DS Part I I I
Inferential Statistics
(Hypothesis Testing, Anova, Chi Squared)

Sampling Distribution

Central Limit Theorem

What is Sampling Distribution ?

Suppose that we draw all possible samples of size *n* from a given population and then we compute a statistic (e.g., a mean, proportion, standard deviation) for each sample. The probability distribution of this statistic is called a **sampling distribution**. And the standard deviation of this statistic is called the **standard error**.

Central Limit Theorem

Sampling distribution of the sample mean will be approximately normal for a sufficiently large sample size. The larger the sample size, the more closely the sampling distribution of x bar will resemble a normal distribution (even if the underlying population is not normal).

Confidence Interval

A confidence interval is always centered around the mean of your sample. In order to construct this, add a margin of error. It can be found by multiplying the standard error of the mean by the z score of the percent confidence interval

$$\text{confidence interval} = \bar{x} \pm \text{margin of error}$$

$$\text{margin of error} = Z \times \frac{\sigma}{\sqrt{n}}$$

Note

Common choices for C.I are 90%, 95% , and 99% . The interval constructed with 99% confidence will have higher chance of containing the true mean than an interval constructed with 95% confidence

C.I - Exercise

- Suppose the average height of a sample of 100 women is 5’5’’, in other words, X bar = 5’5’’. Within what range of heights can we expect the population mean to be, with 95% confidence? Assume a standard deviation for the population of 1.5’’

Our sample mean is 5’5”

The standard deviation of the population is 1’5“

Our sample size is 100.

We are asked for 95% confidence, so we want to use z = ± 1.96.

5.5 ± 1.96 ∗ 1.5/ √ 100 = 5.5±.294 = (5.21,5.79)

We can conclude, with 95% confidence, that the true average height for women is between 5’2" and 5’8".

Hypothesis Testing

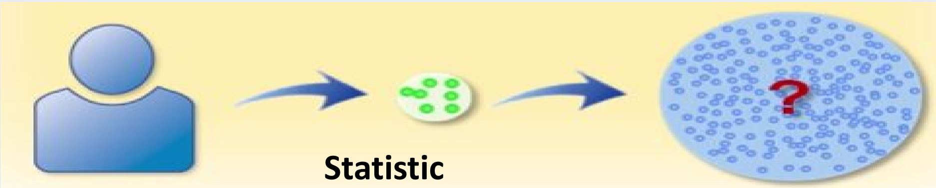
Agenda

- Introduction to Hypothesis testing
- Significance Levels
- Steps in Hypothesis Testing
- One and Two tailed tests
- Type I and Type II errors

Hypothesis Testing

What is Hypothesis Testing ?

- It uses sample data to evaluate a question about a population
- Hypothesis (prediction) testing is the process of determining whether or not a given hypothesis is true. Is also called *significance testing*
- A statistical **Hypothesis** is an assumption about a population parameter
- It examines two opposing **hypotheses** about a population:
  the null **hypothesis** and the alternative **hypothesis**
  - NULL Hypothesis **– A hypothesis to be tested**
  - Alternative Hypothesis **– A hypothesis that we use as an alternative to the null hypothesis**



APC company claims that their UPS backup time last for 6 hours

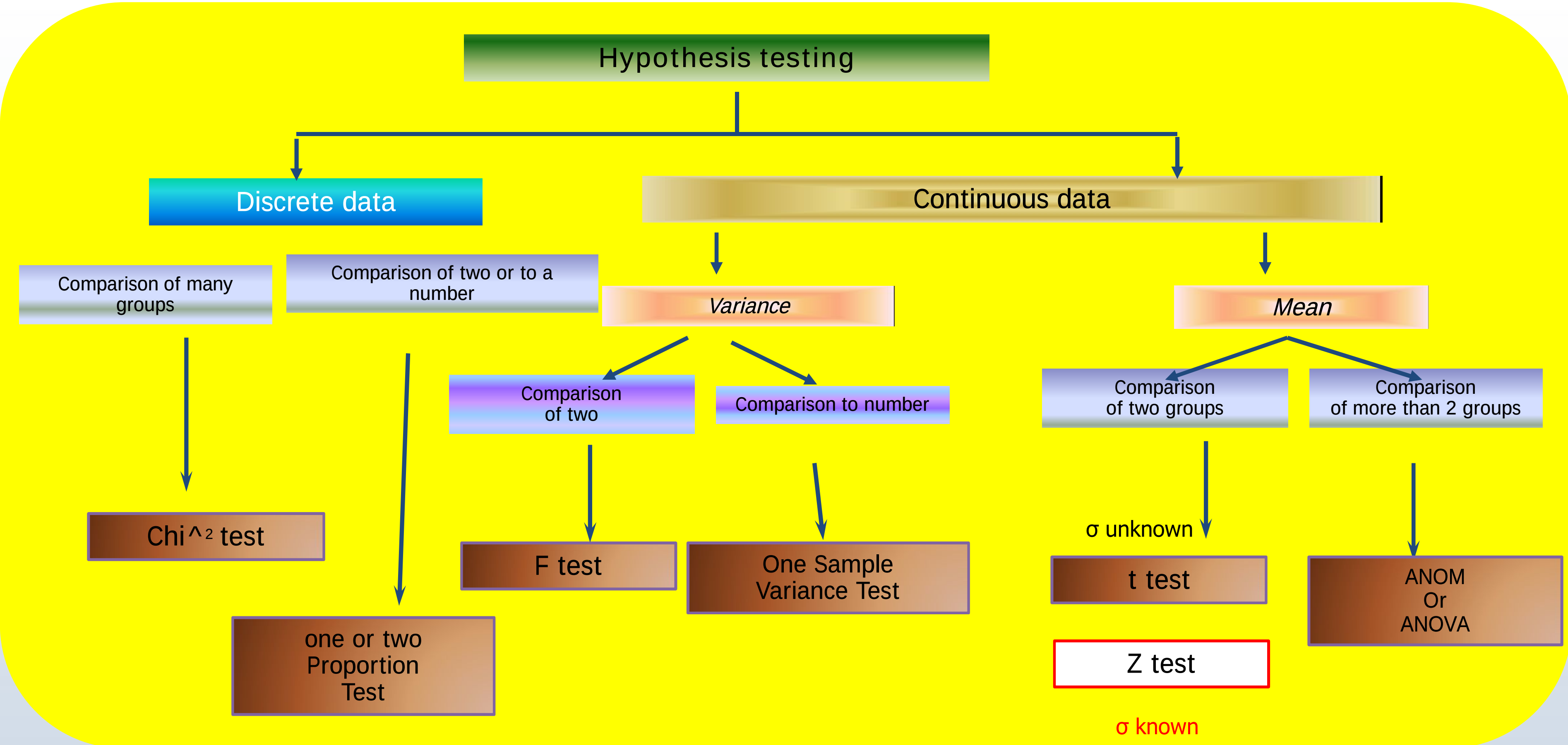
Randomly select 10 APC UPS and check if the average backup time last for 6 hours

HYPOTHESIS TESTING

Types of Hypothesis Tests:

- **One Tail and Two Tail Tests**
- **One sample and Two Sample Tests**
  - **Single Sample Z/T tests, Independent Sample T tests, Paired Sample T tests**
- **Multiple sample tests**
  - **ANOVA**
- **Non-parametric Tests**
  - **Chi Square Tests**

Hypothesis Testing Comparisons



Hypothesis Testing

Average Quantity of chips in these packets = 235 g

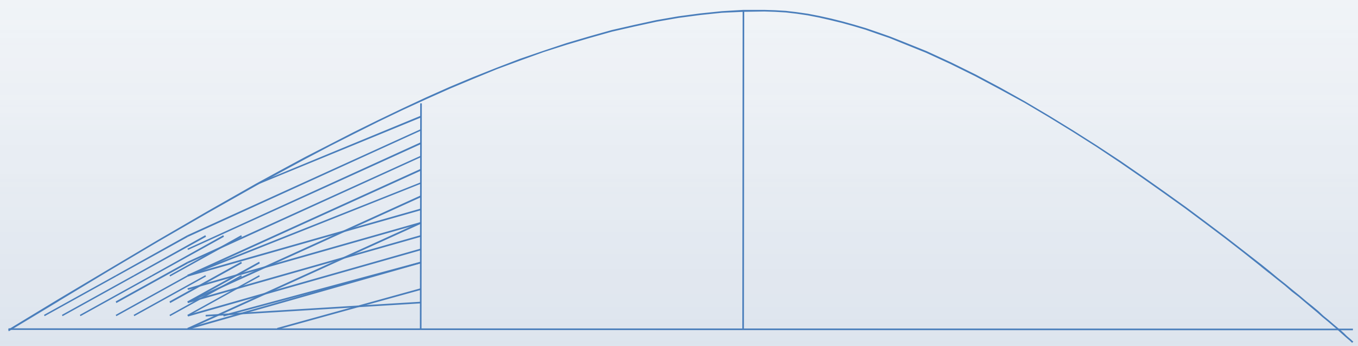
Random sample of 25 packets

Claim : Average Quantity of chips in these packets = 235 g

Scenario 1 : Average amount of chips in these 25 packets is 200 g

Scenario 2 : Average amount of chips in theses 25 packets is 230 g

Hypothesis Testing



How can I be more certain that the claim is false, before rejecting it?

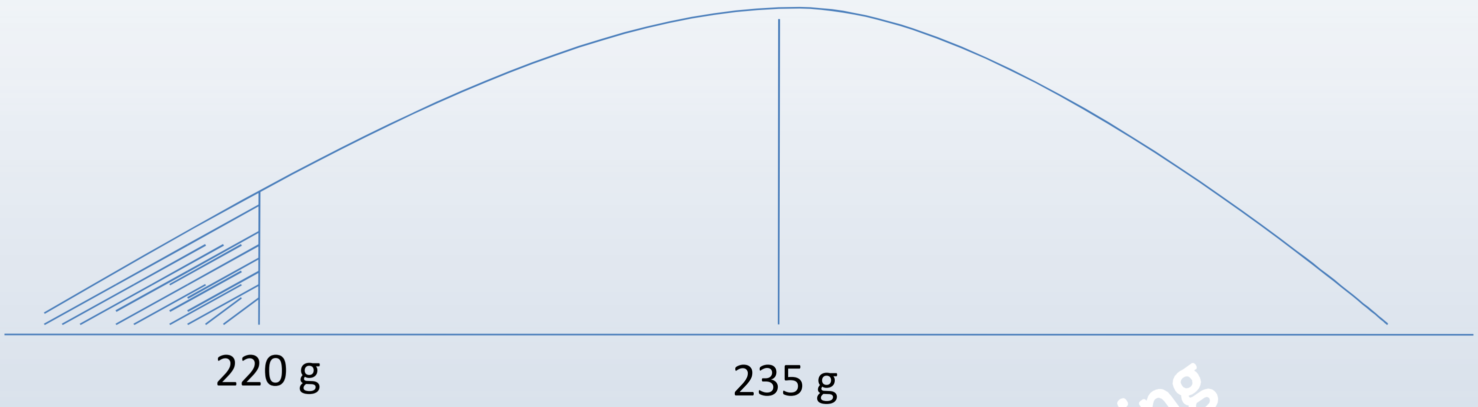
Define a lower cut off

Reduce the cut off value to 200 g

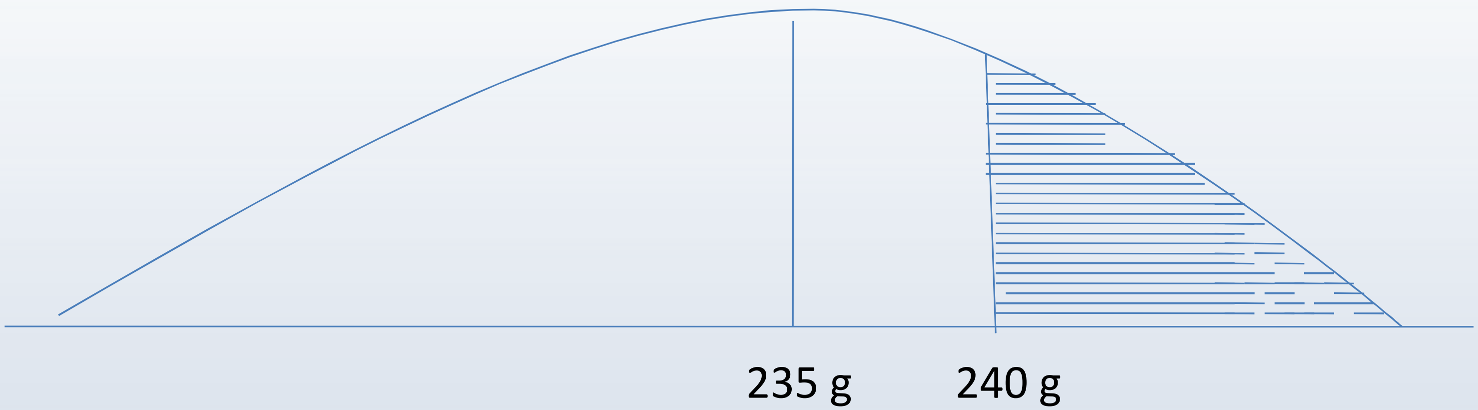
Hypothesis testing – Left tail test

In Hypothesis testing, we begin by making tentative assumption about a population parameter

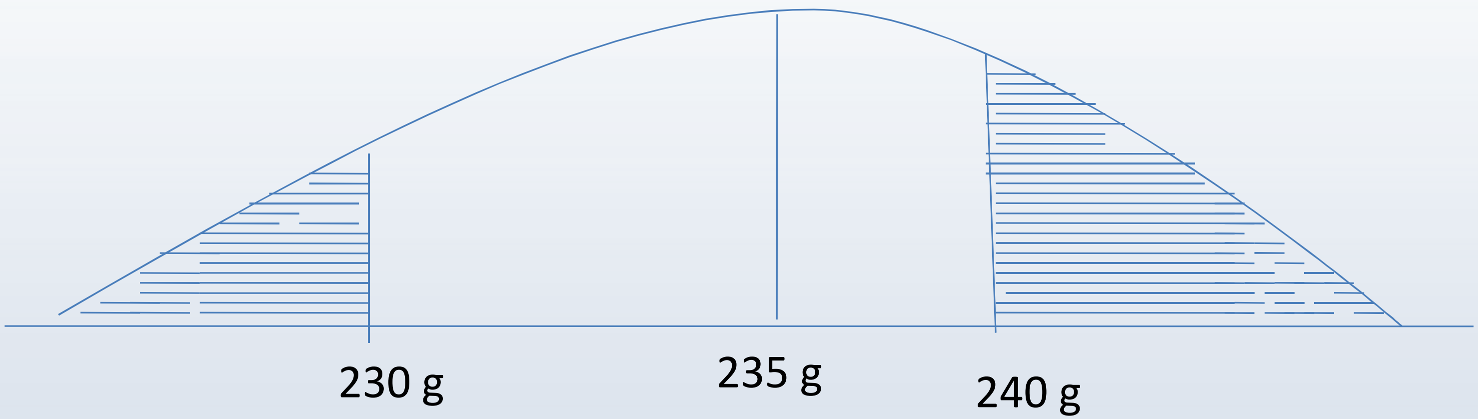
H_0 : Average amount of chips in the packets is 235 g
H_a : Average amount of chips in the packets is < 235 g



Hypothesis testing - Right tail test



Hypothesis testing - Two tail test



Steps in Hypothesis Testing

- Define null and alternative hypotheses
- Collect sample data
- Calculate test statistic
- Reject or fail to reject the null hypotheses
- Draw Conclusion

Hypothesis Testing

H_0 : null hypothesis

what you assume to be true

The null hypothesis is usually (always uses equal sign) a hypothesis of equality.

H_a or H_1: alternative hypothesis

what you attempt to demonstrate

The alternative hypothesis is usually an inequality

Null and Alternative Hypothesis

A manufacturer states that the average life of a car battery is 24 months.

Write the null and alternative hypothesis

Claim : = 24
Counter Claim : ≠ 24

Null Hypothesis

Alternative Hypothesis

H_0: = 24
H_1: ≠ 24

Null and Alternative Hypothesis

- A company claims that on an average the snack packet contain at least 100 g.
- Write the Null and Alternative Hypothesis

Claim :	≥ 100	$\Rightarrow$	Null Hypothesis
Counter Claim :	<100	$\Rightarrow$	Alternative Hypothesis

H_0: ≥ 100

H_1: <100

Null and Alternative Hypothesis

A ups manufacturer claims that the number of backup hours provided is more than 6 hours per day

Claim : > 6
Counter Claim : ≤ 6

Null Hypothesis
Alternative Hypothesis

H_0: ≤ 6
H_1: > 6

Null and Alternative Hypothesis

It's a general practice to use ≥ or ≤ as '=' sign in the null hypothesis

H_0: ≤ 6 H_1: > 6	Can also be written as	H_0: = 6 H_1: > 6
H_0: ≥ 6 H_1: < 6	Can also be written as	H_0: = 6 H_1: < 6

Type of tests and Rejection Region

The type of tests to be conducted based on the alternative hypothesis

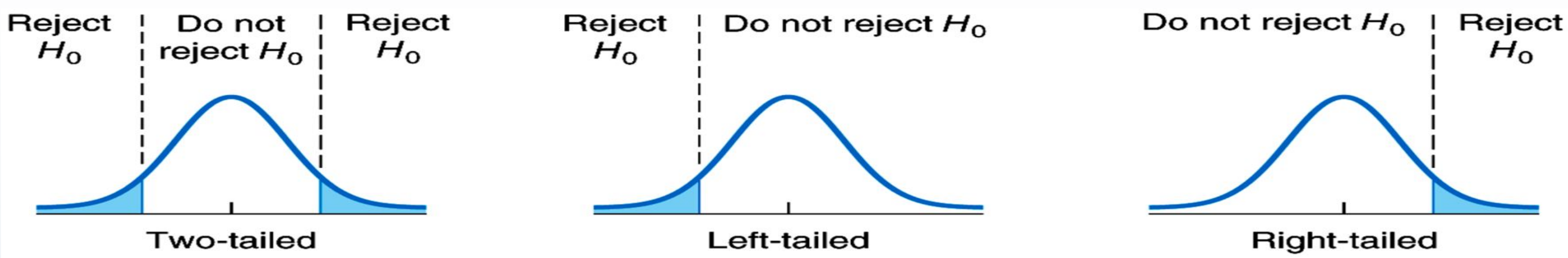
Left – tailed
Right - tailed
Two – tailed

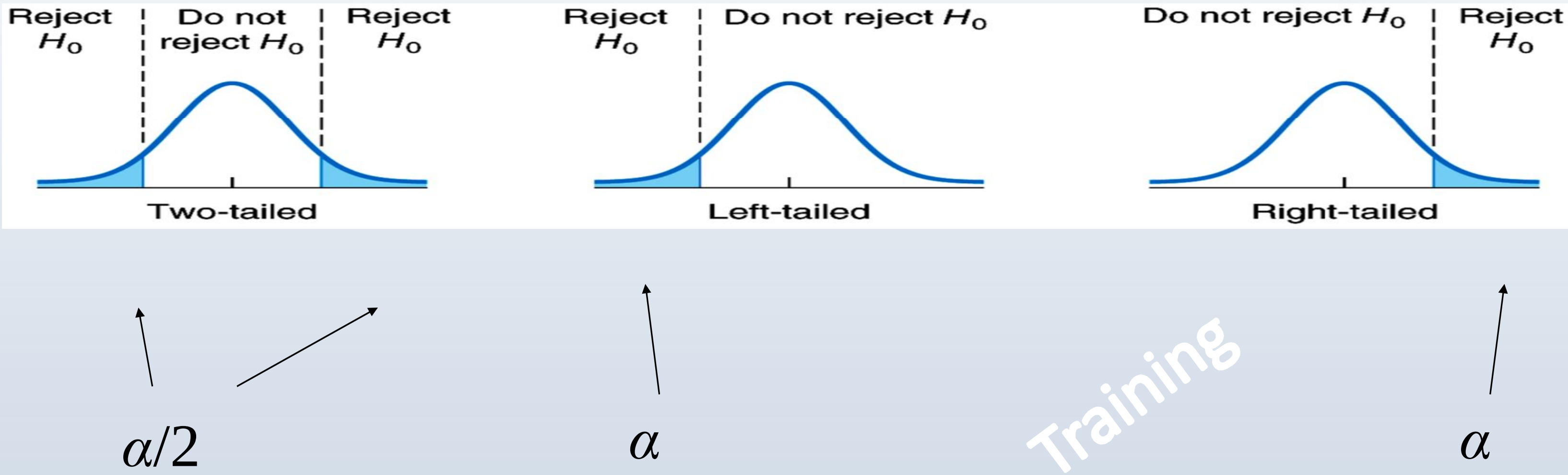
	Two-tailed test	Left-tailed test	Right-tailed test
Sign in H_a	$\neq$	$<$	$>$
Rejection region	Both sides	Left side	Right side

Hypothesis Testing for Single Population Mean

- **Two-Tailed Test**
  - Null hypothesis will be rejected if the sample mean is extremely large (upper tail) or extremely small (lower tail).
  - The alpha level would be split evenly between the two tails.
- **Left-Tailed Test**
  - Null hypothesis will be rejected if the sample mean is extremely small (lower tail).
  - The alpha level would be less than the left tail.
- **Right-Tailed Test**
  - Null hypothesis will be rejected if the sample mean is extremely large (upper tail).
  - The alpha level would be higher than the right tail.

Graphical display of rejection regions for left and right tailed plus two-tailed





Single Sample Hypothesis Z Test

When population sigma is known

Rejection Regions

	Alternative Hypotheses		
	Lower-Tailed	Upper-Tailed	Two-Tailed
$\alpha = .10$	$z < -1.28$	$z > 1.28$	$z < -1.645$ or $z > 1.645$
$\alpha = .05$	$z < -1.645$	$z > 1.645$	$z < -1.96$ or $z > 1.96$
$\alpha = .01$	$z < -2.33$	$z > 2.33$	$z < -2.575$ or $z > 2.575$

Hypotheses Testing – Detailed Steps

1. Define Null and Alt. Hypotheses
2. State alpha
3. Calculate degrees of freedom (applicable to T-Test only)
4. State Decision Rule
5. Calculate Test statistic
6. State Results
7. Draw conclusion

Single Sample Hypothesis z-test

As per the historical data, 6 years ago indicated that the average gross salary for a lead developer was $69,873. Since this data is now outdated., the human resource dept wishes to test this figure against current salaries to see if the current salaries are statistically different from the old ones. Sigma value is σ #13,985, number of samples taken are 112 and sample mean value is 72180.

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

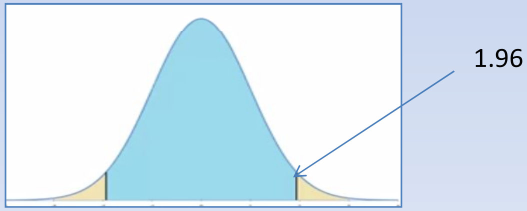
Step 1: Establish Hypothesis

H_0 : = $69,873 and H1 : ≠ $69,873

Step 2 : Determine appropriate Statistical Test and Sampling distribution

This will be a two-tailed test. Salaries could be higher OR lower . Since **σ** is known , we will use the z-distribution.

Step 3 : Specify the Type1 error rate(significance level)



(it is the probability of making a type 1 error that the user is willing to accept)

α = .05,

Subtract α from 1

1 - .05 = .95/2 = .475 (refer the z table to get the c.v)

Step 4 : State decision rule

If z > 1.96 , reject

If z < -1.96 , reject

Step 5 : Gather data

n = 112, $\overline{x}$ = \$72,180

=

Step 6 : Calculate test statistic

x = \$72,180, μ0 = \$69,873

σ = \$13,985
n = 112

$$z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

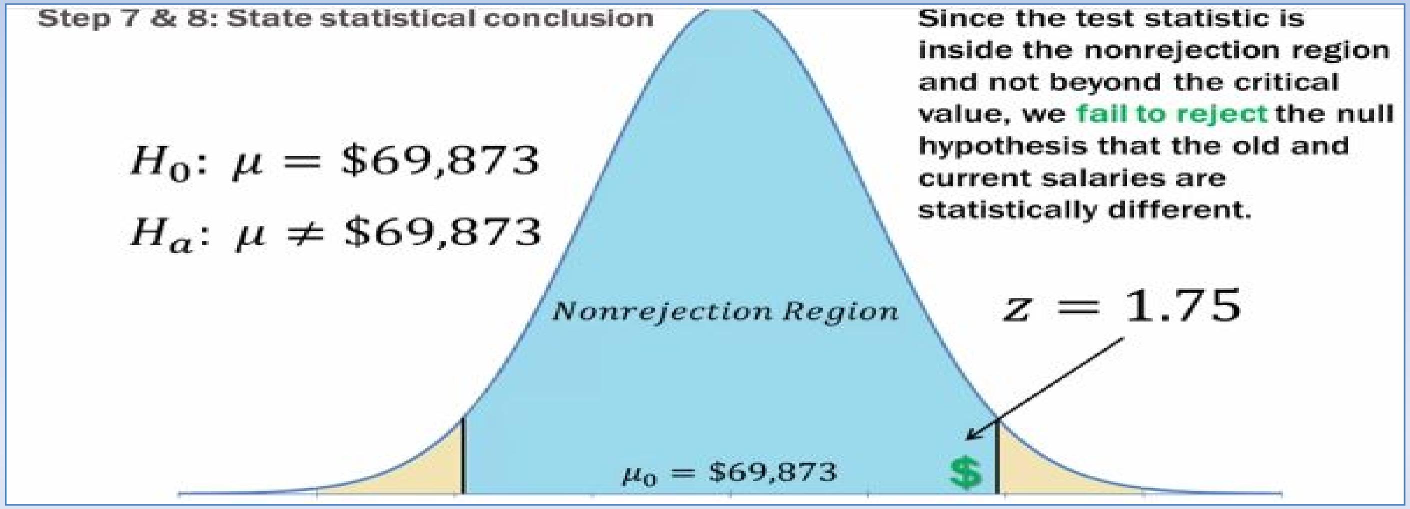
Z Distribution Table

Table of the Standard Normal (z) Distribution

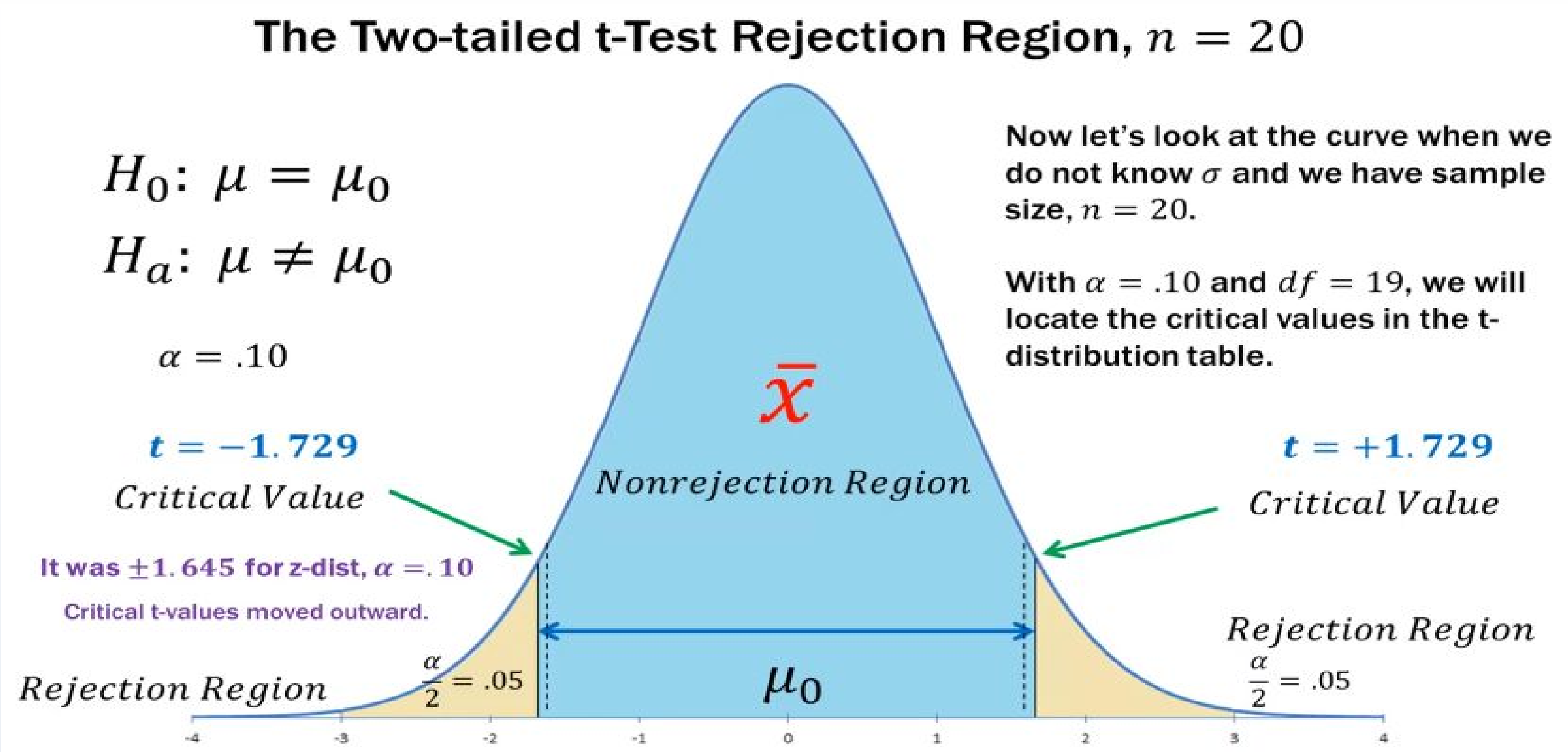
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1[illegible]72	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2[illegible]23	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2[illegible]54	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2[illegible]64	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3[illegible]51	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3[illegible]15	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3[illegible]54	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3[illegible]70	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3[illegible]62	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.41[illegible]31	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4[illegible]79	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.44[illegible]06	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.45[illegible]15	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	[illegible]	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	[illegible]	0.4693	0.4699	0.4706
1.9	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	[illegible]	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Single Sample Hypothesis z-test

- Z = \$72,180 - \$69,873 / \$13,985/√112
- Z = 1.75



Two-tailed t-Test



p value method

Assignment

A leading battery manufacturer claims the batteries life time will last for 30 hours , with a std deviation of 2.95 hours. However, many customers raised concerns that the batteries were lasting less than 30 hours. As a consumer you have drawn a random sample 38 of the specific brands batteries, whose mean life was 29.3 hours.

Assuming the life of the batteries in hours to be approximately normally distributed, decide whether the manufacturer's claim is valid

Solution

Let us take the null hypothesis that the mfr's claim is valid , that is,

Step 1 : Define Null and Alternative Hypothesis **Left tailed test**

$H_0 : \mu = 30$ hours

$H_a : \mu < 30$ hours

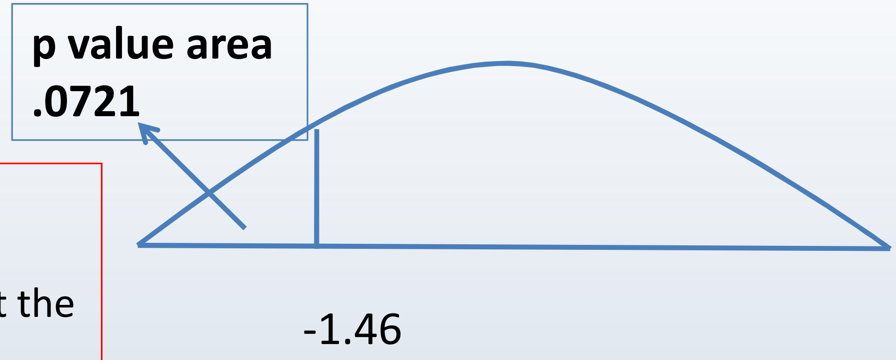
Step 2 : Level of Significance

$\alpha = .05$

Hypothesis Testing – p value method

Step 3 : Test statistic
Test statistic =z= -1.46
Step 4 : p-value 0.0722
Step 5 : Draw conclusion

Look up on a –z score table for -1.46

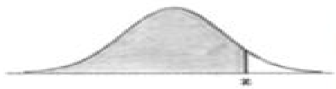


p is not less than alpha

Hence, there is not sufficient evidence to support the claim, at the 5% level of significance.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Normal Standard Distribution Table



Probability Content from -oo to Z

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

Hypothesis T Test

When population sigma is Not Known

t test (Student's t-test)

- The general rule of thumb for *when* to use a t score is when your sample size meets the following two requirements:
    - The sample size is below 30
    - The population standard deviation is unknown (estimated from your sample data)
- In other words, you **must** know the standard deviation of the **population** *and* your sample size **must** be above 30 in order for you to be able to use the z-score. Otherwise, use the t-score

Test Statistic for t test

$$t = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{standard error}}$$

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

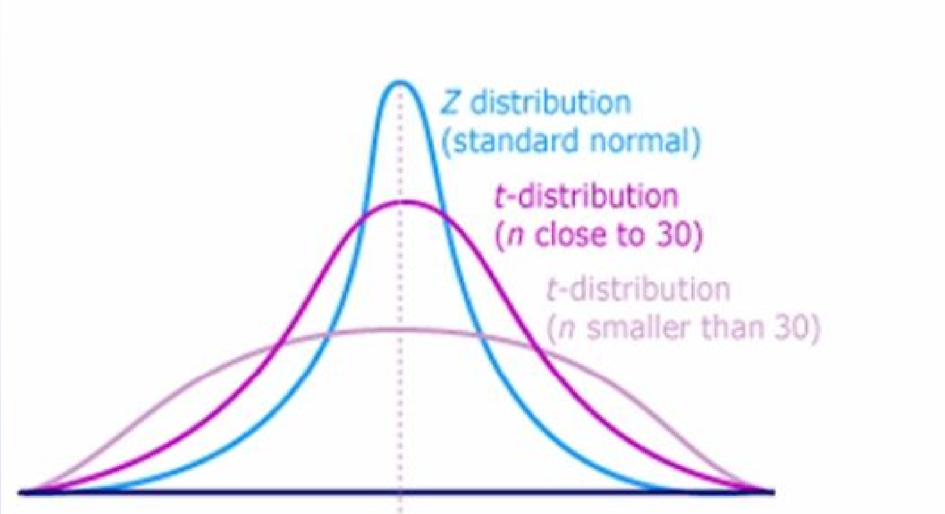
Hypothesis Testing – One Sample T-Test

For a random sample of size n (less than 30)

- The distribution of sample means has a t-distribution with n-1 degrees of freedom
- As a sample size increases and approaches 30, the t-dist approximates a normal distribution

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

Hypothesis Testing T- Test



1. Define Null and Alt. Hypotheses
2. State alpha
3. Calculate degrees of freedom
4. State Decision Rule
5. Calculate Test statistic
6. State Results
7. Draw conclusion

Single Sample Hypothesis t-test

As per the historical data, 6 years ago indicated that the average gross salary for a lead developer was $69,873. Since this data is now outdated., the human resource dept wishes to test this figure against current salaries to see if the current salaries are statistically different from the old ones. Number of samples taken are 12 and sample mean 79180.

Step 1: Establish Hypothesis

H_0 : = $69,873 and H1 : ≠ $69,873

Step 2 : Determine appropriate Statistical Test and Sampling distribution

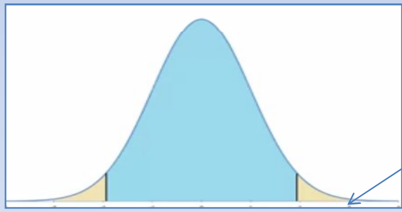
This will be a two-tailed test. Salaries could be higher OR lower . Since **σ** is unknown , we will use the t-distribution.

Step 3 : Specify the Type1 error rate(significance level)
(it is the probability of making a type 1 error that the user is willing to accept)

α = .05,

Step 4 : State decision rule

If t > 2.201 , reject
If t < -2.201 , reject



1.96

Step 5 : Gather data

n = 12, df= 11 (n-1) $\bar{x}$ = $79,180

Step 6 : Calculate test statistic

x = **$79,180, μ0 = $69,873**

σs = $14,985

n = 12

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

Single Sample Hypothesis t-test

$$t = \frac{79{,}180 - 69873}{14{,}985/\text{sqrt}(12) \Rightarrow 3.46} = \frac{9307}{4325.79} = 2.15$$

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

t = 2.15

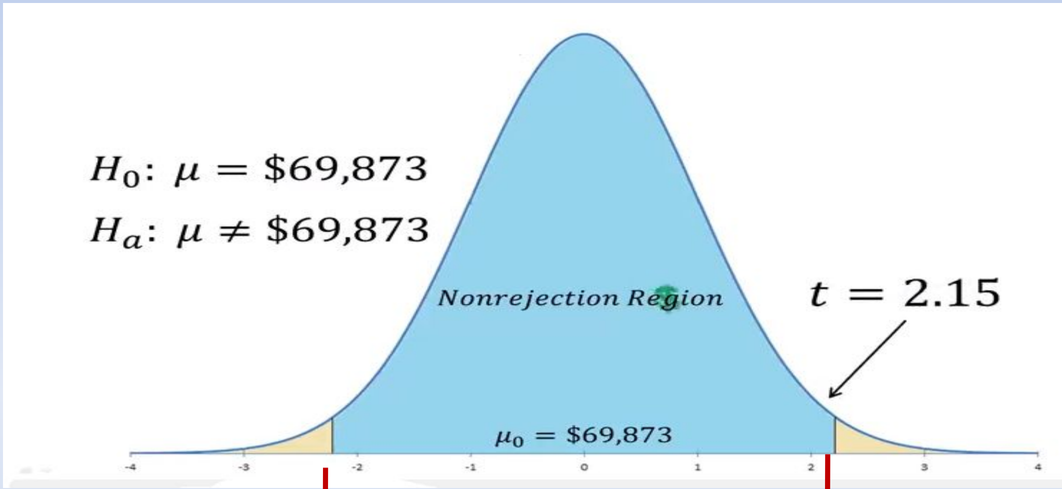
n = 12

df = n-1 = 11

C.value

t > 2.201 and

t < -2.201



- 2.201 **2.201**

T Table

Table T Critical Values of the *t* Distribution

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745

T-Test

X bar = $79,180

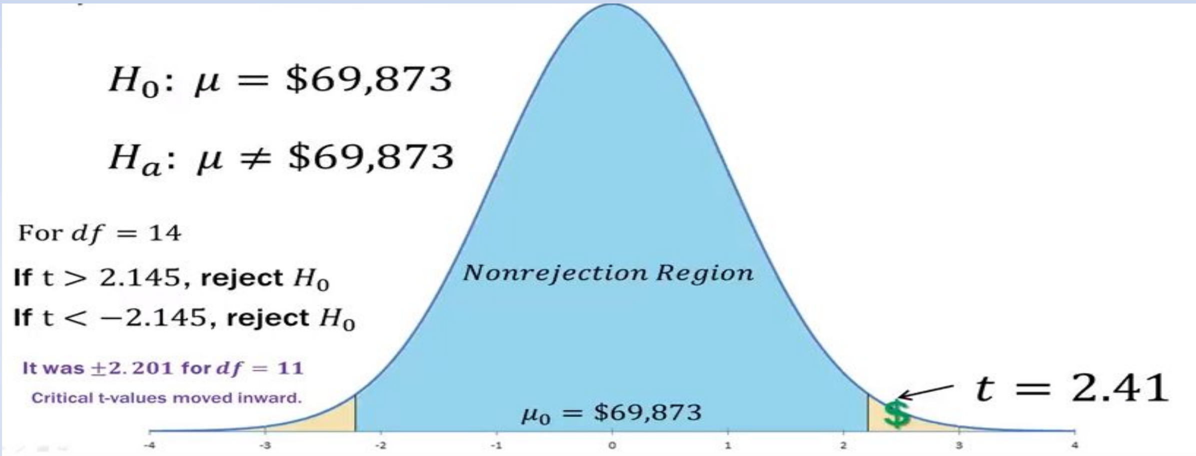
mu 0 = $69873

s = $14985

n = 15

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

t= 2.41



T Table

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883

Quiz

What is true about the t-distribution ?

(a) Increasingly resembles the normal distribution as degrees of freedom increase

(b) Assumes the population is normally distributed

(c) It has a greater spread than the normal distribution

(d) All of the above

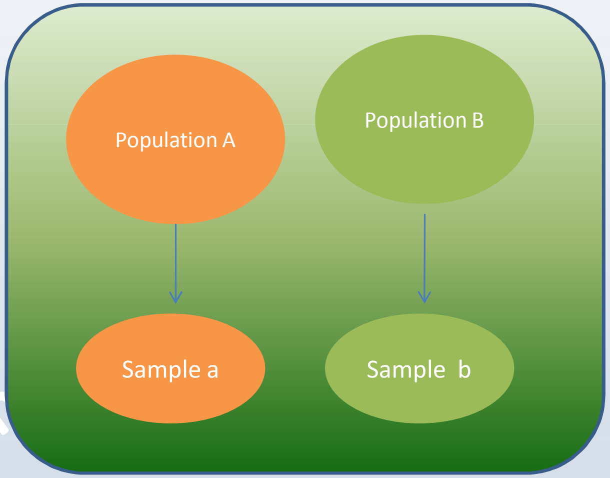
Two Sample Independent T-Test

Two sample Independent T test

To compare the mean values of two normally distributed populations, draw independent random samples of sizes n1 and n2 from the two populations

Estimate the mean value of difference between $_1$ and $_2$ ($_1 -$ 2) two Populations

- We are interested in the difference between two independent groups. As such, we are comparing Two populations by evaluating the mean difference
- To evaluate the mean difference between two populations , we sample from each population and compare the sample means on a given sample
- Must have two independent groups and dependent variable that is continuous to compare them



Two sample independent T test

Do males and females significantly differ on the amount of time spent on the website ?

Do South and North region significantly differ on the amount of sales achieved ?

The test statistic we use here is

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

$$SE(\overline{Y}_1 - \overline{Y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}}$$

Here Y1 , Y2 are group 1 , group 2 respectively

Assumptions

- The two groups are independent of one another
- The dependent variable is normally distributed.
  - Examine skewness and kurtosis (peak) of distribution
- The two groups have approximately equal variance on the dependent variable (when n1=n2 [equal sample sizes], the violation of this assumption has been shown to be unimportant.)

Independent Samples T- test

A Marketing manager wants to compare two of his Sales groups to see if the both performed differently in
The last month. Group A had 25 sales persons with an average sales of 70, standard deviation 15. Group B had 20
sales persons with an average sales of 74, standard deviation of 25. Using alpha 0.05 , did these two groups
Perform differently on the sales ?

1. Define Null and Alt. Hypotheses
2. State alpha
3. Calculate degrees of freedom
4. State Decision Rule
5. Calculate Test statistic
6. State Results
7. Draw conclusion

Independent 2 samples t-test

$s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$

Step 1 : H_0 : 1 = 2

H1 : 1 != 2

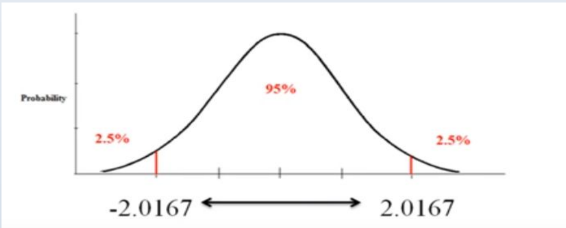
Step 2 : Alpha = 0.05

Step 3 : df = (n-1) + (n-2)

df = (25-1) +(20-1)

= 43

Step 4 : State Decision Rule



If t is less than -2.0167 or greater than 2.0167, reject the H0

Step 5 : Calculate Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

variance

Whereby:

n: Sample size s^2 = variance df

S_p^2 Pooled Variance

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

$$df_1 = n_1 - 1 = 25 - 1 = 24$$

$$df_2 = n_2 - 1 = 20 - 1 = 19$$

$$SS_1 = s_1^2(df_1) = (15^2)(24) = 5400$$

$$SS_2 = s_2^2(df_2) = (25^2)(19) = 11875$$

$$s_p^2 = \frac{5400 + 11875}{24 + 19} = 401.74$$

$s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ = sqrt(401.74) * sqrt(1/25 +1/20) = (70-74)/ (20.04* 0.3)

= -0.67

Step 6 : t = 0.67 , Do not reject H_0,

Hence there was no sig. diff between the Sales perf. of both group a and group b

Calculate confidence interval

70 – 74 +- 2.02 * 6.01 = -4 + (2.02 * 6.01) = -48.56

= -4 – (2.02 * 6.01) = 48.56

We are 95% confident that , on average, Group A made between -48.56 and 48.56 sales than Group B

➢ t Table

df	Critical Values				
1-tailed	0.1	0.05	0.025	0.01	0.005
2-tailed	0.2	0.1	0.05	0.02	0.01
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969

df	Critical Values				
1-tailed	0.1	0.05	0.025	0.01	0.005
2-tailed	0.2	0.1	0.05	0.02	0.01
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045
41	1.3025	1.6829	2.0195	2.4208	2.7012
42	1.3020	1.6820	2.0181	2.4185	2.6981
43	1.3016	1.6811	2.0167	2.4163	2.6951
44	1.3011	1.6802	2.0154	2.4141	2.6923
45	1.3006	1.6794	2.0141	2.4121	2.6896
46	1.3002	1.6787	2.0129	2.4102	2.6870
47	1.2998	1.6779	2.0117	2.4083	2.6846
48	1.2994	1.6772	2.0106	2.4066	2.6822
49	1.2991	1.6766	2.0096	2.4049	2.6800
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259

Two sample independent test

Sales achieved by 2 different sales persons in the 9 months are as follows:

Sales Person 1 : 18 20 36 50 49 36 34 49 41

Sales Person 2 : 29 28 26 35 30 44 46 0 0

Using alpha 0.05 , did these two persons perform differently with respect to the sales

H0 : $_1 - \; 2 = 0$ or $_1 = \; 2$

Ha : $_1 \neq \; 2$

T Table

df			Critical Values		
1-tailed	0.1	0.05	0.025	0.01	0.005
2-tailed	0.2	0.1	0.05	0.02	0.01
1	3.0777	6.3138	12.7062	31.8205	63.6567
2	1.8856	2.9200	4.3027	6.9646	9.9248
3	1.6377	2.3534	3.1824	4.5407	5.8409
4	1.5332	2.1318	2.7764	3.7469	4.6041
5	1.4759	2.0150	2.5706	3.3649	4.0321
6	1.4398	1.9432	2.4469	3.1427	3.7074
7	1.4149	1.8946	2.3646	2.9980	3.4995
8	1.3968	1.8595	2.3060	2.8965	3.3554
9	1.3830	1.8331	2.2622	2.8214	3.2498
10	1.3722	1.8125	2.2281	2.7638	3.1693
11	1.3634	1.7959	2.2010	2.7181	3.1058
12	1.3562	1.7823	2.1788	2.6810	3.0545
13	1.3502	1.7709	2.1604	2.6503	3.0123
14	1.3450	1.7613	2.1448	2.6245	2.9768
15	1.3406	1.7531	2.1314	2.6025	2.9467
16	1.3368	1.7459	2.1199	2.5835	2.9208
17	1.3334	1.7396	2.1098	2.5669	2.8982
18	1.3304	1.7341	2.1009	2.5524	2.8784
19	1.3277	1.7291	2.0930	2.5395	2.8609
20	1.3253	1.7247	2.0860	2.5280	2.8453
21	1.3232	1.7207	2.0796	2.5176	2.8314
22	1.3212	1.7171	2.0739	2.5083	2.8188
23	1.3195	1.7139	2.0687	2.4999	2.8073
24	1.3178	1.7109	2.0639	2.4922	2.7969
25	1.3163	1.7081	2.0595	2.4851	2.7874
26	1.3150	1.7056	2.0555	2.4786	2.7787
27	1.3137	1.7033	2.0518	2.4727	2.7707
28	1.3125	1.7011	2.0484	2.4671	2.7633
29	1.3114	1.6991	2.0452	2.4620	2.7564

df			Critical Values		
1-tailed	0.1	0.05	0.025	0.01	0.005
2-tailed	0.2	0.1	0.05	0.02	0.01
32	1.3086	1.6939	2.0369	2.4487	2.7385
33	1.3077	1.6924	2.0345	2.4448	2.7333
34	1.3070	1.6909	2.0322	2.4411	2.7284
35	1.3062	1.6896	2.0301	2.4377	2.7238
36	1.3055	1.6883	2.0281	2.4345	2.7195
37	1.3049	1.6871	2.0262	2.4314	2.7154
38	1.3042	1.6860	2.0244	2.4286	2.7116
39	1.3036	1.6849	2.0227	2.4258	2.7079
40	1.3031	1.6839	2.0211	2.4233	2.7045
41	1.3025	1.6829	2.0195	2.4208	2.7012
42	1.3020	1.6820	2.0181	2.4185	2.6981
43	1.3016	1.6811	2.0167	2.4163	2.6951
44	1.3011	1.6802	2.0154	2.4141	2.6923
45	1.3006	1.6794	2.0141	2.4121	2.6896
46	1.3002	1.6787	2.0129	2.4102	2.6870
47	1.2998	1.6779	2.0117	2.4083	2.6846
48	1.2994	1.6772	2.0106	2.4066	2.6822
49	1.2991	1.6766	2.0096	2.4049	2.6800
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
110	1.2893	1.6588	1.9818	2.3607	2.6213
120	1.2886	1.6577	1.9799	2.3578	2.6174
130	1.2881	1.6567	1.9784	2.3554	2.6142
140	1.2876	1.6558	1.9771	2.3533	2.6114
150	1.2872	1.6551	1.9759	2.3515	2.6090

Paired (Dependent) Sample T-Test

Paired Sample T-Test

What is Paired Sample t-test

Paired sample t-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated.

When to use ?

Paired sample t-test is used in ‘before-after’ studies, or when the samples are the matched pairs, or when it is a case-control study.

Assumptions:

- Only the matched pairs can be used to perform the test.
- Normal distributions are assumed.
- The variance of two samples is equal.
- Cases must be independent of each other.

$$t = \frac{\bar{d}}{S_d / \sqrt{n}}$$

Mean of the difference between paired observations

Sample std deviation of the distribution of the difference
Between the paired observations

Paired T test – Example 1

Is there a significant difference in students score when taught through online and class room training ?

What is an independent variable ?

Mode of training

What is a dependent variable ?

Score

Paired T Test – Example 2

Is there a significant difference in people's pre-exercise weight and post-exercise weight ?

What is an independent variable ?

Time : pre and post

What is a dependent variable ?

Time : Continuous

An Enterprise gave training to its employee and want to know whether or not the training had any impact on the efficiency of the employee, we could use the paired sample test. We collect data from the employee on a seven scale rating, before the training and after the training. By using the paired sample t-test, we can statistically conclude whether or not training has improved the efficiency of the employee.

By using the paired sample t-test, we can statistically conclude whether or not training has improved The efficiency of the employee.

Paired sample T Test

Training Participant	Before Training	Post Training
Joe	50	62
Ashok	42	40
John	51	61
Madhu	26	35
Raj	35	30
Angelin	42	52
Swetha	60	68
Bob	41	51
Mike	70	84
Jones	55	63
Milea	62	72
Sahoo	38	50

$$\overline{D} = \frac{\sum D}{n}$$ = 96/12 = 8

Sd = Sqrt(∑d^2/(n-1) – (∑d)^2/n (n-1))

Sd= sqrt((1122/11)– (96)^2/(12*11) = 5.673

Training Participant	Before Training	Post Training	Difference in score , d	d^2
Joe	50	62	12	144
Ashok	42	40	-2	4
John	51	61	10	100
Madhu	26	35	9	81
Raj	35	30	-5	25
Angelin	42	52	10	100
Swetha	60	68	8	64
Bob	41	51	10	100
Mike	70	84	14	196
Jones	55	63	8	64
Milea	62	72	10	100
Sahoo	38	50	12 → 96	144 → 1122

$$t = \frac{\overline{d}}{S_d / \sqrt{n}} = \frac{8}{5.673/\text{sqrt}(12)} = 4.885$$

df = 11 and alpha/2 = 0.025 , critical value is = 2.20 ,
t value is 2.20 is more than the critical value.

Summary

One sample t-test
It is used in measuring whether a sample value significantly differs from a hypothesized value.

Independent Samples t-test
It is used in comparing the means for two different groups.

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

Types of Errors and Power

Actual		
DECISION	**H_0 is True**	**Ha is true**
Reject Null (Fail to accept the null Hypothesis) – We claim something happened	**Type I Error (False Positive α)**	Correct Outcome True Positive We reject the Null Hyp.
Fail to Reject Null (We claim nothing happened) OR (Do not reject Ho)	Correct Outcome True negative Accept the Ho	**Type II Error (False Negative β)**

Reject a true hypothesis (Type 1 error)

Accept a false hypothesis (Type II error)

Example Good or Bad customer ? H_0: They're Good customer H_A : bad customer

Population ➔	**H_0 is TRUE**	**Ha is TRUE**
we say it's , if it's →	**Good**	**Bad**
Bad (Reject the null hypothesis)	**Type I – False Positive α)**	OK- True Positive
Good (Do not reject)	Ok (True Negative)	**Type II – (False Negative β)**

Z Distribution Table

Table of the Standard Normal (z) Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	[illegible]	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	[illegible]	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

CASE STUDY - RETAIL COMPANY

A large retail chain is interested in assessing customer satisfaction in the Indian city of Mumbai and Chennai. To conduct the study, mightyvel asked 225 customers in the city:

“Compared to other retail shops in Mumbai, would you say the customer service at mightyvel is much better than average(5), better than average(4), average (3), worse than average(2), or much worse than average(1)?” (Likert scale)

The mean rating was determined to be 3.50. Based on previous studies done by the company, the std devn s = 1.4

CASE STUDY - RETAIL

Step 1: Establish Hypothesis

H_0 : <= 3 and Ha : > 3

Step 2: Determine appropriate Statistical Test and Sampling Distribution

This will be one-tailed test. Mightyvel is interested in a better than average rating.

Since sigma is known, we will use the z-distribution

t = (x bar – mu not) / s / sqrt (n)

Step 3: Specify the Type 1 error rate (significance level)

α = .01

Step 4 : State the decision rule

df = 24

If t > 2.492 , reject H_0

Step 5 : Gather data

n = 25, x bar = 3.5

Step 6 : Calculate test statistic

x bar = 3.50 , mu no = 3 sigma s = 1.4 n = 25 =→>> 1.79

Step 7 & 8 : State Statistical conclusion

Since the test statistic is in the rejection region and beyond the critical value, we reject the null hypothesis that customer satisfaction is at or below average

What is mean customer satisfaction ?

t = 3.50 – 3 / 1.4/ sqrt 25 = >> 2.492 = x bar – 3 / 1.4/sqrt(25) → x bar = 3.698

Therefore any sample n = 25 sample with x bar > 3.698 would lead to a rejection of H_0 assuming a constant sigma & alpha.

Sa

CASE STUDY – Hypothesis Testing in RETAIL - using R

Step 1: Establish Hypothesis

H_0 : <= 3 and Ha : > 3

Step 2: Determine appropriate Statistical Test and Sampling Distribution

This will be one-tailed test. Mightyvel is interested in a better than average rating.

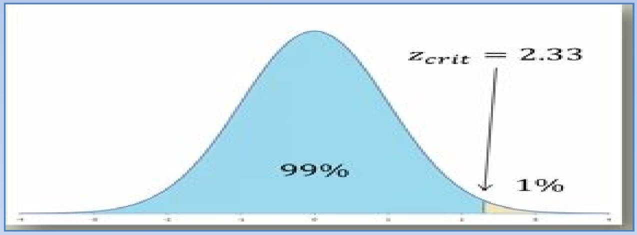
Since sigma is known, we will use the z-distribution

z = (xbar – muo) / (sigma / sqrt (n)) z = sample_mean – population_mean/ std_err_of_mean

Step 3: Specify the Type 1 error rate (significance level)

α = .01

alpha = .01



Step 4 : State the decision rule

If z_α > 2.33 , reject H_0

Step 5 : Gather data

n = 255, x bar = 3.25

Step 6 : Calculate test statistic

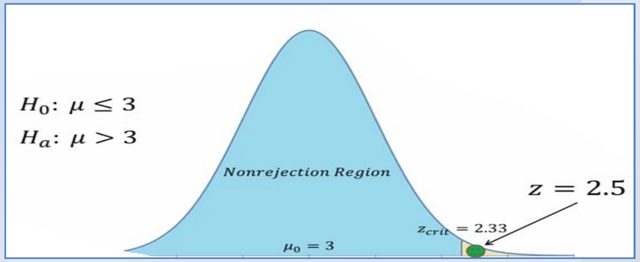
x bar = 3.25 , muo = 3 sigma = 1.5 n = 225 =➔>> 2.5

Compute the critical value

alpha = .01

z.alpha = qnorm(1-alpha)

-z.alpha # critical value



Step 7 & 8 : State Statistical conclusion

Since the test statistic is in the rejection region and beyond the Critical value, we reject the null hypothesis that customer satisfaction is at or below average

What is mean customer satisfaction ?

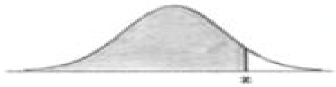
z = 3.25 – 3 / 1.5/ sqrt 225 = >> 2.33 = x bar – 3 / .1 ➔ x bar = 3.233

Therefore any sample n = 225 sample with x bar > 3.233 would lead to a rejection of H_0 assuming a constant sigma & alpha.

p-value method

This area is called p value (look up from Normal distribution table)
(1-0.9633) => .0367

Normal Distribution Table



Probability Content
from -oo to Z

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

CASE STUDY – Hypothesis Testing in RETAIL - using R

RGui (32-bit)

File Edit View Misc Packages Windows Help

R Console

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> xbar = 3.25
> alpha = .01
> mu0 = 3
> sigma = 1.5
> n = 225
> z = (xbar - mu0)/ (sigma / sqrt(n))
> z
[1] 2.5
> z.alpha = qnorm(1-alpha)
> z.alpha
[1] 2.326348
> z.alpha      # Critical Value
[1] 2.326348
> z # test statistic
[1] 2.5
>
```

T Distribution table

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896

T Distribution table

t Distribution

Degrees of freedom	.005 (one tail) .01 (two tails)	.01 (one tail) .02 (two tails)	.025 (one tail) .05 (two tails)	.05 (one tail) .10 (two tails)	.10 (one tail) .20 (two tails)	.25 (one tail) .50 (two tails)
	α					
1	63.657	31.821	12.706	6.314	3.078	1.000
2	9.925	6.965	4.303	2.920	1.886	.816
3	5.841	4.541	3.182	2.353	1.638	.765
4	4.604	3.747	2.776	2.132	1.533	.741
5	4.032	3.365	2.571	2.015	1.476	.727
6	3.707	3.143	2.447	1.943	1.440	.718
7	3.500	2.998	2.365	1.895	1.415	.711
8	3.355	2.896	2.306	1.860	1.397	.706
9	3.250	2.821	2.262	1.833	1.383	.703
10	3.169	2.764	2.228	1.812	1.372	.700
11	3.106	2.718	2.201	1.796	1.363	.697
12	3.054	2.681	2.179	1.782	1.356	.696
13	3.012	2.650	2.160	1.771	1.350	.694
14	2.977	2.625	2.145	1.761	1.345	.692
15	2.947	2.602	2.132	1.753	1.341	.691
16	2.921	2.584	2.120	1.746	1.337	.690
17	2.898	2.567	2.110	1.740	1.333	.689
18	2.878	2.552	2.101	1.734	1.330	.688
19	2.861	2.540	2.093	1.729	1.328	.688
20	2.845	2.528	2.086	1.725	1.325	.687
21	2.831	2.518	2.080	1.721	1.323	.686
22	2.819	2.508	2.074	1.717	1.321	.686
23	2.807	2.500	2.069	1.714	1.320	.685
24	2.797	2.492	2.064	1.711	1.318	.685
25	2.787	2.485	2.060	1.708	1.316	.684

Hypothesis Testing (Summary)

Type Of Test	Purpose	Example	Equation	Comment	Excel Function
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average	$Z = \frac{\bar{x} - u_0}{\frac{\sigma}{\sqrt{n}}}$	Z test does not need df σ = population standard deviation	=Ztest(array,x,sigma)
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?	$t = \frac{\bar{x} - u_0}{\frac{s}{\sqrt{n}}}$ $df = n - 1$	s = sample standard deviation	no built in equation use =STDEVA for standard deviation use =AVERAGE for mean use =T.DIST.RT to get 1 tailed confidence use =T.DIST.2T to get 2 tailed confidence
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?	$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$ $df = n - 1$	d bar = average difference between samples s = sample deviation of the difference n = count of one set of the pairs (don't double count)	=TTEST(Array1,Array2,*,1) * -> 1 for 1 tailed, 2 for 2 tailed
2 Sample T-Test Equal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food	$df = n_1 + n_2 - 2$ $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$	n1, n2 = count of sample 1, 2	=TTEST(Array1,Array2,*,2)
2 Sample T-Test Unequal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$	=TTEST(Array1,Array2,*,3)

Hypothesis Testing - Quiz

In a leading soft drink manufacturing unit , the machine fills the cans with an average of 100 ml. If the machine is either overfilling or under filling the machine has to be shut down.

As a QC staff you need to ensure the appropriate quantity of soft drink has to be filled in each can.

(a) State the null and alternative hypothesis

(b) Explain if a Type 1 error occurs what would be potential consequences

(c) If you make a Type 2 error , in the context of this problem, explain the potential consequences

(d) When do you suppose to shut down and fix the machine; when the H0 is rejected or is not rejected ?

Anova (Analysis of Variance)

ANOVA

Anova helps us to compare the means of two or more groups of observations, or treatment

Anova

The overall variance of the population is divided into two groups
Within group variance
Between group variance

Within Group Variance	Sum of squared differences between each observation and the mean of the group it belongs to	Sum of squares within SSW
Between Groups Variance	Sum of squared differences between each group mean and the overall mean	Sum of squares between SSB
Overall Variance	Total sum of squared differences between observation the overall mean of all observations	Total sum of squares SST

ANOVA – One Way (also called as One Factor Anova)

Variation Within Groups

Shelve Top	Shelve mid	Front Shelve
2	10	10
3	8	13
7	7	14
2	5	13
6	10	15
4	**8**	**13**
22	**18**	**14**

Variation Between Groups

Sum of Squares

Sum of Squares Between Groups

Dependent Variable : Sales
Independent Variable : Shelf position

Dependent Variable : depends on other factors , for instance, shelf height etc.,
Independent Variable : doesn't changed by other variables

Null Hypothesis would be that all means are equal

H_0 : mean1=means2=mean3

H_A: mean1<>mean2<>mean3

Total sum of squares = Sum of Squares Between Groups + Sum of Squares within groups
SS = Sum of Squares
Df = Degrees of Freedo
MS = Mean Sum of Squares = SS/df
Fstatistics = Between Groups SS/ Within Groups SS

Anova – Manual

Sales
2
3
7
2
6
10
8
7
5
10
10
13
14
13
15
8.3
257.3

Total Sum of Squares

=

Variation Within Groups

Shelve Top	Shelve mid	Front Shelve
2	10	10
3	8	13
7	7	14
2	5	13
6	10	15
4	**8**	**13**
22	**18**	**14**

+

Variation Between Groups

Shelve Top	Shelve mid	Front Shelve
2	10	10
3	8	13
7	7	14
2	5	13
6	10	15
4	**8**	**13**
22	**18**	**14**

54 + 203.3

Total sum of squares = Sum of squares within groups + sum of squares Between groups

Sum of squares within groups = 22 + 18 + 14 = 54

Sum of squares between groups = (4-8.3)^2 + (8-8.3)^2 + (13-8.3) ^2

= 18.8 + .1 + 21.8 => 40.7 * 5 => 203.3

257.3 = 203.3 + 54

$\frac{\text{Sum of squares between groups}}{\text{degrees of freedom}}$ = 203.3/2 = 101.667 , $\frac{\text{S.S.W.G}}{\text{df}}$ = (54/12) = **4.5**

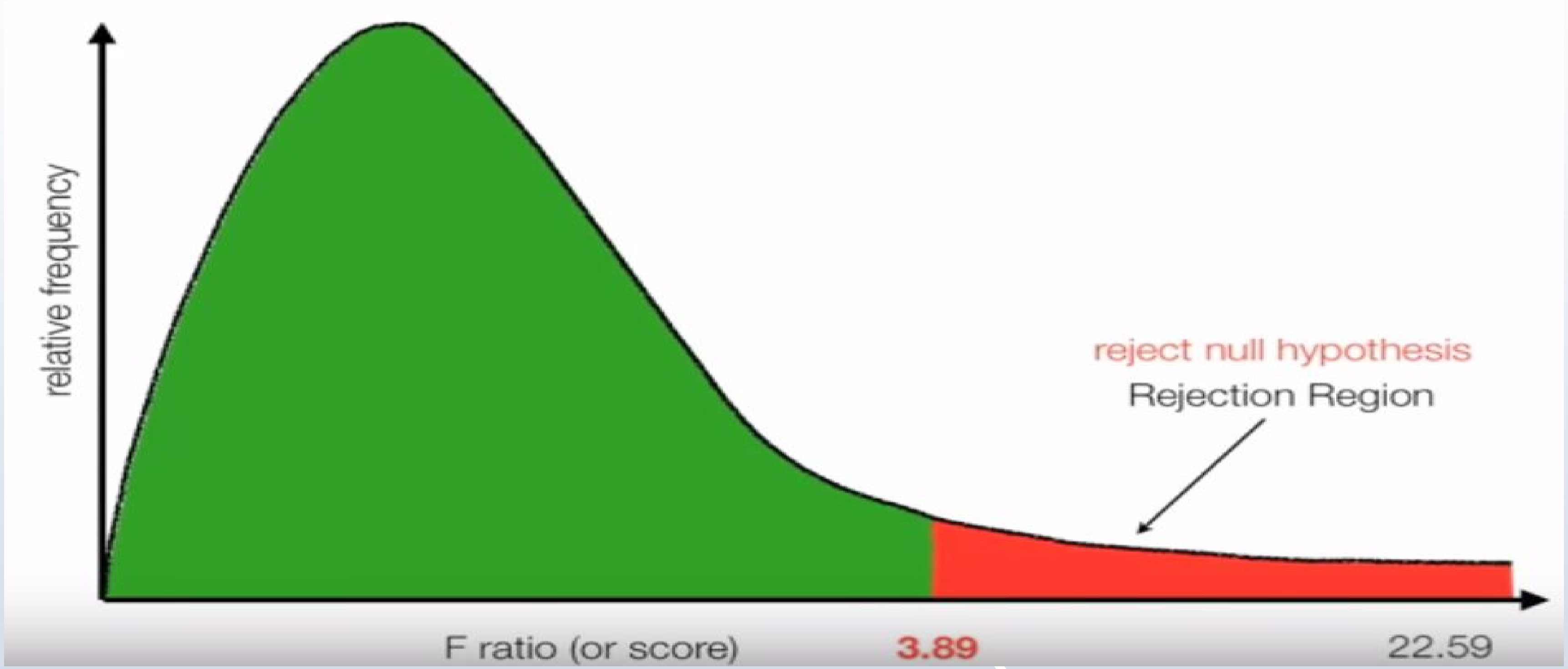
F = (101.667 / 4.5) ➔ 22.59 F(2,12) = 22.59 C.V = 3.89

F Table

F critical values (continued)

Degrees of freedom in the denominator	p	Degrees of freedom in the numerator: 1	2	3	4	5	6	7	8	9
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
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.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
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58

F Distribution



Anova Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed
- The samples must be independent
- The variances of the populations must be equal

Non Parametric Testing – Chi Square χ^2 Test

Chi Square χ^2 Testing

- It's a statistical method assessing the goodness of fit between a set of observed values and those expected theoretically

Test formula for goodness of fit

$$\chi^2 = \sum \frac{(o-e)^2}{e}$$

o is observed frequency
e is expected frequency

Properties

Its based on frequencies and not on the parameters like mean and standard Deviation

Applications of Chi Square test

- **Association Analysis** - Enables us to explain whether or not two attributes are associated
  For instance, to know whether a new drug is effective in controlling blood pressure or not , chi sq test is useful
- **Goodness of fit of distributions**
- **Test of homogenity** - to test whether the occurrence of events follow uniformity or not

Chi Square Test - Example

Males	Females
yes	no
yes	yes
no	no
yes	yes
no	yes
no	yes
yes	no
yes	yes
no	yes
no	no
yes	no
yes	no

Null Hyp : No relationship between gender having mobile device
Alt Hyp : There is a relationship

Step1

OBSERVED			
Genders	yes	no	
Males	7	5	12
Females	6	6	12
Total	13	11	24

Step 2 - Derive the expected value

EXPECTED			
Genders	yes	no	
Males	6.5	5.5	12
Females	6.5	5.5	12
Total	13	11	24

Expected Value is obtained by using below
E.V = (Column total * Row Total) / Grand Total

Step 3 – Compute the p value

=CHITEST(I5:J6,I12:J13) ➔ p value is 0.68
Which is less than alpha value which is .05

Step 4 – Draw Conclusion

Conclusion : There is no relationship between gender having mobile device

Chi Square - Quiz

Type of data do you need for a chi-square test?

(A) Scales
(B) Ordinal
(C) Parametric
(D) Categorical
(E) Ratio
(F) Interval

What does a significant result in a chi-square test imply?

(A) That there is a significant difference between the three categorical variables included in the analysis
(B) That there is a significant negative relationship
(C) That homogeneity of variance has not been established
(D) That there is a significant positive relationship
(E) It implies that the sample is not representative of the population
(F) All of these are possible

What would a chi-square significance value of P > 0.05 suggest?

(A) That there is no significant difference between categories
(B) That there is no significant difference between the sample and the population
(C)That there is a significant relationship between the sample and the population
(D) That there is a significant relationship between categorical variables
(E) That there is a significant difference between the sample and the population
(F) That there is no significant difference between time 1 and time 2

Chi Square - Quiz

Chi-square test for independence assesses which of the following?

- It assesses whether there is a relationship between the population and the sample
- It assesses whether the minimum number of cases exceeds recommended boundaries
- It assesses whether there is a significant difference between two categorical variables
- It assesses whether there is a relationship between two categorical variables
- It assesses whether there is significant difference between scores taken at time 1 and those taken at time 2
- None of these

Case Study – Hypothesis Testing, Anova One-way , Two- way and Chi Square test

Hypothesis Testing

Lets say you are drawing a sample of 200 retail chain customers from a population in order to perform spend analysis but you can't afford using the entire Data set.

The sample which you have drawn must be a representative of your population , in order to confirm the same you take average on the attribute AmountSpent and compare the same with the population average.

If the average amount spent is not same as your population data , what should be done ?

Case Study –Two- way Anova (Two-Factor Anova)

Anova

Anova is used when we want to find out the impacts of discrete factors on a continuous outcome variable

For instance, a retail chain's marking analyst wants to understand, Is there any impact of Sales persons on sales ?

He takes a sample of sales persons performance, and looks at average sales achieved by different sales persons.

Two–Way Anova without Replication

More than one factor variable influence a dependent variable, you can use Two-way analysis of variance

Regions	Rep1	Rep2	Rep3	Rep4
South	2	4	12	14
North	18	14	18	16
East	18	23	24	28
West	22	12	19	24
South East	24	24	39	34

Here the **dependent variable** is **sales** & the **Independent variables** have **two factors**

Null Hyp : All the group mean sales achieved are identical
Alt Hyp : All the group mean sales are not identical

Two–Way Anova without Replication

Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance	
Row 1	4	32	8	34.66666667	-11.45
Row 2	4	66	16.5	3.666666667	16.5
Row 3	4	93	23.25	16.91666667	23.25
Row 4	4	77	19.25	27.58333333	19.25
Row 5	4	121	30.25	56.25	30.25
Column 1	5	84	16.8	75.2	16.8
Column 2	5	77	15.4	68.8	15.4
Column 3	5	112	22.4	104.3	22.4
Column 4	5	116	23.2	69.2	23.2

ANOVA

urce of Variati	SS	df	MS	F	P-value	F crit
Rows	1083.7	4	270.925	17.45088567	6.09811E-05	3.259167
Columns	230.95	3	76.98333333	4.958668814	0.01823832	3.490295
Error	186.3	12	15.525			
		std devn ===>>	**3.940177661**			
Total	1500.95	19				

Two–Way Anova without Replication

- **SST =** Sum of squared differences between each individual data value minus the grand mean $\bar{x}$
- **SSE =** sum of squared differences between individual values and their cell mean

Two–Way Anova without Replication – Predicting Sales

Predicted Sales = Overall Average + Rep (effect) + (Region effect)

- IF the Region effect was significant and the sales representative effect was not, what is the predicted sales for Rep 2 in North region ?
  would be =====>>> 35.95
- IF the Region was not significant and the sales representative effect was significant, What is the predicted sales for Rep 2 in North region ?
  would be =====>>> 34.85
- 95% confident that if Rep 2 is assigned to North Region, monthly sales will be between 43.46 & 59.23

Two-Way Anova (Factor) with Replication

In two-way Anova the dependent variable which observes more than one time for each combination of the two factors is called Two way Anova with replication

Lets take a look at how the advertising and price level influence the monthly sales of a retail store.. Here we have 3 replications for each price advertising combination.

Price

Advertisement

	Low	Medium	High	Effect
Low	44	23	12	-6.07407
	26	19	13	
	24	17	10	
Medium	30	30	13	-1.18519
	32	28	26	
	34	20	19	
High	37	28	22	7.259259
	47	42	27	
	49	34	22	
	47	32	20	

Overall Average = 26.96

Two-Way Anova with Replication

We need to conduct 3 hypothesis test

H0: Price has no impact on the sales
Ha: Price has an impact on sales

H0: Advertisement no has an impact on the sales
Ha : Ad has an impact on the sales

H0: Combination both Price & Advertisement has no impact on the sales
Ha : Both has an impact on the sales

Price

Advertisement

	Low	Medium	High	Effect
Low	44	23	12	-6.07407
	26	19	13	
	24	17	10	
Medium	30	30	13	-1.18519
	32	28	26	
	34	20	19	
High	37	28	22	7.259259
	47	42	27	
	49	34	22	
	8.925926	-0.18519	-8.74074	

Overall Average = 26.96

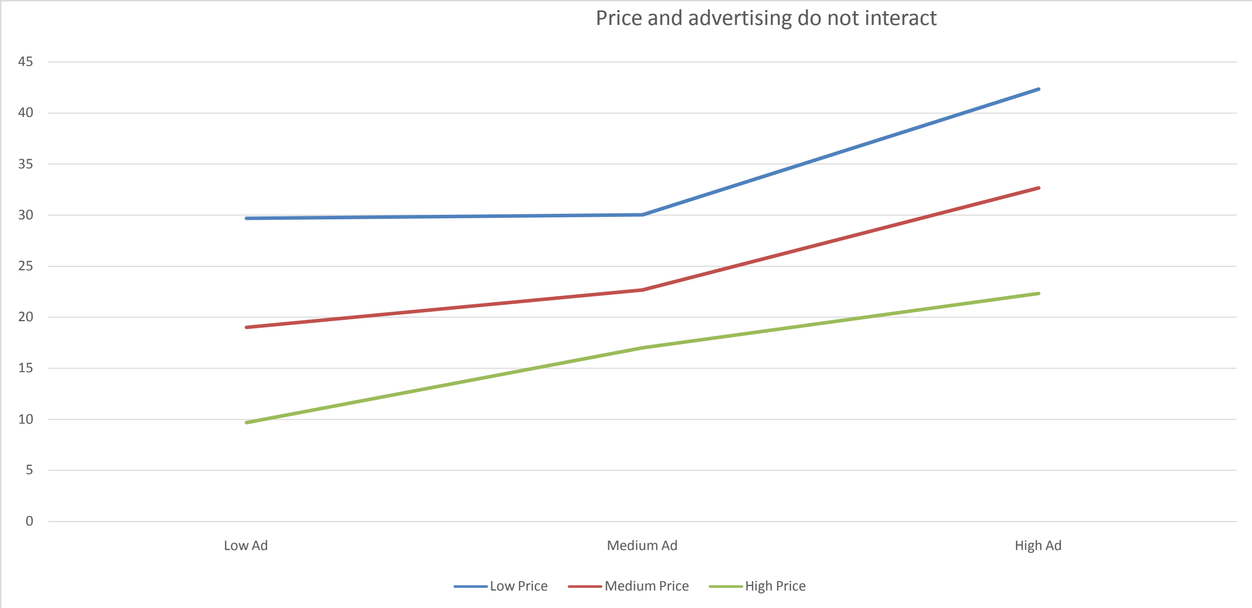
Two-Way Anova with Replication

Anova: Two-Factor With Replication

SUMMARY	Low	Medium	High	Total
Low				
Count	3	3	3	9
Sum	94	59	35	188
Average	31.33333	19.66667	11.66667	20.88889
Variance	121.3333	9.333333	2.333333	106.6111
Medium				
Count	3	3	3	9
Sum	96	78	58	232
Average	32	26	19.33333	25.77778
Variance	4	28	42.33333	48.69444
High				
Count	3	3	3	9
Sum	133	104	71	308
Average	44.33333	34.66667	23.66667	34.22222
Variance	41.33333	49.33333	8.333333	104.9444
Total				
Count	9	9	9	
Sum	323	241	164	
Average	35.88889	26.77778	18.22222	
Variance	81.86111	64.19444	40.94444	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
Sample	818.963	2	409.4815	12.03047	0.000481	3.554557146	Ad effect
Columns	1404.963	2	702.4815	20.63874	2.2E-05	3.554557146	Price effect
Interaction	64.37037	4	16.09259	0.472797	0.755113	2.927744173	
Within	612.6667	18	34.03704				
	std devn		5.834127				
Total	2900.963	26					



As the ad increases, sales also increase at roughly the same rate
Irrespective of the price

Two-Factor with replication

Predicted Sales = Overall Average + [row (ad) effect] + [column (Price) effect]

When Price is high and ad is medium

$$= 26.96 + (-1.185) + (-8.74)$$

$$= 17.035$$

With 95% confidence that the sales during a month with high price and medium advertising will be between 5.368 AND 17.03

We fail to reject the Null hypothesis of the 3rd test and conclude that the combination of Price and Advertisement has no impact on sales

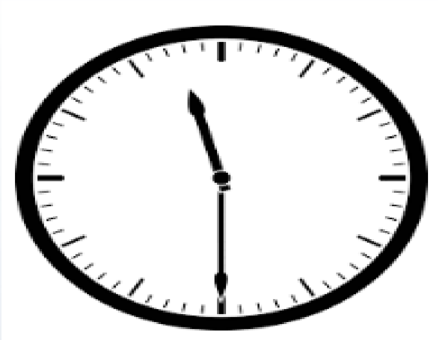
χ^2 Testing

To devise an appropriate marketing strategy, sample has been collected from a total of 50 respondents to ascertain if there is any association between age and the choice of mobile phone brand

Observed	Choice			
Brands	M 20-30	M 31-44	M 45-55	Total
Iphone	38	50	10	98
Samsung	46	8	45	99
MotoG	89	27	12	128
Total	173	85	67	325

Expected	Choice			
Brands	M 20-30	M 31-44	M 45-55	Total
Iphone	52.16615385	25.63076923	20.203077	98
Samsung	52.69846154	25.89230769	20.409231	99
MotoG	68.13538462	33.47692308	26.387692	128
Total	173	85	67	325

Breakout Exercise



Determine all the Xs that impact the loading time of app

Determine the vital Xs that have significant impact on the loading time

ANOVA

A Test engineer while doing the analysis he found that the users were trying to load An app on three different mobile platforms (Android, Windows, and IOS). He had an apprehension that loading time of the app was driven by the type of OS.

To validate the same , he has drawn a random sample to ensure that there is no issue because of the same

H0 mu = No relationship between app loading time and the OS type

Ha = relationship between the two exists.

```
loadtime <- c(4,4.5,5.5,4.9,5.4,5.72,4.92,5.42,6.62,5.72,6.33,6.33,6.23,6.43,6.43,6.73)
type <- c(rep("Ubuntu",5),rep("OSSS",5),rep("Linux",6))
Linux <- c(4,4.5,5.5,4.9,5.4)
mean(Linux)
Ios <- c(5.72,4.92,5.42,6.62,5.72)
data1 <- data.frame(LoadTime,type)
```