ACE 261
Fall 2002
Prof. Katchova

Lecture 10

Statistical Inference About Means and Proportions with Two Populations

Comparisons Involving Means: Outline

- Interval Estimation and Hypothesis Testing of Differences in Means
  - For independent samples
  - For matched samples
- Inferences about the Difference between the Proportions

What problems are we doing to solve?

- Interval estimation
  - Lecture 8: What is the interval estimate mean height of people?
  - This lecture: What is the interval estimate of the difference of mean heights of men and women?
- Hypothesis testing
  - Lecture 9: Is the average height of people 5'7"?
  - This lecture: Is the average height of men and women different?

Interval Estimation

- Lecture 8: Interval estimation for a mean

$$\mu = \bar{x} \pm z_{\alpha/2} s_{\bar{x}}$$

- This lecture: Interval estimation for differences in means

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2} s_{\bar{x}_1 - \bar{x}_2}$$

Hypothesis Testing

- Lecture 8: Hypothesis Testing for Means

$$z = \bar{x} - \mu_0$$

- This lecture: Hypothesis Testing for Differences between Means

$$z = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$$

Sampling Distribution of $x_1^{bar} - x_2^{bar}$

Expected value of $x_1^{bar} - x_2^{bar}$

$$E(\bar{x}_1 - \bar{x}_2)$$

Standard deviation of $x_1^{bar} - x_2^{bar}$

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_1^2 + s_2^2}$$

Interval Estimate of μ_1 - μ_2: Large-Sample Case ($n_1 \geq 30$ and $n_2 \geq 30$)

- Interval Estimate with σ_1 and σ_2 Known

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}}$$

- Interval Estimate with σ_1 and σ_2 Unknown

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}s_{\bar{x}_1-\bar{x}_2}}$$

(1 - α) is the confidence coefficient

Hypothesis Tests About the Difference Between the Means of Two Populations: Independent Samples

- Hypotheses:

H_0: $\mu_1 - \mu_2 \leq 0$	H_0: $\mu_1 - \mu_2 \geq 0$	H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 > 0$	H_a: $\mu_1 - \mu_2 < 0$	H_a: $\mu_1 - \mu_2 \neq 0$

- Test Statistics

Large-Sample

$$\boxed{z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}}$$

Small-Sample

$$\boxed{t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2(1/n_1 + 1/n_2)}}}$$

- Reject the null hypothesis if the test statistic is in the rejection region.

Interval Estimation of the Difference between Means

- Question: What is the interval estimate of the difference between the means of these two populations?

	Sample #1	Sample #2
Sample Size	n_1 = 120 balls	n_2 = 80 balls
Mean	x_1^{bar} = 235 yards	x_2^{bar} = 218 yards
Standard Dev.	σ_1 = 15 yards	σ_2 = 20 yards

Interval Estimation of the Difference between Means

- Solution:
- The point estimate of the difference is

$x_1^{bar} - x_2^{bar} =$

- The interval estimate of the difference is

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 17 \pm 1.96\sqrt{\frac{(15)^2}{120} + \frac{(20)^2}{80}}$$

= 17 $\pm$ 5.14 or 11.86 yards to 22.14 yards.

We are 95% confident that the difference between the means of these two populations is in the interval of 11.86 to 22.14 yards.

Hypothesis Tests About the Difference Between the Means of Two Populations: Large-Sample Case

- Question: Can we conclude, using a .01 level of significance, that the mean driving distance for the first company's balls is greater than the mean driving distance of the second company's balls?
- Hypothesis

H_0: $\mu_1 - \mu_2 \leq 0$

H_a: $\mu_1 - \mu_2 > 0$

Hypothesis Tests About the Difference Between the Means of Two Populations: Large-Sample Case

- Reject H_0 if $z > z_{0.01} = 2.33$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(235 - 218) - 0}{\sqrt{\frac{(15)^2}{120} + \frac{(20)^2}{80}}} = \frac{17}{2.62} = 6.49$$

- Reject H_0. We are at least 99% confident that the mean driving distance of the first company's golf balls is greater than the mean driving distance of the second company's golf balls.

Interval Estimate of μ_1 - μ_2: Small-Sample Case (n_1 < 30 and/or n_2 < 30)

- If σ^2 is known, we use the same method as the large sample method (i.e. the standard normal distribution)

$$\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sigma_{\bar{x}_1-\bar{x}_2}$$

where:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Interval Estimate of μ_1 - μ_2: Small-Sample Case (n_1 < 30 and/or n_2 < 30)

- If σ^2 is unknown, then we estimate s^2 and use the t-distribution.

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}s_{\bar{x}_1-\bar{x}_2}$$

where:

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Note that s^2 is the weighted average of the two sample variances s_1^2 and s_2^2 with weights (n_1-1) and (n_2-1)

Assumptions made in the small sample case

- Both populations have normal distributions.
- The variances of the population are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$)
- If sample sizes are equal ($n_1 = n_2$), then results are acceptable even if variances are not equal.

Interval Estimation of the Differences between Means: Small Sample Case

- Question: Two types of cars are being tested to compare miles-per-gallon (mpg) performance. What is the interval estimate of the population difference?

	Sample #1 Ford	Sample #2 Nissan
Sample Size	n_1 = 12 cars	n_2 = 8 cars
Mean	x_1^{bar} = 29.8 mpg	x_2^{bar} = 27.3 mpg
Standard Deviation	s_1 = 2.56 mpg	s_2 = 1.81 mpg

Interval Estimation of the Differences between Means: Small Sample Case

- Point estimate of μ_1 - μ_2 = $x_1^{bar} - x_2^{bar}$ = 29.8 - 27.3 = 2.5 mpg
- Since it's small sample case, use the t distribution with $n_1 + n_2$ - 2 = 18 degrees of freedom and find that $t_{.025}$ = 2.101.
- Estimate s^2 as the weighted average of two sample variances

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{11(2.56)^2+7(1.81)^2}{12+8-2} = 5.28$$

Interval Estimation of the Differences between Means: Small Sample Case

- Substitute results in the formula for the interval estimate

$$\bar{x}_1 - \bar{x}_2 \pm t_{.025}\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 2.5 \pm 2.101\sqrt{5.28\left(\frac{1}{12}+\frac{1}{8}\right)}$$

= 2.5 $\pm$ 2.2 or .3 to 4.7 miles per gallon.

We are 95% confident that the difference between the mean mpg ratings of the two car types is from .3 to 4.7 mpg.

Hypothesis Tests About the Difference Between the Means of Two Populations: Small -Sample Case

- Question: Can we conclude, using a .05 level of significance, that the miles-per-gallon (*mpg*) performance for Ford cars is greater than the miles-per-gallon performance for Nissan cars?

μ_1 = mean *mpg* for the population of Ford cars
μ_2 = mean *mpg* for the population of Nissan cars

$$H_0\text{: } \mu_1 - \mu_2 \le 0$$
$$H_a\text{: } \mu_1 - \mu_2 > 0$$

Hypothesis Tests About the Difference Between the Means of Two Populations: Small -Sample Case

- Reject H_0 if $t > 1.734$ ($\alpha = .05$, d.f. = 18)
- Test statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s^2(1/n_1 + 1/n_2)}}$$

where:

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

Inference About the Difference Between the Means of Two Populations: Matched Samples

- With a matched-sample design each sampled item provides a pair of data values.
- This design often leads to a smaller sampling error than the independent-sample design because variation between sampled items is eliminated as a source of sampling error.
- We consider only the differences for each pair d^{bar} and the analysis is the same as in chapter 9, when d^{bar} replaces x^{bar} in all formulas.

Matched Sample Example

	Delivery Time (Hours)		
District Office	UPX	INTEX	Difference
Seattle	32	25	7
Los Angeles	30	24	6
Boston	19	15	4
Cleveland	16	15	1
New York	15	13	2
Houston	18	15	3
Atlanta	14	15	-1
St. Louis	10	8	2
Milwaukee	7	9	-2
Denver	16	11	5

Matched Sample Example

- Do the data indicate a difference in mean delivery times for the two services, at the 5% significance level?
- Let μ_d = the mean of the difference values for the two delivery services for the population of district offices
  - Hypothesis H_0: $\mu_d = 0$, H_a: $\mu_d \ne 0$
  - Rejection rule: Assuming the population of difference values is approximately normally distributed, the *t* distribution with *n* - 1 degrees of freedom applies.

    With $\alpha = .05$, $t_{.025} = 2.262$ (9 degrees of freedom).

    Reject H_0 if $t < -2.262$ or if $t > 2.262$

Matched Sample Example

$$\bar{d} = \frac{\sum d_i}{n} = \frac{(7+6+\ldots+5)}{10} = 2.7$$

$$s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{76.1}{9}} = 2.9$$

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} = \frac{2.7 - 0}{2.9/\sqrt{10}} = 2.94$$

- Conclusion: reject H_0.

  There is a significant difference between the mean delivery times for the two services.

Proportions: Sampling Distribution of $p_1^{bar} - p_2^{bar}$

- Expected Value

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

- Standard Deviation

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

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

25

Interval Estimation of $p_1 - p_2$

- Distribution Form
  If the sample sizes are large (n_1p_1, $n_1(1 - p_1)$, n_2p_2, and $n_2(1 - p_2)$ are all greater than or equal to 5), the sampling distribution of $p_1^{bar} - p_2^{bar}$ can be approximated by a normal probability distribution.
- The interval estimate is

$$\bar{p}_1 - \bar{p}_2 \pm z_{\alpha/2}\sigma_{\bar{p}_1-\bar{p}_2}$$

26

Example

- Before an advertising campaign, 60 of the 150 households surveyed said that they will buy a new product. After the advertising campaign, 120 of 250 households said that they will buy the product.
- Do the data support the position that the advertising campaign increased customers interest in buying the product?

H_0: $p_1 - p_2 \le 0$

H_a: $p_1 - p_2 > 0$

Where sample 1 is after the campaign and sample 2 is before the campaign.

27

Hypothesis Tests about $p_1 - p_2$

- Test statistic

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sigma_{\bar{p}_1-\bar{p}_2}}$$

where:

$$s_{\bar{p}_1-\bar{p}_2} = \sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}$$

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

- Reject H_0 if $z > 1.645$

28

Hypothesis Testing Calculations

$p_1 - p_2 = \bar{p}_1 - \bar{p}_2 = 120 - 60 = .48 - .40 = .08$

$\bar{p} = 250(.48) + 150(.40) = 180 = .45$

$s_{\bar{p}_1-\bar{p}_2} = \sqrt{.45(.55)(1/ + 1/)} = .0514$

$z = (.48 - .40) - 0 = .08 = 1.56$

- Since z =1.56<1.645, we do not reject H_0.

29

Interval Estimate for Proportions Differences

- Interval estimate for $\alpha = .05$, $z_{.025} = 1.96$:

$.48 - .40 + 1.96\sqrt{.48(.52) + .40(.60)}$

$.08 \pm 1.96(.0510)$

$.08 \pm .10$

or -.02 to +.18

- At a 95% confidence level, the interval estimate of the difference between the proportion of households aware of the client's product before and after the new advertising campaign is -.02 to +.18.

30