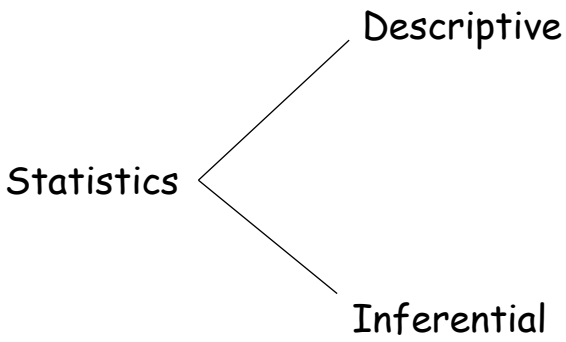
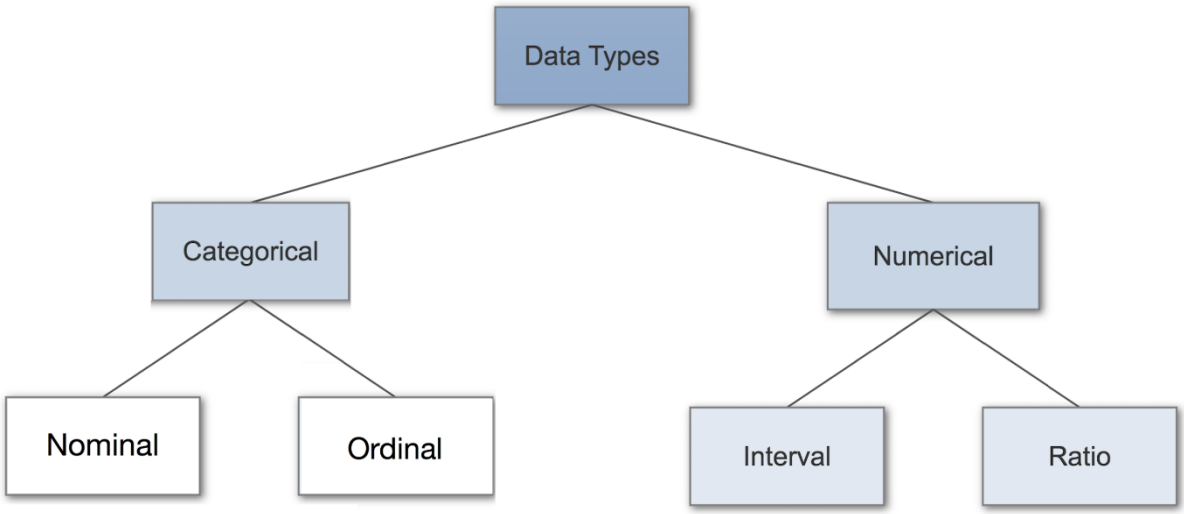
STATISTICS

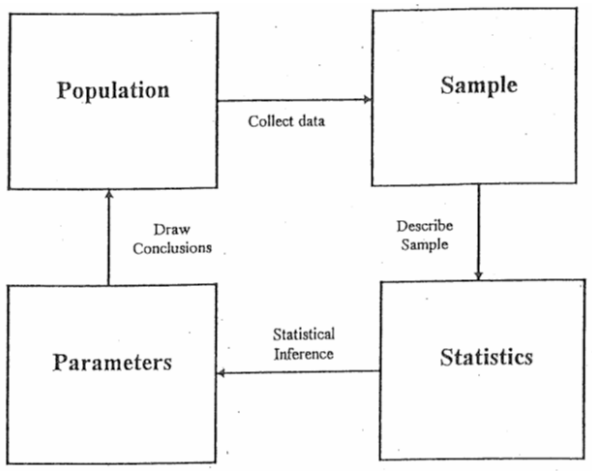




Statistics is the science of collecting, organizing, summarizing, analysing, and interpreting information.

Good statistics come from good samples, and are used to draw conclusions or answer questions about a population. We use sample statistics to estimate population parameters (the truth). So let's begin there…..

A statistic is a characteristic of a sample. If you collect a sample and calculate the mean and standard deviation, these are sample statistics. Inferential statistics allow you to use sample statistics to make conclusions about a population. However, to draw valid conclusions, you must use particular sampling techniques. These techniques help ensure that samples produce unbiased estimates. Biased estimates are systematically too high or too low. You want unbiased estimates because they are correct on average.



Types of Statistics

Descriptive

Statistics used to describe things, frequently groups of people.

- Central Tendency
- Variability
- Relative Standing
- Relationship

Inferential

Statistics used to make inferences and draw conclusions.

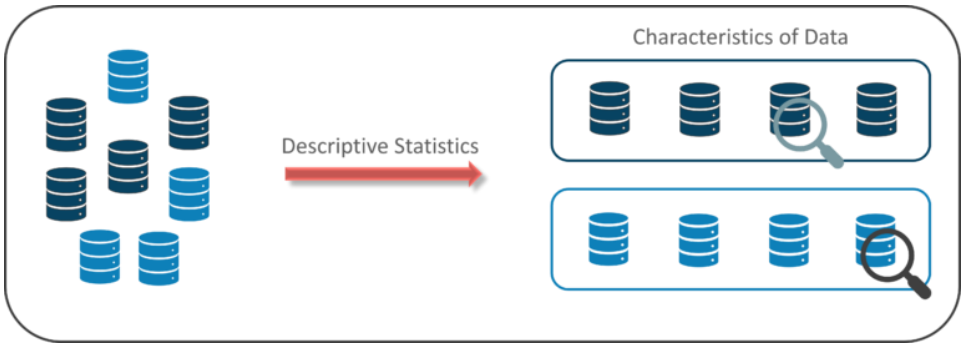
- Parametric (t-test, ANOVA, multiple regression)
- Non-Parametric (chi-square)

Descriptive Statistics

- It uses data to provide description of the population, either through numerical calculations or graphs or tables.
- It helps organize data and focuses on characteristics of data providing parameters.

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analysed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

They are very important because if we simply presented our raw data it would be hard to visualize what the data was showing, especially if there was a lot of it. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.



Types of Variable:

A. Discrete
- Nominal
- Ordinal

B. Continuous
- Countable
- Measurable

Discrete Data can only take certain values.

- ✓ A nominal variable is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence.
- ✓ An ordinal variable is a categorical variable. Observations can take a value that can be logically ordered or ranked.

Continuous Data can take any value (within a range).

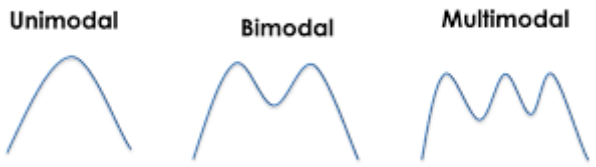
- ✓ A countable variable is a numeric variable. Observations can take a value based on a count from a set of distinct whole values.
- ✓ A measurable variable is a numeric variable. Observations can take any value between a certain set of real numbers

Characteristics of Frequency Distribution:

1. Modality

- Unimodal
- Bimodal

The modality of a distribution is determined by the number of peaks it contains. ... A bimodal distribution has two values that occur frequently (two peaks) and a multimodal has two or several frequently occurring values.

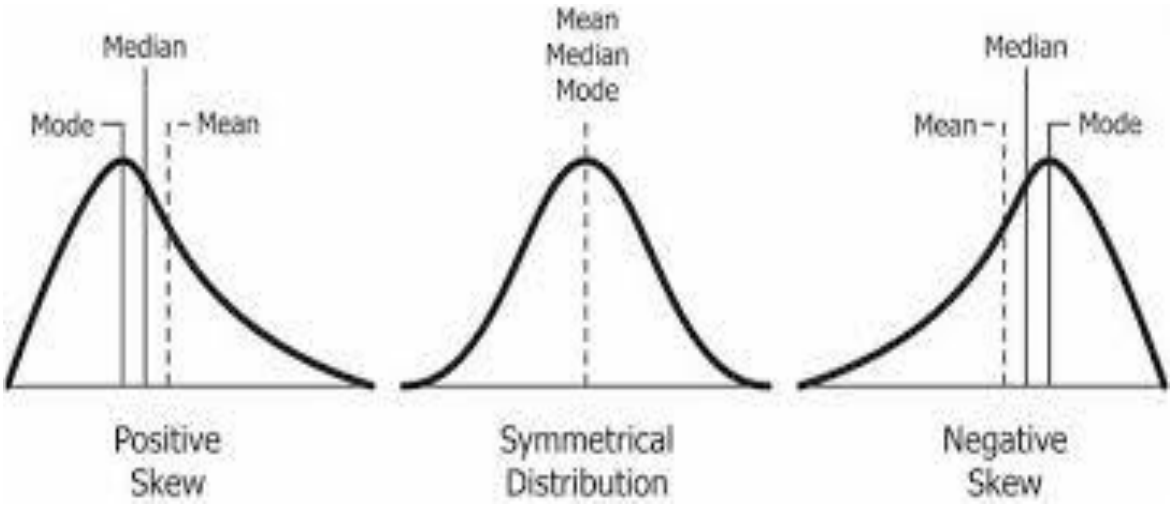


2. Symmetry

- Symmetric
- Asymmetric

a. Positive Skewness
b. Negative Skewness

Symmetry is an attribute used to describe the shape of a data distribution. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other. A non-symmetric distribution cannot.



3. Central Tendency

- Mean
- Median
- Mode

Mean - The arithmetic mean of a variable, often called the average, is computed by adding up all the values and dividing by the total number of values. However, the mean is influenced by extreme values (outliers) and may not be the best measure of centre with strongly skewed data.

Median - The median of a variable is the middle value of the data set when the data are sorted in order from least to greatest. It splits the data into two equal halves with 50% of the data below the median and 50% above the median. The median is resistant to the influence of outliers, and may be a better measure of center with strongly skewed data.

Mode - The mode is the most frequently occurring value and is commonly used with qualitative data as the values are categorical. Categorical data cannot be added, subtracted, multiplied or divided, so the mean and median cannot be computed. The mode is less commonly used with quantitative data as a measure of center.

4. Variability

- Standard Deviation
- Range

Standard Deviation - The Standard Deviation is a measure of how spread out numbers are. it is the square root of the Variance. The Variance is defined as:

The average of the squared differences from the Mean.

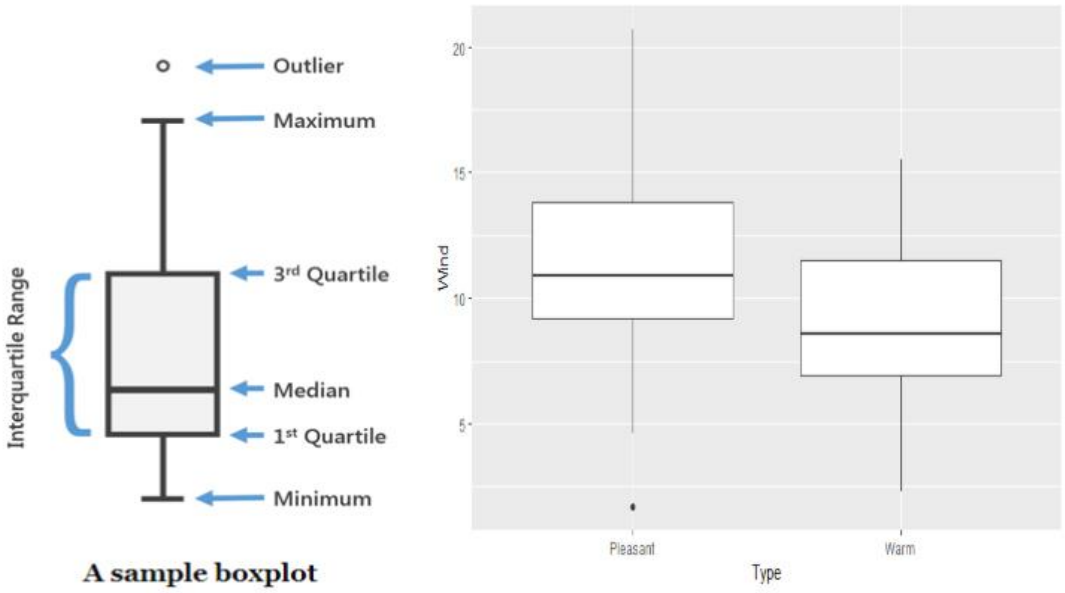
Range - The range of a variable is the largest value minus the smallest value. It is the simplest measure and uses only these two values in a quantitative data set.

Measures of Spread / Dispersion:

- Range
- Variance
- Standard Deviation
- Interquartile Range (IQR)/ 5 number summary

InterQuartile Range (5 number Summary):

The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles.
Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.
Min is the least value in the data set.
Q1 is the "middle" value in the first half of the rank-ordered data set.
Q2 is the median value in the set.
Q3 is the "middle" value in the second half of the rank-ordered data set.
Max is the maximun value in the data set.



Measures of Association b/w two variables:

- Covariance
- Correlation
  - Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.
  - Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called correlation analysis.

Some examples of data that have a high correlation:

- Your caloric intake and your weight.
- The amount of time your study and your GPA.

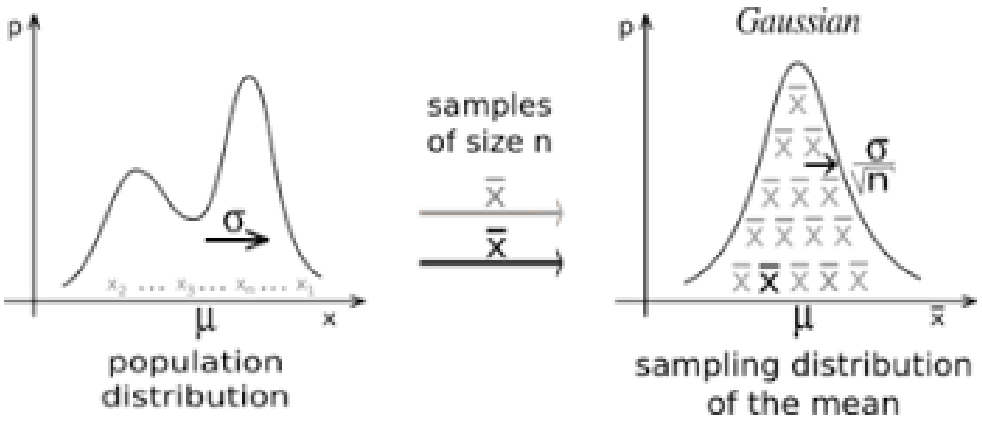
Some examples of data that have a low correlation (or none at all):

- A dog's name and the type of dog biscuit they prefer.
- The cost of a car wash and how long it takes to buy a soda inside the station.

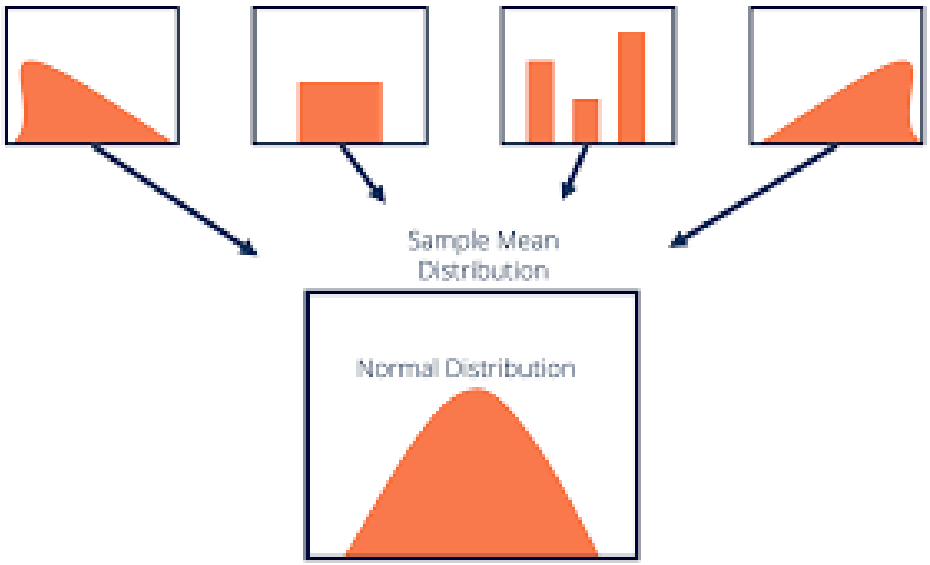
Central Limit Theorem:

The sampling distribution of sample means approaches a "Normal Distribution" as sample size gets large no matter what the shape of population distribution.

In other words, CLT states that regardless of variables distribution in the population, the sampling distribution of mean will tend to approximate Normal Distribution.



The theorem is vital for 2 main reasons- the normality assumption and precision of estimates.



Assumptions of Central Limit Theorem-

- Data must follow randomization condition, i.e- must be sampled randomly.
- Samples should be independent of each other.
- Sample size should not be more than 10% of the population.
- Sample size should be sufficiently large (N>30).

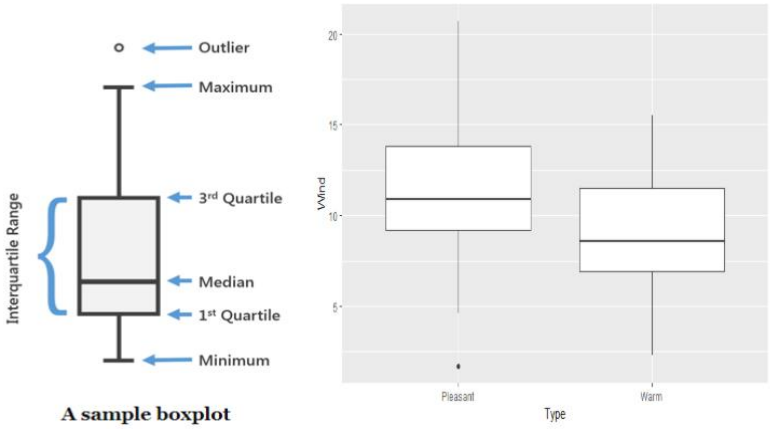
Data Visualization:

- Box Plot
- Scatter Plot
- Density Plot

Box Plot –

Boxplots use the 5-number summary (minimum and maximum values with the three quartiles) to illustrate the center, spread, and distribution of your data. When paired with histograms, they give an excellent description, both numerically and graphically, of the data.

With symmetric data, the distribution is bell-shaped and somewhat symmetric. In the boxplot, we see that Q1 and Q3 are approximately equidistant from the median, as are the minimum and maximum values. Also, both whiskers (lines extending from the boxes) are approximately equal in length.



Scatter Plot –

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.

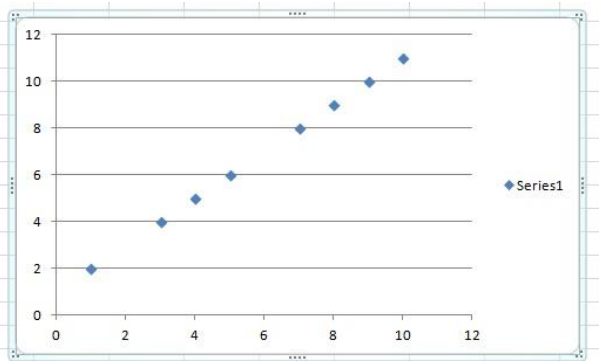
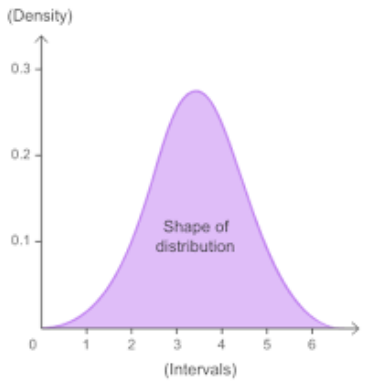


Fig - Discrete Variables on a Scatter Plot

Density Plot –

The Density Plot shows the smoothed distribution of the points along the numeric axis. The peaks of the density plot are at the locations where there is the highest concentration of points.

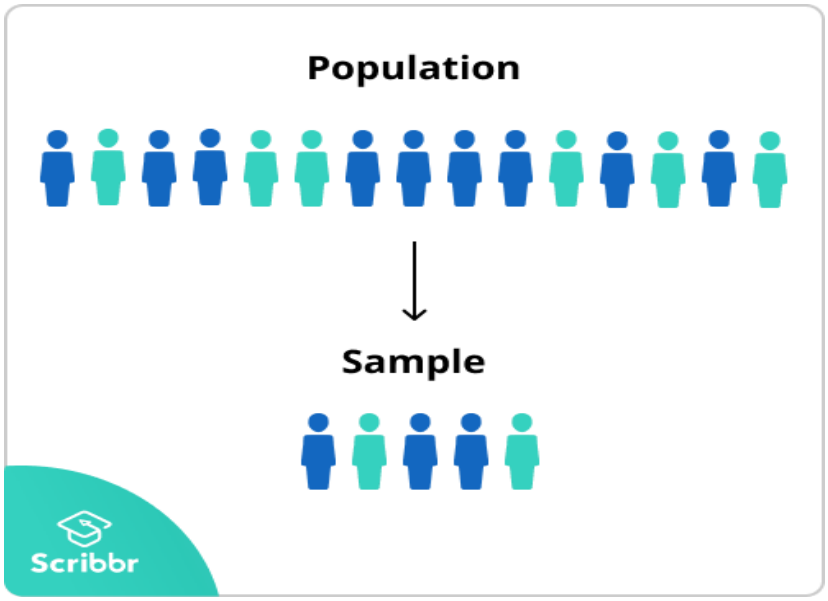


Various Sampling Methods: -

1. Probability Sampling
- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

2. Non – Probability Sampling
- Convenience Sampling
- Voluntary Response Sampling
- Purposive Sampling
- Snowball Sampling

1. Probability Sampling – involves random selection, allowing to make statistical inference about the whole group.



a. Simple random sampling-

Each individual is chosen entirely by chance and each member of population has an equal chance / probability of being selected.

E.g. - You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

b. Systematic Sampling-

It is similar to simple random sampling, but is usually slightly easier to conduct. Every member of population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

E.g. - All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.

c. Stratified Sampling-

This method is appropriate when population has mixed characteristics and you want to ensure that every characteristic is proportionally represented in sample.

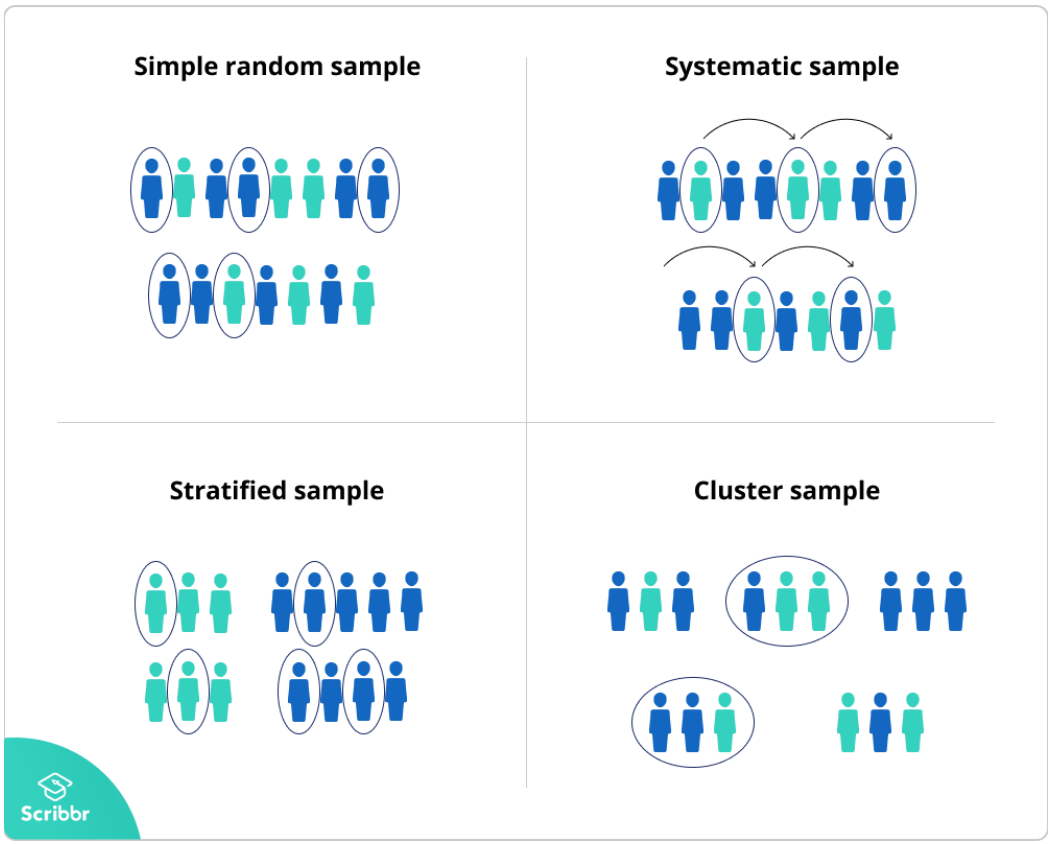
Population is divided into subgroups (Strata) based on relevant characteristics. From overall proportions of population, how many people should be sampled from each sub group is calculated. Then random / systematic sampling is done to select a sample from each subgroup.

E.g.- The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

d. Cluster Sampling-

Cluster sampling also involves dividing population into sub – groups, but each sub-group should have similar characteristics to whole sample. Instead of sampling individuals from sub groups, you randomly select entire sub groups. This method is good for dealing with large and dispersed populations, but there is more risk of error in sample as there should be substantial difference b/w clusters. It's difficult to guarantee that sample clusters are really representative of whole population.

E.g.- The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.



2.Non – Probability Sampling – Involves non – random selection based on convenience/ other criteria, allowing to easily collect initial data. Individuals are selected based on non – random criteria and not every individual has a chance of being included. This type of sampling is easier to access, but you cant use it to make a valid statistical inferences about the whole population.

a. Convenience Sampling-

Simply includes the individuals who happen to be most accessible to researcher. It is an easy and inexpensive way to gather initial data, but there is no way to tell if sample is representative of population, so it can't produce generalized results.

E.g. - You are researching opinions about student support services in your university, so after each of your classes, you ask your fellow students to complete a survey on the topic. This is a convenient way to gather data, but as you only surveyed students taking the same classes as you at the same level, the sample is not representative of all the students at your university.

b. Voluntary Response Sampling-

Similar to Convenience sampling, a voluntary response sample is mainly based on ease of access. Instead of researcher choosing participants and directly contacting the, people volunteer themselves.

E.g.- You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can't be sure that their opinions are representative of all students.

c. Purposive Sampling-

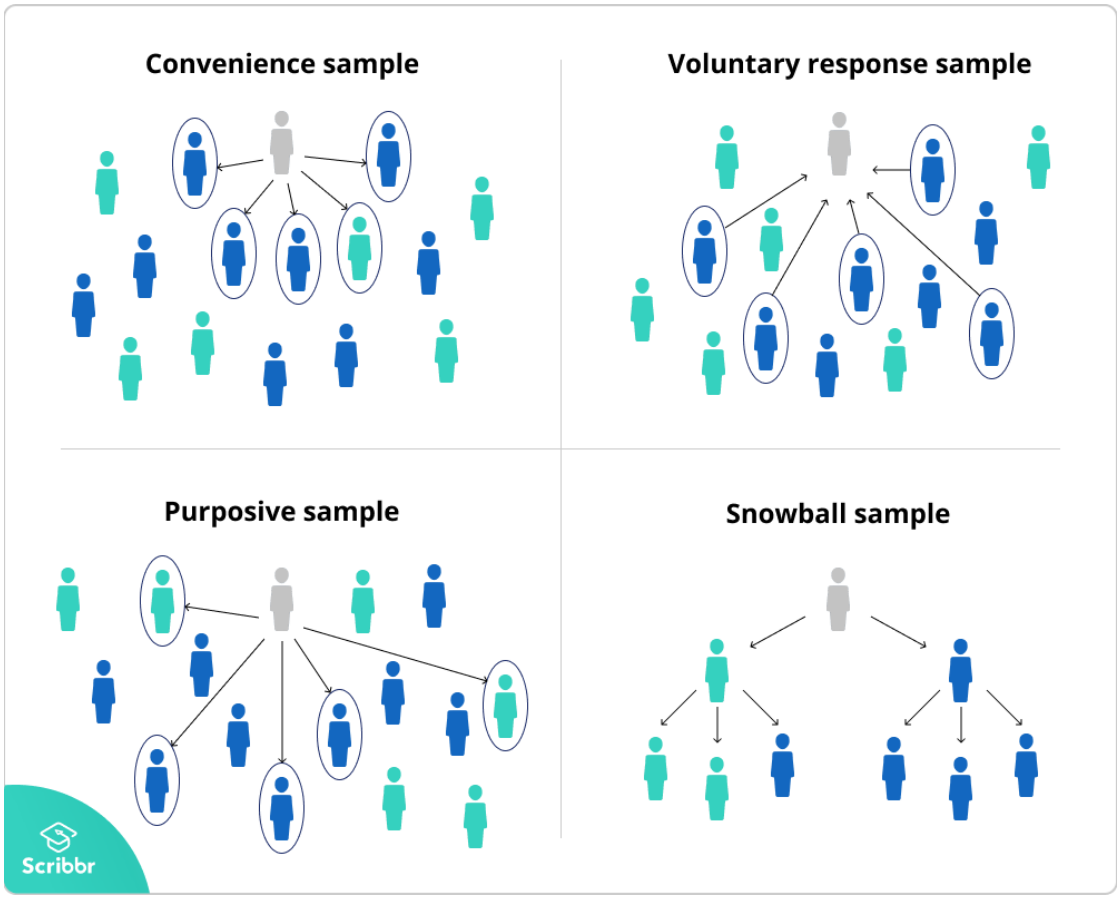
This type of sampling involves researcher using their judgement to select a sample that is most useful to purposes of research. It is often used in qualitative research, where researcher wants to gain detailed knowledge about a specific phenomenon rather than make statistical inference.

E.g.- You want to know more about the opinions and experiences of disabled students at your university, so you purposefully select a number of students with different support needs in order to gather a varied range of data on their experiences with student services.

d. Snowball Sampling-

If population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to as you get in contact with more population.

E.g. - You are researching experiences of homelessness in your city. Since there is no list of all homeless people in the city, probability sampling isn't possible. You meet one person who agrees to participate in the research, and she puts you in contact with other homeless people that she knows in the area.



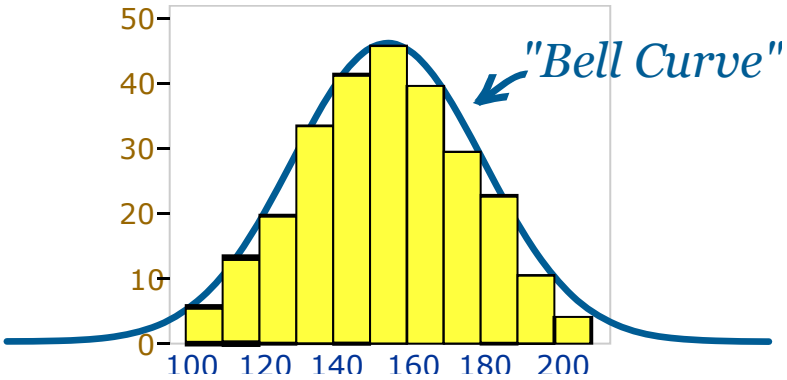
Population Parameters versus Sample Statistics:

A parameter is a value that describes a characteristic of an entire population, such as the population mean. Because you can almost never measure an entire population, you usually don't know the real value of a parameter. In fact, parameter values are nearly always unknowable. While we don't know the value, it definitely exists.

For example, the average height of adult women in the United States is a parameter that has an exact value—we just don't know what it is!

The population mean and standard deviation are two common parameters. In statistics, Greek symbols usually represent population parameters, such as μ (mu) for the mean and σ (sigma) for the standard deviation.

Normal Distribution (Bell Curve):



The following is the plot of the standard normal probability density function.

The general formula for probability density function of normal distribution is

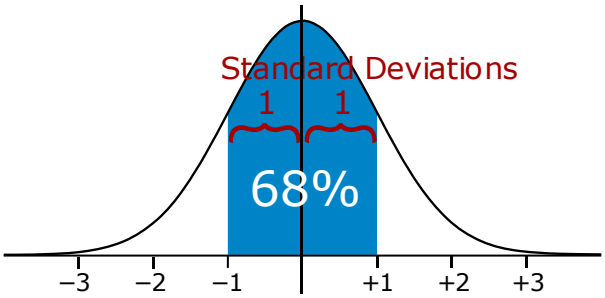
$$f(x) = \frac{e^{-(x-\mu)2/(2\sigma 2)}}{\sigma\sqrt{2\pi}}$$

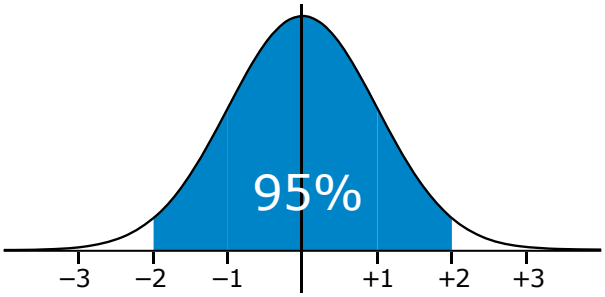
where μ is the location parameter and σ is the scale parameter. The case where μ = 0 and σ = 1 is called the standard normal distribution.

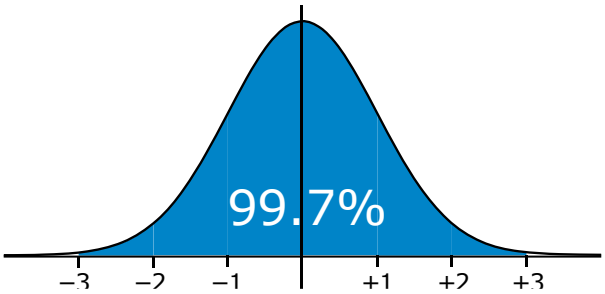
The standard deviation is a measure of how spread out numbers are.

When we calculate Standard Deviation, we can infer that

- 68% of values are within 1 standard deviation of the mean
- 95% of values are within 2 standard deviations of the mean
- 99.7% of values are within 3 standard deviations of the mean
- Rest of the 0.3 values lie beyond that.

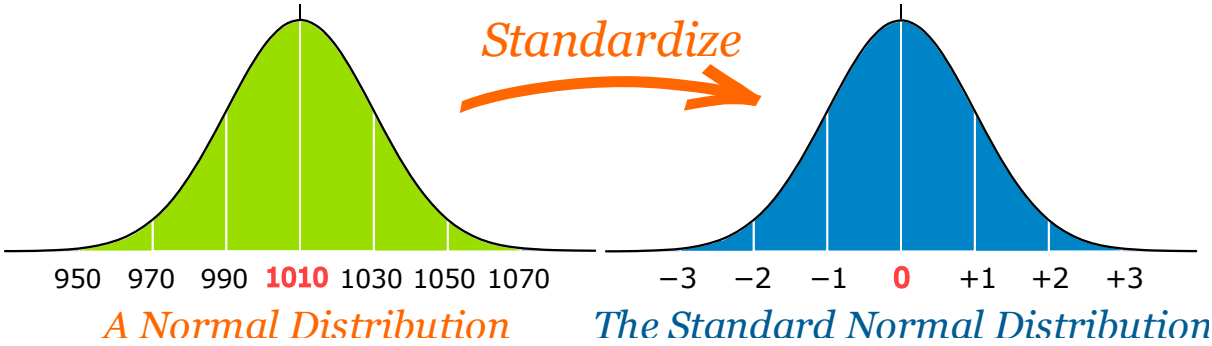






The number of standard deviations from the mean is also called "**z-score**".

So, to convert a value to standard score, when should first subtract the value from the mean and then divide it by standard deviation. The process of doing this is called Standardization.



The z-score formula that we have been using is –

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

Why Standardize ...?

It can help us make decisions about our data.

Example: Professor Willoughby is marking a test.

Here are the student's results (out of 60 points):

20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17

Most students didn't even get 30 out of 60, and **most will fail**.

The test must have been really hard, so the Prof decides to Standardize all the scores and only fail people 1 standard deviation below the mean.

The **Mean is 23**, and the **Standard Deviation is 6.6**, and these are the Standard Scores:

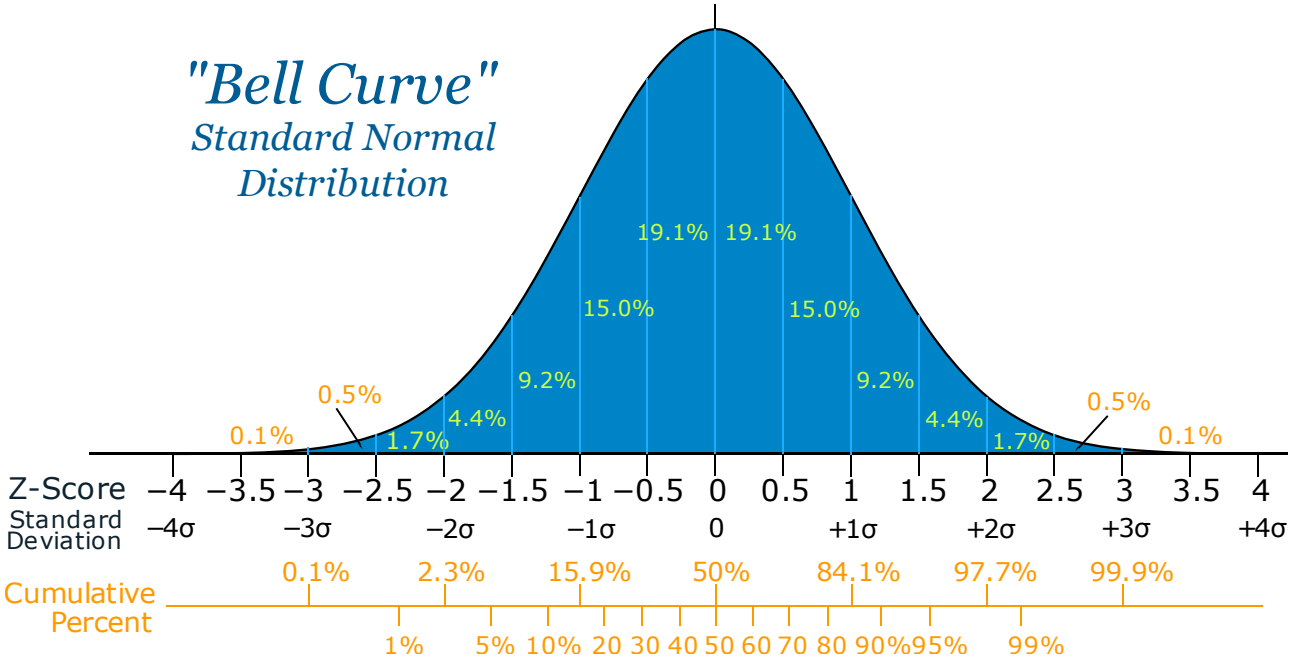
-0.45, -1.21, 0.45, 1.36, -0.76, 0.76, 1.82, -1.36, 0.45, -0.15, -0.91

Now only 2 students will fail (the ones lower than −1 standard deviation)

Much fairer!!

In More Detail:

Here is the Standard Normal Distribution with percentages for every half of a standard deviation, and cumulative percentages:



Now, let's see what an Empirical Rule is ...

The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all data falls within three standard deviations (denoted by σ) of the mean (denoted by μ).

Broken down, the empirical rule shows that,

- 68% falls within the first standard deviation (μ ± σ),
- 95% within the first two standard deviations (μ ± 2σ),
- and 99.7% within the first three standard deviations (μ ± 3σ).

Confidence Interval - A confidence interval is the range of values we are fairly sure our true value lies in. Let's see how to calculate the confidence interval through an example.

We measure the heights of 40 randomly chosen men, and get a mean height of 175cm,

We also know the standard deviation of men's heights is 20cm.

The 95% Confidence Interval is calculated as…

n= 40,

$\bar{x}$ = 175

s= 20

For the corresponding confidence level 95% - take the z value and apply in the formula for confidence interval.

Confidence Interval	z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

$$= \frac{\bar{x} + S}{N^{1/2}}$$

= $175 \pm 1.960 \times 20 / \sqrt{40}$

= $175\text{cm} \pm 6.20\text{cm}$

In other words: from 168.8cm to 181.2cm

This says the true mean of ALL men (if we could measure all their heights) is likely to be between 168.8cm and 181.2cm.

But it might not be!

The "95%" says that 95% of experiments like we just did will include the true mean, but 5% won't.

So there is a 1-in-20 chance (5%) that our Confidence Interval does NOT include the true mean.

Inferential Statistics

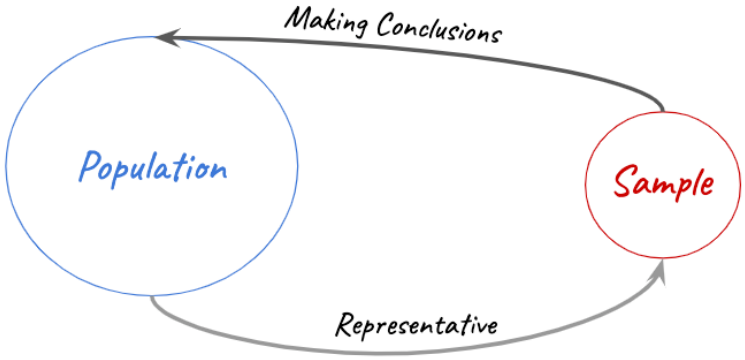
- Generalizes a large data set and applies probability to arrive at conclusion.
- It allows you to infer parameters of population based on sample stats and build models on it.

Statisticians use hypothesis testing to formally check whether the hypothesis is accepted or rejected. Hypothesis testing is an Inferential Statistical technique used to determine whether there is enough evidence in a data sample to infer that a certain condition holds true for an entire population.

To under the characteristics of a general population, we take a random sample and analyse the properties of the sample. We test whether or not the identified conclusion represents the population accurately and finally we interpret their results. Whether or not to accept the hypothesis depends upon the percentage value that we get from the hypothesis.

We have seen that descriptive statistics provide information about our immediate group of data. For example, we could calculate the mean and standard deviation of the exam marks for the 100 students and this could provide valuable information about this group of 100 students.

Any group of data like this, which includes all the data you are interested in, is called a population. A population can be small or large, as long as it includes all the data you are interested in. Descriptive statistics are applied to populations, and the properties of populations, like the mean or standard deviation, are called parameters as they represent the whole population (i.e., everybody you are interested in).



Often, however, you do not have access to the whole population you are interested in investigating, but only a limited number of data instead.

Properties of samples, such as the mean or standard deviation, are not called parameters, but statistics. Inferential statistics are techniques that allow us to use these samples to make generalizations about the populations from which the samples were drawn.

It is, therefore, important that the sample accurately represents the population. The process of achieving this is called sampling (sampling strategies are discussed in detail in the section, Sampling Strategy, on our sister site). Inferential statistics arise out of the fact that sampling naturally incurs sampling error and thus a sample is not expected to perfectly represent the population.

The methods of inferential statistics are -

(1) the estimation of parameter(s) and
(2) testing of statistical hypotheses.

Hypothesis Testing

A statistical hypothesis is an assumption about a population parameter. This assumption may/ may not be true. Hypothesis testing refers to formal procedures used by staticians to accept / reject statistical hypotheses.

Terms:

1. Null Hypothesis: the null hypothesis is denoted by H_o, is usually the hypothesis that sample observations result purely from chance.

2. Alternate Hypothesis: the alternate hypothesis denoted by H_1, is the hypothesis that sample observations are influenced by some non-random cause.

3. Level of significance: Refers to the degree of significance in which we accept or reject the null-hypothesis. 100% accuracy is not possible, so we therefore select a level of significance that is usually 5%.

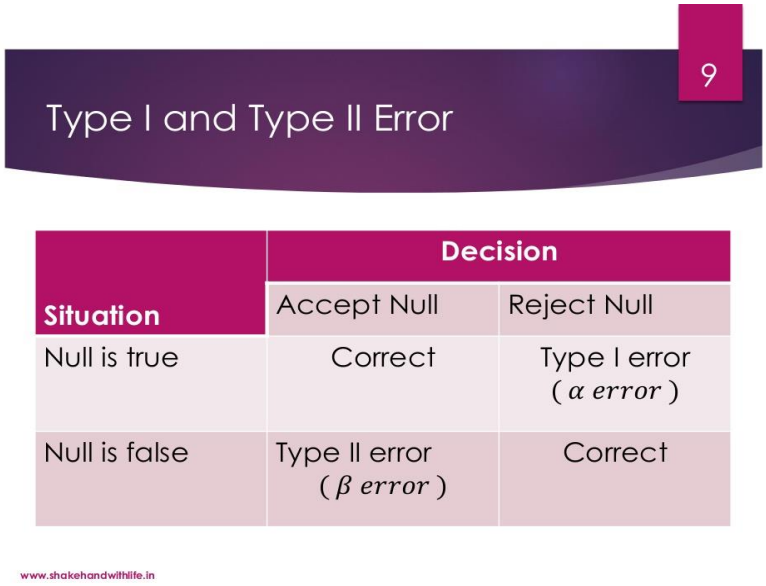
4. Type I error: When we reject the null hypothesis, although that hypothesis was true. Type I error is denoted by alpha. In hypothesis testing, the normal curve that shows the critical region is called the alpha region.

5. Type II error: When we accept the null hypothesis but it is false. Type II errors are denoted by beta. In Hypothesis testing, the normal curve that shows the acceptance region is called the beta region.

6. Power: Usually known as the probability of correctly accepting the null hypothesis. 1-beta is called power of the analysis.

7. One-tailed test: When the given statistical hypothesis is one value like H0: $\mu 1 = \mu 2$, it is called the one-tailed test.

8. Two-tailed test: When the given statistics hypothesis assumes a less than or greater than value, it is called the two-tailed test.



Step to follow:

- State the hypothesis – it involves stating null and alternate hypotheses. The hypotheses are stated in such a way that they are mutually exclusive.
- Choose the level of significance and critical value
- Find test value
- Testing the hypothesis
- Interpret results and draw conclusions

How to make a Decision after the hypothesis testing is over?

In statistical analysis, we have to make decisions about the hypothesis. These decisions include deciding if we should accept the null hypothesis or if we should reject the null hypothesis.

Every test in hypothesis testing produces the significance value for that particular test. In Hypothesis testing,

✓ If the significance value of the test is greater than the predetermined significance level, then we accept the null hypothesis.

✓ If the significance value is less than the predetermined value, then we should reject the null hypothesis.

Types of Hypothesis Testing:

1. Z-test
2. T-test
3. Chi-squared test
4. F-test

Z-Test:

Z-tests are among the most basic of statistical hypothesis testing methods. It is used to compare the mean of a normal random variable to a specified value, μ0. In Z-test, the inference is made from the standard normal distribution, and "Z" is the traditional symbol used to represent a standard normal random variable.

Assumptions:

- ✓ The number of samples taken should be greater than 30 (N>30).
- ✓ The standard deviation of the population should be known.
- ✓ Variable should be a continuous variable.

Steps in following Z-Test:

- State the hypothesis – it involves stating null and alternate hypotheses. The hypotheses are stated in such a way that they are mutually exclusive.
- Choose the level of significance
- Find critical value (level of confidence)
- Find test value
- Find test statistics
- Draw conclusions

Let's see an example……

Boys of a certain age are known to have a mean weight of μ = 85 pounds. A complaint is made that the boys living in a municipal children's home are underfed. As one bit of evidence, n = 25 boys (of the same age) are weighed and found to have a mean weight of $\bar{x}$ = 80.94 pounds. It is known that the population standard deviation σ is 11.6 pounds (the unrealistic part of this example!). Based on the available data, what should be concluded concerning the complaint?

Solution-

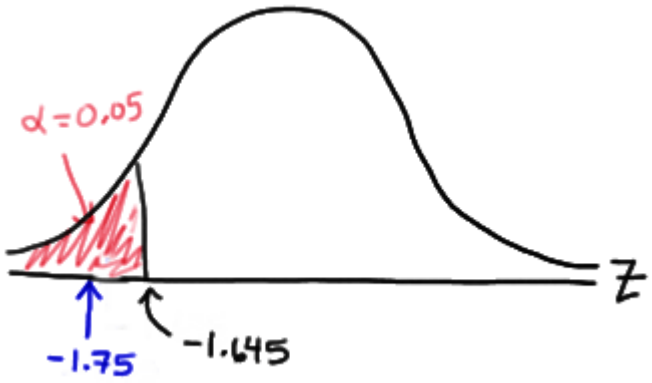
Null hypothesis, H_o= 85, and the alternative hypothesis is H_1: $\mu < 85$.

In general, we know that if the weights are normally distributed, then:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

$$Z = \frac{80.94 - 8511.6}{11.65/\sqrt{25}} = -1.75$$

The critical region approach tells us to reject the null hypothesis at the α = 0.05 level if Z < −1.645. Therefore, we reject the null hypothesis because Z = −1.75 < −1.645 and therefore falls in the rejection region:



T-Test:

T-tests are statistical hypothesis tests that you use to analyze one or two sample means. Depending on the t-test that you use, you can compare a sample mean to a hypothesized value, the means of two independent samples, or the difference between paired samples.

Types:

- ✓ One sample t-test
- ✓ Independent t-test
- ✓ Dependent t-test

Assumptions:

- ✓ The number of samples taken is less than 30 (N<30).
- ✓ The standard deviation of the population is not known.
- ✓ Variable should be a continuous variable.

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

- One sample t-test

It determines whether the sample is a part of the population or sample and population do not have relation between them.

- Independent t-test

It determines whether there is a statistically significant difference between means in two unrelated groups.

$H_0 = \mu_1 = \mu_2$

$H_0 = \mu_1 \neq \mu_2$

To do this, we need to set a significant level that allows us to reject / accept H_1.
An independent t-test helps to compare whether 2 groups have different average values. It asks whether a difference between 2 groups averages is unlikely to have occurred because of random chance in sample selection. A difference is more likely to be meaningful and real if

- Difference between averages is large,
- Sample size is large,
- Responses are consistently close to average values and not widely spread out.

Statistical significance indicates whether difference between sample average is likely to represent an actual difference between population and effect size indicates whether the difference is large enough to be practically meaningful.

Paired t-test is used when each observation in one group is paired with related observation in other group.

Dependent t-test

One dependent variable that is measured on a interval / ratio scale and one categorical variable that has only two related groups.
A dependent t-test is an example of a "within-subjects" / repeated measures. This indicates that same participants are tested more than once. Thus, in dependent t-tests, "related groups" indicates that the same participants are present in both groups. The reason that is possible to have same participants in each group is because each participant has been measured on two occasions on same dependent variable.
The dependent t-test can be used to test either a "change" / "difference " 96in means between two relates groups, but not both at the same time.

Whether you are measuring a "change" / "difference" between means of two relates groups depends on your study design.