5.3 Hypothesis Testing

Let's return to the example of determining voter preference in the 2016 presidential election. Suppose we suspected that the proportion of voters who preferred Hillary Clinton was greater than $\frac{1}{2}$ (indeed, it was!), and that we took n samples, denoted $\{X_i\}_{i=1}^n$ from the U.S. population. Based on these samples, can we support or reject our hypothesis that Hillary Clinton is more popular? And how confident are we in our conclusion? Hypothesis testing is the perfect tool to help answer these questions.

5.3.1 Constructing a Test

A hypothesis in this context is a statement about a parameter. In the presidential election example, the parameter of interest was p, the proportion of the population who supported Hillary Clinton. A hypothesis could then be that $p > 0.5$, i.e. that more than half of the population supports Hillary.

There are four major components to a hypothesis test.

1. The *alternative hypothesis*, denoted H_a, is a claim we would like to support. In our previous example, the alternative hypothesis was $p > 0.5$.

2. The *null hypothesis*, denoted H_0 is the opposite of the alternative hypothesis. In this case, the null hypothesis is $p \leq 0.5$, i.e. that less than half of the population supports Hillary.

3. The *test statistic* is a function of the sample observations. Based on the test statistic, we will either accept or reject the null hypothesis. In the election example, the test statistic was the sample mean $\bar{X}$. The sample mean is often the test statistic for many hypothesis tests.

4. The *rejection region* is a subset of our sample space Ω that determines whether or not to reject the null hypothesis. If the test statistic falls in the rejection region, then we reject the null hypothesis. Otherwise, we accept it. In the presidential election example, the rejection region would be

$$\text{RR: } \{(x_1, \ldots, x_n) : \bar{X} > k\}$$

   This notation means we reject if $\bar{X}$ falls in the interval (k, ∞), where k is some number which we must determine. k is determined by the Type I error, which is defined in the next section. Once k is computed, we reject or accept the null hypothesis depending on the value of our test statistic, and our test will be complete.

5.3.2 Types of Error

There are two fundamental types of errors in hypothesis testing. They are denoted Type I and II error.

Definition 5.2. *A **Type I error** is made when we reject H_0 when it is in fact true. The probability of Type I error is typically denoted as α.*

In other words, α is the probability of a false positive (rejecting H_0 when it is true is the same as accepting H_a when it is false).

Definition 5.3. *A **Type II error** is made when we accept H_0 when it is in fact false. The probability of Type II error is typically denoted as β.*

In other words, β is the probability of a false negative (accepting H_0 when it is false is the same as rejecting H_a when it is true).

In the context of hypothesis testing, α will determine the rejection region. If we restrict the probability of a false positive to be less than 0.05, then we have

$$P(\bar{X} \in \text{RR} \mid H_0) \leq 0.05$$

i.e. our test statistic falls in the rejection region (meaning we reject H_0), given that H_0 is true, with probability less than 0.05.

Notation: The notation $P(\bar{X} \in \text{RR} \mid H_0)$ refers to the probability that $\bar{X} \in \text{RR}$ given that H_0 is true.

Continuing along our example of the presidential election, the rejection region was of the form $\bar{X} > k$, and the null hypothesis was that $p \leq 0.5$. Our above expression then becomes

$$P(\bar{X} > k \mid p \leq 0.5) \leq 0.05$$

If $n > 30$, we can apply the CLT to say,

$$P(\frac{\bar{X} - p}{S/\sqrt{n}} > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5) = P(Y > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5)$$

where Y is a $N(0, 1)$ random variable. Since $p \leq 0.5$ implies $\frac{k-p}{S/\sqrt{n}} \geq \frac{k-0.5}{S/\sqrt{n}}$, we must also have

$$Y > \frac{k - p}{S/\sqrt{n}} \Rightarrow Y > \frac{k - 0.5}{S/\sqrt{n}}$$

Hence,

$$P(Y > \frac{k - p}{S/\sqrt{n}} \mid p \leq 0.5) \leq P(Y > \frac{k - 0.5}{S/\sqrt{n}})$$

So if we bound the probability on the right side of the inequality by 0.05, then we also bound the probability on the left (the Type I error, α) by 0.05. Since Y is distributed $N(0,1)$, we can look up a z table to find that $z_{0.05} = -1.64$, so

$$P(Y > 1.64) = P(Y < -1.64) = 0.05$$

Letting $\frac{k-0.5}{S/\sqrt{n}} = 1.64$, we can solve for k to determine our rejection region. Rearranging to get k on one side of the equation, we find that

$$k = 0.5 + 1.64 \cdot \frac{S}{\sqrt{n}}.$$

Since our rejection region was of the form $\bar{X} > k$, we simply check whether or not $\bar{X} > 0.5 + 1.64 \cdot \frac{S}{\sqrt{n}}$. If this is true, then we reject the null, and conclude that more than half the population favors Hillary Clinton. Since we set $\alpha = 0.05$, we are at least $1 - \alpha = 0.95$ confident in our conclusion. This would complete the hypothesis test (we can't actually reject or accept in this text since we haven't taken any real examples).

In the above example, we determined the rejection region by plugging in 0.5 for p, even though the null hypothesis was $p \leq 0.5$. It is almost as though our null hypothesis was $H_0 : p = 0.5$ instead of $H_0 : p \leq 0.5$. In general, we can simplify H_0 and assume the border case ($p = 0.5$ in this case) when we are determining the rejection region.

5.3.3 p-Values

As we saw in the previous section, a particular α determined the rejection region so that the probability of a false positive was less than α. Now suppose we observe some test statistic, say, the sample proportion of voters $\bar{X}$ who prefer Hillary Clinton. We then ask the following question. Given $\bar{X}$, what is the smallest value of α such that we still reject the null hypothesis? This leads us to the following definition.

Definition 5.4. *The p-value, denoted p, is defined*

$$p = \min\{\alpha \in (0,1) : \textit{Reject } H_0 \textit{ using an } \alpha \textit{ level test}\}$$

i.e. the smallest value of α for which we still reject the null hypothesis.

This definition isn't that useful for computing p-values. In fact, there is a more intuitive way of thinking about them. Suppose we observe some sample mean $\bar{X}_1$. Now suppose we draw a new sample mean, $\bar{X}_2$. The p-value is just the probability that our new sample mean is more *extreme* than the one we first observed, assuming the null hypothesis is true. By "extreme" we mean, more different from our null hypothesis.

Below we go through an example which verifies that the intuitive definition given above agrees with Definition 5.4.

Example 5.5. Suppose that we sampled n people and asked which candidate they preferred. As we did before, we can represent each person as an indicator function,

$$X_i = \begin{cases} 1 & \text{if person } i \text{ prefers Hillary} \\ 0 & \text{otherwise.} \end{cases}$$

Then $\bar{X}$ is the proportion of the sample that prefers Hillary. After taking the n samples, suppose we observe that $\bar{X} = 0.7$. If we were to set up a hypothesis test, our hypotheses, test statistic, and rejection region would be

$$\begin{aligned} &H_0 : q \leq 0.5 \\ &H_a : q > 0.5 \\ &\text{Test statistic: } \bar{X} \\ &\text{RR: } \{(x_1, \ldots, x_n) : \bar{X} > k\} \end{aligned}$$

where q is the true proportion of the entire U.S. population that favors Hillary. Using the intuitive definition, the p value is the probability that we observe something more extreme than 0.7. Since the null hypothesis is that $q \leq 0.5$, "more extreme" in this case means, "bigger than 0.7". So the p-value is the probability that, given a new sample, we observe that the new $\bar{X}$ is greater than 0.7, assuming the null hypothesis, i.e. that $q \leq 0.5$. Normalizing $\bar{X}$, we have

$$\begin{aligned} P(\bar{X} > 0.7 \mid H_0) &= P\Big(\frac{\bar{X} - 0.5}{S/\sqrt{n}} > \frac{0.7 - 0.5}{S/\sqrt{n}} \mid H_0\Big) \\ &\approx P\Big(Y > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big) \doteq p \end{aligned} \tag{1}$$

where $Y \sim N(0, 1)$. We would then compute the value $z_p \doteq \frac{0.7-0.5}{S/\sqrt{n}}$ by plugging in the sample standard deviation, S, and the number of samples we took, n. We would then look up a z table and find the probability corresponding to z_p, denoted p (this is our p value).

We now claim that this p is equal to the smallest α for which we reject the null hypothesis, i.e. that our intuitive definition of a p-value agrees with Definition 5.4. To show that

$$p = \min\{\alpha \in (0, 1) : \text{Reject } H_0 \text{ using an } \alpha \text{ level test}\},$$

i.e. that p is the smallest α for which we reject H_0, we need to show that for any $\alpha \leq p$, we accept the null hypothesis. We also need to show that for any $\alpha > p$, we reject the null hypothesis.

Case 1: Suppose $\alpha \leq p$. We need to show that the test statistic $\bar{X} = 0.7$ doesn't fall into the rejection region determined by α. Using a z table, we could find z_α such that

$$\alpha = P(Y > z_\alpha) \approx P(\frac{\bar{X} - 0.5}{S/\sqrt{n}} > z_\alpha \mid H_0) = P(\bar{X} > z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 \mid H_0)$$

Since the RHS of the above expression is the probability of Type I error, the rejection region is determined by

$$\text{RR: } \{(x_1, \ldots, x_n) : \bar{X} > k_\alpha\} \qquad \text{where } k_\alpha \doteq z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5$$

Since $\alpha \leq p$, the corresponding z_p such that $p = P(Y > z_p)$ also satisfies $z_p \leq z_\alpha$. By expression (1),

$$p = P\Big(Y > \frac{0.7 - 0.5}{S/\sqrt{n}}\Big)$$

which implies $z_p = \frac{0.7-0.5}{S/\sqrt{n}} \Rightarrow z_p \cdot \frac{S}{\sqrt{n}} + 0.5 = 0.7$. This implies that

$$0.7 = z_p \cdot \frac{S}{\sqrt{n}} + 0.5 \leq z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 = k_\alpha.$$

Therefore $\bar{X} = 0.7 \leq k_\alpha$ implies $\bar{X} = 0.7$ is in the acceptance region determined by α. Hence, we accept the null hypothesis for any $\alpha \leq p$.

Case 2: Suppose $\alpha > p$. We need to show that the test statistic $\bar{X} = 0.7$ falls in the rejection region determined by α. By reasoning similar to the kind in Case 1, we would have $z_\alpha < z_p$. This implies

$$k_\alpha \doteq z_\alpha \cdot \frac{S}{\sqrt{n}} + 0.5 < z_p \cdot \frac{S}{\sqrt{n}} + 0.5 = 0.7$$

Hence $\bar{X} = 0.7 > k_\alpha$ implies that the outcome $\bar{X} = 0.7$ is in the rejection region determined by α. Hence, we reject the null hypothesis for any $\alpha > p$.

Example 5.5 (above) justifies the definition of p-values which gives an easy way to compute them. Given some observation of our test statistic $\bar{X}$, we compute the p-value by calculating the probability of seeing something *more* different or "extreme" than our observed $\bar{X}$, assuming H_0 is true. By the argument in Example 5.5, this value is the same as the smallest α level for which we reject H_0.