Chapter 6 – Intro to Hypothesis Testing

In this chapter, we'll discuss the fundamentals of Hypothesis Testing (HT) using the Sign, Binomial Exact, and Fisher's Exact Tests. *Hypothesis Testing is one of the real strengths of Statistics.*

HT's entail our following six steps (p.173):

- **Giving the *Null Hypothesis* (H_0)**
- **Giving the *Alternative Hypothesis* (H_A)**
- **Setting the *level of significance* (α) *or Decision Rule****
- **Calculating the *Test Statistic* (TS)**
- **Finding the *p-value***
- **Making our decision (in plain English)**

In Statistics, when we decide which *testing methodology* is 'better', we usually base it on *statistical power*: the winner is the method which is 'most powerful', if one exists.

Whenever we make a decision, we can make two types of errors:

- **H_0 is true and we declare that it is false (*Type I error*)**
- **H_0 is false (so that is H_A true) but we decide to retain it (*Type II error*)**

The *courtroom analogy* is the presumption of innocence of the defendant, so we set H_0: the defendant is innocent. The alternative hypothesis (H_A) is that s/he is guilty. Committing a Type I error is declaring this innocent defendant guilty; committing a Type II error is letting a guilty defendant go free (declaring him/her not guilty).

The level of significance = α = Pr{Type I error} is set – and set low – since it is usually the more serious of the two mistakes. Also, we define β to be Pr{Type II error} and so the power is $1 - \beta$. Note that the power is the probability of rejecting H_0 when it is in fact false (i.e., when H_A is true).

Example 6.2 (p.166) Medical Treatments. Treatment vs. Placebo, n = 12 pairs of subjects – one randomized to Treatment, the other to Placebo. Let π = the population proportion of pairs in which the treated subject does better (health-wise) than the placebo subject.

H_0: $\pi = \frac{1}{2}$
H_A: $\pi > \frac{1}{2}$ (treatment is better than placebo)

Let K denote the number of the 12 pairs in which the MD declares that treated subject is healthier than the placebo subject. Suppose we adopt the *decision rule* that we reject H_0 if $K \geq 10$, and retain H_0 otherwise (i.e., $K \leq 9$). A Binomial calculation then shows that α then equals 0.0193. (Usually, we choose $\alpha = 0.05$ or 0.01, but that is not possible here.) So now we just look at our data (the observed value of K) and make our decision on whether the new treatment is better.

Calculating power is not straightforward in this example since it is the probability of rejecting H_0 when H_A is true. This is challenging since if H_A is true then we only know that $\pi > \frac{1}{2}$, but we don't know what value π is equal to. In this case, we get a whole power curve (bottom right of p.167), where power is on the Y-axis and π (above ½) is on the X-axis. Note that this curve starts at a height of $\alpha = 0.0193$ for $\pi = \frac{1}{2}$ and rises to 1 for $\pi = 1$. For example, the power "at $\pi = \frac{3}{4}$" is

$$\Pr\{K \geq 10 \text{ given } \pi = \tfrac{3}{4}\}$$
$$= 66\,(0.75)^{10}\,(0.25)^{2} + 12\,(0.75)^{11}\,(0.25)^{1} + (0.75)^{12}\,(0.25)^{0} = 0.390675;$$

this is the same value indicated in Figure 6.3 on p.167.

(A similar example is given on p.168.)

But is the idea of using K (the count of the number of times that the health of the treated subject exceeds that of the placebo subject) as a testing method a good or optimal (highest power) one?

We'll next cover section 6.3, but skip 6.4, 6.5, and 6.6 for now.

The Sign Test – is a test related to the population median, and it can relate to one measurement per subject (Ex. 6.4 on p.170) or two measurements (before and after) per subject (Ex. 6.6 on p.174).

Example 6.4 Heights of Men –

Notice that in contrast with the above, no decision rule is set here – rather, we set an *α-level*, that is, an acceptable level for the probability of a Type I error. Usually, we set $\alpha = 5\%$ (a priori), but that's not set in stone. We're testing here whether the median height has increased (in the alternative hypothesis), and this is translated again into:

$$H_0: \pi = \tfrac{1}{2} \text{ versus } \quad H_A: \pi > \tfrac{1}{2}$$

where π is the proportion of men whose weights exceed 69.5 (the old median). The 2 ties are omitted, so n = 18, and our test statistic is $k_s = 14$. The p-value is then calculated to be p = 0.0154. This is then usually reported, and since the p-value is $< \alpha$, we conclude that it appears that the median height of men has increased since 1970.

Example 6.6 Weight Gains of Women –

We can also use the Sign Test to test in a paired sample of data (as in a before and after study). If there is no change, then we expect there to be as many +'s (increases in weight) as –'s (decreases in weight). So, we again use the Binomial Distribution to find the p-value. Again, we are testing here

$$H_0: \pi = \tfrac{1}{2} \text{ versus } \quad H_A: \pi > \tfrac{1}{2}$$

since we are looking for an *increase* in women's weight. Tossing out the one tie gives n = 19, and here, our test statistic is $k_s = 17$. The p-value is then calculated to be p = 0.0004. We conclude here that there is strong evidence that the median weight of women has increased 10 years after pregnancy.

Homework due next Tuesday on Chapters 6 and 7.
Thursday we will cover Chapter 7.