Data Analytics

Yong Zheng
Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Schedule

- Quick Reviews
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

Schedule

- Quick Reviews
  - Use Sample to Estimate Population
    - Input: Sample data and confidence level
    - Output: confidence interval
  - Hypothesis Testing
    - Elements, steps and methods to make decisions
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Schedule

- Quick Reviews
- One-Sample Hypothesis Testing
- Two-Sample Hypothesis Testing

What is a hypothesis

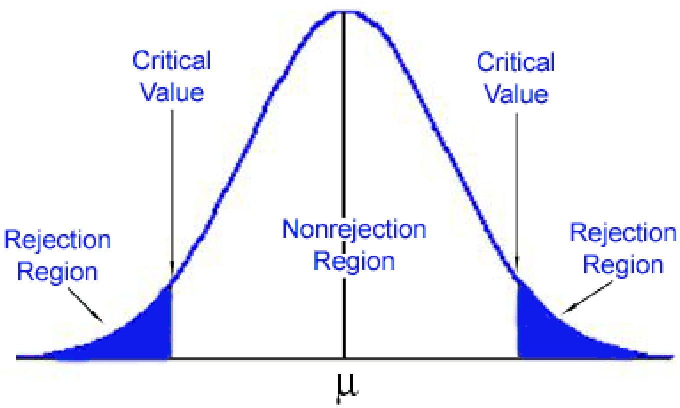
- Hypothesis is a claim or assumption
- Example
  - Average age is 30
  - Average age is no more than 30
  - Average age in NYC is larger than the one in Chicago
- Hypothesis Testing is used to validate an hypothesis is true or false based on a confidence level

Elements in Hypothesis Testing

- Null Hypothesis, H_0
  This is the hypothesis we have doubts
- Alternative Hypothesis, H_a or H_1
  This is the hypothesis which is counter to the null hypothesis. Usually it is what we want to support
- Test Statistics
  It is used to make decisions
- Level of significance, α
  The probability of rejecting H_0 giving H_0 is true

Elements in Hypothesis Testing

- Rejection Region



If our test statistics faill into rejection region, we reject null hypothesis and accept the alternative hypothesis.

- P-value

It is a probability value between 0 and 1 as evidence to reject the null hypothesis.

95% confidence level, we reject H_0 if p-value<0.05

P-value = area under normal curve based on the test statistics

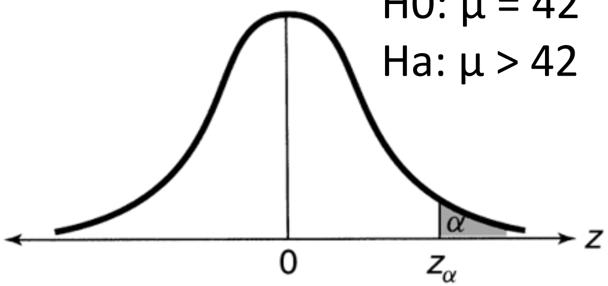
Types Hypothesis Testing: Based on Samples

- Hypothesis testing on one sample mean
  Monthly cell bill is $42
  I do not think this is true
- Hypothesis testing on two sample means
  Monthly cell bill by ATT and T-Mobile is the same
  ATT is more expensive than T-Mobile

Elements in Hypothesis Testing: Based on Ha

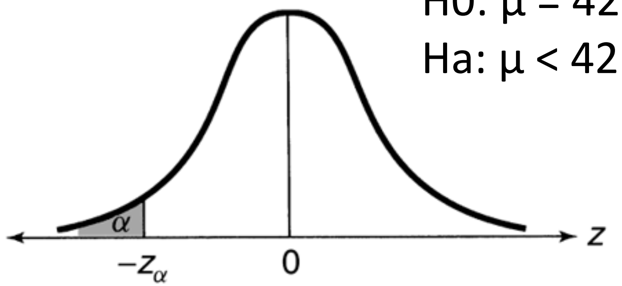
- One-sided or one-tailed statistical test

H0: μ = 42
Ha: μ > 42



(a) $\mu > \mu_0$

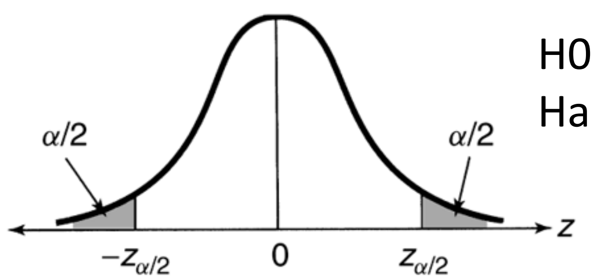
H0: μ = 42
Ha: μ < 42



(b) $\mu < \mu_0$

- Two-sided or two-tailed statistical test

H0: μ = 42
Ha: μ ≠ 42



(c) $\mu \neq \mu_0$

Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_a
2. Based on Ha, decide it is one-tailed or two-tailed test
3. Choose the level of significance, α. Or, you can claim statistical confidence level, $\alpha = 1 -$ confidence level
4. Determine the appropriate test statistic and sampling distribution – depends on sample size
5. Determine the critical values that divide the rejection and non-rejection regions

Steps in Hypothesis Testing

5. Make the statistical decision and state the managerial conclusion.
   - By using test statistics
     If it falls in the rejection area, we reject H0
   - By using p-value
     If the p-value < α, we reject Ho and accept Ha
   - By using confidence interval
     Note, the acceptance region is the confidence interval based on the confidence level

Hypothesis testing on one sample mean

Hypothesis testing on one sample mean

Hypothesis Tests for μ

σ Known, n>=30

(z test)

The test statistic is:

$$Z_{STAT} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ Unknown, n<30

(t test)

The test statistic is:

$$t_{STAT} = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Hypothesis testing on one sample mean

Large-Sample ($n \geq 30$) Test of Hypothesis About μ

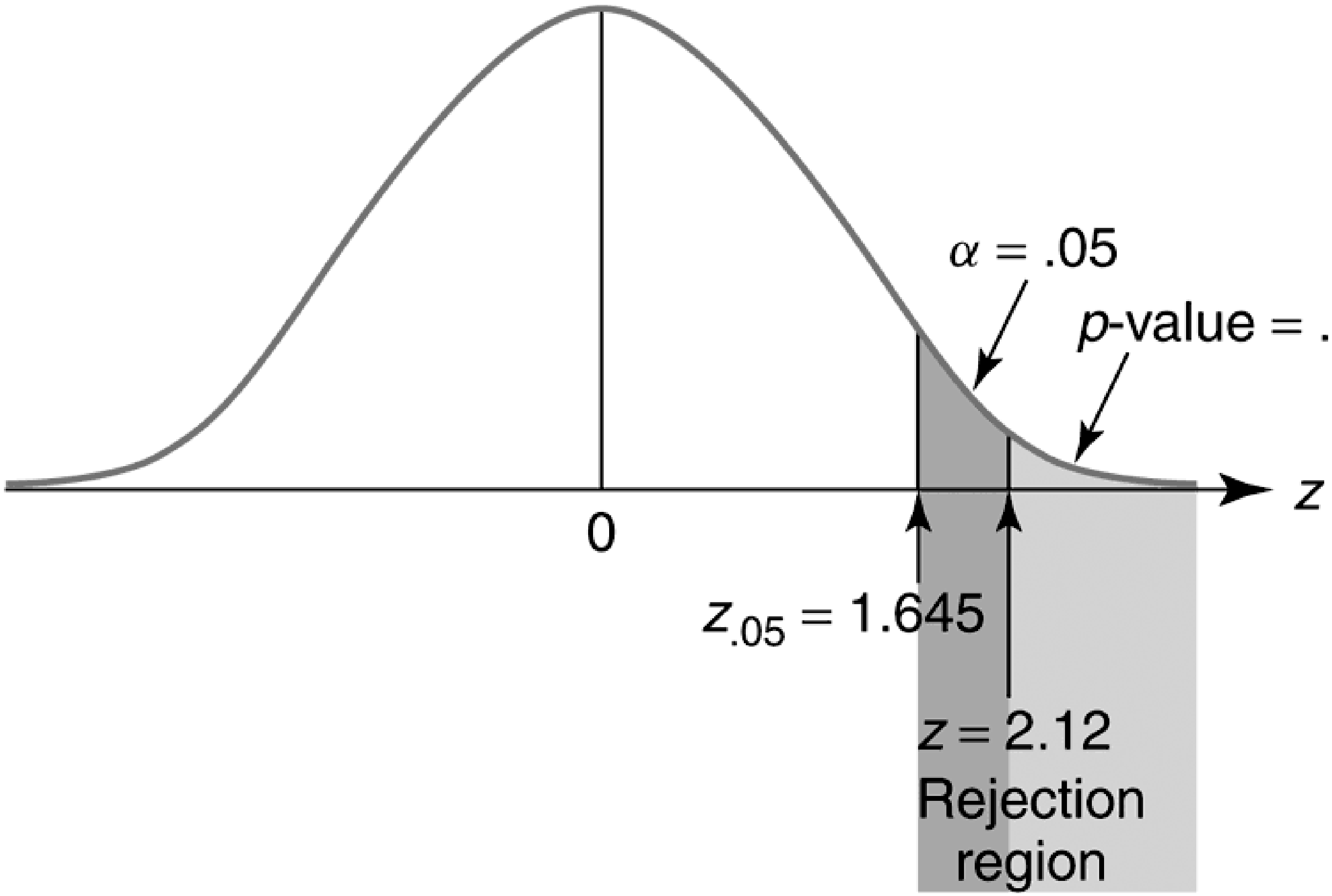
Test statistic: $z = (\bar{y} - \mu_0)/\sigma_{\bar{y}} \approx (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$
Rejection region:	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
p-*value*:	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c is positve $2P(z < z_c)$ if z_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or if test statistic falls in rejection region

where $P(z > z_\alpha) = \alpha$, $P(z > z_{\alpha/2}) = \alpha/2$, z_c = calculated value of the test statistic, and $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ true})$.

Hypothesis testing on one sample mean



Hypothesis testing on one sample mean

Small-Sample Test of Hypothesis About μ

Test statistic: $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve $2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or if test statistic falls in rejection region

where $P(t > t_\alpha) = \alpha$, $P(t > t_{\alpha/2}) = \alpha/2$, t_c = calculated value of the test statistic, and α = P(Type I error) = P(Reject $H_0 | H_0$ true).

Assumption: The population from which the random sample is drawn is approximately normal.

Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_a
2. Based on Ha, decide it is one-tailed or two-tailed test
3. Choose the level of significance, α. Or, you can claim statistical confidence level, $\alpha = 1 -$ confidence level
4. Determine the appropriate test statistic and sampling distribution – depends on sample size
5. Determine the critical values that divide the rejection and non-rejection regions

Steps in Hypothesis Testing

5. Make the statistical decision and state the managerial conclusion.

- By using test statistics
  If it falls in the rejection area, we reject H0
- By using p-value
  If the p-value < α, we reject Ho and accept Ha
- By using confidence interval
  Note, the acceptance region is the confidence interval based on the confidence level

Example: Room Price

The average cost of a hotel room in New York is said to be $168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an $\bar{X}$ of $172.50 and an S of $15.40. Test the appropriate hypotheses at $\alpha = 0.05$.

(Assume the population distribution is normal)

H_0: $\mu = 168$
H_1: $\mu \neq 168$

Hypothesis testing on one sample mean

Hypothesis Tests for μ

σ Known, n>=30

(z test)

The test statistic is:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

σ Unknown, n<30

(t test)

The test statistic is:

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

Hypothesis testing on one sample mean

Small-Sample Test of Hypothesis About μ

Test statistic: $t = (\bar{y} - \mu_0)/(s/\sqrt{n})$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$	$H_a: \mu \neq \mu_0$
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positve $2P(t < t_c)$ if t_c is negative

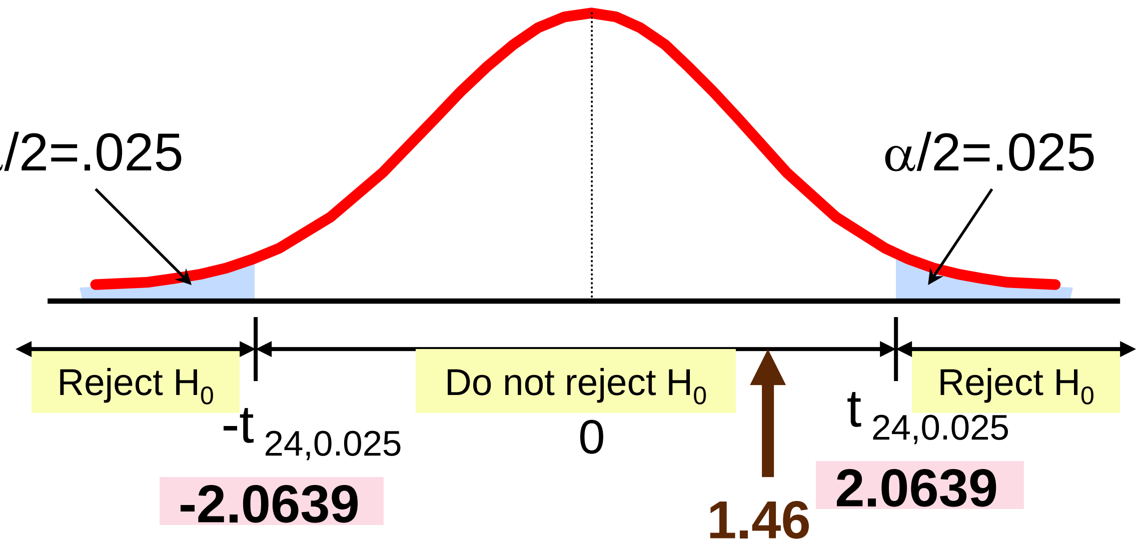
Decision: Reject H_0 if $\alpha > p$-value, or if test statistic falls in rejection region

where $P(t > t_\alpha) = \alpha$, $P(t > t_{\alpha/2}) = \alpha/2$, t_c = calculated value of the test statistic, and α = P(Type I error) = P(Reject $H_0 | H_0$ true).

Assumption: The population from which the random sample is drawn is approximately normal.

Example: Room Price

H_0: $\mu = 168$
H_1: $\mu \neq 168$



- $\alpha = 0.05$
- $n = 25$, $df = 25-1=24$
- σ is unknown, so
- use a **t statistic**
- Critical Value:
- $\pm t_{24,0.025} = \pm\ 2.0639$

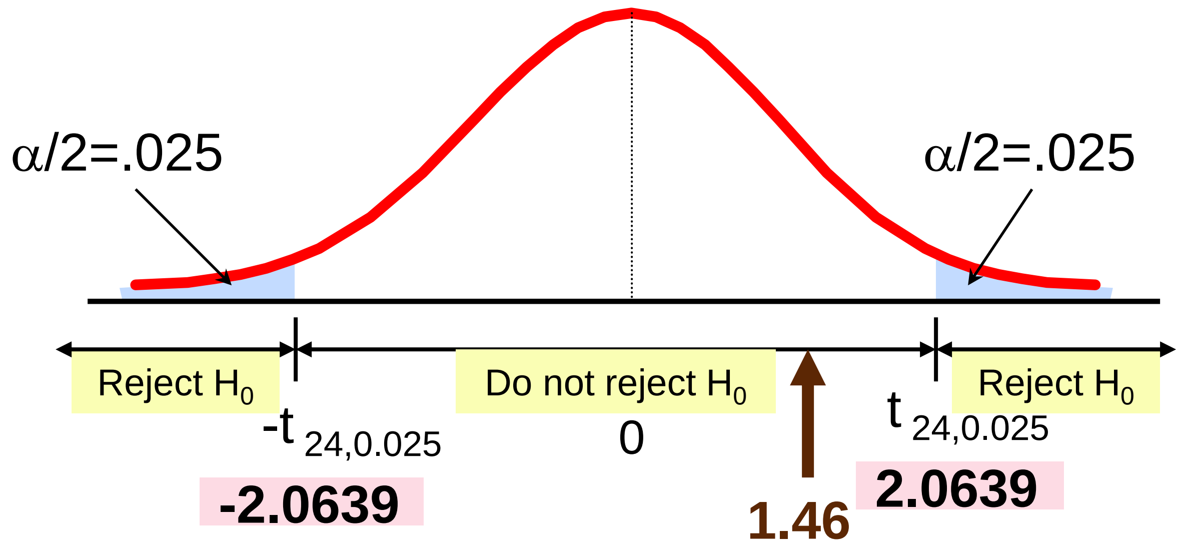
$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0: insufficient evidence that true mean cost is different from \$168

Example: Room Price

H_0: $\mu = 168$

H_1: $\mu \neq 168$



- α = **0.05**
- **n = 25, df = 25-1=24**
- **σ is unknown, so**
- **use a t statistic**
- **Critical Value:**
- $\pm t_{24,0.025} = \pm\ 2.0639$

$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

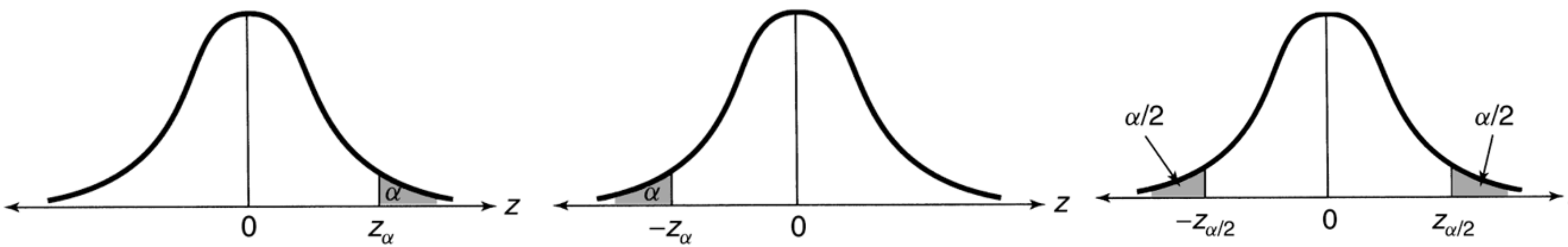
P-value = 0.1572 by two-tailed t-test and t = 1.46, df = 24; Do not reject H_0: Not statistically significant to reject H0

Note About T-test

- As long as the sample size is not very small (even if n<30) and the population is not very skewed, the t-test can be used. Basically you must make sure the population follows normal distribution
- To evaluate the normality assumption:
  - Determine how closely sample statistics match the normal distribution's theoretical properties.
  - Construct a histogram or stem-and-leaf display or boxplot or a normal probability plot.
  - We will learn more (normal test or QQ-Plot) in the next week
- Online tools for you to calculate p-value http://www.socscistatistics.com/pvalues/Default.aspx

Confusions

1. α and $\alpha/2$



(a) $\mu > \mu_0$ (b) $\mu < \mu_0$ (c) $\mu \neq \mu_0$

	ONE-TAILED TESTS		TWO-TAILED TEST
	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$	H_0: $\mu = \mu_0$
	H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$
Rejection region:	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
p-*value*:	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c is positve $2P(z < z_c)$ if z_c is negative

Confusions

1. α and $\alpha/2$

1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

2) Otherwise, t distribution, t value

$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

Confusions

2. z_α and p-value

1) If n >= 30, normal distribution, z value

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

2) Otherwise, t distribution, t value

$$\bar{y} \pm t_{\alpha/2}s_{\bar{y}} = \bar{y} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$, α = 1 – confidence level

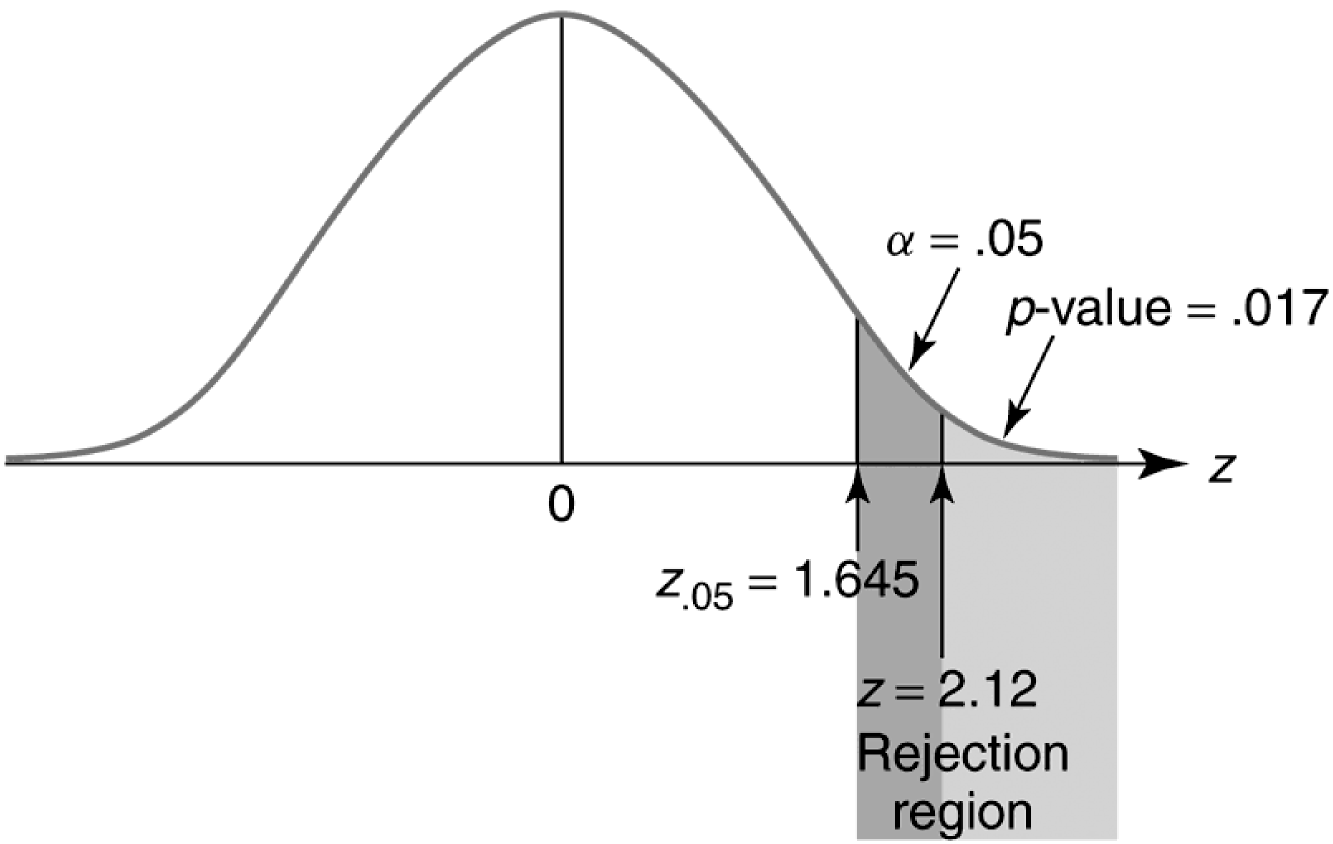
Confusions

2. z_α and p-value

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

Confusions

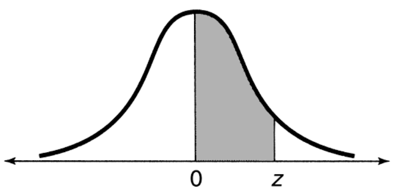
2. z_α and p-value



Confusions

2. z_α and p-value

Table 1.7 Reproduction of part of Table 1 of Appendix D



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

Example: Case Study 1
Student Grades

Hypothesis testing on two sample means

Types Hypothesis Testing

- Hypothesis testing on one sample mean
  Monthly cell bill is $42
  I do not think this is true
- Hypothesis testing on two sample means
  Monthly cell bill by ATT and T-Mobile is the same
  ATT is more expensive than T-Mobile

Dependency Between Two Samples

- Two Samples: Independent
  Students in ITMD525 are better than ones in ITMD527
  Population-1: students in ITMD525 with size n1
  Population-2: students in ITMD527 with size n2
  Note: n1 and n2 could be the same or different
- Two Samples: Paired
  Students perform better in ITMD525 than in ITMD527
  Population-1: Students enrolled in ITMD525
  Population-2: Students also enrolled in ITMD527
  Note: n1 and n2 must be the same – same students!!

Dependency Between Two Samples

- If they are paired two samples
  - Sample size must be the same
  - You can organize the data in two columns, each column represents a list of values for a sample
  - Each row in the data can be referred to a same standard, such as the data related to a same person
  - Different rows have different standards

Two-Sample Test

Two-Sample Tests

- Population Means, Independent Samples
  - Such as, Group 1 vs. Group 2
- Population Means, Paired Samples
  - Such as, Same group before vs. after treatment

Sample Statistics in Two-Sample Test

Table 1.13 Two-sample notation

	Population	
	1	2
Sample size	n_1	n_2
Population mean	μ_1	μ_2
Population variance	σ_1^2	σ_2^2
Sample mean	$\bar{y}_1$	$\bar{y}_2$
Sample variance	s_1^2	s_2^2

Two-Sample Test: Paired Samples

If you found it as two paired sample hypothesis testing, you should convert it to one sample hypothesis testing

Two-Sample Test: Paired Samples

Two Population Means, Paired Samples

Paired Difference Confidence Interval for $\mu_d = \mu_1 - \mu_2$

Large Sample

$$\bar{y}_d \pm z_{\alpha/2}\frac{\sigma_d}{\sqrt{n_d}} \approx \bar{y}_d \pm z_{\alpha/2}\frac{s_d}{\sqrt{n_d}}$$

Assumption: Sample differences are randomly selected from the population.

Small Sample

$$\bar{y}_d \pm t_{\alpha/2}\frac{s_d}{\sqrt{n_d}}$$

where $t_{\alpha/2}$ is based on $(n_d - 1)$ degrees of freedom

Assumptions:

1. Population of differences has a normal distribution.
2. Sample differences are randomly selected from the population.

Two-Sample Test: Paired Samples

Paired Difference Test of Hypothesis for $\mu_d = \mu_1 - \mu_2$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu_d = D_0$	$H_0: \mu_d = D_0$	$H_0: \mu_d = D_0$
	$H_a: \mu_d < D_0$	$H_a: \mu_d > D_0$	$H_a: \mu_d \neq D_0$

Large Sample

$$\textit{Test statistic: } z = \frac{\bar{y}_d - D_0}{\sigma_d/\sqrt{n_d}} \approx \frac{\bar{y}_d - D_0}{s_d/\sqrt{n_d}}$$

Rejection Region:	$z < -z_\alpha$	$z > z_d$	$\lvert z\rvert > z_{\alpha/2}$
p-*value*:	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c positive $2P(z < z_c)$ if z_c negative

Assumption: The differences are randomly selected from the population of differences.

Small Sample

$$\textit{Test statistic: } t = \frac{\bar{y}_d - D_0}{s_d/\sqrt{n_d}}$$

Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(t > t_c)$	$2P(t > t_c)$ if t_c is positive $2P(t < t_c)$ if t_c is negative

Assumptions:

1. The relative frequency distribution of the population of differences is normal.
2. The differences are randomly selected from the population of differences.

Case Study 2: Complaints

- Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? You collect the following data:

Salesperson	Number of Complaints: Before (1)	After (2)	(2) - (1) Difference, D_i
C.B.	6	4	- 2
T.F.	20	6	-14
M.H.	3	2	- 1
R.K.	0	0	0
M.O.	4	0	- 4
			-21

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

$$= -4.2$$

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

$$= 5.67$$

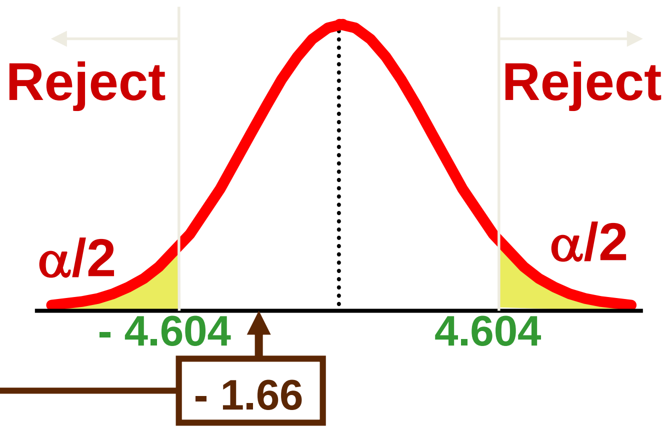
Case Study 2: Complaints

- Has the training made a difference in the number of complaints (at the 0.01 level)?

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

$\alpha = .01 \qquad \overline{D} = -4.2$

$t_{0.005} = \pm 4.604$
d.f. = n - 1 = 4



Test Statistic:

$$t_{STAT} = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} = \frac{-4.2 - 0}{5.67/\sqrt{5}} = \boxed{-1.66}$$

Decision: **Do not reject H_0**
(t_{stat} is not in the rejection region)

Conclusion: **There is insufficient of a change in the number of complaints.**

Hypothesis in Two-Sample Test

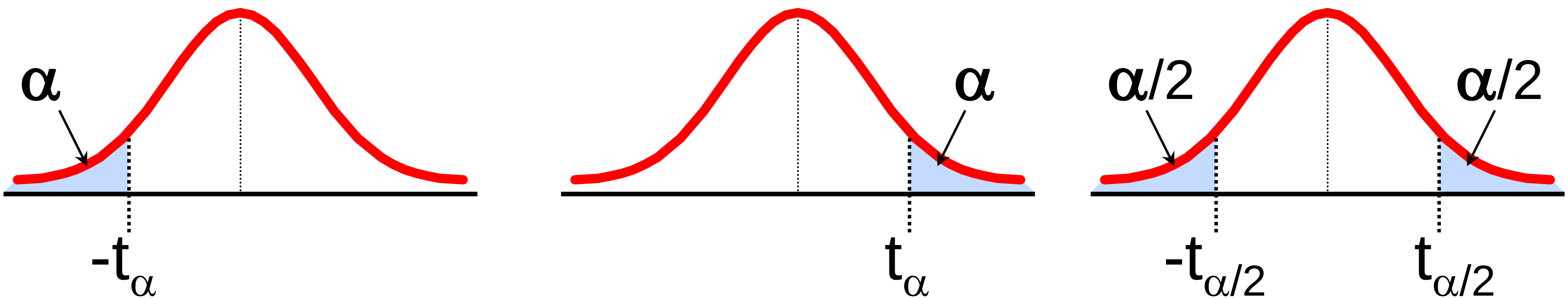
Two Population Means, Independent Samples

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

Hypothesis in Two-Sample Test

Two Population Means, Independent Samples

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



Reject H_0 if $t_{STAT} < -t_{\alpha}$

Reject H_0 if $t_{STAT} > t_{\alpha}$

Reject H_0 if $t_{STAT} < -t_{\alpha/2}$
or $t_{STAT} > t_{\alpha/2}$

Two-Sample Test: Independent Samples

- If they are independent, then follow the statistical way. But it is complicated.
- You do not need to remember the formula. We have statistical software to do that.

Two-Sample Test: Independent Samples

Two Population Means, Independent Samples

Large-Sample Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples

$$(\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sigma_{(\bar{y}_1 - \bar{y}_2)^*} = (\bar{y}_1 - \bar{y}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Assumptions: The two samples are randomly and independently selected from the two populations. The sample sizes, n_1 and n_2, are large enough so that $\bar{y}_1$ and $\bar{y}_2$ each have approximately normal sampling distributions and so that s_1^2 and s_2^2 provide good approximations to σ_1^2 and σ_2^2. This will be true if $n_1 \geq 30$ and $n_2 \geq 30$.

Two-Sample Test: Independent Samples

Two Population Means, Independent Samples

Large-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Independent Samples

$$\textit{Test statistic: } z = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sigma_{(\bar{y}_1 - \bar{y}_2)}} = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \mu_1 - \mu_2 = D_0$	$H_0: \mu_1 - \mu_2 = D_0$	$H_0: \mu_1 - \mu_2 = D_0$
	$H_a: \mu_1 - \mu_2 < D_0$	$H_a: \mu_1 - \mu_2 > D_0$	$H_a: \mu_1 - \mu_2 \neq D_0$
Rejection region:	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
p-*value*:	$P(z < z_c)$	$P(z > z_c)$	$2P(z > z_c)$ if z_c is positve $2P(z < z_c)$ if z_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or, if the test statistic falls in rejection region

where D_0 = hypothesized difference between means, $P(z > z_\alpha) = \alpha$, $P(z > z_{\alpha/2}) = \alpha/2$, z_c = calculated value of the test statistic, and α = P(Type I error) = P(Reject $H_0 | H_0$ true).

Assumptions: Same as for the previous large-sample confidence interval

Two-Sample Test: Independent Samples

Two Population Means, Independent Samples

Small-Sample Confidence Interval for $(\mu_1 - \mu_2)$: Independent Samples

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

where

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

is a "pooled" estimate of the common population variance and $t_{\alpha/2}$ is based on $(n_1 + n_2 - 2)$ df.

Assumptions:

1. Both sampled populations have relative frequency distributions that are approximately normal.
2. The population variances are equal.
3. The samples are randomly and independently selected from the populations.

Two-Sample Test: Independent Samples

Two Population Means, Independent Samples

Small-Sample Test of Hypothesis About $(\mu_1 - \mu_2)$: Independent Samples

$$\textit{Test statistic: } t = \frac{(\bar{y}_1 - \bar{y}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0\colon \mu_1 - \mu_2 = D_0$	$H_0\colon \mu_1 - \mu_2 = D_0$	$H_0\colon \mu_1 - \mu_2 = D_0$
	$H_a\colon \mu_1 - \mu_2 < D_0$	$H_a\colon \mu_1 - \mu_2 > D_0$	$H_a\colon \mu_1 - \mu_2 \neq D_0$
Rejection region:	$t < -t_\alpha$	$z > t_\alpha$	$\lvert t \rvert > z_{\alpha/2}$
p-*value*:	$P(t < t_c)$	$P(z > t_c)$	$2P(t > t_c)$ if t_c is positve $2P(t < t_c)$ if t_c is negative

Decision: Reject H_0 if $\alpha > p$-value, or, if test statistic falls in rejection region

where $D_0 =$ hypothesized difference between means, $P(t > t_\alpha) = \alpha$, $P(t > t_{\alpha/2}) = \alpha/2$, $t_c =$ calculated value of the test statistic, and $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ true})$.

Assumptions: Same as for the previous small-sample confidence interval.

Case Study 1: Stocks

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

	NYSE	NASDAQ
Number	**21**	**25**
Sample mean	**3.27**	**2.53**
Sample std dev	**1.30**	**1.16**

Assuming both populations are approximately normal with equal variances, is there a difference in mean yield ($\alpha = 0.05$)?

Case Study 1: Stocks

H0: $\mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$
H1: $\mu_1 - \mu_2 \neq 0$ i.e. $(\mu_1 \neq \mu_2)$

The test statistic is:

$$t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\left(3.27 - 2.53\right) - 0}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

$$S_p^2 = \frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 - 1) + (n_2 - 1)} = \frac{(21 - 1)1.30^2 + (25 - 1)1.16^2}{(21-1) + (25 - 1)} = 1.5021$$

Case Study 1: Stocks

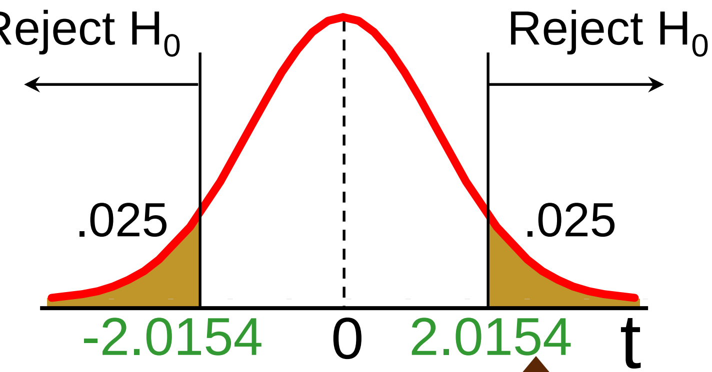
$H_0: \mu_1 - \mu_2 = 0$ i.e. $(\mu_1 = \mu_2)$

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

$\alpha = 0.05$

df = 21 + 25 - 2 = 44

Critical Values: t = ± 2.0154



2.040

Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021\left(\frac{1}{21} + \frac{1}{25}\right)}} = \boxed{2.040}$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.