Cuyahoga Community College MATH 1410 – Elementary Probability and Statistics I – Summer 2025 Review

Timothy Masso

August 15, 2025

Contents

1 Introduction and Background

In pursuit of my career change to data analytics, I found it necessary to pursue classes and certifications in applied mathematics and statistics. The class Elementary Probability and Statistics I - Summer 2025, is one class in the TRI-C *Data* Analytics - Post-Degree Professional Certificate program. Which I am taking to add to my application into a Masters of applied Math and Statistics at Bowling Green State University for fall 2026. This Review is my first LaTeX document, and I am using this paper to get comfortable in the medium.

In 2024, I graduated with a degree in Jazz and Contemporary Music from the College of Performing Arts at The New School. While completing my musical studies, I realized that I had a passion for data when I took a liberal arts class that had the students work with different types of data visualization formats and programs used to make those visualizations. e.g., using leaflet and open street maps to display geojson data taken from studies and surveys. We were approaching the projects in an artistic way rather than an analytical way; as we did not do any formal analysis on the data we gathered from various sources. After graduating and moving back from New York City to Cleveland, I had a good time to evaluate my career and the options to pursue my desired future To get my feet wet and a general overview of data analytics I saw that as a resident of Ohio I had the benefit to take qualifying ”microcredentials”. To my fortune, there were credentials in different levels of Excel, Minitab, Power BI, and data analytics as a whole. Most of these were short day or multi day courses with lectures from industry professionals working at these professional training companies.

After taking multiple classes I decided that I want to fully commit and go back to school. The post degree data analytics program at Tri-C seemed like a perfect start. As the classes were not too expensive, and I could take them starting when I signed up. During the first couple weeks I realized that to plan for a masters in my near future I needed to get into contact with the schools as soon as possible to find how to patch the holes I have in my application considering I have no college level math classes during my music studies. The first task was to find programs I want to apply to. Such as the applied mathematics and statistics masters a t Bowling Green State University. After talking to the program coordinators they gave me the classes I needed to show proficiency in moving forward. This review is not about those classes but about a class, but some context is needed in potential errors in this review as it is also a teaching moment for me to internalize these topics.

All of this context brings us to the class at hand; Elementary Probability and Statistics I. The textbook the class was given was *Statistics: Informed Decisions Using Data 6E* written by Michael Sullivan, III. Published by Pearson. This course was a 5 week accelerated class so all the material and finals were taken within 5 weeks. This review mainly serves as a way for me to get familiar with LaTeX and fully internalize the class. I will hopefully continue to do this for every class after.

Another thing to get out of the way, as the audience for this paper is anyone that wants to read it; is in the context of statistics, *Data* is plural. It took me some getting used to, just as you will have to get used to it if you have only ever referred to *Data* as singular. The overall goal of this is to review and fully internalize the class and have practice writing in LaTeX.

2 Chapter 1

The subject of Chapter 1 was *Data* collection.

Introduced to us were 4 main components of data collection.

- Statistics is the science of collecting, organizing, summarizing and analyzing information to draw conclusion or answer questions.
- Statistics is about providing a measure of confidence in any conclusions
- We must report a measure of our confidence in our results because we do not have 100
- Using subsets of an entire group the analyst is studying leads to error by sampling bias

The thing being collected, organized, summarized and analyzed, is *Data*. *Data* is a fact or a proposition used to draw a conclusion or make a decision. *Data* describe characteristics of an individual. *Data* varies in the individual just as within a group. For example; humans have data, their height, weight, eye color, etc.

The goals of using statistics to describe *Data* is to:

- Describe the variability
- Understand sources of variability

The process and approach are the most important.

Certain terms must be familiar to move forward in *Statistics*.

- The entire group to be studied is called the *Population*
- A subset of the *Population* being studied is called a *Sample*
- A person or object that is a member of the *Population* being studied is an *Individual*
- A *Statistic* is a numerical summary of a *Sample*
- A *Parameter* is a numerical summary of a *Population*
- *Variables* are the characteristics of the Individuals within the Population
- The list of observed values for a variable is *Data*

An oversimplification of a process is taking a *Sample* from the *Population.*

Descriptive *Statistics* consist of organizing and summarizing *Data.* Descriptive *Statistics* describe *Data* through numerical summaries, tables and graphs. While we can describe *Sample*, we can not make conclusions about the *Population.*

Differential *Statistics* uses methods that take a result from a *Sample*, extend it to the Population, and measure the reliability of the result.

The way this is done is to identify Variables. Variables have two main distinctions. Qualitative and Qualitative. Qualitative or Categorical variables allow for classification of individuals based on some attributes or characteristic. While Quantitative variables provide numerical measure of individuals. The values of Quantitative variables can have arithmetic processes on them and provide meaningful results. Each Variable corresponds to their data type as well. Qualitative variable corresponds to qualitative data and so on

Quantitive variables have two further distinctions:

- Discrete Variables have either a finite number if possible values or a countable number of possible values.
- Continuous Variables have an infinite number of possible vales that are not countable

A quick way to distinguish between the two is that discrete variables are counted and continuous are measured.

The type of variable dictates the methods that can be used to analyze data.

Within *Data* collection and Variables. The levels or measurement of variables are divided into four categories.

- *Nominal*
- *Ordinal*
- *Interval*
- *Ratio*

Variables are classified as *Nominal* if the values of the variable name, label, or category in addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order. Think name, as the word *Nominal* comes from the Latin word ”Nomen”, which means name.

Variables are classified as *Ordinal* when they have properties of the *Nominal* level, along with that the naming scheme allows for the values of the variable top be arranged in a ranked or specific order. Think order for *Ordinal.*

Variables are classified as *Interval* when they have the properties of the *Ordinal* level of measurement and the different values of the variable have meaning. Along with that a value of zero does not mean the absence of the quantity. Addition and subtraction can also be performed on the values of the variable.

Variables are classified as *Ratio* when they have the properties of the *Interval* level of measurement and the ratios of the values of the variable have meaning. Along with that a value of zero means the absence of the quantity. Multiplication and division can be performed on these variables.

It is noted that Nominal Or *Ordinal* Variables are also Qualitative Variables and *Interval* or ratio Variables are also Quantitative Variables. Some examples of each level of measurement are:

- *Nominal*: gender, Male or female (dated example but still works if thought in binary)

- *Ordinal*: grade in a class, can categorize them from high to low, but if you only get and an A and someone else gets a B you don't know how much better you did than them, all you know is that it's a category away.

- *Interval*: temperature, we can say $60° > 50°$, but 0° is still a temperature

- *Ratio*: the number of days a college student studies in a week, 0 days studied means no studying and the student can say that they studied more days than someone else

2 Observational And Designed Experiments are the experiments done to make statistics about a *Population.* To quickly define experiment in this context. An Experiment is a controlled study conducted to determine the effect of varying one or more explanatory variables or sometimes called factors, has on a response variable. Any combination of values of the factors is called a treatment. The Experimental unit is a person, object or some other well-defined item upon which a treatment is applied.

An Observational Study measures the value of a response variable and explanatory variables (both of which will be discussed later). The researcher observes the behavior of the individuals without trying to influence the outcome of the study. In Observational studies, we are not allowed to make statements of causality, we can not say that changes to the explanatory variable causes changes in the response variable. We can only say changes in the explanatory variable are associated with changes in the response variable. In short the researcher is only allowed to claim association, not causation.

A Designed Experiment consists of a researcher randomly assigning the individuals in a study to groups, intentionally manipulates the value of the explanatory variable while controlling other explanatory variables at fixed values. Proceeding with recording the value of the response variable for each individual. In Observational Studies and Designed Experiments the goal is to see how an explanatory variable affects a response variable. An explanatory variable is the characteristic that is controlled or distinguished. While the response variable is the characteristic that will change depending on the value of the explanatory variable.

With either experiment there are things to look out for. Confounding is the first scenario to look out for. Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted by the study.

With the context of Confounding, A lurking variable is an explanatory variable that was not considered in a study, but affects the value of the response variable. It typically is related to the explanatory variables that are considered in the study. Following the Lurking variable, Confounding Variables are explanatory variables that are not considered in a study whose effect can not be distinguished from a second explanatory variable in the study.

Observational Studies are divided into three types

- Cross-Sectional
- Case-Control
- Cohort

Cross Sectional studies collect information about individuals at a specific point in time, or over a very short period of time.

Case-Control Studies are retrospective, meaning that they require individuals to look back in time or require the researcher to look at existing records. In these types of studies individuals that have certain characteristics are matched with those that do not, along with that there is a chance the records may be flawed or missing parts.

Cohort Studies first identifies a group of individuals to participate in the study, referred to as the cohort. The cohort is then observed over a period of time. Over this time period characteristics about the individuals are recorded. Since *Data* are collected over time, cohort studies are prospective.

Moving forward it is necessary to define the term used to describe the group taken from the *Population.* That group is called a Sample. Now to make the most unbiased and most representative *Sample* from a *Population.* Researchers must create a Simple Random Sample using the process of random sampling. The process of Random Sampling uses chance to select individuals from a *Population* to be included in the *Sample.* To do this effectively one must create a Frame. A list of all the individuals in the *Population* of interest. Then from a Population of size N use Random Sampling defined by every possible *Sample* size of n having an equally likely chance of occurring to create Simple Random Samples of size n.

Within Experiments the researcher has to decide the degree of blindness. Blinding refers to the nondisclosure of the treatment an experimental unit is receiving. There are two degrees of blindness. A single Blind experiment is one in which the experimental unit (or subject) does not know which treatment it is receiving. But the researcher does know what the experimental unit is receiving. A Double-blind experiment is one in which neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.

Steps Of A Designed Experiment

To conduct an organized Designed experiment the steps to follow are defined.

1. Identify the problem to be solved
2. Determine factors that affect that response variable
   (a) Factors are explanatory variables the researchers conducting the experiment believe impact the value of the response variable. Once the factors are identified it must be determined which factors are to be fixed at some predetermined level. In other words which factors will be manipulated and which factors will be uncontrolled.
3. Determine the number of experimental units
4. Determine the level of explanatory variables
   (a) Control: two ways to control factors
      i. Fix the level of a factor at once value throughout the experiment this is a factor whose impact on the response variable does not have interest in the experiment
      ii. Set the level of a factor at predetermined levels. This is a factor whose impact on the response variable does interest the experiment

   (b) Randomize: Randomly assign experimental units to various treatment groups. This "averages out" the effects of uncontrolled explanatory variables.
5. Conduct the experiment
   (a) Replication occurs when each treatment is applied to more than one experimental. This helps to assure that the effect of a treatment is NOT due to some characteristic of a single experimental unit. It is recommended that each treatment group have the same number of experimental units.

   (b) Collect and process the data by measuring the value of the response variable for each replication. Any differences in the value of the response variable can be attributed to the differences in the level of treatment.
6. Test the claim

Experimental Designs have group designs, they consist of Matched pairs and Randomized Block Design. A Matched Pairs design is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way. For example can be comparing an individual before and after. Within the Matched pairs design there are only two levels of treatment. The term that identifies the next process is Blocking. When similar experimental units are grouped together and then randomly assigning the experimental unit within each group to a treatment. Inside that, each group of homogeneous individuals is called a Block. The other design that takes is process is the Randomized Block Design. It is used when the experimental units are divided into the homogeneous groups (Blocks). Within each Block the experimental unis are randomly assigned to treatments.

3 Chapter 2

With the knowledge of chapter 1 we can go into the details of defining *Samples.* Known as organizing Quantitative data. To review there are two types of *Data.* First we will organize Qualitative data. Qualitative data provide measures that categorize or classify an individual. When raw qualitative data are collected, we often first determine the number of individuals within each category.

A Frequency Distribution lists each category of data and number of occurrences for each category of data.

Category	Number of Occurrences
0	5
1	3
2	1
3	10
4	8
5	0

(Frequency Distribution Example)

The relative frequency is the *Population* or percent of observations within a categroy and is found using the formula:

$$\text{Relative Frequency}_i = \frac{f_i}{\sum f_i}$$

Where $_i$ is the individual values of the category. f_i is the number of occurrences of the category. $\sum f_i$ is the sum of all number of occurrences.

A Relative frequency Distribution lists each category of data together with the relative frequency. The sum of all the values is 1. It can be displayed in the Freq distribution table as shown with $\sum f_i = 20$:

Category	Number of Occurrences	Relative Frequency of Category
0	5	$.25 = \frac{5}{20}$
1	4	$.20 = \frac{4}{20}$
2	1	$.05 = \frac{1}{20}$
3	2	$.10 = \frac{2}{20}$
4	8	$.40 = \frac{8}{20}$
5	0	$.0 = \frac{0}{20}$

(Frequency Distribution with Relative frequency distribution Example)

We can construct a bar Graph by labeling each category of data on either the horizontal(X) or vertical(Y) axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represent the categories frequency or relative frequency.

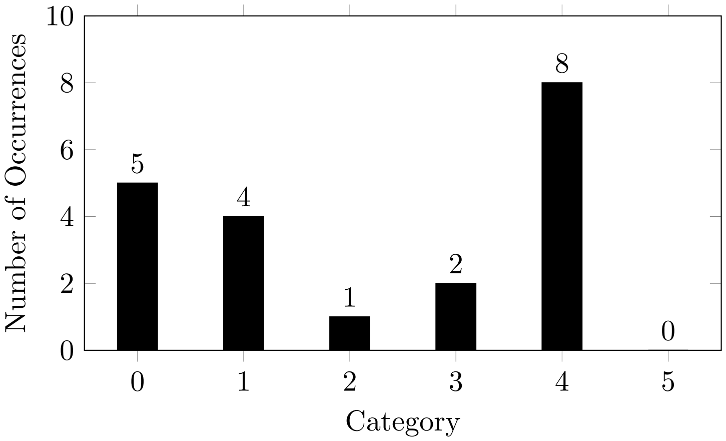


Figure 1: Category vs Number of Occurrences

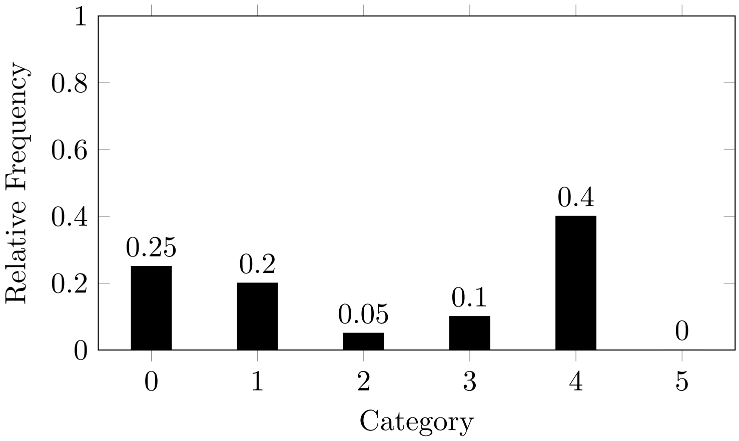


Figure 2: Category vs Relative Frequency

The specific bar chart used is a Pareto Chart, which is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

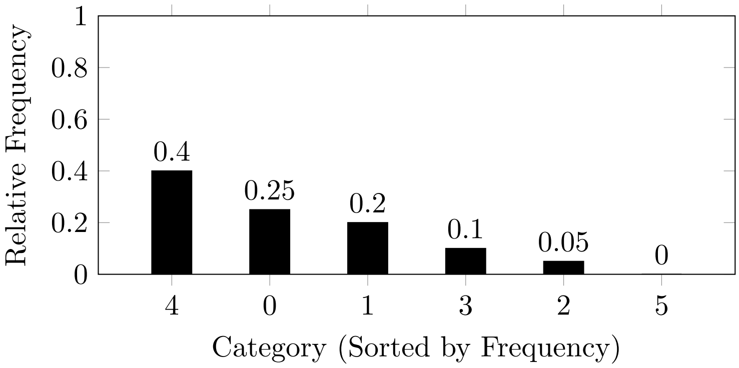
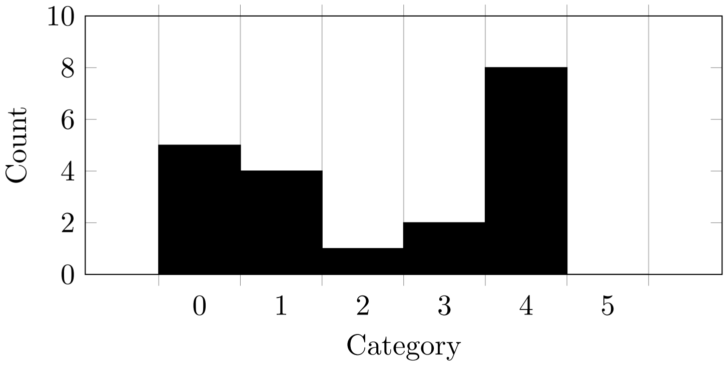
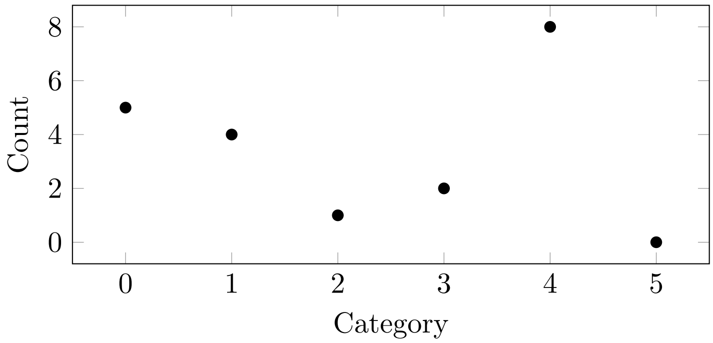


Figure 3: Pareto Chart: Relative Frequency by Category (Sorted)

Another chart used is a Histogram. It is constructed by drawing rectangles for each class of data. The height Of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.



Another plot used frequently is a dot plot, which is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.



We organize Quantitative Data into Discrete or continuous data. If the data is discrete and a few values of the variable then the category (class) will be the observation. If the data is continuous or discrete with a lot of values for the variable, then the classes must be created using intervals of numbers.

Within the data we categorize the data into groups when a data set consists of many different discrete data values or when a data set consists of continuous data. These groups are called Classes.

Classes have a lower class limit, an upper class limit and a class width.

- Lower Class Limit: the smallest value within a class
- Upper Class Limit: the largest value within a class
- Class width: the difference between consecutive lower class limits

An exception to the requirement of equal class widths occurs in open-ended tables. Which is defined by if the first class has no lower class limit or has no upper class limit.

To determine the lower class limit of the first class we should choose the smallest observation in the data set or a convenient number slightly lower than it. For the class width, there should generally be between five and twenty classes. The smaller the data set the fewer the classes there should be. This relates to class width as class width is defined by this equation.

$$\text{Class Width} = \frac{\text{(Largest data value)}-\text{(Smallest data value)}}{\text{Number of classes}}$$

For convenience we round up to the nearest whole number, but may result in fewer classes then were originally intended.

Now that we know the graphs used in these chapters we can look into identifying the shape of plotting distributions on the graphs.

Figure 4: Uniform distribution

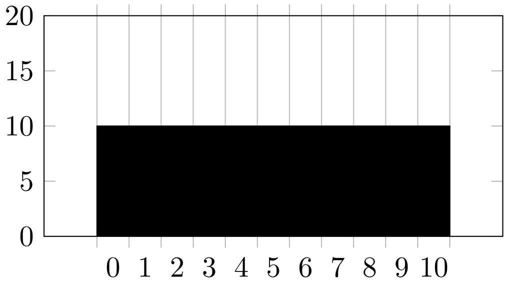
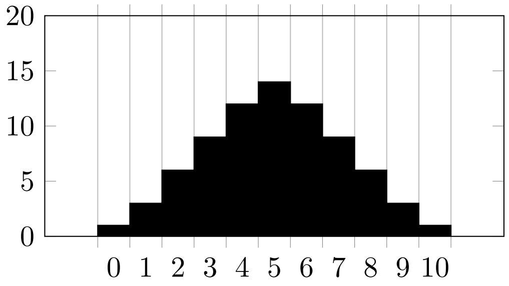


Figure 5: Bell-shaped distribution



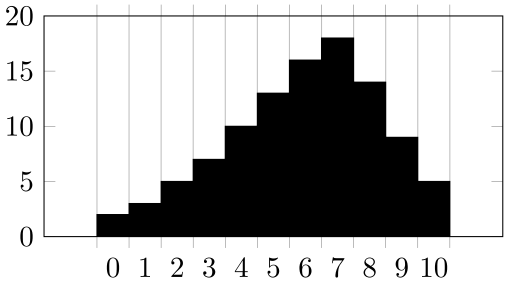


Figure 6: Skewed left distribution

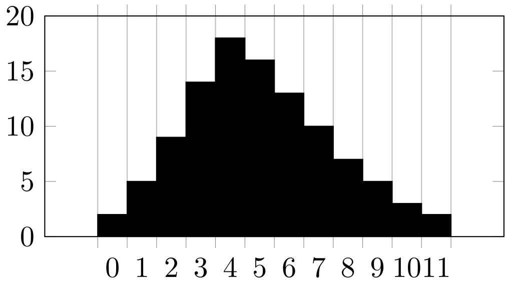


Figure 7: Skewed right distribution

4 Chapter 3

Chapter 3 is all about central tendency. When numerically summarizing data, we look at the characteristics of the data; such as the shape, center, and spread. The center of the data set is commonly called the average. A measure if central tendency numerically describes the "average" or "typical" data value. Other terms that describe central tendency are the proper term for average; Mean, along with median, and mode.

Arithmetic mean is generally referred to as the mean. The Arithmetic mean of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations. The population arithmetic mean, μ, is computed using all the individuals in a population. The population mean is a parameter. It is given by this formula:

$$\mu = \frac{x_1+x_2+...+x_n}{N} \text{ simplified as } \mu = \frac{\sum x_i}{N}$$

The Sample arithmetic mean, $\bar{x}$, is computed using the sample data. The sample mean is a statistic.

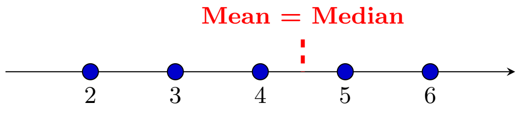
The median of a variable is the value that lies in the middle of the data when arranged in ascending order. The median is represented by the capital letter M. More detailed steps are as follows:

1. Arrange Data in ascending order
2. Determine the number of observations, n
3. Determine the observation in the middle of the data set
   (a) If the number of n is odd, then the median is the data value exactly in the middle of the data set, precisely at the $\frac{n+1}{2}$ position
   (b) If n is even, the median is the mean of the observations that lie in the $\frac{n}{2}$ position and the $\frac{n}{2}+1$ position

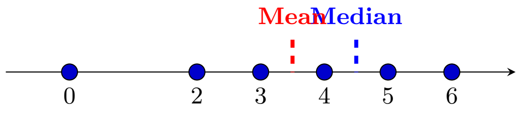
Mean VS Median

To understand the differences between mean and median when they are used to describe central tendency we need to go over Resistance. Resistance is the degree to a metric is affected by extreme values, big or small. The mean is *not* resistant while the median *is* resistant

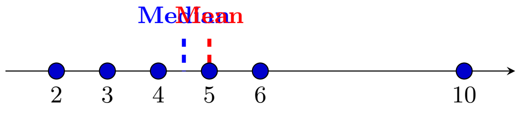
Original Data (No Outlier)



With Left Outlier (0)



With Right Outlier (10)



The plots demonstrate that the **median remains stable** regardless of the outlier position, while the **mean shifts** toward the outlier. This shows that the median is a *resistant* measure of center, unlike the mean.

The Range, R of a variable is the difference between the largest and smallest data value. Written out looks like this:

$$\text{Range} = R = \text{Largest data value} - \text{Smallest data value}$$

The population standard deviation of a Variable is the square root of the sum of the squared deviations about the population mean divided by the number of observations in the population, N. The Standard of deviation measures how spread out the values are from the population mean. It's represented by this formula:

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

Where x_i are the N observations in the population and μ is the population mean.

The computational formula is represented as:

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

Where $\sum x_i^2$ means to square each observation and then sum these squared values. $(\sum x_i)^2$ means to add up all the observations and then square the sum.

The Sample Standard Deviation of variable is the square root of the sum of the squared deviations about the sample mean divided by $n-1$, where n is the sample size. It is represented by the formula:

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

Where x_i is are the n observation in the sample and $\bar{x}$ is the sample mean. For some context that will be expanded upon later is that, $n-1$ has to do with degrees of freedom and bias, n has one observation that has no degree of freedom so you have to take it away, so every observation has a degree of freedom. The nth value has no freedom. It must be whatever value forces the sum of the deviations about the mean to be 0.

The Standard deviation represents the "typical" deviation from the mean. As, such the standard deviation may be used to judge whether a particular observation is far away from the mean of the data set. For example is a measure of 31 cm is far from 25 cm it depends, if the standard deviation of the data is 6 cm then the answer is no because 31 cm would only be 1 standard deviations away from 25 cm. But if the standard deviation is 2 cm the 31 cm would be far as it would be 3 standard deviations away from 25 cm. The rule of a thumb is an observation is far if it is more than 2 Standard deviations away from the other observations such as mean. When comparing two populations the larger the standard deviation. The greater the dispersion or spread of the distribution provided the variable of the interest from the two population has the same unit of measurement.

Another way to describe the data is variance. The variance of a variable is the square of the standard deviation. The population variance is σ^2 and the sample variance is s^2. The variance is in squared units.

If a distribution is roughly bell shaped, then we can approximate the following:

1. Approximately 68% of the data will lie within 1 standard deviation of the mean
   (a) Written out looks like

$$68\% \text{ lie between } \mu - 1\sigma \text{ and } \mu + 1\sigma$$

2. Approximately 95% of the data will lie within 2 standard deviations of the mean
   (a) Written out looks like

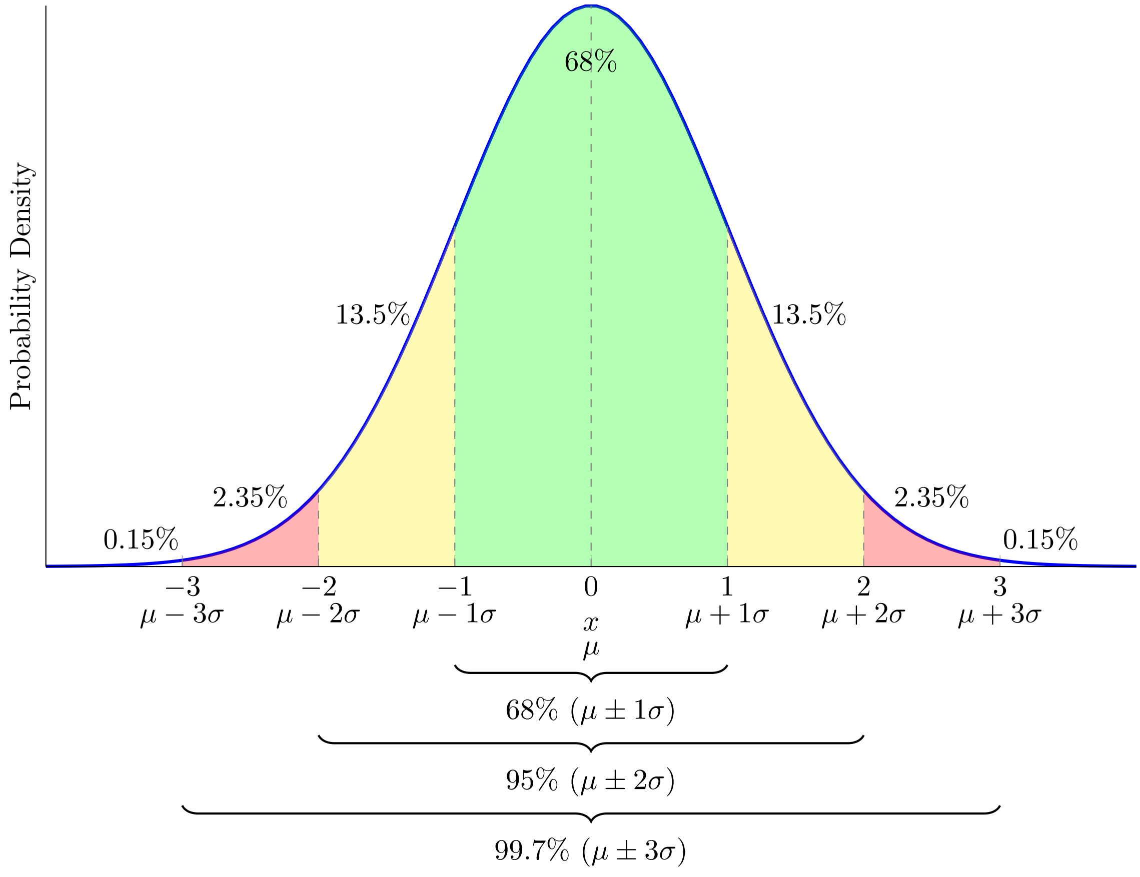
$$95\% \text{ lie between } \mu - 2\sigma \text{ and } \mu + 2\sigma$$

3. Approximately 99.7% of the data will lie within 3 standard deviation of the mean
   (a) Written out looks like

$$99.7\% \text{ lie between } \mu - 3\sigma \text{ and } \mu + 3\sigma$$

These approximations form the empirical rule. Which work on population distributions and sample distributions.

Empirical Rule with $\mu \pm \sigma$ and Percentages



To find the standard of deviation of an observation

$$\frac{x - (\mu \text{ or } \bar{x})}{\sigma \text{ of the whole distribution}}$$

Chebyshev's Inequality

Chebyshev's Inequality is used to determine the minimum percentage of observations that lie within k standard deviations of the mean where $k > 1$. It can be used on any distribution shape. For any data set or distribution, at least $(1 - \frac{1}{k^2}) * 100\%$ of observations lie within k standard deviation of the mean. $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$

When observations have different importance or weight associated with them, we compute the weighted mean. For example GPA is a weighted mean, with the weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value. The weights mean $\bar{x_w}$ of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this um by the sum of the weights. The formula is:

$$\bar{x_w} = \frac{\sum w_i x_i}{\sum w_i}$$

Where w_i is the weight of the ith observations and x_i is the value of the ith observation

Suppose a student has the following grades and credit hours:

Grade Point Scale

A=4.0, B=3.0, C=2.0, D=1.0, F=0.0

Example Table

Course	Letter	x_i	w_i
Math	A	4.0	4
History	B	3.0	3
Biology	C	2.0	3
Art	A	4.0	2
Economics	D	1.0	3

x_i is the grade point value for course i, and w_i is the number of credit hours for course i.

Calculation

$$\sum w_i x_i = (4 \times 4.0) + (3 \times 3.0) + (3 \times 2.0) + (2 \times 4.0) + (3 \times 1.0)$$

$$= 16 + 9 + 6 + 8 + 3 = 42$$

$$\sum w_i = 4 + 3 + 3 + 2 + 3 = 15$$

$$\bar{x}_{\text{GPA}} = \frac{42}{15} \approx 2.80$$

The student's GPA is **2.80**.

Measures of Central Tendency and Dispersion From Grouped Data

We approximate the mean of a variable from grouped data, because raw data cannot be retrieved from a frequency table, we assume that within each class the mean of the data is equal to the class mid-point. We then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. We repeat this process for each class and add the results. This sum approximates the sum of the data.

The formula for the grouped population mean is as follows:

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$ Where x_i is the midpoint of the value of the ith class. And f_i is the frequency of the ith class.

The Grouped Sample Mean is as follows:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$ Where x_i is the midpoint of the value of the ith class. And f_i is the frequency of the ith class

We approximate the Standard deviation of a variable from grouped data. The procedure for approaching the standard of deviation from grouped data is similar to the of finding the mean from grouped data, because we don't have access to the original data the standard of deviation is approximate. The formula for population and sample standard of deviation of grouped data are as follows:

$$\text{Population} = \sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}} \text{ and } \text{Sample} = s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}}$$

Where x_i is the midpoint of the value of the ith class, f_i is the freq of the ith class.
An algebraically equivalent formula for the population standard deviation is:

$$\text{Population} = \sigma = \sqrt{\frac{\sum (x_i^2) - \frac{\sum (x_i f_i)^2}{\sum f_i}}{\sum f_i}}$$

Approximated medians are found by:

1. Construct a cumulative frequency distribution
2. identify the class in which the median lines
   (a) remember, the median can be obtained by determining the observation that lies in the middle
3. Interpolate the median using the formula

The formula is as follows:

$$\text{Median} = M = L + \frac{\frac{n}{2} - CF}{f} * i$$

Where L is the lower class limit of the class containing the median. n is the number of data values in the frequency distribution. CF is the cumulative frequency of the class immediately preceding the class containing the median. F is the frequency of the median class. i is the class width of the class containing the median

Measures of Position

We use Z-scores to represent the distance that a data value is from the mean in terms of the number of standard of deviations. We find it by subtracting the mean from the data value and dividing this result by the standard of deviation. There is both a population Z-score and a sample Z-score.

The population Z-score is found the by the following formula:

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

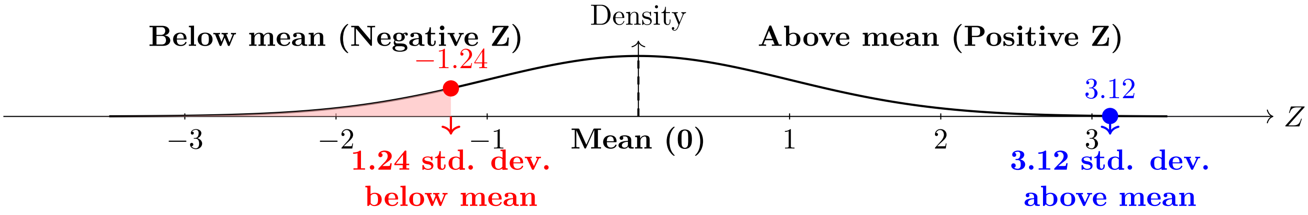
The Sample Z-score is found the by the following formula:

$$Z = \frac{x - \bar{x}}{s}$$

With for both equations, x is the data value. We can rearrange this formula algebraically to find x with knowing $Z, \mu, \bar{x}, \sigma$

The Z-score is unitless and has a mean of 0 and a Standard deviation of 0. If the data value is larger than the mean, the Z-score is positive. If the data value is smaller than the mean, then the Z-score is negative. If the data value equals the mean, the Z-score is 0. A Z-score measures the number of standard deviations an observation is above or below the mean. For a practical example, before this we could only say that a value is far or close to the mean, but now we can quantify this value.

If a Z-score of data value is -1.24 the data value is 1.24 standard deviations below the mean. A Z-score of 3.12 is 3.12 standard deviations above the mean.



Intercept Percentiles

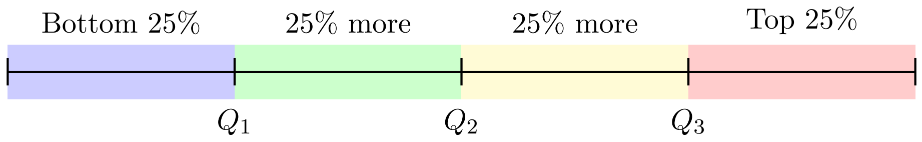
We can recall that the median divided the lower 50% of the set data from the upper 50%. The median is a special case of a general concept called Percentile. The kth percentile denoted P_k of a set of data is a value such that k percent of the observations are less than or equal to the value. So percentiles divide a set of data that is written in ascending order into 100 parts. Thus, 99 percentiles can be determined. For example p_1 divides the bottom 1% of the observations from the top 99%. P_2 divides the bottom 2% of the observations from the top 98% and so on.

An example a score of 600 in the SAT being in the 74th percentile mean 74% of the SAT scores are less than or equal to 600 and 26% of the scores are greater. So 26% of the students who took the SAT scored better than 600. Each year a score of 600 will be a different percentile because the score population changes.

Intercept Quartiles

The most common percentiles are the quartiles. Quartiles divide data sets into fourths, or four equal parts.

1. The first quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile
2. The second quartile, denoted Q_2, divides the bottom 50% of the data from the top 50%. Therefore, the first quartile is equivalent to the 50th percentile
3. The third quartile, denoted Q_3, divides the bottom 75% of the data from the top 25%. Therefore, the first quartile is equivalent to the 75th percentile



To find these Quartiles we need to:

1. Arrange the data in ascending order
2. determine the median, M or the second quartile
3. divide the data set in halves
   (a) Q_2 the observations below (left) M and the observations above (right) M.
   (b) The first quartile Q_1, is the median of the bottom half of the data and the third quartile Q_3, is the median of the top half of the data

If the number of observations is odd, do not include the median when determining Q_1 and Q_3 by hand.

Range standard deviation and variance are measures of dispersion, but they are *not* resistant. Quartiles *are* resistant. For this reason, quartiles are used to define a fourth measure of dispersion.

The Interquartile Range, IQR if the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartile and is found using the formula:

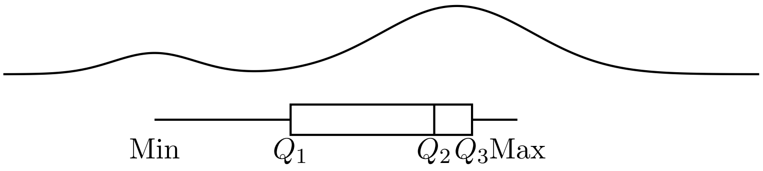
$$IQR = Q_3 - Q_1$$

The interpretation of the interquartile range is similar to that of the range and standard deviation. That is, the more spread a set of data has, the higher the IQR will be.

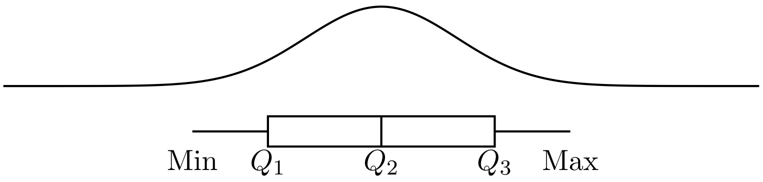
We can find Skewness based on the dimensions of the quartiles.

1. Left Skewed
   - $Q_2 - Q_1 > Q_3 - Q_2$
2. Symmetric
   - $Q_2 - Q_1 \approx Q_3 - Q_2$
3. Right Skewed
   - $Q_2 - Q_1 < Q_3 - Q_2$

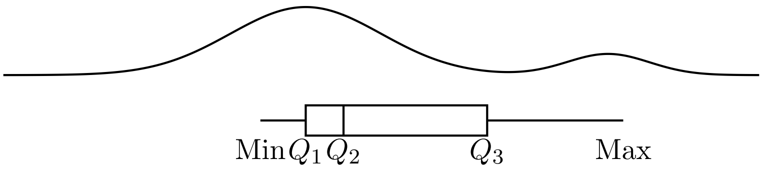
Left-skewed



Symmetric



Right-skewed



Checking a set of data for outliers (extreme observations). Outliers can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling. They aren't always errors, sometimes extreme observations are common within a population. However, we need to heed to caution, outliers distort both the mean and the standard deviation because neither is resistant. Because these measures often form the basis for most statistical inference. Any conclusions drawn from a set of data that contains outliers can be flawed.

Checking for Outliers consist of 4 steps:

1. Determine the first and third quartiles
2. compute the interquartile range
3. Determine the fences, the fences serve as cut off points for determining outliers
   - Upper fence = $Q_3 + 1.5(IQR)$
   - Lower fence = $Q_1 - 1.5(IQR)$
4. If the data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

The Five Number Summary and Box Plots

The previous sections have shown us how to summarize the data to see what it can tell us. We explore the data to see if they contain interesting information that may be useful in our research. The summaries make this exploration much easier because these summaries represent an exploration. A term to define it is the material exploratory data analysis. So far we have only collected info and presented summaries, not reached any conclusions.

Compute Five Number Summary

Median is a measure of central tendency that divides the upper 50% from the bottom 50%. It is resistant to outliers and is the preferred measure of central tendency when data are skewed right or left. As range, standard deviation and variance are not resistant. Interquartile Range is resistant but Q_1, median and Q_3 does not give us info about the extremes of the data.

The five number summary of the data consists of the smallest data value, Q_1, median, Q_3, and the largest data value. It looks as follows:

$$\text{Minimum}, Q_1, M, Q_3, \text{Maximum}$$

Draw and Interpret Box Plots

The five number summary can be used to create another graph called a box plot.

Drawing Box Plot

1. Determine the lower and upper fence
   - Upper fence = $Q_3 + 1.5(IQR)$
   - Lower fence = $Q_1 - 1.5(IQR)$
     - Where $IQR = Q_3 - Q_1$
2. Draw a number line long enough to include the minimum and maximum values.

3. Insert vertical lines at Q_1, M, Q_3 and enclose those vertical lines in a box

4. Label the upper and lower fences

5. Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called Whiskers

6. Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk $*$

Below are examples:

Figure 8: Bell-shaped

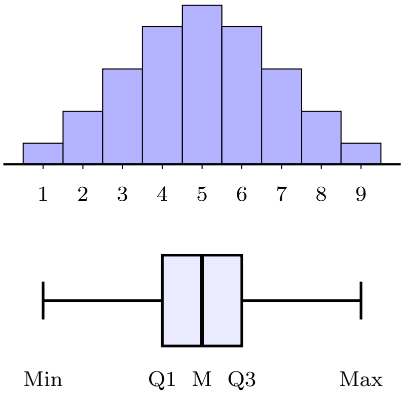


Figure 9: Skewed right

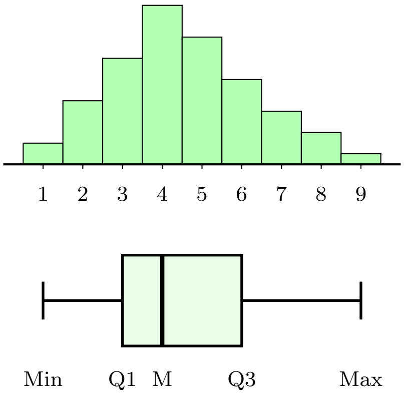
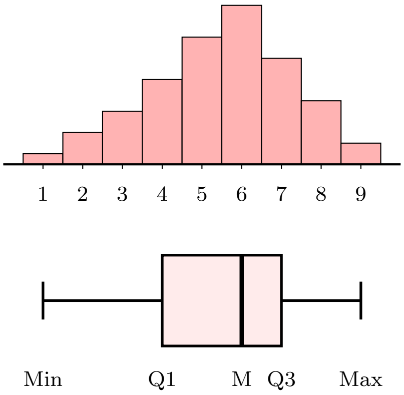


Figure 10: Skewed left



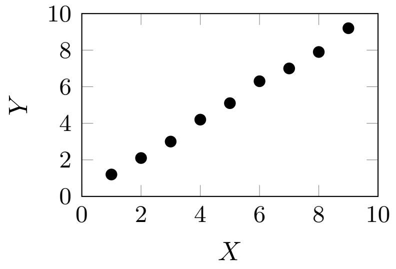
CAUTION: Identifying the shape of a distribution from a box plot or histogram is subjective. When identifying the shape of a distribution from a graph, be sure to support your opinion.

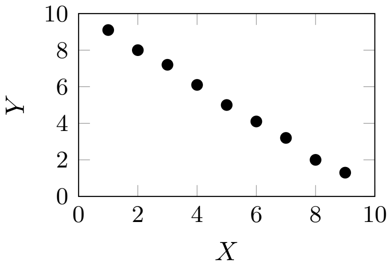
5 Chapter 4

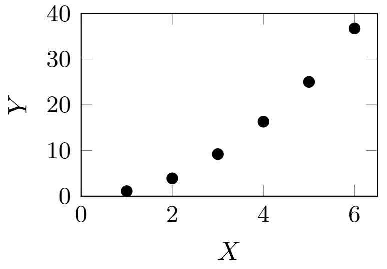
4.1 Describing the relation between two variables

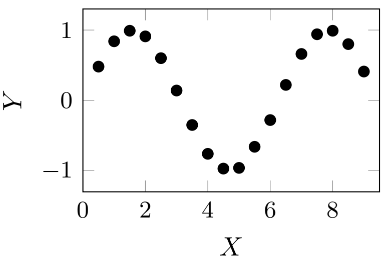
BI-Variance and Scatter Plots

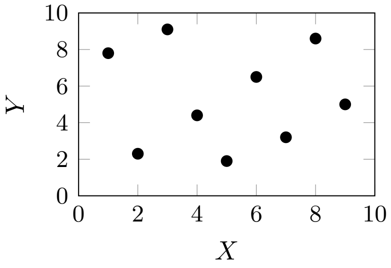
The response variable is the variable whose value can be explained by the value of the explanatory or predictor variable. A Scatter Diagram is a graph that shows the relationship between two Quantitative variables measured on the same individual. Each individual in the data set is represented by a point on the scatter diagram. The explanatory variable is plotted on the horizontal axis (x). The response variable is plotted on the vertical axis (y). We do not connect points in a scatter plot.


Positive linear


Negative linear


Nonlinear: quadratic


Nonlinear: sinusoidal


No relation

Two variables that are linearly related are positively associated when above average values of one variable are associated with above average values of the other variable and below average values of one variable are associated with below average values of the other variable. Positively associated if, whenever the value of one variable increases the value of the other variable also increases.

Two variables that are linearly related are negatively associated when above average values of one variable are associated with below average values of the other variable.

The Linear Correlation Coefficient is the measure of the strength and direction of the linear relation between two quantitative variables. p represents the population Correlation Coefficient, and r represents the sample Correlation Coefficient.

We only need to know Sample Correlation Coefficient right now:

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

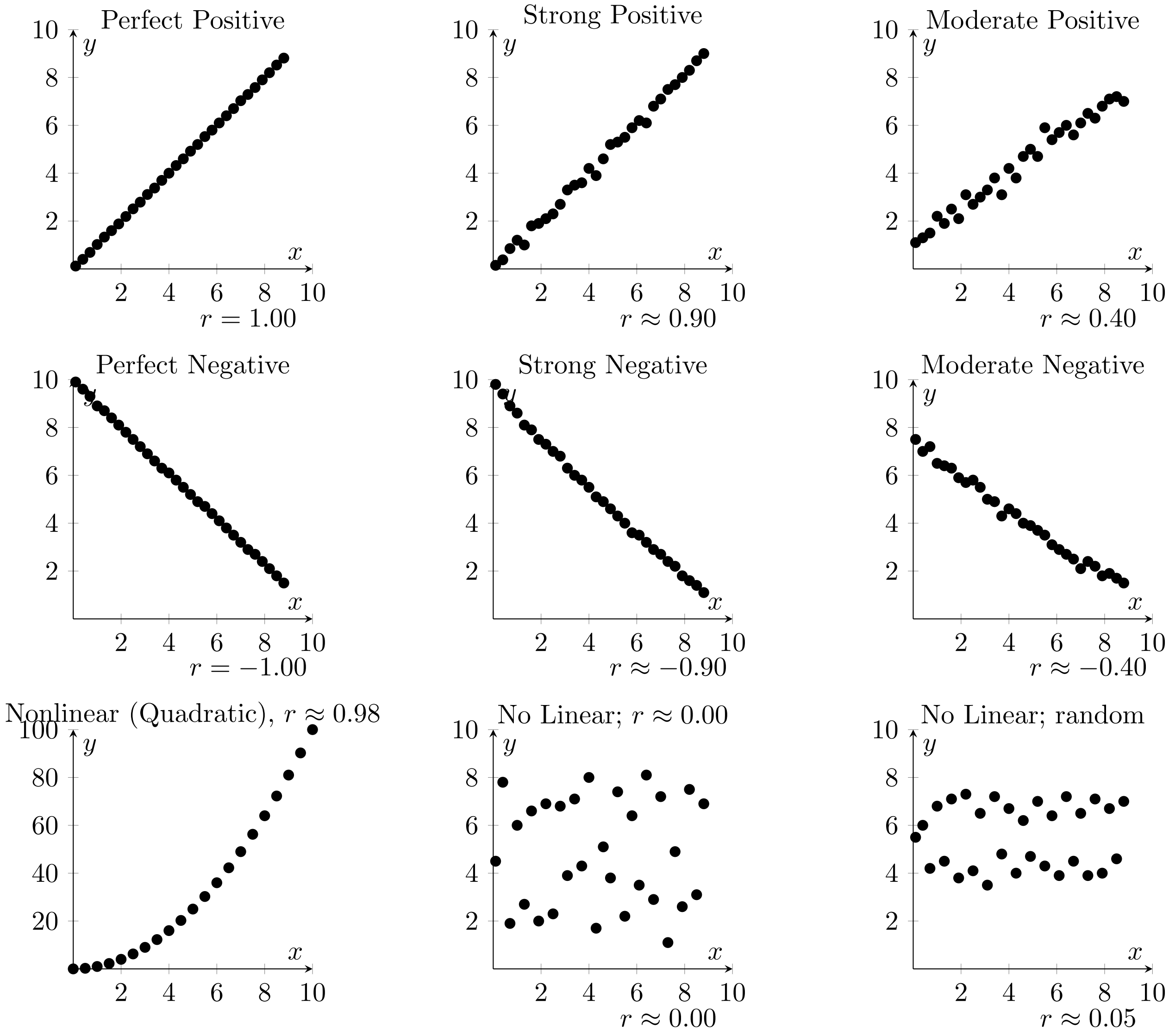
$$r = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]\left[\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right]}}$$

Equivalent computational equation

The properties of the linear Correlation Coefficient are as follows:

1. The L.C.C is always between -1 and 1, inclusive, $-1 \leq r \leq 1$
2. If $r = +1$, than a perfect positive linear relation exists between the two variables
3. If $r = -1$, than a perfect negative linear relation exists between the two variables
4. The closer r is to $+1$, the stronger the evidence of positive association
5. The closer r is to -1, the stronger the evidence of negative association
6. if r is close to 0, then little or no evidence exists of a linear relation.
   - but it DOES NOT imply no relation, just no LINEAR relation
7. L.C.C is unitless
8. L.C.C is not resistant

Below are some examples of different r values:



Perfect Positive $r = 1.00$

Strong Positive $r \approx 0.90$

Moderate Positive $r \approx 0.40$

Perfect Negative $r = -1.00$

Strong Negative $r \approx -0.90$

Moderate Negative $r \approx -0.40$

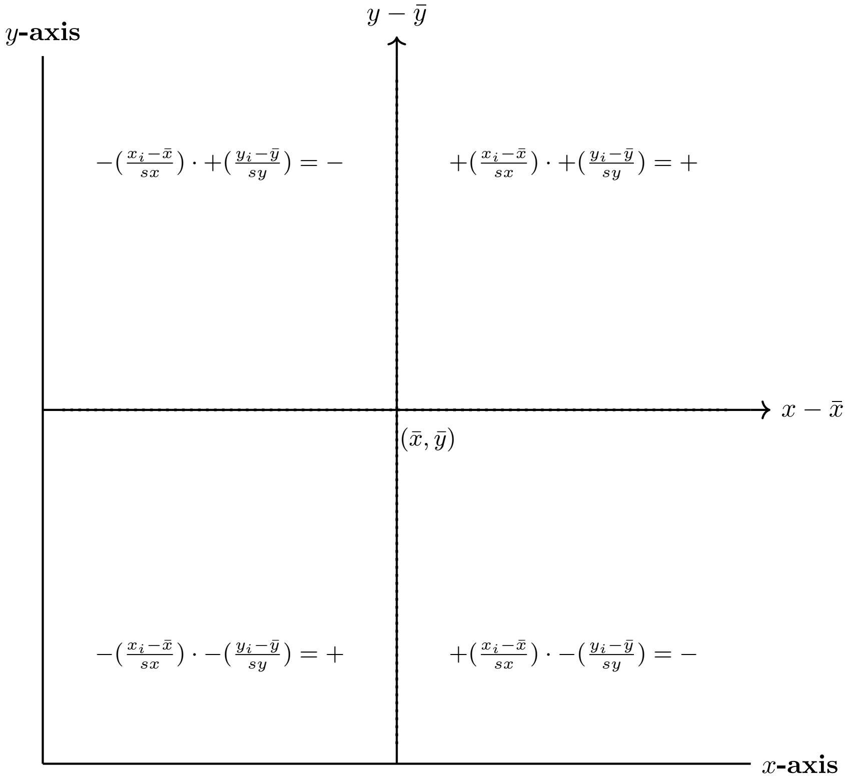
Nonlinear (Quadratic), $r \approx 0.98$

No Linear; $r \approx 0.00$ $r \approx 0.00$

No Linear; random $r \approx 0.05$

$r \approx 0.98$

In the formula for the formula for the L.C.C, notice that the numerator is the sum of the products of Z-scores for the explanatory (x) and the response (y) variables. A positive L.C.C. means that the sum of the products of the Z-scores for x and y must be positive. This occurs because if a certain x value is above its mean, $\bar{x}$, then the corresponding y value will be above its mean, $\bar{y}$. If a certain x value is below its mean, $\bar{x}$, than the corresponding y value will be below its mean, $\bar{y}$. Therefore, the sum of these products are positive so the L.C.C is positive. Below is a graph divided into quadrants that show the relation of the sum of the products.


y-axis
$y - \bar{y}$
$-(\frac{x_i-\bar{x}}{sx}) \cdot +(\frac{y_i-\bar{y}}{sy}) = -$
$+(\frac{x_i-\bar{x}}{sx}) \cdot +(\frac{y_i-\bar{y}}{sy}) = +$
$x - \bar{x}$
$(\bar{x}, \bar{y})$
$-(\frac{x_i-\bar{x}}{sx}) \cdot -(\frac{y_i-\bar{y}}{sy}) = +$
$+(\frac{x_i-\bar{x}}{sx}) \cdot -(\frac{y_i-\bar{y}}{sy}) = -$
x-axis

The outer **x-axis** and **y-axis** mark the boundaries of the plot. The lines through the center represent the means $\bar{x}$ and $\bar{y}$, which divide the plane into four quadrants. Points in Quadrants I and III have positive products of z_x and z_y, while points in Quadrants II and IV have negative products.

Determine whether a linear relation exists between two variables

To test for a linear relation we have to:

1. Determine the absolute value of the Correlation Coefficient
2. Find the critical value in the table of values for the given sample size
3. If the absolute value of the Correlation Coefficient is greater than the critical value, then a linear relation exists between the two variables. Otherwise, no linear reaction exists

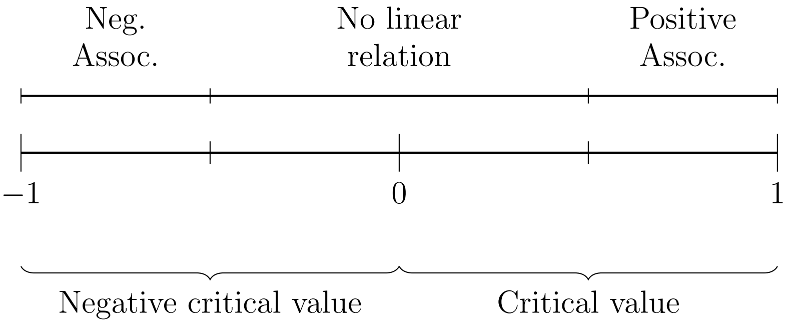


Table 1: Critical Values of the Correlation Coefficient r for Different Sample Sizes and Significance Levels

Sample Size n	**df** $= n - 2$	$\alpha = 0.05$	$\alpha = 0.01$
3	1	0.997	0.999
4	2	0.950	0.988
5	3	0.878	0.959
6	4	0.811	0.917
7	5	0.754	0.888
8	6	0.707	0.816
9	7	0.666	0.834
10	8	0.632	0.765

Correlation Critical Values

Purpose: Determine if a correlation r is statistically significant.

Table Columns:

- **n:** Sample size
- **df:** Degrees of freedom $= n - 2$
- $\alpha = 0.05$: Critical value for 5% significance
- $\alpha = 0.01$: Critical value for 1% significance

How to Use:

1. Find your sample size n.
2. Look up the critical value.
3. Compare $|r|$ to the critical value:
   - $|r| \geq$ critical value $\Rightarrow$ significant
   - $|r| <$ critical value $\Rightarrow$ not significant

Notes:

- Two-tailed test (check both positive and negative correlations)
- Use $|r|$ for negative correlations

Difference between causation and correlation is if the data is used in a study are observational we can not conclude the two correlated variables have a casual relationship. An L.C.C. that implies a strong positive or negative association does not imply causation if it was computed using observational data.

4.2 Least Squares Regression

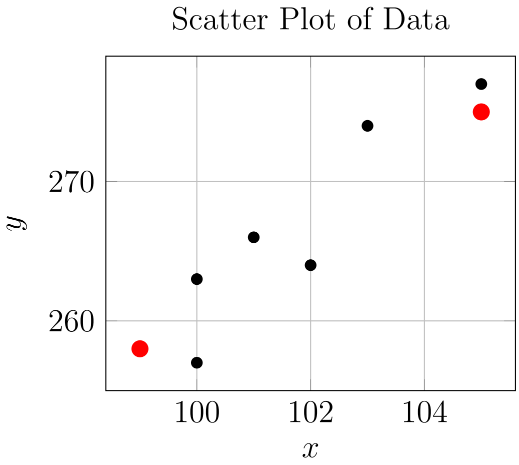
Once the scatter diagram and L.C.C. show that two variables have a linear relation, we can find a linear equation that describes this relation. One way to do this is to select two points from the data that appears to give a good fit of the relation.

To fins it by we follow these steps:

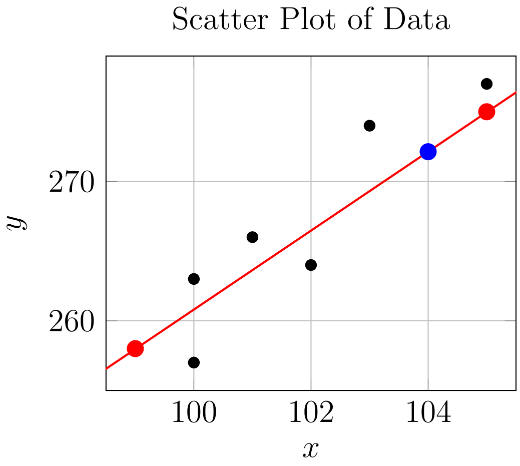
1. Find the equation of the line containing the two points that appear to give a good fit of the relation
   - This is just the slope, m, formula: $\frac{y_2-y_1}{x_2-x_1}$
2. Graph the line on the scatter diagram
3. Use point-slope form to find the y value given an x value
   - $y - y_1 = m(x - x_1)$

Example:

x	y	x, y
100	257	100,257
102	264	102,264
103	274	103,274
101	266	101,266
105	277	105,277
100	263	100,263
99	258	99,258
105	275	105,275



For this example we will take the points **(99,258)** and **(105,275)**. Find the slope of the line between them: slope $= x = \frac{275-258}{105-99} = 2.833$. Then to find a y value for a given x lets use $x = 104$. Point slope for to find y: $y - 258 = 2.833(x - 99)$, solved looks like this: $y = 2.833x - 22.497$, we plug in the given x: $y = 2.833(104) - 22.497$, solved looks like this: $y = 272.132$, so the coordinate given $x = 104$ is **(104,272.132)**. Below is the line plotted with the new coordinate marked in blue.



Find The Least Squares Regression Line And Use The Line To Make Predictions

The line that was found in the example appears to describe the relation between x and y quite well. However, we can still find a line that fits the data the best. A criterion is needed for determining the best of something. So there needs to be a criterion to determine the best like describing the relation between two variables. The difference between the observed and predicted values of y is the error, or residual for x or y. The residual equals observed y - the predicted y. To make a visual at an x value of 103, the y is predicted to be 269.3 but when it is actually observed the y value is 274, so the residual is $274 - 269.3 = 4.7$. The criterion to determine the line that bets describes the relation between two variables is based on the residuals. So we use the Least Squares regression. The least square regression line that minimizes the sum of the sum of the squared error (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$, we represent this as "minimize $\sum$ residuals".

The advantage of the least squares criterion is that it allows for statistical inference on the predicted value and slope. Another advantage is that the least squares regression criterion leads to the following formula for obtaining the least squares regression line also called the formal regression line which is as follows:

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

- Where $b_1 = r(\frac{sy}{sx})$ is the slope of the least squares regression line
- Where $b_0 = \bar{y} - b_1\bar{x}$ is the y intercept of the least squares regression line

Note: $\bar{x}$ is the sample mean, and sx is the sample standard deviation of the explanatory variable x; $\bar{y}$ is the sample mean and the sample standard deviation of the response variable.