Hypothesis Testing
Two Sample Sets

Dr. Supaporn Erjongmanee

Department of Computer Engineering
Kasetsart University
fengspe@ku.ac.th

1

Outline

- Population Mean Test
  - Normal and Known variance
  - Large sample size
  - Normal and Small sample size

2

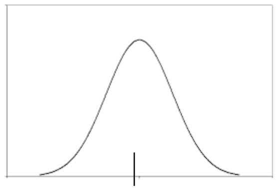
Hypothesis Test : General Idea

- Let θ be parameter
- Let $\hat{\theta}$ be estimate
- Let θ_0 be null value
- Test statistic = $\dfrac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}$
- Hypothesis
  - Null hypothesis (H_0): $\hat{\theta} = \theta_0$
  - Alternative hypothesis (H_a): $\hat{\theta} > \theta_0$; $\hat{\theta} < \theta_0$; $\hat{\theta} \neq \theta_0$

To find where test statistic lies on standard normal

Standard Normal

N (μ=0, σ^2=1)



Supaporn Erjongmanee
fengspe@ku.ac.th

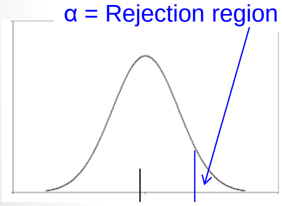
Department of Computer Engineering
Kasetsart University

3

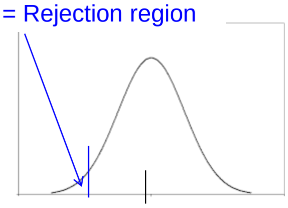
Hypothesis Test : General Idea (cont.)

Null hypothesis: $\hat{\theta} = \theta_0$

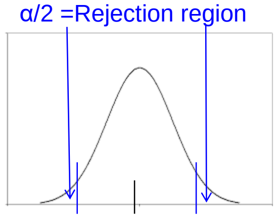
Alternative Hypothesis	Rejection Region at α level	Test
$H_a : \hat{\theta} > \theta_0$	$z \geq z_\alpha$	**Upper-tailed test**
$H_a : \hat{\theta} < \theta_0$	$z \leq -z_\alpha$	**Lower-tailed test**
$H_a : \hat{\theta} \neq \theta_0$	Either $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	**Two-tailed test**



Upper-tailed Test



Lower-tailed Test



Two-tailed Test

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

4

Inferences with Two Samples

- Assume we have two sample data sets: X and Y
- Basic assumptions:

1. $\{X_1, X_2, \ldots, X_m\}$ = set of m random samples with population mean = μ_1, standard deviation = σ_1
2. $\{Y_1, Y_2, \ldots, Y_n\}$ = set of n random samples with population mean = μ_2, standard deviation = σ_2
3. X and Y are independent of each other

$$E(\bar{X}) = \mu_1 \qquad V(\bar{X}) = \frac{\sigma_1^2}{m}$$

$$E(\bar{Y}) = \mu_2 \qquad V(\bar{Y}) = \frac{\sigma_2^2}{n}$$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

5

Mean Difference

- The sample statistic (estimator) for $\mu_1 - \mu_2 = \bar{X}_1 - \bar{Y}_2$
- The standard deviation of $\mu_1 - \mu_2 = \sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
- <u>Proof</u> Since X and Y are independent,

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$$

$$\sigma_{\bar{X}-\bar{Y}}^2 = E[(\bar{X} - \bar{Y})^2] - (E[\bar{X} - \bar{Y}])^2$$

$$= E[\bar{X}^2 - 2\bar{X}\bar{Y} + \bar{Y}^2] - [(E(\bar{X}))^2 - 2E(\bar{X})E(\bar{Y}) + (E(\bar{Y}))^2]$$

$$= E[\bar{X}^2] \cancel{- 2E[\bar{X}]E[\bar{Y}]} + E[\bar{Y}^2] - (E(\bar{X}))^2 \cancel{+ 2E(\bar{X})E(\bar{Y})} - (E(\bar{Y}))^2$$

$$= (E[\bar{X}^2] - (E(\bar{X}))^2) + (E[\bar{Y}^2] - (E(\bar{Y}))^2)$$

$$= V(\bar{X}) + V(\bar{Y})$$

$$= \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n} \Rightarrow \sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$$

$$E(\bar{X}) = \mu_1 \qquad V(\bar{X}) = \frac{\sigma_1^2}{m}$$

$$E(\bar{Y}) = \mu_2 \qquad V(\bar{Y}) = \frac{\sigma_2^2}{n}$$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

6

Population Mean Test : Normal with Known Variances

For any $A \sim N(\mu, \sigma^2)$,
we have $\frac{A-\mu}{\sigma} \sim N(0,1)$

- Thus, $\bar{X} - \bar{Y}$ has normal distribution with
  - mean = $\mu_1 - \mu_2$
  - standard deviation = $\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
- Then, Z = $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$ is standard normal
- Hypothesis
  - Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$

$H_a : \mu_1 - \mu_2 > \Delta_0$

$H_a : \mu_1 - \mu_2 < \Delta_0$

$H_a : \mu_1 - \mu_2 \neq \Delta_0$

If $\Delta_0 = 0$, $\mu_1 = \mu_2$

- Test statistics = Z = $\frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$

Supaporn Erjongmanee
fengspe@ku.ac.th

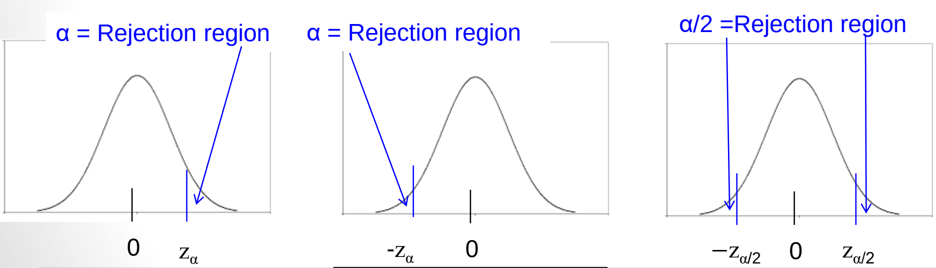
Department of Computer Engineering
Kasetsart University

7

Population Mean Test : Normal with Known Variances (cont.)

Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$

Alternative Hypothesis	Rejection Region at α level	Test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$	**Upper-tailed test**
$H_a : \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$	**Lower-tailed test**
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	Either $z \leq -z_{\alpha/2}$ or $z \geq z_{\alpha/2}$	**Two-tailed test**



Upper-tailed Test **Lower-tailed Test** **Two-tailed Test**

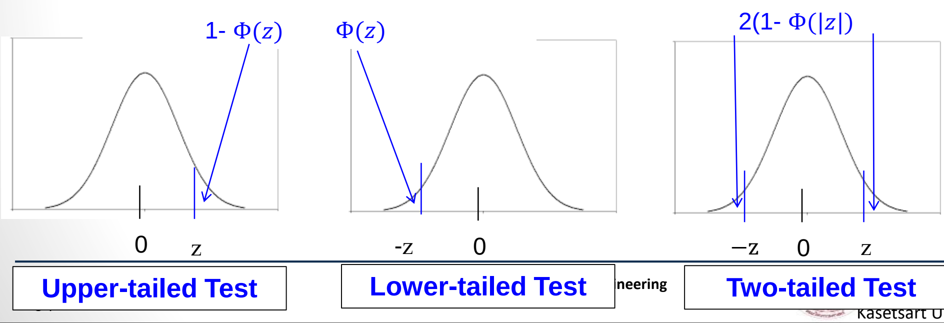
Department of Computer Engineering
Kasetsart University

8

Population Mean Test : Normal with Known Variances (cont.)

Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$

Alternative Hypothesis	P-value	Test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$1-\Phi(z)$	**Upper-tailed test**
$H_a : \mu_1 - \mu_2 < \Delta_0$	$\Phi(z)$	**Lower-tailed test**
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	$2(1-\Phi(\lvert z \rvert))$	**Two-tailed test**



9

Example

- Assume *GPAs for all students are normally distributed* with population standard deviation of GPAs for all students = 0.6
- Two groups of students
  - One group of 10 students who studied less than 10 hours/week

2.80	3.40	4.00	3.60	2.00	3.00	3.47	2.80	2.60	2.00

$\bar{x} = 2.97$

  - Other group of 11 students who studied more than or at least 10 hours/week

3.00	3.00	2.20	4.00	2.96	3.41	3.27	3.80	3.10	2.50	2.40

$\bar{y} = 3.06$

- Using 0.05 significance level, is there difference in average GPAs between these two groups of students?

10

Example (cont.)

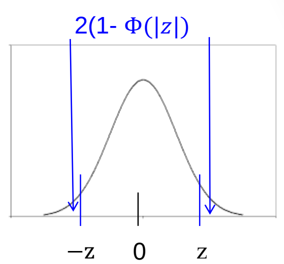
- Our goal is to check difference of average GPAs for these two groups
  - $\mu_1 - \mu_2$ = average GPA difference
  - $\Delta_0 = 0$
  - H_0: $\mu_1 - \mu_2 = 0$
  - H_a: $\mu_1 - \mu_2 \neq 0$
- Compute test statistic

$$Z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} = \frac{2.97-3.06-0}{\sqrt{\frac{0.6^2}{10}+\frac{0.6^2}{11}}} = -0.34$$

Example (cont.)

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

$$Z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} = \frac{2.97-3.06-0}{\sqrt{\frac{0.6^2}{10}+\frac{0.6^2}{11}}} = -0.34$$



Two-tailed Test

- Given $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$
  - Rejection region: $z \geq 1.96$ or $z \leq -1.96$
- Test statistic z falls outside rejection region
  - Null hypothesis is not rejected
  - There is no difference in average GPAs between 2 groups of students

Example (cont.)

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

Alternative Hypothesis	P-value
$H_a : \mu > \mu_0$	$1-\Phi(z)$
$H_a : \mu < \mu_0$	$\Phi(z)$
$H_a : \mu \neq \mu_0$	$2(1-\Phi(\lvert z\rvert))$ or $2(\Phi(-\lvert z\rvert))$

$$Z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}} = \frac{2.97-3.06-0}{\sqrt{\frac{0.6^2}{10}+\frac{0.6^2}{11}}} = -0.34$$

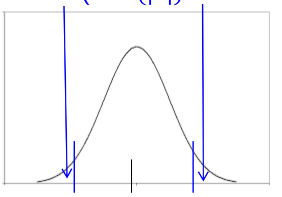
- When z is negative

  P-value = $2 * (\Phi(-0.34))$

  $= 2 * (0.3669)$ = 0.7338
- At $\alpha = 0.05 <$ p-value = 0.7338
  - Test statistic falls outside rejection region
  - We do not reject null hypothesis
  - No difference between two groups of students

Two-tailed Test

$2(1-\Phi(\lvert z\rvert))$



Or using formula from the table.
For two-tailed test:
P-value $=2 * (1-\Phi(\lvert z\rvert))$
$=2 * (1-\Phi(\lvert -0.34\rvert))$
= 2*(1-0.6331)=0.7338

Outline

- Population Mean Test
  - Normal and Known variance
  - Large sample size
  - Normal and Small sample size

Population Mean Test : Large Samples

- When sample sizes are large, CLT states that
  - $\bar{X} - \bar{Y}$ has normal distribution
  - S_1 and S_2 are close to σ_1 and σ_2 respectively
- Therefore, Z = $\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}}$ is approximately standard normal
- <u>Follow test like normal with known variance, but use S_1 and S_2 instead of σ_1 and σ_2</u>
- Test statistics = Z = $\frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}}$

Example

- Two groups of students with different teaching styles
  - One group of 79 students with traditional style
  - Other group of 85 students with experimental style (allow students more involved: more homework, more quizzes)
- Statistics of both groups' scores are :
  - 79 students: $\bar{X} = 23.87$, S_1 = 11.60
  - 85 Students: $\bar{Y} = 27.34$, S_2 = 8.85
- Using 0.05 significance level, is there any suggestion the new style improves more than the traditional?

Let group 1 = group with traditional style (79 students)
group 2 = group with experimental style (85 students)

Example (cont.)

group 1 = group with traditional style (79 students)
group 2 = group with experimental style (85 students)

- Our goal is to check <u>the new style improves more than the traditional</u>?
  - $\mu_1 - \mu_2$ = average test scores
  - $\Delta_0 = 0$
  - $H_0: \mu_1 - \mu_2 = 0$
  - $H_a: \mu_1 - \mu_2 < 0$

Second group get better scores

- Sample statistics
  - $\bar{X} = 23.87$, $S_1 = 11.60$, $\bar{Y} = 27.34$, $S_2 = 8.85$
  - m = 79, n = 85
- Compute test statistic

$$Z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} = \frac{23.87-27.34-0}{\sqrt{\frac{11.60^2}{79}+\frac{8.85^2}{85}}} = \frac{-3.47}{1.620} = -2.14$$

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Example (cont.)

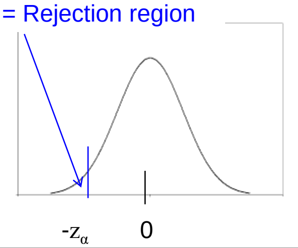
$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 < 0$

Lower-tailed Test

- Compute test statistic

$$Z = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}} = \frac{23.87-27.34-0}{\sqrt{\frac{11.60^2}{79}+\frac{8.85^2}{85}}} = -2.14$$

- Given $\alpha = 0.05$:
  - $-z_\alpha = -z_{0.05} = -1.645$
  - Rejection region: $z \le -1.645$
- Test statistic z falls inside rejection region
  - Null hypothesis is rejected
  - Test scores were improved with new teaching style



P-value = $\Phi(-2.14)$ =

$\alpha = 0.05 >$ p-value = 0.0162

Reject null hypothesis

Supaporn Erjongmanee
fengspe@ku.ac.th

Department of Computer Engineering
Kasetsart University

Outline

- Population Mean Test
  - Normal and Known variance
  - Large sample size
  - Normal and Small sample size

Population Mean Test: Normal and Small Samples

- $X_1, X_2, \ldots, X_m$ are m random samples from normal distribution
- $Y_1, Y_2, \ldots, Y_n$ are n random samples from normal distribution
- Variable $T = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$ has approximately t-distribution

  with degree of freedom = $v^* = \frac{\left(\frac{S_1^2}{m}+\frac{S_2^2}{n}\right)^2}{\frac{(S_1^2/m)2}{m-1}+\frac{(S_2^2/n)2}{n-1}}$

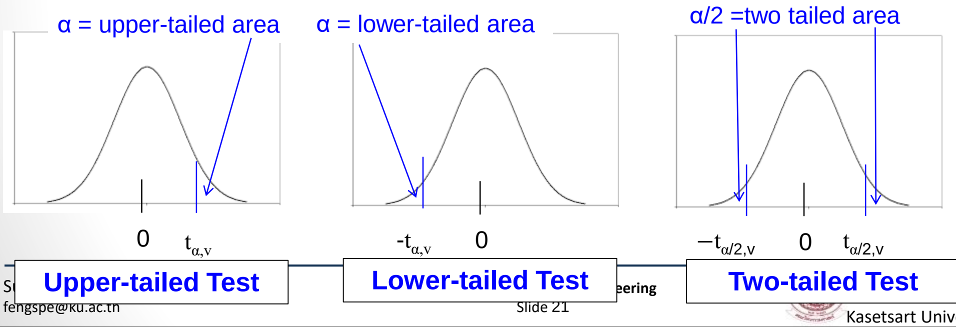
- Test statistic: $t = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}}$

* Round v down in nearest integer. Proof of this requires lots of details

Population Mean Test : Normal with Small Samples (cont.)

Null hypothesis: $\mu_1 - \mu_2 = \Delta_0$

Alternative Hypothesis	Rejection Region at α level	Test
$H_a : \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha, v}$	**Upper-tailed test**
$H_a : \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha, v}$	**Lower-tailed test**
$H_a : \mu_1 - \mu_2 \neq \Delta_0$	Either $t \leq -t_{\alpha/2, v}$ or $t \geq t_{\alpha/2, v}$	**Two-tailed test**



21

Example

- How to pour champagne: traditionally vertical or tilted to preserve gas bubbles?
- Assume CO_2 is dissolved with normal distribution
- Measure average dissolved CO_2 loss

	n	Sample Mean (g/L)	S
Traditional	4	4.0	0.5
Tilted	4	3.7	0.3

- Compute 0.01 significance level

22

Example (cont.)

- How to pour champagne: traditionally vertical or tilted to preserve gas bubbles?
- Solution.
  - Set up hypothesis
    - $\mu_1 - \mu_2$ = difference of gas bubbles
    - $\Delta_0 = 0$
    - H_0: $\mu_1 - \mu_2 = 0$
    - H_a: $\mu_1 - \mu_2 < 0$
  - Compute test statistic

Tilted preserves more gas bubbles

	n	Sample Mean (g/L)	S
Traditional	4	4.0	0.5
Tilted	4	3.7	0.3

$$t = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}} = \frac{4.0-3.7-0}{\sqrt{\frac{0.5^2}{4}+\frac{0.3^2}{4}}} = \frac{0.30}{0.29} = 1.03$$

23

Example (cont.)

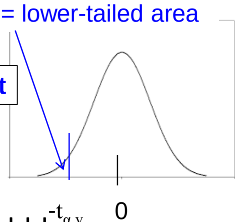
- Compute test statistic

$$t = \frac{\bar{x}-\bar{y}-\Delta_0}{\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}} = \frac{4.0-3.7-0}{\sqrt{\frac{0.5^2}{4}+\frac{0.3^2}{4}}} = 1.03$$

- Find rejection region
  - Find degree of freedom v

$$v = \frac{\left(\frac{S_1^2}{m}+\frac{S_2^2}{n}\right)^2}{\frac{(S_1^2/m)2}{m-1}+\frac{(S_2^2/n)2}{n-1}} = \frac{\left(\frac{0.5^2}{4}+\frac{0.3^2}{4}\right)^2}{\frac{(0.5^2/4)^2}{3}+\frac{(0.3^2/4)^2}{3}} = \frac{0.0072}{0.00147} = 4.91 \sim 4$$

α = lower-tailed area



Lower-tailed Test

  - At $\alpha = 0.01$, $-t_{0.01,4} = -3.747$
  - Rejection region: $t \le -3.747$
- Test statistic falls outside rejection region
  - We do not reject null hypothesis
  - Either traditional vertical or tilted pouring preserves same bubbles

How about using p-value?

24

Example (cont.)

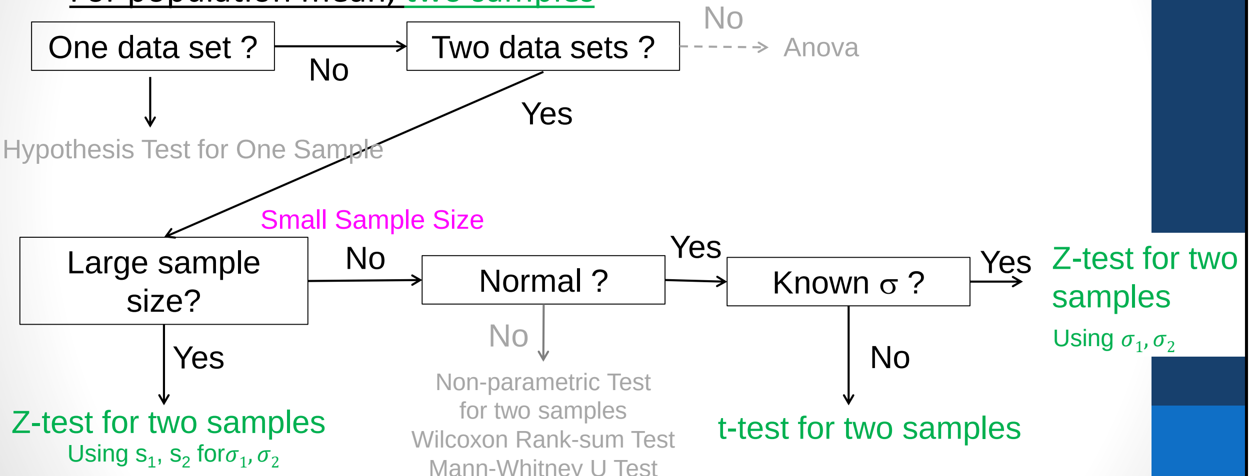
- Compute test statistic

$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{4.0 - 3.7 - 0}{\sqrt{\frac{0.5^2}{4} + \frac{0.3^2}{4}}} = 1.03$$

- P-value = (Left tailed area of t = 1.03) = 0.8194
- At α = 0.01 < p-value = 0.8194
  - Test statistic falls outside rejection region
  - We do not reject null hypothesis
  - Either traditional vertical or tilted pouring preserves same bubbles

Which Test to Choose

- How to choose test to fit data
- For population mean, two samples



References

1. J.L. Devore and K.N.Berk, Modern Mathematical Statistics with Applications, Springer, 2012.