STAT 515 -- Chapter 9: Two-Sample Problems

Paired Differences (Section 9.3)

Examples of Paired Differences studies:

• Similar subjects are paired off and one of two treatments is given to each subject in the pair.

or

• We could have two observations on the same subject.

The key: With paired data, the pairings cannot be switched around without affecting the analysis.

We typically wish to perform inference about the mean of the differences, denoted μ_D.

Example 1: Six students are given two tests, one after being fed, and one on an empty stomach. Is there evidence that students perform better on a full stomach? (Assume normality of data, and use $\alpha = .05$.)

	Student					
Scores	1	2	3	4	5	6
X_1 (with food)	74	71	82	77	72	81
X_2 (without food)	68	71	86	70	67	80

Calculate differences: D = $X_1 - X_2$

D:

Example 2: Find a 98% CI for the mean difference in arm strength for right-handed people (measured by the number of seconds a certain weight can be held extended).

	Person						
	1	**2**	**3**	**4**	**5**	**6**	**7**
X_1 (Right)	**26**	**35**	**17**	**47**	**22**	**16**	**32**
X_2 (Left)	**20**	**31**	**10**	**38**	**23**	**16**	**29**
D:							

Interpretation: With 98% confidence, the mean right-arm strength is between 0.336 seconds <u>less</u> and 8.336 seconds <u>greater</u> than the mean left-arm strength. (We are 98% confident the mean difference is between -0.336 and 8.336 seconds.)

Note: With paired data, the two-sample problem really reduces to a one-sample problem on <u>the sample of differences</u>.

Two Independent Samples (Section 9.2)

Sometimes there's no natural pairing between samples.

Example 1: Collect sample of males and sample of females and ask their opinions on whether capital punishment should be legal.

Example 2: Collect sample of iron pans and sample of copper pans and measure their resiliency at high temperatures.

No attempt made to pair subjects – we have two independent samples.

We could rearrange the order of the data and it wouldn't affect the analysis at all.

Comparing Two Means

Our goal is to compare the mean responses to two treatments, or to compare two population means (we have two separate samples).

We assume both populations are normally distributed (or "nearly" normal).

We're typically interested in the difference between the mean of population 1 (μ_1) and the mean of population 2 (μ_2).

We may construct a CI for $\mu_1 - \mu_2$ or perform one of three types of hypothesis test:

H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$	H_a: $\mu_1 < \mu_2$	H_a: $\mu_1 > \mu_2$

Note: H_0 could be written H_0: $\mu_1 - \mu_2 = 0$.

The parameter of interest is

Notation:

$\bar{X}_1$ **= mean of Sample 1**

$\bar{X}_2$ **= mean of Sample 2**
σ_1 **= standard deviation of Population 1**
σ_2 **= standard deviation of Population 2**
s_1 **= standard deviation of Sample 1**

s_2 = standard deviation of Sample 2
n_1 = size of Sample 1
n_2 = size of Sample 2

The point estimate of $\mu_1 - \mu_2$ is

This statistic has standard error

but we use since σ_1, σ_2 unknown.

Since the data are normal, we can use the t-procedures for inference.

<u>Case I</u>: Unequal population variances ($\sigma_1^2 \neq \sigma_2^2$)

In the case where the two populations have <u>different variances</u>, the t-procedures are <u>only approximate</u>.

Formula for $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

where the d.f. = the smaller of $n_1 - 1$ and $n_2 - 1$.

To test H_0: $\mu_1 = \mu_2$, the test statistic is:

<u>H_a</u>	<u>Rejection region</u>	<u>P-value</u>
$\mu_1 \neq \mu_2$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	2*(tail area)
$\mu_1 < \mu_2$	$t < -t_{\alpha}$	left tail area
$\mu_1 > \mu_2$	$t > t_{\alpha}$	right tail area

where the d.f. = the smaller of $n_1 - 1$ and $n_2 - 1$.

<u>Case II</u>: Equal population variances ($\sigma_1^2 = \sigma_2^2$)

In the case where the two populations have equal variances, we can better estimate this population variance with the <u>pooled</u> sample variance:

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

Our t-procedures in this case are exact, not approximate.

Formula for $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is:

where the d.f. = $n_1 + n_2 - 2$.

To test H_0: $\mu_1 = \mu_2$, the test statistic is:

H_a	Rejection region	P-value
$\mu_1 \neq \mu_2$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	2*(tail area)
$\mu_1 < \mu_2$	$t < -t_{\alpha}$	left tail area
$\mu_1 > \mu_2$	$t > t_{\alpha}$	right tail area

where the d.f. = $n_1 + n_2 - 2$.

Example: What is the difference in mean DVD prices at Best Buy and Walmart?

Let μ_1 = mean DVD price at Best Buy and let μ_2 = mean DVD price at Walmart.

Find 99% CI for $\mu_1 - \mu_2$.

Randomly sample 28 DVDs from Best Buy:

$\bar{X}_1 = 17.93$, $s_1 = 10.22$, $s_1^2 = 104.45$, $n_1 = 28$.

Randomly sample 20 DVDs from Walmart:

$\bar{X}_2 = 25.70$, $s_2 = 11.35$, $s_2^2 = 128.82$, $n_2 = 20$.

Does $\sigma_1^2 = \sigma_2^2$? Could test this formally using an F-test (Sec. 9.5) or could simply compare spreads of box plots for samples 1 and 2.

When in doubt, assume $\sigma_1^2 \neq \sigma_2^2$. Let's assume $\sigma_1^2 \neq \sigma_2^2$ here.

99% CI for $\mu_1 - \mu_2$:

Interpretation: We are 99% confident that Best Buy's mean DVD price is between \$16.89 lower and \$1.35 higher than Walmart's mean DVD price.

Test: $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 < \mu_2$ (at $\alpha = .10$)

Test statistic:

<u>Inference about Two Proportions</u> (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

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

So a (1 – α)100% CI for $p_1 - p_2$ is

$\hat{p}_1$ **= 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" <u>and</u> number of "failures" to be 5 or more in <u>both</u> samples.**

Test of H_0: $p_1 = p_2$

Test statistic:

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

Pooled sample proportion

$\hat{p}$ =

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.

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

Interpretation: We are 95% confident that the proportion of males who park at USC is between .110 <u>lower</u> and .256 <u>higher</u> 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?