8.1 Tests of Hypotheses Revisited

As discussed earlier, a standard practice in research consists of:

1. Select null and alternative hypotheses.
2. Specify the significance level of the test, \( \alpha \). (Remember that the significance level is the probability the test rejects a true null hypothesis.)
3. Collect and analyze data; decide whether or not to reject the null.

In this chapter we will consider the following two questions. In practice, the first question is hugely more applicable than the second.

1. Just because we fail to reject the null, how good should we feel about continuing to assume it is true?
2. Just because we reject the null, does it really mean the null is no good?

Rather than try to explore these questions in a general mathematical way, I will begin with several examples.

To fix ideas, consider our Fisher’s test to investigate whether \( p \) is constant in a sequence of trials that might be BT. Suppose we obtain the following data.

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
<th>( \hat{p} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Half</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0.80</td>
</tr>
<tr>
<td>Second Half</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>0.20</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

The P-value for Fisher’s test is 0.2063 and with any popular choice of \( \alpha \) the decision would be to fail to reject. But note that the \( \hat{p} \)’s are very different! Why did we get such a large P-value when the \( \hat{p} \)’s are so different? Well, because we do not have much data. In fact, if the amount of data is small, it can be difficult or impossible to reject the null. For example, the following table gives the largest possible difference between \( \hat{p} \)’s, yet the P-value is 0.1000; too large to reject for \( \alpha < 0.10 \).
### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
<th>(\hat{p})</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Half</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.00</td>
</tr>
<tr>
<td>Second Half</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

At the other extreme (and note that these data are a bit silly in practice), consider the following table.

### Table 3

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
<th>(\hat{p})</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Half</td>
<td>50,500</td>
<td>49,500</td>
<td>100,000</td>
<td>0.505</td>
</tr>
<tr>
<td>Second Half</td>
<td>50,000</td>
<td>50,000</td>
<td>100,000</td>
<td>0.500</td>
</tr>
<tr>
<td>Total</td>
<td>100,500</td>
<td>99,500</td>
<td>200,000</td>
<td></td>
</tr>
</tbody>
</table>

This table gives a P-value of 0.0256, which would lead to rejecting the null for \(\alpha = 0.05\). But I cannot think of any scientific problem for which I would want to conclude that \(p\) has changed! The \(\hat{p}\)'s are so very close that I believe the assumption of constant \(p\) would be scientifically useful.

So, what can we learn from these examples? Well, it will help to introduce some notation. It is standard practice that if the P-value is \(\leq 0.05\) one says that the data are statistically significant. The flaw/weakness with the concept of statistical significance is that it pays no attention to whether, as in the above examples, the \(\hat{p}\)'s are close to each other or wildly different. This is because statistical significance is an objective mathematical summary of the data; mathematics is not designed to deal with issues of ‘closeness;’ at least, not in the test of hypotheses.

So, what should we do? Many people find it useful to introduce the notion of practical significance. Here is the idea. Forget the P-value and just look at the data; in our examples above, this means to look at the two \(\hat{p}\)'s. Imagine that they are exactly equal to the underlying \(p\)'s. Then ask yourself, “Is this difference in \(p\)'s important scientifically? If the answer is yes, then the data are of practical significance; if the answer is no, then the data are not of practical significance.

Think about this for a moment. The researcher or consumer of the research results makes a subjective interpretation of the data. Is it any wonder that some Math people don’t like this? If one’s view of Math is that every question has a unique correct answer, then one is not going to be happy with this chapter!

So, let’s review the above examples. Only Table 3 has statistical significance and only Table 3 fails to achieve (IMHO) practical significance. I have always found the following display to be useful for me.

<table>
<thead>
<tr>
<th>Practical Significance</th>
<th>Statistical Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>A</td>
</tr>
<tr>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>Yes</td>
<td>C</td>
</tr>
<tr>
<td>No</td>
<td>D</td>
</tr>
</tbody>
</table>

I will now discuss the meanings of A–D.

- **A:** This is a happy situation if you want to reject the null. The data are statistically significant and the departure from the null is deemed to be of practical importance.
• B: This is a very common situation in practice. The pattern in the data looks interesting, but it might not be real. This is incredibly frustrating to the scientist. It is desirable to have more data.

• C: This occurs rarely in practice. The pattern in the data is not interesting, but you have so much data that even this dull departure from the null is statistically significant.

• D: This is a happy situation if you prefer the null. The pattern in the data does not look interesting and it can reasonably be attributed to chance variation.

In my earlier phony data sets, Tables 1 and 2 are examples of B, while Table 3 is an example of C.

8.2 Statistical Power

We have a difficulty. In order to appreciate statistical Power we need a real scientific problem and we need some expertise about the problem. Also, we cannot really do this with data already collected; Power is about planning a study. As a result, I will give you a simple, yet highly artificial, example.

Bob has a work-out schedule that involves running on a treadmill for 30 minutes and shooting free throws. He needs to decide whether to run first or shoot first. Bob wants to ‘look good’ while shooting; i.e. the higher the proportion of free throws that he makes, the happier he is.

Bob has a bit of a tight schedule and can devote only two days to study. So, one day he will run and then shoot \( n = 20 \) free throws; the other day he will shoot \( n = 20 \) free throws and then run. He wonders whether this whole idea is a waste of time. Statistical Power will help him to decide whether it is a waste of his time.

Bob thinks, “Suppose that under one condition my true ability is \( p = 0.60 \) and under the other condition my true ability is \( p = 0.40 \). Will I be able to detect this difference with my test?” It is not surprising, but neither is it obvious, that Fisher’s test can be used to analyze the data that Bob will obtain. (We will learn about this later; trust me for now.)

As usual in this course, Bob’s question above cannot be answered with a simple yes or no. The test might detect the difference (i.e. it might reject the null that the two conditions have the same \( p \)) or it might not (it might fail to reject; i.e. it might make a type 2 error). In other words, there is a probability that the test will detect the difference and this probability is called “The Power of the test when the \( p \)’s are 0.40 and 0.60.”

It is, in general, difficult to calculate the Power, so I usually resort to a simulation study. I performed a simulation study with 100 runs. Each run consisted of 20 BT with \( p = 0.60 \) and 20 BT with \( p = 0.40 \). The resulting data are analyzed with Fisher’s test. If I obtain a P-value of 0.05 or smaller, I reject the null. In my 100 runs, the null was rejected a total of 16 times. Thus, the estimated Power is 0.16.

Now, to me, for \( p \) to change by 0.20 is a huge amount for a free throw shooter. What we have learned is that even if the \( p \)’s differ by this (unrealistically?) huge amount, the test has only a 16% chance of detecting it! It seems to me that performing this study is likely to be a waste of time because it is almost preordained that the test is not going to reject the null.
I repeated the above simulation study—100 runs, 20 shots on each condition—with the two
p’s equal to 0.70 and 0.30. This time, 66 data sets led to rejection, giving an estimated Power of
0.66. Finally, I repeated the simulation study with the two p’s equal to 0.80 and 0.20. This time,
my estimated Power is 0.97.

It is convenient to end this chapter with some notation that we will use later in the course.

For a fixed test of hypotheses, let TS denote the test statistic and let CR denote the critical
region. For example, for the Chi-Squared Goodness of Fit test, the TS is $X^2$ which has observed
value $\chi^2$. The CR is $X^2 \geq \chi^2_\alpha(k - 1)$.

We have defined the significance level as the probability of rejecting the null hypothesis given
that it is true. It is convenient and clarifying to include all of this in the probability statement,
which we do as follows:

$$\alpha = P(TS \text{ is in CR} \mid H_0 \text{ is true}).$$

Recall that within a probability statement, the symbol ‘|’ is read, “given that.”

Recall that a type 2 error is the failure to reject a false null. It is often denoted by the Greek
letter beta: $\beta$. Thus,

$$\beta = P(TS \text{ is not in CR} \mid H_0 \text{ is not true}). \quad (8.1)$$

The Power is

$$\text{Power} = P(TS \text{ is in CR} \mid H_0 \text{ is not true}). \quad (8.2)$$

We do not have a symbol for ‘Power’ because:

$$\text{Power} + \beta = P(TS \text{ is in CR} \mid H_0 \text{ is not true}) + P(TS \text{ is not in CR} \mid H_0 \text{ is not true}) = 1.$$ 

Thus, ‘Power’ is simply $(1 - \beta)$.

It is important to note that Equations 8.1 and 8.2 are deliberately and greatly simplified. If you
think back to my artificial example of shooting free throws before or after running, there are many
ways in which the null can be false and each way can and usually does have a different Power. For
example, we saw three very different estimated Powers for three different ways the null could be
false.

Because Power will be mostly a side issue in this course, we will not bother to define more
precise expressions than Equations 8.1 and 8.2. Just remember that Power and $\beta$ are always a
function of the particular way the null is false, even though our notation ignores this fact.