Chapter 9

Hypothesis testing

In this chapter we will introduce the ideas behind the use of statistics to make decisions – in particular, decisions about whether a particular hypothesis is supported by the data.

9.1 Null Hypothesis Statistical Testing (NHST)

The specific type of hypothesis testing that we will discuss is known (for reasons that will become clear) as *null hypothesis statistical testing* (NHST).

If you pick up almost any scientific or biomedical research publication, you will see NHST being used to test hypotheses, their introductory psychology textbook, Gerrig & Zimbardo (2002) referred to NHST as the "backbone of psychological research". Thus, learning how to use and interpret the results from hypothesis testing is essential to understand the results from many fields of research.

It is also important for you to know, however, that NHST can be problematic and that many statisticians and researchers (including me) think that it has been the cause of some problems in science.

For more than 50 years, there have been calls to abandon NHST in favor of other approaches (like those that we will discuss in the following chapters):

- "The test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research" (Bakan, 1966)
- Hypothesis testing is "a wrongheaded view about what constitutes scientific progress" (Luce, 1988)

NHST is also widely misunderstood, largely because it violates our intuitions about how statistical hypothesis testing should work. Let's look at an example to see.

9.2 Null hypothesis statistical testing: An example

You might be curious whether GRE test-prep training helps students score higher on the GRE. To answer this question, we need experimental evidence -- preferably a randomized controlled trial of the effectiveness of GRE test-prep training. We can randomly assign 100 students to take GRE test-prep training and randomly assign another 100 students to not take GRE test-prep training and then see how well each group does on the GRE.

Before we get to the example, let's ask how you think the statistical analysis might work. Remember, we want to test the hypothesis of whether GRE test-prep training helps students perform better on the GRE. The randomized controlled trial provides us with the data to test the hypothesis. The next obvious step is to look at the data and determine the data provide compelling evidence for or against the hypothesis.

It turns out that this is *not* how null hypothesis testing works. Instead, we first take our hypothesis of interest (i.e., whether GRE test-prep training leads to higher GRE scores), and flip it on its head, creating a *null hypothesis* – in this case, the null hypothesis would be that test-prep training does not lead to higher GRE scores.

Importantly, we then assume that the null hypothesis is true. We then look at the data, and determine whether the data are sufficiently unlikely under the null hypothesis that we can reject the null in favor of the *alternative hypothesis* which is our hypothesis of interest. If there is not sufficient evidence to reject the null, then we say that we "failed to reject" the null.

Understanding some of the concept of NHST (Null Hypothesis Statistical Testing), particularly the notorious "p-value," is invariably challenging the first time one encounters is, because it is so counter-intuitive.

As we will see later, there are other approaches that provide a more intuitive way to address hypothesis testing (but they have their own complexities). However, before we get to those, it's important for you to have a deep understanding of how hypothesis testing works, because it's clearly not going to go away any time soon.

9.3 The process of null hypothesis testing

We can break the process of null hypothesis testing down into a number of steps:

1. Formulate a hypothesis that embodies our prediction (*before seeing the data*)
2. Collect some data relevant to the hypothesis
3. Specify null and alternative hypotheses
4. Fit a model to the data that represents the alternative hypothesis and compute a test statistic
5. Compute the probability of the observed value of that statistic assuming that the null hypothesis is true
6. Assess the "statistical significance" of the result

Table 9.1: Summary of BMI data for active versus inactive individuals

PhysActive	N	mean	sd
No	131	30	9.021
Yes	119	27	5.233

For a hands-on example, let's use the NHANES data to ask the following question: Is physical activity related to body mass index? In the NHANES dataset, participants were asked whether they engage regularly in moderate or vigorous-intensity sports, fitness or recreational activities (stored in the variable $PhysActive$). The researchers also measured height and weight and used them to compute the *Body Mass Index* (BMI):

$$BMI = \frac{weight(kg)}{height(m)^2}$$

9.3.1 Step 1: Formulate a hypothesis of interest

For step 1, we hypothesize that BMI is greater for people who do not engage in physical activity, compared to those who do.

9.3.2 Step 2: Collect some data

For step 2, we collect some data. In this case, we will sample 250 individuals from the NHANES dataset. Figure 9.1 shows an example of such a sample, with BMI shown separately for active and inactive individuals.

9.3.3 Step 3: Specify the null and alternative hypotheses

For step 3, we need to specify our null hypothesis (which we call H_0) and our alternative hypothesis (which we call H_A). H_0 is the baseline against which we test our hypothesis of interest: that is, what would we expect the data to look like if there was no effect?

The null hypothesis always involves some kind of equality ($=$, $\leq$, or $\geq$).

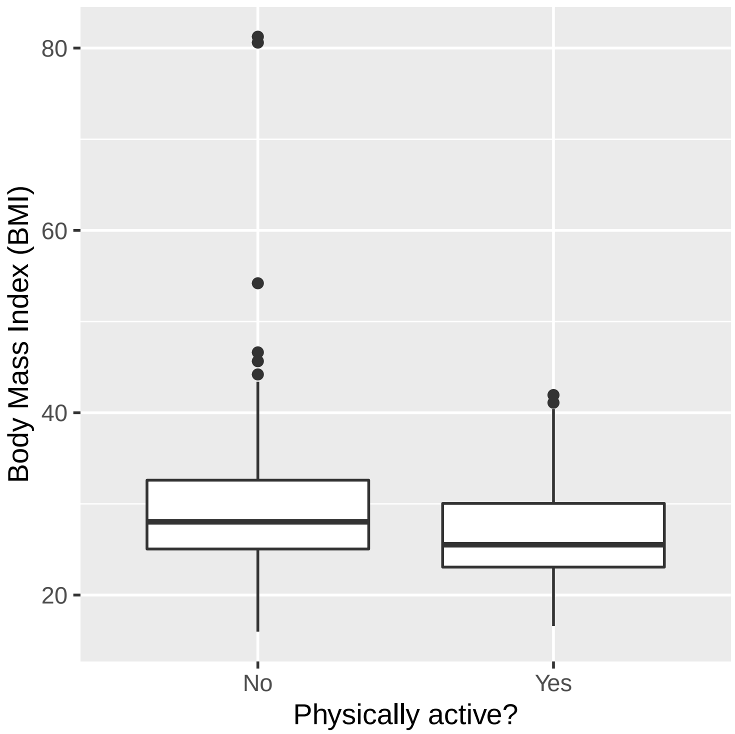

Figure 9.1: Box plot of BMI data from a sample of adults from the NHANES dataset, split by whether they reported engaging in regular physical activity.

The alternative hypothesis always involves some kind of inequality ($\neq$, $>$, or $<$). **Importantly, null hypothesis testing operates under the assumption that the null hypothesis is true unless the evidence shows otherwise.**

We also have to decide whether to use *directional* or *non-directional* hypotheses. A non-directional hypothesis simply predicts that there will be a difference, without predicting which direction it will go. For the BMI/activity example, a non-directional null hypothesis would be:

$H0 : BMI_{active} = BMI_{inactiveI}$

In English: The null hypothesis predicts that the BMI of adults who are active will not differ from the BMI of adults who are inactive.

The corresponding non-directional alternative hypothesis would be:

$HA : BMI_{active} \neq BMI_{inactive}$

In English: The alternative hypothesis predicts that the BMI of adults who are active will differ from the BMI of adults who are inactive.

A directional hypothesis, on the other hand, predicts which direction the difference will go. For example, we might have strong prior knowledge to predict that people who engage in physical activity should weigh less than those who do not, so we would propose the following directional null hypothesis:

$H0 : BMI_{active} \geq BMI_{inactive}$

In English: The null hypothesis predicts that the BMI of adults who are active will be greater than or equal to the BMI of adults who are inactive.

The corresponding directional alternative hypothesis would be:

$HA : BMI_{active} < BMI_{inactive}$

In English: The alternative hypothesis predicts that the BMI of adults who are active will be less than the BMI of adults who are inactive.

As we will see later, testing a non-directional hypothesis is more conservative, so this is generally to be preferred unless there is a strong *a priori* reason to hypothesize an effect in a particular direction. Any directional hypotheses should be specified BEFORE looking at the data!