f_i \quad .$$

Example:
A convenience store recorded a sample of the groceries bought by customers with the following results:

Purchase Value ($)	f	$<Cf$
0 - 10	12	12
10 - 20	30	42
20 - 30	35	77
30 - 40	20	97
40 - 50	23	120
	$\sum f =$ 120	

Notice: The $<Cf_3$ value, 77, in the third class means that 77 of the purchase values are less than $30.

Always sum the f column as a check on the last entry in the $<Cf$ column. (Boxed in the example.)

Cumulative frequency is useful for finding the number of observations within a given interval of the range. For instance in the above example the number of observations greater than (or equal) to $10 but less than $40 is $97 - 12 = 85$. The cumulative frequency column will also be useful for constructing certain types of graphs and for calculating medians and fractiles of data presented as a frequency distribution.

A **greater than cumulative frequency column** could also be constructed (symbol $>Cf$) which proceeds through the data starting at the *highest value* of x and culminating with the total frequency at the *lowest* value. We will not be using the $>Cf$ in our calculations. The $<Cf$ column and the $>Cf$ column provide the same information so in practice reference to a cumulative frequency column usually means the $<Cf$ column.

A total is meaningless at the bottom of cumulative frequency columns.

1.6.2 Cumulative Relative Frequency (*<CP*)

A (less than) **cumulative relative frequency** (symbol $<CP$) could also be added:[7]

$$<CP_i = P_1 + P_2 + \ldots + P_i \quad .$$

[7] Also sometimes referred to as a **relative cumulative frequency** .

Example:
Egg production over one week was measured for a sample of chickens with the following results:

Production (eggs/wk)	# of Chickens	P	$<CP$
0	1	0.05	0.05
1	2	0.10	0.15
2	3	0.15	0.30
3	4	0.20	0.50
4	6	0.30	0.80
5	3	0.15	0.95
7	1	0.05	1.00
	$\sum f = 20$	$\sum P =$ 1.00	

The fifth entry (x=4 eggs/wk) having $<CP = .80$ indicates that 80% of the chickens laid 4 eggs or less.

As with the introduction of relative frequency, cumulative relative frequency makes the result effectively independent of the number of observations; the final entry will always be 1.0 (within rounding error).

Summing cumulative relative frequency is meaningless.

As an aside, note that it is meaningless to create a cumulative relative frequency *density* column involving p since we would have to multiply by the class width Δx to get a useful proportion before summing and this is just the same as $<CP$. Our motive for introducing p was to get a property independent of class width and it turns out that $<CP$ is already independent of class width. This may be verified by adding the $<CP$ column to the first two tables in Section 1.5.2. This said, it is the case that for relative frequency density it is common to replace the class values p_i with a continuous function $p(x)$ that goes through them.[8]

Assignment: For each case study that is a frequency distribution append columns for cumulative frequency, $<Cf$, and cumulative relative frequency, $<CP$.

[8] We will see in the next section a frequency polygon that is just such a function. Alternately one could try fitting a curve to the points. For the student who has calculus: If we had such a function $p(x)$ one would be able to define the *continuous* cumulative relative frequency function $<CP(x)$ by the integral:

$$<CP(x) = \int_{-\infty}^{x} p(x')\, dx'$$

which would approximate our class $<CP_i$ values at similar x. Here one sees the differential relative frequency $dP = p\, dx$ showing up which replaces our $P = p \cdot \Delta x$ when one moves to the continuum limit. The integral replaces our discrete summation. When we look at our normal distribution table we will see it is a tabulated version of just such a beast.

1.7 Graphical Representation of Data

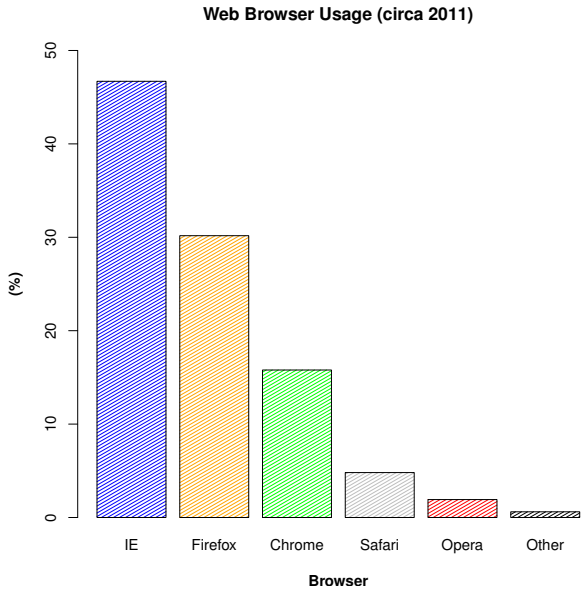
Another method of analyzing the variation in the statistical variable is by constructing a picture of the variation. This can be done by statistical chart or by statistical graph depending on the data type. The basis for both of these graphics is a statistical summary table.

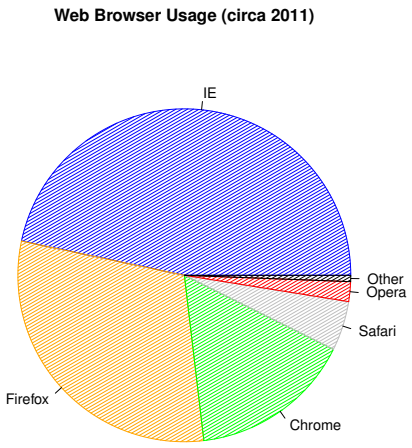
Statistical charts are a visually appealing method of presenting data. They are often used in connection with presenting data to a wide audience such as is done in annual reports of companies and budgets of municipal agencies. Many people do not have the artistic skill to present data this way. Computer packages, such as many spreadsheet programs, can perform the graphic techniques required to create statistical charts. Charts are most often used to present qualitative (categorical) statistical information.

Example:
Represent the following web browser usage data from 2011 as a bar chart and as a pie chart.[9]

Web Browser	Usage (%)
Internet Explorer	46.7
Firefox	30.2
Chrome	15.8
Safari	4.8
Opera	1.9
Other	0.6

Solution:
For bar charts the categories are placed along the horizontal axis. The bars are made of equal width with space between them while the height is plotted as the frequency or relative frequency. To make a pie chart multiply the relative frequency by 360° to get the angle subtended by the category. (For example, $(0.467) \cdot (360°) = 168°$ is the angle of the wedge for Internet Explorer.) In each case the area of the region is proportional to the (relative) frequency of the category. Bar charts often appear oriented horizontally.





We will not be emphasizing chart construction in this course.

The construction of statistical graphs is more important in an introductory statistics class because a graph is the basis for understanding statistical distributions. Statistical graphs differ from charts in that the graphs are plotted on an axis system with two axes. The axes are treated as being continuous in a mathematical sense. That is, a ratio scale can be set up on both axes. As such graphs are used for quantitative data.

Consider the following grouped frequency distribution representing the payments on satellite phone contracts made by a group of surveyors. Additional columns for relative frequency (P), relative frequency density (p), less than cumulative frequency ($<Cf$), and less than cumulative relative frequency ($<CP$) have already been appended. (The former were introduced in Section 1.5 and the latter in (optional) Section *1.6 .)

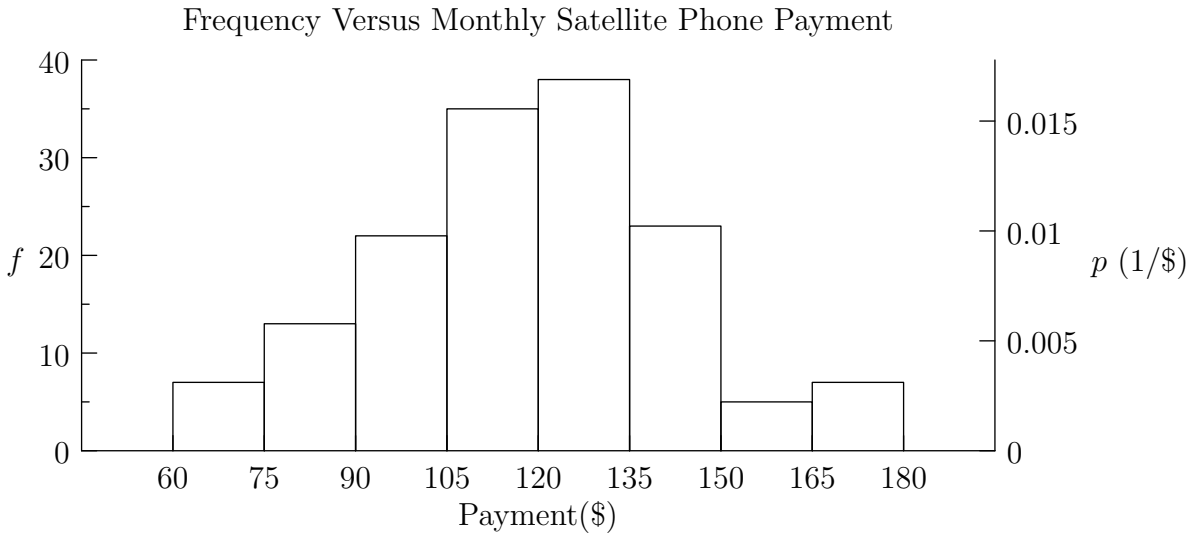
Monthly Satellite Phone Payments					
Payment ($)	f	P	p (1/$)	$<Cf$	$<CP$
60 - 75	7	0.047	0.0031	7	0.047
75 - 90	13	0.087	0.0058	20	0.134
90 - 105	22	0.147	0.0098	42	0.281
105 - 120	35	0.233	0.0156	77	0.514
120 - 135	38	0.253	0.0169	115	0.767
135 - 150	23	0.153	0.0102	138	0.920
150 - 165	5	0.033	0.0022	143	0.953
165 - 180	7	0.047	0.0031	150	1.000
	$\sum f = 150$	$\sum P = 1.000$			

We will now consider graphs involving frequency versus x and cumulative frequency versus x for this example.

[9] Data courtesy StatCounter (gs.statcounter.com). Charts produced using **R** (www.r-project.org). Note that this table is an example of a qualitative (categorical) relative frequency distribution. If instead we had listed the total number of hits with each browser on the counting servers over the measurement time period this would have been a qualitative frequency distribution. What is the variable and what is its level of measurement?

1.7.1 The Histogram and Frequency Polygon

One way of graphing frequency f versus the variable x is to make a **histogram**. The following is a histogram for the previous data.



Rather than plotting points, the histogram is drawn with rectangles of height given by the frequency of the class f and width given by the class width Δx. Unlike a bar chart, no space is placed between the rectangles. On the one hand, plotting a rectangle rather than a point makes sense since the frequency is associated with an entire class. The reason for plotting rectangles is also motivated by the desire to associate not the height but rather the **area** under the curve with the number of observations. This is reasonable since the area of a particular rectangle in the histogram is proportional to the frequency because the area is simply $f \cdot \Delta x$ and Δx is constant. If one rectangle represents the number of observations for the given class it follows that the the **total area** under the frequency histogram represents the total number of observations.

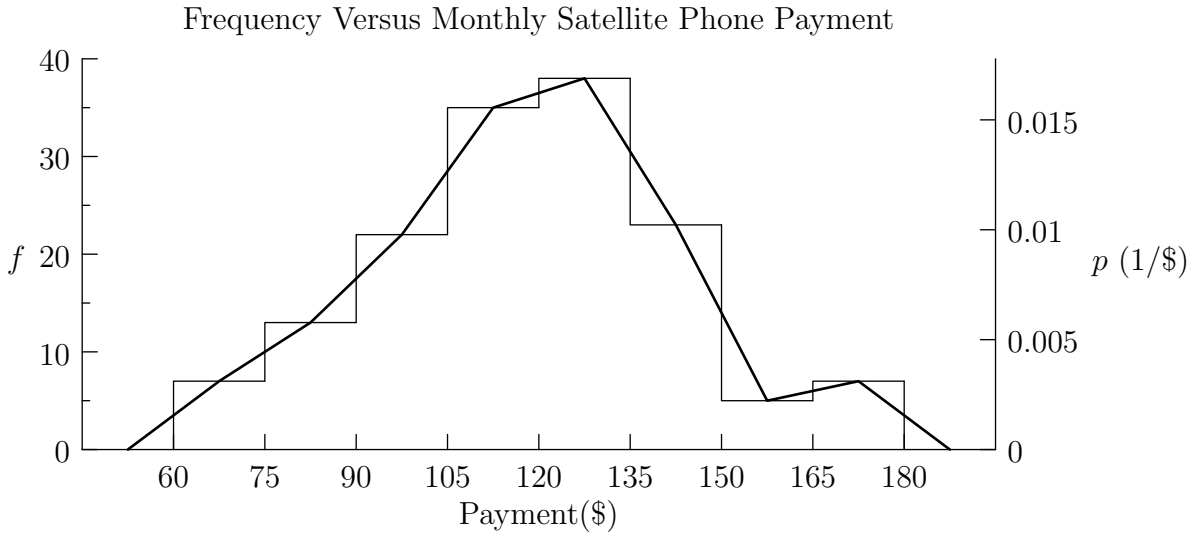
On the right vertical axis is a scale for the relative frequency density p. The graph of p versus x looks the same as f versus x because p is proportional to f. A histogram of relative frequency density p versus x is especially useful since the area of a particular rectangle now becomes $p \cdot \Delta x$ which is just the relative frequency (proportion) P associated with that class. The total area of all the rectangles is then the sum of the relative frequencies which is exactly 1.

If one is interested in the fraction of observations that fall between two values such as \$110 and \$130, simply calculate the area under the histogram curve between the two values. A rough rectangular estimate in this case would see $p = .015(1/\$)$ times $\Delta x = \$130 - \$110 = \$20$ equalling .30 or 30% of the observations. A more precise method of calculating this proportion graphically can be found by plotting cumulative relative frequency versus x. (Section *1.7.2.)

Histograms give a quick visual representation of the data, especially for large data sets. One can immediately discern the shape of the shape of the data and notice any distinct features such as outliers. Some of this information can be dependent upon the class width Δx chosen. For instance in the above histogram if we had made the class width twice as wide (\$30 rather than \$15) we would no longer discern the slight dip then rise in the frequency of satellite phone payments at the high end of the data.

One defect of the histogram is how boxy it is. This is an artifact of our (arbitrary) choice of class width Δx when we made our grouped frequency distribution. An improvement that smooths out this

artifact is the **frequency polygon** shown in the following graph for the same data. The frequency polygon has been superimposed over the original histogram for comparison.



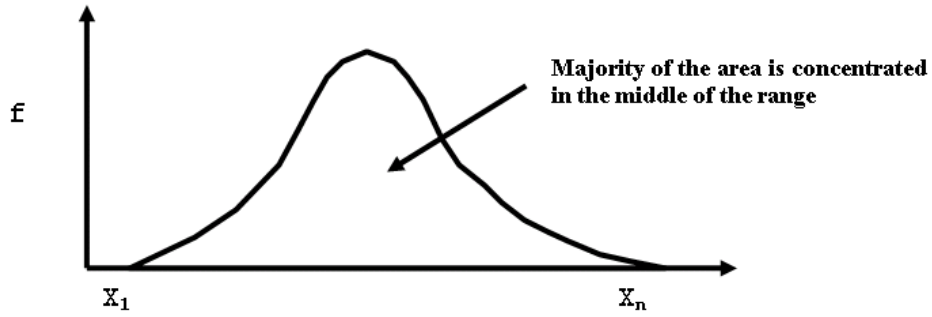
The frequency polygon is plotted by plotting the frequency f (or relative frequency density p) versus the **midpoint** of each class. Note one must add an empty class before the first and after the last class so that the frequency polygon can be extended to zero. If one compares the frequency polygon to the original histogram one sees that it preserves the area under the curve. Any triangle that is under the histogram which is not under the frequency polygon is exactly compensated by a neighbouring triangle of identical area that is under the frequency polygon but not under the histogram. For this reason we do not attempt to further smooth out our line joining the points as we would for another nonlinear graph. When we talk about arbitrary frequency distributions of continuous variables we will, in future, show a perfectly smooth curve which we imagine would result from a frequency polygon for a large amount of data with small class width.

The frequency polygon is an attempt to replace the histogram by a smooth mathematical curve whose equation can be analyzed. Since the boxy nature of the histogram is largely an artifact of the chosen class width this smoothing is desirable. It is said that the mathematical curve models the distribution.

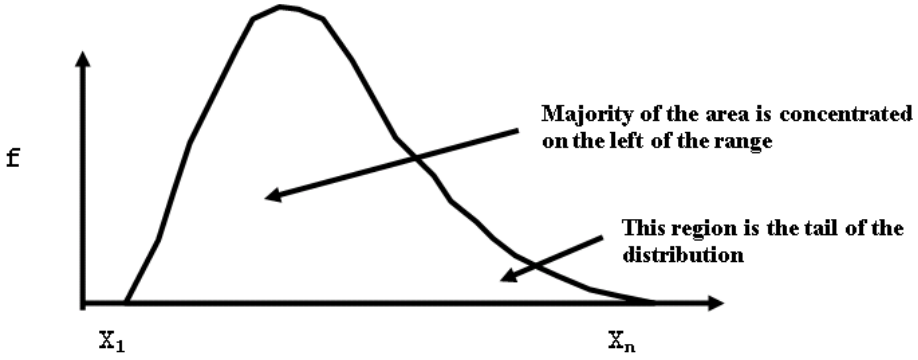
The Shape Of A Distribution

The **shape** of a distribution refers to the shape of the frequency polygon. Some general categories are as follows.

1. Many distributions have the bulk of the observations in the middle of the range with a few observations located at the extremities of the range. As a result, the distribution curve takes the following **symmetric** shape.

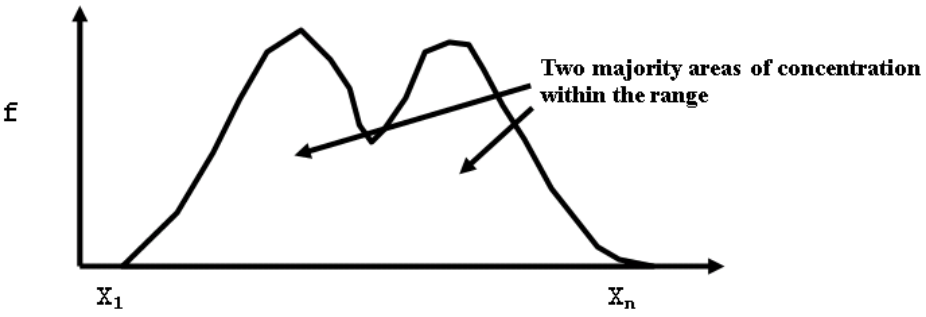


2. Sometimes the areas of concentration are not located at the centre but are located more to one side or the other of the range. As a result the distribution curve takes the following **skewed** shape. If the tail is on the right side the distribution is **positively skewed** or **skewed to the right**. If the tail of the distribution is on the left side we say that it is **negatively skewed**, or **skewed to the left**.



The above distribution is positively skewed.

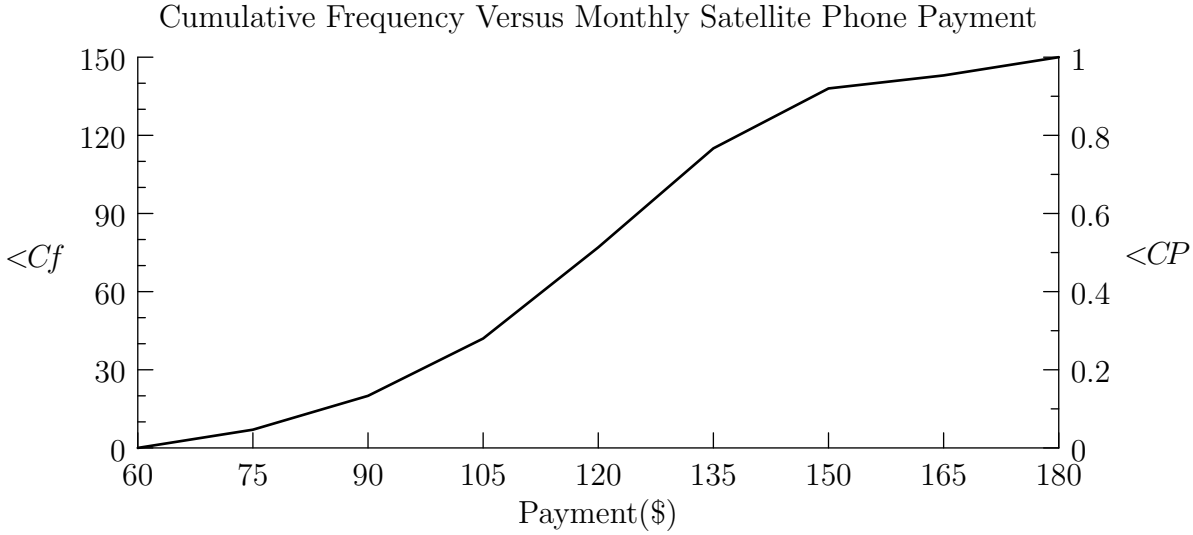
3. The examples so far have had only a single area of concentration reflected by the frequency polygon having a single peak. Such distributions are called **unimodal**. Sometimes there is more than one area of concentration in the distribution. This is reflected in the frequency polygon by several peaks. If the distribution has two peaks, it is referred to as a **bimodal** shape. A bimodal shape indicates that there are two underlying trends within one data set.



Assignment: Sketch a histogram and frequency polygon for each case study that is a grouped frequency distribution. Describe their shapes.

*1.7.2 Cumulative Frequency Polygons (Ogives)

A plot of cumulative frequency $<Cf$ versus x is called a **cumulative frequency polygon** or **ogive** (pronounced "oh jive"). The ogive for the previous data is as follows:

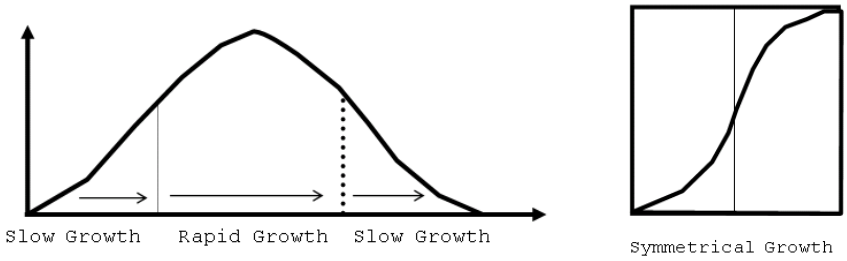


For a less than cumulative frequency ogive one plots $<Cf$ versus the **upper limit** of each class. The lower limit of the first class is given the value zero. That this should be the case follows from the fact that at the very lower end of the class none of the frequency for the given class has accumulated but by the time the upper limit occurs all the observations up to that point have to have been included. By the time the upper limit of the last class has arrived all the observations have been accounted for so its vertical value is $\sum f$.

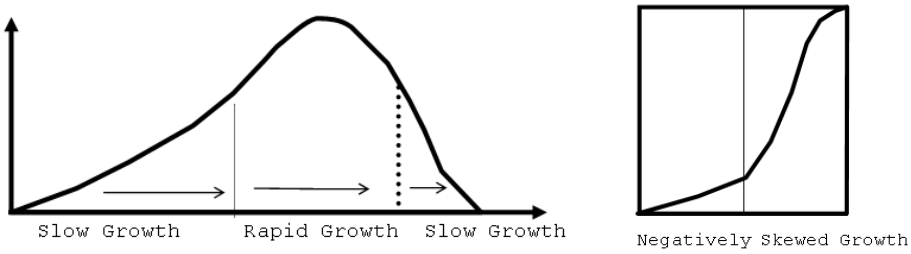
On the right axis has been added the scale for the less than cumulative relative frequency $<CP$. An ogive plotting cumulative relative frequency $<CP$ versus x is useful if one is interested in the fraction of observations occurring up to a point rather than the actual number of observations. In this case the upper limit of the last class is associated with the value 1.0 since 100% of the observations have been accounted for by this point. Note that when plotting an ogive one can restrict oneself to plotting $<Cf$ since one can always add a $<CP$ axis afterward just by ensuring 1.0 (100%) is placed opposite to the total number of observations $\sum f$. (i.e. one does not need to actually calculate the $<CP$ values as we did above.)

The Path Traced Out By The Ogive Curve

If a distribution is perfectly symmetric, the ogive curve will be perfectly "S" shaped. This is because observations build slowly in the extremities of the range and quickly at the centre of the range.



If the distribution is negatively skewed, it will build area slowly for a greater distance at the beginning of the range and the ogive curve will have distorted "S" shape.

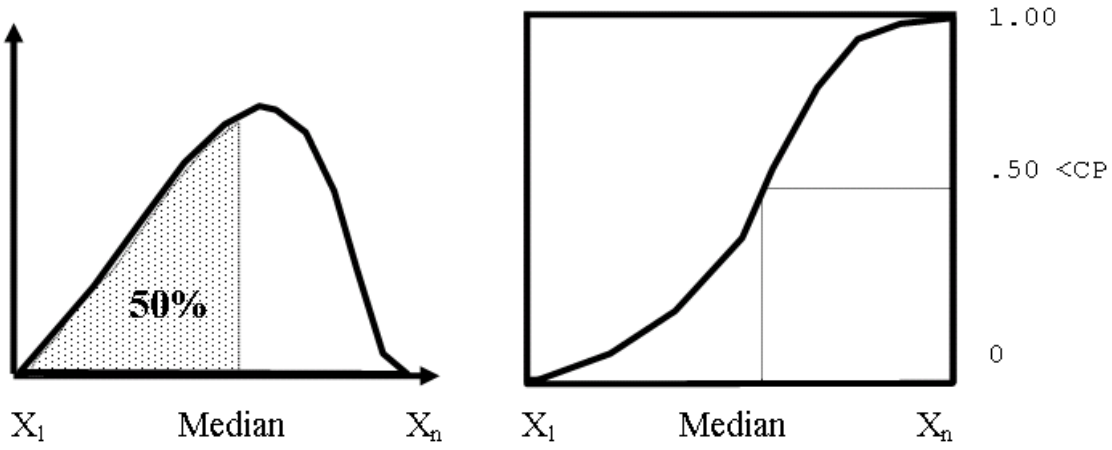


A positively skewed distribution will similarly produce an ogive with a plateau on the right side of the graph.

Utility of an Ogive

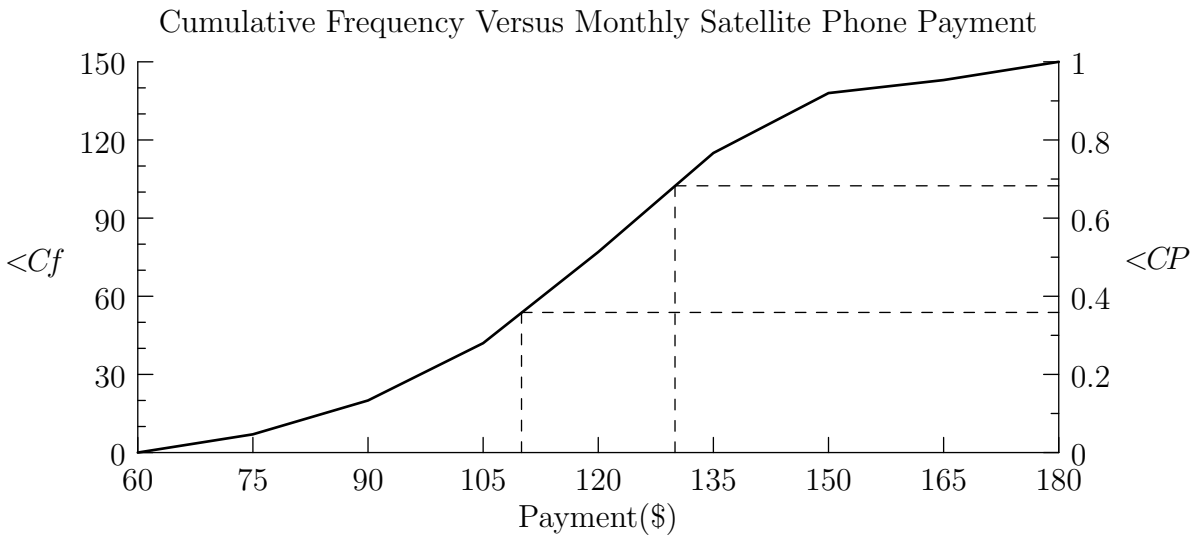
The ogive curve, from a geometrical viewpoint, finds areas under the frequency polygon curve.

Using the ogive, the exact centre of the data can be determined. Start at the the .5 (50%) figure on the *<CP* axis and extend a horizontal line to the left until the ogive curve is intersected. The value of x to which this corresponds is in the exact middle of the data with half the observations below it and half above it. We will see that this value is called the **median**.

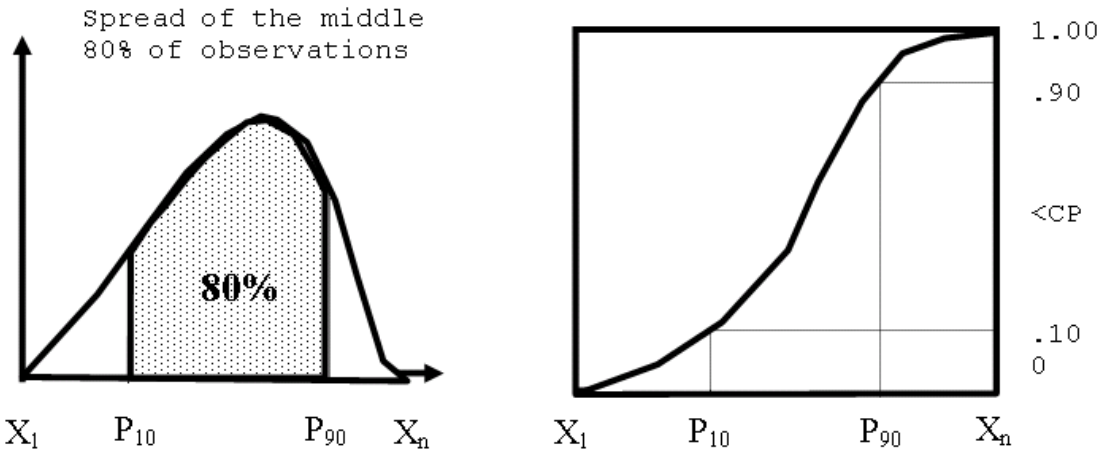


For our satellite phone example one may verify graphically that the median is approximately $120.00 .

The proportion of observations between any two values in the range can be found by subtracting the corresponding *<CP* values on the *<CP* axis. For instance, in our previous example, if we were interested in knowing the fraction of observations that lie between $110 and $130 we could extend vertical lines upward from these values on the x axis until they intersect the ogive curve. From there we can extend horizontal lines to the right to find what the cumulative proportion of observations are up to $110 and $130 respectively. The values are (see following graph) 0.36 and 0.68 respectively. The **difference** of these values (0.32) indicates that 32% of the observations lie between $110 and $130. This may be compared with our rough estimate of 30% for the area under the histogram over this interval from before.



Finally we can invert this argument and ask between what two values of x the middle 80% of the data lie. This involves finding what x values correspond to $<CP = 0.10$ and $<CP = 0.90$ respectively as shown in the following diagram. These x values are known at the 10^{th} and 90^{th} percentiles (P_{10} and P_{90}).



One may confirm that in the satellite phone example the approximate tenth and ninetieth percentiles are $P_{10} \approx \$84.00$ and $P_{90} \approx \$148.00$ respectively.

Assignment: Sketch a cumulative frequency polygon (ogive) for each case study that is a grouped frequency distribution.

1.8 Measures of Central Tendency

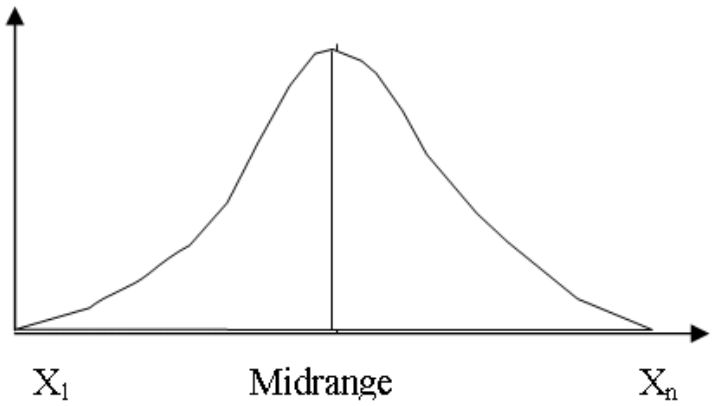
Consider an array of data for a statistical variable x. The following is a partial list of the 105 observations:

2300	2000	2300
2100	2300	2700
2300	2600	2300
⋮	⋮	⋮
1900	2400	2300

We have seen how to condense statistical data to tables and graphs. For descriptive purposes, statisticians sometimes use a single number to represent an entire array of observations. A statistical measure of the centre of a distribution is a value that is representative of the entire array of observations. Another name for a measure of the centre is an average value. As the name suggests, there is a tendency for a collection of observations to cluster around some central value.

On a perfectly symmetric distribution, a possible central value is the midrange.

$$\boxed{\text{Midrange} = \frac{x_1 + x_n}{2}}$$



The midrange in general is a poor central value because it is sensitive to **outliers** (values of the variable that are a great distance from the majority of the observations). It also is independent of most of the data.

When a person unfamiliar with statistical calculations uses the word "average", it is usually understood that the person means the following:

$$\boxed{\text{"average"} = \frac{\text{sum of all observations}}{\text{number of observations}}}$$

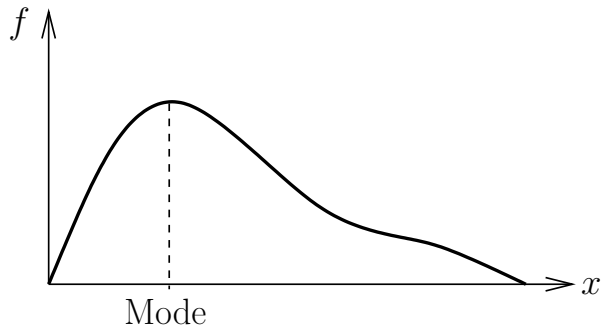
To statisticians there are a number of types of averages used to represent data arrays. Three of these averages have special properties that are related to the distribution curves of the observations. These are the:

1. Mode
2. Median
3. Arithmetic Mean

Each of these averages represents the data array best depending upon what is to be portrayed and the shape of the distribution. This will be more apparent when these measures are calculated.

1.8.1 The Mode

The mode is that value of the variable that occurs the most often. Since it occurs the most often, it is the x value with the greatest frequency on the frequency polygon.



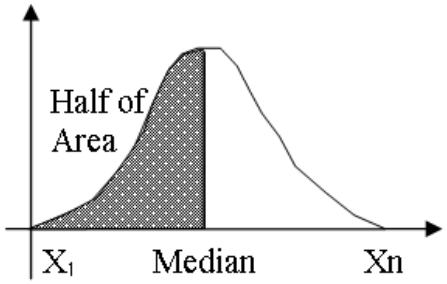
Some of the important characteristics of the mode are:

1. The mode may *not be unique* since a distribution may have more than one mode.
2. There is no calculation required to find the mode since it is *obtained by inspection* of the data.
3. If the mode is used as a representative of individual values in the array, it will be in error less often than any other average used. If the size of this error is important such as in sizes in the manufacturing business, the mode is a good representative of the data array.
4. In a *negatively skewed distribution*, the mode is *to the right of the midrange*. In a *positively skewed* distribution the mode is *to the left of the midrange*. Draw sketches to illustrate this relationship.
5. If what is to be portrayed is a *typical* value in an array, it is most typical because no value occurs more often.
6. For data measured at the nominal level, it is the only average that can be found.

An example where the mode would be an appropriate measure of the centre would be for sizes of shoes sold to men. The mode of the data, say size 10, would answer the question of what size of shoe a typical adult male customer bought. (Compare this to the average (mean) size, say 10.2, which would not even be available as an actual shoe size!)

1.8.2 The Median

The median is that value in the distribution such that half of the observations are less than this value and half are greater than this value. Because the area under the distribution curve represents the total number of observations, the median is that $x$value on the frequency polygon such that half of the area under the curve lies to the right and half of the area lies to the left of the value. In (optional) Section *1.6 it was shown that the median can be read directly off the ogive curve by locating the variable value corresponding to $<CP=.50=50\%$.



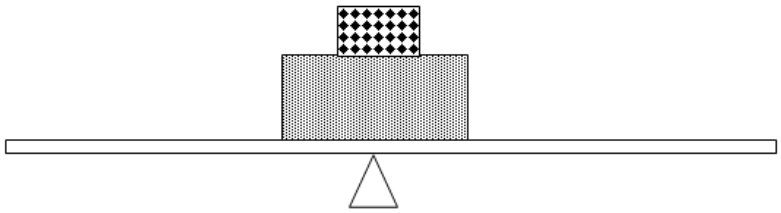
Some important characteristics of the median are:

1. The median is representative of the data array because it is in the geometrical centre of the distribution. It is the exact halfway point. Half the observations are less and half are greater than the median value.
2. In a positively skewed data set, the median will be to the right of the mode and on negatively skewed data sets to the left. Draw sketches to illustrate this relationship.
3. The median is *always unique* for a data set.
4. The median is useful for descriptive purposes when the data set is *skewed* because of the constancy of its location. It is always exactly the middle observation in the data array when the array is *rank ordered* and it is insensitive to outliers.
5. The median can be found for data that has an ordinal level of measurement or higher.

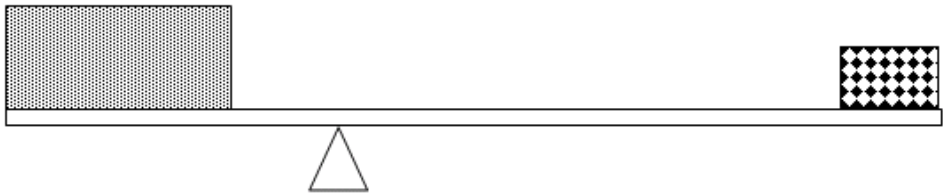
An example of when the median is an appropriate measure of the centre is the reporting of individual annual income in a population. The median income will accurately reflect what a common person in a country earns. The average (mean) income, as shown next, is sensitive to outliers which does not occur for the median. As such the mean income of a population will give a false representation of the earnings of individuals if, for instance, wealth (income) is concentrated in a small subgroup of individuals in the population.

1.8.3 The Arithmetic Mean

If the word "average" is used without a qualifier, the arithmetic mean is the average meant. Its location on a distribution curve is more abstract. It has the property of being the x value at the "balance point" or "centre of gravity" of the distribution curve. Think of the teeter-totter example:

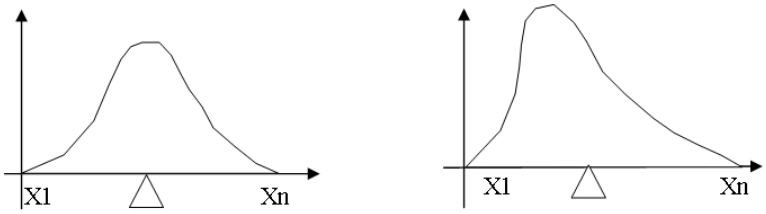


The balance point of the teeter-totter arrangement above is at the centre of the system.



To make the teeter totter arrangement above balance, the balance point must be moved away from the centre of the system.

For a statistical distribution, the arithmetic mean by definition is the value of the variable that is at the balance point on a frequency distribution graph.



The diagram on the left shows that the balance point of this distribution is located at the same place as the median and the mode, namely at the middle of the range. The balance point of a skewed distribution is shifted away from the middle of the range.

Some important characteristics of the arithmetic mean are:

1. The arithmetic mean is the average which is found by the procedure:

$$\text{"average"} = \frac{\text{sum of all observations}}{\text{number of observations}}$$

2. The location of the mean is dependent upon the shape of the distribution. It may not always be representative of the centre.
3. The arithmetic mean has mathematical properties that the other two averages do not have. (Balance point property).
4. Because of its mathematical properties, the mean lends itself to further mathematical analysis of data sets unlike the other two averages.
5. The mean is used as a measure of the centre in statistical inference because these mathematical properties are important in that instance.
6. The mean requires a quantitative variable, typically at the ratio level of measurement.

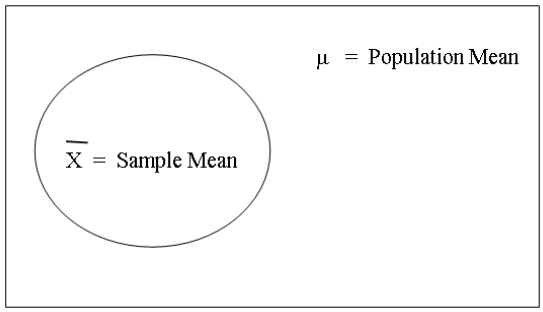
Since the arithmetic mean is the familiar average that one is acquainted with, examples of its usage abound (mean height, mean fuel consumption, mean grade, etc.) However, as we have seen in the examples for the mode and median, there are cases where the arithmetic mean is an inappropriate central measure to use. That said, the mean will dominate our later discussion due to its convenient mathematical properties.

The Arithmetic Mean in Samples and Populations

Because the mean is used as a measure of the centre in statistical inference problems, there is a different symbol that is used to represent the mean in a sample and the mean in a population.

$$\left.\begin{array}{l}\mu = \text{Population Mean} \\ \bar{x} = \text{Sample Mean}\end{array}\right\} = \frac{\text{sum of all observations}}{\text{number of observations}}$$

Visualize the relationship between these two symbols with the following diagram:



If the distribution is symmetric all averages are located at midrange. On a skewed distribution curve there is a spread of the three averages. It is approximately true that the median is one third of the way from the mean towards the mode in a skewed distribution.

Assignment: For case studies that are grouped frequency distributions, estimate the mode and mean from your frequency polygon. If you plotted their ogives, also estimate the median.

We next discuss how to find the three measures of the centre for each of the three methods of data presentation.

1.9 Determining the Mode

1.9.1 Raw Data

No calculation is required. Determine which value occurs the most often by rank ordering the data.

Example:
The ages of people in a certain occupation were observed to be: (yrs)

23, 67, 26, 32, 26, 45, 26, 54

The array rank ordered is:

23, 26, 26, 26, 32, 45, 54, 67

The mode of the data is 26.0 years.

1.9.2 Ungrouped Frequency Distributions

No calculation is required. Determine the value that has the highest frequency of occurrence by inspection of the table.

Example:

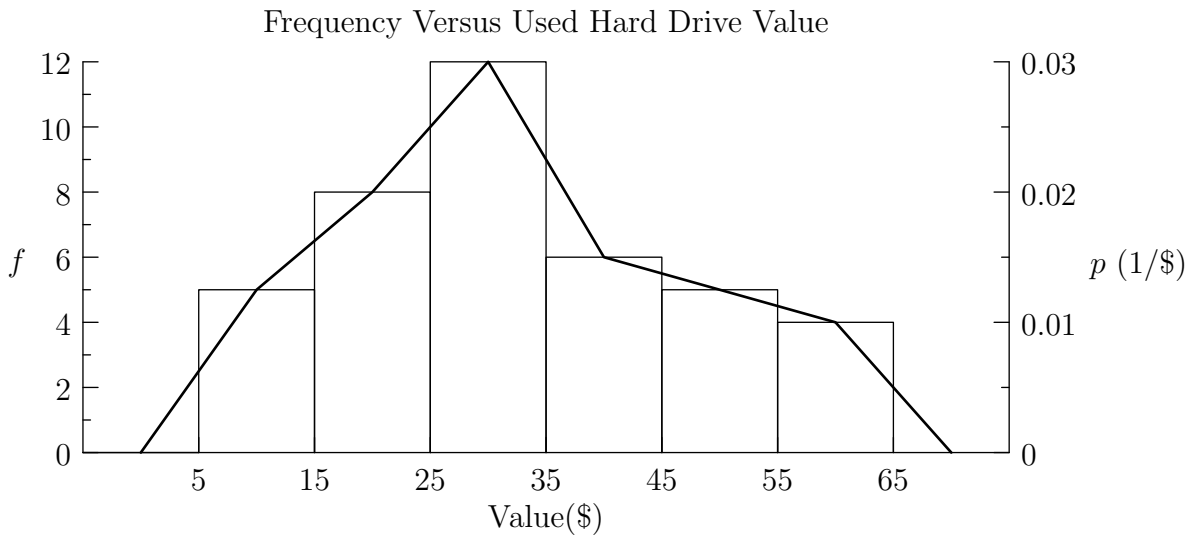
Wage ($/hr)	f
9.25	4
10.50	10
12.50	5
14.50	1
	$\sum f = 20$

The wages of employees on a summer work project were as shown. The mode of the data is 10.50 $/hr since that value has the greatest frequency.

1.9.3 Grouped Frequency Distributions

Recall that the individual observations *are lost* when data are summarized by grouping. It is therefore *impossible* to locate the *exact* value within the array that is at the centre. *Approximations* to these averages are possible by comparing the definitions of these values to the frequency polygon (for mode and mean) and the ogive curve (for the median).

The mode is the value of the variable observed the most often (highest frequency). Consider the following frequency polygon for a grouped frequency distribution of used computer hard drive prices.



For a grouped frequency distribution the **midpoint of the class with the highest frequency** is therefore taken as the mode because, by looking at the frequency polygon, the value of the variable under the peak is the midpoint of the class with the highest frequency.

Example:
Find the mode of the following grouped data which was plotted in the previous histogram:

Value of Used Computer Hard Drive	
Value ($)	f
5.00 - 15.00	5
15.00 - 25.00	8
25.00 - 35.00	12
35.00 - 45.00	6
45.00 - 55.00	5
55.00 - 65.00	4
	$\sum f = 40$

Solution:
Since the class with highest frequency is \$25.00 - \$35.00 the mode is its midpoint:

$$\text{mode} = \frac{\$25.00 + \$35.00}{2} = \$30.00 \ ,$$

which agrees with the frequency polygon.

Assignment: For each case study find the mode. Remember the mode will have the same units as the variable (as will all measures of central tendency).

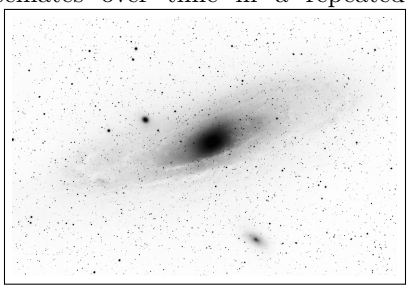
1.10 Calculating the Arithmetic Mean

1.10.1 Raw Data

The arithmetic mean is calculated as the sum of all observations divided by the number of observations. Place the observations in rank order with the sum of all observations at the foot of the column. Use the appropriate captions and footing symbols.

Example:

A *Cepheid Variable* is a type of star whose brightness oscillates over time in a repeated fashion. By measuring this period of oscillation, astronomers are able to determine how bright the star is in absolute terms. By comparing this with the star's apparent brightness and knowing that intensity falls of with distance squared, astronomers can calculate the actual distance to these stars. Using this technique an astronomer calculates the distance to five such stars found in the *Andromeda Nebula*[10] (right) with the following results (in millions of light-years[11]):



The Andromeda Nebula

$$2.6,\ 2.4,\ 2.5,\ 2.4,\ 2.7 \text{ (Mly)} .$$

Calculate the mean distance, and hence estimate how far the Andromeda Nebula is from our sun.

Solution:
Proceeding in a tabular format:

x(Mly)
2.4
2.4
2.5
2.6
2.7
$\sum x = 12.6$

The arithmetic mean of a sample is calculated as:

$$\boxed{\bar{x} = \frac{\sum x}{n}}$$

In our case:

$$\bar{x} = \frac{\sum x}{n} = \frac{12.6 \text{ Mly}}{5} = 2.520000 \text{ Mly} = 2.52 \text{ Mly}$$

It is helpful to rank order the data for purposes of checking the answer. The mean is a central value. The calculated value can be compared to the range of values. *It must lie inside the range.* Pay attention to the symbol used for the sample mean.

1.10.2 Ungrouped Frequency Distributions

To calculate the sum of all observations in this case, the frequency of occurrence of each observation must be taken into account when summing all observations.

Example:
A statistician played the lottery every week for four years and tabulated all of his resulting winnings in the following table for his article *Why I decided to stop playing the lottery after four years.*

[10] The Andromeda Nebula is often referred to by its number in the Messier astronomical catalogue, M31. Original photo by Boris Štromar altered to grayscale with black-white inversion.

[11] One light-year is the distance light can travel in one year, about 9.46×10^{15} metres.

Winnings x(\$)	Tickets f	xf(\$)
0.00	172	0.00
5.00	20	100.00
10.00	6	60.00
100.00	2	200.00
	$\sum f = 200$	$\sum xf = 360.00$

For ungrouped data the arithmetic mean formula is:

$$\boxed{\mu = \frac{\sum xf}{\sum f}}$$

In our case:

$$\mu = \frac{\sum xf}{\sum f} = \frac{\$360.00}{200} = \$1.80000 = \$1.80$$

If each ticket cost \$5.00, how much, on average, did the statistician donate to the lottery company each time he bought a ticket?

1.10.3 Grouped Frequency Distributions

Consider the grouped frequency distribution introduced in Section 1.9.3. By examining the frequency polygon, the point plotted for a class is at the midpoint of the class. In calculating the mean, the assumption is made that the values in the class are approximately equal to the midpoint of the class. An approximation to the mean can be found by replacing the interval by the midpoint and proceeding to do the calculation as done for an ungrouped frequency distribution.

Example:
Calculate the mean of the grouped frequency distribution introduced in Section 1.9.3:

Value of Used Computer Hard Drive	
Value (\$)	f
5.00 - 15.00	5
15.00 - 25.00	8
25.00 - 35.00	12
35.00 - 45.00	6
45.00 - 55.00	5
55.00 - 65.00	4
	$\sum f = 40$

Solution:
We replace each class interval by its midpoint and proceed as in the ungrouped case.

x(\$)	f	xf(\$)
10.00	5	50.00
20.00	8	160.00
30.00	12	360.00
40.00	6	240.00
50.00	5	250.00
60.00	4	240.00
	$\sum f = 40$	$\sum xf = 1300.00$

Approximate arithmetic mean:

$$\boxed{\bar{x} = \frac{\sum xf}{\sum f}} = \frac{\$1300.00}{40} = \$32.500000 = \$32.50$$

The Number of Decimal Places to Retain in a Calculation

Use **one decimal place more precision** in the calculated value than is found in the precision of the data array. An exception is if the data consist of exact dollar amounts, in which case the precision of the calculation will be **to the nearest cent**. The example in Section 1.10.1 illustrates the general rule, while those of Sections 1.10.2 and 1.10.3 illustrate the exceptional case for money. Rounding of final values is important as it communicates how accurately the data was measured.

In later units where a calculated statistic like the sample mean is used in further calculations the unrounded calculator value should be used in the formula to avoid introducing unnecessary error into the calculation. Rounding only occurs as a final step before reporting a result.

Populations and Samples

In applied problems the mean is the average that is most often used in statistical inference situations. Because of this, there are two different symbols used to designate the mean, $\bar{x}$ and μ. For samples use the symbol $\bar{x}$ as was done in Sections 1.10.1 and 1.10.3. For populations use the symbol μ as demonstrated in Section 1.10.2. The procedure is the same for calculating a population mean as for a sample mean. The two formulae use different symbols to show that the data sets represent different situations. The following table summarizes the formulae for the mean.

	Population	Sample
Raw Data	$\mu = \dfrac{\sum x}{N}$	$\bar{x} = \dfrac{\sum x}{n}$
Frequency Distribution	$\mu = \dfrac{\sum xf}{\sum f}$	$\bar{x} = \dfrac{\sum xf}{\sum f}$

Statistical Keys on the Calculator

The mean can be calculated directly on a calculator with the use of its statistical keys. The calculator must be placed in statistical mode before the data is entered as discussed in Section 1.4. The calculator has only one symbol to represent a mean, often $\bar{x}$ because typically when a mean is calculated it is done on sample data. Use this same key to calculate the mean value for populations but replace the symbol with the μ symbol when writing the answer.

Assignment: For each case study find the mean. Include the following steps as demonstrated in the previous examples.

Requirements for a Complete Solution (F.S.A.R.U.)

Formula State the formula used, including appropriate symbols.

Substitution Substitute the values for your problem. Remember to add any columns to a table that are required for the calculation of those values.

Answer Write your answer including sufficient extra decimals of significance to allow for rounding.

Roundoff Round to the appropriate number of decimal places.

Units Include the appropriate units.

1.11 Calculating the Median

1.11.1 Raw Data

Rank order the data and pick the value in the middle position. To identify the median, two procedures are necessary, first finding the position of the median and then its actual value at that position:

$$\boxed{\text{Median Position} = \frac{1}{2}(n+1)}$$

$$\boxed{\text{Median Value} = x_{\frac{1}{2}(n+1)}}$$

Examples:

- Haircuts at barbershops in a community were observed to have the the following prices: (\$)

  20.00, 12.00, 7.50, 6.50, 8.00, 9.50, 8.00

  $$\begin{aligned} \text{Median Position} &= \frac{1}{2}(n+1) = \frac{1}{2}(7+1) = 4^{\text{th}} \\ \text{Median Value} &= x_4 \end{aligned}$$

  The **fourth observation** in **rank order** is the Median value.

  6.50, 7.50, 8.00, $\boxed{8.00}$, 9.50, 12.00, 20.00.

  The median is \$8.00.

- Suppose an eighth barbershop was accidentally omitted. It charges \$9.00 for a haircut.

  $$\begin{aligned} \text{Median Position} &= \frac{1}{2}(n+1) = \frac{1}{2}(8+1) = 4.5^{\text{th}} \\ \text{Median Value} &= x_{4.5} \end{aligned}$$

  Since there are 8 observations, the value halfway between the fourth and fifth observation in rank order will be the median.

  6.50, 7.50, 8.00, $\boxed{8.00,\ 9.00,}$ 9.50, 12.00, 20.00

  Interpolating halfway between x_4 and x_5, amounts to taking the average:

  $$\text{Median Value} = x_{4.5} = \frac{x_4 + x_5}{2} = \frac{\$8.00 + \$9.00}{2} = \$8.500000 = \$8.50\ .$$

  The median is \$8.50 .

*1.11.2 Ungrouped Frequency Distributions

The method of calculating the median is the same as for raw data except a $<Cf$ column is required to locate the position.

Example:
Fishery biologists want to test the efficacy of a fish ladder around a dam. They electronically tag a number of fish and count the number of fish that cross the fish ladder in a given day with the following results:

Catch (fish)	Days	$<Cf$
1	3	3
2	8	11
3	8	19
5	1	20
	$\sum f = 20$	

Remember that this table is a shorthand method of writing a list of 20 values in rank order, so

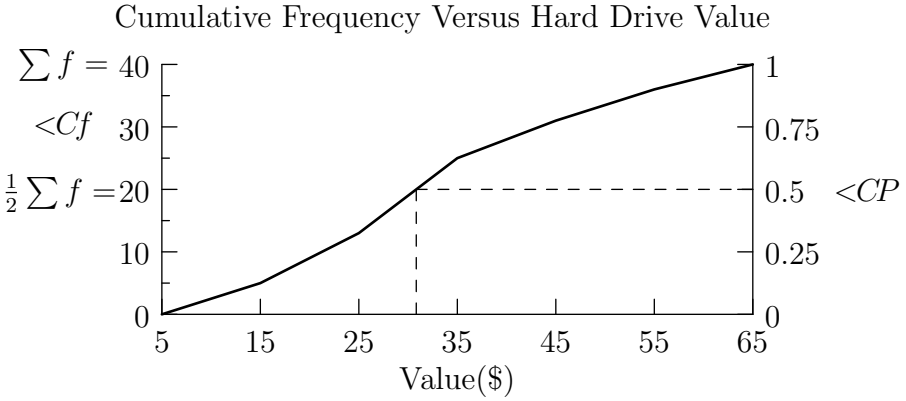
$$\boxed{\text{Median Position} = \frac{1}{2}\left(\sum f + 1\right)} = \frac{1}{2}(20+1) = 10.5^{\text{th}}$$

$$\boxed{\text{Median Value} = x_{\frac{1}{2}(\sum f+1)}} = x_{10.5} = 2.0 \text{ fish}$$

The median is halfway between the tenth and eleventh observations. From the $<Cf$ column, we see that x_1 to x_3 equal 1 fish, x_4 to x_{11} equal 2 fish, x_{12} to x_{19} equal 3 fish and x_{20} is 5 fish. So both desired observations x_{10} and x_{11} are 2 fish so the median is 2.0 fish. Had the values x_{10} and x_{11} been different, we would have averaged them as in our raw data example.

*1.11.3 Grouped Frequency Distributions

For a grouped frequency distribution the median is still that value in the middle of the ranked array of data the distribution represents, however now that value can only be approximated. If the cumulative frequency polygon (ogive) is plotted (Section *1.7.2) the approximate median can be found by locating the value of the variable corresponding to the 0.5 mark (50%) on the $<CP$ axis. (Equivalently, as can be seen on the ogive, this corresponds to the point $\frac{1}{2}\sum f$ on the $<Cf$ axis.) For example consider the hard drive example of Section 1.9.3:



Graphically we see that the median for the data is approximately \$31.00 . If a more precise scale were available the approximate median would be found to be \$30.83 .

To calculate the approximate median once can interpolate from the frequency distribution directly.

Example:
Find the median for the grouped frequency distribution introduced in Section 1.9.3.

Solution:
As with an ungrouped frequency distribution we still require the addition of the less than cumulative frequency column ($<Cf$):

Value of Used Computer Hard Drive		
Value ($)	f	$<Cf$
5.00 - 15.00	5	5
15.00 - 25.00	8	13
25.00 - 35.00	12	25
35.00 - 45.00	6	31
45.00 - 55.00	5	36
55.00 - 65.00	4	40
	$\sum f = 40$	

We then will use interpolation:

First, find the position of the median:

$$\boxed{\text{Median Position} = \frac{1}{2}\left(\sum f\right)} = \frac{1}{2}(40) = 20^{th}$$

* A common error is to find $\frac{1}{2}\left(\sum f + 1\right)$ as was done for ungrouped data. This is not done here because the horizontal line on the ogive curve meets the left axis at $\frac{1}{2}\sum f$.

	Value ($)	f	$<Cf$
	5.00 - 15.00	5	5
	15.00 - 25.00	8	13
Median Class $\longrightarrow$	25.00 - 35.00	12	25
	35.00 - 45.00	6	31
	45.00 - 55.00	5	36
	55.00 - 65.00	4	40
		$\sum f = 40$	

Second, find the value of the median. Go down the $<Cf$ column until the first value equal to or exceeding the location is found. In our case, this is a $<Cf$ value of 25. This means that the median (value) is located in the third class between \$25.00 and \$35.00. This is called the **median class**. Interpolate within this class:

Variable Value		Position
\$10 (Median − \$25 (	\$25	13) 7) 12
	Median	20
	\$35	25

Here the numbers on the sides of the parentheses are just the differences between the lower and higher numbers. (e.g. \$10=\$35-\$25). The difference involving the unknown median value can be solved by constructing the fractional equation:

$$\frac{(\text{Median Value}) - \$25}{\$10} = \frac{7}{12}$$

Solving for the median (by multiplying both sides by \$10 followed by adding \$25 to both sides) gives:

$$\text{Median Value} = \$25 + \frac{7}{12} \cdot (\$10) = \$30.83333333 = \$30.83$$

Equivalently, here is a formula that interpolates within the grouped data table once the position is found. Normally interpolating directly is easier than remembering the terms of the formula.

$$\boxed{\text{Median Value} = L_i + \frac{\left\{\frac{1}{2}\left(\sum f\right) - {<}Cf_{i-1}\right\}}{f_i} \cdot \Delta x}$$

Where:

$$\begin{aligned} i &= \text{number of the median class (here } 3^{\text{rd}}\text{)} \\ L_i &= \text{lower limit of the median class (here \$25)} \\ \Delta x &= \text{class width (here } \$35 - \$25 = \$10\text{)} \\ f_i &= \text{frequency of the median class (here 12)} \\ {<}Cf_{i-1} &= \text{cumulative frequency of the class prior to the median class (here 13)} \end{aligned}$$

Assignment: For each case study calculate the median. Remember F.S.A.R.U. for all calculations.

1.12 Measuring Dispersion in a Distribution

1.12.1 The Range as a Measure of Dispersion

Recall that variability is the basis for all statistical analysis. Data arrays have a tendency to cluster around some central value called an average. The usefulness of this average as a representative depends upon the strength of this tendency to cluster. If many of the data values are spread out a long distance from centre, then the average is a poor representative of the distribution. The tendency for values in the data array to be spread or dispersed from the average is measured in several ways.

Figure 1

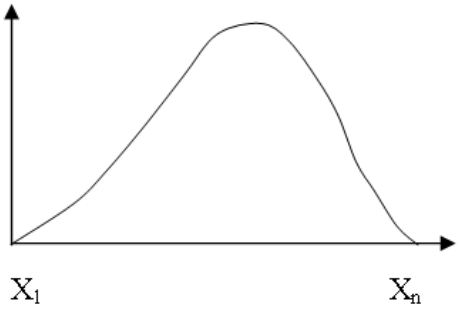
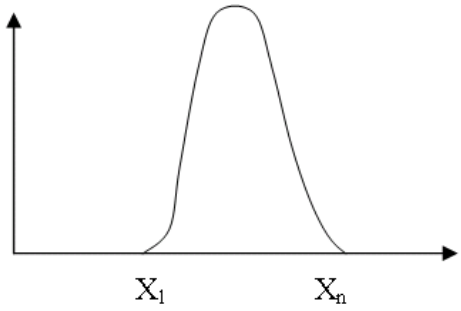


Figure 2



In the above diagrams note:

- There is more variation in distribution 1 than in 2.
- The range $\boxed{R = x_n - x_1}$ of distribution 1 is larger than in 2.

Figure 3

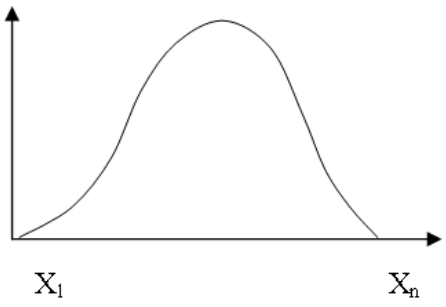
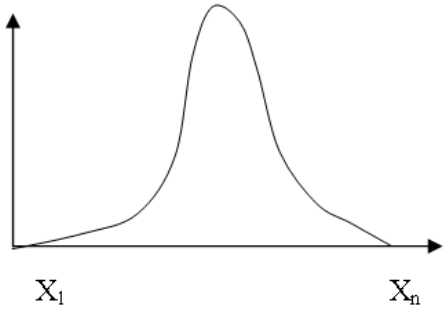


Figure 4



In the above diagrams note:

- There is more variation in distribution 3 than in 4.
- The ranges of the distributions are equal.

In conclusion, variation can be measured absolutely by comparing ranges or it can be measured by the amount of variation within the range. Often the latter is of greater value, and, as shown in the second set of figures the range then is a coarse and often misleading measure of dispersion because of its sensitivity to outliers and its insensitivity to the internal frequency distribution.

1.12.2 The Difference from the Mean

When measuring the amount of variation within a range, statisticians measure the dispersion from the centre. The arithmetic mean (μ or $\bar{x}$) is used as this central point. The mathematical properties of the mean must be taken into account when computing this measure.

As a first attempt for measuring the dispersion of a distribution one might consider calculate the *difference from the mean*, $x - \mu$, for each data value and taking the average of them ($\frac{\sum(x-\mu)}{N}$). Calculating this for the population data (kg) 20, 30, 35, 51 produces:

x(kg)	$x - \mu$ (kg)
20	-14
30	-4
35	1
51	17
$\sum x = 136$	$\sum(x - \mu) = 0$

Here, before tabulating the second column, we had to first calculate the mean:

$$\mu = \frac{\sum x}{N} = \frac{136 \text{ kg}}{4} = 34.000000 \text{ kg} = 34.0 \text{ kg} .$$

One finds that the sum of the difference $x - \mu$ is found to be zero. The reason this sum is zero is because the mean is the balance point of the data and one can prove[12] that it is always true that

$$\sum(x - \mu) = 0 .$$

The sum of the negative differences and the positive differences is zero. Because differences from the mean sum to zero, the average difference from the mean will always be zero no matter how close or how far the individual values are spread from centre.

For this reason, the average deviation is not calculated by averaging these differences.

Statisticians make the deviations from the mean positive in one of two ways:

1. Take the absolute value of the differences (that is take the *distance* from the mean) and compute the average of these deviations.
2. Square these differences and compute the average squared value.

1.12.3 The Average Deviation (a.d.)

A measure of dispersion called the **Average Deviation** is calculated by averaging the **absolute value** of the difference (the distance) from the mean.

$$\boxed{\text{a.d.} = \frac{\sum |x - \bar{x}|}{n}} \quad \text{or} \quad \boxed{\text{a.d.} = \frac{\sum |x - \mu|}{N}}$$

[12] Proof: $\sum(x - \mu) = \sum x - \sum \mu = \sum x - \mu \sum 1 = \sum x - \frac{\sum x}{N} \cdot N = \sum x - \sum x = 0$

The units on this measurement are the same as that of the variable. The only significance of the average deviation is its size. The smaller the value, the more closely are the numbers grouped around the mean within the range of observations. The average deviation is not used for statistical inference purposes so its use is limited to a descriptive measure.[13]

Example:
Calculate the average deviation for the previous data.

Solution:

x(kg)	$x-\mu$ (kg)	$\lvert x-\mu\rvert$ (kg)
20	-14	14
30	-4	4
35	1	1
51	17	17
$\sum x = 136$	*Do not sum	$\sum \lvert x-\mu\rvert = 36$

$$\text{a.d.} = \frac{\sum |x-\mu|}{N} = \frac{36 \text{ kg}}{4} = 9.000000 \text{ kg} = 9.0 \text{ kg}$$

Statistically this means that we should see the majority of the observations within 9.0 kg of centre (the mean). Between $\mu - \text{a.d.} = 34 - 9 = 25$ kg and $\mu + \text{a.d.} = 34 + 9 = 43$ kg lie two of the four or 50% of the data values.

1.12.4 The Population Variance (σ^2) and Standard Deviation (σ)

A measure of dispersion called the **Variance** is calculated by averaging the **squared deviations** from the mean. For a *population*:[14]

$$\boxed{\sigma^2 = \frac{\sum (x-\mu)^2}{N}}$$

The symbol for the variance for a population data array is the Greek letter lower case sigma, σ, with an exponent 2.

Example:
Calculate the variance for the previous population.

Solution:

x(kg)	$x-\mu$ (kg)	$(x-\mu)^2$ (kg^2)
20	-14	196
30	-4	16
35	1	1
51	17	289
$\sum x = 136$	*Do not sum	$\sum (x-\mu)^2 = 502$

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} = \frac{502 \ \left(\text{kg}^2\right)}{4} = 125.5 \ \left(\text{kg}^2\right)$$

[13] The reason the average deviation is not used much in statistics is that the presence of the absolute value | | in the definition makes it mathematically unpleasant. For one thing the derivative of the absolute value function $|x|$ at $x = 0$ is discontinuous.

[14] We will see that the variance and standard deviation for a sample have different formulae than the population formula.

Statistically, this is a bit difficult to interpret at this point, but think about this value as a measure of the variability of the numbers within their range.

As the example shows, the units on the variance will be in units of the **variable squared**. This, compared to the average deviation, is a serious drawback since one cannot say that the majority of measurements are some fraction of variance away from the mean. The fact that the variance has square units of the variable means that one cannot create such an interval directly from the variance. However, one still may compare the variance of two comparable sets of data to see which one has greater dispersion.

To place the units on this measure into the same value as the variable we take the square root. This quantity is called the **standard deviation**. Its symbol for a population is lower case sigma without the square:

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

Now, as with the average deviation, one expects the majority of measured values to lie within one standard deviation from the mean.

Example:
Calculate the standard deviation of the above data.

Solution:

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} = \sqrt{\frac{502 \ \left(\text{kg}^2\right)}{4}} = 11.2026 \text{ kg} = 11.2 \text{ kg}$$

As with the average deviation this also means that statistically we should see the majority of the observations within 11.2 kg of centre (the mean). Between $\mu - \sigma = 34.0 - 11.2 = 22.8$ kg and $\mu + \sigma = 34.0 + 11.2 = 45.2$ kg lie two of the four or 50% of the data values.

Significance of the Two Measures of Dispersion

Note that the size of the standard deviation found in the example (11.2 kg) is roughly the same as that of the average deviation (9.0 kg) in the sense that the interval around the mean captured the same fraction (50%) of the data. This, coupled with its preferable mathematical properties, means the standard deviation is the preferred measure of dispersion. We will only calculate the average deviation for data presented in raw format.

Because of their mathematical properties, the mean and standard deviation will be the most important measures of the centre and dispersion of a distribution respectively for the remainder of the course. These two measures are found in theoretical distribution curves from which statistical inferences will be drawn.

Statistical Keys on the Calculator

Recall the procedure for summing data using the statistical keys on the calculator in Section 1.4. To find the standard deviation, the data is keyed into the calculator in the same way as was done there. Locate the key on your calculator that finds the population standard deviation. The population standard deviation sometimes has an n in the symbol, such as σ_n, which reflects the N that appears in the denominator of the formula. Other calculators have a population mode which must be set before

selecting the standard deviation to ensure the proper formula is being used. There is no key that finds average deviation.

After entering all the data values, the standard deviation for a population can be found by pushing this key. Try it for the previous example. Once you have the standard deviation on your display, the variance can be found by pushing the key that squares numbers.

Assignment: For each case study that is raw data calculate the average deviation. For raw data which is also population data calculate the standard deviation and variance. Remember F.S.A.R.U. Check the standard deviation results on your calculator.

1.13 Computing the Standard Deviation

1.13.1 The Computing Formula for Pop. Variance and Standard Deviation

It can be proven, by expanding $(x-\mu)^2$ algebraically,[15] that

$$\sum (x-\mu)^2 = \sum x^2 - N\mu^2$$

This leads to a mathematically equivalent method for computing the variance and standard deviation that has certain advantages over using the formula based on the definition. The new formula is called the **computing formula**. (The previous formula will be referred to as the **definitional formula**)

Replace $\sum(x-\mu)^2$ by $\sum x^2 - N\mu^2$ in the variance formula $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$:

$$\sigma^2 = \frac{\sum x^2 - N\mu^2}{N} = \frac{\sum x^2}{N} - \mu^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$

The computing formula for population variance for raw data is therefore:

$$\boxed{\sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

Taking the square root gives the computing formula for population standard deviation:

$$\boxed{\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}}$$

Example:
Using the example from Section 1.12, compute the standard deviation by the computing formula. Show that the answer is identical to that obtained by the definitional formula.

Solution:

x(kg)	x^2 (kg^2)
20	400
30	900
35	1225
51	2601
$\sum x = 136$	$\sum x^2 = 5126$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2} = \sqrt{\frac{5126\ \left(\text{kg}^2\right)}{4} - \left(\frac{136\ \text{kg}}{4}\right)^2} = \sqrt{1281.5\ \left(\text{kg}^2\right) - (34\ \text{kg})^2} \\ &= \sqrt{125.5\ \left(\text{kg}^2\right)} = 11.2\ \text{kg}\ ,\end{aligned}$$

the same value as was obtained by the formula based on the definition.

[15] One has $\sum (x-\mu)^2 = \sum \left(x^2 - 2\mu x + \mu^2\right) = \sum x^2 - 2\mu \sum x + \mu^2 \sum 1 = \sum x^2 - 2\mu(N\mu) + \mu^2(N) = \sum x^2 - N\mu^2$, where we used that $\mu = \frac{\sum x}{N}$ is constant and can be pulled out of the two sums.

*Note that the summation $\sum x^2$ appearing in the computing formulae requires calculation of the data values squared followed by their summation. It is **not** equal to $(\sum x)^2$, i.e. the summation of the variable values followed by squaring.

In statistical analysis, you will find both the definitional and computational formulae used to compute σ. The definitional formula is often used if the mean works out to be an integer value producing a difference that is easily squared. The computational formula lessens the number of calculations and allows for easier adjustments if more data points are added to the data array. You should be able to use both formulae.

1.13.2 The Standard Deviation of Population Frequency Distributions

The most effective way of calculating the standard deviation and variance of a frequency table is by the computing formulae. The various quantities that appear in the raw data formulae and their equivalents in a frequency distribution are given in the following table:

Quantity	Raw Data	Frequency Distribution
number of observations	N	$\sum f$
sum of observations	$\sum x$	$\sum xf$
sum of squares of observations	$\sum x^2$	$\sum x^2 f$

If the raw data symbols are replaced with their frequency distribution counterparts in the computing formula, it takes the following form for the **population variance**:

$$\boxed{\sigma^2 = \frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2}$$

The computing formula for **population standard deviation** for a frequency distribution is similarly:

$$\boxed{\sigma = \sqrt{\frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2}}$$

If the data are ungrouped, the x values are the values of the variable listed in the table. If the data are grouped data the x values are the **class midpoints**.

Example:
Suppose all 100 households of a community were surveyed. It was found the number of rooms per household was distributed as follows. Calculate the standard deviation and variance.

Solution:

x(rooms)	f	xf(rooms)	$x^2 f$ (rooms^2)
1	2	2	2
2	15	30	60
3	20	60	180
4	40	160	640
5	20	100	500
8	2	16	128
12	1	12	144
	$\sum f = 100$	$\sum xf = 380$	$\sum x^2 f = 1654$

The standard deviation is:

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2} = \sqrt{\frac{1654 \text{ (rooms}^2\text{)}}{100} - \left(\frac{380 \text{ rooms}}{100}\right)^2} \\ &= \sqrt{2.1 \text{ (rooms}^2\text{)}} = 1.44913 \text{ rooms} \\ &= 1.4 \text{ rooms}\end{aligned}$$

The variance is just the square of this value (the last term under the square root):

$$\sigma^2 = 2.1 \text{ (rooms}^2\text{)}$$

*Note that the $x^2 f$ column is **not** the square of the preceding xf column since the frequency is not squared (i.e. $x^2 f \neq (xf)^2 = x^2 f^2$). Some students may wish to add an x^2 column to the table to facilitate the calculation of $x^2 f$.

Since the population mean is $\mu = \frac{\sum xf}{\sum f} = 3.8$ rooms, it may be observed that the majority of observations (80 of the 100) really do fall between $\mu - \sigma = 2.4$ rooms and $\mu + \sigma = 5.2$ rooms.

It is possible to calculate the standard deviation and variance by the formula based on the definition in frequency tables but because of the number of arithmetic operations required we will always use the computational formula for frequency distributions.

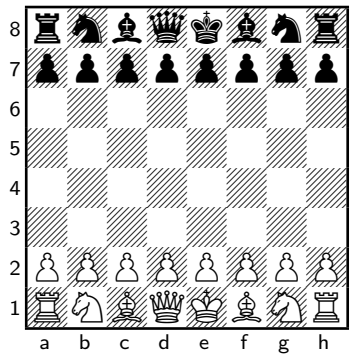
Statistical Keys on the Calculator

As seen in Section 1.4, the summation of data values can be found by using the $\sum x$ function on the calculator after keying in data. Statistical calculators similarly have a key for obtaining the sum of the squares of the data, $\sum x^2$, which can be used to check tabular results. Remember that frequency distribution data may be entered on a calculator as discussed in Section 1.4. In that case the n, $\sum x$, and $\sum x^2$ calculator keys will generate the results for $\sum f$, $\sum xf$, and $\sum x^2 f$ respectively.

Assignment:

1. For each case study that is population data (including recalculating the result for raw data) calculate the standard deviation and variance using the computational formulae. Remember F.S.A.R.U. Check your results using the calculator.

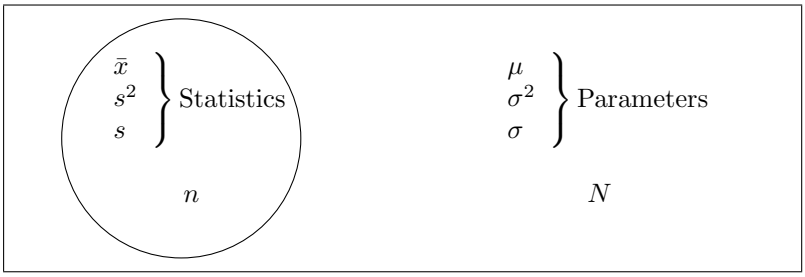
2. The following diagram shows the setup of a chess board. Because of the ways the pieces move, all pieces are not created equal! If Pawns (♙) are worth 1 point, then Bishops (♗) and Knights (♘) are worth roughly 3 points, Rooks (♖) are 5 points, and Queens (♕) are worth 9 points. The Kings (♔) have infinite value because if you lose your king you lose the game!



Create an ungrouped frequency distribution for the **value** of the chess pieces (excluding the kings) at the start of the game. Calculate its mean, standard deviation and variance. (Use the computational formula.) Sketch the histogram and see that your values make sense. Check your work on your calculator.

1.14 Sample Standard Deviation (s) and Variance (s^2)

The purpose of sampling is to make inferences about a population. Statistical descriptors of a population are called **parameters**. The calculated values found using sample data for the purpose of drawing inferences about population parameters are called **statistics**.



Because population data is usually inaccessible, the population values exist in theory but are not available in practice. Their values must be inferred from sample information. Letters of the Greek alphabet are reserved for parameters in statistical analysis to reflect the fact that they are theoretical values.

1.14.1 Sample Standard Deviation by the Definitional Formula

When estimating a population mean, μ, the best available estimate is the sample mean, $\bar{x}$. As may be shown, this is because the average value of all possible sample means is μ. Recall that there is no difference in the procedure for computing the sample statistic, $\bar{x}$, or the population parameter, μ. This is not the case with the **sample variance**, s^2. If the sample variance is calculated by the same procedure as the population variance, the sample variance is not the best estimator of the population variance. $\sum (x-\bar{x})^2$ tends to underestimate $\sum (x-\mu)^2$, on the average, in the variance formula. This places the numerator in the variance formula too low. To correct this, the denominator is reduced by 1. An **unbiased**[16] estimator of σ^2 is s^2 calculated by the **definitional formula for sample variance**:

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

The **definitional formula for sample standard deviation** is its square root:

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

As with populations, we will only use the sample definitional formula with raw data in this course.

Example:
Calculate the standard deviation and variance of the stellar distances of the Andromeda Nebula in the example from Section 1.10.1. Use the definitional formula.

Solution:
Continuing with our previous table we have:

[16] We say an estimator of a parameter is unbiased if the mean of its distribution equals the value of the parameter. So if we took all possible samples, calculated s^2 for each one, and took their average it would be σ^2. Sampling distributions will be considered in Unit 4 .

Distance (Mly)	$(x-\bar{x})$ (Mly)	$(x-\bar{x})^2$ (Mly^2)
2.4	-0.12	0.0144
2.4	-0.12	0.0144
2.5	-0.02	0.0004
2.6	0.08	0.0064
2.7	0.18	0.0324
$\sum x = 12.6$		$\sum (x-\bar{x})^2 = 0.068$

Here we required our sample mean from before:

$$\bar{x} = \frac{\sum x}{n} = \frac{12.6 \text{ Mly}}{5} = 2.52 \text{ Mly}$$

The sample variance is:

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{0.068 \text{ (Mly}^2)}{5-1} = 0.017 \text{ (Mly}^2)$$

The sample standard deviation is its square root:

$$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}} = \sqrt{\frac{0.068 \text{ (Mly}^2)}{5-1}} = 0.130384 \text{ Mly} = 0.13 \text{ Mly}$$

Check that the standard deviation is reasonable with respect to the data. Does a standard deviation of $s = 0.13$ Mly in the distance to the stars tell you anything useful physically about the Andromeda Nebula?

1.14.2 Sample Standard Deviation By the Computing Formula

Similar to the discussion in Section 1.13.1, a mathematical equivalent to $\sum (x-\bar{x})^2$ is $\sum x^2 - n\bar{x}^2$. If this is substituted into the definitional formula, the **computing formula for sample (raw) data standard deviation** results:

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

The computing formula for the sample variance s^2 is just the expression under the root sign.

Example:
Compute the standard deviation and variance for the previous sample using the computational formula.

Solution:

Distance (Mly)	x^2 (Mly^2)
2.4	5.76
2.4	5.76
2.5	6.25
2.6	6.76
2.7	7.29
$\sum x = 12.6$	$\sum x^2 = 31.82$

The sample standard deviation is:

$$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}} = \sqrt{\frac{31.82\ (\text{Mly}^2) - \frac{(12.6\ \text{Mly})^2}{5}}{5-1}} = \sqrt{0.017\ (\text{Mly}^2)} = 0.130384\ \text{Mly} = 0.13\ \text{Mly}$$

The sample variance is just the square of this value:

$$s^2 = 0.017\ (\text{Mly}^2)$$

If the sample data is from a **frequency distribution** the sums must be adjusted to reflect the frequency of occurrence of the listed data values as was done in Section 1.13.2. The **computing formula** for the **standard deviation of a frequency distribution of sample data** is:

$$\boxed{s = \sqrt{\frac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}}}$$

The formula for the sample variance s^2 is the same without the square root.

Example:
A random sample of household appliances showed the following power consumption while nominally switched "off". Calculate the standard deviation and variance of these *phantom* electrical power losses.

Power Consumption (Watts)	# of appliances	x(Watts)	xf(Watts)	$x^2 f\ (\text{Watts}^2)$
0 - 2	16	1	16	16
2 - 4	4	3	12	36
4 - 6	6	5	30	150
6 - 8	2	7	14	98
8 - 10	2	9	18	162
	$\sum f = 30$		$\sum xf = 90$	$\sum x^2 f = 462$

The sample standard deviation is

$$\begin{aligned} s &= \sqrt{\frac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}} = \sqrt{\frac{462\ \left(\text{Watts}^2\right) - \frac{(90\ \text{Watts})^2}{30}}{30-1}} = \sqrt{\frac{192\ \left(\text{Watts}^2\right)}{29}} \\ &= \sqrt{6.62068\ \left(\text{Watts}^2\right)} = 2.573\ \text{Watts} \\ &= 2.6\ \text{Watts} \end{aligned}$$

The sample variance is just the square of this value:

$$s^2 = 6.62068\ \left(\text{Watts}^2\right) = 6.6\ \left(\text{Watts}^2\right)$$

If one mistakenly thought this were population data, one would have found a standard deviation of $\sigma = 2.5$ Watts versus the true value of $s = 2.6$ Watts. This reflects the fact that for a large sample (here 30) the difference between the two values diminishes.

Statistical Keys on the Calculator

Locate the two keys that calculate standard deviation on your calculator. On some calculators the keys are distinguished by σ (or σ_x) for population data and s (or s_x) for sample data. On other calculators since the population standard deviation uses the denominator N in its procedure while the sample standard deviation uses the denominator $n-1$ in its procedure the keys may be distinguished by the subscript and look like σ_n and σ_{n-1} respectively. Other calculators will have the user select a population or sample mode before data entry so the correct formula is applied when standard deviation is selected. Finally other calculators may have one select the procedure from a menu of options. Once you have figured out how your calculator does the calculation, confirm the text examples done with it.

Assignment: For each case study that is raw sample data, calculate the standard deviation and variance using the definitional formula. For each case study that is sample data (including recalculating the result for raw data) calculate the standard deviation and variance using the computational formulae. Remember F.S.A.R.U. Check your results using the calculator.

1.15 Uses of the Standard Deviation

The standard deviation is a key measure of dispersion in a distribution because it is useful in describing and analyzing many different properties of a data array.

1.15.1 Chebyshev's Theorem

As we have seen, the arithmetic mean is a measure of the centre of the data so data lies on either side of the mean. How close is the data to the mean? Chebyshev's theorem gives us a measure of the amount of data near the mean measured in units of the standard deviation.

Chebyshev's Theorem: *The fraction of any data set that lies within k standard deviations of the mean is at least* $\boxed{1 - \frac{1}{k^2}}$ *where k is any number greater than* 1.

According to the theorem within $k = \sqrt{2} \approx 1.41$ standard deviations of the mean there is at least

$$1 - \frac{1}{(\sqrt{2})^2} = 1 - \frac{1}{2} = 0.50 = 50\%$$

of all the data elements. Within $k = 2$ standard deviations of the mean there is at least $1 - \frac{1}{2^2} = \frac{4}{4} - \frac{1}{4} = \frac{3}{4} = 0.75 = 75\%$ of all the data elements. Similarly within $k = 3$, $k = 4$, and $k = 10$ standard deviations one must have at least $\frac{8}{9} = 89\%$, $\frac{15}{16} = 94\%$, and $\frac{99}{100} = 99\%$ of all the data values respectively.

Example:
Astronomers estimate that the population of stars in our Milky Way galaxy have distances (x) from galactic centre (in thousands of light-years) having arithmetic mean $\bar{x} = 13$ kly and standard deviation of $s = 9$ kly. According to Chebyshev's Theorem using $k = 1.41$ at least 50% of the stars must have distances to the centre of the galaxy between

$$\bar{x} - ks = 13 - (1.41)(9) = 0.3 \text{ kly}$$

and

$$\bar{x} + ks = 13 + (1.41)(9) = 25.7 \text{ kly} .$$

For $k = 2$ one has at least 75% of stellar distances lying between $\bar{x} - ks = 13 - 2(9) = -5$ kly and $\bar{x} + ks = 13 + 2(9) = 31$ kly. (Since our variable x is a distance, this can only take on positive values so this means at least 75% of galaxy stars are less than 31 kly from the galactic centre.)

1.15.2 Standard Score (z)

The **standard score** or z **value** measures the distance that an observation is from the mean in units of standard deviations.

$$\boxed{z = \frac{x - \mu}{\sigma}} \leftarrow \text{For a population, or} \quad \boxed{z = \frac{x - \bar{x}}{s}} \leftarrow \text{For a sample}$$

The **sign** of the standard score indicates whether the data element lies above ($z > 0$) or below ($z < 0$) the mean.

Chebyshev's Theorem allows us to conclude, based on the **magnitude** of z how far the observation lies away from the mean:[17]

- $|z| \approx 0$: x is **approximately equal** to the mean
- $|z| \approx 1$: x is **slightly** removed from the mean
- $|z| \approx 2$: x is **moderately** removed from the mean
- $|z| \gtrsim 3$: x is **extremely** removed from the mean

Example:
Continuing our previous example, if stars are distributed with distances from the centre of the galaxy with arithmetic mean $\bar{x} = 13$ kly and standard deviation of $s = 9$ kly, find the standard score z of:

1. Our sun which is at a distance of $x = 27$ kly from the centre of the Milky Way.

   $$z = \frac{x - \bar{x}}{s} = \frac{27 \text{ kly} - 13 \text{ kly}}{9 \text{ kly}} = 1.56$$

   Interpretation: Our sun is +1.56 standard deviations from the mean, so slightly to moderately above the average distance.

2. A star in the central bulge of our galaxy at a distance of $x = 4$ kly from galactic centre.

   $$z = \frac{x - \bar{x}}{s} = \frac{4 \text{ kly} - 13 \text{ kly}}{9 \text{ kly}} = -1.00$$

   Interpretation: This bulge star is -1.00 standard deviations from the mean, so slightly below the average distance.

3. A typical star in the Andromeda Nebula. From Section 1.10.1 such a star has an approximate distance from our sun of 2.52 Mly = 2520 kly. Given our sun is $x = 27$ kly from galactic centre, geometrical considerations show that the closest an Andromeda Nebula star could be to galactic centre is $x = 2520 - 27 = 2493$ kly. The corresponding z value for such a distance is:

   $$z = \frac{x - \bar{x}}{s} = \frac{2493 \text{ kly} - 13 \text{ kly}}{9 \text{ kly}} = 275.56$$

   Interpretation: Stars in the Andromeda Nebula are more than +275 standard deviations from the mean, so extremely, extremely, extremely above the average distance from galactic centre. According to Chebyshev's theorem one has (with $k = 275$) at least $1 - \frac{1}{275^2} = 99.9987\%$ of all Milky Way stars lying within 275 standard deviations of the mean, so these Andromeda stars are highly improbable! What's going on? (Hint: After Edwin Hubble first determined accurate distances to the Cepheid Variable stars of the Andromeda Nebula in 1925, we started referring to it as the Andromeda *Galaxy*.)

[17] The terms used here (slight, moderate, extreme) are subjective in nature, but will be used in this course. Quoting the actual z value is unambiguous.

Besides giving a quick method of interpreting where a data value is with respect to the mean and how typical it is, the z-score is also useful for comparing observations from two different populations.[18]

Example:
The stars in the Andromeda Galaxy have a distance from **their** galactic centre with arithmetic mean $\bar{x} = 17$ kly and standard deviation $s = 12$ kly. If a star in the disk of Andromeda is measured to be at a distance of $x = 30$ kly from the centre of its galaxy, which is further from the centre of its own galaxy that star or our sun?

Solution:
Calculating the z-score of the Andromeda star gives:

$$z = \frac{x - \bar{x}}{s} = \frac{30 \text{ kly} - 17 \text{ kly}}{12 \text{ kly}} = 1.08$$

Interpretation: In absolute terms the Andromeda star is further out than our sun since $30 \text{ kly} > 27 \text{ kly}$. However in relative terms (with respect to the rest of its galaxy), the Andromeda star is closer to its centre since it is only 1.08 standard deviations above the mean, compared to 1.56 standard deviations above the mean for our sun.

Assignment: For each case study answer the standard score or Chebyshev Theorem question found there.

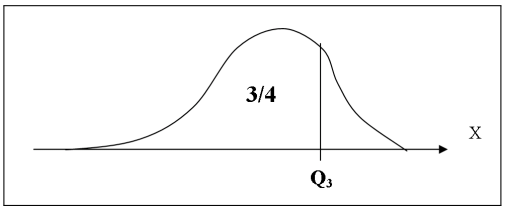
[18] Due to its definition, z will always be dimensionless so one can even compare two variables values of different dimension, such as height (in cm) and weight (in kg), by finding their corresponding standard scores.

1.16 Fractional Measures of Position

1.16.1 Fractiles

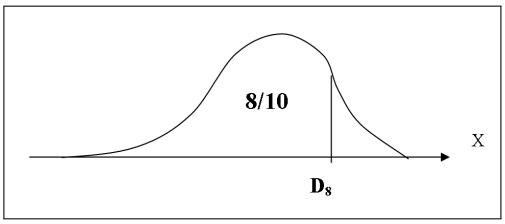
Fractiles are positional measures within a data set. Just as the median measures the halfway point, there are other fractional positions that can be computed. The fractions in common use are quarters, tenths and hundredths. Quartiles are quarter divisions, deciles are tenths divisions and percentiles are hundredth divisions. A sample of these fractiles is shown below.

Q_3 represents the third quartile. It is that value of the variable such that $\frac{3}{4}$ of the distribution falls below this value if the data array is ranked.



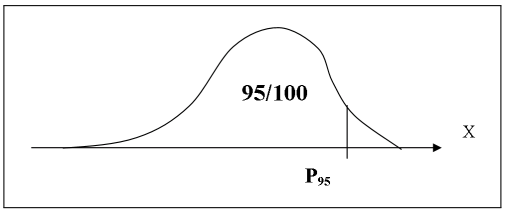
The Third Quartile

D_8 represents the eighth decile. It is that value of the variable such that $\frac{8}{10}$ of the distribution falls below this value if the data array is ranked.



The Eighth Decile

P_{95} represents the ninety-fifth percentile. It is that value of the variable such that $\frac{95}{100}$ of the distribution falls below this value if the data array is ranked.



The Ninety-fifth Percentile

What do Q_2, D_5, and P_{50} equal?

Like the z-score, a percentile can be used to measure a position or rank within a data set. For instance if on a standardized test your score equalled P_{95}, that is your score was in the 95^{th} percentile, then 95% of the other scores were below your score.

1.16.2 Calculating Fractiles for Raw Data

Fractile values are calculated in an **identical manner** to the way in which the **median** was calculated.[19] (See Section 1.11.1.) The only difference is that the fraction used to find the position is not $\frac{1}{2}$ but depends on the fraction in the fractile. On the formula sheet, we will use the formulae for the median calculation but we will change the fraction to correspond to the fractile. For a fractile F_i the denominator will be 4, 10, or 100 depending on whether F is Q, D, or P while the numerator becomes i.

[19] Note that there are minor variations in how fractiles are defined and hence how they are calculated. For this class they should be calculated as shown here. When comparing with other sources the answers may not be exactly the same but they should be close.

Example:
A sample of young children were observed to have the following heights (cm)

$$56,\ 70,\ 93,\ 65,\ 62,\ 71,\ 56,\ 63,\ 67,\ 70$$

Find the **10^{th}** and **90^{th} percentiles**.

Solution:
In rank order we have:

$$56,\ 56,\ 62,\ 63,\ 65,\ 67,\ 70,\ 70,\ 71,\ 93$$

$$\text{Position } P_{10} = \frac{\mathbf{10}}{\mathbf{100}}(n+1) = \frac{10}{100}(10+1) = 1.1^{\text{st}}$$

Since $x_1 = 56$ cm and $x_2 = 56$ cm we have for the value of P_{10}:

$$P_{10} = x_{1.1} = 56.0 \text{ cm}$$

In this last case no interpolation was required because $x_1 = x_2$. Had they been different, in the median case the only possible position in between would have been 1.5 and we would have just averaged the values. New in the case for fractiles for raw data is that interpolation is often required as the calculation of P_{90} shows:

$$\text{Position } P_{90} = \frac{\mathbf{90}}{\mathbf{100}}(n+1) = \frac{90}{100}(10+1) = 9.9^{\text{th}}$$

Since $x_9 = 71$ cm and $x_{10} = 93$ cm we must interpolate to find $P_{90} = x_{9.9}$:

Variable Value		Position
22 cm	$P_{90} - 71$ cm	9
	P_{90}	9.9
	93 cm	10

Here the numbers beside the parentheses are just the differences between the higher and lower number. (e.g. 22 cm = 93 cm − 71 cm). Interpolating requires creating the fractional equation of the differences:

$$\frac{P_{90} - 71 \text{ cm}}{22 \text{ cm}} = \frac{0.9}{1}$$

Solving for the ninetieth percentile value (by multiplying both sides by 22 and then adding 71 to each side) gives:

$$P_{90} = x_{9.9} = 71 \text{ cm} + \frac{0.9}{1} \cdot (22 \text{ cm}) = 90.80000 \text{ cm} = 90.8 \text{ cm}\ .$$

The interpolation turns out to be easy in the raw data case because of the 1 which (always) shows up in the denominator on the right hand side. This allows a shortcut for calculation. After working out the position as usual (here 9.9) the last equation shows:

$$\boxed{P_{90} = x_{9.9} = x_9 + 0.9\,(x_{10} - x_9)} = 71 \text{ cm} + 0.9\,(22 \text{ cm}) = 90.80000 \text{ cm} = 90.8 \text{ cm}$$

In words, to find $x_{9.9}$ we start at x_9 and go 0.9 of the distance (difference) to x_{10}. Note this shortcut will not work in the grouped frequency case shown in the next section.

*1.16.3 Calculating Fractiles for Frequency Distributions

For frequency distributions one calculates fractiles in a manner similar to the median (Sections *1.11.2 and *1.11.3). One difference, however, is that now in the ungrouped frequency case one may also need to interpolate, as occurred in the raw data case.

Ungrouped Frequency Distributions

Example:
Over a two month period the number of cancelled flights in a given day were recorded for an airport with the results at left. Find the **eighth decile** for the data.

Cancellations (flights)	Days	$<Cf$
0	29	29
1	19	48
3	8	56
4	3	59
5	1	60
	$\sum f = 60$	

First we find the position by modifying the median formula for ungrouped data:

$$\text{Position } D_8 = \frac{\mathbf{8}}{\mathbf{10}}\left(\sum f + 1\right) = \frac{8}{10}(60+1) = 48.8^{\text{th}}$$

From the cumulative frequency column we see that $x_{48} = 1$ flight and $x_{49} = 3$ flights, so we interpolate.

Variable Value		Position
2 flights		
	$D_8 - 1$ flight	
	1 flight	48
	D_8	48.8
	3 flights	49

This gives the fractional equation:

$$\frac{D_8 - 1 \text{ flight}}{2 \text{ flights}} = \frac{0.8}{1}$$

Solve for the eighth decile value:

$$D_8 = x_{48.8} = 1 \text{ flight} + \frac{0.8}{1} \cdot (2 \text{ flights}) = 2.60000 \text{ flights} = 2.6 \text{ flights} .$$

Note, as the last equation shows, the shortcut method illustrated in the previous raw data example can be applied for ungrouped frequency distributions as well.

Grouped Frequency Distributions

Examples:
A survey of the price of a certain size of refrigerator was made with the results at left below. Find the price at the **third quartile** and **first quartile**.

Price ($)	f	$<Cf$
500 - 600	6	6
600 - 700	12	18
700 - 800	18	36
800 - 900	10	46
900 - 1000	4	50
	$\sum f = 50$	

First find the position:

$$\text{Position } Q_3 = \frac{\mathbf{3}}{\mathbf{4}}\left(\sum f\right) = \frac{3}{4}(50) = 37.5^{\text{th}}$$

Examining the cumulative frequency column indicates that Q_3 lies in the \$800-\$900 class, so we must interpolate.

Variable Value	Position
$\$100 \left(Q_3 - \$800 \left(\begin{array}{c} \$800 \\ Q_3 \\ \$900 \end{array} \right. \right.$	$\left. \left. \begin{array}{c} 36 \\ 37.5 \\ 46 \end{array} \right) 1.5 \right) 10$

This gives the fractional equation:

$$\frac{Q_3 - \$800}{\$100} = \frac{1.5}{10}$$

Solve for the third quartile value:

$$Q_3 = \$800 + \frac{1.5}{10} \cdot (\$100) = \$815.0000 = \$815.00$$

To find the first quartile, Q_1, let's use the formula approach.

First find the position:

$$\text{Position } Q_1 = \mathbf{\frac{1}{4}} \left(\sum f \right) = \frac{1}{4} (50) = 12.5^{th}$$

Examining the cumulative frequency column indicates that Q_1 lies in the \$600-\$700 class.

Using the median formula for grouped frequency distributions suitably modified for the first quartile gives:

$$Q_1 = L_i + \frac{\left\{ \mathbf{\frac{1}{4}} (\sum f) - <Cf_{i-1} \right\}}{f_i} \cdot \Delta x = \$600 + \left\{ \frac{\frac{1}{4}(50) - 6}{12} \right\} \cdot \$100 = \$654.166666 = \$654.17$$

1.16.4 The Five-Number Summary

The **five-number summary** lists the smallest data value, the first quartile, the median, the third quartile, and the maximum data value in that order. In our symbols this list is:

$$x_1,\ Q_1,\ \text{Median},\ Q_3,\ x_n$$

By construction there is one quarter of the data between each successive pair of values.

Example:
Calculate the five-number summary for the height of children data of Section 1.16.2.

Solution:
The data array in rank order was found to be:

$$56,\ 56,\ 62,\ 63,\ 65,\ 67,\ 70,\ 70,\ 71,\ 93$$

The positions of Q_1 , the median, and Q_3 respectively are:

$$\begin{aligned} \frac{1}{4}(n+1) &= \frac{1}{4}(10+1) = 2.75^{\text{th}} \\ \frac{1}{2}(n+1) &= \frac{1}{2}(10+1) = 5.5^{\text{th}} \\ \frac{3}{4}(n+1) &= \frac{3}{4}(10+1) = 8.25^{\text{th}} \end{aligned}$$

The five numbers are then

$$\begin{aligned} x_1 &= 56 \text{ cm} \\ Q_1 &= x_{2.75} = 56 + .75(62 - 56) = 60.5000 = 60.5 \text{ cm} \\ \text{Median} &= x_{5.5} = (65 + 67)/2 = 66.0000 = 66.0 \text{ cm} \\ Q_3 &= x_{8.25} = 70 + .25(71 - 70) = 70.2500 = 70.3 \text{ cm} \\ x_n &= x_{10} = 93 \text{ cm} \end{aligned}$$

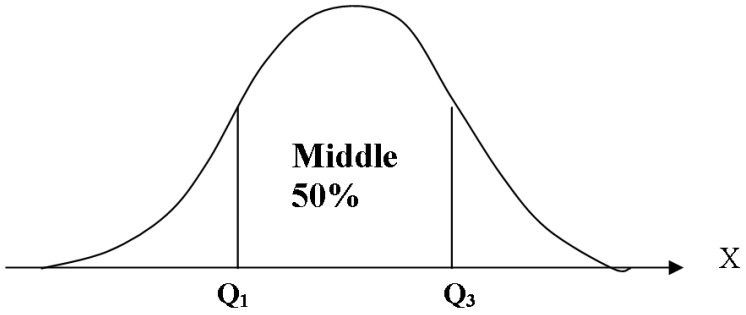
Thus the five-number summary for this data is:

$$56,\ 60.5,\ 66.0,\ 70.3,\ 93 \text{ (cm)}$$

1.16.5 Using Fractiles to Measure Dispersion

There are various types of ranges that are calculated on the basis of these fractional measures. The effect of these calculations is to trap a given proportion of the array about centre and to cut off the values in the extremities of the distribution.

Interquartile Range (IQR)



The **interquartile range** captures the middle 50% of the data and is defined by

$$\boxed{IQR = Q_3 - Q_1}\ .$$

Examples:
Find the **interquartile ranges** for the example of Section 1.16.2 and the grouped frequency example of Section *1.16.3.

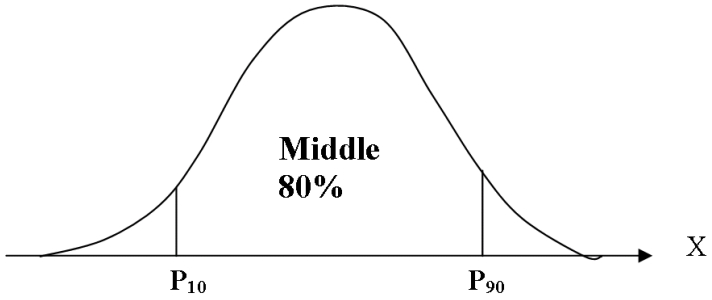
Solutions:
Using the results calculated in Section 1.16.4, the interquartile range for the former is:

$$IQR = Q_3 - Q_1 = 70.3 - 60.5 = 9.8 \text{ cm}$$

For the grouped frequency example, using the results of Section *1.16.3, the interquartile range is:

$$IQR = Q_3 - Q_1 = \$815.00 - \$654.17 = \$160.83$$

10-90 Percentile Range (10-90 PR) or Interdecile Range (IDR)



The **10-90 percentile range** is defined by

$$\boxed{\text{10-90 } PR = P_{90} - P_{10}} \ .$$

This is also sometimes called the **interdecile range** from D_1 to D_9 since $P_{10} = D_1$ and $P_{90} = D_9$.

$$\boxed{IDR = D_9 - D_1} \ .$$

These ranges capture the middle 80% of the data.

Example:
Calculate the **10-90 percentile range** for the example in Section 1.16.2 .

Solution:
Using the results of that example, the 10-90 percentile range is

$$\text{10-90 } PR = P_{90} - P_{10} = 90.8 \text{ cm} - 56.0 \text{ cm} = 34.8 \text{ cm} \ .$$

The interdecile range, IDR, has the same value.

Note:

- If an ogive graph is available, all fractiles may be easily read from it (Section *1.7.2) and these ranges estimated.
- The interquartile and 10-90 percentile ranges both give a distance measuring how spread out the data is (dispersion) in a way similar to how we saw the standard deviation could be used.
- We could have defined an interpercentile range by $P_{99} - P_1$. Why might that not be so useful?

1.16.6 Outliers and Box Plots

Outliers are data values that differ markedly from the majority of values in the data set. By definition the middle 50% of the data lie between Q_1 and Q_3. The distance between these values is the interquartile range, IQR. A quantitative and unbiased way of identifying an outlier is to create a **lower fence** that is 1.5 times the interquartile range below Q_1 and an **upper fence** that is 1.5 times the interquartile range above Q_3. In symbols:

$$\boxed{\text{Lower Fence} = Q_1 - 1.5(IQR)}$$

$$\boxed{\text{Upper Fence} = Q_3 + 1.5(IQR)}$$

A data value below the lower fence or above the upper fence is an outlier.

Example:
Identify any outliers in the height data set example of Section 1.16.2.

Solution:
The interquartile range was found to be $IQR = 9.8$ cm in Section 1.16.5. The fences are then:

$$\text{Lower Fence} = Q_1 - 1.5(IQR) = 60.5 - 1.5(9.8) = 45.8 \text{ cm}$$

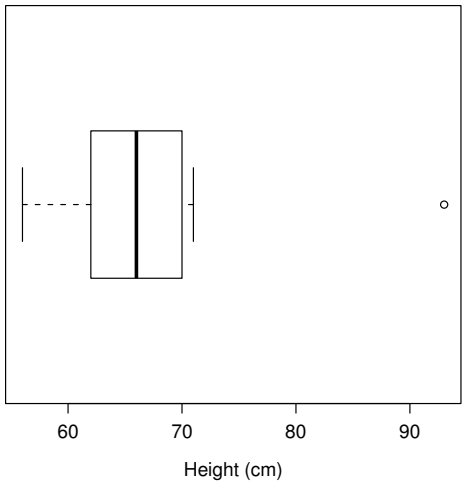
$$\text{Upper Fence} = Q_3 + 1.5(IQR) = 70.3 + 1.5(9.8) = 85.0 \text{ cm}$$

We see there are no data values below the lower fence but 93 is above the upper fence. Thus 93 cm is an outlier.

One could now work with the data set excluding the outliers. For instance we have seen that the mean is particularly susceptible to outliers and taking the mean excluding outliers might be a more useful measure of central tendency in some cases.[20]

A **box plot** is a graphical way of displaying a five-number summary as well as any outliers.

Example:
The following is a box plot for the children's height data of Section 1.16.2:



[20]The **trimmed mean** or **truncated mean** also seeks to exclude outliers before taking a mean. In this case one removes a fixed number or percentage of the data values from the top and bottom of the data set before taking the mean.

Starting in the center (the box) the thick vertical line marks the value of the median. The left side of the box is at Q_1, while the right side of the box is at Q_3. On either side of the box are the "whiskers" which extend from the lowest data value on the left to the highest data value on the right **excluding outliers**. If there are no outliers the end of the whiskers would be at x_1 and x_n, the two remaining numbers of our five-numbers summary. In our case, however, $x_n = 93$ cm, so instead we use the largest non-outlier from the data set, in this case 71 cm, for the end of the whisker on the right. Finally we mark any outliers (here 93 cm) with a circle or asterisk on the horizontal axis.[21] A box plot allows a visual method to rapidly convey the range, spread, and any outliers of a data set. For instance, in the above plot, if we reject the outlier, there appears to be a negative skew to the remaining data due to the pronounced whisker on the left.

Assignment: For each case study determine the fractiles, fractile ranges, outliers, and box plots, requested.

[21] This box plot was generated by the powerful **R** free software environment for statistical computing and graphics. (See http://www.r-project.org/). An accessible interface to R is **R-commander** (http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/). Note that a different convention is used by default in R for calculation of the fractiles resulting in a slightly different five number summary (and hence plot): 56.0, 62.3, 66.0, 70.0, 93.0

1.17 Case Studies

1. A student interested in buying a used textbook from a friend does a quick survey to find out its value. She goes online and finds the following data for used book prices for the text:

Price ($)				
5.00 7.50 7.50 12.00 17.60 50.00 50.00				

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) What is the **number of data elements**?

(i) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(j) Determine the **mode**.

(k) Calculate the **arithmetic mean**.

(l) Calculate the **median**.

(m) Calculate the **average deviation**.

(n) Calculate the **standard deviation** and **variance**. (Use the **definitional** formula.)

(o) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

(p) Calculate the **standard score** of a purchase of a book costing $7.00. Interpret your answer.

(q) Calculate the **first quartile**, the **third quartile**, and the **interquartile range**.

2. A student in a math class received the following scores on his exams for the course:

Score (%)				
65 70 70 79 81 87				

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) What is the **number of data elements**?

(i) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(j) Determine the **mode**.

(k) Calculate the **arithmetic mean**.

(l) Calculate the **median**.

(m) Calculate the **average deviation**.

(n) Calculate the **standard deviation** and **variance**. (Use the **definitional** formula.)

(o) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

(p) Find the **percentage of data** and the **data interval** corresponding to Chebyshev's Theorem with $k = 1.3$. Show that the **actual percentage** of data elements in the data set within this interval satisfies the Chebyshev inequality.

(q) Calculate the **seventh decile**.

(r) Calculate the **five-number summary**, determine any **outliers**, and sketch a **box plot** for the data.

3. A designer of children's games decides to use "3-sided" dice by labelling a regular 6-sided die with two 1's, two 2's, and two 3's. In his game he decides that movement will be determined by rolling two such dice. In order to determine the likelihood of a given roll he rolls two dice repeatedly with the following results for the sum on the dice:

Roll	Occurrences					
2	5					
3	12					
4	17					
5	10					
6	6					

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) Identify the **frequency** column with the appropriate symbol.

(i) What is the **number of data elements**?

(j) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(k) Add a column for **relative frequency**(P). (If the distribution is grouped also add **relative frequency density**(p).) Remember to **sum** columns for which it is appropriate.

*(l) Add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$).

(m) Determine the **mode**.

(n) Calculate the **arithmetic mean**.

*(o) Calculate the **median**.

(p) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

* Questions in the case studies denoted with an asterisk depend on material in optional sections of the course.

(q) Calculate the **standard score** of a roll of 6. Interpret your answer.

*(r) Calculate the **tenth percentile**, the **ninetieth percentile**, and the **10-90 percentile range**.

4. A preschool completes a report for the government indicating the age of students in their care for the past year:

Age (years)	# of children					
1	2					
2	8					
3	4					
4	4					
6	2					

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) Identify the **frequency** column with the appropriate symbol.

(i) What is the **number of data elements**?

(j) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(k) Add a column for **relative frequency**(P). (If the distribution is grouped also add **relative frequency density**(p).) Remember to **sum** columns for which it is appropriate.

*(l) Add columns for **cumulative frequency** $(<Cf)$ and **cumulative relative frequency** $(<CP)$.

(m) Determine the **mode**.

(n) Calculate the **arithmetic mean**.

*(o) Calculate the **median**.

(p) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

(q) Find the **percentage of data** and the **data interval** corresponding to Chebyshev's Theorem with $k = 2$. Show that the **actual percentage** of data elements in the data set within this interval satisfies the inequality.

*(r) Calculate the **sixty-sixth percentile**.

5. A web startup company, *cheapjunk.com*, reported in their annual report the following purchases from their website for the first year of operation:

Value ($)	Purchases							
0.00 – 40.00	20							
40.00 – 80.00	29							
80.00 – 120.00	51							
120.00 – 160.00	28							
160.00 – 200.00	22							

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) Identify the **frequency** column with the appropriate symbol.

(i) What is the **number of data elements**?

(j) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(k) Add a column for **relative frequency**(P). (If the distribution is grouped also add **relative frequency density**(p).) Remember to **sum** columns for which it is appropriate.

*(l) Add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$).

(m) In the space below sketch a **histogram** and add a **frequency polygon** to it. *Also draw an **ogive** for the data if you covered that optional material.

(n) Is your histogram (frequency polygon) **symmetric** or **skewed**? If it is skewed is it **positively** or **negatively** skewed? (Circle the correct answers.)

(o) By looking at only your histogram estimate the **mode** and the **arithmetic mean**. *Also estimate the **median** from your ogive if you covered that optional material. Indicate your reasoning on the graphs.

(p) Determine the **mode**.

(q) Calculate the **arithmetic mean**.

*(r) Calculate the **median**.

(s) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

(t) Calculate the **standard score** of a purchase of value \$100.80. Interpret your answer.

*(u) Calculate the **first decile**, the **ninth decile**, and the **interdecile range**. Verify your fractiles using your ogive.

6. In order to evaluate the viability of a wind turbine installation near Swift Current, SaskPower measured the daily peak wind speed at the location over a month with the following results:

Speed (km/h)	Days							
0 – 20	8							
20 – 40	12							
40 – 60	4							
60 – 80	4							
80 – 100	2							

(a) What is the **population** (i.e. **experimental unit**) under consideration?

(b) Is the data for the whole population (**population data**) or for a sample of it (**sample data**)? (Circle the correct answer.)

(c) What is the **statistical variable**? Label the column with the appropriate symbol.

(d) Is the variable **qualitative** or **quantitative**. (Circle the correct answer.)

(e) Is the variable **continuous** or **discrete**. (Circle the correct answer.)

(f) Is the level of measurement of the variable **nominal**, **ordinal**, **interval**, or **ratio**? (Circle the correct answer.)

(g) Is the data presented as **raw data**, an **ungrouped frequency distribution** or a **grouped frequency distribution**? (Circle the correct answer.)

(h) Identify the **frequency** column with the appropriate symbol.

(i) What is the **number of data elements**?

(j) What is the **sum of observations**? Is your answer approximate? (Circle **yes** or **no**)

(k) Add a column for **relative frequency**(P). (If the distribution is grouped also add **relative frequency density**(p).) Remember to **sum** columns for which it is appropriate.

*(l) Add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$).

(m) In the space below sketch a **histogram** and add a **frequency polygon** to it. *Also draw an **ogive** for the data if you covered that optional material.

(n) Is your histogram (frequency polygon) **symmetric** or **skewed**? If it is skewed is it **positively** or **negatively** skewed? (Circle the correct answers.)

(o) By looking at only your histogram estimate the **mode** and the **arithmetic mean**. *Also estimate the **median** from your ogive if you covered that optional material. Indicate your reasoning on the graphs.

(p) Determine the **mode**.

(q) Calculate the **arithmetic mean**.

*(r) Calculate the **median**.

(s) Calculate the **standard deviation** and **variance**. (Use the **computing** formula.)

(t) Calculate the **standard score** of a wind speed of 150 km/h. Interpret your answer.

*(u) Calculate the **third quartile**. Verify the fractile using your ogive.

Descriptive Statistical Measures

Raw Data	Ungrouped Frequency Distribution	Grouped Frequency Distribution
Measures of the Centre:		
$\mu = \dfrac{\sum x}{N}$ or $\bar{x} = \dfrac{\sum x}{n}$	$\mu, \bar{x} = \dfrac{\sum xf}{\sum f}$	$\mu, \bar{x} = \dfrac{\sum xf}{\sum f}$
$\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}(N+1)$ or $\dfrac{1}{2}(n+1)$	$*\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}\left(\sum f + 1\right)$	$*\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}\left(\sum f\right)$
$\underset{\text{Value}}{\text{Median}} = x_{\frac{1}{2}(N+1)}$ or $x_{\frac{1}{2}(n+1)}$	$*\underset{\text{Value}}{\text{Median}} = x_{\frac{1}{2}(\sum f+1)}$	$*\underset{\text{Value}}{\text{Median}} = L_i + \dfrac{\left\{\frac{1}{2}(\sum f) - <Cf_{i-1}\right\}}{f_i} \cdot \Delta x$
Measures of Dispersion:		
a.d. $= \dfrac{\sum \lvert x-\mu \rvert}{N}$ or $\dfrac{\sum \lvert x-\bar{x} \rvert}{n}$		
$\sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{N}}$		
$\sigma = \sqrt{\dfrac{\sum x^2}{N} - \left(\dfrac{\sum x}{N}\right)^2}$	$\sigma = \sqrt{\dfrac{\sum x^2 f}{\sum f} - \left(\dfrac{\sum xf}{\sum f}\right)^2}$	$\sigma = \sqrt{\dfrac{\sum x^2 f}{\sum f} - \left(\dfrac{\sum xf}{\sum f}\right)^2}$
$s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$		
$s = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$	$s = \sqrt{\dfrac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}}$	$s = \sqrt{\dfrac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}}$
$R = x_n - x_1$	$IQR = Q_3 - Q_1$	$IDR = D_9 - D_1$	10-90 $PR = P_{90} - P_{10}$
Lower Fence $= Q_1 - 1.5(IQR)$		Upper Fence $= Q_3 + 1.5(IQR)$	
Standard Scores			
$z = \dfrac{x-\mu}{\sigma}$		$z = \dfrac{x-\bar{x}}{s}$	
$x = \mu + z\sigma$		$x = \bar{x} + zs$	
Miscellaneous Formulae:			
$P = \dfrac{f}{\sum f}$	$p = \dfrac{P}{\Delta x}$	$1 - \dfrac{1}{k^2}$	

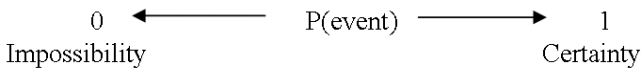
*Formulae denoted with an asterisk are from optional sections of the course.

Unit 2: Basic Probability

2.1 Definitions of Probability

2.1.1 Subjective Probability

The **subjective** approach to probability is a method to express the strength of one's belief in the future occurrence of an event. The beliefs are usually based upon past experience or prior knowledge. Probability is assigned to an event as a number on a continuum between 0 and 1:



Example:
A biologist estimates that there is a 75% chance that if burrowing owls are reintroduced to an area of southern Saskatchewan that they will survive. What does the number mean and how would a biologist arrive at this number?

2.1.2 Experimental Probability

The **relative frequency** definition of probability relates the expected future rate of occurrence to what has happened in the past. The difference between this and subjective probability is that in the relative frequency definition an experiment is done and we refer to it therefore as **experimental probability**. To find the experimental probability of an event, repeat an experiment F times. Observe the number of times the event happened, the frequency f, out of the total number of times the experiment was repeated, F. The probability of the event is then

$$P(\text{event}) = \frac{\text{Frequency of occurrence of event}}{\text{Number of repetitions}} = \frac{f}{F}$$

Thinking back to our discussion of frequency distributions in Unit 1 we note that $F = \sum f$ where the sum counts the frequencies of all the possible outcomes, and hence the number of experiments done.

Example:
In a talk radio phone survey, the question was put to 10 different people: "Are you in favour of the government extracting higher resource royalties from corporations?" 7 people replied that they were. Based on these results, what is the probability that the next person asked would be in favour?

$$P(\text{in favour}) = \frac{f}{F} = \frac{7}{10} = 0.7$$

Notice that we have already done these calculations extensively in Unit 1. There we would have calculated the relative frequency or proportion P of an outcome. The calculation (and symbol) is identical for experimental probability but the interpretation is different. Now we use the proportion to infer the probability that an individual shopper chosen at random in the mall will be in favour of extended store hours. In the following exercise we illustrate the meaning of probability as a relative frequency of occurrence (experimental probability).

Example:
Take 4 coins and shake them fairly by allowing them to drop unobstructed on smooth surface. Repeat this experiment 64 times. In the 8×8 grid below record the number of heads that occur on a repetition of the experiment in each cell. Count the frequency of each value and record your results in the table on the right:

$\Rightarrow$

x(heads)	f	$P(x)$
0		
1		
2		
3		
4		
	$\sum f =$	$\sum P(x) =$

Graph the results below with 7 intervals on the horizontal axis and the number of heads as the midpoint of each interval. Plot $P(x)$ on the vertical axis.

Interpretation:

1. If $P(x)$ means the probability of the occurrence of x heads, what does $P(1)$ mean?
2. What is $P(1)$ specifically in your table?
3. How many times would you expect to get 1 head in the future, based on what has happened in this experiment, if you flipped 4 coins 200 times?
4. What is the shape of the distribution curve?
5. By examining the distribution curve, what is an estimate of the mean of the distribution of heads upon flipping 4 coins?
6. Calculate the mean and standard deviation of the distribution of heads from the data in the table. (Use $\bar{x}$ and s . Why?)
7. If 4 coins are flipped, what is the probability of getting 2 heads? Why isn't it 0.5?
8. What is the area under the $P(x)$ histogram?
9. Would someone else's determination of the experimental probability be identical with yours?

2.1.3 Theoretical Probability

The **theoretical**[1] approach to probability is based upon determining the total number of possibilities in advance of taking the action. The set of possibilities is called the **sample space**[2] In this case let n be the number of possibilities associated with the event while N is the total number of outcomes associated with the action to be taken (i.e. the size of the sample space). If the events in the sample space are **equally likely** then the theoretical probability of an event is:

$$P(\text{event}) = \frac{\text{Possible ways for event to happen}}{\text{Total number of possibilities}} = \frac{n}{N}$$

Example:
A group of four contains two males and two females. A committee of 2 is to be struck by drawing lots. What is the probability that the committee consists of two females? Count the possibilities.

N = 6

John, Bill	Bill, Sue
Bill, Mary	Mary, Sue
John, Sue	John, Mary

n = 1

Mary, Sue

$$P(\text{two female committee}) = \frac{n}{N} = \frac{1}{6} = 0.17$$

The theoretical probability is unique; assuming we do our calculations correctly we will all get the same answer. One could have calculated an experimental probability for this event by writing the six possible pairs on slips of paper and drawing them out of a hat repeatedly. In that case different experimental probabilities for the event would not likely have been exactly 1/6 but would have been close to the true theoretical probability. The experimental probability will converge to the theoretical probability as F, the number of experiments performed, increases. In symbols the actual (theoretical) probability satisfies

$$P(\text{event}) = \lim_{F \to \infty} \frac{f}{F}$$

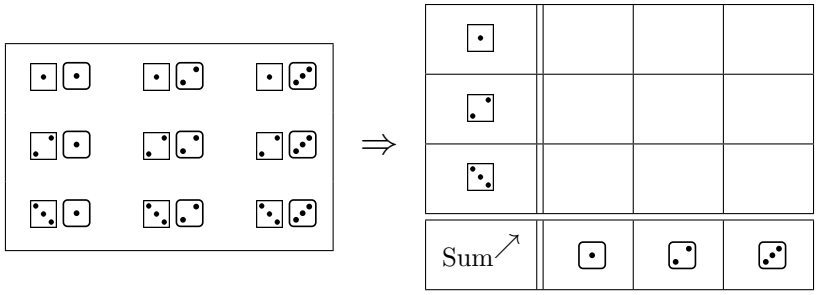
where f is the number of occurrences of the event in the F experiments and ∞ means positive infinity.

If the probability is calculable theoretically, then this is preferable to an experimental probability. However many systems are too complicated to have probabilities calculated theoretically. For instance there is no way a theoretical calculation of probability may be done to find the probability a citizen is in favour of corporations paying higher resource royalties to the government.

In Case Study 3 of Unit 1 a game designer was interested in determining the proportion (probability) of the sum on two 3-sided dice. The designer effectively calculated an experimental probability for each of the possible events (2,3,4,5,6). The object of the next exercise is to illustrate the theoretical meaning of probability as a proportion in the sample space of all possibilities.

[1] Theoretical probability is also referred to as **classical** probability.
[2] The sample space is also sometimes called the **universe**.

Example:
Suppose two three-sided dice[3] are to be rolled so that they land fairly. We are interested in the probability of obtaining a specific sum on the two faces. The sample space consists of the following equally likely outcomes (left). List all of the possible sums on the faces of the dice by filling the appropriate sum inside the square on the grid to the right.



Summarize your findings in the table below and draw the graph of the observations to the right. You may want to compare your results to the relative frequencies (P) found experimentally in Case Study 3 of Unit 1.

x(Sum)	n	$P(x)$
2		
3		
4		
5		
6		
	$N =$	$\sum P(x) =$

Interpretation:

1. If $P(x)$ means the probability of rolling a sum of x, what does $P(3)$ mean? What is your value for $P(3)$?
2. If two dice were rolled 180 times, how many times would you expect to see a sum of 3?
3. Describe the shape of this distribution.
4. What is the mean of the distribution of sums? Use μ not $\bar{x}$. (Here treat the n column as frequency f and use our formulae from Unit 1.)
5. What is the standard deviation of the distribution of sums? Use σ not s.
6. How does the area under the curve compare to the sum of the $P(x)$ column?
7. What is wrong with the argument, "There are five possible sums (2, 3, 4, 5, 6) so the probability of any one of them is $\frac{1}{5}$" ?

[3]Recall our *three-sided* die is created by labelling two sides of a six-sided die with a 1, two sides with a 2, and two-sides with a 3. Would the above analysis be appropriate if three sides were labelled with a 1, two sides with a 2, and one side with a 3? How would the above analysis have to change?

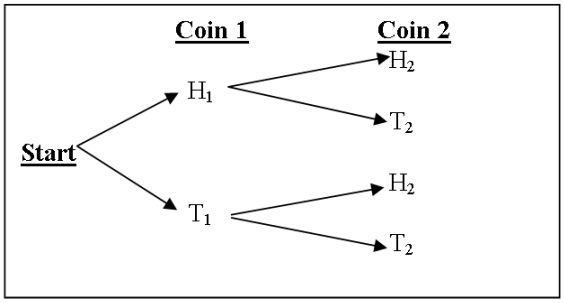
2.2 Counting Outcomes

We already know from Unit 1 how to determine experimental probability, it is just the relative frequency, P. To compute the probability of an event theoretically it is necessary to count the number of **outcomes** associated with both the **event** and the **sample space** of equally likely outcomes. There are various mathematical methods to assist in counting outcomes.

2.2.1 Tree Diagrams

Tree Diagrams list, in a tree structure, each of the steps in an experiment together with all possible outcomes at that step. Each path represents a possible outcome of the total experiment.

Example:
Two coins are flipped. Show all possible outcomes.



2.2.2 The Counting Theorem

The **Counting Theorem** gives the total number of possibilities, without listing each possibility, in a tree structure. This is useful in cases where it is not practical to draw the tree structure because it has too many paths. List each step in the experiment and determine the number of possible ways each step can be done. The number of ways of doing the whole experiment is found by multiplying the number of possibilities at each step in the experiment.

Example:
For the experiment of flipping two coins:

$$\underbrace{(2)}_{\boxed{\text{Coin 1}}} \cdot \underbrace{(2)}_{\boxed{\text{Coin 2}}} = 4 \text{ possibilities in total}$$

2.2.3 Factorial

A **factorial** is a mathematical formula encountered in counting outcomes. It is denoted as $n!$ and is calculated as follows:

$$\boxed{n! = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 3 \cdot 2 \cdot 1}$$

The factorial $n!$ is the number of ways of **arranging** n items since, using the counting theorem, after the first slot is filled in one of n ways, the second slot can be filled in $(n-1)$ ways, etc.

Example:
In how many ways can five children be arranged in five different seats in a minivan?

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120 \text{ ways}$$

Find the $\boxed{n!}$ key on your calculator and show that $5! = 120$.

Notice that factorials increase in value very quickly. $20! = 2.4329020 \times 10^{18}$ on your calculator. (i.e. move the decimal 18 places to the right.) 800! exists but is too large for calculation purposes on your calculator. Finally note that by definition $0! = 1$ not 0 as may be verified on your calculator.

2.2.4 Permutations

An **arrangement** of the letters a to e could be written (c, e, b, a, d), where parentheses have been used, just as when we plot points, to remind us that the **order matters**. Such an arrangement is called a **permutation** and we have seen that $n!$ gives us the number of ways of arranging n objects, or in other words, the number of permutations of the n objects. Sometimes we want to count the number of ways of arranging fewer than all the objects. A permutation where we have selected only 3 of the 5 letters from a to e could be (c, e, b). The **permutation formula** ${}_n P_r$ counts the number of different permutations of r objects taken from n distinct objects. It equals[4]

$$\boxed{{}_n P_r = \frac{n!}{(n-r)!}} \leftarrow \text{ no repetitions are allowed}$$

No repetitions of the same element are allowed in a permutation of distinct objects.

If repetitions of the same element are allowed in an arrangement, then the number of arrangements of r objects taken from the n objects is, by the counting theorem,

$$\boxed{\underbrace{(n) \cdot (n) \cdot \ldots \cdot (n)}_{r \text{ times}} = n^r} \leftarrow \text{ repetitions are allowed}$$

In such a situation we would be allowed to count the arrangement (a, a, d) if we were counting arrangements of 3 objects taken from a to e with repetition.

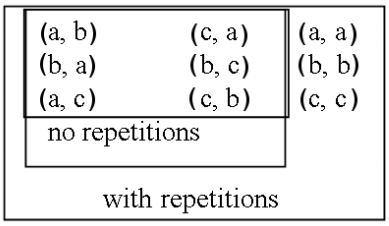
Example:
Suppose arrangements of two objects are to be made, taken from the collection $\{a, b, c\}$. How many arrangements are possible?

$${}_3 P_2 = \frac{3!}{(3-2)!} = \frac{3 \cdot 2 \cdot 1}{1} = 6 \text{ with no repetitions}$$

$$3^2 = 9 \text{ with repetitions}$$

Find your calculator key $\boxed{{}_n P_r}$ and evaluate the above expression directly. As a check the following diagram shows all the possible arrangements.

[4] To see why this is the formula, note by the counting theorem that there are n ways to pick the first item, $(n-1)$ ways to pick the second, ..., and $(n-r+1)$ ways to pick the r^{th}. The number of ways to pick an arrangement is therefore $n \cdot (n-1) \cdot \ldots \cdot (n-r+1)$ which equals the permutation formula since the denominator cancels the lower terms from $(n-r)$ to 1 in the $n!$.



It is always possible to bypass the permutation formula and use the counting theorem.

3 X 2 = 6 ← with no repetitions.
3 X 3 = 9 ← with repetitions.

2.2.5 Combinations

Sometimes order is not important in a counting problem. For instance, if you wanted to know in how many ways you could select three of your five friends to go on a trip with you, the order of the three selected would be meaningless. A **combination** is such a selection. A combination of 3 items selected from the letters a to e would be $\{b, c, e\}$. We use braces here to remind us that order is not meaningful. (The same combination could be written $\{c, b, e\}$.) The **combination formula** counts the number of different **selections** of r objects taken from n objects. It is written ${}_nC_r$ and calculated as[5]

$$\boxed{{}_nC_r = \frac{n!}{r!(n-r)!}}$$

Example:
Suppose two objects are to be selected from the collection {a, b, c}. How many selections are possible?

$${}_3C_2 = \frac{3!}{2!(3-2)!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3$$

Find your calculator key $\boxed{{}_nC_r}$ and evaluate the above expression directly. As a check here is the list of the possible selections:

$$\boxed{\{a, b\},\ \{a, c\},\ \{b, c\}}$$

You cannot bypass the combination formula with the counting theorem.

2.2.6 Contingency Tables

Contingency Tables cross tabulate a data array where the outcomes consist of two factors into a summary table.

Example:
A die is rolled and a coin is tossed at the same time. A person interested in determining the probability of the joint result experimentally summarizes the results of 100 experiments in the following (frequency) contingency table:

		DIE						
		1	**2**	**3**	**4**	**5**	**6**	Total
COIN	**Heads**	9	8	8	9	10	2	46
	Tails	9	10	7	6	12	10	54
	Total	18	18	15	15	22	12	100

[5]The extra $r!$ in the denominator of ${}_nC_r$ versus ${}_nP_r$ deals with the fact that the $r!$ permutations of the selected objects are equivalent as a combination where order is irrelevant. Other combination notation used is $\binom{n}{r}$ or $C(n, r)$.

Notes:

1. The totals at the side and bottom of the table are called **marginal frequencies**. The outcomes included in a marginal frequency have only one attribute. For example, of the hundred trials, heads occurred 46 times. 46 is a marginal frequency.
2. The numbers inside the table are called **joint frequencies**. The outcomes included in a joint frequency have two attributes. For example, of the hundred trials, on 9 occasions the coin turned up heads and the face on the die was a 4.
3. Note that we could have placed these results in a single column frequency table with 12 entries where the first variable value was $x = (\text{heads}, 1)$ with frequency 9 and the last was $x = (\text{tails}, 6)$ with frequency 10. Obviously the contingency table for displaying the frequencies is preferable.
4. Later we will use contingency tables for theoretical probability where the entries correspond to the number of elements of the sample space corresponding to the event. If we had done that here, what would the table have looked like?

Assignment:

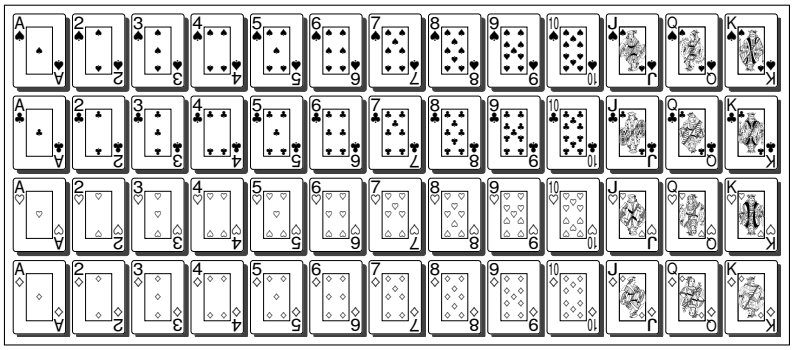
1. A white jar and a blue jar each contain a collection of chips. Some of the chips are red and some are green. A jar is chosen at random and a chip is randomly chosen from it. Construct a tree diagram of the sample space of outcomes for this experiment.
2. A company is selecting a new president and vice president. There are 15 females and 100 males on staff. How many possible ways does the company have of selecting a president and vice president if the president must be female but there are no restrictions on the choice of a vice president? (Hint: Use the counting theorem.)
3. Three newly hired people are to be assigned to 3 different regions. How many different arrangements are possible if a person can be assigned to any (but only one) of the regions?
4. Six newly hired people are to be paired into teams. How many different pairs can be formed?
5. A high school science fair has 20 participants. In how many ways can
   (a) Four students be chosen to move on to the regional science fair?
   (b) The four awards – 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} place – be awarded?
   (c) The four awards – *best new idea*, *best use of math*, *best presentation*, and *best sweet-talking of judges* – be awarded?
6. A survey of 100 business people showed that 30 were for and 70 were against a bylaw. Ten of those for the bylaw were gas station operators, 20 against the bylaw were fast food outlet managers. If only gas station operators and fast food outlet owners were included in the survey, construct a contingency table of the survey results.
7. A builder builds 3 different models of homes. Each home has the option of two different floor plans and comes with or without an attached garage. Draw a tree diagram that lists all possible arrangements of homes built by the builder.
8. The integers from 1 to 49 are written on a sheet of paper. A person is asked to cross out 6 of the integers. In how many ways can this task be completed?
9. Calculate the value of: ${}_{900}C_{898}$

2.3 Set Theory and Counting Outcomes

2.3.1 Events are Sets!

Set Theory is a useful means of identifying events, combinations of events and numbers of outcomes in the sample space. A **set** is just a collection of objects called **elements**.

Example:
The set of playing cards in a standard deck would be, graphically,[6]



and an element in the set would be the the king of spades, $K\spadesuit$.

An **experiment** involves making a measurement or observation. The outcome of a single experiment is called a **simple event**. For example, if we consider the experiment of drawing a single card from the deck, then a **simple event** could be drawing the king of spades. Hence we can identify a simple event with a set element. The **sample space**, $\boldsymbol{S}$, is the set of all such simple events, in this case the set containing the entire deck.

A subset of S would also be a set. For example, the set $\boldsymbol{Q}$ of queens in a deck of cards is

$$\boldsymbol{Q} = \{Q\spadesuit, Q\clubsuit, Q\heartsuit, Q\diamondsuit\} \ .$$

Since order is irrelevant for a set we use braces { } when writing it down here as we did for listing combinations. The connection between sets and probability is that we can represent the **event** of drawing a queen by its **set** of corresponding simple events in the sample space. The **number of elements** in a set we will write as $\boldsymbol{n(\ )}$. Here $n(Q) = 4$.

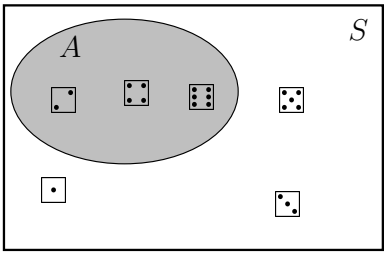
Pictorially it is useful to represent the set S of all simple events in the sample space with a rectangular box. For the experiment of drawing a single card from a deck of cards the sample space is just the deck shown in the box above. A particular event A (i.e. its set) will be some subset of these outcomes, which we can represent with a circle or oval inside the box. Such a diagram is called a **Venn diagram**.

In a Venn diagram one does not usually draw the simple events like the cards above. Rather one draws the events as circles and visualizes their contents and potential overlap with other sets. The dice in the following examples will be shown explicitly in this section to aid in understanding the concepts but a real Venn diagram would not contain them.

[6]Cards are commonly used in discussing probability. Note that a "standard deck" means no extra cards (wild cards/jokers). A "black" card is either from the "suit" of clubs($\clubsuit$) or spades($\spadesuit$) while a "red" card is either from the suit of hearts($\heartsuit$) or diamonds($\diamondsuit$). A "facecard" is a Jack, Queen, or King of any suit. Aces may be counted low (like a 1 before 2) or high (after kings) depending on the game.

Example:
Suppose a die is rolled. Let $\boldsymbol{S}$ represent the sample space. Let $\boldsymbol{A}$ **be the event the die is even**. Draw a Venn diagram and count the number of simple events in the sample space and event A.

Solution:



$\boldsymbol{S}$ = all outcomes in the sample space = $\{1, 2, 3, 4, 5, 6\}$
$\boldsymbol{A}$ = all outcomes that are even = $\{2, 4, 6\}$
$\boldsymbol{n(S)}$ = number of outcomes in the sample space = 6
$\boldsymbol{n(A)}$ = number of outcomes in event A = 3

The event $\boldsymbol{A}$ is shaded in the Venn diagram.

2.3.2 Complement of an Event

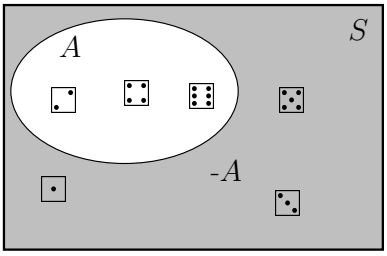
The **complement of an event**, written $\boldsymbol{\text{-}A}$, is the collection of all simple events in the sample space that are **not** in A. The number of simple events in the complement of an event must satisfy

$$\boxed{n(A) + n(\text{-}A) = n(S)}$$

since all outcomes in the sample space belong either to A or its complement, and these do not overlap.

Example:
Identify the complement $\text{-}A$ of A of the previous example and find the number of its elements directly and by formula.

Solution:



$\boldsymbol{\text{-}A}$ = the nonoccurrence of event A = $\{1, 3, 5\}$
$\boldsymbol{n(\text{-}A)}$ = number of outcomes in $\text{-}A$ = 3 (counting directly)
$\boldsymbol{n(\text{-}A)}$ = $n(S) - n(A)$ = $6 - 3$ = 3 (by formula)

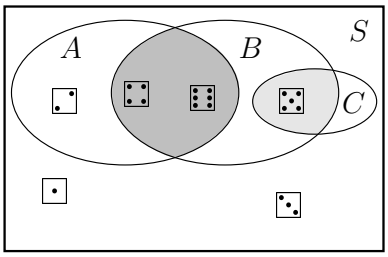
The event $\boldsymbol{\text{-}A}$ is shaded in the Venn diagram.

2.3.3 Intersection of Events

Just as we were able to create a new event by taking the complement of an event, we can also combine several events to create a new event. The **intersection of events** is the region of overlap of events. For two events it is written ***A* and *B*** . It is sometimes called the **joint event** because two things have to happen at the same time for this event to happen.

Example:
Consider the previous problem of rolling the die. Suppose A is the event previously defined (an even number). Let ***B* be the event event of rolling a number greater than 3**. Let ***C* be the event of rolling a 5**. Draw a Venn diagram of the sets, find their mutual intersections, and count the elements in each set.

Solution:



$\boldsymbol{B}$ = all outcomes in event B = $\{4, 5, 6\}$
$\boldsymbol{C}$ = all outcomes in event C = $\{5\}$
A* and *B = all outcomes in both A and B = $\{4, 6\}$
B* and *C = all outcomes in both B and C = $\{5\}$
A* and *C = all outcomes in both A and C = $\{\}$

Note that the event ***B* and *A*** is the same as ***A* and *B*** so there are no more possible intersections between two events.

The counts in each set are:

$\boldsymbol{n(B)}$ = number of outcomes in event B = 3
$\boldsymbol{n(C)}$ = number of outcomes in event C = 1
$\boldsymbol{n(A \text{ and } B)}$ = number of outcomes in event A and B = 2
$\boldsymbol{n(B \text{ and } C)}$ = number of outcomes in event B and C = 1
$\boldsymbol{n(A \text{ and } C)}$ = number of outcomes in event A and C = 0

*The event ***A* and *B*** is shaded dark while ***B* and *C*** is shaded light in the Venn diagram.*

2.3.4 Mutually Exclusive Events

In the last example the intersection of events A and C was empty, A and $C = \{\}$. If an event has no outcome, it is called the **empty set** and is denoted by the symbol $\varnothing$. If two events have as their intersection $\varnothing$ they are called **mutually exclusive**. The following are all equivalent:

1. Event A and event C are **mutually exclusive**.
2. Event A and event C are **disjoint**.
3. Event A and event C have no overlap.
4. Event A and event C cannot happen at the same time, $P(A \text{ and } C) = 0$.
5. A and $C = \{\} = \varnothing$
6. $n(A \text{ and } C) = 0$

2.3.5 Union of Events

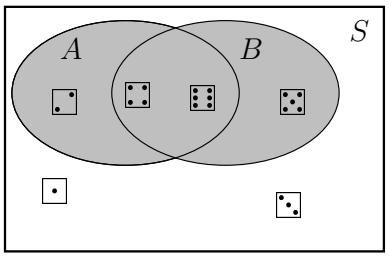
Another way to create new events from other events is to take their union. The **union of events** is the collection of all the simple events in the events combined. For two events it is written ***A* or *B*** . It contains all the simple events in A together with all the simple events in B **including the joint simple events in both**. The number of outcomes in A or B can be written in terms of the number simple events in the events A, B, and A and B as follows:

$$\boxed{n(A \text{ or } B) = n(A) + n(B) - n(A \text{ and } B)}$$

Here we have to subtract the number of elements in A and B because otherwise the elements that are in both A and B get counted twice.

Example:
Consider the previous experiment of rolling a single die where A is rolling an even number and B is rolling a number greater than 3. Find the union of the events, the number of simple events it contains, and show that the above formula for calculating the number of simple events holds.

Solution:



$$\begin{aligned} \textbf{\textit{A} or \textit{B}} &= \text{all outcomes in either event } A \text{ or event } B = \{2,4,6,5\} \\ \textbf{\textit{n}(\textit{A} or \textit{B})} &= \text{number of outcomes in } A \text{ or } B = 4 \text{ (counting directly)} \\ \textbf{\textit{n}(\textit{A} or \textit{B})} &= n(A) + n(B) - n(A \text{ and } B) = 3 + 3 - 2 = 4 \text{ (by formula)} \end{aligned}$$

Note that the event ***B* or *A*** is the same as ***A* or *B***.

*The event ***A* or *B*** is shaded in the Venn diagram.*

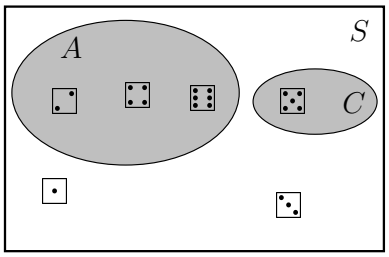
In the case of disjoint or mutually exclusive events we have the following special case formula, since $n(A \text{ and } C) = 0$,

$$\boxed{n(A \text{ or } C) = n(A) + n(C) \quad (A, C \text{ mutually exclusive})}$$

There is no overlap so we can just add the number of events.

Example:
Consider the previous experiment of rolling a single die where A is rolling an even number and C is rolling a 5. Find the union of the events, the number of simple events it contains, and show that the above formula for calculating the number of outcomes holds.

Solution:



$$\begin{aligned} \textbf{\textit{A} or \textit{C}} &= \text{all outcomes in either event } A \text{ or event } C = \{2,4,6,5\} \\ \textbf{\textit{n}(\textit{A} or \textit{C})} &= \text{number of outcomes in } A \text{ or } C = 4 \text{ (counting directly)} \\ \textbf{\textit{n}(\textit{A} or \textit{C})} &= n(A) + n(C) = 3 + 1 = 4 \text{ (by formula for m.e. events)} \end{aligned}$$

*The event ***A* or *C*** is shaded in the Venn diagram.*

Note that the union event A or B includes the outcomes for which both A and B occur. In words our "or" is being used **inclusively**. If one means to exclude the overlap of events, then the term **exclusive or** may be used. We will never mean this when we use the word *or* in this course. Thus if you have a BLT sandwich for lunch and you are asked if you had bacon or lettuce for lunch the answer is logically yes; you are included in ***B* or *L***.[7]

2.3.6 Other Notation

We list here other common notation which occur equivalent to ours above:

complement	-A	A^c, A'
intersection	A and B	$A \cap B$
union	A or B	$A \cup B$

The notation on the right for intersection and union comes from set theory. Because we wish to emphasize the logic in events we will use "and" and "or" directly in our notation.

[7] One source of confusion in the correct inclusive interpretation of or is that our formula for counting $n(A \text{ or } B)$ subtracts $n(A \text{ and } B)$. However recall this subtraction is not being done because the region A and B is not being included, but because we do not want to count the region twice.

Assignment:

Use proper set notation and Venn diagrams to do the following questions.

1. In a standard deck of 52 playing cards:
   (a) How many cards are sixes?
   (b) How many cards are not sixes?
   (c) How many cards are spades and sixes?
   (d) How many cards are spades or sixes?
2. A graduate of a course is looking for a job. H is the event that the job has a high salary; P is the event that the job has a good pension and F is the event that the job has a good future.
   (a) Draw a Venn diagram of all possible jobs and place an integer in each separate region within the Venn diagram. (Check the answer at the back of the study guide to label your regions the same, and then proceed to do the next part of the question.)
   (b) List the numbers of the regions that correspond to the following combinations of events:

$$\text{-}H, (F \text{ and } P), (H \text{ or } F), \text{-}(H \text{ or } F \text{ or } P), (F \text{ and } \text{-}H), (H \text{ and } F \text{ and } P)$$

3. A survey of 100 tourists entering international customs shows the following: 60 visited London, 30 visited Rome and 10 visited London as well as Rome.
   (a) Draw a Venn diagram of the situation. When you draw the diagram, keep in mind that some of the 60 who visited London could have also visited Rome, etc. Use "L" for having visited London and "R" for having visited Rome.
   (b) How many tourists visited London or Rome?
   (c) How many tourists visited neither place?
   (d) How many tourists visited only London?
   (e) What proportion of the tourists who visited Rome also visited London?

2.4 Rules Governing Probability

To calculate probabilities more easily, there are some relationships that can be used that are based upon the previous methods of counting outcomes.

2.4.1 Probability of An Event

Recall the basic rule for theoretical probability, written in terms of our new notation is:

$$\boxed{P(A) = \frac{n(A)}{n(S)}},$$

provided *all simple events in the sample space S are equally likely.*

Since the number of outcomes corresponding to an event must be at least zero and at most the size of the entire sample space, one has $0 \leq n(A) \leq n(S)$. Dividing each term by $n(S)$ and using the above formula gives

$$\boxed{0 \leq P(A) \leq 1}.$$

Make sure that when probabilities are calculated, the answer is in this range. While we have used equally likely simple events in deriving this result, this result and the others derived below do not require this to be the case.

2.4.2 Complement Rule

Dividing our complement counting rule,

$$n(A) + n(\text{-}A) = n(S),$$

by the number of outcomes in the sample space, $n(S)$, gives the following rule for the probability of a complement event:

$$\boxed{P(A) + P(\text{-}A) = 1}$$

Often the probability of the complement of an event is easier to calculate than the event itself. Calculate that event first and then use this rule to find the event of interest.

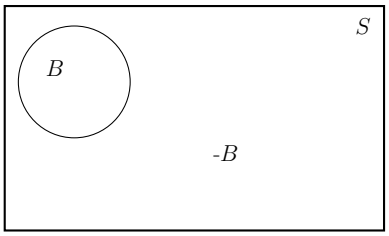
Example:
A biologist samples the fish found in the Frenchman River in south-western Saskatchewan with the following results:

Species	f
Northern Redbelly Dace (N)	2
Fathead Minnow (F)	6
White Sucker (W)	5
Brook Stickleback (B)	3
Iowa Darter (I)	1
	$\sum f = 17$

1. If a fish is selected at random from the sample, what is the probability it is a Brook Stickelback?
2. What is the probability it is not a Brook Stickelback?

Draw a Venn diagram to identify and count outcomes.

Solution:



1. $\boxed{P(B)} = \dfrac{n(B)}{n(S)} = \boxed{\dfrac{3}{17}}$

2. $\boxed{P(\text{-}B)} = 1 - P(B) = \boxed{\dfrac{14}{17}}$

Notice that because we were asked the likelihood of drawing the fish **from the sample** and not **the river** this is actually a theoretical probability calculation.

2.4.3 Addition Rules

Calculating the probabilities of combinations of events, usually involves either adding or multiplying probabilities. Addition is required when the events are combined by the OR operation as will now be shown. Multiplication is required when the events are combined by the AND operation as will be shown in Section 2.5.

Our two addition rules for counting events

$$n(A \text{ or } B) = n(A) + n(B) - n(A \text{ and } B)$$

$$n(A \text{ or } B) = n(A) + n(B) \quad (A, B \text{ mutually exclusive})$$

divided by $n(S)$ give the following two addition probability rules, one general and one special case:

$$\boxed{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$$

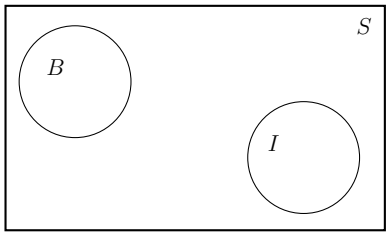
$$\boxed{P(A \text{ or } B) = P(A) + P(B) \quad (A, B \text{ mutually exclusive})}$$

These could be extended to include more than two events. For example the second rule for the union of three mutually exclusive events is $P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$.

Venn diagrams and contingency tables are helpful in applying the above rules.

Example:
In the previous example, find the probability that a fish randomly selected from the sample is a Brook Stickleback (B) or an Iowa Darter (I). Draw a Venn diagram for the situation.

Solution:



The two events are mutually exclusive as a selected fish cannot belong to both species at once. Using the special case addition rule gives:

$$\boxed{P(B \text{ or } I)} = P(B)+P(I) = \frac{n(B)}{n(S)}+\frac{n(I)}{n(S)} = \frac{3}{17}+\frac{1}{17} = \boxed{\frac{4}{17}}$$

Example:
The biologist in the previous example also examines the sex of each fish in the sample and presents the data the following contingency table:

		Species					
		N	F	W	B	I	Total
Sex	Male	1	4	3	1	0	9
	Female	1	2	2	2	1	8
	Total	2	6	5	3	1	17

If a fish is selected at random from the sample:

1. What is the probability that it is female or a Brook Stickelback (B)?
2. What is the probability that it is a Northern Redbelly Dace (N) or a White Sucker (W)?
3. Draw a Venn diagram.

Use the symbols M and -M for Male and Female, since F is already being used for a fish species. (Also it is desirable to use the complement in cases like this where there are only two possibilities.)

Solution:

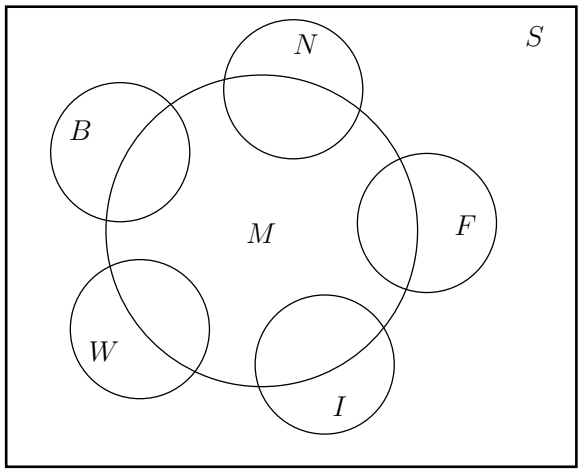
1. One notices that it is possible to draw a fish that is both a Brook Stickelback (B) and Female ($-M$) so the events are **not** mutually exclusive. Using the general addition rule gives:

$$\begin{aligned} \boxed{P(B \text{ or } \text{-}M)} &= P(B) + P(\text{-}M) - P(B \text{ and } \text{-}M) \\ &= \frac{3}{17} + \frac{8}{17} - \frac{2}{17} \\ &= \boxed{\frac{9}{17} = .529} \end{aligned}$$

2. As seen in the earlier example, different species are mutually exclusive events. (In general two rows or two columns on a contingency table will always describe mutually exclusive events.) Use the special case addition rule:

$$\boxed{P(N \text{ or } W)} = P(N) + P(W) = \frac{n(N)}{n(S)} + \frac{n(W)}{n(S)} = \frac{2}{17} + \frac{5}{17} = \boxed{\frac{7}{17} = .412}$$

3. The following is a Venn diagram for the entire situation:



Note:

- Each event labelled is associated with the interior of a particular circle or rectangle. Its complement is the exterior region.
- The species circles have been drawn disjoint because we know all the species types are mutually exclusive. (For a general Venn Diagram where nothing is known about mutual exclusivity all the regions would be drawn to overlap.)
- The circles slice up the sample space into disjoint regions in which one should imagine there lie all possible outcomes (here fish). Label each such region with the number of outcomes it contains (the frequency) from the contingency table. Use the joint frequencies from inside the table, not the marginal frequencies. (Why?)
- Similarly label each disjoint region with the probability it contains.
- Any event of interest corresponds to some union of regions in the Venn diagram for which you can now add up the probability. Consider the following event
  (not a male Fathead Minnow) and (not a female Brook Stickelback)
  Shade (or visualize) the region and find its probability.

Assignment:

1. A bookkeeper records transactions from 40 invoices to a journal. 3 of the invoices were completed inaccurately.
   (a) An auditor selects one of these transactions at random. What is the probability that the transaction is in error?
   (b) What is the probability that the transaction is not in error?
2. 60% of all students at a school take Chemistry and 50% take Biology. 10% of all students take both Chemistry and Biology.
   (a) Draw a Venn diagram of the situation.
   (b) Are the events mutually exclusive?
   (c) What is the probability that a student takes either Chemistry or Biology? (Shade the area)
   (d) What is the probability that a student takes neither Chemistry nor Biology? (Shade the area)
   (e) What is the probability that a student takes Chemistry but not Biology? (Shade the area)
3. A card is selected at random from a standard deck of 52 playing cards. What is the probability that the card is either a red five or a black six?

4. Galaxies in the local universe are observed. It is found that 75% have a disk. It is found that 60% have interstellar matter (dust and gas between stars). It is found that 50% have both a disk and interstellar matter.

   (a) Draw a Venn diagram for this information. Use D for labelling the event that a galaxy has a disk and I for the event it has interstellar matter.

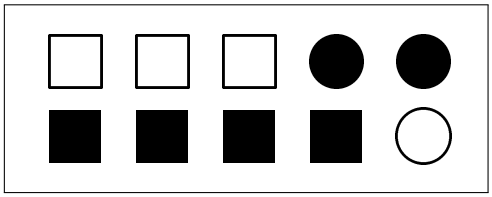
   (b) **Spiral** galaxies have a disk and interstellar matter. **Lenticular** galaxies have a disk and no interstellar matter. **Irregular** galaxies have interstellar matter and no disk. **Elliptical** galaxies have no disk and no interstellar matter. Label the distinct regions of your Venn diagram with the corresponding galaxy subtype name.

   (c) Assuming that the only galaxies in each region are of each galactic subtype, find the probability that a randomly selected galaxy is of each subtype (spiral, lenticular, irregular, and elliptical). Read the probabilities off of your Venn diagram. Label each of these event types in terms of D, I and the appropriate logical symbols.

2.5 Conditional Probability and the Multiplication Rules

2.5.1 Conditional Probability

A container contains a group of black and white cubes and spheres:



One object is selected at random. Label the following events as shown:

B = The object is Black.

-B = The object is white (not Black).

C = The object is a Cube.

-C = the object is a sphere (not Cube).

Using our basic probability rules we see that the probability an object is black is:

$$P(B) = \frac{n(B)}{n(S)} = \frac{6}{10} = \frac{3}{5}$$

One may similarly verify that $P(-B) = \frac{2}{5}$, $P(C) = \frac{7}{10}$, $P(-C) = \frac{3}{10}$.

What is the probability the object is a black cube? This means the **joint** probability of it being black **and** a cube:

$$P(B \text{ and } C) = \frac{n(B \text{ and } C)}{n(S)} = \frac{4}{10} = \frac{2}{5}$$

Note that since it is logically the same to be black and a cube (B and C) as it is to be a cube and black (C and B), we have that it is always true that

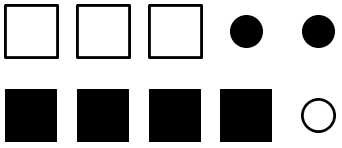
$$P(C \text{ and } B) = P(B \text{ and } C)$$

for any two events B, C. We will next develop a rule for evaluating joint probability, but first we must introduce the concept of conditional probability.

If we **know** that the object we have drawn is a cube, what is the probability that it is black? Such a probability is known as a **conditional probability** and we can write it in shorthand as:

$$P(B|C) = \text{Probability Black Given Cube}$$

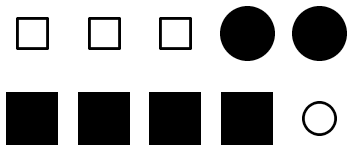
Notice the event that is known, here that the object is a cube (C), is placed second. The vertical bar " | " should be read as "*given*". To calculate the probability, observe that in this case the sample space of events is no longer all the objects, rather the sample space is restricted to the cubes (enlarged):



so that we have:

$$P(B|C) = \frac{n(B \text{ and } C)}{n(C)} = \frac{4}{7}$$

What is the probability that the object is a cube if we know it is black? This is the conditional probability of C given B, $P(C|B)$. Now the colour is known so the sample space is restricted to the black objects (enlarged):



$$P(C|B) = \frac{n(B \text{ and } C)}{n(B)} = \frac{4}{6} = \frac{2}{3}$$

Comparing $P(C|B)$ to $P(B|C)$ we see that unlike joint probability (and), it is **not** true that $P(C|B) = P(B|C)$. Furthermore note that neither of these equal $P(B \text{ and } C)$.

2.5.2 Multiplication Rule (General)

In general, for two events A and B, we have, as just shown, that

$$P(B|A) = \frac{n(A \text{ and } B)}{n(A)}$$

By dividing the numerator and denominator on the right hand side each by the total number of events in the sample space, $n(S)$ we have

$$P(B|A) = \frac{\frac{n(A \text{ and } B)}{n(S)}}{\frac{n(A)}{n(S)}} = \frac{P(A \text{ and } B)}{P(A)}$$

A useful formula for conditional probability is therefore:

$$\boxed{P(B|A) = \frac{P(A \text{ and } B)}{P(A)}}$$

Solving for the joint probability (and) we get the **multiplication rule**:

$$\boxed{P(A \text{ and } B) = P(A) \cdot P(B|A)}$$

Example:
Verify that our joint probability result $P(C \text{ and } B) = \frac{2}{5}$ can be obtained using the conditional probability formula.

Solution:

$$P(C \text{ and } B) = P(C) \cdot P(B|C) = \frac{7}{10} \cdot \frac{4}{7} = \frac{4}{10} = \frac{2}{5}$$

Since $P(C \text{ and } B) = P(B \text{ and } C)$ we could have also used $P(C|B)$ instead to get the same answer:

$$P(B \text{ and } C) = P(B) \cdot P(C|B) = \frac{3}{5} \cdot \frac{2}{3} = \frac{2}{5}$$

2.5.3 Multiplication Rule (Independent Events)

Our general multiplication rule has a special case which can be used if the two events A, B are **independent**. Two events are independent if the probability of one happening is not changed by the occurrence of the other (or vice versa). This is the case when two coins are flipped – the probability of getting heads on one is not affected by the result (heads or tails) that occurred on the other. The individual coin flip events are independent. In symbols we can define two events A, B to be independent if $P(B|A) = P(B)$. This makes sense because in words it means that the probability of B occurring did not change if A had occurred. In the special case of independent events we see, using this definition in the general multiplication rule, that:

$$\boxed{P(A \text{ and } B) = P(A) \cdot P(B) \quad (A, B \text{ independent}) \ .}$$

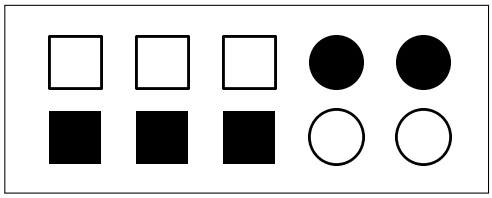
If two events are not independent we say they are **dependent**. For dependent events, or in the case where one does not know whether the events are independent or not, the general multiplication rule must be used.

To test whether two events are independent we can test the conditional probability expression above, $P(B|A) = P(B)$, or equivalently we can check if $P(A|B) = P(A)$. Finally if we can show that the multiplication rule for independent events actually holds this also proves independence.[8] In summary

$$\boxed{\text{Events } A, B \text{ are } \textbf{independent} \text{ if and only if } \begin{cases} P(B|A) = P(B), \ \textit{or} \\ P(A|B) = P(A), \ \textit{or} \\ P(A \text{ and } B) = P(A) \cdot P(B) \end{cases} .}$$

Example:

1. Are the events of drawing a black object (B) and drawing a cube (C) from the previous example dependent or independent?
2. Repeat the last question if the original set of objects had been:



Solution:

1. We saw above for the original box of objects that $P(B) = 3/5$ while $P(B|C) = 4/7$. Thus $P(B) \neq P(B|C)$ so the events B, C by the first definition are not independent, and hence are dependent. (Try testing the other two definitions of independence and show they also fail.)
2. With the new objects lets check the third definition of independence. We have

$$P(B \text{ and } C) = \frac{n(B \text{ and } C)}{n(S)} = \frac{3}{10} \ ,$$

[8] Why? If $P(A \text{ and } B) = P(A) \cdot P(B)$, then $P(B) = \frac{P(A \text{ and } B)}{P(A)} = P(B|A)$ where the last equality holds as we saw earlier for conditional probability. Hence $P(B|A) = P(B)$ and the events are independent.

while

$$P(B) \cdot P(C) = \frac{5}{10} \cdot \frac{6}{10} = \frac{30}{100} = \frac{3}{10} .$$

Hence $P(B \text{ and } C) = P(B) \cdot P(C)$ and the third definition proves the two events B, C are now actually independent. (Try testing the other two definitions of independence and show they also hold.)

The last example shows a case where two events are independent by definition which perhaps was not obvious. Typically we can recognize two events as not having an effect on each other's likelihood and hence being independent so that we can use the simplified multiplication rule. Unlike mutually exclusivity, independence of events cannot be illustrated by a Venn diagram. Tree diagrams are helpful for sequencing the events when applying the multiplication rules as the next example shows. Finally note that we can generalize the special case multiplication rule for 3 (or more) events that are all independent as:

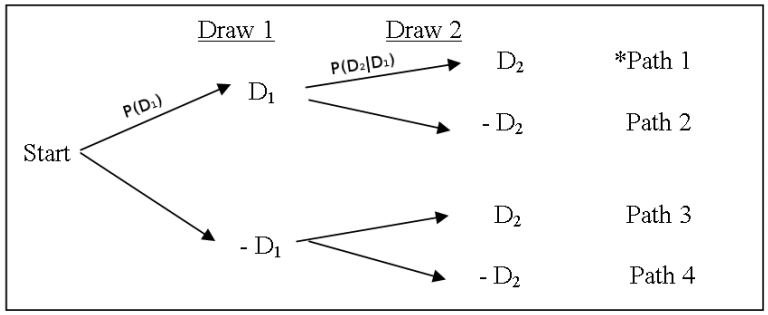
$$P(A \text{ and } B \text{ and } C) = P(A) \cdot P(B) \cdot P(C) \quad (A, B, C \text{ independent}).$$

The general multiplication rule can also be generalized to more than two events as shown in the examples below.

Example:
A box holds 20 items. 5 of the items are defective.

1. If an item is drawn from the box and **not replaced** before the next is drawn, what is the probability of getting two defective items?

   The following tree diagram illustrates the experiment with its possible outcomes being each of the four paths.



   Selecting two defective items corresponds to Path 1 (D_1 and D_2). Since the draws are being done **without replacement**, each time a draw is made the proportion of defective items in the box are dependent upon what was drawn on previous draws. This is the case of **dependent** events. Using the general Multiplication Rule:

$$\boxed{P(D_1 \text{ and } D_2)} = P(D_1) \cdot P(D_2|D_1)$$

$$= \frac{5}{20} \cdot \frac{4}{19} = \frac{20}{380} = \boxed{\frac{1}{19}}$$

   Notice on the tree diagram that we can associate probabilities with each branch, those of Path 1 having been labelled. $P(D_2|D_1)$ is associated with the second branch because at that point we know a defective item has been drawn on the first draw. (D_1 has occurred.) The entire Path 1 corresponds to the event (D_1 and D_2) and its probability is just the product of the probabilities of the branches of which it is composed.

2. If an item is drawn from the box and then **replaced** before the next item is drawn, what is the probability of selecting two defective items?

   The path of interest in the tree remains the same. Since the items are drawn **with replacement** each time a draw is made the proportion of defective items in the box remains the same and consequently the probability of drawing a defective item on the second draw is the same as on the first. This is the case of **independent events**. Using the Multiplication Rule for independent events.

$$\boxed{P(D_1 \text{ and } D_2)} = P(D_1) \cdot P(D_2) = \frac{5}{20} \cdot \frac{5}{20} = \frac{25}{400} = \boxed{\frac{1}{16}}$$

   Notice that this joint event is slightly more likely to occur when the items are drawn with replacement as there are more defective items available on the second draw in this case.

3. If two items are drawn from the box **at once**, what is the probability that exactly one is defective?

   Drawing two at once is equivalent to drawing two items without replacement because once the first item is chosen it is not available to be picked as the second item. The second and third paths from the top both represent the occurrence of exactly one defective item. So the event could happen by path two **or** path three. There individual probabilities are:

$$P(\text{Path 2}) = P(D_1 \text{ and } \text{-}D_2) = P(D_1) \cdot P(\text{-}D_2|D_1) = \frac{5}{20} \cdot \frac{15}{19} = \frac{15}{76}$$

$$P(\text{Path 3}) = P(\text{-}D_1 \text{ and } D_2) = P(\text{-}D_1) \cdot P(D_2|\text{-}D_1) = \frac{15}{20} \cdot \frac{5}{19} = \frac{15}{76}$$

   Paths on a tree diagram are mutually exclusive outcomes so using the M.E. addition rule:

$$\begin{aligned} \boxed{P(\text{exactly one defective})} = P(\text{Path 2 or Path 3}) &= P(\text{Path 2}) + P(\text{Path 3}) \\ &= \frac{15}{76} + \frac{15}{76} = \frac{30}{76} = \boxed{\frac{15}{38}} \end{aligned}$$

4. If three items are drawn **without replacement**, what is the probability that all three are defective?

   Imagine a tree diagram with three steps. This joint event corresponds to one path in the tree where a defective item is picked at each each step. The steps in this case are dependent since we are picking without replacement. The general multiplication rule for two events generalizes to three events as follows:[9]

$$\boxed{P(D_1 \text{ and } D_2 \text{ and } D_3)} = P(D_1) \cdot P(D_2|D_1) \cdot P(D_3|D_1 \text{ and } D_2) = \frac{5}{20} \cdot \frac{4}{19} \cdot \frac{3}{18} = \boxed{\frac{1}{114}}$$

5. If three items are drawn **with replacement**, what is the probability that all three items are defective?

   This is represented by the same path in the tree as in the last example. The steps are independent since we are drawing with replacement.

   The multiplication rule for two independent events generalizes to three events as follows:[10]

$$\boxed{P(D_1 \text{ and } D_2 \text{ and } D_3)} = P(D_1) \cdot P(D_2) \cdot P(D_3) = \frac{5}{20} \cdot \frac{5}{20} \cdot \frac{5}{20} = \boxed{\frac{1}{64}}$$

6. If three items are drawn **with replacement**, what is the probability of at least one good item?

   In this case we could calculate the probability of all the paths with at least one not defective (-D) item, which is seven paths! The quick way is to notice this event is the complement of the event that all are defective (last example) and use the complement rule.

$$\boxed{P(\text{at least one not defective})} = P(\text{-}[\underbrace{D_1 \text{ and } D_2 \text{ and } D_3}_{A}]) = 1 - P(\underbrace{D_1 \text{ and } D_2 \text{ and } D_3}_{A})$$

$$= 1 - \frac{1}{64} = \boxed{\frac{63}{64}}$$

Assignment:

1. In the following experiments, identify the steps in the sequence and determine if the steps are dependent or independent.
   (a) The outcome is observed on each occasion when a coin is flipped 3 times.
   (b) A carton of 40 bolts has 4 that are defective. A handful of three bolts are randomly selected and inspected for defects.
   (c) An opinion poll is conducted in a large shopping centre. Five people are selected at random and asked their opinion about the type of service they prefer in a restaurant.
2. A shipment of eggs contains 25 cartons of eggs. 6 of the cartons contain cracked eggs. A customer buys some of the cartons without inspecting the contents. What is the probability that 2 of the cartons contain cracked eggs if:
   (a) The customer buys two cartons.
   (b) The customer buys three cartons.
3. A card is dealt from a standard deck of 52 playing cards.
   (a) What is the probability that the card is black and a six?
   (b) Are the events of drawing a black card and drawing a six independent?
4. Of the 20 members who attended an organizational meeting, 5 are women. A secretary and a treasurer are to be chosen from the group by a random draw.
   (a) If the same person can hold both offices, what is the probability that both offices are held by women?
   (b) If the same person cannot hold both offices, what is the probability that both offices are held by women?
5. In the example above, that starts on page 110, repeat question 6 but now assume the three items are drawn **without replacement**.

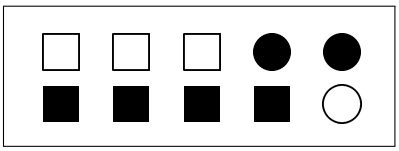
[9] $P(\underbrace{D_1 \text{ and } D_2}_{A} \text{ and } \underbrace{D_3}_{B}) = P(\underbrace{D_1 \text{ and } D_2}_{A}) \cdot P(\underbrace{D_3}_{B} \mid \underbrace{D_1 \text{ and } D_2}_{A}) = P(D_1) \cdot P(D_2|D_1) \cdot P(D_3|D_1 \text{ and } D_2)$

[10] $P(\underbrace{D_1 \text{ and } D_2}_{A} \text{ and } \underbrace{D_3}_{B}) = P(\underbrace{D_1 \text{ and } D_2}_{A}) \cdot P(\underbrace{D_3}_{B}) = P(D_1) \cdot P(D_2) \cdot P(D_3)$ (independent events)

2.6 Using Contingency Tables

2.6.1 Frequency Contingency Tables

Contingency tables can be used to represent certain frequency distributions when outcomes involve multiple events, as shown in Section 2.2. Similarly they are useful for the frequency of occurrence of events in theoretical probability calculations as shown in Section 2.4. The notion of conditional probability and the multiplication rules can often be made clearer by constructing a contingency table of the observable outcomes. For example, we can represent our probability problem of Section 2.5 involving the random selection of a single item from among black and white cubes and spheres



with a contingency table:

		Shape		
		C	$\text{-}C$	
Colour	B	4	2	6
	$\text{-}B$	3	1	4
		7	3	10

		Shape		
		C	$\text{-}C$	
Colour	B	$n(B \text{ and } C)$	$n(B \text{ and } \text{-}C)$	$n(B)$
	$\text{-}B$	$n(\text{-}B \text{ and } C)$	$n(\text{-}B \text{ and } \text{-}C)$	$n(\text{-}B)$
		$n(C)$	$n(\text{-}C)$	$n(S)$

On the right are shown the symbols corresponding to the table entries.

Joint probabilities, such as finding the probability of a white cube ($\text{-}B$ and C), are easily interpreted, since we restrict ourselves to the row and column of the events in question. We can read the answer directly from the table because within the table are the joint frequencies of all combinations of pairs of outcomes.

	C	$\text{-}C$	
B	4	2	6
$\text{-}B$	[3]	1	4
	7	3	[10]

$$P(\text{white cube}) = P(\text{-}B \text{ and } C) = \frac{n(\text{-}B \text{ and } C)}{n(S)} = \frac{3}{10}$$

Similarly interpreting a conditional probability is straightforward. The known constraint restricts consideration to its row or column. For instance, what is the probability of getting a cube if you know the picked object is black, $P(C|B)$?

	C	$\text{-}C$	
B	[4]	2	[6]
$\text{-}B$	3	1	4
	7	3	10

$$P(\text{cube given black}) = P(C|B) = \frac{n(B \text{ and } C)}{n(B)} = \frac{4}{6} = \frac{2}{3}$$

Using the contingency table, it is possible to see exactly what it means to say that the sample space is limited to the condition under conditional probability. Note that the denominator in the proportion is reduced to 6 and does not contain the entire group of objects observed since not all 10 objects are black.

Example:
The local humane society currently has 65 dogs and 89 cats. 33 of the dogs and 46 of the cats required medical treatment when they were admitted. Use a **frequency** contingency table in answering the following questions. If an animal is selected at random, what is the probability:

1. It did not require medical treatment upon its arrival? (Answer: 0.4870)
2. It is a dog that did not require treatment? (Answer: 0.2078)
3. It is a cat, if it is known the animal had been given treatment? (Answer: 0.5823)

Use D for the event the animal is a dog, -D for cat, and T for the event the animal required treatment.

Solution:

	T	-T	
D			
-D			

1. $P\left(\quad\right) = \dfrac{n\left(\quad\right)}{n\left(\quad\right)} = \dfrac{\quad}{\quad} =$
2. $P\left(\quad\right) = \dfrac{n\left(\quad\right)}{n\left(\quad\right)} = \dfrac{\quad}{\quad} =$
3. $P\left(\quad\right) = \dfrac{n\left(\quad\right)}{n\left(\quad\right)} = \dfrac{\quad}{\quad} =$

2.6.2 Probability Contingency Tables

Up to this point we have used frequency contingency tables whose entries are either experimental frequencies or theoretical counts of possible outcomes. We can create a **probability contingency table** by dividing every entry in an experimental frequency contingency table by the sum of frequencies, $\sum f$, or by dividing every entry in a theoretical probability calculation by the number of elements in the sample space, $n(S)$. As an example of the second case, divide every entry in the contingency table of the last section by $n(S) = 10$ and use $P(A) = n(A)/n(S)$ to get:

		Shape		
		C	$\text{-}C$	
Colour	B	$\frac{4}{10}$	$\frac{2}{10}$	$\frac{6}{10}$
	$\text{-}B$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{4}{10}$
		$\frac{7}{10}$	$\frac{3}{10}$	1

		Shape		
		C	$\text{-}C$	
Colour	B	$P(B \text{ and } C)$	$P(B \text{ and } \text{-}C)$	$P(B)$
	$\text{-}B$	$P(\text{-}B \text{ and } C)$	$P(\text{-}B \text{ and } \text{-}C)$	$P(\text{-}B)$
		$P(C)$	$P(\text{-}C)$	1

On the right are shown the symbols corresponding to the table entries. Note the following:

1. Joint probabilities lie inside the table. For example $P(\text{black sphere}) = P(B \text{ and } \text{-}C) = \frac{2}{10}$. Recall that $P(B \text{ and } \text{-}C) = P(\text{-}C \text{ and } B)$ for a joint probability.
2. Marginal probabilities lie on the sides of the table. For example $P(\text{cube}) = P(C) = \frac{7}{10}$.
3. The bottom right entry for a probability contingency table is always 1 since the total probability is 1.
4. All rows and columns add up as usual for a contingency table.
5. Conditional probabilities **do not appear** within a probability contingency table. To calculate a conditional probability one uses the inverted general multiplication rule:

$$\boxed{P(B|A) = \frac{P(A \text{ and } B)}{P(A)}}$$

   and reads the probabilities on the right hand side from the contingency table. For instance, using the above contingency table, the probability of getting a cube given the object is black, $P(C|B)$ is:

$$P(C|B) = \frac{P(B \text{ and } C)}{P(B)} = \frac{\frac{4}{10}}{\frac{6}{10}} = \frac{4}{10} \cdot \frac{10}{6} = \frac{4}{6} = \frac{2}{3}$$

   Similarly you may often be given a conditional probability which can be converted to a table entry using the multiplication rule. For instance if we did not know the contents of our bag of objects but rather had been told that the probability of picking a black object ($P(B)$) was 6/10 and that the 2/3 of the black objects are cubes ($P(C|B)$) then we could have calculated that

$$P(B \text{ and } C) = P(B) \cdot P(C|B) = \frac{6}{10} \cdot \frac{2}{3} = \frac{4}{10} = \frac{2}{5} .$$

   The joint probability $\frac{2}{5}$ could then have been put in the contingency table.
6. In general, to use a contingency table to help with a problem, first complete as many entries as possible in the contingency table using the information given. Next use the mathematical properties of the contingency table to complete the remaining entries. Finally use the contingency table entries to determine the answers of what was requested.

Example:
A statistics class contains 55% first-year students, 30% second-year students, while the remainder are third-year students. It is known that 60% of the first-year students, 50% of the second-year students and 80% of the third-year students have majors in the faculty of arts. Use a **probability** contingency table to aid answering the following questions. What is the probability a randomly chosen student in the class

1. Is an arts major? (Answer: 0.60)
2. Is a third-year student who is not an arts major? (Answer: 0.03)
3. Is a second-year student if it is known the student is an arts major? (Answer: 0.25)

Use A for the event the student is an arts major and F, S, and T for the events a student is first, second or third-year respectively. (Note that conversion of probabilities from percent to decimal form must be done before they may be used in the multiplication rule or its inverted form for conditional probability.)

Solution:

	F	S	T	
A				
-A				
				1.0

1. $P\left(\quad\right) =$
2. $P\left(\quad\right) =$
3. $P\left(\quad\right) =$

Are the events of being a second-year student (S) and being an arts major (A) independent? Justify your answer.

Assignment:

1. Parts are allocated to bins in a warehouse by people on two different shifts. Shift one allocates 2000 parts while shift two allocates 3500 parts. 200 of the parts allocated by shift one are placed in the wrong bins. 250 of the parts allocated by shift two are placed in the wrong bins. Suppose a part is randomly selected from those allocated by both shifts. What is the probability that it was:

   (a) In the correct bin?

   (b) Allocated by shift one and was in the correct bin?

   (c) Allocated by shift two and was in the wrong bin?

   (d) In the wrong bin given that it was allocated by shift one?

   (e) Allocated by shift two given it was in the correct bin?

   Use a relative frequency contingency table in your answer.

2. In a residential neighbourhood 25% of houses have a cat, the remainder do not. In those houses which have a cat, the likelihood that someone in the house has an allergy is 20% . In those houses which do not have a cat, the likelihood of someone in the house having an allergy is 40% . Answer the following questions using a probability contingency table to aid you. For symbols, let C be the event a house contains a cat, and A the event that someone in a house has an allergy.

   (a) What is the probability a house contains a cat and someone with an allergy?

   (b) What is the probability a house does not contain a cat and has someone with an allergy?

   (c) What is the probability a house contains no one with an allergy?

   (d) A house is selected at random and is found to have a person with allergies. What is the probability that the house has a cat?

   Use a probability contingency table in your answer.

3. A marble is to be randomly drawn out of one of three urns. Urn A has 10 black and 10 white marbles. Urn B has 5 white and 15 black marbles. Urn C has 15 white and 5 black marbles. The urn is to be chosen by flipping two coins. If the coins are both heads, urn A is chosen. If the coins are both tails urn B is chosen. If the coins are of unlike sides, urn C is chosen.

   (a) What proportion of the time will urn B be chosen?

   (b) What proportion of the time will a white marble be chosen?

   (c) Given that a white marble was chosen what is the probability that it came from urn B?

   ** Hint: To tackle this fairly complex problem first construct a probability contingency table with the Urn (A, B, or C) on one axis and marble colour (W = white and -W = black) on the other axis. Systematically complete the probabilities in the table using the information given and only then answer the questions. **

   The bend-in-the-road symbol,  , will be used to denote certain problems that are more challenging. While these problems will not be typical of an exam question, puzzling them out will help solidify concepts that have been introduced.

2.7 Review Exercises on Basic Probability

Use proper probability symbols to solve the following questions.

1. An analysis of a collection of loans made to 28 agriculture industries, (A), and 32 food stores (F), revealed the following loan assessments. (low risk loans = L)

		LOAN ASSESSMENT	
		Low Risk	Moderate to High Risk
Industry	Agriculture	6	22
	Food Retailers	15	17

   Interpret the meaning and give the value of the following probability symbols: (a) $P(A)$ (b) $P(\text{-}A)$ (c) $P(A \text{ and } L)$ (d) $P(A|L)$ (e) $P(\text{-}A \text{ and } -L)$ (f) $P(L|A)$ (g) $P(A|\text{-}L)$

2. A card is drawn from a deck of 52 playing cards. What is the probability that the card is black or a six?

3. A card is drawn from a deck and then replaced. The deck is shuffled and another card is drawn.

   (a) What is the probability that both cards are face cards? (i.e., Jack, Queen, King)

   (b) If both cards are drawn at the same time, what is the probability that both cards are face cards?

4. If 3 cards are drawn from a deck, what is the probability that they are all face cards?

5. A card is drawn from a deck, what is the probability that it is a face card given that it is a spade?

6. In a recent survey of merchants in the downtown core, 80% favoured greater powers of arrest be given to police, 60% favoured a curfew for persons under 16 years of age, 50% favoured both proposals.

   (a) What percent of the merchants favoured at least one of the proposals?

   (b) What percent of the merchants favoured none of the proposals?

   Hint: Draw a Venn diagram.

7. Transport Canada estimates that seat belts were worn by drivers in 60% of all reported accidents. In 8% of the accidents where seat belts were worn, the accident was fatal. In 20% of the accidents where seat belts were not worn the accident was fatal.

   (a) What proportion of all reported accidents involved drivers who were wearing seat belts and were killed?

   (b) What proportion of all reported accidents involved drivers who were not wearing seat belts and were killed?

8. Twenty percent of home buyers who have mortgages with a certain bank default on their payments. 45% of those who default were initially given a good credit rating. 80% of those who do not default initially received a good credit rating. Based on this analysis:

(a) What is the probability that a person will default given that they were given a good credit rating initially?

(b) Comment on the validity of the initial credit rating.

Hint: Use a probability contingency table.

9. A company has two suppliers for a certain chemical that it uses in its production process. Firm X supplies 80% of the chemical while firm Y supplies the remainder. 95% of the chemical from Y is good while 85% of the chemical from X is good. A certain production run is spoiled because of poor chemical. Who most likely supplied the chemical given that it is poor?

   Hint: Use a probability contingency table.

10. A community is serviced by a major air carrier. 80% of all its flights arrive on time. On three separate occasions during the year, a person travels via this carrier.

    (a) What is the probability that the airline is late on all three occasions?

    (b) What is the probability that the airline is late on exactly one occasion?

    Hint: Draw a tree diagram.

11. To be hired by a company, applicants granted interviews must pass two separate tests. Based on experience, only 80% of those interviewed pass the first test. To take the second test an applicant must pass the first test. Only 30% of those who take the second test pass it.

    (a) What proportion of the interviewees are hired?

    (b) What proportion of the people not hired failed test two?

    Hint: Draw a tree diagram.

Basic Probability Formulae

Counting Rules

${}_n P_r = \dfrac{n!}{(n-r)!}$	n^r	${}_nC_r = \dfrac{n!}{r!(n-r)!}$

Probability Rules

$P(A) = \dfrac{n(A)}{n(S)}$	$0 \leq P(A) \leq 1$	$P(A) + P(\text{-}A) = 1$
$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$		$P(A \text{ and } B) = P(A) \cdot P(B\|A)$
$P(A \text{ or } B) = P(A) + P(B)$		$P(A \text{ and } B) = P(A) \cdot P(B)$
	$P(B\|A) = \dfrac{P(A \text{ and } B)}{P(A)}$	

Unit 3: Probability Distributions

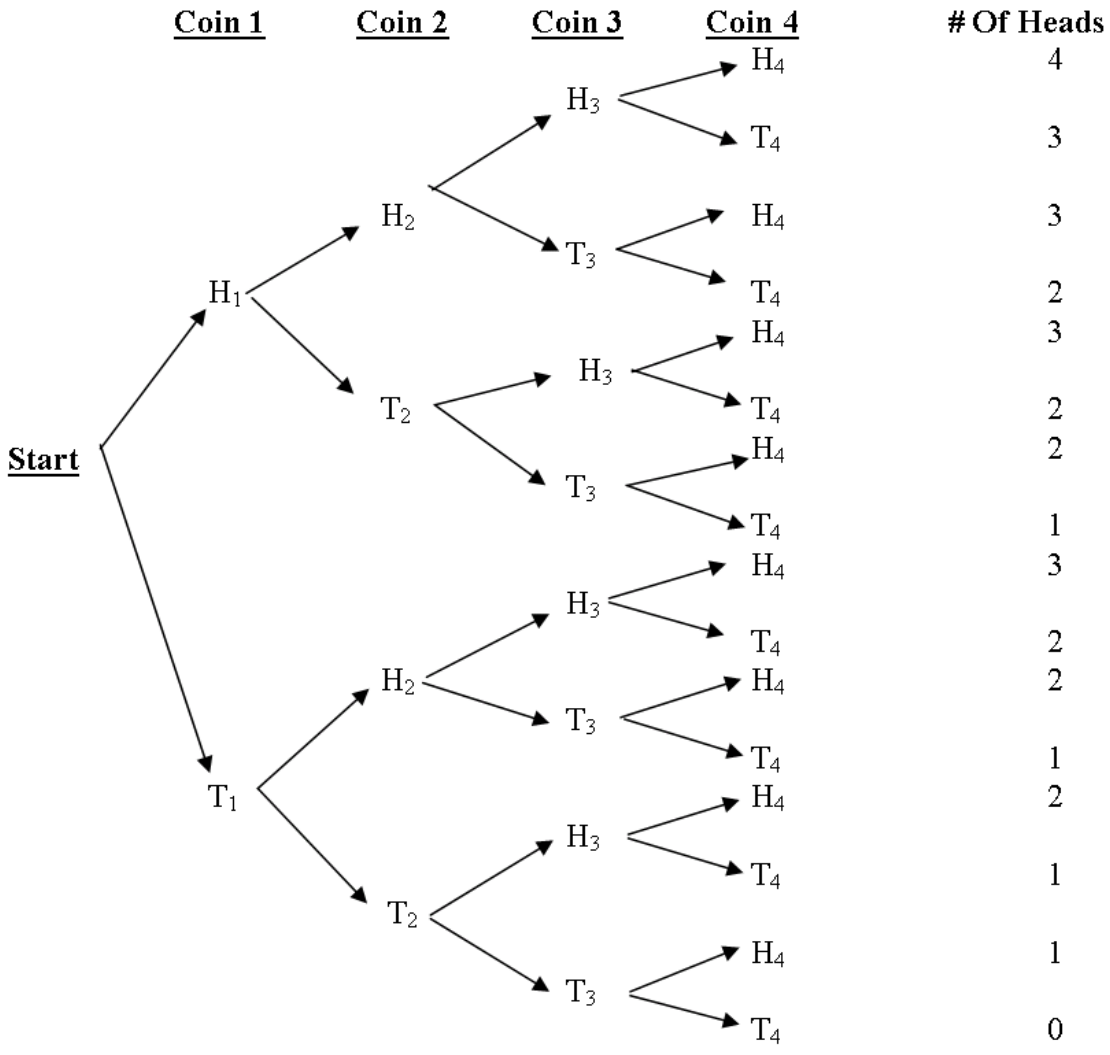
3.1 Discrete Probability Distributions

3.1.1 Definition

Probability distributions give the probabilities of all possible values (outcomes) of a random variable. If the observations result from a count (or the possible outcomes are countable) the distribution is called a **discrete probability distribution**. A discrete probability distribution lists all possible values of the random variable together with their probability of occurrence.

Example:
4 coins are tossed. Construct a probability distribution of the number of heads observed among the 4 coins.

Here is a tree diagram with all the equally likely possibilities:



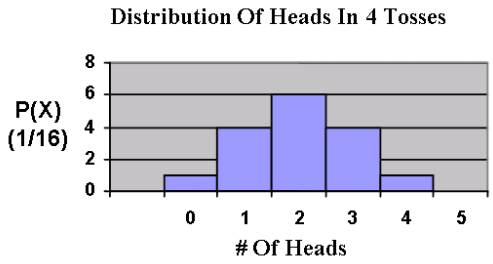
Here is the discrete probability distribution. Because each path is equally likely, divide the number of outcomes that make up an event by 16 to find its probability.

x(heads)	P(x)
0	1/16
1	4/16
2	6/16
3	4/16
4	1/16
	$\sum P(x) = 16/16 = 1$

Here we are writing $P(x)$ to indicate the probability P depends upon the discrete value x.

In Unit 1 of this course we learned that every statistical distribution has three important characteristics. Probability distributions are theoretical distributions with these same features.

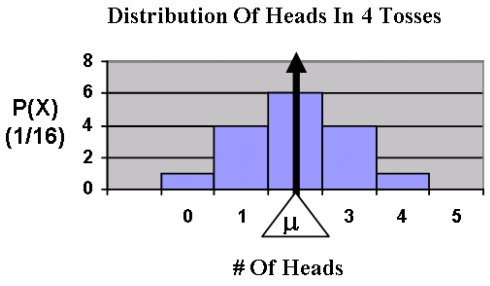
1. **Shape**



   The distribution of number of heads has a symmetric shape. Compare the theoretical graph to your experimental result in Section 2.1.2 .

2. **Centre**

   The centre of a probability distribution is called its **expected value**. The expected value is located at the balance point of the distribution.



   The formula for finding the mean (or expected) value is:[1]

$$\boxed{\mu = E(x) = \sum xP}$$

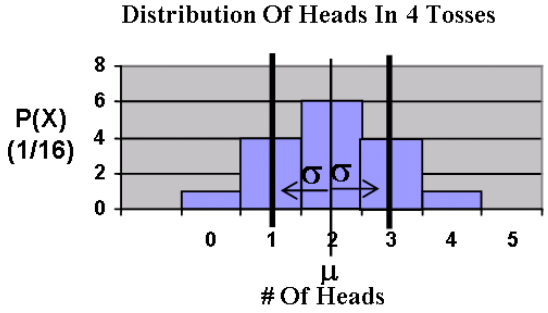
[1] In Unit 1 we could have written the formula for μ in terms of relative frequency P:

$$\mu = \frac{\sum xf}{\sum f} = \left(\frac{1}{\sum f}\right)\sum xf = \sum\left(\frac{1}{\sum f}\right)xf = \sum x\left(\frac{f}{\sum f}\right) = \sum xP \quad .$$

Since relative frequency is just experimental probability, the formula has to be the same. A similar proof can be made for the formula for σ below. (See problem 5 on page 125 .)

3. **Variability**

   The variance of a distribution is defined as its expected squared deviation from the mean and the square root of the variance is the standard deviation. The standard deviation measures the distance from centre in which the majority of possible outcomes are located.



The formulae for finding the variance and standard deviation in a probability distribution are:

$$\sigma^2 = \sum x^2 P - \left(\sum xP\right)^2 \quad \leftarrow \text{ Distribution Variance}$$

$$\boxed{\sigma = \sqrt{\sum x^2 P - \left(\sum xP\right)^2}} \quad \leftarrow \text{ Distribution Standard Deviation}$$

Example:
Here are the calculations done in a tabular format for the mean and standard deviation on the coin flipping example

x (heads)	$P(x)$	xP (hds)	x^2 (hds^2)	x^2P (hds^2)
0	1/16	0/16	0	0/16
1	4/16	4/16	1	4/16
2	6/16	12/16	4	24/16
3	4/16	12/16	9	36/16
4	1/16	4/16	16	16/16
	$\sum P(x) = 16/16 = 1$	$\sum xP = 32/16 = 2$		$\sum x^2P = 80/16 = 5$

Calculate the expected (mean) number of heads: $\mu = E(x) = \sum xP = 2.0$ heads

Calculate the standard deviation: $\sigma = \sqrt{\sum x^2 P - \left(\sum xP\right)^2} = \sqrt{5 - 2^2} = \sqrt{1} = 1.0$ head

Refer back to Section 2.1.2 where we estimated this probability experimentally. Notice that when the theoretical results are compared to what was observed experimentally, the theoretical parameters closely match the experimental statistics. We say that the probability distribution models the physical situation. On the basis of the probability distribution, predictions can be made about a physical situation.

Calculator Note

Some modern calculators allow you to calculate means and standard deviations of discrete probability distributions. Try entering your data like you would for a frequency distribution but in place of the frequency enter the probability associated with the value.

Assignment:

1. A shoe store owner has kept track of the proportion of times a customer at the store buys x pairs of shoes.

x	0	1	2	3	4
$P(x)$	0.17	0.56	0.20	0.05	0.02

   (a) If this pattern continues to hold, what is the probability the next customer will buy anything?

   (b) How many pairs of shoes can the next customer to the store be expected to buy?

   (c) What is the distribution standard deviation?

   (d) Plot the histogram of the probability distribution. What shape is it?

2. A drive-through restaurant owner maintains records of the proportion of vehicles containing a given number of occupants.

x(occupants)	Proportion of Vehicles
1	0.05
2	0.45
3	0.02
4	0.30
5	0.01
6	0.17

   (a) Is this a probability distribution? How can we verify this? What is the significance of the sum of the $P(x)$ column?

   (b) What is the mean of this distribution?

   (c) What is the standard deviation of this distribution?

   (d) If in the future 100 cars pass through the drive-through, how many people can we expect to have passed through?

3. A businessperson wishes to start a business venture. The person estimates that there is a 60% chance of making a \$150,000 profit, a 30% chance of earning an \$80,000 profit and a 10% chance of losing \$50,000 on the venture. What is the expected profit from this venture? (Note that a loss will need to be indicated by a negative value of the variable in your probability distribution.)

4. A game is played with a 10-sided die. 5 of the sides are red, 4 are green and 1 is blue. If a blue turns up the player wins \$10. If a red turns up the player wins \$1.

   (a) If the operator charges \$5 if a green turns up, how much will the operator gain per play in the long run?

   (b) What should the player pay if a green turns up in order that this be a fair game? (i.e. one in which the expected value over the long run is \$0)

   Hint: To do this problem set up a probability distribution. For part (b) introduce a variable for what the player must pay and solve.

5. Starting with the formula for the population standard deviation of a frequency distribution (Section 1.13.2), derive the probability distribution standard deviation formula in a manner similar to what was done to prove the mean formula in footnote 1 on page 123 .

3.1.2 The Binomial Distribution

Definition

A **binomial (or Bernoulli)** experiment satisfies the following criteria:

1. The **experiment** consists of a **fixed sequence of trials**.
2. Each **trial** results in either a **success** or a **failure**. A success means: "What we were looking for happened".
3. The **probability of a success in each trial** is **constant**.
4. The variable $\boldsymbol{x}$ is the discrete **number of successes** in the sequence of trials.
5. The **trials** in the sequence are **independent**.

Note: These criteria are equivalent to a sequence using replacement sampling.

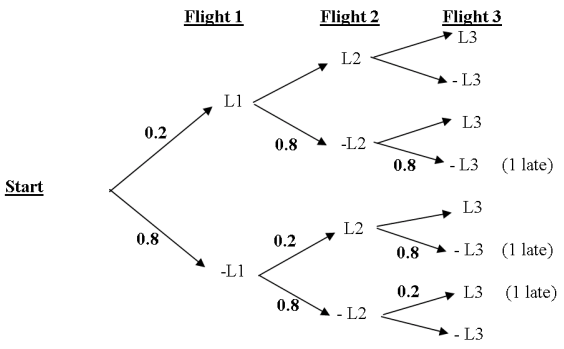
When these conditions exist, the resulting probabilities for x form a discrete distribution called the **binomial probability distribution**. A mathematical formula exists to calculate these probabilities.

Example:
Verify that these conditions hold for problem 10 of Section 2.7 .

Solution:
Recall: A community is serviced by a major air carrier. 80% of all its flights arrive on time. On three separate occasions during the year, a person travels via this carrier. Find the probability that the airliner is late on exactly 1 occasion.

1. The experiment has 3 trials (=flights) in it.
2. Each flight is a success (=late) or a failure (=on time)
3. The probability of success on a trial (a late flight) is fixed at 0.2 .
4. The variable is the number of successes (late flights) among three flights and that is discrete.
5. Whether any flight is late or not does not depend on previous flights so the trials are independent.

This situation satisfies all five conditions so it is called a **binomial experiment**. We will use this example to show how the mathematical equation is developed to compute binomial probabilities. The tree diagram for this experiment shows that not all outcomes for the 3 flights are equally likely.



On a path with 1 late, a joint sequence of 3 things happens. Each of these paths has 1 late flight and 2 not late flights. To find the probability of any of these paths, we must use the multiplication rule for independent events. $P(\text{Path 4}) = (.2) \cdot (.8) \cdot (.8)$ and $P(\text{Path 6}) = (.8) \cdot (.2) \cdot (.8)$ and $P(\text{Path 7}) = (.8) \cdot (.8) \cdot (.2)$. Notice that all of these calculations have one factor of .2 and two of .8 so they produce the same value on multiplication, $(0.2)^1 \cdot (0.8)^2$. Adding the probabilities of the mutually exclusive paths gives

$$\begin{aligned} \boxed{P(\text{exactly 1 late})} &= P(\text{Path 4 or Path 6 or Path 7}) \\ &= P(\text{Path 4}) + P(\text{Path 6}) + P(\text{Path 7}) \\ &= \underbrace{(3)}_{\substack{\boxed{\text{Number of paths with 1 late}}}} \cdot \underbrace{(0.2)^1 \cdot (0.8)^2}_{\substack{\boxed{\text{Probability of path with 1 late}}}} = \boxed{0.384} \end{aligned}$$

We can generalize this into the **binomial probability function**:

The probability of **x successes** among a sequence of **n independent trials** where the **probability of a success on a trial** is $\boldsymbol{\pi}$ and the probability of a failure is $\mathbf{1 - \pi}$ can be calculated as:[2]

$$\boxed{P(x) = {}_nC_x \cdot \pi^x \cdot (1 - \pi)^{n-x}}$$

You may have encountered the symbol π in previous mathematics courses with respect to circle calculations where π was about 3.14. π is a lower case Greek letter roughly equivalent to the letter p in our alphabet. We will use π to represent a parameter, the population proportion.

Example:
In the experiment of observing the number of heads in the toss of 4 coins, verify that the binomial probability rule can be used to calculate the probabilities. Calculate the probability of observing 2 heads in the flip of 4 coins.

Solution:
This is a binomial experiment because 1) We have a sequence of four trials (coin flips), 2) Each coin turns up in one of two states, heads (success) or tails (failure), 3) The probability of a success remains fixed at 0.5, 4) The number of successes is what is being counted, and 5) The outcome on each coin is independent from the outcomes on the other coins.

Because this is a binomial experiment we can use the binomial probability function to do the calculation. Substitute $n = 4$, $\pi = 0.5$, $1 - \pi = 0.5$, and $x = 2$.

$$P(2) = {}_4C_2 \cdot 0.5^2 \cdot (1 - 0.5)^{4-2} = 0.375$$

We have calculated the value for $x = 2$, but x can have any integer value from 0 to 4. If we applied this equation to the other x values, we would have the probability distribution for this experiment without resorting to drawing the tree diagram or applying the basic probability rules.

[2]The combination ${}_nC_x$ counts the number of paths with exactly x successes. To see this note that identifying a specific path with x successes requires stating which x of the n branches have successes. This is done by selecting x numbers without replacement from $1 \ldots n$, which can be done in ${}_nC_x$ ways.

x	$P(x)$
0	0.0625
1	0.2500
2	0.3750
3	0.2500
4	0.0625
	$\sum P(x) = 1.0000$

Binomial Tables

To make it easier to work with binomial probabilities it is often convenient to tabulate the calculations once and for all for a range of values of n, π, and x. Instead of tabulating these probabilities however (as some do) we will tabulate the **cumulative** binomial probabilities. Consider the coin-flipping experiment for $n = 4$ trials above. One first selects the table corresponding to this number of trials, $n = 4$ from among all the binomial tables starting on page 274):

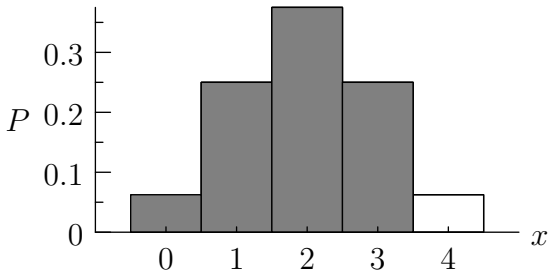
n=4

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.961	0.815	0.656	0.410	0.240	0.130	0.062	0.026	0.008	0.002	0.000	0.000	0.000
1	0.999	0.986	0.948	0.819	0.652	0.475	0.312	0.179	0.084	0.027	0.004	0.000	0.000
2	1.000	1.000	0.996	0.973	0.916	0.821	0.688	0.525	0.348	0.181	0.052	0.014	0.001
3	1.000	1.000	1.000	0.998	0.992	0.974	0.938	0.870	0.760	0.590	0.344	0.185	0.039
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Each column represents a binomial experiment with a particular probability of success. In our case we want the column of numbers with $\pi = 0.50$ (boxed). Next, the rows are indexed from k equal 0 to n. On a cumulative probability table the entry is the probability of getting an outcome between 0 and the value k (including 0 and k). So for instance for row $k = 3$ in our $\pi = 0.50$ column we see[3]

$$P(x \leq 3) = 0.938$$

Graphically the probability is the total shaded area:



where we have drawn the probability of an individual x-value as a rectangle of width 1. The cumulative binomial tables make it easy to quickly retrieve any required probability, especially for ranges of x.

[3]This is precisely what we called $<CP(3)$ when discussing cumulative relative frequency (Section 1.6.2) now written in probability notation.

Example:
For our coin flipping experiment above, find the following probabilities using the cumulative binomial table: (Answers: 1.) 0.938 2.) 0.062 3.) 0.876 4.) 0.250)

1. The probability of at most 3 heads: $P(x \leq 3) =$
2. The probability of more than three heads $P(3 < x) = 1 - P(x \leq 3) =$
3. The probability of between 1 and 3 heads (including 1 and 3):
   $P(1 \leq x \leq 3) = P(x \leq 3) - P(x \leq 0) =$
4. The probability of exactly 3 heads (i.e. what we calculate with the binomial distribution):
   $P(x = 3) = P(x \leq 3) - P(x \leq 2) =$

What would change if our coins were not fair but rather had a probability of getting heads of 60%?

Sometimes the probability tables cannot be used as demonstrated in the following example.

Example:
Compute the probability of getting exactly 3 sixes in 5 rolls of a six-sided die.

Solution:
First verify that this is a binomial experiment. In this case $n = 5$, but $\pi = 1/6$ (which is not one of the probability values listed in the table), $1 - \pi = 5/6$, and $x = 3$. Substitute these values into the binomial function:

$$P(x = 3) = {}_5C_3 \cdot \left[\frac{1}{6}\right]^3 \left[\frac{5}{6}\right]^2 = 0.0322 \leftarrow \text{Must be calculated by calculator}$$

** Notice here that at first glance this may not have seemed to be a binomial problem since each roll of the die has six possible outcomes not two. The problem becomes binomial once it is recognized that success is getting a six while failure is not getting a six. **

Binomial Parameters

In Section 3.1 we learned that once a probability distribution is constructed, the mean (expected value) and standard deviation parameters can be calculated by the formulae:

$$\mu = E(x) = \sum xP \qquad \sigma = \sqrt{\sum x^2 P - \left(\sum xP\right)^2}$$

This is a time consuming procedure especially when n is large. In the case of the binomial distribution, because there is a known function $P(x)$ for calculating individual probabilities, these formulae can be evaluated once and for all to find a formula for finding the mean and the standard deviation of a binomial distribution.

$$\boxed{\mu = E(x) = n\pi} \qquad \boxed{\sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)}}$$

Example:
Calculate the mean and standard deviation for the coin flipping experiment and compare the results with those found in Section 3.1.

Solution:
In this case $n = 4$, $\pi = 0.5$, and $1 - \pi = 0.5$, so

$$\mu = n\pi = (4) \cdot (0.5) = 2.0 \text{ heads}$$

$$\sigma = \sqrt{n \cdot \pi \cdot (1 - \pi)} = \sqrt{(4) \cdot (0.5) \cdot (0.5)} = 1.0 \text{ head}$$

These results are identical to those found previously but with considerably less effort. These shortcut formulae are only applicable for binomial probability distributions.

Large Population Limit

Imagine counting successes in a sample of size n drawn from a population **without replacement**. This would typically involve dependent events. However, if the population is very large compared to the size of the sample drawn from it, drawing the sample does not significantly change the proportion of successes within the population from one draw to the next. In this case the binomial distribution can be used to approximate probabilities with very little error since the trials are effectively independent. Statisticians judge a population to be large when the sample size n is less than 5% of the size of the population N (so $n/N < .05$).

Example:
Of the 50,000 voters in a constituency, 35,000 would vote for candidate Smith. In a random sample of 10 voters selected from the constituency what is the probability that exactly 2 voters would vote for Smith?

Solution:
The trials are really not independent because we are sampling without replacement. However $n/N = 10/50,000 = .0002$ is far less than .05 so the population is "large" compared to the sample size. In this case the probability a voter was in favour of Smith would stay very close to $\pi = 35,000/50,000 = .70 = 70\%$ on each draw regardless of the preference of the voter on the previous draw. Calculate the probability using $n = 10$, $\pi = 0.7$ and $x = 2$ with the binomial probability distribution. (Answer: $P(x = 2) = 0.0014467005 \approx 0.0014$)

The exact answer to this problem (and similar problems when sampling without replacement) requires use of another discrete probability distribution, the **hypergeometric distribution**.

Assignment:

1. A class of students wrote a 25 question true and false test. Each student guessed at the answer to the questions, what is the probability that a student achieved:
   (a) none correct
   (b) exactly 13 correct
   (c) more than 10 correct (use the binomial table).
   (d) What would the mean mark in the class be?
   (e) What would the standard deviation of the set of marks in the class be?
2. If the above test in Problem 1 had been a multiple-choice test with 5 selections on each question where only one selection is correct, what would the answers to question 1 be?
3. 70% of all drivers wear seat belts. A random sample of 4 cars are stopped.
   (a) Construct the probability distribution for the number of drivers wearing seat belts. (Use the binomial tables to find the individual entries similar to the method illustrated in question 4 in the example on page 129.)
   (b) Compute the mean and standard deviation of the distribution.
   (c) What is the probability that 2 or fewer drivers are wearing seat belts?
4. (a) In February 2011 Microsoft released a software update for Windows mobile phones and it was found that the update had problems for 1 in 10 users. If Microsoft tested the update on 30 randomly selected phones running the software before its release, what is the probability at least one would have shown problems? (*Hint: Use the complement event.*)
   (b) Out of the 30 phones how many would you expect to fail?
   (c) Are the 30 trials *really* independent so a binomial distribution is justified?
   (d) Assuming Microsoft did not notice the problem before the release, what might account for this given the high probability of detection?
5. Suppose a handful of 3 marbles is to be randomly drawn from a box containing 200 marbles. If 1/5 of the marbles are white and the rest are coloured, what is the probability of drawing 1 white marble in the sample? **Hint: 3 is less than 5% of 200.**
6. A committee of 10 is to be chosen at random from a community of 1000. 80 members of the community are receiving social assistance. Find the probability that exactly 2 of the committee members are on social assistance. Justify your use of the binomial distribution.

3.1.3 Review Exercises on Discrete Probability Distributions

1. A grocery store keeps track of the proportion of times that a given number of rotten oranges are found in boxes supplied by a certain produce wholesaler. They observed:

x	0	1	2	3	4	5	6	7
$P(x)$	0.08	0.12	0.16	0.26	0.17	0.10	0.06	0.05

   (a) On average, how many rotten oranges are found per box?

   (b) What is the standard deviation of the number of rotten oranges found per box?

2. Of all the members of a large credit union, 20% have loans with the credit union. A random sample of 6 members are polled regarding their loan status.

   (a) What is the probability that of the 6 :

      i. None have loans with the credit union?

      ii. All six have loans with the credit union?

      iii. At least 1 has a loan with the credit union?

   (b) Of the six what is

      i. The expected number of members with a loan?

      ii. The standard deviation in the number of members with a loan?

3.2 Continuous Probability Distributions

3.2.1 Definition

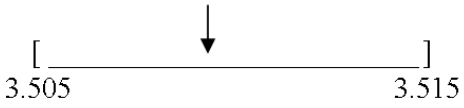
Continuous distributions model situations where the variable's value is obtained by a measuring process. In the cases examined so far, the variable's value has always been obtained by counting and that is why, for example, the binomial distribution was referred to as a discrete distribution.

A continuous variable might represent a quantity such as:

Length, Weight, Test Scores, Volume, Fuel Economy, etc.

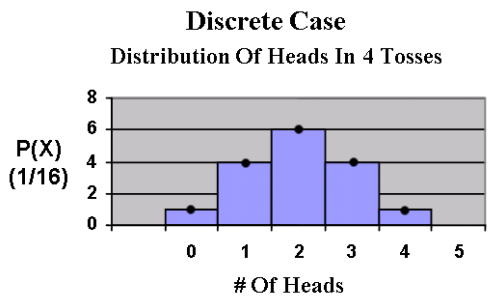
Any measured value, by the nature of the measuring process, is inaccurate. A measuring process is incapable of coming up with an exact value unlike a counting process that usually comes up with an exact value.

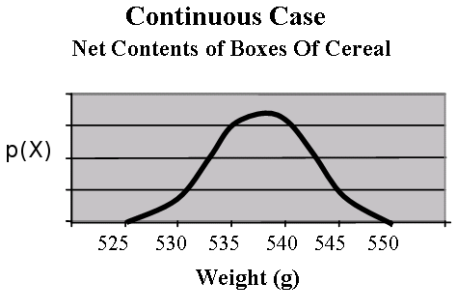
Example:
If an object has a reported length of 3.51 m, this is interpreted as a length being in the interval:



In the last example the true value for the length is unknown. This notion applies to every measurement. The reported value is only as accurate as the device used to take the measurement. The probability that a continuous variable assumes some exact value is 0. So, for instance, $P(x = 3.51000000\ldots) = 0$. Probabilities for continuous variables must therefore be specified over an **interval**. For example we wish to know $P(3.505 \text{ m} < x < 3.515 \text{ m})$.

On a discrete distribution curve, individual probabilities P are plotted at distinct locations x on the horizontal axis. We often choose to plot these as bars with width one, both for appearance but also so that the area of the bar also equals the probability (Area $= P \cdot 1 = P$). However, in the discrete case, the intermediate values between the discrete outcomes are not possible measurements. For instance, in our discrete coin-toss histogram below a value of $x = 1.23$ is not a possible outcome. On a continuous distribution curve, there is a point plotted for every value on the horizontal axis since the value of the variable could be anywhere.

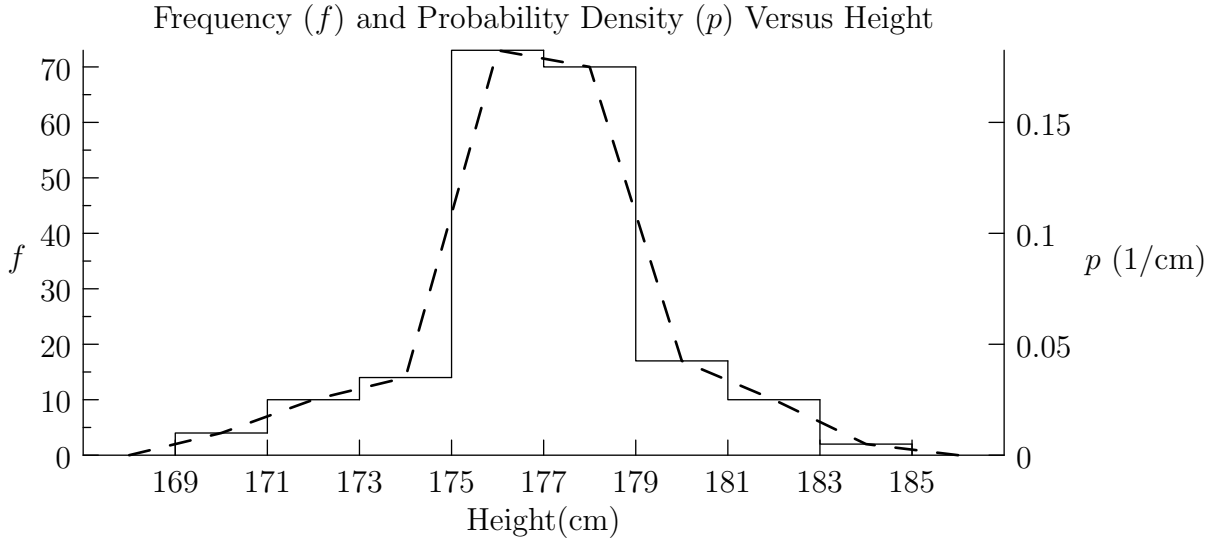




Smooth continuous curves can be represented by a mathematical expression called a **probability density function**, **p(x)**. The function does not calculate probabilities but calculates the y-values on the curve. The $p(x)$ curve has the property that the **area under the curve is 1**. The probability of an event is found by determining the **area under the curve** and above the x-axis associated with the **interval** of the event.

3.2.2 Experimental Probability Densities

It should be noted that in Unit 1 we already created our own probability density functions when we plotted the relative frequency density histogram p versus continuous variable x, or even better the smoothed out relative frequency density polygon $p(x)$ versus x. For example if we created these plots for the distribution of basketball player heights introduced in Section 1.5.2 we have:



In Unit 1 if we were interested in measuring the proportion of observations between the two variable values $x = 173$ cm and $x = 179$ cm we would have calculated the area under the relative frequency density polygon (using the heights measured off the p-axis) over the interval $(173 \text{ cm}, 179 \text{ cm})$. To consider $p(x)$ as a probability density one could simply rephrase the question as, "What is the probability that a basketball player in the sample, chosen at random, has height within the interval $(173 \text{ cm}, 179 \text{ cm})$?" The answer would be the same. Finally, the more useful probabilistic interpretation, however, is one of inference, namely, if the height of any high-school basketball player in Western Canada is measured, what is the likelihood it lies within $(173 \text{ cm}, 179 \text{ cm})$? Since the sample relative frequency density should be a good approximation of that for the population, the answer, once again, is just the area under the curve over the interval. As such our work in constructing relative frequency density polygons $p(x)$ amounts to creating probability density functions experimentally; only a change of interpretation is required.

Example:
Find the probability a randomly selected basketball player is between 173 cm and 179 cm by estimating the area under the curve. Shade in the area. (Remember to use the probability density axis, p, for the height. Using the histogram line rather than the frequency polygon will be easier since it gives rectangles with area base times height.) (Answer: ≈ 0.8)

Solution:
$P(173 \text{ cm} < x < 179 \text{ cm}) =$

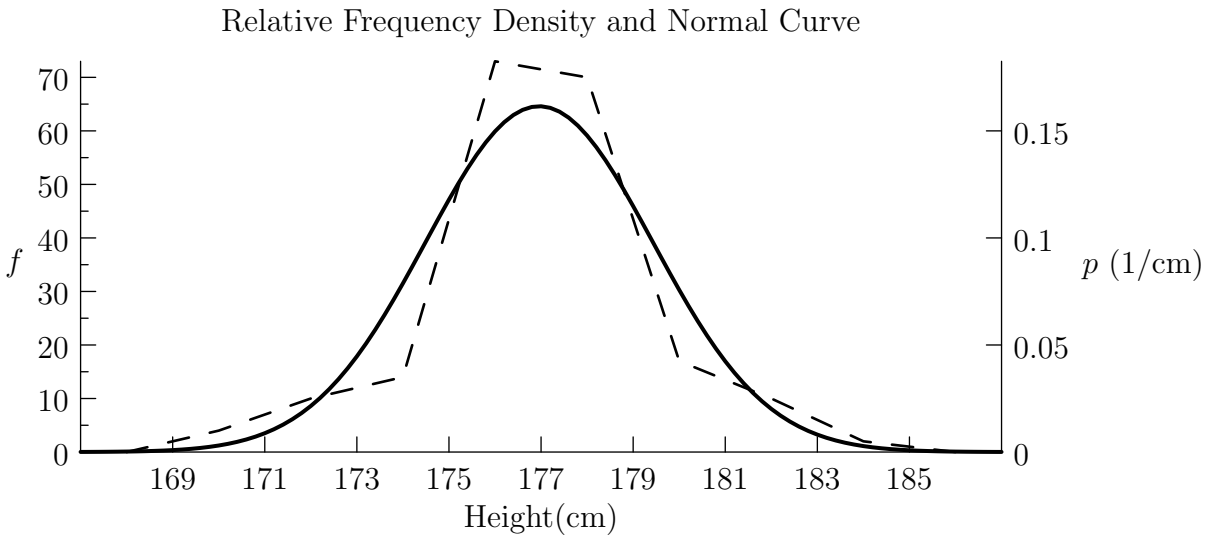
3.2.3 The Normal Curve

Definition

Any function of a variable with total area under the curve equalling one could be interpreted as a probability density function, and, as seen above, we can create our own experimentally that have this property. We have already seen that special cases of discrete probability distributions occur, for example the binomial distribution. Similarly the **normal curve** is a special case of a continuous probability density function.[4] The probability density function for the normal curve is:

$$\boxed{p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \quad \leftarrow \begin{cases} \pi = 3.14159\ldots \\ e = 2.71828\ldots \end{cases}$$

It is not necessary to know this equation to work with the distribution. The normal curve is useful for modeling certain types of measurement distributions and also shows up theoretically. As an example of the former if we take the parameters $\mu = 177.0$ cm and $\sigma = 2.5$ cm, which are the actual mean and standard deviation of the basketball player data, a plot of the normal curve (solid line) is as follows:



We see that our frequency polygon (dashed line) derived from the experimental data is quite close to the normal curve and we would say our data is approximately **normally distributed**. It is important to note that the graph of the normal curve is completely specified by its mean μ and standard deviation σ.[5] Different values of μ and σ will produce different curves but all of these curves have essential properties in common. If we made a table of values for x and $y = p(x)$ for any normal distribution with a given mean and standard deviation and plotted the ordered pairs, the curve that results always has some standard features in common with our example:

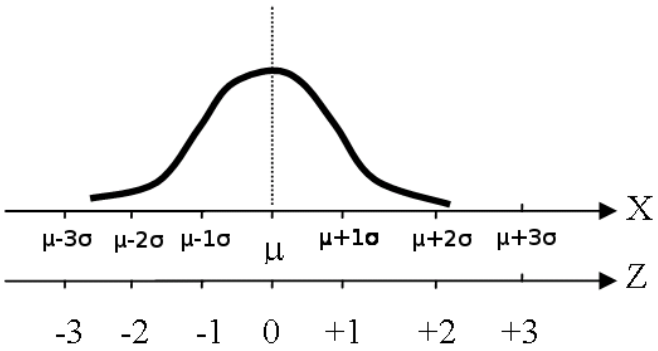
1. The curve is bell-shaped.
2. The curve is symmetric about μ.
3. The curve is asymptotic to the x-axis.

[4]The normal curve is also called a **bell curve** or a **Gaussian distribution**; the former name is due to it's shape and the latter to its developer Carl Friedrich Gauss.

[5]To prove that the two parameters of our normal distribution, μ and σ, really are its mean and standard deviation requires the introduction of mean and standard deviation formulae for continuous probability distributions (similar to Section 3.1.1 for discrete distributions) and the use of integration (calculus).

4. A fixed proportion of observations (probability) lies within a given number of standard deviations of the mean. The distance from the mean is plotted in units of standard score z, already introduced in Section 1.15.2: $\boxed{z = \frac{x - \mu}{\sigma}}$

Because of this last property, it is useful to add a standard score axis z below the x-axis:



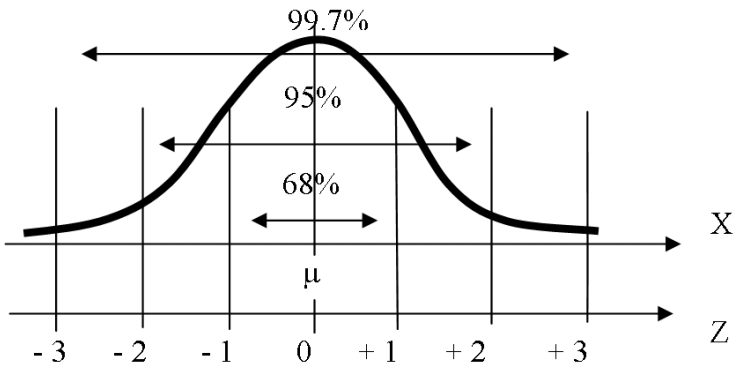
The Empirical Rule

In order to evaluate probabilities involving normally distributed variables we need to be able to find the area under the curve for intervals of interest. Because of standard feature 4 of normal variables detailed in the last section we can state the following known as the **Empirical Rule**:

If a variable is **normally distributed** a measurement has a

- **68%** or roughly **2/3** chance of falling within **1** standard deviation of the mean.
- **95%** chance of falling within **2** standard deviations of the mean.
- **99.7%** chance of falling within **3** standard deviations of the mean. (Effectively all the measurements fall within this interval.)

The following diagram illustrates a normal curve with this dispersion about the mean.



In symbols the Empirical Rule states:

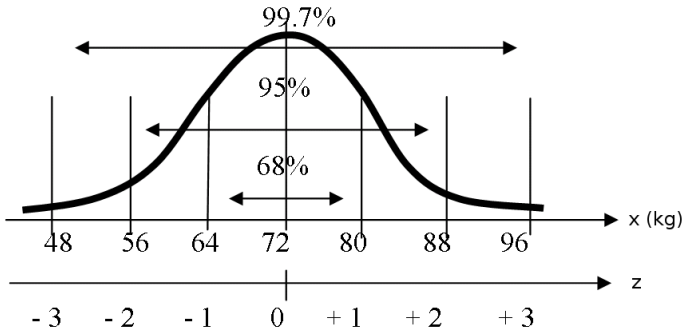
- $P(\mu - 1\sigma < x < \mu + 1\sigma) = P(-1 < z < 1) \approx 0.68$
- $P(\mu - 2\sigma < x < \mu + 2\sigma) = P(-2 < z < 2) \approx 0.95$
- $P(\mu - 3\sigma < x < \mu + 3\sigma) = P(-3 < z < 3) \approx 0.997$

The Empirical Rule should be memorized to allow quick interpretations of normally distributed variables. Note that the Empirical Rule can be used to interpret the likelihood of a measurement of a normally distributed variable in the same way we used Chebyshev's Theorem before in Unit 1 but now more specifically.[6] A value of x with $|z| \approx 0$ is approximately equal to the mean, with $|z| \approx 1$ is only slightly removed from the mean, with $|z| \approx 2$ is moderately removed from the mean while greater than three standard deviations different ($|z| \gtrsim 3$) is extremely far from the mean (and hence improbable).

Example:
If the weights of marathon runners are normally distributed with a mean of 72 kg and a standard deviation of 8 kg,

1. What is the range of weights?
2. Between what two weights will the middle 2/3 of the runners fall?
3. Above what weight do the top 2.5% of the runners lie?
4. Is a weight of 40 kg a typical weight?

Solution:
A diagram illustrating the Empirical Rule for this problem is as follows:



Here we labelled the x-axis just by starting at the mean (72 kg) and going up and down in units of the standard deviation (8 kg).

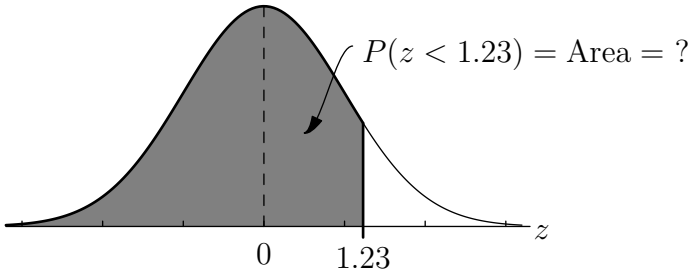
With the aid of the diagram one has:

1. $R = x_n - x_1 \approx (\mu + 3\sigma) - (\mu - 3\sigma) = 96 \text{ kg} - 48 \text{ kg} = 48 \text{ kg}$
2. Between $\mu - 1\sigma = 64$ kg and $\mu + 1\sigma = 80$ kg
3. We need a tail of area 2.5%, so above $\mu + 2\sigma = 88$ kg
4. $z = \dfrac{x - \mu}{\sigma} = \dfrac{40 \text{ kg} - 72 \text{ kg}}{8 \text{ kg}} = -4$, which is not typical ($|z| > 3$)

[6] See problem 9 page 145.

Areas Under the Normal Curve

Sometimes we are interested in the probability associated with an arbitrary interval of a normal variable x so the Empirical Rule is of no use. If x has mean μ and standard deviation σ, the procedure used to calculate the area under the normal curve above such an interval is to first transform it into its equivalent interval in terms of the standard score z which has mean 0 and standard deviation 1. The normal curve with this transformed axis is called the **standard normal curve**. Areas under this curve can be calculated[7] and are tabulated in the table appendix. Just as with our cumulative binomial tables, our cumulative normal distribution tables tabulate the area (probability) under the curve up to a particular value of z. For example, suppose we want the probability that z is lower than the value 1.23 . Graphically this is the following shaded area:



The normal distribution table on page 282 is organized with z-values on the outside of the table and areas (probabilities) on the inside of the table. In our case we want the area corresponding to $z = 1.23$. The row the value is in is determined by the digit to the left of the decimal which here is 1, and the first digit after the decimal (here 2). Find the row labelled 1.2 . The column is found by the second digit after the decimal, here 3 . Look down the 0.03 column and you will find 0.8907 . Therefore the desired probability is:

$$P(z < 1.23) = 0.8907$$

Note:

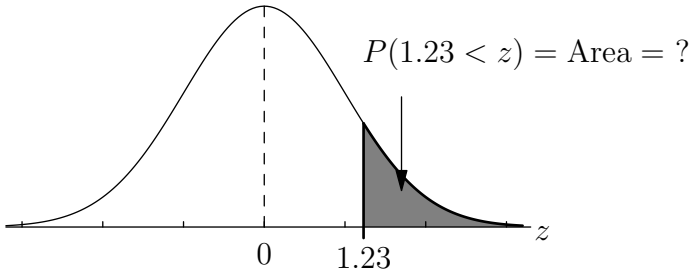
- The first digit which determines the row may be negative or zero and that the column for 0.03 is actually the fourth column because the first column is for 0.00 .
- The area to the left of $z = 0.00$ is 0.5000 because the total area (probability) under the curve is 1, but the normal curve is symmetric so half must lie on each side of 0.00 . Find the area 0.5000 on the inside of the table.
- For a continuous probability $P(z < 1.23) = P(z \leq 1.23)$ since the probability at an exact point (here 1.23) is 0.

All other normal probabilities may be found by using these tabulated values and the properties of the normal distribution.

Example:
Find the probability z is greater than 1.23 .

Solution:
We now desire a right tail:

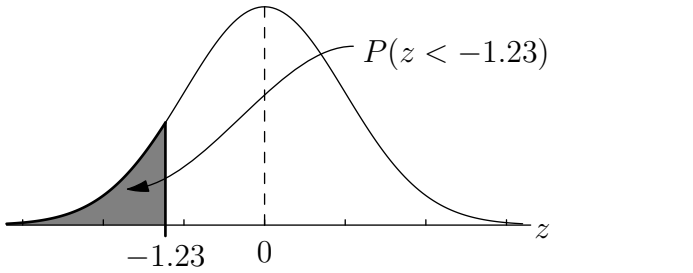
[7] The student familiar with calculus will recognize that the problem is one of evaluating a definite integral of the form $\int_a^b e^{-z^2/2}\, dz$. Despite its simplicity, the integrand has no antiderivative in terms of simple functions and the integral must be evaluated numerically.



Since the total area is 1.0 we can subtract the tabulated area we just found (the complement event) from 1.0 to get our desired area:

$$P(1.23 < z) = 1 - P(z < 1.23) = 1 - 0.8907 = 0.1093$$

Note here the inequality ("z greater than 1.23") was written so that a less than sign was employed again. Another quicker solution to find this area is to recognize that, due to the symmetry of the normal distribution, the tail on the right, $P(1.23 < z)$, exactly equals the tail to the left of $z = -1.23$, which is on the table:



Hence

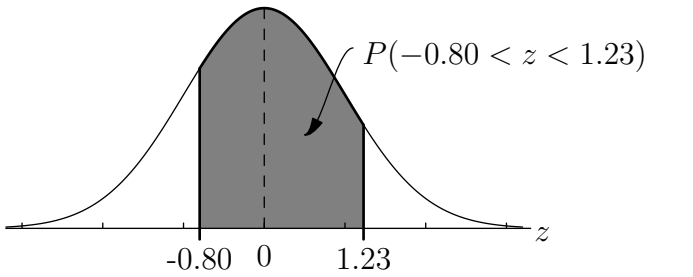
$$P(1.23 < z) = P(z < -1.23) = 0.1093$$

as before.

Another common area we will often require is between two z-values.

Example:
Find the probability z is between -0.80 and 1.23

Solution:
The desired area is the following shaded region:



In this case we just need to subtract the area going up to the lower value (-0.80) which is, from the table, 0.2119, from the area going up to the greater value (1.23) which we found above to

be 0.8907:

$$P(-0.80 < z < 1.23) = P(z < 1.23) - P(z < -0.80) = 0.8907 - 0.2119 = 0.6788$$

Note the identical procedure would have been employed if both z-values had been to the left of 0 or to the right of it. Sketch those situations.

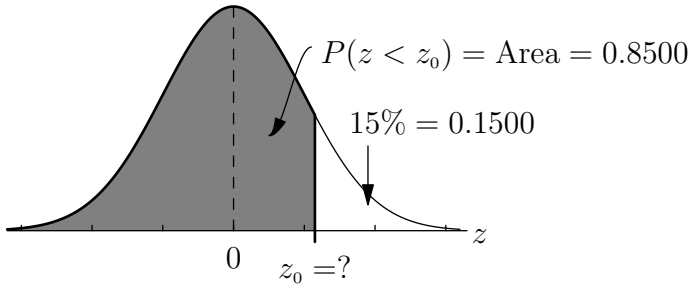
The final problem we may be interested in is the inverse, namely given an area (probability), find the z-value that corresponds with it.

Example:
Find the z-value above which 15% of the measured values (=area=probability) lie.

Solution:
In this case we are given a probability or area of $15\% = 0.1500$ which sits in a tail on the right. Since our table tabulates areas on the left, we will work with an area of:

$$\text{Area} = 1.0 - 0.1500 = 0.8500$$

Diagrammatically the problem is shown here:



Now the cumulative probability 0.8500 lies **inside the table**. The probabilities increase from row to row so scanning inside the table we find the closest numbers are 0.8485 corresponding to $z = 1.03$ and 0.8508 corresponding to $z = 1.04$. Since 0.8500 is closer to 0.8508 we'll use $z_0 = 1.04$. (We could interpolate to be more accurate. For this class that is unnecessary. If the desired area lies an equal distance from the two neighbouring areas, then choosing z exactly between the two values is acceptable.)

Now that we know how to evaluate areas on the normal curve in terms of z, real problems just require the additional initial step of converting from x-values to the z-scores.

Example:
Continuing the marathon runner example, if their weights are normally distributed with a mean of 72 kg and a standard deviation of 8 kg,

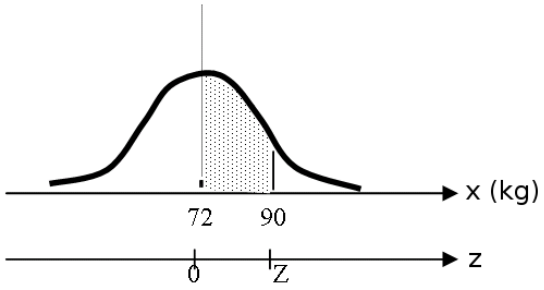
1. What proportion of runners' weights fall between 72 kg and 90 kg? (As a probability question: What is the probability a randomly selected runner will weigh between 72 kg and 90 kg?)
2. What proportion of runners' weights exceed 90 kg? (As a probability question: What is the the probability a randomly selected runner will weigh above 90 kg?)
3. Below what weight do the lightest 10% of the runners fall?

Solution:

For all normal distribution problems a properly labelled diagram with correct scales on the x and z-axes allows easy checking of the reasonability of solutions.

Variable x is the weight so draw a normal curve with a weight axis as we did in the original Empirical Rule diagram. To find areas under this curve, a second horizontal axis with the standard score must be included. (Note the scale ticks have been omitted here for clarity.)

1.



Since $x = 72$ kg is the mean, it corresponds to $z = 0$:

$$z = \frac{x - \mu}{\sigma} = \frac{72 \text{ kg} - 72 \text{ kg}}{8 \text{ kg}} = 0.00$$

Transform $x = 90$ kg to its corresponding z-score:

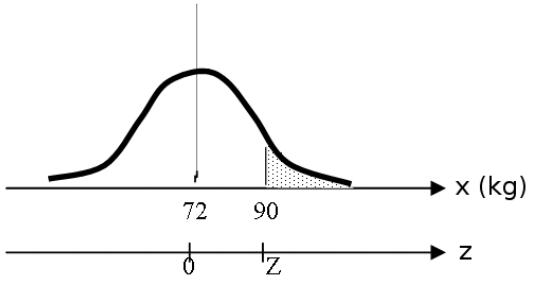
$$z = \frac{x - \mu}{\sigma} = \frac{90 \text{ kg} - 72 \text{ kg}}{8 \text{ kg}} = 2.25$$

The area in the shaded region is:

$P(72 \text{ kg} < x < 90 \text{ kg}) = P(0.00 < z < 2.25) = P(z < 2.25) - P(z < 0.00) = 0.9878 - 0.5000 = 0.4878$

from the tables.

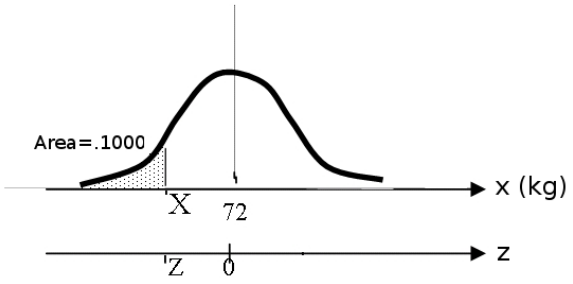
2.



The area in the right tail is:

$$P(90 \text{ kg} < x) = P(2.25 < z) = P(z < -2.25) = 0.0122$$

3.



In this case we work in reverse. The left tail has an area of 10%=0.1000 and we look for it inside our table. The closest value inside the table is 0.1003 which corresponds to a z-value of -1.28 . Inverting our z-score formula gives:

$$\boxed{x = \mu + z\sigma} = 72 \text{ kg} + (-1.28) \cdot (8 \text{ kg}) = 61.76 \text{ kg} = 61.8 \text{ kg}$$

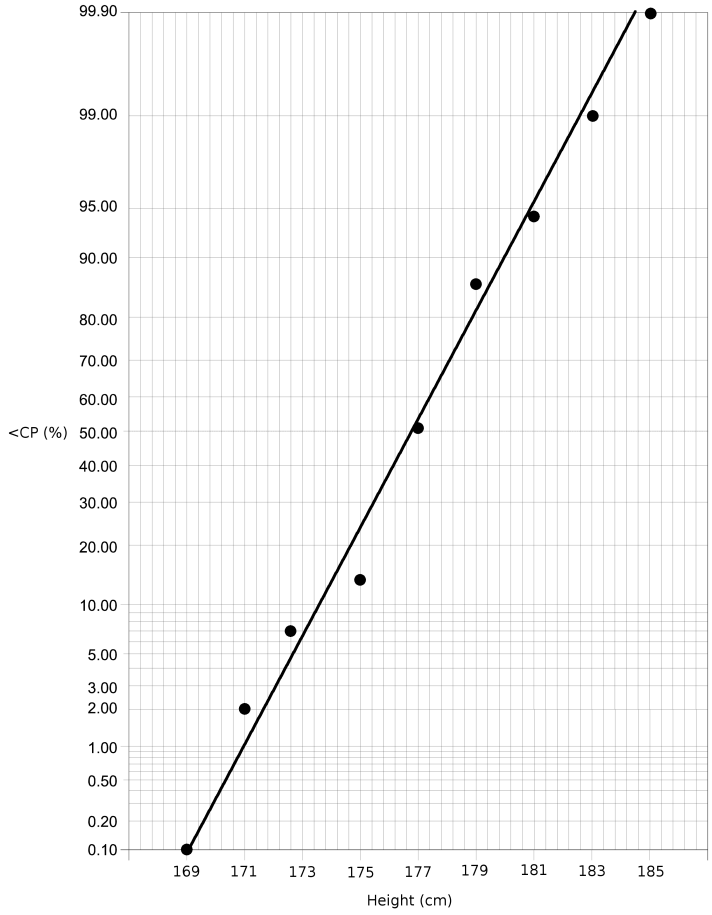
*3.2.4 Tests for Normality

One of the most important objectives of statistics is to determine a goodness of fit of a set of data to a theoretical model. Once we fit a set of data to a particular theoretical model in statistics, we can use that model to solve the problems pertaining to the analysis of that data.

The normal probability model is the most commonly applied model. In most applied problems it would be good to know that the normal distribution is a close approximation to the actual unknown distribution we are working with, particularly because it is easy to use. We have already considered some methods to compare a distribution to the normal distribution:

1. Plot histograms/frequency polygons and compare the appearance observed to that of the normal curve (using the data's mean and standard deviation) for shape.
2. Compute the descriptive summary measures (mean, median, mode, standard deviation) and compare the characteristics of the data with those of the theoretical properties of the normal probability distribution.
3. Compare the proportions of data within units of standard deviations from the mean with those prescribed by the empirical rule.

4. Another approach to checking for normality is to plot a cumulative relative frequency ogive ($<CP$ versus x, Section *1.7.2) on **normal probability paper**. This type of graph paper has a regular arithmetic horizontal axis for x, but the vertical axis indicates cumulative probabilities associated with a normal distribution. It is a transformed normal scale. If the plot of data makes a relatively straight line, the normal distribution is a good approximation and may be used for solving the problem at hand. Note that the idea of linearizing data is a common method for recognizing trends in observations. As an example if we plot an ogive for the basketball player heights of Section 1.5.2 on normal probability paper we see that the points lie quite close to a straight line, suggesting the data is well-described by a normal distribution. This was also suggested by our earlier frequency polygon plot.



Other methods exist for testing the normality of data but are beyond the scope of this presentation.

The normal curve will be used repeatedly for the remainder of this course. Students are encouraged to master this topic before proceeding.

Assignment:

For each question part below make sure to draw a labeled diagram of the normal curve with the appropriate region shaded in. Pay attention to the proper use of symbols.

1. The distribution of wages in a community is approximately normal with a mean of \$37,500 and a standard deviation of \$3,000. * This question requires only use of the Empirical Rule. *
   (a) Draw a labeled diagram to illustrate this information.
   (b) What are the standard scores of incomes of i) \$42,000 ii) \$28,000 iii) \$60,000? Which of these wages would you typically find in the community?
   (c) What proportion of all wage earners in the community earn between \$31,500 and \$43,500?
   (d) What is the probability a wage earner in the community earns more than \$31,500?
   (e) What is the lowest wage likely to be observed in the community?

2. The heights of males in the general population is approximately normal with a mean of 172 cm and a standard deviation of 10 cm. If a male is selected at random, what is the probability his height falls within the range
   (a) $172 \text{ cm} < x < 192 \text{ cm}$?
   (b) $x > 180$ cm ?
   (c) $160 \text{ cm} < x < 180 \text{ cm}$?

3. Scores on an aptitude test have been observed to be approximately normal with a mean of 76% and a standard deviation of 5%.
   (a) What proportion of all scores exceed the mean?
   (b) What proportion of the scores are between 65% and 85%?
   (c) Below what score do the bottom 10% of all scores fall?
   (d) Above what score do the top 15% of all scores fall?
   (e) If 1000 people took the test, how many would you expect to score above 80%?

4. Delinquent accounts for a chain store are normally distributed with a mean of \$308.45 and a standard deviation of \$43.09. If an account is selected at random from among those that are delinquent, what is the probability that it falls in the range?
   (a) $x > \$250$
   (b) $x < \$400$
   (c) $\$250 < x < \$400$

5. A personnel office hires only those persons who score in the top 5% of an applicant test. Scores on the test vary about an average of 65% with a standard deviation of 10% and are normally distributed. What is the cutoff score for being hired?

6. A machine produces parts that are circular in cross section. The diameter of the cross section can be controlled by the production process to vary normally about a mean of 10 cm with a standard deviation of 2 mm. Parts whose diameters differ from the average too significantly are rejected as scrap. If the extreme 10% of the parts are rejected, what is an acceptable range for the diameter of a part?

7. A set of test scores are normally distributed with a mean of 70% and a standard deviation of 10%. Suppose a person receives a percentile of 62. What was the person's actual score on the test?

8. Confirm the Empirical Rule results using areas from the normal distribution table.

9. Using the Empirical Rule, verify that Chebyshev's Theorem applies for the normal distribution with $k = 2$ and $k = 3$.

10. In Section 3.2.2 you found the probability a basketball player was between 173 cm and 179 cm. In Section 3.2.3 that probability distribution was then modelled with a normal distribution with standard deviation $\mu = 177.0$ cm and standard deviation $\sigma = 2.5$ cm. Recalculate the probability using the normal model and compare it to your original result.

11. You are interested in going south for a week during Spring Break. The following table gives a sample of past years data (up to 2011) for the high temperature on March 1st for various locations.

Destination	March 1st Daily High (°C)									
Cancun, Mexico	28	27	28	28	31	30	28	31	31	29
Los Angeles, California	17	22	29	14	19	19	19	17	18	21
Malibu, California	13	17	22	14	14	15	17	15	15	21
Mazatlan, Mexico	25	26	28	26	27	29	26	23	25	27
Miami, Florida	31	21	29	26	28	26	23	26	29	24
Phoenix, Arizona	24	19	30	27	18	24	22	19	16	22
Puerto Vallarta, Mexico	24	27	29	26	26	27	27	25	27	27
Varadero, Cuba	29	22	28	33	28	24	28	33	29	29

Pick one of the above destinations and do the following calculations. (You can compare the destinations by looking at the answer key when you're done.)

(a) Calculate the mean and standard deviation for the March 1st high temperature for your location. (Check your answer before proceeding to the next question.)

(b) A good day at the beach or lying by an outdoor pool absolutely requires a daily high of at least 25°C. If we assume the daily high temperature for your location is **normally distributed** with the mean and standard deviation you calculated from the last question:

   i. Calculate and interpret the z-score for 25°C at your location.

   ii. What is the probability that a given day at the beginning of March will have beach weather (high at least 25°C) at your location? (Check your answer before proceeding to the next question.)

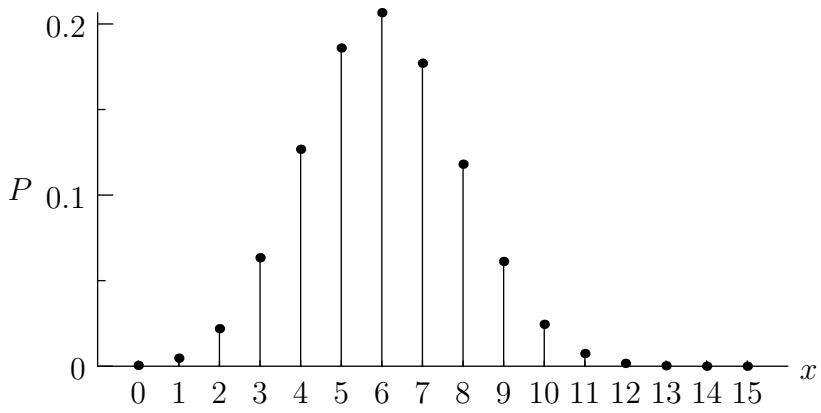
(c) Next, assume you will be staying for one week. If the weather on each day could be considered to be independent – a bad approximation (why?) – use your probability from the last question and the **binomial distribution** to calculate:

   i. The expected (mean) number of days out of the seven that will have beach weather (high at least 25°C).

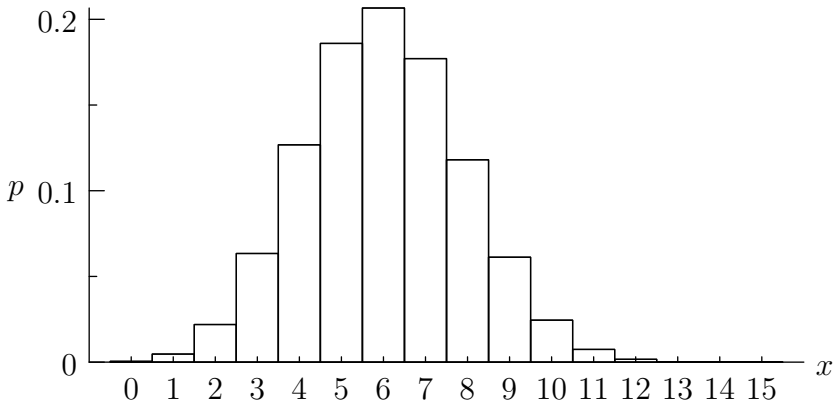
   ii. The probability that you will have at least 5 days of beach weather for your week.

3.3 Normal Approximation to the Binomial Distribution

In some cases the binomial distribution may have too many trials (n large) to be tabulated usefully. We have seen that the normal distribution is useful for modelling many continuous distributions. It can also be used for approximating the discrete binomial distribution under certain circumstances. At first glance, the idea of approximating a discrete distribution with a continuous one may seem impossible. Let us consider the binomial distribution with $n = 15$ and $\pi = 0.4$:

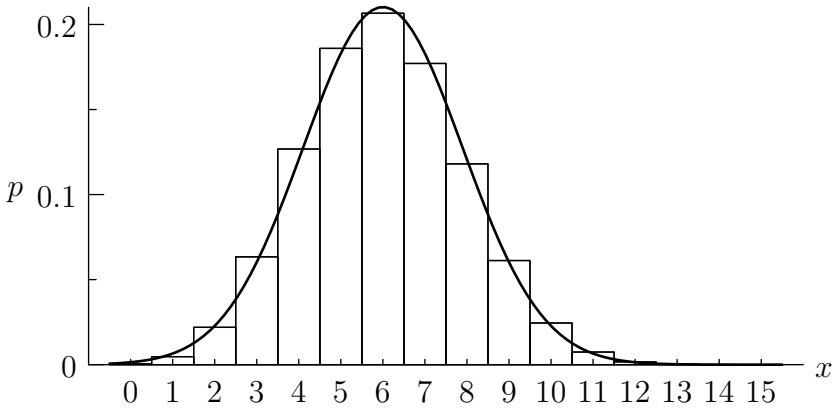


The distribution is discrete so only integer values have any finite probability. We can, however, give each line a width of 1 so that the area of the bar still equals the probability at that value as shown in the following diagram.

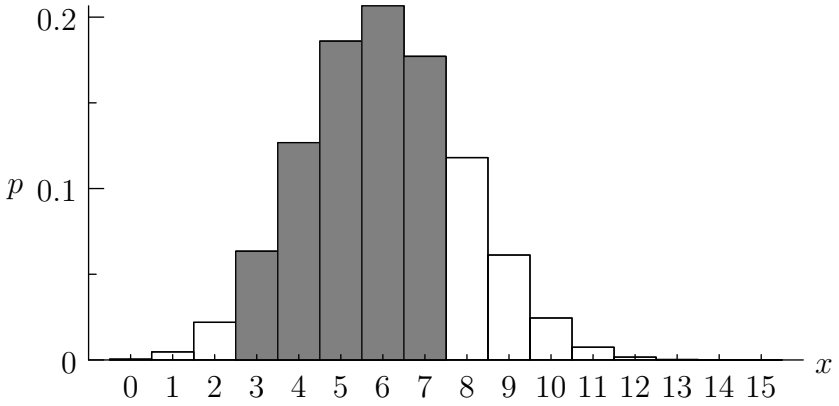


Note we have effectively converted our curve into a probability density $p(x)$. To get the probability for say $x = 4$ we would just find the area under the curve from $x = 3.5$ to $x = 4.5$.

Since the binomial distribution now looks just like a lump, with mean $\mu = n\pi = 6.0$ and standard deviation $\sigma = \sqrt{n\pi(1-\pi)} = 1.897 \approx 1.9$, let us superimpose a normal distribution with those parameters. We see that the normal curve does a good job of approximating the binomial curve:



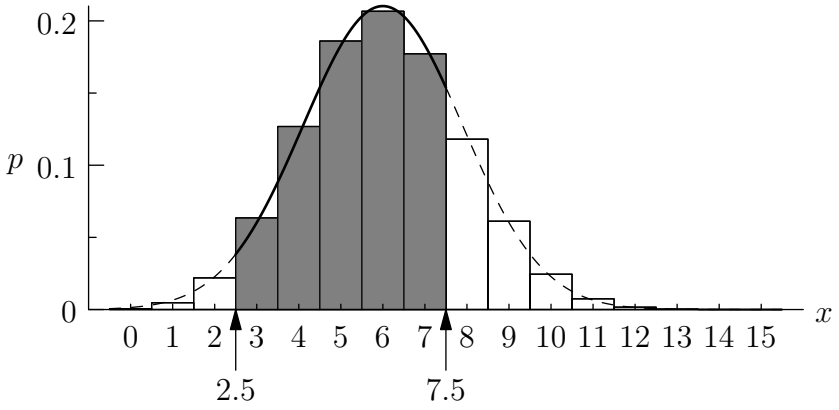
Now suppose we are interested in the probability that x lies between 3 and 7. This is the shaded area on the original distribution:



We can calculate this exactly using our cumulative binomial table with $n = 15$ and $\pi = .4$. We get

$$P(x = 3, 4, 5, 6 \text{ or } 7) = P(x \leq 7) - P(x \leq 2) = 0.787 - 0.027 = 0.760$$

However, as shown in the following diagram, this is approximately equal to the area under the normal curve over the corresponding interval:

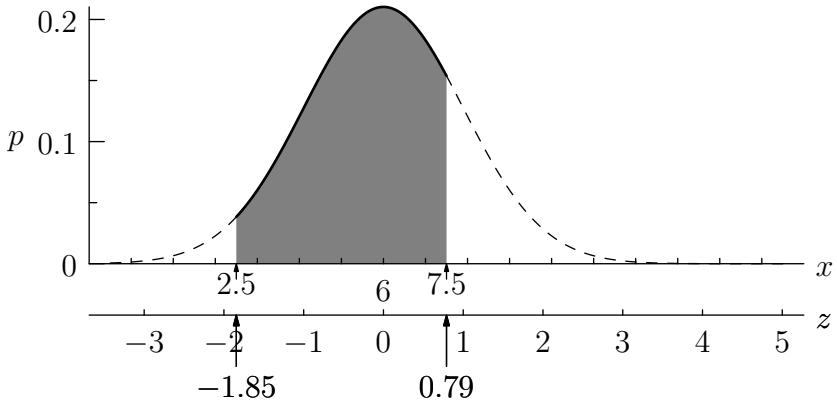


Note that the desired normal distribution interval for x is not between 3 and 7 but rather 2.5 and 7.5 since we want to include all the area of the end bars. This is called the **continuity correction**. The

approximation for the area under the curve is then, using $z = (x - \mu)/\sigma$:

$$P(2.5 < x < 7.5) = P\left(\frac{2.5-6}{1.897} < z < \frac{7.5-6}{1.897}\right) = P(-1.85 < z < 0.79) = 0.7852 - 0.0322 = 0.753$$

This is quite close to our exact value of 0.760 ! The approximate area and normal calculation is summarized on the following diagram.



Validity of the Approximation

So when is the normal approximation to the binomial distribution valid? Since the normal distribution is always symmetric while the binomial distribution is only symmetric for $\pi = 0.5$ the approximation cannot be expected to work all the time. If both the following conditions on n and π hold the approximation will be good:[8]

$$n\pi > 5 \quad \text{and} \quad n(1-\pi) > 5$$

Example:
Check that the normal approximation to the binomial distribution was appropriate in the previous problem.

Solution:
We have

$$n\pi = (15)(.4) = 6$$

$$n(1-\pi) = (15)(.6) = 9$$

Since both these numbers are greater than 5 we expect the approximation to be good which is what was found.

[8]The cutoff used is somewhat subjective. Some texts use 10 rather than 5. A cutoff of 5 will mean a greater error in the approximation is acceptable.

Assignment:

In the following questions use the normal approximation to the binomial distribution. Remember to first check that doing so is valid.

1. A statistician flips a fair coin every day for an entire year (365 days). What is the probability that heads will show up 160 or fewer times?
2. Fruit flies have a certain genetic trait 2% of the time. If 1000 fruit flies are examined in a lab, find the probability of finding between 15 and 25 fruit flies with the trait.
3. At a school 30% of children wanted a dog, 25% wanted a cat, and 45% wanted a horse for their next pet. If 50 children from the school are chosen at random what is the probability that at least half of them would want a horse for their next pet.

Probability Distributions Formulae

Binomial Probability Distribution Function

$$P(x) = {}_nC_x \cdot \pi^x \cdot (1-\pi)^{n-x}$$

Parameters of Probability Distributions

General Formulae for Discrete Distributions

$$\mu = E(x) = \sum xP \qquad \sigma = \sqrt{\sum x^2 P - \left(\sum xP\right)^2}$$

Shortcuts for the Binomial Distribution

$$\mu = E(x) = n\pi \qquad \sigma = \sqrt{n \cdot \pi \cdot (1-\pi)}$$

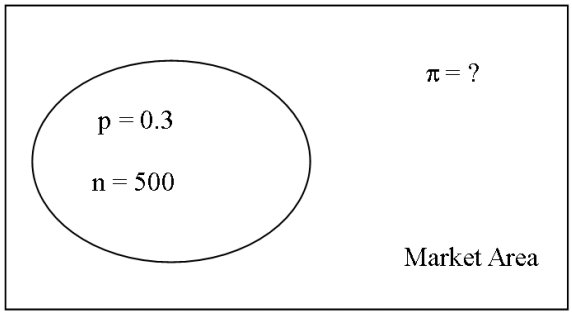
Unit 4: Sampling Distributions

4.1 Introduction

One objective of statistical sampling is to estimate the value of a population parameter. Sample statistics provide the information for making the estimates. To be a useful estimate of the parameter, each estimate must satisfy two conditions:

1. It must provide a **maximum error of estimate**. (How close do we think we are?)
2. It must provide a level of **confidence**. (What is the probability we are that close?)

Example:
A market analyst examined a sample of 500 households in a market area. The research shows that $p = 30\%$ of the households surveyed plan to buy a new vehicle in the next four years. The analyst has concluded that if this percentage were projected to the whole market area as an estimate of the actual proportion (π) of households in the market area planning to buy a new vehicle in the next four years, this value would be within 4% of the correct value in 95 samples out of 100.



In this example the maximum error of estimate is 4%; the level of confidence is 95%.

Sampling always produces estimates. It is unlikely that the sample value obtained exactly equals the population value to be estimated. Why are samples used instead of gathering all of the population data?

The term **error of estimation**[1] is used to represent the difference between the observed sample statistic and the population parameter.

Example:
In the previous market study on buying intentions of homeowners, suppose that 35% of the households in the market area actually intend to buy a new vehicle. Our sample results based on 500 households showed a proportion of 30%. In this case the error of estimation is:

$$\text{error of estimation} = p - \pi = 0.30 - 0.35 = -.05 = -5\% \ .$$

This means that we are 5% too low in our estimate.

[1] Also referred to as the **sampling error**.

In many applied sampling situations, all the items in the population are not available. Only the sample observations are known.

Example:
A sample of 10 families from a new large suburb are surveyed regarding the balance owing on their house mortgages. The following balances were observed: ($)

145,000	280,000	282,000	290,000	350,000
358,000	402,000	466,000	664,000	714,000

1. What is the estimate of the population mean?
2. If another sample of size 10 were drawn, do you think that the same sample estimate of μ would occur?
3. What is an estimate of the population standard deviation?
4. If another sample of size 10 were drawn, do you think that the same sample estimate of σ would occur?
5. Draw a sketch of the distribution of balances in the community based on a normal curve assuming the estimates of the parameters are accurate.

6. Without the assumption of normality is the sketch accurate?
7. Where is a balance of $450,000 located in the distribution?
8. What is the error of estimation in the estimate of μ if we know the population mean, μ, is $396,000?

9. Why is it not always possible when sampling to find the exact error of estimation in a sampling procedure as was done in the last question?

4.2 Unbiased Samples

The quality of your predictions from sample data depend critically on the quality of the method used to obtain it.[2] As we have seen, errors of estimation are a natural part of any statistical endeavour and cannot disappear. However one of our goals must be not to introduce further systematic errors in our predictions through bias. An **unbiased sample** is a sample drawn that is free from bias. One element of this freedom from bias lies within the method of its selection.

There are many different ways of drawing a sample from a population. Some of these methods are referred to as probability sampling and some are referred to as non-probability sampling. Probability methods of drawing samples are important for producing an unbiased sample. First consider some poor ways of constructing a sample.

A **non-probability sample** is chosen not by chance but based on some other criteria. Such a sample can result in certain measurements being over or under-estimated within a study. This in turn will bias the study toward certain outcomes. Some examples of non-probability samples include:

Convenience samples: You draw the sample from individuals in the population who are close at hand. (Example: surveying your classmates). Convenience samples typically suffer from **under-coverage** – whole sections of the population are under-represented or left out.

Voluntary samples: The sample is constructed from individuals in the population who self-select to be part of it. (Example: A website or phone-in poll.)

Example:
What type of non-probability sample is each of the following:

- You wish to estimate the proportion of the population who would like your new blueberry pie recipe so you put up a sign-up sheet to get test candidates to taste your pie.
- You wish to determine your chances at getting elected for city council so you take a poll of your neighbours to see whether they would vote for you.

A **probability sample** is one in which each member of the population has a nonzero probability of being included in the sample. The probability is known prior to sample selection and the choice is made by chance alone.

A **simple random sample** is a probability sample in which each item in the population has the same probability of being included in the sample.

A procedure like drawing names from a hat guarantees a simple random sample. Before selections can be made like this, each item in the population must be identified in order to give it a chance of being selected in the sample. Rather than draw items from a hat, statisticians simulate the process by constructing a **sampling frame** and using a table of random numbers to draw from the sample frame. A small set of random[3] digits is given below:

[2] One critical issue, how large the sample taken should be, will be considered in Section 5.1.3 .

[3] Most computer languages include a random number generator so this procedure can be done by machine. Modern calculators also have a random number button to produce such numbers. Where do computers and calculators get such random numbers? Computers are deterministic so there are algorithms which take a *seed* number which you provide which they then manipulate repeatedly by a function to produce pseudo-random numbers, namely digits that appear random and hopefully are equally likely. Additionally computers will often keep an *entropy pool* which contains random numbers linked to random events such as when you last turned on your computer, or touched your mouse and keyboard. More recently yet, some computer chip designers have been able to make quantum random number generators that exploit the inherent probabilities underpinning reality to produce true random numbers.

56599	62463	25114	61055	45618	73993	60743
85197	30682	77780	08002	57545	96111	23842
58835	10840	63210	56254	73053	09915	32766
13532	75531	83167	22578	12146	51981	73807
64682	84233	72523	21601	57214	52660	19001
12013	26749	84512	64112	28201	27741	84974
44599	13230	91202	95529	98096	05285	47421
35584	27278	03848	34905	85168	65804	68606
13434	46602	48712	58533	53769	48494	87451
47749	64307	69180	85351	02772	97869	93451
32188	88035	98488	61690	70573	78592	68315
93405	81747	44672	91838	87334	35692	47561
40334	93806	58107	33344	45968	82226	11441
79081	35905	94398	94027	40312	91620	41141
72754	96584	17577	34772	53925	52740	97393

Example:
Construct a sampling frame of a class of 35 students and use the table of random numbers to draw a simple random sample of size 5.

Solution:
Construct a sampling frame for the class by assigning each class member a two digit number starting at 01 and sequentially ending with the last class member, student 35. Read down the first two digits of the first column selecting the first five numbers that appear in sequence that are in the range from 01 to 35 and rejecting those numbers outside of the range.

A **stratified random sample** is a probability sample in which the population is divided into relatively homogeneous groups (strata) and a random sample is taken from within each strata.

Example:
In a political survey it is desirable to reflect the diversity of the population. The pollster may deliberately include participants based on strata organized upon lifestyle (urban/rural), regional location (BC, AB, Prairies, ON, QC, Atlantic), race, or religion, the number belonging to a strata in his sample based on its proportion within the overall population. Such a stratified survey might be considered to be more representative than a simple random sample.

Usually national political pollsters will not include Alberta in the Prairies (MB+SK) strata, why not?

In addition to the method of selection it is important that an unbiased sample be free of bias of procedure. Some procedural biases to look out for that are common involving studies of people are:

Non-response error: People refuse to answer.

Response error: People answer but are unsure or lie.

Design of Questions: Questions may be unclear, ambiguous, or leading.

Interviewer: The race, sex, age, or attitude of the interviewer can influence response.

So far the source of data considered has largely been **observation**. Here, in principle, there is no interaction between the researcher and the data. Rather measurements, such as observing cards drawn from a deck or inquiring regarding a political preference, are taken. When the data involves **experimentation** the researcher is instead manipulating one or more variables (the **factors**) to see

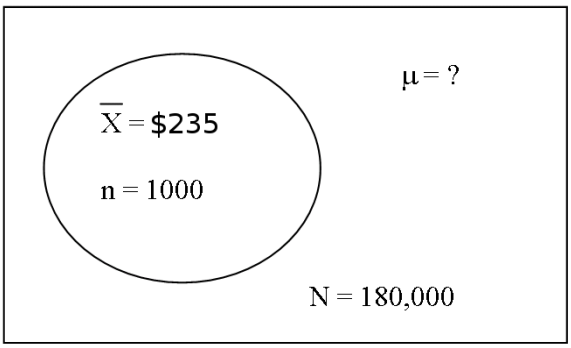
the effect on some other variable (the **response**). For example, the effect of lack of sleep on the ability to perform some task (measured in time taken) could be considered. In such a case there would be two populations and hence two samples required - a group who receive regular sleep and a group that are sleep-deprived. If we allow for different **levels** of sleep rather than just two or we additionally allowed variation of some other factor, say, coffee consumption, among the participants the experiment is more complicated yet. **Confounding variables** which were hidden in the experiment and not accounted for by the experimenter can lead to erroneous conclusions. (Example: social class in the traditional IQ test.) Obtaining experimental data therefore has further pitfalls for bias which can be mitigated through techniques such as having control groups, randomized experiment designs and double-blind experiments.

The goal here is not to throw up our hands and think that statistics is a hopeless endeavour. Rather we should think critically and ask questions about how the statistics we read came about and we should look for potential problems with our own methodologies. Obviously whole courses can be taken with regard to data collection and experimental design. In this class we will assume our data comes from simple random samples. The inference methods we will be considering cannot be applied to nonrandom sample data.

4.3 The Central Limit Theorem

In an applied sampling problem, usually one sample is drawn for purposes of determining the population parameter. In theory, it is possible to draw many more samples.

Example:
In Saskatchewan there are about 180,000 residents in the age group between 15 and 25 years of age. A market analyst would like to make projections about this group based on a sample of size 1000 individuals. Imagine that the parameter of interest is related to the sporting industry and concerns the average expenditure per individual on sporting equipment per year.



A list of 1000 expenditures is drawn without replacement from 180,000 expenditures. To have 100% confidence that we have the exact value of μ, all 180,000 values would have to be identified. The value that we observe in the sample depends on which values are drawn in the sample. If we sample without replacement, there are ${}_{180,000}C_{1000}$ different samples that could be obtained each with its own mean. If we sample with replacement then there are $180,000^{1000}$ possible samples. The size of this combination or power is too large for purposes of visualizing all the different possible sample possibilities. The error of estimation in our estimate depends on which sample we happen to select.

The previous example illustrates the following points which will be true when we try to estimate any population parameter with a sample statistic:

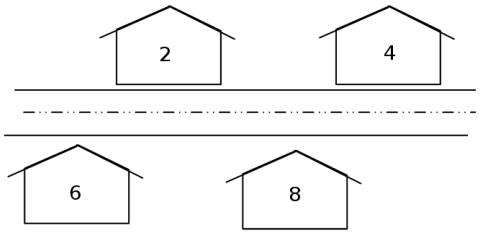
- There are usually a large number of samples that are possible (${}_NC_n$ without replacement, N^n with replacement).
- Each sample will generate a sample statistic (e.g. $\bar{x}$) which will estimate the population parameter (e.g. μ).
- Each such sample statistic will be in error; it will differ from the population parameter by its error of estimation. The sample statistics in general do not equal each other and therefore neither will their errors.

These points make it clear that for an individual sample mean, $\bar{x}$, to be useful one has to have some idea how close it is to the true population mean, μ. To analyze the problem we consider a **sampling distribution**: Take all possible samples of fixed size n from the population and measure the sample statistic (e.g. $\bar{x}$). The frequency of the sample statistic values (the $\bar{x}$'s) themselves form the distribution.[4]

[4]Note the difference here between a sampling distribution and a sample distribution. A sample distribution is just the distribution of the values (the x's) in one particular sample.

The above example is too large to study the sampling distribution of the mean so let us consider a simple example where the total number of possible samples is manageable. This situation is too small to actually use sampling techniques to analyze but we will use this example because it is small enough to visualize all of the possible samples that can occur.

Example:
Suppose a social scientist drives through a small village consisting of 4 households. Each household has a different number of occupants specified by the number inside the graphic below.[5] Imagine this to be a sampling situation where the scientist can only observe some of the households. The scientist is interested in the mean number of occupants per household in the village. Suppose 2 households are selected.



How many samples of size 2 can the analyst draw from this village? This depends upon how the sampling is done.

- If sampling is done **without replacement**, the answer is ${}_4C_2$. Confirm that ${}_4C_2 = 6$.
- If sampling is done **with replacement**, the answer is $4^2 = 16$.

In this type of problem sampling would likely be done without replacement but in problems where the sample size is a small fraction of the population the calculations can be done assuming sampling is done with replacement and the population looks the same from one draw to the next. For our purposes we will sample with replacement.[6]

The following table shows the 16 possible samples based on sampling with replacement. (Place the mean to the right of the sample for each possibility).

Sample	Sample Mean, $\bar{x}$	$\bar{x}^2$
(2,2)		
(2,4)		
(2,6)		
(2,8)		
(4,2)		
(4,4)		
(4,6)		
(4,8)		
(6,2)		
(6,4)		
(6,6)		
(6,8)		
(8,2)		
(8,4)		
(8,6)		
(8,8)		
	$\sum \bar{x} =$	$\sum \bar{x}^2 =$

The last table shows that when drawing random samples, the sample mean is itself a variable.[7] This collection of observations for the sample mean is called the **sampling distribution of the mean**.

Questions on this Sampling Distribution Example

1. On your calculator (or otherwise) calculate the mean and standard deviation of the population data values (2, 4, 6, and 8).

   $\mu =$ $\sigma =$

2. Calculate the mean of the sample mean column.

   $\mu_{\bar{x}} = \dfrac{\sum \bar{x}}{16} =$

3. How does the mean of all sample means, $\mu_{\bar{x}}$ compare to the mean of the x values in the population, μ?

4. If the third sample is the one randomly chosen, what is the error of estimation in estimating μ?

   error of estimation $= \bar{x} - \mu =$

5. Calculate the standard deviation of the sample mean column.

   $\sigma_{\bar{x}} = \sqrt{\dfrac{\sum \bar{x}^2}{16} - \left(\dfrac{\sum \bar{x}}{16}\right)^2} =$

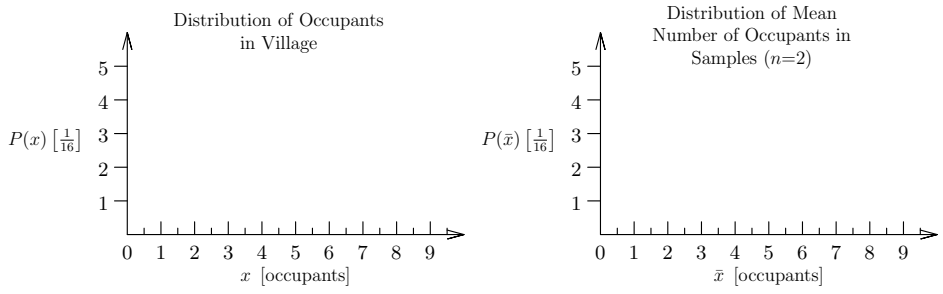
6. How does the standard deviation of the $\bar{x}$ values, $\sigma_{\bar{x}}$, compare to the standard deviation of the x values in the population, σ?

7. Complete the following tables for the probability distribution of x and $\bar{x}$.

x	P(x)
2	/4
4	/4
6	/4
8	/4
	$\sum P(x) =$

$\bar{x}$	P($\bar{x}$)
2	/16
3	/16
4	/16
5	/16
6	/16
7	/16
8	/16
	$\sum P(\bar{x}) =$

8. Now make histograms for these two probability distributions. Add a frequency polygon to your distribution of $P(\bar{x})$.



9. What is the shape of the distribution of sample means?

The behaviour of the sampling distribution of the means is summarized by the **Central Limit Theorem (C.L.T.)**. It states the following:

1. The **shape** of the distribution of sample means approaches a **normal curve**, as the size of the sample grows larger when sampling from a large population. For sample sizes of $\boxed{n \geq 30}$ the distribution of $\bar{x}$ will be approximately normal.

   ** Look at the shape of the frequency polygon in your diagram on the right, it is symmetric but it does not have enough points to produce a smooth bell shape. If our sample size had been 30 or more rather than 2 the shape would be very close to a bell curve. **

2. The mean of the sampling distribution is the same as that of the population.

$$\mu_{\bar{x}} = \frac{\sum \bar{x}}{{}_N C_n \text{ or } N^n} = \mu$$

   ** Examine your two graphs. Their balance points are identical. **

3. The standard deviation of the sampling distribution, called the **standard error of the mean**, is less than the standard deviation in the population. If we are sampling from a distribution that is very large compared to the size of the sample ($n/N < 0.05$) or we are sampling with replacement as above, the standard error is found by:

$$\boxed{\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

   ** Verify this holds for the standard deviations you found above. **

As n becomes large, the standard error becomes small due to the $\sqrt{n}$ in the denominator which also gets larger. A small standard error means that little variation can be expected from one sample mean to the next. Any sample mean we observe for purposes of estimating μ is expected to have little estimation error and so is a good estimate.[8]

Example:
Suppose a population variable has a standard deviation σ of 12 cm and a mean μ of 115 cm. Random samples are drawn from this population for the purpose of analyzing the expected variability in sample means from 115 cm. Compute the standard error of the mean, $\sigma_{\bar{x}}$, if samples of size 50, 100, 500, and 5000 are drawn from the population. Assume we are sampling without replacement but that the population is large with respect to the sample size, $n/N < 0.05$.

Solution:
Since all of our sample sizes (n) are 30 or more, the C.L.T. states that the samples will have means that are normally distributed. The fact that

$$\frac{n}{N} = \frac{n}{\text{“large”}} \approx 0 < 0.05$$

means we can use the formula above for the standard error of the mean.

[5] In this example having a different number of occupants in each household will make our visualization of the samples easier, but it is unnecessary. Try to think how you would approach the analysis if two of the houses had the same number of occupants.

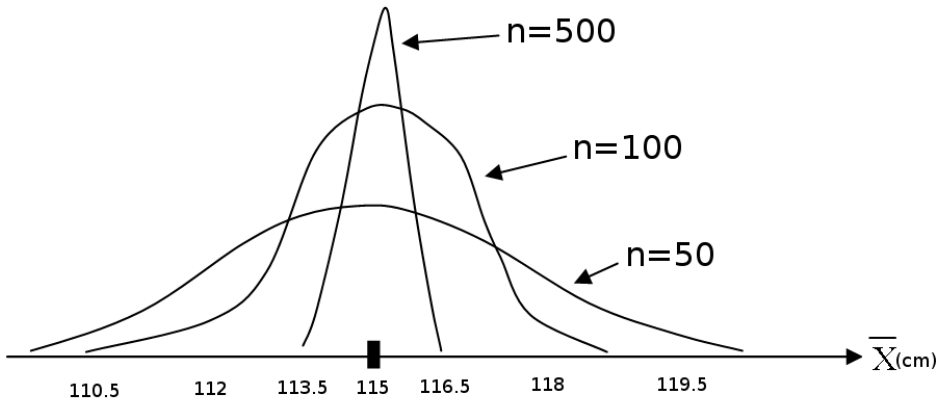
[6] See problem 4 on page 172 for an analysis of the problem sampling without replacement.

[7] In a sampling distribution, if the sample statistic $\bar{x}$ is itself a *variable*, what is the *population* it is measuring?

[8] In other words, the larger the sample size, the better the estimate. This is known as the **law of large numbers.**

n	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (cm)
50	$\frac{12}{\sqrt{50}} = 1.70$
100	$\frac{12}{\sqrt{100}} = 1.20$
500	$\frac{12}{\sqrt{500}} = 0.54$
5000	$\frac{12}{\sqrt{5000}} = 0.17$

If we plotted several of these distributions they would look as follows:



As n becomes large the standard error becomes small. A small standard error means that little variation can be expected from one sample mean to the next. The larger sample size ensures that any sample mean we measure is much more likely to lie near the unknown population mean we are trying to estimate. In other words, any sample mean we observe for purposes of estimating μ is expected to have little estimation error. This is why we prefer larger samples.

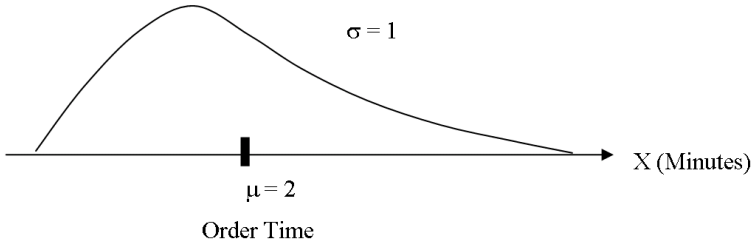
4.4 The Central Limit Theorem Applied

To aid in understanding the C.L.T. we can solve conceptual problems involving the likelihood of finding a given sample statistic if the population parameters are known. In practice the latter is not usually the case – we are taking a sample to estimate an unknown population parameter. We will see in later sections how we can use the C.L.T. to our advantage in this estimation and for hypothesis testing. For now, however, consider the following simpler conceptual problem.

Example:
A fast food service has a very large clientele. On the average it takes 2 minutes with a standard deviation of 1 minute to take an order from a customer. Suppose a simple random sample of 100 customers is observed, what is the probability that the average time $\bar{x}$ it takes to take a customer's order within the sample is:

1. Between 2 and 2.1 minutes?
2. Less than 1.8 minutes?
3. Over 3 minutes?

Solution:
Notice that there is no information about the shape of the distribution of times (x) to take a customer's order among all of the clientele. It may be skewed or bimodal. It could look like this:



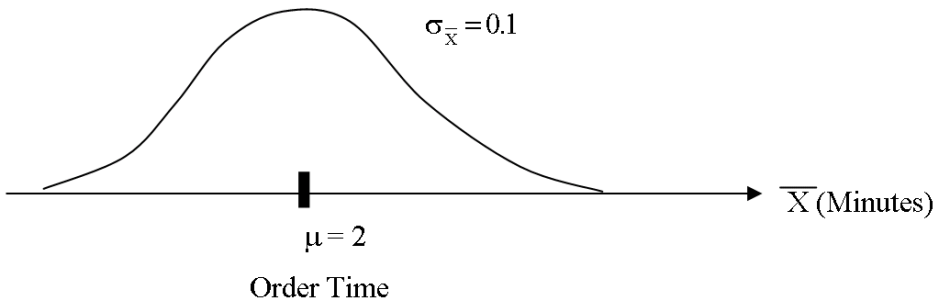
Since $n = 100 \geq 30$ the C.L.T. guarantees us that if all possible samples of size 100 are taken from this population of times, the distribution of mean times from sample to sample will be a normal curve. The mean and standard deviation of this sampling distribution are known. The mean of the sampling distribution is the same as that of the population, namely

$$\mu_{\bar{x}} = \mu = 2 \text{ min} .$$

The standard deviation of the sampling distribution (the standard error of the mean) is less than the standard deviation of the population.[9]

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1 \text{ min}}{\sqrt{100}} = 0.1 \text{ min}$$

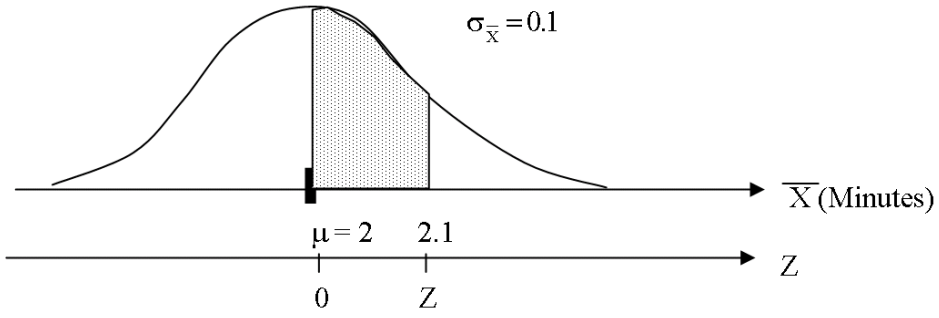
The sampling distribution of mean times would look like this:



Since the shape, mean, and standard deviation of the sampling distribution curve are known, the probability associated with an interval can be found by using the table of areas under the normal curve, where z on the sampling curve is found by:

$$\boxed{z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}}$$

1. For example, to find $P(2 \text{ min} < \bar{x} < 2.1 \text{ min})$, find the z-values for $\bar{x}$'s of 2 and 2.1 minutes. Since 2 minutes is the mean, its z-value is 0.00, i.e. $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{2 \text{ min} - 2 \text{ min}}{0.1 \text{ min}} = 0.00$.
   If $\bar{x} = 2.1$ minutes then $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{2.1 \text{ min} - 2 \text{ min}}{0.1 \text{ min}} = 1.00$.

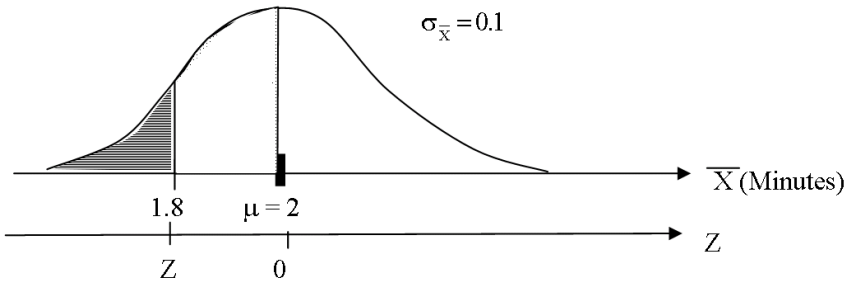


We have:

$P(2 \text{ min} < \bar{x} < 2.1 \text{ min}) = P(0.00 < z < 1.00) = P(z < 1.00) - P(z < 0.00) = 0.8413 - 0.5000 = 0.3413$

Caution: A common error is to use the population standard deviation rather than the standard error for the standard deviation in the z-value calculation.

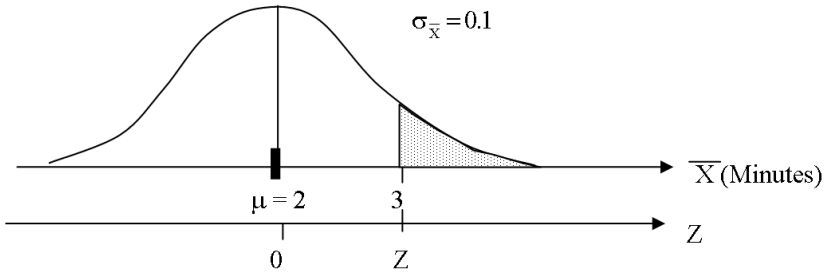
2.



If $\bar{x} = 1.8$ minutes then $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{1.8 \text{ min} - 2 \text{ min}}{0.1 \text{ min}} = -2.00$ and so:

$$P(\bar{x} < 1.8 \text{ min}) = P(z < -2.00) = 0.0228$$

3.



If $\bar{x} = 3$ minutes then $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \dfrac{3 \text{ min} - 2 \text{ min}}{0.1 \text{ min}} = 10.00$, therefore:

$$P(3 \text{ min} < \bar{x}) = P(10.00 < z) = P(z < -10.00) \approx 0.0000$$

Notice that if we had just asked the probability that a single order (x) was longer than 3 minutes the answer would have been, if the underlying population variable x were normally distributed,

$$P(3 \text{ min} < x) = P(1.00 < z) = 0.1587 \ .$$

The result here is significantly different than that of the sample mean due to the use of the standard deviation when calculating z. I.e. $z = \dfrac{x - \mu}{\sigma} = \dfrac{3 \text{ min} - 2 \text{ min}}{1 \text{ min}} = 1.00$.

[9]The C.L.T. standard error of the mean formula applies here because it is stated that the population is very large so the sample is of insignificant size ($< 5\%$) relative to the population. (It does not matter then whether we sample with replacement or not; in practice we probably would not.) Note that the underlying population here is the orders, not the customers so a large clientele implies a large number of orders. Even a pub with only 20 regular customers would still have a large number of orders (effectively infinite) from which to draw our sample.

4.5 Extending the Central Limit Theorem

The central limit theorem describes the behaviour of the distribution of sample means. In the previous examples, it has been demonstrated that the sample mean is itself a random variable. When one sample is drawn from a population, the only information that is known about the population mean is that the mean of this sample is one of the possible sample means. If the conditions of the C.L.T. are met, then this sample mean will lie on a normal distribution curve that has the population mean as its centre and standard deviation equaling the standard error of the mean.

The central limit theorem as stated requires a sample size of $n \geq 30$. When sampling without replacement we also have required a population that is much larger ("infinite") compared to the sample size ($n/N < 0.05$). If these conditions do not hold then some adjustments must be made to the statements in the central limit theorem.

The sampling distribution of the mean is always symmetric in shape regardless of the size of the sample or the shape of the population from which the sample is drawn. If the sample size is thirty or more and the population is infinite the distribution of sample means will also be approximately normal regardless of the shape of the population distribution from which the sample is drawn. However sometimes large samples are very expensive or impossible to get.

If the sample size is less than 30, the sampling distribution of the mean is no longer normal. However if the variable x itself is known to have an approximate normal distribution then the sample mean $\bar{x}$ also will be normally distributed. However, for reasons that will be discussed in Section 5.1.4 the practical small sample size problems we will be interested in ultimately require the use of what is called a ***t* distribution**.

If the C.L.T.'s conditions are met (large sample), one could wonder why we do not use a t distribution even then if the results are more accurate. In principle we could and some do. However, in this course the use of a normal distribution for large sample size will be done both because it is easier and also to give a clearer understanding of what is going on.

Now consider the second assumption of an infinite population ($n/N < 0.05$). Since the C.L.T. as stated will always apply if we sample with replacement, regardless of how big the underlying population is, we will, for simplicity in this course, **always assume that we are sampling with replacement** unless otherwise stated. However this assumption is not practical in many cases. If you were taking a sample of order wait times it is unlikely you would select them in such a way that you could pick the same order more than once. Fortunately, as in the latter case, sampling without replacement would still be fine since the population size is effectively infinite ($n/N < 0.05$) so our C.L.T. still applies to good approximation. However, what happens when we do want to sample without replacement and the population is not infinite in size. It turns out that a slight modification to the C.L.T. is required. We need to introduce a finite correction factor (*F.C.F.*) when calculating the standard error. This is discussed in Section *4.7 . That section is made optional since our assumption of always sampling with replacement ensures we will never need to use the *F.C.F.* in the course, in the event our populations are finite ($n/N \geq 0.05$). Practical considerations in real-world problems may require it however if the sample is drawn without replacement.

Assignment:

For any conceptual problem involving application of the C.L.T. :

⇒ Draw a Venn diagram with labeled parameters and statistics.
⇒ Draw a sampling curve with appropriate labels and the region that answers the question shaded.
⇒ Use the proper symbols for sampling distributions and probability.

1. The average income in a certain country is $9600 with a standard deviation of $640. A simple random sample of 1024 of the citizens of this country is selected and their average income is computed.
   (a) Draw a diagram of the sampling distribution of the mean.
   (b) What is the probability of observing a sample average within $50 of the population mean? (i.e. between $9550 and $9650)
   (c) What is the probability of observing a sample average that is more than $60 higher than the population mean? (i.e. above $9660)
2. The average weight of a vehicle traveling on a municipal road is 2500 kg with a standard deviation of 700 kg. A simple random sample of 40 vehicles traveling the road is selected.
   (a) Draw a diagram of the sampling distribution of the mean.
   (b) What is the probability of observing a sample mean within 200 kg of the average weight? (i.e. between 2300 kg and 2700 kg)
   (c) What is the probability of observing a sample mean that differs from the average weight by more than 500 kg? (i.e. below 2000 kg or above 3000 kg)
3. A personnel test is administered to a large number of applicants. A simple random sample of 100 scores is selected from the group of all test scores. The average test score for all people writing the test is 75% with a standard deviation of 10%.
   (a) What is the probability the sample mean will be above 76%?
   (b) Between what two values would the average for the sample fall in 95 cases out of 100?

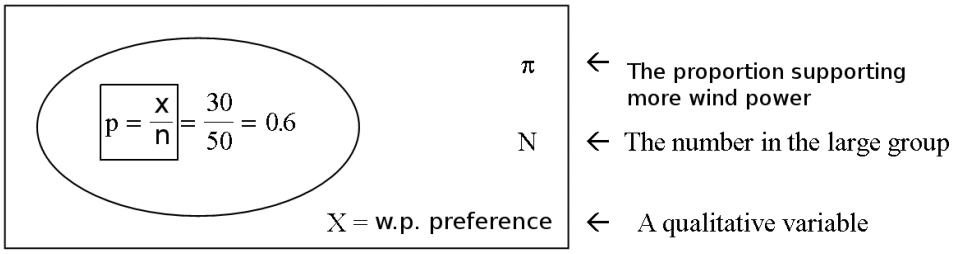
4.6 Sampling Distribution of Proportions

Every sample statistic has a sampling distribution. That is, for any given sample statistic, there is bound to be some variation observed from sample to sample in the sample statistic. The C.L.T. describes the behaviour of the variation among sample means. The distribution of sample modes, sample medians or sample standard deviations could be analyzed in a similar manner. Each of these distributions would have the same attributes related to shape, centre and variability.

Another statistic of importance whose value varies from sample to sample is the sample proportion p. The symbol for the population parameter is π.

Example:
Suppose a simple random sample of 50 people is selected from a large group of people. It is found that 30 people in the sample are in favour of more wind power generation. Estimate the proportion of the population in favour of more wind power generation.

Solution:



Is it likely that $p = \pi$? If we took another sample of size 50 are we likely to observe the same sample proportion p?

Large samples are always used to estimate population proportions. For samples of fixed size n, the distribution of the sample proportion has the following three features.

1. The distribution of p approaches a normal curve as n increases. The normal approximation is valid provided $n\pi > 5$ and $n(1-\pi) > 5$.
2. The mean of all sample proportions is the population proportion, π:
$$\mu_p = \pi$$
3. If the population size is large ($n/N < 0.05$) or we are sampling with replacement, the standard deviation of sample proportions, called the **standard error of proportion**, is found by[10]
$$\boxed{\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}} \quad \Leftarrow \text{ Note same } \sqrt{n} \text{ in denominator as for means!}$$

[10]The similarity to the standard deviation formula of the binomial probability distribution is no coincidence. We could have considered this as a binomial problem where we were counting x, the number in favour of wind power, out of n selected. This x would have mean $\mu = n\pi$ and standard deviation $\sigma = \sqrt{n\pi(1-\pi)}$. The proportion in favour, p, is just x/n. Dividing the variable x by a constant (here n) results in the new variable (here p) having a new mean and standard deviation equal to the old values divided by that same constant. We thus have $\mu_p = \mu/n = (n\pi)/n = \pi$. Similarly $\sigma_p = \sqrt{n\pi(1-\pi)}/n = \sqrt{n\pi(1-\pi)/(n^2)} = \sqrt{\pi(1-\pi)/n}$. That p is normally distributed for $n\pi > 5$ and $n(1-\pi) > 5$ likewise follows from the fact that the binomial distribution for x is approximately normal under these constraints. The connection between a binomial probability and a sample proportion is demonstrated in problem 3 on page 169 .

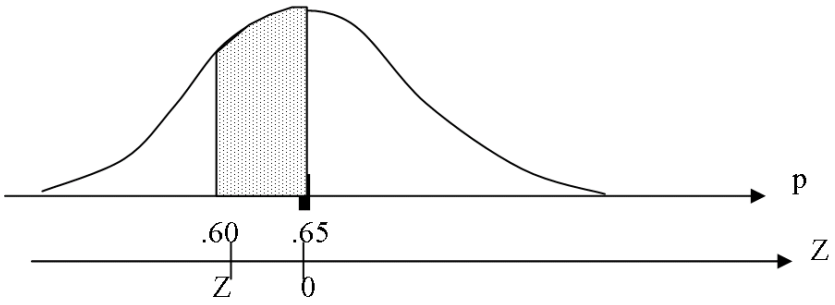
Example:
In the previous example, suppose that π, the proportion of the population who support more wind power, is 0.65. What is the probability of finding a sample proportion, p, between 0.60 and 0.65?

Solution:
Since we have that $n\pi = (50)(.65) = 32.5 > 5$ and $n(1-\pi) = (50)(.35) = 17.5 > 5$, p should be approximately normally distributed.

The mean of the sampling distribution curve is the same as the population proportion, namely 0.65 . Since the population is assumed large, the standard error of proportion is:

$$\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = \sqrt{\frac{(0.65)(0.35)}{50}} = 0.0675$$

Drawing the sampling distribution curve for the statistic we have:



Next we compute the z-values that correspond to $p = 0.60$ and $p = 0.65$. For $p = 0.60$,

$$\boxed{z = \frac{p-\pi}{\sigma_p}} = \frac{0.60 - 0.65}{0.0675} = -0.74 \ .$$

Since 0.65 is the mean, its z-value is just 0.00:

$$z = \frac{p-\pi}{\sigma_p} = \frac{0.65 - 0.65}{0.0675} = 0.00$$

Using proper probability symbols the answer is:

$$P(0.60 < p < 0.65) = P(-0.74 < z < 0.00) = P(z < 0.00) - P(z < -0.74) = 0.5 - 0.2296 = 0.2704$$

Caution: A common confusion here is calculating the value of the standard deviation of this sampling distribution using p rather than π.

When sampling **without replacement** if N is not "infinite" (when $n/N \geq 0.05$) we must introduce a finite correction factor when calculating σ_p as we did before for the standard error of the mean $\sigma_{\bar{x}}$.[11] This will be discussed in Section *4.7 . For simplicity for this course **we will assume**, as was done for means, that in the event a population is finite ($n/N \geq 0.05$) that we have been **sampling with replacement**. As such no such *F.C.F.* correction will be required in the course.

[11] Recall that when sampling without replacement binomial trials are only approximately independent if the population is large ($n/N < 0.05$). When the population is finite this is no longer the case and we have to consider x as having a hypergeometric distribution. Using the formulae for its standard deviation one may show that upon division by n to get σ_p, the *F.C.F.* appears as given in Section *4.7 .

Means or Proportions?

For the rest of the course we will be dealing with problems involving either sample means, $\bar{x}$, or sample proportions, p. To identify the type of problem that you are working with, it is helpful to note that in sample mean problems the underlying variable x is always quantitative, since it is something that has to be averaged. We are measuring, say, a height, or weight. For a proportion problem, the underlying variable x is qualitative since we are calculating the proportion of the population that satisfies some criterion. We are asking the individual, say, whether they are in favour of a war, or not. As such the values of the variable are like a binomial trial, either $x =$ success (satisfies the criterion) or $x =$ failure (does not satisfy the criterion).[12] Finally note that as a proportion, it must be the case that p be dimensionless and satisfy $0 \leq p \leq 1$.

Assignment:

1. In a recent plebescite, 40.0% of the citizens of Saskatoon supported a new animal control bylaw. Suppose a survey of 200 of the voters had been taken prior to the vote.
   (a) If the proportion who support the bylaw of the 200 polled is calculated, what is the standard error of proportion of such sample proportions?
   (b) What is the probability that the poll of 200 would have shown more than 50.0% support for the bylaw? (Include a sketch of the sampling distribution of p in your solution, labeling the p and z axes and shading the relevant area.)
2. Currently the ruling political party in a large country is supported by 45% of the voters. If a simple random sample of 1024 voters is chosen in this country,
   (a) What is the probability that the proportion in the sample in favour of the ruling party is less than 46%?
   (b) What is the probability that the proportion in the sample in favour of the ruling party differs by more than 2% from the population proportion? (i.e. p less than 43% or greater than 47%)

3. A large box of batteries contains 20% that are defective. A simple random sample of 30 batteries is chosen. What is the probability of finding from 3 to 6 batteries in the sample defective
   (a) Using the normal approximation for a sample proportion? (Remember $p = x/n$.)
   (b) Using an exact binomial probability formula calculation?

   Your answers should differ considerably. One reason is that we are approximating the discrete binomial probability distribution with a continuous distribution. For instance if we wanted to evaluate the probability of exactly $x = 3$ defective batteries we would need to evaluate the area under the continuous normal curve for the proportion range corresponding to 2.5 to 3.5 defective batteries.
   (c) To improve on our approximation, redo part (a) using the proportions for 2.5 to 6.5 defective batteries instead and compare this with the actual answer in (b). Modifying intervals in this way is another example of a *continuity correction*.

[12] To simplify the notation when dealing with proportions we will tend to write $p = \frac{x}{n}$ where now x means the number of sample elements satisfying the criterion. As such our usage of x here reflects the count that would appear in a binomial distribution.

*4.7 Finite Populations

If the population cannot be considered to be infinite, that is when the **sample size constitutes 5% or more of the population** ($n/N \geq 0.05$), and if we are sampling **without replacement** we must modify our previous formulae for the standard error of the mean, $\sigma_{\bar{x}}$, and the standard error of proportion, σ_p. In this case the **finite correction factor (*F.C.F.*)** is applied to the relevant standard error:

$$\boxed{\sigma_{\bar{x}} = (F.C.F.) \cdot \frac{\sigma}{\sqrt{n}}} \qquad \boxed{\sigma_p = (F.C.F.)\sqrt{\frac{\pi \cdot (1-\pi)}{n}}},$$

where

$$\boxed{F.C.F. = \sqrt{\frac{N-n}{N-1}}}.$$

Since the $F.C.F. \leq 1$, its inclusion makes the standard error smaller. This makes sense since, in sampling without replacement, we are guaranteed to have measured the same number or more distinct elements of the population than when sampling with replacement for the same sample size n. Note that $n/N < 0.05$ for a population to be considered "infinite" or "large" is the same condition that was used when we discussed the large population limit of the binomial distribution in Section 3.1.2 .

Example:
Repeat the example from page 160 ($\sigma = 12$ cm, $\mu = 115$ cm) but now with a population of size $N = 10,000$. Compute the standard error of the mean if samples of size 50, 100, 500, and 5000 are drawn without replacement from the population.

Solution:

n	n/N	$F.C.F. = \sqrt{\frac{N-n}{N-1}}$	$\sigma_{\bar{x}} = (F.C.F.) \cdot \frac{\sigma}{\sqrt{n}}$ (cm)
50	.005	$\sqrt{\frac{9950}{9999}} = .9975$	$\sqrt{\frac{9950}{9999}} \cdot \frac{12}{\sqrt{50}} = 1.69$
100	.010	$\sqrt{\frac{9900}{9999}} = .9950$	$\sqrt{\frac{9900}{9999}} \cdot \frac{12}{\sqrt{100}} = 1.19$
500	$\boxed{.050}$	$\sqrt{\frac{9500}{9999}} = \boxed{.9747}$	$\sqrt{\frac{9500}{9999}} \cdot \frac{12}{\sqrt{500}} = \boxed{0.52}$
5000	$\boxed{.500}$	$\sqrt{\frac{5000}{9999}} = \boxed{.7071}$	$\sqrt{\frac{5000}{9999}} \cdot \frac{12}{\sqrt{5000}} = \boxed{0.12}$

For n small relative to the size of N, that is for $n/N < 0.05$, the size of the *F.C.F.* is approximately 1, and as a result makes very little difference to the value of the standard error from the original example. In these two cases ($n = 50$, $n = 100$) the population is effectively infinite and we can neglect the *F.C.F.* For the cases where $n/N \geq 0.05$ the *F.C.F.* differs significantly from 1 and introduces a significant modification to $\sigma/\sqrt{n}$ as can be seen by comparing the entries (boxed) for $n = 500$ and $n = 5000$ to their values in the original example. In these cases the finite size of the population is noticeable.

Note that had we sampled with replacement, no *F.C.F.* would have been required regardless of the sample size.

The following proportion example illustrates the *F.C.F.* being used in an actual problem involving sampling without replacement.

Example:
Repeat the windpower example on page 168, but now assume the population under consideration is a town of 800 that is contemplating putting up wind turbines and that we are taking the sample of size 50 without replacement. If 65% of the town are actually in favour of more wind power, what is the probability of finding a sample proportion, p, lying between 0.60 and 0.65?

Solution:
The true proportion in favour is still $\pi = 0.65$. Now, however, the population is not infinite since

$$\frac{n}{N} = \frac{50}{800} = 0.0625 > 0.05 \ .$$

Since we are sampling without replacement the required finite correction factor is

$$F.C.F. = \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{800-50}{800-1}} = \sqrt{\frac{750}{799}} = 0.9689 \ ,$$

so the standard error in proportion is

$$\sigma_{_p} = (F.C.F.) \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = (0.9689)\sqrt{\frac{(0.65)(.35)}{50}} = 0.0654,$$

and our probability becomes

$$\begin{aligned} P(0.60 < p < 0.65) &= P\left(\frac{0.60-0.65}{0.0654} < z < \frac{0.65-0.65}{0.0654}\right) \\ &= P(-0.76 < z < 0.00) = 0.5 - 0.2236 \\ &= 0.2764 \end{aligned}$$

Note that when sampling without replacement you can always include the $F.C.F.$ to be more accurate, even when the population is effectively infinite. In the latter case, however, the modification to the answer will be slight. For many problems, such as a measured result of an experiment, the population size N really is infinite, since one can in theory repeat the experiment as many times as one wishes. In that case the $F.C.F.$ is exactly 1 and can be ignored.

Since we will be assuming sampling with replacement for the rest of the course no $F.C.F.$'s will appear in our solutions. In practice, however, many of the problems (confidence intervals and hypothesis tests) where the population is finite in Unit 5 would require an $F.C.F.$ if we sampled without replacement. Some answers at the back of the text will also indicate the solution had the problem involved sampling without replacement.

Assignment:

The following Section *4.7 problems illustrate the need for a finite correction factor (*F.C.F.*) when sampling without replacement from finite populations.

1. Repeat problem 3 on page 166 but now assume the test was administered to only 500 applicants and assume the sample of 100 was taken without replacement. Take other aspects of the problem to be the same.

2. Repeat problem 2 on page 169 but modify it so that now rather than talking about "the ruling political party in a large country" we are talking about "the ruling faction of Melville town council". Assume all other aspects of the problem are the same and we are sampling without replacement. Take the eligible voter population of Melville, Saskatchewan to be 3800 .

3. What happens to the standard error of the mean, $\sigma_{\bar{x}}$, when

   (a) The sample size is just one, $n = 1$?

   (b) The sample is the whole population, $n = N$? (Assume sampling without replacement.)

   Explain your answers.

4. Recall the example on page 158 of sampling two of the four houses in the village to measure the number of occupants. Repeat the analysis but now sample without replacement:

   (a) What are the ${}_4C_2 = 6$ possible samples?

   (b) What is the mean of the sample means, $\mu_{\bar{x}}$? Compare it to the population mean, μ.

   (c) What is the standard error of the mean, $\sigma_{\bar{x}}$? Show that it satisfies the equation for $\sigma_{\bar{x}}$ of the C.L.T. if the finite correction factor is used.

   (d) Which of the conclusions of the C.L.T. are not true in this example?

Sampling Distribution Formulae

Standard Error Formulae for Single Means and Proportions

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \sigma_{p} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

$$^{*}\textbf{Finite Correction Factor} \quad \Rightarrow \quad {}^{*}F.C.F. = \sqrt{\frac{N-n}{N-1}}$$

Standard Scores for Single Means and Proportions

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad z = \frac{p - \pi}{\sigma_{p}}$$

Other Formulae

$${}_{N}C_{n} \quad \text{or} \quad N^{n} \qquad p = \frac{x}{n}$$

Unit 5: One Population Inferences

5.1 Point Estimates and Confidence Intervals

5.1.1 Confidence Intervals for the Population Mean (Large Samples)

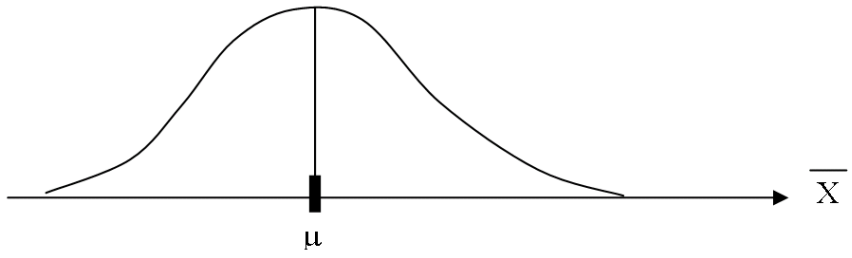
The previous unit has demonstrated the theoretical working of the C.L.T. for sample means and the equivalent theorem for sample proportions. We will modify this slightly to apply the theory in a practical setting.

In all of the sampling theory developed so far, the assumption was made that the population parameters were known. For example, to compute $\mu_{\bar{x}}$ or $\sigma_{\bar{x}}$, the assumption is made that μ and σ are known. If this were the case, there would be no need to sample because one would never use an estimate if an exact value were available.

In practice only one sample is selected. This means that we have exactly one point known to us on the whole sampling distribution curve. The exact position of this point on that curve is also unknown to us.

Example:
A simple random sample of 50 wage earners from a population of 10,000 is selected. The average wage in the sample is \$42,000 with a standard deviation of \$5000. Suppose we wish to estimate μ. If another sample of size 50 was chosen, is it likely that we would see the same sample statistics?

Solution:
Because the sample size exceeds 30, we know that $\bar{x}$ is approximately normally distributed:



We do not know whether to place the $\bar{x}$ we observed to the right or to the left of μ. The best we can say is that $\mu \approx \bar{x} = \$42,000$. Another sample would lie elsewhere.

A sample statistic like $\bar{x}$ is called a **point estimator** of the population parameter μ because it generates estimates of the latter from samples. Several population parameters with corresponding unbiased point estimators are:

Parameter	Point Estimator
μ	$\bar{x}$
σ	s
π	p

The estimator is the symbol and its corresponding formula. A value of the point estimator (i.e. the number in a particular situation) is called a **point estimate**. (e.g. $\bar{x} = \$42,000$ is a point estimate of the mean.)

The problem with a point estimate is that there is no way to assess how close it is to the population value nor how much confidence to place in it. To satisfy these criteria, we need **interval estimators**

with their corresponding **interval estimates**.

Example:
A **confidence interval estimate** for the last example is:

$$P(\,\$40,614.06\ < \mu <\ \$43,385.94\,) = 0.95$$

In words there is a 0.95 probability that the population mean μ is between \$40,614.06 and \$43,385.94 . Some terms that are used in connection with the confidence interval are:

Lower Limit: \$40,614.06

Upper Limit: \$43,385.94

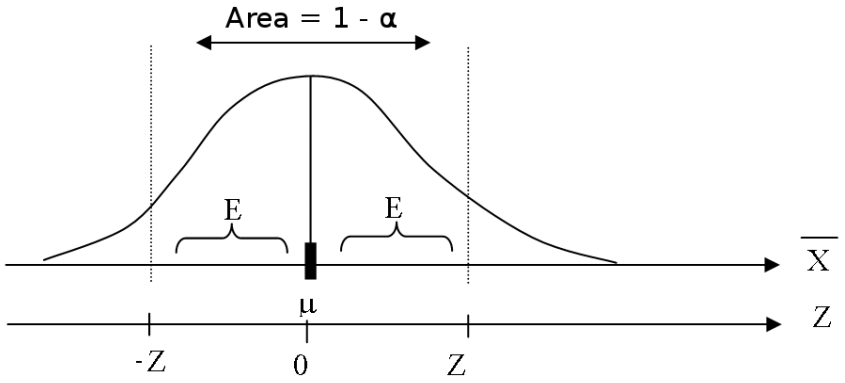
Confidence Coefficient: 0.95 ($\Leftarrow$ How confident probability-wise we are that the population parameter lies in the interval.)

To make the point estimate (the centre of the interval) explicit we may also write:

$$P\,(\,\$42,000 - \$1385.94\ < \mu <\ \$42,000 + \$1385.94) = 0.95$$

The number \$1385.94 is called the **maximum error of estimate** and indicates how close we expect our estimate to be.

We now examine the process of creating an interval estimate.



The above diagram shows that $1-\alpha$ of all sample means are within a distance of $E = z \cdot \sigma_{\bar{x}}$ of the mean.[1] In probability symbols:

$$\boxed{P\,(\,\bar{x} - E < \mu < \bar{x} + E\,) = 1 - \alpha}$$

This probability event is called a confidence interval estimator for the population mean. This interval estimator $\boldsymbol{\bar{x} - E < \mu < \bar{x} + E}$ has a **maximum error of estimate**[2] of $\boldsymbol{E}$ and a **confidence coefficient** of $\boldsymbol{1-\alpha}$.[3]

[1] That $E = z \cdot \sigma_{\bar{x}}$ follows from solving the z-score formula $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ with $\bar{x} = \mu + E$ (see diagram) for E.

[2] The maximum error of estimate is also called the **precision**, **maximum sampling error**, or the **margin of error**. Since it depends on the actual confidence level, as we shall see, it may be more precisely preceded by that percentage. For example the 98% maximum error of estimate.

[3] The term **confidence level** or **level of confidence** is used if the confidence coefficient (say $1-\alpha = 0.95$) is reported as a percentage (95%) rather than a proportion (0.95).

The odd choice of symbol for the confidence coefficient, $1-\alpha$, rather than say just "C", reflects the fact that we will often wish to know to the total area in the two tails (α). If the confidence coefficient (the central area) is, say, 98% so $1-\alpha = 0.98$, then it follows that $\alpha = 1-0.98 = 0.02$ is the total area in the two tails. A deeper motivation for this notation will be revealed in Section 5.3. For now just think of $1-\alpha$ as a single symbol for the confidence coefficient.

Example:
Using the information given in the income example on page 176, construct the 95% confidence interval for μ found in the last example.

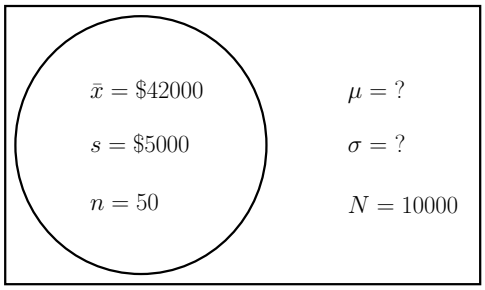
Solution:
Keeping in mind that the goal is to fill in the blanks of the confidence interval statement,

$$P(\,\bar{x}-E<\mu<\bar{x}+E\,)=1-\alpha\ ,$$

proceed with the following steps for any confidence interval:

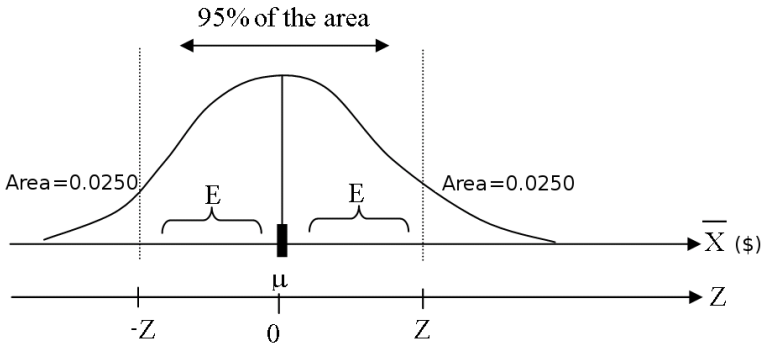
Step 1) Identify all given information with symbols.

This is conveniently done with a Venn diagram:



Step 2) Draw a diagram of the sampling distribution.

Since $n=50\geq 30$ the distribution of $\bar{x}$ is approximately normal by the C.L.T. .



Step 3) Determine the z-value.

Use the confidence, here $1-\alpha=0.95$, to find the total area in the two tails, $\alpha = 1-0.95 = 0.05$. The area in a single tail is

$$\frac{\alpha}{2}=\frac{0.05}{2}=0.0250\ ,$$

and the corresponding z-value is (for the right tail)

$$z = 1.96 \ ,$$

using the normal table.

Step 4) Calculate the standard error.

To calculate the maximum error of estimate E, we require the standard error of the mean, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, but σ is unknown so replace it by its point estimator, s. We designate the approximate standard error of the mean as

$$\boxed{s_{\bar{x}} = \frac{s}{\sqrt{n}}} = \frac{\$5000}{\sqrt{50}} = \$707.11$$

Since $\frac{n}{N} = \frac{50}{10000} = 0.005 < 0.05$ the standard error formula is valid even had we sampled without replacement.

Step 5) Calculate the maximum error of estimate.

$$\boxed{E = z \cdot s_{\bar{x}}} = (1.96)(\$707.11) = \$1385.94$$

Step 6) Make the confidence interval statement.

$$\begin{aligned} P\,(\,\bar{x} - E \ &< \ \mu \ < \ \bar{x} + E\,) = 1 - \alpha \\ P\,(\,\$42,000 - \$1385.94 \ &< \ \mu \ < \ \$42,000 + \$1385.94\,) = 0.95 \\ P(\,\$40,614.06 \ &< \ \mu \ < \ \$43,385.94\,) = 0.95 \end{aligned}$$

Again, in words, there is a 0.95 probability that the population mean μ is between \$40,614.06 and \$43,385.94.

Notes:

1. The reason we will always write a confidence interval statement which includes the interval along with the confidence is because the interval **depends on the confidence coefficient chosen**. To see this, repeat the above calculation at a 99% level of confidence.
(Answer: $P(\$40,179.19 < \mu < \$43,820.81) = 0.99$)

   Step 1) Identify all given information with symbols on a Venn diagram.

   Step 2) Draw a diagram of the sampling distribution.

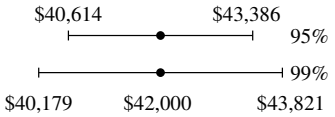
Step 3) Determine the z-value.

Step 4) Calculate the standard error.

Step 5) Calculate the maximum error of estimate.

Step 6) Make the confidence interval statement.

2. One does not get something for nothing when creating confidence intervals with the same sample data. Comparing our 99% confidence interval with the 95% confidence interval we see that if we want to be more confident about catching the population mean we have to have a wider interval (larger E).



   If you wanted a narrower interval with the same level of confidence what would you have to do?

3. When creating a confidence interval we choose the level of confidence. Usually we want to be fairly confident we have caught the population mean in the interval.[4] Standard confidence coefficients are in the following table, along with their corresponding z-value for normally distributed sample statistics.[5]

Confidence, $1-\alpha$	z	Confidence, $1-\alpha$	z
0.80	1.282	0.98	2.326
0.90	1.645	0.99	2.576
0.95	1.960	0.999	3.291

[4] A level of confidence of 30% would not be very useful information since the probability of being outside the stated interval would be 70%!

[5] These may be verified up to two digits after the decimal from the normal distribution table.

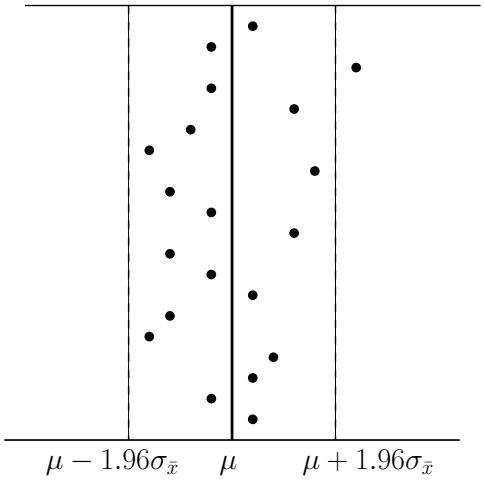
4. We stated that the generic confidence interval for the population mean is

$$P\,(\,\bar{x} - E < \mu < \bar{x} + E\,) = 1 - \alpha\ ,$$

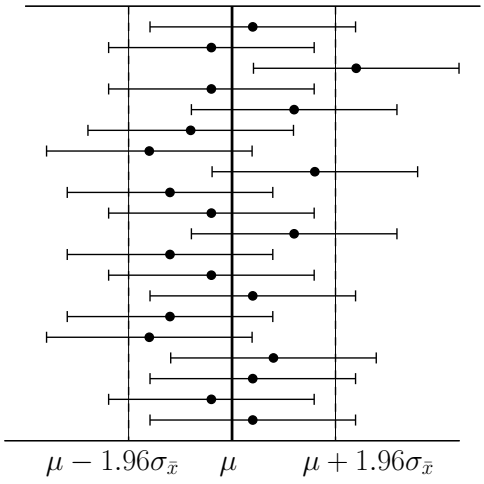
but looking at the sampling distribution it ought to read

$$P\,(\,\mu - E < \bar{x} < \mu + E\,) = 1 - \alpha\ \ !$$

What gives? To see why we can exchange the population and sample means consider the following diagram illustrating means of 20 possible samples. The 95% confidence interval around the population mean is indicated and as expected there is one sample out of 20 (the third) which does not lie in this interval while 19/20 or 95% lie within the interval.



In the next diagram we draw the confidence intervals centred on each sample mean.



One notices that 19 of the 20 sample confidence intervals (95%) contain the true population mean μ while only one (again the third) does not. As such for a sample chosen at random we have a 95% chance of containing the actual population mean within the confidence interval centred on the sample mean which is the statement we wanted to make. Note however that when drawing the sampling distribution, the variable is the sample statistic $\bar{x}$ not the population parameter μ. The latter is a fixed, if unknown, constant.

Assignment:

1. Mercury levels were tested in a sample of 36 Walleye at Threepoint Lake in northern Manitoba in 2005 and found to have a mean of 0.300 ppm (parts per million) with a standard deviation of 0.040 ppm. Construct a 98% confidence interval for the average concentration of mercury in Walleye in that lake for that year.

2. To estimate the average amount owing in a large journal of receivables, a random sample of 100 entries is selected from the entries. The average amount owing in the sample is \$512 with a standard deviation of \$58. Construct a 90% confidence interval for the average amount of an entry in the journal.

3. A community has 500 family units. An estimate of the average amount spent per household winterizing a family dwelling is required. 80 families are randomly selected. The average amount in the sample spent on winterizing a dwelling is \$85 with a standard deviation of \$15.

   (a) Construct a 95% confidence interval for the average amount of money spent on winterizing a dwelling.

   (b) Repeat (a) but for a 99% confidence interval.

4. A taxi company wishes to obtain an estimate of an average fare. A random sample of 20 customers shows that the average fare is \$19.50 with a standard deviation of \$4.80. Why can our knowledge of the C.L.T. not be used to construct a 90% confidence interval or any other confidence interval in this case?

5. A training seminar has been offered many times over the course of the life of a business to many of its employees. At the end of the seminar an exam is administered to measure the knowledge level of an employee at the end of the seminar. An estimate is to be obtained of the average score on the test from the large bank of test scores of all employees who have taken the test. A random sample of 180 scores from the files of the company show the average in the sample to be 82% with a standard deviation of 6%. Construct a 99% confidence interval for the average of all test scores.

5.1.2 Confidence Intervals for the Population Proportion

The same confidence interval idea also holds for proportions in addition to means. Ultimately we wish to make a confidence interval statement of the form

$$\boxed{P\,(\,p - E < \pi < p + E\,) = 1 - \alpha}\,,$$

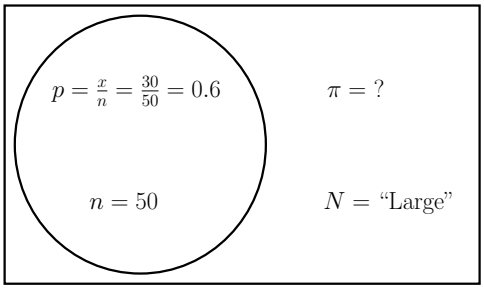
from which the steps below logically follow. The maximum error of estimate now depends on the standard error of proportion, $E = z \cdot \sigma_p$.

Example:
The example on page 167 dealt with a sample of size 50 randomly selected from a large group for the purpose of estimating the proportion of the group who support more wind power generation. Construct a 99% confidence interval for the proportion in the population who support more wind power if it is observed that there were 30 supporters in the sample.

Solution:

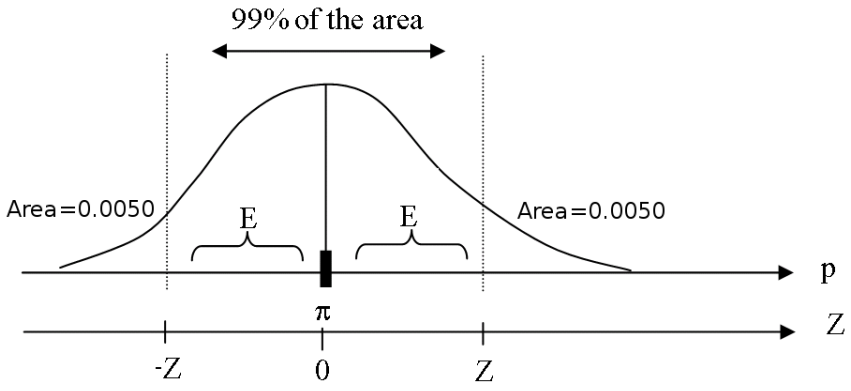
Step 1) Draw a Venn diagram labelling the information known.

It is also convenient to calculate the sample proportion here.



Step 2) Construct a sampling distribution diagram.

Since we are now assuming we do not know what the true population proportion is, we can use the the point estimate of $\pi \approx p = 0.6$ to confirm that, $n\pi \approx (50)(0.6) = 30 > 5$ and $n(1-\pi) \approx (50)(0.4) = 20 > 5$ so the distribution of sample p should be approximately normal.



Step 3) Determine the z-value.

Use the confidence, here $1-\alpha = 0.99$, to find the total area in the two tails, $\alpha = 1 - 0.99 = 0.01$. The area in a single tail is

$$\text{Area} = \frac{\alpha}{2} = \frac{0.01}{2} = 0.0050 \ ,$$

so from the normal table (selecting the value midway between $z = -2.57$ and $z = -2.58$ (since their areas are 0.0051 and 0.0049 respectively) the corresponding z-value is (for the right tail)

$$z = 2.575 \ .$$

Step 4) Calculate the standard error.

The standard error of proportion, in theory is found by $\sigma_{_p} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$. Since π is unknown in this case, π is replaced by its point estimator, p, and the approximate standard error of proportion, designated by $s_{_p}$, replaces $\sigma_{_p}$:

$$\boxed{s_{_p} = \sqrt{\frac{p \cdot (1-p)}{n}}} = \sqrt{\frac{(0.60)(1-0.60)}{50}} = 0.069 \ .$$

This formula is valid since, by assumption, we are sampling with replacement. However, even had we been sampling without replacement the formula is still approximately true since $\frac{n}{N} = 50/\text{“Large”} \approx 0 < 0.05$.

Step 5) Calculate the maximum error of estimate.

$$\boxed{E = z \cdot s_{_p}} = (2.575)(0.069) = 0.18$$

Step 6) Make the confidence interval statement.

$$\begin{aligned} P\,(\,p - E \ &< \ \pi \ < \ p + E\,) = 1 - \alpha \\ P\,(\,0.60 - 0.18 \ &< \ \pi \ < \ 0.60 + 0.18\,) = 0.99 \\ P(\,0.42 \ &< \ \pi \ < \ 0.78\,) = 0.99 \end{aligned}$$

In words there is a 0.99 probability that the population proportion supporting more wind power, π, is between 0.42 and 0.78 .

Assignment:

1. The Saskatchewan Ministry of the Environment has a surveillance program to monitor the incidence of Chronic Wasting Disease (CWD). Between 1997 and 2007 around 34,000 cervids (which include deer, elk, and moose) in Saskatchewan have been tested for CWD with 197 testing positive.

   (a) Will the sample proportion p of infected animals be approximately normally distributed?

   (b) Construct a 90% confidence interval for the proportion of cervids in Saskatchewan with CWD. (Assume the actual population size of cervids in Saskatchewan is in the millions.)

   (c) According to the Ministry of Environment (MOE) website:

   > "The CWD surveillance program is based primarily on the testing of hunter-killed animals and to a lesser extent on the testing of cervids submitted through the passive surveillance of sick or dead animals in the province. In areas where CWD was detected, MOE implemented herd reduction programs, primarily through increased hunting quotas and reduced hunting licence fees, in an attempt to increase hunter harvest to reduce deer densities and limit the spread of CWD."

   Does this affect the conclusion in (b)? If so, how?

2. A credit union branch has 650 members some of which have more than one type of account. A random sample of 50 of the members shows that 26 have a secondary account.

   (a) What is a point estimate for the proportion of members who have a secondary account at the credit union?

   (b) Construct a 99% confidence interval for the proportion of members who have a secondary account at the credit union.

3. A poll of 1000 people taken from a large population asking their voting intention gave the following results:

Response	Number
Pink Party	560
Blue Party	320
Green Party	90
Violet Party	30
	1000

   (a) Construct a 95% confidence interval for the proportion of people who support the Pink Party.

   (b) Based on your result in (a) can we conclude that the majority (more than half) of the population support the Pink Party?

   (c) If you wanted to be 99% confident that the Pink Party had majority support would your result in (b) still be true?

4. Astronomers now have techniques to observe planets orbiting other stars than our sun. In the nearby region of our galaxy, 55 stars were observed to see if they had a planet that could potentially support life (i.e. a rocky planet with the ability to have liquid water on the surface.) Suppose 22 of these stars were found to have such a "habitable" planet.

   (a) Construct a 95% confidence interval estimate for the proportion of stars in our galaxy which have a habitable planet.

   (b) Since $\pi = \chi/N$ where χ (Greek letter chi) is the actual number of habitable stars in the galaxy of $N \approx 200$ billion stars, multiply your confidence interval in (a) by N to arrive at a 95% confidence interval for the number of habitable planets, χ, in our galaxy.

5. In problem 4 part (b) it was shown that it is relatively easy to convert a proportion confidence interval into a confidence interval for χ, the total number of elements in the population having the attribute. Similarly, since $\pi = \chi/N$, we can also construct a confidence interval for population size N by solving our proportion interval for N as follows:[6]

$$p - E < \pi < p + E$$
$$p - E < \frac{\chi}{N} < p + E$$
$$\frac{1}{p-E} > \frac{N}{\chi} > \frac{1}{p+E}$$
$$\frac{\chi}{p-E} > N > \frac{\chi}{p+E}$$

And so our confidence interval for N becomes, writing the inequalities as less than:

$$P\left(\frac{\chi}{p+E} < N < \frac{\chi}{p-E}\right) = 1 - \alpha$$

Use this last result to solve the following problem.

Fisheries biologists are interested in determining the number of adult *yellow perch* in Christopher Lake, a lake found in the central part of Saskatchewan. The biologists capture and tag 200 of the adult yellow perch from the lake and then return them to it. Later they take a sample of 600 adult yellow perch from the lake and find that 8 have the tags. Construct a 90% confidence interval for the number of adult yellow perch found in the lake.

[6] Notice in the third line that when reciprocating an inequality involving like-signed terms (here positive) the inequality changes direction. Also when multiplying by χ in the following line the inequalities did not change direction because χ is positive.

5.1.3 Sample Size Determination

In all of the previous problems, it was assumed that a sample had already been taken and an estimate was to be based upon this sample. In a practical situation, a sample must first be selected before any of the sample statistics can be computed. The size of the sample is not just arbitrarily selected. There are three factors that determine the size of the sample to be selected both from a statistical point of view and from an economic point of view.

From an economic point of view, the sample size turns out to be costly because it involves the time, resources, and funds of the firm. The method of choosing the sample should minimize these costs while at the same time guaranteeing the statistical reliability and precision of the estimate provided.

From a statistical point of view, the size of the sample depends upon three things:

1. The maximum error of estimate, E, required. (The interval width is $2E$.)
2. The degree of confidence to be placed on the estimate.
3. The amount of variation in the population.

Recall that:

- In estimating **means**, the maximum error of estimate is found by:

$$E = z \cdot \sigma_{\bar{x}} = z \cdot \frac{\sigma}{\sqrt{n}}$$

  Solving this equation for n gives:

$$\boxed{n = \left[\frac{z \cdot \sigma}{E}\right]^2}$$

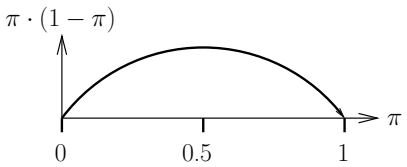
- In estimating **proportions**, the maximum error of estimate is found by:

$$E = z \cdot \sigma_p = z \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

  Solving this equation for n gives:

$$\boxed{n = \pi \cdot (1-\pi) \cdot \left[\frac{z}{E}\right]^2}$$

Note here that E is the desired maximum error of estimate for the interval; E is not calculated. The value of z is determined from the desired confidence for the interval. Finally, notice that in both the case of estimating means and of estimating proportions, the formula developed requires the value of a parameter in the population. The parameters in the population are unknown to us so an estimate of these parameters must be available before the sample size is determined and the sample taken. This estimating is often done by using a small pilot survey or in the case of proportions the worst case can be assumed, namely where $\pi = 0.5$. The latter maximizes the function $\pi \cdot (1-\pi)$:



Note the sample size n must be an integer and to be conservative we **always round up** for n regardless of the decimal fraction.

Example:
A survey is planned to determine the amount of time computer-owning adults in Regina spend on their home computers during a week. A pilot survey of such adults found that 8 hours were spent on average with a standard deviation of 3 hours. If we want to estimate the actual average weekly home computer use to within 1/4 of an hour, how many should be surveyed if we want to be 95% confident of the result? (Answer: $n = 554$ computer-owning Regina adults)

Solution:

Example:
In the example on page 183 a sample of 50 was taken from a large population and it was found that 30 were supporters of more wind power generation. The 99% confidence interval for the proportion of support based on this small sample was quite wide, namely
$([0.60 - 0.18] < \pi < [0.60 + 0.18]) = (0.42 < \pi < 0.78)$.

1. How large a sample should be taken if we desire to know the true proportion to within 5%, still at a 99% level of confidence. Use the small sample proportion 0.60 as an estimate of π. (Answer: $n = 637$ people, if $z = 2.575$ is used.)
2. How large a sample should be used if no estimate of π had been available? (Answer: $n = 664$ people)

Solution:

1.

2.

Assignment:

1. A manufacturing company produces a component for a product on an assembly line. The company wishes to know the average time taken for the component to travel from one workstation on the line to the next. By taking a handful of such measurements a worker has estimated that the mean time taken is 202.0 seconds with a standard deviation of 5.0 seconds. What sample size should be used in the actual analysis in order to be 95% certain that the mean transfer time is estimated to within 1.0 seconds in the final confidence interval?

2. What sample size should a government official employ if he wishes to estimate the proportion of citizens of Quebec who wish to secede from Canada to become a sovereign country to within 1% at a 99% level of confidence. (The last Quebec referendum in 1995 was 49.42% in favour of leaving Canada.)

3. In order to assess the cost of commuting daily between Regina and Moose Jaw the average mileage of carpool vehicles is required. A pilot survey of six such vehicles found the following mileages:

   $$6.8,\ 8.1,\ 8.5,\ 9.4,\ 12.3,\ 13.6 \quad \text{(km/litre)}$$

   If the average mileage is desired to be known to within 0.5 km/litre at a 90% level of confidence, how many car mileages should be measured in the actual survey?

4. In an attempt to assess the status of women in nineteenth century Paris, a researcher wishes to know the proportion of burial plots occupied by women in the Père Lachaise cemetery, a large affluent necropolis in Paris containing over 300,000 burial plots.

   (a) If the researcher wishes to know the fraction of female burials to within 4% at a 95% level of confidence, how many nineteenth century burial plots will she need to examine in the cemetery?

   (b) In an attempt to lessen the number of plots requiring examination, the researcher chooses 20 nineteenth century plots at random, looks at the names, and determines that 6 are women. How large a sample size would she need now to achieve a confidence interval with the requirements given in (a)?

   (c) The researcher decides that this is still a larger sample size than she wishes to take. What other options does she have to reduce the sample size required?

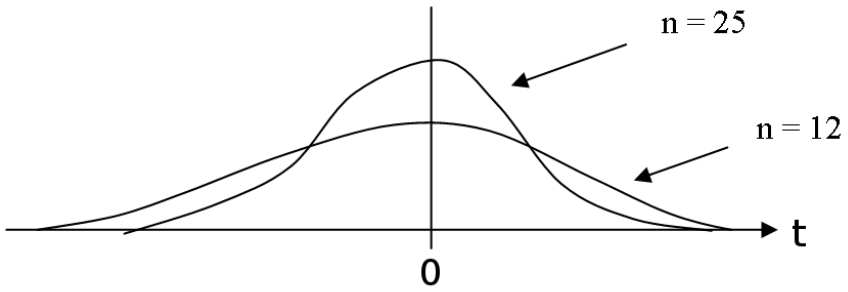
5.1.4 Determining the Mean Using Small Samples (The *t* Distribution)

In some instances it is not practical to take the large sample ($n \geq 30$) required for the C.L.T. to be applied. (Name some!) However, if a random variable x is **known to be approximately normally distributed** (something not required by the C.L.T.) it can be shown theoretically that both $\bar{x}$ and $z = (\bar{x} - \mu)/\sigma_{\bar{x}} = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ are normally distributed even for small samples ($n < 30$). The problem that arises for small samples, however, is that, as before, we typically do not know the population standard deviation σ but wish to approximate it by the sample standard deviation s . Now the test statistic, call it t, defined in the same way as was done for z,

$$\boxed{t = \frac{\bar{x} - \mu}{s_{\bar{x}}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \ ,$$

is no longer normally distributed as it was before for large samples. It has what is called a **t distribution**. Essentially when $n < 30$ the variation of s about σ causes the above value to depart from normality. By definition, the t-score measures the number of standard deviations from the mean of a given value, just as with the z-score.

The shape of the t distribution depends upon the sample size n. The larger the sample, the closer the shape is to a normal curve. The smaller the size of the sample, the more the distribution tends to be dispersed from the centre and the flatter the sampling distribution curve.



The total area under any of these curves, is, as usual for a continuous probability distribution, equal to one. Rather than label the different t curves by their associated sample size n, statisticians use a related number called the **number of degrees of freedom**, defined by

$$\boxed{df = n - 1} \ .$$

This is the same quantity that is the denominator in the sample variance formula:

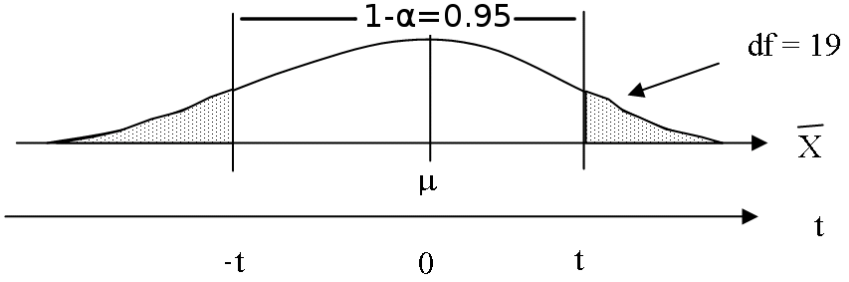
$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{\sum (x - \bar{x})^2}{df}$$

Essentially one has a different t distribution for each different value of df (or n), and we could make an area under the curve table for each one as was done for the normal (z) distribution. In practice, since we are only interested in applying the t distribution to find confidence intervals (and later for hypothesis testing) this is unnecessary; we do not need 30 tables. Recall only a handful of numbers off the normal distribution table were required for these problems. We will grab the same set of numbers for each t distribution and put them on a single table. Furthermore our t table will illustrate the area under the t distribution in a different format than the z table; we will use the area in the centre (or tails) rather than the area up to the statistic value. Finally, whereas the normal table has the area (probability) inside the table, since we will be interested in doing reverse lookups (using area to get t),

it is the t-value which is inside the table. See the t distribution table. Notice that the values for the z (i.e. normal) distribution are given at the bottom of the right column ($df = \infty$).[7]

Example:
Find the t-value based on a sampling distribution with samples of size 20 that would be used for constructing a 95% confidence interval. Draw a sketch of the sampling curve.

Solution:
If $n = 20$ then $df = n - 1 = 20 - 1 = 19$.



The confidence coefficient is $1 - \alpha = 0.95$ so read down the appropriate column on the t-table to find $t = 2.093$.[8]

Confidence intervals for the mean are done exactly as before but now we need to be careful to identify which t distribution is the correct one to find t by using df as just shown. Also since we are using a t distribution rather than the normal distribution our maximum error of estimate for our confidence interval is $\boxed{E = t \cdot s_{\bar{x}}}$. Finally we note that the assumption that x be normal is not overly restrictive as many distributions are normal. Moreover the t statistic is **robust** which means that its distribution remains largely unchanged even if the underlying distribution is not normal. In practice if x has a distribution that merely has a mound in the centre this analysis will still typically work.

Example:
A simple random sample of 10 small cars is tested for fuel economy. Here are the results:

$$5.3,\ 6.2,\ 8.2,\ 6.1,\ 5.9,\ 7.3,\ 9.7,\ 5.5,\ 10.3,\ 7.5 \text{ (litres/100 km)}$$

Construct a 98% confidence interval for the mean fuel economy of small cars based upon this evidence if the fuel economy of cars is known to be approximately normally distributed.

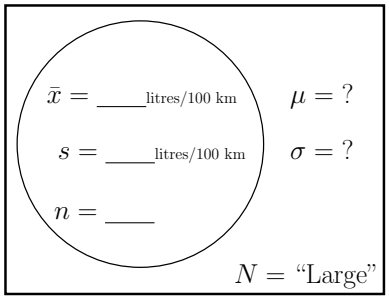
(Answer: $P(7.20 - 1.56 < \mu < 7.20 + 1.56) = P(5.64 \text{ L/100 km} < \mu < 8.76 \text{ L/100 km}) = 0.98$)

Solution:
We follow the same steps as before, guided by what we need to fill in the confidence interval statement.

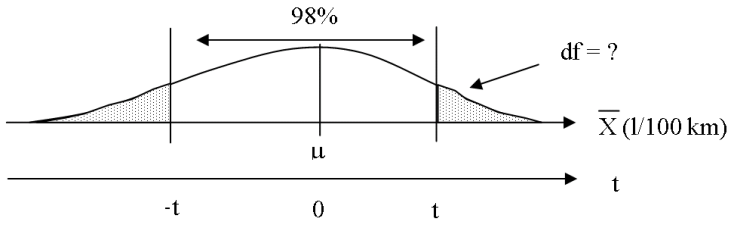
[7] The ∞ symbols means infinity, or "very large" df. Recall as n gets really large (≥ 30) the t distribution converges to the normal curve.

[8] Note that some texts use the notation t_x to indicate the t-value that has x area in the single tail beyond it. In this example one could write $t_{0.0250} = 2.093$, since half of 0.05 or 0.0250 is to the right of $t = 2.093$. The tables in these notes do not use this notation.

Step 1) Identify all given information with symbols on a Venn diagram. Compute $\bar{x}$ and s using your calculator and the appropriate statistical keys.



Step 2) Draw a diagram of the sampling distribution for samples of this size. Since $n = 10$ is less than thirty and x is known to be normal we will need a t distribution.[9]



Step 3) Determine the t-value.

- Identify the confidence coefficient, $1 - \alpha =$ _____ (the column).
- Decide on the number of degrees of freedom, $df =$ _____ $- 1 =$ _____ (the row).
- The t-value from the tables is $t =$ _____.

Step 4) Calculate the standard error.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} =$$

Step 5) Compute the maximum error of estimate.

$$E = t \cdot s_{\bar{x}} =$$

Step 6) Make the confidence interval statement.

$$P\,(\,\bar{x} - E \;<\; \mu \;<\; \bar{x} + E\,) = 1 - \alpha$$

$$P\,(\,\_\_\_\_\_ - \_\_\_\_\_ \;<\; \mu \;<\; \_\_\_\_\_ + \_\_\_\_\_\,) = \_\_\_\_\_$$

$$P(\,\_\_\_\_\_ \;<\; \mu \;<\; \_\_\_\_\_\,) = \_\_\_\_\_$$

[9]The perceptive student should note that in drawing this distribution the two curves are different since $\bar{x}$ has a normal distribution while the statistic we are manipulating $t = (\bar{x} - \mu)/s_{\bar{x}}$ has a t distribution. For our purposes we are only sketching a lump-shaped distribution in either case. The point of the diagram is to indicate important areas and the fact that $\bar{x}$ values and the dimensionless t values are interrelated.

Assignment:

Note in the confidence interval problems below that the "fun" part of using small samples is that the instructor can ask the student to calculate the descriptive statistics required ($\bar{x}$ and s) since there are few enough data elements (less than 30) that you can do this yourself on a calculator or otherwise. After that step is taken, however, these problems are no more complicated that the large sample ones except use of the t distribution is required. The sample statistics ($\bar{x}$,s) for the raw data are given in the solutions for checking. Use these as a starting point to the problems if you do not wish to calculate them.

1. Find t-values required to solve:
   (a) a 98% confidence interval with sample size 15.
   (b) a 90% confidence interval with sample size 28.
   (c) a 99% confidence interval with sample size 6.
   (d) a 99.9% confidence interval with sample size 560.

2. According to Environment Canada, the average water discharge for the month of May of the South Saskatchewan River at Medicine Hat for the years 2000-2009 was measured to be (m^3/s):

69	51	64	272	87
82	238	294	317	89

   (a) If we could assume that the discharges are normally distributed, what would be the 95% confidence interval for the mean average water discharge for the month of May of the South Saskatchewan River at Medicine Hat.
   (b) If the underlying distribution is not approximately normally distributed, what could we do to construct the confidence interval?

3. The distributions of purchases at a snack food outlet is thought to be normal. A random sample of 12 purchases shows the following amounts ($):

7.50	17.80	42.50	5.00	23.40	32.50
9.80	22.20	17.50	15.00	8.80	41.50

   (a) Construct a 90% confidence interval for the average value of a purchase at the outlet.
   (b) A taxation officer observes the outlet over the course of a week and notices that 800 customers have entered the cash register line of the outlet. He seizes the records from the cash register and adds the total value of all purchases over this time. If purchases have been correctly accounted for, between what two values should the total of all purchases lie?

4. After completing the fuel economy example in this section you go to sleep for a well-earned rest. In a dream an angel appears to you and reveals that the standard deviation that you calculated for your sample, $s = 1.74$ litres/100 km, is not only your sample standard deviation but the actual population standard deviation σ as well! The angel commands you to recalculate your confidence interval for the population mean. After you awake you rush back to your notes. What has to change in the calculation? Compare your new confidence interval to your old one. (See the answers if you are mystified about what is required here.)

5.1.5 Confidence Interval Review Exercises

1. Suppose that there are currently 1,200,000 unemployed members of Canada's labour force. The mean length of time that a person is unemployed is 12.0 weeks with a standard deviation of 4.0 weeks. The distribution of time-spent unemployed is approximately normal.

   (a) A person is randomly chosen from among those unemployed in the labour force. What is the probability that the person has been unemployed in excess of 18 weeks?

   (b) Suppose that a simple random sample of 100 people is selected from those unemployed in the labour force. What is the probability that the mean length of time unemployed for the people in the sample exceeds 18 weeks?

   (c) Suppose we do not know what the mean or standard deviation in the labour force is but we do know that a simple random sample of 1000 unemployed workers yields an average of 11.80 weeks with a standard deviation of 3.75 weeks.

      i. Based on the 99% confidence level, what is the maximum error of estimate?

      ii. What is the confidence interval for the mean length of time spent unemployed by a person in the labour force?

   (d) Recalculate your answers to part (c) based upon a sample size of 20 rather than a sample size of 1000 using the same point estimates.

   (e) A small pilot survey was selected from the 1,200,000 unemployed. It was found that the standard deviation in the population was 4.2 weeks. How large should the random sample be in order to estimate the population mean based upon a 92% level of confidence if we wish to be within 0.5 weeks in our estimate?

2. A company has over 5000 salespeople spread across Canada. A random survey of business travel for these people showed that on the basis of 200 of these salespeople, the average distance travelled was 35,000 km with a standard deviation of 2,500 km. Construct a 98% confidence interval for the average travel of the 5000 salespeople.

3. A company wishes to purchase a compact vehicle to obtain the best fuel economy for urban deliveries. They chose 10 compact vehicles from sales lots and drove them under similar conditions. The average fuel economy achieved was 7.2 1/100 km with a standard deviation of 1.2 1/100 km. On the basis of the sample results, construct a 90% confidence interval for the average fuel economy one can expect from compact cars driven under these conditions.

4. In 2011 the Saskatchewan government decided not to hold a province-wide vote on whether to introduce daylight savings time (DST) arguing that in a recent government poll, 66 per cent of Saskatchewan residents opposed switching to DST, while only 27 per cent were in favour. Seven per cent had no opinion. Further details of the poll were as follows:

   > "The province-wide poll of 1,012 Saskatchewan residents was conducted by Fast Consulting between Jan. 10 and Jan. 24. It has a margin of error of 3.1%, 19 times out of 20."

   (a) Calculate the maximum error of estimate for the proportion of the population in favour of switching to DST at the 95% level of confidence. Why is there a discrepancy between your answer and the 3.1% reported?

   (b) Find the confidence interval for the proportion in favour of switching to the DST at a 99.9% level of confidence. (You can find the appropriate z-value on the bottom of the t table with $df = \infty$.) State your conclusion using the same statistical language of the quotation. Would the 99.9% confidence interval have been more useful to report than the 95%?

5. A shaker box contains 10,000 beads of which some are white and some are coloured. A random sample of 500 of these beads are chosen. 100 are found to be coloured. Construct a 99% confidence interval for the proportion of beads in the box that are coloured.

6. To test the durability of a new paint for white centre lines, a highway department painted test strips across heavily travelled roads in eight different locations, and electronic counters showed that they deteriorated after having been crossed by (to the nearest hundred vehicles):

   142,600 167,800 136,500 108,300 126,400 133,700 162,000 149,400

   Construct a 95% confidence interval for the average amount of traffic (vehicles) the new white paint can withstand before it deteriorates. Assume that the data are approximately normally distributed.

7. A political party is interested in its current status among the electorate. A random sample of voters is to be polled regarding their current political preference. How many voters should be included in the poll so that the party will be off in its estimate from the true proportion by no more than 2% with a 95% confidence level.

8. An auditor is checking the accounts receivable journal of a firm. There are 10,000 entries in the journal. The auditor wishes to know the average value of an entry to the journal to within \$40. How many entries should be included in the sample to have a 99% degree of confidence in the answer if he estimates the standard deviation of entries to be about \$800?

Steps for a Confidence Interval

Step 1) Identify all given information with symbols, preferably on a Venn diagram.

Step 2) Draw a diagram of the sampling distribution.

Step 3) Determine the z-value or t-value.

Step 4) Calculate the standard error.

Step 5) Calculate the maximum error of estimate, E .

Step 6) Make the confidence interval statement.

5.2 Hypothesis Testing

5.2.1 The Logic Of Hypothesis Testing

The Null and Alternative Hypotheses

Statisticians use a very precise definition of the word "hypothesis". From a statistical point of view a hypothesis is an assumption made about a population parameter.

Grammatically, a hypothesis is always phrased as a declarative sentence.

Example:
The following are examples of **Declarative Sentences**:

- It is sunny outside.
- The Passenger Pigeon is extinct.

The following are examples of sentences which are **Not Declarative**:

- Hi, how are you?
- Have a nice day!

From the above examples, we see that a declarative sentence can be either true or false but other types of sentences have no truth-value. A **statement**[10] is a declarative sentence which has a truth-value. Hypotheses are statements as they are either true or false but not both.

Hypothesis testing is a decision-making procedure that uses sample evidence to test the truth of a statement about a population parameter.

Example:
The following are examples of some hypotheses we may want to test:

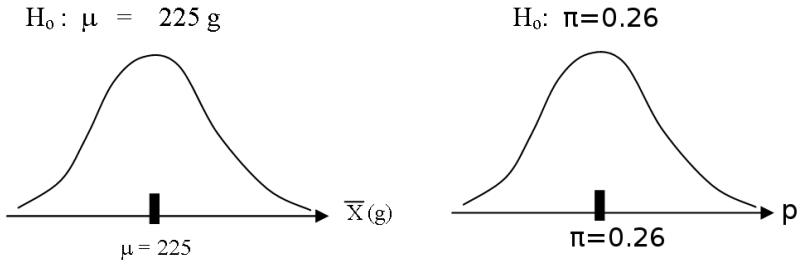
1. The net contents of a can of tuna fish coming off of the production line averages 225 g.
2. Historically, the average mark scored by applicants on a personnel test is 75%.
3. The proportion of overweight children in Canada is 26% .
4. Most widgets produced in Canadian factories are within 5 grams of the average weight.

Sampling theory is used in hypothesis testing for the purpose of avoiding bias in making the decision. It is an important part of the process that the decision be arrived at based upon an initial impartial point of view. It has been said that statistics can be used to prove any position that you wish to take in a discussion. That can be the case if statistics procedures are used incorrectly.

The decision to be tested is called the **null hypothesis**, $\mathbf{H_0}$. In the null hypothesis, the parameter to be tested is always set equal to some assumed value. This assumed value is the centre of the sampling distribution curve on which we will base our decision. The word *null* means *no difference* in statistics. It is another way of saying, *equal to*.

[10] Statements in logic are also called **propositions**.

Example:
The null hypotheses from cases 1 and 3 of the previous example would imply the following sampling distributions, assuming large enough sample sizes are taken:



The **alternative hypothesis**, $\mathbf{H_a}$, is a hypothesis that will be accepted if the statistical evidence leads to the conclusion that the null hypothesis is probably false. There are three different arguments that can be put up in opposition to the statement that a parameter is equal to a certain value. We could conclude that it is *less than*, *greater than* or simply *not equal to* that value depending on the circumstances. The alternative hypothesis will always have one of these three mathematical relation symbols: $<$, $>$, or $\neq$.

Example:
Possible alternative hypotheses for case 1 of the earlier example are:

$$H_a : \mu < 225 \text{ g} \qquad H_a : \mu > 225 \text{ g} \qquad H_a : \mu \neq 225 \text{ g}$$

The first step in decision-making is to formulate these two hypotheses so that we can gather some sample evidence and use it to examine how the sample observations compare to the assumed population parameter value. If the sample statistic deviates too far from the assumed population parameter, we say the difference is **statistically significant** and we reject H_0. When we make our decision, it will either be that the evidence suggests that we **reject $\mathbf{H_0}$ and accept $\mathbf{H_a}$** or that, based on the evidence, we **fail to reject $\mathbf{H_0}$**.

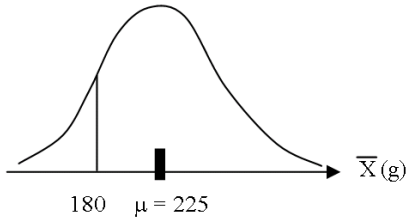
The Risk of Making a Wrong Decision Based upon Incomplete Information

One of the objectives of decision-making is to minimize the risk of making the wrong decision. Here is a diagram that illustrates the pitfalls that can occur when making a decision based on sample evidence.

		Decision Made About H_0	
		Fail to Reject	**Reject**
Unknown State of H_0	**True**	Correct	Incorrect *(Type I error)*
	False	Incorrect *(Type II error)*	Correct

Sample evidence drawn from the population is used to decide the truth of the null hypothesis about the population parameter. The probability of **rejecting $\mathbf{H_0}$ when it is true** is called the **α risk**. This risk, also called the **level of significance**, is set arbitrarily low. Common values of α are 0.01, 0.05, and 0.10. These numbers represent the probability of making a **Type I error** in decision-making.

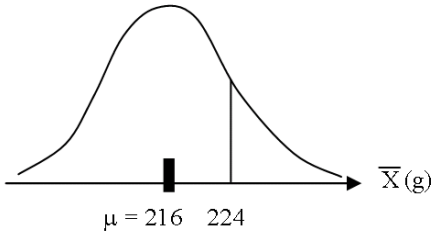
Example:
Continuing the previous example (case 1), suppose we take a sample of 50 cans of tuna fish and find the mean weight in the sample is 180 grams. We might reject the null hypothesis that the mean weight of all cans of tuna fish is 225 grams because there is such a large difference between the values. It is possible, however, just due to the random nature of sampling that we picked 50 cans of lighter weight among all cans packaged and the true mean is in fact 225 grams for all cans. After all, as can be seen from the sampling distribution curve, we expect half the samples to have means lower than 225 grams.



If, based on our evidence, we chose to reject H_0, i.e. that the mean is 225 grams, when it was in fact true, we made a Type I error.

The β **risk** is the probability of **accepting a false null hypothesis** on the basis of sample evidence. This type of error is called a **Type II error** in hypothesis testing.

Example:
Continue the previous example (case 1) with a sample of 50 cans of tuna fish. Suppose the mean weight in the sample is 224 grams. We might fail to reject the null hypothesis that the mean weight of all cans of tuna fish is 225 grams because there is such a small difference between the values. However, the mean weight of all cans might in fact be 216 grams and through random selection we just happened to pick 50 cans that were very heavy relative to the distribution.



If, based on our evidence, we failed to reject the null hypothesis ($\mu = 225$ grams) when it is in fact false ($\mu = 216$ grams) we have made a Type II error.

The interrelationship between the errors may be understood with a law metaphor. Suppose we are wishing to the evaluate the guilt of an accused person standing trial. In an English court of law the defendant is innocent until proven guilty. Let the null hypothesis H_0 be the hypothesis that the defendant is innocent. The alternative hypothesis H_a is that the defendant is guilty. In this case a Type I error, rejecting the null hypothesis when it is true, corresponds to finding the defendant guilty when he is innocent. This is considered a more serious error legally than a Type II error, failing to reject the null hypothesis when it is false, which corresponds to finding the defendant innocent when he is guilty. For purposes of this course, we need to be able to describe what a Type II error is. We will not deal with calculations involving the probability (β) of making this type of error. The risk of making a Type I error (α) will, however, inform our decision making process.

Evaluating the Evidence

Evaluation of the evidence is always the last consideration in making a statistical decision. The strategy for how a decision is to be made is set forth and then the sample evidence is gathered from the population based on random sampling techniques. The evidence is always weighed against the theoretical assumptions in order to judge the probability of observing these sample results using the theoretical statistical model. Clearly the degree of separation between the sample statistic and the hypothesized parameter will be critical in our utlimate decision.

Example:
In the previous example (case 1), the null hypothesis is that the mean weight of all cans of tuna fish is 225 grams. We may decide to reject this only if the observed sample mean is in the bottom 1% of possible observations for the sample mean. The risk of rejecting a true null hypothesis would then be $\alpha = 0.01$, a very small risk. Having established this process, we would then go and examine a random selection of 50 cans of tuna fish and see where its sample mean lies.

Assignment:

Formulate the null and alternative hypotheses for the following situations. (This is the first step of any hypothesis test.) The alternative hypothesis, though written second, is usually worth identifying first as it typically reflects the motivation for the test.

1. An ambulance service is considering replacing its ambulances with new equipment. If \$52.00 is the average weekly maintenance cost of one of the old ambulances and μ is the average weekly maintenance cost it can expect for one of the new ones, what hypotheses should be used if the service wants to buy the ambulances only if it can be shown that this will significantly reduce the average weekly maintenance cost?
   - H_0:
   - H_a:
2. What hypotheses should the ambulance service use in problem 1 if the ambulance service is anxious to buy the new ambulances (which have some other nice features) unless it can be shown that they will significantly increase the average weekly maintenance cost?
   - H_0:
   - H_a:
3. A major Canadian brewery has held an 18% share of the market. However, because of an increased marketing effort, company officials believe the brewery's market share is now greater than 18%. What hypotheses should the company officials use if they want to prove that their market share has significantly increased?
   - H_0:
   - H_a:
4. What hypotheses should a quality improvement team use if it wants to make sure that wire leads are meeting a specification of 10 cm in length?
   - H_0:
   - H_a:

Note that the claim which we are asked to test while often appearing in H_a, sometimes shows up in H_0, as, for instance, the last example shows.

5.2.2 Testing Hypotheses About a Single Mean

Our first type of hypothesis testing involves a single parameter. We begin by testing a population mean, μ.

The theory that was presented in Section 5.2.1 is formulated into a formal six step procedure. All tests of hypothesis follow this process.

Example:
A promotional agency markets chocolate almonds through service clubs and charities. The agency claims that the packages contain on average 350 grams. Evaluate this claim at the 0.01 level of significance if in a random sample of 49 boxes a mean of 340 grams with a standard deviation of 28 grams was measured.

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\mathbf{H_0 : \mu = 350\ g}$ $\Leftarrow$ Remember the null hypothesis must always have an "=" sign.
- $\mathbf{H_a : \mu < 350\ g}$ $\Leftarrow$ The only point of contention is whether the customer is short changed.

Step 2) State the level of significance.

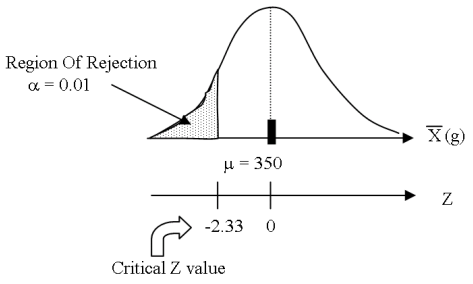
$\boldsymbol{\alpha = 0.01}$ $\Leftarrow$ This is some low value, here given as 1%.

Step 3) Determine the test statistic.

We will examine a large sample of boxes ($n \geq 30$) and determine their average weight so a ***z*-test** applies because the distribution of the mean for large samples is a normal curve. If we were confident that the population of weights of all boxes was itself approximately normally distributed and we examined a small sample of boxes ($n < 30$), then a t test would be required.

Step 4) Establish a decision rule.

Draw a **sampling distribution curve** and determine the **critical *z*-value** which identifies the region in which the null hypothesis is rejected.

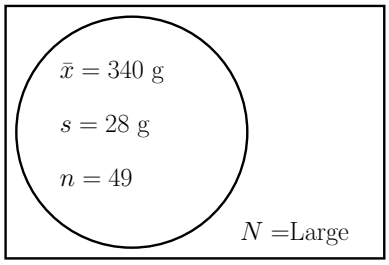


Here since $\alpha = 0.01$ is the tail area on the left side of a normal curve we find $z_{\text{critical}} = -2.33$.

This is called a **one-tail test** because the region of rejection falls in one tail. If the alternative hypothesis has a "$\neq$" sign, there will be two tails of rejection and the probability of rejection (the significance) would be split equally between them.

A useful trick to identify the region of rejection is to look at the alternative hypothesis and consider the inequality as pointing to the region. (So $<$ means left, $>$ means right, and $\neq$ means left and right.)

Step 5) Evaluate the evidence. (Use a Venn diagram for the statistics.)



In order to determine how extreme our sample data is we find its **calculated *z*-score**. The goal is therefore to evaluate

$$\boxed{z = \frac{\bar{x} - \mu}{s_{\bar{x}}}}$$

Given the above data the required estimated standard error of the mean is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{28 \text{ g}}{\sqrt{49}} = 4 \text{ g}$$

(This formula for $s_{\bar{x}}$ is valid even if sampling without replacement since the population size N of all boxes produced by the factory is much larger than the sample so $n/N = 49/\text{Large} \approx 0 < 0.05$.)

The calculated z-score for our measured sample mean is:

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{340 \text{ g} - 350 \text{ g}}{4 \text{ g}} = -2.50$$

Step 6) State your decision.

Since the calculated z-value (-2.50) lies in the region of rejection ($z < -2.33$), the evidence suggests, at a 0.01 level of significance, that we reject H_0 and accept H_a. In words, the mean weight is less than 350 grams.

Example:
Repeat the previous example but now at the **0.005** level of significance. The other details are the same. A promotional agency markets chocolate almonds through service clubs and charities. The agency claims that the packages contain 350 g. Evaluate this claim at the 0.005 level of significance if in a random sample of 49 boxes a mean of 340 grams with a standard deviation of 28 grams was measured. (Answer: $z_{\text{critical}} = -2.575$ so fail to reject H_0)

Solution:

Step 1) Formulate the null and alternative hypotheses.

- **H_0 :**
- **H_a :**

Step 2) State the level of significance.

$\alpha =$

Step 3) Determine the test statistic.

Step 4) Establish a decision rule. (draw diagram, find critical value)

Step 5) Evaluate the evidence. (Use a Venn diagram for the statistics.)

Step 6) State your decision.

Comparing with the first hypothesis test we see that if we want a lower probability of making an error (Type I) this is a smaller α. This more stringent requirement meant in this case that we could no longer reject the claim being made (H_0).

5.2.3 The *P*-value Approach to Hypothesis Testing

The previous section outlined the **critical value** approach to hypothesis testing. As we saw in the examples, the rejection or failure of rejection of H_0 depends on the level of significance, α. Indeed we could always choose an α small enough that we would fail to reject H_0 . Alternatively a large enough α will result in a rejection of H_0.[11] So how do you find the cutoff value of α between the two realms of rejection and failure of rejection so you can choose a level of significance to get the desired conclusion you want? **YOU DON'T!** We're not in the propaganda business here! The level of significance α is chosen at a stage before evaluation of the evidence precisely to avoid this type of bias.

However this line of thinking is valuable as a second approach to hypothesis testing as we will now see. Define the P-value to be just this cutoff point:

The ***P*-value** is the smallest value of α which would result in the rejection of H_0 .

The P-value will be determined by the evidence. Once you know it, the result of the hypothesis test is evaluated as follows:

- If the P-value is less than or equal to the significance (***P*-value** $\leq \boldsymbol{\alpha}$) then **Reject H_0** .
- If the P-value is greater than the significance (***P*-value** $> \boldsymbol{\alpha}$) then **Fail to Reject H_0** .

The P-value approach to hypothesis testing requires the same first three steps as the critical-value approach. The remaining steps require using the evidence to find the calculated value as before but now this gets converted into a P-value as shown in the following example.

Example:
Repeat the example from Section 5.2.2 using the P-value approach. Recall: A promotional agency markets chocolate almonds through service clubs and charities. The agency claims that the packages contain on average 350 grams. Evaluate this claim at the 0.01 level of significance if in a random sample of 49 boxes a mean of 340 grams with a standard deviation of 28 grams was measured.

Solution:
The first three steps are the same and we evaluate the evidence one step earlier:

Step 1) Formulate the null and alternative hypotheses.

- $\mathbf{H_0 : \mu = 350\ g}$
- $\mathbf{H_a : \mu < 350\ g}$

Step 2) State the level of significance.

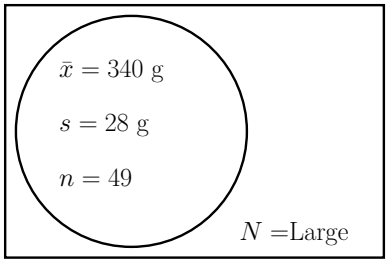
$\boldsymbol{\alpha = 0.01}$

Step 3) Determine the test statistic.

We will examine a large sample of boxes ($n \geq 30$) and determine their average weight so a ***z*-test** applies because the distribution of the mean for large samples is a normal curve.

[11] This assumes the sample statistic lies on the same side of the parameter as the region of rejection.

Step 4) Evaluate the evidence. (Use a Venn diagram for the statistics.)



In order to determine how extreme our sample data is we find its **calculated *z*-score**. The goal is therefore to evaluate

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

Given the above data the required estimated standard error of the mean is

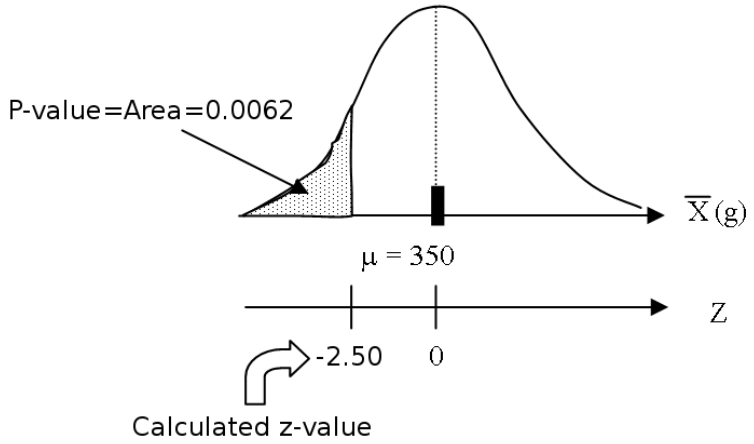
$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{28 \text{ g}}{\sqrt{49}} = 4 \text{ g}$$

The calculated z-score for our measured sample mean is:

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{340 \text{ g} - 350 \text{ g}}{4 \text{ g}} = -2.50$$

Step 5) Find the P-value.

The P-value is just the area in the tail (or tails for a two-sided test) corresponding to our calculated test statistic $z = -2.50$. Draw a diagram of the sampling distribution to identify the area and look it up on the appropriate (here normal) table.



Step 6) State your decision.

Since P-value $= 0.0062$ is less than $\alpha = 0.01$ we reject H_0 at $\alpha = 0.01$. (Our P-value tail would be smaller than the α tail if we had drawn them on the same graph.) The evidence suggests the mean weight is less than 350 g.

Note:

1. We got the same result as the critical value approach.

2. If we wanted to repeat the question as we did before at a level of significance of $\alpha = 0.005$ the only changes required would be in Step 2 where we just stated the level of significance and Step 6 where now we have P-value $= 0.0062$ greater than $\alpha = 0.005$ so we would fail to reject the null hypothesis.

3. For a two-tailed test the P-value would require the area in two tails. So for instance if H_a here had been $\mu \neq 350$ g, all other things being equal, then we would have P-value $= (2) \cdot (0.0062) = 0.0124$. We would have then failed to reject H_0 because this probability is bigger than $\alpha = 0.01$. Note, however, that in this different problem the α area would have been split into two tails of area 0.005 each had we used the critical value approach.

4. From an evaluation point of view the critical value approach requires taking α (an area or probability) and converting it to a z or t value. In the P-value approach we had to go the other way around and take the statistic and convert it to an area (the P-value). In principle this is six of one or half-a-dozen of the other. In practice however we can only go both ways with our tables using the normal (z) distribution. For the t distribution we only have a table which can convert a specific area into the corresponding t-value. As such we will only be able to use the P-value approach for normal sampling distribution problems in this course. (Of course on a computer one could readily calculate P-values for problems requiring the t distribution.) Also note that if your calculated z-value is too large in magnitude ($|z| > 3.49$) our normal table cannot be used to calculate the required P-value area; it will be approximately 0. (Once again a computer could be used to get a more precise value.)

5. An advantage of the P-value approach is that a statistician can immediately determine whether the decision to reject or fail to reject was close (P-value $\approx \alpha$) or not (P-value differing substantially from α). In our example above we had P-value $= 0.0062 = 0.62\%$ which is not substantially lower than α=0.01=1% With the critical value approach the strength of the decision is less obvious.

6. With just the P-value being reported a statistician can use a sliding scale approach to classify the significance of the result based on the value as shown in the following table:

P-value Range	Decision	Result Status
P-value < 0.01	reject H_0 and accept H_a	highly significant
$0.01 < P$-value < 0.05	reject H_0 and accept H_a	statistically significant
$0.05 < P$-value < 0.10	fail to reject H_0 (usually)	tending towards significance
$0.10 < P$-value	fail to reject H_0	not statistically significant

   Obviously this table is somewhat subjective – one may always return to the P-value to be clear about what the result is. In this course we will always state a level of significance α for hypothesis tests against which the P-value is judged. This sliding scale table can give you a feel for what size of α is appropriate for a test, however.

7. Finally, do not to confuse the level of significance α and the P-value . The former (α) is identified by the experimenter before data is even collected as the risk she is prepared to take in making the error of rejecting H_0 when it is in fact true. The P-value, on the other hand, depends upon the data collected in the experiment.

Assignment:

The objective of this exercise is to gain familiarity with the formal testing procedure for hypothesis testing. Make sure to follow the steps outlined in the examples. Use the critical value method for all problems. Find P-values where possible with our tables.

1. An accountant reads a summary about a large journal which states that the average value of a journal entry is \$560. If she takes a sample of 50 entries and finds a mean of \$595 with a standard deviation of \$100, evaluate the claim at a 10% level of significance.

2. Repeat the previous question, but now the accountant takes a sample size of only 20 and finds a mean of \$595 with a standard deviation of \$100. What assumption do we have to make about the population of entries before we can test the hypothesis?

3. A merchant sells butter which is advertised as weighing 454 grams (on average). Establish a testing procedure to test the merchant's claim based on large samples and a level of significance of 0.05. Evaluate his claim using sample evidence which shows that a random sample of 60 bricks averaged 447 grams with a standard deviation of 10 grams.

4. What is the probability of committing a Type I error in problem 3?

5. What would it mean to commit a Type II error in problem 3 in terms of the words of the problem?

6. An astronomer observes a group of over one thousand stars that appear to be clustered together in a small area of the sky. She believes that rather than being due to the random chance of stars being lined up along her line of sight, the stars appear clustered because they are actually near each other and likely have a common formation history. To test her theory she measures the *metallicity*[12]of several stars in the group with the following results (dex):

-0.50	0.00	0.02	0.06	0.08	0.09	0.11	0.12
0.14	0.19	0.19	0.20	0.22	0.25	0.30	0.38

   The astronomer consults a journal article that shows that stars in a larger region of the sky in which the cluster lies have a metallicity that is normally distributed with a mean of -0.301 . Test the the claim, at a level of significance of 1%, that the cluster stars have this mean metallicity. What is the likely explanation of the star with metallicity -0.50?

7. According to the Canadian Mortgage and Housing Corporation the average rent for a two-bedroom apartment in October, 2010 in Regina was \$881. Suspecting that the average rent for such dwellings has increased, a renter surveys 10 two-bedroom apartments and finds the following monthly rents (\$):

750	800	810	880	910
925	990	1000	1010	1030

   Evaluate the claim, at the 0.05 level of significance using this data that the rent is still \$881.

[12]Astronomers consider *metals* to be any element other than hydrogen and helium! After the Big Bang, the universe was primarily hydrogen and helium, metals only forming later in stars, such as in supernova events. A high metallicity therefore suggests a later generation star, such as our sun, which has formed out of such metal-enriched matter. Metallicity is written as [Fe/H], which represents the logarithm of the ratio of a star's iron abundance, easily identified in the star's light spectrum, compared to that of the Sun. A star with metallicity 0 has the same iron abundance as our sun, while metallicity of 1 would be 10 times solar, and -1 would be $1/10^{\text{th}}$ solar.

5.2.4 Testing Hypotheses About a Single Proportion

We next consider tests involving a single population proportion, π. For our purposes sample sizes must be large, as discussed in Section 4.6, so that we may assume that sample proportions are normally distributed. The six step hypothesis testing procedure applies here as well.

Example:
In a recent municipal election for a large city the votes were counted using a *Demonold*-brand electronic voting machine and the winning candidate for mayor received 77.1% of the vote. A meddling elections official wanting to test the accuracy of the result takes a simple random sample of 1000 of the valid ballots, counts them by hand, and finds that 792 supported the winning candidate. At a level of significance of 2% test the hypothesis that the actual proportion was different than the 77.1% originally reported. (That is, test the hypothesis that the electronic count was inaccurate.)

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\mathrm{H}_0 : \pi = 0.771$ $\Leftarrow$ The candidate reportedly received 77.1% support.
- $\mathrm{H}_\mathrm{a} : \pi \neq 0.771$ $\Leftarrow$ The machine could err in either direction.

Step 2) State the level of significance.

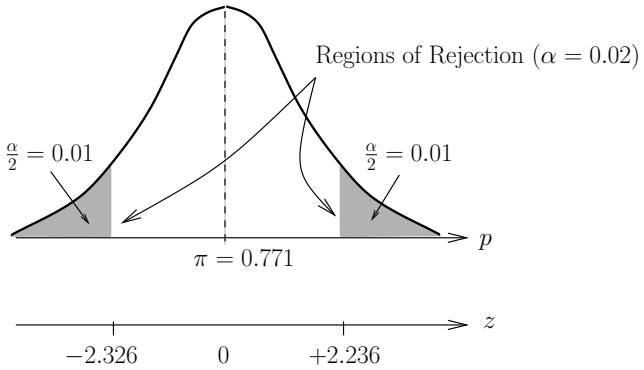
$\alpha = 0.02$ $\Leftarrow$ As directed in the problem.

Step 3) Determine the test statistic.

The official plans on using a sample of size $n = 1000$ for the hypothesis test so, assuming the null hypothesis is true, $n\pi = (1000)(0.771) = 771$ and $n(1-\pi) = (1000)(0.229) = 229$ which are both greater than 5. We can therefore assume that p will be approximately normally distributed and thus a z test applies. (For this course in tests about proportions this will always be the case because large size samples will always be used.)

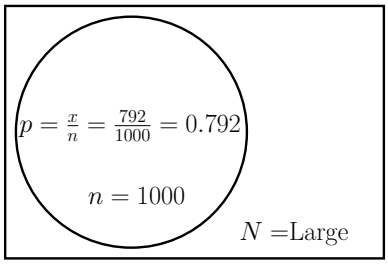
Step 4) Establish a decision rule.

Draw a sampling distribution curve and determine the critical z-value which identifies the region in which the null hypothesis is rejected. Since the alternative hypothesis is $\neq$ this is a two-tailed test:



Here for the given level of significance, $z_{\text{critical}} = \pm 2.326$ from the t-table with $\alpha = 0.02$ for a **two-tailed** test and $df = \infty$ (normal). (It is ± 2.33 if the normal table is used.)

Step 5) Evaluate the evidence (Use a Venn diagram for the statistics).



In order to determine how extreme our sample data is we find its **calculated *z*-score**. The goal is therefore to evaluate

$$\boxed{z = \frac{p - \pi}{\sigma_p}}$$

The required standard error of proportion is:

$$\sigma_p = \sqrt{\frac{\pi \cdot (1 - \pi)}{n}} = \sqrt{\frac{(0.771) \cdot (1 - 0.771)}{1000}} = 0.01328755$$

(This formula for σ_p is valid even if sampling without replacement as the population of votes is assumed much larger than 1000, $n/N = 1000/\text{Large} \approx 0 < 0.05$.)

Note that we have used the population parameter π of the null hypothesis in the calculation of the standard error since we are assuming the null hypothesis is true and knowing π is sufficient to calculate σ_p .

The **calculated *z*-value** is: $z = \dfrac{p - \pi}{\sigma_p} = \dfrac{0.792 - 0.771}{0.01328\ldots} = 1.580426723 = 1.580$

Step 6) State your decision.

Since the calculated z-value (1.580) does not lie in the region of rejection ($z < -2.326$ or $z > 2.326$) we fail to reject the null hypothesis H_0. At a level of significance of 0.02 the evidence does not support that the vote count was inaccurate.

Example:
Repeat the above example using the P-value approach.
(Answer: P-value $= 0.1142$ greater than $\alpha = 0.02 \Rightarrow$ fail to reject H_0 .)

Solution:
Steps 1-3 are the same as above.

Step 4) Evaluate the evidence. (Use a Venn diagram for the statistics.)

Same as Step 5) above, calculated $z = 1.580$

Step 5) Find the P-value.

Step 6) State your decision.

Assignment:

1. A television network sells advertising based on the proportion of the large metropolitan area it reaches that will watch a given program. A company buys advertising during a particular show with the expectation that 40% of households will be watching the program. After the program airs the company samples 100 households and finds that only 30 watched the show. If the 10% level of significance is used, can the null hypothesis that 40% of the households would watch the program be rejected?

2. In 2008, 68% of Canadian women were unable to identify that the leading cause of death among women was stroke and heart disease. From 2008 onward the Heart and Stroke Foundation ran *The Heart Truth* campaign to raise awareness of the severity of the issue and the warning signs associated with stroke and heart disease. In 2011 if a Harris-Decima poll of 1,013 Canadian women found 53% of Canadian women unable to identify their leading cause of death, can we conclude the proportion of women unable to identify their leading cause of death decreased over the three year interim period at the 0.02 level of significance?

3. A counselor claims that at least 30 percent of the students attending a local college hold a part-time job. Doubting the claim, a college official does a random sample of 75 students finds that 15 have part-time jobs. Is there adequate evidence to suggest that the counselor is incorrect in her claim at the 0.05 level of significance. The college has approximately 2000 students.

4. In an effort to cut costs a computer chip manufacturer replaces a semiconducting material in an existing component with a cheaper material. The manufacturer wishes to test the effect, if any, of such a change. With the old material the failure rate of the component was 7%. If in a batch of 200 of the new components there are 16 failures, test the claim that the new failure rate is still 7% at the 2% level of significance.

5. To meet a government regulation a gasoline retailer claims to contain at least 6% ethanol by volume in its gasoline. A government regulator wishes to test this claim at the 5% level of significance. Would she use a proportion hypothesis test? Explain your answer.

6. In a civic election mayoral candidate Mike "the slasher" promises to "slash" public spending in order to "slash" taxes. To get the scoop on the actual election results a local television station does an exit poll on election day. Out of 1000 voters who willingly responded after leaving the polling station, Mike has only 430 supporters versus 570 supporters for "compassionate Joe" the only other candidate running. The station reports that Mike will lose the election only to find out, once the votes are counted, that he wins!

   (a) Using the results of the exit poll, test whether Mike the slasher had less than the required 50% to win the election at the 0.05% level of significance.

   (b) Assuming the exit poll was an actual random sample of voters in the city, is chance the likely explanation of the discrepancy between the poll conclusion and the actual result?

5.2.5 One Parameter Hypothesis Testing Review Exercises

Use the formal six step procedure (critical value approach) for hypothesis testing to do the following problems. Determine P-value's where possible.

1. A building products association sets standards for building sizes of products used in building construction. To set the standard for door heights, the association is interested in knowing if the average height of adult males in North America has increased from its previous value of 178 cm. A sample of 121 males shows an average height of 181 cm with a standard deviation of 5 cm. Can an increase in the average height be supported at the 5% level of significance?
2. A trucking firm claims that the weights of trucks it sends over a municipal road is 25,000 kg. Test the claim at the 0.01 level of significance if a sample of 100 random weighings of the trucks shows a mean weight of 28,000 kg with a standard deviation of 3000 kg.
3. It is thought that half of all winners of major prizes in raffles held at summer fairs are males. A random sample of 100 prize winners shows that 40 are male. At the 0.10 level of significance can we conclude that less than half of all prize winners are male?
4. A drink dispenser is supposed to be set to fill an average of 300 ml of beverage in a cup. Using a 1% level of significance, should the drink dispenser be adjusted if a random sample of 8 cups filled showed the following volumes (ml):

   280 290 321 315 302 275 318 302

   (Assume drink dispensations are approximately normally distributed.)
5. A large manufacturing company investigated the service it received from suppliers and discovered that, in the past, 68% of all materials shipments were received on time. However, the company recently installed a just-in-time system in which suppliers are linked more closely to the manufacturing process. A random sample of 110 deliveries since the just-in-time system was installed reveals that 81 deliveries were on time. Use this sample information to test whether the proportion of on time deliveries has significantly increased at the 5% level.

Steps for a Hypothesis Test

Critical Value Approach	***P*-value Approach**
Step 1) Formulate the null and alternative hypotheses	Step 1) Formulate the null and alternative hypotheses
Step 2) State the level of significance.	Step 2) State the level of significance.
Step 3) Determine the test statistic.	Step 3) Determine the test statistic.
Step 4) Establish a decision rule. (draw sampling distribution, find critical value)	Step 4) Evaluate the evidence. (identify statistics on a Venn diagram, find the calculated value)
Step 5) Evaluate the evidence. (identify statistics on a Venn diagram, find the calculated value)	Step 5) Find the P-value. (draw sampling distribution)
Step 6) State your decision.	Step 6) State your decision.

5.3 The Relationship Between Confidence Intervals and Hypothesis Tests

It should not be surprising that there is a relationship between the construction of confidence intervals and the testing of hypotheses. For a single parameter test they use similar data; a sample is taken and a statistic such as the sample mean or proportion is calculated. In one case this statistic is used to estimate the population parameter, while in the other it is used to test the value of the population parameter. To see the connection between confidence intervals and hypothesis tests let us return to our first confidence interval. In Section 5.1.1 we measured a sample of 50 wage earners and found a mean wage of $\bar{x} = \$42,000$ with a standard deviation of $s = \$5000$. We used this to construct a 95% confidence interval of the population mean of:

$$P(\$40,614.06 \ < \ \mu \ < \ \$43,385.94) = 0.95$$

Suppose now that rather than using the data to estimate the value of μ, we used it to test the value of μ equalling some value. For instance we could have tested the hypothesis:

- H_0: $\mu = \$41,000$
- H_a: $\mu \neq \$41,000$

at some level of significance α. Recall that when stating our confidence we involved the same symbol α writing our confidence as $1 - \alpha$. So let us choose $\alpha = 0.05$, the same probability in the two tails that our confidence interval had. (Solving $1 - \alpha = 0.95$ for α gives $\alpha = 0.05$.) For a test of any value like \$41,000 that sits in the above confidence interval it will turn out that we will fail to reject H_0 at $\alpha = 0.05$. However, for any test of a value such as \$40,000 which lies outside the confidence interval, we will end up rejecting H_0 at $\alpha = 0.05$.

To see this note that for a two-tailed test with $\alpha = 0.05$ the critical value for this z-test is $z_{\text{critical}} = \pm 1.960$. The following table shows the calculated values for a hypothesis test, such as that for \$41,000 above, for four different values of the mean that lie just inside and just outside the interval. (The student is encouraged to verify these results):

H_0 Test	Relationship to C. I.	$z_{\text{calculated}}$	Result of test
$\mu = \$40,613$	outside	1.962	Reject H_0 and Accept H_a
$\mu = \$40,615$	inside	1.959	Fail to Reject H_0
$\mu = \$43,384$	inside	-1.957	Fail to Reject H_0
$\mu = \$43,387$	outside	-1.962	Reject H_0 and Accept H_a

We see that the calculated z-values lie just inside the region of rejection for tests of values that lie outside the confidence interval.

This shows how a confidence interval of the mean is equivalent to a two-tailed hypothesis test. What about a one-tailed hypothesis test? This can be related to a **one-sided confidence bound**, a construction not covered in our course. A one-sided confidence bound is used to place a lower or upper limit on a population parameter.

As a final comment, the correspondence between confidence intervals and hypothesis tests for proportions is not as exact as for means. If we try to play the same game we did above with the proportion confidence interval in Section 5.1.2 we will see that the boundaries do not exactly line up. To find the discrepancy, look carefully how we calculate the standard error of proportion with the confidence interval and with the hypothesis test respectively.

Assignment:

1. In Exercise 1 on Page 206 it was found that the two-sided hypothesis test for $\mu = \$560$ was rejected at $\alpha = 0.10$. If a confidence interval for the population mean is created for the same measured data ($\bar{x} = \$595$, $s = \$100, n = 50$) at a confidence of $1 - \alpha = 0.90$, would you expect \$560 to lie inside it or not? Confirm your answer by calculating the confidence interval.

One Population Inferences Formulae

Standard Error for Single Means and Proportions

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \sigma_{p} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \qquad s_{p} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$^{*}\textbf{Finite Correction Factor} \quad \Rightarrow \quad {}^{*}F.C.F. = \sqrt{\frac{N-n}{N-1}}$$

Maximum Error of Estimate for Single Means And Proportions

$$E = z \cdot s_{\bar{x}} \qquad E = t \cdot s_{\bar{x}} \qquad E = z \cdot s_{p}$$

Confidence Intervals for Population Means and Proportions

$$P\,(\,\bar{x} - E < \mu < \bar{x} + E\,) = 1 - \alpha \qquad P\,(\,p - E < \pi < p + E\,) = 1 - \alpha$$

Sample Sizes for Estimating Means And Proportions

$$n = \left[\frac{z \cdot \sigma}{E}\right]^{2} \qquad n = \pi \cdot (1-\pi) \cdot \left[\frac{z}{E}\right]^{2}$$

Standard Scores for Single Means and Proportions

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad z = \frac{p - \pi}{\sigma_{p}}$$

Other Formulae

$${}_{N}C_{n} \;\text{ or }\; N^{n} \qquad p = \frac{x}{n} \qquad df = n - 1$$

Unit 6: Multiple Population Inferences

6.1 Difference Between Means

6.1.1 Introduction

In Unit 5 we were interested in whether a single population mean has a particular stated value (a hypothesis test) or in finding an estimate of that value (a confidence interval). Sometimes instead we are interested the relationship between the means of two different populations (or more). It is conceptually important to distinguish whether samples taken are **dependent** or **independent**. To have independent samples requires that the selection of one sample is in no way affected by the selection of the other. The following examples distinguish these cases.

dependent PSA[1] levels in a group of prostate cancer sufferers before and after radiotherapy is measured. In this case the two samples involve the same people (hence the dependence). Note that the underlying populations are distinguished by time (patients before or after treatment).

independent The average rent for a 3-bedroom apartment in Vancouver versus that in Regina is measured. In this case the samples taken from the two populations (3-bedroom apartments in Vancouver and Regina respectively) can be treated as independent since taking a sample of Vancouver apartments does not affect taking a sample in Regina.

Both independent and dependent samples can have their uses. We will start by looking at an example using dependent samples since these can be reworked into a single parameter problem which we know how to handle already.

*6.1.2 Paired-Difference Experiments for Dependent Samples

Our PSA example illustrates the workings of a paired-difference experiment. The measurements are paired, in this case "before" and "after", for each patient as (x_1, x_2). In the paired-difference experiment we may do a hypothesis test or a confidence interval involving the difference d between the pair values, $\boxed{d = x_1 - x_2}$. The mean value of this new variable is just the difference of the means both for populations ($\mu_d = \mu_1 - \mu_2$) or samples ($\bar{d} = \bar{x}_1 - \bar{x}_2$). After determination of the paired difference values this becomes in its essentials a single parameter hypothesis test or confidence interval involving this new variable d . What sort of hypothesis should be made for the parameter μ_d? Usually (and always for this class) the test will be against the value 0 :

$$\mathrm{H}_0 : \mu_d = 0$$

Since $\mu_d = \mu_1 - \mu_2$ this test is equivalent to

$$\mathrm{H}_0 : \mu_1 - \mu_2 = 0$$

which in turn is equivalent to

$$\mathrm{H}_0 : \mu_1 = \mu_2$$

That is we are testing whether the means of the two populations are equal. The alternative hypothesis can be less than, greater than, or unequal. For instance, following the same logic, if we are interested in testing if μ_1 is less than μ_2, i.e. $\mathrm{H_a} : \mu_1 < \mu_2$ or equivalently $\mathrm{H_a} : \mu_1 - \mu_2 < 0$, then our test on the difference is

$$\mathrm{H_a} : \mu_d < 0 \ .$$

The following example illustrates a paired-difference hypothesis test.

[1] Prostate specific antigen (PSA) is a protein manufactured exclusively by the prostate gland. Elevated levels of PSA is a potential indicator of prostate cancer.

Example:
In order to assess the value of an employee training course a manager in a large corporation compared the weekly sales of a small group of its employees before and after having them attend the course. The following results were observed.

	Employee				
	Bob	Bill	Brita	Babs	Biff
Weekly Sales After Training ($)	5300	4900	4800	5200	4600
Weekly Sales Before Training ($)	4900	4400	4600	4700	4500

Can we conclude, at a 0.05 level of significance, that an employee's weekly sales will increase as a result of the course?

Solution:
Let A denote after and B denote before. If we calculate the paired-difference $d = x_A - x_B$ we have:

	Employee				
	Bob	Bill	Brita	Babs	Biff
x_A ($)	5300	4900	4800	5200	4600
x_B ($)	4900	4400	4600	4700	4500
$d = x_A - x_B$ ($)	400	500	200	500	100

We now forget about the original pairs and focus on the differences for our hypothesis test.

Step 1) Formulate the null and alternative hypotheses.

- $H_0 : \mu_d = 0$ equivalent to $\mu_A - \mu_B = 0$ which is equivalent to $\mu_A = \mu_B$
- $H_a : \mu_d > 0$ equivalent to $\mu_A - \mu_B > 0$ which is equivalent to $\mu_A > \mu_B$

(We suspect the weekly sales after are better than before.)

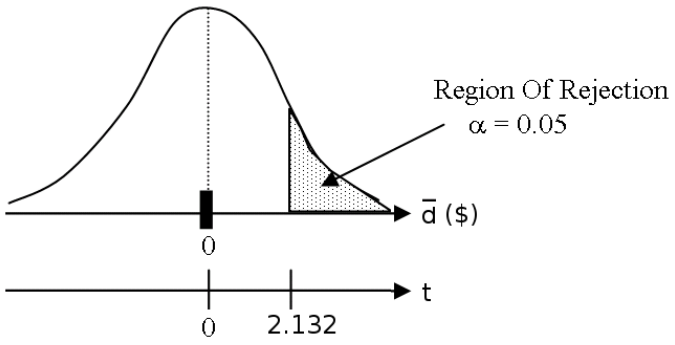
Step 2) State the level of significance.

$\alpha = 0.05$

Step 3) Determine the test statistic.

A t-test will apply since the sample size is $n = 5$ which is less than 30.

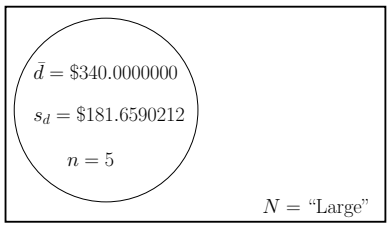
Step 4) Establish a decision rule.



Our t distribution has $df = 5 - 1 = 4$ degrees of freedom. The critical value for $\alpha = 0.05$ for the right-tailed test is therefore, from the t-table, $t_{\text{critical}} = 2.132$.

Step 5) Evaluate the evidence. (Use a Venn diagram for the statistics.)

For the sample d data values the (unrounded) mean and standard deviation are:



In order to determine how extreme our sample data is we find its **calculated *t*-score**. The goal is therefore to evaluate

$$\boxed{t = \frac{\bar{d} - 0}{s_{\bar{d}}}}$$

The required standard error of the difference is:

$$\boxed{s_{\bar{d}} = \frac{s_d}{\sqrt{n}}} = \frac{\$181.659\ldots}{\sqrt{5}} = \$81.24038405$$

The calculated t-score is: $t = \dfrac{\bar{d} - 0}{s_{\bar{d}}} = \dfrac{\$340.00 - 0}{\$81.24\ldots} = 4.185$

(Here we used that the mean of the sampling distribution is zero ($\mu_d = 0$) by hypothesis.)

Step 6) State your decision.

Since the calculated t-value lies in the region of rejection ($4.185 > 2.132$) we reject H_0 and accept H_a. Evidence suggests a significant increase in the weekly sales for an employee who receives the training.

Note that in the previous example a t distribution was required due to the size of the sample being small. As such we must assume that the underlying d distribution is approximately normal. If the sample size was large enough ($n \geq 30$) a normal distribution could be used for the difference. In this course expect the pair-difference problems to be for small samples since otherwise the difference calculation becomes unwieldy.

We already have seen in Unit 5 that the same data can be used both for testing a hypothesis about a population parameter and for making a confidence interval estimate of a parameter. In that case we considered the two concepts separately to establish the fundamentals. For two population experiments we will illustrate the two concepts together for brevity and contrast. In the case of a paired difference experiment the confidence interval statement will have the form:

$$\boxed{P\left(\bar{d} - E < \mu_1 - \mu_2 < \bar{d} + E\right) = 1 - \alpha}$$

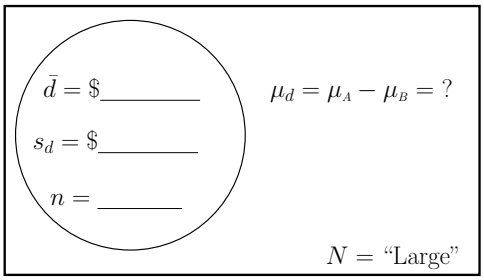
where we have used the fact that $\mu_d = \mu_1 - \mu_2$. The maximum error of estimate here is given by $\boxed{E = t \cdot s_{\bar{d}}}$ where we are assuming our problems will involve t distributions.

Example:
Having determined in the last example that the training improved the weekly sales of employees the manager now wants to know by how much. Create a 95% confidence interval for the difference in weekly sales generated by training the five employees in the previous example.

Answer:
$P(\,\$340 - \$225.52 < \mu_A - \mu_B < \$340 + \$225.52\,) = P(\,\$114.48 < \mu_A - \mu_B < \$565.52\,) = 0.95$

Solution:

Step 1) Identify all given information with symbols on a Venn diagram. (This requires calculating the differences as before.)



Step 2) Draw a diagram of the sampling distribution. (Note: Here a t distribution is required since $n < 30$.)

Step 3) Determine the t-value.

- Identify the confidence coefficient, $1 - \alpha =$ _____
- Decide on the number of degrees of freedom, $df =$ _____.
- The t-value from the tables is $t =$ _____.

Step 4) Calculate the standard error.

$s_{\bar{d}} = \dfrac{s_d}{\sqrt{n}} =$

Step 5) Calculate the maximum error of estimate.

$E = t \cdot s_{\bar{d}} =$

Step 6) Make the confidence interval statement.

$$P\left(\bar{d} - E \;<\; \mu_A - \mu_B \;<\; \bar{d} + E\right) = 1 - \alpha$$

$$P(\;\_\_\_\_\_ - \_\_\_\_\_ \;<\; \mu_A - \mu_B \;<\; \_\_\_\_\_ + \_\_\_\_\_\;) = \_\_\_\_\_$$

$$P(\;\_\_\_\_\_ \;<\; \mu_A - \mu_B \;<\; \_\_\_\_\_\;) = \_\_\_\_\_$$

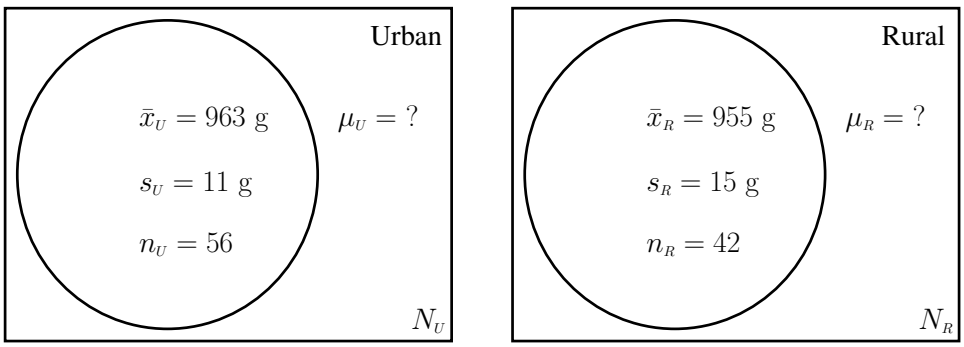
In words, what is the meaning of your interval?

6.1.3 Large Independent Samples

When we turn our attention to independent samples we no longer can reduce the problem to a single variable. There is no meaningful pairing between the measurements in the two samples. Indeed the two sample sizes are in general going to be different. We must instead consider the relationship between the two independent sample statistics $\bar{x}_1$ and $\bar{x}_2$ measured on samples taken from each population and consider the sampling distribution of their difference, $\bar{x}_1 - \bar{x}_2$.

Example:
Imagine that two biologists want to test the "city mouse versus country mouse" conjecture that there may be significant differences between urban and rural rodents of the same species. They decide to test this by considering the mean weight of adult male fox squirrels in summer. Suppose the biologists suspect that the mean weight of a fox squirrel living in an urban (U) setting, μ_U, is greater than the mean weight of a fox squirrel living in a rural (R) setting, μ_R. Obviously one cannot test every fox squirrel in each region to calculate their respective population mean weights to see if they are the same, so a sample would be taken in each region and its mean value calculated. These sample means would have some estimate error from their actual population means, so a difference in the sample values would not necessarily imply a difference in the population means of the different squirrel types. How can we resolve this? We need to know how the difference of the sample means, $\bar{x}_U - \bar{x}_R$, is distributed.

Specifically, suppose that two biologists trap and release adult male squirrels in each region over a set summer period. The biologist in the urban region catches 56 squirrels while the biologist in the rural region catches 42 squirrels. Graphically we could have the following results:



Assuming the squirrels were sampled with replacement, then the biologist in the urban region observed one of the N_U^{56} possible urban sample means while the biologist in the rural region observed one of the N_R^{42} possible rural sample means. Given the large size of the samples (≥ 30) the C.L.T. describes the behaviour of each of these distributions ($\bar{x}_U$ and $\bar{x}_R$) independently. If the biologists measured repeated samples of the same sizes (56 and 42 respectively) and we were able to list the joint possibilities and calculate the difference in their results we would see a table such as this:

$\bar{x}_U$ (g)	$\bar{x}_R$ (g)	$\bar{x}_U - \bar{x}_R$ (g)
963	955	8
964	952	12
960	961	-1
- - -	- - -	- - -
- - -	- - -	- - -

Here the first difference entry, 8 g, is the one which was measured out of many possible such measurements. There will be the same number of observable possible differences as there are

individual possible paired samples. The difference of sample means $\bar{x}_U - \bar{x}_R$ is itself a variable over the population comprised of paired samples.

As just demonstrated, the difference of sample means is itself a variable over the population comprised of paired samples. If two large ($n_1 \geq 30$, $n_2 \geq 30$) independent and random samples are taken from two populations, the difference of the sample means $\bar{x}_1 - \bar{x}_2$ has these three characteristics:[2]

1. The shape of the distribution will be very close to normal.
2. The mean difference will be the difference of the population means:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

3. The standard error of the difference between means is found by this formula:[3]

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In practice, the standard deviations of the population are unknown. In the case of large samples each sample standard deviation is used as an estimator of the corresponding population value. In practice the standard error estimator for the difference of sample means therefore is:

$$\boxed{s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

With this knowledge we can now do hypothesis tests and confidence intervals with large independent samples. For a hypothesis test the required z-score will be

$$\boxed{z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}}}$$

Example:
Continuing the fox squirrel example, if the actual measured sample statistics were:

Sample	Mean Weight (g)	Std Dev (g)	Number of Squirrels
Urban	963.0	11.0	56
Rural	955.0	15.0	42

,

Test whether the weight of the urban squirrels is significantly higher than that of the rural ones at the 0.01 level of significance. Use the critical value approach.

[2]The sum (difference) of two normal variables x and y is itself normally distributed with a mean which is the sum (difference) of the means, $\mu_{x \pm y} = \mu_x \pm \mu_y$, and a variance satisfying $\sigma^2_{x \pm y} = \sigma^2_x + \sigma^2_y$. In our case the two variables are $\bar{x}_1$ and $\bar{x}_2$. Since our samples are of large size, these are, by the C.L.T., normally distributed with means equalling μ_1 and μ_2 respectively and standard deviations approximated by the sample standard errors in the mean, $s_1/\sqrt{n_1}$ and $s_2/\sqrt{n_2}$, from which our results above follow. It is to be noted that the general property, namely that the variance of the sum or difference of two normal variables is added in quadrature, is also useful in combining error estimates in calculations.

[3]If we were sampling without replacement and the populations were finite (i.e. $n_1/N_1 \geq 0.05$ or $n_2/N_2 \geq 0.05$ an $(F.C.F.)^2$ term would need to be introduced in front of each of finite population terms under the root sign. (See Section *4.7.) We continue our assumption in this unit of replacement sampling so that even in the case of finite populations no such modification will be required. Most of our examples will involve infinite populations for which no modifications are ever required irrespective of whether the sampling is done with replacement or not.

Solution:

Step 1) Formulate the null and alternative hypotheses.

The null hypothesis would take the view that the populations of squirrels were the same. That is that there should be no difference between the means in the populations. The hypotheses statements must be formulated as differences so that we can use the sampling distribution for the test. Let the subscript U refer to urban and R to rural.

- $\mathrm{H}_0 : \mu_U - \mu_R = 0 \;\Leftarrow$ Equivalent to $\mu_U = \mu_R$
- $\mathrm{H}_\mathrm{a} : \mu_U - \mu_R > 0 \;\Leftarrow$ We suspect $\mu_U > \mu_R$ which implies this.

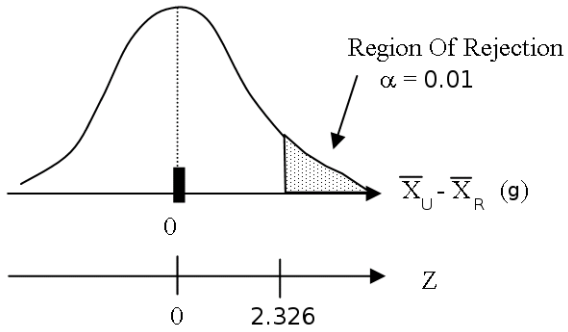
Step 2) State the level of significance.

Use an $\alpha = 0.01$ as directed in the problem.

Step 3) Determine the test statistic.

A z test applies because the decision will be based on large ($n_U \geq 30$, $n_R \geq 30$), independent samples so the difference in sample means is approximately normal.

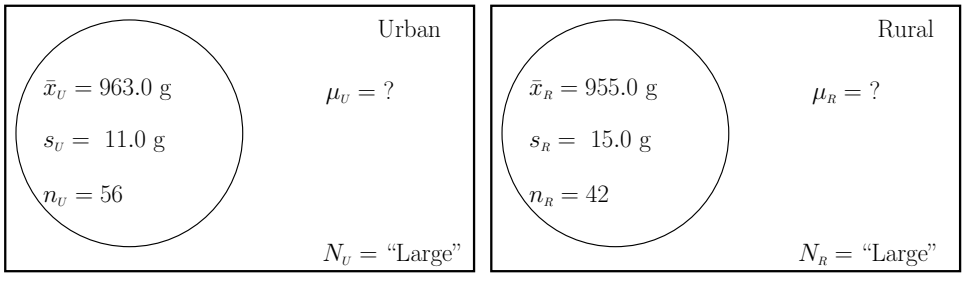
Step 4) Establish a decision rule.



For the given tail area we have $z_{\text{critical}} = 2.326$.

Step 5) Evaluate the evidence.

Using Venn diagrams to label our statistics we have:



In order to determine how extreme our sample data is we find its **calculated z-score**. The goal is therefore to evaluate

$$z = \frac{(\bar{x}_U - \bar{x}_R) - 0}{s_{\bar{x}_U - \bar{x}_R}}$$

The observed difference in means is:

$$\bar{x}_U - \bar{x}_R = 963.0 \text{ g} - 955.0 \text{ g} = 8 \text{ g}$$

The standard error of the difference of means is:

$$s_{\bar{x}_U - \bar{x}_R} = \sqrt{\frac{s_U^2}{n_U} + \frac{s_R^2}{n_R}} = \sqrt{\frac{(11 \text{ g})^2}{56} + \frac{(15 \text{ g})^2}{42}} = 2.741871103 \text{ g}$$

The z-value for our measured difference is calculated by:

$$z = \frac{(\bar{x}_U - \bar{x}_R) - 0}{s_{\bar{x}_U - \bar{x}_R}} = \frac{8 - 0}{2.741\ldots} = 2.918$$

Step 6) State your decision.

Because the calculated z-value (2.918) lies in the region of rejection ($z > 2.326$) we reject H_0 and accept H_a. Based on the evidence, at a 0.01 level of significance, the urban adult squirrels weigh more on average than their rural counterparts.

(If we had calculated a z-value like 0.87, the decision would be to fail to reject H_0 . There is no evidence to suggest a difference in the average weights between the two populations of squirrel.)

Example:
Repeat the previous hypothesis test using the P-value approach.
(Answer: P-value = 0.0018 less than $\alpha = 0.01 \Rightarrow$ reject H_0 and accept H_a.)

Solution:
Steps 1-3 are the same as above.

Step 4) Evaluate the evidence. (Use a Venn diagram for the statistics.)

Same as step 5) above to get calculated $z = 2.918 \approx 2.92$

Step 5) Find the P-value.

Step 6) State your decision.

A confidence interval for the difference of means for large independent samples has the general form[4]

$$\boxed{P(\,[\bar{x}_1 - \bar{x}_2] - E < \mu_1 - \mu_2 < [\bar{x}_1 - \bar{x}_2] + E\,) = 1 - \alpha}$$

where the maximum error of estimate is $\boxed{E = z \cdot s_{\bar{x}_1 - \bar{x}_2}}$.

Example:
Using the same data as before,

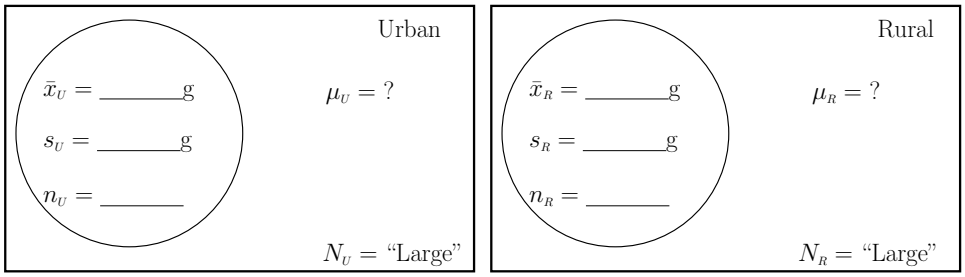
Sample	Mean Weight (g)	Std Dev (g)	Number of Squirrels
Urban	963.0	11.0	56
Rural	955.0	15.0	42

,

construct a 99% confidence interval for the difference in the mean weight of the rural and urban adult male fox squirrels, $\mu_U - \mu_R$.
Answer: $P(8.0 \text{ g} - 7.1 \text{ g} < \mu_R - \mu_U < 8.0 \text{ g} + 7.1 \text{ g}) = P(0.9 \text{ g} < \mu_R - \mu_U < 15.1 \text{ g}) = 0.99$

Solution:

Step 1) Identify all given information with symbols on a Venn diagram.



The point estimate for the difference of means is $\bar{x}_U - \bar{x}_R =$ _____ g

Step 2) Draw a diagram of the sampling distribution. (Note: Here a z distribution is used since $n_U \geq 30$ and $n_R \geq 30$.)

Step 3) Determine the z-value.

[4]Note the square brackets around the means in $[\bar{x}_1 - \bar{x}_2]$ are not absolute value bars since the difference can be negative. The square brackets are there simply to indicate that this is the point estimate of the population mean difference and that it should be calculated first.

Step 4) Calculate the standard error.

$s_{\bar{x}_U - \bar{x}_R} =$

Step 5) Calculate the maximum error of estimate.

$E = z \cdot s_{\bar{x}_U - \bar{x}_R} =$

Step 6) Make the confidence interval statement.

$$P\left([\bar{x}_U - \bar{x}_R] - E \;<\; \mu_U - \mu_R \;<\; [\bar{x}_U - \bar{x}_R] + E \right) = 1 - \alpha$$

$$P\left(\_\_\_\_\_ - \_\_\_\_\_ \;<\; \mu_U - \mu_R \;<\; \_\_\_\_\_ + \_\_\_\_\_ \right) = \_\_\_\_\_$$

$$P\left(\_\_\_\_\_ \;<\; \mu_U - \mu_R \;<\; \_\_\_\_\_ \right) = \_\_\_\_\_$$

In words, what is the meaning of your interval?

6.1.4 Small Independent Samples

In certain cases it is too costly or time consuming to draw large samples in order to test assertions about the difference between population means. If smaller sample sizes are used ($n_1 < 30$ and/or $n_2 < 30$), a different sampling distribution is used in the decision rule.

In this case the difference of sample means will not have a normal distribution because the sample sizes are not large enough. However, we do know from Section 5.1.4 that if the underlying distribution of the population variables, x_1 and x_2 respectively, are approximately normal then the sample means, $\bar{x}_1$ and $\bar{x}_2$, will each have a t distribution. It turns out that this will also be the case for the difference of the sample means, $\bar{x}_1 - \bar{x}_2$, under the further assumption that the underlying populations have the same standard deviation, $\sigma_1 = \sigma_2 = \sigma$. In this case if the samples are drawn randomly and independently the sampling distribution of the difference between sample means $\bar{x}_1 - \bar{x}_2$ has these three characteristics:

1. The shape of the distribution will be a t distribution with

$$\boxed{df = n_1 + n_2 - 2}$$

   degrees of freedom.

2. The mean difference will be the difference of the population means:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

3. The standard error of the difference between means is found by the formula:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2 \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

where σ is the common standard deviation of the two populations.

In practice, the standard deviation of the populations, σ, is unknown. In the case of small samples, the variability in each sample is pooled into one large sample to obtain an estimator of the population standard deviation. The pooled estimator of σ is s_{Pool}. Recall from page 190 that the variance of a sample can be defined as:

$$s^2 = \frac{\sum (x - \bar{x})^2}{df}$$

The appropriate pooled estimator of variance, s^2_{Pool}, reflects the above definition:

$$s^2_{\text{Pool}} = \frac{\sum_{\text{Sample 1}} (x - \bar{x}_1)^2 + \sum_{\text{Sample 2}} (x - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} .$$

Substituting the pooled variance estimator s^2_{Pool} for σ^2 in the theoretical standard error difference formula gives the standard error of the difference of means estimator:

$$\boxed{s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} .$$

For hypothesis tests the needed calculated statistic will be

$$\boxed{t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{\bar{x}_1 - \bar{x}_2}}} .$$

Example:
Platinum is a very expensive metal. A purchaser wishes to know if there is any significant difference in the mean weight of a platinum item supplied by Supplier A versus those supplied by supplier B. It is assumed that the weight of platinum parts supplied by the suppliers is distributed normally with the same standard deviation. Because of the expense involved, a small random sample of parts will be tested from each supplier. Test the assumption that there is no difference in the weights of parts supplied by these two suppliers at the 10% level of significance if 12 items are tested from each supplier with the following results:

Supplier	Mean Weight (g)	Standard Deviation (g)	Number of Items
A	17.0	2.0	12
B	20.0	4.0	12

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\mathrm{H}_0 : \mu_A - \mu_B = 0$
- $\mathrm{H}_\mathrm{a} : \mu_A - \mu_B \neq 0$ $\Leftarrow$ We suspect $\mu_A \neq \mu_B$ which implies this.

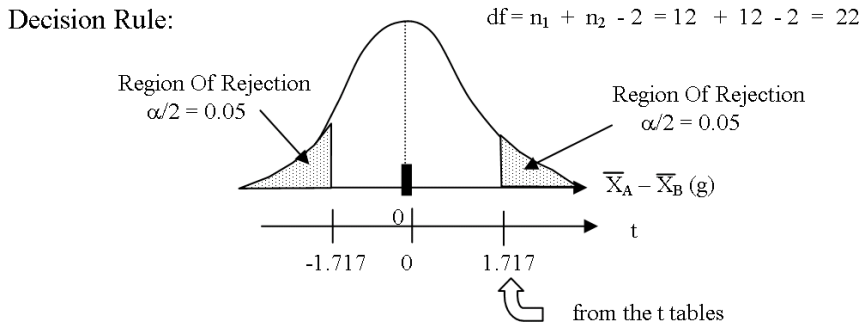
Step 2) State the level of significance.

Use an $\alpha = 0.10$ as directed in the problem.

Step 3) Determine the test statistic.

A t test applies because the decision will be based on small, independent samples drawn randomly from normal populations with equal standard deviations.

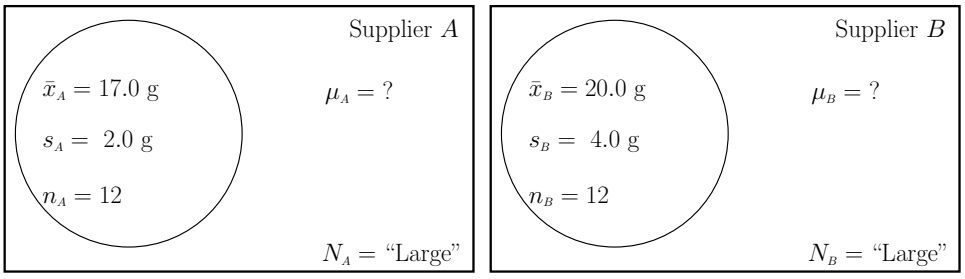
Step 4) Establish a decision rule.



For our given tail area and using a t distribution we have $t_{\text{critical}} = \pm 1.717$.

Step 5) Evaluate the evidence.

The provided data is summarized in the following Venn diagrams:



(Note if we had been given the actual 12 measurements from A and B we would have to have calculated the sample means and standard deviations ourselves.)

Our goal is to calculate t for the data, $t = \dfrac{(\bar{x}_A - \bar{x}_B) - 0}{s_{\bar{x}_A - \bar{x}_B}}$.

The observed difference in means is

$$\bar{x}_A - \bar{x}_B = 17.0 \text{ g} - 20.0 \text{ g} = -3.0 \text{ g}\,.$$

The required standard error of the sampling distribution is

$$\begin{aligned} s_{\bar{x}_A - \bar{x}_B} &= \sqrt{\frac{(n_A - 1)\cdot s_A^2 + (n_B - 1)\cdot s_B^2}{n_A + n_B - 2} \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]} \\ &= \sqrt{\frac{(12 - 1)\cdot (2 \text{ g})^2 + (12 - 1)\cdot (4 \text{ g})^2}{12 + 12 - 2} \cdot \left[\frac{1}{12} + \frac{1}{12}\right]} \\ &= 1.290994449 \text{ g}\,. \end{aligned}$$

Finally the calculated t-value for the observed sample difference is

$$t = \frac{(\bar{x}_A - \bar{x}_B) - 0}{s_{\bar{x}_A - \bar{x}_B}} = \frac{-3.0 - 0}{1.29\ldots} = -2.324\,.$$

Step 6) State your decision.

Because the calculated t-value (-2.324) lies in the region of rejection ($t < -1.717$ or $t > 1.717$) the decision is to reject H_0 and accept H_a. Based on the evidence, at a 0.10 level of significance, there is a difference in the mean weight of the platinum parts supplied by these two suppliers.

We can use the same data to estimate the amount of the difference between mean weights from the two suppliers. The confidence interval will be the same as that for the two random and independent large samples but the maximum error of estimate will now require a t-score, $\boxed{E = t \cdot s_{\bar{x}_1 - \bar{x}_2}}$.

Example:
Using the same data as before,

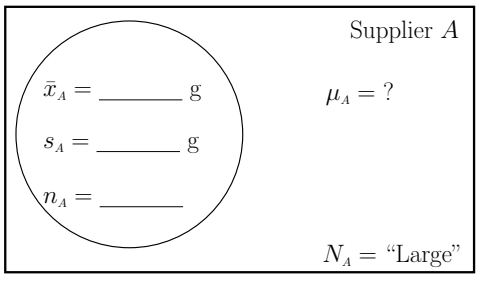
Supplier	Mean Weight (g)	Standard Deviation (g)	Number of Items
A	17.0	2.0	12
B	20.0	4.0	12

,

find a 90% confidence interval for the difference of the mean weights, $\mu_A - \mu_B$, from the two suppliers. Answer:
$P(-3.0 \text{ g} - 2.2 \text{ g} < \mu_A - \mu_B < -3.0 \text{ g} + 2.2 \text{ g}) = P(-5.2 \text{ g} < \mu_A - \mu_B < -0.8 \text{ g}) = 0.90$

Solution:

Step 1) Identify all given information with symbols on a Venn diagram.



The point estimate for the difference of means is $\bar{x}_A - \bar{x}_B =$ _____ g

Step 2) Draw a diagram of the sampling distribution. (Note: Here a t distribution is required since $n_A < 30$ and/or $n_B < 30$.)

Step 3) Determine the t-value.

- Identify the confidence coefficient, $1 - \alpha =$ _____.
- Decide on the number of degrees of freedom, $df =$ _____.
- The t-value from the tables is $t =$ _____.

Step 4) Calculate the standard error.

$s_{\bar{x}_A - \bar{x}_B} =$

Step 5) Calculate the maximum error of estimate.

$E = t \cdot s_{\bar{x}_A - \bar{x}_B} =$

Step 6) Make the confidence interval statement.

$$P\left([\bar{x}_A - \bar{x}_B] - E \;<\; \mu_A - \mu_B \;<\; [\bar{x}_A - \bar{x}_B] + E \right) = 1 - \alpha$$

$$P\left(\_\_\_\_\_ - \_\_\_\_\_ \;<\; \mu_A - \mu_B \;<\; \_\_\_\_\_ + \_\_\_\_\_ \right) = \_\_\_\_\_$$

$$P\left(\_\_\_\_\_ \;<\; \mu_A - \mu_B \;<\; \_\_\_\_\_ \right) = \_\_\_\_\_$$

A few things to think about:

1. Given the same α was involved in both the confidence interval and the two-sided hypothesis test, how can we check the results are consistent?
2. What is the significance of the fact that the confidence interval involves negative values? Shouldn't weights (technically masses) always be positive?
3. We stated that for this t-test to be valid the standard deviations had to be equal, but the standard deviations given are 2.0 g and 4.0 g which are not equal. Are our results therefore invalid?

Assignment:

Be careful in identifying the type of mean difference problem in the following. Also use the formal six step procedures for confidence intervals and hypothesis tests.

1. A researcher wishes to test if there is a difference between the average monthly household income in two large communities. A random sample of 40 households in the first community has a mean of \$5,900 with a standard deviation of \$540. For the second community a sample of 30 households has a mean of \$5,600 with a standard deviation of \$420.

   (a) Using a level of significance of 0.05, test the hypothesis that there is no difference between the average monthly household income in the two communities.

   (b) Create a 95% confidence interval for the difference of the mean monthly household incomes.

2. A firm hires a large number of professional people to carry out the services it offers. At salary negotiation time, the question is always raised as to whether the professionals at this firm earn similar wages on average to professionals in other firms who do similar duties. Two random samples, one of the internal (I) and one of the external (E) professionals, are taken with the following results:

Professionals	Average Wage(\$/yr)	Std Dev(\$/yr)	Number
Internal	55000	5000	15
External	58000	4000	17

   (a) Assuming the underlying wage distributions are approximately normal with similar standard deviations, can the firm conclude at the 1% level of significance that its professionals earn less based on these sample results?

   (b) Construct a 95% confidence interval for the difference in the mean population wages between the internal and external professionals $(\mu_I - \mu_E)$.

*3 After playing a round of golf at the *Dutch Treat Mini Golf* at Greig Lake Saskatchewan, a group of statisticians relaxed for several rounds of beer. To test the effect, if any, of the alcohol on their performance, they returned and played a second round of mini golf. The results are summarized on the following table.

Statistician	Score Sober	Score Less than Sober
Freundin	62	79
Lindley	68	77
Engel	102	163
Murphey	52	72
Roberts	88	109
March	75	101

   (a) Construct a 95% confidence interval for the difference $(\mu_L - \mu_S)$ between a statistician's score at the mini golf establishment when they are less than sober (L) and when they are sober (S).

   (b) Test the hypothesis, at the 0.05 level of significance, that there is a difference in the ability of a statistician to play mini golf sober (S) and after a few beers (L).

*Questions denoted with an asterisk depend on material in optional sections of the course.

4. In an attempt to evaluate the hypothesis that a "good night's sleep" gives better performance, a researcher created two groups of subjects. Those with regular sleep (group R) were allowed to sleep up to 7 hours maximum at night, while the second group with enhanced sleep (group E) were allowed up to 9 hours of sleep. After 1 week all individuals in the study were timed to see how long it took each of them to accomplish the same task involving motor skills and coordination. The 40 individuals in the regular sleep group had a mean time of 53.1 seconds with a standard deviation of 8.0 seconds, while the 50 in the enhanced sleep group took an average time of 51.2 seconds with a standard deviation of 9.0 seconds to accomplish the task. Can we concluded that the average time required by a person with enhanced sleep (E) is significantly less than that of a person with regular sleep (R) at the 10% level of significance?

5. A manager in a corporation evaluates a new sales training course by evaluating the average sales performance between salespeople who took the training course compared to salespeople who did not receive the training with the following results:

	Mean Weekly Sales ($)	Standard Deviation ($)	Number In Group
Trained Group	5000	800	60
Untrained Group	4700	1100	80

   (a) Test whether the average sales of the employees who have taken the training course is significantly higher than that of those who have not at the 0.05 level of significance. Also find a P-value for your test.

   (b) Find a 98% confidence interval for the difference of the weekly sales of the two groups of employees, $\mu_T - \mu_{NT}$.

   (c) This problem used large independent samples to solve essentially the same problem given in the example in Section *6.1.2 where a paired difference approach was used. Which manager's approach to analyzing the value of the training is better?

*6 The Moose Jaw Sporting Goods Company wishes to evaluate the use of advertising to increase sales. To do so they placed an advertisement on television that was shown three times during the live broadcast of the Saskatchewan Roughriders' first away game which occurred on a Sunday evening. The sales for each of the seven days after the ad was placed is to be compared with the sales for the seven days immediately before running the ad. The following data representing the total dollar sales each day were collected:

Day	Sales Before Ad ($)	Sales After Ad ($)
Mon	1,765	2,045
Tue	1,543	2,456
Wed	2,867	2,590
Thu	1,490	1,510
Fri	2,800	2,850
Sat	1,379	1,255
Sun	2,097	2,255

   (a) Based on the sample data, would you conclude that there was a significant increase in sales after the advertising? Test using a level of significance of 0.01 .

   (b) Construct a 95% confidence interval for the difference in daily sales ($\mu_A - \mu_B$) before and after advertising.

7. Recall in question 6 on page 206 an astronomer measured the metallicity of 16 stars that appeared to be clustered together (C) with the following results (dex):

-0.50	0.00	0.02	0.06	0.08	0.09	0.11	0.12
0.14	0.19	0.19	0.20	0.22	0.25	0.30	0.38

   In the original question she tested the hypothesis that the mean metallicity of the cluster stars equalled the published measurement of the mean star metallicity in a larger region of the sky. Rather than trust this value, the astronomer decides to measure the metallicity of 10 background stars (B) nearby, but not within, the observable cluster with the following results (dex):

-0.62	-0.04	-0.21	-0.45	-0.19
-0.58	-0.09	-0.50	-0.30	-0.36

   (a) Is the mean metallicity of the cluster stars equal to that of the background? Test the claim at the 1% level of significance.

   (b) Construct a 99% confidence interval for the difference in the metallicity of the cluster stars and that of the background ($\mu_C - \mu_B$).

   (c) For these small sample calculations to be valid the underlying metallicity distributions must be approximately normal. What further assumption is required in our analysis and is it reasonable given the data?

8. An insurance company wishes to know whether foreign cars are more expensive to repair than domestic cars on average. Can their claim be supported at the 5% level of significance if the following sample results are observed?

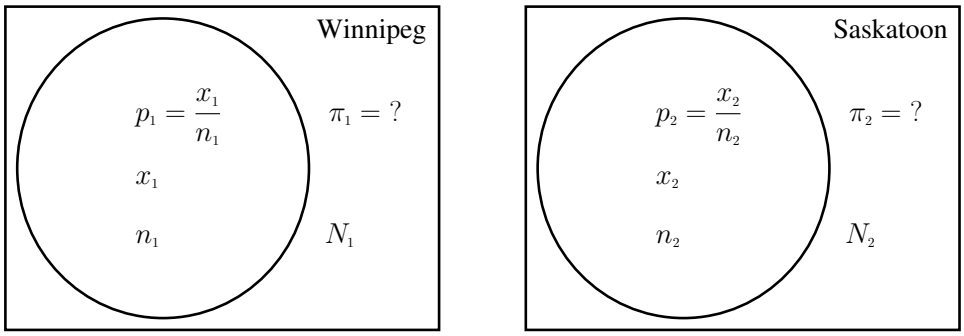
	Average Repair ($)	Std Dev ($)	Number of Cars
Domestic	8000	2000	50
Foreign	9500	3100	60

*6.2 Difference Between Proportions (Large Samples)

6.2.1 Introduction

As with the comparison of two means, we may also be interested in comparing two proportions. For instance, we might be interested in knowing if two different groups have a different incidence of diabetes. If we could test each member of the two groups to see if they had diabetes we could work out the proportions to see what the difference was, or compare the proportions to see if they are equal or not. However this is too costly so we want, instead, to sample each group and get estimates of the proportions and compare these. Due to error in estimation, each such proportion differs from its true group's proportion and thus a difference in sample proportions does not imply a difference in the two actual proportions. We need to know how the difference between two sample proportions drawn from the two populations is distributed.

Example:
Consider the case of two market analysts one in Winnipeg and one in Saskatoon. They wish to sample their respective populations of households to estimate the proportion of households that consume more than two packages of potato chips per week.



Here the underlying qualitative variable is x, the consumption behaviour of the household. If in Winnipeg there are x_1 respondents who consume more than two packages of potato chips per week, then, assuming the sample size was n_1, the estimate would be $p_1 = x_1/n_1$ of the unknown proportion of households π_1 in Winnipeg. A similar analysis applies for Saskatoon. What can we say about the distribution of the difference of the estimates, $p_1 - p_2$?

Sampler 1 could find any of $N_1^{n_1}$ possible sample proportions while Sampler 2 could find any of $N_2^{n_2}$ possible sample proportions (assuming replacement sampling). Assuming the sample sizes are large enough these sample proportions will each be normally distributed. If we were able to list the joint possibilities from the samplers and calculate the difference in their results we could see a table like this:

p_1	p_2	$p_1 - p_2$
0.32	0.29	0.03
0.22	0.26	0.04
0.31	0.30	0.01
- - -	- - -	- - -
- - -	- - -	- - -

The difference between proportions $p_1 - p_2$ is itself a variable over the population of paired samples.

If two random and independent samples are drawn from two populations and the sample sizes are sufficiently large so that $n_1\pi_1$, $n_1(1-\pi_1)$, $n_2\pi_2$, $n_2(1-\pi_2)$ are all greater than 5, the distribution of the difference between the sample proportions $p_1 - p_2$ has these three characteristics:[5]

1. The difference $p_1 - p_2$ will be normally distributed.
2. The mean of the proportion differences will be the difference of the population proportions:

$$\mu_{p_1-p_2} = \pi_1 - \pi_2$$

3. The standard error of the difference between proportions will be:[6]

$$\sigma_{p_1-p_2} = \sqrt{\frac{\pi_1 \cdot (1-\pi_1)}{n_1} + \frac{\pi_2 \cdot (1-\pi_2)}{n_2}}$$

6.2.2 Confidence Intervals

If we are interested in knowing how great a difference there is between two population proportions, $\pi_1 - \pi_2$ can be estimated with the point estimator $p_1 - p_2$ which we can measure by taking two samples. The more useful interval estimator with confidence coefficient $1-\alpha$ is:

$$\boxed{P(\,[p_1 - p_2] - E < \pi_1 - \pi_2 < [p_1 - p_2] + E\,) = 1 - \alpha}\,.$$

In calculating the required maximum error of estimate, E, we will need the following point estimator of the standard error of proportion since the true population proportions in the theoretical formula are in practice unknown:

$$\boxed{s_{p_1-p_2} = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}}}\,.$$

The maximum error of estimate is then $\boxed{E = z \cdot s_{p_1-p_2}}$.

Example:
A biologist believes that the population of elk in the boreal forest north of Prince Albert has a higher proportion of adult males than a southern Saskatchewan population of elk due to potentially several factors (different hunting pressure, etc.). The following data are collected (adult elk only).

Location	Number of Males	Number of Elk Observed
Boreal	91	350
Southern	32	200
Total	123	550

Find a 95% confidence interval for the difference in the proportion of adult males between the boreal and southern populations. Assume the underlying populations in each region are much larger than the samples taken.

[5] These results follow from the properties of combining normal variables outlined in the footnote on page 220 and the known standard deviation and mean of the proportion sampling distributions.

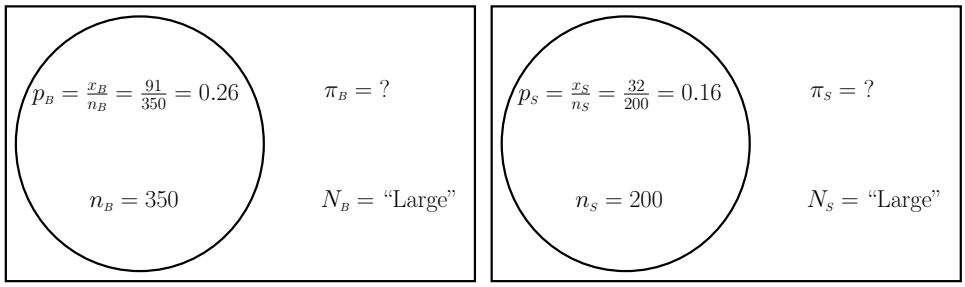
[6] Here if we wished to sample without replacement we would have to know that the populations were "infinite" with respect to their samples (i.e. $n_1/N_1 < 0.05$, $n_2/N_2 < 0.05$). If this were not the case an $(F.C.F.)^2$ term would need to be introduced in front of each of the terms under the root sign as discussed in Section *4.7. Since we assume replacement sampling in all our problems such a modification will never be required in this course, even for finite populations.

Solution:
The goal here is to fill in the confidence interval statement for the difference of population proportions $\pi_B - \pi_S$ whose point estimator is $p_B - p_S$:

$$P(\,[p_B - p_S] - E < \pi_B - \pi_S < [p_B - p_S] + E\,) = 1 - \alpha$$

Note we will use the meaningful subscripts B and S to denote the boreal and southern regions respectively rather than the 1 and 2 which appear above.

Step 1) Identify all given information with symbols on a Venn diagram.

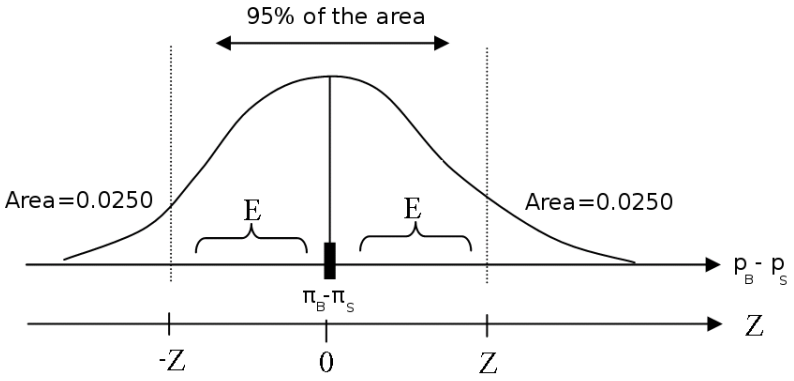


Note that the point estimate of $\pi_B - \pi_S$ is $p_B - p_S = 0.26 - 0.16 = 0.10$.

Step 2) Draw a diagram of the sampling distribution.
The variable of interest is $p_B - p_S$. Using point estimates of the the unknown population proportions we have that

- $n_B\pi_B \approx (350)(.26) = 91$
- $n_B(1 - \pi_B) \approx (350)(.74) = 259$
- $n_S\pi_S \approx (200)(.16) = 32$
- $n_S(1 - \pi_S) \approx (200)(.84) = 168$

Since all of these values exceed 5 we know that $p_B - p_S$ is approximately normally distributed with (unknown) mean $\pi_B - \pi_S$:



Step 3) Determine the z-value.
Use the confidence, here $1 - \alpha = 0.95$, to find the total area in the two tails, $\alpha = 1 - 0.95 = 0.0500$. The area in a single tail is

$$\text{Area} = \frac{\alpha}{2} = \frac{0.0500}{2} = 0.0250\ ,$$

so from the normal table (or the t-table using $df = \infty$), the corresponding z-value is (for the right tail)

$$z = 1.96\ .$$

Step 4) Calculate the standard error.

$$\begin{aligned} s_{p_B - p_S} &= \sqrt{\frac{p_B \cdot (1-p_B)}{n_B} + \frac{p_S \cdot (1-p_S)}{n_S}} \\ &= \sqrt{\frac{(0.26)(0.74)}{350} + \frac{(0.16)(0.84)}{200}} \\ &= 0.034953029 \end{aligned}$$

(Since the two populations, N_B and N_S are "Large" the above formula with no finite corrections was appropriate.)

Step 5) Calculate the maximum error of estimate.

$$E = z \cdot s_{p_B - p_S} = (1.96)(0.03495\ldots) = 0.068507938 = 0.069$$

Step 6) Make the confidence interval statement.

$$\begin{aligned} P(\,[p_B - p_S] - E \;<\; & \pi_B - \pi_S \;<\; [p_B - p_S] + E\,) = 1 - \alpha \\ P(\,0.10 - 0.069 \;<\; & \pi_B - \pi_S \;<\; 0.10 + 0.069\,) = 0.95 \\ P(\,0.031 \;<\; & \pi_B - \pi_S \;<\; 0.169\,) = 0.95 \end{aligned}$$

In words, the boreal male elk proportion is 95% likely to be between 3.1% and 16.9% higher than the corresponding southern proportion.

6.2.3 Hypothesis Tests

The parameter being tested is the difference of two population proportions, $\pi_1 - \pi_2$. As with any hypothesis test we are interested in whether this parameter equals some value or not. The most common case is for this value to be zero since $\pi_1 - \pi_2 = 0$ is an equivalent way of saying we are testing if $\pi_1 = \pi_2$. For this course we will only ever test this hypothesis for the difference of proportions. In symbols we therefore have the null hypothesis:

- $\mathrm{H}_0 : \pi_1 - \pi_2 = 0$ (equivalent to $\pi_1 = \pi_2$)

The alternative hypotheses depend on whether one suspects one proportion is larger, smaller or simply different than the other:

- $\mathrm{H}_\mathrm{a} : \pi_1 - \pi_2 > 0$ (equivalent to $\pi_1 > \pi_2$)
- $\mathrm{H}_\mathrm{a} : \pi_1 - \pi_2 < 0$ (equivalent to $\pi_1 < \pi_2$)
- $\mathrm{H}_\mathrm{a} : \pi_1 - \pi_2 \neq 0$ (equivalent to $\pi_1 \neq \pi_2$)

When doing a hypothesis test with this H_0 we proceed slightly differently than with the confidence interval when we calculate the standard error. The reason is that if we are hypothesizing that the two population proportions have the same value, so $\pi_1 = \pi_2 = \pi$, then the theoretical standard error becomes

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\pi \cdot (1-\pi)}{n_1} + \frac{\pi \cdot (1-\pi)}{n_2}} = \sqrt{\pi \cdot (1-\pi) \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}\,.$$

Since by hypothesis the two sample proportions are equal the variability in each sample is pooled into one large sample to obtain an estimator of the unknown π . The pooled estimator of π is

p_{Pool} given by

$$\boxed{p_{\text{Pool}} = \frac{x_1 + x_2}{n_1 + n_2}}.$$

Substitute the pooled estimator p_{Pool} for π in the standard error formula to obtain

$$\boxed{s_{p_1 - p_2} = \sqrt{p_{\text{Pool}} \cdot (1 - p_{\text{Pool}}) \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}.$$

Finally the calculated z-value for our test will be

$$\boxed{z = \frac{(p_1 - p_2) - 0}{s_{p_1 - p_2}}}$$

because the mean of the sampling distribution, $\pi_1 - \pi_2$, is zero by hypothesis.

Example:
Using the previous elk data, test the hypothesis, at the 1% level of significance, that the proportion of male elk in the adult population is greater in the boreal region than the southern region. Use the critical-value approach.

Solution:
Step 1) Formulate the null and alternative hypotheses.

- $\text{H}_0 : \pi_B - \pi_S = 0$ $\Leftarrow$ There is no difference in the proportions (equivalent to $\pi_B = \pi_S$)
- $\text{H}_\text{a} : \pi_B - \pi_S > 0$ $\Leftarrow$ The biologist suspects $\pi_B > \pi_S$ which implies this.

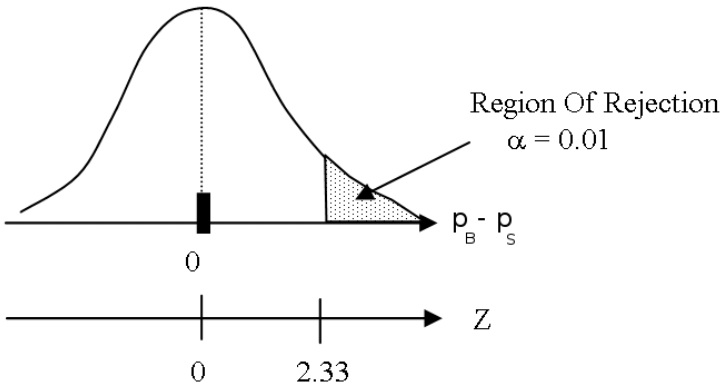
Step 2) State the level of significance.

Use an $\alpha = 0.01$ as directed in the problem.

Step 3) Determine the test statistic.

A z test applies because the decision will be based on large sample observations of proportion. Specifically we check (see evidence below) that $n_B\pi_B \approx (350)(0.26) = 91$, $n_B(1 - \pi_B) \approx (350)(0.74) = 259$, $n_S\pi_S \approx (200)(0.16) = 32$, $n_S(1 - \pi_S) \approx (200)(0.84) = 168$ are all greater than 5 .

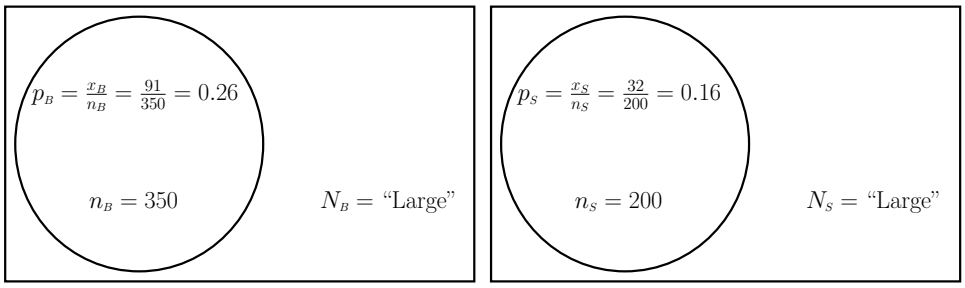
Step 4) Establish a decision rule.



For our given tail area and using the normal distribution we have $z_{\text{critical}} = 2.33$, or 2.326 by the t table. (Note that the mean of $p_B - p_S$ is 0 by hypothesis as shown in the diagram.)

Step 5) Evaluate the evidence. (Use a Venn diagram for the statistics.)

The biologist observed elk in the two regions with the following results:



Our goal is to use the data to calculate $z = \dfrac{(p_B - p_S) - 0}{s_{p_B - p_S}}$.

The observed difference in proportions is: $p_B - p_S = \dfrac{91}{350} - \dfrac{32}{200} = 0.26 - 0.16 = 0.10$

Calculate the pooled estimate of the population proportion.

$$p_{\text{Pool}} = \frac{x_B + x_S}{n_B + n_S} = \frac{91 + 32}{350 + 200} = \frac{123}{550} = 0.223636363$$

Then calculate the standard error of the sampling distribution.

$$\begin{aligned} s_{p_B - p_S} &= \sqrt{p_{\text{Pool}} \cdot (1 - p_{\text{Pool}}) \cdot \left[\frac{1}{n_B} + \frac{1}{n_S}\right]} \\ &= \sqrt{(0.223\ldots) \cdot (1 - 0.223\ldots) \cdot \left[\frac{1}{350} + \frac{1}{200}\right]} \\ &= 0.036934832 \end{aligned}$$

Finally, calculate the z-value for the observed sample difference.

$$z = \frac{(p_B - p_S) - 0}{s_{p_B - p_S}} = \frac{0.10 - 0}{0.0369\ldots} = 2.71$$

Step 6) State your decision.

Because the calculated z-value (2.71) lies in the region of rejection ($z > 2.33$) we decide to reject H_0 and accept H_a. Based on the evidence, at a 0.01 level of significance, the proportion of adult male elk is greater in the boreal region than in the southern region.

Note that the standard errors for the confidence interval and the hypothesis test, while being close (0.03495... versus 0.03693... respectively) are not equal.

Example:
Repeat the previous hypothesis test using the P-value approach.
(Answer: P-value $= 0.0034$ less than $\alpha = 0.01 \Rightarrow$ reject H_0 and accept H_a.)

Solution:
Steps 1-3 are the same as above.

Step 4) Evaluate the evidence. (Use a Venn diagram for the statistics.)
Same as Step 5) above to get calculated $z = 2.71$.

Step 5) Find the P-value.

Step 6) State your decision.

Assignment:

Be careful to choose the correct standard error formulae below. Also use the formal six step procedures for confidence intervals and hypothesis tests.

1. A counselor at a college suspects that a higher proportion of business students holds part-time jobs than do science students. To test the hypothesis, a random sample of 100 business students was taken and 30 had part-time jobs, while a sample of 150 science students found 36 with part-time jobs. The two programs have large enrolments. Can the counselor confirm his suspicion at the 0.05 level of significance?

2. Using the data from problem 1, find a 98% confidence interval for the difference in the proportion of business and science students holding part-time jobs at the college.

3. In the 1975 comedy "Monty Python and the Holy Grail", it is suggested that there would be a difference between the ability of an African swallow and a European swallow to carry a coconut over a large distance. To test the claim, an ornithologist (with too many financial resources and no ethics) attaches coconuts to 40 African swallows and 36 European swallows and releases them from a common location. Using transmitters he has inserted in the coconuts the researcher is able to determine that 18 of the African and 9 of the European swallows were able to travel a distance of over 100 km with coconuts intact.

   (a) Can the researcher conclude at the 20% level of significance that there is a difference in the ability of African and European swallows to travel laden with coconuts?

   (b) How would the problem change if one had hypothesized that African swallows were better coconut carriers than European ones (at the 20% level of significance)?

4. Using the data from problem 3, find a 90% confidence interval for the difference in the proportion of African and European swallows that can travel successfully coconut-laden.

5. A study of financial planning services showed that 380 of 1160 people who were over 45 years of age preferred an accountant for assistance in their financial planning. Of those 45 years of age or less, 120 of 470 surveyed preferred an accountant for assistance in their financial planning.

   (a) Construct a 99% confidence interval for the difference in the proportion of older (> 45 years) and younger (≤ 45 years) people who prefer an accountant for financial planning.

   (b) Using a level of significance of 0.01, is there a difference between the two groups in the proportion who prefer an accountant for financial planning?

   (c) Comment on the relationship between your answers in (a) and (b) .

6.3 Analysis of Variance (ANOVA)

Up to this point we have been interested in comparing a single population mean to a known value (Section 5.2.2) or testing whether means from two populations are equal (Section 6.1). Now we wish to test whether the means of several (k) populations are all equal or if some differ by taking random independent samples from each population. The following illustrates the problem.

Example:
A hospital is interested in quantifying the cost effectiveness of three pain relievers. Patients recovering from the same type of surgery were given the ability (by pressing a button) to self-administer pain medication intravenously after surgery. They were randomly selected to receive one of the three types of pain relief and the hospital kept track of how long it took for a fixed value of drug to be consumed. The results were as follows (in hours):

Pain Reliever 1	Pain Reliever 2	Pain Reliever 3
15	20	9
8	13	13
8	16	7
7	14	9
13	19	7
3	9	10
	14	12
		7

The treatment populations are known to be normally distributed with a common (but unknown) standard deviation σ. For the same dollar value of each drug does the data suggest that the mean pain relief provided by each drug is the same (or do some of them differ) at a level of significance of 5%?

The samples have been randomly selected independently from the $k = 3$ populations. The problem is an example of an experimental design that is **completely randomized**. Furthermore it is referred to as a **one-way classification** because only one **factor** (the drug type) is changing. (See discussion at the end of Section 4.2 .) Our factor has three **levels** determined by the drug type (1, 2, and 3) which designate the populations. If a second relevant factor were included to distinguish the population (say the sex of the patient) there would be six **treatments** (1M, 2M, 3M, 1F, 2F, 3F) in the experiment. In our simple one-way classification problem where only one factor is involved the treatments are the same as that factor.[7] For this course we will only be considering the analysis of problems such as the one above, a completely randomized one-way classification with the additional assumptions, familiar from our analysis of small independent means (Section 6.1.4):

- Each treatment population's observations are normally distributed
- The standard deviations (or equivalently variances) of each population are all equal. ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$)

Analysis of more complicated classification and sampling schemes are beyond the scope of this course.

The hypothesis we wish to test (the null hypothesis) is

[7]The term **treatment** arises from an agricultural context where different chemicals are treatments applied to the soil. For our patient example the terminology of the three different treatments is also appropriate. When, however, statisticians refer to a nationality or a location as a treatment the technical term no longer reflects its common usage.

- $\mathbf{H_0 : \mu_1 = \mu_2 = \ldots = \mu_k}$,

where $\boldsymbol{k}$ is the **number of treatments** involved in the experiment. For our specific example above with three treatments this is:

- $H_0 : \mu_1 = \mu_2 = \mu_3$.

What is the alternative hypothesis? We are only interested in means differing ($\neq$). One way this could happen, to be sure, would be if all the means were different from each other. However other ways this could happen are also possible because *some* of the means may yet be equal. In our specific case we could have (logically):

- $H_a : (\mu_1 \neq \mu_2$ and $\mu_1 \neq \mu_3$ and $\mu_2 \neq \mu_3)$ or $\mu_1 = \mu_2 \neq \mu_3$ or $\mu_1 = \mu_3 \neq \mu_2$ or $\mu_2 = \mu_3 \neq \mu_1$

Even in our simple example with only $k = 3$ treatments the null hypothesis could fail in many ways! As such we will write our alternative hypothesis in words as:

- **$\mathbf{H_a}$: Not all treatment means ($\boldsymbol{\mu_i}$) are equal.**

How do we test the hypothesis? If we calculate the sample means for each treatment we get $\bar{x}_1 = 9.0$, $\bar{x}_2 = 15.0$, and $\bar{x}_3 = 9.25$ hour (unrounded) respectively. This suggests the second treatment population mean might differ from the others. However we have done enough statistics to know that, due to error in estimation, a difference in sample statistics does not mean the underlying population parameters differ. A more rigorous approach then might be to look for any significant difference between each possible pair of treatments using our two-mean test for small independent samples (Section 6.1.4). For three treatments this would require three tests ($H_0 : \mu_1 = \mu_2$, $H_0 : \mu_1 = \mu_3$, and $H_0 : \mu_2 = \mu_3$). This would be a bit of work for our example but in general it will be even worse since there are ${}_kC_2$ such pairs. (For $k = 5$ treatments there are 10 pairs!) A deeper reason not to pursue this course however is the probability of making a Type I error for the overall procedure. If we assign a level of significance of $\alpha = 0.05$ to each pair test, the probability of not making a Type I error on each test is $1 - \alpha = 0.95$. The probability, however, of not making such an error on all three tests, assuming independence, is only $(0.95)(0.95)(0.95) = 0.8574$, so α for our overall hypothesis test using this pairwise procedure is not 0.05 but much larger. (The situation only gets worse for larger k.)

Motivation for ANOVA

Obviously what is desired is, just as before, the calculation of a single test statistic which can be compared to some critical value to determine the test. The following discussion is an attempt to motivate what will ultimately become our **(one-way) Analysis Of Variance (ANOVA)** technique. (If the following discussion seems too theoretical, do not worry as everything will be boiled down to a relatively straightforward procedural method.)

As the ANOVA name suggests, the idea is not to focus on the means of our treatment samples but rather their variance. Recall we know that all the treatments share a common variance of σ^2 in our experiment. Our approach will be to consider estimators for σ^2 and determine their behaviour when H_0 is assumed (all means equal) and when H_0 occurs (at least one mean differs). If H_0 is true (our assumption in a hypothesis test) then not only does each treatment population have a common standard deviation but also a common mean μ. In that case all of the data (irrespective of treatment) is being drawn from a normal population with mean μ and variance σ^2. If this is the case we have the usual unbiased estimator $s^2 = \sum(x - \bar{x})^2/(n-1)$ for the variance from Unit 1:

$$\text{Total } MS = \frac{\text{Total } SS}{df_{TOT}} = \frac{\sum\limits_{\text{All Treatments}} (x - \bar{x})^2}{n-1} \ .$$

Here we have defined the numerator of the expression to be the **total sum of squares**, Total SS, and the denominator to be the **total degrees of freedom**, df_{TOT}. Note the difference in the numerator (Total SS) is with respect to the **grand mean** (the mean of the data irrespective of treatment):

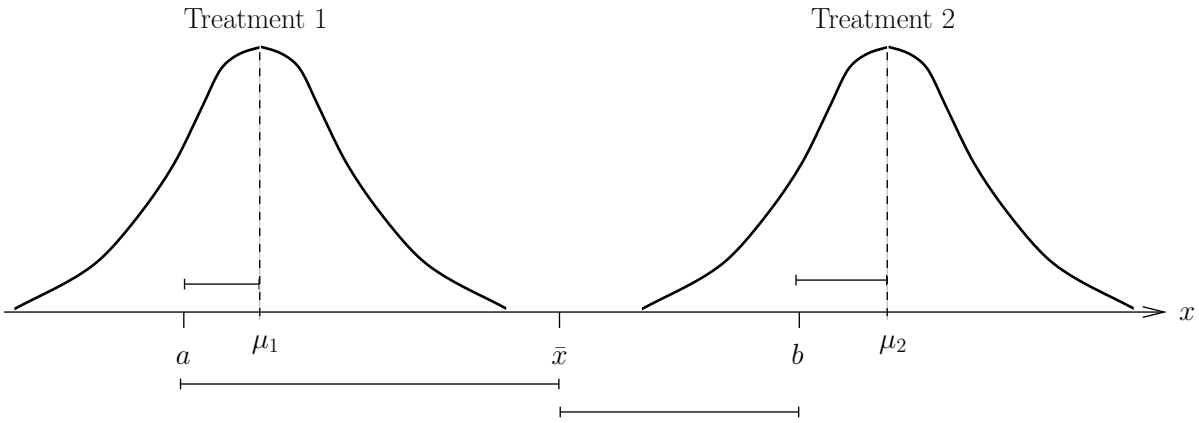
$$\bar{x} = \frac{\sum_{\text{All Treatments}} x}{n} \ .$$

Here n in the total degrees of freedom formula ($df_{TOT} = n - 1$) is the total number of observations (irrespective of treatment):

$$n = n_1 + n_2 + \ldots + n_k = \sum n_i \ .$$

The overall expression labelled Total MS above will be called the **total mean square**, rather than s^2 because, as we will now show, Total MS is only an estimator of the common σ^2 if H_0 is true.

If at least one of the means differs from another (H_a true) then the Total MS will not approximate the common σ^2, rather the Total MS will be typically too large. To see this imagine having just two normal treatment populations with common variance but different means ($\mu_1 \neq \mu_2$):

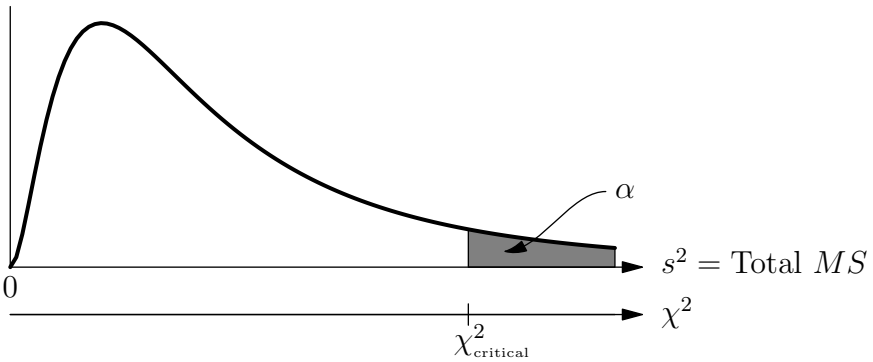


Assuming treatment sizes are roughly equal, the grand mean $\bar{x}$ will be near the balancing point of the two treatment distributions (i.e. the middling point shown). A typical data element from Treatment 1 might sit at the point labelled a. The distance from a to $\bar{x}$ which appears squared in the Total SS is shown below the x-axis. However the appropriate distance to estimate σ^2 in this case is not this distance but rather the much shorter distance shown above the x-axis from a to μ_1. Similarly an element at b from Treatment 2 also contributes an overestimated square distance to the Total SS. Thus Total MS = Total SS/df_{TOT} is expected to overestimate σ^2 in the event any means differ. (If the two means had been the same here then the two probability density curves would exactly overlap and the grand mean $\bar{x}$ would lie close to $\mu_1 = \mu_2$.) The observation that the Total SS gives an unbiased estimator of the common treatment variance σ^2 when H_0 is true but differs (is too large) when some means differ gives us a method to test the hypothesis. Assuming, as usual, that H_0 is true, calculate the Total MS and see where on the distribution of s^2 it lies. If it is too extreme (large) we will reject H_0 and accept H_a, namely conclude that the means are not all equal.

We need to know how s^2 is distributed. The answer is that for random samples of size n taken from an approximate normal distribution the chi-square statistic χ^2, defined to be proportional to s^2 by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \ ,$$

follows a **chi-square distribution** with $\boldsymbol{df = n - 1}$ degrees of freedom. The distribution of s^2 and correspondingly the (dimensionless) χ^2 statistic looks as follows:



Notice that χ^2 can only take on positive values due to the variances always being positive. So now our procedure is clear. Find Total MS and get the calculated χ^2 statistic via

$$\chi^2 = \frac{(n-1)\cdot(\text{Total } MS)}{\sigma^2}$$

and compare it to χ^2_{critical} which we can look up on a table for our degrees of freedom df_{TOT} and given level of significance α.[8] If the calculated χ^2 exceeds χ^2_{critical} then we know our Total MS looks too extreme to have occurred by chance. Rather we reject H_0 and accept H_a, namely believing that the result is due to the treatment means not all being equal.

While this procedure seems theoretically workable it is flawed because of the presence of the common treatment *population* variance, σ^2, which, in practice we typically do not know![9] Logically our plan should be to replace σ^2 with yet another sample estimate s^2 of the variance drawn from the data. This, recall, is what was done for our z-score formula where unknown σ was replaced by sample s. Just like then we need to worry about small sample sizes. If the sample was too small our test statistic required not the normal, but a t distribution. Since we only have 21 measurements in all our data here we will also have to worry about the randomness introduced by approximating σ^2. It turns out that we must abandon the chi-square distribution in favour of the ***F* distribution**.

Given two independent random samples of size n_1 and n_2 from populations that have approximate normal distributions with common variance ($\sigma_1^2 = \sigma_2^2$) the test statistic:

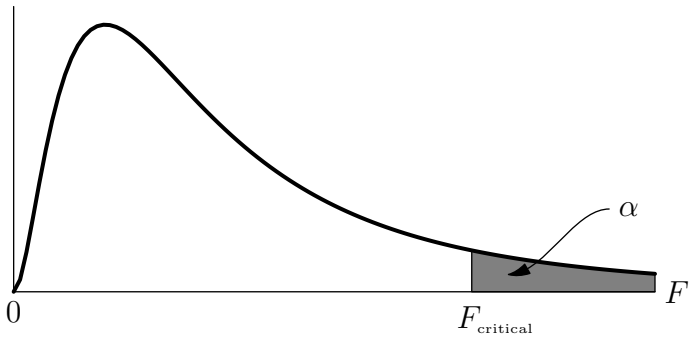
$$F = \frac{s_1^2}{s_2^2} \ ,$$

which is the ratio of the two sample variances, has an ***F* distribution** with $\boldsymbol{df_1 = n_1 - 1}$ and $\boldsymbol{df_2 = n_2 - 1}$ degrees of freedom. The former are the degrees of freedom of the numerator and the latter the denominator. Thus, just like our t distribution, there are many F curves, except now there is a separate one for each possible (df_1, df_2) pair. Aside from the multiplicative constant $n-1$ in the definition, the chi-square distribution is essentially the limiting case of the F distribution when $df_2 = \infty$.[10] It should be no surprise that the latter has the same appearance:

[8] These notes do not actually provide a χ^2 table. As will be seen in a moment the χ^2 distribution will not be adequate to solve our problem. The χ^2 critical values are the F distribution critical values multiplied by $df = n-1$ using $df_1 = df$ and $df_2 = \infty$. The need to multiply by df arises due to the presence of $n-1$ in the definition of the χ^2 statistic that is absent in the definition of F. Note that actual χ^2 tables will also have values (unlike our F table) indicating critical values for left tail areas because, as the distribution diagram indicates, the χ^2 distribution is not symmetric. The F distribution is not symmetric either, but such left tail critical values are not required for our purposes.

[9] As an aside, since the data for this example was randomly generated by the author, we *do know* the population variance used, it was $\sigma^2 = 9.0$ (hour^2). For this example we can actually perform the test! We will show that Total $SS = 375.809\ldots$ so our total mean square is Total MS = Total $SS/df_{TOT} = 375.809\ldots/(21-1) = 18.7904\ldots$. The calculated chi-square statistic is $\chi^2 = (21-1)(18.7904\ldots)/9.0 = 41.756$ which is greater than the tabulated critical value for $df_{TOT} = 20$ and $\alpha = 0.05$ of $\chi^2_{\text{critical}} = 31.410$. Thus we would reject H_0 and accept H_a. At $\alpha = 0.05$ the treatment means appear to be not all equal. Note that our χ^2_{critical} can actually be found on our F tables using the given $\alpha = 0.05$, $df_1 = df_{TOT} = 20$, and $df_2 = \infty$ and multiplying the value (1.571) by $df_{TOT} = 20$. Shortly we will introduce MST, a second estimator of σ^2, which could also be used in the chi-square test. It is likely more sensitive to variation in the means than Total MS due to the former not including the statistical estimation errors of the latter.

[10] Just like the normal (z) distribution was the limit of the t distribution when $df = \infty$.



Just as with the t distribution we only tabulate critical F-values but now for given α, df_1, and df_2. The tables begin on page 285. Note that df_1 (along the top) determines the column, df_2 (along the side) gives the appropriate row, while α (along the bottom) determines the appropriate subcolumn.

Example:
Use the F tables to find the appropriate critical values for a right tail test with:
(Answers: 1.) 2.711 2.) 4.558 3.) 12.90 4.) 2.039)

1. $df_1 = 5$, $df_2 = 20$, $\alpha = 0.05$:
2. $df_1 = 20$, $df_2 = 5$, $\alpha = 0.05$:
3. $df_1 = 20$, $df_2 = 5$, $\alpha = 0.005$:
4. $df_1 = 15$, $df_2 = \infty$, $\alpha = 0.01$:

So returning to our ANOVA problem if we used $s_1^2 =$ Total MS we still need another estimator of σ^2 to be s_2^2 in the denominator of F. We can hardly use Total MS again as it would not be independent of the first (as required for use of the F-distribution). In fact the F statistic would always be 1! Furthermore for the denominator variance we would also like this estimate to be always a good estimator, even when some means differed (H_a true). If the denominator variance estimator also failed in the event of differing means then it might just fail in the same way as the numerator and we would not detect an extreme F with our test. Our discussion above regarding why Total MS fails points us in the right direction for an estimator of σ^2 that is always valid, namely use the distances between the treatment values and their respective means. This approach is essentially a generalization of s_{Pool}^2 from Section 6.1.4 from two sample means to k treatments. Thus our pooled variance there,

$$s_{\text{Pool}}^2 = \frac{\sum_{\text{Sample 1}} (x - \bar{x}_1)^2 + \sum_{\text{Sample 2}} (x - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} ,$$

becomes[11]

$$\boxed{MSE = \frac{SSE}{df_E}} = \frac{\sum_{\text{Treatment 1}} (x - \bar{x}_1)^2 + \sum_{\text{Treatment 2}} (x - \bar{x}_2)^2 + \ldots + \sum_{\text{Treatment } k} (x - \bar{x}_k)^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1)}$$

$$= \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2 + \ldots + (n_k - 1) \cdot s_k^2}{n - k} ,$$

[11] Note that in this equation s_1^2, s_2^2, and s_k^2 refer to the variances of the individual treatment data and should not be confused with the unfortunate appearance of the same expressions s_1^2, s_2^2 in our previous variance ratio for F.

where we simplified the denominator to $\boxed{df_{\scriptscriptstyle E} = n - k}$ using

$$(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = (n_1 + n_2 + \ldots + n_k) - \underbrace{(1 + 1 + \ldots + 1)}_{k \text{ times}} .$$

The expression labelled MSE, which is an unbiased estimator for σ^2 regardless of the validity of H_0, will be called the **error mean square**. It measures the pooled variation within the k samples. As expected, it is the ratio of the **error sum of squares**, SSE, and the **error degrees of freedom**, $df_{\scriptscriptstyle E}$. The term "error" in this context refers to the experimental error (natural chance fluctuations) that occur in the sampling process, to be distinguished from the variation that will arise if the treatment means actually differ.

So now can we use as our test statistic $F = (\text{Total } MS)/(MSE)$? Unfortunately not. While we have a ratio of two estimators for the same population variance they are not independent as required for using the F distribution. That this should be the case is not surprising considering they are being drawn from the same set of sample data. Furthermore our total mean square Total MS has already used up all the $df_{\scriptscriptstyle TOT} = k - 1$ available degrees of freedom suggesting that there is nothing from which to construct a further independent variance estimate.

A solution to our conundrum is the observation that the total sum of squares can be broken into two pieces:[12]

$$\text{Total } SS = SSE + SST .$$

Here Total SS and SSE are defined previously while the **treatment sum of squares**, SST, is defined by

$$SST = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \ldots + n_k(\bar{x}_k - \bar{x})^2 = \sum n_i(\bar{x}_i - \bar{x})^2 ,$$

where $\bar{x}$ is the grand mean from before and $\bar{x}_i$ is the mean of the i^{th} set of treatment data. Since the MSE is constructed from SSE this suggests that if we could construct an estimator of the variance from the treatment sum of squares SST it would be independent of the former. This is true and the correct answer is:

$$\boxed{MST = \frac{SST}{df_{\scriptscriptstyle T}}} ,$$

where MST is called the **treatment mean square** and $\boxed{df_{\scriptscriptstyle T} = k - 1}$ is the **treatment degrees of freedom**. As its definition suggests, MST measures the variation among the k sample means. Our MST has the same behaviour as the Total MS. It only is an estimator of σ^2 when all the treatment means are equal (H_0 true). When they are not, just like Total MS, MST will tend to be too large.

Why is MST as defined a valid estimator of σ^2 when all the means are equal? To see this, suppose, for example that all our treatment sizes were equal, $n_i = \text{constant}$, for $1 \le i \le k$. (So in our example rather than the three treatment sizes being 6, 7, and 8 respectively suppose they all had $n_i = 7$ patients.) In this case each treatment is being drawn from a normal distribution with the same variance (equivalently standard deviation) and, by assumption, the same mean. In other words they are just k samples of the same size n_i being effectively drawn from the same normal population. Now recall by the C.L.T. that the standard error of the mean is, in this case,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n_i}} ,$$

[12] This can be proven by first writing Total SS as sums over the individual treatments, $\sum_{\text{Tr.}1}(x - \bar{x})^2 + \sum_{\text{Tr.}2}(x - \bar{x})^2 + \ldots + \sum_{\text{Tr.}k}(x - \bar{x})^2$. The expression for the first treatment sum (and similarly for the others) can then be written $(x - \bar{x})^2 = [(x - \bar{x}_1) + (\bar{x}_1 - \bar{x})]^2 = (x - \bar{x}_1)^2 + 2(x - \bar{x}_1)(\bar{x}_1 - \bar{x}) + (\bar{x}_1 - \bar{x})^2$. Upon summation over the elements of treatment 1, the first term gives the contribution for SSE, while the last term, being constant with respect to the elements in treatment 1, becomes $n_1(\bar{x}_1 - \bar{x})$ which is the contribution to SST. The middle term vanishes under summation because $\sum_{\text{Tr.}1} 2(x - \bar{x}_1)(\bar{x}_1 - \bar{x}) = 2(\bar{x}_1 - \bar{x}) \sum_{\text{Tr.}1}(x - \bar{x}_1)$ and the final sum vanishes due to the definition of the mean. (i.e. $\sum_{\text{Tr.}1}(x - \bar{x}_1) = \sum_{\text{Tr.}1} x - \sum_{\text{Tr.}1} \bar{x}_1 = n_1\bar{x}_1 - \bar{x}_1 \sum_{\text{Tr.}1}(1) = n_1\bar{x}_1 - \bar{x}_1 n_1 = 0$).

so squaring and solving for σ^2 gives

$$\sigma^2 = n_i \sigma_{\bar{x}}^2 \ .$$

But an estimator of $\sigma_{\bar{x}}^2$ is $s_{\bar{x}}^2$ which is just the (sample) variance of the k sample means in question, given by

$$s_{\bar{x}}^2 = \frac{\sum(\bar{x}_i - \bar{x})^2}{k-1} \ .$$

Inserting this point estimator for $\sigma_{\bar{x}}^2$ into the above formula for σ^2 gives the following point estimator for σ^2:

$$s^2 = n_i \frac{\sum(\bar{x}_i - \bar{x})^2}{k-1} = \frac{n_i \sum(\bar{x}_i - \bar{x})^2}{k-1} = \frac{\sum n_i(\bar{x}_i - \bar{x})^2}{k-1} \ .$$

This is precisely our formula for MST above with SST identified with the numerator and df_T the denominator! That the estimator should still work when the treatment sizes (n_i) are not equal is plausible. Furthermore if we consider the sum of the error and treatment degrees of freedom we have

$$df_E + df_T = (n-k) + (k-1) = k - 1 = df_{TOT} \ .$$

In other words, together our two variance estimators SSE and SST do not exceed the total number of degrees of freedom available which would have been symptomatic of their dependence.

We have finally arrived! The one-way ANOVA calculated test statistic will now involve not Total MS but rather MST in the numerator:

$$\boxed{F = \frac{MST}{MSE}} \ .$$

When all treatment means are equal (H_0 true) both MST and MSE are valid estimators of σ^2 and F should have a value around 1, the error due only to the random variation involved in sampling. If, however, H_0 is false then MST is no longer a valid estimator of σ^2 but will tend to be too large, while MSE still estimates σ^2. The result in that case will be that F will be too large. Thus comparing our calculated F above with the known critical value of the F distribution allows us to evaluate the hypothesis test for given α. Note that as MST is in the numerator and MSE is in the denominator the appropriate degrees of freedom to use for the F table are $\boxed{df_1 = df_T}$ and $\boxed{df_2 = df_E}$.

Computational Formulae

The sums of squares can be simplified through algebraic expansion similar to what was done for variance and standard deviation in Unit 1. In particular the total sum of squares has the computational formula

$$\boxed{\text{Total } SS = \sum x^2 - CM} \ ,$$

where $\sum x^2$ is just a sum of all the squares of the data values (irrespective of treatment) and the **correction of the mean** is defined by[13]

$$\boxed{CM = \frac{\left(\sum x\right)^2}{n}} \ .$$

Defining T_i as the sum of x for the i^{th} treatment allows us to write this also as

$$\boxed{CM = \frac{\left(\sum T_i\right)^2}{n}} \ .$$

[13]The student should be careful here with terminology. The total sum of squares is **not** equal to $\sum_{\text{All Treatments}} x^2$. As just mentioned, Total $SS = \sum x^2 - (\sum x)^2/n$.

Remember that $n = \sum n_i$ is the total number of observations, n_i being the size of the i^{th} treatment.

A computational formula for SST also exists:

$$\boxed{SST = \sum \frac{T_i^2}{n_i} - CM}\,.$$

Note only the first term is being summed, the correction of the mean being subtracted after summing. Finally to find SSE we manipulate the earlier additive relationship between the three sums of squares to get:

$$\boxed{SSE = \text{Total } SS - SST}\,.$$

The ANOVA Table

The relevant interim calculations required to determine our precious F statistic are conveniently summarized on an **ANOVA Table** whose entries are as follows:

Source	*df*	Sum of Squares	Mean Squares	F statistic
Treatments	df_T	SST	MST	F
Error	df_E	SSE	MSE	
Total	df_{TOT}	Total SS		

ANOVA tables for a particular problem can be generated by statistical software. For our purposes the table is convenient for organizing the calculation of the F statistic. One typically has to complete the calculations from left to right on the table. The formula sheet for this unit organizes all the previous formulae directly within an ANOVA table for convenience:

Source	*df*	Sum of Squares	Mean Squares	F statistic
Treatments	$df_T = k - 1$	$SST = \sum \frac{T_i^2}{n_i} - CM$	$MST = \frac{SST}{df_T}$	$F = \frac{MST}{MSE}$
Error	$df_E = n - k$	$SSE = \text{Total } SS - SST$	$MSE = \frac{SSE}{df_E}$	
Total	$df_{TOT} = n - 1$	Total $SS = \sum x^2 - CM$		$\left(F_{\text{crit}} : \begin{array}{l} df_1 = df_T \\ df_2 = df_E \end{array}\right)$

Here $CM = \dfrac{\left(\sum x\right)^2}{n} = \dfrac{\left(\sum T_i\right)^2}{n}$ where T_i is the sum of x for the i^{th} treatment of size n_i.

Having completed our theoretical discussion for testing multiple means, let us now do a complete hypothesis test of the original example presented in this section.

Example:
A hospital is interested in quantifying the cost effectiveness of three pain relievers. Patients recovering from the same type of surgery were given the ability (by pressing a button) to self-administer pain medication intravenously after surgery. They were randomly selected to receive one of the three types of pain relief and the hospital kept track of how long it took for a fixed value of drug to be consumed. The results were as follows (in hours):

Pain Reliever 1	Pain Reliever 2	Pain Reliever 3
15	20	9
8	13	13
8	16	7
7	14	9
13	19	7
3	9	10
	14	12
		7

The treatment populations are known to be normally distributed with a common (but unknown) standard deviation σ. For the same dollar value of each drug does the data suggest that the mean pain relief provided by each drug is the same (or do some of them differ) at a level of significance of 5%?

Solution:

Step 1) Formulate the null and alternative hypotheses

- $\mathbf{H_0 : \mu_1 = \mu_2 = \mu_3}$
- $\mathbf{H_a}$ **: Not all treatment means ($\boldsymbol{\mu_i}$) are equal.** $\Leftarrow$ Recall this can occur in several ways.

Step 2) State the level of significance.

$\boldsymbol{\alpha = 0.05}$ $\Leftarrow$ As given in the problem.

Step 3) Determine the test statistic.

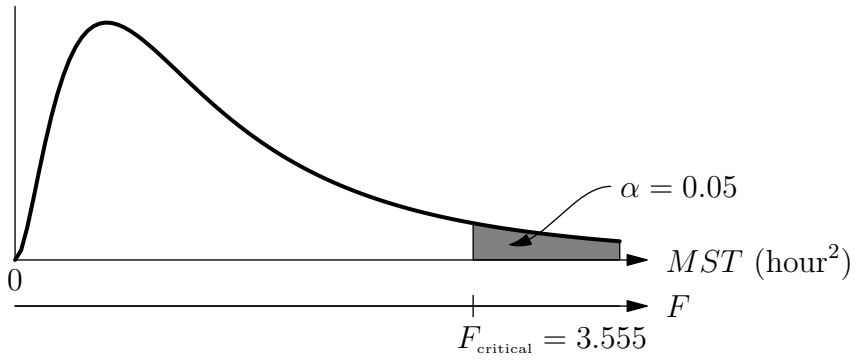
We will be drawing from several (more than 2) small independent normally distributed populations with common standard deviation using ANOVA to compare their means so an ***F* test** applies.

Step 4) Establish a decision rule.
(draw sampling distribution, find critical value)

The total number of measurements is $n = \sum n_i = 6 + 7 + 8 = 21$ and the number of treatments is $k = 3$. Our particular F curve is determined by the degrees of freedom:

- $df_1 = df_T = k - 1 = 3 - 1 = 2$
- $df_2 = df_E = n - k = 21 - 3 = 18$

(These values also will be needed in our ANOVA table.)



Here since $\alpha = 0.05$ is the tail area on the right side of the F distribution curve with $df_1 = 2$ and $df_2 = 18$ we find $F_{\text{critical}} = 3.555$ from the F tables.

Step 5) Evaluate the evidence.
(complete an ANOVA table to find the calculated value)

The goal is to find the calculated F-value which is the final column on our ANOVA table. We can systematically complete the calculation by filling in our ANOVA table **from left to right** using the formulae for each of its elements laid out in the formula sheet.

For the *degrees of freedom* (df) column we have already found $df_T = 2$ and $df_E = 18$. Then the total degrees of freedom is just the sum $df_{TOT} = 2 + 18 = 20$, which, as a check, should also equal $n - 1 = 21 - 1 = 20$.

In the *Sum of Squares* column there are three terms to evaluate (see formula sheet), but the second (SSE) depends on the other two so we do it last. The first (SST) and the third (Total SS) require the sum of each treatment T_i, the size of each treatment, n_i, the sum of the squares of all the elements $\sum x^2$, and the correction for the mean, CM, which also depends on the T_i. Proceeding tabularly to evaluate first the sums we get:

Pain Reliever 1		Pain Reliever 2		Pain Reliever 3	
x	x^2	x	x^2	x	x^2
15	225	20	400	9	81
8	64	13	169	13	169
8	64	16	256	7	49
7	49	14	196	9	81
13	169	19	361	7	49
3	9	9	81	10	100
		14	196	12	144
				7	49
$T_1 = 54$	580	$T_2 = 105$	1659	$T_3 = 74$	722

The sizes of the treatments are $n_1 = 6$, $n_2 = 7$, $n_3 = 8$ and their total is $n = 21$. We now find the *correction for the mean*:

$$CM = \frac{\left(\sum T_i\right)^2}{n} = \frac{(54 + 105 + 74)^2}{21} = \frac{233^2}{21} = \frac{54289}{21} = 2585.19\ldots\,(\text{hour}^2)\,.$$

Since $\sum \frac{T_i^2}{n_i} = \frac{54^2}{6} + \frac{105^2}{7} + \frac{74^2}{8} = 2745.5$ we get the *treatment sum of squares*:

$$SST = \sum \frac{T_i^2}{n_i} - CM = 2745.5 - 2585.19\ldots = 160.309\ldots\,(\text{hour}^2)$$

Adding up the x^2 columns gives $\sum x^2 = 580 + 1659 + 722 = 2961$ (hour^2) so the *total sum of squares* is

$$\text{Total } SS = \sum x^2 - CM = 2961 - 2585.19\ldots = 375.809\ldots\,(\text{hour}^2)$$

Subtract these last two to get the *error sum of squares*:

$$SSE = \text{Total } SS - SST = 375.809\ldots - 160.309\ldots = 215.5\ (\text{hour}^2)$$

The *Mean Squares* column (see formula sheet) is just the Sum of Squares (the previous column) divided by the degrees of freedom (the column before that):

$$\begin{aligned} MST &= \frac{SST}{df_T} = \frac{160.309\ldots}{2} = 80.15476\ldots\,(\text{hour}^2) \\ MSE &= \frac{SSE}{df_E} = \frac{215.5}{18} = 11.97222\ldots\,(\text{hour}^2) \end{aligned}$$

Finally our calculated F *statistic*, the last column, is just the fraction of the two entries in the previous column:

$$F = \frac{MST}{MSE} = \frac{80.15476\ldots(\text{hour}^2)}{11.97222\ldots(\text{hour}^2)} = 6.695061\ldots$$

Placing all these results (rounded off) in an ANOVA table gives:

Source	*df*	Sum of Squares (h^2)	Mean Squares (h^2)	F
Treatments	2	160.31	80.1548	6.69506
Error	18	215.50	11.9722	
Total	20	375.81		

Step 6) State your decision.

Since our calculate F-value (6.695) lies in the region of rejection ($F > 3.555$) we reject H_0 and accept H_a at $\alpha = 0.05$. Evidence supports that not all the means are equal. (At least one mean differs from the others.)

Notes:

- Using computer software the P-value for the test which corresponds to the calculated $F = 6.695$ is 0.0067. Since this is less than our $\alpha = 0.05$ we see this affirms our decision to reject H_0 and accept H_a using the P-value approach. Had we tested at $\alpha = 0.005$ however we would have failed to reject H_0 since this is less than 0.0067. You can confirm this result by looking up F_{critical} for $\alpha = 0.005$ to find 7.215. Since our calculated F (6.695) is less than this we would have failed to reject H_0 at $\alpha = 0.005$ as well using the critical value approach.

- In the above example the standard deviations of the various treatment samples are $s_1 = 4.3$ hour, $s_2 = 3.7$ hour, and $s_3 = 2.3$ hour. Given our small sample sizes it is plausible that the actual population standard deviations they estimate are equal ($\sigma_1 = \sigma_2 = \sigma_3 = \sigma$) as required for the test. Notice that $\sqrt{MSE} = 3.5$ hour is approximately of the same size as these values which it should be since, as discussed, the MSE is a valid estimator for σ^2 even if the treatment means are different. Moreover 3.5 hour is expected to be a better estimate than the individual treatment standard deviations as it is calculated over all treatments. On the other hand, $\sqrt{MST} = 9.0$ hour is not close to these values as MST fails to estimate σ^2 in the case of different means.

- If H_0 is not rejected then the evidence does not support that any of the treatment means differ. However if we reject H_0 and accept H_a like we did in the example we now know at least one mean differs from another. We obviously would like to determine which ones differ. Problem 4 in the assignment considers this for the above example.

- If the treatment sample sizes are equal and the treatment population distributions have a mound-shape (i.e. approximately normal) the one-way ANOVA test is fairly robust. It should be no surprise, however, that if the common variance assumption does not hold the consequences are more serious as this is precisely the statistic being analyzed to reach our conclusion.

- In our plot of the sampling distribution curve in the last example we have labelled the x-axis as MST. Note that this placing of a double-axis here on the same curve is technically inaccurate as a single estimator of σ^2 like MST would have a chi-square distribution not an F distribution as already mentioned. The point of the diagram is simply to show the correspondence that exists between measurements of MST and the actual statistic (F) required, and that both curves have roughly the same shape. The numerator variance (MST) actually determines the test since it is

only a valid estimator when H_0 is true. The denominator variance (MSE), being a valid estimator all the time and reflecting only the statistical error of estimation, is therefore suppressed. The same graphing inaccuracy was introduced when discussing the t distribution where we likewise suppressed the sample standard deviation. Technical accuracy is being sacrificed for (hopefully) conceptual clarity!

- It should be noted that the χ^2 and F distributions discussed in our context of one-way ANOVA analysis have broader application to statistical inference regarding population variance. The chi-square distribution can be used to find confidence intervals and do hypothesis tests for a single population variance. This would be useful in the calibration of machines like our drink dispenser in Problem 4 on page 210. There we tested whether the average amount of drink being dispensed from the machine was 300 ml. However even if this were true surely the machine would also need adjusting if the variance were too large. If the standard deviation were say 50 ml then cups could be being receiving as low as 150 ml or as high as 450 ml! A test of the variance of the data would allow us to determine whether the amounts were widely varying (and the machine likely required adjustment) or if the variance was statistically equal to some specified variance. A confidence interval for the variance could similarly be found using the chi-square distribution. If we were choosing between buying two automatic drink dispensers we may want to compare two population variances for which the F distribution would be useful. It can also be used to test whether the standard deviations needed to be equal for our hypothesis test of Section 6.1.4 really are equal given the variation found in the sample statistics.

Assignment:

1. A botanist wishes to test the effect of several chemicals (A, B, C, D) on the number of tomatoes produced by Roma tomato plants. He treats four series of plants each with one of the chemicals. Of the plants that survived to the stage of producing tomatoes, the following number of tomatoes were recorded on each plant:

Chemical A	Chemical B	Chemical C	Chemical D
22	26	15	17
19	16	19	25
30	18	19	27
19	21	22	14
20	19	13	

   (a) Can the botanist conclude that the differently treated plants all produce on average the same number of tomatoes at a level of significance of 0.025?

   (b) What assumptions must be made for the analysis to be valid and do they seem reasonable given the data?

2. A movie theatre is interested in knowing if there is a difference in the amount of money spent by a moviegoer at the concession stand based on the time of day the movie is run (afternoon matinee, early evening, or late show). The theatre randomly selects moviegoers who make purchases at the concession stand at these times and finds they spend the following amounts:

Matinee ($)	Evening ($)	Late Show ($)
14	7	13
12	6	11
5	6	11
10	5	7
12	7	10
8	10	14
	3	

   Can we conclude there is any difference between the average amount spent at the concession stand at these different movie times? Test at the 5% level of significance. (Assume the treatment populations are normal with common standard deviation.)

3. An archaeologist measures the lengths of a random selection of arrowheads found at three different sites in southern Alberta with the following results (cm):

Site 1	Site 2	Site 3
5.4	5.3	8.3
5.8	6.1	7.7
4.6	7.1	8.1
6.0	6.9	8.1
5.1	5.7	6.9

   Assuming the populations of arrowheads have lengths that are normally distributed with a common standard deviation, can the archaeologist conclude that mean arrowhead lengths at the sites at all differ at a level of significance of $\alpha = 0.01$?

4. If an ANOVA test ends up rejecting H_0 and accepting H_a then at least one of the treatment means must differ from another, but which ones are different? One way of evaluating this is to construct confidence intervals for means and for their differences. The procedure is largely the same as we have seen before involving small means except that now our point estimator for the standard deviation is replaced by $\sqrt{MSE}$. (Recall that MSE is a valid estimator for the common variance σ^2 of the treatments even if the means are not equal.) Since it is the error mean square we are using it should be no surprise that the other change required is to use $df_E = n - k$ for the number of degrees of freedom when finding t. The following two questions illustrate this in detail.

   (a) One constructs a confidence interval for each mean (μ_i) exactly likely we did in Section 5.1.4 for a single mean except now the standard error becomes

   $$s_{\bar{x}_i} = \sqrt{\frac{MSE}{n_i}}$$

   and we use df_E for the number of degrees of freedom when finding t. Find a 95% confidence interval for the mean time of effectiveness for each of the three pain relievers for the example in the main text. Draw them on a number line and identify which means differ by seeing which do not overlap.

   (b) One may construct a confidence interval for the difference of a pair of means ($\mu_i - \mu_j$) as done in Section 6.1.4 but once again we use the MSE as our point estimator for variance to get a standard error of:

   $$s_{\bar{x}_i - \bar{x}_j} = \sqrt{MSE \cdot \left[\frac{1}{n_i} + \frac{1}{n_j}\right]}\ .$$

   Also therefore use df_E when finding t. Find a 95% confidence interval for the difference in each of the three pairs of sample means for the pain relief example done in this section.

Exam Strategy for Identifying Problems:

The challenge on exams when doing inference problems is to pick out what sort of problem you are dealing with. Ask yourself the following questions when approaching a problem:

1. Am I dealing with one population, two populations, or more?
2. Am I dealing with means ($\bar{x}$ and μ) or proportions (p and π)?
3. Am I doing a confidence interval or a hypothesis test?
4. Are the sample(s) involved large or small?
5. If more than one sample is involved, are they independent or dependent?

Try categorizing sample exam problems using this checklist. Also go through the formula sheet categorizing their application with this checklist.

Multiple Population Inferences Formulae

Standard Error Formulae for Differences of Means and Proportions

$$^*s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \qquad s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

$$^*s_{p_1-p_2} = \sqrt{\frac{p_1\cdot(1-p_1)}{n_1} + \frac{p_2\cdot(1-p_2)}{n_2}} \qquad ^*s_{p_1-p_2} = \sqrt{p_{\text{Pool}}\cdot(1-p_{\text{Pool}})\cdot\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

Maximum Error of Estimate for Differences of Means and Proportions

$$^*E = t\cdot s_{\bar{d}} \qquad E = z\cdot s_{\bar{x}_1-\bar{x}_2} \qquad E = t\cdot s_{\bar{x}_1-\bar{x}_2} \qquad ^*E = z\cdot s_{p_1-p_2}$$

Confidence Intervals for Differences of Means and Proportions

$$^*P\left(\bar{d}-E < \mu_1-\mu_2 < \bar{d}+E\right) = 1-\alpha$$

$$P(\,[\bar{x}_1-\bar{x}_2]-E < \mu_1-\mu_2 < [\bar{x}_1-\bar{x}_2]+E\,) = 1-\alpha$$

$$^*P(\,[p_1-p_2]-E < \pi_1-\pi_2 < [p_1-p_2]+E\,) = 1-\alpha$$

Standard Scores For Differences Between Means And Proportions[†]

$$^*t = \frac{\bar{d}-0}{s_{\bar{d}}} \qquad z = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}} \qquad t = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}} \qquad ^*z = \frac{(p_1-p_2)-0}{s_{p_1-p_2}}$$

[†]Here we are assuming hypothesis tests of equality so $\mu_1-\mu_2=0$ and $\pi_1-\pi_2=0$ in these formulae already.

Other Formulae

$$^*d = x_1-x_2 \qquad df = n_1+n_2-2 \qquad ^*p_{\text{Pool}} = \frac{x_1+x_2}{n_1+n_2}$$

ANOVA Table

Source	df	Sum of Squares	Mean Squares	F statistic
Treatments	$df_T = k-1$	$SST = \sum\frac{T_i^2}{n_i} - CM$	$MST = \frac{SST}{df_T}$	$F = \frac{MST}{MSE}$
Error	$df_E = n-k$	$SSE =$ Total $SS - SST$	$MSE = \frac{SSE}{df_E}$	
Total	$df_{TOT} = n-1$	Total $SS = \sum x^2 - CM$		$\left(F_{\text{crit}}: \begin{array}{l} df_1 = df_T \\ df_2 = df_E \end{array}\right)$

Here $CM = \frac{\left(\sum x\right)^2}{n} = \frac{\left(\sum T_i\right)^2}{n}$ where T_i is the sum of x for the i^{th} treatment with size n_i.

Unit 7: Bivariate Data Analysis

7.1 Bivariate Data and Linear Regression

Bivariate data arises when two variables are measured on the same experimental unit. In **bivariate data analysis** we seek to determine if there are any trends or relationships between the variables.

Example:
Consider the following questions:

1. Is there any relationship between the variation in time spent training salespeople and their sales performance?
2. Is there any relationship between the variation in money a government spends on the military (per capita) and the number of times the government goes to war over the same time period?
3. Is there any relationship between the variation in the amounts of money spent on advertising in an area and the sales volume in that area?
4. Is there any relationship between the variation in the frequency of numbers that are played in a lottery and the relative frequency of numbers that win in the lottery?
5. Is there any relationship between the period of oscillation of a Cepheid variable star and its intrinsic brightness (absolute magnitude)?

Each of the above examples contains two quantities which vary.[1] (For each example identify the two variables and the experimental unit upon which they are being measured.) The variability in these two quantities is pictured by way of a mathematical graph. In bivariate data analysis a theoretical mathematical curve is assumed to occur between variables. When a set of bivariate data is analyzed by the technique of **linear regression**, the assumption is made that there is a straight line which best describes the relationship between the data sets. We will restrict ourselves to this type of curve fitting.[2]

Example:
A municipal parks department is interested in knowing the relationship between the price of a lawnmower in their fleet of small lawnmowers and the number of breakdowns it has over a five year period. A random sample of 10 lawnmowers are studied. The observations are summarized as follows:

Price($)	600	300	200	500	100	700	100	400	100	200
# of Breakdowns	1	2	2	1	4	0	3	2	1	4

The attributes of each lawnmower are identified by two numbers which combine to make an **ordered pair**. In statistical analysis, as in other disciplines, the ordered pairs are plotted on a **coordinate plane**. The variable plotted on the horizontal axis is called the **independent** or **explanatory variable** while the variable that is plotted on the vertical axis is called the **dependent** or **response**

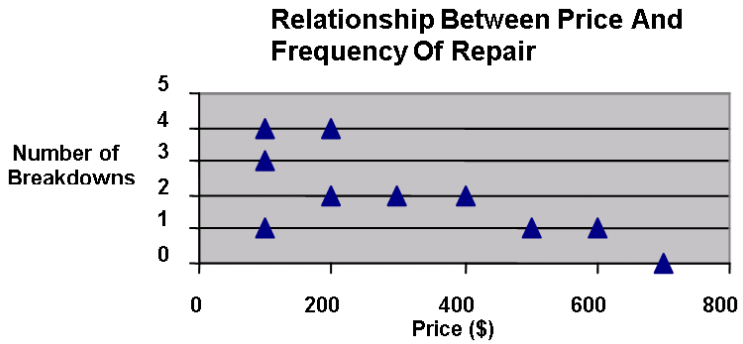
[1] The first question looks superficially like the two population examples of Unit 6. There however it was the same variable x =net sales being measured on two populations. In the current example there are two distinct variables x = training time and y = net sales measured on the same individuals, hence a new variable (y) must be introduced.

[2] Variables may have a strong correlation even if it is not linear. They may follow an exponential or some other curve which may become evident from a scatterplot (introduced below). A problem may sometimes be linearized by taking a function of one variable to produce a new variable. See for instance problem 3 on page 270 where the variable is the logarithm of time rather than time itself which is linearly related to brightness.

variable. For purposes of later calculations, the independent variable is always identified by the symbol x while the dependent variable is always identified by the symbol y. The ordered pairs formed by the bivariate data are plotted in the form of a **scatterplot** on the coordinate axes. It is important that the dependent variable be plotted on the vertical axis.

In regression analysis one inferential objective is to predict the value of y given a value of x. The dependent variable can usually be identified because it is the quantity that is to be predicted in the relationship.[3] The predicted value of y will be designated by y_p to distinguish from the y coordinate of a potential data point or points with the same x coordinate.

Example:
In the previous example the variable x is the price while y is the number of breakdowns. This is because the purchaser is in control of how much he wishes to spend on the lawnmower which is the nonrandom input, while the number of breakdowns is that which is uncertain or random, perhaps dependent upon the input price, and what we wish to predict. The ordered pair corresponding to the third lawnmower is $(x, y) = (\$200, 2 \text{ breakdowns})$. A scatterplot of the data is



The statistical population data in bivariate data analysis is the collection of all possible ordered pairs that could be observed. The characteristics of this population data are described by its **parameters**, just as before we found parameters for describing data belonging to a single variable x such as the mean, μ.

For data with a linear trend, such as the lawnmower data above, it is possible to characterize the population by a best fit straight line which runs through the data. Since the data consists of more than two points, any straight line can only approximate the points. Note, however, that our goal is not to reconstruct the random scatter of the data with our line, but rather to give an indication of what we expect for a given input value x.

A straight line can be described completely by two parameters; the **slope** and the **vertical axis intercept**.

The vertical axis intercept can be determined by inspection of the graph or by calculating the value of y when $x = 0$. The intercept parameter is designated by the symbol α (alpha) in statistics. Recall parameters are designated by Greek letters to show their theoretical nature.

The definition of the slope of a line is the ratio of the rise to the run between any two points along the line. The slope parameter is designated by the symbol β (beta) in statistics, and satisfies the relation:

$$\beta = \frac{\Delta y}{\Delta x}$$

[3]Consider the example at the top of page 258 and decide which of the two variables in each case should be considered the independent (x) and which the dependent (y) variable. In some cases, especially the second, it may be unclear which to make which, however one can still be interested in the relationship between the variables.

The equation, in terms of its parameters, which governs the relationship between variable x and variable y along a straight line always takes the form:

$$y_p = \alpha + \beta x$$

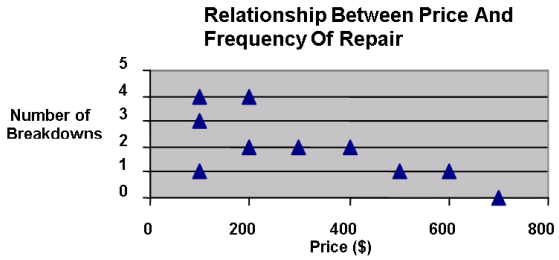
When analyzing single variables we saw how a sample from the population produced an estimate, such as $\bar{x}$ of the true, and usually unknown, population parameter μ. In the same way a sample data set, such as for our lawnmowers, will only produce an estimate of the best-fit line that would theoretically exist through the entire population of lawnmower data. As such a best fit line to the sample data will be written

$$\boxed{y_p = a + bx}$$

where a estimates α, and b estimates β.[4]

Example:
For the lawnmower scatterplot, use a ruler to draw a best fit straight line through the data. Approximate a by examining the intersection with the vertical axis and b by drawing a triangle of appropriate size and calculating its rise and the run. Fill in your equation for y_p and use it to predict the number of breakdowns that the municipality would expect if it pays \$400 for a lawnmower.

Solution:



- $a \approx$
- $b \approx \dfrac{\Delta y}{\Delta x} =$
- $y_p \approx \left(\quad\right) + \left(\quad\right) x$
- $y_p \approx \left(\quad\right) + \left(\quad\right) (\$400) =$

Note that your predicted value for the number of breakdowns, y_p, at $x = \$400$ likely is not the same as the data y-value, $y = 2$ breakdowns, found at that same x.

The Least-Squares Criteria

Cases where data points do not line up along a straight line are said to be random in nature. Through the random scatter in our lawnmower example we can discern a downward linear trend. The regression equation

$$\boxed{y_p = a + bx}$$

provides us with a model for the underlying straight-line trend through the random scatter. We have seen how we could by eye estimate a linear relationship for our data by minimizing how far off the data points were from our line. One quantitative way to do this is to consider the **residual** of each point, namely the vertical difference between the data point (x, y) and the predicted value at x, (x,y_p),

$$\text{residual} = y - y_p \ .$$

Statisticians draw the best fit straight line using what is called **the least-squares criteria**. The **least-squares line** or **regression line** is that straight line through the data that minimizes the sum

[4]The student should note that our statistical notation deviates from that commonly used for the slope-intercept form of the line, namely $y = mx + b$. Our b in statistics is the slope not the y-intercept of the line.

of the squared residuals.[5] This criteria will be met if we use the statistics a and b, estimates of α and β, where a and b are calculated as follows:

$$\boxed{a = \frac{\left(\sum x^2\right) \cdot \left(\sum y\right) - \left(\sum x\right) \cdot \left(\sum xy\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2}}$$

$$\boxed{b = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2}}$$

Here n is the number of pairs in our sample. These two calculations may be done tabularly.

Example:
For the lawnmower example find the linear regression line. Plot the line on the scatterplot. Use the regression line to find the predicted number of breakdowns for x=\$400.

Solution:
Noting the required summations of our intercept and slope formulae, the following table will provide us with the information to do the two calculations.

x(\$)	y(breakdowns)	xy(\$ · breakdowns)	x^2 (\$^{2})
600	1	600	360000
300	2	600	90000
200	2	400	40000
500	1	500	25000
100	4	400	10000
700	0	0	490000
100	3	300	10000
400	2	800	160000
100	1	100	10000
200	4	800	40000
$\sum x = 3200$	$\sum y = 20$	$\sum xy = 4500$	$\sum x^2 = 1460000$

Substitute the totals into the two statistics formulae to find estimates of the regression line parameters.

$$a = \frac{\left(\sum x^2\right) \cdot \left(\sum y\right) - \left(\sum x\right) \cdot \left(\sum xy\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(1460000) \cdot (20) - (3200) \cdot (4500)}{10 \cdot (1460000) - (3200)^2} = 3.3945 \text{ breakdowns}$$

$$b = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{10 \cdot (4500) - (3200) \cdot (20)}{10 \cdot (1460000) - (3200)^2} = -0.0044 \text{ breakdowns/\$}$$

The least-squares line is then:

$$y_p = 3.3945 - 0.0044x \quad \text{(breakdowns)}$$

To show the underlying linear trend, one can plot this equation on the scatterplot by identifying the y-intercept and calculating a second y value near the end of the range of x values:

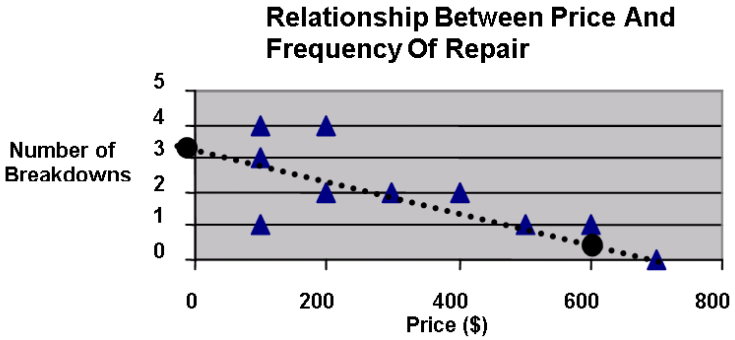
By examining the equation, the Y-intercept= $a = 3.3945$ breakdowns.

An x value at the far end of the range is x = \$600, where

$$y_p = 3.3945 - 0.0044(600) = 0.75 \text{ breakdowns}$$

Connecting these points on our plot gives the regression line.

[5]This is a mathematical criteria whose solution can be found using mathematics beyond the scope of this course. For our purposes it is enough to know that a unique solution to the problem always exists and is given by the formulae provided.



Notice that the points are scattered approximately evenly on either side of this line. If the predictability had been perfect the points would all lie exactly on the line.

We can predict the number of breakdowns at \$400 using our regression line to get:

$$y_p = 3.3945 - 0.0044(400) = 1.63 \text{ breakdowns} .$$

Note that this value, as well as a and b should compare favourably to your rough estimates done by eye.[6]

A property of the regression line is that $(\bar{x}, \bar{y})$ will always lie upon it. In other words the regression line predicts the mean y value perfectly on the basis of the mean x value. To see this, note that for the above example:

$$\bar{x} = \frac{\sum x}{n} = \frac{3200}{10} = \$320 , \text{ and}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{20}{10} = 2 \text{ breakdowns} ,$$

while placing $\bar{x} = \$320$ in the regression line formula gives:

$$y_p = 3.3945 - 0.0044(320) = 1.9865 = 2.0 \text{ breakdowns} .$$

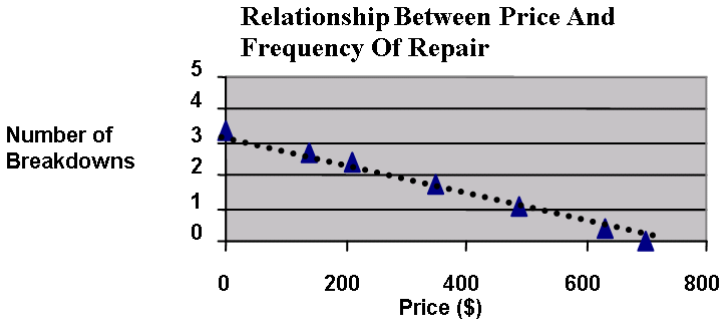
Because $(\bar{x}, \bar{y})$ always lies on the linear regression line, it is another second point, after the y-intercept, that may be used in plotting the best-fit line.

Be careful not to use your best fit line to make predictions too far outside of the range of the measured values of x since there is no guarantee that the linear relationship is still valid for those values. For instance if we used our regression line to predict the number of breakdowns for a lawnmower whose purchase price was \$1200 we would find $y_p = -1.89$ breakdowns, an absurd result. Clearly the actual curve must flatten out for larger purchase prices to keep the number of breakdowns positive.

[6] That the regression line and your fit by eye should be close may not always be the case. For instance a data point that is an outlier may significantly alter the position of the regression line due to the weight of its residual-squared in the sums. (That this may be significant follows from, among other things, the fact that $(\bar{x}, \bar{y})$ always lies on the regression line.) When fitting by eye you may instinctively ignore an outlier. It may be the case that an outlier should be re-examined and potentially removed before doing the least-squares fit.

7.2 Correlation Analysis

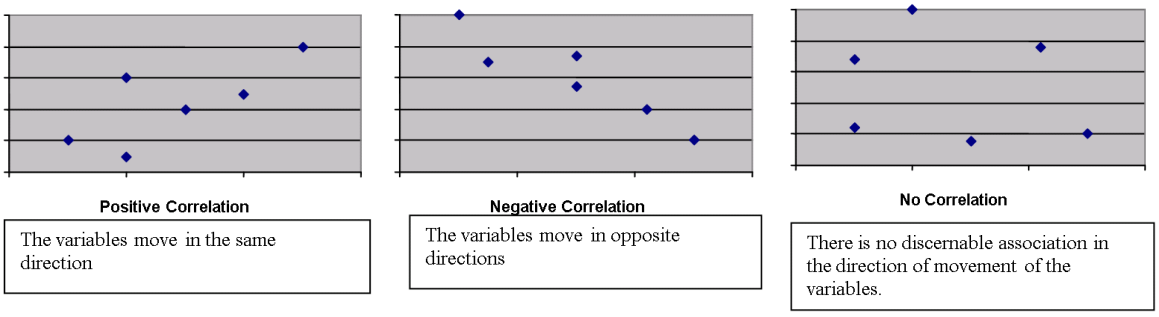
A situation in which there is perfect predictability is said to be a **deterministic model**. In a deterministic model all ordered pairs fall exactly along a straight line. For example, suppose our data for the lawnmowers had looked as follows:



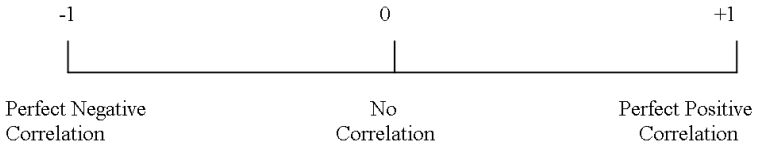
In this case after finding the regression line parameters, predictions about frequency of repair could be made perfectly with the regression line equation.

In this case we would say that all of the variability observed in number of breakdowns is related to the variability in price paid. Nothing is left to chance. In practice it would be very rare to see a perfect situation like this as was shown in our actual data.

Correlation analysis measures the strength of the relationship between variables. It is useful for assessing how much of the variation in the y values is related to variation in the x values. If a scatterplot is analyzed it is possible to qualitatively analyze strong and weak correlation as well as data trends.



The **correlation coefficient**[7] gives a quantifiable measure of the degree to which the points fit a straight line and the direction of movement of the variables. It is designated by the symbol $\boldsymbol{r}$. The correlation coefficient, r, is found in this range of values:



[7]Technically it is called the **Pearson product moment correlation coefficient**.

The **sign** of r indicates whether the data are positively or negatively correlated, while the **magnitude** of r indicates the extent of the correlation. If r is either +1 or -1, predictability is perfect and all ordered pairs will lie exactly on a straight line of either positive or negative slope respectively. If r is close to 0 there is no discernible association between the variables and the points will lie very far from a straight line. In general we will interpret the magnitude $|r|$ as follows:

- 0.00 to 0.40: insignificant (negligible) correlation
- 0.40 to 0.70: substantial correlation
- 0.70 to 0.80: high correlation
- 0.80 to 0.90: very high correlation
- 0.90 to 1.00: extremely high correlation

These guidelines are somewhat subjective and will vary depending on the field of study. In practice a value of $|r|$ that is very high (0.80 to 0.90) is useful for predicting the behaviour of a group of items or individuals, but a larger value is required to predict the behaviour of a single individual.

Look at the three data sets above and qualitatively estimate r on the given number line. The exact formula for the correlation coefficient is the following:

$$r = \frac{n \cdot (\sum xy) - (\sum x) \cdot (\sum y)}{\sqrt{n \cdot (\sum x^2) - (\sum x)^2} \cdot \sqrt{n \cdot (\sum y^2) - (\sum y)^2}}$$

Example:
Using the data for the lawnmower example calculate the correlation coefficient and interpret the result.

Solution:
Looking at the formula, there is one additional column required from our regression calculation, the y^2 column:

x(\$)	y(breakdowns)	xy(\$ · breakdowns)	x^2 (\$^{2})	y^2 (breakdowns^2)
600	1	600	360000	1
300	2	600	90000	4
200	2	400	40000	4
500	1	500	25000	1
100	4	400	10000	16
700	0	0	490000	0
100	3	300	10000	9
400	2	800	160000	4
100	1	100	10000	1
200	4	800	40000	16
$\sum x = 3200$	$\sum y = 20$	$\sum xy = 4500$	$\sum x^2 = 1460000$	$\sum y^2 = 56$

Substitute:

$$r = \frac{n \cdot (\sum xy) - (\sum x) \cdot (\sum y)}{\sqrt{n \cdot (\sum x^2) - (\sum x)^2} \cdot \sqrt{n \cdot (\sum y^2) - (\sum y)^2}} = \frac{10 \cdot (4500) - (3200) \cdot (20)}{\sqrt{10 \cdot (1460000) - (3200)^2} \cdot \sqrt{10 \cdot (56) - 20^2}} = -0.72$$

Interpretation:

- The negative sign indicates a negative correlation; as the price of a lawnmower increases the number of breakdowns tends to decrease.
- The magnitude 0.72 means that the correlation is high.

This should not be interpreted as cause and effect. A low price does not cause a high frequency of repair. The two variables simply move together. In other words **correlation does not imply causation**.

*7.3 Alternative Formulae

We have seen expressions for calculating a, b, and $r$which are not particularly intuitive. Properties of the regression line such as the fact $(\bar{x}, \bar{y})$ lies upon it and the fact that it is somehow optimized by measuring the dispersion of points around it, suggest that our regression and correlation formulae might have forms that reflect our means and standard deviations from descriptive statistics. This is indeed the case. The correlation coefficient may be written in terms of standard deviations via:

$$\boxed{r = \frac{s_{xy}}{s_x \cdot s_y}}$$

Here s_x is the standard deviation of all the x-values of the data points while s_y is the standard deviation of all the y-values of the data points.[8] These may be calculated using the usual formula from Unit 1 for s in the x case and the same formula, with y substituted for x to obtain s_y. The numerator, s_{xy}, is called the **covariance** and may be computed via:

$$\boxed{s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}} \text{ or } \boxed{s_{xy} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{n-1}}$$

The left formula is the definitional form for the covariance and clearly shows that the covariance is like the variance s^2 but with $(x-\bar{x})^2 = (x-\bar{x})(x-\bar{x})$ replaced by $(x-\bar{x})(y-\bar{y})$. The second formula is the more convenient computational formula for the covariance, found by manipulating sums in the usual way.

Once r is calculated using the new formula (or our original one), we can calculate the slope b by[9]

$$\boxed{b = r\left(\frac{s_y}{s_x}\right)}.$$

Finally the intercept can be calculated using these two values with

$$\boxed{a = \bar{y} - b\bar{x}}.$$

Here $\bar{x}$, and $\bar{y}$ are just the means of the x-values and y-values of the data points as usual.

Compared to our original formulae, we will see in the next example that this new approach might appear to be more work. However this method using pre-calculated values has the advantage that it produces useful numbers such as means and standard deviations that help in the visualization of the data set and may have already been calculated.

[8] That sample standard deviations should appear in the formula is expected since we are using sample data. Recall a primary motivation for linear regression analysis is to allow prediction; with bivariate *population* data there would be nothing left to predict.

[9] Note that since s_x and s_y are both positive this formula implies that slope b and r have the same sign, a property we have already seen.

Example:
Find the correlation coefficient and regression line for the lawnmower data using the new formulae.

Solution:
We have already calculated all necessary sums on the previous table (units can be found there) for our $n = 10$ data points:

$$\sum x = 3200 \quad \sum y = 20 \quad \sum xy = 4500 \quad \sum x^2 = 1,460,000 \quad \sum y^2 = 56$$

From these calculate the means, standard deviations, and covariance required. Use the computational formulae so we can use the calculated sums. Do not round these values as these are intermediate results.

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

$$\bar{y} = \frac{\sum y}{n} =$$

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

$$s_y = \sqrt{\frac{\sum y^2 - \frac{(\sum y)^2}{n}}{n-1}} =$$

$$s_{xy} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{n-1} =$$

Finally use these values to calculate r, b, and a, in that order:

$$r = \frac{s_{xy}}{s_x \cdot s_y} =$$

$$b = r\left(\frac{s_y}{s_x}\right) =$$

$$a = \bar{y} - b\bar{x} =$$

Finally the regression line is:

$$y_p = a + bx =$$

Calculator Keys and Software

Modern statistical calculator are capable of processing data as ordered pairs in the statistical mode. Typically one goes into the statistical mode for linear regression and entering ordered pairs separated by a comma. Once the data is keyed in, the three statistics a, b, and r can be retrieved from the storage registers with the appropriate label. Also note that the intermediate sums required in the formulae can be retrieved by appropriate keystrokes. Finally in linear regression mode the calculator is also able to produce means and standard deviations of both the x and y values.

On a computer, spreadsheet programs can plot scatterplots, produce the least-squares equation and also calculate the correlation coefficient. Googling **regression calculator** will find sites where you can enter data and do a regression analysis. For more sophisticated statistical work there are statistical programs available, many of them free, such as the **R** project for statistical computing. Statistical libraries accessible from within computer programs such as those included in **Scipy** for the **Python** programming language are also available. Google **statistics free software**.

Assignment:

1. The table below presents data for a random sample of eight students showing the amount of outside class study time spent and the grade earned at the end of a one month statistics course.

Study Time (h)	Grade (%)			
20	64			
16	61			
34	84			
23	70			
27	88			
32	92			
18	72			
22	77			

Answer the following questions by completing the table and check your work on your calculator.

   (a) Draw a scatterplot of the data and a best-fit straight line by hand. **Estimate** the intercept a, the slope b, and the correlation coefficient r from your line and the scatter about it.

   (b) Determine the regression equation for estimating the examination grade given the hours of study. Carry your calculations to one place beyond the decimal place.

   (c) Use the regression equation to estimate the examination grade for a student who devotes 30 hours to study.

   (d) Compute and interpret the correlation coefficient.

2. The following gives the peak power load in megawatts for a power plant and the daily high temperature for a random sample of 10 days:

Day	1	2	3	4	5	6	7	8	9	10
Daily High Temp (°F)	95	82	90	81	99	100	93	95	93	87
Peak Load (MW)	214	152	156	129	254	266	210	204	213	150

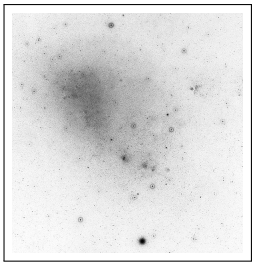
Do the following in a tabular format.

   (a) Determine the best-fitting straight line (linear regression equation) to predict peak load, y, based on the daily high temperature, x.

   (b) Using the prediction equation determined in (a), find the peak load for a daily high temperature of 85°F.

   (c) What is the coefficient of correlation? What two things does the coefficient of correlation tell us about the relationship between daily high temperature, x, and the peak load, y?

3. Cepheid variable stars have the interesting property that their brightness changes in a periodic (regular) way over time. Measuring the time between when a star is at its brightest gives its period. In 1908, Henrietta S. Leavitt published the following data concerning sixteen Cepheid variable stars in a nearby dwarf galaxy called the Small Magellanic Cloud (SMC). The following data gives the (natural logarithm of the) period in days, x, of each variable star along with its apparent magnitude, y, when it is at its maximum brightness.



The SMC (photo courtesy ESA/Hubble)

x	y	x	y
2.34	13.6	1.51	14.3
4.84	11.2	0.51	14.8
3.46	12.2	0.57	14.8
4.19	11.4	0.23	14.8
2.60	13.4	0.63	15.1
1.46	14.6	1.67	14.4
2.13	13.9	1.67	14.3
1.89	14.1	1.61	14.3

(a) Using your calculator, determine the correlation coefficient for the data and interpret your result.

(b) Using your calculator, determine the linear regression line and use it to determine the expected maximum apparent magnitude of a variable star in the SMC of period 60.2 days. (Remember x is the natural logarithm of this time.)

Bivariate Data Analysis

Linear Regression

$$y_p = a + bx$$

$$a = \frac{\left(\sum x^2\right) \cdot \left(\sum y\right) - \left(\sum x\right) \cdot \left(\sum xy\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} \qquad {}^*a = \bar{y} - b\bar{x}$$

$$b = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} \qquad {}^*b = r\left(\frac{s_y}{s_x}\right)$$

Correlation Analysis

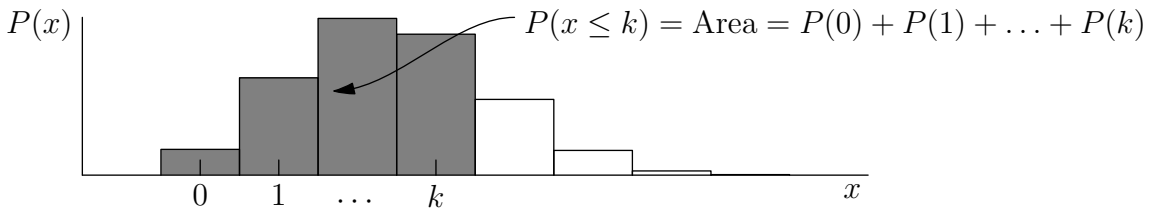
$$r = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{n \cdot \left(\sum x^2\right) - \left(\sum x\right)^2} \cdot \sqrt{n \cdot \left(\sum y^2\right) - \left(\sum y\right)^2}} \qquad {}^*r = \frac{s_{xy}}{s_x \cdot s_y}$$

Covariance

$${}^*s_{xy} = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{n-1}$$

Statistical Tables

Cumulative Binomial Probabilities, $P(x \leq k)$



n=1

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.990	0.950	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050	0.010
1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=2

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.980	0.902	0.810	0.640	0.490	0.360	0.250	0.160	0.090	0.040	0.010	0.003	0.000
1	1.000	0.998	0.990	0.960	0.910	0.840	0.750	0.640	0.510	0.360	0.190	0.098	0.020
2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=3

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.970	0.857	0.729	0.512	0.343	0.216	0.125	0.064	0.027	0.008	0.001	0.000	0.000
1	1.000	0.993	0.972	0.896	0.784	0.648	0.500	0.352	0.216	0.104	0.028	0.007	0.000
2	1.000	1.000	0.999	0.992	0.973	0.936	0.875	0.784	0.657	0.488	0.271	0.143	0.030
3	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=4

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.961	0.815	0.656	0.410	0.240	0.130	0.062	0.026	0.008	0.002	0.000	0.000	0.000
1	0.999	0.986	0.948	0.819	0.652	0.475	0.312	0.179	0.084	0.027	0.004	0.000	0.000
2	1.000	1.000	0.996	0.973	0.916	0.821	0.688	0.525	0.348	0.181	0.052	0.014	0.001
3	1.000	1.000	1.000	0.998	0.992	0.974	0.938	0.870	0.760	0.590	0.344	0.185	0.039
4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=5

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.951	0.774	0.590	0.328	0.168	0.078	0.031	0.010	0.002	0.000	0.000	0.000	0.000
1	0.999	0.977	0.919	0.737	0.528	0.337	0.188	0.087	0.031	0.007	0.000	0.000	0.000
2	1.000	0.999	0.991	0.942	0.837	0.683	0.500	0.317	0.163	0.058	0.009	0.001	0.000
3	1.000	1.000	1.000	0.993	0.969	0.913	0.812	0.663	0.472	0.263	0.081	0.023	0.001
4	1.000	1.000	1.000	1.000	0.998	0.990	0.969	0.922	0.832	0.672	0.410	0.226	0.049
5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=6

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.941	0.735	0.531	0.262	0.118	0.047	0.016	0.004	0.001	0.000	0.000	0.000	0.000
1	0.999	0.967	0.886	0.655	0.420	0.233	0.109	0.041	0.011	0.002	0.000	0.000	0.000
2	1.000	0.998	0.984	0.901	0.744	0.544	0.344	0.179	0.070	0.017	0.001	0.000	0.000
3	1.000	1.000	0.999	0.983	0.930	0.821	0.656	0.456	0.256	0.099	0.016	0.002	0.000
4	1.000	1.000	1.000	0.998	0.989	0.959	0.891	0.767	0.580	0.345	0.114	0.033	0.001
5	1.000	1.000	1.000	1.000	0.999	0.996	0.984	0.953	0.882	0.738	0.469	0.265	0.059
6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=7

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.932	0.698	0.478	0.210	0.082	0.028	0.008	0.002	0.000	0.000	0.000	0.000	0.000
1	0.998	0.956	0.850	0.577	0.329	0.159	0.062	0.019	0.004	0.000	0.000	0.000	0.000
2	1.000	0.996	0.974	0.852	0.647	0.420	0.227	0.096	0.029	0.005	0.000	0.000	0.000
3	1.000	1.000	0.997	0.967	0.874	0.710	0.500	0.290	0.126	0.033	0.003	0.000	0.000
4	1.000	1.000	1.000	0.995	0.971	0.904	0.773	0.580	0.353	0.148	0.026	0.004	0.000
5	1.000	1.000	1.000	1.000	0.996	0.981	0.938	0.841	0.671	0.423	0.150	0.044	0.002
6	1.000	1.000	1.000	1.000	1.000	0.998	0.992	0.972	0.918	0.790	0.522	0.302	0.068
7	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=8

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.923	0.663	0.430	0.168	0.058	0.017	0.004	0.001	0.000	0.000	0.000	0.000	0.000
1	0.997	0.943	0.813	0.503	0.255	0.106	0.035	0.009	0.001	0.000	0.000	0.000	0.000
2	1.000	0.994	0.962	0.797	0.552	0.315	0.145	0.050	0.011	0.001	0.000	0.000	0.000
3	1.000	1.000	0.995	0.944	0.806	0.594	0.363	0.174	0.058	0.010	0.000	0.000	0.000
4	1.000	1.000	1.000	0.990	0.942	0.826	0.637	0.406	0.194	0.056	0.005	0.000	0.000
5	1.000	1.000	1.000	0.999	0.989	0.950	0.855	0.685	0.448	0.203	0.038	0.006	0.000
6	1.000	1.000	1.000	1.000	0.999	0.991	0.965	0.894	0.745	0.497	0.187	0.057	0.003
7	1.000	1.000	1.000	1.000	1.000	0.999	0.996	0.983	0.942	0.832	0.570	0.337	0.077
8	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=9

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.914	0.630	0.387	0.134	0.040	0.010	0.002	0.000	0.000	0.000	0.000	0.000	0.000
1	0.997	0.929	0.775	0.436	0.196	0.071	0.020	0.004	0.000	0.000	0.000	0.000	0.000
2	1.000	0.992	0.947	0.738	0.463	0.232	0.090	0.025	0.004	0.000	0.000	0.000	0.000
3	1.000	0.999	0.992	0.914	0.730	0.483	0.254	0.099	0.025	0.003	0.000	0.000	0.000
4	1.000	1.000	0.999	0.980	0.901	0.733	0.500	0.267	0.099	0.020	0.001	0.000	0.000
5	1.000	1.000	1.000	0.997	0.975	0.901	0.746	0.517	0.270	0.086	0.008	0.001	0.000
6	1.000	1.000	1.000	1.000	0.996	0.975	0.910	0.768	0.537	0.262	0.053	0.008	0.000
7	1.000	1.000	1.000	1.000	1.000	0.996	0.980	0.929	0.804	0.564	0.225	0.071	0.003
8	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.990	0.960	0.866	0.613	0.370	0.086
9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=10

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.904	0.599	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000	0.000	0.000
1	0.996	0.914	0.736	0.376	0.149	0.046	0.011	0.002	0.000	0.000	0.000	0.000	0.000
2	1.000	0.988	0.930	0.678	0.383	0.167	0.055	0.012	0.002	0.000	0.000	0.000	0.000
3	1.000	0.999	0.987	0.879	0.650	0.382	0.172	0.055	0.011	0.001	0.000	0.000	0.000
4	1.000	1.000	0.998	0.967	0.850	0.633	0.377	0.166	0.047	0.006	0.000	0.000	0.000
5	1.000	1.000	1.000	0.994	0.953	0.834	0.623	0.367	0.150	0.033	0.002	0.000	0.000
6	1.000	1.000	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013	0.001	0.000
7	1.000	1.000	1.000	1.000	0.998	0.988	0.945	0.833	0.617	0.322	0.070	0.012	0.000
8	1.000	1.000	1.000	1.000	1.000	0.998	0.989	0.954	0.851	0.624	0.264	0.086	0.004
9	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.994	0.972	0.893	0.651	0.401	0.096
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=11

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.895	0.569	0.314	0.086	0.020	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.995	0.898	0.697	0.322	0.113	0.030	0.006	0.001	0.000	0.000	0.000	0.000	0.000
2	1.000	0.985	0.910	0.617	0.313	0.119	0.033	0.006	0.001	0.000	0.000	0.000	0.000
3	1.000	0.998	0.981	0.839	0.570	0.296	0.113	0.029	0.004	0.000	0.000	0.000	0.000
4	1.000	1.000	0.997	0.950	0.790	0.533	0.274	0.099	0.022	0.002	0.000	0.000	0.000
5	1.000	1.000	1.000	0.988	0.922	0.753	0.500	0.247	0.078	0.012	0.000	0.000	0.000
6	1.000	1.000	1.000	0.998	0.978	0.901	0.726	0.467	0.210	0.050	0.003	0.000	0.000
7	1.000	1.000	1.000	1.000	0.996	0.971	0.887	0.704	0.430	0.161	0.019	0.002	0.000
8	1.000	1.000	1.000	1.000	0.999	0.994	0.967	0.881	0.687	0.383	0.090	0.015	0.000
9	1.000	1.000	1.000	1.000	1.000	0.999	0.994	0.970	0.887	0.678	0.303	0.102	0.005
10	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.980	0.914	0.686	0.431	0.105
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=12

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.886	0.540	0.282	0.069	0.014	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.994	0.882	0.659	0.275	0.085	0.020	0.003	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.980	0.889	0.558	0.253	0.083	0.019	0.003	0.000	0.000	0.000	0.000	0.000
3	1.000	0.998	0.974	0.795	0.493	0.225	0.073	0.015	0.002	0.000	0.000	0.000	0.000
4	1.000	1.000	0.996	0.927	0.724	0.438	0.194	0.057	0.009	0.001	0.000	0.000	0.000
5	1.000	1.000	0.999	0.981	0.882	0.665	0.387	0.158	0.039	0.004	0.000	0.000	0.000
6	1.000	1.000	1.000	0.996	0.961	0.842	0.613	0.335	0.118	0.019	0.001	0.000	0.000
7	1.000	1.000	1.000	0.999	0.991	0.943	0.806	0.562	0.276	0.073	0.004	0.000	0.000
8	1.000	1.000	1.000	1.000	0.998	0.985	0.927	0.775	0.507	0.205	0.026	0.002	0.000
9	1.000	1.000	1.000	1.000	1.000	0.997	0.981	0.917	0.747	0.442	0.111	0.020	0.000
10	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.980	0.915	0.725	0.341	0.118	0.006
11	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.986	0.931	0.718	0.460	0.114
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=13

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.878	0.513	0.254	0.055	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.993	0.865	0.621	0.234	0.064	0.013	0.002	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.975	0.866	0.502	0.202	0.058	0.011	0.001	0.000	0.000	0.000	0.000	0.000
3	1.000	0.997	0.966	0.747	0.421	0.169	0.046	0.008	0.001	0.000	0.000	0.000	0.000
4	1.000	1.000	0.994	0.901	0.654	0.353	0.133	0.032	0.004	0.000	0.000	0.000	0.000
5	1.000	1.000	0.999	0.970	0.835	0.574	0.291	0.098	0.018	0.001	0.000	0.000	0.000
6	1.000	1.000	1.000	0.993	0.938	0.771	0.500	0.229	0.062	0.007	0.000	0.000	0.000
7	1.000	1.000	1.000	0.999	0.982	0.902	0.709	0.426	0.165	0.030	0.001	0.000	0.000
8	1.000	1.000	1.000	1.000	0.996	0.968	0.867	0.647	0.346	0.099	0.006	0.000	0.000
9	1.000	1.000	1.000	1.000	0.999	0.992	0.954	0.831	0.579	0.253	0.034	0.003	0.000
10	1.000	1.000	1.000	1.000	1.000	0.999	0.989	0.942	0.798	0.498	0.134	0.025	0.000
11	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.987	0.936	0.766	0.379	0.135	0.007
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.990	0.945	0.746	0.487	0.122
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=14

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.869	0.488	0.229	0.044	0.007	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.992	0.847	0.585	0.198	0.047	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.970	0.842	0.448	0.161	0.040	0.006	0.001	0.000	0.000	0.000	0.000	0.000
3	1.000	0.996	0.956	0.698	0.355	0.124	0.029	0.004	0.000	0.000	0.000	0.000	0.000
4	1.000	1.000	0.991	0.870	0.584	0.279	0.090	0.018	0.002	0.000	0.000	0.000	0.000
5	1.000	1.000	0.999	0.956	0.781	0.486	0.212	0.058	0.008	0.000	0.000	0.000	0.000
6	1.000	1.000	1.000	0.988	0.907	0.692	0.395	0.150	0.031	0.002	0.000	0.000	0.000
7	1.000	1.000	1.000	0.998	0.969	0.850	0.605	0.308	0.093	0.012	0.000	0.000	0.000
8	1.000	1.000	1.000	1.000	0.992	0.942	0.788	0.514	0.219	0.044	0.001	0.000	0.000
9	1.000	1.000	1.000	1.000	0.998	0.982	0.910	0.721	0.416	0.130	0.009	0.000	0.000
10	1.000	1.000	1.000	1.000	1.000	0.996	0.971	0.876	0.645	0.302	0.044	0.004	0.000
11	1.000	1.000	1.000	1.000	1.000	0.999	0.994	0.960	0.839	0.552	0.158	0.030	0.000
12	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.992	0.953	0.802	0.415	0.153	0.008
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.993	0.956	0.771	0.512	0.131
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=15

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.860	0.463	0.206	0.035	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.990	0.829	0.549	0.167	0.035	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.000	0.964	0.816	0.398	0.127	0.027	0.004	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.995	0.944	0.648	0.297	0.091	0.018	0.002	0.000	0.000	0.000	0.000	0.000
4	1.000	0.999	0.987	0.836	0.515	0.217	0.059	0.009	0.001	0.000	0.000	0.000	0.000
5	1.000	1.000	0.998	0.939	0.722	0.403	0.151	0.034	0.004	0.000	0.000	0.000	0.000
6	1.000	1.000	1.000	0.982	0.869	0.610	0.304	0.095	0.015	0.001	0.000	0.000	0.000
7	1.000	1.000	1.000	0.996	0.950	0.787	0.500	0.213	0.050	0.004	0.000	0.000	0.000
8	1.000	1.000	1.000	0.999	0.985	0.905	0.696	0.390	0.131	0.018	0.000	0.000	0.000
9	1.000	1.000	1.000	1.000	0.996	0.966	0.849	0.597	0.278	0.061	0.002	0.000	0.000
10	1.000	1.000	1.000	1.000	0.999	0.991	0.941	0.783	0.485	0.164	0.013	0.001	0.000
11	1.000	1.000	1.000	1.000	1.000	0.998	0.982	0.909	0.703	0.352	0.056	0.005	0.000
12	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.973	0.873	0.602	0.184	0.036	0.000
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.833	0.451	0.171	0.010
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.965	0.794	0.537	0.140
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=16

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.851	0.440	0.185	0.028	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.989	0.811	0.515	0.141	0.026	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.999	0.957	0.789	0.352	0.099	0.018	0.002	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.993	0.932	0.598	0.246	0.065	0.011	0.001	0.000	0.000	0.000	0.000	0.000
4	1.000	0.999	0.983	0.798	0.450	0.167	0.038	0.005	0.000	0.000	0.000	0.000	0.000
5	1.000	1.000	0.997	0.918	0.660	0.329	0.105	0.019	0.002	0.000	0.000	0.000	0.000
6	1.000	1.000	0.999	0.973	0.825	0.527	0.227	0.058	0.007	0.000	0.000	0.000	0.000
7	1.000	1.000	1.000	0.993	0.926	0.716	0.402	0.142	0.026	0.001	0.000	0.000	0.000
8	1.000	1.000	1.000	0.999	0.974	0.858	0.598	0.284	0.074	0.007	0.000	0.000	0.000
9	1.000	1.000	1.000	1.000	0.993	0.942	0.773	0.473	0.175	0.027	0.001	0.000	0.000
10	1.000	1.000	1.000	1.000	0.998	0.981	0.895	0.671	0.340	0.082	0.003	0.000	0.000
11	1.000	1.000	1.000	1.000	1.000	0.995	0.962	0.833	0.550	0.202	0.017	0.001	0.000
12	1.000	1.000	1.000	1.000	1.000	0.999	0.989	0.935	0.754	0.402	0.068	0.007	0.000
13	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.982	0.901	0.648	0.211	0.043	0.001
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.974	0.859	0.485	0.189	0.011
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.972	0.815	0.560	0.149
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=17

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.843	0.418	0.167	0.023	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.988	0.792	0.482	0.118	0.019	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.999	0.950	0.762	0.310	0.077	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.991	0.917	0.549	0.202	0.046	0.006	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.999	0.978	0.758	0.389	0.126	0.025	0.003	0.000	0.000	0.000	0.000	0.000
5	1.000	1.000	0.995	0.894	0.597	0.264	0.072	0.011	0.001	0.000	0.000	0.000	0.000
6	1.000	1.000	0.999	0.962	0.775	0.448	0.166	0.035	0.003	0.000	0.000	0.000	0.000
7	1.000	1.000	1.000	0.989	0.895	0.641	0.315	0.092	0.013	0.000	0.000	0.000	0.000
8	1.000	1.000	1.000	0.997	0.960	0.801	0.500	0.199	0.040	0.003	0.000	0.000	0.000
9	1.000	1.000	1.000	1.000	0.987	0.908	0.685	0.359	0.105	0.011	0.000	0.000	0.000
10	1.000	1.000	1.000	1.000	0.997	0.965	0.834	0.552	0.225	0.038	0.001	0.000	0.000
11	1.000	1.000	1.000	1.000	0.999	0.989	0.928	0.736	0.403	0.106	0.005	0.000	0.000
12	1.000	1.000	1.000	1.000	1.000	0.997	0.975	0.874	0.611	0.242	0.022	0.001	0.000
13	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.954	0.798	0.451	0.083	0.009	0.000
14	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.923	0.690	0.238	0.050	0.001
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.981	0.882	0.518	0.208	0.012
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.977	0.833	0.582	0.157
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=18

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.835	0.397	0.150	0.018	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.986	0.774	0.450	0.099	0.014	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.999	0.942	0.734	0.271	0.060	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.989	0.902	0.501	0.165	0.033	0.004	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.998	0.972	0.716	0.333	0.094	0.015	0.001	0.000	0.000	0.000	0.000	0.000
5	1.000	1.000	0.994	0.867	0.534	0.209	0.048	0.006	0.000	0.000	0.000	0.000	0.000
6	1.000	1.000	0.999	0.949	0.722	0.374	0.119	0.020	0.001	0.000	0.000	0.000	0.000
7	1.000	1.000	1.000	0.984	0.859	0.563	0.240	0.058	0.006	0.000	0.000	0.000	0.000
8	1.000	1.000	1.000	0.996	0.940	0.737	0.407	0.135	0.021	0.001	0.000	0.000	0.000
9	1.000	1.000	1.000	0.999	0.979	0.865	0.593	0.263	0.060	0.004	0.000	0.000	0.000
10	1.000	1.000	1.000	1.000	0.994	0.942	0.760	0.437	0.141	0.016	0.000	0.000	0.000
11	1.000	1.000	1.000	1.000	0.999	0.980	0.881	0.626	0.278	0.051	0.001	0.000	0.000
12	1.000	1.000	1.000	1.000	1.000	0.994	0.952	0.791	0.466	0.133	0.006	0.000	0.000
13	1.000	1.000	1.000	1.000	1.000	0.999	0.985	0.906	0.667	0.284	0.028	0.002	0.000
14	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.967	0.835	0.499	0.098	0.011	0.000
15	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.992	0.940	0.729	0.266	0.058	0.001
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.986	0.901	0.550	0.226	0.014
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.982	0.850	0.603	0.165
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=19

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.826	0.377	0.135	0.014	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.985	0.755	0.420	0.083	0.010	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.999	0.933	0.705	0.237	0.046	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.987	0.885	0.455	0.133	0.023	0.002	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.998	0.965	0.673	0.282	0.070	0.010	0.001	0.000	0.000	0.000	0.000	0.000
5	1.000	1.000	0.991	0.837	0.474	0.163	0.032	0.003	0.000	0.000	0.000	0.000	0.000
6	1.000	1.000	0.998	0.932	0.666	0.308	0.084	0.012	0.001	0.000	0.000	0.000	0.000
7	1.000	1.000	1.000	0.977	0.818	0.488	0.180	0.035	0.003	0.000	0.000	0.000	0.000
8	1.000	1.000	1.000	0.993	0.916	0.667	0.324	0.088	0.011	0.000	0.000	0.000	0.000
9	1.000	1.000	1.000	0.998	0.967	0.814	0.500	0.186	0.033	0.002	0.000	0.000	0.000
10	1.000	1.000	1.000	1.000	0.989	0.912	0.676	0.333	0.084	0.007	0.000	0.000	0.000
11	1.000	1.000	1.000	1.000	0.997	0.965	0.820	0.512	0.182	0.023	0.000	0.000	0.000
12	1.000	1.000	1.000	1.000	0.999	0.988	0.916	0.692	0.334	0.068	0.002	0.000	0.000
13	1.000	1.000	1.000	1.000	1.000	0.997	0.968	0.837	0.526	0.163	0.009	0.000	0.000
14	1.000	1.000	1.000	1.000	1.000	0.999	0.990	0.930	0.718	0.327	0.035	0.002	0.000
15	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.977	0.867	0.545	0.115	0.013	0.000
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.995	0.954	0.763	0.295	0.067	0.001
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.990	0.917	0.580	0.245	0.015
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.986	0.865	0.623	0.174
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

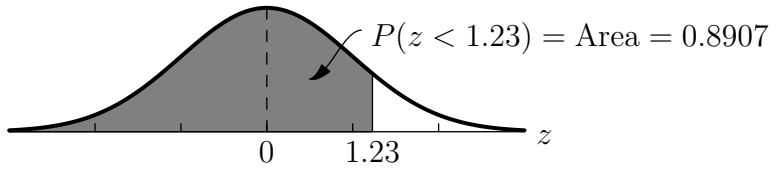
n=20

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.818	0.358	0.122	0.012	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.983	0.736	0.392	0.069	0.008	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.999	0.925	0.677	0.206	0.035	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.984	0.867	0.411	0.107	0.016	0.001	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.997	0.957	0.630	0.238	0.051	0.006	0.000	0.000	0.000	0.000	0.000	0.000
5	1.000	1.000	0.989	0.804	0.416	0.126	0.021	0.002	0.000	0.000	0.000	0.000	0.000
6	1.000	1.000	0.998	0.913	0.608	0.250	0.058	0.006	0.000	0.000	0.000	0.000	0.000
7	1.000	1.000	1.000	0.968	0.772	0.416	0.132	0.021	0.001	0.000	0.000	0.000	0.000
8	1.000	1.000	1.000	0.990	0.887	0.596	0.252	0.057	0.005	0.000	0.000	0.000	0.000
9	1.000	1.000	1.000	0.997	0.952	0.755	0.412	0.128	0.017	0.001	0.000	0.000	0.000
10	1.000	1.000	1.000	0.999	0.983	0.872	0.588	0.245	0.048	0.003	0.000	0.000	0.000
11	1.000	1.000	1.000	1.000	0.995	0.943	0.748	0.404	0.113	0.010	0.000	0.000	0.000
12	1.000	1.000	1.000	1.000	0.999	0.979	0.868	0.584	0.228	0.032	0.000	0.000	0.000
13	1.000	1.000	1.000	1.000	1.000	0.994	0.942	0.750	0.392	0.087	0.002	0.000	0.000
14	1.000	1.000	1.000	1.000	1.000	0.998	0.979	0.874	0.584	0.196	0.011	0.000	0.000
15	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.949	0.762	0.370	0.043	0.003	0.000
16	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.984	0.893	0.589	0.133	0.016	0.000
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.965	0.794	0.323	0.075	0.001
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.992	0.931	0.608	0.264	0.017
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.988	0.878	0.642	0.182
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

n=25

	π												
k	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0	0.778	0.277	0.072	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.974	0.642	0.271	0.027	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.998	0.873	0.537	0.098	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	1.000	0.966	0.764	0.234	0.033	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	1.000	0.993	0.902	0.421	0.090	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	1.000	0.999	0.967	0.617	0.193	0.029	0.002	0.000	0.000	0.000	0.000	0.000	0.000
6	1.000	1.000	0.991	0.780	0.341	0.074	0.007	0.000	0.000	0.000	0.000	0.000	0.000
7	1.000	1.000	0.998	0.891	0.512	0.154	0.022	0.001	0.000	0.000	0.000	0.000	0.000
8	1.000	1.000	1.000	0.953	0.677	0.274	0.054	0.004	0.000	0.000	0.000	0.000	0.000
9	1.000	1.000	1.000	0.983	0.811	0.425	0.115	0.013	0.000	0.000	0.000	0.000	0.000
10	1.000	1.000	1.000	0.994	0.902	0.586	0.212	0.034	0.002	0.000	0.000	0.000	0.000
11	1.000	1.000	1.000	0.998	0.956	0.732	0.345	0.078	0.006	0.000	0.000	0.000	0.000
12	1.000	1.000	1.000	1.000	0.983	0.846	0.500	0.154	0.017	0.000	0.000	0.000	0.000
13	1.000	1.000	1.000	1.000	0.994	0.922	0.655	0.268	0.044	0.002	0.000	0.000	0.000
14	1.000	1.000	1.000	1.000	0.998	0.966	0.788	0.414	0.098	0.006	0.000	0.000	0.000
15	1.000	1.000	1.000	1.000	1.000	0.987	0.885	0.575	0.189	0.017	0.000	0.000	0.000
16	1.000	1.000	1.000	1.000	1.000	0.996	0.946	0.726	0.323	0.047	0.000	0.000	0.000
17	1.000	1.000	1.000	1.000	1.000	0.999	0.978	0.846	0.488	0.109	0.002	0.000	0.000
18	1.000	1.000	1.000	1.000	1.000	1.000	0.993	0.926	0.659	0.220	0.009	0.000	0.000
19	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.971	0.807	0.383	0.033	0.001	0.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.910	0.579	0.098	0.007	0.000
21	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.967	0.766	0.236	0.034	0.000
22	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.991	0.902	0.463	0.127	0.002
23	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.998	0.973	0.729	0.358	0.026
24	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.928	0.723	0.222
25	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

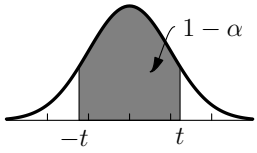
Normal Curve Areas



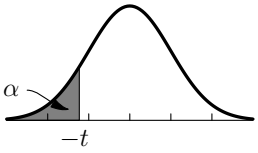
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

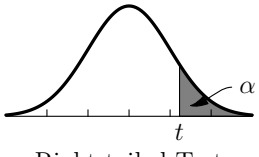
Student's t Distribution (Critical Values)



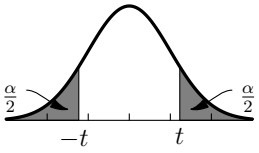
Confidence Interval



Left-tailed Test



Right-tailed Test



Two-tailed Test

df	Confidence Coefficient, $1-\alpha$					
	0.80	0.90	0.95	0.98	0.99	0.999
	Level of Significance for One-Tailed Test, α					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646

df	Confidence Coefficient, $1-\alpha$					
	0.80	0.90	0.95	0.98	0.99	0.999
	Level of Significance for One-Tailed Test, α					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
200	1.286	1.653	1.972	2.345	2.601	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.291

The F Distribution (Critical Values)

df_1

df_2	1							2							3							4							5						
1	39.86	161.4	647.8	4052	16211			49.50	199.5	799.5	5000	20000			53.59	215.7	864.2	5403	21615			55.83	224.6	899.6	5625	22500			57.24	230.2	921.8	5764	23056	
2	8.526	18.51	38.51	98.50	198.5			9.000	19.00	39.00	99.00	199.0			9.162	19.16	39.17	99.17	199.2			9.243	19.25	39.25	99.25	199.2			9.293	19.30	39.30	99.30	199.3	
3	5.538	10.13	17.44	34.12	55.55			5.462	9.552	16.04	30.82	49.80			5.391	9.277	15.44	29.46	47.47			5.343	9.117	15.10	28.71	46.19			5.309	9.013	14.88	28.24	45.39	
4	4.545	7.709	12.22	21.20	31.33			4.325	6.944	10.65	18.00	26.28			4.191	6.591	9.979	16.69	24.26			4.107	6.388	9.605	15.98	23.15			4.051	6.256	9.364	15.52	22.46	
5	4.060	6.608	10.01	16.26	22.78			3.780	5.786	8.434	13.27	18.31			3.619	5.409	7.764	12.06	16.53			3.520	5.192	7.388	11.39	15.56			3.453	5.050	7.146	10.97	14.94	
6	3.776	5.987	8.813	13.75	18.63			3.463	5.143	7.260	10.92	14.54			3.289	4.757	6.599	9.780	12.92			3.181	4.534	6.227	9.148	12.03			3.108	4.387	5.988	8.746	11.46	
7	3.589	5.591	8.073	12.25	16.24			3.257	4.737	6.542	9.547	12.40			3.074	4.347	5.890	8.451	10.88			2.961	4.120	5.523	7.847	10.05			2.883	3.972	5.285	7.460	9.522	
8	3.458	5.318	7.571	11.26	14.69			3.113	4.459	6.059	8.649	11.04			2.924	4.066	5.416	7.591	9.596			2.806	3.838	5.053	7.006	8.805			2.726	3.687	4.817	6.632	8.302	
9	3.360	5.117	7.209	10.56	13.61			3.006	4.256	5.715	8.022	10.11			2.813	3.863	5.078	6.992	8.717			2.693	3.633	4.718	6.422	7.956			2.611	3.482	4.484	6.057	7.471	
10	3.285	4.965	6.937	10.04	12.83			2.924	4.103	5.456	7.559	9.427			2.728	3.708	4.826	6.552	8.081			2.605	3.478	4.468	5.994	7.343			2.522	3.326	4.236	5.636	6.872	
11	3.225	4.844	6.724	9.646	12.23			2.860	3.982	5.256	7.206	8.912			2.660	3.587	4.630	6.217	7.600			2.536	3.357	4.275	5.668	6.881			2.451	3.204	4.044	5.316	6.422	
12	3.177	4.747	6.554	9.330	11.75			2.807	3.885	5.096	6.927	8.510			2.606	3.490	4.474	5.953	7.226			2.480	3.259	4.121	5.412	6.521			2.394	3.106	3.891	5.064	6.071	
13	3.136	4.667	6.414	9.074	11.37			2.763	3.806	4.965	6.701	8.186			2.560	3.411	4.347	5.739	6.926			2.434	3.179	3.996	5.205	6.233			2.347	3.025	3.767	4.862	5.791	
14	3.102	4.600	6.298	8.862	11.06			2.726	3.739	4.857	6.515	7.922			2.522	3.344	4.242	5.564	6.680			2.395	3.112	3.892	5.035	5.998			2.307	2.958	3.663	4.695	5.562	
15	3.073	4.543	6.200	8.683	10.80			2.695	3.682	4.765	6.359	7.701			2.490	3.287	4.153	5.417	6.476			2.361	3.056	3.804	4.893	5.803			2.273	2.901	3.576	4.556	5.372	
16	3.048	4.494	6.115	8.531	10.58			2.668	3.634	4.687	6.226	7.514			2.462	3.239	4.077	5.292	6.303			2.333	3.007	3.729	4.773	5.638			2.244	2.852	3.502	4.437	5.212	
17	3.026	4.451	6.042	8.400	10.38			2.645	3.592	4.619	6.112	7.354			2.437	3.197	4.011	5.185	6.156			2.308	2.965	3.665	4.669	5.497			2.218	2.810	3.438	4.336	5.075	
18	3.007	4.414	5.978	8.285	10.22			2.624	3.555	4.560	6.013	7.215			2.416	3.160	3.954	5.092	6.028			2.286	2.928	3.608	4.579	5.375			2.196	2.773	3.382	4.248	4.956	
19	2.990	4.381	5.922	8.185	10.07			2.606	3.522	4.508	5.926	7.093			2.397	3.127	3.903	5.010	5.916			2.266	2.895	3.559	4.500	5.268			2.176	2.740	3.333	4.171	4.853	
20	2.975	4.351	5.871	8.096	9.944			2.589	3.493	4.461	5.849	6.986			2.380	3.098	3.859	4.938	5.818			2.249	2.866	3.515	4.431	5.174			2.158	2.711	3.289	4.103	4.762	
21	2.961	4.325	5.827	8.017	9.830			2.575	3.467	4.420	5.780	6.891			2.365	3.072	3.819	4.874	5.730			2.233	2.840	3.475	4.369	5.091			2.142	2.685	3.250	4.042	4.681	
22	2.949	4.301	5.786	7.945	9.727			2.561	3.443	4.383	5.719	6.806			2.351	3.049	3.783	4.817	5.652			2.219	2.817	3.440	4.313	5.017			2.128	2.661	3.215	3.988	4.609	
23	2.937	4.279	5.750	7.881	9.635			2.549	3.422	4.349	5.664	6.730			2.339	3.028	3.750	4.765	5.582			2.207	2.796	3.408	4.264	4.950			2.115	2.640	3.183	3.939	4.544	
24	2.927	4.260	5.717	7.823	9.551			2.538	3.403	4.319	5.614	6.661			2.327	3.009	3.721	4.718	5.519			2.195	2.776	3.379	4.218	4.890			2.103	2.621	3.155	3.895	4.486	
25	2.918	4.242	5.686	7.770	9.475			2.528	3.385	4.291	5.568	6.598			2.317	2.991	3.694	4.675	5.462			2.184	2.759	3.353	4.177	4.835			2.092	2.603	3.129	3.855	4.433	
26	2.909	4.225	5.659	7.721	9.406			2.519	3.369	4.265	5.526	6.541			2.307	2.975	3.670	4.637	5.409			2.174	2.743	3.329	4.140	4.785			2.082	2.587	3.105	3.818	4.384	
27	2.901	4.210	5.633	7.677	9.342			2.511	3.354	4.242	5.488	6.489			2.299	2.960	3.647	4.601	5.361			2.165	2.728	3.307	4.106	4.740			2.073	2.572	3.083	3.785	4.340	
28	2.894	4.196	5.610	7.636	9.284			2.503	3.340	4.221	5.453	6.440			2.291	2.947	3.626	4.568	5.317			2.157	2.714	3.286	4.074	4.698			2.064	2.558	3.063	3.754	4.300	
29	2.887	4.183	5.588	7.598	9.230			2.495	3.328	4.201	5.420	6.396			2.283	2.934	3.607	4.538	5.276			2.149	2.701	3.267	4.045	4.659			2.057	2.545	3.044	3.725	4.262	
30	2.881	4.171	5.568	7.562	9.180			2.489	3.316	4.182	5.390	6.355			2.276	2.922	3.589	4.510	5.239			2.142	2.690	3.250	4.018	4.623			2.049	2.534	3.026	3.699	4.228	
40	2.835	4.085	5.424	7.314	8.828			2.440	3.232	4.051	5.179	6.066			2.226	2.839	3.463	4.313	4.976			2.091	2.606	3.126	3.828	4.374			1.997	2.449	2.904	3.514	3.986	
60	2.791	4.001	5.286	7.077	8.495			2.393	3.150	3.925	4.977	5.795			2.177	2.758	3.343	4.126	4.729			2.041	2.525	3.008	3.649	4.140			1.946	2.368	2.786	3.339	3.760	
80	2.769	3.960	5.218	6.963	8.335			2.370	3.111	3.864	4.881	5.665			2.154	2.719	3.284	4.036	4.611			2.016	2.486	2.950	3.563	4.029			1.921	2.329	2.730	3.255	3.652	
100	2.756	3.936	5.179	6.895	8.241			2.356	3.087	3.828	4.824	5.589			2.139	2.696	3.250	3.984	4.542			2.002	2.463	2.917	3.513	3.963			1.906	2.305	2.696	3.206	3.589	
∞	2.706	3.841	5.024	6.635	7.879			2.303	2.996	3.689	4.605	5.298			2.084	2.605	3.116	3.782	4.279			1.945	2.372	2.786	3.319	3.715			1.847	2.214	2.567	3.017	3.350	
	0.100	0.050	0.025	0.010	0.005			0.100	0.050	0.025	0.010	0.005			0.100	0.050	0.025	0.010	0.005			0.100	0.050	0.025	0.010	0.005			0.100	0.050	0.025	0.010	0.005	

α

df_2	df_1 = 6					7					8					9					10				
1	58.20	234.0	937.1	5859	23437	58.91	236.8	948.2	5928	23715	59.44	238.9	956.7	5981	23925	59.86	240.5	963.3	6022	24091	60.19	241.9	968.6	6056	24224
2	9.326	19.33	39.33	99.33	199.3	9.349	19.35	39.36	99.36	199.4	9.367	19.37	39.37	99.37	199.4	9.381	19.38	39.39	99.39	199.4	9.392	19.40	39.40	99.40	199.4
3	5.285	8.941	14.73	27.91	44.84	5.266	8.887	14.62	27.67	44.43	5.252	8.845	14.54	27.49	44.13	5.240	8.812	14.47	27.35	43.88	5.230	8.786	14.42	27.23	43.69
4	4.010	6.163	9.197	15.21	21.97	3.979	6.094	9.074	14.98	21.62	3.955	6.041	8.980	14.80	21.35	3.936	5.999	8.905	14.66	21.14	3.920	5.964	8.844	14.55	20.97
5	3.405	4.950	6.978	10.67	14.51	3.368	4.876	6.853	10.46	14.20	3.339	4.818	6.757	10.29	13.96	3.316	4.772	6.681	10.16	13.77	3.297	4.735	6.619	10.05	13.62
6	3.055	4.284	5.820	8.466	11.07	3.014	4.207	5.695	8.260	10.79	2.983	4.147	5.600	8.102	10.57	2.958	4.099	5.523	7.976	10.39	2.937	4.060	5.461	7.874	10.25
7	2.827	3.866	5.119	7.191	9.155	2.785	3.787	4.995	6.993	8.885	2.752	3.726	4.899	6.840	8.678	2.725	3.677	4.823	6.719	8.514	2.703	3.637	4.761	6.620	8.380
8	2.668	3.581	4.652	6.371	7.952	2.624	3.500	4.529	6.178	7.694	2.589	3.438	4.433	6.029	7.496	2.561	3.388	4.357	5.911	7.339	2.538	3.347	4.295	5.814	7.211
9	2.551	3.374	4.320	5.802	7.134	2.505	3.293	4.197	5.613	6.885	2.469	3.230	4.102	5.467	6.693	2.440	3.179	4.026	5.351	6.541	2.416	3.137	3.964	5.257	6.417
10	2.461	3.217	4.072	5.386	6.545	2.414	3.135	3.950	5.200	6.302	2.377	3.072	3.855	5.057	6.116	2.347	3.020	3.779	4.942	5.968	2.323	2.978	3.717	4.849	5.847
11	2.389	3.095	3.881	5.069	6.102	2.342	3.012	3.759	4.886	5.865	2.304	2.948	3.664	4.744	5.682	2.274	2.896	3.588	4.632	5.537	2.248	2.854	3.526	4.539	5.418
12	2.331	2.996	3.728	4.821	5.757	2.283	2.913	3.607	4.640	5.525	2.245	2.849	3.512	4.499	5.345	2.214	2.796	3.436	4.388	5.202	2.188	2.753	3.374	4.296	5.085
13	2.283	2.915	3.604	4.620	5.482	2.234	2.832	3.483	4.441	5.253	2.195	2.767	3.388	4.302	5.076	2.164	2.714	3.312	4.191	4.935	2.138	2.671	3.250	4.100	4.820
14	2.243	2.848	3.501	4.456	5.257	2.193	2.764	3.380	4.278	5.031	2.154	2.699	3.285	4.140	4.857	2.122	2.646	3.209	4.030	4.717	2.095	2.602	3.147	3.939	4.603
15	2.208	2.790	3.415	4.318	5.071	2.158	2.707	3.293	4.142	4.847	2.119	2.641	3.199	4.004	4.674	2.086	2.588	3.123	3.895	4.536	2.059	2.544	3.060	3.805	4.424
16	2.178	2.741	3.341	4.202	4.913	2.128	2.657	3.219	4.026	4.692	2.088	2.591	3.125	3.890	4.521	2.055	2.538	3.049	3.780	4.384	2.028	2.494	2.986	3.691	4.272
17	2.152	2.699	3.277	4.102	4.779	2.102	2.614	3.156	3.927	4.559	2.061	2.548	3.061	3.791	4.389	2.028	2.494	2.985	3.682	4.254	2.001	2.450	2.922	3.593	4.142
18	2.130	2.661	3.221	4.015	4.663	2.079	2.577	3.100	3.841	4.445	2.038	2.510	3.005	3.705	4.276	2.005	2.456	2.929	3.597	4.141	1.977	2.412	2.866	3.508	4.030
19	2.109	2.628	3.172	3.939	4.561	2.058	2.544	3.051	3.765	4.345	2.017	2.477	2.956	3.631	4.177	1.984	2.423	2.880	3.523	4.043	1.956	2.378	2.817	3.434	3.933
20	2.091	2.599	3.128	3.871	4.472	2.040	2.514	3.007	3.699	4.257	1.999	2.447	2.913	3.564	4.090	1.965	2.393	2.837	3.457	3.956	1.937	2.348	2.774	3.368	3.847
21	2.075	2.573	3.090	3.812	4.393	2.023	2.488	2.969	3.640	4.179	1.982	2.420	2.874	3.506	4.013	1.948	2.366	2.798	3.398	3.880	1.920	2.321	2.735	3.310	3.771
22	2.060	2.549	3.055	3.758	4.322	2.008	2.464	2.934	3.587	4.109	1.967	2.397	2.839	3.453	3.944	1.933	2.342	2.763	3.346	3.812	1.904	2.297	2.700	3.258	3.703
23	2.047	2.528	3.023	3.710	4.259	1.995	2.442	2.902	3.539	4.047	1.953	2.375	2.808	3.406	3.882	1.919	2.320	2.731	3.299	3.750	1.890	2.275	2.668	3.211	3.642
24	2.035	2.508	2.995	3.667	4.202	1.983	2.423	2.874	3.496	3.991	1.941	2.355	2.779	3.363	3.826	1.906	2.300	2.703	3.256	3.695	1.877	2.255	2.640	3.168	3.587
25	2.024	2.490	2.969	3.627	4.150	1.971	2.405	2.848	3.457	3.939	1.929	2.337	2.753	3.324	3.776	1.895	2.282	2.677	3.217	3.645	1.866	2.236	2.613	3.129	3.537
26	2.014	2.474	2.945	3.591	4.103	1.961	2.388	2.824	3.421	3.893	1.919	2.321	2.729	3.288	3.730	1.884	2.265	2.653	3.182	3.599	1.855	2.220	2.590	3.094	3.492
27	2.005	2.459	2.923	3.558	4.059	1.952	2.373	2.802	3.388	3.850	1.909	2.305	2.707	3.256	3.687	1.874	2.250	2.631	3.149	3.557	1.845	2.204	2.568	3.062	3.450
28	1.996	2.445	2.903	3.528	4.020	1.943	2.359	2.782	3.358	3.811	1.900	2.291	2.687	3.226	3.649	1.865	2.236	2.611	3.120	3.519	1.836	2.190	2.547	3.032	3.412
29	1.988	2.432	2.884	3.499	3.983	1.935	2.346	2.763	3.330	3.775	1.892	2.278	2.669	3.198	3.613	1.857	2.223	2.592	3.092	3.483	1.827	2.177	2.529	3.005	3.377
30	1.980	2.421	2.867	3.473	3.949	1.927	2.334	2.746	3.304	3.742	1.884	2.266	2.651	3.173	3.580	1.849	2.211	2.575	3.067	3.450	1.819	2.165	2.511	2.979	3.344
40	1.927	2.336	2.744	3.291	3.713	1.873	2.249	2.624	3.124	3.509	1.829	2.180	2.529	2.993	3.350	1.793	2.124	2.452	2.888	3.222	1.763	2.077	2.388	2.801	3.117
60	1.875	2.254	2.627	3.119	3.492	1.819	2.167	2.507	2.953	3.291	1.775	2.097	2.412	2.823	3.134	1.738	2.040	2.334	2.718	3.008	1.707	1.993	2.270	2.632	2.904
80	1.849	2.214	2.571	3.036	3.387	1.793	2.126	2.450	2.871	3.188	1.748	2.056	2.355	2.742	3.032	1.711	1.999	2.277	2.637	2.907	1.680	1.951	2.213	2.551	2.803
100	1.834	2.191	2.537	2.988	3.325	1.778	2.103	2.417	2.823	3.127	1.732	2.032	2.321	2.694	2.972	1.695	1.975	2.244	2.590	2.847	1.663	1.927	2.179	2.503	2.744
∞	1.774	2.099	2.408	2.802	3.091	1.717	2.010	2.288	2.639	2.897	1.670	1.938	2.192	2.511	2.744	1.632	1.880	2.114	2.407	2.621	1.599	1.831	2.048	2.321	2.519
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005
	α																								

	df_1 = 11					12					13					14					15				
1	60.47	243.0	973.0	6083	24334	60.71	243.9	976.7	6106	24426	60.90	244.7	979.8	6126	24505	61.07	245.4	982.5	6143	24572	61.22	245.9	984.9	6157	24630
2	9.401	19.40	39.41	99.41	199.4	9.408	19.41	39.41	99.42	199.4	9.415	19.42	39.42	99.42	199.4	9.420	19.42	39.43	99.43	199.4	9.425	19.43	39.43	99.43	199.4
3	5.222	8.763	14.37	27.13	43.52	5.216	8.745	14.34	27.05	43.39	5.210	8.729	14.30	26.98	43.27	5.205	8.715	14.28	26.92	43.17	5.200	8.703	14.25	26.87	43.08
4	3.907	5.936	8.794	14.45	20.82	3.896	5.912	8.751	14.37	20.70	3.886	5.891	8.715	14.31	20.60	3.878	5.873	8.684	14.25	20.51	3.870	5.858	8.657	14.20	20.44
5	3.282	4.704	6.568	9.963	13.49	3.268	4.678	6.525	9.888	13.38	3.257	4.655	6.488	9.825	13.29	3.247	4.636	6.456	9.770	13.21	3.238	4.619	6.428	9.722	13.15
6	2.920	4.027	5.410	7.790	10.13	2.905	4.000	5.366	7.718	10.03	2.892	3.976	5.329	7.657	9.950	2.881	3.956	5.297	7.605	9.877	2.871	3.938	5.269	7.559	9.814
7	2.684	3.603	4.709	6.538	8.270	2.668	3.575	4.666	6.469	8.176	2.654	3.550	4.628	6.410	8.097	2.643	3.529	4.596	6.359	8.028	2.632	3.511	4.568	6.314	7.968
8	2.519	3.313	4.243	5.734	7.104	2.502	3.284	4.200	5.667	7.015	2.488	3.259	4.162	5.609	6.938	2.475	3.237	4.130	5.559	6.872	2.464	3.218	4.101	5.515	6.814
9	2.396	3.102	3.912	5.178	6.314	2.379	3.073	3.868	5.111	6.227	2.364	3.048	3.831	5.055	6.153	2.351	3.025	3.798	5.005	6.089	2.340	3.006	3.769	4.962	6.032
10	2.302	2.943	3.665	4.772	5.746	2.284	2.913	3.621	4.706	5.661	2.269	2.887	3.583	4.650	5.589	2.255	2.865	3.550	4.601	5.526	2.244	2.845	3.522	4.558	5.471
11	2.227	2.818	3.474	4.462	5.320	2.209	2.788	3.430	4.397	5.236	2.193	2.761	3.392	4.342	5.165	2.179	2.739	3.359	4.293	5.103	2.167	2.719	3.330	4.251	5.049
12	2.166	2.717	3.321	4.220	4.988	2.147	2.687	3.277	4.155	4.906	2.131	2.660	3.239	4.100	4.836	2.117	2.637	3.206	4.052	4.775	2.105	2.617	3.177	4.010	4.721
13	2.116	2.635	3.197	4.025	4.724	2.097	2.604	3.153	3.960	4.643	2.080	2.577	3.115	3.905	4.573	2.066	2.554	3.082	3.857	4.513	2.053	2.533	3.053	3.815	4.460
14	2.073	2.565	3.095	3.864	4.508	2.054	2.534	3.050	3.800	4.428	2.037	2.507	3.012	3.745	4.359	2.022	2.484	2.979	3.698	4.299	2.010	2.463	2.949	3.656	4.247
15	2.037	2.507	3.008	3.730	4.329	2.017	2.475	2.963	3.666	4.250	2.000	2.448	2.925	3.612	4.181	1.985	2.424	2.891	3.564	4.122	1.972	2.403	2.862	3.522	4.070
16	2.005	2.456	2.934	3.616	4.179	1.985	2.425	2.889	3.553	4.099	1.968	2.397	2.851	3.498	4.031	1.953	2.373	2.817	3.451	3.972	1.940	2.352	2.788	3.409	3.920
17	1.978	2.413	2.870	3.519	4.050	1.958	2.381	2.825	3.455	3.971	1.940	2.353	2.786	3.401	3.903	1.925	2.329	2.753	3.353	3.844	1.912	2.308	2.723	3.312	3.793
18	1.954	2.374	2.814	3.434	3.938	1.933	2.342	2.769	3.371	3.860	1.916	2.314	2.730	3.316	3.793	1.900	2.290	2.696	3.269	3.734	1.887	2.269	2.667	3.227	3.683
19	1.932	2.340	2.765	3.360	3.841	1.912	2.308	2.720	3.297	3.763	1.894	2.280	2.681	3.242	3.696	1.878	2.256	2.647	3.195	3.638	1.865	2.234	2.617	3.153	3.587
20	1.913	2.310	2.721	3.294	3.756	1.892	2.278	2.676	3.231	3.678	1.875	2.250	2.637	3.177	3.611	1.859	2.225	2.603	3.130	3.553	1.845	2.203	2.573	3.088	3.502
21	1.896	2.283	2.682	3.236	3.680	1.875	2.250	2.637	3.173	3.602	1.857	2.222	2.598	3.119	3.536	1.841	2.197	2.564	3.072	3.478	1.827	2.176	2.534	3.030	3.427
22	1.880	2.259	2.647	3.184	3.612	1.859	2.226	2.602	3.121	3.535	1.841	2.198	2.563	3.067	3.469	1.825	2.173	2.528	3.019	3.411	1.811	2.151	2.498	2.978	3.360
23	1.866	2.236	2.615	3.137	3.551	1.845	2.204	2.570	3.074	3.475	1.827	2.175	2.531	3.020	3.408	1.811	2.150	2.497	2.973	3.351	1.796	2.128	2.466	2.931	3.300
24	1.853	2.216	2.586	3.094	3.497	1.832	2.183	2.541	3.032	3.420	1.814	2.155	2.502	2.977	3.354	1.797	2.130	2.468	2.930	3.296	1.783	2.108	2.437	2.889	3.246
25	1.841	2.198	2.560	3.056	3.447	1.820	2.165	2.515	2.993	3.370	1.802	2.136	2.476	2.939	3.304	1.785	2.111	2.441	2.892	3.247	1.771	2.089	2.411	2.850	3.196
26	1.830	2.181	2.536	3.021	3.402	1.809	2.148	2.491	2.958	3.325	1.790	2.119	2.451	2.904	3.259	1.774	2.094	2.417	2.857	3.202	1.760	2.072	2.387	2.815	3.151
27	1.820	2.166	2.514	2.988	3.360	1.799	2.132	2.469	2.926	3.284	1.780	2.103	2.429	2.871	3.218	1.764	2.078	2.395	2.824	3.161	1.749	2.056	2.364	2.783	3.110
28	1.811	2.151	2.494	2.959	3.322	1.790	2.118	2.448	2.896	3.246	1.771	2.089	2.409	2.842	3.180	1.754	2.064	2.374	2.795	3.123	1.740	2.041	2.344	2.753	3.073
29	1.802	2.138	2.475	2.931	3.287	1.781	2.104	2.430	2.868	3.211	1.762	2.075	2.390	2.814	3.145	1.745	2.050	2.355	2.767	3.088	1.731	2.027	2.325	2.726	3.038
30	1.794	2.126	2.458	2.906	3.255	1.773	2.092	2.412	2.843	3.179	1.754	2.063	2.372	2.789	3.113	1.737	2.037	2.338	2.742	3.056	1.722	2.015	2.307	2.700	3.006
40	1.737	2.038	2.334	2.727	3.028	1.715	2.003	2.288	2.665	2.953	1.695	1.974	2.248	2.611	2.888	1.678	1.948	2.213	2.563	2.831	1.662	1.924	2.182	2.522	2.781
60	1.680	1.952	2.216	2.559	2.817	1.657	1.917	2.169	2.496	2.742	1.637	1.887	2.129	2.442	2.677	1.619	1.860	2.093	2.394	2.620	1.603	1.836	2.061	2.352	2.570
80	1.653	1.910	2.158	2.478	2.716	1.629	1.875	2.111	2.415	2.641	1.609	1.845	2.071	2.361	2.577	1.590	1.817	2.035	2.313	2.520	1.574	1.793	2.003	2.271	2.470
100	1.636	1.886	2.124	2.430	2.657	1.612	1.850	2.077	2.368	2.583	1.592	1.819	2.036	2.313	2.518	1.573	1.792	2.000	2.265	2.461	1.557	1.768	1.968	2.223	2.411
∞	1.570	1.789	1.993	2.248	2.432	1.546	1.752	1.945	2.185	2.358	1.524	1.720	1.903	2.130	2.294	1.505	1.692	1.866	2.082	2.237	1.487	1.666	1.833	2.039	2.187
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005

df_2 (rows) $\quad$ α (columns)

df_1

df_2	16					17					18					19					20				
1	61.35	246.5	986.9	6170	24681	61.46	246.9	988.7	6181	24727	61.57	247.3	990.3	6192	24767	61.66	247.7	991.8	6201	24803	61.74	248.0	993.1	6209	24836
2	9.429	19.43	39.44	99.44	199.4	9.433	19.44	39.44	99.44	199.4	9.436	19.44	39.44	99.44	199.4	9.439	19.44	39.45	99.45	199.4	9.441	19.45	39.45	99.45	199.4
3	5.196	8.692	14.23	26.83	43.01	5.193	8.683	14.21	26.79	42.94	5.190	8.675	14.20	26.75	42.88	5.187	8.667	14.18	26.72	42.83	5.184	8.660	14.17	26.69	42.78
4	3.864	5.844	8.633	14.15	20.37	3.858	5.832	8.611	14.11	20.31	3.853	5.821	8.592	14.08	20.26	3.849	5.811	8.575	14.05	20.21	3.844	5.803	8.560	14.02	20.17
5	3.230	4.604	6.403	9.680	13.09	3.223	4.590	6.381	9.643	13.03	3.217	4.579	6.362	9.610	12.98	3.212	4.568	6.344	9.580	12.94	3.207	4.558	6.329	9.553	12.90
6	2.863	3.922	5.244	7.519	9.758	2.855	3.908	5.222	7.483	9.709	2.848	3.896	5.202	7.451	9.664	2.842	3.884	5.184	7.422	9.625	2.836	3.874	5.168	7.396	9.589
7	2.623	3.494	4.543	6.275	7.915	2.615	3.480	4.521	6.240	7.868	2.607	3.467	4.501	6.209	7.826	2.601	3.455	4.483	6.181	7.788	2.595	3.445	4.467	6.155	7.754
8	2.455	3.202	4.076	5.477	6.763	2.446	3.187	4.054	5.442	6.718	2.438	3.173	4.034	5.412	6.678	2.431	3.161	4.016	5.384	6.641	2.425	3.150	3.999	5.359	6.608
9	2.329	2.989	3.744	4.924	5.983	2.320	2.974	3.722	4.890	5.939	2.312	2.960	3.701	4.860	5.899	2.305	2.948	3.683	4.833	5.864	2.298	2.936	3.667	4.808	5.832
10	2.233	2.828	3.496	4.520	5.422	2.224	2.812	3.474	4.487	5.379	2.215	2.798	3.453	4.457	5.340	2.208	2.785	3.435	4.430	5.305	2.201	2.774	3.419	4.405	5.274
11	2.156	2.701	3.304	4.213	5.001	2.147	2.685	3.282	4.180	4.959	2.138	2.671	3.261	4.150	4.921	2.130	2.658	3.243	4.123	4.886	2.123	2.646	3.226	4.099	4.855
12	2.094	2.599	3.152	3.972	4.674	2.084	2.583	3.129	3.939	4.632	2.075	2.568	3.108	3.909	4.595	2.067	2.555	3.090	3.883	4.561	2.060	2.544	3.073	3.858	4.530
13	2.042	2.515	3.027	3.778	4.413	2.032	2.499	3.004	3.745	4.372	2.023	2.484	2.983	3.716	4.334	2.014	2.471	2.965	3.689	4.301	2.007	2.459	2.948	3.665	4.270
14	1.998	2.445	2.923	3.619	4.200	1.988	2.428	2.900	3.586	4.159	1.978	2.413	2.879	3.556	4.122	1.970	2.400	2.861	3.529	4.089	1.962	2.388	2.844	3.505	4.059
15	1.961	2.385	2.836	3.485	4.024	1.950	2.368	2.813	3.452	3.983	1.941	2.353	2.792	3.423	3.946	1.932	2.340	2.773	3.396	3.913	1.924	2.328	2.756	3.372	3.883
16	1.928	2.333	2.761	3.372	3.875	1.917	2.317	2.738	3.339	3.834	1.908	2.302	2.717	3.310	3.797	1.899	2.288	2.698	3.283	3.764	1.891	2.276	2.681	3.259	3.734
17	1.900	2.289	2.697	3.275	3.747	1.889	2.272	2.673	3.242	3.707	1.879	2.257	2.652	3.212	3.670	1.870	2.243	2.633	3.186	3.637	1.862	2.230	2.616	3.162	3.607
18	1.875	2.250	2.640	3.190	3.637	1.864	2.233	2.617	3.158	3.597	1.854	2.217	2.596	3.128	3.560	1.845	2.203	2.576	3.101	3.527	1.837	2.191	2.559	3.077	3.498
19	1.852	2.215	2.591	3.116	3.541	1.841	2.198	2.567	3.084	3.501	1.831	2.182	2.546	3.054	3.465	1.822	2.168	2.526	3.027	3.432	1.814	2.155	2.509	3.003	3.402
20	1.833	2.184	2.547	3.051	3.457	1.821	2.167	2.523	3.018	3.416	1.811	2.151	2.501	2.989	3.380	1.802	2.137	2.482	2.962	3.347	1.794	2.124	2.464	2.938	3.318
21	1.815	2.156	2.507	2.993	3.382	1.803	2.139	2.483	2.960	3.342	1.793	2.123	2.462	2.931	3.305	1.784	2.109	2.442	2.904	3.273	1.776	2.096	2.425	2.880	3.243
22	1.798	2.131	2.472	2.941	3.315	1.787	2.114	2.448	2.908	3.275	1.777	2.098	2.426	2.879	3.239	1.768	2.084	2.407	2.852	3.206	1.759	2.071	2.389	2.827	3.176
23	1.784	2.109	2.440	2.894	3.255	1.772	2.091	2.416	2.861	3.215	1.762	2.075	2.394	2.832	3.179	1.753	2.061	2.374	2.805	3.146	1.744	2.048	2.357	2.781	3.116
24	1.770	2.088	2.411	2.852	3.201	1.759	2.070	2.386	2.819	3.161	1.748	2.054	2.365	2.789	3.125	1.739	2.040	2.345	2.762	3.092	1.730	2.027	2.327	2.738	3.062
25	1.758	2.069	2.384	2.813	3.151	1.746	2.051	2.360	2.780	3.111	1.736	2.035	2.338	2.751	3.075	1.726	2.021	2.318	2.724	3.043	1.718	2.007	2.300	2.699	3.013
26	1.747	2.052	2.360	2.778	3.107	1.735	2.034	2.335	2.745	3.067	1.724	2.018	2.314	2.715	3.031	1.715	2.003	2.294	2.688	2.998	1.706	1.990	2.276	2.664	2.968
27	1.736	2.036	2.337	2.746	3.066	1.724	2.018	2.313	2.713	3.026	1.714	2.002	2.291	2.683	2.990	1.704	1.987	2.271	2.656	2.957	1.695	1.974	2.253	2.632	2.928
28	1.726	2.021	2.317	2.716	3.028	1.715	2.003	2.292	2.683	2.988	1.704	1.987	2.270	2.653	2.952	1.694	1.972	2.251	2.626	2.919	1.685	1.959	2.232	2.602	2.890
29	1.717	2.007	2.298	2.689	2.993	1.705	1.989	2.273	2.656	2.953	1.695	1.973	2.251	2.626	2.917	1.685	1.958	2.231	2.599	2.885	1.676	1.945	2.213	2.574	2.855
30	1.709	1.995	2.280	2.663	2.961	1.697	1.976	2.255	2.630	2.921	1.686	1.960	2.233	2.600	2.885	1.676	1.945	2.213	2.573	2.853	1.667	1.932	2.195	2.549	2.823
40	1.649	1.904	2.154	2.484	2.737	1.636	1.885	2.129	2.451	2.697	1.625	1.868	2.107	2.421	2.661	1.615	1.853	2.086	2.394	2.628	1.605	1.839	2.068	2.369	2.598
60	1.589	1.815	2.033	2.315	2.526	1.576	1.796	2.008	2.281	2.486	1.564	1.778	1.985	2.251	2.450	1.553	1.763	1.964	2.223	2.417	1.543	1.748	1.944	2.198	2.387
80	1.559	1.772	1.974	2.233	2.425	1.546	1.752	1.948	2.199	2.385	1.534	1.734	1.925	2.169	2.349	1.523	1.718	1.904	2.141	2.316	1.513	1.703	1.884	2.115	2.286
100	1.542	1.746	1.939	2.185	2.367	1.528	1.726	1.913	2.151	2.326	1.516	1.708	1.890	2.120	2.290	1.505	1.691	1.868	2.092	2.257	1.494	1.676	1.849	2.067	2.227
∞	1.471	1.644	1.803	2.000	2.142	1.457	1.623	1.776	1.965	2.101	1.444	1.604	1.751	1.934	2.064	1.432	1.587	1.729	1.905	2.031	1.421	1.571	1.708	1.878	2.000
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005

α

	df_1																									
	21					22					23					24					25				
1	61.81	248.3	994.3	6216	24866	61.88	248.6	995.4	6223	24892	61.95	248.8	996.3	6229	24917	62.00	249.1	997.2	6235	24940	62.05	249.3	998.1	6240	24960
2	9.444	19.45	39.45	99.45	199.5	9.446	19.45	39.45	99.45	199.5	9.448	19.45	39.45	99.46	199.5	9.450	19.45	39.46	99.46	199.5	9.451	19.46	39.46	99.46	199.5
3	5.182	8.654	14.16	26.66	42.73	5.180	8.648	14.14	26.64	42.69	5.178	8.643	14.13	26.62	42.66	5.176	8.639	14.12	26.60	42.62	5.175	8.634	14.12	26.58	42.59
4	3.841	5.795	8.546	13.99	20.13	3.837	5.787	8.533	13.97	20.09	3.834	5.781	8.522	13.95	20.06	3.831	5.774	8.511	13.93	20.03	3.828	5.769	8.501	13.91	20.00
5	3.202	4.549	6.314	9.528	12.87	3.198	4.541	6.301	9.506	12.84	3.194	4.534	6.289	9.485	12.81	3.191	4.527	6.278	9.466	12.78	3.187	4.521	6.268	9.449	12.76
6	2.831	3.865	5.154	7.372	9.556	2.827	3.856	5.141	7.351	9.526	2.822	3.849	5.128	7.331	9.499	2.818	3.841	5.117	7.313	9.474	2.815	3.835	5.107	7.296	9.451
7	2.589	3.435	4.452	6.132	7.723	2.584	3.426	4.439	6.111	7.695	2.580	3.418	4.426	6.092	7.669	2.575	3.410	4.415	6.074	7.645	2.571	3.404	4.405	6.058	7.623
8	2.419	3.140	3.985	5.336	6.578	2.413	3.131	3.971	5.316	6.551	2.409	3.123	3.959	5.297	6.526	2.404	3.115	3.947	5.279	6.503	2.400	3.108	3.937	5.263	6.482
9	2.292	2.926	3.652	4.786	5.803	2.287	2.917	3.638	4.765	5.776	2.282	2.908	3.626	4.746	5.752	2.277	2.900	3.614	4.729	5.729	2.272	2.893	3.604	4.713	5.708
10	2.194	2.764	3.403	4.383	5.245	2.189	2.754	3.390	4.363	5.219	2.183	2.745	3.377	4.344	5.195	2.178	2.737	3.365	4.327	5.173	2.174	2.730	3.355	4.311	5.153
11	2.117	2.636	3.211	4.077	4.827	2.111	2.626	3.197	4.057	4.801	2.105	2.617	3.184	4.038	4.778	2.100	2.609	3.173	4.021	4.756	2.095	2.601	3.162	4.005	4.736
12	2.053	2.533	3.057	3.836	4.502	2.047	2.523	3.043	3.816	4.476	2.041	2.514	3.031	3.798	4.453	2.036	2.505	3.019	3.780	4.431	2.031	2.498	3.008	3.765	4.412
13	2.000	2.448	2.932	3.643	4.243	1.994	2.438	2.918	3.622	4.217	1.988	2.429	2.905	3.604	4.194	1.983	2.420	2.893	3.587	4.173	1.978	2.412	2.882	3.571	4.153
14	1.955	2.377	2.828	3.483	4.031	1.949	2.367	2.814	3.463	4.006	1.943	2.357	2.801	3.444	3.983	1.938	2.349	2.789	3.427	3.961	1.933	2.341	2.778	3.412	3.942
15	1.917	2.316	2.740	3.350	3.855	1.911	2.306	2.726	3.330	3.830	1.905	2.297	2.713	3.311	3.807	1.899	2.288	2.701	3.294	3.786	1.894	2.280	2.689	3.278	3.766
16	1.884	2.264	2.665	3.237	3.707	1.877	2.254	2.651	3.216	3.682	1.871	2.244	2.637	3.198	3.659	1.866	2.235	2.625	3.181	3.638	1.860	2.227	2.614	3.165	3.618
17	1.855	2.219	2.600	3.139	3.580	1.848	2.208	2.585	3.119	3.555	1.842	2.199	2.572	3.101	3.532	1.836	2.190	2.560	3.084	3.511	1.831	2.181	2.548	3.068	3.492
18	1.829	2.179	2.543	3.055	3.471	1.823	2.168	2.529	3.035	3.446	1.816	2.159	2.515	3.016	3.423	1.810	2.150	2.503	2.999	3.402	1.805	2.141	2.491	2.983	3.382
19	1.807	2.144	2.493	2.981	3.375	1.800	2.133	2.478	2.961	3.350	1.793	2.123	2.465	2.942	3.327	1.787	2.114	2.452	2.925	3.306	1.782	2.106	2.441	2.909	3.287
20	1.786	2.112	2.448	2.916	3.291	1.779	2.102	2.434	2.895	3.266	1.773	2.092	2.420	2.877	3.243	1.767	2.082	2.408	2.859	3.222	1.761	2.074	2.396	2.843	3.203
21	1.768	2.084	2.409	2.857	3.216	1.761	2.073	2.394	2.837	3.191	1.754	2.063	2.380	2.818	3.168	1.748	2.054	2.368	2.801	3.147	1.742	2.045	2.356	2.785	3.128
22	1.751	2.059	2.373	2.805	3.149	1.744	2.048	2.358	2.785	3.125	1.737	2.038	2.344	2.766	3.102	1.731	2.028	2.331	2.749	3.081	1.726	2.020	2.320	2.733	3.061
23	1.736	2.036	2.340	2.758	3.089	1.729	2.025	2.325	2.738	3.065	1.722	2.014	2.312	2.719	3.042	1.716	2.005	2.299	2.702	3.021	1.710	1.996	2.287	2.686	3.001
24	1.722	2.015	2.311	2.716	3.035	1.715	2.003	2.296	2.695	3.011	1.708	1.993	2.282	2.676	2.988	1.702	1.984	2.269	2.659	2.967	1.696	1.975	2.257	2.643	2.947
25	1.710	1.995	2.284	2.677	2.986	1.702	1.984	2.269	2.657	2.961	1.695	1.974	2.255	2.638	2.939	1.689	1.964	2.242	2.620	2.918	1.683	1.955	2.230	2.604	2.898
26	1.698	1.978	2.259	2.642	2.941	1.690	1.966	2.244	2.621	2.917	1.683	1.956	2.230	2.602	2.894	1.677	1.946	2.217	2.585	2.873	1.671	1.938	2.205	2.569	2.853
27	1.687	1.961	2.237	2.609	2.900	1.680	1.950	2.222	2.589	2.876	1.673	1.940	2.208	2.570	2.853	1.666	1.930	2.195	2.552	2.832	1.660	1.921	2.183	2.536	2.812
28	1.677	1.946	2.216	2.579	2.863	1.669	1.935	2.201	2.559	2.838	1.662	1.924	2.187	2.540	2.815	1.656	1.915	2.174	2.522	2.794	1.650	1.906	2.161	2.506	2.775
29	1.668	1.932	2.196	2.552	2.828	1.660	1.921	2.181	2.531	2.803	1.653	1.910	2.167	2.512	2.780	1.647	1.901	2.154	2.495	2.759	1.640	1.891	2.142	2.478	2.740
30	1.659	1.919	2.178	2.526	2.796	1.651	1.908	2.163	2.506	2.771	1.644	1.897	2.149	2.487	2.748	1.638	1.887	2.136	2.469	2.727	1.632	1.878	2.124	2.453	2.708
40	1.596	1.826	2.051	2.346	2.571	1.588	1.814	2.035	2.325	2.546	1.581	1.803	2.020	2.306	2.523	1.574	1.793	2.007	2.288	2.502	1.568	1.783	1.994	2.271	2.482
60	1.534	1.735	1.927	2.175	2.360	1.526	1.722	1.911	2.153	2.335	1.518	1.711	1.896	2.134	2.311	1.511	1.700	1.882	2.115	2.290	1.504	1.690	1.869	2.098	2.270
80	1.503	1.689	1.866	2.092	2.259	1.495	1.677	1.850	2.070	2.233	1.487	1.665	1.835	2.050	2.210	1.479	1.654	1.820	2.032	2.188	1.472	1.644	1.807	2.015	2.168
100	1.485	1.663	1.830	2.043	2.199	1.476	1.650	1.814	2.021	2.174	1.468	1.638	1.798	2.001	2.150	1.460	1.627	1.784	1.983	2.128	1.453	1.616	1.770	1.965	2.108
∞	1.410	1.556	1.689	1.854	1.971	1.401	1.542	1.672	1.831	1.945	1.392	1.529	1.655	1.810	1.921	1.383	1.517	1.640	1.791	1.898	1.375	1.506	1.626	1.773	1.877
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005
	α																								

df_2

df_2	df_1 = 26					27					28					29					30				
1	62.10	249.5	998.8	6245	24980	62.15	249.6	999.6	6249	24997	62.19	249.8	1000	6253	25014	62.23	250.0	1001	6257	25029	62.26	250.1	1001	6261	25044
2	9.453	19.46	39.46	99.46	199.5	9.454	19.46	39.46	99.46	199.5	9.456	19.46	39.46	99.46	199.5	9.457	19.46	39.46	99.46	199.5	9.458	19.46	39.46	99.47	199.5
3	5.173	8.630	14.11	26.56	42.56	5.172	8.626	14.10	26.55	42.54	5.170	8.623	14.09	26.53	42.51	5.169	8.620	14.09	26.52	42.49	5.168	8.617	14.08	26.50	42.47
4	3.826	5.763	8.492	13.89	19.98	3.823	5.759	8.483	13.88	19.95	3.821	5.754	8.476	13.86	19.93	3.819	5.750	8.468	13.85	19.91	3.817	5.746	8.461	13.84	19.89
5	3.184	4.515	6.258	9.433	12.73	3.181	4.510	6.250	9.418	12.71	3.179	4.505	6.242	9.404	12.69	3.176	4.500	6.234	9.391	12.67	3.174	4.496	6.227	9.379	12.66
6	2.811	3.829	5.097	7.280	9.430	2.808	3.823	5.088	7.266	9.410	2.805	3.818	5.080	7.253	9.392	2.803	3.813	5.072	7.240	9.374	2.800	3.808	5.065	7.229	9.358
7	2.568	3.397	4.395	6.043	7.603	2.564	3.391	4.386	6.029	7.584	2.561	3.386	4.378	6.016	7.566	2.558	3.381	4.370	6.003	7.550	2.555	3.376	4.362	5.992	7.534
8	2.396	3.102	3.927	5.248	6.462	2.392	3.095	3.918	5.234	6.444	2.389	3.090	3.909	5.221	6.427	2.386	3.084	3.901	5.209	6.411	2.383	3.079	3.894	5.198	6.396
9	2.268	2.886	3.594	4.698	5.689	2.265	2.880	3.584	4.685	5.671	2.261	2.874	3.576	4.672	5.655	2.258	2.869	3.568	4.660	5.639	2.255	2.864	3.560	4.649	5.625
10	2.170	2.723	3.345	4.296	5.134	2.166	2.716	3.335	4.283	5.116	2.162	2.710	3.327	4.270	5.100	2.159	2.705	3.319	4.258	5.085	2.155	2.700	3.311	4.247	5.071
11	2.091	2.594	3.152	3.990	4.717	2.087	2.588	3.142	3.977	4.700	2.083	2.582	3.133	3.964	4.684	2.080	2.576	3.125	3.952	4.668	2.076	2.570	3.118	3.941	4.654
12	2.027	2.491	2.998	3.750	4.393	2.022	2.484	2.988	3.736	4.376	2.019	2.478	2.979	3.724	4.360	2.015	2.472	2.971	3.712	4.345	2.011	2.466	2.963	3.701	4.331
13	1.973	2.405	2.872	3.556	4.134	1.969	2.398	2.862	3.543	4.117	1.965	2.392	2.853	3.530	4.101	1.961	2.386	2.845	3.518	4.087	1.958	2.380	2.837	3.507	4.073
14	1.928	2.333	2.767	3.397	3.923	1.923	2.326	2.758	3.383	3.906	1.919	2.320	2.749	3.371	3.891	1.916	2.314	2.740	3.359	3.876	1.912	2.308	2.732	3.348	3.862
15	1.889	2.272	2.679	3.264	3.748	1.885	2.265	2.669	3.250	3.731	1.880	2.259	2.660	3.237	3.715	1.876	2.253	2.652	3.225	3.701	1.873	2.247	2.644	3.214	3.687
16	1.855	2.220	2.603	3.150	3.600	1.851	2.212	2.594	3.137	3.583	1.847	2.206	2.584	3.124	3.567	1.843	2.200	2.576	3.112	3.553	1.839	2.194	2.568	3.101	3.539
17	1.826	2.174	2.538	3.053	3.473	1.821	2.167	2.528	3.039	3.457	1.817	2.160	2.519	3.026	3.441	1.813	2.154	2.510	3.014	3.426	1.809	2.148	2.502	3.003	3.412
18	1.800	2.134	2.481	2.968	3.364	1.795	2.126	2.471	2.955	3.347	1.791	2.119	2.461	2.942	3.332	1.787	2.113	2.453	2.930	3.317	1.783	2.107	2.445	2.919	3.303
19	1.777	2.098	2.430	2.894	3.269	1.772	2.090	2.420	2.880	3.252	1.767	2.084	2.411	2.868	3.236	1.763	2.077	2.402	2.855	3.221	1.759	2.071	2.394	2.844	3.208
20	1.756	2.066	2.385	2.829	3.184	1.751	2.059	2.375	2.815	3.168	1.746	2.052	2.366	2.802	3.152	1.742	2.045	2.357	2.790	3.137	1.738	2.039	2.349	2.778	3.123
21	1.737	2.037	2.345	2.770	3.110	1.732	2.030	2.335	2.756	3.093	1.728	2.023	2.325	2.743	3.077	1.723	2.016	2.317	2.731	3.063	1.719	2.010	2.308	2.720	3.049
22	1.720	2.012	2.309	2.718	3.043	1.715	2.004	2.299	2.704	3.026	1.711	1.997	2.289	2.691	3.011	1.706	1.990	2.280	2.679	2.996	1.702	1.984	2.272	2.667	2.982
23	1.705	1.988	2.276	2.671	2.983	1.700	1.981	2.266	2.657	2.966	1.695	1.973	2.256	2.644	2.951	1.691	1.967	2.247	2.632	2.936	1.686	1.961	2.239	2.620	2.922
24	1.691	1.967	2.246	2.628	2.929	1.686	1.959	2.236	2.614	2.912	1.681	1.952	2.226	2.601	2.897	1.676	1.945	2.217	2.589	2.882	1.672	1.939	2.209	2.577	2.868
25	1.678	1.947	2.219	2.589	2.880	1.672	1.939	2.209	2.575	2.863	1.668	1.932	2.199	2.562	2.847	1.663	1.926	2.190	2.550	2.833	1.659	1.919	2.182	2.538	2.819
26	1.666	1.929	2.194	2.554	2.835	1.660	1.921	2.184	2.540	2.818	1.656	1.914	2.174	2.526	2.802	1.651	1.907	2.165	2.514	2.788	1.647	1.901	2.157	2.503	2.774
27	1.655	1.913	2.171	2.521	2.794	1.649	1.905	2.161	2.507	2.777	1.645	1.898	2.151	2.494	2.761	1.640	1.891	2.142	2.481	2.747	1.636	1.884	2.133	2.470	2.733
28	1.644	1.897	2.150	2.491	2.756	1.639	1.889	2.140	2.477	2.739	1.634	1.882	2.130	2.464	2.724	1.630	1.875	2.121	2.451	2.709	1.625	1.869	2.112	2.440	2.695
29	1.635	1.883	2.131	2.463	2.722	1.630	1.875	2.120	2.449	2.705	1.625	1.868	2.110	2.436	2.689	1.620	1.861	2.101	2.423	2.674	1.616	1.854	2.092	2.412	2.660
30	1.626	1.870	2.112	2.437	2.689	1.621	1.862	2.102	2.423	2.672	1.616	1.854	2.092	2.410	2.657	1.611	1.847	2.083	2.398	2.642	1.606	1.841	2.074	2.386	2.628
40	1.562	1.775	1.983	2.256	2.464	1.556	1.766	1.972	2.241	2.447	1.551	1.759	1.962	2.228	2.431	1.546	1.751	1.952	2.215	2.416	1.541	1.744	1.943	2.203	2.401
60	1.498	1.681	1.857	2.083	2.251	1.492	1.672	1.845	2.068	2.234	1.486	1.664	1.835	2.054	2.217	1.481	1.656	1.825	2.041	2.202	1.476	1.649	1.815	2.028	2.187
80	1.465	1.634	1.795	1.999	2.149	1.459	1.626	1.783	1.983	2.131	1.453	1.617	1.772	1.969	2.115	1.448	1.609	1.762	1.956	2.099	1.443	1.602	1.752	1.944	2.084
100	1.446	1.607	1.758	1.949	2.089	1.440	1.598	1.746	1.934	2.071	1.434	1.589	1.735	1.919	2.054	1.428	1.581	1.725	1.906	2.039	1.423	1.573	1.715	1.893	2.024
∞	1.368	1.496	1.612	1.755	1.857	1.361	1.486	1.600	1.739	1.839	1.354	1.476	1.588	1.724	1.821	1.348	1.467	1.577	1.710	1.805	1.342	1.459	1.566	1.696	1.789
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005
	α																								

df_2 \ df_1	40					60					80					100					∞				
1	62.53	251.1	1006	6287	25148	62.79	252.2	1010	6313	25253	62.93	252.7	1012	6326	25306	63.01	253.0	1013	6334	25337	63.33	254.3	1018	6366	25464
2	9.466	19.47	39.47	99.47	199.5	9.475	19.48	39.48	99.48	199.5	9.479	19.48	39.49	99.49	199.5	9.481	19.49	39.49	99.49	199.5	9.491	19.50	39.50	99.50	199.5
3	5.160	8.594	14.04	26.41	42.31	5.151	8.572	13.99	26.32	42.15	5.147	8.561	13.97	26.27	42.07	5.144	8.554	13.96	26.24	42.02	5.134	8.526	13.90	26.13	41.83
4	3.804	5.717	8.411	13.75	19.75	3.790	5.688	8.360	13.65	19.61	3.782	5.673	8.335	13.61	19.54	3.778	5.664	8.319	13.58	19.50	3.761	5.628	8.257	13.46	19.32
5	3.157	4.464	6.175	9.291	12.53	3.140	4.431	6.123	9.202	12.40	3.132	4.415	6.096	9.157	12.34	3.126	4.405	6.080	9.130	12.30	3.105	4.365	6.015	9.020	12.14
6	2.781	3.774	5.012	7.143	9.241	2.762	3.740	4.959	7.057	9.122	2.752	3.722	4.932	7.013	9.062	2.746	3.712	4.915	6.987	9.026	2.722	3.669	4.849	6.880	8.879
7	2.535	3.340	4.309	5.908	7.422	2.514	3.304	4.254	5.824	7.309	2.504	3.286	4.227	5.781	7.251	2.497	3.275	4.210	5.755	7.217	2.471	3.230	4.142	5.650	7.076
8	2.361	3.043	3.840	5.116	6.288	2.339	3.005	3.784	5.032	6.177	2.328	2.986	3.756	4.989	6.121	2.321	2.975	3.739	4.963	6.088	2.293	2.928	3.670	4.859	5.951
9	2.232	2.826	3.505	4.567	5.519	2.208	2.787	3.449	4.483	5.410	2.196	2.768	3.421	4.441	5.356	2.189	2.756	3.403	4.415	5.322	2.159	2.707	3.333	4.311	5.188
10	2.132	2.661	3.255	4.165	4.966	2.107	2.621	3.198	4.082	4.859	2.095	2.601	3.169	4.039	4.805	2.087	2.588	3.152	4.014	4.772	2.055	2.538	3.080	3.909	4.639
11	2.052	2.531	3.061	3.860	4.551	2.026	2.490	3.004	3.776	4.445	2.013	2.469	2.974	3.734	4.391	2.005	2.457	2.956	3.708	4.359	1.972	2.404	2.883	3.602	4.226
12	1.986	2.426	2.906	3.619	4.228	1.960	2.384	2.848	3.535	4.123	1.946	2.363	2.818	3.493	4.069	1.938	2.350	2.800	3.467	4.037	1.904	2.296	2.725	3.361	3.904
13	1.931	2.339	2.780	3.425	3.970	1.904	2.297	2.720	3.341	3.866	1.890	2.275	2.690	3.298	3.812	1.882	2.261	2.671	3.272	3.780	1.846	2.206	2.595	3.165	3.647
14	1.885	2.266	2.674	3.266	3.760	1.857	2.223	2.614	3.181	3.655	1.843	2.201	2.583	3.138	3.602	1.834	2.187	2.565	3.112	3.569	1.797	2.131	2.487	3.004	3.436
15	1.845	2.204	2.585	3.132	3.585	1.817	2.160	2.524	3.047	3.480	1.802	2.137	2.493	3.004	3.427	1.793	2.123	2.474	2.977	3.394	1.755	2.066	2.395	2.868	3.260
16	1.811	2.151	2.509	3.018	3.437	1.782	2.106	2.447	2.933	3.332	1.766	2.083	2.415	2.889	3.279	1.757	2.068	2.396	2.863	3.246	1.718	2.010	2.316	2.753	3.112
17	1.781	2.104	2.442	2.920	3.311	1.751	2.058	2.380	2.835	3.206	1.735	2.035	2.348	2.791	3.152	1.726	2.020	2.329	2.764	3.119	1.686	1.960	2.247	2.653	2.984
18	1.754	2.063	2.384	2.835	3.201	1.723	2.017	2.321	2.749	3.096	1.707	1.993	2.289	2.705	3.042	1.698	1.978	2.269	2.678	3.009	1.657	1.917	2.187	2.566	2.873
19	1.730	2.026	2.333	2.761	3.106	1.699	1.980	2.270	2.674	3.000	1.683	1.955	2.237	2.630	2.946	1.673	1.940	2.217	2.602	2.913	1.631	1.878	2.133	2.489	2.776
20	1.708	1.994	2.287	2.695	3.022	1.677	1.946	2.223	2.608	2.916	1.660	1.922	2.190	2.563	2.861	1.650	1.907	2.170	2.535	2.828	1.607	1.843	2.085	2.421	2.690
21	1.689	1.965	2.246	2.636	2.947	1.657	1.916	2.182	2.548	2.841	1.640	1.891	2.148	2.503	2.786	1.630	1.876	2.128	2.475	2.753	1.586	1.812	2.042	2.360	2.614
22	1.671	1.938	2.210	2.583	2.880	1.639	1.889	2.145	2.495	2.774	1.622	1.864	2.111	2.450	2.719	1.611	1.849	2.090	2.422	2.685	1.567	1.783	2.003	2.305	2.545
23	1.655	1.914	2.176	2.535	2.820	1.622	1.865	2.111	2.447	2.713	1.605	1.839	2.077	2.401	2.658	1.594	1.823	2.056	2.373	2.624	1.549	1.757	1.968	2.256	2.484
24	1.641	1.892	2.146	2.492	2.765	1.607	1.842	2.080	2.403	2.658	1.590	1.816	2.045	2.357	2.603	1.579	1.800	2.024	2.329	2.569	1.533	1.733	1.935	2.211	2.428
25	1.627	1.872	2.118	2.453	2.716	1.593	1.822	2.052	2.364	2.609	1.576	1.796	2.017	2.317	2.553	1.565	1.779	1.996	2.289	2.519	1.518	1.711	1.906	2.169	2.377
26	1.615	1.853	2.093	2.417	2.671	1.581	1.803	2.026	2.327	2.563	1.562	1.776	1.991	2.281	2.508	1.551	1.760	1.969	2.252	2.473	1.504	1.691	1.878	2.131	2.330
27	1.603	1.836	2.069	2.384	2.630	1.569	1.785	2.002	2.294	2.522	1.550	1.758	1.966	2.247	2.466	1.539	1.742	1.945	2.218	2.431	1.491	1.672	1.853	2.097	2.287
28	1.592	1.820	2.048	2.354	2.592	1.558	1.769	1.980	2.263	2.483	1.539	1.742	1.944	2.216	2.427	1.528	1.725	1.922	2.187	2.392	1.478	1.654	1.829	2.064	2.247
29	1.583	1.806	2.028	2.325	2.557	1.547	1.754	1.959	2.234	2.448	1.529	1.726	1.923	2.187	2.391	1.517	1.710	1.901	2.158	2.357	1.467	1.638	1.807	2.034	2.210
30	1.573	1.792	2.009	2.299	2.524	1.538	1.740	1.940	2.208	2.415	1.519	1.712	1.904	2.160	2.358	1.507	1.695	1.882	2.131	2.323	1.456	1.622	1.787	2.006	2.176
40	1.506	1.693	1.875	2.114	2.296	1.467	1.637	1.803	2.019	2.184	1.447	1.608	1.764	1.969	2.125	1.434	1.589	1.741	1.938	2.088	1.377	1.509	1.637	1.805	1.932
60	1.437	1.594	1.744	1.936	2.079	1.395	1.534	1.667	1.836	1.962	1.372	1.502	1.625	1.783	1.900	1.358	1.481	1.599	1.749	1.861	1.291	1.389	1.482	1.601	1.689
80	1.403	1.545	1.679	1.849	1.974	1.358	1.482	1.599	1.746	1.854	1.334	1.448	1.555	1.690	1.789	1.318	1.426	1.527	1.655	1.748	1.245	1.325	1.400	1.494	1.563
100	1.382	1.515	1.640	1.797	1.912	1.336	1.450	1.558	1.692	1.790	1.310	1.415	1.512	1.634	1.723	1.293	1.392	1.483	1.598	1.681	1.214	1.283	1.347	1.427	1.485
∞	1.295	1.394	1.484	1.592	1.669	1.240	1.318	1.388	1.473	1.533	1.207	1.273	1.333	1.404	1.454	1.185	1.243	1.296	1.358	1.402	1.000	1.000	1.000	1.000	1.000
	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005	0.100	0.050	0.025	0.010	0.005
	α																								

Answers

page 51:

1. See individual case study answers.
2. $\mu = 2.6$ points, $\sigma = 2.2$ points, $\sigma^2 = 4.9$ points^2

page 67: Case study 1:

(a) The population is used copies of the particular textbook available online.

(b) sample data

(c) $x =$ book price

(d) quantitative

(e) discrete

(f) ratio

(g) raw data

(h) $n = 7$

(i) $\sum x = \$149.60$, no

(j) Two modes, \$7.50 and \$50.00

(k) $\bar{x} = \$21.37$

(l) median = \$12.00

(m) a.d. = \$16.36

(n) $s = \$19.98$, $s^2 = 399.02$ ($\$^2$)

(o) $s = \$19.98$, $s^2 = 399.02$ ($\$^2$)

(p) $z = -0.72$ Price is .72 standard deviations (slightly) below the mean.

(q) $Q_1 = \$7.50$, $Q_3 = \$50.00$, $IQR = \$42.50$

page 70: Case study 2:

(a) Population is the math exams taken by the student for his math course.

(b) population data

(c) $x =$ score on the exam

(d) quantitative

(e) discrete (assuming no fractional percents given)

(f) ratio

(g) raw data

(h) $N = 6$

(i) $\sum x = 452\%$, no

(j) mode = 70.0%

(k) $\mu = 75.3\%$

(l) median = 74.5%

(m) a.d. = 7.0%

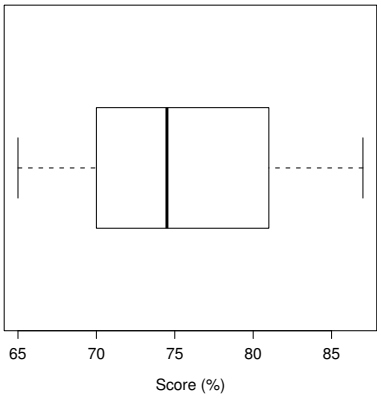
(n) $\sigma = 7.6\%$, $\sigma^2 = 57.6$ ($\%^2$)

(o) $\sigma = 7.6\%$, $\sigma^2 = 57.6$ ($\%^2$)

(p) at least 41% of data between 65.4% and 85.2%, $\frac{4}{6} = 67\% > 41\%$

(q) $D_7 = 80.8\%$

(r) 65, 68.8, 74.5, 82.5, 87 (%)
lower fence = 48.2 %, upper fence = 103.1 %
$\Rightarrow$ no outliers



page 74: Case study 3:

(a) Population is the rolls of two dice.

(b) sample data

(c) $x =$ sum of the two dice

(d) quantitative

(e) discrete

(f) ratio

(g) ungrouped frequency distribution

(h) f is the *Occurrences* column.

(i) $\sum f = 50$

(j) $\sum xf = 200$, no

(k) Add P . $\sum P = 1.00$

(l) Add $<Cf$, $<CP$.

(m) mode = 4.0

(n) $\bar{x} = 4.0$

(o) median = 4.0

(p) $s = 1.2$, $s^2 = 1.3$

(q) $z = 1.67$ Roll is 1.67 standard deviations (moderately) above the mean.

(r) $P_{10} = 2.1$, $P_{90} = 6.0$, 10-90 $PR = 3.9$

page 77: Case study 4:

(a) Population is all students at this particular preschool in the last year.

(b) population data

(c) $x =$ student age

(d) quantitative

(e) discrete (since human age means number of birthdays)

(f) ratio

(g) ungrouped frequency distribution

(h) f is the *# of children* column.

(i) $\sum f = 20$

(j) $\sum xf = 58$ years, no

(k) Add P . $\sum P = 1.00$

(l) Add $<Cf$, $<CP$.

(m) mode = 2.0 years

(n) $\mu = 2.9$ years

(o) median = 2.5 years

(p) $\sigma = 1.4$ years, $\sigma^2 = 1.9$ (years^2)

(q) at least 75% of data between 0.1 yr and 5.7 yr, $\frac{18}{20} = 90\% > 75\%$

(r) $P_{66} = 3.0$ yr

page 80: Case study 5:

(a) Population is items sold on *cheapjunk.com* over the previous year.

(b) population data

(c) x = selling price

(d) quantitative

(e) discrete (since price is to nearest penny)

(f) ratio

(g) grouped frequency distribution

(h) f is the *Purchases* column.

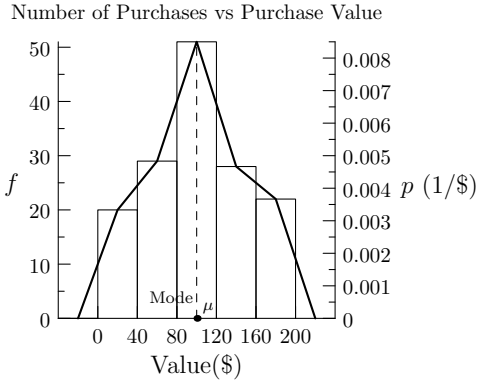
(i) $\sum f = 150$

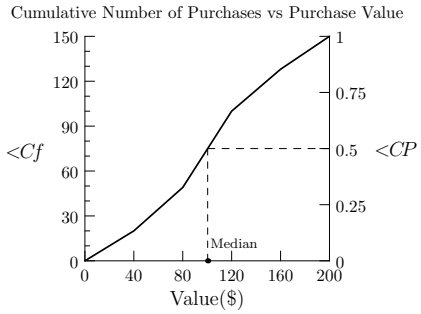
(j) $\sum xf = \$15120.00$, yes

(k) Add P, p . $\sum P = 1.000$

(l) Add $<Cf$, $<CP$.

(m)





(n) symmetric

(o) From histogram, mode $\approx \$100.00$, $\mu \approx \$100.00$. From ogive, median $\approx \$100.00$

(p) mode = \$100.00

(q) $\mu = \$100.80$

(r) median = \$100.39

(s) $\sigma = \$48.98$, $\sigma^2 = 2399.36$ (\$^{2})

(t) $z = 0.00$ Purchase price equals the mean.

(u) $D_1 = \$30.00$, $D_9 = \$172.73$, $IDR = \$142.73$

page 83: Case study 6:

(a) Population is wind near Swift Current on a given day.

(b) sample data

(c) x = peak wind speed

(d) quantitative

(e) continuous

(f) ratio

(g) grouped frequency distribution

(h) f is the *Days* column.

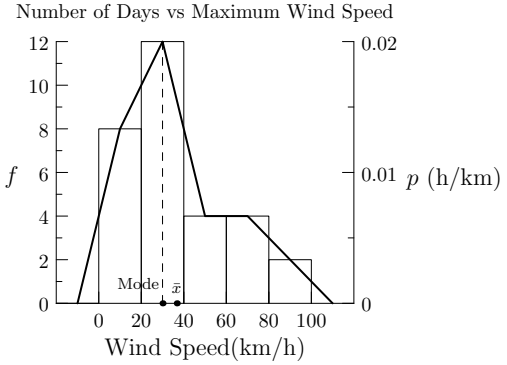
(i) $\sum f = 30$

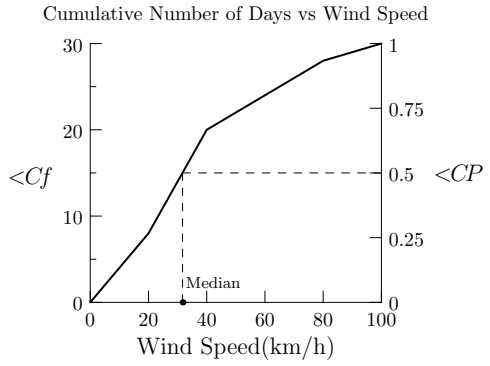
(j) $\sum xf = 1100.0$ km/h, yes

(k) Add P, p . $\sum P = 1.000$

(l) Add $<Cf$, $<CP$.

(m)





(n) positively skewed

(o) From histogram, mode $\approx$ 30 km/h, $\mu \approx$ 36 km/h. From ogive, median $\approx$ 32 km/h

(p) mode = 30.0 km/h

(q) $\bar{x} = 36.7$ km/h

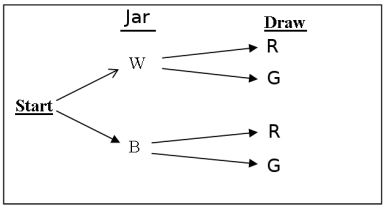
(r) median = 31.7 km/h

(s) $s = 24.3$ km/h, $s^2 = 588.5$ km^2/h^2

(t) $z = 4.66$ Wind speed is 4.66 standard deviations (extremely) above the mean.

(u) $Q_3 = 52.5$ km/h

page 95:

1. 

2. 1710

3. 6

4. 15

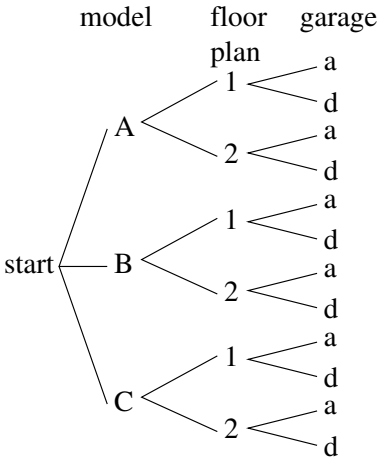
5. (a) ${}_{20}C_4 = 4845$

   (b) ${}_{20}P_4 = 116,280$

   (c) $20^4 = 160,000$

6.

	Gas Station Operators	Fast Food Outlets	Totals
For Bylaw	10	20	30
Against Bylaw	50	20	70
Totals	60	40	100

7. 

   The tree has 12 paths.

8. 13,983,816

9. 404,550

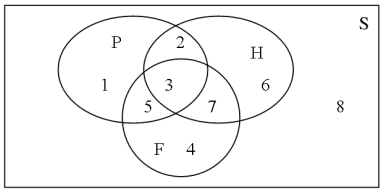
page 101:

1. (a) $n(Sx) = 4$

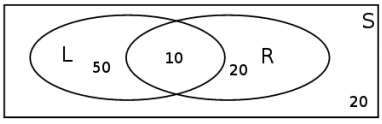
   (b) $n(\text{-}Sx) = 48$

   (c) $n(Sp \text{ and } Sx) = 1$

   (d) $n(Sp \text{ or } Sx) = 16$

2. (a) 

   (b) {1,4,5,8}, {3,5}, {2,3,4,5,6,7}, {8}, {4,5}, {3}

3. (a) 

   (b) 80

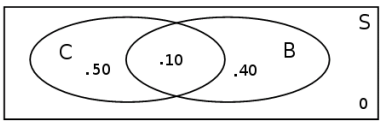
   (c) 20

   (d) 50

   (e) $\frac{10}{30} = \frac{1}{3} = 0.3333$

page 105:

1. (a) 0.075 (b) 0.925

2. (a) 

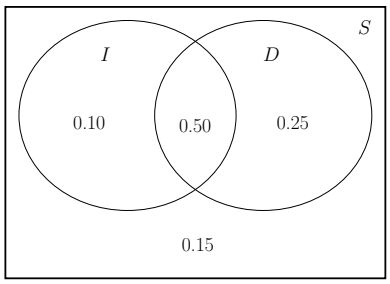
   (b) no since $P(C \text{ and } B) \neq 0$

   (c) 1.00

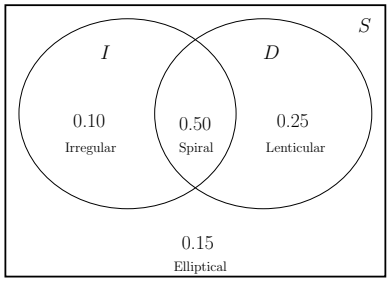
   (d) 0.00

   (e) 0.5

3. $\frac{1}{13} = 0.0769$
4. (a)



(b)



(c)
- $P(\text{Spiral}) = P(D \text{ and } I) = 0.50$
- $P(\text{Lenticular}) = P(D \text{ and } \text{-}I) = 0.25$
- $P(\text{Irregular}) = P(\text{-}D \text{ and } I) = 0.10$
- $P(\text{Elliptical}) = P(\text{-}D \text{ and } \text{-}I) = 0.15$

page 112:

1. (a) independent
   (b) dependent
   (c) Effectively independent since, even selecting without replacement, the probability of a preference will not change substantially due to the large population size. This will be discussed further in Section 3.1.2 .
2. (a) $\frac{1}{20} = 0.05$ (b) $\frac{57}{460} = 0.1239$
3. (a) $\frac{1}{26} = 0.0385$ (b) yes
4. (a) $\frac{1}{16} = 0.0625$ (b) $\frac{1}{19} = 0.0526$
5. $\frac{113}{114} = 0.9912$

page 117:

1. (a) $\frac{101}{110} = 0.9182$ (b) $\frac{18}{55} = 0.3273$
   (c) $\frac{1}{22} = 0.0455$ (d) $\frac{1}{10} = 0.1$
   (e) $\frac{65}{101} = 0.6436$
2. (a) 0.05 (b) 0.30 (c) 0.65 (d) $\frac{1}{7} = 0.1429$
3. (a) $\frac{1}{4} = 0.25$ (b) $\frac{9}{16} = 0.5625$ (c) $\frac{1}{9} = 0.1111$

page 118:

1. (a) $P(A) = \frac{7}{15} = 0.4667$
   (b) $P(\text{-}A) = \frac{8}{15} = 0.5333$
   (c) $P(A \text{ and } L) = \frac{1}{10} = 0.1$
   (d) $P(A|L) = \frac{2}{7} = 0.2857$
   (e) $P(\text{-}A \text{ and } \text{-}L) = \frac{17}{60} = 0.2833$
   (f) $P(L|A) = \frac{3}{14} = 0.2143$
   (g) $P(A|\text{-}L) = \frac{22}{39} = 0.5641$
2. $\frac{7}{13} = 0.5385$
3. (a) $\frac{9}{169} = 0.0533$ (b) $\frac{11}{221} = 0.0498$
4. $\frac{11}{1105} = 0.00995$
5. $\frac{3}{13} = 0.2308$
6. (a) 90% (b) 10%
7. (a) 0.048 (b) 0.08
8. (a) $\frac{9}{73} = 0.1233$
   (b) about 12% of the bad credit risks slip through the screening process.
9. $P(X|\text{-}G) = \frac{12}{13}$ while $P(Y|\text{-}G) = \frac{1}{13}$, so X most likely supplied it.
10. (a) 0.008 (b) 0.384
11. (a) 0.24 (b) $\frac{0.56}{0.76} = \frac{14}{19} = 0.7368$

page 125:

1. (a) $P(1 \le x) = 0.83$
   (b) $\mu = 1.1900\ldots \approx 1.2$ pair Note: obviously a customer can't buy 1.2 pairs of shoes. However given a large number of customers this is the average number of pairs that would be expected to be bought.
   (c) $\sigma = 0.8449\ldots \approx 0.8$ pairs
   (d) positively skewed
2. (a) Yes, all possible outcomes and the probability of each are listed because $\sum P(x) = 1$.
   (b) $\mu = 3.2800\ldots \approx 3.3$ occupants
   (c) $\sigma = 1.5625\ldots \approx 1.6$ occupants
   (d) 328
3. \$109,000
4. (a) $\mu = \$0.50$ (b) \$3.75
5. $$\begin{aligned}\sigma &= \sqrt{\frac{\sum x^2 f}{\sum f} - \left(\frac{\sum xf}{\sum f}\right)^2} \\ &= \sqrt{\frac{1}{\sum f}\sum x^2 f - \left(\frac{1}{\sum f}\sum xf\right)^2} \\ &= \sqrt{\sum \frac{1}{\sum f} x^2 f - \left(\sum \frac{1}{\sum f} xf\right)^2} \\ &= \sqrt{\sum x^2 \frac{f}{\sum f} - \left(\sum x \frac{f}{\sum f}\right)^2} \\ &= \sqrt{\sum x^2 P - (\sum xP)^2}\end{aligned}$$

page 131:

1. (a) 2.9802×10^{-8}
   (b) 0.1550
   (c) 0.788
   (d) 12.5 correct
   (e) 2.5 correct
2. (a) 0.0038
   (b) 0.0003
   (c) 0.006

(d) 5 correct

(e) 2.0 correct

3. (a)

x	0	1	2	3	4
$P(x)$	0.008	0.076	0.264	0.412	0.240

(b) $\mu = 2.8$ drivers,
$\sigma = 0.9165\ldots \approx 0.9$ drivers

(c) 0.348

4. (a) $P(1 \leq x) = 0.9576$

(b) $\mu = 3.0$ failures.

(c) In theory the trials are dependent assuming the phones were tested without replacement. However given the large number of phones to choose from in the population the probability of failure would not change appreciably after a phone was removed from it and so a binomial probability calculation is justified
$(n/N = 30/\text{"Large"} \approx 0.00 < .05)$. The exact calculation would require a hypergeometric distribution but the answers would be very close.

(d) The issue may have been related to Microsoft having not used a random sample. (The problem was reported to have affected largely the Omnia 7 model of Samsung headsets. Microsoft also suggested that those affected may have had a bad Internet connection or too little storage on the computer from which the update was installed, which also may not have been reflected in the testing lab.)

5. 0.384

6. 0.1478 (approx. binomial since $n/N = 10/1000 = .01 < .05$)

page 132:

1. (a) $\mu = E(x) = 3.11 \approx 3.1$ oranges

(b) $\sigma = 1.804965 \approx 1.8$ oranges

2. (a) i. $P(x = 0) = 0.2621$
ii. $P(x = 6) = 0.000064 \approx 0.0001$
iii. $P(x \geq 1) = 0.7379$

(b) i. $\mu = E(x) = 1.2$ members
ii. $\sigma = 0.979795 \approx 1.0$ members

page 144:

1. (a) Range $\approx \$46,500 - \$28,500 = \$18,000$, approx 68% between \$34,500 and \$40,500, approx 95% between \$31,500 and \$43,500

(b) i) $z = +1.50$ (typical)
ii) $z = -3.17$ (unlikely)
iii) $z = +7.50$ (very unlikely)

(c) 95%

(d) 97.5%

(e) \$28,500.00

2. (a) 0.4772

(b) 0.2119

(c) 0.6730

3. (a) 0.50

(b) 0.9502

(c) 69.6%

(d) 81.2%

(e) 212 people

4. (a) 0.9131

(b) 0.9830

(c) 0.8961

5. 81.45%

6. 96.71 mm to 103.29 mm

7. P_{62}=73.1%

8. $P(-1.00 < z < 1.00) = .6826$,
$P(-2.00 < z < 2.00) = .9544$,
$P(-3.00 < z < 3.00) = .9974$

9. $k = 2$: 95% > 75%, $k = 3$: 99.7% > 89%

10. $P(173 \text{ cm} < x < 179 \text{ cm}) = .7333$ (normal)
$\approx$ .785 (original)

11. (a)

Destination	$\bar{x}$ (°C)	s (°C)
Cancun, Mexico	29.1	1.5
Los Angeles, California	19.5	4.0
Malibu, California	16.3	3.0
Mazatlan, Mexico	26.2	1.7
Miami, Florida	26.3	3.1
Phoenix, Arizona	22.1	4.3
Puerto Vallarta, Mexico	26.5	1.4
Varadero, Cuba	28.3	3.4

(b)

Destination	z	Interp. of z	$P(x \geq 25°\text{C})$
Cancun	-2.73	extremely low	0.9968
Los Angeles	1.38	slightly high	0.0838
Malibu	2.90	extremely high	0.0019
Mazatlan	-0.71	slightly low	0.7611
Miami	-0.42	approx. average	0.6628
Phoenix	0.67	slightly high	0.2514
Puerto Vallarta	-1.07	slightly low	0.8577
Varadero	-0.97	slightly low	0.8340

Here x is the daily high temperature on March 1st.

(c)

Destination	μ (beach days)	$P(x \geq 5)$
Cancun	7.0	$0.999998864 \approx 1$
Los Angeles	0.6	$7.5098 \times 10^{-05} \approx 0$
Malibu	0.0	$5.1836 \times 10^{-13} \approx 0$
Mazatlan	5.3	0.7791
Miami	4.6	0.5617
Phoenix	1.8	0.0132
Puerto Vallarta	6.0	0.9354
Varadero	5.8	0.9052

Here x is the number of beach days and $P(x \geq 5) = P(5) + P(6) + P(7)$. Note independence from day to day for weather is a bad approximation since weather usually comes in systems that go for multiple days. As such you might expect the distribution for x to be more polarized, with all bad and all good days being more likely than indicated by the binomial distribution.

page 149:

1. $n\pi = 182.5 > 5$, $n(1-\pi) = 182.5 > 5$ so normal approximation valid.
   $P(x < 160.5) = P(z < -2.30) = 0.0107$
2. $n\pi = 20 > 5$, $n(1-\pi) = 980 > 5$ so normal approximation valid. $P(14.5 < x < 25.5) = P(-1.24 < z < 1.24) = 0.7850$
3. $n\pi = 22.5 > 5$, $n(1-\pi) = 27.5 > 5$ so normal approximation valid.
   $P(24.5 < x) = P(0.57 < z) = 0.2843$

page 166:

1. (a) $P(\$9550 < \bar{x} < \$9650)$
   $= P(-2.50 < z < 2.50) = 0.9876$
   (b) $P(\$9660 < \bar{x})$
   $= P(3.00 < z) = 0.0013$
2. (a) $P(2300 \text{ kg} < \bar{x} < 2700 \text{ kg})$
   $= P(-1.81 < z < 1.81) = 0.9298$
   (b) $P(\bar{x} < 2000 \text{ kg or } 3000 \text{ kg} < \bar{x})$
   $= P(\bar{x} < 2000 \text{ kg}) + P(3000 \text{ kg} < \bar{x})$
   $= P(z < -4.52) + P(4.52 < z) \approx 0$
3. (a) $P(76\% < \bar{x}) = P(1.00 < z) = 0.1587$
   (b) 95% of all sample means would fall between 73.04% and 76.96% . (Here we need to find the z-scores which trap the middle 0.9500 of area, i.e. which correspond to cumulative areas of 0.0250 and 0.9750. We get $z = \pm 1.96$ which is the approximate ± 2 of the Empirical Rule. Then solve for $\bar{x}$ in the z-score formula.)

page 169:

1. (a) $\sigma_p = 0.034641016 \approx 0.035$
   (b) $P(0.50 < p) = P(2.89 < z) = 0.0019$
2. (a) $P(p < 0.46) = P(z < 0.64) = 0.7389$
   (b) $P(p < 0.43 \text{ or } 0.47 < p)$
   $= P(p < 0.43) + P(0.47 < p)$
   $= P(z < -1.29) + P(1.29 < z) = 0.1970$
3. (a) $P(0.1 < p < 0.2)$
   $= P(-1.37 < z < 0.00) = 0.4147$
   (b) $P(3 \leq x \leq 6)$
   $= P(3) + P(4) + P(5) + P(6)$
   $= 0.0785 + 0.1325 + 0.1723 + 0.1795$
   $= 0.5628$
   (c) $P(0.0833 < p < 0.2166)$
   $= P(-1.60 < z < 0.23) = 0.5362$
   Comparing to (b) shows a smaller difference than between (a) and (b). Note that $n\pi = 6$ which shows we are close to being outside of the region of validity for p to be approximately normal so some difference is expected.

page 172:

1. (a) $P(76\% < \bar{x}) = P(1.12 < z) = 0.1314$
   (b) 95% of all sample means would fall between 73.25% and 76.75%
2. (a) $P(p < 0.46) = P(z < 0.75) = 0.7734$
   (b) $P(p < 0.43 \text{ or } 0.47 < p)$
   $= P(p < 0.43) + P(0.47 < p)$
   $= P(z < -1.50) + P(1.50 < z) = 0.1336$
3. (a) $\sigma_{\bar{x}} = \sigma$. The mean of a sample of size 1 is the same as measuring x.
   (b) $\sigma_{\bar{x}} = 0$ since $F.C.F. = 0$. No uncertainty as the sample mean is the population mean as the sample is the entire population if sampling without replacement with $n = N$.
4. (a) {2,4}, {2,6}, {2,8}, {4,6}, {4,8}, {6,8}
   Note that when sampling without replacement we use selections (order does not matter), while when sampling with replacement the samples are permutations. (Hence our use of braces here rather than parentheses.)
   (b) $\mu_{\bar{x}} = 5 \text{ occupants} = \mu$
   (c) $\sigma_{\bar{x}} = \sqrt{\frac{5}{3}} \text{ occupants} = \sqrt{\frac{4-2}{4-1}} \cdot \frac{\sqrt{5}}{\sqrt{2}}$
   $= (F.C.F.) \cdot \frac{\sqrt{\sigma}}{\sqrt{n}}$
   (d) The distribution $P(\bar{x})$ is not normal. (Graph it!)

page 182:

1. $P(0.300 - 0.016 \text{ ppm} < \mu < 0.300 + 0.016 \text{ ppm})$
   $= P(0.284 \text{ ppm} < \mu < 0.316 \text{ ppm}) = 0.98$
   *Note: This result would still be valid had the fish been sampled without replacement (instead of our assumed replacement sampling) provided we assume the size of the Walleye population in the lake is significantly larger than the sample size (i.e. $n/N < 0.05$). (See Section *4.7.) Similar considerations apply to the remainder of the problems except for problem 3.
2. $P(\$512 - \$9.54 < \mu < \$512 + \$9.54)$
   $= P(\$502.46 < \mu < \$521.54) = 0.90$
3. (a) $P(\$85 - \$3.29 < \mu < \$85 + \$3.29)$
   $= P(\$81.71 < \mu < \$88.29) = 0.95$

(b) $P(\$85 - \$4.32 < \mu < \$85 + \$4.32)$
$= P(\$80.68 < \mu < \$89.32) = 0.99$

*Note here because
$n/N = 80/500 = 0.16 \geq 0.05$, had we sampled without replacement rather than with replacement (see Section *4.7) the results would have been:

(a) $P(\$85 - \$3.02 < \mu < \$85 + \$3.02)$
$= P(\$81.98 < \mu < \$88.02) = 0.95$

(b) $P(\$85 - \$3.96 < \mu < \$85 + \$3.96)$
$= P(\$81.04 < \mu < \$88.96) = 0.99$

4. n must be 30 or greater to use the C.L.T. .

5. $P(82\% - 1.15\% < \mu < 82\% + 1.15\%)$
$= P(80.85\% < \mu < 83.15\%) = 0.99$

page 185:

1. (a) Yes. $\pi \approx p = 0.005794$, so $n\pi \approx 197 > 5$ and $n(1-\pi) \approx 33803 > 5$.

(b) $P(0.0058 - 0.0007 < \pi < 0.0058 + 0.0007)$
$= P(0.0051 < \pi < 0.0065) = 0.90$

(c) Yes, since the selection process is not random. Sick animals are selected and infected areas have increased hunting quotas. This would mean our estimate will be too high. (On the other hand, CWD tests are likely difficult with infected animals not having observable symptoms resulting in false negatives.)

2. (a) $\pi \approx p = 0.520$

(b) $P(0.520 - 0.182 < \pi < 0.520 + 0.182)$
$= P(0.338 < \pi < 0.702) = 0.99$
*Note that had we been sampling without replacement rather than our assumed replacement sampling the finite population ($n/N \geq 0.05$) results in the different answer (see Section *4.7):
$P(0.520 - 0.175 < \pi < 0.520 + 0.175)$
$= P(0.345 < \pi < 0.695) = 0.99$

3. (a) $P(0.56 - 0.03 < \pi < 0.56 + 0.03)$
$= P(0.53 < \pi < 0.59) = 0.95$
Note a pollster would say "Support of the PP is 56%. This result is accurate to within ±3%, 19 times out of 20."

(b) Yes, since 0.50 is not in the interval.

(c) Yes, $P(0.56 - 0.04 < \pi < 0.56 + 0.04)$
$= P(0.52 < \pi < 0.60) = 0.99$
which still does not contain 0.50 .

4. (a) $P(0.40 - 0.13 < \pi < 0.40 + 0.13)$
$= P(0.27 < \pi < 0.53) = 0.95$

(b) $P[(0.27)(200 \text{ bil}) < \pi N < (0.53)(200 \text{ bil})]$
$= P(54 \text{ billion} < \chi < 106 \text{ billion}) = 0.95$

5. $P\left(\frac{200}{0.0133\ldots+0.0077\ldots} < N < \frac{200}{0.0133\ldots-0.0077\ldots}\right)$
$= P(9507 < N < 35520) = 0.90$
Note that this problem oversimplifies the difficulties of this procedure. At its simplest level the tagged fish have to have had sufficient time to disperse in the population before catching happens again but not so long that a significant number migrate out of the lake (or die). As well a large lake may have different subpopulations within the lake each of which would require adequate sampling both within the initial tagging procedure and also in the subsequent catch.

page 189:

1. 97 measurements

2. 16,588 Quebec citizens (if the 1995 result is used as an estimate)

3. 75 vehicles (Use $\sigma \approx 2.624055386$ km/litre, the (unrounded) pilot survey sample standard deviation.)

4. (a) 601 burial plots

(b) 505 burial plots (Use $\pi \approx \frac{6}{20}$.)

(c) Lower the confidence and/or increase max sampling error. For example, with $1 - \alpha = 0.90$ and $E = 10\% = 0.10$, we need only $n = 57$ burial plots.

page 193:

1. (a) $t = 2.624$

(b) $t = 1.703$

(c) $t = 4.032$

(d) $z = 3.291$ (This is effectively normal as $n = 560 > 30$. Use $df = \infty$ on t table.)

2. (a) $P(156.3 - 78.0\ \frac{\text{m}^3}{\text{s}} < \mu < 156.3 + 78.0\ \frac{\text{m}^3}{\text{s}})$
$= P(78.3\ \frac{\text{m}^3}{\text{s}} < \mu < 234.3\ \frac{\text{m}^3}{\text{s}}) = 0.95$
Note the sample statistics are (unrounded):
$\bar{x} = 156.3\ \text{m}^3/\text{s},\ s = 108.9995413\ \text{m}^3/\text{s}$

(b) Use 30 or more years of data so the C.L.T. applies and the underlying distribution does not matter.

3. (a) $P(\$20.29 - \$6.61 < \mu < \$20.29 + \$6.61)$
$= P(\$13.68 < \mu < \$26.90) = 0.90$
Note the sample statistics are (unrounded):
$\bar{x} = \$20.29166667,\ s = \$12.7455845$

(b) $\$10944 < \text{total} < \$21520$

4. In the original problem the test statistic we manipulated to construct our confidence interval was $(\bar{x} - \mu)/s_{\bar{x}}$ which satisfies a t distribution provided the underlying distribution (x) is normal. If instead we work

with $(\bar{x}-\mu)/\sigma_{\bar{x}}$ then this satisfies a normal (z) distribution provided the underlying distribution x is still normal. Thus we revert to using the normal distribution for our problem and find for the given 98% confidence, $z = 2.326$. Since by uncanny coincidence we have $\sigma = s$, the standard error $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ equals $s_{\bar{x}} = s/\sqrt{n} = 0.55 \ \frac{\text{litres}}{100 \text{ km}}$. However now the maximum error of estimate is $E = z \cdot \sigma_{\bar{x}} = 1.28 \ \frac{\text{litres}}{100 \text{ km}}$ and the confidence interval becomes $P(7.20 - 1.28 \ \frac{\text{litres}}{100 \text{ km}} < \mu < 7.20 + 1.28 \ \frac{\text{litres}}{100 \text{ km}})$ $= P(5.92 \ \frac{\text{litres}}{100 \text{ km}} < \mu < 8.48 \ \frac{\text{litres}}{100 \text{ km}}) = 0.98$ Comparing this confidence interval to the original example answer, $P(5.64 \ \frac{\text{litres}}{100 \text{ km}} < \mu < 8.76 \ \frac{\text{litres}}{100 \text{ km}}) = 0.98$, we see that the new confidence interval is smaller, despite using the same mean and standard deviation. The difference is that in the original problem there is additional uncertainty introduced by having to use a sample standard deviation which results in a less accurate estimate. As a final note, knowledge of the true population standard deviation for an interval estimate rarely occurs since if you actually had the data to calculate the true population standard deviation you would simply press the button to calculate the true population mean as well rather than estimate it. However, if σ is given to you in a revelation you may as well use it ...

page 194:

1. (a) $P(18 \text{ wks} < x) = P(1.50 < z) = 0.0668$
   (b) $P(18 \text{ wks} < \bar{x}) = P(15.00 < z) \approx 0$
   (c) i. $E = 0.31$ wks
   ii. $P(11.80 - 0.31 \text{ wks} < \mu < 11.80 + 0.31 \text{ wks})$ $= P(11.49 \text{ wks} < \mu < 12.11 \text{ wks}) = 0.99$
   (d) i. $E = 2.39$ wks
   ii. $P(11.80 - 2.39\text{wks} < \mu < 11.80 + 2.39\text{wks})$ $= P(9.41 \text{ wks} < \mu < 14.19 \text{ wks}) = 0.99$
   (e) $n = 217$ unemployed people
2. $P(35,000 - 412 \text{ km} < \mu < 35,000 + 412 \text{ km})$ $= P(34,588 \text{ km} < \mu < 35,412 \text{ km}) = 0.98$
3. $P\left(7.2 - 0.7 \ \frac{l}{100 \text{ km}} < \mu < 7.2 + 0.7 \ \frac{l}{100 \text{ km}}\right)$ $= P\left(6.5 \ \frac{l}{100 \text{ km}} < \mu < 7.9 \ \frac{l}{100 \text{ km}}\right) = 0.90$
4. (a) $E = 2.7\%$ for π in favour. The maximum error of estimate depends upon the proportion asked. For the proportion against changing to DST for instance $E = 2.9\%$ while for the proportion with no opinion it is $E = 1.5\%$. The largest sampling error for any possible proportion will occur when $p = 0.5$ is used and this results in $E = 3.1\%$ which is larger than all of them and quoted.
   (b) $P(0.270 - 0.046 < \pi < 0.270 + 0.046)$ $= P(0.224 < \pi < 0.316) = 0.999$ "The percentage in favour of switching to DST was 27%. The result is accurate to within 5.2%, 999 times out of 1000." In some sense this is more useful because the 50% needed for a successful plebiscite is still not in the interval thereby demonstrating the unlikelihood of it succeeding. There is only a 0.1% chance of the population value lying outside of this interval versus 5% in the one quoted. That said, bias is being introduced here if you fiddle the confidence level after seeing what the data is in this manner. Finally, it would make sense to drop the undecided from the sample size before doing the calculation as their inclusion causes the percent in favour on an actual plebescite to be underestimated.
5. $P(0.200 - 0.046 < \pi < 0.200 + 0.046)$ $= P(0.154 < \pi < 0.246) = 0.99$ *Note that if we had sampled without replacement rather than the assumed replacement sampling the answer becomes (see Section *4.7): $P(0.200 - 0.045 < \pi < 0.200 + 0.045)$ $= P(0.155 < \pi < 0.245) = 0.99$
6. $P(140,840 - 16,080 < \mu < 140,840 + 16,080)$ $= P(124,760 \text{ veh.} < \mu < 156,920 \text{ veh.}) = 0.95$ Note the sample statistics are (unrounded): $\bar{x} = 140837.5$ veh., $s = 19228.47126$ veh.
7. $n = 2401$ voters
8. $n = 2653$ entries if $z = 2.575$ is used from the normal table, $n = 2655$ entries if the more accurate $n = 2.576$ is used from the t table.

page 199:

1. $H_0 : \mu = \$52.00/\text{wk}$, $H_a : \mu < \$52.00/\text{wk}$
2. $H_0 : \mu = \$52.00/\text{wk}$, $H_a : \mu > \$52.00/\text{wk}$
3. $H_0 : \pi = 0.18$, $H_a : \pi > 0.18$
4. $H_0 : \mu = 10$ cm, $H_a : \mu \neq 10$ cm

page 206:

1. $z_{\text{critical}} = \pm 1.645$, $z_{\text{calculated}} = 2.475$, therefore reject H_0 and accept H_a at $\alpha = 0.10$. Evidence suggests the population mean differs from \$560 . Alternatively, use P-value $= 0.0136$ less than $\alpha = 0.10$.
2. $t_{\text{critical}} = \pm 1.729$, $t_{\text{calculated}} = 1.565$, therefore fail to reject H_0 at $\alpha = 0.10$. Evidence does not suggest the population mean differs from \$560 . Assume population normality.

3. $z_{\text{critical}} = -1.645$, $z_{\text{calculated}} = -5.42$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. Evidence suggests the population mean is less than 454 g. Alternatively, use P-value ≈ 0 less than $\alpha = 0.05$.
4. $\alpha = 0.05$
5. The butter weighs less than 454 grams on average and the sample average we randomly chose led us to believe that the it does weigh 454 grams on average. That is, we accepted the null hypothesis when it was false.
6. $t_{\text{critical}} = \pm 2.947$, $t_{\text{calculated}} = 8.639$, therefore reject H_0 and accept H_a at $\alpha = 0.01$. Evidence suggests the cluster mean metallicity differs from the general surrounding stellar population. The -0.50 star is an outlier (its z value in the cluster data is -3.19) which may not be a member of the cluster but happens to lie along the line of sight or perhaps migrated into the region of the cluster at some time. Note the sample statistics (not rounded) are $\bar{x} = 0.115625$ dex, $s = 0.192906497$ dex.
7. $t_{\text{critical}} = 1.833$, $t_{\text{calculated}} = 0.945$, therefore fail to reject H_0 at $\alpha = 0.05$. The evidence does not support the conclusion that the Regina rent is greater than \$881 . Note the sample statistics (not rounded) are $\bar{x} = \$910.5$, $s = \$98.72661467$.

page 209:

1. $z_{\text{critical}} = -1.28$, $z_{\text{calculated}} = -2.04$, therefore reject H_0 and accept H_a at $\alpha = 0.10$. Evidence suggests that less than 40% of households watched the program. Alternatively, use P-value $= 0.0207$ less than $\alpha = 0.10$.
2. $z_{\text{critical}} = -2.05$, $z_{\text{calculated}} = -10.23$, therefore reject H_0 and accept H_a at $\alpha = 0.02$. Evidence suggests a (very) significant decrease in the percent of women unaware occurred since 2008. Alternatively, use P-value ≈ 0 less than $\alpha = 0.02$.
3. $z_{\text{critical}} = -1.645$, $z_{\text{calculated}} = -1.89$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. The counselor is incorrect; evidence suggests that less than 30% of students hold a part-time job. Alternatively, use P-value $= 0.0294$ less than $\alpha = 0.05$.
4. $z_{\text{critical}} = \pm 2.326$, $z_{\text{calculated}} = 0.554$, therefore fail to reject H_0 at $\alpha = 0.02$. The evidence does not support a change in the failure rate. Alternatively, use P-value $= 0.5824$ greater than $\alpha = 0.02$.
5. No, this is not a proportion hypothesis test since the underlying variable is not a binomial (yes/no) qualitative variable. Here the underlying variable x itself is the fraction of a litre of gas that is ethanol which is quantitative. The regulator could take a random sample of n measurements (buying one litre of gasoline at n different outlets and/or times) and measure the fraction of ethanol in each one and take its average $\bar{x}$ and then test the alternative hypothesis regarding the population mean that $\mu < 0.06$.
6. (a) $z_{\text{critical}} = -3.291$, $z_{\text{calculated}} = -4.427$, therefore reject H_0 and accept H_a at $\alpha = 0.0005$. The evidence suggests Mike the slasher does not have the required support to win the election (as the T.V. station reported). Alternatively, use P-value ≈ 0 less than $\alpha = 0.0005$.
   (b) Given the incredibly small risk of error ($\alpha = 0.0005$!) the inaccuracy was more likely due to non-response error (voters for Mike were more likely to not divulge how they voted) or response error (more voters for Mike actually said they voted for Joe than the other way around). Given Mike's platform reflected crass narrow self-interest his supporters may have been less than willing to divulge how they voted publicly. Not many people get mad at you if you tell them you voted for "compassionate Joe"!

page 210:

1. $z_{\text{critical}} = 1.645$, $z_{\text{calculated}} = 6.600$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. The evidence suggests an increase. Alternatively, use P-value ≈ 0 less than $\alpha = 0.05$.
2. $z_{\text{critical}} = 2.326$, $z_{\text{calculated}} = 10.000$, therefore reject H_0 and accept H_a at $\alpha = 0.01$. Evidence suggests the trucks weigh more than 25,000 kg. Alternatively, use P-value ≈ 0 less than $\alpha = 0.01$.
3. $z_{\text{critical}} = -1.282$, $z_{\text{calculated}} = -2.000$, therefore reject H_0 and accept H_a at $\alpha = 0.10$. Evidence suggests less than 50% are male. Alternatively, use P-value $= 0.0228$ less than $\alpha = 0.10$.
4. $t_{\text{critical}} = \pm 3.499$, $t_{\text{calculated}} = 0.061$, therefore fail to reject H_0 at $\alpha = 0.01$. Evidence does not support an adjustment be made. Note the sample statistics (unrounded) are $\bar{x} = 300.375$ ml, $s = 17.41048535$ ml .
5. $z_{\text{critical}} = 1.645$, $z_{\text{calculated}} = 1.267$, therefore fail to reject H_0 at $\alpha = 0.05$. Evidence does not support an increase. Alternatively, use P-value $= 0.1020$ greater than $\alpha = 0.05$.

page 212:

1. Since our two-sided hypothesis test for $\mu = \$560$ failed at $\alpha = 0.10$ we expect \$560 to lie outside the $1 - \alpha = 0.90 = 90\%$ confidence interval. That C.I. is

$P(\$595 - \$23.26 < \mu < \$595 + \$23.26)$
$= P(\$571.74 < \mu < \$618.26) = 0.90$ which does indeed not include \$560.

page 230:

1. (a) $z_{\text{critical}} = \pm 1.960$, $z_{\text{calculated}} = 2.614$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. The evidence suggests that there is a difference between the average monthly household incomes in the two communities. Alternatively, use P-value $= 0.0090$ less than $\alpha = 0.05$.

   (b) $P(\$300 - \$224.93 < \mu_1 - \mu_2 < \$300 + \$224.93)$
   $= P(\$75.07 < \mu_1 - \mu_2 < \$524.93) = 0.95$

2. (a) $t_{\text{critical}} = -2.457$, $t_{\text{calculated}} = -1.884$, therefore fail to reject H_0 at $\alpha = 0.01$. The evidence does not support that the firm's professionals earn less.

   (b) $P(-\$3000.00 - \$3251.14 < \mu_I - \mu_E < -\$3000.00 + \$3251.14)$
   $= P(-\$6251.14 < \mu_I - \mu_E < \$251.14)$
   $= 0.95$

3. Note this is a paired-difference problem as we are comparing the same golfers sober and not and have dependent samples.

   (a) $P(25.7 - 19.1 < \mu_L - \mu_S < 25.7 + 19.1)$
   $= P(6.6 < \mu_L - \mu_S < 44.8) = 0.95$

   (b) $t_{\text{critical}} = \pm 2.571$, $t_{\text{calculated}} = 3.455$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. The evidence suggests statisticians under the influence of alcohol play mini golf differently than when sober. Note that this agrees with part (a) since this is a two-sided test and 0 is not in the confidence interval for the difference.

4. No. $z_{\text{critical}} = -1.282$, $z_{\text{calculated}} = -1.059$, therefore fail to reject H_0 at $\alpha = 0.10$. The evidence does not support the hypothesis that a person with enhanced sleep performed the task in less time with statistical significance. Alternatively, use P-value $= 0.1446$ greater than $\alpha = 0.10$.

5. (a) $z_{\text{critical}} = 1.645$, $z_{\text{calculated}} = +1.868$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. Evidence suggests training increased sales. P-value $= 0.0307$.

   (b) $P(\$300.00 - \$373.55 < \mu_T - \mu_{NT} < \$300.00 + \$373.55)$
   $= P(-\$73.55 < \mu_T - \mu_{NT} < \$673.55)$
   $= 0.98$ (using $z = 2.326$)
   Note we ended up rejecting the null hypothesis of $\mu_T - \mu_{NT} = 0$ in part (a) but the confidence interval contains 0 in part (b) which may appear to contradict the discussion of Section 5.3 . No contradiction exists because part (a) did not involve a two-sided test, nor was the same α involved.

   (c) The paired difference approach is better because randomness is reduced by comparing the same employee's performance before and after.

6. Note this is a paired-difference problem as we are comparing same day sales and hence have dependent samples.

   (a) $t_{\text{critical}} = 3.143$, $t_{\text{calculated}} = 1.005$, therefore fail to reject H_0 at $\alpha = 0.01$. The evidence does not support an increase in sales.

   (b) $P(\$145.71 - \$354.87 < \mu_A - \mu_B < \$145.71 + \$354.87)$
   $= P(-\$209.16 < \mu_A - \mu_B < \$500.58)$
   $= 0.95$

7. This problem first requires the calculation of the sample mean and standard deviation of the two data sets which are (unrounded) $\bar{x}_C = 0.115625$ dex, $s_C = 0.192906497$ dex, $\bar{x}_B = -0.334$ dex, $s_B = 0.2020011$ dex. You can use these numbers and the sample sizes of $n_C = 16$ and $n_B = 10$ as a starting point to the problem if you do no wish to calculate the statistics.

   (a) $t_{\text{critical}} = \pm 2.797$, $t_{\text{calculated}} = 5.680$, therefore reject H_0 and accept H_a at $\alpha = 0.01$. The evidence suggests that the mean metallicity of the cluster and background stars differ.

   (b) $P(0.450 - 0.221 \text{ dex} < \mu_C - \mu_B < 0.450 + 0.221 \text{ dex})$
   $= P(0.229 \text{ dex} < \mu_C - \mu_B < 0.671 \text{ dex})$
   $= 0.99$
   Note that this agrees with part (a) since that was a two-sided test and 0 is not in the confidence interval for the difference.

   (c) Also need to assume $\sigma_C = \sigma_B$. Comparing the point estimates $s_C = 0.193$ dex and $s_B = 0.202$ dex this appears reasonable.

8. $z_{\text{critical}} = 1.645$, $z_{\text{calculated}} = 3.061$, therefore reject H_0 and accept H_a at $\alpha = 0.05$. Evidence suggests foreign cars are more expensive to repair on average. Alternatively, use P-value $= 0.0011$ less than $\alpha = 0.05$.

page 240:

1. $z_{\text{critical}} = 1.645$, $z_{\text{calculated}} = 1.054$, therefore fail to reject H_0 at $\alpha = 0.05$. There is no evidence

to suggest that a higher proportion of business students have part-time jobs than science students. Alternatively, use P-value $= 0.1469$ greater than $\alpha = 0.05$.

2. $P(0.06 - 0.134 < \pi_B - \pi_S < 0.06 + 0.134)$
$= P(-0.074 < \pi_B - \pi_S < 0.194) = 0.98$

3. (a) $z_{\text{critical}} = \pm 1.282$, $z_{\text{calculated}} = 1.819$, therefore reject H_0 and accept H_a at $\alpha = 0.20$. The evidence supports a difference in the ability of the two swallow types to carry coconuts. Alternatively, use P-value $= 0.0688$ less than $\alpha = 0.20$.

   (b) It becomes a right-tailed test ($H_a : \pi_A - \pi_E > 0$) now with $z_{\text{critical}} = 0.84$ (which must be found from the normal table) while $z_{\text{calculated}} = 1.819$ still. We still reject H_0 and accept H_a but now evidence supports that African swallows are better carriers than European ones. Alternatively, now use P-value $= 0.0344$ less than $\alpha = 0.20$.

4. $P(0.20 - 0.176 < \pi_A - \pi_E < 0.20 + 0.176)$
$= P(0.024 < \pi_A - \pi_E < 0.376) = 0.90$

5. (a) $P(0.072-0.063 < \pi_O-\pi_Y < 0.072+0.063)$
$= P(0.009 < \pi_O - \pi_Y < 0.135) = 0.99$

   (b) $z_{\text{critical}} = \pm 2.576$, $z_{\text{calculated}} = 2.866$, therefore reject H_0 and accept H_a at $\alpha = 0.01$. Evidence suggests that the proportions preferring an accountant differ. Alternatively, use P-value $= 0.0042$ less than $\alpha = 0.01$.

   (c) The confidence interval does not contain 0 which suggests the hypothesis that the two proportions would be equal, $\pi_O - \pi_Y = 0$, should fail for a two-tailed test with the same α .

page 253:

1. (a) $F_{\text{critical}} = 4.153$, $F_{\text{calculated}} = 0.817$, so fail to reject H_0 at $\alpha = 0.025$. The evidence does not suggest that the mean number of tomatoes was any different between the chemical treatments. The ANOVA table is:

Source	*df*	Sum of Squares	Mean Squares	*F*
Treatments	3	50.9974	16.9991	0.817397
Error	15	311.9500	20.7967	
Total	18	362.9474		

   (b) The tomato production in all treatments must be approximately normally distributed with common standard deviation. Given the small sample sizes (≤ 5) it is hard to evaluate normality with the tools developed in this course (plotting frequency distributions, etc.) The sample standard deviations for the treatments are 4.6, 3.8, 3.6, and 6.2 tomatoes respectively which are reasonably close given the tiny sample sizes. (They only approximate their population standard deviations which are plausibly equal.)

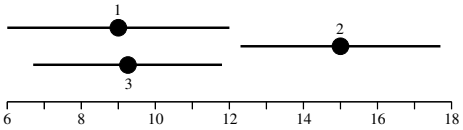
2. $F_{\text{critical}} = 3.634$, $F_{\text{calculated}} = 6.076$, so reject H_0 and accept H_a at $\alpha = 0.05$. Evidence suggests that moviegoers at at least one of the movie times spend a different amount on average than at another time. The ANOVA table is:

Source	*df*	Sum of Squares	Mean Squares	*F*
Treatments	2	83.7381	41.86905	6.07558
Error	16	110.2619	6.89137	
Total	18	194.0000		

3. $F_{\text{critical}} = 6.927$, $F_{\text{calculated}} = 18.954$, so reject H_0 and accept H_a at $\alpha = 0.01$. Evidence suggests that at least one site has a mean arrowhead length that is significantly different than another's. The ANOVA table is:

Source	*df*	Sum of Squares	Mean Squares	*F*
Treatments	2	15.3653	7.682667	18.9539
Error	12	4.8640	0.405333	
Total	14	20.2293		

4. (a) $P(9.0\text{h} - 3.0\text{h} < \mu_1 < 9.0\text{h} + 3.0\text{h})$
$= P(6.0 < \mu_1 < 12.0\text{h}) = 0.95$
$P(15.0\text{h} - 2.7\text{h} < \mu_2 < 15.0\text{h} + 2.7\text{h})$
$= P(12.3 < \mu_2 < 17.7\text{h}) = 0.95$
$P(9.25\text{h} - 2.57\text{h} < \mu_3 < 9.25\text{h} + 2.57\text{h})$
$= P(6.7 < \mu_3 < 11.8\text{h}) = 0.95$



So 1 and 3 appear equal within error but different from 2.

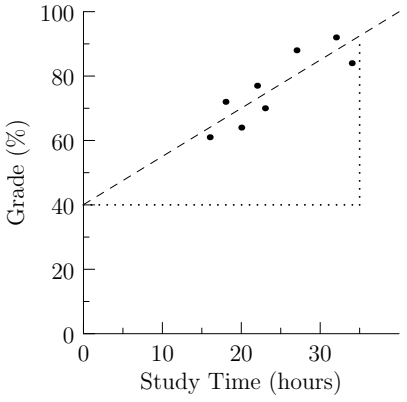
   (b) $P(-6.0\text{h} - 4.0\text{h} < \mu_1 - \mu_2 < -6.0\text{h} + 4.0\text{h})$
$= P(-10.0\text{h} < \mu_1 - \mu_2 < -2.0\text{h}) = 0.95$
$P(-0.25\text{h} - 3.93\text{h} < \mu_1 - \mu_3 <$
$-0.25\text{h} + 3.93\text{h})$
$= P(-4.2\text{h} < \mu_1 - \mu_3 < 3.7\text{h}) = 0.95$
$P(5.75\text{h} - 3.76\text{h} < \mu_2 - \mu_3 < 5.75\text{h} + 3.76\text{h})$
$= P(2.0\text{h} < \mu_2 - \mu_3 < 9.5\text{h}) = 0.95$
Here the absence of zero in the first and third intervals suggests that the means 1 and 2, and the means 2 and 3 differ.

The methods presented for comparing means in this question are problematic to the extent that they consider a given mean or pair of means in isolation of the rest of the treatments. In part (b) for instance our various pairs of intervals individually can identify a significant difference in means between two treatments at $\alpha = 0.05$ (by looking for whether zero occurs in the interval or not). However

the overall probability of error in the procedure of distinguishing which pairs of means differ will not be $\alpha = 0.05$ since error accrues for each test. A method for doing paired comparisons with an overall error of α of declaring at least one pair of means has a difference when none actually exists is *Tukey's method for paired comparisons*. In addition to our usual assumptions Tukey's method requires all treatments sizes to be equal so it could not be used in our current circumstance.

page 269:

1. (a)



   $a \approx 40$ %, $b \approx \frac{\Delta y}{\Delta x} = \frac{90-40}{35-0} = 1.4$ %/h, $r \approx +0.8$

   (b) $a = 40.08$ %, $b = 1.497$ %/h, $y_p = a + bx = 40.1 + 1.5x$ (%)

   (c) $y_p = 85\%$

   (d) $r = 0.8621$, very high positive correlation

2. (a) $y_p = -419.8 + 6.7x$ (MW)

   (b) $y_p = 151.1$ MW

   (c) $r = 0.9441$; magnitude $|r| = 0.9441$ implies extremely high correlation; sign of r positive implies positive correlation (As temperature (x) increases, peak load (y) increases.)

3. (a) $r = -0.9722$, an extremely high negative correlation.

   (b) $y_p = 15.58 - 0.8966x$; For $x = \ln(60.2) = 4.10$, $y_p = 11.9$.

   Note: Cepheid variable stars are of critical importance to astronomers precisely because one can measure their period to determine their intrinsic brightness (absolute magnitude). Then measuring their apparent brightness allows an exact measurement of their distance from us. They are one of the *standard candles* for measuring distances far beyond our galaxy.

Descriptive Statistical Measures

Raw Data	Ungrouped Frequency Distribution	Grouped Frequency Distribution
Measures of the Centre:		
$\mu = \dfrac{\sum x}{N}$ or $\bar{x} = \dfrac{\sum x}{n}$	$\mu, \bar{x} = \dfrac{\sum xf}{\sum f}$	$\mu, \bar{x} = \dfrac{\sum xf}{\sum f}$
$\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}(N+1)$ or $\dfrac{1}{2}(n+1)$	$*\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}\left(\sum f + 1\right)$	$*\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}\left(\sum f\right)$
$\underset{\text{Value}}{\text{Median}} = x_{\frac{1}{2}(N+1)}$ or $x_{\frac{1}{2}(n+1)}$	$*\underset{\text{Value}}{\text{Median}} = x_{\frac{1}{2}(\sum f+1)}$	$*\underset{\text{Value}}{\text{Median}} = L_i + \dfrac{\left\{\frac{1}{2}(\sum f) - <Cf_{i-1}\right\}}{f_i} \cdot \Delta x$
Measures of Dispersion:		
a.d. $= \dfrac{\sum \lvert x-\mu \rvert}{N}$ or $\dfrac{\sum \lvert x-\bar{x} \rvert}{n}$		
$\sigma = \sqrt{\dfrac{\sum (x-\mu)^2}{N}}$		
$\sigma = \sqrt{\dfrac{\sum x^2}{N} - \left(\dfrac{\sum x}{N}\right)^2}$	$\sigma = \sqrt{\dfrac{\sum x^2 f}{\sum f} - \left(\dfrac{\sum xf}{\sum f}\right)^2}$	$\sigma = \sqrt{\dfrac{\sum x^2 f}{\sum f} - \left(\dfrac{\sum xf}{\sum f}\right)^2}$
$s = \sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$		
$s = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$	$s = \sqrt{\dfrac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}}$	$s = \sqrt{\dfrac{\sum x^2 f - \frac{(\sum xf)^2}{\sum f}}{\sum f - 1}}$
$R = x_n - x_1$	$IQR = Q_3 - Q_1$	$IDR = D_9 - D_1$	10-90 $PR = P_{90} - P_{10}$
Lower Fence $= Q_1 - 1.5(IQR)$		Upper Fence $= Q_3 + 1.5(IQR)$	

Standard Scores

$z = \dfrac{x-\mu}{\sigma}$	$z = \dfrac{x-\bar{x}}{s}$
$x = \mu + z\sigma$	$x = \bar{x} + zs$

Miscellaneous Formulae:

$P = \dfrac{f}{\sum f}$	$p = \dfrac{P}{\Delta x}$	$1 - \dfrac{1}{k^2}$

*Formulae denoted with an asterisk are from optional sections of the course.

Basic Probability Formulae

Counting Rules

$${}_n P_r = \frac{n!}{(n-r)!} \qquad n^r \qquad {}_nC_r = \frac{n!}{r!(n-r)!}$$

Probability Rules

$$P(A) = \frac{n(A)}{n(S)} \qquad 0 \leq P(A) \leq 1 \qquad P(A) + P(\text{-}A) = 1$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \qquad P(A \text{ and } B) = P(A) \cdot P(B|A)$$

$$P(A \text{ or } B) = P(A) + P(B) \qquad P(A \text{ and } B) = P(A) \cdot P(B)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Probability Distributions Formulae

Binomial Probability Distribution Function

$$P(x) = {}_nC_x \cdot \pi^x \cdot (1-\pi)^{n-x}$$

Parameters of Probability Distributions

General Formulae for Discrete Distributions

$$\mu = E(x) = \sum xP \qquad \sigma = \sqrt{\sum x^2 P - \left(\sum xP\right)^2}$$

Shortcuts for the Binomial Distribution

$$\mu = E(x) = n\pi \qquad \sigma = \sqrt{n \cdot \pi \cdot (1-\pi)}$$

One Population Inferences Formulae

Standard Error for Single Means and Proportions

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad \sigma_{p} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \qquad s_{p} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$^{*}\textbf{Finite Correction Factor} \quad \Rightarrow \quad {}^{*}F.C.F. = \sqrt{\frac{N-n}{N-1}}$$

Maximum Error of Estimate for Single Means And Proportions

$$E = z \cdot s_{\bar{x}} \qquad E = t \cdot s_{\bar{x}} \qquad E = z \cdot s_{p}$$

Confidence Intervals for Population Means and Proportions

$$P\left(\bar{x} - E < \mu < \bar{x} + E\right) = 1 - \alpha \qquad P\left(p - E < \pi < p + E\right) = 1 - \alpha$$

Sample Sizes for Estimating Means And Proportions

$$n = \left[\frac{z \cdot \sigma}{E}\right]^{2} \qquad n = \pi \cdot (1-\pi) \cdot \left[\frac{z}{E}\right]^{2}$$

Standard Scores for Single Means and Proportions

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad z = \frac{p - \pi}{\sigma_{p}}$$

Other Formulae

$$_{N}C_{n} \text{ or } N^{n} \qquad p = \frac{x}{n} \qquad df = n - 1$$

Bivariate Data Analysis

Linear Regression

$$y_{p} = a + bx$$

$$a = \frac{\left(\sum x^{2}\right) \cdot \left(\sum y\right) - \left(\sum x\right) \cdot \left(\sum xy\right)}{n \cdot \left(\sum x^{2}\right) - \left(\sum x\right)^{2}} \qquad {}^{*}a = \bar{y} - b\bar{x}$$

$$b = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{n \cdot \left(\sum x^{2}\right) - \left(\sum x\right)^{2}} \qquad {}^{*}b = r\left(\frac{s_{y}}{s_{x}}\right)$$

Correlation Analysis

$$r = \frac{n \cdot \left(\sum xy\right) - \left(\sum x\right) \cdot \left(\sum y\right)}{\sqrt{n \cdot \left(\sum x^{2}\right) - \left(\sum x\right)^{2}} \cdot \sqrt{n \cdot \left(\sum y^{2}\right) - \left(\sum y\right)^{2}}} \qquad {}^{*}r = \frac{s_{xy}}{s_{x} \cdot s_{y}}$$

Covariance

$${}^{*}s_{xy} = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{n-1}$$

Multiple Population Inferences Formulae

Standard Error Formulae for Differences of Means and Proportions

$$^*s_{\bar{d}} = \frac{s_d}{\sqrt{n}} \qquad s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \qquad s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2}{n_1+n_2-2}\cdot\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

$$^*s_{p_1-p_2} = \sqrt{\frac{p_1\cdot(1-p_1)}{n_1}+\frac{p_2\cdot(1-p_2)}{n_2}} \qquad ^*s_{p_1-p_2} = \sqrt{p_{\text{Pool}}\cdot(1-p_{\text{Pool}})\cdot\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

Maximum Error of Estimate for Differences of Means and Proportions

$$^*E = t\cdot s_{\bar{d}} \qquad E = z\cdot s_{\bar{x}_1-\bar{x}_2} \qquad E = t\cdot s_{\bar{x}_1-\bar{x}_2} \qquad ^*E = z\cdot s_{p_1-p_2}$$

Confidence Intervals for Differences of Means and Proportions

$$^*P\left(\bar{d}-E<\mu_1-\mu_2<\bar{d}+E\right)=1-\alpha$$

$$P(\,[\bar{x}_1-\bar{x}_2]-E<\mu_1-\mu_2<[\bar{x}_1-\bar{x}_2]+E\,)=1-\alpha$$

$$^*P(\,[p_1-p_2]-E<\pi_1-\pi_2<[p_1-p_2]+E\,)=1-\alpha$$

Standard Scores For Differences Between Means And Proportions[†]

$$^*t = \frac{\bar{d}-0}{s_{\bar{d}}} \qquad z = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}} \qquad t = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_{\bar{x}_1-\bar{x}_2}} \qquad ^*z = \frac{(p_1-p_2)-0}{s_{p_1-p_2}}$$

[†]Here we are assuming hypothesis tests of equality so $\mu_1-\mu_2=0$ and $\pi_1-\pi_2=0$ in these formulae already.

Other Formulae

$$^*d = x_1-x_2 \qquad df = n_1+n_2-2 \qquad ^*p_{\text{Pool}} = \frac{x_1+x_2}{n_1+n_2}$$

ANOVA Table

Source	df	Sum of Squares	Mean Squares	F statistic
Treatments	$df_T = k-1$	$SST = \sum\frac{T_i^2}{n_i} - CM$	$MST = \frac{SST}{df_T}$	$F = \frac{MST}{MSE}$
Error	$df_E = n-k$	$SSE =$ Total $SS - SST$	$MSE = \frac{SSE}{df_E}$	
Total	$df_{TOT} = n-1$	Total $SS = \sum x^2 - CM$		$\left(F_{\text{crit}}:\ \begin{matrix}df_1=df_T\\df_2=df_E\end{matrix}\right)$

Here $CM = \frac{\left(\sum x\right)^2}{n} = \frac{\left(\sum T_i\right)^2}{n}$ where T_i is the sum of x for the i^{th} treatment with size n_i.

Requirements for a Complete Solution (F.S.A.R.U.)

Formula State the formula used, including appropriate symbols.

Substitution Substitute the values for your problem. Remember to add any columns to a table that are required for the calculation of those values.

Answer Write your answer including sufficient extra decimals of significance to allow for rounding.

Roundoff Round to the appropriate number of decimal places.

Units Include the appropriate units.

Steps for a Confidence Interval

Step 1) Identify all given information with symbols, preferably on a Venn diagram.

Step 2) Draw a diagram of the sampling distribution.

Step 3) Determine the z-value or t-value.

Step 4) Calculate the standard error.

Step 5) Calculate the maximum error of estimate, E .

Step 6) Make the confidence interval statement.

Steps for a Hypothesis Test

Critical Value Approach

Step 1) Formulate the null and alternative hypotheses

Step 2) State the level of significance.

Step 3) Determine the test statistic.

Step 4) Establish a decision rule. (draw sampling distribution, find critical value)

Step 5) Evaluate the evidence. (identify statistics on a Venn diagram, find the calculated value)

Step 6) State your decision.

P-value Approach

Step 1) Formulate the null and alternative hypotheses

Step 2) State the level of significance.

Step 3) Determine the test statistic.

Step 4) Evaluate the evidence. (identify statistics on a Venn diagram, find the calculated value)

Step 5) Find the P-value. (draw sampling distribution)

Step 6) State your decision.

GNU Free Documentation License

Version 1.3, 3 November 2008
Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

<http://fsf.org/>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.