Worksheet 04: Two-sample T-Test

Consider observing two different random samples from two potentially different underlying distributions. We will write this as $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{G}_X$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{G}_Y$. A common task is to determine the difference in the expected values between the two groups. We will work with the assumption that both $\mathcal{G}_X$ and $\mathcal{G}_Y$ are normal and that they have a shared common (but unknown) variance σ^2. Our concern will be determining the difference $\mu_X - \mu_Y$. The central limit theorem can be used to extend our results to the case where the distributions are not normal. In R, we will see a variant that further extends the result to where the groups have different variances.

Q01. What is the distribution of $\bar{X} - \bar{Y}$? You should be able to get an answer that is only in terms of μ_X, μ_Y, σ^2, n, and m. We know that the sample means are normal, so their difference should also be a normal distribution. All that is left is to compute the expected value and variance. We have:

$$\begin{aligned} \mathbb{E}\left[\bar{X} - \bar{Y}\right] &= \mathbb{E}\bar{X} - \mathbb{E}\bar{Y} \\ &= \mu_X - \mu_Y \end{aligned}$$

Using the fact that the samples are independent, the variance is given by:

$$\begin{aligned} \text{Var}\left[\bar{X} - \bar{Y}\right] &= \text{Var}\bar{X} + \text{Var}\bar{Y} \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{m} \end{aligned}$$

These fully characterized the distribution of $\bar{X} - \bar{Y}$.

The **pooled variance** S_p^2 is defined as the following combination of the sample variance of X and Y:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

If we divide both sides by $n+m-2$ and σ^2, we get a more clear view of how these three sample variances relate to one another:

$$\frac{(n+m-2)S_p^2}{\sigma^2} = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_p^2}{\sigma^2}$$

Q02. What is the distribution of the left-hand side of the equation above? Notice that this does not depend on the expected value of $\mathcal{G}_X$ and $\mathcal{G}_Y$ being the same. We know that right-hand side are two independent chi-squared distributions of $n-1$ and $m-1$ degrees of freedom, respectively. So, the left-hand side must be a chi-squared distribution with $(n-1)+(m-1) = n+m-2$ degrees of freedom.

Q03. Take the expected value of the left-hand side of the previous equation and show that the expected value of S_p^2 is σ^2. This is very straightforward from the previous result:

$$\begin{aligned} \mathbb{E}\left[\frac{(n+m-2)S_p^2}{\sigma^2}\right] &= n+m-2 \\ \mathbb{E}S_p^2 &= \sigma^2 \end{aligned}$$

Putting this all together, we see that the following random variable is the ratio between a standard normal and a chi-squared distribution with $n+m-2$ degrees of freedom divided by its degrees of freedom. In other words, the following should have a T distribution with $n+m-2$ degrees of freedom:

$$T = \frac{\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}}{\sqrt{\frac{(n+m-2)S_p^2}{\sigma^2}\cdot\frac{1}{n+m-2}}} = \frac{\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\sqrt{S_p^2}} = \frac{\bar{X}-\bar{Y}}{S_p\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}.$$

We can use this as a test statistic for an hypothesis test or as the basis of a confidence interval for the difference in means between two groups.

Let's end with an example. Consider a larger replication study of the potato diet. Here, instead of trying to see if the potato diet does result in statistically significant weight loss, we want to investigate whether there is a difference between a diet based on potatoes or a diet based on only eating pasta. We have a dataset with $n = 55$ samples from a potato diet, with a sample mean of $\mu_X = 4.5$ and sample variance $S_X^2 = 2.2$. We also have a set of $m = 28$ participants eating only pasta, with a sample mean of $\mu_Y = 2.3$ and $S_Y^2 = 2.5$. **Q04.** Compute the T statistic to test the hypothesis that there is no difference in the weight loss between these two diets. We have:

$$S_p^2 = \frac{(54)\cdot 2.2 + (27)\cdot 2.5}{54+27} = 2.3$$

And then:

$$T = \frac{4.5-2.3}{\sqrt{2.3}\cdot\sqrt{\frac{1}{54}+\frac{1}{27}}} = 2.79$$

Q05. A T distribution with 81 degrees of freedom can be approximated well by a standard normal distribution. Does your result above have a p-value that is more or less than 0.01? The cut-off for would be 2.58, so the p-value should be less than 0.01.