Handout 04: Hypothesis Testing

An hypothesis test is a way of establishing whether a random sample offers support for a particular conclusion about the population distribution $\mathcal{G}$. The mechanics of hypothesis testing are similar to those confidence intervals, but the interpretation and terminology differs. For an hypothesis test, we start with a **null hypothesis** H_0 and see if the data support rejecting this baseline assumption in favor of an **alternative hypothesis** H_A. A common example is to have a null hypothesis in which some unknown parameter θ is assumed to be zero and we want to see if the data support the alternative hypothesis that $\theta \neq 0$.

In its formal specification, to do an hypothesis test we first select a **significant level** $(1 - \alpha)$ and construct an event R, called a **rejection region**, such that $\mathbb{P}[R^2 | H_0] \leq \alpha$. If we then observe some data and R occurs, we **reject** the null hypothesis in favor of the alternative. Otherwise, we **retain** (or fail to reject) the null hypothesis. The goal is to create a rejection region that has a much higher probability of occuring under the alternative hypothesis than it does under the null.

The standard approach to creating a rejection region is to start with a pivot, plug in the parameters of the null distribution, and then form a **test statistic** with a known distribution under H_0. Then, we can create something similar to a confidence interval around the test statistic and use the complement of that as the rejection region. Let's see an example! Assume that you have a random sample from a normal distribution with unknown mean and unknown variance. We will do an hypothesis test with $H_0 : \mu = 1$ and $H_A : \mu \neq 1$. We form a test statistic T using the pivot quantity that we saw last time, but with μ replaced with the null hypothesis of 1:[1]

$$T = \frac{\bar{X} - 1}{\sqrt{S_X^2 / n}}$$

[1] I have reversed the numerator from the previous notes to follow the standard convention that a positive T corresponds to a mean higher than the null hypothesis. Due to the symmetry of the t-distribution, T still has the same t-distribution either way.

The rejection region R is given by the following event:

$$R = \left\{ t_{\alpha/2} \leq T \leq t_{1-\alpha/2} \right\}^c .$$

Due to the way it is defined, this will have the desired property that $\mathbb{P}[R | H_0] \leq 1 - \alpha$. A common shorthand for the rejection region is to call the quantity $t_{1-\alpha/2}$ a **critical value**, and then simply reject the null hypothesis if the (absolue value) of the test statistic exceeds the critical value. Hypothesis test procedures are oftened named based on the distribution of the test statistic.[2] Here, we have derived what is called the **one-sample T-test**.

[2] This is not an ideal convention because many different tests can have the same test statistic distribution. The naming convention causes constant confusion. We can talk more about ways to mitigate these issues in your own work.

There is another related way of doing hypothesis testing that, at least initially, forgoes the language about rejecting or retaining the null hypothesis. Instead of having a fixed critical value, we could ask the question: what is the smallest value of α such that corresponding test statistic would be equal to (or greater than) the critical value? This quantity is called a **p-value**. A p-value is usually the form that statsitical software will report the results of an hypothesis test and is increasingly the way that results are communicated through. Note that we can recover the concept of statistical significance by simply checking the p-value is less than the significant level. If it is, we have a statistically significant result and would reject the null hypothesis in favor of the alternative hypothesis.

We have already fully derived the one-sample T-test. On today's worksheet, we will introduce the **two-sample T-test**. The setup and final results are repeated here for easy reference. 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 $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{G}_Y$. We will assume that both $\mathcal{G}_X$ and $\mathcal{G}_Y$ are normal and that they have a shared common (but unknown) variance σ^2. We want to produce an hypothesis test that the difference in means $\theta = \mu_X - \mu_Y$ is equal to some fixed value θ_0 (typically zero). First, we define the pooled sample variance as follows:

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

Then, the following is a valid pivot:

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

With the null value of θ plugged in, we get a valid test statistic. The corresponding confidence inverval for the difference means is:

$$(\bar{X} - \bar{Y}) \pm t_{1-\alpha/2} \cdot \sqrt{\frac{S_p^2}{\frac{1}{n} + \frac{1}{m}}}$$

The central limit theorem can be used to extend this result to the case where the distributions are not normal. In R, we will see a variant that further extends this to the situation where the groups have different variances.