Hypothesis Testing

March 23, 2020

1 Hypothesis Testing

We now take a different approach to statistical analysis. Instead of trying to estimate a parameter, we will begin with a hypothesis to test. Often times these 2 ideas are related. For example, if we have an unknown parameter, θ, we can hypothesize that θ lies in a certain interval $[a,b]$. Then we run a series of experiments and record the results, to see if they support our hypothesis. We will see that mathematically this is similar to checking if our estimator for θ lies in the interval.

When we make an observation that supports our hypothesis, there will be two ways this could happen. The first is our hypothesis is correct and the second is our hypothesis is false but the noise and inherent randomness of the measurement happened to cause a recording near our hypothesis. We would like to choose a test that minimizes the second possibility.

We now begin to make this more precise by stating the definition of a statistical test:

DEFINITION 1.1. *A statistical test consists of the following 4 elements:*

- *A **Null Hypothesis** - H_0 - This is corresponds to the uninteresting case, where our Hypothesis is false*
- *An **Alternative Hypothesis** - H_a -This is corresponds to the interesting case, where our Hypothesis is true*
- *A **Test Statistic** - Just like in the last section a statistic is a function of our measurement data.*
- ***Rejection Region** - RR - If our test statistic lands in the region, we reject the Null Hypothesis, in favor of our Alternative Hypothesis*

Let's look at an example:

EXAMPLE 1.2. *You suppose that a coin is unfair, that is the probability of being heads on a flip is not 1/2. In order to test this hypothesis you design the following experiment:*
You flip the coin 100 times and let $X_i = 1$ if the i^{th} flip is a heads and 0 otherwise. Let p be the probability that $X_1 = 1$.
Let H_0, the null hypothesis, be that $p = .5$, the coin is fair.
Let H_a, the alternative hypothesis, be that $p \neq .5$, the coin is not fair.
Let $\sum_{i=1}^{100} X_i$ be the test statistic.
Let $|\sum_{i=1}^{100} X_i - 50| > 10$ be the Rejection Region (RR).

In the above example the rejection region was chosen somewhat arbitrarily, later we'll learn how to use the distribution of the test statistic to choose a rejection region with the desired properties. But we need to account for the fact that even when H_a is true, we don't expect to see exactly 50 heads, but it is very unlikely to see much more or less than 50 heads. Therefore, if we observe a very large or small number of heads, we can confidently reject the null hypothesis.

RR can be "one" or "two"-sided. In the two-sided case (as above), H_a will be of the form $\theta \neq \theta_0$. In the one-sided case H_a will be of the form $\theta > \theta_0$ or of the form $\theta < \theta_0$.

Associated to a RR, we have the possibility of misclassification, rejecting H_0 when it is true, or failing to reject H_0 when it is false. We call these types of errors type I and II.

DEFINITION 1.3. *A **type I error** is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted α. The value of α is called the **level** of the test.*
*A **type II error** is made if H_0 is accepted when H_a is true. The probability of a type II error is denoted β.*

The names are admittedly not very creative, but are quite standard. We mostly focus on type I errors, because when you reject H_0, you are concluding the interesting thing happened. Note that we speak of accepting or rejecting H_0, not H_a.

The probabilities of these errors is computed from the distribution of the test statistic under the various situations.

EXAMPLE 1.4. *In the above example, compute the probability of a Type I error in the above example.*

Solution: Since we are computing a Type I error, we are for a moment assuming the null hypothesis is true, that $p = 1/2$. In this case, the test statistic, $\sum_{i=1}^{100} X_i$, is a Binomial random variable with $n = 100$ and $p = 1/2$. We want to compute the probability H_0 is rejected, in other words, the probability the test statistic lands in the rejection region. So

$$\alpha = \sum_{k>60} \binom{100}{k} .5^k (1-.5)^{100-k} + \sum_{k<40} \binom{100}{k} .5^k (1-.5)^{100-k}.$$

Note that we can't compute the probability of a Type II error without making additional assumptions, in H_a, p takes on a range of values. For computational purposes we can assume p takes on a specific value that satisfies H_a. More generally, there exist Theorems controlling the worst case scenario.

EXAMPLE 1.5. *Assuming $p = .3$ in the above example, compute the probability of a Type II error.*

Solution: This is similar to above but now assume $p = .3$ and compute the probability that our statistic lies in the complement of the previous set. The test statistic, $\sum_{i=1}^{100} X_i$, is now a Binomial random variable with $n = 100$ and $p = .3$. We want to compute the probability we fail to reject H_0, in other words, the probability the test statistic lands outside the rejection region. So

$$\alpha = \sum_{40 \leq k \leq 60} \binom{100}{k} .3^k (1-.3)^{100-k}.$$

One more example:

Example 1.6. *We wish to test the null hypothesis, H_0, that the proportion p of problems in a homework set with errors is equal to .05 versus the alternative H_a, that the proportion is larger than .05. We use the following test: randomly choose 2 problems and count the number with errors, if both are error free, we reject H_0. If one or more contains an error, we randomly choose a third problem, if the third problem is error free, we reject H_0. In all other cases we accept H_0 (later we'll see that it is better to say fail to reject H_0).*

What is the probability of a type I error?
What is the probability of a type II error, as a function of p?

Assume that number of problems is large.

Solution: Since the number of problems is large, we can model the probability that each randomly chosen problem is wrong to be p, independently.

For this problem a type I error, is when $p = .05$ (%5 percent of the problems have an error), but we reject H_0. We reject H_0 when either: both of the first two problems don't have an error or the third problem doesn't have an error. Since these two events are mutually exclusive (at most one of them can occur) we compute the probability of each event separately and add the probabilities together. The probability of this event (using that $p = .05$) is:

$$\alpha = \binom{2}{0} .95^2 + (1 - \binom{2}{0} .95^2) .95$$

The first term is the probability both pages don't have an error, the second is probability of at least one error (computed as 1 minus probability of no errors) times the probability of an no errors on the third problem. This is a relatively high number because this is not a very extensive test.

For this problem a type II error, is when $p > 0$ but we accept H_0 (or really fail to reject H_0). This happens if $p > .05$, but either the first two problems have an error and the third problem has an error. We can compute this probability similarly to in the previous part.

$$\beta = \left(1 - \binom{2}{0} (1-p)^2\right) p$$

Admittedly, this last example was not a very good statistical test, but it is nevertheless a statistical test, as it has an H_0, H_a, a test statistic and a rejection region. In the next sections we investigate how to get and use better tests.