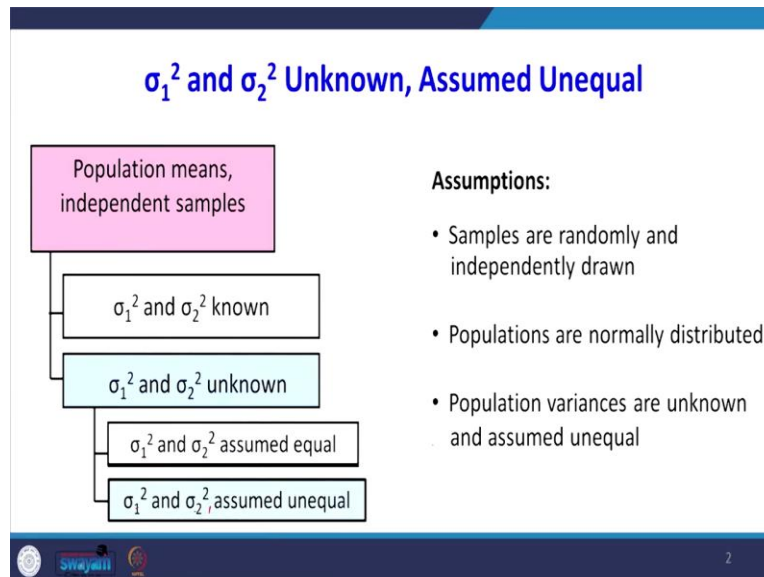


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 21
Hypothesis Testing: Two Sample Test-II

Dear students in the previous class we have seen the problem in comparing hypothesis testing in to population where σ_1^2 σ_2^2 is known. Then we have seen that the next problem that is σ_1^2 σ_2^2 is unknown but assumed equal variance. In this class we are going to take another category of the problem where σ_1^2 σ_2^2 square unknown but assumed unequal.

(Refer Slide Time: 00:59)



What are the assumptions we are having the samples are randomly and independently drawn the populations are normally distributed population variance are unknown and assumed unequal. The population variances are assumed unequal so your pooled variance is not appropriate. So, use here we have to use a p-value with new deals of freedom the formula for degrees of freedom is this one $\nu = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}$

(Refer Slide Time: 01:30)

Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

The test statistic for $\mu_1 - \mu_2$ is:

σ_1^2 and σ_2^2 unknown

- σ_1^2 and σ_2^2 assumed equal
- σ_1^2 and σ_2^2 assumed unequal

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where t has ν degrees of freedom:

$$\nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left(\frac{s_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2}\right)^2 / (n_2 - 1)}$$

The t statistics is $\bar{X}_1 - \bar{X}_2$ - the difference root of s_1^2 square by $n_1 + s_2^2$ square by n_2 you see the previous problem we have used SP^2 square by $n_1 + SP^2$ square by n_2 where the variances are equal. But here it is unequal we cannot use SP^2 in both the places so we have to use only s_1^2 square, the corresponding formula for degree of freedom already which we explained.

(Refer Slide Time: 01:57)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

- Arsenic concentration in public drinking water supplies is a potential health risk.
- An article in the Arizona Republic (Sunday, May 27, 2001) reported drinking water arsenic concentrations in parts per billion (ppb) for 10 metropolitan Phoenix communities and 10 communities in rural Arizona.
- The data as shown:

	Metro Phoenix		Rural Arizona	
Phoenix,	3	Rimrock,	48	
Chandler,	7	Goodyear,	44	
Gilbert,	25	New River,	40	
Glendale,	10	Apache Junction,	38	
Mesa,	15	Buckeye,	33	
Paradise Valley,	6	Nogales,	21	
Peoria,	12	Black Canyon City,	20	
Scottsdale,	25	Sedona,	12	
Tempe,	15	Payson,	1	
Sun City,	7	Casa Grande,	18	

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

$$\bar{x}_1 = 12.5 \quad \bar{x}_2 = 27.5$$

$$s_1 = 7.63 \quad s_2 = 15.3$$

We will take you one sample problem will solve this one the problem is arsenic concentration in public drinking water supplies is a potential health risk. An article in Arizona Republic Sunday May 27 2001 reporter drinking water arsenic concentration in parts per billion ppb for 10

metropolitan fornic communities and 10 communities in rural Arizona are given in the table. We can know what is the \bar{X}_1 that is a 12.5 s_1 is 7.63 \bar{X}_2 is 27.5 s_2 is 15.3.

(Refer Slide Time: 02:38)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

- We wish to determine if there is any difference in mean arsenic concentrations between metropolitan Phoenix communities and communities in rural Arizona.

6

We wish to determine if there is any difference in mean arsenic concentration between metropolitan phonic communities and communities in rural Arizona.

(Refer Slide Time: 02:48)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

- The parameters of interest are the mean arsenic concentrations for the two geographic regions, say, μ_1 and μ_2 , and we are interested in determining whether $\mu_1 - \mu_2 = 0$.
- $H_0: \mu_1 - \mu_2 = 0$, or $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- $\alpha = 0.05$ (say)
- The test statistic is

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

7

So what are the steps in hypothesis the testing as usual the first step is the parameter of interest are the mean arsenic concentration for the 2 regions say μ_1 and μ_2 and we are interested in determining whether $\mu_1 - \mu_2$ equal to 0. So, what will be about null hypothesis null hypothesis is $\mu_1 - \mu_2$ equal to 0 otherwise μ_1 equal to μ_2 . Alternative hypothesis μ_1

not equal to mu 2 because the signs are complementary so alpha is 5% but is not given we have to assume it is a 5% is the formula for test statistics is this one $t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

(Refer Slide Time: 03:40)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

6. The degrees of freedom

$$v = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right]^2}{\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1)} = \frac{\left[\frac{7.63^2}{10} + \frac{15.3^2}{10} \right]^2}{\left(\frac{7.63^2}{10} \right)^2 / (10 - 1) + \left(\frac{15.3^2}{10} \right)^2 / (10 - 1)} = 13.2 = 13$$

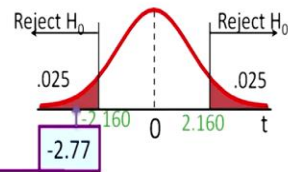
Therefore, using $\alpha = 0.05$, we would reject $H_0: \mu_1 = \mu_2$ if $t_0^* > t_{0.025,13} = 2.160$ or if $t_0^* < -t_{0.025,13} = -2.160$.

So, the degrees of freedom be all the data is given in the previous slide we can supply that value $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$ and so on. So, we are getting 13.2 approximately the degrees of freedom is 13 therefore using alpha 5% we would reject $H_0: \mu_1 = \mu_2$ if the p value the calculated T value is greater than 2.160 or p value is less minus because these values there are 2 ways we can get this value we can refer the T table but we can use Python also directly you can use the Python to get the critical value when alpha equal to 0.02 that means when the probability is 0.025 when degrees of freedom is 13.

(Refer Slide Time: 04:31)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

7. Computations:



$$t = \frac{(12.5 - 27.5) - 0}{\sqrt{\left(\frac{7.63^2}{10} + \frac{15.3^2}{10}\right)}} = -2.77$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.

So, we have done the t by using the calculated t values -2.77 obviously -2.77 is lying on the rejection side so we have to reject null hypothesis. So, what we are concluding there is evidence of difference in the means that means it is not the equal amount of arsenic is available there is it in some cities it is more in other cities it is less.

(Refer Slide Time: 05:00)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

8. Conclusions: Because $t_0^* = -2.77 < t_{0.025, 13} = -2.160$,

- Reject the null hypothesis.
- There is evidence to conclude that mean arsenic concentration in the drinking water in rural Arizona is different from the mean arsenic concentration in metropolitan Phoenix drinking water.

So, the conclusion because $t_0 = -2.77$ is less than you were -2.160 we have to reject a null hypothesis there is a evidence to conclude that the mean arsenic concentration in the drinking water in rural Arizona is different from the arsenic concentration in metropolitan Phoenix drinking water it is not the same.

(Refer Slide Time: 05:22)

Problem: Test Statistic: σ_1^2 and σ_2^2 Unknown, Unequal

```
In [17]: stats.t.ppf(0.025,13) #critical t value
Out[17]: -2.160368656461013

In [18]: metro = [3,7,25,10,15,6,12,25,15,7]
         rural = [48,44,40,38,33,21,20,12,1,18]

In [20]: stats.ttest_ind(metro,rural, equal_var = False)
Out[20]: Ttest_indResult(statistic=-2.7669395785560558, pvalue=0.015827284816100885)
```



So, we will use Python to solve this problem we can see the p value as I told you stat stat t dot ppf when in the t distribution when this area is 0.025 because a 2-tailed any area equal to 0.025 when the degrees of freedom is 13 we are getting it is - 2.160 so it is a - 2.160 our calculated t value is how much - 2.77 so - 2.77 will be on the left-hand side obviously we have to reject it. Instead of doing that it is very simple in Python you take array 1 as the values which is given for Metro there are another one array 2 that is call it is rural.

The value is given a rural area so stat start t-test underscore call this to array metro, rural equal underscore variance you have to type equal to equal to false do you remember for a previous one we have written is it true. Now simply write false you'll get the your t value your p value obviously it is a 2-tailed test, so, the p value the alpha is it is very small the p value is very small so we have to reject a null hypothesis when compared to alpha it is only 0.01 so we have to reject our null hypothesis.

(Refer Slide Time: 06:42)

Dependent Samples

Tests Means of 2 Related Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use difference between paired values:

$$d_i = x_i - y_i$$

- Assumptions:
 - Both Populations Are Normally Distributed



Now we will go to another setup problem where there is a samples are dependent. So, the test of 2 related populations they are called paired sample or match the samples so it is repeated measures. The same population we are collecting the data before and after so we have to find use the difference between paired sample $d_i = x_i - y_i$. So, what the logic is, is this is say population 1 this is population this is also population 1 same population before what was the this before any treatment suppose the we can see a lot of hair oil advertisements are coming before applying oil what was the length of you hair.

Here you can see some example some values who take some sample mean after sometime up after applying hail oil you can see what was the say this is \bar{X}_1 this is \bar{Y}_1 \bar{X}_1 \bar{Y}_1 before and after we find when you plot the difference, so you take X_1 from the sample one before then you take Y_1 from the same sample because it is the independent sample when you plot the difference that when you keep on collect different pair from the same sample when you plot the difference that will follow normal distribution. So both our populations are normally distributed.

(Refer Slide Time: 08:14)

Test Statistic: Dependent Samples

The test statistic for the mean difference is a t value, with $n - 1$ degrees of freedom:

$$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}} \quad \bar{d} = \frac{\sum d_i}{n} = \bar{x} - \bar{y}$$

D_0 = hypothesized mean difference
 s_d = sample standard dev. of differences
 n = the sample size (number of pairs)

The test statistics for the mean difference is the t value with $n - 1$, degrees of freedom. So, here you see previously we would write X bar here the mean of the difference there is a d bar you add all the difference divided by n there is nothing but equal to X bar - Y bar this was difference in the population sd is for that data for the difference the data what was the standard deviation the root of n . so, we will get the t -value.

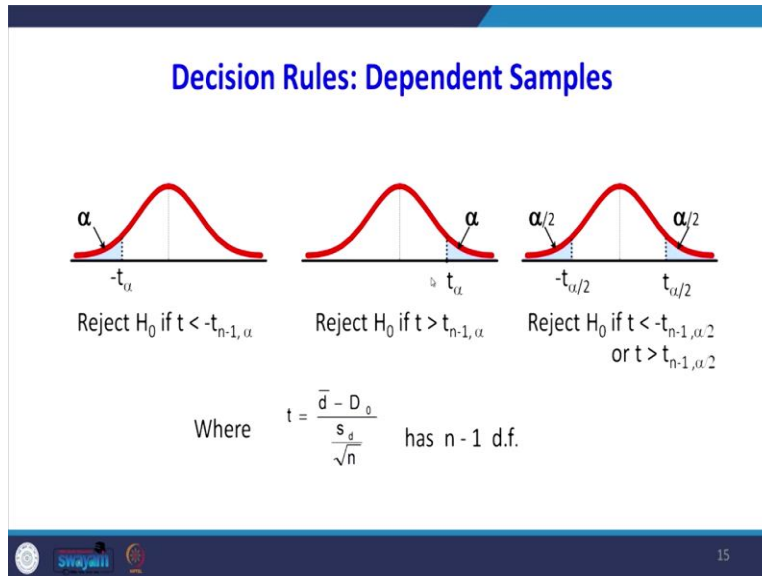
(Refer Slide Time: 08:46)

Decision Rules: Dependent Samples

Lower-tail test:	Upper-tail test:	Two-tail test:
$H_0: \mu_1 - \mu_2 \geq 0$	$H_0: \mu_1 - \mu_2 \leq 0$	$H_0: \mu_1 - \mu_2 = 0$
$H_1: \mu_1 - \mu_2 < 0$	$H_1: \mu_1 - \mu_2 > 0$	$H_1: \mu_1 - \mu_2 \neq 0$

Here also see this is your left tailed test the second one is right tailed test this is 2-tailed test.

(Refer Slide Time: 08:51)



Left tail right tail 2 tail test but only the t is $\bar{d} - D_0 / (s_d / \sqrt{n})$ with $n - 1$ degrees of freedom.
(Refer Slide Time: 09:04)

Dependent Samples: Example

- An article in the Journal of Strain Analysis (1983, Vol. 18, No. 2) compares several methods for predicting the shear strength for steel plate girders.
- Data for two of these methods, the Karlsruhe and Lehigh procedures, when applied to nine specific girders, are shown in Table .
- We wish to determine whether there is any difference (on the average) between the two methods.

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

We will take one example for dependent sample an article in the journal of strain analysis compares that is volume 18 and number 22 compare several methods of predicting the shear strength of steel plate girders. Data for 2 of these methods one method is Karlsruhe another method is Lehigh procedures when applied to nine specific graders are shown in the table. I think these this is 2 methods are the different way of measuring the shear strength.

We wish to determine whether there is any difference on the average value between 2 methods because the populations are same 2 difference are conducted.

(Refer Slide Time: 09:49)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S11	1.186	1.061	0.119
S21	1.151	0.992	0.159
S31	1.322	1.063	0.259
S41	1.339	1.062	0.277
S51	1.200	1.065	0.138
S21	1.402	1.178	0.224
S22	1.365	1.037	0.328
S23	1.537	1.086	0.451
S24	1.559	1.052	0.507

So, called Karlsruhe method Lehigh method in Karlsruhe method this was the values this is Lehigh method this was the value. You are finding the differences look at this here the difference are positive there is a possibility the difference may be negative also that will subtracted from the positive value there is no problem.

(Refer Slide Time: 10:09)

Inferences About the Difference Between Two Population Means: Matched Samples

1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. $H_0: \mu_D = 0$
3. $H_1: \mu_D \neq 0$
4. $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}}$$

So, the first step is the parameter of interest is the difference in the mean shear strength between 2 methods there's $\mu_D = \mu_1 - \mu_2$ equal to see rather we call it as difference is equal to 0 the third would third step is $\mu_D \neq 0$ so α equal to 5% those the tested statistics is \bar{S}_D divided by S_D root of n it is nothing but the same thing previously what was the t formula if

it is the if it is not paired sample what was the t formula $\bar{X} - \mu$ divided by S by root n but the \bar{X} is nothing but the mean of the difference.

This is nothing but the standard deviation of the difference this was the difference in the mean all others are same. Because what I am saying every statistical test has some link once that is why you have to follow the order of learning this statistics because if you in between if you are going for some lectures that may require certain prerequisites. So, when you learn this one so you have to follow this sequence so that will be very easy for connecting with other statistical test.

(Refer Slide Time: 11:28)

**Inferences About the Difference Between Two Population Means:
Matched Samples**

6. Reject H_0 if $t_0 > t_{0.025,8} = 2.306$ or if $t_0 < -t_{0.025,8} = -2.306$.

7. Computations: The sample average and standard deviation of the differences d_j are $\bar{d} = 0.2736$ and $s_D = 0.1356$, so the test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}} = \frac{0.2736}{0.1356/\sqrt{9}} = 6.05$$

19

When you look at the table when the Alpha value is 0.025 that is half of the Alpha values 0.025 80 degrees of freedom so that value is 2.3, so if the calculated 2 value is it is like this so what will you do this value on positive side we are getting 2.306 and negative side we are getting – 2.306 this is the value which you got from the table. The calculated t values lies on either side of this limit it will be rejected. so, what we got the mean of the difference is 0.2736 the standard deviation is 0.1356 when you input this data we are getting 6.05 that is far away.

So we got to reject the null hypothesis when you reject null hypothesis the μ_1 what was null hypothesis that $\mu_1 - \mu_2$ equal to 0, so H_1 is $\mu_1 - \mu_2$ not equal to 0 when you reject that there is a difference if it is an hair oil example yes there is a effect of hair oil that help you to grow the hair.

(Refer Slide Time: 12:42)

8. Conclusions: Since $t_0 = 6.05 > 2.306$,

we conclude that the strength prediction methods yield different results.

The P -value for $t_0 = 6.05$ is $P = 0.0002$.

20

So, we are rejecting so we conclude that this strength prediction method yield different result we look at the p-value because with the help of statistical table especially t table find if the p-value is very difficult, but we will use Python to see what is the p-value.

(Refer Slide Time: 12:58)

```
In [37]: KARL= [1.186,1.151,1.322,1.339,1.200,1.402,1.365,1.537,1.559]
        LEH= [1.061,0.992,1.063,1.062,1.065,1.178,1.037,1.086,1.052]

In [38]: stats.ttest_rel(KARL,LEH)

Out[38]: Ttest_relResult(statistic=6.0819394375848255, pvalue=0.00029529546278604066)
```

21

So, you take call it as array 1 call second one Lehigh stats dot t-test you see that this is underscore rel so that means dependent sample so call the 2 variable you will get this is your t value this is a p value less than alpha. So, we have to reject the null hypothesis so that means there is a different.

(Refer Slide Time: 13:31)

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$

- Expected Value

$$E(\bar{p}_1 - \bar{p}_2) = p_1 - p_2$$
- Standard Deviation (Standard Error)

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where: n_1 = size of sample taken from population 1
 n_2 = size of sample taken from population 2

Next we will go to another problem that is the inferences about the differences between 2 population proportions. So, far we talk population proportion means whenever there is a categorical variable is there so far we have measured the continuous variable about from the population. If there is a categorical variable obviously the count is taken care that is nothing but the population proportions. So, inferences about the difference between 2 population proportions here also we can estimate the population proportion $p_1 - p_2$ we will do the hypothesis test about the difference of $p_1 - p_2$.

So, what is the expected value before going to this expected value say this is population 1 this is population 2, I take some sample from population 1 I am finding p_1, p_2, p_3 and so on. I am taking some sample from this population to from population 2 there are different sample. If every time if it take p_1 minus that is sample which is taken from sample 1 population 1 and population 2 if I find this difference $p_1 - p_2$ every time of a finding $p_1 - p_1$, so that difference if we plot that that will follow a normal distribution.

The same logic there if you want to know the difference of the variance for example here what was the variance you remember there p_1q_1 by n_1 Sigma square is p_1q_1 by n_1 Sigma 1 square Sigma 2 square is when you call it as p_2q_2 by n_2 here it is p_2q_2 by n_2 if you want to know the difference in the variance you to add the variance. So, what will happen p_1q_1 by $n_1 + p_2q_2$

buy into this is the variance if you want to know those standard deviation just to take square root of that, that is why we have got this one.

So, the expected value nothing but the mean value of $\bar{p}_1 - \bar{p}_2$ is it is $p_1 - p_2$ standard deviation is σ of $\bar{p}_1 - \bar{p}_2$ is root of $p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2$ this $p_1 - p_2$ you see that previously you have taken this is nothing but this sample proportion $\bar{p}_1 - \bar{p}_2$ also sample proportion for population 2 n_1 is size of the sample taken from population 1 n_2 is size of the sample taken from population 2.

(Refer Slide Time: 16:15)

Sampling Distribution of $\bar{p}_1 - \bar{p}_2$

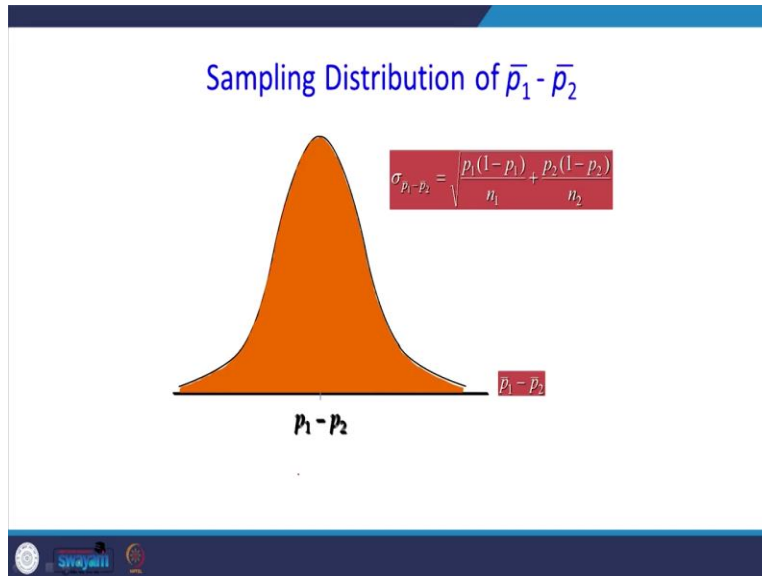
- If the sample sizes are large, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a normal probability distribution.
- The sample sizes are sufficiently large if all of these conditions are met:

$n_1 p_1 \geq 5$	$n_1(1 - p_1) \geq 5$
$n_2 p_2 \geq 5$	$n_2(1 - p_2) \geq 5$

The slide features a blue header and footer. The footer contains logos for 'Sri Jayanti' and other institutional symbols.

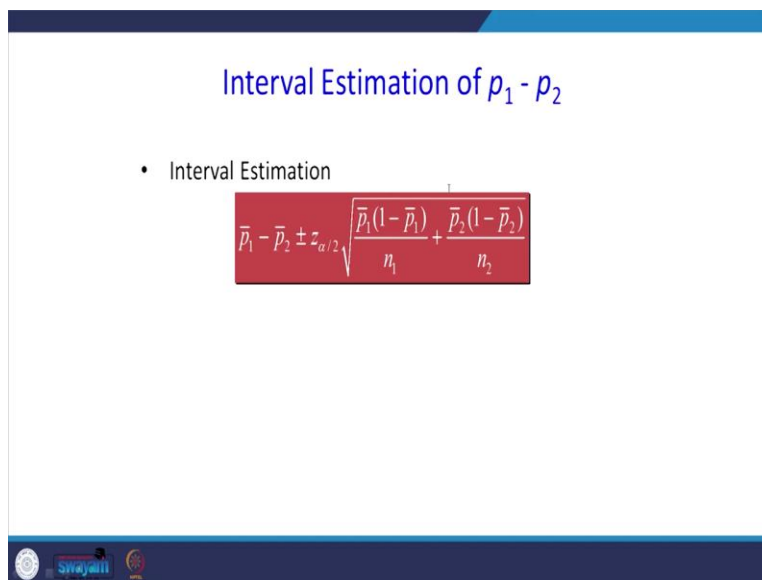
If the sample sizes are large the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a normal probability distribution. The sample sizes are sufficiently large if all the conditions are met when **when** np is greater than or equal to 5 or nq is greater than 5 then only we can approximate this one to the normal distribution.

(Refer Slide Time: 16:41)



You see that the mean of the mean of this non-word distribution is $p_1 - p_2$ the standard deviation is root of p_1 into $1 - p_1$ by n_1 + p_2 into $1 - p_2$ by n_2 .

(Refer Slide Time: 16:57)



The interval estimation is as usual $\bar{p}_1 - \bar{p}_2 \pm Z_{\alpha/2} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$ if it is a single sample you remember can you recollect what was the formula for interval estimation we evidently this way $\bar{p} \pm Z_{\alpha/2} \sqrt{\frac{\bar{p}q}{n}}$ what is happening this pq is population proportion but we have to assume we have to approximate it with the sample proportion $\bar{p}q$ that is a small $\bar{p}q$ by n . So what I am saying everything is came from work previous single sample hypothesis testing.

(Refer Slide Time: 17:41)

Point Estimator of the Difference Between Two Population Proportions

- p_1 = proportion of the population of households "aware" of the product after the new campaign
- p_2 = proportion of the population of households "aware" of the product before the new campaign
- \bar{p}_1 = sample proportion of households "aware" of the product after the new campaign
- \bar{p}_2 = sample proportion of households "aware" of the product before the new campaign

$$\bar{p}_1 - \bar{p}_2 = \frac{120}{250} - \frac{60}{150} = .48 - .40 = .08$$

We will take one problem point estimator of the difference between 2 population proportions say p_1 the proportion of population of households aware of the product after new campaign, p_2 is the proportion of population of households aware of the product before new campaign. So, we are going for new promotions we have to see the effectiveness of that promotion advertisement so \bar{p}_1 is the sample proportion of households aware of the product after the new campaign \bar{p}_2 is sample proportion of households aware of the product before the new campaigns.

So we will find the difference is any impact on the campaign new campaign on awareness. So, $\bar{p}_1 - \bar{p}_2$ is we know that \bar{p}_1 is you know that this 120 divided by 250 because it is given so out of 250, 120 people are aware after the campaign, so, before the campaign out of 150 only 60 people are aware.

(Refer Slide Time: 18:55)

Hypothesis Tests about $p_1 - p_2$

- Hypothesis

We focus on tests involving no difference between the two population proportions (i.e. $p_1 = p_2$)

$H_0: p_1 - p_2 \geq 0$	$H_0: p_1 - p_2 \leq 0$	$H_0: p_1 - p_2 = 0$
$H_a: p_1 - p_2 < 0$	$H_a: p_1 - p_2 > 0$	$H_a: p_1 - p_2 \neq 0$
Left-tailed	Right-tailed	Two-tailed

So, the $p_1 - p_2$ is 8% so hypothesis we focus on test involving no difference between 2 population proportions. Here what is happening here they are also left tail test right tailed test 2 tail test even in the 2 sample population proportion also it can be left tail test right tailed test or 2 tail test.

(Refer Slide Time: 19:15)

Hypothesis Tests about $p_1 - p_2$

- Standard Error of $\bar{p}_1 - \bar{p}_2$ when $p_1 = p_2 = p$

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- Pooled Estimator of p when $p_1 = p_2 = p$

$$\bar{p} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2}$$

The standard error is root of $p(1-p)$ divided by $\frac{1}{n_1} + \frac{1}{n_2}$ we have seen that one here also we can use pooled estimate of p when p_1 and p_2 equal to p . What is the meaning of this one is if the if you assume that the 2 population proportions are same then we can pool that so p bar is $n_1 \bar{p}_1 + n_2 \bar{p}_2$ it away $n_1 + n_2$.

(Refer Slide Time: 19:45)

Hypothesis Tests about $p_1 - p_2$

- Test Statistic

$$z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Test statistics is $p_1 - p_2$ into n so is the \bar{p} into $1 - \bar{p}$ into $\frac{1}{n_1} + \frac{1}{n_2}$.
(Refer Slide Time: 19:54)

Problem: Hypothesis Tests about $p_1 - p_2$

- Extracts of St. John's Wort are widely used to treat depression.
- An article in the April 18, 2001 issue of the *Journal of the American Medical Association* ("Effectiveness of St. John's Wort on Major Depression: A Randomized Controlled Trial") compared the efficacy of a standard extract of St. John's Wort with a placebo in 200 outpatients diagnosed with major depression.
- Patients were randomly assigned to two groups; one group received the St. John's Wort, and the other received the placebo.
- After eight weeks, 19 of the placebo-treated patients showed improvement, whereas 27 of those treated with St. John's Wort improved.
- Is there any reason to believe that St. John's Wort is effective in treating major depression? Use $\alpha = 0.05$.

Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

We will take on problem hypothesis test about $p_1 - p_2$ extract of St. John's Wort are widely used to treat depression this Jones what is a plant or medicine for treating depression. An article in April 18 2001 issue of Journal of American Medical Association the journal the article title is effectiveness of St. John's Wort on major depression a randomized control trial. Compare the efficiency of standard extract of St. John's Wort with the placebo in 200 outpatients diagnosed with major depression. Patients were randomly assigned to groups one group received the St. John's Wort and other received the placebo.

After 8 weeks nine of the placebo treated patients showed improvement whereas 27 of those treated with St. John's Worton improved is there any reason to believe that St. John's Worton is effective curing major depression. Assume alpha equal to 5%. Now we have to see effect of this medicine and curing their depression.

(Refer Slide Time: 21:19)

Problem: Hypothesis Tests about $p_1 - p_2$

1. The parameters of interest are p_1 and p_2 , the proportion of patients who improve following treatment with St. John's Wort (p_1) or the placebo (p_2).
2. $H_0: p_1 = p_2$
3. $H_1: p_1 \neq p_2$
4. $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}_1 = 27/100 = 0.27$, $\hat{p}_2 = 19/100 = 0.19$, $n_1 = n_2 = 100$, and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{19 + 27}{100 + 100} = 0.23$$

The parameter of interest are p_1 and p_2 the proportion of patients who improve following treatment with the st. John's what p_1 our placebo so the null hypothesis is there is no effect of this new medicine p_1 so we are going to assume p_1 equal to p_2 then alternative hypothesis it is not equal to p_2 okay it does need not be 2-tailed test it is up to you to decide whether it is one tail or 2 tail test at present we are assuming we go that there is no difference in the medicine on the improvement of the patients.

The test statistics is $\hat{p}_1 - \hat{p}_2$ root of $\hat{p}(1 - \hat{p})$ into $1/n_1 + 1/n_2$ so where \hat{p}_1 is 27 by 100 \hat{p}_2 19 by 100 $n_1 = n_2 = 100$, so it is the pooled one so we see since the population proportions are same we are find out the pooled proportion $19 + 27 + 100 + 100$, 0.23.

(Refer Slide Time: 22:30)

Problem: Hypothesis Tests about $p_1 - p_2$

6. Reject $H_0: p_1 = p_2$ if $z_0 > z_{0.025} = 1.96$ or if $z_0 < -z_{0.025} = -1.96$.
7. Computations: The value of the test statistic is

$$z_0 = \frac{0.27 - 0.19}{\sqrt{0.23(0.77)\left(\frac{1}{100} + \frac{1}{100}\right)}} = 1.35$$

8. Conclusions: Since z_0 1.35 does not exceed $z_{0.025}$, we cannot reject the null hypothesis. The P -value is $P \cong 0.177$. There is insufficient evidence to support the claim that St. John's Wort is effective in treating major depression.

So we have to reject our null hypothesis if it is greater than + 1.96 otherwise less than - 1.96. so, the z value when you substitute it is 1.35 so what is happening 1.96 is here so this is the rejection region our 1.35 is lying on the acceptance region. So, what we are concluding since Z_0 1.35 does not exceed $Z_{0.025}$ that is 1.96 we cannot reject hypothesis. When you look at the p-value it is 0.177 so 0.177 is it is more than 0.5.

So we have to accept the null hypothesis there is insufficient evidence to support the claim that the Saint John's Worton is effective in treating major depression. So, we would accept our null hypothesis that means there is no evidence that Saint John's Worton is effective.

(Refer Slide Time: 23:44)

```
In [29]: import math
def two_samp_proportion(p1,p2,n1,n2):
    p_pool = ((p1*n2)+(p2*n1))/(n1+n2)
    x = (p_pool*(1- p_pool)*((1/n1)+(1/n2)))
    s = math.sqrt(x)
    z = (p1- p2)/s
    if (z < 0):
        p_val = stats.norm.cdf(z)
    else:
        p_val = 1 - stats.norm.cdf(z)
    return z, p_val*2

In [30]: two_samp_proportion(0.27,0.19,100,100)
Out[30]: (1.3442056254198995, 0.17888190308175567)

In [27]: stats.norm.cdf(1.3442056254198995)
Out[27]: 0.9105590484591222
```

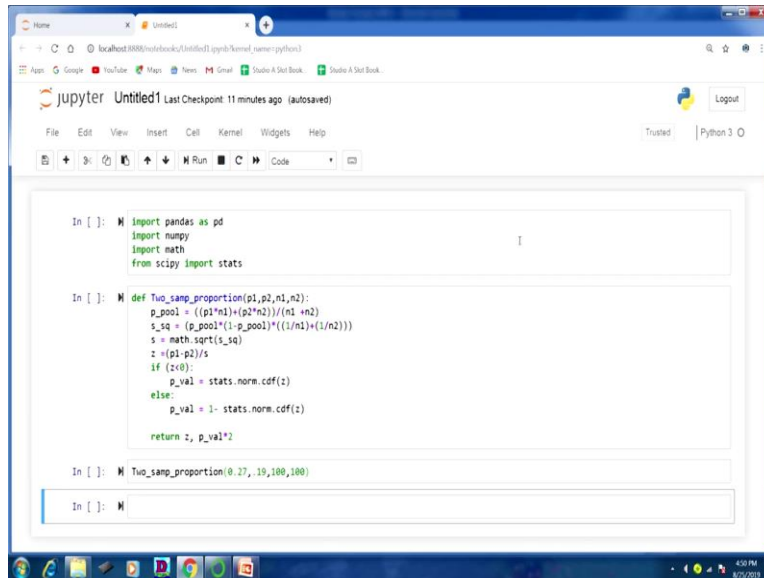
This will do with the help of you can type in Jupiter this command then you have to verify this import math will make a function to sample proportions p1 p2 n1 n2 first we will find out the pooled proportion with the help of Python we learn how to use 2 sample proportion test. So, import math we define a new function to underscore sample underscore proportion p 1 p 2 n1 n2 first we will find out the pool to proportion by n 1 p 1 + n 2 p 2 divided by n 1 + n 2 will solve with the help of Python 2 sample proportions hypothesis testing.

We know what is the formula? It is $\bar{p}_1 - \bar{p}_2$ divided by root of $p q$ into 1 by $n_1 + 1$ by n_2 . First there \bar{p} we call it as a \bar{p} underscore pool is nothing but pooled proposition where $\bar{p} = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$ they will find the variance, variance is $p q$ multiplied by $\frac{1}{n_1 + 1}$ by n_2 then you will take a square root of that that will be the denominator of this formula that is a square root of X .

The value of $Z = \frac{\bar{p}_1 - \bar{p}_2 - (P_1 - P_2)}{\text{root of } pq \text{ into } \frac{1}{n_1 + 1} \text{ by } n_2}$ since the value of $\bar{p}_1 - \bar{p}_2$ whatever assumed to 0 showing there so the Z value is $\bar{p}_1 - \bar{p}_2$ divided by your standard error. If the Z value is less than 0 so if the Z value is less than 0 we can find out the left side probability that is nothing but our p value. If it is greater than 0 we were to substrate from 1 so you will get the p value.

So in the given problem the p_1 population proportion is 0.27 p_2 population proportion is 1/9 so n_1 is 100 n_2 is 100. So, after getting we are getting the Z varies 1.3 the p value is 0.17 that is more than our 0.5, so we would accept our null hypothesis. Since stats suppose if you want to know what was the Z critical value, so stat stat norm dot cdf 1.35 where we got this 1.35 so the corresponding probability 0.91 from this side this side is 0.91 will use Python to solve a 2 sample proportion test.

(Refer Slide Time: 26:59)



```

In [ ]: import pandas as pd
import numpy
import math
from scipy import stats

In [ ]: def two_samp_proportion(p1,p2,n1,n2):
    p_pooled = ((p1*n1)+(p2*n2))/(n1 +n2)
    s_sq = (p_pooled*(1-p_pooled))*((1/n1)+(1/n2))
    s = math.sqrt(s_sq)
    z = (p1-p2)/s
    if (z<0):
        p_val = stats.norm.cdf(z)
    else:
        p_val = 1 - stats.norm.cdf(z)
    return z, p_val*2

In [ ]: two_samp_proportion(0.27, 19,100,100)

In [ ]:

```

We import pandas as pd import numpy import math from scipy we will import stats, so we will make your function, function name is to underscore samp underscore proportions p 1 p 2 n 1 n 2 first we will find the sample pooled the proportion by using this formula $n_1 p_1 + n_2 p_2$ divided by $n_1 + n_2$ then find out the variance in says t into $1 - p$ multiplied by 1 by $n_1 + 1$ by n_2 so that will be the variance to get the standard deviation otherwise standard error will take square root of our variance that is s underscore sq .

So Z is $p_1 - p_2$ Z if the Z value is less than 0 the p value from the table we can treat as it is if the p value is positive we have to subtract from 1 . So, when you the way we are going to call this function is by so the function will returns it p value that has to be multiplied by 2 because it is a 2-sample t -tests. So, we run this to sample proportion p_1 is 0.27 p_2 is 0.19 n_1 is hundred n_2 100 we will run it.

So we got the t value is 1.33 for the p value 0.17 , so it is more than our alpha value, so we are to accept null hypothesis. Where this will conclude this will summarize this class we have seen 2 sample hypothesis testing when σ_1^2 σ_2^2 is unknown but not equal. Then we have seen 2 sample Z -test we have taken some problems then we solved it the next class will go for comparing 2 population variance using F test.