CHAPTER 9: INFERENCES FROM TWO SAMPLES (LESSON 29.5)

Only Part A of this Lesson is fair for your Final.

PART A: MEANS FROM DEPENDENT SAMPLES: MATCHED PAIRS (SECTION 9-4)

A matched pair of data values may correspond to two different measures for one individual (as in Example 1 below), the same measures for a husband/wife couple, etc.

When comparing the means for the two measures, we perform our usual tests on the **differences** (d) between the measures for each matched pair.

The population data

Let D be the distribution of the population of differences between all the matched pairs.

Let μ_d be the mean of the D distribution.

The sample data

Let $\bar{d}$ and s_d be the sample mean and the sample standard deviation, respectively, for the differences between the paired sample data values.

Let n be the number of matched pairs of sample data values.

CLT Assumptions:

We require:

- $n > 30$, or
- D is approximately normally distributed.

Example 1

We take a random sample of three men from the participants in a men's weight loss program. Use the sample data below to test the claim that participants in the program lose weight on average. Use a significance level of 0.05. Assume that the weight changes of the participants in the program are approximately normally distributed.

Subject #	Before Weight (lbs.)	After Weight (lbs.)
1	230	225
2	250	248
3	210	211

Solution to Example 1

We calculate the differences, *d*, from the given table.

Subject #	Before Weight (lbs.)	After Weight (lbs.)	Differences (*d* in lbs.)
1	230	225	5
2	250	248	2
3	210	211	−1

Here, we take differences "Before" – "After." More on this later.

Sample statistics:

$$(n = 3)$$

$$\bar{d} = 2.0 \text{ lbs.}$$
$$s_d = 3.0 \text{ lbs.}$$

Setup:

$$H_0:\ \mu_d = 0\ (\text{lbs.})$$
$$H_A:\ \mu_d > 0\ (\text{lbs.})\quad (\Rightarrow \text{Right-tailed test})\quad (\text{Claim})$$
$$\alpha = 0.05$$

Observe that a positive value of μ_d corresponds to weight loss (on average). If differences had been taken the other way, "After" – "Before," then we would have $H_A: \mu_d < 0$ (lbs.) and we would conduct a left-tailed test.

Test statistic

We're testing for a population mean [difference between matched pairs], we're assuming that the population of differences is approximately normally distributed, and the population standard deviation of the differences is presumed unknown, so we use the *t* test statistic.

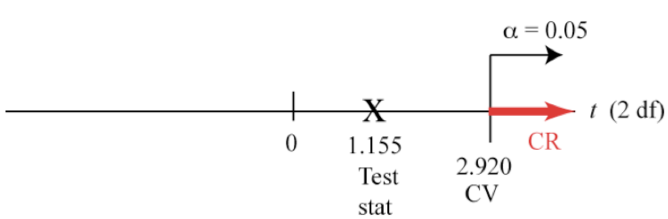
$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \quad (\mu_d = 0 \text{ under } H_0)$$

$$= \frac{2.0}{\frac{3.0}{\sqrt{3}}}$$

$$\approx 1.155$$

Critical value (CV) and critical region (CR); this is a right-tailed test.

We need to use the *t* distribution on $n-1=3-1=2$ degrees of freedom. We want to use the 0.05 (one tail) column.



Decision

The test statistic value is not in the critical region (CR), so we do not reject H_0.

Final Conclusion

There is insufficient evidence for (in support of) the claim that participants in the program lose weight on average.

PART B: COMPARING MEANS FROM INDEPENDENT SAMPLES (SECTION 9-3) (NOT ON THE FINAL)

	Population 1 (Ex.: Weights of U.S. Men in lbs.)	**Population 2 (Ex.: Weights of U.S. Women in lbs.)**
Population means (unknown)	μ_1	μ_2
Population standard deviations (unknown)	σ_1	σ_2
	Sample 1 (from Population 1)	**Sample 2 (from Population 2)**
Sample sizes	n_1	n_2
Sample standard deviations	s_1	s_2

Example (Medical study): Control (Placebo) vs. Treatment (Drug).

CLT Assumptions:

We require:

- $n_1 > 30$ and $n_2 > 30$, or
- the populations are approximately normally distributed.

The null hypothesis is written as $H_0: \mu_1 = \mu_2$.

The "two sample *t*" test statistic formula is:

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \cancelto{0 \text{ under } H_0}{\left(\mu_1 - \mu_2\right)}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where # df = (smaller of n_1 and n_2) – 1.

Note: Another formula for # df is given by Formula 9-1 on p.450.

Note 1: If the population variances σ_1^2 and σ_2^2 are known, then use the analogous z test statistic formula that uses them instead of s_1^2 and s_2^2:

$$z = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \overbrace{\left(\mu_1 - \mu_2\right)}^{0 \text{ under } H_0}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Note 2: If we assume that the population variances are equal ($\sigma_1^2 = \sigma_2^2$), then some people use a pooled sample variance s_p^2 that estimates that common variance, though some statisticians warn against this. s_p^2 is a weighted average of s_1^2 and s_2^2; the larger of the two samples has more of an impact on the value of s_p^2.

The revised test statistic formula is then:

$$t = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \overbrace{\left(\mu_1 - \mu_2\right)}^{0 \text{ under } H_0}}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}},$$

where $s_p^2 = \dfrac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{\left(n_1 - 1\right) + \left(n_2 - 1\right)}$, or $s_p^2 = \dfrac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}$

df $= n_1 + n_2 - 2$.

Note: This higher # df corresponds to a t distribution that looks more like a z distribution; the SD for the t distribution is lower now than before. It is generally "easier" to reject H_0 now, and we have a more "powerful" test.

PART C: COMPARING PROPORTIONS FROM INDEPENDENT SAMPLES (SECTION 9-2) (NOT ON THE FINAL)

	Population 1	**Population 2**
Population proportions (unknown)	p_1	p_2
	Sample 1 (from Population 1)	**Sample 2 (from Population 2)**
Sample sizes	n_1	n_2
Numbers of successes	x_1	x_2
Sample proportions	$\hat{p}_1 = \frac{x_1}{n_1}$	$\hat{p}_2 = \frac{x_2}{n_2}$
Required	$x_1 \left(\text{or } n_1\hat{p}_1\right) \geq 5$	$x_2 \left(\text{or } n_2\hat{p}_2\right) \geq 5$

The null hypothesis is written as $H_0: \ p_1 = p_2$.

The z test statistic formula is:

$$z = \frac{\left(\hat{p}_1 - \hat{p}_2\right) - \cancel{\left(p_1 - p_2\right)}^{\text{0 under } H_0}}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}},$$

where $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$, and $\bar{q} = 1 - \bar{p}$.

$\bar{p}$ is called the pooled sample proportion. Under H_0, the two population proportions are equal, and $\bar{p}$ is our point estimate for this common population proportion. It equals the total number of successes from both samples divided by the total number of trials from both samples.

PART D: COMPARING STANDARD DEVIATIONS OR VARIANCES FROM INDEPENDENT SAMPLES FROM NORMAL POPULATIONS
(NOT ON THE FINAL)

"Population 1" corresponds to the sample with the larger sample variance, denoted by s_1^2. The population variance is denoted by σ_1^2.

"Population 2" corresponds to the sample with the smaller sample variance, denoted by s_2^2. The population variance is denoted by σ_2^2.

The null hypothesis is written as $H_0: \sigma_1 = \sigma_2$, or $H_0: \sigma_1^2 = \sigma_2^2$.

The F test statistic formula is given by:

$$F = \frac{s_1^2}{s_2^2}$$, where s_1^2 is the larger of the two sample variances.

In order to use the F distribution table, we require two numbers of degrees of freedom:

• The # df for the numerator, denoted by # df_1, is given by $n_1 - 1$, where n_1 is the sample size for the sample with the larger variance.

• The # df for the denominator, denoted by # df_2, is given by $n_2 - 1$, where n_2 is the sample size for the sample with the smaller variance.

If we agree that "Population 1" corresponds to the sample with the larger sample variance, then it would not make sense to conduct a left-tailed test; we would not use H_1 or $H_A: \sigma_1 < \sigma_2$, or H_1 or $H_A: \sigma_1^2 < \sigma_2^2$.

• For all practical purposes, we can act as though we have a right-tailed test with only one critical value.

• If the original test was a two-tailed test, with H_1 or $H_A: \sigma_1 \neq \sigma_2$, or H_1 or $H_A: \sigma_1^2 \neq \sigma_2^2$, then the right tail of the critical region corresponds to a probability (or area) of $\alpha / 2$.

• If the original test was a right-tailed test, with H_1 or H_A : $\sigma_1 > \sigma_2$, or H_1 or H_A : $\sigma_1^2 > \sigma_2^2$, then the right-tailed critical region corresponds to a probability (or area) of α.