BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

Inferential Statistics

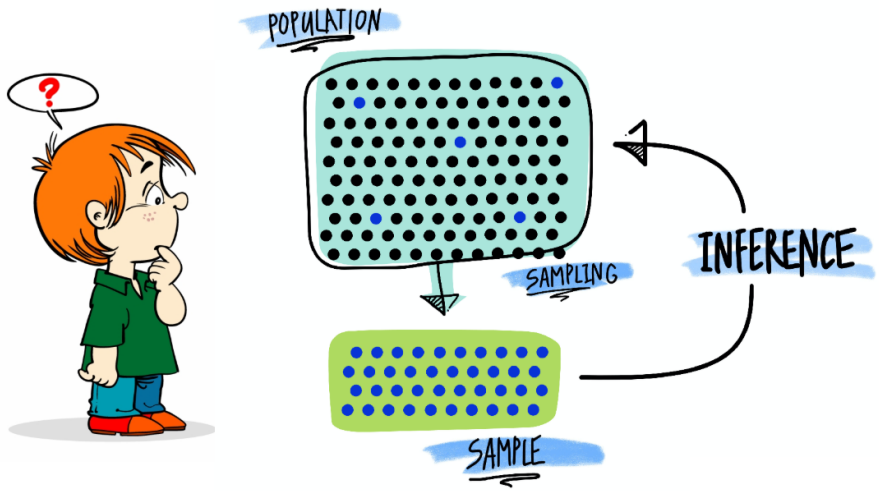
Revision-1.0

Prof Vineet Garg

BITS Pilani Work Integrated Learning Programmes (WILP)
Bangalore Professional Development Center

Introduction

- We have reviewed Descriptive Statistics, Probability and Probability Distributions in the first few modules.
- We have an idea that *Inferential Statistics* involves taking a sample from a population, computing a *statistic* on the sample, and then inferring from this statistic the value of the corresponding *parameter* of the population.
- Many times the sample statistics are going to be used to *verify the claim* of population parameters (Hypothesis Testing).
- Sampling from the population is done for several reasons; primarily because of feasibility and to save time and resources.
- Below is the proposed path to study inferential statistics:
  - ✓ Sampling and Sampling Distribution:
    - Sampling Techniques: ways to select a smaller part of the population.
    - Sampling Distribution: when several samples are picked up, they might provide different statistics e.g. different values of mean and standard deviation. What is the distribution of theses statistics?
  - ✓ Point and Interval Estimation of the Population Parameter:
    - Point Estimation: single value estimation of population parameter from the sample statistic.
    - Interval Estimation: estimation of an interval (range) where the population parameter is likely to fall.
  - ✓ Hypothesis Testing: accept or reject a claim about a population parameter gathering the evidence from the sample statistic.
- This module focuses on few of the above topics.



Sampling Techniques

Random Sampling

The sampling technique where every unit of the population has the same probability of being selected into the sample, is called the **Random Sampling**. Few methods:

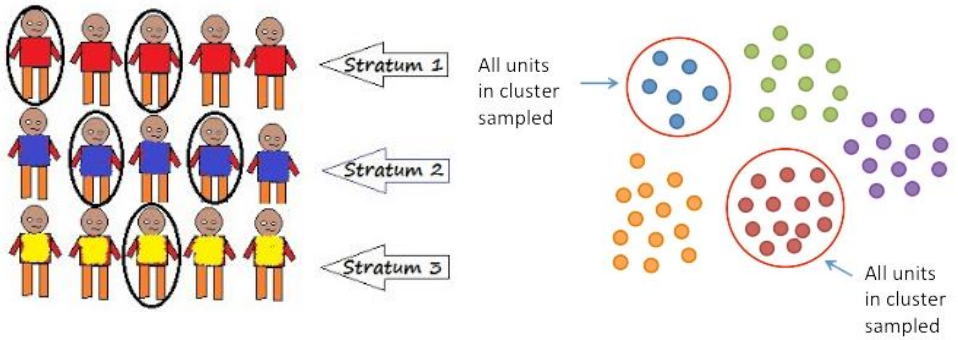
- **Simple Random Sampling**: a list of 30 companies are provided out of which 6 are to be selected randomly. A single digit random number generator output is used to perform the sampling. The random numbers are arranged in a group of 5 for ease of reading, otherwise they can be treated as a sequence of individual digits horizontally.
  - Since companies are numbered in two digits from 01 to 30, the pair of digits ranging from 01 to 30 are of interest in the random numbers. Other pairs will be ignored.
  - First, second and third pairs are 91, 56 and 74 respectively. They are ignored. The next pair is 25 which is acceptable so the company Occidental Petroleum is picked up in the sample. Continuing in the same way other companies that would be selected in the sample are: Procter & Gamble, Alaska Airlines, Bank of America, Alcoa and Sears.
- **Stratified Random Sampling:** the population is divided into non-overlapping sub-populations called strata and then a random sample from each stratum is drawn. It is advantageous in situations when representation of the population is required in the sample. Size of random sample from each stratum can be proportionate to the size of stratum in the population or disproportionate.
- **Systematic Sampling:** Every k^{th} item is to be selected in the sample (size = n) from the population (size = N) where k = N/n. It may not be a good sample if population is arranged in some order and the value of k is in syncopation with it. If population is random, then it could be a good approach.
- **Cluster Sampling:** The population is divided into clusters then randomly selected elements from all clusters are taken into the sample. It is a cost effective approach but does not guarantee the true representation of the population in the sample.

01 Alaska Airlines	11 DuPont	21 Lubrizol
02 Alcoa	12 ExxonMobil	22 Mattel
03 Ashland	13 General Dynamics	23 Merck
04 Bank of America	14 General Electric	24 Microsoft
05 BellSouth	15 General Mills	25 Occidental Petroleum
06 Chevron	16 Halliburton	26 JCPenney
07 Citigroup	17 IBM	27 Procter & Gamble
08 Clorox	18 Kellogg	28 Ryder
09 Delta Air Lines	19 Kmart	29 Sears
10 Disney	20 Lowe's	30 Time Warner

List of 30 Companies

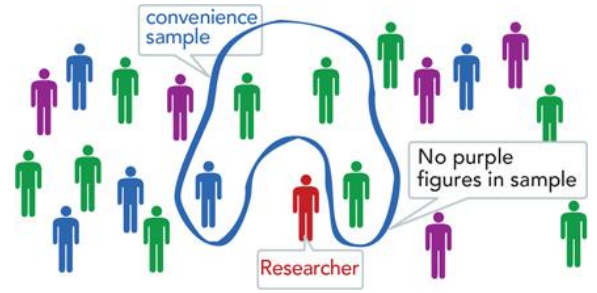
91567	42595	27958	30134	04024	86385	29880	99730
46503	18584	18845	49618	02304	51038	20655	58727
34914	63974	88720	82765	34476	17032	87589	40836
57491	16703	23167	49323	45021	33132	12544	41035
30405	83946	23792	14422	15059	45799	22716	19792
09983	74353	68668	30429	70735	25499	16631	35006
85900	07119	97336	71048	08178	77233	13916	47564

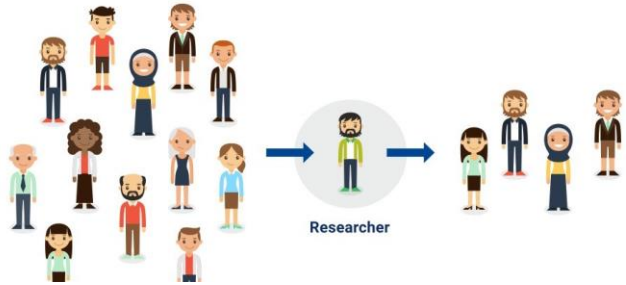
Random Numbers

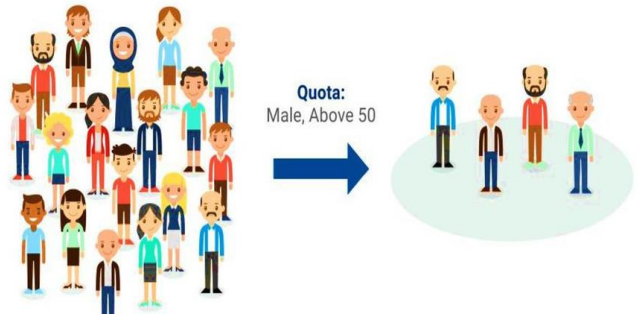


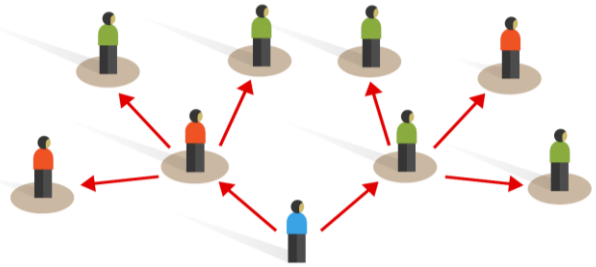
Non-Random Sampling

- The sampling technique where not every unit of the population has the same probability of being selected into the sample is called the **Non-Random Sampling.** It is also called the non-probability sampling.
- Non-Random Sampling is **NOT RECOMMENDED** for inferential statistics. Few of its variants are described here so that as the data scientists we are alert for such situations:
- **Convenience Sampling:** typically those elements are sampled that are readily available, nearby or willing to participate.
- **Judgmental Sampling:** sampling technique where the researcher selects elements to be sampled based on his own existing knowledge or his professional judgment. For example if researcher finds that males of age 50 and above are good candidates for his survey.
- **Quota Sampling:** it sounds similar to stratified random sampling but here the researcher selects the elements from each strata non-randomly. For example, in-place of randomly collecting data about Bengalis in New Delhi, the researcher visits Chitranjan Park in New Delhi itself (a populated place for Bengalis in New Delhi) and surveys.
- **Snowball Sampling:** The researcher identifies a person who fits the profile for the sampling. The researcher then asks this person for the names and locations of others who would also fit the profile of subjects.









Exercise

1. Unknowingly, best 50 performing stocks are kept on the top in the list of 150 companies of a stock exchange and 3 such stock exchanges' lists are prepared. If a random sample of 3 companies is to created from the population of these lists, explain why systematic sampling may not be a good technique.
2. An advertising company interested in determining the impact of TV advertisements conducts a survey in a geographical area. The area has three towns A, B and C. Town A is built around a factory and mostly factory workers with young kids live there. Town B contains mainly old retired people and the town C has mainly farmers. Assume all households have at least one TV set. There are 155 households in town A, 62 in town B and 93 in town C. The company wants to pickup up a sample of 40 households. Suggest them a suitable random sampling technique.

Sampling Distribution

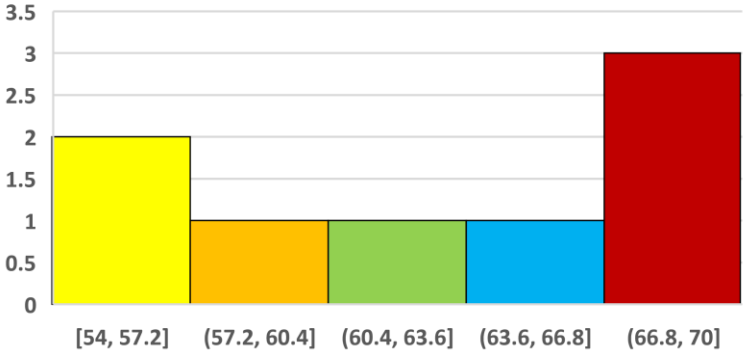
Central Limit Theorem (CLT)

- There is a small finite population of N = 8 numbers as: 54, 55, 59, 63, 64, 68, 69 and 70 with μ as mean and σ as standard deviation.
- The histogram is drawn, where [ ] indicates inclusive and () indicates exclusive values in the x-axis bins. Right-end inclusion method is followed here.
- The population size is small so its distribution cannot be ascertained.
- Different samples of size n = 2 (as x and y) from this population are drawn with replacement (8^2 = 64 possibilities).
- For these samples, means are also calculated. The mean of these samples ($\overline{X}$) will be different every time. May be the variations in the mean are not high.
- The questions arises here, if $\overline{X}$ is considered as a random variable, what kind of distribution it would exhibit?
- A histogram of the means of different samples is also drawn. It is normal in distribution.
- **Central Limit Theorem (CLT):** if there is a population with mean μ and standard deviation σ and if several random samples from the population with replacement are taken, then the ***distribution of the sample means*** will be approximately normally distributed. This will be true regardless of distribution of the population.
- The mean of the sample means ($\mu_{\overline{X}}$) and the standard deviation ($\sigma_{\overline{X}}$) is given by:

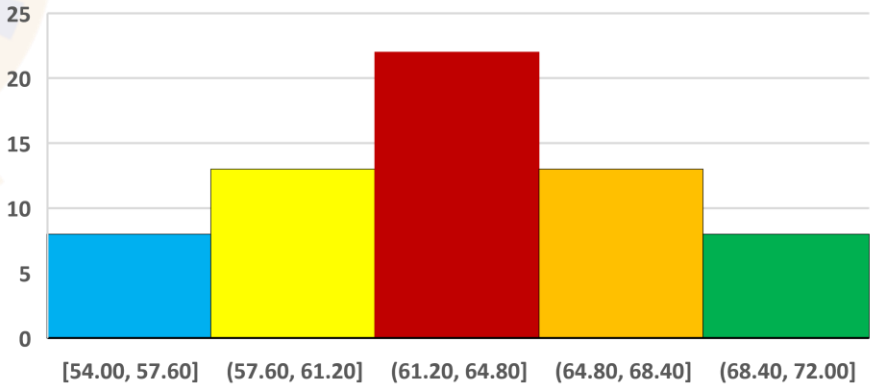
$$\mu_{\overline{X}} = \mu$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- The mean of the sample means = 62.75 and standard deviation = 4.12.



Population Histogram



Histogram of Sample Means

x	y	Mean
54	54	54.00
54	55	54.50
54	59	56.50
54	63	58.50
54	64	59.00
54	68	61.00
54	69	61.50
54	70	62.00
55	54	54.50
55	55	55.00
55	59	57.00
55	63	59.00
55	64	59.50
55	68	61.50
55	69	62.00
55	70	62.50
59	54	56.50
59	55	57.00
59	59	59.00
59	63	61.00
59	64	61.50
59	68	63.50
59	69	64.00
59	70	64.50
63	54	58.50
63	55	59.00
63	59	61.00
63	63	63.00
63	64	63.50
63	68	65.50
63	69	66.00
63	70	66.50
64	54	59.00
64	55	59.50
64	59	61.50
64	63	63.50
64	64	64.00
64	68	66.00
64	69	66.50
64	70	67.00
68	54	61.00
68	55	61.50
68	59	63.50
68	63	65.50
68	64	66.00
68	68	68.00
68	69	68.50
68	70	69.00
69	54	61.50
69	55	62.00
69	59	64.00
69	63	66.00
69	64	66.50
69	68	68.50
69	69	69.00
69	70	69.50
70	54	62.00
70	55	62.50
70	59	64.50
70	63	66.50
70	64	67.00
70	68	69.00
70	69	69.50
70	70	70.00

Samples with Means

Central Limit Theorem: Features

- Central Limit Theorem holds for any type of population distribution for large sample sizes (n >= 30).
- If the population itself is normally distributed then CLT holds even for smaller values of n.
- The expression for standard deviation is true for infinite size of population.

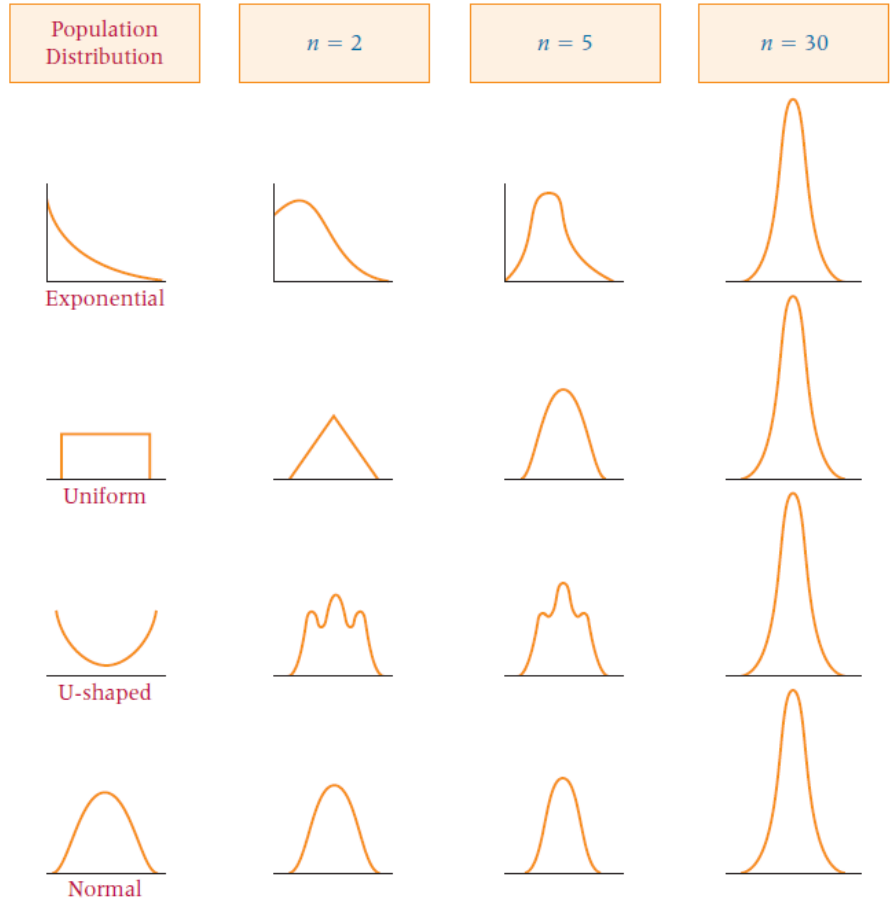
$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- If the population is finite (N) then correction factor is applied to it (normally not used because N will be still be >> n):

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

- If $\overline{x}$ (where $\overline{x} \in \overline{X}$)is the mean of a random sample of size n, taken from a population having the mean (μ) and standard deviation (σ), then the z-transformation for the sample mean will be:

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$



Example

In a large shopping mall the average number of customers/hour is 448, with a standard deviation of 21 customers. This data is collected over several years. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 customers?

Given that N = ∞, μ = 448, σ = 21, n = 49, $\overline{x1} = 441$ and $\overline{x2} = 446$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ so:}$$

$$z1 = \frac{441 - 448}{21/\sqrt{49}} = \frac{-7}{3} = -2.33, \text{Probability} = -0.4901$$

$$z2 = \frac{446 - 448}{21/\sqrt{49}} = \frac{-2}{3} = -0.67, \text{Probability} = -0.2486$$

So, P(441 <= $\bar{X}$ <= 446) = -0.2486 – (-0.4901) = **0.2415 = 24.15%**

	SECOND DECIMAL PLACE IN z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

z-Distribution Table

The t-distribution

- In the Central Limit Theorem (CLT), we reviewed that the means of samples follow a normal distribution for large sample sizes and the theory of standard normal distribution can be applied for some analysis.

$$\mu_{\bar{X}} = \mu, \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, \qquad z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- The issue arises when:
  - Population is known to be normally distributed.
  - It is not feasible to draw several random samples.
  - It is not feasible to draw a large-sized sample.
  - The population standard deviation (σ) and mean (μ) are not known.
  - The estimation of population mean (μ) is to be done or the claim about population mean (μ) is to be ascertained.
- We have reviewed that even for small sample sizes, the mean of samples follow the normal distribution if these samples are drawn from normally distributed population.
- In such situations, **Student's t-distribution** can be used. Student's t-distribution was developed by William S Gosset, a 19th century English statistician. His pen name was **Student**. The t-statistic can be calculated using sample mean ($\bar{x}$) and standard deviation (s) as:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

μ is not known, so how t-statistic will be calculated?

- The assumption while calculating the t-statistic that population is to be normally distributed is not so severe. The results are fairly close for non-normally distributed populations also unless they are not too skewed.

t-distribution: Characteristics

- Like normal distribution, t-distribution is also bell shaped and symmetrical. But they are **flatter** in the middle and have **more area in the tails**.
- The mean of t-distribution is also zero (like standard normal distribution) but variance depends on (n-1) which is called the **Degree of Freedom (DF)** and denoted by the Greek alphabet **v (nu).**
- The variance of t-distribution approaches 1 as $n \to \infty$ or in another words the t-distribution approaches to standard normal distribution when $n \to \infty$.
- The term **DF (v)** refers to the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation. Here there are n observations in the sample and one parameter (the mean of population) is being estimated. So v = n-1.
- The value $t_{\alpha;v}$ on the x-axis for which the area (tail area) under the t-curve with v DF to right is α is called the **t-critical value**.
- In other words, the probability that some t_v will be greater than $t_{\alpha;v}$ is α for v DF. The Student's t-distribution table provides the values of $t_{\alpha;v}$ for few α and v.
- In the table for a given v, when α decreases, the value of $t_{\alpha;v}$ increases. That means moving away from the mean towards right.
- In the table for a given α, when v increases, the value of $t_{\alpha;v}$ gradually decreases.
- For v = ∞, the $t_{\alpha;v}$ values are the corresponding z values for (1 - α) in the standard normal distribution table. E.g. $t_{\alpha;v}$ = 1.960 for v = ∞ and α = 0.025. The z-distribution table value for z = 1.960 is 0.9750 that is (1 - α).

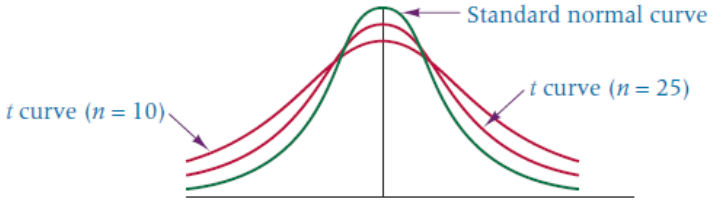
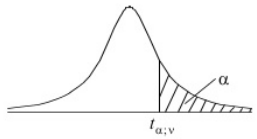


Table of the Student's *t*-distribution



v \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Student's t-distribution Table

Example-1

No inference is made; only table calculations are done in this example!

An electric equipment manufacturer claims that on 20% overload its circuit breakers can sustain for 12.40 minutes on an average.

To test the claim a QA engineer selected a random sample of 20 circuit breakers and put them under 20% overload. The mean time of the sustenance was 10.63 minutes and standard deviation was 2.48 minutes.

Calculate the t-statistic. What is the range of α expected from this data?

Given that:

$\mu = 12.40$

$n = 20$, $v = 20 - 1 = 19$, $\bar{x} = 10.63$ and $s = 2.48$

So

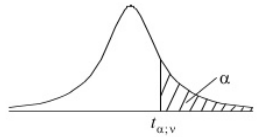
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{10.63 - 12.40}{\frac{2.48}{\sqrt{20}}}$$

$$= -3.19$$

From the Student's t-distribution table, we can observe that the t value -3.19 will occur in the range of **$0.005 > \alpha > 0.001$** for v = 19 for the left half of the t-curve.

Table of the Student's *t*-distribution



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Student's t-distribution Table

Example-2

No inference is made; only table calculations are done in this example!

The population mean of the heights of five-year old boys is claimed to be 100 cm. A teacher measures the height of her 25 students, obtaining a mean height of 105 cm and standard deviation 18 cm.

Calculate the t-statistic. What is the range of α expected from this data?

Given that:

$\mu = 100$

$n = 25$, $v = 25 - 1 = 24$, $\bar{x} = 105$ and $s = 18$

So

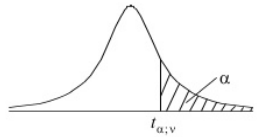
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{105 - 100}{\frac{18}{\sqrt{25}}}$$

$$= 1.39$$

From the Student's t-distribution table, we can observe that the t value 1.39 will occur in the range of **0.1 > α > 0.05** for v = 24 for the right half of the t-curve.

Table of the Student's *t*-distribution



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Student's t-distribution Table

Exercise

1. The tensile strength of a new composite material can be modelled as a normal distribution. A random sample of size 25 has mean of 45.3 and standard deviation as 7.9. In case the population mean is 40.5, calculate the t-value and the range of α.

   (Answer t = 3.038, 0.005 > α > 0.001)

2. The following are the times between 6 calls for an ambulance of a specific hospital and the patient's arrival at that hospital: 27, 15, 20, 32, 18 and 26 minutes. The hospital's claim is average 20 minutes. Assume duration is normally distributed, find out the value of t and the range of α.

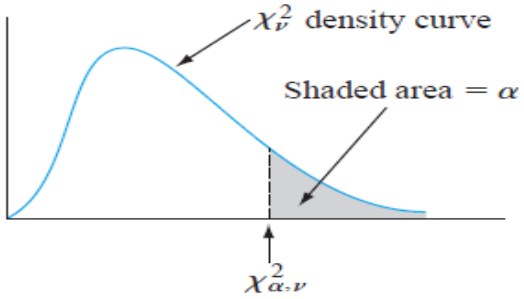
   (Answer t = 1.15, α > 0.1)

Sampling Distribution of the Variance

- To win a new contract, a manufacturer who supplies few parts needs to show the continuous reduction in the variation of the supplied parts.
- To measure the altitude, the airplanes are equipped with altimeter. The variation among the altimeter readings should be minimum for a given altitude.
- In the above situations, the distribution of the mean will not be helpful. The distribution of variance is required.
- The relationship of the sample variance to the population variance is captured by the **chi-square distribution**. The ratio of the sample variance (s^2) multiplied by (n – 1) to the population variance (σ^2) is approximately chi-square distributed. The samples are assumed to be drawn from the normal population.
- s^2 cannot be negative and this sampling distribution is related to Gamma Distribution with $\alpha = v/2$ and $\beta = 2$. Where v is the Degree of Freedom (DF) and is equal to (n - 1) where n is the sample size.
- Chi-square is denoted by the square of Greek alphabet chi (χ) and is defined as:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}$$

- The value $\chi^2_{\alpha;v}$ on the x-axis for which the area under the χ^2 curve with v DF to the right is α is called the **χ^2-critical value**.



Example

No inference is made; only table calculations are done in this example!

An optical firm purchases glasses to be ground for lenses and it is known from the past experiences that the variation in the refractive index of this kind of glass is 1.26 x 10^{-4}. The firm rejects a shipment if the sample variance of 20 pieces selected at random exceeds 2.0 x 10^{-4}. Calculate the value of χ^2 and the range of α from it.

Given that:

$\sigma^2 = 1.26 \times 10^{-4}$

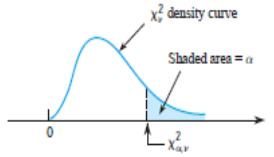
$n = 20$ and $s^2 = 2.0 \times 10^{-4}$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$$= \frac{(20-1).\ 2.0 \times 10{-}4}{1.26 \times 10{-}4}$$

$= 30.16$

For given v = 20 - 1 = 19, the range is $0.05 > \alpha > 0.025$.



	α									
ν	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

Chi-Square Distribution Table

Exercise

1. Hard disks of computers must spin evenly and one departure from the acceptable level is called a pitch. Samples are regularly taken to ensure the quality through the measurement of pitches. From the data over a period shows that pitches are normally distributed with variance 0.065. A sample of 10 is collected each week. The process will be declared out of control if the sample variance exceeds 0.122. Calculate the value of χ^2 and the range of α from it.

   (Answer: $\chi^2 = 16.89$, $0.10 > \alpha > 0.05$)

2. A random sample of 10 observations is taken from a normal population having the variance as 42.5. Find the approximate probability of obtaining a sample standard deviation between 3.14 and 8.94.

   (Answer: 0.95)

3. From the data provided in the last example of Optical Firm, verify the calculations using the Gamma Distribution function.

4. Using the Gamma Distribution function, find the probability that the variance of a random sample of size 5 from a normal population with standard deviation 12 will exceed 180.

   (Answer = 0.2873)

Parameter Estimation Techniques

Interval Estimation of Population Mean

Using z-statistic

- First we will review the methods to calculate and report an entire interval (range) of reasonable values for the parameter. This is called an **Interval Estimate.** It is calculated by setting up a **Confidence Level.**
- Let us say it is needed to estimate the interval of the population mean (μ) with 95% Confidence Level. What does it imply?
- It implies, that 95% of all samples would give an interval that includes μ and only 5% of all samples would yield an erroneous interval for the population mean.
- In other words, if we have to estimate the interval of the population mean with some Confidence Level (1 - α)% and the sampling and interval estimate is done Y times, then Y.(1 - α)% times the calculated intervals will contain the population mean.
- In the Central Limit Theorem, we reviewed that the means of the samples follow a normal distribution. The area under the normal curve for the calculated interval will be (1 - α)% and under both the tails it will be (α/2)% each.
- From the Central Limit Theorem and corresponding z-statistic for the significantly very large population and n sample size we know that:

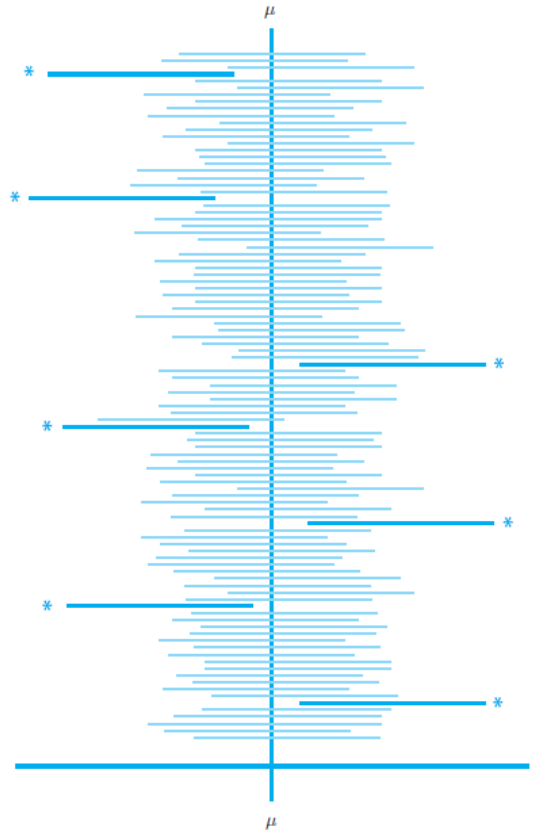
$$-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

- The above expression can be arranged for the population mean as:

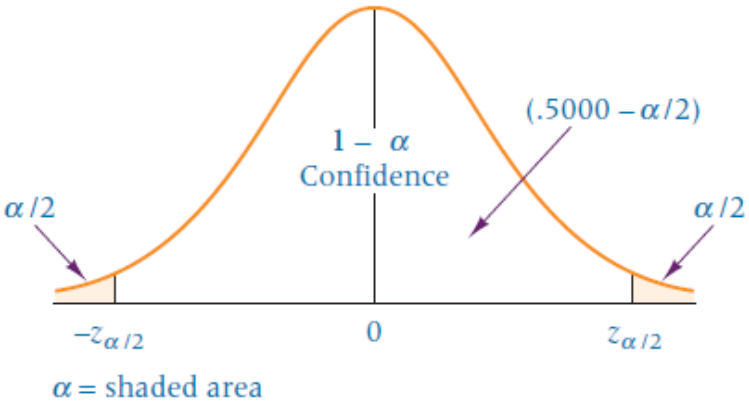
$$\mu = \bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Or the interval estimate of the population mean can be written as:

$$\bar{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$



Asterisk (*) shows: μ is not in the Interval



Example-1

For a random sample of 36 items and a sample mean of 211, calculate a 95% confidence interval for population mean if the population standard deviation is 23.

Given that:

n = 36, $\overline{x}$ = 211

σ = 23

(1-α) = 0.95, so α = 0.05 and α/2 = 0.025

$z_{\alpha/2}$ = the value of z that gives the normal curve area

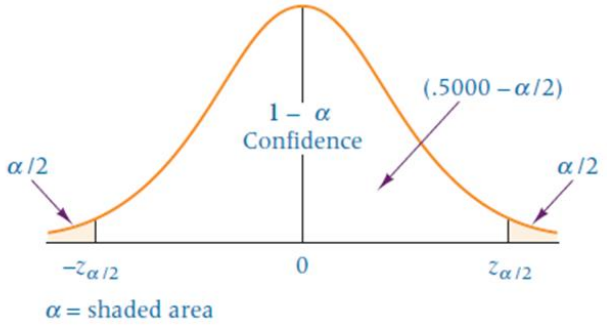
after the mean that is (0.5000 - α/2) = (0.5000 - 0.025) = 0.4750

= 1.96

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$211 - 1.96.\left(\frac{23}{\sqrt{36}}\right) \leq \mu \leq 211 + 1.96.\left(\frac{23}{\sqrt{36}}\right)$$

$203.49 \leq \mu \leq 218.51$



*Explore **Inverse Normal** Feature in the Calculator*

SECOND DECIMAL PLACE IN z										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

z-Distribution Table

Example-2

A survey was conducted for US multinational companies that do business with the companies in India. One of the questions in the survey was: around how many years has your company been working with the companies in India? A random sample of 44 responses to this question gave a mean of 10.50 years. Suppose the population standard deviation is 7.70 years. Construct a 90% confidence interval for the mean number of years that a company has been working with Indian companies India for the population.

Given that:

n = 44, $\overline{x}$ = 10.50

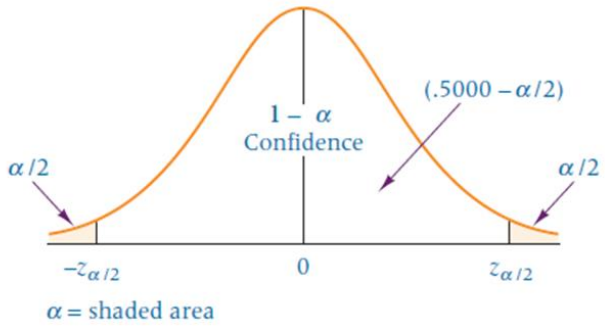
σ = 7.70

(1-α) = 0.90, so α = 0.10 and α/2 = 0.05

$z_{\alpha/2}$ = the value of z that gives the normal curve area

after the mean that is (0.5000 - α/2) = (0.5000 - 0.05) = 0.4500

= (1.64 + 1.65) / 2 = 1.645



*Explore **Inverse Normal** Feature in the Calculator*

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$10.50 - 1.645.\left(\frac{7.70}{\sqrt{44}}\right) \leq \mu \leq 10.50 + 1.645.\left(\frac{7.70}{\sqrt{44}}\right)$$

$8.59 \leq \mu \leq 12.41$

SECOND DECIMAL PLACE IN z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

z-Distribution Table

Exercise

1. Construct an 80% confidence interval for the population mean given that: $\overline{x} = 56.7$, $\sigma = 12.1$, N = 500 and n = 47.

   (Answer: $54.4381 \leq \mu \leq 58.9619$)

2. A random sample of size 39 is taken from a population of 200 members. The sample mean is 66 and the population standard deviation is 11. Construct a 96% confidence interval to estimate the population mean?

   (Answer: $62.3825 \leq \mu \leq 69.6175$)

3. A community health association is interested in estimating the average number of maternity days women stay in the local hospital. A random sample is taken of 36 women who had babies in the hospital during the past year. The numbers below show the maternity days each woman in the sample was in the hospital. Use this data and a population standard deviation of 1.17 to construct a 98% confidence interval to estimate the average maternity stay in the hospital for all women who have babies in this hospital: 3, 3, 4, 3, 2, 5, 3, 1, 4, 3, 4, 2, 3, 5, 3, 2, 4, 3, 2, 4, 1, 6, 3, 4, 3, 3, 5, 2, 3, 2, 3, 5, 4, 3, 5, 4.

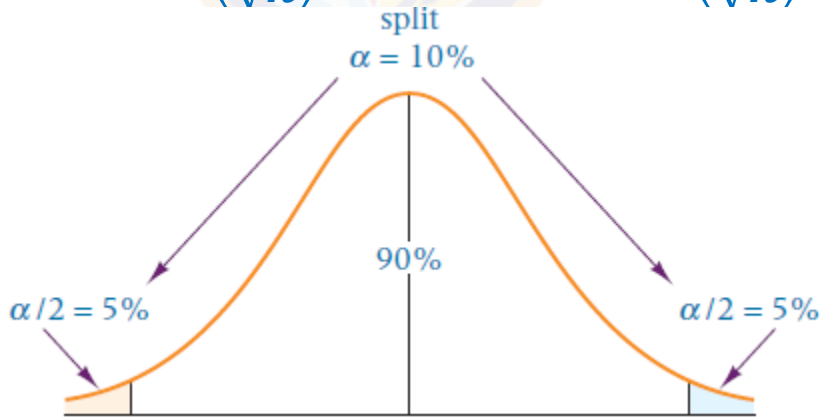
   (Answer: $2.85 \leq \mu \leq 3.76$)

Interval Estimation of Population Mean

Using t-statistic

- We have reviewed the situations when t-statistic can be used to observe the distribution of the population mean (μ) using sample mean ($\bar{x}$) and sample standard deviation (s).
- Correspondingly, the t-distribution can be utilized to estimate the interval of population mean for a given confidence level.
- The method is similar to what is used in estimating the interval of population mean using z-statistic when population standard deviation (σ) was known, except in this method the sample standard deviation (s) will be used.
- Using the t-statistic for v Degree of Freedom (DF) and (1-α) confidence interval:

$$\bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$



t-distribution with 90% Confidence Level

Example

If a random sample of 41 items produces $\bar{x}$ = 128.4 and s = 20.6, what is the 98% confidence interval for μ? Assume x is normally distributed for the population

Given that:

n = 41 so v = (n – 1) = 40

$\bar{x}$ = 128.4 and s = 20.6

(1 – α) = 0.98, so α = 0.02 and α/2 = 0.01

From the t-distribution table:

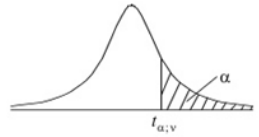
$t_{0.01} = 2.423$

$$\bar{x} - t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$128.4 - 2.423.\left(\frac{20.6}{\sqrt{41}}\right) \leq \mu \leq 128.4 + 2.423.\left(\frac{20.6}{\sqrt{41}}\right)$$

$$\mathbf{120.60 \leq \mu \leq 136.20}$$

Table of the Student's *t*-distribution



ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Student's t-distribution Table

Interval Estimation of Population Variance

Using $\chi 2$ distribution

- The relationship of the sample variance to the population variance is captured by the **chi-square distribution**. The ratio of the sample variance (s^2) multiplied by (n – 1) to the population variance (σ^2) is approximately chi-square distributed. The samples are assumed to be drawn from the normal population.
- s^2 cannot be negative and this sampling distribution is related to Gamma Distribution with $\alpha = v/2$ and $\beta = 2$. Where v is the Degree of Freedom (DF) and is equal to (n - 1) where n is the sample size.
- Chi-square is denoted by the square of Greek alphabet chi (χ^2) and is defined as:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

- The above expression can be re-written for the population variance as:

$$\sigma^2 = \frac{(n-1)s^2}{\chi^2}$$

- Therefore, interval estimate for the population variance for v DF and (1 - α) confidence level can be written as:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}$$

- The meaning of population variance interval estimate is discussed in the next slide.

Population Variance: Interval Estimation

- From the chi-square distribution, we reviewed that the variance and chi-square value are inversely proportional.

$$\sigma^2 = \frac{(n-1)s^2}{\chi^2}$$

- So as the value of χ^2 progresses towards the right from the origin, the corresponding value of σ^2 decreases.

- Chi-square distribution table is arranged in such a way that it tells the value of χ^2 for v DF and area under the curve (α) to the right of χ^2.

- When it is required to estimate the population variance interval for the confidence level (1 - α), the following areas (shaded) are excluded from the distribution curve:

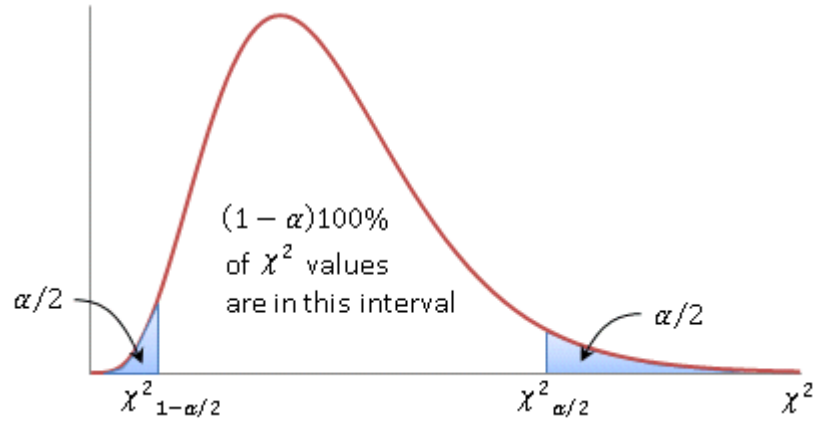
  i. α/2 to the right of $\chi^2_{\alpha/2}$ — *$\chi^2_{\alpha/2}$ can be read from the table*

  ii. α/2 to the left of $\chi^2_{1-(\alpha/2)}$ — *how will $\chi^2_{1-(\alpha/2)}$ be read; area α/2 is on the left?*

- Notice that for the point (ii) above the value of $\chi^2_{1-(\alpha/2)}$ will be equal to χ^2 value for which the area on the right is 1 – (α/2).

- Therefore, interval estimate for the population variance for v DF and (1 - α) confidence level can be written as:

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}$$



Example-1

A company manufactures aluminium cylinders and when a random sample of 8 cylinders is taken, the variance of the diameters comes out as 0.0022125. Estimate the interval for the diameter variance for the cylinders that are produced by the company for the 90% confidence level.

Given that:

$s^2 = 0.0022125$

n = 8 so v = (n - 1) = 7

(1 – α) = 0.90 so α = 0.10 and α/2 = 0.05, 1 – (α/2) = 0.95

From the table:

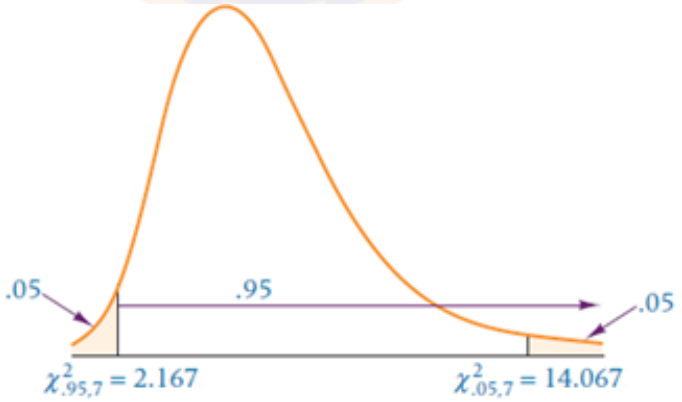
$\chi^2_{\alpha/2} = \chi^2_{0.05} = 14.067$

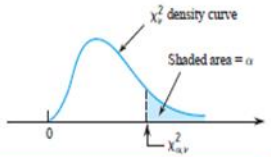
$\chi^2_{1-(\alpha/2)} = \chi^2_{0.95} = 2.167$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}$$

$$\frac{7 \times 0.0022125}{14.067} \leq \sigma^2 \leq \frac{7 \times 0.0022125}{2.167}$$

$0.001101 \leq \sigma^2 \leq 0.007147$





	α									
ν	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

Chi-Square Distribution Table

Example-2

Refractive indices of 20 random pieces of glass selected from a large shipment that is purchased from an optical firm have a variance as 1.20×10^{-4}. Construct a 95% confidence interval for the population variance.

Given that:

$s^2 = 1.20 \times 10^{-4}$

n = 20 so v = (20 - 1) = 19

(1 – α) = 0.95 so α = 0.05 and α/2 = 0.025, 1 – (α/2) = 0.975

From the table:

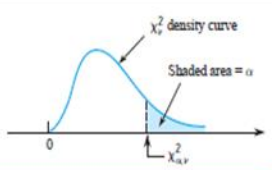
$\chi^2_{\alpha/2} = \chi^2_{0.025} = 32.852$

$\chi^2_{1-(\alpha/2)} = \chi^2_{0.975} = 8.906$

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}$$

$$\frac{19 \times 1.20 \times 10{-}4}{32.852} \leq \sigma^2 \leq \frac{19 \times 1.20 \times 10{-}4}{8.906}$$

$0.000069402 \leq \sigma^2 \leq 0.0002560$



	α									
ν	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.843	5.025	6.637	7.882
2	0.010	0.020	0.051	0.103	0.211	4.605	5.992	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.344	12.837
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.832	15.085	16.748
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.440	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.012	18.474	20.276
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.534	20.090	21.954
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.022	21.665	23.587
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.724	26.755
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.735	27.687	29.817
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.600	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.577	32.799
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.407	7.564	8.682	10.085	24.769	27.587	30.190	33.408	35.716
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.843	7.632	8.906	10.117	11.651	27.203	30.143	32.852	36.190	38.580
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.033	8.897	10.283	11.591	13.240	29.615	32.670	35.478	38.930	41.399
22	8.643	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289	42.796
23	9.260	10.195	11.688	13.090	14.848	32.007	35.172	38.075	41.637	44.179
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.558
25	10.519	11.523	13.120	14.611	16.473	34.381	37.652	40.646	44.313	46.925
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.807	12.878	14.573	16.151	18.114	36.741	40.113	43.194	46.962	49.642
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.120	14.256	16.147	17.708	19.768	39.087	42.557	45.772	49.586	52.333
30	13.787	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.457	15.655	17.538	19.280	21.433	41.422	44.985	48.231	52.190	55.000
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.814	17.073	19.046	20.866	23.110	43.745	47.400	50.724	54.774	57.646
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.191	18.508	20.569	22.465	24.796	46.059	49.802	53.203	57.340	60.272
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.584	19.960	22.105	24.075	26.492	48.363	52.192	55.667	59.891	62.880
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.994	21.425	23.654	25.695	28.196	50.660	54.572	58.119	62.426	65.473
40	20.706	22.164	24.433	26.509	29.050	51.805	55.758	59.342	63.691	66.766

Chi-Square Distribution Table

Exercise

1. A random sample of 14 families produced the monthly household incomes that is shown below in the table. Construct a 95% confidence interval for the population variance assuming household income is normally distributed.

Monthly Income for 14 Sample Families (₹)						
37,500	44,800	33,500	36,900	42,300	32,400	28,000
41,200	46,600	38,500	40,200	32,000	35,500	36,800

(Answer: $14084380.17 \leq \sigma^2 \leq 69550238.25$)

2. During the Covid-19 pandemic it was reported that the total weekly work hours are reducing primarily because of several distractions during work from home. To do some analysis, a random sample of 20 employees was collected from different IT companies and it was found that the distribution of weekly hours they work has the standard deviation as 4.3 hours. Assume work hours worked per week are normally distributed in the population. Use this sample information to develop a 98% confidence interval for the population variance of the number of work hours worked per week for an employee.

(Answer: $9.71 \leq \sigma^2 \leq 46.03$)

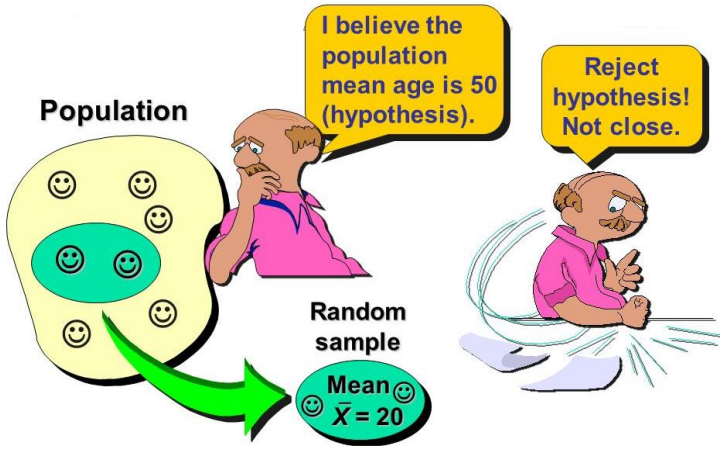
Hypothesis Testing

Hypothesis Testing

Plural: Hypotheses

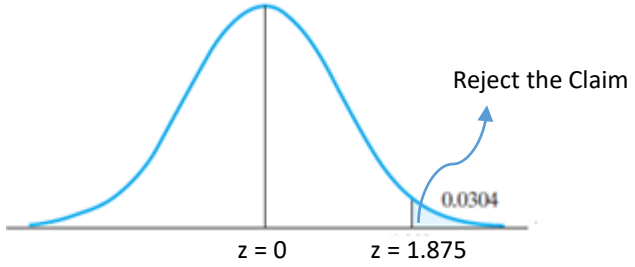
There are many problems in which, rather than estimate the value or interval of a parameter, we must decide whether a **statement** concerning a parameter is True or False. Statistically speaking we test a hypothesis (claim) about a parameter. Few example statements are:

- Average time spent by women on social media is more than men.
- Children who drink health drinks (Horlicks etc.) grow taller.
- Average salary of people with PhD in engineering is more than the salary of people with just the masters degrees.



Recap Example-1: A paint manufacturer claims that the mean drying time of their new paint is 20 minutes with standard deviation of 2.4 minutes. A researcher decided to reject the claim if a sample of 36 new paint boxes yield a mean that exceeds 20.75 minutes of drying time.

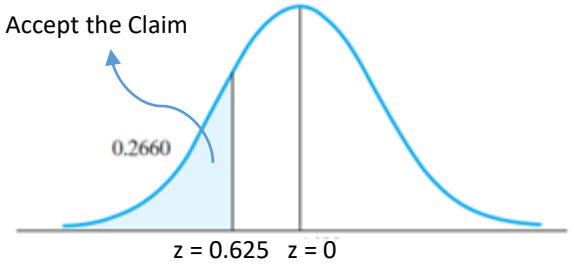
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{20.75 - 20}{2.4/\sqrt{36}} = 1.875$$



From the z-distribution table the shaded area = 0.0304, which is the **probability of rejecting the claim**.

Recap Example-2: In the above example, if the claim were 21 minutes:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{20.75 - 21}{2.4/\sqrt{36}} = -0.625$$



From the z-distribution table the shaded area = 0.2660, which is the **probability of accepting the claim.**

Is that all? Shouldn't there be a criteria to reject or accept a claim like probability more than or less than a certain threshold?

Steps to Set up a Hypothesis Test

1. Describe the Hypothesis (claim) in words in the form of a statement. E.g. average time spent by women on social media is more than men.
2. Based on the claim:
   - Define **Null (H_0)** and **Alternative Hypothesis (H_A or H_1)**.
   - In general null hypothesis means that there is no relationship between the two variables under consideration. E.g. there is no relationship between the gender and the time spent on social media.
   - Initially it is believed that the null hypothesis is true.
3. Identify the statistic for testing the validity of the Null Hypothesis. E.g. if it is about the mean time spent on social media, then the statistic is z-statistic or t-statistic.
4. Decide the criteria to reject or accept the null hypothesis. This is called the **significance value** and it is denoted by **α.** This is also called **Type-I Error**, that will be elaborated in the subsequent slides.
5. Calculate the probability value (p-value) which is the *conditional probability* observing the statistic given null hypothesis is true.

$$\text{p} - \text{value} = \text{P}\left(\text{observing the test statistic value} \mid \text{null hypothesis is true}\right)$$

6. p-value is the probability value for the calculated statistic. E.g. you calculate the probability for the calculated z-statistic.
7. Null hypothesis is **rejected when p-value is less than significance value (α)**, otherwise accepted.
8. α is the threshold conditional probability. The statistic value for α is called the **critical value** (e.g. corresponding z-statistic for α).
9. Alternative hypothesis is complement of the null hypothesis which the claimant believes and tries to reject the null hypothesis.
10. In summary, we start with the assumption that null hypothesis is true and try to retain it unless there is an evidence against it.

Examples	
Hypothesis	**Null and Alternative Hypothesis**
Average annual salary is **different** for IT and Data Science (DS) Engineers.	$H_0: \mu_{IT} = \mu_{DS}$ $H_A: \mu_{IT} \neq \mu_{DS}$
Mean height of the children who drink health drinks (HD) is **more** than those who do not drink health drinks (NHD).	$H_0: \mu_{HD} <= \mu_{NHD}$ $H_A: \mu_{HD} > \mu_{NHD}$

One Tailed and Two Tailed Tests

Area under the tail (α), the significance value, shown is the rejection region.

Reject or retain decision will depend on the direction of the deviation of the estimated static from the hypothesis value (population parameter) that is being ascertained.

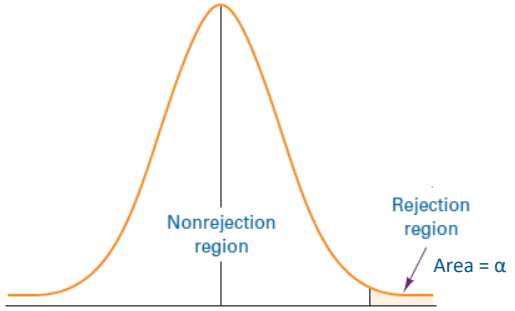
Scenario-1:

Hypothesis: In US, average annual salary of Machine Learning (ML) experts is at least $100,000.

Null Hypothesis (H_0): $\mu_{ML} < 100{,}000$

Alternative Hypothesis (H_A): $\mu_{ML} >= 100{,}000$

In this case, alternative hypothesis is for more than the hypothesis value, so rejection region will be on the right side of the one tailed test.



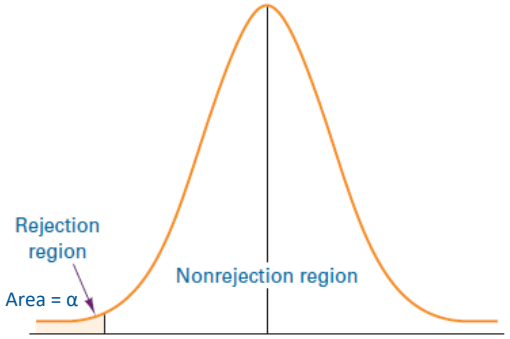
Scenario-2:

Hypothesis: Average waiting time (w) at the Bangalore airport security check is less than 30 minutes.

Null Hypothesis (H_0): $\mu_w >= 30$

Alternative Hypothesis (H_A): $\mu_w < 30$

In this case, alternative hypothesis is for less than the hypothesis value, so rejection region will be on the left side of the one tailed test.



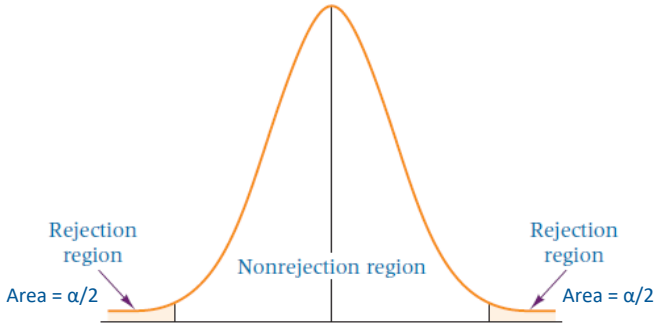
Scenario-3:

Hypothesis: Average annual salaries of Engineering (E) and Business Management (M) graduates are different.

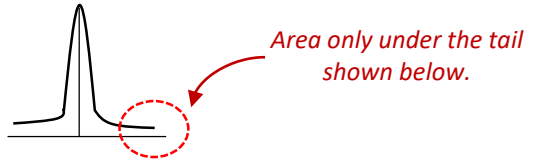
Null Hypothesis (H_0): $\mu_E = \mu_M$

Alternative Hypothesis (H_A): $\mu_E \neq \mu_M$

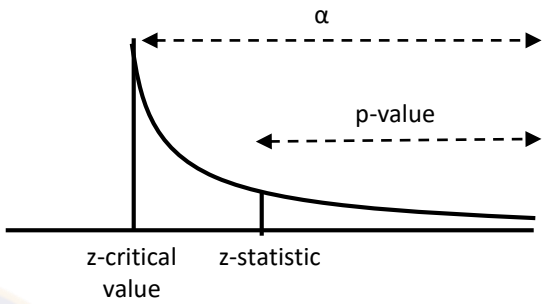
In this case, alternative hypothesis could be less than or more than the hypothesis value, so rejection region will be on the both the sides of the two tailed test. (observe α/2 regions).



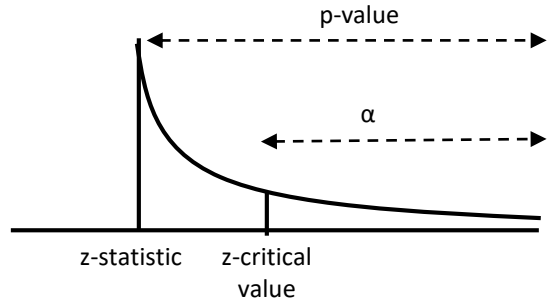
Consolidated View



Type of Test	Condition	Probability	Decision
Right-Tail Test	z-statistic > Critical Value	p-value < α	Reject H_0
	z-statistic <= Critical Value	p-value >= α	Retain H_0

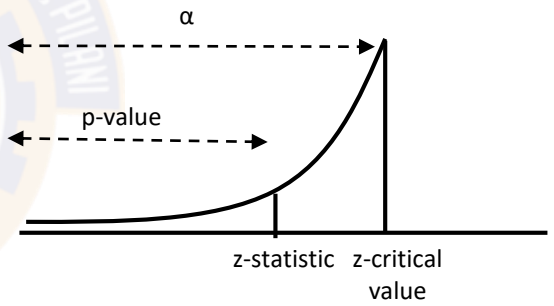


Right-Tail H_0 Rejection

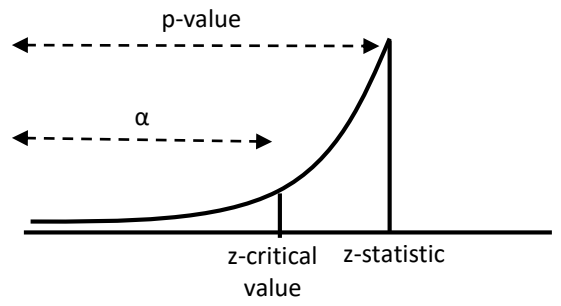


Right-Tail H_0 Retention

Left-Tail Test	z-statistic < Critical Value	p-value < α	Reject H_0
	z-statistic >= Critical Value	p-value >= α	Retain H_0

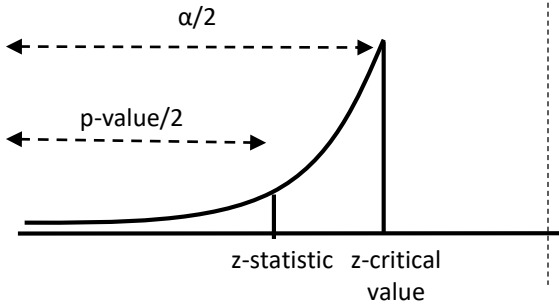


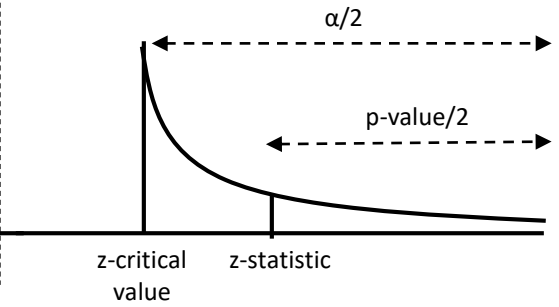
Left-Tail H_0 Rejection



Left-Tail H_0 Retention

Two-Tail Test	|z-statistic| > |Critical Value|	p-value < α	Reject H_0
	|z-statistic| <= |Critical Value|	p-value >= α	Retain H_0





Two-Tail H_0 Rejection

The similar approach applicable for t-statistic.

Example-1

A study claimed that average monthly disposable income of the families in the IT hubs of India is greater than ₹42,000 with a standard deviation of ₹32,000. The random sample of 40,000 families provided a mean of ₹42,500. Conduct an appropriate hypothesis test using the confidence level of 95%.

Given that:

$\mu = 42000$, $\sigma = 32000$, $n = 40000$ and $\bar{x} = 42500$

$(1-\alpha) = 0.95$, so significance value (α) = **0.05**

Critical value for α = **1.645**

Null Hypothesis (H_0): $\mu <= 42000$

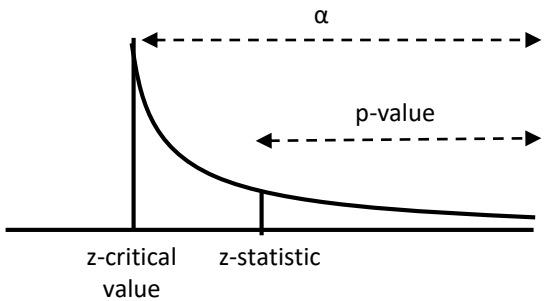
Alternative Hypothesis (H_A): $\mu > 42000$

Alternative hypothesis is for more than the hypothesis value, so rejection region will be on the right side of the one tailed test.

$$z - \text{statistic} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{42500 - 42000}{\frac{32000}{\sqrt{40000}}} = \mathbf{3.125}$$

p-value for z-statistic = **0.0009**

Since, p-value < significance value, so **null hypothesis is rejected.**



Right-Tail H_0 Rejection

Example-2

Passport office claims that passport applications are processed within 30 days on an average when all the papers are complete with standard deviation of 12.5 days. To verify the claim, data of 40 applications are examined and the sample mean was found as 27.05 days. Conduct a hypothesis test for significance value 0.05.

Given that:

$\mu = 30$, $\sigma = 12.5$, $n = 40$ and $\bar{x} = 27.05$

Significance value (α) = **0.05**

Critical value for α = 1.645

Null Hypothesis (H_0): $\mu > 30$

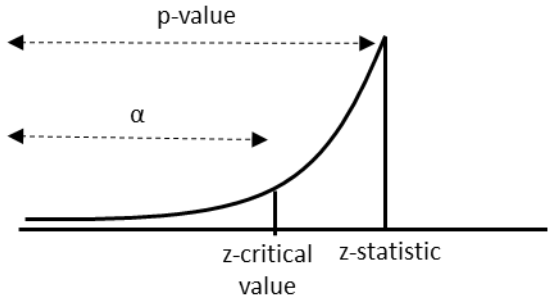
Alternative Hypothesis (H_A): $\mu <= 30$

Alternative hypothesis is for less than the hypothesis value, so rejection region will be on the left side of the one tailed test and the critical value will be with the negative sign (**-1.645**).

$$z - \text{statistic} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{27.05 - 30}{\frac{12.5}{\sqrt{40}}} = -1.49$$

p-value for z-statistic = **0.0681**

Since, p-value > significance value, so **null hypothesis is retained.**



Left-Tail H_0 Retention

Example-3

Based on a research conducted the average IQ is 82 with standard deviation 11.03 for the citizen of a country. When a random sample of 100 citizen was tested, the average IQ was found to be 84. Using the significance value of 0.05, conduct a hypothesis test.

Given that:

$\mu = 82$, $\sigma = 11.03$, $n = 100$ and $\bar{x} = 84$

Significance value (α) = 0.05

Null Hypothesis (H_0): $\mu \neq 82$

Alternative Hypothesis (H_A): $\mu = 82$

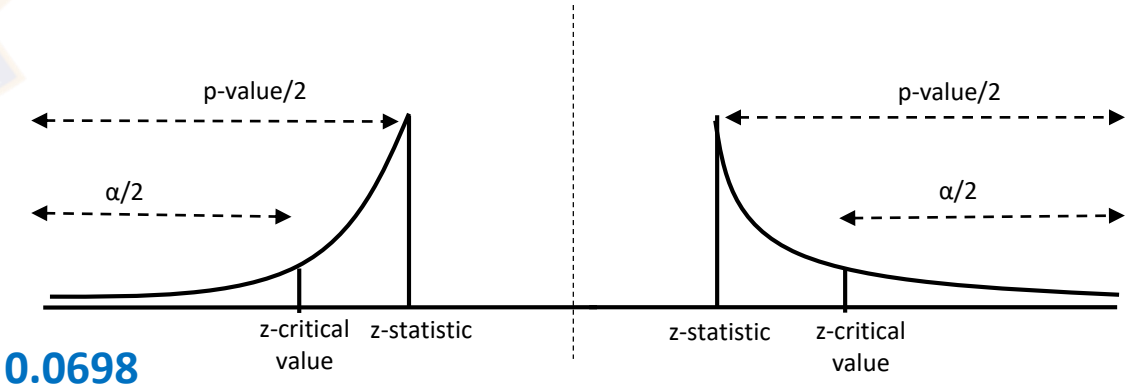
The deviation could be on the either side, so rejection region will be on both the sides of the two tailed test with significance value ($\alpha/2$) = **0.025**

Critical value for $\alpha/2$ = **1.96**

$$z - \text{statistic} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{84 - 82}{\frac{11.03}{\sqrt{100}}} = \mathbf{1.8132}$$

p-value for z-statistic = 0.0349 and total p-value (both sides) = **0.0698**

Since, p-value > significance value, so **null hypothesis is retained.**



Two-Tail H_0 Retention

Example-4

An agricultural researcher believes the average farm size has now increased from the 1990 mean figure of 4.71 acres. To test this notion, he randomly sampled 23 farms across India and ascertained the size of each farm from the county records. The data he gathered follows a mean of 5 acres with standard deviation of 0.47. Use a 5% level of significance to test the hypothesis.

Given that:

$\mu = 4.71$, $s = 0.47$, $n = 23$, $v = 23 - 1 = 22$ and $\bar{x} = 5$

Significance value (α) = 0.05

This problem involves t-statistic because σ is unknown.

Null Hypothesis (H_0): $\mu <= 4.71$

Alternative Hypothesis (H_A): $\mu > 4.71$

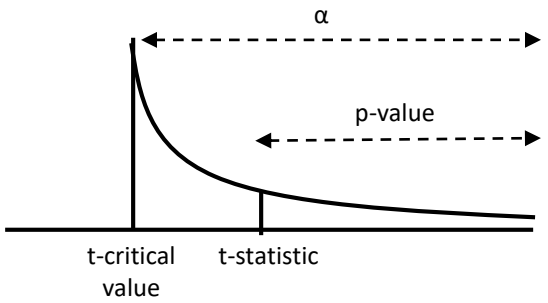
Alternative hypothesis is more than the hypothesis value, so rejection region will be on the right side of the one-tailed test.

Critical value for $\alpha = 0.05$ and v (= 22) = **1.717**

$$t - \text{statistic} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{5 - 4.71}{\frac{0.47}{\sqrt{23}}} = \mathbf{2.96}$$

Since t-statistic > critical value (or, p-value < significance value)

so **null hypothesis is rejected.**



Right-Tail H_0 Rejection

Example-5

A heavy machine part needs to be 25 pounds average in weight. The factory supervisor has doubts that the new parts that are being built are no more of 25 pounds and their manufacturing process has gone out of control. To verify the doubt, 20 parts are sampled which are found to have a mean weight as 25.51 pounds with standard deviation as 2.19 pounds. Using significance value of 0.05, conduct a hypothesis test.

Given that:

$\mu = 25$, $s = 2.19$, $n = 20$, $v = 20 - 1 = 19$ and $\bar{x} = 25.51$

Significance value (α) = 0.05

This problem involves t-statistic because σ is unknown.

Null Hypothesis (H_0): $\mu = 25$

Alternative Hypothesis (H_A): $\mu \neq 25$

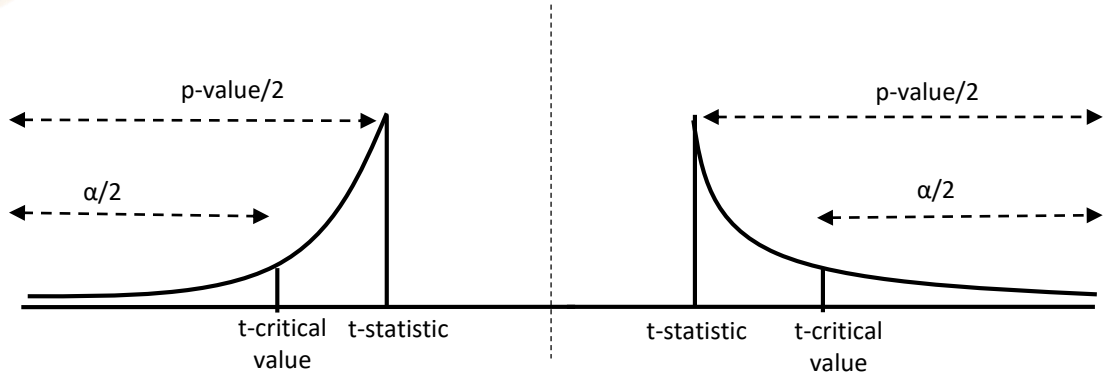
Alternative hypothesis could be less than or more than the hypothesis value, so rejection region will be on both the sides of the two tailed test with significance value ($\alpha/2$) = **0.025**

Critical value for $\alpha/2 = 0.025$ and v (= 19) = **2.093**

$$t - \text{statistic} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{25.51 - 25}{\frac{2.19}{\sqrt{20}}} = \mathbf{1.0415}$$

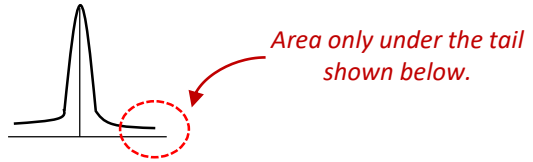
Since t-statistic < critical value (or, p-value > significance value)

so **null hypothesis is retained.**



Two-Tail H_0 Retention

Recap: How Hypothesis Formulation Works?

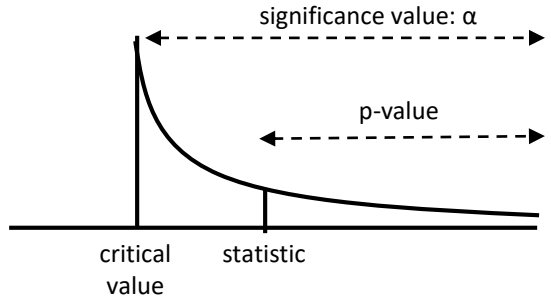


- We have reviewed Hypothesis Testing that involves z-statistic and t-statistic.
- It is a good point, that we take a step back and recap how does the formulation of Null and Alternative Hypothesis work in the context of **significance value (α).**
- Let us say for some problem, we have formed the hypotheses as:

  Null Hypothesis (H_0): $\mu <= 30$

  Alternative Hypothesis (H_A): $\mu > 30$ *that is being claimed*

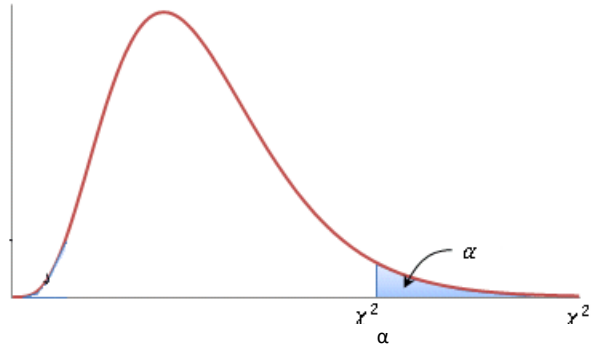
- The Null Hypothesis will be rejected (or alternative hypothesis will be accepted) if our evidence from the sample shows that mean > 30 with respect to the significance value (α).
- We are talking about "*more than*" relationship so it means it is a right-tail test.
- Significance value (α) is area under the curve on the right side. Its corresponding x-axis value is called the **critical value**.
- From the sample data we calculate the **statistic**. It is also a point on the x-axis.
- For the calculated statistic, we find out the area under the curve on the right side (for this problem). This area is called the **p-value**.
- If *p-value < significance value (α)*, we accept the Alternative Hypothesis (or reject the Null Hypothesis).
- Why? Notice that for the right-tail test, if ***p-value < α*** it also means that ***statistic > critical value***. That proves that the alternative hypothesis (claim) is correct.
- In general any one pair either (p-value, α) or (statistic, critical value) can be used to test the hypothesis.

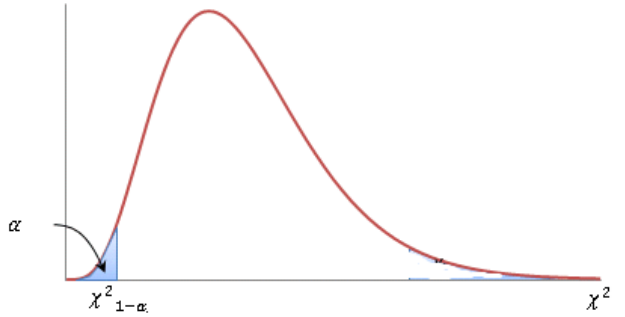


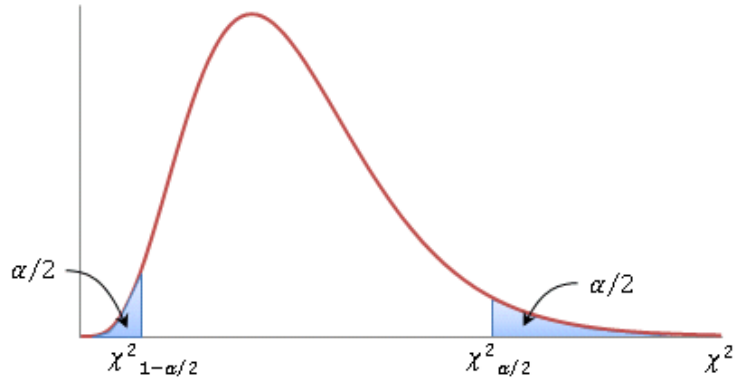
Right-Tail H_0 Rejection

Hypothesis Testing for Variance

Type of Test	Condition	Probability	Decision
Right-Tail Test	χ^2-statistic > Critical Value (χ^2_{α})	p-value < α	Reject H_0
	χ^2-statistic <= Critical Value (χ^2_{α})	p-value >= α	Retain H_0
Left-Tail Test	χ^2-statistic < Critical Value ($\chi^2_{1-\alpha}$)	p-value < α	Reject H_0
	χ^2-statistic >= Critical Value ($\chi^2_{1-\alpha}$)	p-value >= α	Retain H_0
Two-Tail Test	χ^2-statistic < Critical Value ($\chi^2_{1-\alpha/2}$) **OR** χ^2-statistic > Critical Value ($\chi^2_{\alpha/2}$)	p-value < α	Reject H_0
	χ^2-statistic >=Critical Value ($\chi^2_{1-\alpha/2}$) **AND** χ^2-statistic <= Critical Value ($\chi^2_{\alpha/2}$)	p-value >= α	Retain H_0







Example-1

A company's goal is to minimize the number of machine parts that are piled up at the workstation waiting to be installed. The company expects that, on the average, about 20 parts will be at the station. However, the production superintendent suspects that the variance of has increased and it is now greater than 4 at the workstation. On a given day, the number of machine parts piled up at the workstation is determined eight different times and the following number of parts are recorded. 23, 17, 20, 29, 21, 14, 19 and 24. Using the significance value of 5%, test the hypothesis.

Given that:

$\sigma^2 = 4$, $n = 8$, $v = 8 - 1 = 7$

From the given data sample variance (s^2) can be calculated as = 20.9821

Significance value (α) = 0.05

This problem involves χ^2-statistic.

Null Hypothesis (H_0): $\sigma^2 <= 4$

Alternative Hypothesis (H_A): $\sigma^2 > 4$

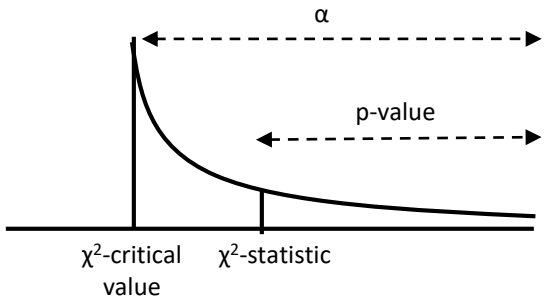
Alternative hypothesis is more than the hypothesis value, so rejection region will be on the right side of the one-tailed test.

Critical value for $\alpha = 0.05$ and v (= 7) = **14.067**

$$\chi^2\text{–statistic} = \frac{(n-1)s^2}{\sigma^2} = \frac{7 \text{ x } 20.9821}{4} = \mathbf{36.719}$$

Since χ^2-statistic > critical value (or, p-value < significance value)

so **null hypothesis is rejected.**



Right-Tail H_0 Rejection

Example-2

A small business has few employees. Because of the uncertain demand for its product, the company usually pays overtime on any given week. The company assumed that about 50 total hours of overtime per week is required and that the variance on this figure is about 25. Company officials want to know whether the variance of overtime hours has changed. A sample of 16 weeks of overtime data (in hours per week) gave a variance of 28.06. Use this data to test the null hypothesis that the variance of overtime data is 25. Significance value is 0.10.

Given that:

$\sigma^2 = 25$, n = 16, v = 16 - 1 = 15

From the given data sample variance (s^2) = 28.06

Significance value (α) = 0.10

This problem involves χ^2-statistic.

Null Hypothesis (H_0): $\sigma^2 = 25$

Alternative Hypothesis (H_A): $\sigma^2 \neq 25$

Alternative hypothesis could be less than or more than the hypothesis value, so rejection region will be on both the sides of the two tailed test with significance value ($\alpha/2$) = **0.05**

Critical value for $\alpha/2 = 0.05$ and v (= 15) = **24.996** (right end tail)

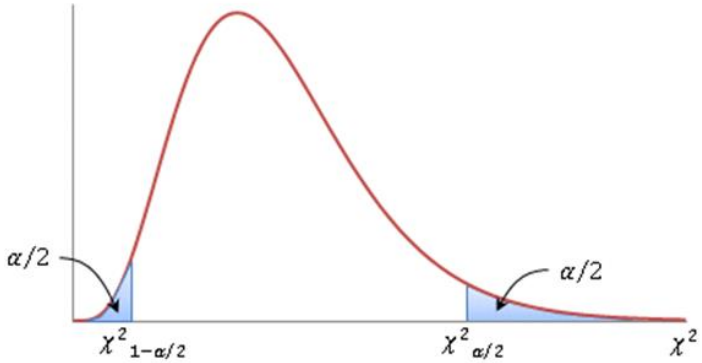
Critical value for $1-(\alpha/2) = 0.95$ and v (= 15) = **7.261** (left end tail)

$$\chi^2\text{–statistic} = \frac{(n-1)s^2}{\sigma^2} = \frac{15 \text{ x } 28.06}{25} = 16.836$$

Since, Critical Value ($\chi^2_{1-\alpha/2}$) < χ^2-statistic < Critical Value ($\chi^2_{\alpha/2}$)

Not in the so **null hypothesis is retained.**

Remember: χ^2 distribution starts from the origin.



Exercise

1. Calculate the p-value to reach to a statistical conclusion: Null Hypothesis (H_0): $\mu = 25$, Alternative Hypothesis (H_A): $\mu \neq 25$, $\bar{x} = 28.1$, $n = 57$, $\sigma = 8.46$ and $\alpha = 0.01$.

   (Answer: p-value = 0.0028 x 2)

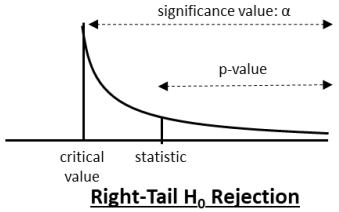
2. A random sample of size 20 is taken, resulting in a sample mean of 16.45 and a sample standard deviation of 3.59. Assume x is normally distributed and $\alpha = 0.05$. Test the following hypotheses: Null Hypothesis (H_0): $\mu = 16$, Alternative Hypothesis (H_A): $\mu \neq 16$.

   (Answer t−statistic = 0.56, Retained)

3. Test the hypothesis when Null Hypothesis (H_0): $\sigma^2 = 20$, Alternative Hypothesis (H_A): $\sigma^2 > 20$ given that $\alpha = 0.05$, $n = 15$, $s^2 = 32$.

   (Answer χ^2−statistic = 22.4, Retained)

Type-I and Type-II Errors



- **Type-I Error**: Conditional probability of rejecting a null hypothesis when it is true is called Type-I Error. It is same as significance value and denoted by **α**.
- **Type-II Error**: Conditional probability of retaining a null hypothesis when null hypothesis is false is called Type-II Error. It is denoted by **β**.
- **Power of Hypothesis Test**: Conditional probability of rejecting a null hypothesis when it is false is called the power of the hypothesis test. It is equal to **1 – β**.

	Retain	Reject
H_0 is True	Correct Decision	Type-I Error (α)
H_0 is False	Type-II Error (β)	Correct Decision

Challenges in Finding Out the Type-II Error: (Example): As per the data the average IQ is 82 with standard deviation 11.03 for the citizen of a country. Ministry does not agree with this data and believes that IQ is now more than 82.

Null Hypothesis (H_0): $\mu <= 82$

Alternative Hypothesis (H_A): $\mu > 82$

- Type-II error will occur when null hypothesis is false (alternative hypothesis is true) and it is retained.
- So if null hypothesis is false, what is the actual value of the mean. It is > 82, but what is the actual value? It could be 82.5, 83, 83.5 or anything else?
- So, Type-II Error is calculated w.r.t. an assumed true mean value.

Example-1

A research firm claimed that average IQ is 82 with standard deviation 11.03 for the citizen of a country. Ministry does not completely agree with the research firm and believes that IQ is more than 82 at around 86. If significance value is 0.05, sample count = 100 and calculate the Type-II error.

Null Hypothesis (H_0): $\mu <= 82$

Alternative Hypothesis (H_A): $\mu > 82$

Ministry believes that IQ is more than 82, so it is a right tail test.

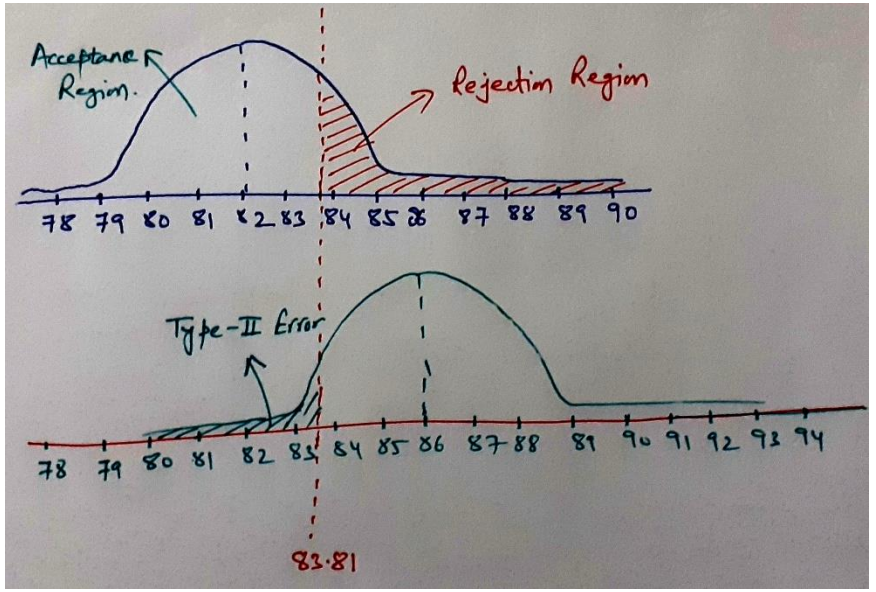
For $\alpha = 0.05$, the z critical value = 1.645

x-critical value = $\mu + z\left(\frac{\sigma}{\sqrt{n}}\right) = 82 + 1.645.\left(\frac{11.03}{\sqrt{100}}\right) = 83.81$

The z-value for 83.81 w.r.t. the assumed true mean = $\frac{83.81-86}{11.03/\sqrt{100}} = -1.98549$

The probability P(z <= -1.98549) = 0.02354

So Type-II error is 0.02354 w.r.t. the assumed mean 86 when null hypothesis is retained but it is false.



Example-2

Suppose a null hypothesis is that the population mean is greater than or equal to 100. Suppose further that a random sample of 48 items is taken and the population standard deviation is 14. Compute the probability of committing a Type-II error if the population mean actually is 96 with significance value as 0.05.

Null Hypothesis (H_0): $\mu >= 100$

Alternative Hypothesis (H_A): $\mu < 100$

Actually mean is 96 which is less than 100, so it is a left tail test.

For $\alpha = 0.05$, the z critical value = 1.645

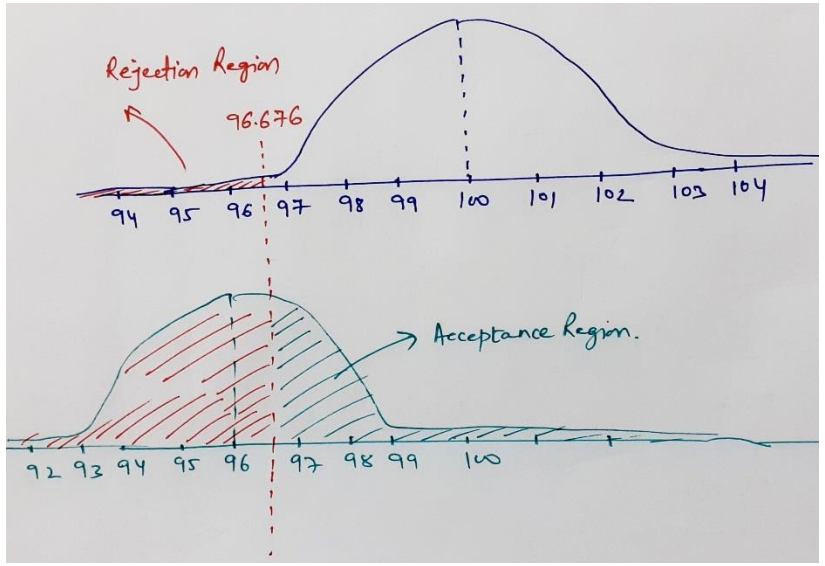
x-critical value = $\mu - z\left(\frac{\sigma}{\sqrt{n}}\right) = 100 - 1.645.\left(\frac{14}{\sqrt{48}}\right) = 96.676$

The z-value for 96.676 w.r.t. the actual mean = $\frac{96.676-96}{14/\sqrt{48}} = 0.3345$

The probability P(z <= 0.3345) = 0.6310

So Type-II error is 1 - 0.6310 = 0.3690 w.r.t. the actual mean 96

when null hypothesis is retained but it is false.



Exercise

The test is conducted on the following hypothesis:

H_0: $\mu >= 12$

H_A: $\mu < 12$

Test was done taking a sample size as 60. Taking significance value as 0.05 and population standard deviation is 0.10, calculate the probability of committing a Type-II error if the assumed true mean is 11.96.

(Answer = 0.073)

Appendix

(Dealing with Two or More Populations/Samples)

For Self Study

Will Not be Included in the Evaluation Components

Comparing Two Populations

Introduction

- It is expensive for a mobile operator when one of its towers breaks down. Depending on the availability, sometimes the operator might need to send a novice to fix the problem. An mobile operator wants to conduct an experiment to compare the average time for an expert to fix a problem and the average time for a novice to fix the problem.
- Petroleum companies advertise the advantages of using premium fuel. A consumer group wants to determine the difference in mileage of cars using regular fuel and cars using premium fuel.
- A consumer group wants to test if two different brands of toothpaste are rated same by the consumers or two different brand of tyres wear differently.
- What is the interval estimation of average saving per month on groceries from the samples of families who buy only online and who buy only visiting the stores?
- Two machines produce the same mechanical part in time shifts. Is the accuracy measurement vary between these two lots of parts?
- Can the stock price for a set of companies be compared for the two successive years?
- Do employees perform better after a training program?
- In the above examples, the objectives are to determine the amount of differences in the two populations or two populations before and after some event. This is done drawing the random samples from these two sets.
- In this topic, we will review what are the different techniques to compare these two sets.

Some Background Math

- Let us say there is a function g(x) = a.x + b where a and b are constants.
- For this function f(x) is the probability distribution function.
- We have reviewed that the first order moment is the mean. So the first order moment of g(x):

$$E(a.x+b)$$

$$= \int_{-\infty}^{\infty} g(x).f(x)dx$$

$$= \int_{-\infty}^{\infty} (a.x+b).f(x)dx$$

$$= a.\int_{-\infty}^{\infty} x.f(x)\,dx + b.\int_{-\infty}^{\infty} f(x)dx$$

$$= a.E(x) + b$$

$$= \boldsymbol{a.\mu_x + b}$$

- We have also reviewed that the second order moment about the mean is the variance. So the variance of g(x):

$$Var(a.x+b)$$

$$= \int_{-\infty}^{\infty} \{(a.x+b) - (a.\mu_x + b)\}^2.f(x)dx$$

$$= a^2 \int_{-\infty}^{\infty} (x-\mu_x)^2.f(x)dx$$

$$= \boldsymbol{a^2.Var(x)}$$

- Let us say Y is a combination of two independent random variables x_1 and x_2 and can be represented with two constants a_1 and a_2 in the following form:

$$Y = a_1x_1 + a_2x_2$$

- The mean of Y can be expressed as:

$$E(a_1x_1 + a_2x_2)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (a_1x_1 + a_2x_2)f_1(x_1)f_2(x_2)dx_1\,dx_2$$

$$= a_1\int_{-\infty}^{\infty} x_1\,f_1(x_1)dx_1.\int_{-\infty}^{\infty} f_2(x_2)dx_2 + a_2.\int_{-\infty}^{\infty} f_1(x_1)\,dx_1.\int_{-\infty}^{\infty} x_2.f_2(x_2)dx_2$$

$$= a_1.E(x_1) + a_2.E(x_2)$$

$$= \boldsymbol{a_1.\mu_1 + a_2.\mu_2}$$

- The variance of Y can be expressed as:

$$Var\,(Y)$$

$$= E\,(Y-\mu_Y)^2$$

$$= E\,(a_1x_1 + a_2x_2 - a_1\mu_1 - a_2\mu_2)^2$$

$$= E[a_1(x_1-\mu_1) + a_2(x_2-\mu_2)]^2$$

$$= E[a_1^2.(x_1-\mu_1)^2 + a_2^2.(x_2-\mu_2)^2 + 2.a_1.a_2.(x_1-\mu_1).(x_2-\mu_2)]$$

$$= a_1^2.E(x_1-\mu_1)^2 + a_2^2.E(x_2-\mu_2)^2 + 2.a_1.a_2.E[(x_1-\mu_1).(x_2-\mu_2)]$$

$$= a_1^2.Var(x_1) + a_2^2.Var(x_2) + 2.a_1.a_2.E[(x_1-\mu_1).(x_2-\mu_2)]$$

- If x1 and x2 are independent their covariance will be 0.

$$E[(x_1-\mu_1).(x_2-\mu_2)] = 0$$

- So:

$$\boldsymbol{Var\,(Y) = a_1^2.Var(x_1) + a_2^2.Var(x_2)}$$

Example

Example-1: Given that if x_1 and x_2 are independent. Let x_1 has mean 4 and variance 9 and let x_2 has mean -2 and variance 6. Find the mean and variance of: $(2x_1 + x_2 - 5)$.

We have reviewed that:

$\text{Mean}\ (ax + b) = a.\mu_x + b$

$\text{Var}\ (ax + b) = a^2.\text{Var}(x)$

if $Y = a_1x_1 + a_2x_2$, then where x_1 and x_2 are independent and a_1 and a_2 are two constants:

$\text{Mean}\ (Y) = a_1\mu_1 + a_2\mu_2$

$\text{Var}\ (Y) = a_1^2\text{Var}(x_1) + a_2^2\text{Var}(x_2)$

$$\text{Mean}\ (2x_1 + x_2 - 5) = 2.\text{Mean}(x_1) + 1.\text{Mean}(x_2) - 5 = 2x4 - 2 - 5 = 1$$

$$\text{Var}\ (2x_1 + x_2 - 5) = 2^2.\text{Var}(x_1) + 1^2.\text{Var}(x_2) = 4x9 + 1x6 = 42$$

Example-2: Given that if x_1 and x_2 are independent. Let x_1 has mean μ_1 and variance σ^2_1 and let x_2 has mean μ_2 and variance σ^2_2.Find the mean and variance of: $(x_1 - x_2)$ and $(x_1 + x_2)$.

Comparing $(x_1 - x_2)$ with $a_1x_1 + a_2x_2$

$a_1 = 1$ and $a_2 = -1$

So:

$\text{Mean}\ (x_1 - x_2) = \mu_1 - \mu_2$

$\text{Var}\ (Y) = \text{Var}(x_1) + \text{Var}(x_2)$

$= \sigma^2_1 + \sigma^2_2$

Comparing $(x_1 + x_2)$ with $a_1x_1 + a_2x_2$

$a_1 = 1$ and $a_2 = 1$

So:

$\text{Mean}\ (x_1 + x_2) = \mu_1 + \mu_2$

$\text{Var}\ (Y) = \text{Var}(x_1) + \text{Var}(x_2)$

$= \sigma^2_1 + \sigma^2_2$

Interval Estimation & Hypothesis Testing

Using z-Statistic

- Let us say two samples (with $\bar{x}_1$ and $\bar{x}_2$ means) are drawn from two populations (with μ_1 and μ_2 means and σ_1 and σ_2 standard deviations) of sizes of sizes n_1 and n_2.
- The central limit theorem states that the difference in two sample means is normally distributed for large sample sizes (both n_1 and n_2 >= 30) regardless of the shape of the populations with the following properties:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2$$

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Why?

- The z-statistic is calculated by:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Confidence interval for the population mean:

$$(\bar{x}_1 - \bar{x}_2) - z.\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z.\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Example

Standard deviation of the mileage on regular fuel is 3.46 miles per gallon (mpg) and on premium fuel is 2.99 mpg. Two sets of 50 cars each are tested on regular and premium fuels that give the mean mileage of 21.45 mpg and 24.60 mpg respectively. Construct a 95% confidence interval for the difference of mean mileage of regular and premium fuel based on this sample experiment.

Given that:

$\sigma_1 = 3.46, \sigma_2 = 2.99, n_1 = n_2 = 50, \bar{x}_1 = 21.45$ and $\bar{x}_2 = 24.60$

The z-value for 95% confidence interval = 1.96

$$(\bar{x}_1 - \bar{x}_2) - z.\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + z.\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(21.45 - 24.60) - 1.96.\sqrt{\frac{3.46^2}{50} + \frac{2.99^2}{50}} \leq (\mu_1 - \mu_2) \leq (21.45 - 24.60) + 1.96.\sqrt{\frac{3.46^2}{50} + \frac{2.99^2}{50}}$$

$$\mathbf{-4.42 \leq (\mu_1 - \mu_2) \leq -1.88}$$

Example

A DTH company sends technicians to resident houses for dish and set top box installations. In 2009, the standard deviation for the overhead expenditure for these trips per technician per day was ₹18.50 and in 2019 it was ₹15.60. The company's auditor believes that the daily overhead expenditure rose significantly between 2009 and 2019. To test this belief, the auditor samples 51 trips from the company's records for 2009. The sample average was ₹190 per day. The auditor selects a second random sample of 47 business trips from the company's records for 2019. The sample average was ₹198 per day. If he uses a risk of committing a Type-I error of 0.01, does the auditor find that the average expenditure has gone up significantly?

Given that:

$\sigma_1 = 18.50, \sigma_2 = 15.60, n_1 = 51, n_2 = 47, \bar{x}_1 = 190$ and $\bar{x}_2 = 198$

Type-1 Error $(\alpha) = 0.01$

Null Hypothesis (H_0): $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

Alternative Hypothesis (H_A): $\mu_1 - \mu_2 < 0$

z-critical value for the given $\boldsymbol{\alpha = -2.3263}$ for the left-end tail test

z-statistic from the given data:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$= \frac{(190 - 198) - 0}{\sqrt{\frac{18.50^2}{51} + \frac{15.60^2}{47}}} = \boldsymbol{-2.3202}$$

Since, z-statistic >= z-critical value for the left-end tail test, the null hypothesis is retained and auditor's belief is tested as incorrect.

Interval Estimation & Hypothesis Testing

Using t-Statistic

- The pooled variance of the two samples (x_1 and x_2) can be given as:

$$S_p^2 = \frac{\sum_{i=1}^{n1}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n2}(x_{i2} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

- If the individual variances of the samples are S_1^2 and S_2^2 respectively, the above expression can be re-written as:

$$S_p^2 = \frac{(n_1 - 1).S_1^2 + (n_2 - 1).S_2^2}{(n_1 + n_2 - 2)}$$

- From the Central Limit Theorem of two populations, the standard deviation for the difference of sample means:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the standard deviations of two populations are not known, the above expression can be used to estimate the standard deviation of the difference of the two sample means:

$$S_{\bar{x}_1 - \bar{x}_2} = S_p . \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- The t-statistic is calculated by:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p . \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Example

Waste industrial material is crushed and used for constructing the roadways. This is not only environmental friendly but also provides better strength and durability of the roads. Six samples each from two locations of waste collection centres are picked up. Their resilience modulus of strength values are listed below. Use the 0.05 significance value and test the null hypothesis if the mean strength of the waste are same at the two locations.

Resilience Modulus						
Location-1	707	632	604	652	669	674
Location-2	552	554	484	630	648	610

Given that, $n_1 = n_2 = 6$

Mean $(\bar{x}_1)$ of location-1 $= 656.33$, variance $(S_1^2) = 1277.87$

Mean $(\bar{x}_2)$ of location-2 $= 579.67$, variance $(S_2^2) = 3739.87$

Significance value $(\alpha) = 0.05$, Degree of Freedom $(v) = 6+6-2 = 10$

Null Hypothesis (H_0): $\mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

Alternative Hypothesis (H_A): $\mu_1 \neq \mu_2$

The t-critical value for $\alpha/2 = 0.025$ for two-tailed test for the given $v =$ **2.228**

$$S_p^2 = \frac{(n_1 - 1).S_1^2 + (n_2 - 1).S_2^2}{(n_1 + n_2 - 2)}$$

$$S_p^2 = \frac{(6-1).1277.87 + (6-1).3739.87}{(6+6-2)} = 2508.87, \text{so } S_p = 50.09$$

$$t - \text{statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t - \text{statistic} = \frac{(656.33 - 579.67) - 0}{50.09.\sqrt{\frac{1}{6} + \frac{1}{6}}} = \mathbf{2.65}$$

Since t-statistic > t-critical value for the right-end tailed test, the null hypothesis is rejected. It also means that the strength of the waste is different at the two locations.

Example

A coffee manufacturer is interested in estimating the difference in the average daily coffee consumption of regular-coffee drinkers and decaffeinated-coffee drinkers. Its researchers sampled 13 and 15 coffee drinkers in each category respectively and collected the data for the number of cups of coffee they drink. The average for the regular-coffee drinkers is 4.35 cups, with a standard deviation of 1.20 cups. The average for the decaffeinated-coffee drinkers is 6.84 cups, with a standard deviation of 1.42 cups. Assuming that the daily consumption is normally distributed, construct a 95% confidence interval to estimate the difference in the averages of the two populations.

Given that:

$S_1 = 1.20, S_2 = 1.42, n_1 = 13, n_2 = 15, \bar{x}_1 = 4.35$ and $\bar{x}_2 = 6.84$

The Degree of Freedom (v) $= 13+15-2 = 26$

The t-value for $\alpha/2 = (1-0.95)/2 = 0.025$ and the given degree of freedom $= 2.056$

$$S_p^2 = \frac{(n_1 - 1).S_1^2 + (n_2 - 1).S_2^2}{(n_1+n_2 - 2)}$$

$$S_p^2 = \frac{(13 - 1).1.20^2 + (15 - 1).1.42^2}{(13 + 15 - 2)} = 1.75, \text{so } S_p = 1.32$$

$$(\bar{x}_1 - \bar{x}_2) - t.S_p.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + t.S_p.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(4.35-6.84) - 2.056 * 1.32.\sqrt{\frac{1}{13} + \frac{1}{15}} \leq (\mu_1 - \mu_2) \leq (4.35-6.84) + 2.056 * 1.32.\sqrt{\frac{1}{13} + \frac{1}{15}}$$

$$\mathbf{-3.52 \leq (\mu_1 - \mu_2) \leq -1.46}$$

Matched Pair Comparison

- There are many situations where before-and-after condition of the data is to be compared. E.g.:
  - Ceramic components of a mechanical part lose weight after the baking process.
  - Price to Earning Ratio (P/E) of selected 10 companies in the two successive years.
  - Market price of the flats in an apartment complex before and during Covid-19 pandemic.
  - Weekly loss of working hours in a factory before and after taking the stringent safety measures.
  - Body Mass Index (BMI) before and after following a strict diet and exercise regime for 3 months.
- This comparison of two populations is popularly known as Matched Pair Comparison or correlated t-tests.
- t-statistic for this test is given by:

$$t - statistic = \frac{\bar{d} - D}{\frac{S_d}{\sqrt{n}}}$$

Where:
$\bar{d}$ = mean sample difference
D = mean population difference
S_d = standard deviation of the sample difference
n = number of pairs with degree of freedom = n-1

- Null hypothesis is tested for mean population difference D = 0.
- The hypothesis D = 0 indicates that the means of the two responses are the same.

Example

A stock market investor thinks that there is a significant difference in the P/E (price to earnings) ratio for 9 companies from a year to the next. Test the hypothesis taking the significance value as 0.01 and the data below.

$$\mathbf{t-statistic} = \frac{\bar{\mathbf{d}} - \mathbf{D}}{\frac{\mathbf{S_d}}{\sqrt{\mathbf{n}}}}$$

Company	Year-1 P/E Ratio	Year-2 P/E Ratio	d
1	8.9	12.7	-3.80
2	38.1	45.4	-7.30
3	43	10	33.00
4	34	27.2	6.80
5	34.5	22.8	11.70
6	15.2	24.1	-8.90
7	20.3	32.3	-12.00
8	19.9	40.1	-20.20
9	61.9	106.5	-44.60

Sum	-45.30
Mean ($\bar{d}$)	-5.03
Standard Deviation (S_d)	21.60

Given that:

Significance value $(\alpha) = 0.01$, Degree of Freedom $(v) = 9\text{-}1 = 8$

Null Hypothesis (H_0): $D = 0$

Alternative Hypothesis (H_A): $D \neq 0$

The t-critical value for $\alpha/2 = 0.005$ for two-tailed test for the given $v = \pm\mathbf{3.355}$

$$t - \text{statistic} = \frac{-5.03 - 0}{\frac{21.60}{\sqrt{9}}} = -0.6986$$

Since t-statistic > t-critical value in the left-end tail, the null hypothesis is retained. It means the experiment suggests there is no significance difference in the P/E ratio in two years.

Exercise

1. Variance for the two populations are 22.74 and 26.65. Given the significance value as 0.02, test the null hypothesis that there is no difference in the population means. The sample data from two populations are shown below:

Sample-1						Sample-2					
90	88	80	88	83	94	78	85	82	81	75	76
88	87	91	81	83	88	90	80	76	83	88	77
81	84	84	87	87	93	77	75	79	86	90	75
88	90	91	88	84	83	82	83	88	80	80	74
89	95	97	95	93	97	80	90	74	89	84	79

(Answer: z-statistic = 5.4813, null hypothesis is rejected)

2. Construct a 98% confidence interval to estimate the difference in population means using the above data.

(Answer: $4.04 \leq (\mu_1 - \mu_2) \leq 10.02$)

3. A sample-1 of size 8 has mean 24.56 and standard deviation 12.40. Another sample-2 of size 11 has mean 26.42 and standard deviation 15.80. Using 1% significance value, test the alternative hypothesis that $\mu_1 - \mu_2 < 0$.

(Answer: t − statistic = −1.05, null hypothesis is retained or alternative hypothesis is rejected)

4. The following are the average weekly losses of worker-hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation. Using the significance value as 0.05, test the null hypothesis that safety program is not effective.

Before	45	73	46	124	33	57	83	34	26	17
After	36	60	44	119	35	51	77	29	24	11

(Answer: t − statistic = 4.03, null hypothesis is rejected)

F-distribution

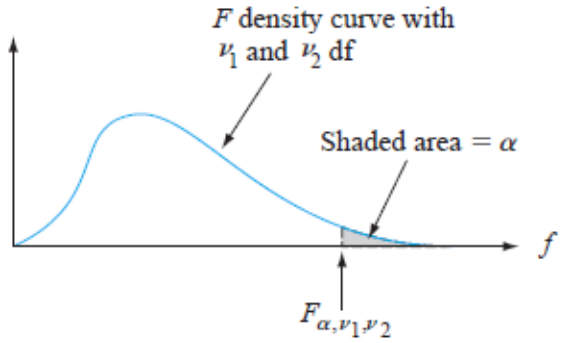
- If S^2_1 and S^2_2 are the variances of independent random samples of size n_1 and n_2 respectively taken from two normally distributed populations having the same variance then:

$$F = \frac{S_1^2}{S_2^2}$$

- If two samples come from the same population (or populations with equal variances), the ratio of the sample variances (F) should be 1. However, because of sampling errors this ratio vary.
- The distribution of F belongs to Beta-distribution, where the value of the random variable takes on a value between 0 and 1.
- F follows a continuous distribution with degrees of freedom (df) for numerator (v_1) = n_1-1 and for denominator (v_2) = n_2-1.
- Statistics books capture the critical values of the F ratio in a table for few values of v_1 and v_2 and α (the area under the tail on the right hand side). This is denoted by $F_{\alpha, v1, v2}$.
- Left hand side F value for the probability (1-α) is the inverse of corresponding F value for α. So:

$$F_{(1-\alpha),v1,v2} = \frac{1}{F_{\alpha,v1,v2}}$$

- F-distribution is sensitive to the assumption that populations are in normal distribution.



Critical Values for *F* Distributions

ν_2 = denominator df	α	ν_1 = numerator df: 1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.010	4052.20	4999.50	5403.40	5624.60	5763.60	5859.00	5928.40	5981.10	6022.50
	.001	405,284	500,000	540,379	562,500	576,405	585,937	592,873	598,144	602,284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77

Hypothesis Testing

Concerning Two Variances

Two machines produce metal sheets that are specified to be 22 mm thick. There is variability in the thickness because of several factors. Operators are concerned about the consistency of the two machines. To test the consistency, few sheets are randomly sampled and their thickness is measured. The details are captured in the table below. Assume that the sheet thickness is normally distributed. Test the null hypothesis that the sheets that are produced by these two machine have the same variability with significance value as 0.05.

Machine-1		Machine-2		
22.30	21.90	22.00	21.70	22.00
21.80	22.40	22.10	21.90	22.10
22.30	22.50	21.80	22.00	
21.60	22.20	21.90	22.10	
21.80	21.60	22.20	21.90	

Given that: $\alpha = 0.05$

$n_1 = 10$, $n_2 = 12$, so $v_1 = 9$ and $v_2 = 11$

Null Hypothesis (H_0): $\sigma^2_1 = \sigma^2_2$

Alternative Hypothesis (H_A): $\sigma^2_1 \neq \sigma^2_2$

For the two tailed test, $\alpha/2 = 0.025$

F-critical value for the right-end tail = **3.59**

F-critical value for the left-end tail = 1/3.59 = **0.28**

From the provided data, the sample variances:
$S_1^2 = 0.11378$ and $S_2^2 = 0.02023$

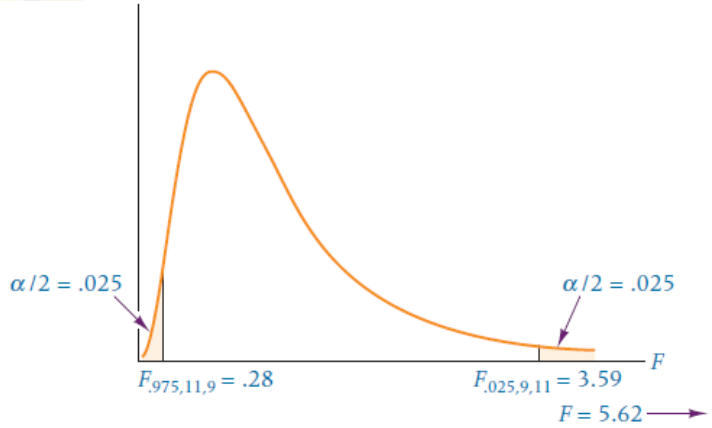
$$F - \text{statistic} = \frac{S_1^2}{S_2^2}$$

$$= \frac{0.11378}{0.02023} = \mathbf{5.62}$$

$\alpha = 0.025$

v_2 \ v_1	Numerator Degrees of Freedom 1	2	3	4	5	6	7	8	9
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59

$F_{.025,9,11}$



Since, F-statistic is in the rejection region on the right-end tail, the null hypothesis is rejected.

Exercise

General Brite is specialized in silver plating on the metal surfaces. 12 samples each from two of its workers are randomly collected that provided the standard deviations of 0.035 mil and 0.062 mil of the thickness of their plating (1 mil = 1/1000 inch). Using 0.05 significance value, test the null hypothesis if $\sigma^2_1 = \sigma^2_2$ against alternative hypothesis $\sigma^2_1 < \sigma^2_2$. Given that $F_{0.05, 11, 11} = 2.82$.

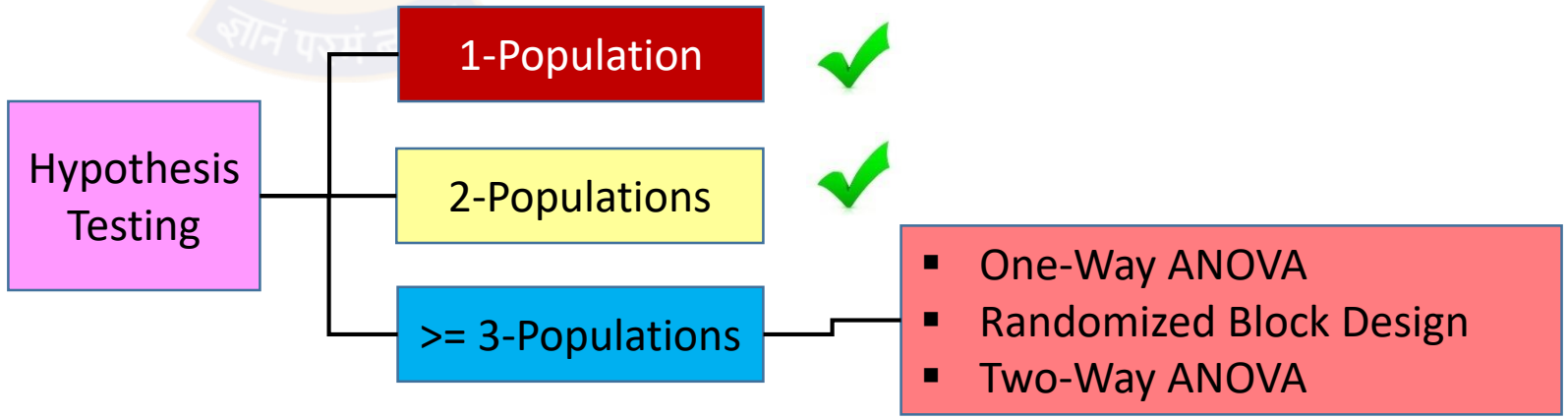
(Answer: F-statistic = 0.3187 and $F_{0.95, 11, 11} = 0.3546$, Null Hypothesis rejected)

Analysis of Variance (ANOVA)

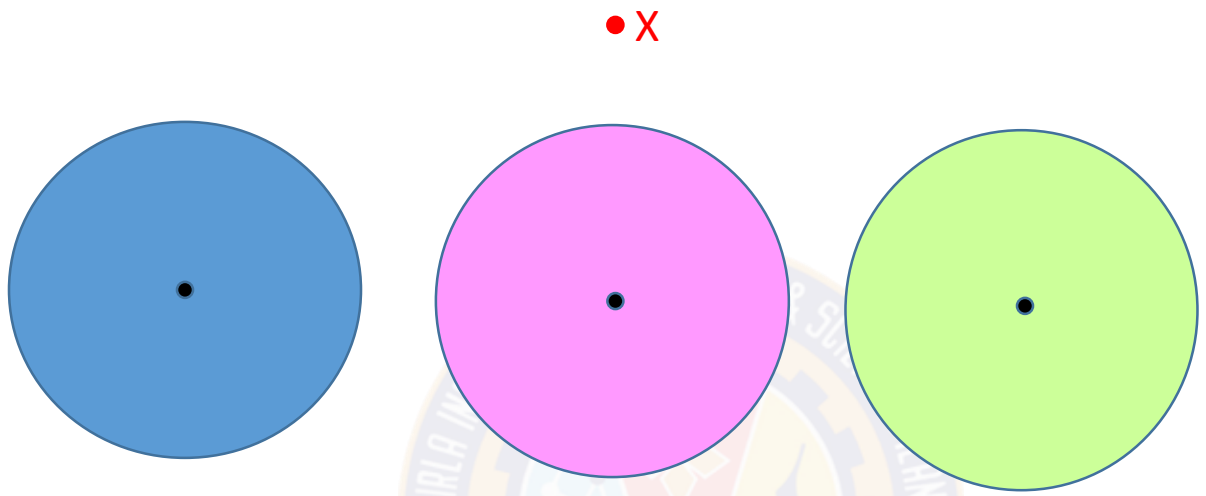
Introduction

- We have reviewed hypothesis testing in the cases of 1-population and 2-populations so far. There could be situations where statisticians want to perform the hypothesis testing for 3 or more populations. Examples:
  - Different quality of tyres in the experiment such as low-quality, medium-quality, and high-quality need to be compared.
  - Sales of garments under different discount.
  - Returns of mutual funds under different categories like large, mid and small cap funds.
- The experiment begins with finalizing few variables:
- Independent Variables (Factors): they could be treatment or classification variables. Examples:
  - 0%, 10% or 20% discount on garments is treatment.
  - Low, medium or high quality tyres is classification.
  - Mutual fund categories.
- Dependent Variables: they are the response to the independent variables in an experiment. Examples:
  - Amount of sale with different discounts.
  - Wear and tear of different quality tyres.
  - Returns from the mutual funds.

- The details of One-Way ANOVA will be reviewed in this module.
- The assumptions are that response variable will be in the normal distribution and population variance will be the same.



Measures of Partitioning



In the conceptual diagram above, assume each circle group consists of several random points (not shown), with marked centre as their group average.
The red dot (X) out the circles is the average of all random points in these groups.

Sum of square of Total Variation (SST) = $\sum(\text{Point} - X)^2$

Sum of square of Between Group Variation (SSB) = $\sum$ Count of group points * $(\text{Group Average} - X)^2$

Sum of square of Within Group Variation (SSW) = $\sum(\text{Point} - \text{Its Group Average})^2$

Measures of ANOVA

k = Number of groups
n_i = Number of observations in the group $-$ i
n = Total observations
Y_{ij} = Observation $-$ j in group $-$ i
μ_i = Mean of group i
μ = Overall mean

Sum of Squares of Total Variation (**SST**) $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu)^2$

Degree of Freedom for SST = (n-1)

Mean of Squares for Total Variation (**MST**) $= \frac{SST}{n-1}$

Sum of Squares of Between Group Variation (**SSB**) $= \sum_{i=1}^{k} n_i.(\mu_i-\mu)^2$

Degree of Freedom for SSB= (k-1)

Mean of Squares for Between Variation (**MSB**) $= \frac{SSB}{k-1}$

Sum of Squares of Within Group Variation (**SSW**) $= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\mu_i)^2$

Degree of Freedom for SSW= (n-k)

Mean of Squares for Within Variation (**MSW**) $= \frac{SSW}{n-k}$

$SST = SSB + SSW$

#	Group-1			Group-2			Group-3		
1	39	37	34	34	47	38	42	41	40
2	32	34	25	41	44	35	43	47	50
3	25	28	33	45	46	34	44	55	52
4	25	36	26	39	38	34	46	55	43
5	37	38	33	38	42	37	41	47	47
6	28	38	26	33	33	39	52	48	55
7	26	34	26	35	37	34	43	41	49
8	26	31	27	41	45	34	42	42	46
9	40	39	32	47	38	36	50	45	55
10	29	36	40	34	44	41	41	48	42

$k=3, n = 90, \mu = 39.06$

$n_1 = 30, \mu_1 = 32$

$n_2 = 30, \mu_2 = 38.77$

$n_3 = 30, \mu_3 = 46.4$

SST $= (39-39.06)^2 + \cdots + (42-39.06)^2 = 5170.72$

Degree of Freedom for **SST** = (n-1) = 89

MST $= \frac{5170.72}{89} = 58.10$

SSB $= 30*(32-39.06)^2 + 30*(38.77-39.06)^2 + 30*(46.4-39.06)^2 = 3114.16$

Degree of Freedom for **SSB**= (k-1) =2

MSB $= \frac{3114.10}{3-1} = 1557.05$

SSW $= (39-32)^2 + \cdots + (34-38.77)^2 + \cdots + (42-46.4)^2 = 2056.56$

Degree of Freedom for **SSW**= (n-k)=90-3 = 87

MSW $= \frac{2056.56}{87} = 23.64$

Cochran's Theorem and F-Test

- From the previous slide, it is reviewed that SST is broken into SSB and SSW.
- Cochran's Theorem states that if:
  - Y1, Y2, …Yn are drawn from a normal distribution with mean μ and standard deviation σ.
  - Sum of squares of total variation (SST) is decomposed into k different sum of squares (SS), with degree of freedom for the i^{th} sum of square SSi as DFi.
  - The ratio of SSi/σ^2 are independent chi-square variables (χ^2) with DFi as degree of freedom.
  - If $\sum_{i=1}^{k} DFi = n - 1$
- In this context, the F-statistic is defined as:

$$F = \frac{MSB}{MSW}$$

- F-critical value from the tables is read for the degrees of freedom for MSB and MSW respectively.
- Null hypothesis in ANOVA is if means for the different groups are equal. ANOVA is a right tail test hypothesis testing as the objective is if the variation between groups is greater than the variation within groups.

Example

The three groups represent the product sale on randomly selected days under 0%, 10% and 20% discounts respectively. Using significance value as 0.05 test the hypothesis if discount has any significant impact on the average sale quantity.

#	Group-1			Group-2			Group-3		
1	39	37	34	34	47	38	42	41	40
2	32	34	25	41	44	35	43	47	50
3	25	28	33	45	46	34	44	55	52
4	25	36	26	39	38	34	46	55	43
5	37	38	33	38	42	37	41	47	47
6	28	38	26	33	33	39	52	48	55
7	26	34	26	35	37	34	43	41	49
8	26	31	27	41	45	34	42	42	46
9	40	39	32	47	38	36	50	45	55
10	29	36	40	34	44	41	41	48	42

Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3$

Alternative Hypothesis (H_A): $\mu_1 \neq \mu_2 \neq \mu_3$

From the previous slide:

MSB = 1557.05, degree of freedom = 2

MSW = 23.64, degree of freedom = 87

F-statistic = MSB/MSW = 1557.05 /23.64 = 65.87

F-critical value for 2 and 87 degrees of freedoms for α(=0.05) = 3.101

Since, F-statistic > F-critical value for the right-tail test, the null hypothesis is rejected. It also means that different discounts have an impact.

Exercise

A company has three manufacturing plants. The following data shows the ages of five randomly selected workers at each plant. Using significance value as 0.01, perform one-way ANOVA to determine whether there is a significant difference in the mean ages of the worker-populations at the three plants. Given that $F_{0.01,\,2,\,12} = 6.93$.

Plant-1	Plant-2	Plant-3
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26

(Answer: F-statistic = 39.80, There is age difference)

Thank You