

**Class 16: The T-Distribution for One and Two Samples (Text: Sections 7.1 and 7.2)****When can we safely use a T-Test? Robustness of T-Procedures**

An inference procedure, such as a  $t$ -test is **robust** if it is not sensitive to violations for the assumptions. Assumptions we need to make for the  $T$ -test:

**Normal**

- The  $t$ -distribution applies if original distribution normal
- Since  $\bar{x}, s$  are not resistant to outliers, this procedure not robust to outliers
- Provided there are no outliers, procedure is robust with respects to skewness unless very extreme. (That is, unless there is a very noticeable tail in one direction.)

**Size of sample:** By CLT larger samples mean  $\bar{x}$  is more normally distributed

- If  $n < 15$ , use  $t$ -procedure if original population is close to normal, no outliers
- If  $n \geq 15$ , use  $t$  unless outliers, strong skewness, or strongly not normal
- If  $n \geq 40$ , use  $t$  even if skewed
- If  $n \geq 120$ , can use  $Z$  because  $Z$  and  $t$  distributions are approximately the same for large df.

**We have a SRS:** It is important that the sampling be well done.

**One versus Two Samples**

When testing a hypothesis, we ask whether the sample mean is significantly different from some particular value—for example, if coached students had significantly higher SAT scores than the average.

- But how do we know that the coached students are like the rest of the population? Indeed, the fact that they signed up for coaching suggests that they may be different.
- In practice, we often study a *control* group and compare two samples

Today we see how to see if there is a significant difference between two samples.

**MATCHED PAIRS (Section 7.1)**

Suppose we had SAT results for a group of students before and after coaching. Then we have *matched pairs*: we have two samples (the pre-coaching scores and post-coaching scores) and they are matched in pairs coming from the same person. We test to see if the gain in scores is significant.

**Ex: Forty students who are coached have SAT scores where the mean gain is 9.3 and the standard deviation of the gain is 8.1. Is the gain significant?**

**Ex: Find a 95% confidence interval for the mean increase in SAT score with coaching**

Notice that even if the increase in scores is *statistically significant*, it may not be of *practical significance*.

**Assumptions for t-test for matched pairs**

- Random variable in original population is normally distributed, so difference is normal

**Matched pairs are useful because:**

- Comparative studies enable us to avoid the avoid effects of confounding (for example, choosing to get coaching with an increase in SAT scores—the students who chose coaching may have been more likely to see an increase in scores anyway)
- Can be used even when randomization is not possible

*But* students who choose to go for coaching may not represent the rest of the population, so we may not be able to generalize the results to other students. In other words, coaching may not help other students in the same way. Better to compare increases in coached students' scores with increases of other students' scores; this is *difference in differences*. (Difference in differences is not in Math 263.)

**COMPARING TWO MEANS : UNMATCHED SAMPLES** (Section 7.2)

Suppose we want to compare the means of two groups, but

- No pairing of individuals (not matched pairs)
- Samples can be different sizes

We will assume that

- We want to compare means of two groups
- Each group is a sample from a different population
- Responses in each group are independent of those from other group

**Notation:**

|              | Variable | Mean        | Standard deviation | Size  |
|--------------|----------|-------------|--------------------|-------|
| Population 1 | $x_1$    | $\mu_1$     | $\sigma_1$         |       |
| Population 2 | $x_2$    | $\mu_2$     | $\sigma_2$         |       |
| Sample 1     |          | $\bar{x}_1$ | $s_1$              | $n_1$ |
| Sample 2     |          | $\bar{x}_2$ | $s_2$              | $n_2$ |

**Goal: To estimate  $\mu_1 - \mu_2$** 

Estimator is  $\bar{x}_1 - \bar{x}_2$ ; we want to know the distribution of  $\bar{x}_1 - \bar{x}_2$

- If original distributions of  $x_1$  and  $x_2$  are both normal, then distribution of  $\bar{x}_1 - \bar{x}_2$  is normal
- Mean is given by

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

(This true even without the normality.)

- $\text{Var}(\bar{x}_1 - \bar{x}_2) = \text{Var}(\bar{x}_1) + \text{Var}(\bar{x}_2)$  if  $\bar{x}_1$  and  $\bar{x}_2$  are independent. Then

$$\text{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

- So standard deviation is given by

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Notice this formula for the SE depends on the two samples being independent.

**Two-Sample Z-statistic for Hypothesis Test**

If we know the standard deviations of both populations, then

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

**Two-Sample T-statistic for Hypothesis Test**

If we use the sample standard deviations instead of the population standard deviations, then we have a statistic that does not have exactly a  $T$ -distribution, but can be approximated by one:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T(k)$$

where the degree of freedom is  $k = \min(n_1 - 1, n_2 - 1)$  if we calculate by hand.

Degrees of Freedom by Computer: (Optional) If we calculate the  $df$  by computer, the formula used is that below, which gives values close to those obtained by hand—but may not be integers:

$$df = k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

**Assumption for Two Sample Test:**

- The two samples must be independent.

**Hypothesis Test for Two Sample Means**

Follows the usual steps; calculate the standard deviation using the new formula.

**Ex: Are the following two schools comparable in SAT scores?**

School 1: A random sample of 43 students has mean SAT of 502, and standard deviation of 60

School 2: A random sample of 35 students has mean SAT of 480, and standard deviation of 50

Show all the steps in the reasoning and include an interpretation of the  $P$ -value.

**Interpretation of the P-value:**

**Confidence Intervals for Difference in Two Population Means**

If we know the population standard deviations, the confidence interval for the difference in population means is given by

$$\left( \bar{x}_1 - \bar{x}_2 - z \cdot \sigma_{\bar{x}_1 - \bar{x}_2}, \bar{x}_1 - \bar{x}_2 + z \cdot \sigma_{\bar{x}_1 - \bar{x}_2} \right)$$

If we do not know the population standard deviations, the confidence interval for the difference in population means is given by

$$\left( \bar{x}_1 - \bar{x}_2 - t \cdot SE_{\bar{x}_1 - \bar{x}_2}, \bar{x}_1 - \bar{x}_2 + t \cdot SE_{\bar{x}_1 - \bar{x}_2} \right)$$

Where

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{and} \quad SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

**Ex:** Find a 95% confidence interval for difference in population means for the two schools in the previous example.

**Ex:** Interpret the 95% confidence interval in the previous example.

**Relationship Between Confidence Intervals and Hypothesis Tests**

**Ex:** What does the confidence interval tell us about the difference in population means of the two schools?

**If 0 is in a confidence interval, the corresponding *two* sided hypothesis test is not significant.**

**Pooled estimate for  $SE_{\bar{x}}$  (Optional)**

If we have two samples from populations that we believe have the same standard deviation (even though we may not know what that value is), we can estimate the shared standard deviation and get a statistics which exactly has the  $T$ -distribution. This is called a **pooled t-test**.

If both populations have same  $\sigma$ , then  $s_1^2$  and  $s_2^2$  both estimate  $\sigma^2$ . We take a weighted average:

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

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad \text{with } df = n_1 + n_2 - 2$$

We can use this standard error in confidence intervals and hypothesis tests.