5.1: Introduction to Hypothesis Testing

Textbook: 10.1 – 10.2

Objectives

- Understand the logic behind hypothesis testing
- Be able to carry out the steps of hypothesis testing
- Know the definitions of Type I error, Type II error, and power

Motivation

This course focuses heavily on the idea of statistical inference: taking a sample from a larger population, analyzing the sample, and using the results of this analysis to make a claim about a parameter θ that is unknown in the population.

In the Unit 3 notes, we introduced the topic of confidence intervals and learned how we could use them to make a claim about an unknown parameter value. In such cases, we started with a sample and then used our sample to create an interval of "likely" values for our unknown parameter θ.

A confidence interval is useful for inference as it allows us to make a claim about a parameter based on a sample drawn from the parent population. These intervals can be especially helpful for inference if you do not have a specific value in mind for the parameter but rather just want to see a range of likely values.

But what if we had a more specific question that we wanted to ask about a given unknown parameter θ?

Examples

- Let X represent the amount of time spent waiting before the #9 bus arrives at the Craigie Hall bus stop. Assume $X \sim exponential(\beta)$. What is the probability that β, the average amount of time spent waiting for the #9 at this stop, is at least five minutes?
- Let X represent the cost (in dollars) of an MRI test in Alberta. Assume $X \sim normal(\mu, \$450)$. What is the probability that μ, the average cost of an MRI test in Alberta, is more than $600?

In this set of notes, we will discuss another very commonly used method of statistical inference, hypothesis testing, which is appropriate for addressing questions similar to those in the examples above. We will give a general introduction to the method of hypothesis testing here, then see some more specifics of it in the 5.2 notes.

Hypothesis Testing

Hypothesis testing is a formal procedure that allows us to choose between two competing hypotheses when we are uncertain about our measurements. The process of hypothesis testing can be broken down into a series of steps, and the easiest way to understand the method is to look at it step-by-step.

The Steps of Hypothesis Testing

While some textbooks and other resources might differ slightly in how these specific steps are defined, the general steps of hypothesis testing are as follows:

[**Step 0:** Determine the level of significance α]
Step 1: Formulate the null and alternative hypotheses
Step 2: Compute the test statistic and/or p-value
Step 3: Compare the test statistic to the rejection region (or compare the p-value to α)
Step 4: State the conclusion about the hypotheses

For now, let's ignore Step 0 (we actually compute/use the level of significance in Steps 3 and 4, so we'll talk about it then) and jump right into Step 1.

Step 1: Formulate the Null and Alternative Hypotheses

As the name "hypothesis testing" suggests, this method of statistical inference involves testing a hypothesis or claim about the value of a population parameter. In order to do so, we set up two mutually exclusive hypotheses about the population parameter of interest.

The null hypothesis (denoted H_0, pronounced "null hypothesis" or "H-naught") represents the hypothesis or claim about the population parameter that is assumed to be true unless convincing evidence is provided to suggest that it is false. You can think of H_0 as representing the "status quo," or, in the context of new research, the claim that there is *no difference* or *no change* from what is currently accepted as truth in the population.

The alternative hypothesis (denoted H_a or H_1, pronounced "H-A" or "H-1") represents the hypothesis or claim about the population parameter that will only be accepted if convincing evidence is provided to suggest that it is true. Sometimes H_a is considered the "researcher's hypothesis" as it often represents the claim about the population parameter for which the researcher wants to demonstrate is true.

A simple hypothesis occurs when a null hypothesis uniquely specifies the distribution of the population from which the sample is taken. A composite hypothesis occurs when a null hypothesis is not a simple hypothesis. Consider an unknown population parameter θ and some claimed or hypothesized value for θ, θ_0 (often called the null value). The following are the three sets of mutually exclusive pairs of hypotheses:

Simple ("two-tailed")	Composite ("left-tailed")	Composite ("right-tailed")
$H_0: \theta = \theta_0$	$H_0: \theta \geq \theta_0$	$H_0: \theta \leq \theta_0$
$H_a: \theta \neq \theta_0$	$H_a: \theta < \theta_0$	$H_a: \theta > \theta_0$

A few things to note about the phrasing of these hypotheses in all three pairs:

- Both hypotheses (H_0 and H_a) are making a claim about the population parameter value θ. Thus, when we are making claims about a population parameter, we should always use θ (rather than a sample statistic, $\hat{\theta}_n$) in our hypothesis statements.
- Both hypotheses should involve a claim about the null value, θ_0.
- The null hypothesis (H_0) statement always has an "=" in it, whether it is written as "=" or "≤" or "≥." One way to remember this is that our H_0 is the claim that our parameter is equal to the null value. The alternative hypothesis (H_a) does not have an "=" in it and should be a complement statement to H_0. One way to remember this is that our H_a is the claim that our parameter differs from the null value θ_0.

Note: is it also acceptable to write H_0 with only the equal sign rather than the inequality sign for a composite hypothesis.

Example 5.1.1
Let X represent the amount of time spent waiting before the #9 bus arrives at the Craigie Hall bus stop. Assume $X \sim exponential(\beta)$. State the appropriate null and alternative hypotheses to test the claim that β, the average amount of time spent waiting for the #9 at this stop, is at least five minutes.

Example 5.1.2
Let X represent the number of heads observed in n flips of a coin and let $\hat{p} = \frac{X}{n}$. State the appropriate null and alternative hypotheses to test the claim that the coin is a "fair" coin.

Step 2: Compute the Test Statistic and/or p-Value
One way to think about hypothesis testing is as a criminal trial where the idea of "innocent until proven guilty" holds. In a criminal trial, you have two hypotheses set for the jury: the defendant is not guilty or the defendant is guilty.

If the defendant is "innocent (not guilty) until proven guilty," then the jury must assume the defendant is not guilty unless evidence overwhelmingly suggests that this is not the case. We can see the similarities to our hypotheses in hypothesis testing as follows. If the null and alternative hypotheses are

H_0: the defendant is not guilty
H_a: the defendant is guilty

then the null hypothesis is, by default, assumed true. It is only when there is overwhelming evidence against the null hypothesis that we can say that we reject the claim of the null hypothesis.

In statistics, this "evidence" takes the form of sample data, and it is in the second step of hypothesis testing where we first take a look at the data. More specifically, we aggregate our data into a test statistic. The test statistic is a function of our sample measurements. The test statistic follows some known distribution for which we can calculate probabilities and it gives us a succinct summary of what our data tells us.

To better understand the test statistic and its role in hypothesis testing, we have to go back to our knowledge about sampling distributions. The sampling distribution of a sample statistic $\hat{\theta}_n$ is the distribution of all possible $\hat{\theta}_n$ values for a given sample size n.

The key here in Step 2 of our hypothesis testing process is **to first assume that H_0 is true.** If we assume that H_0 is true, then under this assumption, the population parameter is equal to our null value, or $\theta = \theta_0$.

A consequence of this is that the mean of our sampling distribution of is also equal to the null value, since we're treating the null value as the "true" population mean under H_0.

So how do we use this information? Since the sampling distribution is a distribution of all possible sample statistics $\hat{\theta}_n$ for a given sample size, this means that the sample statistic from any given sample will fall somewhere in the sampling distribution, which means our specific sample statistic sits somewhere in relation to the null value. The test statistic itself is just the standardized value of our sample statistic under the assumption that H_0 is true.

Example 5.1.2 (continued)
Let X represent the number of heads observed in n flips of a coin and let $\hat{p} = \frac{X}{n}$. State the appropriate null and alternative hypotheses to test the claim that the coin is a "fair" coin. You flip the coin 25 times and find that $\hat{p} = 0.44$. Determine the test statistic for this sample.

Hypotheses
$H_0: p = 0.50$
$H_a: p \neq 0.50$

Test Statistic

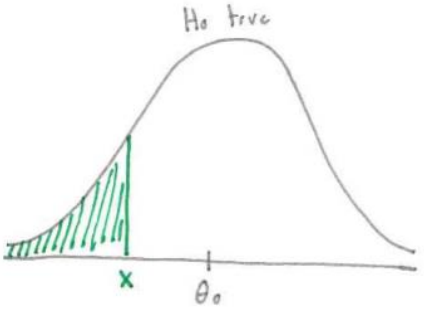
Now that we have a test statistic, what can we do with it? What we want from our test statistic is the "unusualness" of $\hat{\theta}_n$. In other words, we want to determine how "extreme" or "unlikely" our sample statistic is under the assumption that H_0 is true. This is the job of the p-value.

The p-value is the probability of collecting another sample as "extreme" or more "extreme" than the one we've observed from our given sample, assuming H_0 is true. Another way to think of the p-value is that it gives you the probability of repeating the sampling process again and comparing the likelihood of that theoretical result with the result that you got from your actual sample.

What's considered "extreme" or "unusual" depends on the sign (<, >, ≠) in the alternative hypothesis, H_a.

"Left-Tailed" (<) Phrasing
The p-value is calculated as the probability of observing a sample statistic smaller than our actual observed sample statistic. If we let X represent our test statistic,

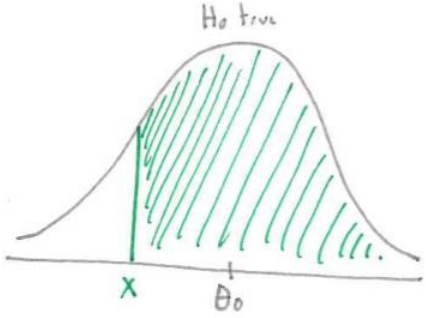
p-value $= \mathrm{P}(f_X(\mathrm{x}) < \mathrm{X} | H_0 \text{ true})$



"Right-Tailed" (>) Phrasing

The p-value is calculated as the probability of observing a sample statistic larger than our actual observed sample statistic. If we let X represent our test statistic,

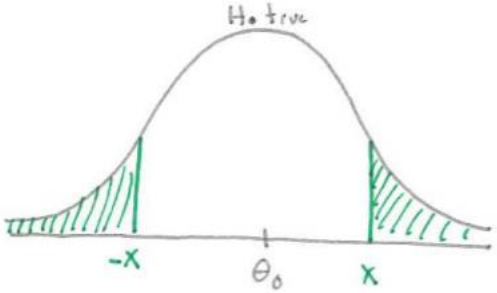
$$\text{p-value} = \mathrm{P}(f_X(\mathrm{x}) > \mathrm{X} | H_0 \text{ true})$$



"Two-Tailed" (≠) Phrasing

The p-value can best be understood as the probability of observing an "extreme" sample statistic in either direction (either very large or very small). If we let X represent our test statistic,

$$\text{p-value} = 2 \times \mathrm{P}(f_X(\mathrm{x}) < -|\mathrm{X}| | H_0 \text{ true})$$



Example 5.1.2 (continued)

Let X represent the number of heads observed in n flips of a coin and let $\hat{p} = \frac{X}{n}$ State the appropriate null and alternative hypotheses to test the claim that the coin is a "fair" coin. You flip the coin 25 times and find that $\hat{p} = 0.44$. Determine the p-value for this sample.

Hypotheses

$H_0: \hat{p} = 0.50$

$H_a: \hat{p} \neq 0.50$

Test Statistic

$Z_{calc} = -0.60$

p-Value

R function: `pnorm(x,mean,sd)`

`2*pnorm(  )`

Step 3: Compare the Test Statistic to the Rejection Region or Compare the p-Value to α

Step 2 is all about computing the test statistic and/or p-value from a given sample and sample statistic $\hat{\theta}_n$. On its own, the test statistic (and p-value) acts as a summary of the "evidence" we have collected in the form of our sample. Step 3 is the part of the hypothesis testing process where we determine what type of evidence – evidence in favor of the null hypothesis or evidence against the null hypothesis – our sample gives us.

In Step 3, we can either compare our test statistic to what's known as a rejection region or we can compare our p-value to a value called α. Let's actually start with the latter, since it is a bit easier to understand, and then we'll see how comparing our test statistic to a rejection region yields the same results.

Comparing the p-Value to α

The value of a sample statistic can fluctuate from sample to sample and thus a single sample's statistic is not a perfect representation of the population parameter. One consequence of this is that any inference we make based on a sample statistic carries with it a certain chance of error – the chance that we make the wrong conclusion about our population parameter's value.

Thus, just as we did with confidence intervals, it is necessary to assign a certain amount of allowable error (or tolerance) to our hypothesis testing procedure to ensure that the implications of our calculations say something of significance. This chance of error is our level of significance (or significance level or α-value). The level of significance is usually denoted α and is formally defined as the probability that we reject H_0 based on our sample when in fact H_0 is actually true in the population.

$$\alpha = P(reject\ H_0 | H_0\ true)$$

The α-value usually selected to be quite small (we don't want a large probability of this type of error!) and is most often set at α = 0.05. We like to "set" or "assign" our α-value before we even begin the formal process of carrying out the test itself; thus, I like to refer to the action of setting the α-value as "Step 0."

Now let's return to our p-value and see how we use it and α together in the context of hypothesis testing. Our p-value is the probability of observing a sample statistic more "extreme" than the sample statistic obtained in our original sample, under the assumption that H_0 is true. So what does this mean?

If the p-value is **large**, it suggests that our sample statistic is a **likely value** if H_0 is indeed true. In other words, if H_0 is true, there is a good probability or chance that we would have observed the sample statistic we calculated in our sample.

If the p-value is **small**, it suggests that our sample statistic is an **unlikely value** if H_0 is indeed true. In other words, if H_0 is true, there is a small probability or chance that we would have observed the sample statistic we calculated in our sample.

Another way to think about it:

- A **large** p-value provides evidence **in favor of** H_0 being true in the population
- A **small** p-value provides evidence **against** H_0 being true in the population

A question you might be asking at this point: what is considered a "large" p-value and what is considered a "small" p-value? In other words, how do we decide we have evidence in favor of or against H_0? To determine if a p-value is "large enough" or "small enough," we compare it to our α-level. This gives us our decision rule.

The p-value approach to the decision rule:

- If p-value < α, reject H_0
- If p-value ≥ α, fail to reject H_0

In other words, we use our α-value to set up a dichotomous decision. If our p-value is "small enough"—smaller than α—we say that our sample statistic value is so unlikely under the assumption of H_0 being true that we actually can use it as evidence against H_0 being true.

If our p-value is "large"—equal to or larger than α—we say that our sample statistic value is not unlikely or unusual under the assumption of H_0 being true, so we can't make a claim against H_0 based on our sample.

Comparing the Test Statistic to a Rejection Region

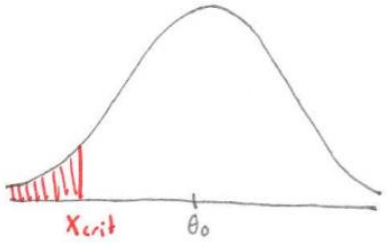
We can also develop an equivalent decision rule without ever having to calculate a p-value. This second method of constructing a decision rule is based on the idea of a critical value and a rejection region.

A critical value is a value in the distribution of our test statistic that corresponds to the α-level. Just like our test statistic (a standardized value based on our sample) corresponds to the p-value, the critical value corresponds to the α-level. This critical value allows us to define a rejection region. The rejection region (RR) is an interval of test statistic values where, if our test statistic falls in the rejection region, we reject H_0.

The critical value and rejection region depend, in part, on the sign (<, >, ≠) in our alternative hypothesis, H_a.

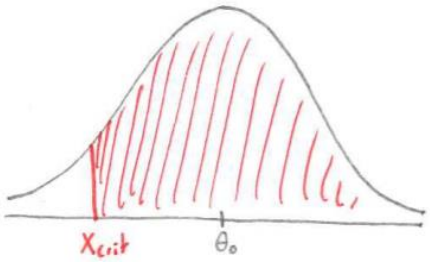
<u>"Left-Tailed" (<) Phrasing</u>
The critical value X_{crit} is the value in the distribution of our test statistic that defines α as a "lower-" or "left-tail" area.

$\mathrm{P}(f_X(\mathrm{x}) < X_{crit} | H_0 \text{ true}) = \alpha$



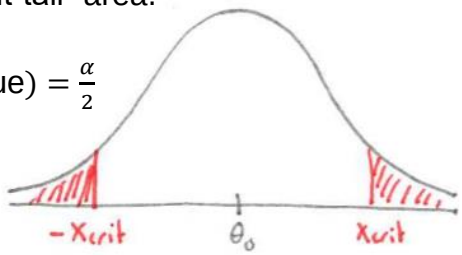
<u>"Right-Tailed" (>) Phrasing</u>
The critical value X_{crit} is the value in the distribution of our test statistic that defines α as an "upper-" or "right-tail" area.

$\mathrm{P}(f_X(\mathrm{x}) > X_{crit} | H_0 \text{ true}) = \alpha$



<u>"Two-Tailed" (≠) Phrasing</u>
The critical values $X_{crit_{upper}}$ and $X_{crit_{lower}}$ are the values in the distribution of our test statistic that define $\frac{\alpha}{2}$ as a "left-tail" area and $\frac{\alpha}{2}$ as a "right-tail" area.

$\mathrm{P}(f_X(\mathrm{x}) < X_{crit_{lower}} | H_0 \text{ true}) = \mathrm{P}(f_X(\mathrm{x}) > X_{crit_{upper}} | H_0 \text{ true}) = \frac{\alpha}{2}$



The test statistic approach to the decision rule:

- If the test statistic is in the rejection region, reject H_0
- If the test statistic is not in the rejection region, fail to reject H_0

Again, these two versions of the decision rule are equivalent! That is, they should lead you to the same conclusion.

p-value < α	…is equivalent to…	the test statistic in RR	…which leads us to…	reject H_0
p-value ≥ α	…is equivalent to…	the test statistic not in RR	…which leads us to…	fail to reject H_0

Step 4: State the Conclusion about the Hypotheses
Once we have made our decision regarding our hypotheses in Step 3, the final step is easy – just state this conclusion formally and in the context of the scenario.

Example 5.1.2 (continued)
Let X represent the number of heads observed in n flips of a coin and let $\hat{p} = \frac{X}{n}$ State the appropriate null and alternative hypotheses to test the claim that the coin is a "fair" coin. You flip the coin 25 times and find that $\hat{p} = 0.44$. State the conclusion of this hypothesis test using $\alpha = 0.05$.

Hypotheses
$H_0: \hat{p} = 0.50$
$H_a: \hat{p} \neq 0.50$

Test Statistic
$Z_{calc} = -0.60$

p-Value
p-value = 0.5485

Conclusion

Errors in Hypothesis Testing

When we conduct a hypothesis test, it leads us to one of two decisions based on the evidence of our sample: we either reject H_0 or fail to reject H_0. However, a hypothesis test is not foolproof—our decision might be incorrect!

A hypothesis test can result in one of two decision errors:

1. A Type I error occurs when we reject H_0 when H_0 is actually true. The probability of committing a Type I error is defined by the significance level, α, and thus is usually denoted as α. The probability expression of a Type I error is:

$$P(reject\ H_0 | H_0\ true) = \alpha$$

2. A Type II error occurs when we fail to reject H_0 when H_0 is actually false. The probability of committing a Type II error is usually denoted as β. The probability expression of a Type II error is:

$$P(fail\ to\ reject\ H_0 | H_0\ false) = \beta$$

Conversely, a hypothesis test can result in one of two *correct* decisions:

1. We can reject H_0 when H_0 is actually false. The probability of rejecting H_0 when H_0 is false is called the power of a test and is computed as $1 - \beta$. The probability expression of power is:

$$P(reject\ H_0 | H_0\ false) = 1 - \beta$$

2. We can fail to reject H_0 when H_0 is actually true. The probability of failing to reject H_0 when H_0 is true is computed as $1 - \alpha$ with a probability expression of:

$$P(fail\ to\ reject\ H_0 | H_0\ true) = 1 - \alpha$$

The relationships amongst these correct decisions and decision errors can be seen by using a table:

		Truth	
		H_0 True	H_0 False
Decision from hypothesis test	Reject H_0		
	Fail to Reject H_0		

*Example 5.1.3**
A political candidate is considering running for public office, but she will only do so if she has some indication that she will receive at least 50% of the votes. In a preliminary sample of 20 voters, the candidate determines that if at most five voters say that they will vote for her, then she will not run for public office.

1. Let X represent the number of voters out of this preliminary sample of $n = 20$ who say that they would vote for this particular candidate. State the appropriate null and alternative hypotheses.

2. Determine α, the probability of making a Type I error.

R function: `pbinom(x,size,p)`

```
pbinom(   , size =     , p =      )
```

3. Unbeknownst to the candidate, she will actually receive 40% of the votes. Given this information, what is the probability of concluding from the preliminary sample that she will receive at least 50% of the votes?

R function: `pbinom(x,size,p)`

```
1 - pbinom(   , size =     , p =      )
```

4. A random sample of 20 voters was taken. In this sample, eight voters said that they would vote for the candidate. From this, what decision can be made about the null hypothesis in part 1.?

R function: `pbinom(x,size,p)`

```
pbinom(   , size =     , p =     )
```