IIT ROORKEE

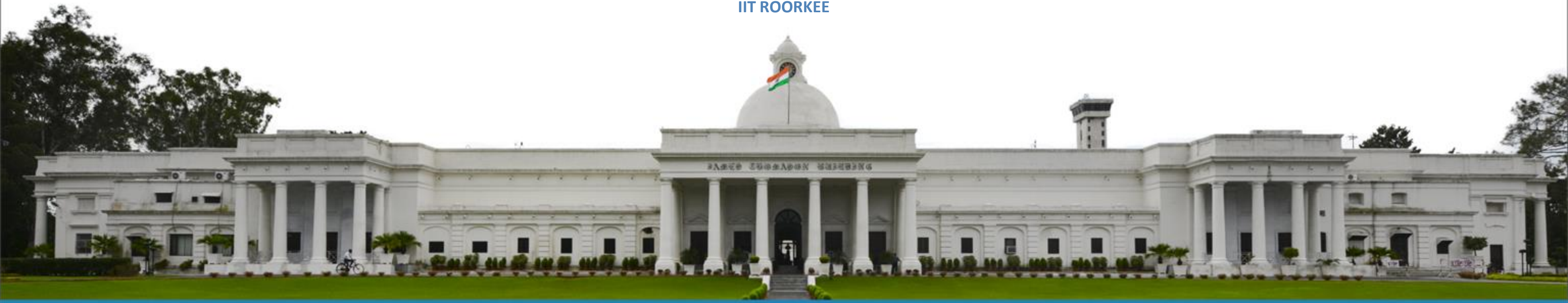
NPTEL ONLINE CERTIFICATION COURSE

Hypothesis Testing: Two sample test

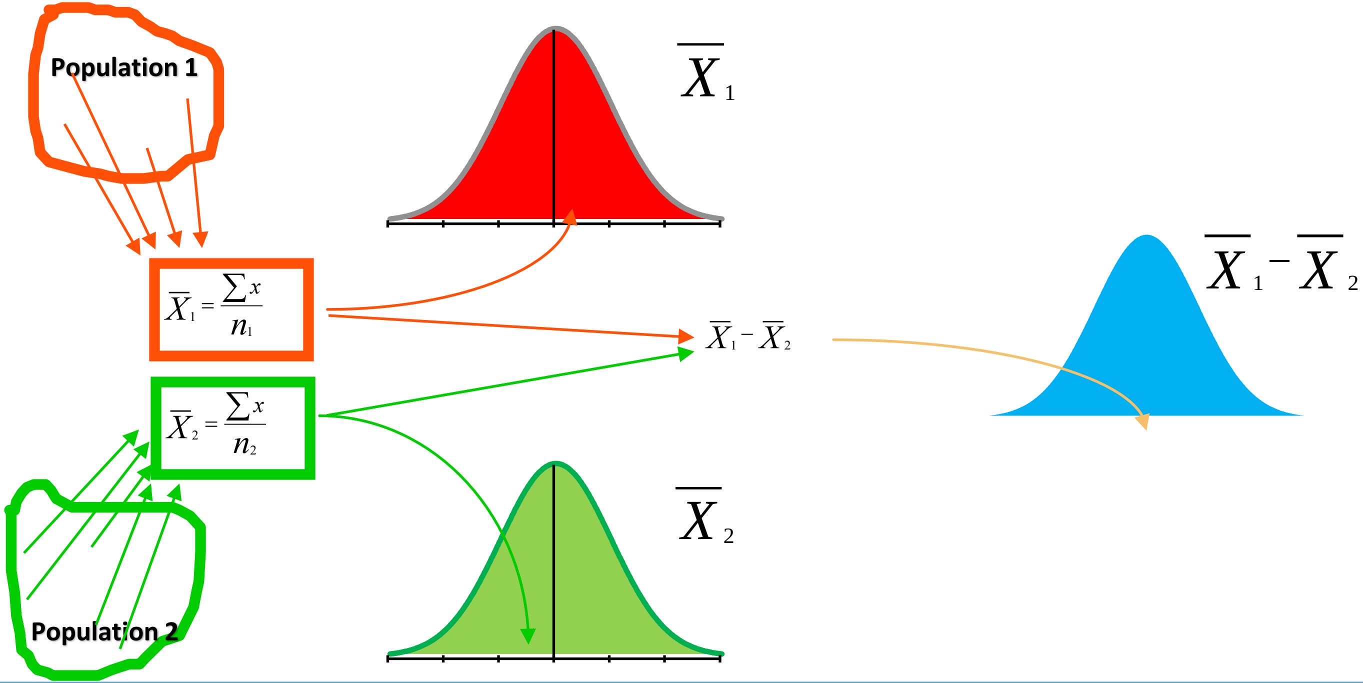
Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT

IIT ROORKEE



Hypothesis Testing about the Difference in Two Sample Means



Two Sample Tests

Two Sample Tests

- Population Means, Independent Samples
- Population Means, Dependent Samples
- Population Proportions
- Population Variances

Examples:

- Group 1 vs. independent Group 2
- Same group before vs. after treatment
- Proportion 1 vs. Proportion 2
- Variance 1 vs. Variance 2

Difference Between Two Means

- Population means, independent samples
  - σ_1^2 and σ_2^2 known → Test statistic is a **z** value
  - σ_1^2 and σ_2^2 unknown
    - σ_1^2 and σ_2^2 assumed equal → Test statistic is a value from the Student's **t** distribution
    - σ_1^2 and σ_1^2 assumed unequal → Test statistic is a value from the Student's **t** distribution

σ_1^2 and σ_1^2 Known

Population means, independent samples

σ_1^2 and σ_2^2 known

σ_1^2 and σ_2^2 unknown

Assumptions:

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

σ_1^2 and σ_2^2 Known

Population means, independent samples

- σ_1^2 and σ_2^2 known
- σ_1^2 and σ_2^2 unknown

When σ_x^2 and σ_y^2 are known and both populations are normal, the variance of 1 – 2 is

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

...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

Test Statistic, σ_1^2 and σ_2^2 Known

Population means, independent samples

- σ_1^2 and σ_2^2 known
- σ_1^2 and σ_2^2 unknown

$$H_0: \mu_1 - \mu_2 = D_0$$

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

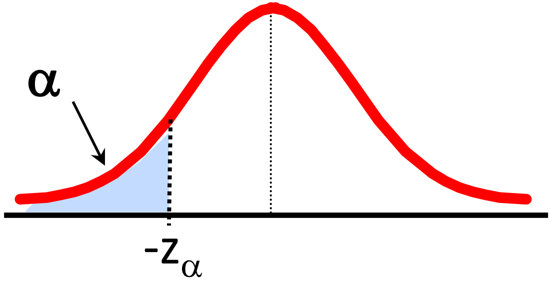
$$z = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Hypothesis Tests for Two Population Means

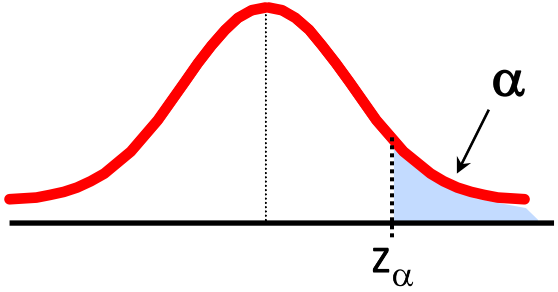
Two Population Means, Independent Samples

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

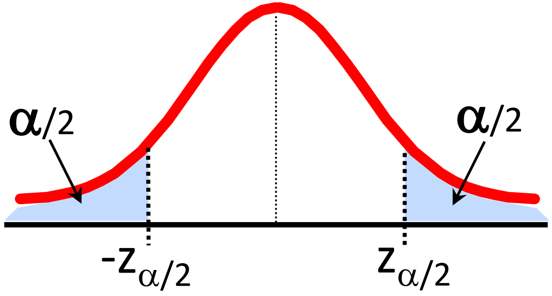
Decision Rules



Reject H_0 if $z < -z_\alpha$



Reject H_0 if $z > z_\alpha$



Reject H_0 if $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$

Hypothesis Testing about the Difference in Two Sample Means

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_1 - \mu_2$$

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

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: σ_1 = standard deviation of population 1

σ_2 = standard deviation of population 2

n_1 = sample size from population 1

n_2 = sample size from population 2

Interval Estimation of μ_1 - μ_2: σ_1 and σ_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

Problem (σ_1 and σ_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 $\overline{x_1} = 121$ minutes and $\overline{x_2}$ = 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*

Problem (σ_1 and σ_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

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

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

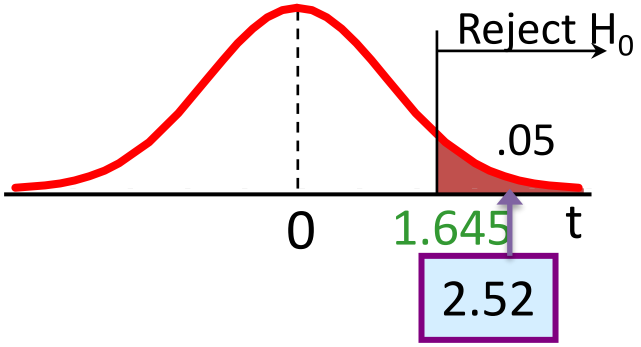
Problem (σ_1 and σ_2 Known)

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

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

Problem (σ_1 and σ_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.

Problem (σ_1 and σ_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 $\alpha \geq 0.0059$.

Problem (σ_1 and σ_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 0.005941894621073 64
```

σ_1^2 and σ_2^2 Unknown, Assumed Equal

- Population means, independent samples
  - σ_1^2 and σ_1^2 known
  - σ_1^2 and σ_2^2 unknown
    - σ_1^2 and σ_1^2 assumed equal
    - σ_1^2 and σ_1^2 assumed unequal

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed
- Population variances are unknown but assumed equal

σ_1^2 and σ_2^2 Unknown, Assumed Equal

- The population variances are assumed equal, so use the two sample standard deviations and pool them to estimate σ

- use a t value with $(n_1 + n_2 - 2)$ degrees of freedom

Test Statistic, σ_1^2 and σ_2^2 Unknown, Equal

The test statistic for

$\mu_1 - \mu_2$ is:

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

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

and

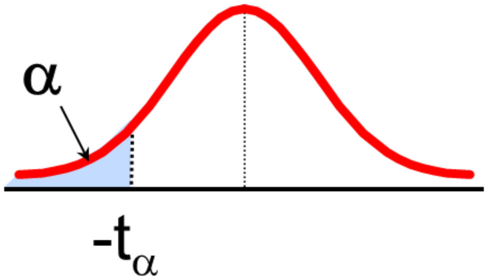
$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Decision Rules

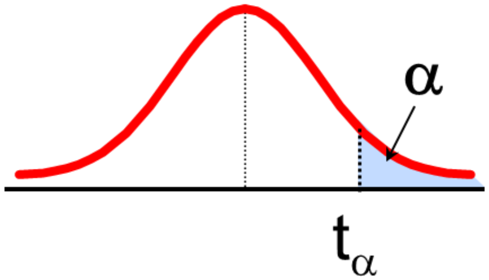
Two Population Means, Independent Samples, Variances Unknown

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

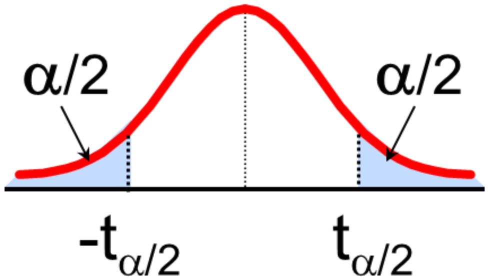
Decision Rules



Reject H_0 if
$t < -t_{(n1+n2-2),\ \alpha}$



Reject H_0 if
$t > t_{(n1+n2-2),\ \alpha}$



Reject H_0 if
$t < -t_{(n1+n2-2),\ \alpha/2}$ or
$t > t_{(n1+n2-2),\ \alpha/2}$

σ_1^2 and σ_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

$\bar{x}_1 = 92.255$ $\bar{x}_1 = 92.733$

$S_1 = 2.39$ $S_2 = 2.98$

σ_1^2 and σ_2^2 Unknown, Assumed equal

1. The parameters of interest are μ_1 and μ_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

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

σ_1^2 and σ_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 $\bar{x}_1 = 92.255$, $s_1 = 2.39$, $n_1 = 8$, $\bar{x}_2 = 92.733$, $s_2 = 2.98$, 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

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

σ_1^2 and σ_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.

σ_1^2 and σ_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
```

Thank You