Hypothesis Testing

Nicholas Kortessis

What is a hypothesis test?

So far in class, we have focused primarily on estimation. Estimation is about trying our best to figure out the value of a population parameter (a mean, a variance, a median, a proportion, etc.). Hypothesis testing is an additional practice done in biology and statistics. While related to estimation, it is actually quite a different exercise.

Hypothesis testing if fundamentally about the evidence in support of a particular probabilistic statement about the world.

Remember that we characterize populations (all individuals that we care about for a particular question) in terms of **population parameters**. Population parameters are characteristics of populations. Hypothesis testing works by making statements about these population parameters and using characteristics of samples to evaluate whether there is sufficient evidence consistent with our statement about population parameters (a hypothesis).

Steps in Hypothesis Testing

We can break hypothesis testing down into just a few steps.

1. Formulate a statement about the world probabilistically. This statement is the *hypothesis.*
2. Consider a sampling design and sample characteristics to measure that provide information about the hypothesis. The sample characteristic is called a *test statistic.*
3. Derive a sampling distribution for the values of test statistic and its associated probabilities *assuming the hypothesis is true.*
4. Decide which values of the sampling distribution are so unlikely so as to *count as evidence inconsistent with the hypothesis.* These outcomes are called *the rejection region.*
5. Sample the population, calculate the test statistic, and check to see if it is in the rejection region. If it is, reject the hypothesis. If it is not, you fail to reject the hypothesis.
6. While not a formal component of traditional hypothesis testing, it is common practice to now calculate and report the probability of finding an outcome **more extreme than the outcome observed under the hypothesis**. This is called a p-value and it represents a measure of the consistency of the data with the hypothesis.

A Worked Example with Sex Ratios

Let's see this in action. We'll begin with an example with sex ratios in animals. For the sake of simplicity, let's begin with an example of a diploid species with genetic sex determination. Many vertebrates (but not all) follow this kind of sexual system; humans are one example.

If viable offspring require two parents, one from each sex, then the genetics of heredity of sex chromosomes puts the probability of a zygote being male at 50% and the probability of a zygote being female at 50%.

So a question to ask is "Is the sex ratio in a species 50%?" The answer is of course yes or no. As such, a hypothesis test is warranted.

1. Formulate a hypothesis about the world probabilisticially.

Each individual has a possible sex, male or female. Hence, the character of interest is categorical with two outcomes. We can thus write the sex of individual i as

$$X_i \sim \text{Bernoulli}(p)$$

and let a "success" count as a male. For this probability distribution for an individual, p is the fraction of individuals in the species that are male. So a basic hypothesis for the sex ratio in a population based on the arguments above is $H_0 : p = 0.5$.

2. Consider the sampling design and a test statistic.

   To sample, we will randomly collect n individuals from the population and measure whether they are male or female. A test statistic that is relevant to the question is the proportion of males in the sample. Let's call the number of males, M. It is a random variable because it is as yet unknown and can take many values, anywhere from 0 to n. To get the proportion of males, we simply need to divide by the number of individuals in the sample. Let's call the proportion of males $P = M/n$, which is also a random variable.

3. Derive a sampling distribution under the hypothesis

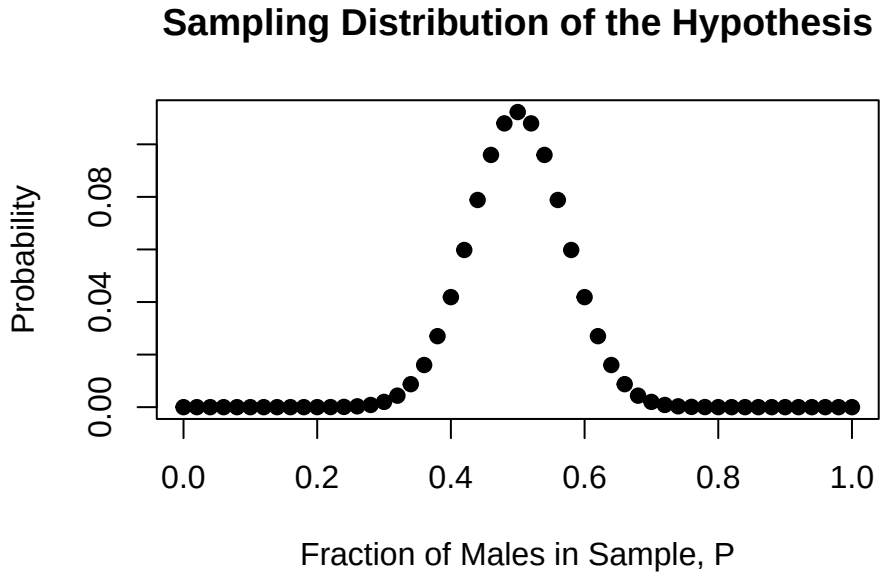
   When we do repeated random sampling from the population, the number of males follows a binomial distribution with parameters n and $p = 0.5$. That is,

   $$M \sim \text{Binomial}(n, p = 0.5)$$

   We can also write this in terms of the test statistic $P = M/n$ by just rearranging the equation for P so that $nP = M$. This means that

   $$nP \sim \text{Binomial}(n, p = 0.5).$$

   Here is what that looks like if we sample $n = 50$ individuals.



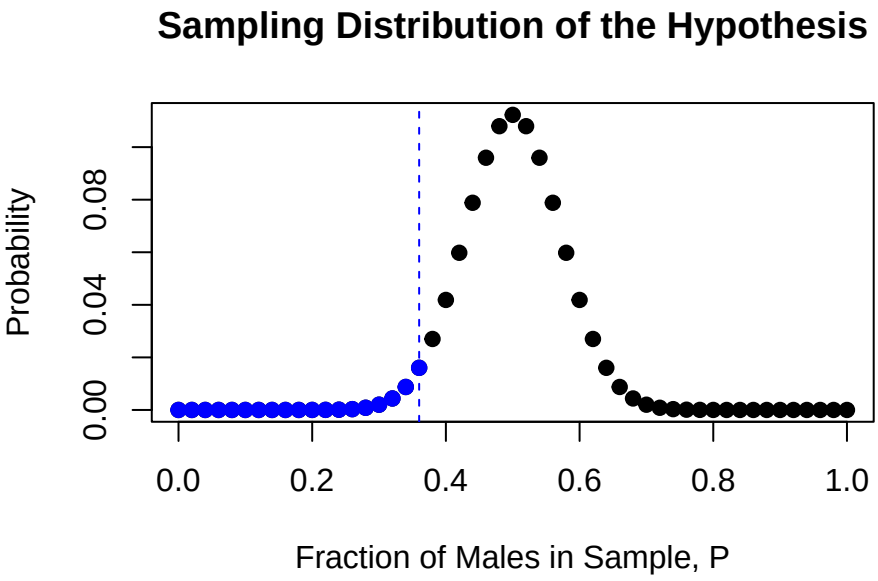
4. Decide which outcomes are sufficiently unlikely to count as inconsistent with the hypothesis.

You can see from the sampling distribution that some outcomes are likely–male fractions between about 40% and 60%, but others are unlikely–anything less than 20% or above 80%, for example.

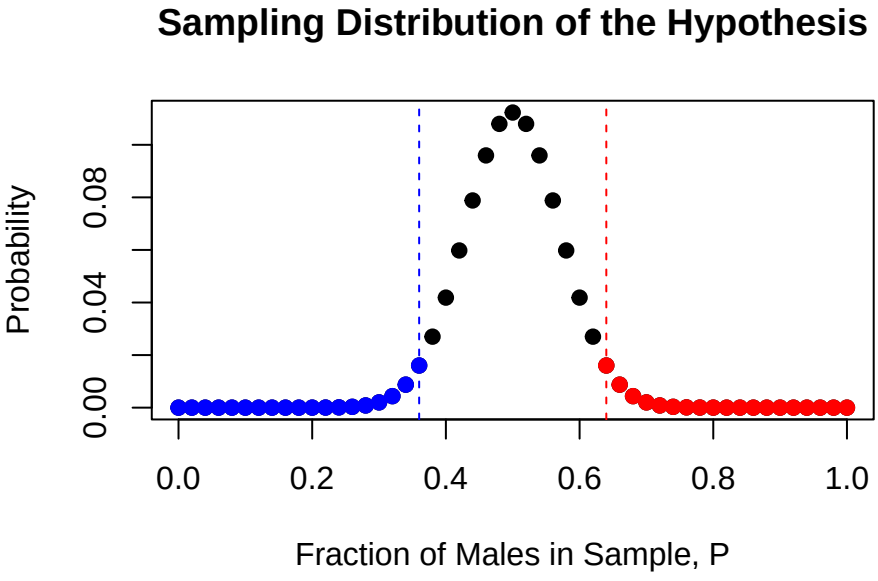
Hypothesis testing works by setting boundaries on what counts as likely versus unlikely. Just like with confidence intervals, we set a level, called an α-level, that quantifies how unlikely the outcomes much be such that we decide they are inconsistent with the hypothesis.

The α level gives the total probability of outcomes that are considered too unlikely to be consistent with the hypothesis. A value of $\alpha = 0.05$ is typical, and means that the least likely 5% of outcomes from the hypothesis are too inconsistent that we could feel comfortable rejecting the hypothesis.

For this example, the most likely outcomes are in the center of the distribution, near $P = 0.5$. The unlikely outcomes are large and small values of P. We can find the $\alpha = 0.05$ least likely outcomes using the cumulative distribution function for the sampling distribution. We just find the $\alpha/2 = 0.05/2 = 0.025$ percent of least likely small values of P using the $\alpha/2$ quantile of P. Here is the cutoff and the outcomes colored in blue.



And we can find the least likely large outcomes by finding the $1 - \alpha/2 = 1 - 0.05/2 = 0.975$ quantile. Any values above this quantile represent 2.5% least likely *large* values of P. In the figure below, the quantile is given by the vertical red line and the outcomes are colored in red.



This now breaks up the sampling distribution into two distinct kinds of regions.

- **Rejection region**: The rejection region is all the blue points and all the red points. Together, the red and blue points are the $\alpha = 0.05$ least likely outcomes under the hypothesis that the sex ratio is even. That should make some sense. If the sex ratio is even, it's pretty unlikely that a sample leads to predominately males (red region) or predominantly females (blue region).
- **Fail to reject region:** The fail to reject region is the black points. These values represent $1-\alpha = 1-0.05$ of the most probable outcomes under the hypothesis. Any outcome in this region is deemed sufficiently similar to the expectations of the hypothesis that we fail to reject the hypothesis.

We now have a simple decision rule. If the observed test statistic from data falls in the rejection region, we reject the hypothesis. If not, we fail to reject the hypothesis and move forward with our science using the hypothesis as a workable model for understanding the system.

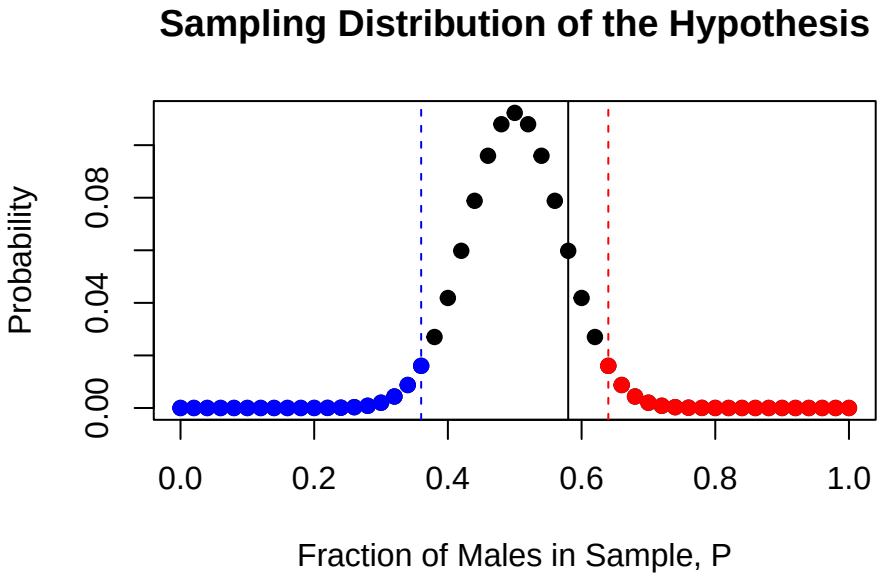
For this hypothesis ($p = 0.5$) and sample size ($n = 50$), our decision rule is

Reject the hypothesis of equal sex ratio if the observed fraction of males is greater than 64% or less than 36%.

5. Collect our sample, calcualte the test statistic, and make a decision

Now that we have a test ready, we can go out and sample 50 individuals from the population. Imagine that in doing so we collect 29 males and 21 females. This corresponds to an observed fraction of males in our sample of $\hat{p} = 29/50 = 0.58$. This is in the region where we fail to reject the hypothesis. As such, we say that we don't have sufficient evidence against the idea that the sex ratio is 50:50. This observation is consistent with normal sampling variability (the fact that sample properties differ from population properties).

Here is a figure to show the data in the context of the expectation. The vertical black line shows the fraction male in the data. This is broadly in the range of expected outcomes.



Null Hypotheses and Alternatives

What hypothesis do you choose when doing hypothesis testing? Typically, we choose what is termed a "null" hypothesis. It is given the name "null" to illustrate that this hypothesis signifies something with "no effect". In the example of sex ratios, the null hypothesis is $p = 0.5$, and is generated based on the arguments related to genetics. Hence, this implies there is "no effect" of any other factor that can influence sex ratio. Sex ratios can be shaped by many selective factors, including the sex ratio generated at birth or differential mortality later in life. In any case, these are factors that we can go investigate using other means. But a first step is to ask whether the standard understanding of sex ratios (50:50) works as a descriptor for a given species. If the answer is no, and we reject the null hypothesis, then we can go ask what factors influenced the non-equal sex ratio.

This applies to all kinds of hypothesis testing. The main goal is to construct a null hypothesis that represents a standard, baseline assumption of what the world looks like.

The other approach is to set up straw man arguments that can be shot down. Imagine you want to give a plant a hormone that is well documented to influence plant growth. To verify that a version of the plant hormone you use actually works as intended, you could set up the straw man argument that the hormone has no effect on plant growth as a null hypothesis. This strength of evidence then must be to reject this straw man. This is how much of hypothesis testing is done in practice.

Null hypothesis are often written as H_0 where the subscripted "0" is pronounced "naught", meaning "nothing". An alternative hypothesis, written as H_A, is anything that is not H_0. For the sex ratio example here, we write

$$H_0 : p = 0.5$$

$$H_A : p \neq 0.5.$$

This notation signifies that the we take the null expectation that sex ratio of the population is 50% and that an alternative is that it is *anything other than 50%.*

Types of Errors

There are two outcomes of a hypothesis test: reject the hypothesis or fail to reject it. Each of these conclusions have the potential to be wrong. For example, we could erroneously reject the hypothesis when it is a good model of the world. Similarly, we could erroneously fail to reject the hypothesis when it is not a good representation about the world.

The table below gives a summary of the kinds of errors one can make when doing hypothesis testing.

	H_0 **True**	H_0 **False**
Reject H_0	Type I Error	Correct Inference
Fail to Reject H_0	Correct Inference	Type II Error

The top of the table gives two possibilities for how the world actually is; either the null hypothesis is true or it is not. On the left of the table shows the two possibilities for the outcome of a test; either you reject the null hypothesis or not. Within the table are two kinds of ways of being correct: either you reject the null when the null is not true or you fail to reject the null when the null is actually true.

There are also two ways to be wrong. We call these Type I and and Type II errors.

The probability of making a Type I error is α. It says, what is the chance that we reject the null even if it is true?

The probability of making a Type II error is typically denoted as β. It quantifies a very different kind of way of being wrong. Some alternative to the null is true, but we don't have enough evidence inconsistent with the null to say the null is not true.

Type II errors are related to an important concept in hypothesis testing called **power**. Power is the probability of correctly rejecting the null when the alternative is true. Please read about power in the notes about power and power analysis.

P-values: a measure of how unlikely data are under the hypothesis

p-values are commonly reported statistics in science when hypothesis testing is done, but they don't show up anywhere here yet. What are they? Despite their ubiquitous use, they are often misunderstood.

They have a very particular meaning. It is NOT the probability that the null hypothesis is true (this a typical misconception). They quantify how likely the observed effect size, or a more extreme effect size, is under the null hypothesis.

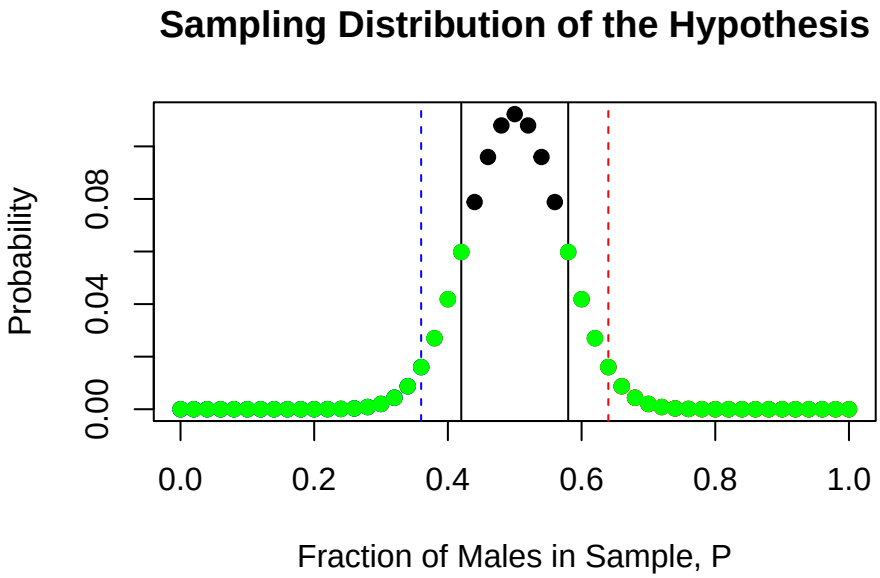
To understand how this works, we need a definition of effect size. It's really just the difference between the hypothesis and the data you collected. In our sex ratio example above, the hypothesis is that the proportion male is 50%. But our sample had 58% male (29/50). This means that the effect size is 0.58 - 0.5 = 0.08. This is just the difference between the hypothesis and the observation.

A p-value takes the probability of a test statistic at least as extreme than this value. A note here is that our observed statistic could be extreme in two ways: (1) we could have 58% males (effect size 0.08) or (2) we could have 58% females (effect size -0.08). Both count as extreme in this context because the hypothesis is only about males and females being equal. Excess males is just as extreme as excess females and vice versa.

Given our test statistic, P, we can define the p-value as

$$p = \Pr(P \geq 0.58) + \Pr(P \leq 0.42).$$

Here these are colored in.



The green points are the outcomes that are more extreme than the outcome we got. The red and blue dashed lines define the boundaries of the rejection regions.

Calculating the p-value is straightforward with cumulative density functions. Like this

$$p = 1 - \Pr(P \leq 0.58) + \Pr(P \leq 0.42).$$

This is just two calls from the cumulative density function for the binomial. When we do the calculation, we get

```
## [1] 0.2624375
```

Relationship between the α value and the p-value

Typically, people look for a p-value below 0.05 to call something statistically significant (more on that later). Since p-values measure the cumulative density of outcomes that are more extreme, the p-value is α when the outcome lands right at the boundary of the rejection region.

This leads to the practice of implicitly defining α (usually at 0.05) and then rejecting the null if $p < \alpha$. This always works. If $p > \alpha$, then the test statistic does not fall in the rejection regions.

Some warnings about hypothesis testing

Here are some warnings about hypothesis testing. You will read about hypothesis testing and some of the known issues with it. Here is a quick summary.

1. Null hypotheses make a very specific statement and the world can be different in very many different ways. Any of these differences count as rejection. Even biologically non-meaningful differences. This is definitely the case when data sources are large.
2. Statistical significance is not biological significance. All hypothesis testing does is reject null hypotheses. It does not say how different an alternative is compared ot the null. Most importantly, p-values don't tell you anything about how biologically meaningful an effect is.
3. P-values are unstable at small sample sizes. Repeating a study with small samples could lead to really big differences in p-values from one experiment to another.