Math 3081 (Probability and Statistics)

Lecture #24 of 27 $\sim$ August 16th, 2021

Two-Sample t Tests

- Student's t Test
- Welch's t Test
- Confidence Intervals

This material represents §4.2.4 from the course notes, and problems 6-9 from WeBWorK 7.

Last lecture, we talked all about one-sample t tests. Today we will discuss two-sample t tests, in which we want to compare the unknown means of two normally-distributed populations with unknown standard deviations.

I gave the motivation last time but let me review it again.

Suppose the two populations are labeled A and B, with respective means μ_A and μ_B and population standard deviations σ_A and σ_B.

- If we don't know σ_A or σ_B, we must use S_A and S_B to estimate them, to find the standard deviation $\sqrt{\sigma_A^2 + \sigma_B^2}$ of the difference in means $\mu_A - \mu_B$.
- However, using the sample standard deviation in place of the population standard deviation changes the underlying distributions.
- The problem is that, although we can write down a model, it does not yield itself to pleasant calculations.

Two-Sample t Tests, III

As I calculated last time, $\mu_A - \mu_B$ is modeled by
$(\hat{\mu}_A - \hat{\mu}_B) + \dfrac{S_A}{\sqrt{n_A}} T_{n_A - 1} - \dfrac{S_B}{\sqrt{n_B}} T_{n_B - 1}....$ (whatever that is!)

- The problem is that we don't have a nice description of what the difference between two (scaled) t distributions looks like.
- For normal distributions, we can use the very convenient fact that the sum or difference of normal random variables is also normal; that is not the case for t distributions!
- In principle, because we know the probability density functions of $T_{n_A - 1}$ and $T_{n_B - 1}$, and they are independent, we could calculate the probability density function of the random variable listed above for particular values of the parameters.
- But that does not solve our problem, because we need to write down a distribution that is independent of the test parameters (i.e., that does not depend on S_A and S_B).

In fact, the problem of finding an appropriate statistic for testing the equality of two sample means from normally distributed random samples is still unsolved!

- This problem is known as the <u>Behrens-Fisher problem</u>. (Various generalizations are also often named with this moniker as well.)

- It may seem dispiriting to have an unsolved problem be an obstacle to doing two-sample t tests.

- What we will do instead is construct various partial or approximate solutions to the problem.

The first approximation is to construct a <u>pooled</u> standard deviation, much like our approach previously when we did two-sample z tests for binomially distributed data. The approach comes from the following theorem:

(the statement is so long I need to put it on the next slide)

Two-Sample t Tests: Equal Variances, II

Theorem (t Distribution With Pooled Variance)

Suppose that $X_1, \ldots, X_n$ are independent and identically normally distributed with mean μ_X and standard deviation σ, and that $Y_1, \ldots, Y_m$ are independent and identically normally distributed with mean μ_Y and standard deviation σ. If $\hat{\mu}_X$, $\hat{\mu}_Y$, S_X, and S_Y denote the sample means and sample standard deviations of $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$, then for the

pooled standard deviation $S_{pool} = \sqrt{\dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}} =$

$\sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)^2 + \sum_{j=1}^{m}(Y_j - \hat{\mu}_Y)^2}{m+n-2}}$, *the distribution of the*

test statistic $\dfrac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{S_{pool}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$ is T_{m+n-2}, the t

distribution with $m + n - 2$ degrees of freedom.

The point of this theorem is that if we assume that the population variances are equal, then the test statistic

$$t = \frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{S_{\text{pool}}\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$$

for the pooled standard deviation

$$S_{\text{pool}} = \sqrt{\frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}}$$

is t-distributed with $m + n - 2$ degrees of freedom. This allows us to do our hypothesis test with these parameters.

<u>Proof</u> (outline):

- The test statistic is the quotient of $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ by $\frac{S_{\text{pool}}}{\sigma}$.

- The first term $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$, is normally distributed with mean 0 and standard deviation 1 (a standard normal).

- The other term has square
  $\frac{S_{\text{pool}}^2}{\sigma^2} = \frac{1}{m+n-2}\left[\sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}_X}{\sigma}\right)^2 + \sum_{j=1}^{m}\left(\frac{Y_j - \hat{\mu}_Y}{\sigma}\right)^2\right]$.

- As noted in the proof of our earlier theorem, the sum $\sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}_X}{\sigma}\right)^2$ can be rewritten as the sum of squares of $n-1$ standard normal variables, and by the same argument, $\sum_{i=1}^{n}\left(\frac{X_i - \hat{\mu}_X}{\sigma}\right)^2$ can be rewritten as the sum of squares of $m-1$ standard normal variables.

- Therefore, the sum of these terms is the sum of squares of $m+n-2$ standard normal variables.

Two-Sample t Tests: Equal Variances, V

Proof (outline) (continued):

- The first term $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ is the sum of squares of
  $m + n - 2$ standard normal variables.

- But as we showed in our earlier theorem, the distribution of a
  ratio $\frac{\overline{x} - \mu}{S/\sqrt{n}} = \frac{(\overline{x} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{N_{0,1}}{S/\sigma}$ is the t distribution
  T_{n-1} with degrees of freedom equal to the number of squares
  of standard normal variables summed in the denominator.

- Since there are $m + n - 2$ standard normal variables, that
  means that our quotient of $\frac{(\hat{\mu}_X - \hat{\mu}_Y) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$ by $\frac{S_{\text{pool}}}{\sigma}$ is
  t-distributed with $m + n - 2$ degrees of freedom, as claimed.

Two-Sample t Tests: Equal Variances, V

This theorem gives us an explicit procedure for performing a two-sample t test with a pooled standard deviation where the population variances are assumed to be equal. This test is known as <u>Student's equal-variances t test</u>.

- First, we select appropriate null and alternative hypotheses and a significance level α.

- Our null hypothesis will be of the form H_0: $\mu_A - \mu_B = c$ for some constant c that is our hypothesized value for the difference of the means (usually 0).

- We take the test statistic $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{pool}}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}$, where

  $S_{\text{pool}} = \sqrt{\dfrac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}}$ is the pooled standard deviation estimate.

From the theorem, the distribution of the test statistic will be the t-distribution $T_{n_A+n_B-2}$ with $n_A + n_B - 2$ degrees of freedom.

- We then find the p-value based on the alternative hypothesis.
- If the hypotheses are $H_0 : \mu_A - \mu_B = c$ and $H_a : \mu_A - \mu_B > c$, then the p-value is $P(T_{n_A+n_B-2} \geq t)$.
- If the hypotheses are $H_0 : \mu_A - \mu_B = c$ and $H_a : \mu_A - \mu_B < c$, then the p-value is $P(T_{n_A+n_B-2} \leq t)$.
- If the hypotheses are $H_0 : \mu_A - \mu_B = c$ and $H_a : \mu_A - \mu_B \neq c$, then the p-value is
$$P(|T_{n_A+n_B-2}| \geq |t|) = \begin{cases} 2P(T_{n_A+n_B-2} \geq t) & \text{if } t \geq \mu \\ 2P(T_{n_A+n_B-2} \leq t) & \text{if } t < \mu \end{cases}.$$

We then compare the p-value to the significance level and then either reject or fail to reject the null hypothesis, as usual.

Two-Sample t Tests: Equal Variances, VII

Before doing an example, I'll briefly motivate why these particular choices of parameters (the pooled standard deviation, the test statistic, and the number of degrees of freedom) are logical.

- If nothing else, it might help you remember them when you're doing the problems!
- The denominator of the test statistic is analogous to the standard deviation $\sqrt{\dfrac{\sigma^2}{n_A} + \dfrac{\sigma^2}{n_B}}$ for the difference of the two averages coming from normal distributions with the same standard deviation σ, and is what we would divide by if we were doing a two-sample z test to get a normalized statistic.
- The number of degrees of freedom of the t distribution is $m + n - 2$, because we have $m + n$ data points, but we lose one degree of freedom by comparing A to its mean, and we lose another by comparing B to its mean.

The pooled standard deviation comes from pooling variances.

- If the two sample means were actually equal to the same number μ, then the variance of $\{X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_m\}$ would be
  $\frac{1}{m+n} \left[(X_1 - \mu)^2 + \cdots + (X_n - \mu)^2 + (Y_1 - \mu)^2 + \cdots + (Y_m - \mu)^2 \right]$.

- However, since the two sets don't have the same mean, we instead measure them relative to their own means.

- Furthermore, the resulting variance estimate is biased (for the same reason that the estimate for the sample variance for a single sample is biased), so we must divide by $\frac{1}{m+n-2}$ rather than $\frac{1}{m+n}$ to unbias it.

- We can then rewrite the complicated sum above more simply in terms of the sample standard deviations S_X and S_Y: this is precisely the pooled standard deviation S_{pool}.

Example: A statistics instructor wants to determine whether students do better on exams in a morning class or in an evening class. They randomly sample 11 exams from the morning class, which have an average score of 84 and a sample standard deviation of 13, and compare to a random sample of 11 exams from the evening class, which have an average score of 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

1. Find the pooled standard deviation and degrees of freedom.
2. Test whether the average score in the morning class is higher.
3. Test whether the average scores in the classes are different.
4. Test whether the average score in the morning class is at least 2 points higher than the evening class.

<u>Example</u>: A statistics instructor wants to determine whether students do better on exams in a morning class or in an evening class. They randomly sample 11 exams from the morning class, which have an average score of 84 and a sample standard deviation of 13, and compare to a random sample of 11 exams from the evening class, which have an average score of 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

1. Find the pooled standard deviation and degrees of freedom.
2. Test whether the average score in the morning class is higher.
3. Test whether the average scores in the classes are different.
4. Test whether the average score in the morning class is at least 2 points higher than the evening class.

- Since the population variances are assumed to be equal, we use Student's equal-variances t test.

<u>Example</u>: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal.

1. Find the pooled standard deviation and degrees of freedom.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal.

1. Find the pooled standard deviation and degrees of freedom.

- The pooled standard deviation is
$$S_{\text{pool}} = \sqrt{\frac{(11-1) \cdot 13^2 + (11-1) \cdot 9^2}{11 + 11 - 2}} = 11.1803.$$
- The number of degrees of freedom is $11 + 11 - 2 = 20$.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

2. The average score in the morning class is higher.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

2. The average score in the morning class is higher.

- Our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e > 0$; we want this one-sided alternative hypothesis since the morning class average is higher than the evening class average.

- With $S_{\text{pool}} = 11.1803$, the test statistic is
$t = \dfrac{(84-77)-0}{11.1803 \cdot \sqrt{\frac{1}{11} + \frac{1}{11}}} = 1.4683$.

- Thus, the p-value is $P(T_{20} \geq 1.4683) = 0.07878$.

- The p-value is less than the first significance level, we reject the null hypothesis there, but fail to reject the null hypothesis in the other cases.

<u>Example</u>: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

3. The average scores in the classes are different.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

3. The average scores in the classes are different.

- Our hypotheses are now $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e \neq 0$; we want this two-sided alternative hypothesis since now we want only to test whether the scores are equal.

- The parameters are the same as the last part; the only difference is that the p-value is now doubled: $2P(T_{20} \geq 1.4683) = 0.1576$.

- Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

4. The average score in the morning class is at least 2 points higher than the evening class.

Example: 11 exams from the morning class have an average 84 and a sample standard deviation of 13, and 11 exams from the evening class have average 77 and a sample standard deviation of 9. Assume that the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels that

4. The average score in the morning class is at least 2 points higher than the evening class.

- Our hypotheses are $H_0 : \mu_m - \mu_e = 2$, $H_a : \mu_m - \mu_e > 2$ (the morning class actually did score more than two points above the evening class).
- We still have $S_{\text{pool}} = 11.1803$.
- The test statistic is $t = \frac{(84-77)-2}{11.1803 \cdot \sqrt{\frac{1}{11}+\frac{1}{11}}} = 1.0488$, giving p-value $P(T_{20} \geq 1.0488) = 0.1534$.
- Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.

<u>Example</u>: To test which ride service costs less (Süper or Hyft), a news team takes identical trips on each service. On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.49, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

1. Find the sample means, sample standard deviations, pooled standard deviation, and degrees of freedom.
2. Test whether the cost on Süper is higher than on Hyft.
3. Test whether the average cost on Süper was more than \$1.50 more expensive than on Hyft.

<u>Example</u>: To test which ride service costs less (Süper or Hyft), a news team takes identical trips on each service. On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.49, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

1. Find the sample means, sample standard deviations, pooled standard deviation, and degrees of freedom.

2. Test whether the cost on Süper is higher than on Hyft.

3. Test whether the average cost on Süper was more than \$1.50 more expensive than on Hyft.

- Since the population variances are assumed to be equal, we use Student's equal-variances t test.

<u>Example</u>: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

1. Find the sample means, sample standard deviations, pooled standard deviation, and degrees of freedom.

Example: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

1. Find the sample means, sample standard deviations, pooled standard deviation, and degrees of freedom.

- We compute $\mu_S = \$24.69$, $S_S = \$3.1019$, $\mu_H = \$21.99$, and $S_H = \$3.3445$.

- Also, $S_{\text{pool}} = \$3.2428$, and the number of degrees of freedom is $4 + 5 - 2 = 7$.

<u>Example</u>: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

2. Test whether the cost on Süper is higher than on Hyft.

Example: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

2. Test whether the cost on Süper is higher than on Hyft.

- Our hypotheses are $H_0 : \mu_S - \mu_H = 0$, $H_a : \mu_S - \mu_H > 0$ (Süper was more expensive than Hyft in the sample).
- We still have $S_{\text{pool}} = \$3.2428$.
- The test statistic is $t = \frac{(\$24.69 - \$21.99) - 0}{\$3.2428 \cdot \sqrt{\frac{1}{4} + \frac{1}{5}}} = 1.2419$, giving p-value $P(T_7 \geq 1.2419) = 0.1273$.
- The p-value just squeaks below 13% so we reject the null there, but not in the other two cases.

<u>Example</u>: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

3. Test whether the average cost on Süper was more than \$1.50 more expensive than on Hyft.

Example: On Süper, the costs were \$21.30, \$22.98, \$28.09, and \$26.39. On Hyft, the costs were \$27.36, \$22.88, \$21.05, \$19.53, and \$19.13. Assume that the costs are normally distributed with equal population variances, and test at the 13%, 6%, and 0.9% significance levels.

3. Test whether the average cost on Süper was more than \$1.50 more expensive than on Hyft.

- Our hypotheses are $H_0 : \mu_S - \mu_H = 1.5$, $H_a : \mu_S - \mu_H > 1.5$ (Süper was more than \$1.50 more expensive in the sample).
- We still have $S_{\text{pool}} = \$3.2428$.
- The test statistic is $t = \frac{(\$24.69 - \$21.99) - \$1.50}{\$3.2428 \cdot \sqrt{\frac{1}{4} + \frac{1}{5}}} = 0.5516$, giving p-value $P(T_7 \geq 0.5516) = 0.2992$.
- The p-value is above all three significance levels, so we fail to reject in each case.

In most situations when we are comparing two populations, it is not reasonable to assume that the population variances are the same. For this reason, various unpooled two-sample t tests have been developed.

- The most popular such test is known as <u>Welch's unequal-variances t test</u>.
- Welch's t-test is generally more accurate than Student's equal-variances t test (described above) in the situation where the two sample variances are far apart, or when the sample sizes differ drastically.

Welch's unequal-variances t-test uses an unpooled version of the standard deviation.

- With null hypothesis H_0: $\mu_A - \mu_B = c$, the test statistic is $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{unpool}}}$, where we use the unpooled standard deviation $S_{\text{unpool}} = \sqrt{\dfrac{S_A^2}{n_A} + \dfrac{S_B^2}{n_B}}$.

- Note that S_{unpool} is the natural standard deviation estimate for the difference in the sample means: μ_A has standard deviation $\sigma_A/\sqrt{n_A}$ while μ_B has standard deviation $\sigma_B/\sqrt{n_B}$.

- Thus, the standard deviation for $\mu_A - \mu_B$ is $\sqrt{\dfrac{\sigma_A^2}{n_A} + \dfrac{\sigma_B^2}{n_B}}$.

- Now estimate the σs by Ss: this gives the unpooled standard deviation above.

Welch's unequal-variances t-test uses an unpooled version of the standard deviation.

- The test statistic from Welch's test does not have an exact distribution we can describe in any convenient way.
- However, as proven by Welch, it turns out to be approximately t-distributed by the t distribution with the number of degrees of freedom equal to the rather complicated formula $df = \dfrac{(S_A^2/n_A + S_B^2/n_B)^2}{\frac{1}{n_A-1}(S_A^2/n_A)^2 + \frac{1}{n_B-1}(S_B^2/n_B)^2}$.
- Most computer systems allow the number of degrees of freedom to be an arbitrary positive real number (in which case one may use the exact value given above); otherwise, such as when using tables or doing the WeBWorK, one usually rounds to the nearest integer.

Welch's result is quite technical, but we can describe roughly where the formula for the degrees of freedom comes from.

- The idea is to rewrite the quotient in the test statistic (in the same way we did in the theorem earlier) and try to write the denominator ratio as the sum of squares of independent standard normals.

- This cannot be done exactly, but if it could, we would then be able to find the number of terms by using the method of moments to compare the means and variances of the two expressions.

- Carefully going through the calculations eventually yields the degree-of-freedom formula given on the previous slide.

So, just to summarize for you the setup of Welch's test:

- We use Welch's unequal-variances t test to compare the sample means for populations whose variances are not assumed to be equal.

- With null hypothesis H_0: $\mu_A - \mu_B = c$, the test statistic is $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{unpool}}}$, where we use the unpooled standard deviation $S_{\text{unpool}} = \sqrt{\dfrac{S_A^2}{n_A} + \dfrac{S_B^2}{n_B}}$.

- The test statistic is approximately t-distributed with degrees of freedom $df = \dfrac{(S_A^2/n_A + S_B^2/n_B)^2}{\dfrac{1}{n_A - 1}(S_A^2/n_A)^2 + \dfrac{1}{n_B - 1}(S_B^2/n_B)^2}$.

<u>Example</u>: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

1. Find the unpooled standard deviation and degrees of freedom.
2. Test whether the average score in the morning class is higher.
3. Test whether the average scores in the classes are different.
4. Test whether the average score in the morning class is at least 2 points higher than the evening class.

- Now we use Welch's test, since we are not assuming the population variances are equal.

<u>Example</u>: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

1. Find the unpooled standard deviation and degrees of freedom.

Example: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

1. Find the unpooled standard deviation and degrees of freedom.

- The unpooled standard deviation is
$$S_{\text{unpool}} = \sqrt{\frac{13^2}{11} + \frac{9^2}{11}} = 4.7673.$$

- The number of degrees of freedom is
$$df = \frac{(13^2/11 + 9^2/11)^2}{\frac{1}{11-1}(13^2/11)^2 + \frac{1}{11-1}(9^2/11)^2} = 17.7951.$$

<u>Example</u>: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

  2. Test whether the average score in the morning class is higher.

Example: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

2. Test whether the average score in the morning class is higher.

- Our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e > 0$; we want this one-sided alternative hypothesis since the morning class average is higher than the evening class average.
- We also have $S_{\text{unpool}} = 4.7673$.
- The test statistic is thus $t = \frac{(84-77)-0}{4.7673} = 1.4683$, giving p-value $P(T_{17.7951} \geq 1.4683) = 0.07973$.
- The p-value is less than the first significance level, we reject the null hypothesis in that case, but we fail to reject the null hypothesis in the other cases.

<u>Example</u>: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

3. Test whether the average scores in the classes are different.

Example: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

3. Test whether the average scores in the classes are different.

- Our hypotheses are $H_0 : \mu_m - \mu_e = 0$, $H_a : \mu_m - \mu_e \neq 0$; we want this two-sided alternative hypothesis since now we want only to test whether the scores are equal.

- The parameters are the same as in (i) above; the only difference is that the p-value is now $2P(T_{17.7951} \geq 1.4683) = 0.1594$.

- Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.

<u>Example</u>: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

4. Test whether the average score in the morning class is at least 2 points higher than the evening class.

Example: Consider again the exams example (morning class with 11 exams of average score 84 and sample standard deviation 13, evening class with 11 of average score 77 and sample standard deviation 9. Now do not assume the population variances are equal, and test at the 11%, 3%, and 0.7% significance levels.

4. Test whether the average score in the morning class is at least 2 points higher than the evening class.

- Our hypotheses are $H_0 : \mu_m - \mu_e = 2$, $H_a : \mu_m - \mu_e > 2$.
- The unpooled standard deviation and degrees of freedom are the same as before.
- The test statistic is $t = \dfrac{(84 - 77) - 2}{4.7673} = 1.0488$, giving p-value $P(T_{17.7951} \geq 1.0488) = 0.1542$.
- Since the p-value is above all three significance levels, we fail to reject the null hypothesis in each case.

We can also adapt the two testing methods to give confidence intervals for the difference of two sample means.

- Using either Student's or Welch's procedure, we simply compute the appropriate t-statistic for the t distribution with the number of degrees of freedom indicated by the method, and take as the standard deviation either
  $S = S_{\text{pool}}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}$ or $S = S_{\text{unpool}}$ respectively.

- Then the desired $100(1 - \alpha)\%$ confidence interval will be given by $(\hat{\mu}_A - \hat{\mu}_B) \pm t_{\alpha/2, df}\, S$.

<u>Example</u>: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels.

1. Find the sample means and standard deviations of each group.

2. Test using Student's equal-variances t test whether the average salaries are equal.

3. Test using Welch's unequal-variances t test whether the average salaries are equal.

4. Find 80% and 95% confidence intervals for the difference between the average salaries using the pooled estimate.

5. Find 80% and 95% confidence intervals for the difference between the average salaries using the unpooled estimate.

<u>Example</u>: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels.

1. Find the sample means and standard deviations of each group.

<u>Example</u>: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels.

1. Find the sample means and standard deviations of each group.

- For the male faculty ($n_m = 6$) the sample mean is $\hat{\mu}_m = \$79583$ and the sample standard deviation is $S_m = \$25315$.

- For the female faculty ($n_f = 5$) the sample mean is $\hat{\mu}_f = \$74620$ and the sample standard deviation is $S_f = \$20875$.

- Our hypotheses for each test are $H_0 : \mu_m - \mu_f = 0$ and $H_a : \mu_m - \mu_f \neq 0$.

Example: The salaries of six randomly-chosen male faculty in a university's math department are $51000, $90500, $46000, $97000, $108000, $85000 and the salaries of five randomly-chosen female faculty in the same department are $56600, $55000, $104000, $70500, $87000. Test at the 20%, 11%, and 2% significance levels.

2. Test using Student's equal-variances t test whether the average salaries are equal.

Two-Sample t Tests: Confidence Intervals, IV

Example: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels.

2. Test using Student's equal-variances t test whether the average salaries are equal.

- We have $S_{\text{pool}} = \sqrt{\frac{(6-1)\cdot\$25315^2 + (5-1)\cdot\$20875^2}{6+5-2}} = \$23446$.

- The test statistic is $t = \dfrac{(\$79583 - \$74620) - 0}{\$23446 \cdot \sqrt{\frac{1}{6} + \frac{1}{5}}} = 0.3496$.

- Thus, the p-value is $2P(T_9 \geq 0.3496) = 0.7347$.

- The p-value is quite large so we fail to reject the null hypothesis in all cases.

<u>Example</u>: The salaries of six randomly-chosen male faculty in a university's math department are $51000, $90500, $46000, $97000, $108000, $85000 and the salaries of five randomly-chosen female faculty in the same department are $56600, $55000, $104000, $70500, $87000. Test at the 20%, 11%, and 2% significance levels.

  3. Test using Welch's unequal-variances t test whether the average salaries are equal.

Example: The salaries of six randomly-chosen male faculty in a university's math department are \$51000, \$90500, \$46000, \$97000, \$108000, \$85000 and the salaries of five randomly-chosen female faculty in the same department are \$56600, \$55000, \$104000, \$70500, \$87000. Test at the 20%, 11%, and 2% significance levels.

3. Test using Welch's unequal-variances t test whether the average salaries are equal.

- We have $S_{\text{unpool}} = \sqrt{\frac{\$25315^2}{6} + \frac{\$20875^2}{5}} = \$13927$.
- The test statistic is $t = \frac{(\$79583 - \$74620) - 0}{\$13927} = 0.3564$, and the number of degrees of freedom is
  $df = \frac{(\$25315^2/6 + \$20875^2/5)^2}{\frac{1}{6-1}(\$25315^2/6)^2 + \frac{1}{5-1}(\$20875^2/5)^2} = 8.9991$ giving p-value
  $P(T_{8.9991} \geq 0.3564) = 0.7298$.
- The p-value is again quite large so we fail to reject the null hypothesis in all cases.

<u>Example</u>: Female: \$56600, \$55000, \$104000, \$70500, \$87000.
Male: \$51000, \$90500, \$46000, \$97000, \$108000, \$85000.

4. Find 80% and 95% confidence intervals for the difference between the average salaries using the pooled estimate.

<u>Example</u>: Female: \$56600, \$55000, \$104000, \$70500, \$87000.
Male: \$51000, \$90500, \$46000, \$97000, \$108000, \$85000.

4. Find 80% and 95% confidence intervals for the difference between the average salaries using the pooled estimate.

- To compute the 80% and 95% confidence intervals, we need to compute the appropriate *t*-statistics.
- The difference in the average salaries is $\hat{\mu}_m - \hat{\mu}_f = \$4963$.
- For the pooled estimate, there are 9 degrees of freedom, so using a *t* table or computer yields $t_{\alpha/2,n} = 1.3830$ for the 80% confidence interval and $t_{\alpha/2,n} = 2.2622$.
- Then $S = S_{\text{pool}}\sqrt{\dfrac{1}{6} + \dfrac{1}{5}} = \$14197$, so our 80% confidence interval is $\$4963 \pm 1.3830 \cdot \$14197 = (-\$14672, \$24598)$ and our 95% confidence interval is
$\$4963 \pm 2.2622 \cdot \$14197 = (-\$27153, \$37079)$.

Example: Female: $56600, $55000, $104000, $70500, $87000.
Male: $51000, $90500, $46000, $97000, $108000, $85000.

5. Find 80% and 95% confidence intervals for the difference between the average salaries using the unpooled estimate.

Example: Female: \$56600, \$55000, \$104000, \$70500, \$87000.
Male: \$51000, \$90500, \$46000, \$97000, \$108000, \$85000.

5. Find 80% and 95% confidence intervals for the difference between the average salaries using the unpooled estimate.

- For the unpooled estimate, there are $df = 8.9991$ degrees of freedom, so using a t table or computer yields $t_{\alpha/2,n} = 1.3830$ for the 80% confidence interval and $t_{\alpha/2,n} = 2.2622$. (The degrees of freedom are so close to 9 in this case it actually doesn't matter if we just round to 9.)

- Then $S = S_{\text{unpool}} = \$13927$, so our 80% confidence interval is $\$4963 \pm 1.3830 \cdot \$13927 = (-\$14298, \$24225)$ and our 95% confidence interval is $\$4963 \pm 2.2622 \cdot \$13927 = (-\$26542, \$36469)$.

- The two estimates are quite close, since the sample variances are not far away from each other.

We make a few brief remarks about when to use these various t tests.

- Most sources still identify Student's t test as the preferred test to use when the sample variances are not far away from each other, and give various approximate rules for deciding what "far away" means (e.g., requiring the variances not to differ by a factor of more than 2).

- When the sample variances are far apart, Welch's t test tends to give more reliable results (in the sense of having lower type I and type II error probabilities). Even when the sample variances are close, Welch's t test is generally not that much worse than Student's t test (which has a higher power in the situations where it should be used).

Two-Sample t Tests: When To Do What?, II

- Neither test is exact (in the sense that it gives exact p-values) except in the case of Student's t test where the population variances are equal. In practice, this means that the type I error rate will deviate somewhat from the desired significance level α.

- Welch's t test tends to maintain a type I error rate closer to the desired significance level α than Student's t test does (although of course there are scenarios in which it is worse).

- It is also worth noting that, as the sample sizes of both groups become large, both tests are very close to the two-sample z test we have previously described. In practice, with samples larger than 100-200 or so, there is a negligible difference between the results of these t tests and the simpler z test.

Summary

We discussed Student's two-sample equal-variances t test.

We discussed Welch's two-sample unequal-variances t test.

We discussed confidence intervals for the difference in means with t statistics.

Next lecture: Matched pairs, more examples, robustness of t tests