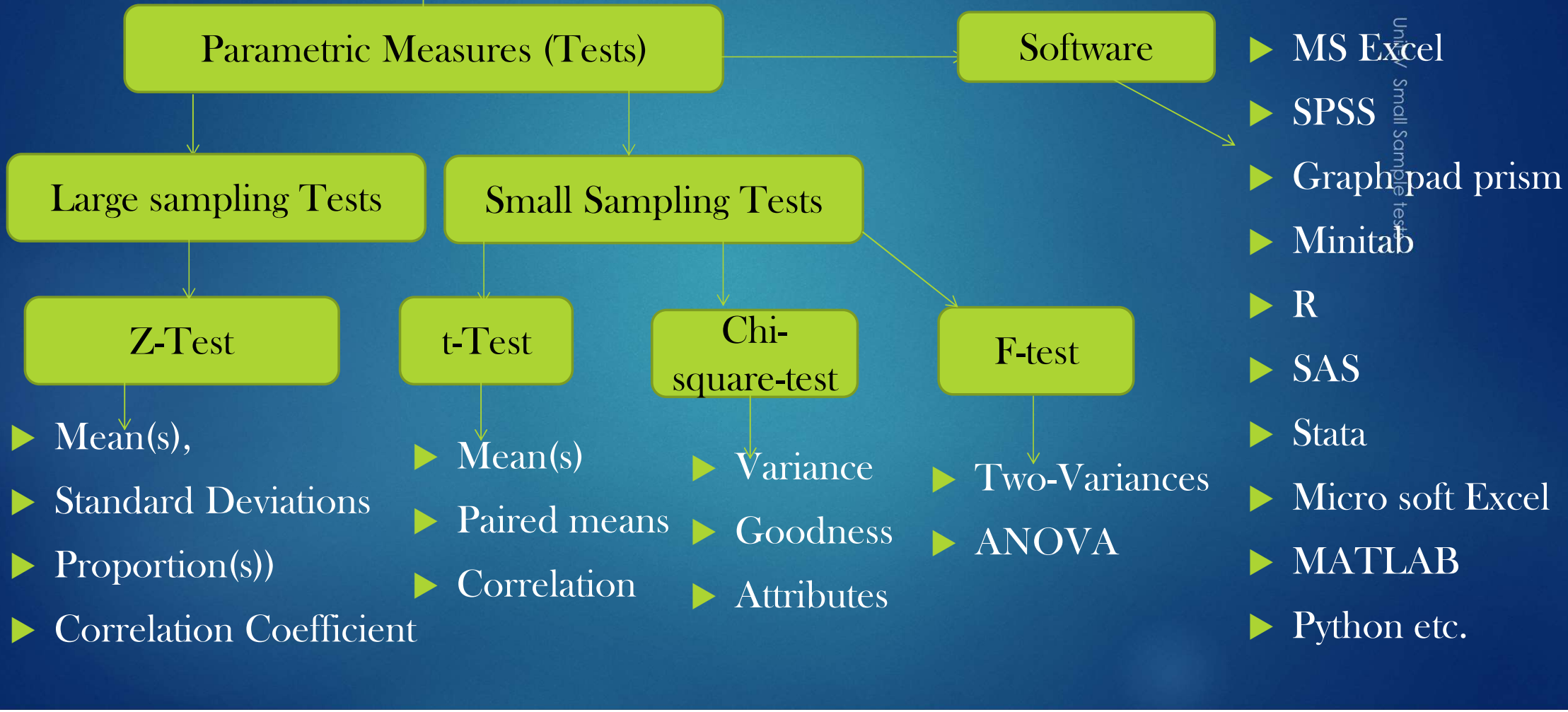
Course
MATH2361 Probability and Statistics
(Common for BSC BTECH, CSE, ECE, BBA)
@GITAM, deemed to be University

UNIT-V **Small Sample Tests**

Dr Malikarjuna Reddy Doodipala
MSc, M.Phil, PGDCA, Ph D,
GITAM Deemed to be University

Testing of Hypothesis



Unit-V

- **Student t-distribution**

  test for single mean, two means and paired t-test
- **Chi-square tests**

  χ2- test for variance

  χ2 - test for goodness of fit,

  χ2 -test for independence of attributes.
- **F-Test:** Testing of equality of variances (F-test)

UNIT V Small sample tests

- Tests of significance based on chi square, t and F.
- chi-square -test for test for independence of attributes,
- t-test for single, double and paired tests,
- Variance Ratio Test(F-test).

08L

UNIT V

Learning outcomes

At the end of this unit, the student will be able to

- Test the significance based on chi-square distribution, t- distribution, and F
- distribution
- Test for independence of attributes
- Test the hypothesis of t-test for single, double and paired

Unit-V

After completion of this unit, the student will be able to

- Analyze the testing of hypothesis for small samples (L4)
- Test for the χ2-goodness of fit and independence of attributes (L4)

Prerequisites

- **One is supposed to have the knowledge of**
  - **Population, Sample, Parameter**
  - **Sampling Distribution**
  - **Estimate or Statistic**
  - **Estimator**
  - **Testing of Hypothesis**

Small Sample Tests(n ≤30)

- To study the general magnitude of variation of the population or to test the population characteristic a sample size at most 30 is drawn and it can be treated as small sample .The test statistic required is called small sample(ing)
- T-tests
- Chi-Square Test
- F-test

Small Sample Tests(n ≤30)

- t-Test
- Chi-Square Test
- F-test

1. t-Test and it's Applications

- At first t-distribution was introduced by W.S.Goset on his pen name student.
- t-distribution has the following applications
- Test for Mean
- Test for Difference of Means
- Paired t-test for difference of Means
- Test for correlation coefficient

t- Test for the Mean (σ Unknown)

- Start with the hypothesis is that H_0 $\mu = \mu o$ is true.
- In applied work σ is usually unknown
- It could be replaced by *S*, the SD of the sample.
- If the population is assumed to be normally distributed, the test statistic ***t*** will follow ***t*** distribution with n–1 degrees of freedom.

Assumptions

- All the observations are independent
- Variance of the population is unknown.
- Parent population assumed to be normal.

t-Test for the Mean (Procedure)

- Using the sample data compute the calculated of t statistic.
- Now compare this calculated value with table or critical value of t with (n-1) degrees of freedom at given level of significance a or confidence level (1-a).
- If tcal value< ttab value then we accept null hypothesis otherwise
- we reject the null hypothesis(or) accept alternate hypothesis.

t -Test Example

The average sales in Co. is 120. A random sample of 12 sales during the past month is:

108.98 152.22 111.45 110.59 127.46 107.26

93.32 91.97 111.56 75.71 128.58 135.11

Is there evidence of a change in the average amount of these sales at $\alpha = 0.05$?

$\mu = 120$, $n = 12$, σ unknown

Using two tailed ***t* test**, $\overline{X} = 112.85$, $S = 20.80$

Statistical null hypothesis is H_0: $\mu = 120$

Statistical Alternate hypothesis is H_1: $\mu \neq 120$

Calculations for Mean & SD

x	$x-\bar{x}$	$(x-\bar{x})^2$
108.98	-3.87	14.9769
152.22	39.37	1549.997
111.45	-1.4	1.96
110.59	-2.26	5.1076
127.46	14.61	213.4521
93.32	-19.53	381.4209
107.26	-5.59	31.2481
91.97	-20.88	435.9744
111.56	-1.29	1.6641
75.71	-37.14	1379.38
128.58	15.73	247.4329
135.11	22.26	495.5076
Total=1354.12	Total	4758.121

From the table n=12

$$\sum x = 1354.12$$

$$\bar{x} = 1354.12/12$$
$$= 112.85$$

$$\sum (x-\bar{x})^2 = 4758.12$$

$$S^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$
$$= \frac{4758.12}{12-1} = 432.56$$
$$S = 20.79$$

t -Test Statistic: Computations &Conclusion

with sample mean weight=112.85
N=12
S=20.80

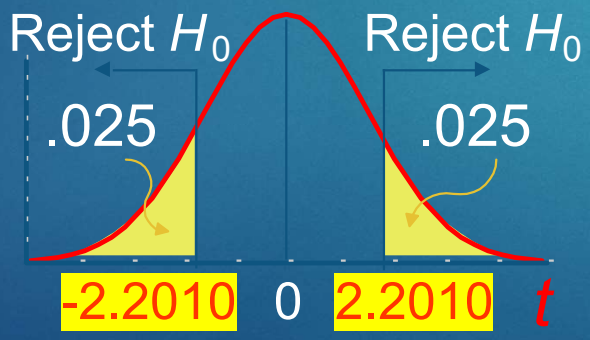
H_0: $\mu = 120$ with sample mean weight=112.85

H_1: $\mu \neq 120$

$\alpha = 0.05$

df = 12 - 1 = 11

Critical Value(s):



Test Statistic:

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

$$t = \frac{112.85 - 120}{20.80/\sqrt{12}} = -1.19$$

Decision:

Do Not Reject at α=.05

Conclusion:

No evidence average weight has changed

Interval estimators (For mean)

- The interval estimators for population mean μ when sample size is small by t-Test for mean

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right)$$

- Random sample of 12 students spent an average of Rs.273.20 on textbooks. Sample standard deviation was 94.40. Find interval estimators at 5%level

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}}\left(\frac{S}{\sqrt{n}}\right) \qquad \mu = 273.20 \pm t_{(0.025,11)}\left(\frac{94.40}{\sqrt{12}}\right)$$

- where $t_{(0.025,11)}$=2.201 (from t-tables)

2.*t* Test for Independent Samples (Variances Unknown)

Assumptions

- Both populations are normally distributed
- Samples are randomly and independently drawn
- Population variances are unknown but assumed equal
- If both populations are not normal, need large sample sizes
- Setting Up the Hypotheses

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

OR

$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 \neq 0$

Two Tail

t Test for Independent Samples

- Compute the Sample Statistic

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

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

Hypothesized difference

$$df = n_1 + n_2 - 2$$

Test Statistic: Procedure

- Compute the Test Statistic by using the sample data
- Now compare with the critical or table value of t with the given degree of freedom(d.f) at level of significance a
- If observed value of t is less than the critical or table value then we accept null hypothesis
- Otherwise Reject the same. (or)
- Accept Alternative hypothesis.

t Test for Independent Samples: Example

- You're a analyst for agriculture . Is there a difference in average dividend yield between two crops listed You collect the following data:

	Crop-1	crop-2
Sample size	**21**	**25**
Sample Mean	**3.27**	**2.53**
Sample Std Dev	**1.30**	**1.16**

- Assuming equal variances, is there a difference in average yield (***a* = 0.05**)?

Solution

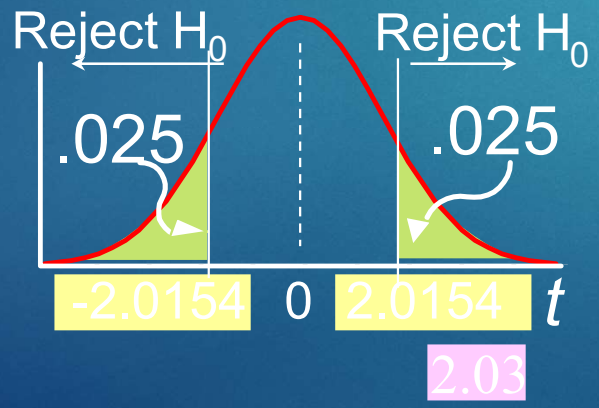
H_0: $\mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$

H_1: $\mu_1 - \mu_2 \neq 0$ i.e. $(\mu_1 \neq \mu_2)$

$\alpha = 0.05$

df = 21 + 25 - 2 = 44

Critical Value(s):



Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.502\left(\frac{1}{21} + \frac{1}{25}\right)}} = 2.03$$

Decision:

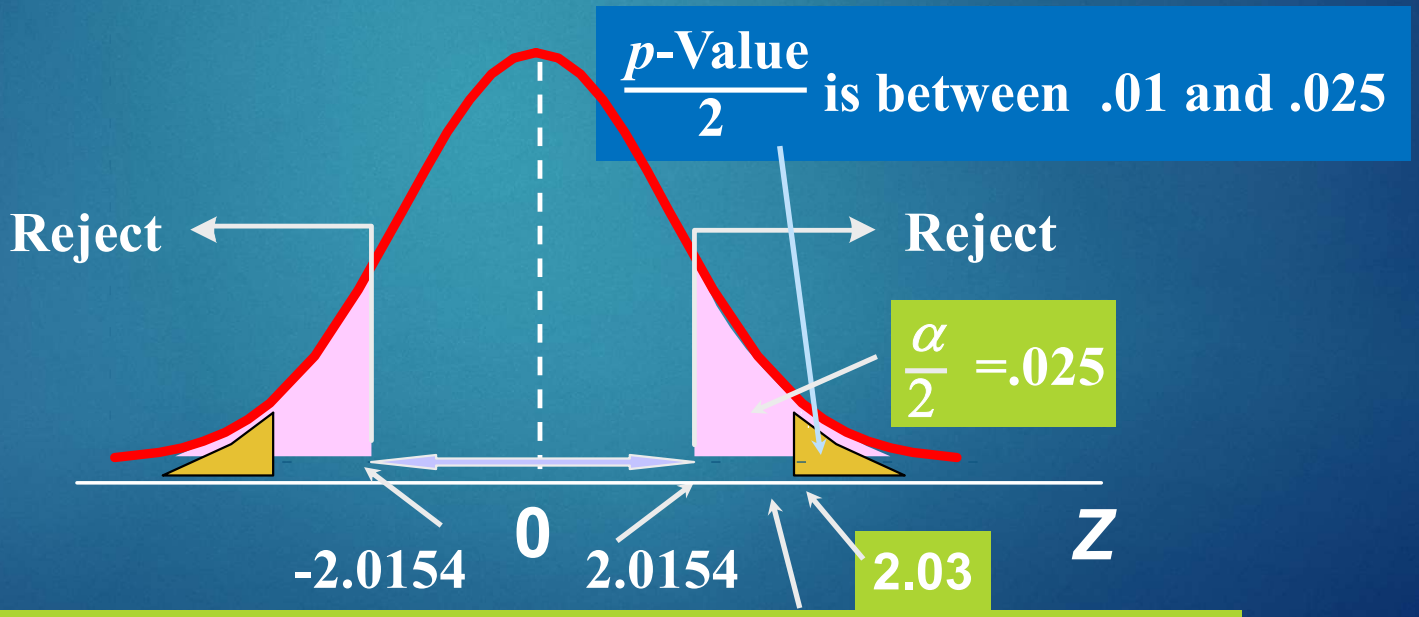
Reject at $\alpha = 0.05$.

Conclusion:

There is evidence of a difference in means.

p -Value Solution

(p-Value is between .02 and .05) < ($\alpha = 0.05$)
Reject.



Test Statistic 2.03 is in the Reject Region

Comparisons of two independent population Means (t-Test): Confidence Interval

- confidence interval:

$$\text{C.I} = (\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}} \sqrt{\frac{(n_1 - 1){S_1}^2}{n_1} + \frac{(n_2 - 1){S_2}^2}{n_2}}$$

- Notice that there is no overall weighted average or
- Combined population variance as there is in a significance test for means.

Example

You're a statistical analyst for agriculture Field.

You collect the following data:

	CropI	CropII
Number	**21**	**25**
Sample Mean	**3.27**	**2.53**
Sample Std Dev	**1.30**	**1.16**

You want to construct a 95% confidence interval for the difference in population average yields of the crops listed on I and II

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)} = \frac{(21-1)1.30^2 + (25-1)1.16^2}{(21-1)+(25-1)} = 1.502$$

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,n_1+n_2-2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$(3.27 - 2.53) \pm 2.0154\sqrt{1.502\left(\frac{1}{21}+\frac{1}{25}\right)}$$

$$0.0088 \le \mu_1 - \mu_2 \le 1.4712$$

Sample Problem -2

- In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on the average than the machine currently used.
- To test that hypothesis, the times it takes each machine to pack ten cartons are recorded.
- The results, in seconds, are shown in the table.
- Do the data provide sufficient evidence to conclude that, on the average, the new machine packs faster?
- Perform the required hypothesis test at the 5% level of significance.

New machine					Old machine				
42.1	41.3	42.4	43.2	41.8	42.7	43.8	42.5	43.1	44.0
41.0	41.8	42.8	42.3	42.7	43.6	43.3	43.5	41.7	44.1

Solution:

- We can thus proceed with the pooled *t*-test.
- **Step 1**. Set up Null & alternate hypothesis

  Let μ_1 denote the mean for the new machine and μ_2 denote the mean for the old machine.

  $H0: \mu_1 = \mu_2$,
  $H1: \mu_1 < \mu_2$
- **Step 2**. given that the sample data(see table)

  $\bar{X} = 42.14$, $s_1 = 0.683$
  $\bar{Y} = 43.23$, $s_2 = 0.750$
- Significance level:$a = 0.05$.

Calculations for means- Sample variances

Sno	X	Y	$X- \bar{X}$	$(X- \bar{X})^2$	$Y- \bar{Y}$	$(Y- \bar{Y})^2$
1	42.10	42.70	-0.04	0.002	-0.53	0.28
2	41.30	43.80	-0.84	0.71	0.57	0.32
3	42.40	42.50	0.26	0.07	-0.73	0.53
4	43.20	43.1	1.06	1.12	-0.13	0.02
5	41.80	44	-0.34	0.12	0.77	0.59
6	41.00	43.60	-1.14	1.30	0.37	0.14
7	41.80	43.30	-0.34	0.12	0.07	0.00
8	42.80	43.50	0.66	0.44	0.27	0.07
9	42.30	41.7	0.16	0.03	-1.53	2.34
10	42.70	44.1	0.56	0.31	0.87	0.76
Sum	**421.40**	**432.3**	**0.00**	**4.20**	**0.00**	**5.06**
Mean	**42.14**	**43.23**	S^2_1	**0.47**	S^2_2	**0.56**

Solution:

- **Step 3**. Compute the *t*-statistic:

$$t = \frac{(\bar{X} - \bar{Y})}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

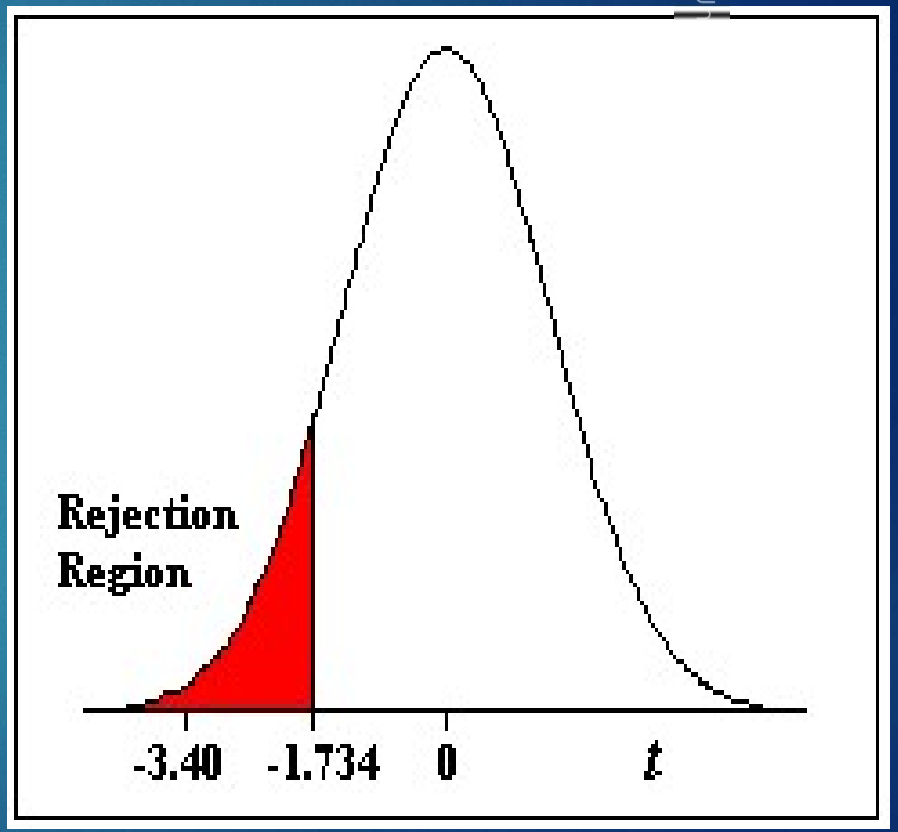
$$where \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

$$= \frac{9 \cdot (0.683)^2 + 9 \cdot (0.750)^2}{10 + 10 - 2} = 0.514, \quad S_p = 0.717$$

$$t_{Cal} = \frac{42.14 - 43.23}{0.717\sqrt{1/10 + 1/10}} = -3.40$$

Solution:

- **Critical value:**
- Left-tailed test
  Critical value = t_{α}=0.05
  Degrees of freedom =10+10−2=18
  $t_{0.05}$=−1.734
  Rejection region t_{tab}=−1.734
- **Step 4**. Check to see if the value of the test statistic falls in the rejection region and decide whether to reject H_o.
- t_{cal}=−3.40< t_{tab} −1.734(one tail)
  Reject H0 at a=0.05
- **conclusion : At 5% level of significance, the data provide sufficient evidence that the new machine packs faster than the old machine on average.**



Comparing Two Related Samples

- Test the Means of Two Related Samples
  - Paired or matched
  - Repeated measures (before and after)
  - Use difference between pairs

$$d_i = X_i - Y_j \text{ then } \bar{d} = \bar{x} - \bar{y}$$

- Eliminates Variation between Subjects

Z Test for Mean Difference (Variance Known)

$$Z = \frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}}$$

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

- Assumptions
  - Both populations are normally distributed
  - Observations are paired or matched
  - Variance known
- Test Statistic

3.Paired t Test for Mean Difference (Variance Unknown)

- Assumptions
  - Both populations are normally distributed
  - Observations are matched or paired
  - Variance unknown
  - If population not normal, need large samples
- Test Statistic

$$t = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

$$S_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1}}$$

Dependent-Sample *t* Test: Example

Assume you work in the pharma department. Is the new drug faster (α=0.05 level)? You collect the following processing times:

sno	Drug1	Drug2	D_i
	Seconds		
1	9.98	9.88	.10
2	9.88	9.86	.02
3	9.84	9.75	.09
4	9.99	9.80	.19
5	9.94	9.87	.07
6	9.84	9.84	.00
7	9.86	9.87	- .01
8	10.12	9.98	.14
9	9.90	9.83	.07
10	9.91	9.86	.05

$$\bar{D} = \frac{\sum D_i}{n} = .072$$

$$S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}} = .06215$$

Dependent-Sample *t* Test: Example Solution

Is the new drug faster (0.05 level)?

H_0: $\mu_D = 0$
H_1: $\mu_D > 0$

$\alpha = .05$ $\bar{D} = .072$

Critical Value = 1.8331
df = n - 1 = 9

Test Statistic

$$t = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = \frac{.072 - 0}{.06215/\sqrt{10}} = 3.66$$

Decision: Reject H_0

t Stat. in the rejection zone.

Conclusion: The new drug is faster.

Confidence Interval Estimate for of Two Dependent Samples

- Assumptions
  - Both populations are normally distributed
  - Observations are matched or paired
  - Variance is unknown
- Confidence Interval Estimate:

$$\mu_D = \bar{D} \pm t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}$$

$$\bar{D} = \frac{\sum D_i}{n} = .072 \qquad S_D = \sqrt{\frac{\sum(D_i - \bar{D})^2}{n-1}}$$

$$= .06215$$

$$t_{\alpha/2,n-1} = t_{0.025,9} = 2.2622$$

$$\bar{D} \pm t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}$$

$$.072 \pm 2.2622\left(\frac{.06215}{\sqrt{10}}\right)$$

$$0.0275 < \mu_D < 0.1165$$

Comparing Two Related Samples

- Test the Means of Two Related Samples
  - Paired or matched
  - Repeated measures (before and after)
  - Use difference between pairs

$$d_i = X_i - Y_j \text{ then } \bar{d} = \bar{x} - \bar{y}$$

- Eliminates Variation between Subjects

Z Test for Mean Difference (Variance Known)

- Assumptions
  - Both populations are normally distributed
  - Observations are paired or matched
  - Variance known
- Test Statistic

$$Z = \frac{\bar{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}}$$

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

Test for correlation coefficient

- Null and alternative hypotheses.

  H_0: ρ=0.
  H_1: ρ is not 0

Under null hypothesis

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

r=sample correlation coefficient of the given data

Using the sample data find t-cal value compare with table value at given level wth d.f n-2

Note for large sample : t test statistic is replaced by Z-test statistic

Test for correlation coefficient

- **Null and alternative hypotheses.**

H_0: ρ=0.
H_1: ρ is not 0

Given that n=27, r =0.6 level=0.05 tab value=2.06

Using the sample data find t-cal value compare with table value.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 3.75$$

$r = $ sample correlation coefficient of the given data

Reject the null hypothesis

Fishers Z-transformation for correlation coefficient

- Null and alternative hypotheses.

H_0: $\rho=0$.
H_1: ρ is not 0

for example, if your correlation coefficient(r) is 0.4, the transformation is:

$z' = .5[\ln(1+0.4) - \ln(1-0.4)]$

$z' = .5[\ln(1.4) - \ln(0.6)]$

$z' = .5[0.33647223662 - -0.51082562376]$

$z' = .5[0.84729786038]$

$z' = 0.4236$.

where ln is the natural log.

NOTE:

Instead of working the formula, you can also refer to the r to z' table.

Fishers Z-transformation for correlation coefficient

- if the Pearson correlation coefficient between two variables is found to be r = 0.55, then we would calculate zr to be:

- zr = ln((1+r) / (1-r)) / 2
- zr = ln((1+.55) / (1-.55)) / 2
- zr = 0.618
- NOTE:

Instead of working the formula, you can also refer to the r to z' table.

Fishers Z-transformation for correlation coefficient

It turns out that the sampling distribution of this transformed variable follows a normal distribution.

This is important because it allows us to calculate a confidence interval for a Pearson correlation coefficient.

Without performing this Fisher Z transformation, we would be unable to calculate a reliable confidence interval for the Pearson correlation coefficient.

Fishers Z-transformation for correlation coefficient

- Fisher's z' is used to find confidence intervals for both r and differences between correlations.
- But it's probably most commonly be used to test the significance of the difference between two correlation coefficients, r1 and r2 from independent samples.
- If r1 is larger than r2, the z-value will be positive; If r1 is smaller than r2, the z-value will be negative.
- While the Fisher transformation is mainly associated with Pearson's r for bivariate normal data, it can also be used for Spearman's rank correlation coefficients in some cases.

Chi-Square Distribution

Introduction:

The square of standard normal viriate is known as Chi-Square variate with one Degree of Freedom
Thus we know that the *Standard Normal Variate*

$Z = \frac{x-\mu}{\sigma}$ then $Z^2 = \left(\frac{x-\mu}{\sigma}\right)^2$
is a Chi − Square variate with one d.f in general
if $X_i (i = 1,2,3,..n)$ are n independent
normal variates with mean μ_i,
var i nce $\sigma_i{}^2$ then
$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$ is a Chi − Square variate with n d.f

Chi-Square Distribution

- **PDF:** A continuous r.v X is said to have follows chi-Square distribution with n Degree of Freedom if its pdf is given by

$$f(x) = \frac{x^{n/2-1}e^{-x/2}}{\gamma(n/2)\ 2^{n/2}} \text{ if } 0 < x < \infty \text{ and } n > 0$$
$$= 0 \text{ otherwise}$$

- The mean and variance is n and 2n
- Mode is n-2
- Coefficient of skewness is 8/n
- Coefficient of kurtosis is12/n+3
- As n is large it tends to normal

CHI-SQUARE TEST (Statistic)

- **Introduction:**
- In the agriculture as well as the biological research, apart from the quantitative characters one has to deal with the qualitative data, like the flower color or seed color in which observations are classified in a particular category, class or group.
- The results of breeding experiments and genetically analysis come under this type of analysis.
- In all genetically studies it becomes necessary to test the significance of over all deviation between the observed and the expected frequencies.
- The significance test of this deviation is known as chi-square test. The symbol of chi-square is χ^2 .it is Greek letter. The chi-square test was first used by Karl Pearson in the year 1900.

Applications of CHI-SQUARE TEST

The chi-square test is applicable to various problems in agriculture and biology besides other statistical analysis. They are

- To test the goodness of fit
- To test the independence of attributes
- To test the homogeneity of independent estimates of
  the population variance
- To test the detection of linkage

Test for Goodness of fit

- The χ^2 test can be applied in various problems of biostatistics.
- **Karl Pearson** developed a test of significance in the year 1900.this test is known as chi-square test for goodness of fit
- It is used to test whether there is a significant difference between an observed data and the theoretical data
- The quantity chi-square describes the magnitude of difference between the observed and the expected frequencies.
- This test is helpful to find out whether such differences are significant or not.

Chi-square Test Statistic

- The Chi-square test statistics can be defined as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$E = \frac{(rowsum)(columnsum)}{grandtotal}$$

$$O = Observed\ frequency$$

$$E = Expected\ frequency$$

- If the observed frequencies and the expected frequencies are identical, the computed chi square value will be zero.
- There fore the possible value of chi-square ranges upward from zero.

Test Statistic

- To determine the chi-square value, the steps required are:
- Calculate the expected frequencies(E)
- Find out the difference between the observed frequencies (O) and the expected frequencies.

  If the deviation (O-E) is large, the square deviation (O-E)2 is also large.

- square the values of (O-E)2 and divide each value by respective value of E and obtain the total (O-E)2 this will be the value of chi-square which ranges from zero to infinity.
- The calculated value of chi-square is compared with the table value for the given degree of freedom at either 5% or 1% level of significance.

Test Statistic

- If the calculated value of chi-square is less than the tabulated value at a particular level of significance, the difference between the observed and the expected frequencies are not significant
- when the calculated value is more than the tabulated value, the difference between the observed and the expected value is significant
- In chi-square analysis while comparing the calculated value of chi-square with the tabulated value
- we must calculate the degree of freedom.
- The degree of freedom are calculated from the number of degrees of freedom in a chi-square test is equal to the number of classes minus one.

Characteristics Of Chi-square Test:

- Chi-square distribution has some important characteristics. They are:
- This test is based on frequencies whereas in theoretical distribution, the test is based on mean and standard deviation.
- this test can be used for testing the difference between the entire set of the expected and the observed frequencies.
- A new chi-square distribution is formed for every increase in the number of degrees of freedom.
- This test is applied for testing the hypothesis but is not useful for estimation.

SSUMPTIONS :

- There are few assumptions for the validity of chi-square test:
- All the observations must be independent, no individual item should be included twice or a number of times in the sample.
- The total number of observations should be large. The chi-square test should not be used if n>50.
- All the events must be mutually exclusive.
- For comparison purposes, the data must be in original units.
- If the theoretical frequency is less than 5, and then we pool it with the preceding or the succeeding frequency, so that the resulting sum is greater than 5.

Sample Problems

- The following figures show the distribution of digits in number chosen at random from a telephone Directory
- Test at 5% whether the digits may be taken to occur equally frequently in the directory
- Chi-square critical value=19.02

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	1026	1107	997	966	1075	933	1107	972	964	853

Solution

- Null Hypothesis H0:
- It is assumed that the digits may be taken to occur equally frequently in the directory.
- Alternate Hypothesis H0:
- It is assumed that the digits may be taken to occur not equally frequently in the directory.
- Now We apply the Following Good ness of fit Formula:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Expected frequencies

- Under null hypothesis Expected frequencies are obtained by using

$$\frac{Totalobservedfrequency}{No.ofdigits} = \frac{10000}{10} = 1000,$$

for i=0,1, ...9

Chi-square table: Calculations under Ho:

Digits	Observed frequency (O)	Expected frequency (E)	O-E	(O-E)2	(O-E)2/E
0	1026	1000	26	676	0.676
1	1107	1000	107	11449	11.449
2	997	1000	-3	9	0.009
3	966	1000	-34	1156	1.156
4	1075	1000	75	5625	5.625
5	933	1000	-67	4489	4.489
6	1107	1000	107	11449	11.449
7	972	1000	-28	784	0.784
8	964	1000	-36	1296	1.296
9	853	1000	-147	21609	21.609
Total chi-square value(observed)					**58.542**

Formula applied: Chi-square test Statistics

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 58.54$$

Degrees of freedom $=10-1=9$,
Tabulated of $\chi^2_{0.05} = 16.919$.(see from tables)
Since calculated χ^2 $>$ tabulated χ^2
we reject the null hypothesis.
Thus the digits may not be taken to occur equally frequently in the directory.

Sample Problem

- A survey of 320 families with 5 children each, revealed the following distribution.

No. of Boys	5	4	3	2	1	0
No. of Girls	0	1	2	3	4	5
No. of families	14	56	110	88	40	12

- Is the result consistent with the hypothesis that male and female births are equally probable at 0.01 significance level.
- Test the goodness of fit.

Solution:

- Let us set up the statistical null and alternate hypothesis
- H0: Male and female births are equal.
  H1: Male and female births are not equal.
  [Null and alternative hypotheses.]
- Let X is male birth : X follows Binomial then
- Probability of male birth, p = 1 / 2; Probability of female birth, q = 1 / 2
- Given that n=5

Binomial formula: Probabilities, Expected frequencies

- $P(x)= nCx\ p^x\ q^{n-x}$ $x=0,1,2,\ldots5$

- The degrees of freedom is 6 - 1 = 5. The critical value at a = 0.05 from the chi-square distribution table is 11.07.
  [Degrees of freedom = number of categories - 1.]
- The probability of having 5 Boys [or 5 Girls using the symmetry of terms] = p(x=5)=(1 / 2) 5 = 1 / 32

  The expected number of families with 5 boys (or 5 girls) = 320 × 1 / 32= 10

Binomial formula: Probabilities, Expected frequencies

- p(x=4) = 5C4 (1 / 2) 4 (1/ 2) 1 = 5 / 32

  The expected number of families with 4 boys and 1 girl (or 4 girls and 1 boy) = 320 × 5 / 32= 50

  p(x=3)= 5C3 (1 / 2) 3(1 / 2) 2 = 10 / 32

  The expected number of families with 3 boys and 2 girls (or 3 girls and 2 boys) = 320 × 10 / 32= 100
- Similarly we get P(x=2)=10/32, Exp.Frequency=100
- P(X=1)=5/32 , =50
- P(X=0)=1/32 =10
-

Chi Square test Formula

- Total Observed frequency (O) 14 56 110 88 40 12
- Expected frequency (E) 10 50 100 100 50 10
- Chi-square test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 7.16$$

- [Expand and substitute the values of O and E from the above.
- Since 7.16< 11.1, the decision is to accept the null hypothesis at a = 0.05 level.Thus male and female births are equal.

Test the hypothesis that the Depression and gender are independent With the following observed data at 5% level

Character	Male	Female	Total
Depressed	10	20	30
Not Depressed	40	30	70
Total	50	50	100

- Sol: Set up the Hypotheses:

  H_o: depression & gender are independent

  H_1: depression and gender are not independent

.

$$\chi^2 = \sum\sum \frac{(O-E)^2}{E}$$

$$O = observed\ frequency$$

$$E = Expected\ frequency$$

$$E = \frac{Row\ total \times Column\ total}{N}$$

$$N = Total\ observed\ frequency$$

- **Expected Frequencies**

  E1 = 30x50/100=15 ,E3 = 70x50/100=35

  E2 = 30x50/100=15, E4 = 30x50/100=35

- $\chi^2 = \frac{(10-15)^2}{15} + \frac{(20-15)^2}{15} + \frac{(40-35)^2}{35} + \frac{(30-15)^2}{35} = 4.76$

$$\chi^2 = \sum\sum \frac{(O-E)^2}{E}$$

$O = observed\ frequency$

$E = Expected\ frequency$

$$E = \frac{Row\ total \times Column\ total}{N}$$

$N = Total\ observed\ frequency$

$\chi^2_{critical}$ $\alpha = 0.05$

df = $(r-1)\ (k-1)$ = $(2-1)(2-1)$ = 1

- 1= 1 from table χ^2 =3.84.

- Reject the Null hypothesis
- Depression is Gender dependent

	Male	Female	Total
Depressed	10(**15**)	20(**15**)	**30**
Not Depressed	40(**35**)	30(**35**)	**70**
Total	**50**	**50**	**100**

Fisher's Exact Test Test for independence of attributes (2x2-table)

$$\chi^2 = \frac{\left[|ad - bc| - \frac{N}{2}\right]^2}{R_1 R_2 C_1 C_2} \quad N = 22.061$$

$R_1 = a + b = 20 + 5 = 25$

$R_2 = c + d = 5 + 25 = 30$

$C_1 = a + c = 20 + 5 = 25$

$C_2 = b + d = 5 + 25 = 30$

$N = a + b + c + d$

$= 20 + 5 + 20 + 5 = 55$

$N/2 = 55/2 = 27.5$

A survey was conducted on of 55 women and the following information is obtained.Test the hyothesis that the crocin tablet control the fever or not at 5%level

Suffering from fever	Treatment		Total
	Effect	Not	
Yes	20	5	25
No	5	25	30
Total	25	30	55

$\chi^2_{tab} = 3.84 \quad at\ 5\%\ with\ 1d.f$

Unit-V Small Sample tests

Problems for Practice

- No of misprints: 0 1 2 3 4

  No of pages : 122 60 15 2 1

  Fit Poisson distribution and test the good ness of fit (Mean =100/200=0.5)

- Day no : Sun Mon Tue Wed Thur Fri Sat Total

  No of accidents: 14 15 11 20 15 15 10 100

  Test the hypothesis that the accidents are uniformly distributed throughout the year at 5%level. (Cal value =4.43 Tab value=12.592 at 5%)

- Fit a Poisson distribution to the following data and test the goodness of fit:

  No.of accidents: 0 1 2 3 4 56

  No. of days : 150 65 45 34 10 6 2

F - Test for Difference in Two Population Variances

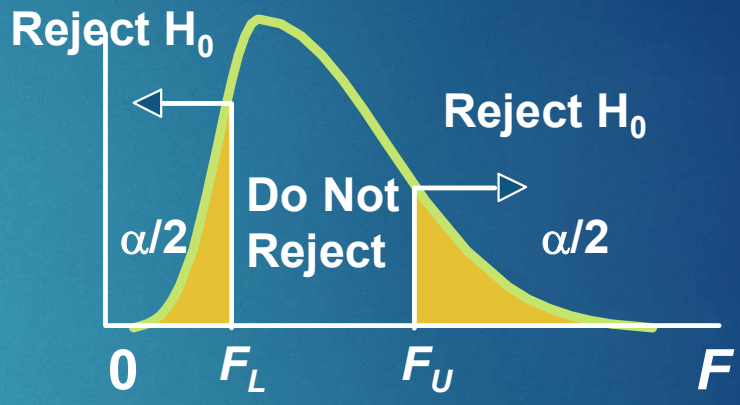
- Test for the Difference in 2 Independent Populations Parametric Test Procedure
- Assumptions Both populations are normally distributed Test is not robust to this violation
- Samples are randomly and independently drawn

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

F - Test for Difference in Two Population Variances

- **Hypotheses**
  - $H_0: \sigma_1^2 = \sigma_2^2$
  - $H_1: \sigma_1^2 \neq \sigma_2^2$
- **Test Statistic**
  - $F = S_1^2 / S_2^2$
  - **Two Sets of Degrees of Freedom**
    - $df_1 = n_1 - 1$; $df_2 = n_2 - 1$
  - **Critical Values:** $F_{L(n_1 - 1,\, n_2 - 1)}$ and $F_{U(n_1 - 1,\, n_2 - 1)}$

$F_L = 1/F_U^*$ (*degrees of freedom switched)



F Test: Application

- You're a analyst for agriculture. Is there a difference in variance dividend yield between two crops listed You collect the following data:

	Crop-1	crop-2
Sample size	**21**	**25**
Sample Mean	**3.27**	**2.53**
Sample Std Dev	**1.30**	**1.16**

- Is there a difference in the variances between the crops
- at $\alpha = 0.05$ level?

F Test: Example Solution

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$\alpha = .05$

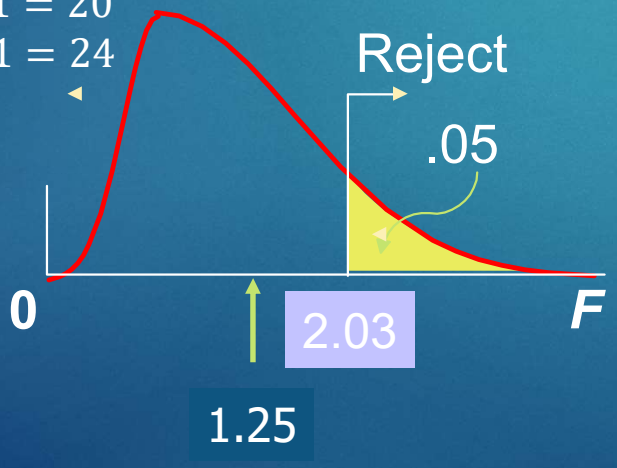
$df_1 = 20$ $df_2 = 24$

Critical Value(s):

$$F_{.05,20,24} = 2.03$$

$df_1 = n_1 - 1 = 21 - 1 = 20$

$df_2 = n_2 - 1 = 25 - 1 = 24$



Test Statistic:

$$F = \frac{S_1^2}{S_2^2} = \frac{1.30^2}{1.16^2} = 1.25$$

Decision: F cal < F tab

Conclusion:

Do not reject at $\alpha = 0.05$.

There is insufficient evidence to prove a difference in variances.

F Test: Example Solution

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

$\alpha = .05$

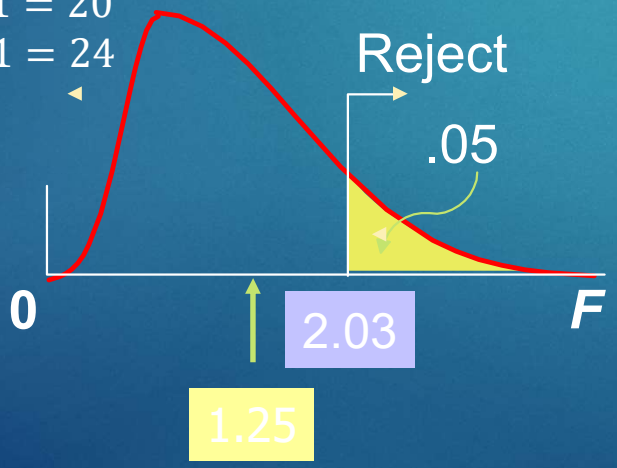
$df_1 = 20$ $df_2 = 24$

Critical Value(s):

$$F_{.05,20,24} = 2.03$$

$df_1 = n_1 - 1 = 21 - 1 = 20$

$df_2 = n_2 - 1 = 25 - 1 = 24$



Test Statistic:

$$F = \frac{S_1^2}{S_2^2} = \frac{1.30^2}{1.16^2} = 1.25$$

Decision: F cal < F tab

Conclusion:

Do not reject at $\alpha = 0.05$.

There is insufficient evidence to prove a difference in variances.

Any Questions? Suggestions?

Thank you

Feedback to

mdoodipa@gitam.edu