SIAST Palliser Campus

Mathematics
STAT 220

Lecture Notes and Examples

Unit 1
Descriptive Statistics

by Blake Friesen and Robert G. Petry

published by the Department of Mathematics, SIAST Palliser Campus

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

- Original document produced in 2011 entitled "Descriptive Statistics" written by principal authors Blake Friesen and Robert G. Petry with contributions from Mike Decorby. Published by the Department of Mathematics, SIAST Palliser Campus. A transparent copy of this document is available via *http://www.campioncollege.ca/about-us/faculty-listing/dr-robert-petry*

Contents

1 Introduction to Statistics

What is statistics?

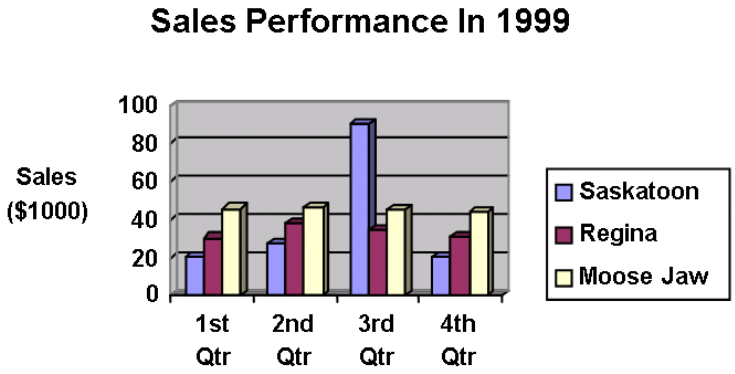
Depending upon one's background, the word statistics has many usages within our language.

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: 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 is statistics used in technology and business?

In a technology/business setting, situations which give rise to analysis of quantities which vary, and as a result involve statistical methods, are activities such as:

- quality control
- forecasting for the purpose of planning
- statistical reports of business activities
- estimating
- testing
- validation of assumptions in planning business
- any procedure which relies on sampling
- simulation

In particular some of the sectors of technology and business who rely on statistical procedures are:

1. Risk taking organizations like the insurance business who analyze risks and found their business on the basis of these risks. Persons 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 engineers and technologists are concerned with 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.

As a technology/business student, how will statistics be useful?

1. It will be useful in order to understand many of the strategies used in marketing, designing and manufacturing.
2. It will be useful in order to understand the technology/ business techniques based on sampling which are used by people like engineers and accountants to make decisions.
3. The statistical data presentation techniques are useful in report writing.
4. To be technically literate in a complex technical world, a person should understand the meaning of the statistical measures on which decisions are based.

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 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.

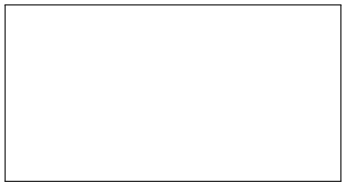
Population (It doesn't necessarily mean people.)

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

Example:

- 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?"

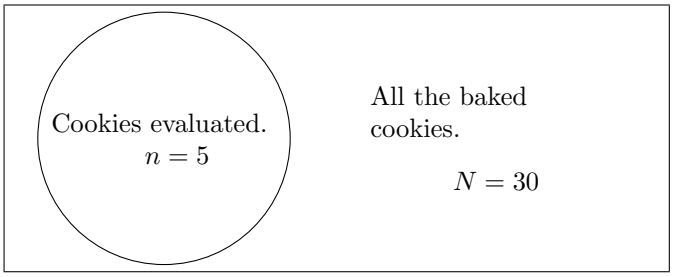
Sample (The everyday definition meaning *to take a portion of* is similar to the statistical one.)

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.



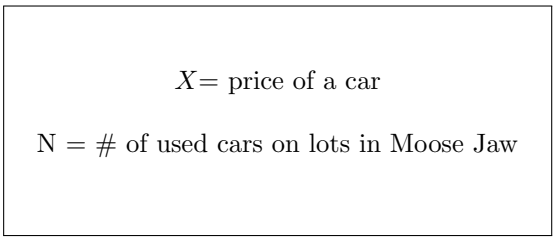
We will use n to refer to the size of a sample and N to refer to the size of a population.

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.

Example:

- In the above example the child could test quality by observing the number of chocolate chips found in each cookie (X=# 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.



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 is 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. 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 Data (Usually non-numerical labels or categories called attributes)

Data 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.
- The ranks of military personnel at an armed forces base are recorded.

Quantitative 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.
- 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.

Discrete Data (Counts)

Data is called discrete when it 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 magazine subscribed to by a household.
- The number of birds visiting a bird feeder in an hour.

Continuous Data (Measurements)

Data is called continuous when it results 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.

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) A large part of statistical analysis involves recording the value of the statistical variable. If the data is not recorded with enough precision, certain types of calculations cannot be done on the observations. Precision refers to the degree of refinement of the observation. To measure the mass of an atom we require a more precise method of measuring than is required to rate customer satisfaction with a product. Four levels of measurement are identified in statistical calculations.

Nominal level is the most crude form of data for analysis. In this form, data can only be categorized such as by colour, religious denomination, political affiliation, and 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 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 ± 3% in 95 samples out of 100.

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

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

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 on page 1 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 in the unit supplement which we will analyze with increasing sophistication. To begin with, read the case studies. For each case 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.

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.

3.1 Raw Data

This is data listed in an array in the order it was observed. This format is used when there are a very few observations to consider in the presentation or when the data is to be keyed into a computer for analysis.

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

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 above information could be summarized in an ungrouped frequency table that looks like this:

# 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.

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.

4 Frequency Distribution Table Construction

Recall frequency distribution tables are appropriate when there is a large number of data elements. When data are summarized by means of a frequency table, the data should be analyzed to determine whether a grouped or an ungrouped distribution summary table is the appropriate format.

4.1 Ungrouped Frequency Table Construction

An ungrouped frequency table is suitable when the number of values taken on by the data is small. Once it is decided that this format suits the data best for summary purposes, ungrouped tables are straightforward to construct. For instance in the example from Section 3.2 one would go through the data and construct the following tally.

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

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

4.2 Grouped Frequency Table Construction

Grouped frequency distributions are useful when the data value assumes a large number of distinct values. The construction of these tables is more complicated because one has choices to make regarding the number of classes, their width and where they should start. One desires that the trends in the data be preserved after the data are grouped. The following example outlines a commonly used systematic procedure for creating a grouped frequency distribution from a data array.

Example:
Weights of the contents of boxes of cereal were: (in g)

228.3	223.4	223.7	225.6	223.6	221.9	229.4	221.4	221.8	240.8
224.9	224.7	224.1	227.8	222.4	225.9	220.9	221.8	222.5	219.9
221.4	224.6	223.9	223.1	227.3	220.9	224.1	224.1	223.4	226.4
224.8	224.6	223.3	220.4	229.5	224.8	227.7	230.3	222.8	220.4
230.8	229.1	221.8	223.2	221.9	228.2	223.9	226.5	224.4	223.9

Construct a grouped frequency distribution suitable for the data.

Step 1) Find the smallest (X_1) and largest (X_n) values in the (ranked) data array. Establish the observed **range** of the variable using the formula

$$\boxed{R = X_n - X_1} = 240.8 \text{ g} - 219.9 \text{ g} = 20.9 \text{ g}$$

There is a spread of 20.9 g between the least and greatest observation.

Step 2) Determine the **approximate number of classes** to use. This is done by *Sturge's formula* which says that there should be a doubling relation between the number of classes (N) and the number of observations (n) to be grouped. Mathematically this means:

$$\boxed{N = 1 + \log(n)/\log(2)} = 1 + \log(50)/\log(2) = 6.64 \text{ classes}$$

This number means that between 6 and 7 classes should be used. Round the answer to two places after the decimal for purposes of the next calculation.

Step 3) Determine the **approximate class width** ΔX by the formula:

$$\boxed{\Delta X = \frac{R}{N}} = \frac{20.9 \text{ g}}{6.64 \text{ classes}} = 3.15 \text{ g/class}$$

Step 4) Determine the **actual class width** and **class limits** of the **first class**. The values calculated by these formulae should be used as guidelines rather than absolutes. The chosen class size and class limits should be values that are close to those calculated but easily interpreted. Multiples of 2, 5, or 10 should be used for the class limits and also for the class interval for ease in interpreting the data. The first class should be chosen so that it includes the first data value. The first class need not start at the first data value.

In our case choose $\Delta X = 4.0$ g/class for ease in interpretation. Although the least observation is 219.9 g, the lower limit for the first class will be chosen to be 218.0 g for clarity. The **precision** of the stated limits is to the nearest tenth of a gram because the original data array had this precision. The first class is therefore [218.0 g − 222.0 g] .

Step 5) Determine the **remaining class limits** by continuing until the maximum value X_n is included.

The class limits at right will result for our data. The final class will be able to include $X_n = 240.8$ g. This will result in 6 classes. This is within the guidelines of the number of classes determined by formula in Step 2. The classes must be mutually exclusive. The first class reads 218.0 g up to but not including 222.0 g. An observation of 222.0 g would fall into the second class.

Weight (g)
218.0 - 222.0
222.0 - 226.0
226.0 - 230.0
230.0 - 234.0
234.0 - 238.0
238.0 - 242.0

Step 6) Complete the grouped frequency table using a tally (or otherwise) of the data as before.

Weight (g)	Tally	f
218.0 - 222.0	𝍸 𝍸 ||	12
222.0 - 226.0	𝍸 𝍸 𝍸 𝍸 𝍸	25
226.0 - 230.0	𝍸 𝍸	10
230.0 - 234.0	||	2
234.0 - 238.0		0
238.0 - 242.0	|	1
		$\sum f = 50$

Note:

1. This six step process for summary table construction is only used for grouped frequency tables. Don't use this process for constructing an ungrouped frequency table.

2. The classes chosen by this method are closed and of equal width. This is desirable for graphical and computational considerations. Sometimes open classes and unequal class widths are found in frequency tables that are used for purely descriptive purposes. Age distributions and income distributions are such examples because they often have several values that are outliers with respect to the other numbers in the distribution.

3. Sometimes grouped frequency tables are created which have class limits chosen so that they do not overlap. For example, why not make the first class [218.0 g – 221.9 g] ? At first glance this seems reasonable to prevent confusion. Can you see any problem with this?

Assignment:

1. Reconstruct the frequency distribution found in case study 4 using the following raw data from which it is derived. (years)

3, 2, 6, 4, 3,
2, 2, 4, 2, 4,
2, 3, 3, 1, 2,
4, 1, 2, 2, 6

2. Reconstruct the frequency distribution found in case study 6 using the following raw data from which it is derived. (km/h)

20, 29, 18, 30, 60, 10, 49, 6, 5, 31,
12, 21, 25, 19, 36, 3, 88, 67, 71, 39,
21, 55, 42, 62, 9, 24, 48, 99, 38, 31

5 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.

5.1 Raw Data

The sum of all data values is given by:

$$\sum X = X_1 + X_2 + X_3 + ... + X_n \, .$$

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
$\sum X = 190$

5.2 Ungrouped Frequency Distribution

In this case the sum of all observations is:

$$\sum fX = f_1X_1 + f_2X_2 + f_3X_3 + ... + f_mX_m \, .$$

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

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

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

5.3 Grouped Frequency Distributions

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 $\sum fX$ where X is the midpoint of the class.

Example:
An exam has the following grade results:

Score (%)	f	$X(\%)$	$fX(\%)$
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
	$\sum f = 50$		$\sum fX = 3000$

Note: There are 50 scores summarized in the table and the approximate sum of all scores is 3000%.

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.[1] 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 (# 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.[2] 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.

[1] 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.

[2] 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.

6 Frequency Distribution Tables Extended

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

6.1 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	.383
2	25	.417
3	5	.083
4	5	.083
5	1	.017
6	1	.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	.125
20 - 40	600	.375
40 - 60	400	.250
60 - 80	300	.188
80 - 100	100	.063
	$\sum f = 1600$	$\sum P = 1.001$

Age Distribution in Bigtown		
Age (yr)	f	P
0 - 20	1500	.025
20 - 40	12000	.200
40 - 60	22000	.367
60 - 80	16000	.267
80 - 100	8500	.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	.129
20 - 40	296	.370
40 - 60	200	.250
60 - 80	152	.190
80 - 100	49	.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. A **relative % frequency** (symbol $\%f$) may be introduced and a conversion to P from $\%f$ made by multiplying P by 100. In this course, however, we will restrict ourselves to relative frequency P and use the decimal and percent notation interchangeably.

6.2 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	.020
171.0-173.0	172.0	10	.050
173.0-175.0	174.0	14	.070
175.0-177.0	176.0	73	.365
177.0-179.0	178.0	70	.350
179.0-181.0	180.0	17	.085
181.0-183.0	182.0	10	.050
183.0-185.0	184.0	2	.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	.100
174.0-178.0	176.0	128	.640
178.0-182.0	180.0	46	.230
182.0-186.0	184.0	6	.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**[3] (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	.010
171.0-173.0	172.0	10	.025
173.0-175.0	174.0	14	.035
175.0-177.0	176.0	73	.183
177.0-179.0	178.0	70	.175
179.0-181.0	180.0	17	.043
181.0-183.0	182.0	10	.025
183.0-185.0	184.0	2	.005
		$\sum f = 200$	

$\Delta X = 4.0$ cm

Height(cm)	X(cm)	f	p(cm^{-1})
170.0-174.0	172.0	20	.025
174.0-178.0	176.0	128	.160
178.0-182.0	180.0	46	.058
182.0-186.0	184.0	6	.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 $\Delta X = 178.5 \text{ cm} - 177.5 \text{ cm} = 1.0$ cm centred on $X = 178.0$ cm. At $X = 178.0$ cm, p is .175 cm^{-1} so

$$P = p \cdot \Delta X = (.175 \text{ cm}^{-1})(1.0 \text{ cm}) = .175$$

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.

6.3 Cumulative Frequency ($<Cf$)

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

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

[3]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.

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.

This column will be useful when the proportions of observations within a given interval of the range are calculated and for constructing certain types of graphs.

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

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 the cumulative frequency columns.

The usefulness of these columns will become apparent as we interpret their meaning graphically.

6.4 Cumulative Relative Frequency ($<CP$)

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

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

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	.05	.05
1	2	.10	.15
2	3	.15	.30
3	4	.20	.50
4	6	.30	.80
5	3	.15	.95
7	1	.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 6.2. 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.[4] In that case one would be able to define the *continuous* cumulative relative frequency function $<CP(X)$ by the integral:

$$<CP(X) = \int^X p\, dX$$

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.

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. Append columns for cumulative frequency, $<Cf$, and cumulative relative frequency, $<CP$, to all frequency distributions.

[4] 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.

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 statistical information. 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 axis.

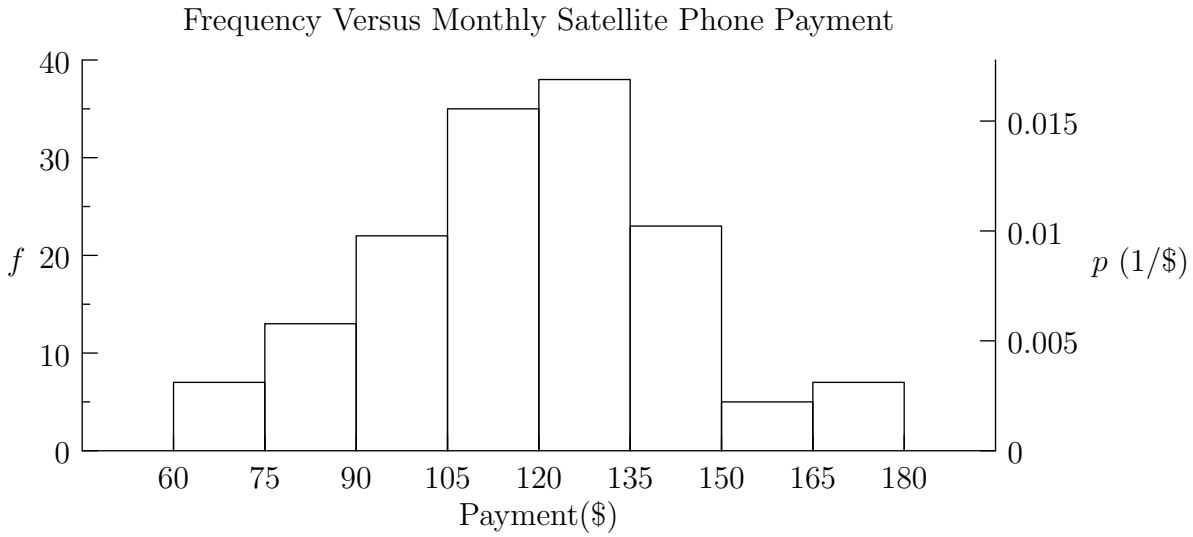
Consider the following grouped frequency distribution representing the payments, on satellite phone contracts, made by a group of surveyors. Additional columns for relative frequency (P), relative frequency density (p), less than cumulative frequency ($<Cf$) and less than cumulative relative frequency ($<CP$) have already been appended.

Monthly Satellite Phone Payments					
Payment (\$)	f	P	p (1/\$)	$<Cf$	$<CP$
60 - 75	7	.047	.0031	7	.047
75 - 90	13	.087	.0058	20	.134
90 - 105	22	.147	.0098	42	.281
105 - 120	35	.233	.0156	77	.514
120 - 135	38	.253	.0169	115	.767
135 - 150	23	.153	.0102	138	.920
150 - 165	5	.033	.0022	143	.953
165 - 180	7	.047	.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.

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.

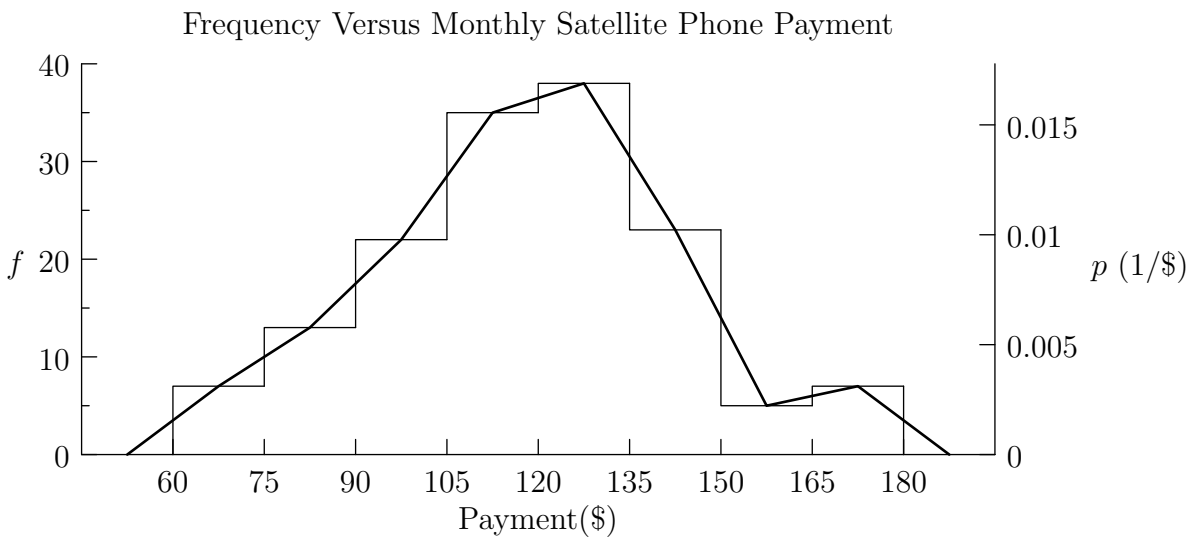


Rather than plotting points, the histogram is drawn with bars of height given by the frequency of the class f and width given by the class width ΔX. On the one hand, plotting a bar rather than a point makes sense since the frequency is associated with an entire class. The reason for plotting bars is 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 bar in the histogram is proportional to the frequency because the area is simply $f \cdot \Delta X$ and ΔX is constant. If one bar 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. A histogram of relative frequency density p versus X is especially useful since the area of a particular bar now becomes $p \cdot \Delta X$ which is just the relative frequency (proportion) P associated with that class. The total area of all the bars 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 will be found when we plot cumulative frequency versus X.

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.



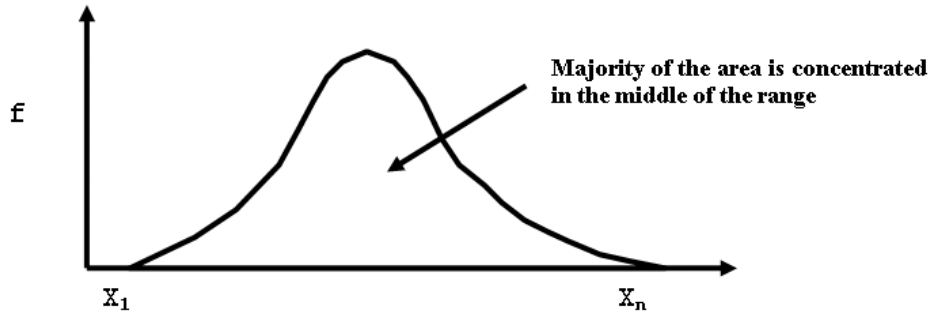
The frequency polygon is plotted by plotting the frequency f (or relative frequency density p) versus the **midpoint** of each class. Note one must add an empty class before the first and after the last class so that the frequency polygon can be extended to zero. If one compares the frequency polygon 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. It is said that the mathematical curve models the distribution.

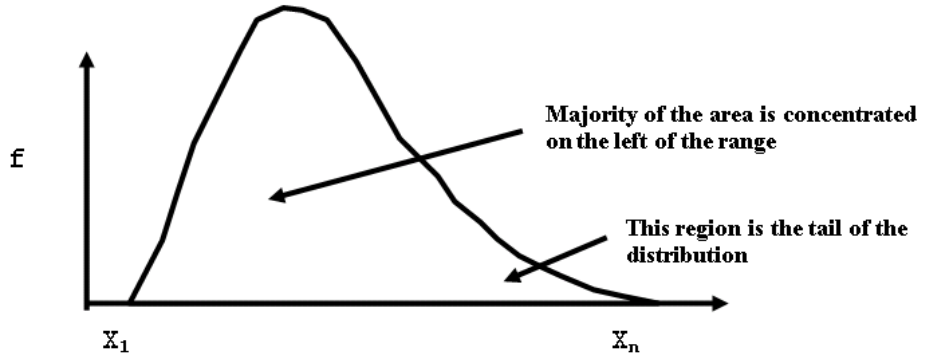
The Shape Of A Distribution

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

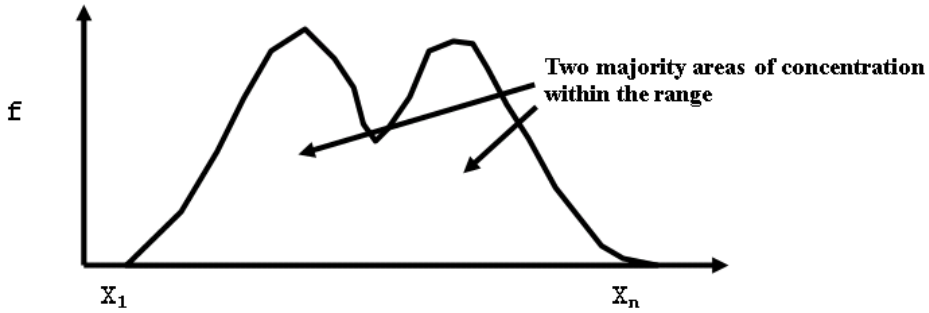
1. Many distributions have the bulk of the observations in the middle of the range with a few observations located at the extremities of the range. As a result, the distribution curve takes the following **symmetrical** shape:



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 of the distribution is on the left side we say that it is **negatively** skewed. If the tail is on the right side the distribution is **positively** skewed.

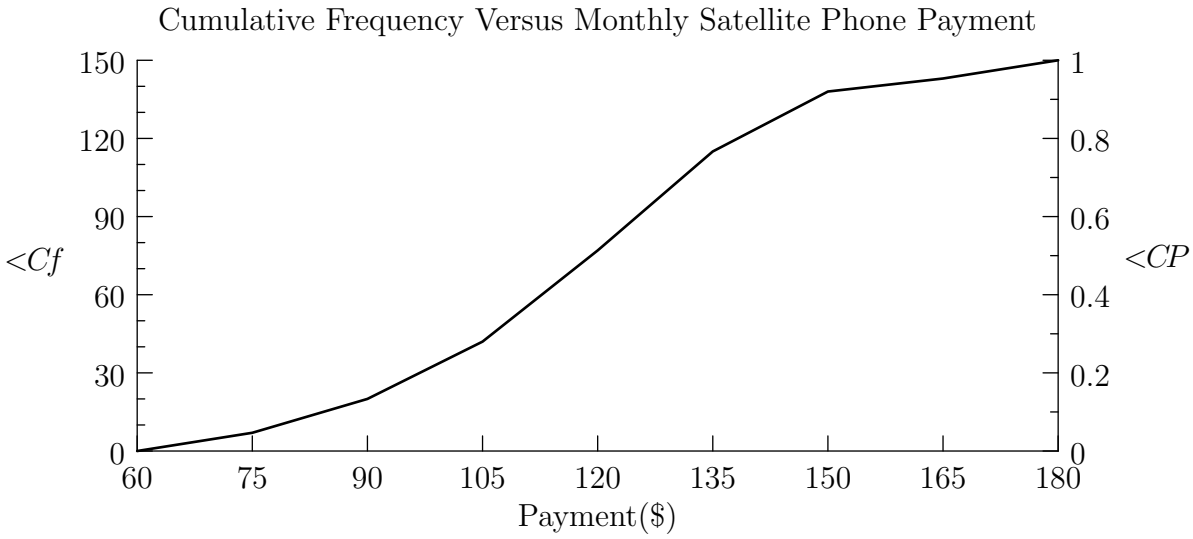


3. Sometimes there is more than one area of concentration in the distribution. This is reflected in the frequency polygon by several peaks. If the distribution has two peaks, it is referred to as a **bimodal** shape. A bimodal shape indicates that there are two underlying trends within one data set.



7.2 Cumulative Frequency Polygon Curves (Ogive Curves)

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:

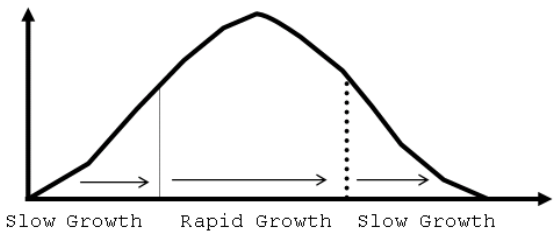


For a less than cumulative frequency ogive one plots $<Cf$ versus the **upper limit** of each class. The lower limit of the first class is given the value zero. That this should be the case follows from the fact that at the very lower end of the class none of the frequency for the given class has accumulated 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 $\sum f$.

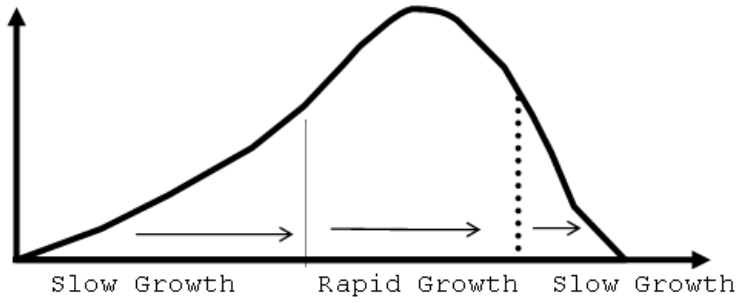
On the right axis has been added the scale for the less than cumulative relative frequency $<CP$. An ogive plotting cumulative relative frequency $<CP$ versus X is useful if one is interested in the fraction of observations occurring up to a point rather than the actual number of observations. In this 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.)

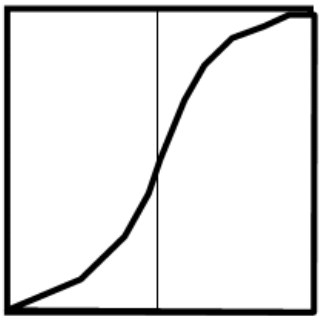
The Path Traced Out By The Ogive Curve

If a distribution is perfectly symmetrical, 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.

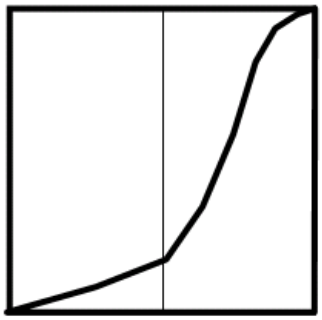


If the distribution is skewed, it will build area slowly for a greater distance at the beginning of the range and the ogive curve will have distorted "S" shape.





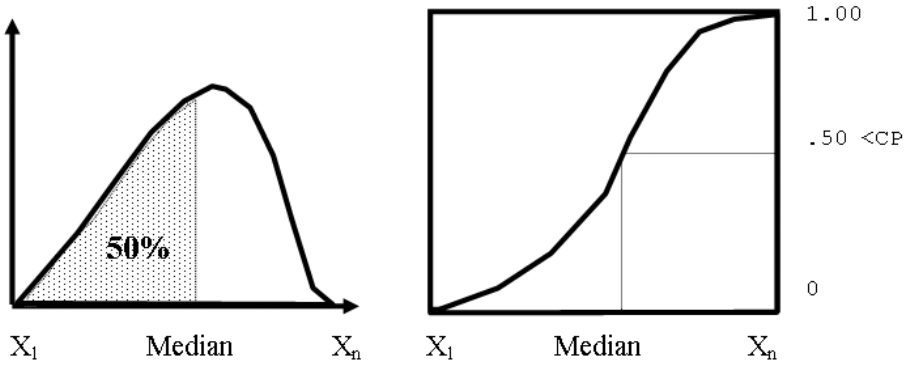
Symmetrical Growth



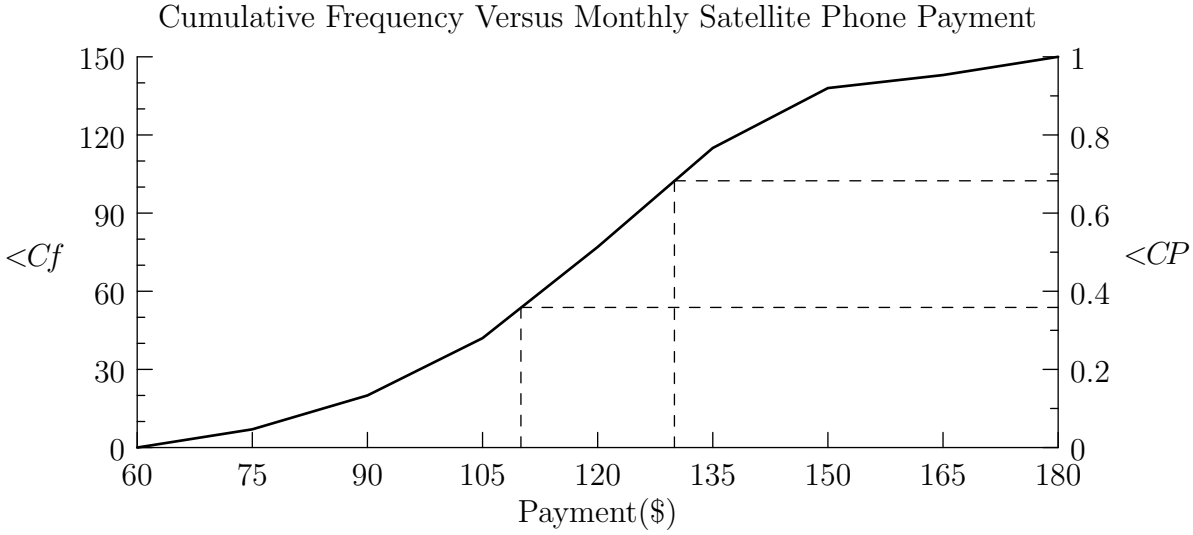
Negatively Skewed Growth

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

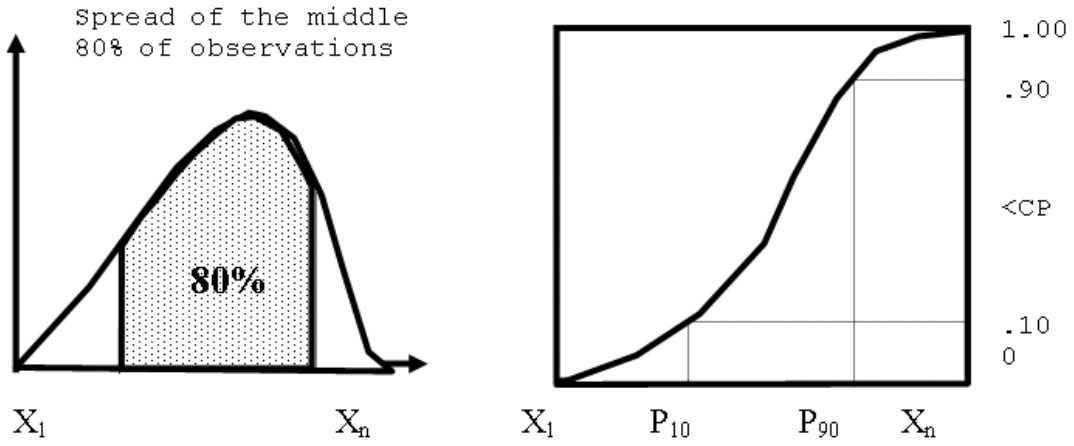
Using the ogive, the exact centre of the data can be determined. Start at the the .5 (50%) figure on the $<CP$ axis and extend a horizontal line to the left until the ogive curve is intersected. The value of X to which this corresponds is in the exact middle of the data with half the observations below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



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 graph) .36 and .68 respectively. The **difference** of these values (.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 $<CP = .10$ and $<CP = .90$ 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}).



Assignment: Sketch a histogram and ogive for each case study that is a grouped frequency distribution. Describe the shape of the histograms.

8 Measures of Central Tendency

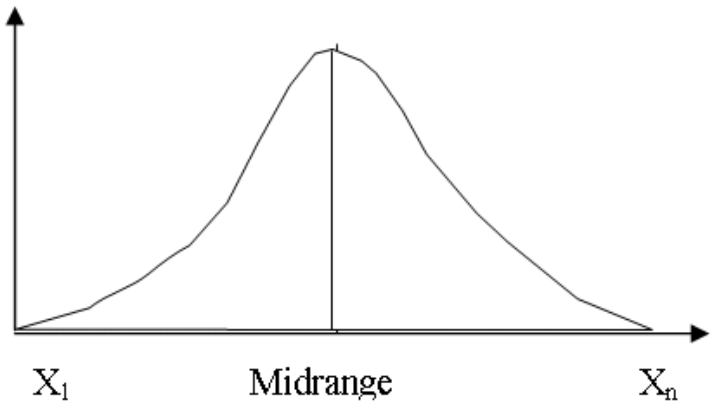
Consider an array of data for a statistical variable X. The following is a partial list of the 105 observations:

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

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 symmetrical distribution, a good 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.

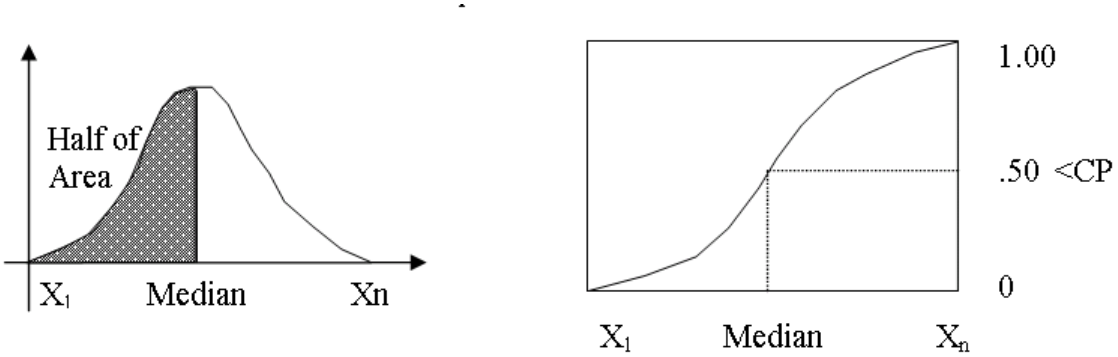
8.1 The Mode

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.

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. The median can be read directly off the ogive curve by locating the variable value corresponding to $<CP=.50=50\%$.

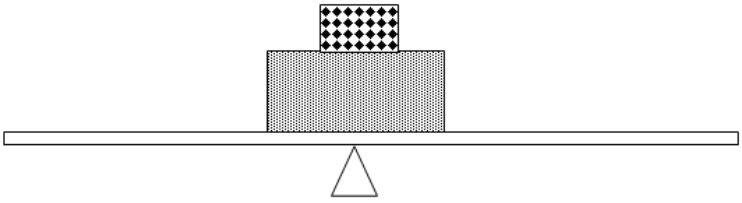


Some important characteristics of the median are:

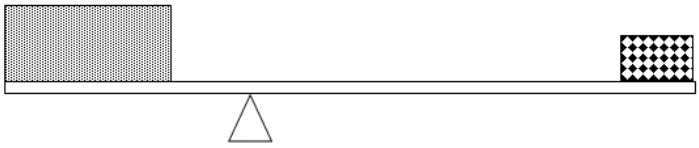
1. The median is representative of the data array because it is in the geometrical centre of the distribution. It is the exact halfway point. Half the observations are less and half are greater than the median value.
2. On the positively skewed data set, the median will be to the right of the mode and 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.

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:

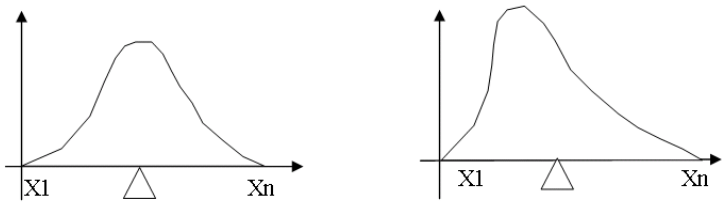


The balance point of the teeter-totter arrangement above is at the centre of the system.



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:

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

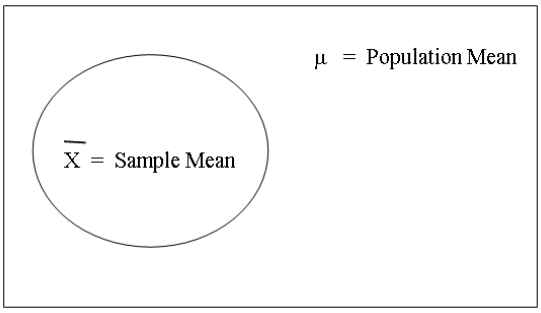
2. The location of the mean is dependent upon the shape of the distribution. It may not always be representative of the centre.
3. The arithmetic mean has mathematical properties that the other two averages do not have. (Balance point property).
4. Because of its mathematical properties, the mean lends 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.

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 symmetrical 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, median, and the mean for the distributions using only your histograms and ogives.

9 Calculating an Average for Ungrouped Data

Because an ungrouped frequency table preserves all the raw data from which it is assembled, the two cases are treated similarly when calculating the measures of central tendency.

9.1 The Mode

9.1.1 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.

9.1.2 Ungrouped Frequency Table

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.

9.2 The Median

9.2.1 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:

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

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

Example:

- 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 .

9.2.2 Ungrouped Frequency Table

The method of calculating the median is the same as for ungrouped 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 40 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 half way between the tenth and eleventh observation. From the $<Cf$ column, we see that both observations 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 before.

9.3 The Arithmetic Mean

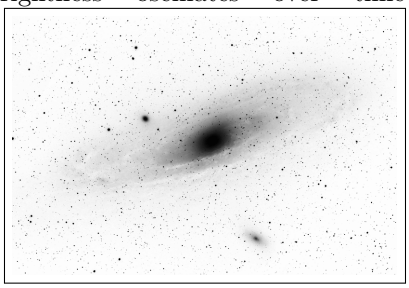
9.3.1 Raw Data

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*[5] (right) with the following results (in millions of light-years[6]):

2.6, 2.4, 2.5, 2.4, 2.7 (Mly) .



The Andromeda Nebula

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.

9.3.2 Ungrouped Frequency Tables

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.*

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

For ungrouped data the arithmetic mean formula is:

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

In our case:

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

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

The Number Of Decimal Place To Retain In An Average Calculation

Use **one place more precision** in the calculated average 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**. See the examples in Sections 9.3.1 and 9.3.2 respectively for application of these rules.

[5]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.

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

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 Section 9.3.1. For populations use the symbol μ as demonstrated in Section 9.3.2. 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 so far for the mean.

	Population	Sample
Raw Data	$\mu = \frac{\sum X}{N}$	$\bar{X} = \frac{\sum X}{n}$
Grouped Frequency	$\mu = \frac{\sum fX}{\sum f}$	$\bar{X} = \frac{\sum fX}{\sum f}$

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 5. The calculator has only one symbol to represent a mean because most often 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.

Assignment: For each case study that is raw data or an ungrouped frequency distribution calculate the mode, the median, and the arithmetic mean. Where a calculation is required (median and mean), remember to include the items below in your answer as demonstrated in the previous examples.

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

Formula State the formula used, including appropriate symbols to identify population or sample.

Substitution Substitute the values for your problem. Remember to add any columns to the table that are required for the calculation of those values (e.g. fX or $<Cf$).

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

Rounding Round to the appropriate number of decimal places discussed above.

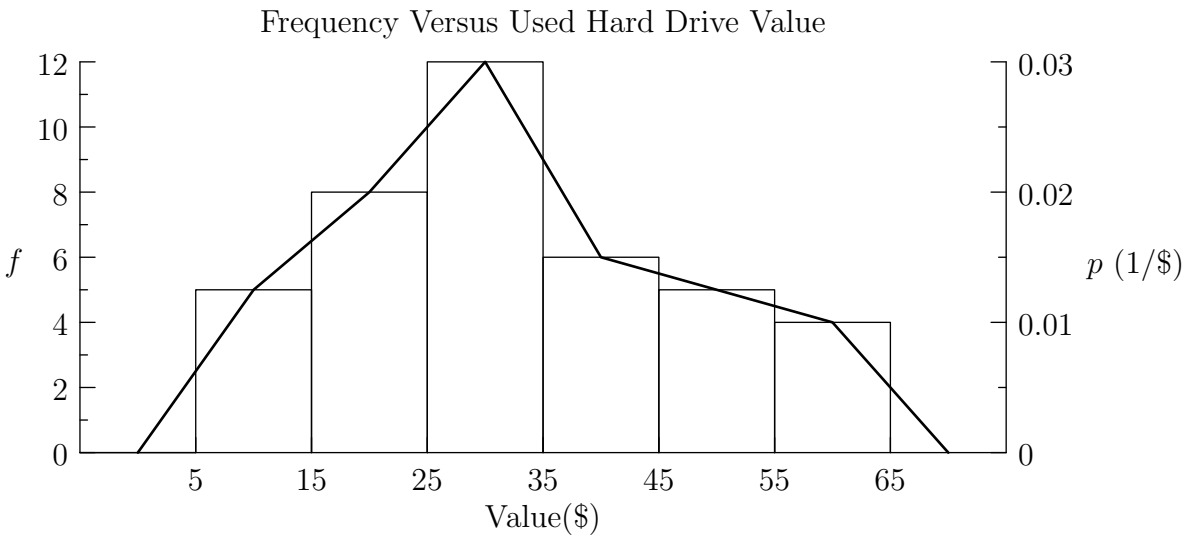
Units Include the appropriate units. Note that all the measures of central tendency have the same units as the statistical variable X.

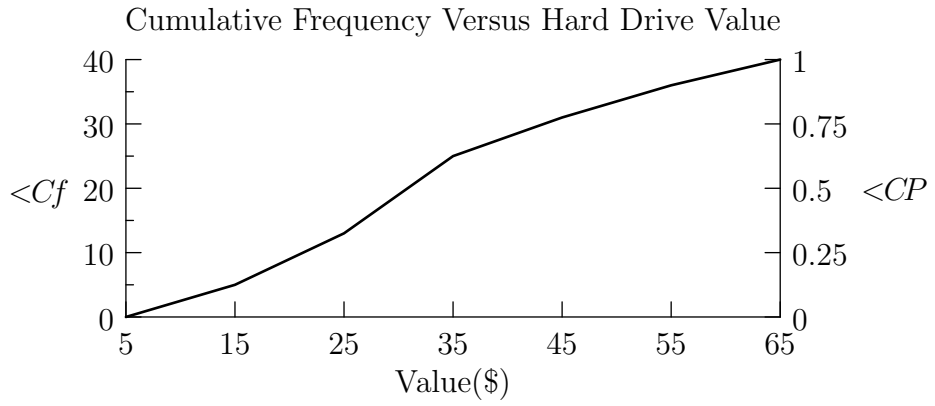
10 Calculating Average Values of Grouped Frequency Distributions

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 and the ogive curve.

Examine the summary table, frequency polygon, and ogive curves below:

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
	$\sum f = 40$	





10.1 The Mode Of Grouped Data

The mode is the value of the variable observed the most often. The **midpoint of the class with the highest frequency** is 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:
For the previous data:

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

10.2 The Arithmetic Mean Of Grouped Data

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 concentrated at 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 it was done for ungrouped summary tables.

Example:
Here is the calculation table for the previous data:

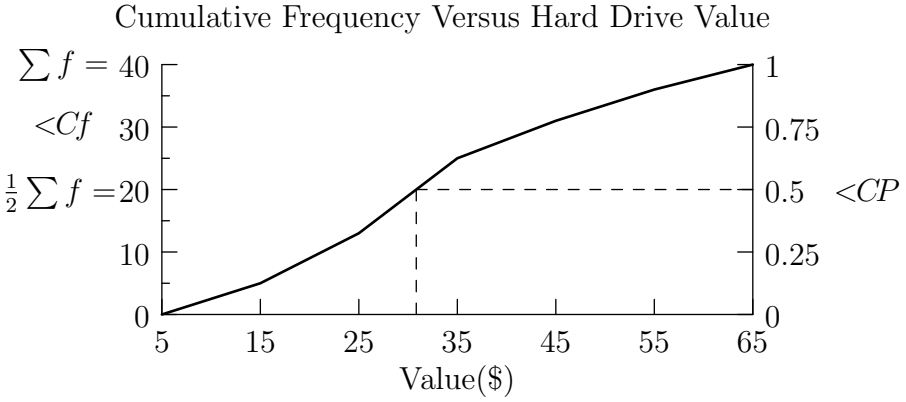
$X(\$)$	f	$fX(\$)$
10.00	5	50.00
20.00	8	160.00
30.00	12	360.00
40.00	6	240.00
50.00	5	250.00
60.00	4	240.00
	$\sum f = 40$	$\sum fX = 1300.00$

Approximate arithmetic mean:

$$\boxed{\bar{X} = \frac{\sum fX}{\sum f}} = \frac{\$1300.00}{40} = \$32.50$$

10.3 The Median Of Grouped Data

The median is that value in the data array located exactly in the middle of an ordered array. If the ogive curve is available, the median can be approximated by locating the value of the variable corresponding to the .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.)



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

If the ogive curve is not available, the median can be approximated by interpolating from the frequency distribution.

Example:
Find the median for the previous grouped frequency distribution.

Solution:
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 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

This gives the fractional equation:

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

Solving for the median:

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

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:

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

Assignment: For each case study that is a grouped frequency distribution calculate the mode, the median, and the arithmetic mean. Remember F.S.A.R.U. for all calculations.

11 The Weighted Mean ($\bar{X}_w$)

Sometimes values in a data array carry more "importance" than others. In this case the arithmetic mean cannot be computed by simply adding up all the values and dividing by the number of values. The importance of each value must be taken into account when summing. This is done by weighting the values according to importance before summing. The arithmetic mean for a frequency distribution can be thought of a weighted mean with the frequency being the weight. The weighted mean generalizes this concept to other weights as shown in the following example.

Example:
A student wrote 3 tests of various time lengths. On a $\frac{1}{2}$ hour test, the score achieved was 62%. On a 1 hour test, the score achieved was 95%. On a 3 hour test the score achieved was 51%. What is the student's mean score?

Set up the calculation tabularly with an X column representing the variable column, i.e. score, and a W column representing the weight of the variable value. The weighted mean score is given by:

$$\boxed{\bar{X}_w = \frac{\sum WX}{\sum W}}$$

X(%)	W(h)	WX(h%)
51	3.0	153
62	0.5	31
95	1.0	95
	$\sum W = 4.5$	$\sum WX = 279$

$$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{279 \text{ h\%}}{4.5 \text{ h}} = 62.00000\% = 62.0\%$$

Note:

- It would not make sense, in this case, to simply add the three marks and divide by 3.
- A common error is to mix up the X and W columns. Remember the X column is the quantity that is to be averaged, i.e. the variable.
- The units of the weight column, here hours, always cancel in the final result.

Assignment: Do the following exercise:

A lumberyard packages timber studs of different quality into a package for retail sale. 50% of the package contains first grade lumber that would sell for \$2.80 per stud. 30% of the package contains construction grade lumber that would sell for \$1.80 per stud. The remainder of the package is utility lumber that would sell for \$1.50 per stud.

1. What is the average value of a stud in the package?
2. Why can the prices not be added up and divided by 3 to give a meaningful average price?
3. If the package contained 50 studs, what should it sell for?

12 The Geometric Mean ($G.M.$)

Up to this point we have considered data that could be reordered (for instance ranked) without significant loss of information because the order has been unimportant. Such is the case for a repeated experiment under controlled conditions; the measurements should not depend on when the experiment is performed. A **time series** is a listing of a data array at specific instances of time where the time is important. For example, the amount of bacteria in a culture grows over time so the hour at which the measurement is made is significant. Usually the time periods at which the variable is measured are equal.

If a time series changes such that the *ratio* or *percentage change* of the variable between any two equal successive time periods is constant, the series is called a **geometric series**. For instance if a bacteria sample was growing such that subsequent measurements at one hour time intervals were always 1.2 times the previous measurement (a constant increase of 20%) this would be a geometric series. A plot of the variable versus time for a geometric series on semilogarithmic graph paper[7] is a straight line.

Example:
A person earns a salary to do a given job which changes over time as follows:

Year	Salary on Jan 1 ($)
1998	10,000
1999	11,200
2000	12,320
2001	12,566

Clearly the salaries constitute a time series as they show a definite trend (increasing) in time. Calculating the factors of increase (F) or equivalently the percent increases we see they are roughly constant as shown in the next table.

Year	Salary on Jan 1 ($)	F	% change
1998	10,000	–	–
1999	11,200	1.12	12
2000	12,320	1.10	10
2001	12,566	1.02	2

Here we used for the factor of increase:

$$F = \frac{\text{New Value}}{\text{Old Value}}$$

and for the percent change:

$$\%\text{ change} = \frac{\text{New Value} - \text{Old Value}}{\text{Old Value}} \cdot 100$$

Consideration of the F and % change columns show the following useful conversion formulae between them:

$$\boxed{F = 1 + \frac{\%\text{ change}}{100}} \quad \Leftrightarrow \quad \boxed{\%\text{ change} = (F-1)\cdot 100}$$

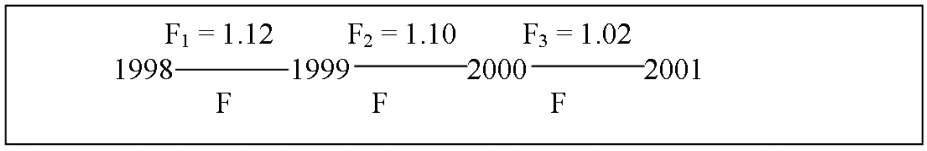
Now since they are roughly constant, we could consider taking the arithmetic mean of our factors of increase ($\frac{1.12+1.10+1.02}{3} = 1.08$) or equivalently the percent changes ($\frac{12+10+2}{3} = 8\%$). Unfortunately the arithmetic mean does not give a useful average when applied to a distribution of ratios or % values. To see why, consider what a constant factor of increase of 1.08 (percent increase of 8%) implies if we start at \$10,000:

Year	Actual Salary on Jan 1 ($)	Constant 8% Increase
1999	11,200	10,800
2000	12,320	11,664
2001	12,566	$\boxed{12{,}597}$

After three years of increase, the arithmetic mean produces a value of the variable (boxed) that differs from the actual variable.

[7] Semilogarithmic graph paper has an arithmetic horizontal scale and a logarithmic vertical scale. A logarithmic scale is one where a number goes on the scale at distance proportional to the logarithm of the number. The vertical scale starts at some chosen power of ten not at zero because the logarithm of zero is not defined. The distance between consecutive powers of ten on a logarithmic axis is referred to as one cycle.

The correct average to use in this case is called the **geometric mean**. By design, if the geometric mean is applied for n time intervals starting with the initial value, one will arrive at the same final value as the original data. The geometric mean assumes a constant factor of increase, F, across each time period. To see what value it must have, convert each annual % increase to its corresponding factor of increase and place the figures on the time line below:



Applying the constant factor F three times should have the same effect as applying the three variable factors.

$$F \cdot F \cdot F = F_1 \cdot F_2 \cdot F_3 \quad \text{or} \quad F^3 = F_1 \cdot F_2 \cdot F_3$$

Solve this for F:

$$F = \sqrt[3]{F_1 \cdot F_2 \cdot F_3} \ .$$

In this case[8]

$$F = \sqrt[3]{1.12 \cdot 1.10 \cdot 1.02} = 1.0791 \ ,$$

which converted to a % change using our formula above is 7.91%. This is the effective average annual % increase.

Generalizing our work above, a formula, which finds the effective % change across n equal time intervals in a time series is:

$$\boxed{G.M. = \left[\sqrt[n]{F_1 F_2 F_3 \ldots F_n} - 1 \right] \cdot 100} \ ,$$

where the F_i values are the variable factors of increase. Make sure to convert the % changes to factors of change before substitution using our above formula.

Example:
The annual sea ice minimum in the arctic ocean occurs at the end of each summer. Scientists have measured the percent change each year in this value from 2000 until 2007 and found the following results:

$$7.4, \ -12.1, \ 5.9, \ 0.0, \ -8.0, \ 4.0, \ -28.8 \ (\%)$$

Calculate the **average percent change per year** in the sea ice extent over this time period.

[8] Note that to evaluate the n^{th} root on your calculator you can either use the $\boxed{\sqrt[n]{\ }}$ key or take the exponent to (1/n). For example, for a cube root, raise the number to the power (1/3). Parentheses will be required for proper evaluation.

Solution:

$$\begin{aligned} G.M. &= \left[\sqrt[n]{F_1F_2F_3\ldots F_n} - 1\right]\cdot 100 \\ &= \left[\sqrt[7]{F_1F_2F_3F_4F_5F_6F_7} - 1\right]\cdot 100 \\ &= \left[\sqrt[7]{(1.074)(.879)(1.059)(1.000)(.920)(1.040)(.712)} - 1\right]\cdot 100 \\ &= [-.0533922]\cdot 100 \\ &= -5.33922\ \%/\text{yr} \\ &= -5.34\ \%/\text{yr} \end{aligned}$$

The sea ice extent **decreased** on average by 5.34 %/yr from 2000 to 2007. (Notice the effect of the signs in this problem both in the conversion to factors and in the interpretation of the answer.)

Another situation that can arise is when the initial amount and the final amount of the time series are given. The problem here is to find the constant % change per period required to affect the overall amount of change. Returning to our original problem, since the constant factor of F applied three times to the initial amount \$10,000 must result in the same final amount \$12,566, one has

$$F^3(\$10,000) = \$12,566 \quad \text{or} \quad F = \sqrt[3]{\frac{\$12,566}{\$10,000}} = 1.0791$$

This converted to a % change is 7.91% as before.

A general formula which will calculate the geometric mean in these types of situations is:

$$\boxed{G.M. = \left(\sqrt[n]{\frac{\text{Final}}{\text{Initial}}} - 1\right)\cdot 100}\ .$$

Example:
The amount of carbon dioxide in the atmosphere has increased steadily over the last 45 years from 315 ppmv (parts per million by volume) in 1960 to 380 ppmv in 2005. Calculate the **average percent change per year** in CO_2 over this time period.

Solution:

$$\begin{aligned} G.M. &= \left(\sqrt[n]{\frac{\text{Final}}{\text{Initial}}} - 1\right)\cdot 100 \\ &= \left(\sqrt[45]{\frac{380\ \text{ppmv}}{315\ \text{ppmv}}} - 1\right)\cdot 100 \\ &= (1.00417756 - 1)\cdot 100 \\ &= .417756\ \%/\text{yr} \\ &= .42\ \%/\text{yr} \end{aligned}$$

CO_2 has **increased** an average of .42 %/yr over this time period.

Assignment:

Do the following exercises involving the geometric mean.

1. A telephone company increased its rates by the following % values over the last 7 years:

   $$15, 2, 13, 10, 6, 7, 1$$

   What is the average annual % increase in phone rates over the past seven years?

2. A widget has had the following % change in prices over the past 4 years.

   $$20, -8, 12, -5$$

   What is the average annual % change in price of the widget? Be careful when converting % decreases into factors of change.

3. Mary had a salary of \$13,500 on July 1, 1982. By July 1, 1988, her salary had increased to \$18,000. What is the average annual % increase in salary that Mary has had over this time period?

4. The population of Averageville was 5800 on January 1, 1980. It has experienced a rapid growth in the last 8 years. On January 1, 1988 the population was 12,500. What is the average annual % increase in population over this time period?

5. Five students wrote an exam. Their marks were (%):

   $$56, 89, 76, 98, 59$$

   (a) Should the geometric mean be used to find the average score in % obtained by the students? Explain.

   (b) What average score in % should be reported for the students?

6. Sales at Ajax Company were \$45,000,000 last year. If total sales for this year are expected to be \$60,000,000, what is the average % increase per quarter expected to be?

13 Measuring Dispersion in a Distribution

13.1 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

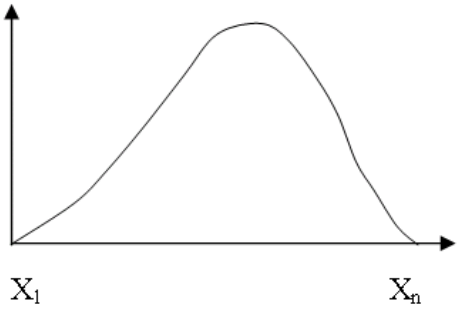
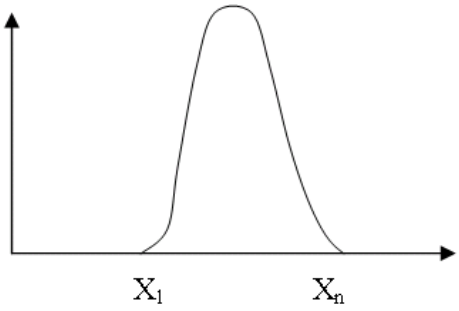


Figure 2



In the above diagrams note:

- There is more variation in distribution 1 than in 2.
- The range $(X_n - X_1)$ of distribution 1 is larger than in 2.

Figure 3

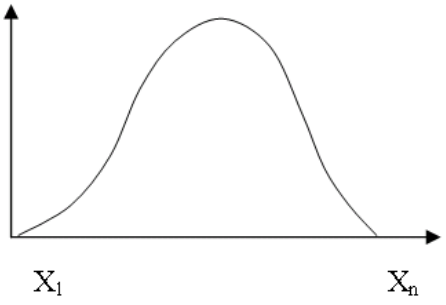
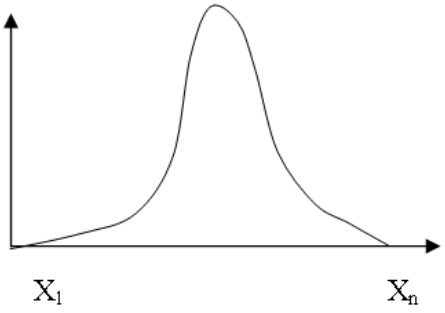


Figure 4



In the above diagrams note:

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

13.2 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, 41 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[9] 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.

13.3 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.

$$\boxed{\text{A.D.} = \frac{\sum |X - \bar{X}|}{n}} \quad \text{or} \quad \boxed{\text{A.D.} = \frac{\sum |X - \mu|}{N}}$$

[9] 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.[10]

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).

13.4 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*:[11]

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

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	$\sum (X-\mu)^2 = 502$

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

[10] 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.

[11] 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:

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

This also means that statistically we should see the majority of the observations within 11.2 kg of centre (the mean).

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). 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 the mathematical properties of the mean, the variance and standard deviation are more important. These two measures are found in theoretical distribution curves from which statistical inferences are drawn.

Statistical Keys On The Calculator

Recall the procedure for summing data using the statistical keys on the calculator in Section 5. 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 often 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 data if there is no separate key for it on your calculator. Report your answer to one place more precision than is found in the data set.

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.

14 Computing the Standard Deviation

14.1 The Computing Formula for Pop. Variance and Standard Deviation

It can be proven, by expanding $(X-\mu)^2$ algebraically,[12] 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 13, 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.

[12] 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.

14.2 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	$\sum X$	$\sum fX$
sum of squares of observations	$\sum X^2$	$\sum fX^2$

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**:

$$\boxed{\sigma^2 = \frac{\sum fX^2}{\sum f} - \left[\frac{\sum fX}{\sum f}\right]^2}$$

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

$$\boxed{\sigma = \sqrt{\frac{\sum fX^2}{\sum f} - \left[\frac{\sum fX}{\sum f}\right]^2}}$$

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	fX(rooms)	fX^2 (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
	$\sum f = 100$	$\sum fX = 380$	$\sum fX^2 = 1654$

The standard deviation is:

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

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

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

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

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

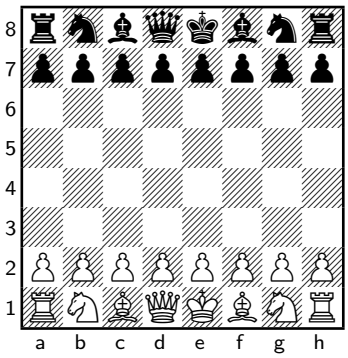
It is possible to calculate the standard deviation and variance by the formula based on the definition in frequency tables but because of the number of arithmetic operations required we will always use the computational formula for frequency distributions.

Statistical Keys on the Calculator

As seen in Section 5, 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 5. In that case the n, $\sum X$, and $\sum X^2$ calculator keys will generate the results for $\sum f$, $\sum fX$, and $\sum fX^2$ respectively.

Assignment:

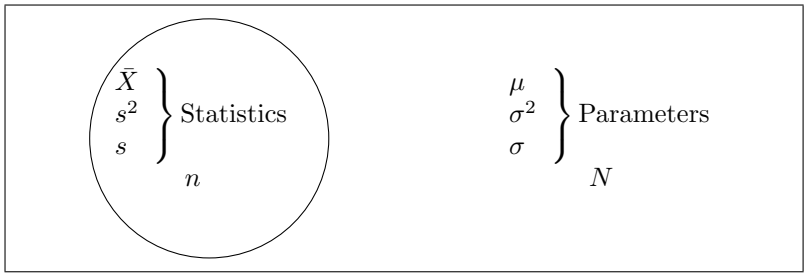
1. For each case study that is population data (including recalculating the result for raw data) calculate the standard deviation and variance using the computational formulae. Remember F.S.A.R.U. Check your results using the calculator.
2. The following diagram shows the setup of a chess board. Because 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.

15 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.

15.1 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. $\sum \left(X - \bar{X}\right)^2$ 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. The best available estimate of σ^2 is s^2 calculated by the **definitional formula for sample variance**:

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

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

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

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 9.3.1. Use the definitional formula.

Solution:
Continuing with our previous table we have:

Distance (Mly)	$\left(X-\bar{X}\right)$ (Mly)	$\left(X-\bar{X}\right)^2$ (Mly^2)
2.4	-.12	.0144
2.4	-.12	.0144
2.5	-.02	.0004
2.6	.08	.0064
2.7	.18	.0324
$\sum X = 12.6$		$\sum\left(X-\bar{X}\right)^2 = .068$

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:

$$s^2 = \frac{\sum\left(X-\bar{X}\right)^2}{n-1} = \frac{.068 \text{ (Mly}^2\text{)}}{5-1} = .017 \text{ (Mly}^2\text{)}$$

The sample standard deviation is its square root:

$$s = \sqrt{\frac{\sum\left(X-\bar{X}\right)^2}{n-1}} = \sqrt{\frac{.068 \text{ (Mly}^2\text{)}}{5-1}} = .130384 \text{ Mly} = .13 \text{ Mly}$$

Check that the standard deviation is reasonable with respect to the data. Does a standard deviation of $s = .13$ Mly in the distance to the stars tell you anything useful physically about the Andromeda Nebula?

15.2 Sample Standard Deviation By the Computing Formula

Similar to the discussion in Section 14.1, a mathematical equivalent to $\sum\left(X-\bar{X}\right)^2$ 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:

$$s = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}} = \sqrt{\frac{31.82\ (\text{Mly}^2) - \frac{(12.6\ \text{Mly})^2}{5}}{5-1}} = \sqrt{.017\ (\text{Mly}^2)} = .130384\ \text{Mly} = .13\ \text{Mly}$$

The sample variance is just the square of this value:

$$s^2 = .017\ (\text{Mly}^2)$$

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 14.2. The **computing formula** for the **standard deviation of a frequency distribution of sample data** is:

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

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)	fX(Watts)	$fX^2\ \left(\text{Watts}^2\right)$
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
	$\sum f = 30$		$\sum fX = 90$	$\sum fX^2 = 462$

The sample standard deviation is

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

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.

16 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.

16.1 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

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

of all the data elements. Within $k = 2$ standard deviations of the mean there is at least $1 - \frac{1}{2^2} = \frac{3}{4} = .75 = 75\%$ 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

$$\bar{X} - ks = 13 - (1.41)(9) = .3 \text{ kly}$$

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.)

16.2 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:[13]

- $|Z| \approx 0$: 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 9.3.1 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*.)

[13] 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.

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.

17 Other Statistics Derived from the Standard Deviation

17.1 The Coefficient of Variation (*C.V.*)

The smaller the value of the standard deviation, the less the variation in the distribution. When judging whether the standard deviation is large or small, the standard deviation must be compared to some value. The value used for comparison purposes is the mean. The statistical measure, which expresses the standard deviation as a percentage of the mean, is called the coefficient of variation.

$$\boxed{C.V. = \frac{\sigma}{\mu} \cdot 100} \leftarrow \text{For a population, or} \quad \boxed{C.V. = \frac{s}{\bar{X}} \cdot 100} \leftarrow \text{For a sample}$$

Example:
The mean age of a group of employees is 49.6 yrs with a standard deviation of 6.0 yrs. Calculate the *C.V.* .

Solution:

$$C.V. = \frac{\sigma}{\mu} \cdot 100 = \frac{6.0 \text{ yr}}{49.6 \text{ yr}} \cdot 100 = 12.1\% \quad \text{(age)}$$

The larger the coefficient of variation, the greater the relative dispersion. One can consider the following values to be typical[14]:

- $C.V. \approx 5\%$: **small** dispersion
- $C.V. \approx 15\%$: **moderate** dispersion
- $C.V. \gtrsim 25\%$: **large** dispersion

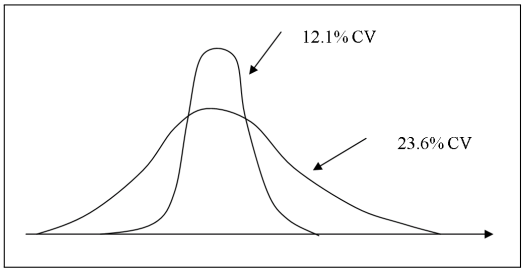
The coefficient of variation is particularly useful for comparing variations in different data sets.

Example:
The mean wage of the same group of employees is \$3600 with a standard deviation of \$850. Which set is more homogeneous, ages or wages?

Solution:
Calculating the coefficient of variation for the employee wages one finds:

$$C.V. = \frac{\sigma}{\mu} \cdot 100 = \frac{\$850}{\$3600} \cdot 100 = 23.6\% \quad \text{(wage)}$$

Comparing coefficients of variation, there is more variety (relative dispersion) in wages than in ages. This means that ages tend to be relatively more similar than wages for this group of people. Since *homogeneous* means similar in the context of data, we conclude that ages are more homogeneous than wages in this case. If the frequency polygon for each distribution were superimposed on the same graph and the scale of the variables chosen so that their



[14] These ranges are subjective. Others consider $C.V. < 100\%$ small and $C.V. > 100\%$ large. For purposes of this course the stated ranges will be used for interpreting the coefficient of variation.

means lay at the same place the curves may appear as in the diagram on the right.

Notice that the standard deviations cannot be directly compared because they have different units. That would be like comparing apples and oranges. 6.0 years would have to be compared to 850 dollars in terms of size.

17.2 The Coefficient of Skewness (S_k)

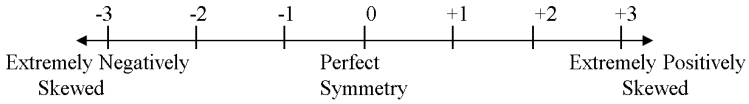
The coefficient of skewness measures the degree to which a distribution deviates from symmetry. Recall that in a perfectly symmetrical distribution the mean and the median are equal. It can be shown that in the worst case of skewness the mean and the median are separated by about 1 standard deviation. A measure of skewness based on this principle is the coefficient of skewness.[15]

$$\boxed{S_k = \frac{3\,(\mu - \text{Median})}{\sigma}} \leftarrow \text{For a population, or} \quad \boxed{S_k = \frac{3\left(\bar{X} - \text{Median}\right)}{s}} \leftarrow \text{For a sample}$$

The quantity, S_k, is a number that ranges from about -3 to +3. The **magnitude** of $S_k$tells you how skewed the distribution is:

- $|S_k| \approx 0$: close to **symmetric** (no skew)
- $|S_k| \approx 1$: **slightly** skewed
- $|S_k| \approx 2$: **moderately** skewed
- $|S_k| \approx 3$: **extremely** skewed

Assuming the distribution has a significant skew the **sign** of S_k tells you whether it is **positively skewed** (S_k positive) or **negatively skewed** (S_k negative). Combining magnitude and sign in one diagram gives:



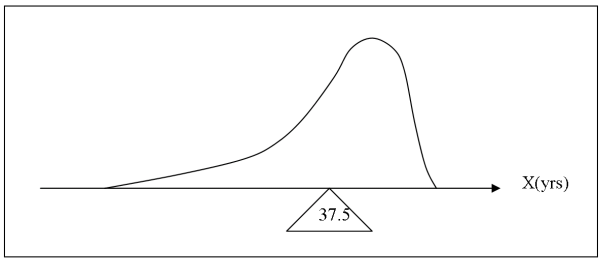
Example:
Suppose a group of 100 people had a mean age of 37.5 years, a median age of 44.0 years and a standard deviation of 9.0 years. Describe the shape of the distribution of ages.

Solution:

$$S_k = \frac{3\,(\mu - \text{Median})}{\sigma} = \frac{3\,(37.5\text{ yr} - 44.0\text{ yr})}{9.0\text{ yr}} = -2.16666 = -2.2$$

[15]This definition is called *Pearson's median or second skewness coefficient.* The more standard definition of skewness is defined as the third standardized moment γ_1 which, for a finite population is $\gamma_1 = \frac{1}{N}\sum\left(\frac{X-\mu}{\sigma}\right)^3$.

This means that the distribution of ages is moderately negatively skewed in shape. If its frequency polygon were drawn it might look like that at the right. It would have a pronounced tail on the left hand side of the range. In terms of the numbers in the distribution, this means that a large majority of the ages in the distribution are above the mean but that there are a few extremely small values that are pulling the mean down.



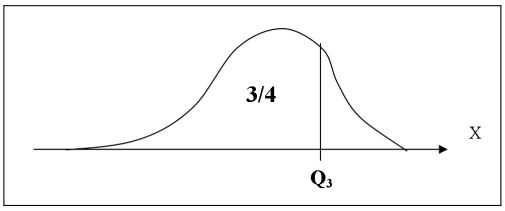
Assignment: For each case study calculate the coefficient of variation and the coefficient of skewness. Interpret your values.

18 Fractional Measures of Position

18.1 Fractiles

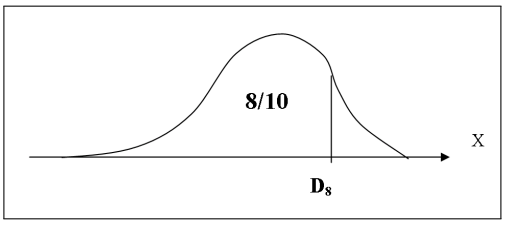
Fractiles are positional measures within a data set. Just as the median measures the halfway point, there are other fractional positions that can be computed. The fractions in common use are quarters, tenths and hundredths. Quartiles are quarter divisions, deciles are tenths divisions and percentiles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



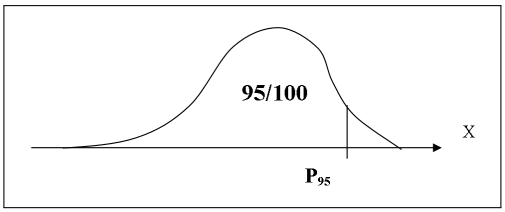
The Third Quartile

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?

18.2 Calculating Fractile Values

Fractile values are calculated in an **identical manner** to the way in which the **median** was calculated.[16] (See Sections 9.2.1, 9.2.2, and 10.3.) The only difference is that the fraction used to find the location is not $\frac{1}{2}$ but depends on the fraction in the fractile. On the formula sheet, we will use the two 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.

[16] 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.

18.2.1 Raw Data and Ungrouped Frequency Distributions

Rank order the data and locate the value occupying the position.

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

$$56,\ 70,\ 78,\ 65,\ 62,\ 71,\ 56,\ 63,\ 65,\ 71$$

Find the $\mathbf{10^{th}}$ **percentile**.

Solution:
In rank order we have:

$$56,\ 56,\ 62,\ 63,\ 65,\ 65,\ 70,\ 71,\ 71,\ 78$$

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

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 example 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 and ungrouped frequency distributions is that a real interpolation is often required, as demonstrated in the following example.

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:

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

18.2.2 Grouped Frequency Distributions

Example:
A survey of the price of a certain size of refrigerator was made with the results at left. Find the price at the **third 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:

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

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

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

This gives the fractional equation:

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

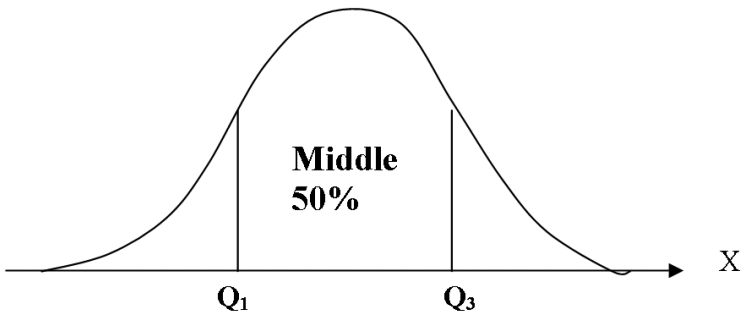
Solve for the third quartile value:

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

18.3 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.

18.3.1 Interquartile Range (IQR)



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

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

Example:
Find the **interquartile range** for the example of Section 18.2.2.

Solution:
We have already found $Q_3 = \$815.00$. We need Q_1. This time we'll use the formula to find it.

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_3 lies in the \$600-\$700 class.

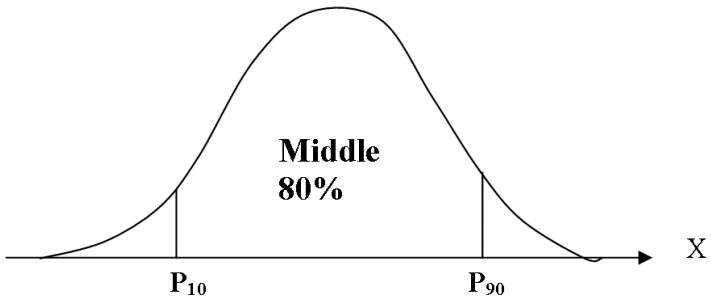
Using the median formula for grouped frequency distributions suitably modified for the first quartile gives:

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

Finally the interquartile range is:

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

18.3.2 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 first example in Section 18.2.1 .

Solution:
We already found $P_{10} = 56.0$ cm. We need P_{90}. First find its position:

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

Since $X_9 = 71$ cm and $X_{10} = 78$ cm we must interpolate to find $P_{90} = X_{9.9}$:

Variable Value	Position
$7 \text{ cm}\left(\begin{array}{l} P_{90} - 71 \text{ cm}\left(\begin{array}{r} 71 \text{ cm} \\ P_{90} \\ 78 \text{ cm} \end{array}\right. \end{array}\right.$	$\left.\begin{array}{r} 9 \\ 9.9 \\ 10 \end{array}\right) 0.9 \Bigg) 1$

This gives the fractional equation:

$$\frac{P_{90} - 71 \text{ cm}}{7 \text{ cm}} = \frac{0.9}{1}$$

Solve for the ninetieth percentile value:

$$P_{90} = 71 \text{ cm} + \frac{0.9}{1} \cdot (7 \text{ cm}) = 77.30000 \text{ cm} = 77.3 \text{ cm} .$$

Finally the 90-10 percentile range is

$$\text{10-90 } PR = P_{90} - P_{10} = 77.3 \text{ cm} - 56.0 \text{ cm} = 21.3 \text{ cm} .$$

The interdecile range, IDR, has the same value.

Note:

- If an ogive graph is available, all fractiles may be easily read from it and these ranges calculated.
- We could have defined an interpercentile range by $P_{99} - P_1$. Why might that not be so useful?

Assignment: For each case study calculate the fractiles and fractile ranges requested.

19 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) Calculate the **mode**.

(k) Calculate the **median**.

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

(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 **coefficient of variation**. Interpret your answer.

(r) Calculate the **coefficient of skewness**. Interpret your answer.

(s) 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) Calculate the **mode**.

(k) Calculate the **median**.

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

(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 dataset within this interval satisfies the inequality.

(q) Calculate the **coefficient of variation**. Interpret your answer.

(r) Calculate the **coefficient of skewness**. Interpret your answer.

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

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).) Also add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$). Remember to **sum** any column for which it is appropriate.

(l) Calculate the **mode**.

(m) Calculate the **median**.

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

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

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

(q) Calculate the **coefficient of variation**. Interpret your answer.

(r) Calculate the **coefficient of skewness**. Interpret your answer.

(s) 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).) Also add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$). Remember to **sum** any column for which it is appropriate.

(l) Calculate the **mode**.

(m) Calculate the **median**.

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

(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 $k = 2$. Show that the **actual percentage** of data elements in the dataset within this interval satisfies the inequality.

(q) Calculate the **coefficient of variation**. Interpret your answer.

(r) Calculate the **coefficient of skewness**. Interpret your answer.

(s) 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).) Also add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$). Remember to **sum** any column for which it is appropriate.

(l) In the space below sketch a **histogram** and an **ogive** for the data. On your histogram also draw a **frequency polygon**.

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

(n) By looking at only your histogram and ogive, estimate the **mode**, the **median**, and the **arithmetic mean**. Indicate your reasoning on the graphs.

(o) Calculate the **mode**.

(p) Calculate the **median**.

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

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

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

(t) Calculate the **coefficient of variation**. Interpret your answer.

(u) Calculate the **coefficient of skewness**. Interpret your answer.

(v) 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).) Also add columns for **cumulative frequency** ($<Cf$) and **cumulative relative frequency** ($<CP$). Remember to **sum** any column for which it is appropriate.

(l) In the space below sketch a **histogram** and an **ogive** for the data. On your histogram also draw a **frequency polygon**.

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

(n) By looking at only your histogram and ogive, estimate the **mode**, the **median**, and the **arithmetic mean**. Indicate your reasoning on the graphs.

(o) Calculate the **mode**.

(p) Calculate the **median**.

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

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

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

(t) Calculate the **coefficient of variation**. Interpret your answer.

(u) Calculate the **coefficient of skewness**. Interpret your answer.

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

Answers

page 15: Frequency Distribution Construction:

1. needs ungrouped: tally and check case study 4
2. needs grouped:

   Step 1) $R = 99 \text{ km/h} - 3 \text{ km/h} = 96 \text{ km/h}$

   Step 2) $N = 1 + \log(30)/\log(2) = 5.91$ classes

   Step 3) $\Delta X = 16.75$ (km/h)/class

   Step 4) Choose $\Delta X = 20$ (km/h)/class, first class: $[0 \text{ km/h}, 20 \text{ km/h}]$

   Step 5) See case study 6 for remaining classes

   Step 6) Tally data and check case study 6

page 45: Weighted Mean:

1. $\bar{X}_w = \$2.24$
2. values are unequally represented
3. \$112.00

page 49: Geometric Mean:

1. 7.60 %/yr
2. 4.11 %/yr
3. 4.91 %/yr
4. 10.07 %/yr
5. (a) not time series

   (b) the arithmetic mean, 75.6%
6. 7.46 %/qtr

page 58:

1. See individual case study answers.
2. $\mu = 2.6$ points, $\sigma = 2.2$ points, $\sigma^2 = 4.9$ points^2

page 74: 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) Median = \$12.00

(l) $\bar{X} = \$21.37$

(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) $C.V. = 93\%$ large variation

(r) $S_k = 2.82$ extreme positive skew

(s) $Q_1 = \$7.50$, $Q_3 = \$50.00$, $IQR = \$42.50$

page 78: 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) median = 74.5%

(l) $\mu = 75.3\%$

(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) $C.V. = 10\%$ moderate variation

(r) $S_k = 0.32$ very slight positive skew

(s) $D_7 = 80.8\%$

page 82: 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) $\sum fX = 200$, no

(k) Add P, $<Cf$, $<CP$. $\sum P = 1.00$

(l) mode = 4.0

(m) median = 4.0

(n) $\bar{X} = 4.0$

(o) $s = 1.2$, $s^2 = 1.3$

(p) $Z = 1.67$ Roll is 1.67 standard deviations (moderately) above the mean.

(q) $C.V. = 30\%$ moderate variation

(r) $S_k = 0.00$ symmetric

(s) $P_{10} = 2.1$, $P_{90} = 6.0$, 10-90 $PR = 3.9$

page 85: 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) $\sum fX = 58$ years, no

(k) Add P, $<Cf$, $<CP$. $\sum P = 1.00$

(l) mode = 2.0 years

(m) median = 2.5 years

(n) $\mu = 2.9$ years

(o) $\sigma = 1.4$ years, $\sigma^2 = 1.9$ (years^2)

(p) at least 75% of data between 0.1 yr and 5.7 yr, $\frac{18}{20} = 90\% > 75\%$

(q) $C.V. = 48\%$ large variation

(r) $S_k = 0.86$ slight positive skew

(s) $P_{66} = 2.9$ yr

page 88: 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) $\sum fX = \$15120.00$, yes

(k) Add P, p, $<Cf$, $<CP$. $\sum P = 1.000$

(l) See Figure 1 .

(m) symmetric

(n) From histogram, mode $\approx \$100.00$, $\mu \approx \$100.00$. From ogive, median $\approx \$100.00$

(o) mode = \$100.00

(p) median = \$100.39

(q) $\mu = \$100.80$

(r) $\sigma = \$48.98$, $\sigma^2 = 2399.36$ (\$^{2})

(s) $Z = 0.00$ Purchase price equals the mean.

(t) $C.V. = 49\%$ large variation

(u) $S_k = 0.03$ symmetric

(v) $D_1 = \$30.00$, $D_9 = \$172.73$, $IDR = \$142.73$

page 92: 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) $\sum fX = 1100.0$ km/h, yes

(k) Add P, p, $<Cf$, $<CP$. $\sum P = 1.000$

(l) See Figure 2 .

(m) positively skewed

(n) From histogram, mode ≈ 30 km/h, $\mu \approx 36$ km/h. From ogive, median ≈ 32 km/h

(o) mode = 30.0 km/h

(p) median = 31.7 km/h

(q) $\bar{X} = 36.7$ km/h

(r) $s = 24.3$ km/h, $s^2 = 588.5$ km^2/h^2

(s) $Z = 4.66$ Wind speed is 4.66 standard deviations (extremely) above the mean.

(t) $C.V. = 66\%$ large variation

(u) $S_k = 0.62$ slight positive skew

(v) $Q_3 = 52.5$ km/h

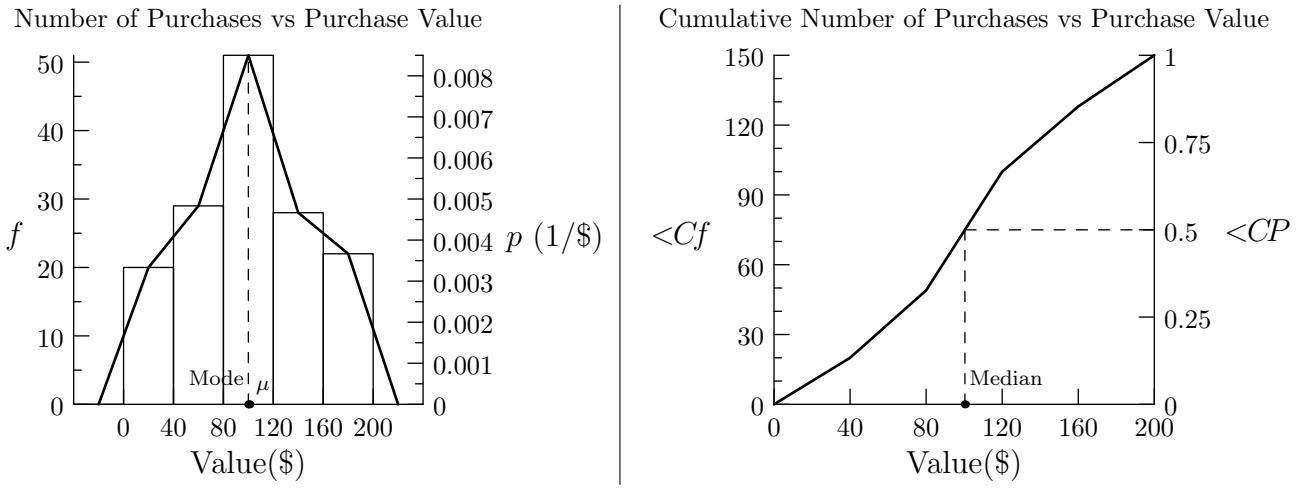


Figure 1: Histogram and ogive for web startup case study.

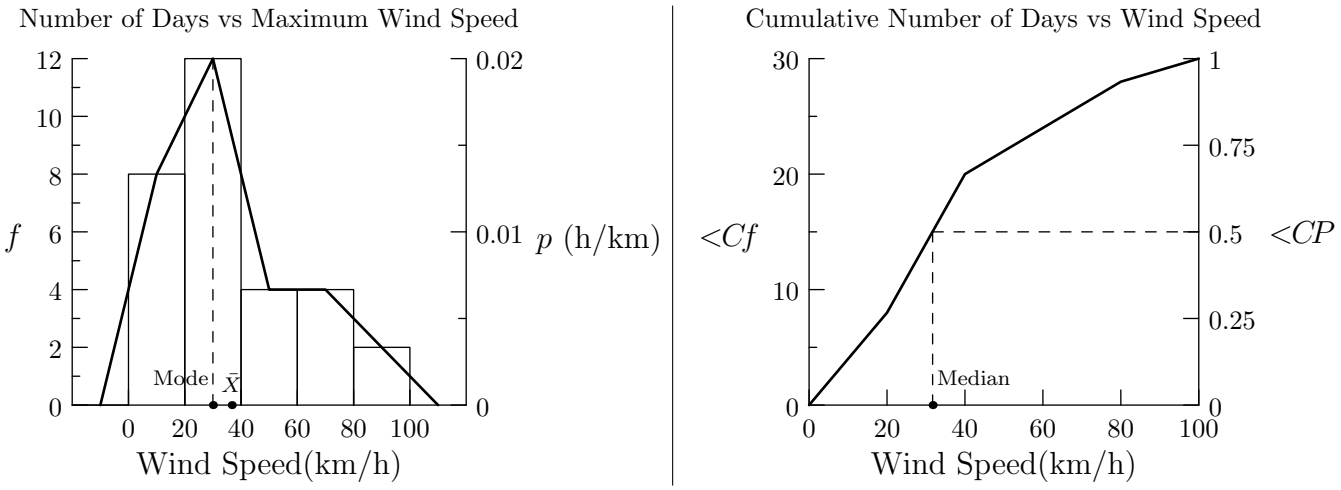


Figure 2: Histogram and ogive for wind speed case study.

Descriptive Statistical Measures

Raw Data	Ungrouped Frequency Distribution	Grouped Frequency Distribution
Measures of the Centre:		
$\mu = \dfrac{\sum X}{N}$ or $\bar{X} = \dfrac{\sum X}{n}$	$\mu, \bar{X} = \dfrac{\sum fX}{\sum f}$	$\mu, \bar{X} = \dfrac{\sum fX}{\sum f}$
$\underset{\text{Position}}{\text{Median}} = \dfrac{1}{2}(N+1)$ 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 $X_{\frac{1}{2}(n+1)}$	$\underset{\text{Value}}{\text{Median}} = X_{\frac{1}{2}(\sum f+1)}$	$\underset{\text{Value}}{\text{Median}} = L_i + \dfrac{\left\{\frac{1}{2}\left(\sum f\right) - <Cf_{i-1}\right\}}{f_i} \cdot \Delta X$
Measures of Dispersion:		
$A.D. = \dfrac{\sum \lvert X-\mu \rvert}{N}$ or $\dfrac{\sum \lvert X-\bar{X} \rvert}{n}$		
$\sigma = \sqrt{\dfrac{\sum (X-\mu)^2}{N}}$		
$\sigma = \sqrt{\dfrac{\sum X^2}{N} - \left[\dfrac{\sum X}{N}\right]^2}$	$\sigma = \sqrt{\dfrac{\sum fX^2}{\sum f} - \left[\dfrac{\sum fX}{\sum f}\right]^2}$	$\sigma = \sqrt{\dfrac{\sum fX^2}{\sum f} - \left[\dfrac{\sum fX}{\sum f}\right]^2}$
$s = \sqrt{\dfrac{\sum \left(X-\bar{X}\right)^2}{n-1}}$		
$s = \sqrt{\dfrac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$	$s = \sqrt{\dfrac{\sum fX^2 - \frac{(\sum fX)^2}{\sum f}}{\sum f - 1}}$	$s = \sqrt{\dfrac{\sum fX^2 - \frac{(\sum fX)^2}{\sum f}}{\sum f - 1}}$
IQR=Q_3-Q_1	IDR=D_9-D_1	10-90 PR=P_{90}-P_{10}

Other Statistical Measures:

$G.M. = \left[\sqrt[n]{F_1F_2F_3\ldots F_n} - 1\right] \cdot 100$	$G.M. = \left(\sqrt[n]{\dfrac{\text{Final}}{\text{Initial}}} - 1\right) \cdot 100$
$\bar{X}_w = \dfrac{\sum WX}{\sum W}$	$Z = \dfrac{X-\mu}{\sigma}$ or $Z = \dfrac{X-\bar{X}}{s}$
$C.V. = \dfrac{\sigma}{\mu} \cdot 100$ or $C.V. = \dfrac{s}{\bar{X}} \cdot 100$	$S_k = \dfrac{3(\mu - \text{Median})}{\sigma}$ or $S_k = \dfrac{3\left(\bar{X} - \text{Median}\right)}{s}$

Miscellaneous Formulae:

$R = X_n - X_1$	$N = 1 + \log(n)/\log(2)$		$\Delta X = \dfrac{R}{N}$
$P = \dfrac{f}{\sum f}$	$p = \dfrac{P}{\Delta X}$	$F = 1 + \dfrac{\%\text{ change}}{100}$	$1 - \dfrac{1}{k^2}$

GNU Free Documentation License

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.