Unit

5

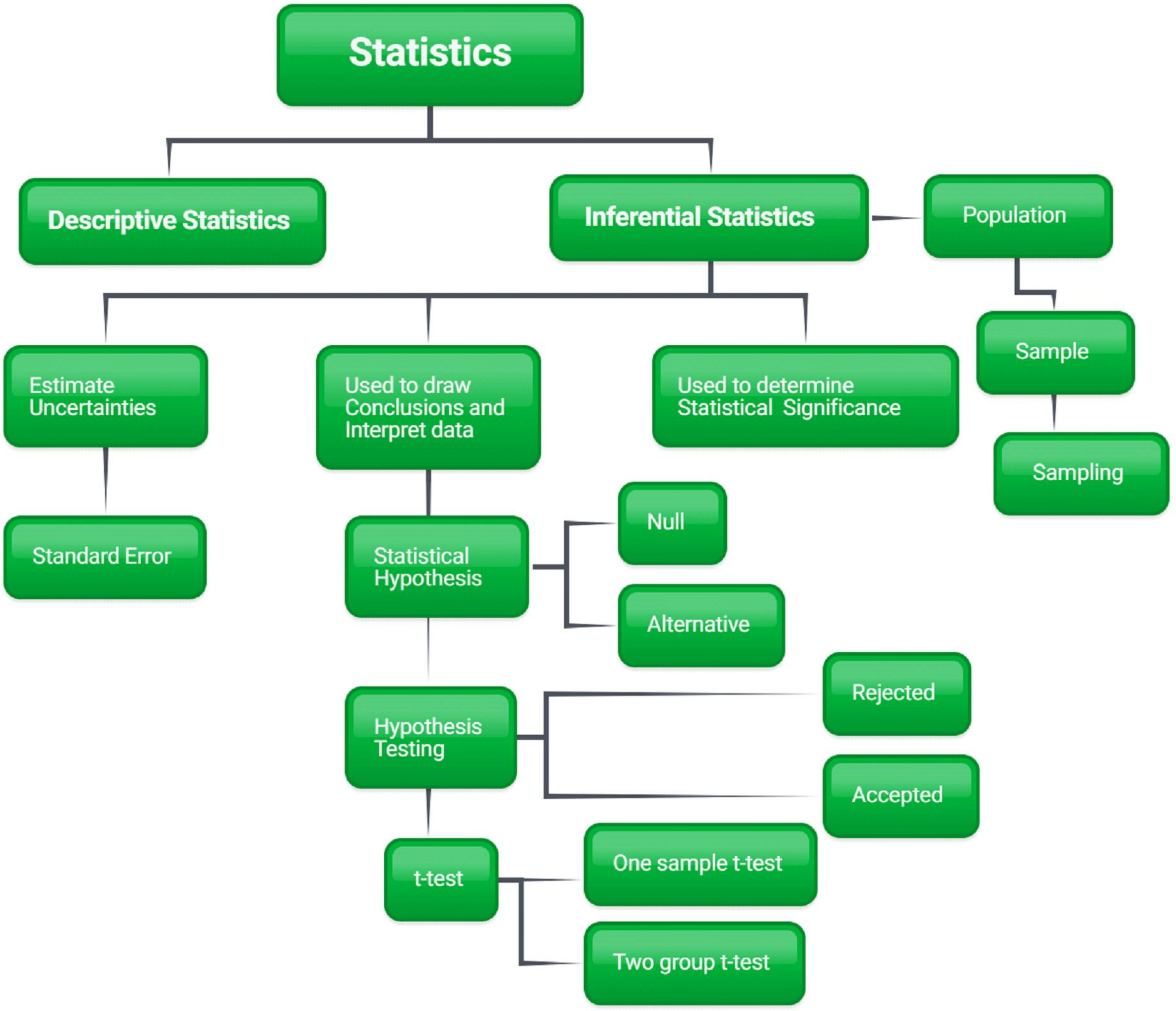
Inferential Statistics

5.0 LEARNING OUTCOMES

After the completion of this unit, the students will be able to :

- develop an understanding of population and sample
- understand the concept of parameter and statistical interferences
- understand the idea of a hypothesis testing
- use and extend the knowledge of inferential statistics and their applications in real-life situations

Concept Map



Introduction

One of the most important application of statistics is making estimations about an entire population based on the information from a small sample. This process is known as **statistical inference**. This can be achieved only if we are confident that our sample accurately reflects the desired population. For example, making exit poll results of public opinions using a small group of thousand voters and exactly predicting the outcome of an election in which millions of votes are cast.

This chapter on *inferential statistics* will take you to see how to draw conclusions from a sample and generalize them to a larger population.

5.1 Population and sample

Several real-life problems are statistical in nature. Let's take some examples;

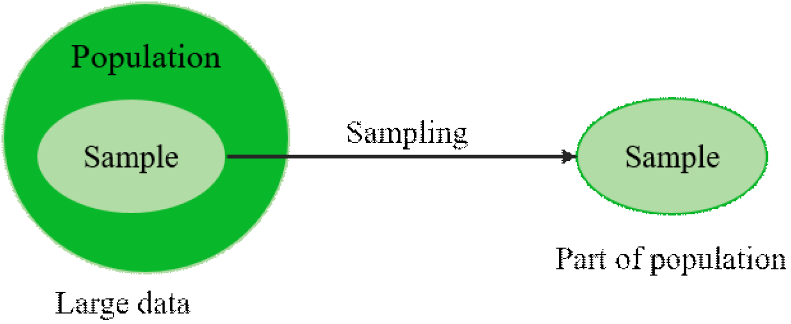
1. You are a part of a fitness campaign in your school. You are concerned about the overall wellbeing of fellow students and want to know that what proportion of students regularly do exercises.

2. As a quality control expert, you want to know what percentage of good computer chips are produced by the manufacturing unit of your company in a week.

In example 1, the population under study is total number of students enrolled in school as you want to conduct study on them. In example 2, the population is the total number of computer chips produced by manufacturing unit in a week then out of it you will see what proportion is good.

Thereby a *population* is a group of all distinct individuals or objects that you want to draw conclusions about. The number of individuals/objects in a population is called population size.

In statistics, we commonly use a *sample* that is a small subset of a larger set of data for making inferences about the large set. Here larger set is population out of which sample is drawn.



NOTE : ● *Every time the sample size is smaller than the population's total size.*

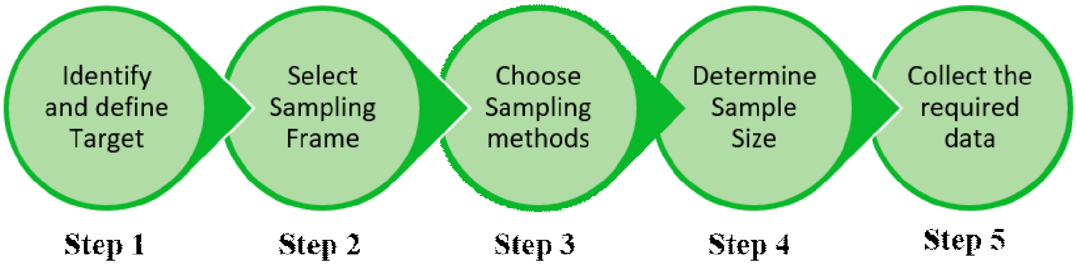
● *The population refers to the entire group from which you want to draw conclusions.*

Sampling

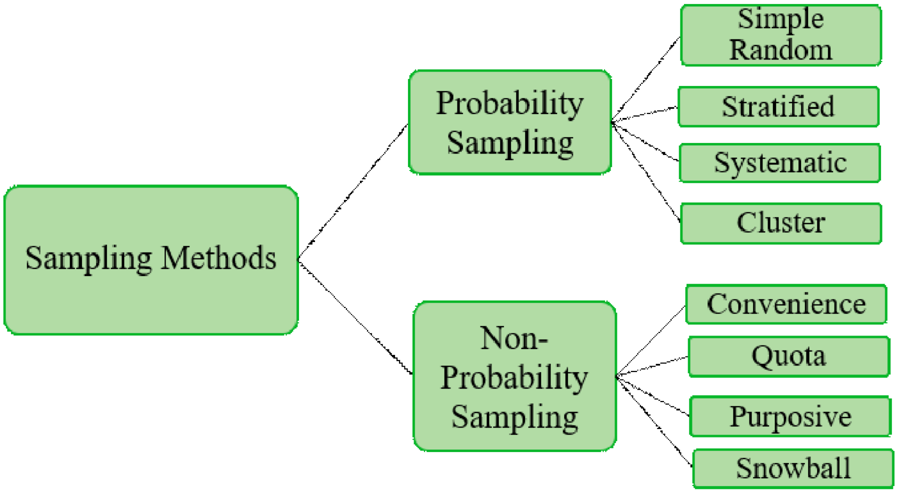
Sampling is a technique of selecting small group (subset) of population for estimating the characteristics, without having to investigate every individual. It includes selecting a group of people, events, behaviors, or other elements with which we are concerned to make our conclusions. We can extend our results obtained from sample group to the entire population.

Let us suppose a vaccine company has manufactured a new vaccine for COVID-19 and would like to see its adverse effects on country's population, then it is almost impossible to perform clinical trials that includes all. So in this scenario, researchers select a group of people from each demographic for conducting the tests on them and estimates the impact on whole population.

Steps involved in Sampling



There are number of ways in which the sampling process can be carried out. But in this chapter, we shall limit ourselves to simple random sampling and systematic random sampling only.



1. **Probability Sampling:** Randomization (choosing something at random) is used in this sampling method to ensure that every member of the population has an equal chance of being included in the selected sample.
2. **Non-Probability Sampling:** Randomization is not used in non-probability sampling. The result of this method can be biased, making it difficult for all the elements of population to be included in the sample equally.

Simple random sampling

As name suggests here every individual is chosen entirely by chance and every member of the population has an equal chance of being included in sample.

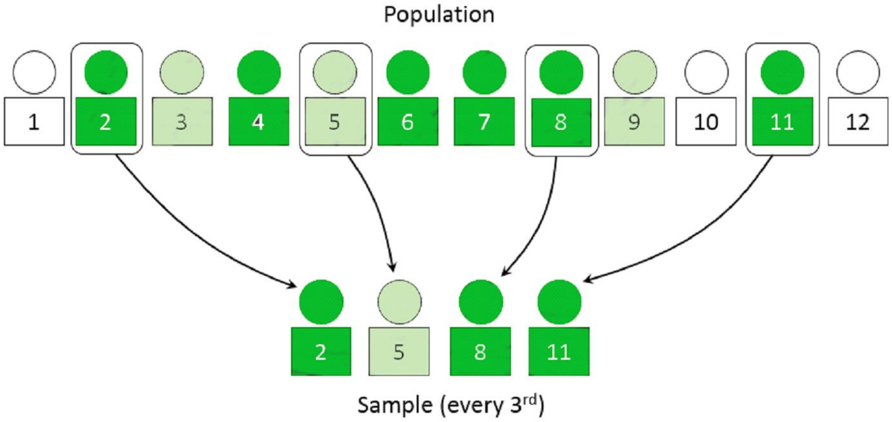
Suppose from a finite population of size N we take a sample of size 'n'. It implies that we will have ${}^{N}C_{n}$ possible samples to choose. A sampling method wherein each of the ${}^{N}C_{n}$ samples has an equal chance to be selected is known as random sampling and the sample attained by this method is called a random sample.

For example: From a class of 50 students randomly selecting 10 students where every student has equal opportunity of getting selected. The probability of every selection is 1/50.

Systematic random sampling

When members of sample are chosen at regular interval of population. This requires starting point of selection and sample size which then follows repetition of the same. The items of population can be first arranged alphabetically, numerically or in any increasing/decreasing order then individuals are chosen at regular interval.

For example: Suppose all twelve students of your class are listed as per their roll number. You randomly select a starting point 2 then 2 onwards, every 3rd student is selected (2,5,8,11) and you end up with a sample of 4 students selected in systematic method.



ACTIVITIES

- *What should be the suitable sample to collect data on how people use smart phones these days?*
- *Use simple random sampling to collect data on dress codes among women. Choose your sample across age, profession and family background.*

Representative and unrepresentative Sample

Representative Sample: *A sample which accurately represents, reflects, or matches with some of the features of your population.* It has to be an unbiased reflection of the population. For example, if you have selected a representative sample of Indian Cricket team fans and found that 75% of your sample are male, then it follows that 75% of your target population will also be male.

In order to get the correct inferences, you should always try to make your sample representative of your target population. However, sometimes you might deliberately choose not to study a representative sample.

Sample needs to fulfill following conditions to be a representative sample:

(*i*) The sampling process must have a component of **random selection**.

(*ii*) The sample size must be large enough to give us a good picture of the **variability** of the population.

Unrepresentative Sample: *When the statistic does not represent the population parameter, it is called unrepresentative sample.* For example, in some cases you might not want to make generalizations about a very large group of people based only on a very small group.

This is also known as biased Sample. The bias that results from an unrepresentative sample is called selection bias.

Suppose, for obtaining a sample of households, a TV rating service dials numbers taken at random from telephone directories. Then it is going to be an unrepresentative sample as some households may have unlisted telephone numbers.

Unbiased and biased sampling

Unbiased Sampling:

If every individual or the elements in the population has an equal chance to be part of the selected sample, then the sampling process is called unbiased. "Probability sampling is unbiased in nature." Some of the unbiased sampling are Simple random, Stratified and Systematic random sampling.

For example :

i. One student is randomly selected by teacher, every week, to review the homework answers with rest of the class. This is unbiased because it uses simple random sampling process.

ii. To know the students' favorite sport, every fifth student who enters the school is asked to tell the name of their most favourite sport. It is unbiased as systematic sampling process is being followed.

Biased Sampling:

If a sampling process systematically favours certain outcomes over others, it is said to be biased. Convenience sampling type is one of the biased sampling. The following example shows how a sample can be biased,

For example;

i. Suppose for making selections for a competition, a teacher selects those students whose roll numbers ends with the digit 2. Then it is not a simple random sampling because every student does not have a chance to be chosen.

Types of biasness

i. Voluntary response bias: When individual has choice to choose to participate.

ii. Undercover: If sample gives less representation of the sample.

iii. Convenience: When a sample is taken from individual that are conveniently available.

iv. Response bias: Anything in a survey that influence responses.

Sampling errors

The difference between a population parameter and a sample statistic is known as a sampling error. Even randomly selected sample also contains sampling errors because random samples are not identical to the population in terms of numerical measures like means and standard deviations. It can be either positive or negative, and the estimated sampling error decreases as the sample size grows.

Sampling error = $\overline{x} - \mu$

Where $\overline{x}$ = *Sample Mean* and μ = *Population mean*

$$Population\ Mean\ (\mu)=\frac{\sum X_i}{N} \qquad Sample\ Mean(\bar{x})=\frac{\sum X_i}{n}$$

Where N = population size and n = sample size.

Reasons for sampling errors:

i. The population parameter is estimated differently by different samples.

ii. Faulty selection of sample.

iii. Small size of sample.

iv. Sample results have potential variability.

5.2 Parameter and Statistics and Statistical Interferences

5.2.1 Parameter and Statistics

Parameter	Statistic
➢ It is a characteristic of a population.	➢ It is a characteristic of a sample.
➢ A parameter is a numerical value that is taken from the entire population, such as the population mean.	➢ A statistic is the numerical value taken from a sample and calculated from the sample observations alone, i.e. some subset of the entire population.
➢ The value of a parameter is computed from all the population observations.	➢ The value of a statistic is computed from portion of population (sample).
➢ Generally denoted by Greek alphabets (mean-μ, S.D.-σ, Variance- σ^2 etc.)	➢ Generally denoted by english alphabets (Mean –X, S.D. –S, Variance –S^2, etc.
Example: Under a study of calculating the average income of people of some specific region, the mean income and standard deviation of these incomes are parameters. Knowing the average height of adults in India is a parameter which is nearly impossible to calculate.	***Example:*** Mean and standard deviation of income of 1000 residents from South Delhi. Mean and standard deviation of height of 50 Indian adults.

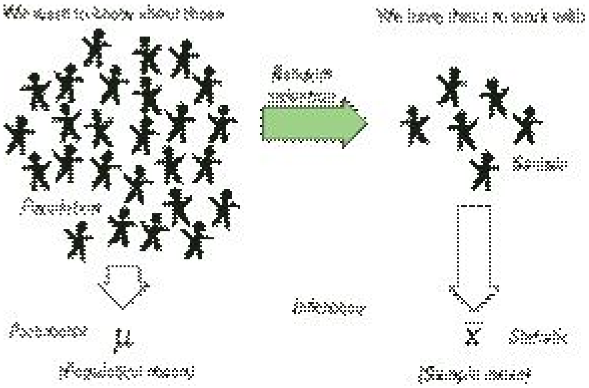


Image Source: https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

Thus, Parameter and statistic both are related yet distinct measures. The first refers to the whole population, while the second refers to part of the population.

5.2.2 Statistical Significance and Sampling distribution

Statistical Significance

Statistical significance is a measure of reliability of findings which establishes that when a finding is significant, it simply means we are confident that it is real and sample was framed wisely.

To decide if a data set's outcome is statistically significant, statistical hypothesis testing is used. When a statistic has high significance then it is thought to be more reliable.

Sampling distribution

The sampling distribution of a statistic is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population. It is a theoretical idea—we do not actually build it.

To put it another way, suppose we are regularly taking samples of the same sample size from the population, compute the statistics (Mean, S.D. mean), and then draw a histogram of those statistics, the distribution of that histogram tends to have is called the sample distribution of that particular statistics (Mean, S.D.).

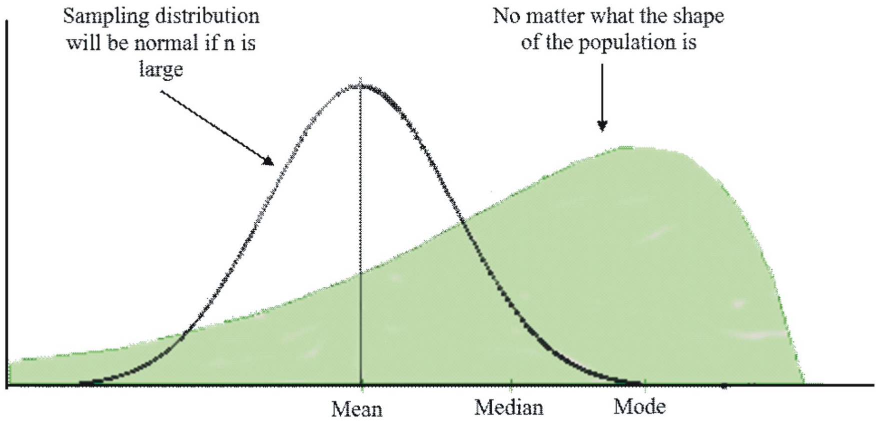
Central Limit Theorem (CLT)

Central limit theorem (CLT) implies that the distribution of a sample leads to become a normal distribution (bell curve shaped) as the sample size becomes larger, considering that all the sizes of samples are identical, whatever be the shape of the population distribution.

A sample size of 30 or more is considered to be sufficient to hold CLT and as the sample size becomes large the prediction of characteristics of population becomes more accurate.

***NOTE** : As per CLT, when sample size increases the mean of a sample of data becomes close to mean of overall population.*

The interesting thing about CLT is that as N increases, the sampling distribution of the mean approaches a normal distribution, regardless of the shape of the parent population.

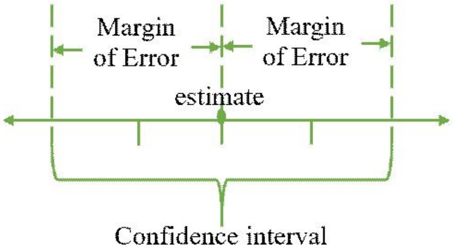


Central Limit Theorem

Confidence Interval (CI):

We use Confidence Interval (CI) to express the precision and uncertainty of a sampling process. A confidence level, a statistic and a margin of error are the three components of it. The margin of error describes the accuracy of a sampling method, while the confidence level explains its uncertainty.

Consider the case where we are computing an interval estimate of a population parameter with a 95% confidence interval. It means that 95% of the time, by using the same sampling method to pick different samples and computing different interval estimates, the true population parameter would fall within the margin of error specified by the sample statistic.



For example: Assume a news channel conducts pre-election survey and predicts that the candidate A will get 30% of the vote. According to news channel the survey had margin of error of 5% and a confidence level of 95%. This means that we are 95% sure that the candidate A will receive between 25% and 35% of the vote.

5.3 t-Test (a test of difference for parametric data)

5.3.1 Hypothesis

In order to make decisions it is useful to make some assumptions about the population. Such assumptions, which may or may not be true, are known as hypothesis. These are the tentative, declarative statement about the relationship between two or more variables. There are two types of statistical hypotheses for each situation: the **null hypothesis** and the **alternative hypothesis**. Both of these hypotheses contain opposite view points.

Null Hypothesis (H_0)

- The **null hypothesis (H_0)** states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters (i.e H_0 is a statement of a no relationship)- It explicitly says that the two groups we are studying are the same.

Alternative Hypothesis (H_1)

- The **alternative hypothesis (H_1)** states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters- In other words it says that the two groups we are studying are different

Symbols used in hypothesis	
H_0	H_1
equal (=)	not equal (≠) or greater than (>) or less than (<)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Example 1

If we want to examine that on an average college student take less than five years to complete their education. The null and alternative hypotheses are:

$$H_0 : \mu \geq 5$$

$$H_1 : \mu < 5$$

Writing null hypothesis

Case 1: Suppose a cake baked through conventional method has an average life span of μ days and it is proposed to test a new process of baking cakes. So, we have two populations of cakes (one by conventional method and other by new process). Here hypothesis can be formed like:

(i) New method is better than conventional method.

(ii) New method is inferior to conventional method.

(iii) There is no difference between the two methods.

Since first two statements display a preferential mentality, they tend to be biased. As a result, adopting the hypothesis of no difference, i.e. a neutral or null attitude toward the outcome, is the safest course of action. Thus, if the average life of cakes baked using the new method is μ_0, the null hypothesis is:

$$H_0: \mu = \mu_0$$

Case 2 : Suppose a departmental store is planning to have its own android application (app) conditioned that new service will be introduced only if more than 60% of its customers use internet to shop. So here null hypothesis would be that % of customers using internet is less or equal to 60% and the alternative hypothesis will be its opposite.

H_0: Proportion of customers using internet for shopping ≤ 60%

H_1: Proportion of customers using internet for shopping > 60%

If the null hypothesis is rejected, then the alternative hypothesis H_1 will be accepted and as a result e-commerce shopping service will be introduced.

ACTIVITY

Choose type of hypothesis from following statements and write them (H_0, H_1) in terms of the appropriate parameter (μ or p).

(i) During COVID-19 pandemic, the chance of getting infected from virus is under 25% for school students.

(ii) Fewer than 7% of students ride two-wheeler to reach the school on time.

(iii) The average salary package for Delhi University graduates is at least ₹ 10,00,000/annum.

Answers:

(i)	H_0: $p \geq 0.25$;	H_1: $p < 0.25$
(ii)	H_0: $p = 0.07$;	H_1: $p < 0.07$
(iii)	H_0: $\mu \geq 10,00,000$;	H_1: $\mu < 10,00,000$

NOTE

- *After stating the hypothesis, the researcher designs the study. The researcher selects the correct statistical test, chooses an appropriate level of significance, and formulates a plan for conducting the study.*
- *If we discard the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.*

Standard Error of Mean (σ_M)

When we take a sample from a population, we pick up one of many samples. Some of them will have the same mean whereas some will have very different means. Standard error of the mean (SEM) measures how much dispersion there is likely to be in a sample's mean compared to the population mean i.e it measures the standard deviation of sampling distribution about the mean.

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

σ = Standard deviation of original distribution

N = Sample size

- **Small SEM:** Having large number of observations and all of them being close to the sample mean (large N, small SD) gives us confidence that our estimation of the population means (i.e., that it equals the sample mean) is relatively accurate.
- **Large SEM:** Having small number of observations and they vary a lot (small N, large SD), then population estimation is likely to be quite inaccurate.

Degrees of freedom

The number of independent pieces of information on which an approximation is based is known as the degrees of freedom. You can also think of it as the number of values that are free to vary as you estimate parameters.

Example 1

Consider a classroom having seating capacity of 30 students. The first 29 students have a choice to sit but the 30th student can only sit on the one remaining seat. Therefore, the degrees of freedom is 29.

Example 2

For scheduling three hour-long tasks (read, eat and nap) between the hours of 5 p.m. and 8 p.m. we have two degrees of freedom as any two tasks can be scheduled at will, but after two of them have been set in time slots, the time slot for the third is decided by default.

Degrees of freedom is some or other way related with the size of the sample because higher the degrees of freedom generally mean larger sample sizes.

Note: *A higher degree of freedom means more power to reject a false null hypothesis and find a significant result.*

$$Df = N-1$$

where: *Df = degrees of freedom and*

N = sample size

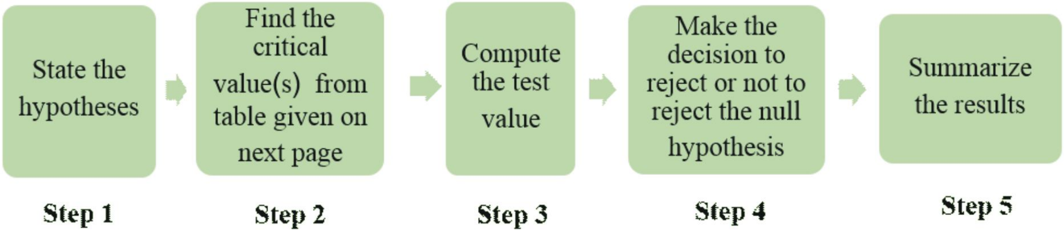
The t-Test (for one sample and two independent groups)

The t test is a statistical test for the mean of a population and is used when the population is normally or approximately normally distributed with an unknown variance.

The inferential statistic calculated in the t-test is called the t-ratio and denoted by "t". The larger the t-ratio (in absolute value), the more likely we will reject the null hypothesis because the more evidence in the data that the two groups differ from each other.

Note: "t" statistic is used to determine whether the null hypothesis should be rejected or not.

Use following procedure for testing the hypotheses by using the t test (traditional method):


State the hypotheses
Step 1
Find the critical value(s) from table given on next page
Step 2
Compute the test value
Step 3
Make the decision to reject or not to reject the null hypothesis
Step 4
Summarize the results
Step 5

NOTE :

- *If the population is roughly normally distributed and the population standard deviation is unknown, then only t test should be used.*
- *Perform a two-tailed t-test if you only want to see if the two populations are different from one another.*
- *Perform one-tailed t-test if you wish to know whether one population mean is greater than or less than the other.*

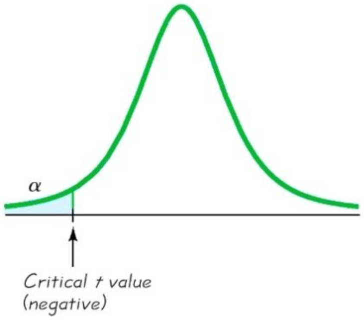
Reading and locating t-value from the t-table;

The t distribution table values are critical values of the t distribution. The column header are the t distribution probabilities (alpha) whereas the row depicts the degrees of freedom (df). This can be used for both one-sided (lower and upper) and two-sided tests using the appropriate value of α.

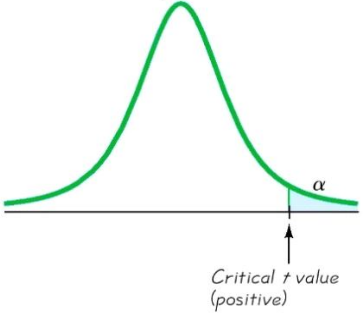
t Distribution: Critical *t* Values

	Area in One Tail				
	0.005	0.01	0.025	0.05	0.10
Degrees of Freedom	Area in Two Tails				
	0.01	0.02	0.05	0.10	0.20
1	**63.657**	31.821	**12.706**	6.314	3.078
2	**9.925**	6.965	**4.303**	2.920	1.886
3	**5.841**	4.541	**3.182**	2.353	1.638
4	**4.604**	3.747	**2.776**	2.132	1.533
5	**4.032**	3.365	**2.571**	2.015	1.476
6	**3.707**	3.143	**2.447**	1.943	1.440
7	**3.499**	2.998	**2.365**	1.895	1.415
8	**3.355**	2.896	**2.306**	1.860	1.397
9	**3.250**	2.821	**2.262**	1.833	1.383
10	**3.169**	2.764	**2.228**	1.812	1.372
11	**3.106**	2.718	**2.201**	1.796	1.363
12	**3.055**	2.681	**2.179**	1.782	1.356
13	**3.012**	2.650	**2.160**	1.771	1.350
14	**2.977**	2.624	**2.145**	1.761	1.345
15	**2.947**	2.602	**2.131**	1.753	1.341
16	**2.921**	2.583	**2.120**	1.746	1.337
17	**2.898**	2.567	**2.110**	1.740	1.333
18	**2.878**	2.552	**2.101**	1.734	1.330
19	**2.861**	2.539	**2.093**	1.729	1.328
20	**2.845**	2.528	**2.086**	1.725	1.325
21	**2.831**	2.518	**2.080**	1.721	1.323
22	**2.819**	2.508	**2.074**	1.717	1.321
23	**2.807**	2.500	**2.069**	1.714	1.319
24	**2.797**	2.492	**2.064**	1.711	1.318
25	**2.787**	2.485	**2.060**	1.708	1.316
26	**2.779**	2.479	**2.056**	1.706	1.315
27	**2.771**	2.473	**2.052**	1.703	1.314
28	**2.763**	2.467	**2.048**	1.701	1.313
29	**2.756**	2.462	**2.045**	1.699	1.311
30	**2.750**	2.457	**2.042**	1.697	1.310
31	**2.744**	2.453	**2.040**	1.696	1.309
32	**2.738**	2.449	**2.037**	1.694	1.309
34	**2.728**	2.441	**2.032**	1.691	1.307
36	**2.719**	2.434	**2.028**	1.688	1.306
38	**2.712**	2.429	**2.024**	1.686	1.304
40	**2.704**	2.423	**2.021**	1.684	1.303
45	**2.690**	2.412	**2.014**	1.679	1.301
50	**2.678**	2.403	**2.009**	1.676	1.299
55	**2.668**	2.396	**2.004**	1.673	1.297
60	**2.660**	2.390	**2.000**	1.671	1.296
65	**2.654**	2.385	**1.997**	1.669	1.295
70	**2.648**	2.381	**1.994**	1.667	1.294
75	**2.643**	2.377	**1.992**	1.665	1.293
80	**2.639**	2.374	**1.990**	1.664	1.292
90	**2.632**	2.368	**1.987**	1.662	1.291
100	**2.626**	2.364	**1.984**	1.660	1.290
200	**2.601**	2.345	**1.972**	1.653	1.286
300	**2.592**	2.339	**1.968**	1.650	1.284
400	**2.588**	2.336	**1.966**	1.649	1.284
500	**2.586**	2.334	**1.965**	1.648	1.283
750	**2.582**	2.331	**1.963**	1.647	1.283
1000	**2.581**	2.330	**1.962**	1.646	1.282
2000	**2.578**	2.328	**1.961**	1.646	1.282
Large	**2.576**	2.326	**1.960**	1.645	1.282

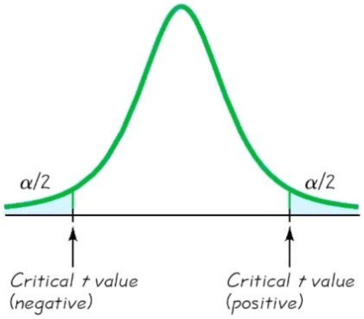
Left tail



Right tail



Two tails



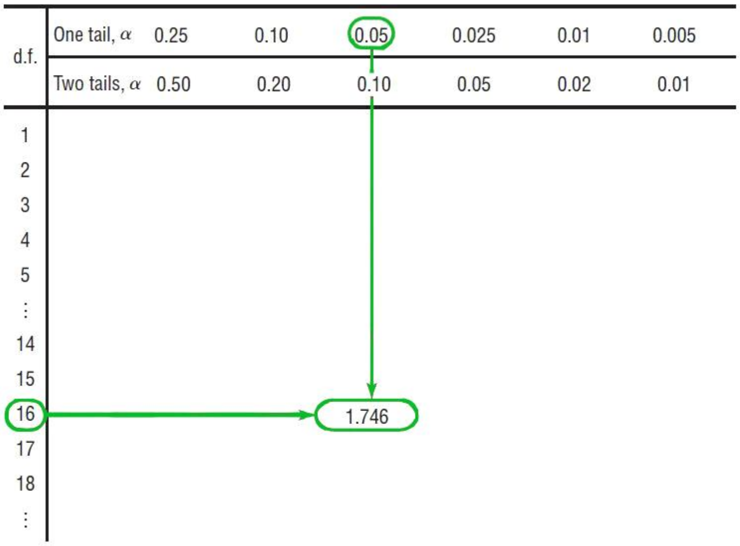
Source: http://www.math.odu.edu/stat130/t-tables.pdf

For a one-tailed test, find α level by looking at the top row of the table and finding the appropriate column. Look down the left-hand column for the degrees of independence.

Example:

Find the critical t value for α = 0.05 with d.f. = 16 for a right-tailed t test.

Solution: Look for 0.05 column in top row and 16 in left hand column. The critical value of found where row and column meet. It is +1.746.



d.f.	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
4							
5							
⋮							
14							
15							
16				1.746			
17							
18							
⋮							

5.3.2 One sample t- test

The one sample t-test is used to compare a sample mean to a specific value. In this test, we draw a random sample from the population and then compare the sample mean with the population mean and make a statistical decision as to whether or not the sample mean is different from the population.

$$t = \frac{\textit{Mean} - \text{Comparison value}}{\textit{Standard Error}}$$

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

μ_0 = The test value, $\bar{x}$ = Sample mean, n = Sample size and S = Sample standard deviation

This t value is compared to the critical t value from the t distribution table with degrees of freedom df = n – 1 and confidence level chosen. We reject the null hypothesis if the measured t value is greater than the critical t value.

Example:

Let us consider the average rainfall in a given area is 8 inches. However, a local meteorologist claims that rainfall was above average from 2016-2020 and argues that average rainfall during this period was significantly different from overall average rainfall. The following is the average rainfall for the observed period of 2016-2020:

Year	2016	2017	2018	2019	2020
Rainfall (inches)	8	5	7	5	6

Solution: Sample mean ($\bar{x}$) = 6.2, Sample size (n) = 5 and

Sample standard deviation (S) = 1.30. Here we are comparing a single sample mean (6.2 inches) to a known population mean (8 inches).

Step 1: ***Null Hypothesis:*** The average annual rainfall from 2016-2020 is the same as the overall average annual rainfall of 8 inches. If any difference is observed it is purely due to the random error.

Alternative Hypothesis: The average annual rainfall from 2016-2020 was not the same as the overall average annual rainfall of 8 inches, but was significantly higher. The observed difference is not solely due to random error, but rather indicates a true difference in average annual rainfall.

Step 2: $$t = \frac{6.2 - 8}{1.3/\sqrt{5}} = -3.10$$

Step 3:

Where $t_{.025}$ is the critical value from the t distribution and is found using:

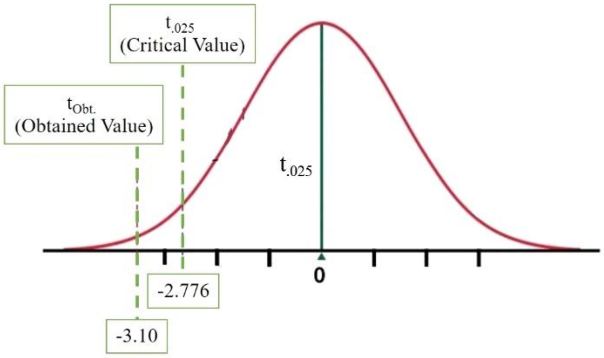
df = N–1 = 5–1 = 4

Step 4 :

Since t (4) = -3.10, p <.05; Reject the null hypothesis.

The null hypothesis is rejected since the obtained value is more extreme than the critical value (p = .05)

Hence, we can say that there was less-than average rainfall 2016-2020. The observed average rainfall for this period does not appear to be due to random error alone, but suggests that the weather pattern for the local area was different during the period studied."



5.3.3 t- test for two independent groups

This compares two groups (experimental and control groups) and helps us to see whether a statistically significant difference exists between the two means-thus, the ***two-sample t test compares two group means.*** For example, we can use t-test to testing the following hypothesis:

"It is expected that boys will have higher Mathematics scores than the girls."

Hypothesis for two independent samples can be expressed following ways:

H_0: $\mu_1 = \mu_2$ ("the two-population means are equal")

H_1: $\mu_1 \neq \mu_2$ ("the two-population means are not equal")

$$t = \frac{\textit{Sample one mean} - \textit{Sample two mean}}{\textit{Standard error of the difference in means}}$$

There are two forms of the test statistic for this test.

Case 1: When Variances are assumed to be equal

When the two independent samples are assumed to be drawn from populations with identical population variances (i.e., $\sigma_1^2 = \sigma_2^2$), the test statistic t is computed as

$$t = \frac{\overline{x}_1 - \overline{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad \text{Where,} \qquad S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

s_1 = standard deviation of first sample

s_2 = standard deviation of second sample

s_p = pooled standard deviation (a combined estimate of the overall standard deviation)

Case 2 : When Variances are assumed to be unequal

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Use $\overline{x}_1 - \overline{x}_2$ if $\overline{x}_1 \succ \overline{x}_2$

Use $\overline{x}_2 - \overline{x}_1$ if $\overline{x}_1 \prec \overline{x}_2$

The calculated t value is then compared to the critical t value from the t distribution table with following formula of degrees of freedom and chosen confidence level:

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1 - 1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2 - 1}(\frac{s_2^2}{n_2})^2}$$

If the calculated t value > critical t value, then we reject the null hypothesis.

Example

Country A has an average farm size of 191 acres, while Country B has an average farm size of 199 acres. Assume the data were attained from two samples with standard deviations of 38 and 12 acres and sample sizes of 8 and 10, respectively. Is it possible to infer that the average size of the farms in the two countries is different at α = 0.05? Assume that the populations are normally distributed.

Solution:

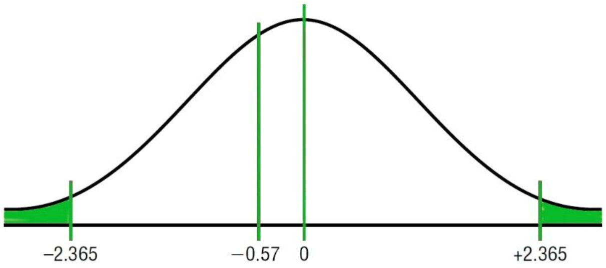
Step 1: Hypothesis H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$ (claim)

Step 2: Find the critical values. The test is two-tailed and α = 0.05, also variances are unequal, the degrees of freedom are the smaller of n_1–1 or n_2–1. In this case, the degrees of freedom are 8 – 1 = 7. Hence, from t-table F, the critical values are –2.365 and –2.365.

Step 3 : $t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}} = \frac{191 - 199}{\sqrt{\frac{38^2}{8} + \frac{12^2}{10}}} = -0.57$

Step 4 : Make the decision.

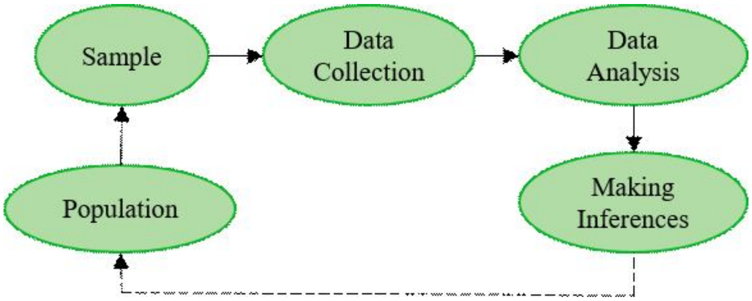
Do not reject the null hypothesis, since - 0.57 > -2.365.



Step 5: Make Conclusion. There is not enough evidence to support the claim that the average size of the farms is different.

5.4 Chapter Summary

- Process of drawing statistical inferences is as follows



- Hypothesis is an educated guess which needs to be tested.
- Sampling distribution: A sampling distribution is a distribution of possible values of a statistic for a given size sample selected from population.
- **Estimation:** The process by which one makes inferences about a population, based on the information obtained from a sample.
- **Confidence Interval:** It is the amount of uncertainty associated with a sample estimate of a population parameter.
- **Hypothesis testing:** It is the procedure used by statisticians to accept or reject statistical hypotheses.
- Sampling error = $\overline{x} - \mu$
- Central Limit Theorem (CLT): Sampling distribution leads to be normal (bell curve shaped) if n is large, no matter what the shape of the population is
- Degree of Freedom (Df) = N–1, where N = sample size
- T-test for one sample

$$t = \frac{\overline{x} - \mu_0}{S/\sqrt{n}}$$

- T- test for two independent groups

$$t = \frac{\overline{x}_1 - \overline{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

When Variance is equal

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When Variances is unequal

5.5 Online resources

1. Virtual Laboratories in Probability and Statistics https://www.randomservices.org/random/
2. Statistics online support
   http://sites.utexas.edu/sos/
3. Website for Statistical Computation
   http://vassarstats.net/

4. Intro to Inferential Statistics (Free course on Udacity)
   https://www.udacity.com/course/intro-to-inferential-statistics--ud201?irclickid=x7HRwbRQBxyLUMFwUx0Mo3QnUkEXcow3a1SlSI0&irgwc=1&utm_source=affiliate&utm_medium=&aff=259799&utm_term=&utm_campaign=_gtc_search_&utm_content=&adid=788805
5. Statistical Resources
   https://sixsigmastats.com/

Exercise– 5.1

1. Identify the below statement as biased or Unbiased statement. Justify your answer.
   "For a survey about daily mobile uses by students, random selection of twenty students from a school"
2. (i) Find the critical t value for α = 0.01 with d.f.= 22 for a left-tailed test.
   (ii) Find the critical t values for α = 0.10 with d.f.=18 for a two-tailed t test.
3. Suppose that a 95% confidence interval states that population mean is greater than 100 and less than 300. How would you interpret this statement?
4. A shoe maker company produces a specific model of shoes having 15 months average lifetime. One of the employees in their R & D division claims to have developed a product that lasts longer. This latest product was worn by 30 people and lasted on average for 17 months. The variability of the original shoe is estimated based on the standard deviation of the new group which is 5.5 months. Is the designer's claim of a better shoe supported by the findings of the trial? Make your decision using two tailed testing using a level of significance of p < .05.
5. An electric light bulbs manufacturer claims that the average life of their bulb is 2000 hours. A random sample of bulbs is tested and the life (x) in hours recorded. The following were the outcomes:
   $$\Sigma x = 127808 \quad \text{and} \quad \Sigma(\overline{x} - x)^2 = 9694.6$$
   Is there sufficient evidence, at the 1% level, that the manufacturer is over estimating the life span of light bulbs?
6. A fertilizer company packs the bags labelled 50 kg and claims that the mean mass of bags is 50 kg with a standard deviation 1kg. An inspector points out doubt on its weight and tests 60 bags. As a result, he finds that mean mass is 49.6 kg. Is the inspector right in his suspicions?
7. The average heart rate for Indians is 72 beats/minute. To lower their heart rate, a group of 25 people participated in an aerobics exercise programme. The group was tested after six months to see if the group had significantly slowed their heart rate. The average heart rate for the group was 69 beats/minute with a standard deviation of 6.5. Was the aerobics program effective in lowering heart rate?

Answers for Exercise 5.1

1. Unbiased because it is random sampling.
2. (i) -2.508 (ii) +1.734 and -1.734
4. Yes. Null hypothesis accepted.
5. No sufficient evidence to reject Null Hypothesis.
6. Yes. Null hypothesis is accepted.
7. Yes. There was significant effect of the aerobics in lowering heart rate.

❑❑❑