Introduction to Statistics

A growing scope of application and importance, more often requires a statistical approach to understanding the world in this 'information age'. Successful applications of statistical analysis demand a philosophy of learning and action that follow these fundamental principles:

- A system of interconnected processes here all work occurs,
- All processes include variation, and
- The keys to success are comprehension and reduction of variation.

Data's role of quantifying variation and measuring the effects provides historical value, a measure of progress, effective process management, and ultimately improvement. This depends on the levels of activity and job responsibility. Just as corporations have executives who strategize direction, managers who guide workers, and workers who do the work, statistics offer an understanding and control of the process.

Two meanings and two aspects to the word statistics are as follows. The more common definition of statistics directs at numerical facts. The second definition is a group of methods used in data collection, analysis, presentation, and interpretation to make decisions. Numbers represent information and Americans obsess on statistics. Solutions to real-world problems may not be definite so statistics helps people to make intelligent decisions in these situations. Theoretical or mathematical statistics develops, derives, and proves theorems, formulas, rules, and laws. Applied statistics applies those theorems, formulas, rules, and laws to solve real-world problems. Two areas of applied statistics are descriptive statistics and inferential statistics. Descriptive statistics helps in the analysis of large data sets by shrinking them to a manageable summary so that decisions can be made. Inferential statistics samples result about a portion of a collection of elements of interest (a population) and uses the sample to predict or decide.

Everyday life requires making decisions hinged on the theory of probability. The probability of an event occurring during a given day can be the determining factor of all decisions made that morning. Interpretation of probability provides a measure of credence to a person's willingness to place bets. A willingness to bet everything on the truth of a proposition with very little to gain if right and a great deal to lose if wrong indicates an absolute certainty of its truth. A person's personality matters in how he or she makes decisions. Suppose it has been raining all week and the weather forecast says it will rain today. Confident that it will rain today, a person may choose to wear a raincoat and take an umbrella to work. Should the day be sunny and without rain, nothing is lost. On the other hand, if a person chose not to wear a raincoat or take an umbrella to work and the day turned out rainy, the person will be wet and miserable. Propositions may be certain in themselves rather than how certain a person feels about them. Betting on predictions depends on the level of confidence in the source, and the willingness to bet. Decision strategies can be misleading and the ability of some people to interpret information correctly depends on uncertain knowledge and the confidence they place in it.

Statistics simply provide vital information in a comprehensible form so that people know how to interpret the data and can draw intelligent conclusions or make intelligent decisions rather than educated guesses. The probability of an event ranges between and includes 0 and 1. If an event cannot occur the result is 0. If an event is certain to occur the result is 1. The sum of all possible outcomes in a sample space is equal to 1. Many problems are solved by using together the probability and counting rules of statistics.

There are several definitions given by the different author regarding Statistics.

Few definitions of the statistics are:

Statistics is the branch of science which deals with the collection, classification and tabulation of numerical facts as the basis for explanations, description and comparison of phenomenon **– Lovitt**

The science which deals with the collection, analysis and interpretation of numerical data **- Corxton& Cowden**

The science of statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates **-King**

Statistics may be called the science of counting or science of averages or statistics is the science of the measurement of social organism, regarded as whole in all its manifestations – **Bowley**

Statistics is a science of estimates and probabilities **-Boddington**

Statistics is a branch of science, which provides tools (techniques) for decision making in the face of uncertainty (probability) - **Wallis and Roberts**

The above definitions clearly focused on the four different aspects of statistics:

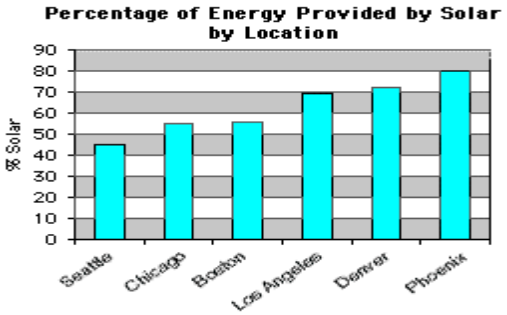
- Data Collection
- Data Presentation
- Data Analysis
- Interpretation of Data

There are basically four division onto which statistics is divided.

- Mathematical or theoretical statistics
- Statistical methods or functions
- Descriptive statistics
- Inferential Statistics

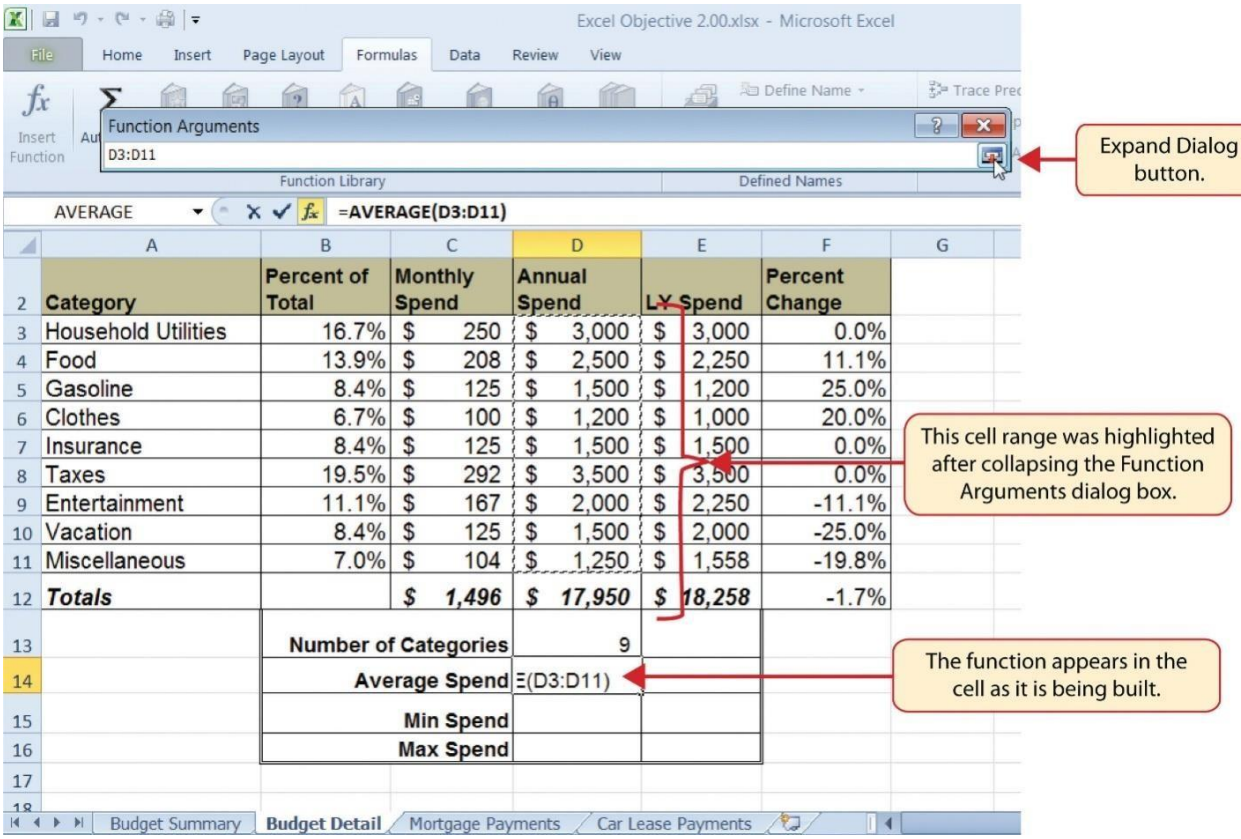
Mathematical Statistics:

It helps in forming the experimental and statistical distribution.



Statistical methods:

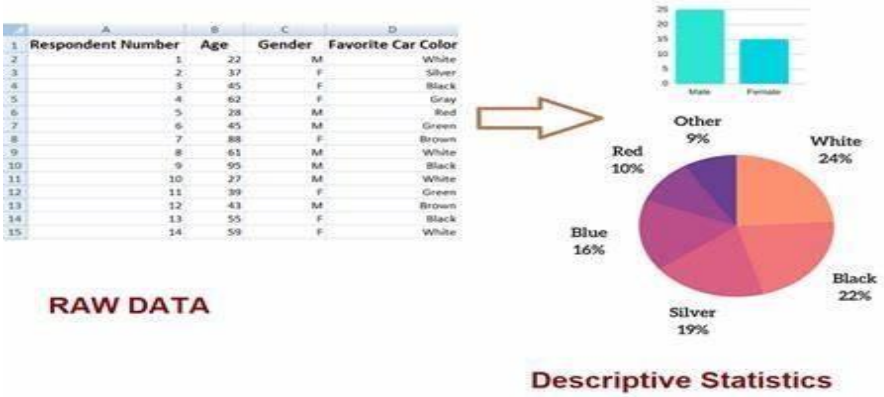
It helps in the collection, tabulation and interpretation of the data. It helps in analyzing the data and returns insight from the data.



	Category	Percent of Total	Monthly Spend	Annual Spend	LY Spend	Percent Change
3	Household Utilities	16.7%	$ 250	$ 3,000	$ 3,000	0.0%
4	Food	13.9%	$ 208	$ 2,500	$ 2,250	11.1%
5	Gasoline	8.4%	$ 125	$ 1,500	$ 1,200	25.0%
6	Clothes	6.7%	$ 100	$ 1,200	$ 1,000	20.0%
7	Insurance	8.4%	$ 125	$ 1,500	$ 1,500	0.0%
8	Taxes	19.5%	$ 292	$ 3,500	$ 3,500	0.0%
9	Entertainment	11.1%	$ 167	$ 2,000	$ 2,250	-11.1%
10	Vacation	8.4%	$ 125	$ 1,500	$ 2,000	-25.0%
11	Miscellaneous	7.0%	$ 104	$ 1,250	$ 1,558	-19.8%
12	***Totals***		***$ 1,496***	***$ 17,950***	***$ 18,258***	-1.7%
13		**Number of Categories**		9		
14		**Average Spend**		=(D3:D11)		
15		**Min Spend**				
16		**Max Spend**				

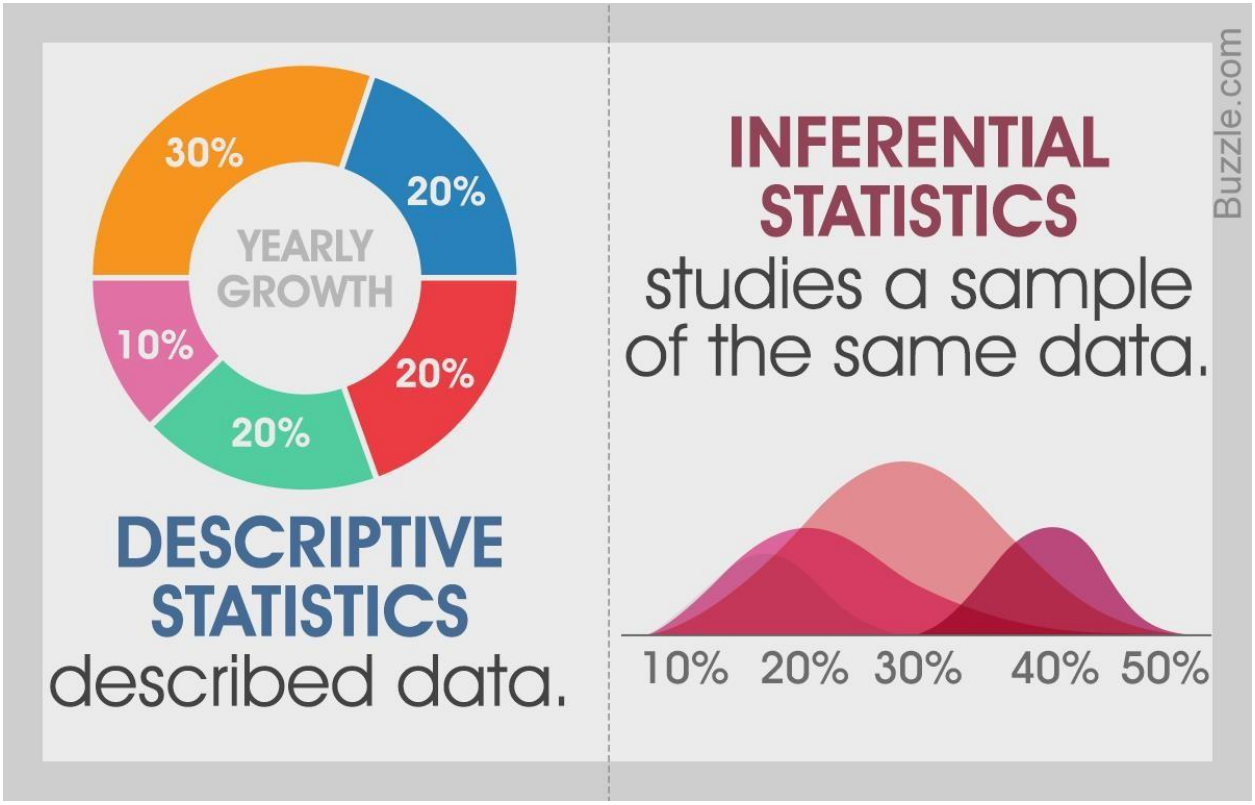
Descriptive Statistics:

It helps in summarizing and organizing any data set characteristics. It also helps in the representation of data in both classification and diagrammatic way.



Inferential Statistics:

It helps in finding the conclusion regarding the population after analysis on the sample drawn from it.



Statistics is widely used in the field of:

- Production Management of an organization
- Quality Control of an organization
- Banking Institution
- Forecasting
- Agriculture Research
- Economic Analysis
- Stock Market Analysis
- Business and commerce
- Panning
- Projection

Importance of statistics:

Statistics plays an important role in our daily life; it is useful in almost all sciences – social as well as physical – such as biology, psychology, education, economics, and business management, agricultural sciences etc.The statistical methods can be and are being followed by both educated and uneducated people. In many instances we use sample data to make inferences about the entire population.

1. Planning is indispensable for better use of nation's resources. Statistics are indispensable in planning and in taking decisions regarding export, import, and production etc. Statistics serves as foundation of the super structure of planning.
2. Statistics helps the business man in the formulation of polices with regard to business. Statistical methods are applied in market and production research, quality control of manufactured products.
3. Statistics is essential in economics. Any branch of economics that require comparison, correlation requires statistical data for salvation of problems.
4. Statistics is helpful in administration in fact statistics are regarded as eyes of administration. In collecting the information about population, military strength etc. Administration is largely depending on facts and figures thud it needs statistics.
5. Bankers, stock exchange brokers, insurance companies all make extensive use of statistical data. Insurance companies make use of statistics of mortality and life premium rates etc., for bankers, statistics help in deciding the amount required to meet day to day demands.
6. Problems relating to poverty, unemployment, food storage, deaths due to diseases, due to shortage of food etc., cannot be fully weighted without the statistical balance. Thus, statistics is helpful in promoting human welfare.
7. Statistics are a very important part of political campaigns as they lead up to elections. Every time a scientific poll is taken, statistics are used to calculate and illustrate the results in percentages and to calculate the margin for error.

Drawbacks of Statistics:

- Fails in Qualitative Phenomenon.
- Fails in individual study.
- Their laws are not exact. It based on hypothesis.
- Fails in giving the entire information.
- Data Security issues.
- They are valid only on the average base.

Population Distribution

A population distribution is made up of all the classes or values of variables which we would observe if we were to conduct a census of all members of the population. For instance, if we wish to determine whether voters "Approve" or "Disapprove" of a particular candidate for president, then all individuals who are eligible voters constitute the population for this variable. If we were to ask every eligible voter his or her voting intention, the resulting two-class distribution would be a population distribution. Similarly, if we wish to determine the number of column inches of coverage of Fortune 500 companies in the Wall Street Journal, then the population consists of the top 500 companies in the US as determined by the editors of Fortune magazine. The population distribution is the frequency with which each value of column inches occurs for these 500 observations. Here is a formal definition of a population distribution:

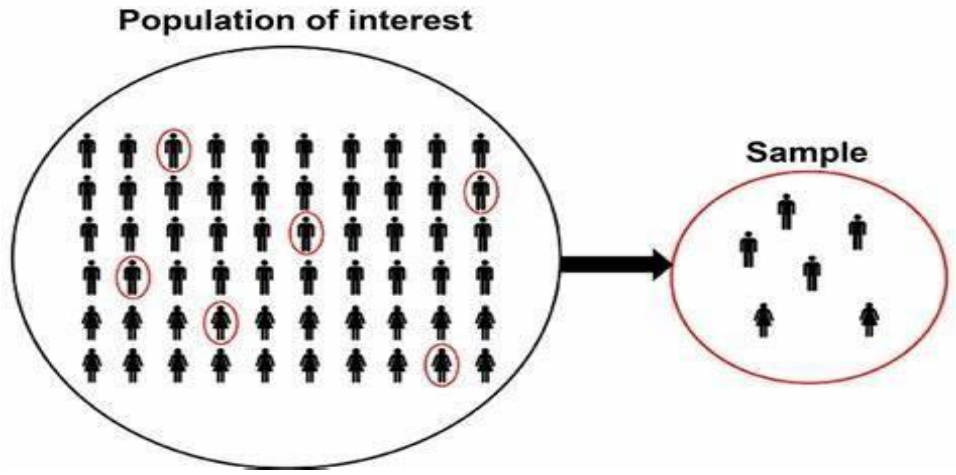
Note: A population distribution is a statement of the frequency with which the units of analysis or cases that together make up a population are observed or are expected to be observed in the various classes or categories that make up a variable.

Sample Distribution:

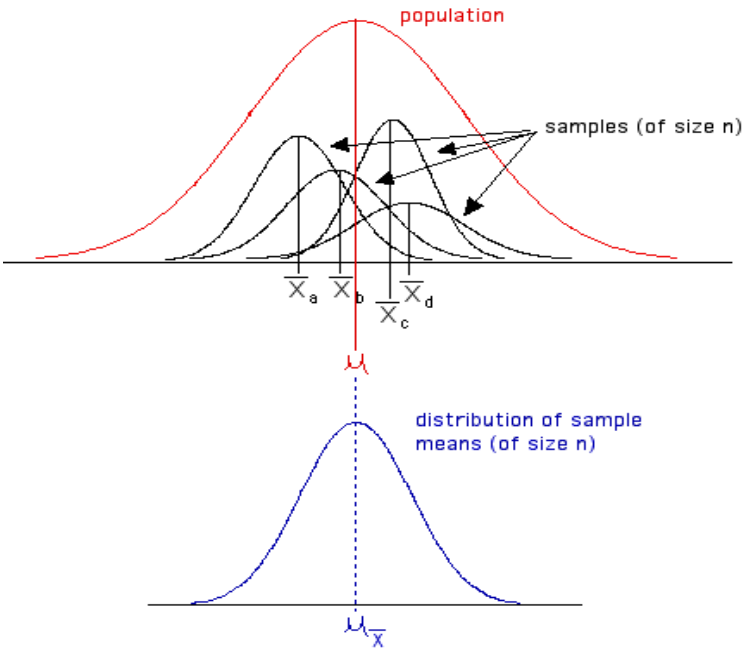
Samples which are representative of the populations from which they have been drawn, so that we can make valid statistical generalizations. This means that we will restrict our discussion to randomly selected samples. These random probability samples were defined in Chapter 6 as samples drawn in such a way that each unit of analysis in the population has an equal chance of being selected for the sample. A sample is simply a subset of all the units of analysis which make up the population. For instance, a group of voters who "Approve" or "Disapprove" of a particular presidential candidate constitute a small subset of all those who are eligible voters (the population). If we wanted to determine the actual number of column inches of coverage given to Fortune 500 companies in the WSJ we could draw a random sample of 50 of these companies. Below is a definition of a sample distribution:

Note: A sample distribution is a statement of the frequency with which the units of analysis or cases that together make up a sample are actually observed in the various classes or categories that make up a variable.

Image showing the difference between Sample and Population:



Graphical Representation:



Symbols related to population:

	Population	Sample
Mean	*M*	*X*
Variance	σ^2	*var*
Standard Deviation	σ	*sd*

Examples: Parameters of a Fictious Population of N=5 Cases.

Person	X_i	$(X-M)$	$(X-M)^2$
A	5	-2	4
B	6	-1	1
C	7	0	0
D	8	1	1
E	9	2	4
	$\sum X = 35$	$\sum (X-M) = 0$	$\sum (X-M)^2 = 10$
Parameter mean $= M = \sum X / N = 35/5 = 7$			
Parameter variance $= \frac{\sum (X-M)^2}{N} = \frac{10}{5} = 2$			
Parameter standard deviation $= \sqrt{2} = 1.41$			

Some samples of N=3 and their associated means, variances and standard deviations.

Sample	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
A	5	-1.00	1.00
	6	0.00	0.00
	7	1.00	1.00
			$\sum(X-\bar{X})^2 = 2.00$
Sample mean = $\bar{X} = 6$			
Sample Variance = *var* = 2.00 / 3 = .66			
Sample Std. Deviation = *sd* = $\sqrt{.66} = .82$			

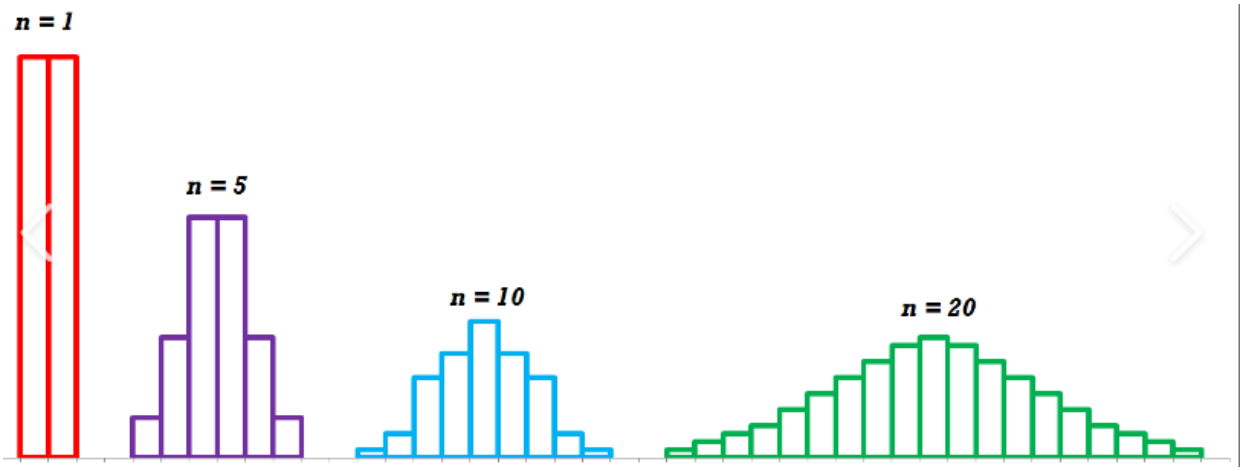
Sample	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
A	5	-1.00	1.00
	6	0.00	0.00
	7	1.00	1.00
			$\sum(X-\bar{X})^2 = 2.00$
Sample mean = $\bar{X} = 6$			
Sample Variance = *var* = 2.00 / 3 = .66			
Sample Std. Deviation = *sd* = $\sqrt{.66} = .82$			

Sample	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
B	5	-2.00	4.00
	8	1.00	1.00
	8	1.00	1.00
			$\sum(X-\bar{X})^2 = 6.00$
$\bar{X} = 7$			
var = 6.00 / 3 = 2.00			
sd = $\sqrt{2.00} = 1.41$			

Sample	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
C	7	.33	.11
	8	1.33	1.77
	5	-1.66	2.76
			$\sum(X-\bar{X})^2 = 4.64$
$\bar{X} = 6.67$			
var = 4.64 / 3 = 1.56			
sd = $\sqrt{1.55} = 1.25$			

Sampling Distribution:

If we draw a number of samples from the same population, then compute sample statistics for each, we can construct a distribution consisting of the values of the sample statistics we've computed. This is a kind of "second-order" distribution. Whereas the population distribution and the sample distribution are made up of data values, the sampling distribution is made up of values of statistics computed from a number of sample distributions. Probably the easiest way to visualize how one arrives at a sampling distribution is by looking at an example. We'll use our running example of mothers' communication with children in which samples of N = 3 were selected. Figure 9-1 illustrates a model of how a sampling distribution is obtained. Figure 9-1 illustrates the population which consists of a set of scores (5, 6, 7, 8 and 9) which distribute around a parameter mean of 7.00. From this population we can draw a number of samples. Each sample consists of three scores which constitute a subset of the population. The sample scores distribute around some statistic mean for each sample. For sample A, for instance, the scores are 5, 6 and 7 (the sample distribution for A) and the associated statistic mean is 6.00. For sample B the scores are 5, 8 and 8, and the statistic mean is 7.00. Each sample has a statistic mean. The statistics associated with the various samples can now be gathered into a distribution of their own. The distribution will consist of a set of values of a statistic, rather than a set of observed values. This leads to the definition for a sampling distribution: A sampling distribution is a statement of the frequency with which values of statistics are observed or are expected to be observed when a number of random samples is drawn from a given population. It is extremely important that a clear distinction is kept between the concepts of sample distribution and of sampling distribution. A sample distribution refers to the set of scores or values that we obtain when we apply the operational definition to a subset of units chosen from the full population. Such a sample distribution can be characterized in terms of statistics such as the mean, variance, or any other statistic. A sampling distribution emerges when we sample repeatedly and record the statistics that we observe. After a number of samples have been drawn, and the statistics associated with each computed, we can construct a sampling distribution of these statistics. The sampling distributions resulting from taking all samples of N=2 as well as the one based taking all samples of N=3 out of the population of mothers of school-age children.



Refer to the below Image:

Sample Size = 2

(1) $\bar{X}$	(2) f	(3) $f\cdot(\bar{X})$	(4) $(\bar{X}-\bar{\bar{X}})$	(5) $(\bar{X}-\bar{\bar{X}})^2$	(6) $f\cdot(\bar{X}-\bar{\bar{X}})^2$
5.0	1	5.00	-2.00	4.00	4.00
5.5	2	11.00	-1.50	2.25	4.50
6.0	3	18.00	-1.00	1.00	3.00
6.5	4	26.00	-.50	.25	1.00
7.0	5	35.00	.00	.00	.00
7.5	4	30.00	.50	.25	1.00
8.0	3	24.00	1.00	1.00	3.00
8.5	2	17.00	1.50	2.25	4.50
9.0	1	9.00	2.00	4.00	4.00
	25	**175.00**			**25.00**

Mean of means: $\bar{\bar{X}} = 175/25 = 7.00$

Sampling Variance: 25/25=1.00

Standard Error: $\sqrt{1.00} = 1.00$

Sample Size = 3

(1) $\bar{X}$	(2) f	(3) $f\cdot(\bar{X})$	(4) $(\bar{X}-\bar{\bar{X}})$	(5) $(\bar{X}-\bar{\bar{X}})^2$	(6) $f\cdot(\bar{X}-\bar{\bar{X}})^2$
5.00	1	5.00	-2.00	4.00	4.00
5.33	3	15.99	-1.67	2.79	8.37
5.66	6	33.96	-1.34	1.79	10.74
0006.00	10	60.00	-1.00	1.00	10.00
6.33	15	94.95	-.67	.45	6.73
6.66	18	119.88	-.34	.12	2.16
7.00	19	133.00	0.00	0.00	0.00
7.33	18	131.94	.34	.12	2.16
7.66	15	114.90	.67	.45	6.73
8.00	10	80.00	1.00	1.00	10.00
8.33	6	49.98	1.34	1.79	10.74
8.66	3	25.98	1.67	2.79	8.37
9.00	1	9.00	2.00	4.00	4.00
	125	**875.00**			**84.00**

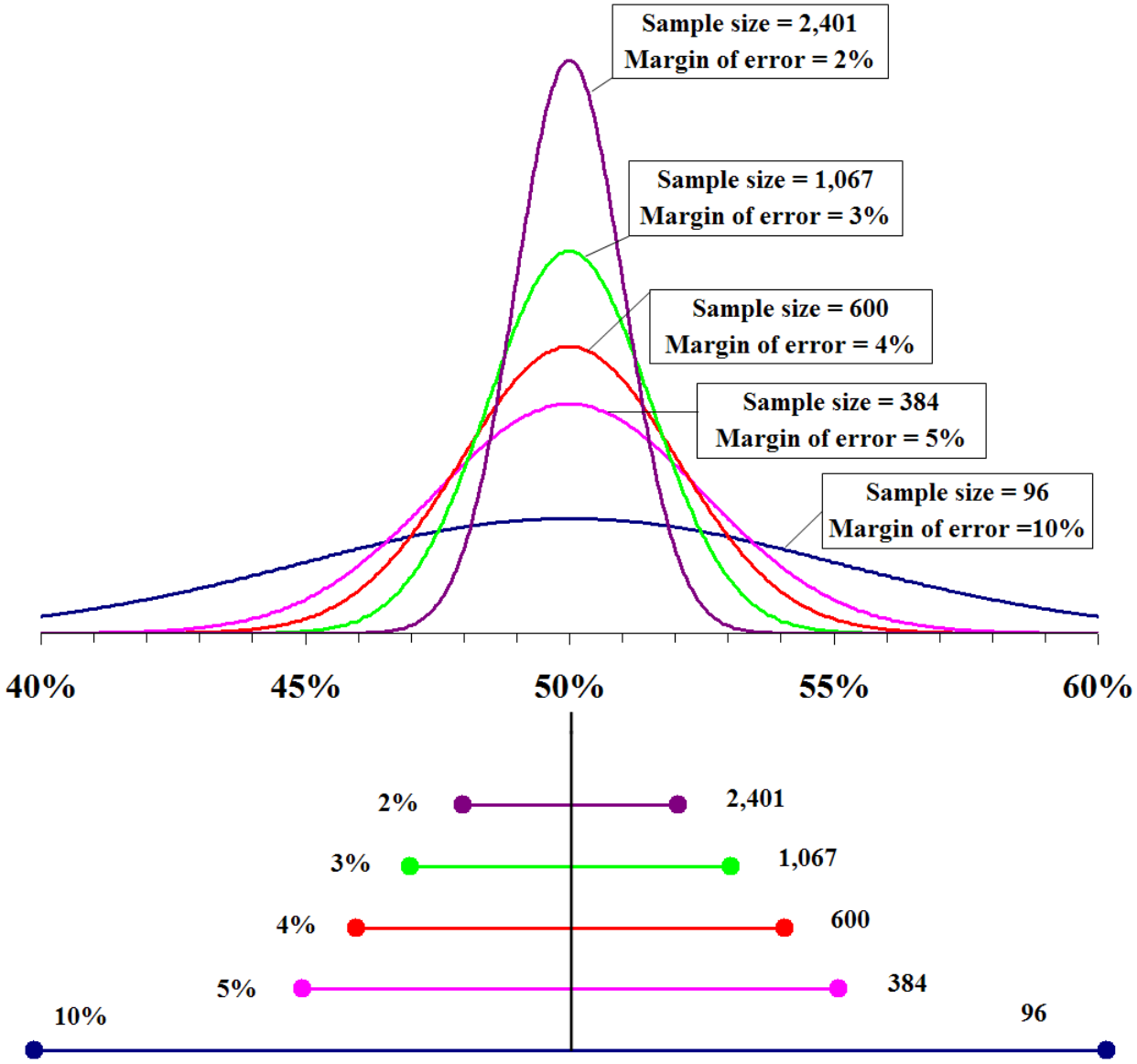
Mean of means: $\bar{\bar{X}} = 875/125 = 7.00$

Sampling Variance: 84/125=.672

Standard Error: $\sqrt{.672} = .819$

Distribution of sampling error:

When we draw random samples from a population there are no guarantees that the sample will indeed be exactly representative of the population. As seen earlier in this chapter it is quite possible that there will be differences between sample characteristics and population characteristics. In fact, sampling error can be defined as the discrepancy between the parameter of a population and the corresponding statistic computed for a sample drawn randomly from that population. It shows a listing of all the sample statistic means that were computed when all possible samples of a given size were drawn from the population. Most of these different statistic means show a certain amount of discrepancy with the population mean; some means show larger discrepancies, others show smaller ones. The sampling distribution that results when we take all samples from a given population is therefore also the distribution of the amounts of sampling error that we encountered as we drew those samples.



Sampling Error and Standard Error:

We know that sampling error is unavoidable, even when we sample randomly. However, let us assume for a moment that we could randomly sample without committing sampling error. In that case, for all the samples that we would draw from a population, there would be no discrepancy between the statistic computed for each sample and the population parameter. Each sample mean would be exactly equal to the population mean. Since all the entries in the sampling distribution would be identical, the sampling distribution would have a mean equal to the population mean. The sampling variance and the standard error are measures of dispersion of a set of statistics about the mean of the sampling distribution of those statistics. As all entries in this distribution are exactly equal to the mean, it follows that under these conditions we would observe the sampling variance and standard error to be equal to zero. A zero standard error indicates that there is no sampling error, which is correct given our original assumption of sampling without error. If we assume small amounts of sampling error, the observed statistics will be quite similar to one another, but not identical. This means that they will be quite similar to the population parameter. As a consequence, the resulting sampling distribution would have a non-zero sampling variance and standard error, but both will be quite small, since all statistics are similar to the population parameter. As increasing amounts of sampling error are introduced, the differences between the individual sample statistics and the population parameter will increase, and the sampling variance and standard error will be larger.

Factor affecting the sampling error:

- Population variance
- Sample size

Population Variance:

The larger the population variance, the larger the sampling error. If you think about it for a moment, you'll probably see why this is the case. A population distribution with a small variance has its scores more tightly clustered around the population mean. This means that any random sample drawn from this population is likely to contain many observed values which are close to the population mean. The mean of such a sample (and the means of others like it) will then be close to the population mean, and there will be little sampling error.

Sample size:

The second factor which determines the magnitude of the sampling error is the size of the sample drawn from the population. The general rule of statistical generalization that we have already developed is that increased sample size reduces the sampling error, all other things remaining equal. Again, this phenomenon can be illustrated simply with an example. We can see the sampling distributions that result when we take samples of two different sizes. If we look at the various sample means that are observed when N = 2 and when N = 3, we see that, as sample size increases, a larger proportion of the sample means will fall near the true population mean.

Refer to the below table for better understanding of Population Variance, Sampling Error and Standard Error:

Population A	*Population B*	*Population C*
Values: 6,7,7,7,8	Values: 5,6,7,8,9	Values: 3,5,7,9,11

PARAMETERS

Population A	*Population B*	*Population C*
M= 7.00	M =7.00	M = 7.00
σ^2= .40	σ^2 =2.00	σ^2 = 8.00
σ = .63	σ =1.41	σ = 2.83

SAMPLING DISTRIBUTIONS

Population A		*Population B*		*Population C*	
Sample $\bar{X}$	f	Sample $\bar{X}$	f	Sample $\bar{X}$	f
6.0	1	5.0	1	3	1
6.5	6	5.5	2	4	2
7.0	11	6.0	3	5	3
7.5	6	6.5	4	6	4
8.0	1	7.0	5	7	5
	25	7.5	4	8	4
		8.0	3	9	3
		8.5	2	10	2
		9.0	1	11	1
			25		**25**

$(\bar{X}-M)$	$f\cdot(\bar{X}-M)^2$	$(\bar{X}-M)$	$f\cdot(\bar{X}-M)^2$	$(\bar{X}-M)$	$f\cdot(\bar{X}-M)^2$
-1.00	1.00	-2.00	4.00	-4.00	16.00
-.50	1.50	-1.50	4.50	-3.00	18.00
0.00	0.00	-1.00	3.00	-2.00	12.00
.50	1.50	-.50	1.00	-1.00	4.00
1.00	1.00	0.00	0.00	0.00	0.00
		.50	1.00	1.00	4.00
		1.00	3.00	2.00	12.00
		1.50	4.50	3.00	18.00
$\sum f\cdot(X-M)^2 = 5.00$		2.00	4.00	4.00	16.00
			25.00		**100.00**

SAMPLING VARIANCE

Population A	*Population B*	*Population C*
5.00/25 =.20	25.00/25 =1.00	100.00/25 = 4.00

STANDARD ERROR

Population A	*Population B*	*Population C*
v.20 = .44	v1.00 = 1.00	v4.00 = 2.00

<u>Refer to the below screenshot for calculating probability using sampling distribution:</u>

Sample Size N = 2

$\bar{X}$	f	probability
5.0	1	1/25 =.04
5.5	2	2/25 =.08
6.0	3	3/25 = .12
6.5	4	4/25 = .16
7.0	5	5/25 = .20
7.5	4	4/25 = .16
8.0	3	3/25 = .12
8.5	2	2/25 = .08
9.0	1	1/25 = .04
	25	1.00

$\bar{\bar{X}} = M = 7.00$

Descriptive Statistics

Types of data:

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

An observation or response is the value taken by a variate for some given unit.
There are various types of variate.

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

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

Quantitative data can be:

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

Displaying data:

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

Discrete data: frequency table and bar chart:

Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.
The frequency of a value is the number of observations taking that value.
A frequency table is a list of possible values and their frequencies.
A bar chart consists of bars corresponding to each of the possible values, whose heights are equal to the frequencies.

Refer to the below Example for a better understanding:

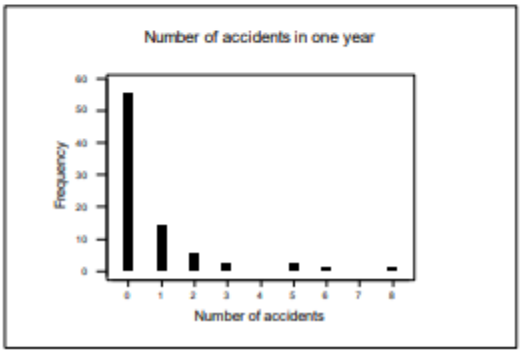
The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below. Construct a frequency table and draw a bar chart.

2	0	0	1	0	3	0	6	0	0	8	0	2	0	1
5	1	0	1	1	2	1	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	5	1	0	0	0	0	0	0	0	0	1	1
0	3	0	0	1	1	0	0	0	2	0	1	0	0	0
0	0	0	0	0										

Solution

Number of accidents	Tallies	Frequency
0	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	55
1	𝍸 𝍸 𝍸	14
2	𝍸	5
3	||	2
4		0
5	||	2
6	|	1
7		0
8	|	1

Barchart



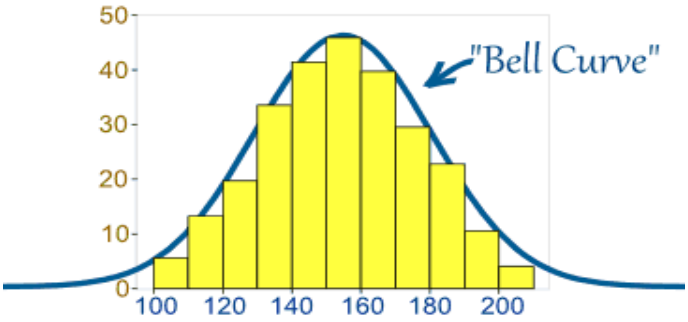
Different Terminologies which is to be noticed in data:

Bar charts and histograms provide an easily understood illustration of the distribution of the data. As well as showing where most observations lie and how variable the data are, they also indicate certain "danger signals" about the data.

Normally Distributed data:

The histogram is bell-shaped, like the probability density function of a Normal distribution. It appears, therefore, that the data can be modelled by a Normal distribution. (Other methods for checking this assumption are available.)

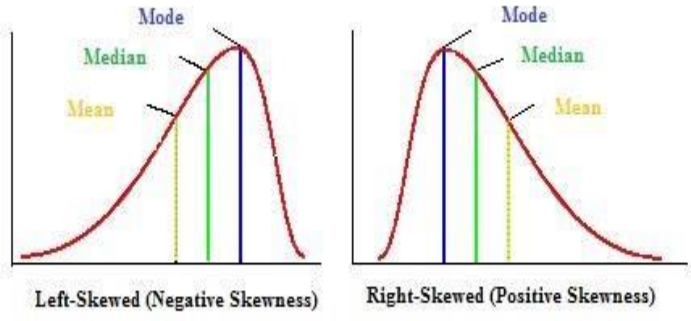
Refer to the below image for better understanding:



Very skew data:

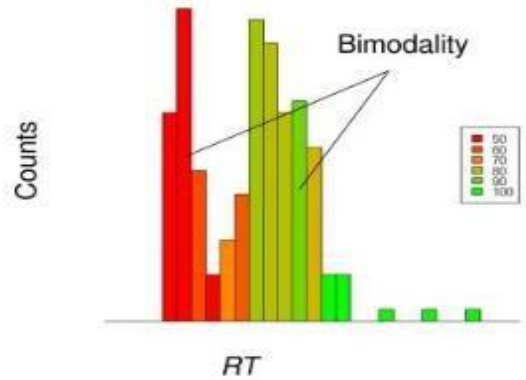
The relatively few large observations can have an undue influence when comparing two or more sets of data. It might be worthwhile using a transformation e.g., taking logarithms.

Refer to the below image for better understanding:



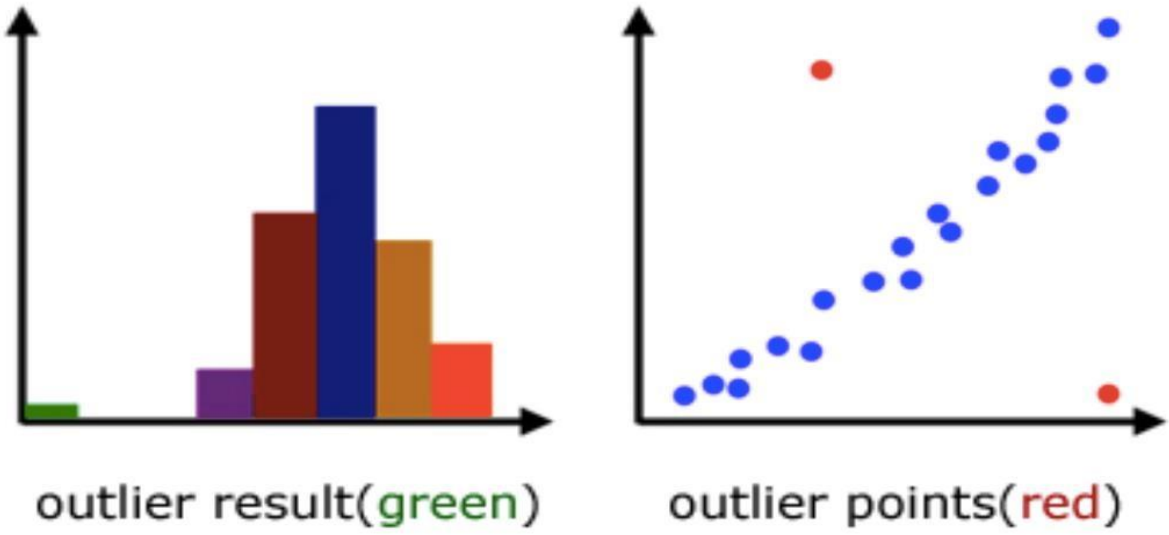
Bimodality:

This may indicate the presence of two subpopulations with different characteristics. If the subpopulations can be identified it might be better to analyze them separately.



Outliers:

The data appear to follow a pattern with the exception of one or two values. You need to decide whether the strange values are simply mistakes, are to be expected or whether they are correct but unexpected. The outliers may have the most interesting story to tell.



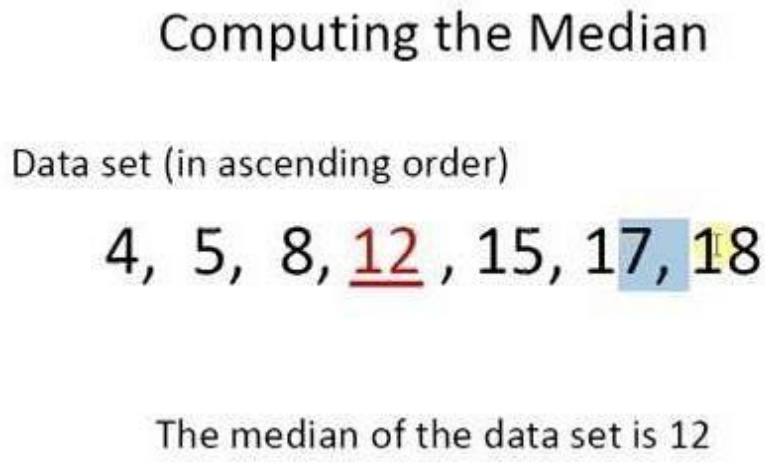
Sample mean:

This is just the average or arithmetic mean of the values. Sometimes the prefix "sample" is dropped, but then there is a possibility of confusion with the population mean which is defined later.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

Sample Median:

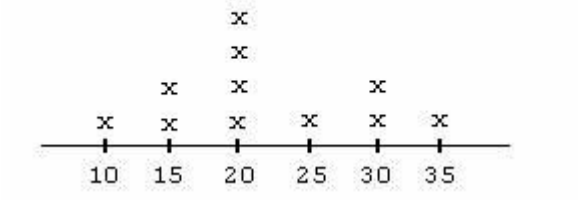
The median is the central value in the sense that there as many values smaller than it as there are larger than it.



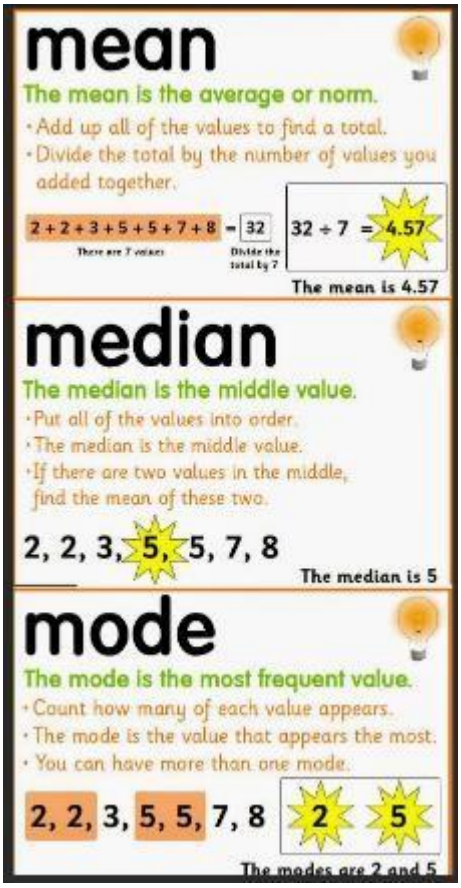
Sample Mode:

The mode, or modal value, is the most frequently occurring value. For continuous data, the simplest definition of the mode is the midpoint of the interval with the highest rectangle in the histogram. (There is a more complicated definition involving the frequencies of neighboring intervals.) It is only useful if there are a large number of observations.

10, 20, 15, 20, 25, 30, 35, 20, 20, 30, 15



Refer to the below screenshot for understanding the difference between mean, median and mode.

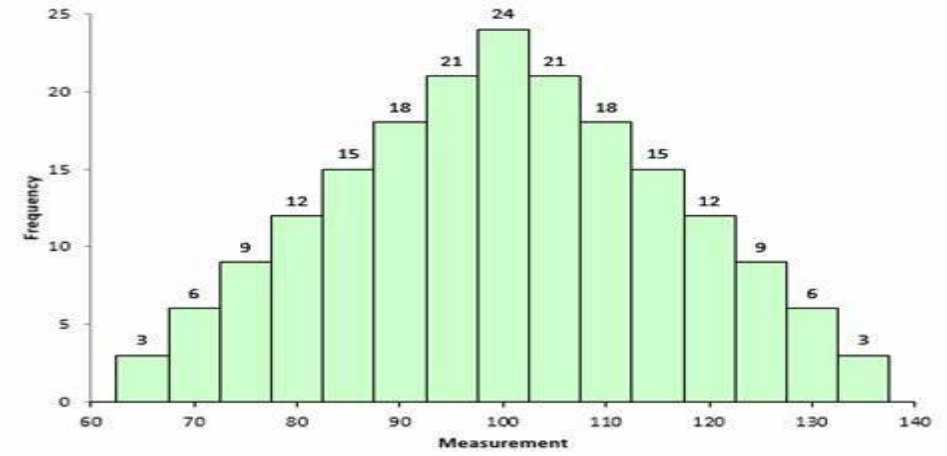


Types of data:

- Symmetric data
- Skew data

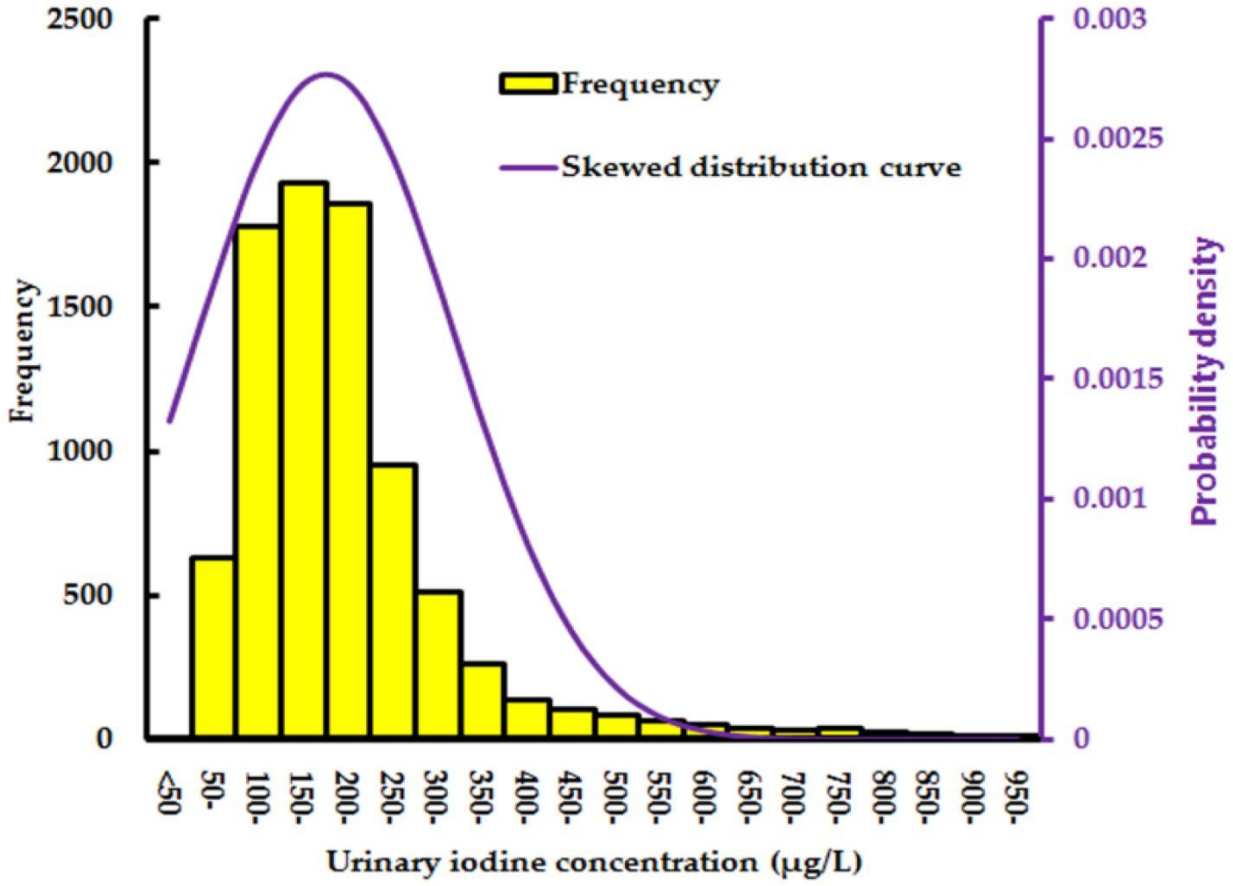
Symmetric Data:

The mean median and mode will be approximately equal.



Skew Data:

The median is less sensitive than the mean to extreme observations. The mode ignores them. The mode is dependent on the choice of class intervals and is therefore not favoured for sophisticated work.



Terminologies related to Statistical Inference:

Probability Theory:

The probability distribution of the population is known; we want to derive results about the probability of one or more values ("random sample") - deduction.

Parametric Estimation:

We assume that we know the type of distribution, but we do not know the value of the parameters θ, say. y. We want to estimate θ, on the basis of a random sample X1,X2…..Xn. Let's call the random sample X1, X2… Xn our data D. We wish to infer P(θ|D) which by Bayes' theorem is,

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(\theta)$ is called the prior, which is the probability distribution from any prior information we had before looking at the data (often this is taken to be a constant). The denominator P(D) does not depend on the parameters, and so is just a normalization constant. $P(D|\theta)$ is called the likelihood: it is how likely the data is given a particular set of parameters.

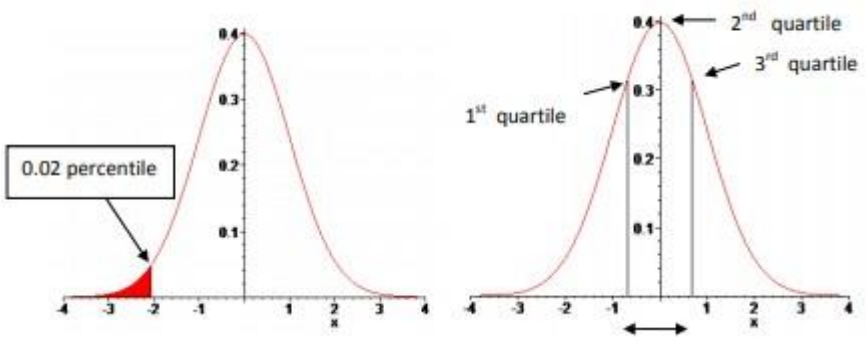
The full distribution $P(\theta|D)$ gives all the information about the probability of different parameters values given the data. However, it is often useful to summaries this information, for example giving a peak value and some error bars.

Maximum likelihood estimator:

The value of θ that maximizes the likelihood is called the maximum likelihood estimate: it is the value that makes the data most likely, and if P(θ) does not depend on parameters (e.g., is a constant) is also the most probable value of the parameter given the observed data. The maximum likelihood estimator is usually the best estimator, though in some instances it may be numerically difficult to calculate. Other simpler estimators are sometimes possible.

Percentiles and the interquartile range:

The kth percentile is the value corresponding to cumulative relative frequency of k/100 on the cumulative relative frequency diagram e.g., the 2nd percentile is the value corresponding to cumulative relative frequency 0.02. The 25th percentile is also known as the first quartile and the 75th percentile is also known as the third quartile. The interquartile range of a set of data is the difference between the third quartile and the first quartile, or the interval between these values. It is the range within which the "middle half" of the data lie, and so is a measure of spread which is not too sensitive to one or two outliers.



The arrow shows interquartile range.

Range:

The range of a set of data is the difference between the maximum and minimum values, or the interval between these values. It is another measure of the spread of the data.

Comparing sample standard deviation, interquartile range and range:

The range is simple to evaluate and understand, but is sensitive to the odd extreme value and does not make effective use of all the information of the data. The sample standard deviation is also rather sensitive to extreme values but is easier to work with mathematically than the interquartile range.

Confidence Intervals:

Estimates are "best guesses" in some sense, and the sample variance gives some idea of the spread. Confidence intervals are another measure of spread, a range within which we are "pretty sure" that the parameter lies.

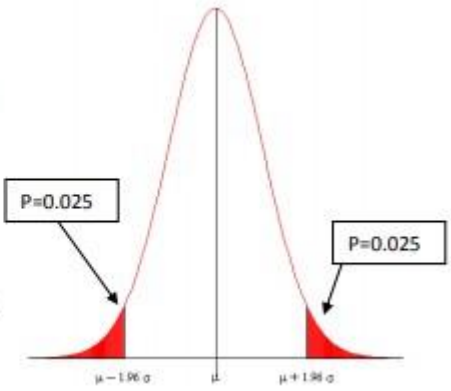
Normal data, variance known

Random sample $X_1, X_2, \ldots X_n$ from $N(\mu, \sigma^2)$, where σ^2 is known but μ is unknown. We want a confidence interval for μ.

Recall:

(i) $\bar{X} \sim N(\mu, {\sigma_{\bar{X}}}^2)$

(ii) With probability 0.95, a Normal random variables lies within 1.96 standard deviations of the mean.



$$P(\mu - 1.96\sigma_{\bar{X}} \leq \bar{X} \leq \mu + 1.96\sigma_{\bar{X}} \,|\mu) = 0.95$$

Since the variance of the sample mean is $\sigma_{\bar{X}}^2 = \sigma^2/n$ this gives

$$P\left(\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \bar{X} \leq \mu + 1.96\sqrt{\frac{\sigma^2}{n}} \,\middle|\, \mu\right) = 0.95$$

To infer the distribution of μ given $\bar{X}$ we need to use Bayes' theorem

$$P(\mu|\bar{X}) = \frac{P(\bar{X}\,|\mu)P(\mu)}{P(\bar{X})}$$

If the prior on μ is constant, then $P(\mu|\bar{X})$ is also Normal with mean $\bar{X}$ so

$$P(\bar{X} - 1.96\,\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}) = 0.95$$

Or

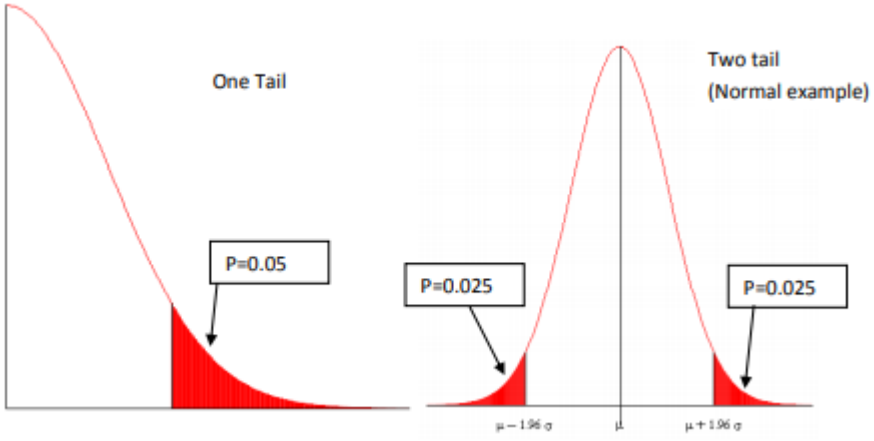
$$P\left(\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}} \,\middle|\, \bar{X}\right) = 095$$

A 95% confidence interval for μ is: $\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}}$ to $\bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}$.

Two tail versus one tail:

When the distribution has two ends (tails) where the likelihood goes to zero, the most natural choice of confidence interval is the regions excluding both tails, so a 95% confidence region means that 2.5% of the probability is in the high tail, 2.5% in the low tail. If the distribution is one sided, a one tail interval is more appropriate.

Example: 95% confidence regions.



Central Tendency

The following are the five measures of average or central tendency that are in common use :

- Arithmetic average or arithmetic mean or simple mean
- Median
- Mode
- Geometric mean
- Harmonic mean

Arithmetic mean, Geometric mean and Harmonic means are usually called Mathematical averages while Mode and Median are called Positional averages.

Arithmetic Mean:

To find the arithmetic mean, add the values of all terms and them divide sum by the number of terms, the quotient is the arithmetic mean. There are three methods to find the mean:

Direct method:

In individual series of observations x1, x2,… xn the arithmetic mean is obtained by following formula.

$$A.M. = \frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots\ldots\ldots x_{n-1} + x_n}{n}$$

Short-cut method:

This method is used to make the calculations simpler. Let A be any assumed mean (or any assumed number), d the deviation of the arithmetic mean, then we have,

$$M. = A + \frac{\sum fd}{N}$$

Step deviation method:

If in a frequency table the class intervals have equal width, say i than it is convenient to use the following formula.

$$M = A + \frac{\sum fu}{n} \times i$$

Where, u=(x-A)/I, and i is the length of the interval and A is the assumed mean.

PROPERTIES OF ARITHMETIC MEAN:

Property 1 The algebraic sum of the deviations of all the variates from their arithmetic mean is zero. Proof. Let X1, X2,… Xn be the values of the variates and their corresponding frequencies be f1, f2, …, fn respectively.

Let xi be the deviation of the variate Xi from the mean M, where i = 1,2, …, n. Then Xi = Xi –M, i = 1,2,…, n.

$$\sum_{i=1}^{n} fixi = \sum_{i-1}^{n} f_i(X_i - M)$$

$$= \mathbf{M} \sum_{i=1}^{n} f_i - M \sum_{i=1}^{n} f_i$$

$$= \mathbf{0}$$

Exercise for practice(Answer of these exercises not given. This is for practice only using above theory):

Q.1) Marks obtained by 9 students in statistics are given below.

52 75 40 70 43 65 40 35 48

calculate the arithmetic mean.

Q.2) Calculate the arithmetic mean of the following distribution

Variate :	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Frequency:	*20*	*43*	*57*	*61*	*72*	*45*	*39*

Q.3) Find the mean of the following distribution

Variate :	*0-10*	*10-20*	*20-30*	*30-40*	*40-50*
Frequency:	*31*	*44*	*39*	*58*	*12*

Median:

The median is defined as the measure of the central term, when the given terms (i.e., values of the variate) are arranged in the ascending or descending order of magnitudes. In other words the median is value of the variate for which total of the frequencies above this value is equal to the total of the frequencies below this value.

Due to Corner, —The median is the value of the variable which divides the group into two equal parts one part comprising all values greater, and the other all values less than the median.

Example:

The marks obtained, by seven students in a paper of Statistics are 15, 20, 23, 32, 34, 39, 48 the maximum marks being 50, then the median is 32 since it is the value of the 4^{th} term, which is situated such that the marks of 1^{st} , 2^{nd} and 3^{rd} students are less than this value and those of 5th, 6th and 7th students are greater than this value.

COMPUTATION OF MEDIAN

Median in individual series.

Let n be the number of values of a variate (i.e., total of all frequencies). First of all, we write the values of the variate (i.e., the terms) in ascending or descending order of magnitudes.

Here two cases arise:

Case 1. If n is odd then value of $\frac{(n+1)}{2}$ term gives the median.

Case2. If n is even then there are two central terms i.e., $\frac{n}{2}$ and $\frac{(n+1)}{2}$.The mean of these two values gives the median.

Median in continuous series (or grouped series). In this case, the median (Md) is computed by the following formula,

$$M_d = l + \frac{\frac{n}{2} - cf}{f} \times i$$

Where, m_d is the median,
l is lower limit of the median class,
cf is the total of all frequencies before median class,
f is the frequency of the median class,
i is the class width of the median class.

Mode:

The word mode is formed from the French word _La mode 'which means in fashion '. According to Dr. A. L. Bowle the value of the graded quantity in a statistical group at which the numbers registered are most numerous, is called the mode or the position of greatest density or the predominant value. '

According to other statisticians:
The value of the variable which occurs most frequently in the distribution is called the mode. ' The mode of a distribution is the value around the items tends to be most heavily concentrated. It may be regarded at the most typical value of the series.

Definition:
The mode is that value (or size) of the variate for which the frequency is maximum or the point of maximum frequency or the point of maximum density. In other words, the mode is the maximum ordinate of the ideal curve which gives the closest fit to the actual distribution.

Method to Compute the mode:
When the values (or measures) of all the terms (or items) are given. In this case the mode is the value (or size) of the term (or item) which occurs most frequently.
In continuous frequency distribution the computation of mode is done by the following formula,

Mode, $M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$
Where, l= lower limit of class,
f_1 =frequency of the modal class,
f_0 =frequency of the class just preceding to the modal class,
f_2 =frequency of the class just following of the modal class,
i= class interval.

Method of determining mode by the method of grouping frequencies.

This method is usually applied in the cases when there are two maximum frequencies against two different size of items. This method is also applied in the cases when it is possible that the effect of neighboring frequencies on the size of item (of maximum frequency) may be greater. The method is as follows:
Firstly the items are arranged in ascending or descending order and corresponding frequencies are written against them. The frequencies are then grouped in two and then in threes and then is fours (if necessary). In the first stage of grouping, they are grouped (i.e., frequencies are added) by taking, first and second, third and fourth, …, After it, the frequencies are added in threes. The frequencies are added in the following two ways:
>(i) First and second, third and fourth, fifth and sixth, seventh and eighth, …
(ii) Second and third, fourth and fifth, …
>(i) First, second and third; fourth, fifth and sixth, …
(ii) Second, third and fourth; fifth, sixth and seventh, …
(iii) Third, fourth and fifth; sixth seventh and eighth, … Now the items with maximum frequencies are selected and the item which contains the maximum is called the mode.

Relationship between Median and Mode:

For moderately asymmetrical distribution (or for asymmetrical curve), the relation
Mean – Mode = 3 (Mean - Median).
In such a case, first evaluate mean and median and then mode is determined by
Mode = 3 Median – 2 Mean.
If in the asymmetrical curve the area on the left of mode is greater than area on the right then
Mean < median < mode, i. e., (M < Md < M0)

Geometric Mean:

If $x_1, x_2, \ldots . x_n$ are n values of the variate x, none of which is zero. Then their geometric mean G is defined by,
$G = (x_1, x_2, \ldots x_n)^{\frac{1}{n}}$.
If $f_1, f_2, \ldots f_n$ are the frequencies of $x_1, x_2, \ldots x_n$ respectively, then geometric mean G is given by,

$$G = (x_1^{f1} x_2^{f2} \ldots x_n^{fn})^{\frac{1}{N}}$$

Harmonic Mean:

The Harmonic mean of a series of values is the reciprocal of the arithmetic means of their reciprocals. Thus if $x_1, x_2, \ldots, x_n$ (none of them being zero) is a series and H is its harmonic mean then,

$$\frac{1}{H} = \frac{1}{N}\left[\left(\frac{1}{x_1}\right) + \left(\frac{1}{x_2}\right) + \cdots + \left(\frac{1}{x_n}\right)\right]$$

Distributions

Random Sampling:

A population is a collection of all the values that may be included in a sample. A numerical value or a classification value may exist in the sample multiple times. A sample is a collection of certain values chosen from the population. The sample size, usually denoted by n, is the number of these values. If these values are chosen at random, the sample is called a random sample.

A sample can be considered a sequence of random variables: $x_1, x_2, \ldots x_n$ (" the first sample variable"," the second sample variable", . . .) that are independent and identically distributed. A concrete realized sample as a result of sampling is a sequence of values (numerical or classification values): $x_1, x_2, \ldots x_n$.
Note: random variables are denoted with upper case letters, realized values with lower case letters.

The sampling considered here is actually sampling with replacement. In other words, if a population is finite (or countably infinite), an element taken from the sample is replaced before taking another element.

Sampling Distributions:

If the expectation of the sampling distribution is μ and its variance is σ^2 , then the expectation of the sample mean is,

$$E(X) = \mu$$

and the variance is,

$$var(X) = \frac{\sigma^2}{n}$$

Here, n is sample size.
The standard deviation of the sample mean or its standard error is σ/√ n and it decreases as the sample size increases. If the population distribution is a normal distribution N (μ, σ^2), then the distribution of the sample mean is also a normal distribution, namely N(μ, σ^2). The distribution of (X) is however almost always normal in other cases, if just n is great enough (and the population distribution has an expected value and a finite variance). This is ensured by a classical approximation result.

Central Limit Theorem:

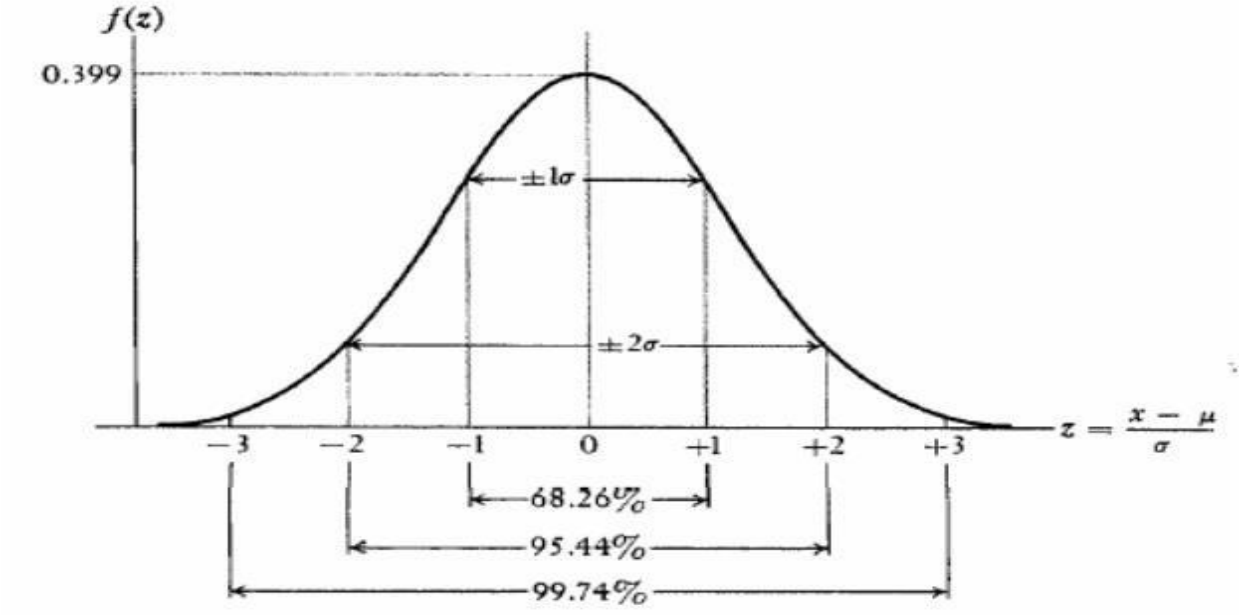
If the expectation of the population distribution is μ and its (finite) variance is σ, then the cumulative distribution function of the standardized random variable,

$$Z = \frac{X_{mean} - \mu}{\sigma/(\sqrt{n})}$$

approaches the cumulative distribution function Φ of the standard normal distribution in the limit as n increases. Usually a sample size of n = 30 is enough to normalize the distribution of accurately enough. If the population distribution X_{mean} is" well-shaped" (unimodal, almost symmetric) to begin with, a smaller sample size is enough (for example n = 5).

Normal Distribution:

The normal distribution is the most widely known and used of all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.



Why Normal Distribution?

Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution.

The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.

There is a very strong connection between the size of a sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

Standard Normal Distributed Curve Approach:

As you might suspect from the formula for the normal density function, it would be difficult and tedious to do the calculus every time we had a new set of parameters for μ and σ. So instead, we usually work with the standardized normal distribution, where μ = 0 and σ = 1, i.e., N (0,1). That is, rather than directly solve a problem involving a normally distributed variable X with mean μ and standard deviation σ, an indirect approach is used.

We first convert the problem into an equivalent one dealing with a normal variable measured in standardized deviation units, called a standardized normal variable. To do this, if $X \sim N(\mu,\sigma^2)$, then

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

If necessary, we can then convert back to the original units of measurement. To do this, simply note that, if we take the formula for Z, multiply both sides by σ, and then add μ to both sides, we get

$$X = Z\sigma + \mu$$

The interpretation of Z values is straightforward. Since σ = 1, if Z = 2, the corresponding X value is exactly 2 standard deviations above the mean. If Z = -1, the corresponding X value is one standard deviation below the mean. If Z = 0, X = the mean, i.e., μ.
It is very important to understand how the standardized normal distribution works, so we will spend some time here going over it. Recall that, for a random variable X,

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

Sampling Distribution of the sample variance:

The sampling distribution of the sample variance is a difficult concept, unless it can be assumed that the population distribution is normal. Let's make this assumption, so the sampling distribution of the sample variance can be formed using the X^2 distribution.
If random variables $U_1, \ldots U_v$ have the standard normal distribution and they are independent, a random variable.

$$V = U_1^2 + \cdots + U_v^2$$

has the X^2 distribution. Here v is a distribution's parameter, the number of degrees of freedom. The density function of the distribution is,

$$g(x) = \begin{cases} \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v-2}{2}} e^{-\frac{x}{2}}, \text{ when } x > 0 \\ 0, \text{ when } x \leq 0, \end{cases}$$

Where, Γ is the gamma-function. Despite its difficult form, the probabilities of the X^2distribution are numerically quite easily computed. Here are presented some density function of the X^2distribution (the number of degrees of freedom is denoted by n, the function is calculated in MATLAB.)

t-Distribution:

Earlier when considering the sample mean, it was required to know the standard deviation σ. If the standard deviation is not known, it is possible to proceed, but instead of a normal distribution, a t-distribution (or Student's distribution) is used. Additionally, the Central limit theorem isn't used, but the population distribution has to be normal.

If random variables U and V are independent, U has the standard normal distribution and V is χ 2 -distributed with v degrees of freedom, a random variable

$$T = \frac{U}{\sqrt{\frac{V}{v}}}$$

has a t-distribution with v degrees of freedom.

F-Distribution:

has the F-distribution with v1 and v2 degrees of freedom. In that case, random variable 1/F has also F-distribution, namely with v2 and v1 degrees of freedom. The formula for the density function of the F-distribution is quite complicated:

$$g(x) = \begin{cases} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} x^{\frac{v_1-2}{2}} \left(1 + \frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}}, \text{ when } x > 0 \\ 0, \text{ when } x \leq 0. \end{cases}$$

Estimators and Estimates

Statistical Interference:

The term "Statistical Interference" means generalizing a sample data to a larger phenomenon considering the degree of certainty. This whole phenomenon is known as Statistical inference.

The two common forms of statistical inference are:

- Estimation
- Null hypothesis tests of significance (NHTS)

There are two forms of estimation:

- Point estimation (maximally likely value for parameter)
- Interval estimation (also called confidence interval for parameter)

Note: Both estimation and NHTS are used to infer parameters. A parameter is a statistical constant that describes a feature about a phenomenon, population, or pdf.

Examples of parameters include:

- Binomial probability of "success" p (also called "the population proportion")
- Expected value μ (also called "the population mean")
- Standard deviation σ (also called the "population standard deviation")

Point Estimates:

The Point estimates are single points that are used to infer parameters directly.
For example:

- Sample proportion pˆ ("p hat") is the point estimator of p
- Sample mean x ("x bar") is the point estimator of μ
- Sample standard deviation s is the point estimator of σ

Notice the use of different symbols to distinguish estimators and parameters. More importantly, point estimates and parameters represent fundamentally different things.

- Point estimates are calculated from the data; parameters are not.
- Point estimates vary from study to study; parameters do not.
- Point estimates are random variables: parameters are constants.

The acronyms and symbols used in the below explanation:

$\hat{q}$= complement of the sample proportion
$\bar{x}$= sample mean
$\hat{p}$= sample proportion
$1-\alpha$= confidence level
CI= Confidence interval
LCL= lower confidence interval
m= margin of error
n= sample size
NHTS= null hypothesis test of significance
p= binomial success parameter
s= sample standard deviation
SDA= sampling distribution of mean
SEM= sampling error of the mean
SEP= Standard error of the proportion
UCL= upper confidence limit
α =**alpha level**
$\boldsymbol{\mu}$ =**expected value**
$\boldsymbol{\sigma}$ =**standard deviation parameter**

Sampling distribution of the mean:

Although point estimate $\bar{x}$ is a valuable reflections of parameter μ, it provides no information about the precision of the estimate. We ask: How precise is $\bar{x}$ as estimate of μ? How much can we expect any given $\bar{x}$ to vary from μ?

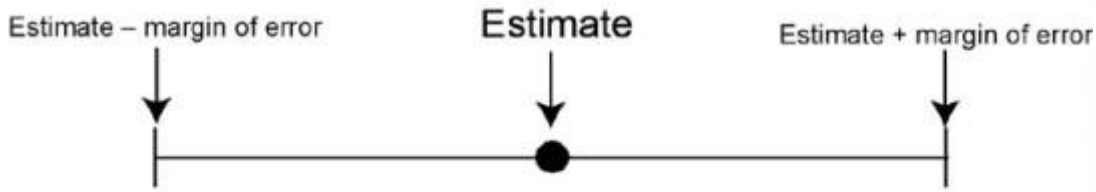
The variability of $\bar{x}$ as the point estimate of μ starts by considering a hypothetical distribution called the sampling distribution of a mean (SDM for short). Understanding the SDM is difficult because it is based on a thought experiment that doesn't occur in actuality, being a hypothetical distribution based on mathematical laws and probabilities. The SDM imagines what would happen if we took repeated samples of the same size from the same (or similar) populations done under the identical conditions. From this hypothetical experiment we "build" a pmf or pdf that is used to determine probabilities for various hypothetical outcomes.

Without going into too much detail, the SDM reveals that:
$\bar{x}$ is an unbiased estimate of μ.
The SDM tends to be normal (Gaussian) when the population is normal or when the sample is adequately large.
The standard deviation of the SDM is equal to $\sigma/\sqrt{n}$. This statistic—which is called the standard error of the mean (SEM)—predicts how closely the $\bar{x}$s in the SDM is likely to cluster around the value of μ and is a reflection of the precision of $\bar{x}$ as an estimate of μ:

$$SEM = \sigma/\sqrt{n}$$

Confidence Interval for μ when σ is known before hand:

To gain further insight into μ, we surround the point estimate with a margin of error:



This forms a confidence interval (CI). The lower end of the confidence interval is the lower confidence limit (LCL). The upper end is the upper confidence limit (UCL).

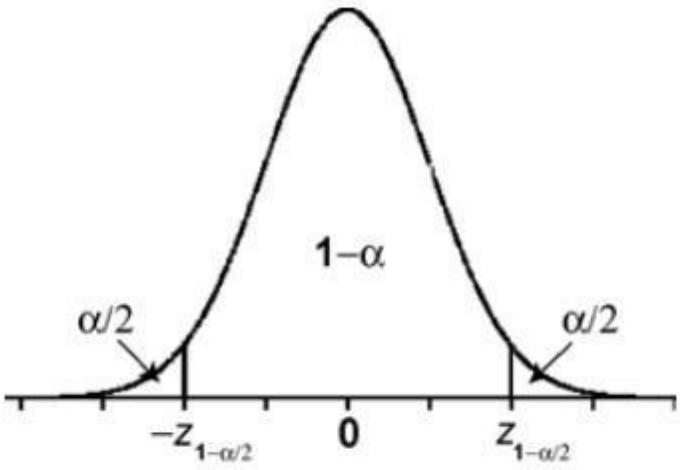
Note: The margin of error is the plus-or-minus wiggle-room drawn around the point estimate; it is equal to half the confidence interval length.

Let (1−α)100% represent the confidence level of a confidence interval. The α ("alpha") level represents the "lack of confidence" and is the chance the researcher is willing to take in not capturing the value of the parameter.

A (1−α)100% CI for μ is given by:

$$\bar{x} \pm (z_{1-\frac{\alpha}{2}})(SEM)$$

The $z_{1-\frac{\alpha}{2}}$ in this formula is the z quantile association with a $(1-\alpha)$ level of confidence. The reason we use $z_{1-\frac{\alpha}{2}}$ instead of $z_{1-\alpha}$ in this formula is because the random error (imprecision) is split between underestimates (left tail of the SDM) and overestimates (right tail of the SDM). The confidence level 1−α area lies between $-z_{1-\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$.



You may use the z/t table on the Stat Primer website to determine z quantiles for various levels of confidence. Here are the common levels of confidence and their associated alpha levels and z quantiles:

(1−α)100%	α	$z_{1-\alpha/2}$
90%	.10	1.64
95%	.05	1.96
99%	.01	2.58

Sample Size Requirements for estimating μ with confidence:

One of the questions we often face is "How much data should be collected?" Collecting too much data is a waste of time and money. Also, by collecting fewer data points we can devote more time and energy into making these measurements accuracy. However, collecting too little data renders our estimate too imprecise to be useful.

To address the question of sample size requirements, let m represent the desired margin of error of an estimate. This is equivalent to half the ultimate confidence interval length.

Note that margin of error $m = z^2_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ Solving this equation for n derives,

$$n = z^2_{1-\frac{\alpha}{2}} \times \frac{\sigma^2}{m^2}$$

We always round results from this formula up to the next integer to ensure that we have a margin of error no greater than m.

Estimating p with confidence:

Sampling distribution of the proportion:

Estimating parameter p is analogous to estimating parameter μ. However, instead of using $\bar{x}$ as an unbiased point estimate of μ, we use $\hat{p}$ as an unbiased estimate of p.

Here, the symbol $\hat{p}$ represents the sample proportion.

$$\hat{p} = (number\ of\ successes\ in\ the\ sample)/n$$

In samples that are large, the sampling distribution of $\hat{p}$ is approximately normal with a mean of p and standard error of the proportion $SEP = \sqrt{(pq/n)}$ where q = 1 – p. The SEP quantifies the precision of the sample proportion as an estimate of parameter p.

Confidence interval for p:

This approach should be used only in samples that are large. "Use this rule to determine if the sample is large enough: if $npq \geq 5$ proceed with this method. (Call this "the npq rule").

An approximate (1−α)100% CI for p is given by,

$$\hat{p} \pm (z_{1-\frac{\alpha}{2}})(SEP)$$

Where the estimated $SEP = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Sample size requirement for estimating p with confidence:

In planning a study, we want to collect enough data to estimate p with adequate precision. Earlier in the chapter we determined the sample size requirements to estimate μ with confidence. We apply a similar method to determine the sample size requirements to estimate p.

Let m represent the margin of error. This provides the "wiggle room" around $\hat{p}$ for our confidence interval and is equal to half the confidence interval length. To achieve margin of error m,

$$n = \frac{z^2_{1-\frac{\alpha}{2}} p^* q^*}{m^2}$$

where p^* represent the an educated guess for the proportion and $q^* = 1 - p^*$.

When no reasonable guess of p is available, use $p^* = 0.50$ to provide a "worst-case scenario" sample size that will provide more than enough data.

Confidence Intervals

The standard deviation of a sampling distribution is called the standard error of the mean (basically they are measures of sampling variability or estimates of dispersion or spread).

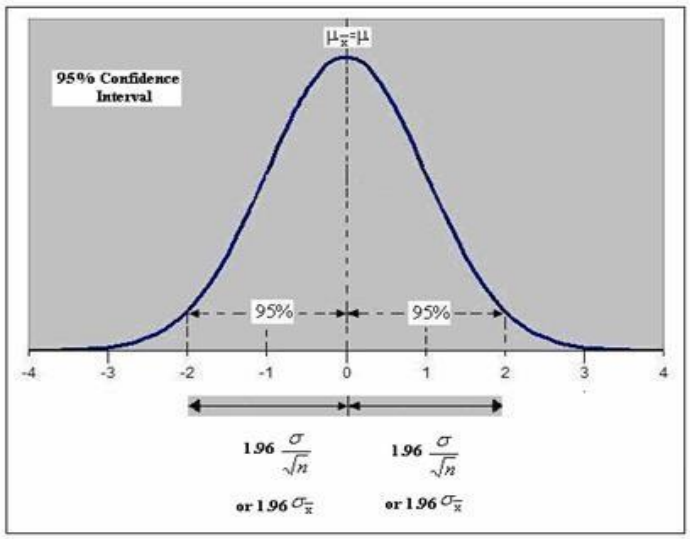
A standard error generally has a level of confidence associated with it. You use the standard error of the mean to determine how close to the true population mean you can expect your sample mean to be and how much confidence you can place in that expectation.

To reduce the amount of sampling variability you can make your sample larger and more homogeneous.

Level of Confidence or Confidence Intervals:

Confidence levels are used when two sets of data are being compared. A confidence level is the likelihood of obtaining a particular result by chance rather than due to a truly significant difference in the two sets of data. How well the sample statistic estimates the underlying population value is always an issue. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

A 95% confidence interval means that there is a 95% chance that the confidence interval contains the population mean.

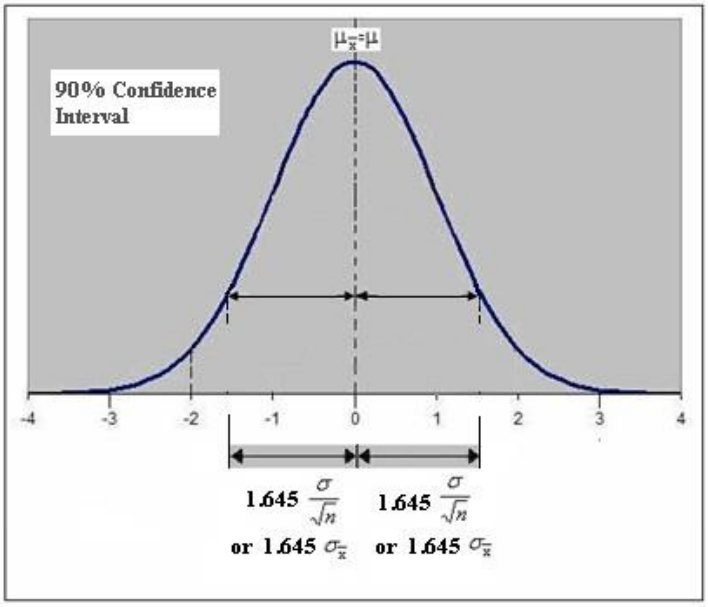


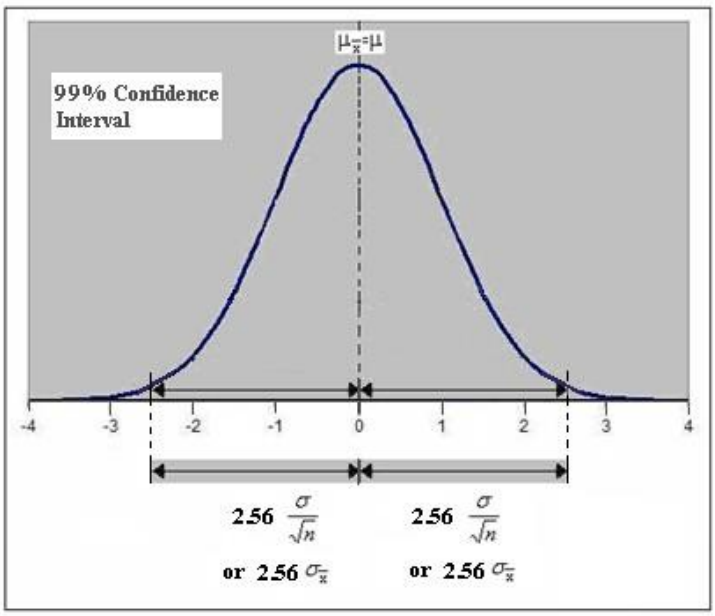
The standard normal distribution is sometimes called the Z distribution. A Z score always reflects the number of standard deviations a particular score is above or below the mean.

If you are calculating a 95% confidence interval, then z = 1.96

If you are calculating a 90% confidence interval, then z = 1.645

If you are calculating a 99% confidence interval, then z = 2.56





When a sample size is large, the confidence interval for the population mean is calculated using the formula:

$$\bar{x} \pm Z(\sigma \backslash (\sqrt{n}))$$

Confidence Intervals When σ (Population Standard Deviation) is Unknown:

In many situations, the population standard deviation is not known. With a large sample size (n ≥ 30) you can replace the σ with the sample standard deviation S_x and solve using the formula as an interval estimator. The margin of error can be determined once the standard deviation and the sample size are known. It represents a statistic expressing the amount of random sampling error in a survey's results.

Here, $Z(S_x\backslash(\sqrt{n}))$ is the margin of error.
So the confidence intervals is $\bar{X}\pm the\ margin\ of\ error$..

$$CI = \bar{X}\pm Z(S_x\backslash\sqrt{n})$$

Confidence intervals are constructed with the following formula:

statistic ± (critical value) * (standard deviation of statistic)

Margin of error:

Once we have a statistic, we want to subtract and add a margin of error, or "wiggle room." This margin of error (ME) has a simple formula:
ME= (critical value) *(standard deviation of statistic)

The standard deviation of a statistic tells us "typically" how off a sample statistic will be from the true parameter (either p or μ). The critical value is the number of these standard deviations that we will subtract from our sample statistic: the larger the critical value, the more standard deviations we subtract/add; the more we subtract/add to our statistic, the more confident we can be that our method will capture the true parameter a high percentage of times. This critical value will depend on the confidence level.

The width of these intervals depends on what is called a confidence level. A confidence level tells us the "hit rate" of a certain method of sampling and constructing intervals.

CONFIDENCE INTERVALS ABOUT MEANS:

The t- distribution:

When we want to construct a confidence interval about a sample mean, we take a sample. This sample will have a standard deviation s_x that is probably a little bit smaller than the population standard deviation σ_x. Therefore, when we are finding our critical value – the number of standard deviations we want to subtract/add to the mean – we need to use a slightly larger value than the traditional z∗. This value is called t∗.

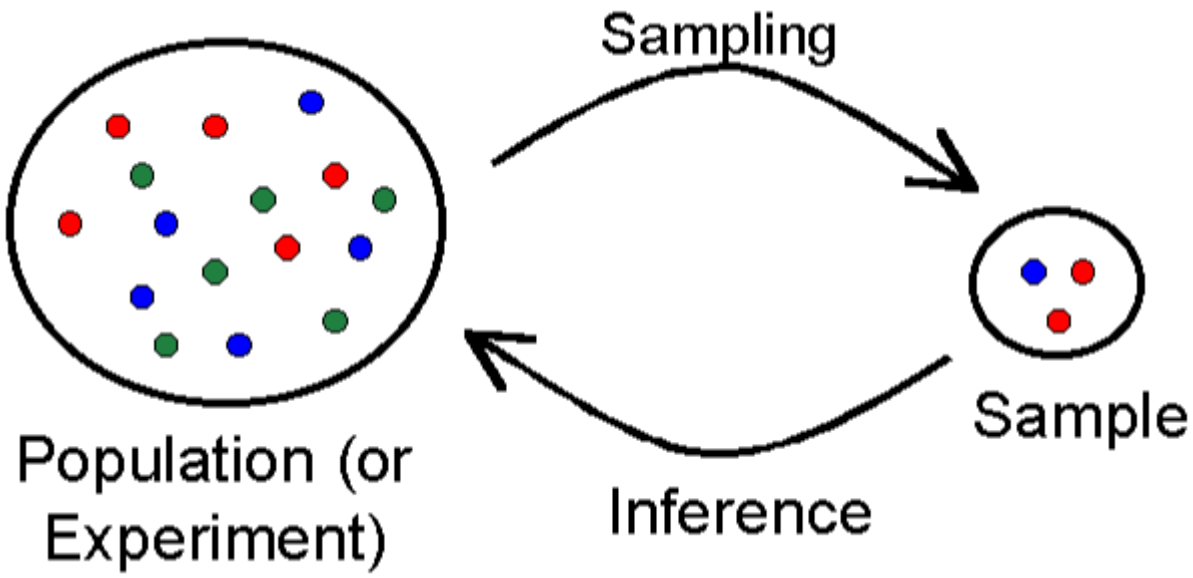
The t-distribution is symmetric, but has more value in the tails than the standard Normal curve. However, as the sample size increases, s_x becomes closer and closer to σ_x; therefore, t∗ will start to become very close to z∗.
The shape of the t-distribution is therefore defined by how big the sample size n is. We define the ﬀ-distribution by the sample size minus 1, or the degrees of freedom.

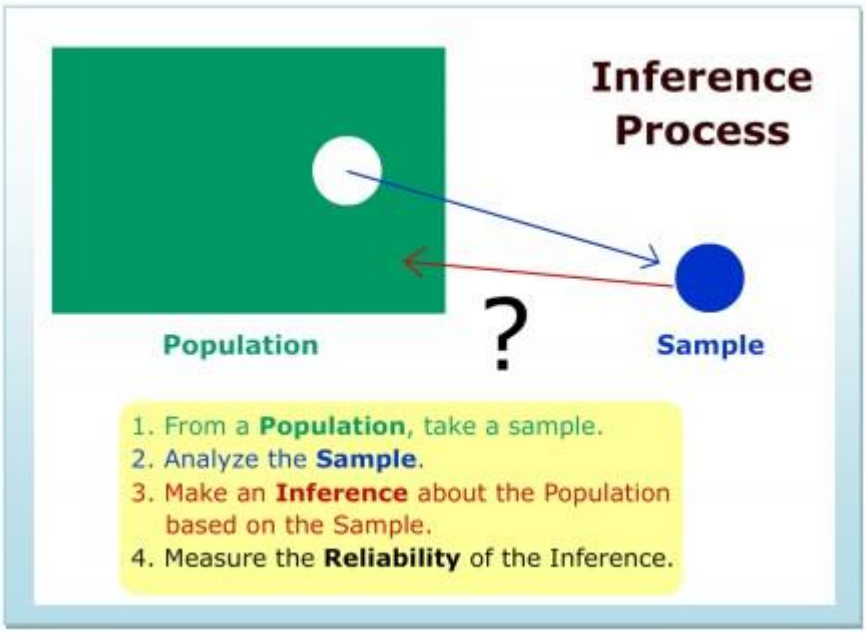
CONSTRUCTING A CONFIDENCE INTERVAL ABOUT A PROPORTION Calculator Command: **STAT -> TESTS -> 8: TInterval**		
Condition	What it gives us	Helpful information
1. Random	Unbiased estimator -> mean of sampling distribution of $\hat{p}$ is p	Don't just say "random." Include *context*, and make sure that you identify WHETHER OR NOT THE CONDITION IS MET. Put a "yes" or a check mark.
2. 10% condition: $n \leq \frac{1}{10}N$	Trials will be reasonably independent of one another.	Trials will be reasonably independent because the probability of success will stay about the same each time (since you didn't sample too much from the population).
3. Large counts/Normality a. Population is Normal, or b. $n \geq 30$ (CLT), or c. Sample data is reasonably symmetric w/ no strong skew or outliers	The sampling distribution of $\hat{p}$ will be approximately Normal.	**If the shape of the pop'n dist. is unknown and $n < 30$, you must look at a *plot* of the sample data.**

Name	Formula
One-proportion z-interval (Plan step)	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
One-sample t-interval (Plan step)	$\bar{x} \pm t^* \left(\frac{s_x}{\sqrt{n}}\right)$

Inferential Statistics:

The reason we conduct statistical research is to obtain an understanding about phenomena in a population. For example, we may want to know if a potential drug is effective in treating a disease. Since it is not feasible or ethical to distribute an experimental drug to the entire population, we instead must study a small subset of the population called a sample. We then analyze the sample and make an inference about the population based on the sample. Using probability theory and the Central Limit Theorem, we can then measure the reliability of the inference.





Point Estimation:

The example above is an example of Estimation, a branch of Inferential Statistics where sample statistics are used to estimate the values of a population parameter. We were trying to estimate the population mean (μ) based on the sample mean ($\bar{X}$)

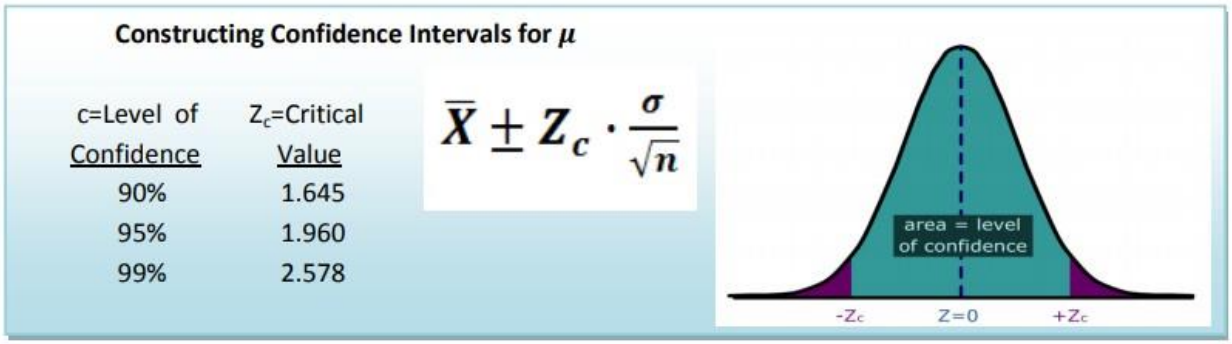
Interval Estimation:

A point estimate is our "best" estimate of a population parameter, but will most likely not exactly equal the parameter. Instead, we will choose a range of values called an Interval Estimate that is likely to include the value of the population parameter.

If the Interval Estimate is symmetric, the distance from the Point Estimator to either endpoint of the Interval Estimate is called the Margin of Error.

Confidence Intervals:

Using probability and the Central Limit Theorem, we can design an Interval Estimate called a Confidence Interval that has a known probability (Level of Confidence) of capturing the true population parameter.
To find a confidence interval for the population mean (μ) when the population standard deviation (σ) is known, and n is sufficiently large, we can use the Standard Normal Distribution probability distribution function to calculate the critical values for the Level of Confidence:



Key Points of Confidence Intervals:

- The confidence interval is constructed from random variables calculated from sample data and attempts to predict an unknown but fixed population parameter with a certain level of confidence.
- Increasing the level of confidence will always increase the margin of error.
- It is impossible to construct a 100% Confidence Interval without taking a census of the entire population.
- Think of the population mean like a dart that always goes to the same spot, and the confidence interval as a moving target that tries to "catch the dart." A 95% confidence interval would be like a target that has a 95% chance of catching the dart.

Confidence Interval for Population Proportion:

Recall from the section on random variables the binomial distribution where p represented the proportion of successes in the population. The binomial model was analogous to coin-flipping, or yes/no question polling. In practice, we want to use sample statistics to estimate the population proportion (p).

The sample proportion ($\hat{p}$) is the proportion of successes in the sample of size n and is the point estimator for p. Under the Central Limit Theorem, if $np > 5$ and $n(1-p) > 5$, the distribution of the sample proportion $\hat{p}$will have an approximately Normal Distribution.

Normal Distribution for $\hat{p}$ if Central Limit Theorem conditions are met.

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Using this information we can construct a confidence interval for p, the population proportion:

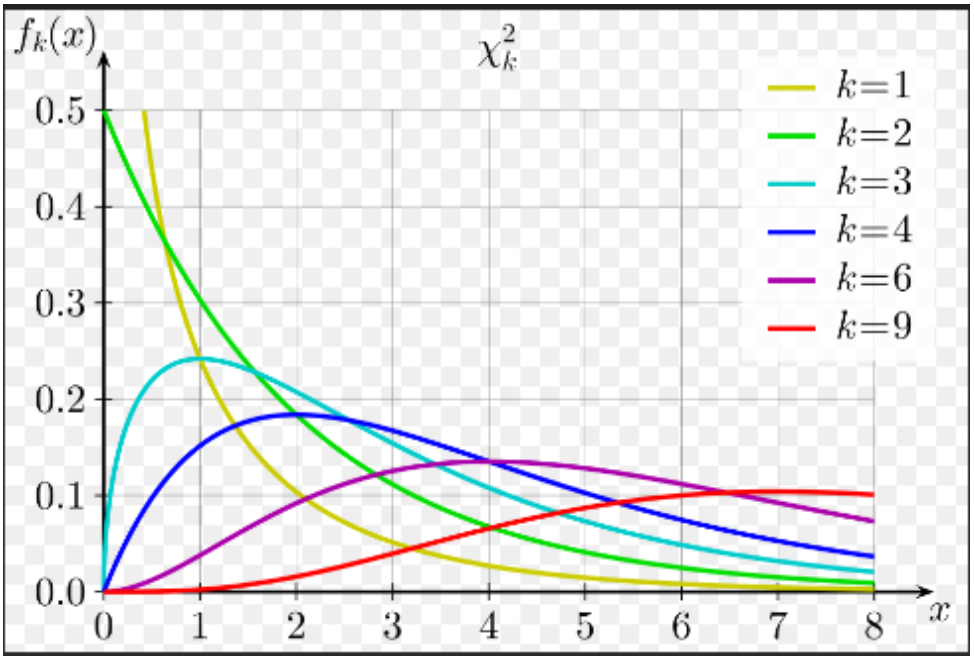
Confidence interval for p: $\hat{p} \pm Z\sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Point Estimator for Population Standard Deviation:

We often want to study the variability, volatility or consistency of a population. For example, two investments both have expected earnings of 6% per year, but one investment is much riskier, having higher ups and downs. To estimate variation or volatility of a data set, we will use the sample standard deviation (s) as a point estimator of the population standard deviation (σ).

The Chi-square □□ Distribution:

The Chi-square distribution is a family of distributions related to the Normal Distribution as it represents a sum of independent squared standard Normal Random Variables. Like the Student's t distribution, the degrees of freedom will be n-1 and determine the shape of the distribution. Also, since the Chi-square represents squared data, the inference will be about the variance rather than the standard deviation.



Confidence Interval for Population Variance and Standard Deviation:

Since the Chi-square represents squared data, we can construct confidence intervals for the population variance ($\sigma 2$), and take the square root of the endpoints to get a confidence interval for the population standard deviation. Due to the skewness of the Chi-square distribution the resulting confidence interval will not be centered at the point estimator, so the margin of error form used in the prior confidence intervals doesn't make sense here.

Hypothesis Testing

In the prior section we used statistical inference to make an estimate of a population parameter and measure the reliability of the estimate through a confidence interval. In this section, we will explore in detail the use of statistical inference in testing a claim about a population parameter, which is the heart of the scientific method used in research.

Procedures of Hypotheses Testing and the Scientific Method:

The actual conducting of a hypothesis test is only a small part of the scientific method. After formulating a general question, the scientific method consists of: the designing of an experiment, the collecting of data through observation and experimentation, the testing of hypotheses, and the reporting of overall conclusions. The conclusions themselves lead to other research ideas making this process a continuous flow of adding to the body of knowledge about the phenomena being studied. Others may choose a more formalized and detailed set of procedures, but the general concepts of inspiration, design, experimentation, and conclusion allow one to see the whole process.

Hypotheses and Hypothesis Testing:

For purposes of testing, we need to design hypotheses that are statements about population parameters. Some examples of hypotheses:

- At least 20% of juvenile offenders are caught and sentenced to prison.
- The mean monthly income for college graduates is $5000.
- The mean standardized test score for schools in Cupertino is the same as the mean scores for Los Altos.
- The lung cancer rates in California are lower than the rates in Texas.
- The standard deviation of the New York Stock Exchange today is greater than 10 percentage points per year.

These same hypotheses could be written in symbolic notation:

p **> 0.20**
μ **> 5000**
μ**1 =** μ**2**
p**1 <** p**2**
σ **> 10**

Hypothesis Testing is a procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected. This hypothesis that is tested is called the Null Hypothesis designated by the symbol Ho. If the Null Hypothesis is unreasonable and needs to be rejected, then the research supports an Alternative Hypothesis designated by the symbol Ha.

Null Hypothesis (Ho):

A statement about the value of a population parameter that is assumed to be true for the purpose of testing.

Alternative Hypothesis (Ha):

A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

From these definitions it is clear that the Alternative Hypothesis will necessarily contradict the Null Hypothesis; both cannot be true at the same time. Some other important points about hypotheses:
>Hypotheses must be statements about population parameters, never about sample statistics.
>In most hypotheses tests, equality (=, ≤, ≥) will be associated with the Null Hypothesis while non-equality (≠,) will be associated with the Alternative Hypothesis.
>It is the Null Hypothesis that is always tested in attempt to "disprove" it and support the Alternative Hypothesis. This process is analogous in concept to a "proof by contradiction" in Mathematics or Logic, but supporting a hypothesis with a level of confidence is not the same as an absolute mathematical proof.

Statistical Model and Test Statistic:

To test a hypothesis we need to use a statistical model that describes the behavior for data and the type of population parameter being tested. Because of the Central Limit Theorem, many statistical models are from the Normal Family, most importantly the Z, t, χ2 , and F distributions. Other models that are used when the Central Limit Theorem is not appropriate are called non-parametric Models and will not be discussed here. Each chosen model has requirements of the data called model assumptions that should be checked for appropriateness. For example, many models require the sample mean has approximately a Normal Distribution, which may not be true for some smaller or heavily skewed data sets. Once the model is chosen, we can then determine a test statistic, a value derived from the data that is used to decide whether to reject or fail to reject the Null Hypothesis.

Statistical Model	Test Statistic
Mean vs. Hypothesized Value	$t = \frac{\bar{X}-\mu_o}{s/\sqrt{n}}$
Proportion vs. Hypothesized Value	$Z = \frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_0)}{n}}}$
Variance vs. Hypothesized Value	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

Errors in Decision Making:

Whenever we make a decision or support a position, there is always a chance we make the wrong choice. The hypothesis testing process requires us to either to reject the Null Hypothesis and support the Alternative Hypothesis or fail to reject the Null Hypothesis. This creates the possibility of two types of error:

Type I Error
Rejecting the null hypothesis when it is actually true.

Type II Error
Failing to reject the null hypothesis when it is actually false.

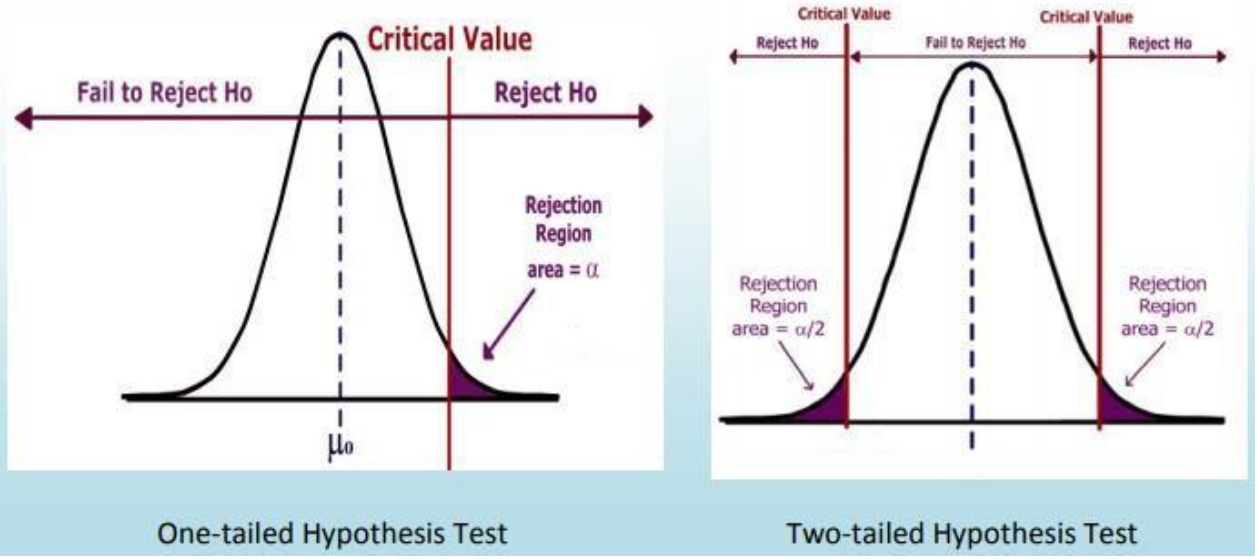
	Fail to Reject Ho	Reject Ho
Ho is true	Correct Decision	**Type I error**
Ho is False	**Type II error**	Correct Decision

In designing hypothesis tests, we need to carefully consider the probability of making either one of these errors.

Critical Value and Rejection Region:

Once the significance level of the test is chosen, it is then possible to find region(s) of the probability distribution function of the test statistic that would allow the Null Hypothesis to be rejected. This is called the Rejection Region and the boundary between the Rejection Region and the "Fail to Reject" is called the Critical Value.

There can be more than one critical value and rejection region. What matters is that the total area of the rejection region equals the significance level α.



One and Two tailed Tests:

A test is one-tailed when the Alternative Hypothesis, Ha, states a direction, such as:

H0: The mean income of females is less than or equal to the mean income of male.
Ha: The mean income of females is greater than males.

Since equality is usually part of the Null Hypothesis, it is the Alternative Hypothesis which determines which tail to test.

A test is two-tailed when no direction is specified in the alternate hypothesis Ha, such as:

H0: The mean income of females is equal to the mean income of males.
Ha: The mean income of females is not equal to the mean income of the males.

In a two tailed-test, the significance level is split into two parts since there are two rejection regions. In hypothesis testing where the statistical model is symmetrical (e.g.: The Standard Normal Z or Student's t distribution) these two regions would be equal. There is a relationship between a confidence interval and a two-tailed test: If the level of confidence for a confidence interval is equal to 1-α, where α is the significance level of the two-tailed test, the critical values would be the same.

Deciding when to conduct a one or two-tailed test is often controversial and many authorities even go so far as to say that only two-tailed tests should be conducted. Ultimately, the decision depends on the wording of the problem. If we want to show that a new diet reduces weight, we would conduct a lower tailed test since we don't care if the diet causes weight gain. If instead, we wanted to determine if mean crime rate in California was different from the mean crime rate in the United States, we would run a two-tailed test, since different means greater than or less than.

H_a: $\mu > \mu_0$ means test the upper tail and is also called a right-tailed test.
H_a: $\mu < \mu_0$ means test the lower tail and is also called a left-tailed test.
H_a: $\mu \neq \mu_0$ means test both tails.

Collect and Analyse Experimental Data:

After designing the experiment, the next procedure would be to actually collect and verify the data. For the purposes of statistical analysis, we will assume that all sampling is either random, or uses an alternative technique that adequately simulates a random sample.

Data Verification:

After collecting the data but before running the test, we need to verify the data. First, get a picture of the data by making a graph (histogram, dot plot, box plot, etc.) Check for skewness, shape and any potential outliers in the data.

Working with Outliers:

An outlier is data point that is far removed from the other entries in the data set. Outliers could be caused by:

- Mistakes made in recording data
- Data that don't belong in population
- True rare events

The first two cases are simple to deal with as we can correct errors or remove data that that does not belong in the population. The third case is more problematic as extreme outliers will increase the standard deviation dramatically and heavily skew the data.

In the Black Swan, Nicholas Taleb argues that some populations with extreme outliers should not be analyzed with traditional confidence intervals and hypothesis testing.9 He defines a Black Swan to be an unpredictable extreme outlier that causes dramatic effects on the population. A recent example of a Black Swan was the catastrophic drop in the value of unregulated Credit Default Swap (CDS) real estate insurance investments which caused the near collapse of international banking system in 2008. The traditional statistical analysis that measured the risk of the CDS investments did not take into account the consequence of a rapid increase in the number of foreclosures of homes. In this case, statistics that measure investment performance and risk were useless and created a false sense of security for large banks and insurance companies.

Example

Here are the quarterly home sales for 10 realtors

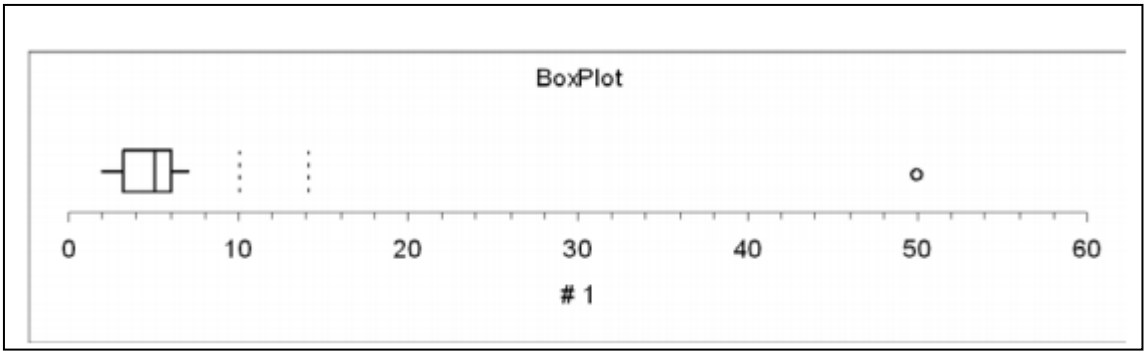
2 2 3 4 5 5 6 6 7 50

	With outlier	Without Outlier
Mean	9.00	4.44
Median	5.00	5.00
Standard Deviation	14.51	1.81
Interquartile Range	3.00	3.50

In this example, the number 50 is an outlier. When calculating summary statistics, we can see that the mean and standard deviation are dramatically affected by the outlier, while the median and the interquartile range (which are based on the ranking of the data) are hardly changed. One solution when dealing with a population with extreme outliers is to use inferential statistics using the ranks of the data, also called non-parametric statistics.

Using Box Plot to find outliers:

- The "box" is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR's from the box (inner fence)
- Probable outliers are more than 3 IQR's from the box (outer fence)
- In the box plot below of the realtor example, the dotted lines represent the "fences" that are 1.5 and 3 IQR's from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



Logic of Hypothesis Testing:

After the data is verified, we want to conduct the hypothesis test and come up with a decision, whether or not to reject the Null Hypothesis. The decision process is similar to a "proof by contradiction" used in mathematics:

- We assume Ho is true before observing data and design Ha to be the complement of Ho.
- Observe the data (evidence). How unusual are these data under Ho?
- If the data are too unusual, we have "proven" Ho is false: Reject Ho and support Ha (strong statement).
- If the data are not too unusual, we fail to reject Ho. This "proves" nothing and we say data are inconclusive. (weak statement).
- We can never "prove" Ho, only "disprove" it.
- "Prove" in statistics means support with (1-α)100% certainty. (example: if α=.05, then we are at least 95% confident in our decision to reject Ho.

Decision Rule – Two methods, Same Decision:

Earlier we introduced the idea of a test statistic which is a value calculated from the data under the appropriate Statistical Model from the data that can be compared to the critical value of the Hypothesis test. If the test statistic falls in the rejection region of the statistical model, we reject the Null Hypothesis.
Recall that the critical value was determined by design based on the chosen level of significance α. The more preferred method of making decisions is to calculate the probability of getting a result as extreme as the value of the test statistic. This probability is called the p-value, and can be compared directly to the significance level.

Comparing p-value to α:

Both the p-value and α are probabilities of getting results as extreme as the data assuming Ho is true. The p-value is determined by the data is related to the actual probability of making Type I error (Rejecting a True Null Hypothesis). The smaller the p-value, the smaller the chance of making Type I error and therefore, the more likely we are to reject the Null Hypothesis. The significance level α is determined by design and is the maximum probability we are willing to accept of rejecting a true H0.

This p-value method of comparison is preferred to the critical value method because the rule is the same for all statistical models: Reject Ho if p-value < α.
Let's see why these two rules are equivalent by analyzing a test of mean vs. hypothesized value.

Decision is Reject Ho

Ho: $\mu = 10$

Ha: $\mu > 10$

Design: Critical value is determined by significance level α.

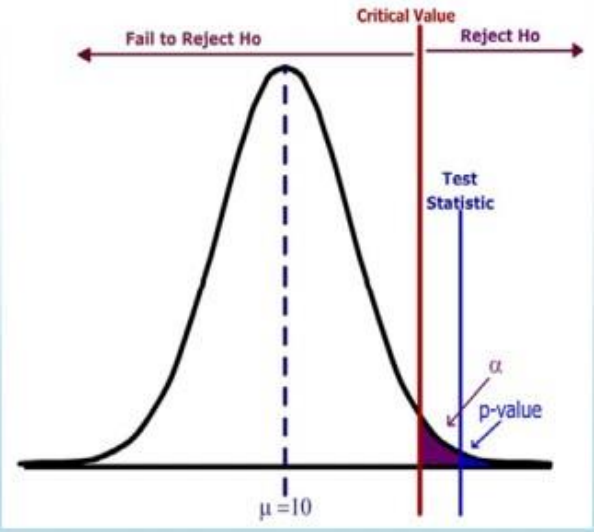
Data Analysis: p-value is determined by test statistic

Test statistic falls in rejection region.

p-value (blue) < α (purple)

Reject Ho.

Strong statement: Data supports the Alternative Hypothesis.



In this example, the test statistic lies in the rejection region (the area to the right of the critical value). The p-value (the area to the right of the test statistic) is less than the significance level (the area to the right of the critical value). The decision is Reject Ho.

Decision is Fail to Reject Ho

Ho: $\mu = 10$

Ha: $\mu > 10$

Design: critical value is determined by significance level α.

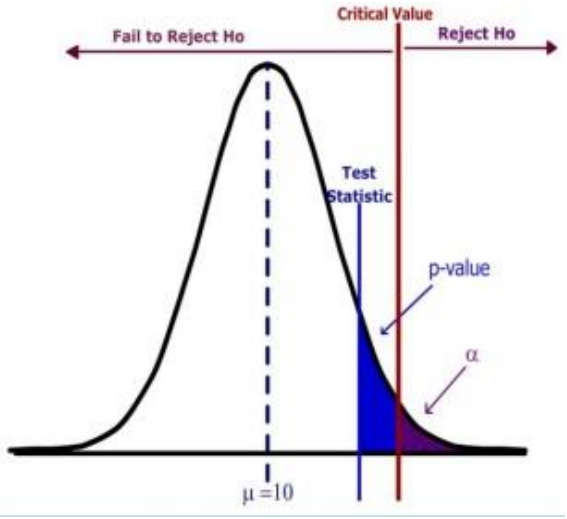
Data Analysis: p-value is determined by test statistic

Test statistic does not fall in the rejection region.

p-value (blue) > α (purple)

Fail to Reject Ho.

Weak statement: Data is inconclusive and does not support the Alternative Hypothesis.



Be consistent with the results of the Hypothesis Test:

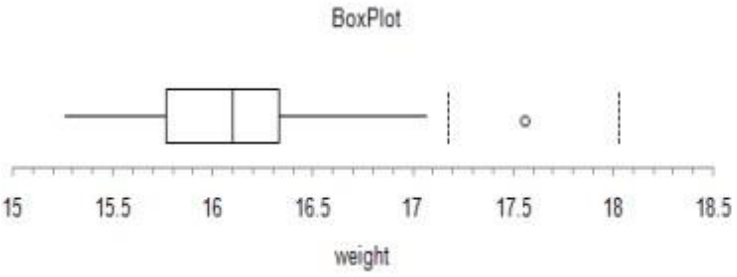
Rejecting Ho requires a strong statement in support of Ha, while failing to reject Ho does NOT support Ho, but requires a weak statement of insufficient evidence to support Ha.

Example: A researcher wants to support the claim that, on average, students send more than 1000 text messages per month and the research hypotheses are Ho: μ=1000 vs. Ha: μ>1000 Conclusion if Ho is rejected: The mean number of text messages sent by students exceeds 1000. Conclusion if Ho is not rejected: There is insufficient evidence to support the claim that the mean number of text messages sent by students exceeds 1000.

Type I Error:

Type I error would be to reject the Null Hypothesis and say the machine is not running properly when in fact it was operating properly. Since the company does not want to needlessly stop production and recalibrate the machine, the statistician chooses to limit the probability of Type I error by setting the level of significance (α) to 5%.

The statistician now conducts the experiment and samples 36 bottles in the last hour and determines from a box plot of the data that there is one unusual observation of 17.56 ounces. The value is rechecked and kept in the data set. Next, the sample mean and the test statistic are calculated.



Next, the sample mean and the test statistic are calculated.

$$\bar{X} = 16.12 \text{ ounces} \qquad Z = \frac{16.12-16}{0.5/\sqrt{36}} = 1.44$$

The decision rule under the critical value method would be to reject the Null Hypothesis when the value of the test statistic is in the rejection region. In other words, reject Ho when Z >1.96 or Z<-1.96.
Based on this result, the decision is fail to reject Ho since the test statistic does not fall in the rejection region.

Alternatively (and preferably) the statistician would use the p-value method of decision rule. The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic, so in this example we find the probability that Z < -1.44 or Z > 1.44 (the area shaded blue). Using a calculator, computer software or a Standard Normal table, the p-value=0.1498. Since the pvalue is greater than α, the decision again is failed to reject Ho.

Finally the statistician must report the conclusions and make a recommendation to the company's management:

Note :- There is insufficient evidence to conclude that the machine that fills 16-ounce soy sauce bottles is operating improperly. This conclusion is based on 36 measurements taken during a single hour's production run. I recommend continued monitoring of the machine during different employee shifts to account for the possibility of potential human error.

Type II Error and Statistical Power:

In the prior example, the statistician failed to reject the Null Hypothesis because the probability of making Type I error (rejecting a true Null Hypothesis) exceeded the significance level of 5%. However, the statistician could have made Type II error if the machine is really operating improperly. One of the important and often overlooked tasks is to analyze the probability of making Type II error (β). Usually statisticians look at statistical power which is the complement of β.

Beta (β): The probability of failing to reject the null hypothesis when it is actually false.
Power (or Statistical Power): The probability of rejecting the null hypothesis when it is actually false.

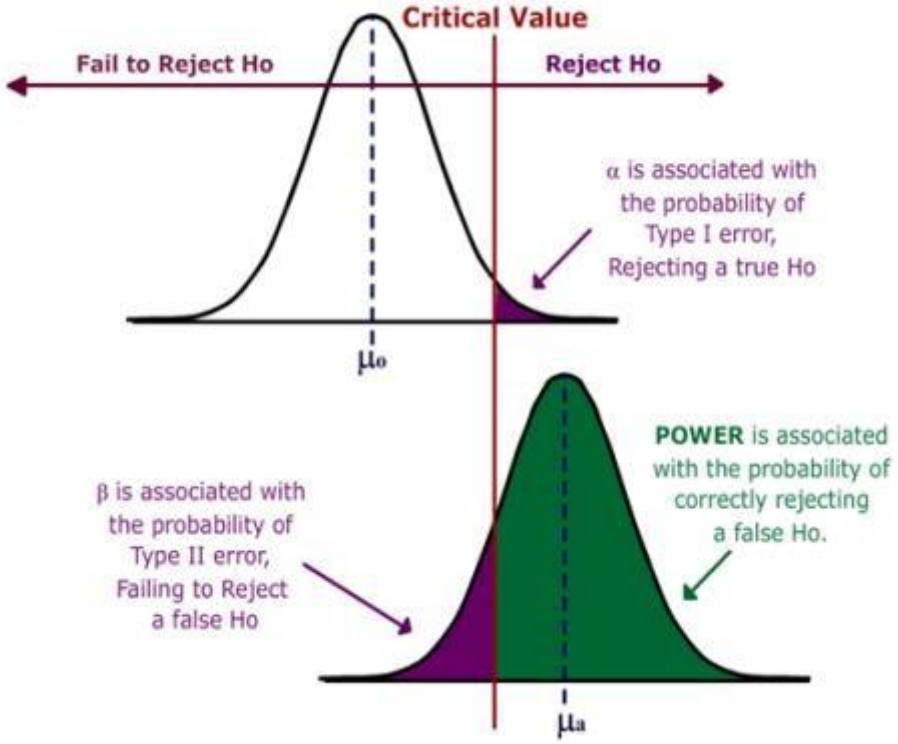
Both beta and power are calculated for specific possible values of the Alternative Hypothesis.

	Fail to Reject Ho	Reject Ho
Ho is true	$1-\alpha$	α Type I error
Ho is False	β Type II error	$1-\beta$ Power

If a hypothesis test has low power, then it would difficult to reject Ho, even if Ho were false; the research would be a waste of time and money. However, analyzing power is difficult in that there are many values of the population parameter that support Ha. For example, in the soy sauce bottling example, the Alternative Hypothesis was that the mean was not 16 ounces. This means the machine could be filling the bottles with a mean of 16.0001 ounces, making Ha technically true. So when analyzing power and Type II error we need to choose a value for the population mean under the Alternative Hypothesis (μa) that is "practically different" from the mean under the Null Hypothesis (μo). This practical difference is called the effect size.

Suppose we are conducting a one-tailed test of the population mean:
Ho: $\mu = \mu 0$ Ha: $\mu > \mu 0$
Consider the two graphs shown to the right. The top graph is the distribution of the sample mean under the Null Hypothesis that we covered in an earlier section. The area to the right of the critical value is the rejection region. We now add the bottom graph which represents the distribution of the sample mean under the Alternative Hypothesis for the specific value μa. We can now measure the Power of the test (the area in green) and beta (the area in purple) on the lower graph. There are several methods of increasing Power, but they all have trade-offs:

Ways to increase power	Trade off
Increase sample size	Increased cost or unavailability of data
Increase significance level (α)	More likely to Reject a True Ho (Type I error)
Choose a value of μ_a further from μ_0	Result may be less meaningful
Redefine population to lower standard deviation	Result may be too limited to have value
Do as a one-tail rather than a two-tail test	May produce a biased result



Test of population proportion vs. hypothesized value:

When our data is categorical and there are only two possible choices (for example a yes/no question on a poll), we may want to make a claim about a proportion or a percentage of the population (p) being compared to a particular value (po). We will then use the sample proportion ($\hat{p}$)to test the claim.

Test of proportion vs. hypothesized value

$\boldsymbol{p}$ = population proportion $\quad$ $\boldsymbol{p_o}$ = population proportion under Ho

$\hat{\boldsymbol{p}}$ = sample proportion $\quad$ $\boldsymbol{p_a}$ = population proportion under Ha

Test Statistic: $Z = \dfrac{\hat{p} - p_o}{\sqrt{\dfrac{p_o(1-p_o)}{n}}}$ $\quad$ **Requirement for Normality Assumption:** $np(1-p) > 5$

One Tail Tests:

In the quality-control problem, we compared $\bar{X}$against a pair of upper and lower control limits. This is an example of a two-tail test. Below, we will discuss an example of a one-tail test.

The manager of a department store is interested in the cost effectiveness of establishing a new billing system for the store's credit customers. After a careful analysis, she determines that the new system is justified only if the mean monthly account size is more than $170 . The manager wishes to find out if there is sufficient statistical support for this.
The manager takes a random sample of 400 monthly accounts. The sample mean turns out to be $178. Historical data indicate that the standard deviation of monthly accounts is about $65.
Observe that what we are trying to find out is whether or not there is sufficient support for the hypothesis that the mean monthly accounts are "more than $170." The standard procedure is then to let μ > 170 be the alternative hypothesis. For this reason, the alternative hypothesis is also often referred to as the research hypothesis.

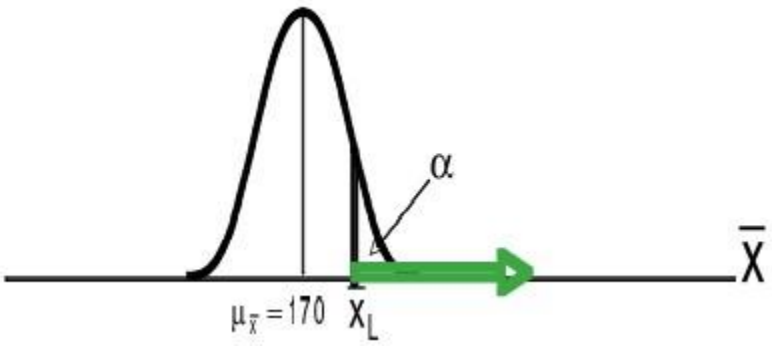
It follows that the null hypothesis should be defined as μ = 170. Note that we do not use μ ≤ 170 as the null hypothesis; this is because the null hypothesis must be precise enough for us to determine a "unique" sampling distribution. The choice μ = 170 also gives H0, our favored assumption, the least probability of being rejected. Thus,

$$H_o: \mu = 170$$
$$H_1: \mu > 170$$

where H_1 is what we want to determine and H_0 specifies a single value for the parameter of interest.

Rejection-Region Approach:

This approach is similar to what we did in the quality control problem. However, we will just have one "upper" control limit, and hence the name one-tail test. Clearly, if the sample mean is "large" relative to 170, i.e., if $\bar{X} > \bar{X}_L$ for a suitably-chosen control limit $\bar{X}_L$, then we should reject the null hypothesis in favor of the alternative. Pictorially, this means that for a given α, we wish to find $\bar{X}_L$ such that:



Referring to the central limit theorem.

$$P(\bar{X} \leq 170 + z_\alpha(\sigma \backslash \sqrt{n}))$$

Here, the value of $\bar{X}_L = 170 + z_\alpha(\sigma \backslash \sqrt{n})$.

Then $P(\bar{X} > \bar{X}_L) = \alpha$ and the rejection region is the interval $(\bar{X}_L, \infty)$.

For $\alpha = 0.05$, we have,

$$\bar{X}_L = 170 + 1.645(65 \backslash \sqrt{400}) = 175.34$$

Since the observed sample mean $\bar{X}$ = 178 is greater than 175.34, we reject the null hypothesis in favor of the research hypothesis (which is what we are investigating). In other words, statistical evidence suggests that the installation of the new billing system will be cost effective.

The critical-region approach can also be implemented via the standardized variable Z, as follows.
Observe that (1) is equivalent to:

$$P\left(Z = ((\bar{X} - 170) \backslash (\sigma \backslash \sqrt{n})) \leq z_\alpha\right) = 1 - \alpha$$

Hence, $P(Z > z_\alpha) = \alpha$ and, in terms of Z, the rejection region is the interval (z_α, ∞).
For $\alpha = 0.05$, we have

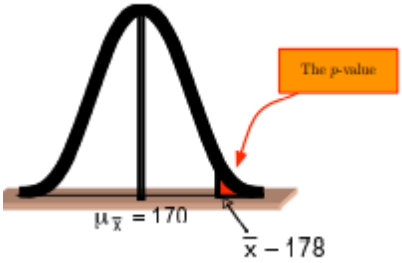
$$Z = ((178 - 170)(65/\sqrt{400})) = 2.46$$

Since 2.46 is greater than $z_\alpha = 1.645$, we reject H_0 in favor of H_1. This conclusion is of course the same as the previous one.

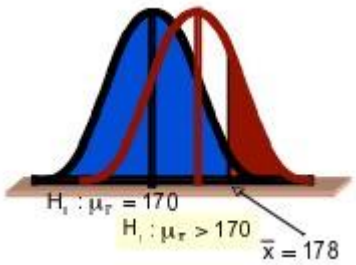
P-value approach:

The concept of p-values introduced earlier can also be used to conduct hypothesis tests. In this context, the p-value is defined as the probability of observing any test statistic that is at least as extreme as the one computed from a sample, given that the null hypothesis is true.
In the monthly account size example, the p-value associated with the given sample mean 178 is:

$$\begin{aligned} P(\bar{X} \geq 178 \mid \mu = 170) &= P(Z \geq (178 - 170) \backslash (65\sqrt{400})) \\ &= P(Z \geq 2.4615) \\ &= 1 - P(Z \leq 2.4615) \\ &= 0.0069 \end{aligned}$$



It is illuminating to observe that under H1 with a μ > 170 but still with the same sample mean 178 (the brown curve below), we would have a higher p-value:



The p-value therefore provides explicit information about the amount of statistical evidence that supports the alternative hypothesis. More explicitly, the smaller the p-value, the stronger the statistical evidence against the null hypothesis is.
For our problem, since the p-value 0.0069 is (substantially) below the specified α = 0.05, the null hypothesis should be rejected in favor of the alternative hypothesis.

This conclusion again is the same as before. In general, we have the following guidelines:

- If the p-value is less than 1%, there is overwhelming evidence that supports the alternative hypothesis.
- If the p-value is between 1% and 5%, there is a strong evidence that supports the alternative hypothesis.
- If the p-value is between 5% and 10% there is a weak evidence that supports the alternative hypothesis.
- If the p-value exceeds 10%, there is no evidence that supports the alternative hypothesis.

Evaluating the Q value:

It is important to have a good understanding of the relationship between Type I and Type II errors; that is, how the probability of a Type II error is calculated and its interpretation.
Consider again the account size problem. Recall that a Type II error occurs when a false null hypothesis is not rejected. In that example, we would not reject the null hypothesis if the sample mean $\bar{X}$ is less than or equal to the critical value $\bar{X}$ L = 175.34. It follows that
β = P ($\bar{X} \leq 175.34$ | H0 is false)

Observe that to compute β, one has to work with a specific value of μ that is greater than what H0 states, i.e., 170. For the sake of discussion, let us pick 180. For this choice, we have
β = P ($\bar{X} \leq 175.34$ | μ = 180).

The central limit theorem then tells us that

$$\begin{aligned} \beta &= P(Z \leq ((175.34 - 180)\backslash(65\backslash\sqrt{400})) \\ &= P(Z \leq -1.43) \\ &= 0.0764 \end{aligned}$$

The conditional probability 1−β is called the power of a test; it is the probability of taking the correct action of rejecting the null hypothesis when it is false. By increasing n, we can improve the power of a test. For the same α and the same n, the power of test is also used to choose between different tests; a "more powerful" test is one that yields the correct action with greater frequency.

Refer to this table only for the p-value.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Refer to the below link for a better understanding.
Hypothesis Testing (utah.edu)

Chi-Square Test:

To determine whether the association between two qualitative variables is statistically significant, researchers must conduct a test of significance called the Chi-Square Test. There are five steps to conduct this test.

Steps Need to follow to do CHI-SQUARE TEST:

Step1:

Null Hypothesis:

H0: There is no significant association between student's educational level and their preference for online or face-to-face instruction.

Or

H0: There is no difference in the distribution of instructional preferences between undergraduate and graduate students.

If there is no association between the two variables, the individuals would be uniformly distributed across the cells of the table.

The alternative hypothesis for a chi-square test is always two-sided. (It is technically multi-sided because the differences may occur in both directions in each cell of the table).

Alternative Hypothesis:

Ha: There is a significant association between students' educational level and their preference for online or face-to-face instruction.

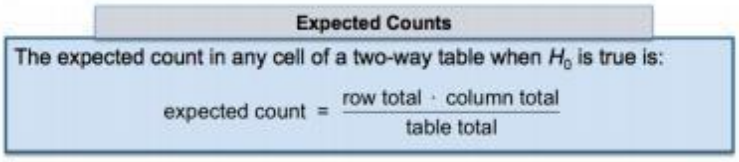
or

Ha: There is a significant difference in the distribution of instructional preferences between undergraduate and graduate students.

Steps 2:

The expected values specify what the values of each cell of the table would be if there was no association between the two variables.

The formula for computing the expected values requires the sample size, the row totals, and the column totals.



Step 3:

To see if the data give convincing evidence against the null hypothesis, compare the observed counts from the sample with the expected counts, assuming H0 is true.

The observed values are the actual counts computed from the sample.

Statistical software will compute both the expected and observed counts for each cell when conducting a chi-square test.

Step 4:

The chi-square statistic compares the observed values to the expected values.

This test statistic is used to determine whether the difference between the observed and expected values is statistically significant.

The **chi-square statistic** is a measure of how far the observed counts are from the expected counts. The formula for the statistic is:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where the sum is over all possible values of the categorical variable.

Uses of Chi-Square Test:

One of the most useful properties of the chi-square test is that it tests the null hypothesis "the row and column variables are not related to each other" whenever this hypothesis makes sense for a two-way variable.

Annova Test (Analysis of Variance):

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance.
ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. The chapter "All Pairwise Comparisons among Means" showed how to test differences among means.

Analysis of Variance Designs:

Factors and Levels:

The section on variables defined an independent variable as a variable manipulated by the experimenter. In the case study "Smiles and Leniency," the effect of different types of smiles on the leniency showed to a person was investigated. Four different types of smiles (neutral, false, felt, miserable, on leniency) were shown. In this experiment, "Type of Smile" is the independent variable. In describing an ANOVA design, the term factor is a synonym of independent variable. Therefore, "Type of Smile" is the factor in this experiment. Since four types of smiles were compared, the factor "Type of Smile" has four levels.

An ANOVA conducted on a design in which there is only one factor is called a one-way ANOVA. If an experiment has two factors, then the ANOVA is called a two-way ANOVA. For example, suppose an experiment on the effects of age and gender on reading speed were conducted using three age groups (8 years, 10 years, and 12 years) and the two genders (male and female). The factors would be age and gender. Age would have three levels and gender would have two levels.

Between- and Within-Subjects Factors:

In the "Smiles and Leniency" study, the four levels of the factor "Type of Smile" were represented by four separate groups of subjects. When different subjects are used for the levels of a factor, the factor is called a between-subjects factor or a between-subjects variable. The term "between subjects" reflects the fact that comparisons are between different groups of subjects.
In the "ADHD Treatment" study, every subject was tested with each of four dosage levels (0, 0.15, 0.30, 0.60 mg/kg) of a drug. Therefore, there was only one group of subjects, and comparisons were not between different groups of subjects but between conditions within the same subjects. When the same subjects are used for the levels of a factor, the factor is called a within-subjects factor or a within subject's variable. Within-subjects variables are sometimes referred to as repeated measures variables since there are repeated measurements of the same subjects.

Multi Factors Design:

It is common for designs to have more than one factor. For example, consider a hypothetical study of the effects of age and gender on reading speed in which males and females from the age levels of 8 years, 10 years, and 12 years are tested. There would be a total of six different groups as shown.

Group	Gender	Age
1	Female	8
2	Female	10
3	Female	12
4	Male	8
5	Male	10
6	Male	12

This design has two factors: age and gender. Age has three levels and gender has two levels. When all combinations of the levels are included (as they are here), the design is called a factorial design. A concise way of describing this design is as a Gender (2) x Age (3) factorial design where the numbers in parentheses indicate the number of levels. Complex designs frequently have more than two factors and may have combinations of between- and within-subjects' factors.

This section shows how ANOVA can be used to analyze a one-factor between subjects design. We will use as our main example the "Smiles and Leniency" case study. In this study there were four conditions with 34 subjects in each condition. There was one score per subject. The null hypothesis tested by ANOVA is that the population means for all conditions are the same. This can be expressed as follows: H0: $\mu 1 = \mu 2 = ... = \mu k$

where H0 is the null hypothesis and k is the number of conditions. In the smiles and leniency study, k = 4 and the null hypothesis is
H0: $\mu false = \mu felt = \mu miserable = \mu neutral$

If the null hypothesis is rejected, then it can be concluded that at least one of the population means is different from at least one other population mean.

Analysis of variance is a method for testing differences among means by analyzing variance. The test is based on two estimates of the population variance (σ^2). One estimate is called the mean square error (MSE) and is based on differences among scores within the groups. MSE estimates σ^2 regardless of whether the null hypothesis is true (the population means are equal). The second estimate is called the mean square between (MSB) and is based on differences among the sample means. MSB only estimates σ^2 if the population means are equal. If the population means are not equal, then MSB estimates a quantity larger than σ^2. Therefore, if the MSB is much larger than the MSE, then the population means are unlikely to be equal. On the other hand, if the MSB is about the same as MSE, then the data are consistent with the hypothesis that the population means are equal.

Before proceeding with the calculation of MSE and MSB, it is important to consider the assumptions made by ANOVA:

1. The populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the values are not independent. The analysis of data with two scores per subject is shown in the section on within-subjects ANOVA later in this chapter.

These assumptions are the same as for a t test of differences between groups except that they apply to two or more groups, not just to two groups. The means and variances of the four groups in the "Smiles and Leniency" case study is shown in Table 1. Note that there are 34 subjects in each of the four conditions (False, Felt, Miserable, and Neutral).

Condition	Mean	Variance
FALSE	5.3676	3.3380
Felt	4.9118	2.8253
Miserable	4.9118	2.1132
Neutral	4.1176	2.3191

Sample Sizes:

The first calculations in this section all assume that there is an equal number of observations in each group. Unequal sample size calculations are shown in the section on sources of variation. We will refer to the number of observations in each group as n and the total number of observations as N. For these data there are four groups of 34 observations. Therefore n = 34 and N = 136.

Computing MSB:

The formula for MSB is based on the fact that the variance of the sampling distribution of the mean is,

$$\sigma_m^2 = \sigma^2 \backslash n$$

where n is the sample size of each group. Rearranging this formula, we have,

$$\sigma^2 = n\sigma_m^2$$

Therefore, if we knew the variance of the sampling distribution of the mean, we could compute σ^2 by multiplying it by n. Although we do not know the variance of the sampling distribution of the mean, we can estimate it with the variance of the sample means. For the leniency data, the variance of the four-sample means is 0.270. To estimate σ2, we multiply the variance of the sample means (0.270) by n (the number of observations in each group, which is 34). We find that MSB = 9.179.

To sum up these steps:

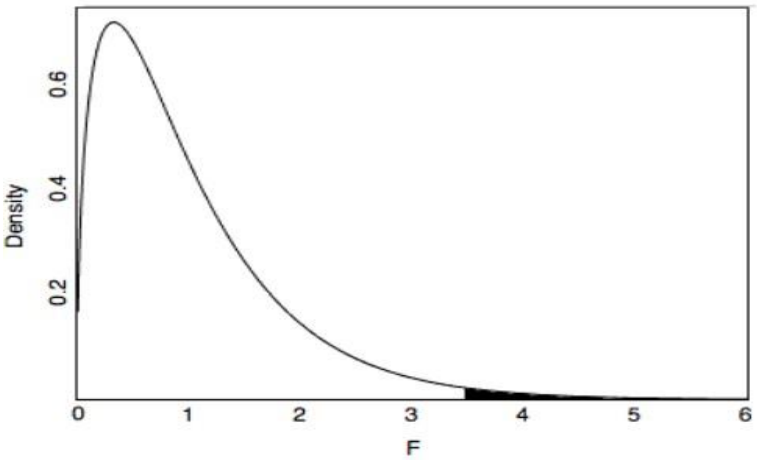
1. Compute the means.
2. Compute the variance of the means.
3. Multiply the variance of the means by n.

Comparing MSE and MSB:

The critical step in an ANOVA is comparing MSE and MSB. Since MSB estimates a larger quantity than MSE only when the population means are not equal, a finding of a larger MSB than an MSE is a sign that the population means are not equal. But since MSB could be larger than MSE by chance even if the population means are equal, MSB must be much larger than MSE in order to justify the conclusion that the population means differ. But how much larger must MSB be? For the "Smiles and Leniency" data, the MSB and MSE are 9.179 and 2.649, respectively. Is that difference big enough? To answer, we would need to know the probability of getting that big a difference or a bigger difference if the population means were all equal. The mathematics necessary to answer this question were worked out by the statistician R. Fisher. Although Fisher's original formulation took a slightly different form, the standard method for determining the probability is based on the ratio of MSB to MSE. This ratio is named after Fisher and is called the F ratio.

For these data, the F ratio is,
F=9.179/2.649=3.465

Therefore, the MSB is 3.465 times higher than MSE. Would this have been likely to happen if all the population means were equal? That depends on the sample size. With a small sample size, it would not be too surprising because results from small samples are unstable. However, with a very large sample, the MSB and MSE are almost always about the same, and an F ratio of 3.465 or larger would be very unusual. Figure shows the sampling distribution of F for the sample size in the "Smiles and Leniency" study. As you can see, it has a positive skew.



The above figure shows distribution of F.

Selection of One-Tailed or Two-tailed:

Is the probability value from an F ratio a one-tailed or a two-tailed probability? In the literal sense, it is a one-tailed probability since, as you can see in Figure 1, the probability is the area in the right-hand tail of the distribution. However, the F ratio is sensitive to any pattern of differences among means. It is, therefore, a test of a two-tailed hypothesis and is best considered a two-tailed test.

Relationship to the t test:

Since an ANOVA and an independent-groups t test can both test the difference between two means, you might be wondering which one to use. Fortunately, it does not matter since the results will always be the same. When there are only two groups, the following relationship between F and t will always hold:

$$F(1, dfd) = t^2(df)$$

where dfd is the degrees of freedom for the denominator of the F test and df is the degrees of freedom for the t test. dfd will always equal df.

Transformations of data:

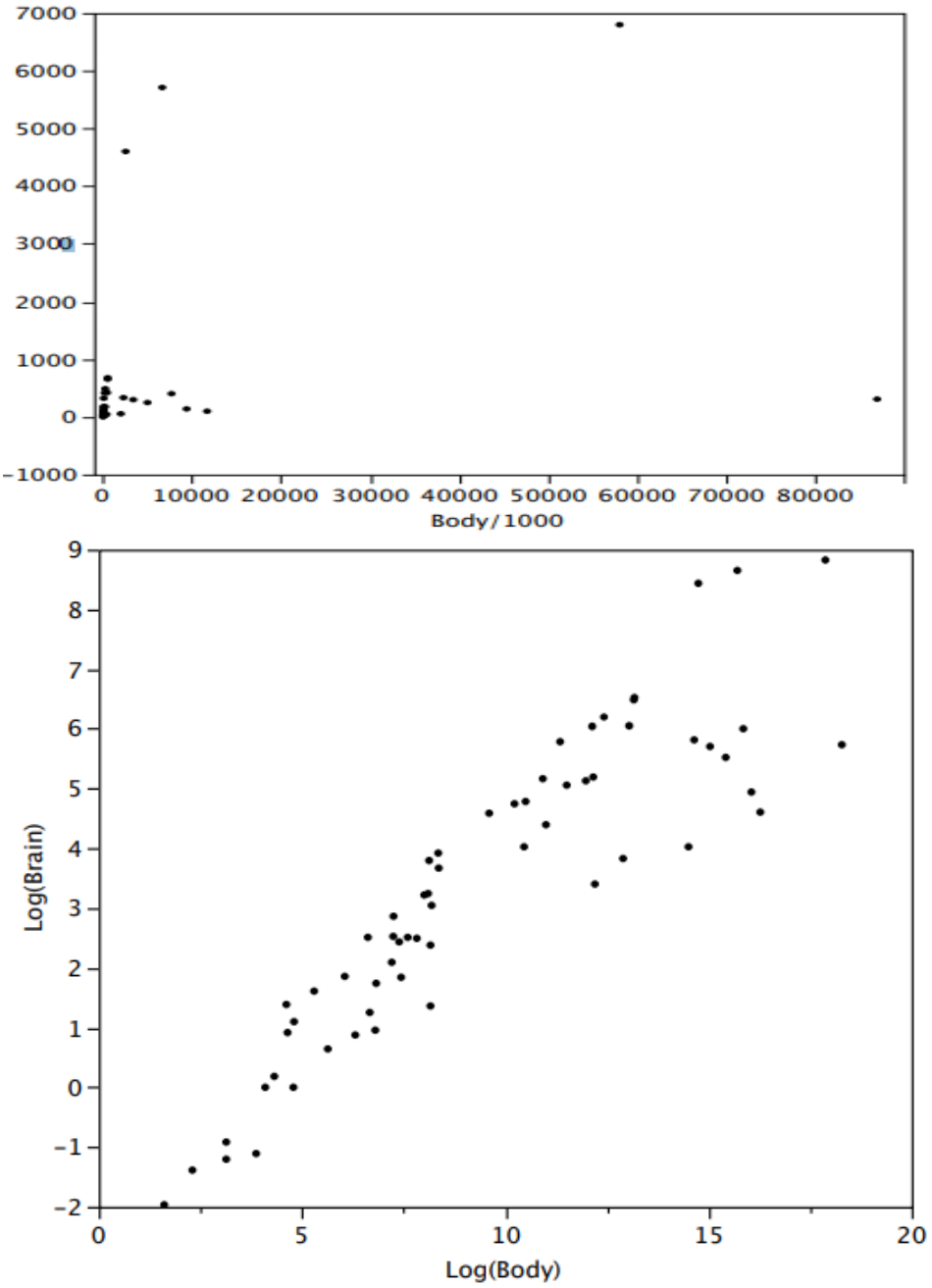
- **Log**
- **Tukey's Ladder of Powers**
- **Box-Cox Transformations**

The focus of statistics courses is the exposition of appropriate methodology to analyze data to answer the question at hand. Sometimes the data are given to you, while other times the data are collected as part of a carefully-designed experiment. Often the time devoted to statistical analysis is less than 10% of the time devoted to data collection and preparation. If aspects of the data preparation fail, then the success of the analysis is in jeopardy. Sometimes errors are introduced into the recording of data. Sometimes biases are inadvertently introduced in the selection of subjects or the mis-calibration of monitoring equipment. In this chapter, we focus on the fact that many statistical procedures work best if individual variables have certain properties. The measurement scale of a variable should be part of the data preparation effort. For example, the correlation coefficient does not require the variables have a normal shape, but often relationships can be made clearer by re-expressing the variables. An economist may choose to analyze the logarithm of prices if the relative price is of interest. A chemist may choose to perform a statistical analysis using the inverse temperature as a variable rather than the temperature itself. But note that the inverse of a temperature will differ depending on whether it is measured in °F, °C, or °K.

The introductory chapter covered linear transformations. These transformations normally do not change statistics such as Pearson's r, although they do affect the mean and standard deviation. The first section here is on log transformations which are useful to reduce skew. The second section is on Tukey's ladder of powers. You will see that log transformations are a special case of the ladder of powers. Finally, we cover the relatively advanced topic of the Box-Cox transformation.

Log Transformation:

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. Figure shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.



The above two figures show the scatter plots of brain weight as a function of body weight in terms of both raw data and log transformed data.
It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs

because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

X	$Log_{10}(X)$
1	0
10	1
100	2

Tukey's Ladder of Powers:(Not Required)

Box-Cox Transformations:

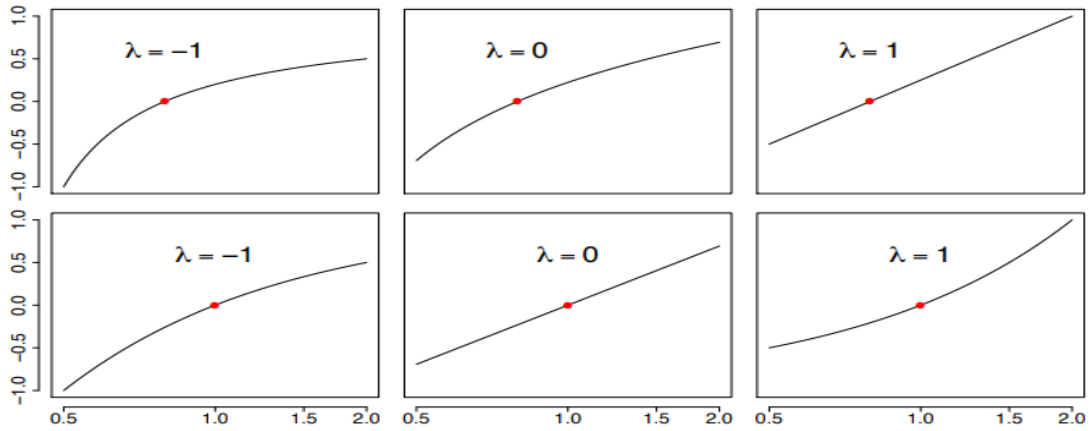
George Box and Sir David Cox collaborated on one paper (Box, 1964). The story is that while Cox was visiting Box at Wisconsin, they decided they should write a paper together because of the similarity of their names (and that both are British). In fact, Professor Box is married to the daughter of Sir Ronald Fisher. The Box-Cox transformation of the variable x is also indexed by λ, and is defined as,

$$x'_\lambda = (x^\lambda - 1)\backslash\lambda$$

At first glance, although the formula is a scaled version of the Tukey transformation x^λ, this transformation does not appear to be the same as the Tukey transformation. However, a closer look shows that when λ < 0, both x_λ and x'_λ change the sign of x_λ to preserve the ordering. Of more interest is the fact that when λ = 0, then the Box-Cox variable is the indeterminate form 0/0. Rewriting the Box-Cox formula as,

$$x'_\lambda = \frac{e^{\lambda \log(x)} - 1}{\lambda} \approx \frac{\left(1 + \lambda \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \cdots\right) - 1}{\lambda} \to \log(x)$$

Here, λ tends to zero.



Regression Analysis

We will discuss the approach for the regression analysis. We will investigate the relation of the features with the target. The tools used to explore this relationship, is the regression and correlation analysis. These tools can be used to find out if the outcome from one variable depends on the value of the other variable, which would mean a dependency from one variable on the other. Regression and correlation analysis can be used to describe the nature and strength of the relationship between two continuous variables.

Two ways to investigate the relationship between the features and the target:

- Scatterplot
- Correlation

Scatterplot:

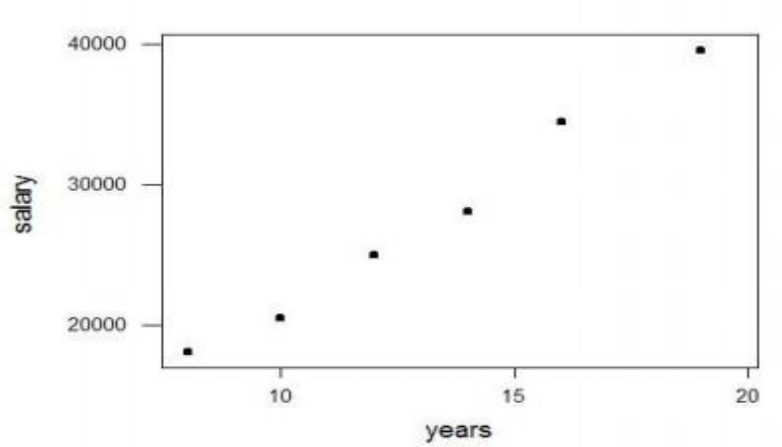
The first step in the investigation of the relationship between two continuous variables is a scatterplot! Create a scatterplot for the two variables and evaluate the quality of the relationship.

Consider an example:

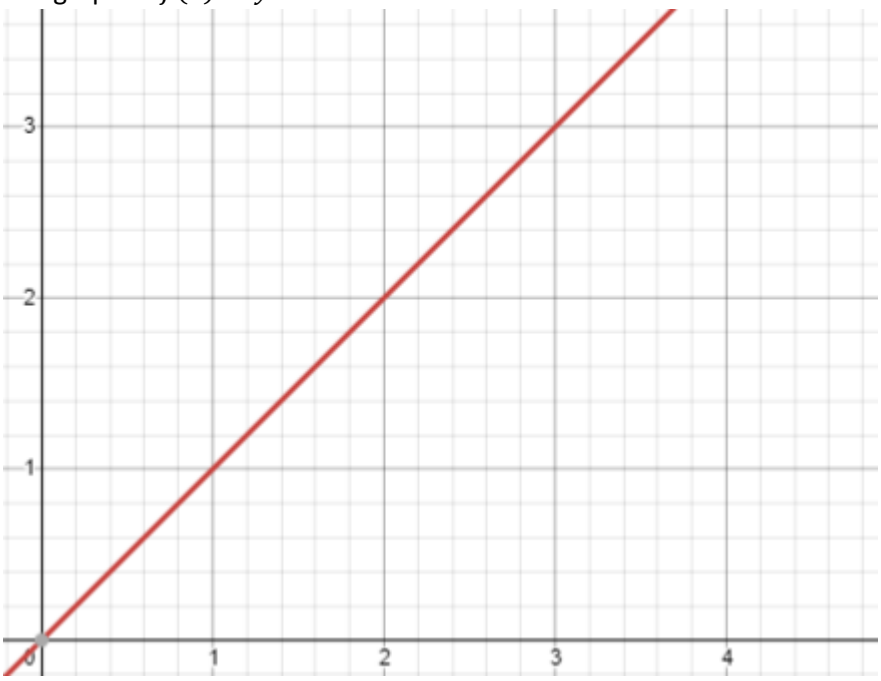
Does the number of years invested in schooling pay off in the job market? Apparently so – the better educated you are, the more money you will earn. The data in the following table give the median annual income of full-time workers age 25 or older by the number of years of schooling completed.

x=Years of Schooling	y=Salary (dollars)
8	18,000
10	20,500
12	25,000
14	28,100
16	34,500
19	39,700

Start off with creating a scatterplot for X and Y.



The scatterplot shows a strong, positive, linear association between years and salary. You can say this by matching the graph of $f(x) = y$.
The graph of $f(x) = y$ is shown below.



Correlation:

If the scatterplot shows a reasonable linear relationship (straight line) calculate Pearson's correlation coefficient to evaluate the strength of the linear relationship.

Notation:
Here, $(x_1, y_1), (x_2, y_2), \ldots\ldots\ldots, (x_n, y_n)$ denote a sample of (x, y) pairs.

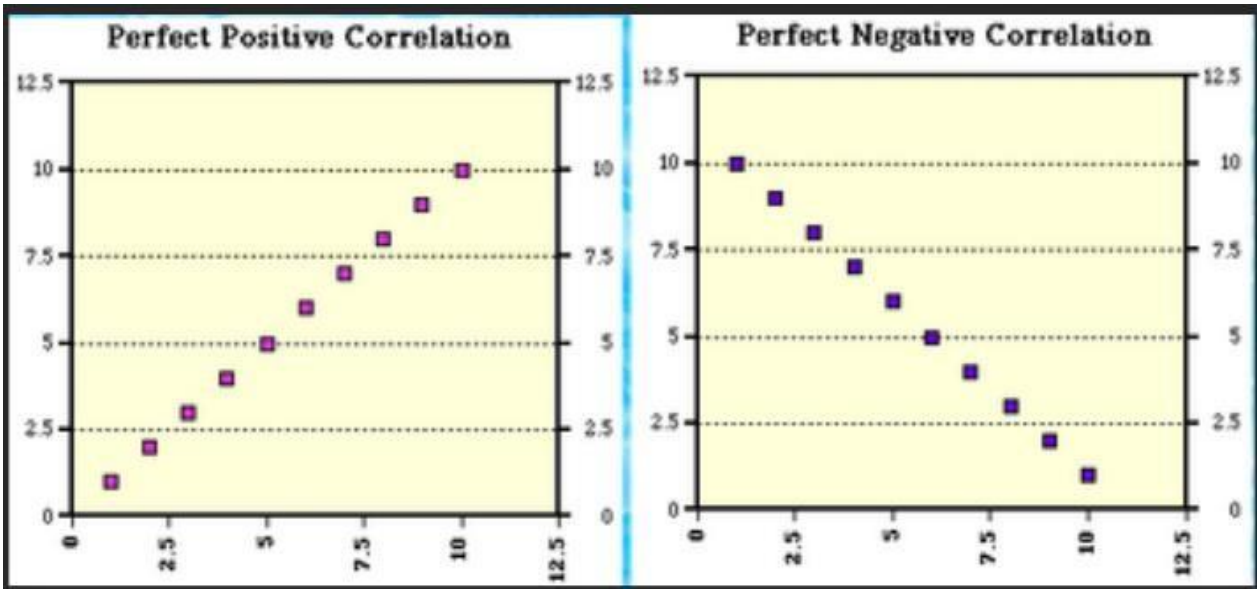
The formula of the correlation can be given by,

$$corr = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

The above formula is given by Pearson's.
In the above formula,
s.
In the above formula,

$$S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$$
$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$
$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Pearson's correlation coefficient (named after Karl Pearson, 1857-1936) is a number between -1 and 1, that measures the strength of a linear relationship between two continuous variables. The absolute value of the coefficient measures how closely the variables are related. The closer it is to 1 the closer the relationship. A correlation coefficient over 0.8 indicates a strong correlation between the variables.



The sign of the correlation coefficient tells you of the trend in the relationship. A positive (negative) coefficient means that one variable increases (decreases), when the other increases.

Linear Regression:

If we want to use a variable x to draw conclusions concerning a variable y: y is called dependent or response variable. x is called independent, predictor, or explanatory variable. If the relationship between two variables is linear is can be summarized by a straight line. A straight line can be described by an equation:

$$y = a + bx$$

a is called the intercept and b the slope of the equation. The slope is the amount by which y increases when x increases by 1 unit.

How fit method works:

Given data points (x_i, y_i), a and b shall now be chosen in that way that the corresponding linear line will have the "best fit" for the given data. The criteria for "best fit" used in regression analysis is the sum of the squared differences between the data points and the line itself, that is the y deviations. For data points (x_i, y_i), $1 \leq i \leq n$ this can be written as,

$$\min_{a,b} \sum^{n} (y_i - (a + bx_i))^2$$

In words: minimize the sum by choosing the appropriate parameters a and b. The resulting line is called the least square line or sample regression line. After the problem is stated it can be solved mathematically and the results are formulas, how to calculate the best parameters.

$b = \frac{S_{xy}}{S_{xx}}$ and $a = \bar{y} - b\,\bar{x}$

Write the equation of the least squares line as,

$$\hat{y} = a + bx$$

Here, $\hat{y}$ gives an estimate for y for a given value of x.

Properties of the regression or least squares line:

The least squares line passes always through the balance point $(\bar{x}, \bar{y})$ of the data set.

The regression line of y on x should not be used to predict x, since it is not the line that minimizes the sum of squared x deviations.

Assessing the fit of a line:

Once the least squares line has been obtained, it is natural to examine how effectively the line summarizes the relationship between x and y.

The first question that has to be answered is, if the line is an appropriate way to summarize the relationship. In order to answer this question, we will calculate the coefficient of determination r^2.

Pearson's Correlation Coefficient:

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

It gives the proportion of variation in y that can be attributed to a linear relationship between x and y.

Is r^2 greater than 0.8, the model has a good fit and can be used to calculate reliable predictions of the dependent variable by using the independent variable. In the example, the variable Years of Schooling explains $r^2 = 98.8\ \%$ of the variation in the variable Salary. Which is very high. The plot showed that the data points are almost on a straight line. Use the least squares line for predicting the annual salary of a person with 13 years of schooling.

This is just an estimate, from the other parts of the class, we know that a confidence interval can be found that gives more information.