Two sample tests – Cases IV, V

D. Case IV: Matched Pairs, σ unknown.

1. Sometimes it happens that subjects are actually sampled in pairs - and the scores of the members of the pair are not necessarily independent.

✓ You might be interested in comparing scores of husbands and wives. Knowing the husband's score probably gives us some information about the wife's score and vice versa. Scores are not independent – on some variables a husband and wife are more likely to be similar to each other than are two randomly selected individuals.

✓ Sometimes matching is done by the experimenter. We pick individuals who are identical (or nearly so) and assign them to treatment groups; hence the variables used for matching are unlikely to be responsible for any observed differences. For example, Richard LaManna of Notre Dame has done studies where blacks and whites of comparable education, occupation and income each went out to rent an apartment; any difference in the way they were treated was assumed to be due to race.

✓ Sometimes people are matched with themselves in a "before and after" comparison. For example, we might compare a person's health before and after receiving some drug. Or, we might see how attitudes change across time.

2. Let $D = X_1 - X_2$, where X_1 is the score of the first member of the pair and X_2 is the score of the second member. Assume X_1 and X2 are normally distributed. Recall that $E(X + Y) = E(X) + E(Y)$ (Expectations Rule 8) and $V(X \pm Y) = V(X) + V(Y) \pm 2\,COV(X,Y) = \sigma^2_{X \pm Y}$ (Expectations Rule 15). Then,

$$E(D) = \mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2,$$

$$V(D) = \sigma_D^2 = V(X_1 - X_2) = V(X_1) + V(X_2) - 2COV(X_1, X_2) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12},$$

$$SD(D) = \sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

Note that, because X_1 and X_2 are not independent, the covariance is <u>not</u> 0. Now, let

$$\overline{D} = \frac{\sum_{i=1}^{N}(X_{1_i} - X_{2_i})}{N} = \frac{\sum_{i=1}^{N} D_i}{N}$$

i.e. $\overline{D}$ equals the average difference.

Also,

$$E(\overline{D}) = \frac{\sum_{i=1}^{N} E(D_i)}{N} = \frac{\sum_{i=1}^{N} \mu_D}{N} = \frac{N * \mu_D}{N} = \mu_D = \mu_{\overline{D}}$$

and (because V[aX] = a^2V[X]),

$$V(\overline{D}) = V\left(\frac{\sum_{i=1}^{N} D_i}{N}\right) = \frac{\sum_{i=1}^{N} V(D_i)}{N^2} = \frac{N\sigma_D^2}{N^2} = \frac{\sigma_D^2}{N} = \frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{N} = \sigma_{\overline{D}}^2$$

3. True Standard error of $\overline{D}$: If we happen to know the population variances and covariances, we can easily compute

$$SD(\overline{D}) = \sqrt{\frac{\sigma_D^2}{N}} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}{N}} = \sigma_{\overline{D}}$$

4. Estimated standard error of $\overline{D}$. More commonly, the population variances and covariances will not be known and will have to be estimated. The sample variance and standard deviation of D are

$$s_D^2 = \frac{\sum_{i=1}^{N}(D_i - \overline{D})^2}{N-1} = \frac{\sum_{i=1}^{N} D_i^2 - N\overline{D}^2}{N-1} = s_1^2 + s_2^2 - 2s_{12}$$

$$s_D = \sqrt{\frac{\sum_{i=1}^{N}(D_i - \overline{D})^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} D_i^2 - N\overline{D}^2}{N-1}} = \sqrt{s_1^2 + s_2^2 - 2s_{12}}$$

Hence, the estimated standard error of $\overline{D}$ is

$$s_{\overline{D}} = \frac{s_D}{\sqrt{N}} = \sqrt{\frac{s_D^2}{N}}$$

Incidentally, note that the above formulas suggest two strategies for finding the sample variance of D. You could first compute D, and then compute its variance. Or, you could compute the sample variances and covariances of X1 and X2. We'll discuss the latter option more in the next handout.

5. Test statistic for matched pairs (assuming σ's are unknown):

$$t = \frac{\bar{d} - \mu_{D_0}}{s_{\bar{D}}} = \frac{\bar{d} - \mu_{D_0}}{\frac{s_D}{\sqrt{N}}} = \frac{\bar{d} - \mu_{D_0}}{\sqrt{\frac{s_D^2}{N}}}$$

When the null hypothesis is true, the test statistic will have a T distribution with N - 1 degrees of freedom.

6. Confidence interval:

$$\bar{d} \pm (t_{\alpha/2,v} * s_D / \sqrt{N}), i.e.$$

$$\bar{d} - (t_{\alpha/2,v} * s_D / \sqrt{N}) \le \mu_D \le \bar{d} + (t_{\alpha/2,v} * s_D / \sqrt{N})$$

7. 2-tailed acceptance region:

$$- t_{\alpha/2,v} \le t \le t_{\alpha/2,v} ; equivalently,$$

$$\mu_{D_0} \pm (t_{\alpha/2,v} * s_D / \sqrt{N}), i.e.$$

$$\mu_{D_0} - (t_{\alpha/2,v} * s_D / \sqrt{N}) \le \bar{d} \le \mu_{D_0} + (t_{\alpha/2,v} * s_D / \sqrt{N})$$

8. COMMENT: Note the strong similarities between 2 sample tests, case IV, and single sample tests, Case III, sigma unknown. Once you construct the variable $D = X_1 - X_2$, or are given information about the variance and mean of D, you can proceed exactly as you would in a single sample case. <u>Indeed, some authors treat the matched pairs problem as a single sample case rather than two sample.</u> (See table on last page of this section.) Also, even though we focus on the case where σ is unknown, it would be very easy to adapt to the case where σ is known.

9. Example:

1. A researcher constructed a scale to measure influence on family decision-making, and collected the following data from 8 pairs of husbands and wives:

Pair #	H score	W score	H - W = D	$(H - W)^2 = D^2$
1	26	30	-4	16
2	28	29	-1	1
3	28	28	0	0
4	29	27	2	4
5	30	26	4	16
6	31	25	6	36
7	34	24	10	100
8	37	23	14	196
Σ	243	212	31	369

(a) Test, at the .05 level, whether there is any significant difference between the average scores of husbands and wives.

(b) Construct the 95% c.i. for the average difference between the husband's and wife's score.

Solution. Let D = (Husband's score - Wife's score). In the sample, n = 8,

$$\bar{d} = \frac{\sum_{i=1}^{8} D_i}{8} = \frac{31}{8} = 3.875$$

$$s_D^2 = \frac{\sum_{i=1}^{N} (D_i - \bar{D})^2}{N - 1} = \frac{\sum_{i=1}^{N} D_i^2 - N\bar{D}^2}{N - 1} = \frac{369 - 8 * 3.875^2}{7} = 35.554$$

$$s_D = \sqrt{35.554} = 5.963$$

Step 1. Using α = .05, We want to test

H_0: E(D) = 0
H_A: E(D) <> 0

Step 2. The appropriate test statistic is

$$t = \frac{\bar{d} - \mu_{D_0}}{\frac{s_D}{\sqrt{N}}} = \frac{\bar{d} - \mu_{D_0}}{\sqrt{\frac{s_D^2}{N}}} = \frac{\bar{d}}{\sqrt{\frac{35.554}{8}}} = \frac{\bar{d}}{2.1081}$$

Step 3. For α = .05, accept H_0 if $-2.365 \le T_7 \le 2.365$, or, equivalently, accept H_0 if $-4.986 \le \hat{\mu}_D \le 4.986$ (since 2.365 * 2.1081 = 4.986).

Step 4. The computed test statistic equals

$$t = \frac{\bar{d} - \mu_{D_0}}{\frac{s_D}{\sqrt{N}}} = \frac{\bar{d} - \mu_{D_0}}{\sqrt{\frac{s_D^2}{N}}} = \frac{\bar{d}}{\sqrt{\frac{35.554}{8}}} = \frac{\bar{d}}{2.1081} = \frac{3.875}{2.1081} = 1.838$$

Step 5. Do not reject H_0.

(b) The 95% c.i. is

$$\bar{d} \pm (t_{\alpha/2,v} * s_D / \sqrt{N}), i.e.$$

$$3.875 - (2.365 * 5.963 / \sqrt{8}) \le \mu_D \le 3.875 + (2.365 * 5.963 / \sqrt{8}), i.e.$$

$$-1.11 \le \mu_D \le 8.86$$

Note that the hypothesized value of E(D) = 0 falls within this interval, hence the c.i. confirms that we should not reject H_0.

E. Case V: Difference between two proportions.

1. *We are often interested in comparing how proportions differ between two populations.* For example, we might want to compare the proportion of women who smoke with the proportion of men who smoke; or, we might ask whether blacks are more likely to support a political candidate than whites are.

2. *Let $\hat{p}_i$ = proportion of observed successes in group i* = X_i/N_i, where X_i = the number of successes in group i.

3. *Test statistic.* Recall that, if N_1 and N_2 are sufficiently large, then $\hat{p}_1 \sim N(p_1, p_1q_1/N_1)$, $\hat{p}_2 \sim N(p_2, p_2q_2/N_2)$. Therefore,

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

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{N_1} + \frac{p_2 q_2}{N_2} = \frac{N_2 p_1 q_1 + N_1 p_2 q_2}{N_1 N_2}$$

Note that, if $p_1 = p_2 = p$, then

$$V(\hat{p}_1 - \hat{p}_2) = \frac{(N_1 + N_2)pq}{N_1 N_2} = \frac{(N_1 + N_2)}{N_1 N_2} * p * q$$

But of course, we don't know what p is. And, within our samples, we have two estimates of p, $\hat{p}_1$ and $\hat{p}_2$. The pooled estimate of p and q is therefore

$$\hat{p} = \frac{N_1 \hat{p}_1 + N_2 \hat{p}_2}{N_1 + N_2} = \frac{X_1 + X_2}{N_1 + N_2}, \quad \hat{q} = 1 - \frac{X_1 + X_2}{N_1 + N_2}$$

This result is very intuitive. It says that the pooled estimate of p equals the total number of successes in the two samples divided by the total number of cases. Substituting the <u>pooled estimates</u> of p and q into the previous formula for the variance, we get

$$V(\hat{p}_1 - \hat{p}_2) = \left(\frac{N_1 + N_2}{N_1 N_2}\right)\left(\frac{X_1 + X_2}{N_1 + N_2}\right)\left(1 - \frac{X_1 + X_2}{N_1 + N_2}\right)$$

Again, recall that we are assuming that N_1 and N_2 are sufficiently large. Otherwise, our use of sample estimates is questionable.

Hence the <u>test statistic for H_0: $p_1 = p_2$</u> is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{N_1 + N_2}{N_1 N_2}\right)\left(\frac{X_1 + X_2}{N_1 + N_2}\right)\left(1 - \frac{X_1 + X_2}{N_1 + N_2}\right)}} = \frac{\frac{X_1}{N_1} - \frac{X_2}{N_2}}{\sqrt{\left(\frac{N_1 + N_2}{N_1 N_2}\right)\left(\frac{X_1 + X_2}{N_1 + N_2}\right)\left(1 - \frac{X_1 + X_2}{N_1 + N_2}\right)}}$$

4. Approximate confidence interval for $\hat{p}_1 - \hat{p}_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{N_1} + \frac{\hat{p}_2\hat{q}_2}{N_2}}$$

5. COMMENT: Note that problems involving binomial proportions always seem annoyingly more complicated, primarily because the mean and variance are not independent of each other. Often problems involving binomial proportions are treated the same way as problems involving σ unknown (Case II). For small samples, this is questionable, but for large samples it probably does not matter that much.

6. EXAMPLES:

1. Two groups, A and B, each consist of 100 randomly assigned people who have a disease. One serum is given to Group A and a different serum is given to Group B; otherwise, the two groups are treated identically. It is found that in groups A and B, 75 and 65 people, respectively, recover from the disease.

(a) Test the hypothesis that the serums differ in their effectiveness using $\alpha = .05$.

(b) Compute the approximate 95% c.i. for $\hat{p}_1 - \hat{p}_2$.

Solution. (a)

Step 1:

H_0: $p_1 - p_2 = 0$ or, $p_1 = p_2$

H_A: $p_1 - p_2 <> 0$ or, $p_1 <> p_2$

Step 2: The appropriate test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{N_1 + N_2}{N_1 N_2}\right)\left(\frac{X_1 + X_2}{N_1 + N_2}\right)\left(1 - \frac{X_1 + X_2}{N_1 + N_2}\right)}} = \frac{\frac{X_1}{100} - \frac{X_2}{100}}{\sqrt{\left(\frac{200}{10{,}000}\right)\left(\frac{X_1 + X_2}{200}\right)\left(1 - \frac{X_1 + X_2}{200}\right)}}$$

Step 3: Accept H_0 if $-1.96 \le Z \le 1.96$

Step 4: The computed value of the test statistic is

$$z = \frac{\frac{X_1}{100} - \frac{X_2}{100}}{\sqrt{\left(\frac{200}{10{,}000}\right)\left(\frac{X_1 + X_2}{200}\right)\left(1 - \frac{X_1 + X_2}{200}\right)}} = \frac{\frac{75}{100} - \frac{65}{100}}{\sqrt{\left(\frac{200}{10{,}000}\right)\left(\frac{75 + 65}{200}\right)\left(1 - \frac{75 + 65}{200}\right)}} = 1.543$$

Step 5: Therefore, do not reject H_0.

(b) The approximate 95% c.i. is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{N_1} + \frac{\hat{p}_2\hat{q}_2}{N_2}} = .75 - .65 \pm 1.96\sqrt{\frac{.75 * .25}{100} + \frac{.65 * .35}{100}}, i.e.$$

$$-.026 \le p_1 - p_2 \le .226$$

Note that 0 falls within the approximate c.i., again suggesting H_0 should not be rejected.

COMMENTS:

(1) Try working this problem as though it fell under Case II. You will find that results are virtually identical.

(2) Again, remember the c.i. is only approximate, but it works pretty well with large samples.

Single sample tests, case III (σ unknown) Compared to Two sample tests, Case IV (Matched pairs)		
	Single sample tests, case III, σ unknown	Two sample tests, case IV, matched pairs
Variable of interest	X	$D = X_1 - X_2$
Sample mean	$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$	$\overline{D} = \frac{\sum_{i=1}^{N}(X_{1_i} - X_{2_i})}{N} = \frac{\sum_{i=1}^{N} D_i}{N}$
Sample standard deviation	$s_X = \sqrt{\frac{1}{N-1}\sum(x_i - \overline{x})^2} =$ $\sqrt{\frac{1}{N-1}(\sum X_i^2 - N\overline{X}^2)}$	$s_D = \sqrt{\frac{1}{N-1}\sum(d_i - \overline{d})^2} =$ $\sqrt{\frac{1}{N-1}(\sum d_i^2 - N\overline{d}^2)}$
Estimated standard error of the estimate	$s_{\overline{X}} = \hat{\sigma}_{\overline{X}} = \frac{s_X}{\sqrt{N}}$	$s_{\overline{D}} = \hat{\sigma}_{\overline{D}} = \frac{s_D}{\sqrt{N}}$
Test statistic	$t = \frac{\overline{x} - \mu_{X_0}}{\frac{s_X}{\sqrt{N}}} = \frac{\overline{x} - \mu_{X_0}}{\sqrt{\frac{s_X^2}{N}}}$	$t = \frac{\overline{d} - \mu_{D_0}}{\frac{s_D}{\sqrt{N}}} = \frac{\overline{d} - \mu_{D_0}}{\sqrt{\frac{s_D^2}{N}}}$