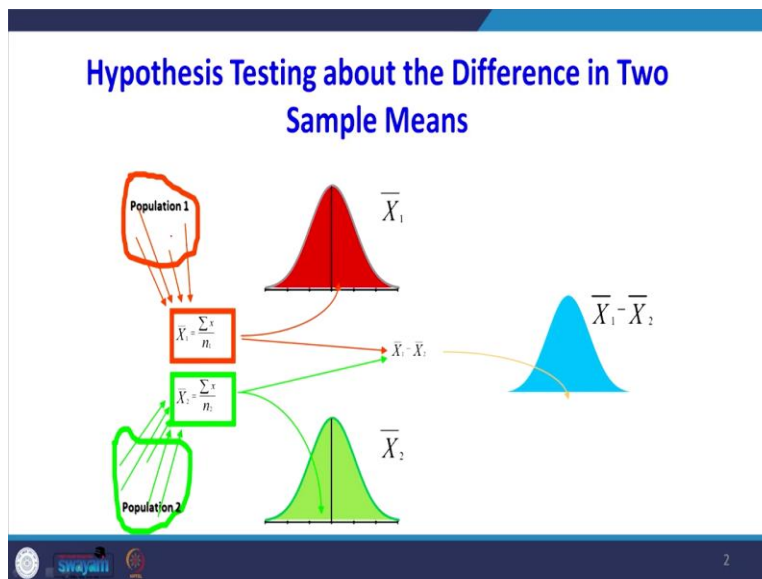


**Data Analytics with Python**  
**Prof. Ramesh Anbanandam**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture – 20**  
**Hypothesis Testing: Two Sample Test-I**

Dear students today we are entering into another topic that is a hypothesis testing for 2 sample tests. First I will explain what is the theory behind this 2 sample test?

**(Refer Slide Time: 00:38)**



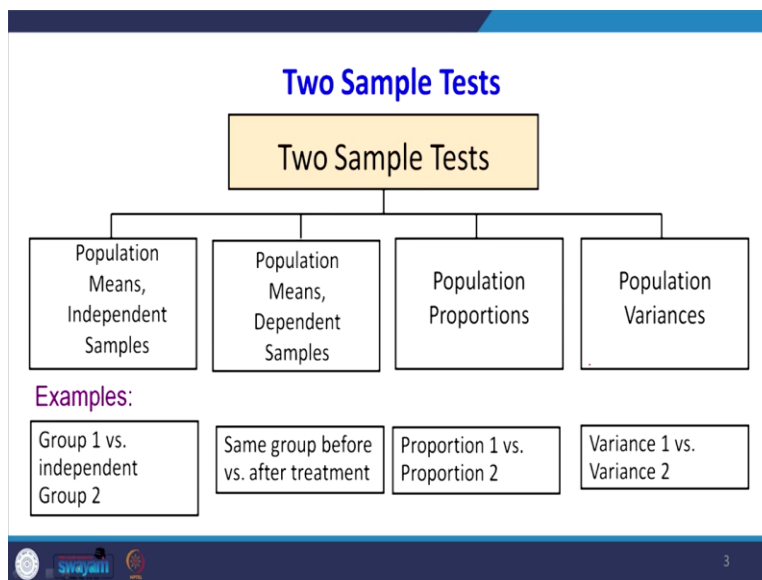
You look at this picture see there are 2 population there are population 1 and population 2 suppose if I take some sample from population 1 I am taking sample and finding sample mean I am calling it as  $\bar{X}_1$  that is a  $\frac{\sum X}{n_1}$  there is another population which is in green color from that I am taking some sample  $\bar{X}_2$  then finding mean of that sample by using this formula  $\frac{\sum X}{n_2}$ .

If I plot this is the sampling distribution of  $\bar{X}_1$  the green one is the sampling distribution of  $\bar{X}_2$ . So, the mean of this sampling distribution of  $\bar{X}_1$  is  $\mu_1$  we can say  $\mu_1$  the mean of sampling distribution of  $\bar{X}_2$  is  $\mu_2$ . The variance is  $\frac{\sigma_1^2}{n_1}$  now what will happen because this result is so the variance actually you know to right this way variance of  $\bar{X}_1$ , this is variance of  $\bar{X}_2$ .

Suppose if I find the difference of their sample mean if I plot the difference that will follow a normal distribution. So, the mean of this population is  $\mu_1 - \mu_2$  the variance of this population is so the variance is we have to find the difference of these variants for population 1 variance is  $\frac{\sigma_1^2}{n_1}$  for population 2 the variance is  $\frac{\sigma_2^2}{n_2}$ . So, the variance is if we want to know the difference of the variance of the 2 population we have to add the variance.

We know that the formula variance of  $A - B$  equal to variance of  $A$  plus variance of  $B$ . So, the formula if I say variance of  $A - B$  you might have studied. So, variance of  $A$  plus variance plus variance of  $B$ . So, for the first population variances  $\frac{\sigma_1^2}{n_1}$  for second population the variance is  $\frac{\sigma_2^2}{n_2}$  so this is the variance of this population which is in blue in color.

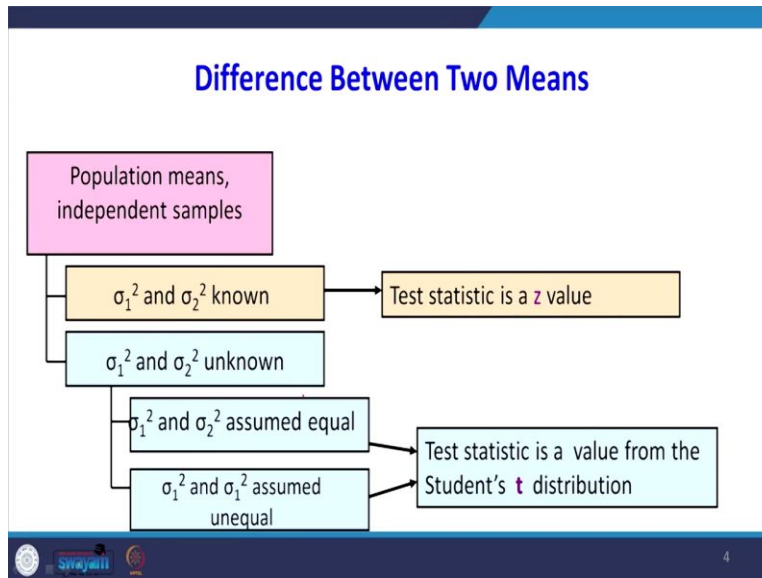
**(Refer Slide Time: 03:14)**



So, will you use this result in coming this one this is the classification of different 2 sample test. One classification is population means for independent samples, population mean for dependent samples, population proportions and population variances. In the population mean for independent variables will compare group 1 versus group 2 both the populations are independent. The population mean for dependent variables same group before and after the treatment so this is dependent samples.

In population proportion there are 2 population we will take proportion one from population one versus variance of proportion 2 we will take another population. Similarly for comparing variances of 2 population we will find the variance of 1 of population 1 versus variance of 2 of population 2.

**(Refer Slide Time: 04:10)**



First you will start with between 2 means so the population means for independent sample there are 2 possibilities there one is the sigma1 square and sigma2 square is known sigma1 square is variance of population 1 Sigma 2 square is variance of population 2. The another category is variance are unknown for population 1 and 2 variants are unknown. When variant are unknown if we can assume it is equal or we can assume it it is not equal.

If variance of population 1 and 2 is known we should go for test statistics is Z value if the variance is unknown we have to go for e statistics.

**(Refer Slide Time: 04:57)**

### $\sigma_1^2$ and $\sigma_2^2$ Known

Population means,  
independent samples

$\sigma_1^2$  and  $\sigma_2^2$  known

$\sigma_1^2$  and  $\sigma_2^2$  unknown

**Assumptions:**

- Samples are randomly and independently drawn
- both population distributions are normal
- Population variances are known

5

We will see what are the assumptions for when a variance of 2 populations are known to you, the first assumption is samples are randomly and independently drawn both the populations are normal. We have seen in the first slide both the populations are normal population variance are known but different.

**(Refer Slide Time: 05:17)**

### $\sigma_1^2$ and $\sigma_2^2$ Known

Population means,  
independent samples

$\sigma_1^2$  and  $\sigma_2^2$  known

$\sigma_1^2$  and  $\sigma_2^2$  unknown

When  $\sigma_1^2$  and  $\sigma_2^2$  are known and both populations are normal, the variance of  $\bar{X}_1 - \bar{X}_2$  is

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

...and the random variable

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

has a standard normal distribution

6

Then sigma1 square and sigma2 square known we will find out what is the variance of that population when sigma1 square and sigma2 square are known both the populations are normal the variance of the difference in the variance of population 1 and 2 is nothing but the summation of their variance that is Sigma 1 square by n 1 plus Sigma 2 square by n 2 which I have already

explained. So, corresponding Z statistics is  $\bar{X}_1 - \bar{X}_2 - \mu_1 - \mu_2$  divided by  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$  has a standard normal distribution.

So, if it is a 1 sample what was our formula if it is a 1 sample our formula was like this  $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$  the denominator  $\frac{\sigma}{\sqrt{n}}$  is called standard error. So, in this formula previously for one sample you have taken only one sample here there is 2 sample that is we are finding the difference of the 2 sample so it should be  $\bar{X}_1 - \bar{X}_2$ . Previously we have assumed one mean population mean now we are going to find the difference of the 2 population mean we are going to assume the difference of 2 population so see  $\mu_1 - \mu_2$  this standard error  $\frac{\sigma}{\sqrt{n}}$  is we got from this one.

So, this you have to take the square root of this so that will become root of  $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ , when the population means our independent samples first we will start from null hypothesis. So, now that  $\mu_1 - \mu_2$  will have some difference  $D_0$ , the test statistics for  $\mu_1 - \mu_2$  is  $\frac{\bar{X}_1 - \bar{X}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$  the previously what we had taken view we wrote  $\mu_1 - \mu_2$  as it is but instead of some  $\mu_1 - \mu_2$  we can write only the difference also here, so root of  $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ .

**(Refer Slide Time: 07:19)**

### Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

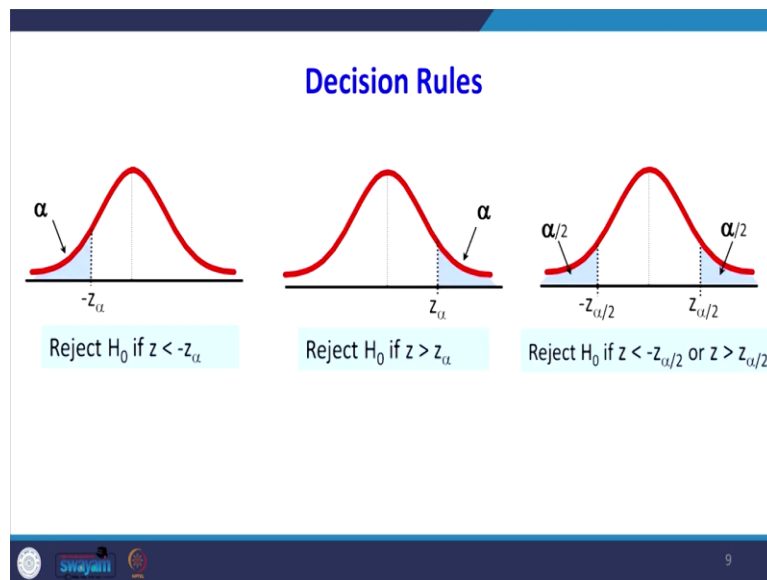
<p style="text-align: center;">Lower-tail test:</p> <p style="text-align: center;"><math>H_0: \mu_1 \geq \mu_2</math>  <math>H_1: \mu_1 &lt; \mu_2</math>              i.e.,  <math>H_0: \mu_1 - \mu_2 \geq 0</math>  <math>H_1: \mu_1 - \mu_2 &lt; 0</math></p>	<p style="text-align: center;">Upper-tail test:</p> <p style="text-align: center;"><math>H_0: \mu_1 \leq \mu_2</math>  <math>H_1: \mu_1 &gt; \mu_2</math>              i.e.,  <math>H_0: \mu_1 - \mu_2 \leq 0</math>  <math>H_1: \mu_1 - \mu_2 &gt; 0</math></p>	<p style="text-align: center;">Two-tail test:</p> <p style="text-align: center;"><math>H_0: \mu_1 = \mu_2</math>  <math>H_1: \mu_1 \neq \mu_2</math>              i.e.,  <math>H_0: \mu_1 - \mu_2 = 0</math>  <math>H_1: \mu_1 - \mu_2 \neq 0</math></p>
--	--	--

When both the Sigma are known to us there are different types of test as possible it may be a lower tailed test how we are saying lower tails is the left one you see that. The null hypothesis is

$\mu_1$  greater than or equal to  $\mu_2$  under hypothesis is  $\mu_1$  less than  $\mu_2$ . So, this will be a left tailed test otherwise you can bring this  $\mu_2$  to the left side so it will be  $\mu_1 - \mu_2$  greater than equal to 0 alternate hypothesis is  $\mu_1 - \mu_2$  less than 0 that means. So,  $H_0$  is  $\mu_1 - \mu_2$   $H_1$  is  $\mu_1$  greater than  $\mu_2$ , we can find the difference  $\mu_1 - \mu_2$  less than or equal to 0  $\mu_1 - \mu_2$  greater than 0 it is a right tailed test.

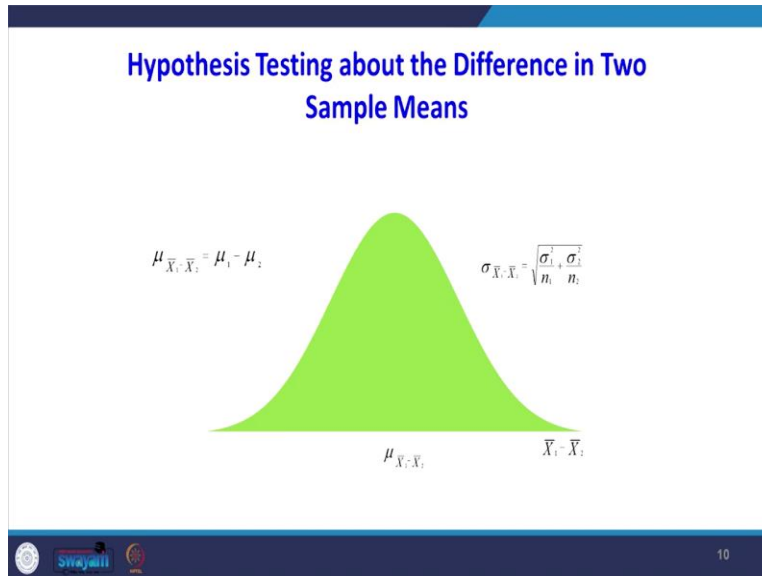
Here we can have an assumption that  $\mu_1$  equal to  $\mu_2$  and  $\mu_1$  not equal to  $\mu_2$ , if you bring on left hands to be  $\mu_1 - \mu_2$  equal to 0 then  $\mu_1 - \mu_2$  not equal to 0. We will see this different test pictorially.

**(Refer Slide Time: 08:28)**



The decision rule for the left tailed test is reject  $H_0$  if the calculated  $Z$  value is less than  $-Z$  alpha. The decision rule for the right tailed test is reject  $H_0$  if the calculated  $Z$  value is greater than  $Z$  alpha, for two tail test reject  $H_0$  if  $Z$  values less than  $-Z$  alpha by 2 are greater than  $Z$  alpha by 2 anyone can happen.

**(Refer Slide Time: 08:58)**



So this distribution is sampling distribution of difference of 2 samples. So, the mean of this distribution is mu of X 1 bar - X 2 bar the standard deviation of this distribution is Sigma 1 square by n 1 + Sigma 2 square by n 2 this already I have explained how we have got this values.

**(Refer Slide Time: 09:23)**

### Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

- Expected Value  $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$
- Standard Deviation (Standard Error)  $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

where:  $\sigma_1$  = standard deviation of population 1  
 $\sigma_2$  = standard deviation of population 2  
 $n_1$  = sample size from population 1  
 $n_2$  = sample size from population 2

For the sampling distribution of difference of 2 population mean the expected value is E of X 1 bar - X 2 bar equal to mu1 - mu2. So, the standard deviation is already we have seen root of Sigma 1 square by n 1 plus Sigma 2 square by n 2.

**(Refer Slide Time: 09:43)**

## Interval Estimation of $\mu_1 - \mu_2$ ; $\sigma_1$ and $\sigma_2$ Known

- Interval Estimate

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where:  $1 - \alpha$  is the confidence coefficient

The interval is  $\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$  this was for 2 sample. If it is a one sample what was the remember that it is  $\bar{X} \pm Z_{\alpha/2} \sigma/\sqrt{n}$ , so equation just as extended for 2 sample population. So, instead of  $\bar{X}$  we are going to write  $\bar{X}_1 - \bar{X}_2 \pm Z_{\alpha/2} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$  so what I am trying to say is that even though it is a 2 sample Z test the logic of extending one sample to 2 sample is very easy you need not remember the formula just intuitively you can extend the formula.

**(Refer Slide Time: 10:52)**

### Problem ( $\sigma_1$ and $\sigma_2$ Known)

- A product developer is interested in reducing the drying time of a primer paint.
- Two formulations of the paint are tested; formulation 1 is the standard chemistry, and formulation 2 has a new drying ingredient that should reduce the drying time.
- From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient.
- Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order.
- The two-sample average drying times are  $\bar{x}_1 = 121$  minutes and  $\bar{x}_2 = 112$  minutes, respectively.
- What conclusions can the product developer draw about the effectiveness of the new ingredient, using  $\alpha = 0.05$ ?

Source: Applied Probability and statistics for Engineers by Douglas C. Montgomery and George C. Runger John Wiley, 3rd Ed. 2003



We will do problem on this when Sigma 1 and Sigma 2 is known the problem is taken from this book applied probability and statistics for engineers by Montgomery. A product developer is interested in reducing the drying time of a primer paint 2 formulations of the painter tested. Formulation one is the standard chemistry and formulation 2 has new drying ingredient that should reduce the drying time. From experience it is known that the standard deviation of drying time is 8 minutes and this inherent variability should be unaffected by the addition of new ingredient.

10 specimens are painted with the formulation 1 and another 10 specimens are painted with the formulation 2. The 20 specimens are painted in random order to sample average drying times are  $\bar{X}_1$  is 121 minutes  $\bar{X}_2$  is 112 minutes respectively. What conclusions can the product developer draw about the effectiveness of the new ingredient?


**(Refer Slide Time: 12:12)**

### Problem ( $\sigma_1$ and $\sigma_2$ Known)

1. The quantity of interest is the difference in mean drying times,  $\mu_1 - \mu_2$ , and  $\Delta_0 = 0$ .
2.  $H_0: \mu_1 - \mu_2 = 0$ , or  $H_0: \mu_1 = \mu_2$ .
3.  $H_1: \mu_1 > \mu_2$ . We want to reject  $H_0$  if the new ingredient reduces mean drying time.
4.  $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where  $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$  and  $n_1 = n_2 = 10$ .


14

Assuming significance level alpha equal to 0.05, the first step in the hypothesis testing is the quantity of interest is the difference in the mean of drying time so  $\mu_1 - \mu_2$  if there is no difference it will become 0. Next we will form null hypothesis  $\mu_1 - \mu_2$  equal to 0 or  $\mu_1$  equal to  $\mu_2$  that means the mean drying time of ingredient 1 and 2 is same. The alternative hypothesis is  $\mu_1$  is greater than  $\mu_2$  what we are going to assume that the new ingredient is more efficient for that the drying time is going to be less.

If it is less will be greater and mu 2 will be lesser so we are writing mu 1 greater than mu 2 that is going to be our alternative hypothesis. We want to reject H 0 if the new ingredient reduces mean drying time for alpha 5% test statistic is  $\bar{X}_1 - \bar{X}_2 - \mu_1 - \mu_2$  that is 0 root of  $\text{Sigma 1 square by } n_1 + \text{Sigma 2 square by } n_2$  standard deviations given 8 so the variance is 64 the sample size is  $n_1 n_2$  equal to 10 when you supply this Sigma 1 square Sigma 2 square  $n_1$  and  $n_2$  in this formula.

**(Refer Slide Time: 13:30)**

**Problem ( $\sigma_1$  and  $\sigma_2$  Known)**

6. Reject  $H_0: \mu_1 = \mu_2$  if  $z_0 > 1.645 = z_{0.05}$ .

7. Computations: Since  $\bar{x}_1 = 121$  minutes and  $\bar{x}_2 = 112$  minutes, the test statistic is

$$z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

So the rejection rule is if  $\mu_1$  equal to  $\mu_2$  that is a null hypothesis reject H 0 if the test statistics is greater than 1.645, how we got this 1.645 because you see that this test is right tailed test so right tailed test means it will be like this. So, when alpha equal to 0.05 corresponding Z value is 1.645 if the calculated Z is greater than 1.645 we have to reject our null hypothesis. Next we will compute our calculate our Z average test statistics.

After supplying X1 and X2 we are getting 2.52, so 2.52 will be in the rejection side so we have reject we have to reject null hypothesis.

**(Refer Slide Time: 14:29)**

### Problem ( $\sigma_1$ and $\sigma_2$ Known)

$$t = \frac{(121 - 112) - 0}{\sqrt{8^2 \left( \frac{1}{10} + \frac{1}{10} \right)}} = 2.52$$

Decision:  
Reject  $H_0$  at  $\alpha = 0.05$

Conclusion:  
There is evidence of a difference in means.

16

You see that 2.52 is lying on the rejection site that decision is we are to reject null hypothesis by comparing the critical value.

**(Refer Slide Time: 14:44)**

### Problem ( $\sigma_1$ and $\sigma_2$ Known)

8. Conclusion: Since  $z_0 = 2.52 > 1.645$ , we reject  $H_0: \mu_1 = \mu_2$  at the  $\alpha = 0.05$  level and conclude that adding the new ingredient to the paint significantly reduces the drying time. Alternatively, we can find the  $P$ -value for this test as

$$P\text{-value} = 1 - \Phi(2.52) = 0.0059$$

Therefore,  $H_0: \mu_1 = \mu_2$  would be rejected at any significance level  $\alpha \geq 0.0059$ .

17

The same problem can be done with help of comparing p values also so what conclusion we can have since the Z calculated is that is 2.52 is greater than 1.645 we reject  $H_0$  that is  $\mu_1 = \mu_2$  at the Alpha equal to 0.05 level and conclude that adding new ingredient to the paint significantly reduce the drying time. Since we reject null hypothesis we are going to accept  $H_1$  hypothesis that says that new ingredient is reducing the drying time.

Alternatively we can find the p-value for this test, so because it is a right-tailed test the p-value should be 1 minus when calculated Z value is 2.5 minus corresponding probability so we will get we got 0.0059 I will verify this with the help of Python how we got this 0.0059, see these 0.0059 we are comparing with alpha. So, when we have to reject they see that therefore  $H_0: \mu_1 = \mu_2$  would be rejected to any significance level alpha is greater than 0.0059 what is happening here the value of p is very less. So we are to reject our null hypothesis.

**(Refer Slide Time: 15:59)**

```

Problem ( $\sigma_1$  and  $\sigma_2$  Known)

In [2]: import pandas as pd
import numpy as np
import math
from scipy import stats

In [6]: def z_and_p(x1,x2,sigma1,sigma2,n1,n2):
z = (x1-x2)/(math.sqrt(((sigma1**2)/n1)+((sigma2**2)/n2)))
if(z < 0):
p = stats.norm.cdf(z)
else:
p = 1 - stats.norm.cdf(z)
print (z,p)

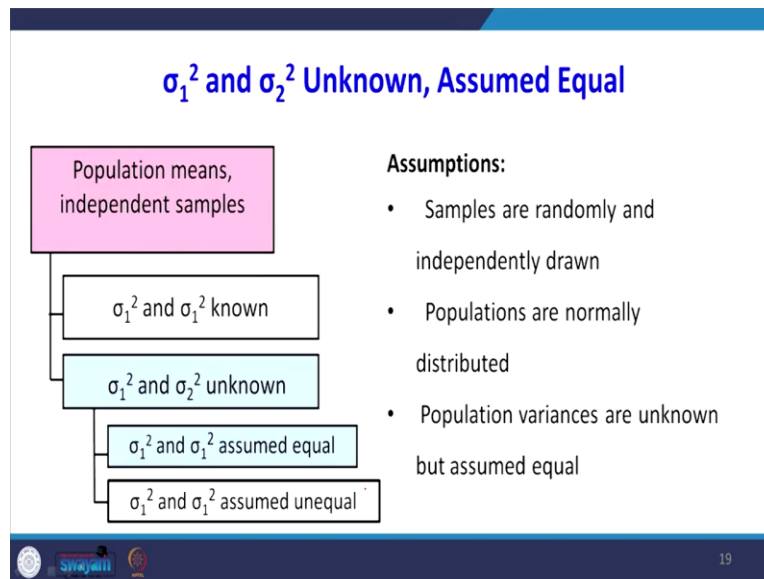
In [7]: z_and_p(121,112,8,8,10,10)
2.5155764746872635 0.00594189462107364
  
```

So we have done your Python code for this import pandas as pd import numpy as np import math from scipy import stats we are going to make one definition here we are going to define your function define def that is a standard syntax Z and underscore and underscore p the variable which are going to take is X 1 X 2 Sigma 1 Sigma 2 n 1 n 2 with useful notations. Then we will find out the Z value Z value is  $X_1 - X_2$  square root of  $\text{Sigma}_1^2 \text{ by } n_1 + \text{Sigma}_2^2 \text{ by } n_2$  the value of Z is less than 0 then p value is nothing but the actual value you can read as it is.

When the Z is it is like this if the Z value is coming on the left hand side the p value you can read as it is but if the if the z values positive we need the right side area so 1 minus the left side area will give you the right side area p equal to 1 - stat stat norm dot cdf upset so print Z p. So, this code in our monitor just you can try this after pausing this video you can verify this answer. So, Z underscore Z underscore and underscore P just and supply all the values.

In the previous problem the  $\bar{X}_1$  is 121  $\bar{X}_2$  is 112  $\sigma_1$  is the 8  $\sigma_2$  is 8  $n_1$  is 10  $n_2$  is 10, what happening we are getting the Z value is 2.51 there is a Z calculated value look at this p value and you go to previous slide say p value here also we got 0.0059, so here we get with the help of Python. When you compare with the alpha this is very small we are rejecting null hypothesis.

**(Refer Slide Time: 18:01)**



Now we will go to the second category of the problem. So, far in the previous problem we know  $\sigma_1^2$   $\sigma_2^2$  but this case  $\sigma_1^2$   $\sigma_2^2$  is unknown but we are assuming it is equal. There is a concept behind why we are assuming it is equal whenever we make comparison we can compare our the comparison is meaningful only when the variance of 2 groups are equal like comparing the performance of third year versus fourth year student is there is no meaning for that.

We can compare only the third year student versus another third year students so that way the variance has to be equal then only the comparison will be meaningful. So, the second case you see the blue 1  $\sigma_1^2$  and  $\sigma_2^2$  unknown when done we are going to make another assumption that it is equal there is another possibility it is unknown but unequal we will come to the that one in the after sometime.

First we will go Sigma 1 square Sigma 2 square unknown but assumed equal. What are the assumption we are making samples are randomly and independently drawn populations are normally distributed population variances are unknown but assumed to be equal.

**(Refer Slide Time: 19:15)**

**$\sigma_1^2$  and  $\sigma_2^2$  Unknown, Assumed Equal**

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate  $\sigma$
- use a t value with  $(n_1 + n_2 - 2)$  degrees of freedom

20

The population variance are assumed equal use the 2 sample standard deviation and fold them to estimate the variance or standard deviation use your t value with  $n_1 + n_2 - 2$  degrees of freedom. Actually what the concept here is assume that there is a group 1 this variance is  $s_1$  square suppose  $n_1$  sample there is a group 2 the variance is  $s_2$  square this sample sizes  $n_2$ . Suppose assuming population variance are equal you can pull the variance how we can do the pull the variance.

We can find out the weighted variance that we are going to called pooled variants the weighted variance is nothing but suppose assume that we are going to find out the weighted mean. So, what is the formula for weighted mean suppose  $W_1 X_1 + W_2 X_2$  divided by sum of weighted  $W_1 + W_2$  this is nothing but your weighted mean. Here the weight is nothing but your degrees of freedom. Suppose for the sample 1 the degrees of freedom is  $n_1 - 1$  here the variance is  $s_1$  square plus sample to the weighted is corresponding degrees of freedom  $s_2$  square.

So, next we have to sum the degrees of freedom  $n_1 - 1 + n_2 - 1$  that is nothing but  $n_1 + n_2 - 2$  so, this is nothing but the pooled variance.


(Refer Slide Time: 20:59)

### Test Statistic, $\sigma_1^2$ and $\sigma_2^2$ Unknown, Equal

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

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

Where t has  $(n_1 + n_2 - 2)$  d.f.,  
and

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



The test statistics for  $\mu_1 - \mu_2$  is say previously we are used as ed now we are using  $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$  since  $s_p^2$  is same we can bring left hand side that is nothing but pooled variance that pooled variances see that  $(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$  divided by  $n_1 + n_2 - 2$  degrees of freedom. You see that here the degrees of freedom is  $n_1 + n_2 - 2$ .

(Refer Slide Time: 21:38)

### Decision Rules

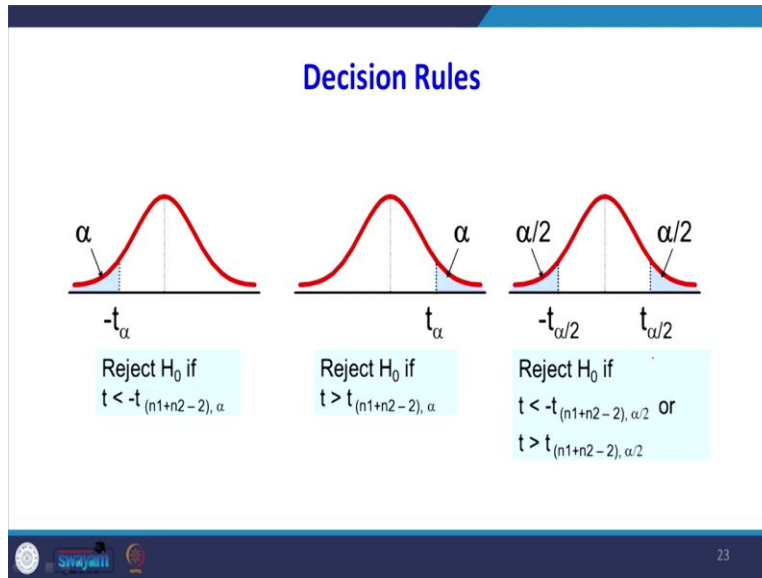
Two Population Means, Independent Samples, Variances Unknown

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



Then we do the population mean standard deviation as unknown the error also there is a possibility to see this test is left tailed test this is right tailed test this is 2 tailed test.

(Refer Slide Time: 21:46)



The next slides see that if the - t alpha no alpha this one this is a left tailed test this is the right tail test middle one the right hand side was is the 2 tailed test.

**(Refer Slide Time: 21:57)**

### $\sigma_1^2$ and $\sigma_2^2$ Unknown, Assumed equal

- Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process.
- Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable.
- Since catalyst 2 is cheaper, it should be adopted, providing it does not change the process yield.
- A test is run in the pilot plant and results in the data shown in table.
- Is there any difference between the mean yields?
- Use 0.05, and assume equal variances.

Observation Number	Catalyst 1	Catalyst 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$

We will take one problem where Sigma 1 square and Sigma 2 square unknown assumed equal, 2 catalyst are being analyzed to determine how they affect the mean yield of a chemical process. Specifically catalyst 1 is currently in use but catalyst 2 is acceptable. Since catalyst 2 is cheaper it should be adapted providing it does not change the process yield a test run in the pilot plant and the result in the data shown in the table.



Is there any difference between mean yields use alpha equal to 0.05 and assume equal variances. By looking at this problem you see that how it is given the variance are equal, no where the population variance is given so we should go for sample t-test assuming equal variance.

**(Refer Slide Time: 22:58)**

**$\sigma_1^2$  and  $\sigma_2^2$  Unknown, Assumed equal**

1. The parameters of interest are  $\mu_1$  and  $\mu_2$ , the mean process yield using catalysts 1 and 2, respectively, and we want to know if  $\mu_1 - \mu_2 = 0$ .
2.  $H_0: \mu_1 - \mu_2 = 0$ , or  $H_0: \mu_1 = \mu_2$
3.  $H_1: \mu_1 \neq \mu_2$
4.  $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

25

As usual the step one is we have to see which parameter of the population we are studying whether mean or variance. Now it is a mean so the parameter of interest are  $\mu_1$  and  $\mu_2$  the mean process yield using catalyst one and 2 respectively and we want to know if  $\mu_1 - \mu_2$  equal to 0. So,  $\mu_1 - \mu_2$  equal to 0, so  $H_1$  is  $\mu_1 \neq \mu_2$  alpha 0.05 there is the test status is  $t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$  so  $s_p$  square was inside root we brought left side so it is the pooled standard deviation root of 1 by  $n_1 + 1$  by 2.

**(Refer Slide Time: 23:47)**

## $\sigma_1^2$ and $\sigma_2^2$ Unknown, Assumed equal

6. Reject  $H_0$  if  $t_0 > t_{0.025,14} = 2.145$  or if  $t_0 < -t_{0.025,14} = -2.145$ .
7. Computations: From Table 10-1 we have  $\bar{x}_1 = 92.255, s_1 = 2.39, n_1 = 8, \bar{x}_2 = 92.733, s_2 = 2.98, \text{ and } n_2 = 8$ . Therefore

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(7)(2.39)^2 + 7(2.98)^2}{8 + 8 - 2} = 7.30$$

$$s_p = \sqrt{7.30} = 2.70$$

and

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{2.70 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70 \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$



So, what will happen when we look at the say statistical table when 14 degrees of freedom because it is a 2 tile, so since it is 2 tile but this area is 0.025 this area is 0.025 when 14 degrees of freedom the right hand side value is 2.145 I am writing here at the bottom it is 2.145 the left hand side it is  $-2.145$  it is symmetric so positive or negative from the previous slide we have the mean of sample one is 92.255 and standard deviation of the sample one is 2.39 for  $n_1$  equal to 8.

Similarly for the sample 2 the sample mean is 92.733 a standard deviation is 2.98 and  $n_2$  is 8 therefore first we will find the pooled variance by using the formula  $n_1 - 1 s_1^2 + n_2 - 1 s_2^2$  divided by  $n_1 + n_2 - 2$  after substituting this value we are getting 7.30 we could take the square root of that will get the standard deviation so use 2.70.

**(Refer Slide Time: 25:10)**

## $\sigma_1^2$ and $\sigma_2^2$ Unknown, Assumed equal

8. Conclusions: Since  $-2.145 < t_0 = -0.35 < 2.145$ , the null hypothesis cannot be rejected. That is, at the 0.05 level of significance, we do not have strong evidence to conclude that catalyst 2 results in a mean yield that differs from the mean yield when catalyst 1 is used.

In this t formula we are getting - 0.35 obviously you have to locate this is the rejection region. So, what we are concluding since  $-2.145$  is less than that what we are got the value is going on the left hand side that is - 0.35 the calculated t value is - 0.35 so in this the - 0.35 will be the acceptance side. So, we have to accept null hypothesis so what is happening - 0.35 the null this cannot be rejected that is at the 5% level of significance we do not have strong evidence to conclude that the catalyst 2 result in a mean yield that differs from mean yield when catalyst 1 is used.

**(Refer Slide Time: 25:59)**

## $\sigma_1^2$ and $\sigma_2^2$ Unknown, Assumed equal

```
In [12]: b = [ 89.19,90.95,90.46,93.21,97.19,97.04,91.07 , 92.75]
In [13]: a = [ 91.5, 94.18,92.18,95.39,91.79,89.07,94.72,89.21]
In [14]: stats.ttest_ind(a, b, equal_var = True)
Out[14]: Ttest_indResult(statistic=-0.3535908643461798, pvalue=0.7289136186068217)
In [21]: stats.t.ppf(0.025,14) #critical t value
Out[21]: -2.1447866879169277
```

So, now with the help of Python when Sigma 1 square Sigma 2 square unknown assuming equal variance will solve the problem. Previously I am taking the be equal to I am assigning into an

objective b I have taken an array a, I have taken the next one. So, when he stats dot t-test underscore independent you call that array a, b equal variance and equal to true right it can be true or false the next after sometime we will solve that problem when it is a true directly we are getting you see that the test statistics - 0.35, so the p value is 0.72.

So we have to accept our null hypothesis we can see how we got  $-2.144$  also stat at t dot cdf if you want to know the key value 0.025 when 40 degrees of freedom so we can compare t values also see that so when you see that one so the t value is  $-2.14$  but our test statistics - 0.35 it is lying on the acceptance region we are accepting our null hypothesis. Dear students so for what we have seen we are comparing hypothesis testing for 2 sample.

We have seen three types of problems number one is Sigma 1 square Sigma 2 square is known then we have compared to the mean. Another type of problem is Sigma 1 square and Sigma 2 square is unknown and we have assumed equal variance. We have done the Z test and we have done the t-test and also I have explained the concept behind of standard deviation of the difference of 2 population variance.

So what is the concept there if I want to know the difference of the 2 population variance you have to add the variance. If you want to know the difference of the 2 population mean just you can find the difference of the population mean. In the next class we will take a new problem where Sigma 1 square Sigma 2 square unknown but if it is unequal variance with that we will start the next class, thank you very much.