Comparing Two Groups

Comparing Two Means

We now know how to compare two different proportions from two independent samples. In this section, we will lean how to compare two means from two independent samples or say things about the difference of those means. If we have two independent samples, so they are not matched or paired up in any way, we would observe the following sample sizes, means and variances

Sample #1: $n_1, \bar{x}_1, s_1^2$
Sample #2: $n_2, \bar{x}_2, s_2^2$

We would like to be able to say something about the distributions so that we can compute confidence intervals and do hypothesis tests. We can show that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{df}$$

To calculate the degrees of freedom in this case, we will use the Welch-Satterthwaite approximation

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

which gives us the approximate number of degrees of freedom and does not have to be integer-valued. Note that the quantity above only approximately has a t-distribution.

To compute a 95% confidence interval for the difference in true means, we have the following formula

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example 1
For the skeleton dataset, recall that we had 400 skeletons in total and each one of them had an age difference calculated as the estimated minus the actual age at the time of death using the DiGangi method. Suppose we wanted to compare the age differences depending on the gender of the skeleton.

Here are the summary statistics of the age differences by gender

Males: $n_1 = 281, \bar{x}_1 = -12.9, s_1^2 = 181.5$
Females: $n_2 = 119, \bar{x}_2 = -17.1, s_2^2 = 231.5$

This means the difference in the sample means was equal to 4.2 years, but the question is what does that say, if anything, about the difference of the true means? That is, we are interested in looking at the true age difference for male skeletons minus the true age difference for female skeletons ($\mu_1 - \mu_2$). Using the formula for the degrees of freedom above, we obtain $df = 200.09$ and using a statistical software to compute the corresponding critical value from the $t_{200.09}$ distribution we obtain $t_{0.025,200.09} = 1.97$. Thus, our 95% confidence interval is

$$\begin{aligned}
&\bar{x}_1 - \bar{x}_2 \ \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
&= -12.9 - (-17.1) \ \pm \ 1.97 \ \sqrt{\frac{181.5}{281} + \frac{231.5}{119}} \\
&= \ [1.03, 7.37]
\end{aligned}$$

We can conclude that the age difference using the DiGangi method is indeed higher on average for males by 1.03 to 7.37 years.

In hypothesis testing, we will typically want to test whether the true means of the two different samples are equal to each other or not. Thus, the null and alternative hypothesis is

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_a : \mu_1 \neq \mu_2$$

From the discussion above, we know that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{df}$$

where the degrees of freedom are calculated using the Welch-Satterthwaite approximation. We can now use this fact to compute p-values.

Example 2
Referring back to Example 1, we can test whether the average age difference for the male skeletons is equal to the average age difference for the female skeletons. That is, we would like to test

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_a : \mu_1 \neq \mu_2$$

Here are the summary statistics of the age differences by gender

The observed difference is $\bar{x}_1 - \bar{x}_2 = 4.2$ years, so the p-value will be the probability that we observe a difference in sample means at least as big as 4.2 years, in absolute value. That is

$$\begin{aligned}
p\text{-value} &= P(|\bar{X}_1 - \bar{X}_2| \geq 4.2) \qquad (\text{under } H_0) \\
&= P\left(\frac{|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \geq \frac{4.2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right) \\
&= P\left(|t_{200.09}| \geq \frac{4.2 - 0}{\sqrt{\frac{181.5}{281} + \frac{231.5}{119}}} \right) \\
&= P\left(|t_{200.09}| \geq 2.61 \right) \\
&\doteq 0.0097
\end{aligned}$$

Since the p-value is small, we can reject H_0 and conclude that the true mean age difference for the male skeletons differs from the true mean age difference for the female skeletons.

Recall in Section 6.3, we said that under the null hypothesis the two proportions are assumed to be equal and therefore we can use a pooled estimate for both population proportions. In a **'pooled' hypothesis test** for difference in means, we can use a similar method but only if we can assume that the two population variances are equal, that is $\sigma_1 = \sigma_2$. We still want to test the same null and alternative hypothesis

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_a : \mu_1 \neq \mu_2,$$

and we already know that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx t_{df}$$

If we assume that $\sigma_1 = \sigma_2$, then under H_0, the two samples have the same mean and variance. Then we can 'pool' the samples together to calculate a **pooled variance** s^2_{pooled}

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

Under the assumption of equal variances, we also have a simpler formula for the degrees of freedom for the t-distribution

$$df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2.$$

Example 3

Referring back to Example 2, and suppose we wanted to test again whether the average age difference for the male skeletons is equal to the average age difference for the female skeletons. That is, we would like to test

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_a : \mu_1 \neq \mu_2$$

Here are the summary statistics of the age differences by gender again

$$\text{Males: } n_1 = 281, \bar{x}_1 = -12.9, s_1^2 = 181.5$$
$$\text{Females: } n_2 = 119, \bar{x}_2 = -17.1, s_2^2 = 231.5$$

Now we are going to assume that the true variances are the same for both genders, in which case we can pool the variance estimator to obtain

$$s^2_{pooled} = \frac{(281-1)181.5 + (119-1)231.5}{(281-1)+(119-1)} \doteq 196.3$$

and calculate the new degrees of freedom for the t-distribution to be $df = (n_1-1)+(n_2-1) = n_1 + n_2 - 2 = 281 + 119 - 2 = 398$.

Again, the observed difference is $\bar{x}_1 - \bar{x}_2 = 4.2$ years, so the p-value will be the probability that we observe a difference in sample means at least as big as 4.2 years, in absolute value. That is

$$\begin{aligned} p\text{-value} &= P(|\bar{X}_1 - \bar{X}_2| \geq 4.2) \qquad (\text{under } H_0) \\ &= P\left(\frac{|(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)|}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \geq \frac{4.2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} \right) \\ &= P\left(|t_{398}| \geq \frac{4.2 - 0}{\sqrt{\frac{196.3}{281} + \frac{196.3}{119}}} \right) \\ &= P\left(|t_{398}| \geq 2.74 \right) \\ &\doteq 0.0064 \end{aligned}$$

This p-value is even smaller than the p-value we obtained when we did not make that assumption of equal variances, but our conclusion remains the same. Since the p-value is small, we can reject H_0 and conclude that the true mean age difference for the male skeletons differs from the true mean age difference for the female skeletons.

Now whether we are doing a pooled or an un-pooled version, we understand how to compare two different means from two independent samples. We can compute confidence intervals, we can compute hypothesis tests without pooling and without assuming the variances are equal, and we can compute hypothesis tests with pooling when we do assume that the variances are equal. In this way we understand a lot more about how to compare two different independent sample means.