Inference about Two Proportions (Sec. 9.4)

We now consider inference about $p_1 - p_2$, the difference between two population proportions.

Point estimate for $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$

For large samples, this statistic has an approximately normal distribution with mean $p_1 - p_2$ and standard deviation $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.

So a $(1 - \alpha)100\%$ CI for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$\hat{p}_1$ **= sample proportion for Sample 1**

$\hat{p}_2$ **= sample proportion for Sample 2**

n_1 **= sample size of Sample 1**

n_2 **= sample size of Sample 2**

Requires large samples:

(1) Need $n_1 \geq 20$ and $n_2 \geq 20$.

(2) Need number of "successes" and number of "failures" to be 5 or more in both samples.

$\Longleftrightarrow$ (2) Need: $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$,

$n_2\hat{p}_2 \geq 5$, $n_2(1-\hat{p}_2) \geq 5$

Test of H_0: $p_1 = p_2$ $\longleftrightarrow$ $H_0: p_1 - p_2 = 0$

Test statistic: $$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(Use pooled proportion because under H_0, p_1 and p_2 are the same.)

Pooled sample proportion

$$\hat{p} = \frac{\text{\# successes in both samples combined}}{\text{\# observations in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

Example: Let p_1 = the proportion of male USC students who park on campus and let p_2 = the proportion of female students who park on campus. Find a 95% CI for the difference in the true proportion of males and the true proportion of females who park at USC.

Popn 1 = males   Popn 2 = females

Take a random sample of 50 males; 32 park at USC.
Take a random sample of 60 females; 34 park at USC.

$$\hat{p}_1 = \frac{32}{50} = .64, \quad \hat{p}_2 = \frac{34}{60} = .567$$

95% CI for $p_1 - p_2$: $1 - \alpha = .95 \Rightarrow \alpha = .05$

$\alpha/2 = .025 \Rightarrow z_{.025} = 1.96$

$$(.64 - .567) \pm 1.96\sqrt{\frac{(.64)(.36)}{50} + \frac{(.567)(.433)}{60}}$$

$\Rightarrow .073 \pm .1828$

$\Rightarrow (-.110, .256)$

Interpretation: We are 95% confident that the proportion of males who park at USC is between .110 lower and .256 higher than the proportion of females who park at USC.

Hypothesis Test: Is the proportion of males who park greater than the proportion of females who park? Use $\alpha = .10$

$H_0: p_1 = p_2$ vs. $H_a: p_1 > p_2$

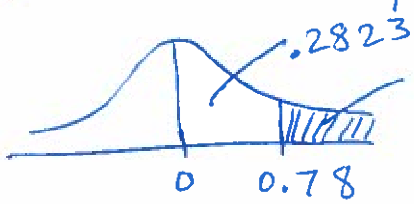
Reject H_0 if $z > z_{.10} = 1.282$

$\hat{p}_1 = \frac{32}{50} = .64$, $\hat{p}_2 = \frac{34}{60} = .567$, $\hat{p} = \frac{32+34}{50+60} = \frac{66}{110}$

$= \boxed{0.6}$

$$z = \frac{.64 - .567}{\sqrt{(.6)(.4)\left(\frac{1}{50} + \frac{1}{60}\right)}} = \frac{.073}{.0938} = 0.78$$

Since $0.78 \not> 1.282$, we fail to reject H_0. At $\alpha = .10$, we cannot conclude the population proportion of males who park at USC is greater than the population proportion of females who park at USC.



p-value $= .5 - .2823$

$\boxed{\text{p-value} = .2177}$

STAT 515 -- Chapter 10: Analysis of Variance

Designed Experiment – A study in which the researcher controls the levels of one or more variables to determine their effect on the variable of interest (called the **response variable** or **dependent variable**).

Response variable: Main variable of interest (continuous)
Factors: Other variables (typically discrete) which may have an effect on the response.

- Quantitative factors are numerical.
- Qualitative factors are categorical.

The **levels** are the different values (for each factor) used in the experiment.

Example 1:
Response variable: College GPA
Factors: Gender (levels: Male, Female)
# of AP courses (levels: 0, 1, 2, 3, 4+)

The **treatments** of an experiment are the different factor level combinations.

Treatments for Example 1: {M, 0} {F, 0}
{M, 1} {F, 1}
{M, 2} {F, 2}
{M, 3} {F, 3}
{M, 4+} {F, 4+}

Experimental Units: the objects on which the factors and response are observed or measured.
Example 1? Students

Designed experiment: The analyst controls which treatments to use and assigns experimental units to each treatment.

Observational study: The analyst simply observes treatments and responses for a sample of units. (like Example 1)

Example 2: Plant growth study:
Experimental Units: A sample of plants
Response: Growth over one month
Factors: Fertilizer Brand (levels: A, B, C)
Environment
(levels: Natural Sunlight, Artificial Lamp)
both Qualitative

There are how many treatments? $6 = 3 \times 2$

{A, NS}, {B, NS}, {C, NS}
{A, AL}, {B, AL}, {C, AL}

(Could also have a quantitative factor…)

Amount of Water

If 5 plants are assigned to each treatment (5 replicates per treatment), there are how many observations in all?

30 observations overall

Completely Randomized Design (CRD)

A Completely Randomized Design is a design in which independent samples of experimental units are selected for each treatment.

Suppose there are k treatments (usually $k \geq 3$).

We want to test for any differences in mean response among the treatments.

Hypothesis Test:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_a: At least two of the treatment population means differ.

Visually, we could compare all the sample means for the different treatments. (Dot plots, p. 512)

If there are more than two treatments, we cannot just subtract sample mean values.

Instead, we analyze the variance in the data:

Q: Is the variance within each group small compared to the variance between groups (specifically, between group means)?

Top figure? No – variance within groups is large

Bottom figure? Yes

How do we measure the variance within each group and the variance between groups?

The Sum of Squares for Treatments (SST) measures variation <u>between</u> group means.

$$\text{SST} = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2$$

n_i **= number of observations in group** i

$\overline{X}_i$ **= sample mean response for group** i

$\overline{X}$ **= <u>overall</u> sample mean response**

SST measures how much each group sample mean varies from the overall sample mean.

The Sum of Squares for Error (SSE) measures <u>variation within groups</u>.

$$\text{SSE} = \sum_{i=1}^{k} (n_i - 1)s_i^2$$

s_i^2 **= sample variance for group** i

SSE is a sum of the variances <u>of each group</u>, weighted by the sample sizes by each group.

To make these measures comparable, we divide by their degrees of freedom and obtain:

Mean Square for Treatments (MST) = $\dfrac{SST}{k-1}$

Mean Square for Error (MSE) = $\dfrac{SSE}{n-k}$

total # of observations in whole study

The ratio $\dfrac{MST}{MSE}$ **is called the ANOVA F-statistic.**

If $F = \dfrac{MST}{MSE}$ **is much bigger than 1, then the variation between groups is much bigger than the variation within groups, and we would reject** H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ **in favor of** H_a.

Example (Table 10.3)
<u>Response</u>: Distance a golf ball travels
<u>4 treatments</u>: Four different brands of ball

$\bar{X}_1 = 250.8$, $\bar{X}_2 = 261.1$, $\bar{X}_3 = 270.0$, $\bar{X}_4 = 249.3$.

=> $\bar{X} = 257.8$.

$n_1 = 10$, $n_2 = 10$, $n_3 = 10$, $n_4 = 10$. => $n = 40$.
Sample variances for each group:
$s_1^2 = 22.42$, $s_2^2 = 14.95$, $s_3^2 = 20.26$, $s_4^2 = 27.07$.