Hypothesis testing

Basic hypothesis testing

We now have learned enough to proceed to one of the most important topics in statistics. This is the idea of hypothesis testing. Simply, we develop a hypothesis about our data, and then we either reject this hypothesis or we fail to reject it.

Let's do an example. Suppose we have an idea about the true average index finger length in humans (note our idea is about μ, not $\bar{y}$), and we think that the true average finger length (μ) is 8.4 cm. How would we test this?

First we need to collect some data, so let's take a sample:

$$n = \qquad \bar{y} = \qquad s = \qquad SE_{\bar{y}} =$$

Now suppose we calculate a 95% confidence interval and we get $(6.2, 7.6)$. Is our confidence interval consistent with our hypothesis? **No**, of course not. We (for some reason) think $\mu = 8.4$, which is *not* in our confidence interval. In this case we would *reject* our hypothesis because we don't believe it is true.

We can do hypothesis tests this way (using confidence intervals), and the answers we get will always match the answers we get with a more formal hypothesis testing approach. But for a number of reasons, statisticians often prefer to think about hypothesis testing in a different way. So let's introduce the hypothesis testing approach.

First, we want to formally identify our hypothesis. We'll call this our *null hypothesis*:

> H_0: The true average population index finger length in humans is 8.4 cm.

Since our null hypothesis (H_0) might not be true, we also have to identify an alternative. We'll call this the *alternative hypothesis*:

> $H_1 (= H_\text{A})$: The true average population index finger length in humans is *not* 8.4 cm.
>
> (Some texts like to use H_A, but we'll stick with H_1)

The two hypotheses above include all possible outcomes (the population mean either is, or is not 8.4 cm). If we want to abbreviate this, we could write the hypotheses as follows:

$H_0 : \mu = 8.4$

$H_1 : \mu \neq 8.4$

One problem with abbreviations like this is that it is easy to loose track of what one is testing. This isn't so apparent here, but will be more obvious when we explore more complicated hypotheses tests. If you do abbreviate, you should always keep track of what the abbreviations stand for (e.g., μ is the true average population index finger length).

As mentioned above, our hypotheses are all written in terms of μ. Making a hypothesis about $\bar{y}$ is silly since we *know* what $\bar{y}$ is. Our hypotheses are *always* about the population parameters.

Now that we have our hypotheses written out, what do we do next? We make a decision about when to reject our null hypothesis (H_0). In other words, we look at our data and ask ourselves "*are the data consistent with* H_0"?

To do this, we use a probability argument. We calculate the probability of getting the data in our sample if our H_0 is true:

If this probability is low, we *reject* our H_0.

If this probability is high, we *fail to reject* our H_0.

This will become more obvious below. For now let's decide that if the probability of getting our sample data if H_0 is true is less than 5%, then we will reject H_0. This value, (5%) is also known as "alpha", designated by the Greek letter α.

So how do we get a probability? We look at our data and from our data we calculate a t-value:

$$|t^*| = \left| \frac{\bar{y} - \mu}{s/\sqrt{n}} \right|$$

We use t^* to differentiate the t-value we calculate from our data from the t-value we derive from the t-tables (more below).

So let's do this for our finger length example:

$$|t^*| = \left| \frac{\square - 8.4}{\square/\sqrt{\square}} \right| = |\square| = \square$$

Finally we use software to get the probability of getting this value of $|t^*|$. Above we decided to use 5% as a cut-off value. In other words:

> If the probability of getting $|t^*|$ is less than (or equal to) 5%, we will *reject* our H_0.
>
> If the probability of getting $|t^*|$ is greater than 5%, we will *fail to reject* our H_0.

This is very similar to looking up probabilities in the z-table, except now we would use the t-table if it were extensive enough (it isn't). Remember, we need to estimate σ using s, which means $|t^*|$ has a t distribution.

Since we can't really use the t-table (but see below), we'll use R to get the probability of our $|t^*|$:

$$Pr\{|t^*| \geq \square\} = \square$$

And from this we can make our decision.

So we now know how to do a hypothesis test. But we have one problem. Without software to give us the probability of getting $|t^*|$ we would have a very difficult time in calculating the probability. In other words, without software, we need a different approach.

The t-table gives us the value of t that corresponds to certain probabilities. For example, if we look at our t-table for 6 degrees of freedom and a probability of 0.05 (two sided) we find that $t_{0.05,6} = 1.943$. So we can write:

$$Pr\{|t| \geq 1.943\} = 0.05$$

So how do we use this for a hypothesis test? If the value of t that we calculate is *larger or equal* to the value of t_{table} we know the probability must be 0.05 or smaller. So we can proceed as follows:

> **If our $|t^*| \geq t$ from the table we *reject* our H_0.**
>
> (Because the probability of getting our $|t^*| \leq 0.05$)
>
> **If our $|t^*| < t$ from the table we *fail to reject* our H_0.**
>
> (Because the probability of getting our $|t^*| > 0.05$)

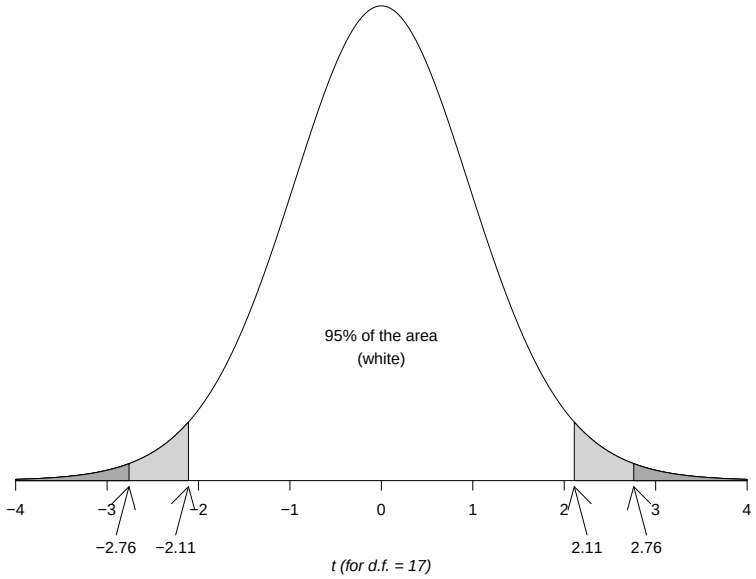
Let's do an example. Suppose we calculate $|t^*| = 2.76$ and have a sample size of n =18.

We look up $t_{0.05,17}$ =2.110.

And since our $|t^*| = 2.76 \geq t_{0.05,17} = 2.110$ we say that the probability of getting $|t^*| = 2.76$ is less than 5%.

> So we say we don't believe the results we got are due to chance, or *we don't believe our null hypothesis (H_0) is true.* **We *reject* H_0.**

Here's what we did on a graph:



The probability of getting $t^* \geq 2.11$ or $t^* \leq -2.11$ is 5% (indicated in light gray on the graph).

So the probability of getting $t^* \geq 2.76$ or $t^* \leq -2.76$ is must be less than 5% (the darker gray area in the graph).

> In other words, the probability of our result is less than 5%, so we *reject* H_0.

What we just learned is a *one-sample t-test*. It is also *two sided* (important: *sample* and *sided* are two unrelated concepts - don't get them confused).

Why is it called a one-sample t test?

We have one sample (e.g., our finger lengths).

We used the t distribution.

Why two sided?

Because we reject if our $\bar{y}$ is either much bigger or much smaller than our hypothesized value for μ.

A one sample t test isn't terribly useful, so it's not always covered in introductory tests, but it serves as a nice lead-in to more complicated tests and it follows naturally from confidence intervals. We will also see this test again in a slightly different guise when we look at paired tests.

Making mistakes: α and β.

So now that we have learned how to do a basic hypothesis test we need to look at some of the details.

When we look up a value in the t-table to compare to our $|t^*|$ we need to know two things: the degrees of freedom ($= d.f. = \nu$) and the level at which we will decide if our H_0 is wrong. In the example above, we used 5% for this value. You might remember that we also referred to this value as *alpha*, designated by the Greek letter α. We need to define α more carefully now:

$$\alpha = Pr\ \{reject\ H_0\ if\ H_0\ is\ true\}.$$

In other words, α is the probability of making a mistake:

You reject H_0 because you don't think it is true.

But if it actually is true - then you've made a mistake.

So why not just make α very small? Wouldn't that stop you from making a mistake?

Unfortunately not. Because you can also make another mistake (designated by the Greek letter β):

$$\beta = Pr\ \{fail\ to\ reject\ H_0\ if\ H_0\ is\ false\}.$$

Here's another way of looking at all of this:

	$\boldsymbol{H_0}$ **is true**	$\boldsymbol{H_1}$ **is true**
We decide $\boldsymbol{H_0}$ is true	we're right	we're wrong *Type II error*
We decide $\boldsymbol{H_1}$ is true	we're wrong *Type I error*	we're right

The problem with making α really small is that the probability of β increases (and vice versa - although we can't control β).

Incidentally, notice that we can also define α and β as follows:

$\alpha = Pr\{type\ I\ error\}$.

$\beta = Pr\{type\ II\ error\}$.

Let's see what happens if we set α too low. We'll use an example based on a study done by Schaller in the Serengeti in 1972. He observed the following litter sizes in 18 ($n = 18$) lionesses:

0 0 0 2 3 3 3 4 4 5 5 5 5 6 8 8 9 9

For these data we can calculate the following:

$\bar{y} = 4.4$, and $s = 2.9$.

Is it reasonable to think that $\mu = 9$? Of course not! But let's test this hypothesis anyway. So we have:

$H_0 : \mu = 9$

$H_1 : \mu \neq 9$

Let's pick a very reasonable level of α and use $\alpha = 0.05$.

So we calculate our $|t^*|$:

$$t^* = \frac{\bar{y} - \mu}{s/\sqrt{n}} = \frac{4.4 - 9}{2.9/\sqrt{18}} = -6.73 \implies |t^*| = 6.73$$

Now we look up t with $d.f. = 17$ and $\alpha = 0.05$:

$$t_{17,0.05} = 2.110$$

And because $|t^*| = 6.73 \geq t_{17,0.05} = 2.110$ we reject our H_0.

This seems reasonable. After all, a litter size of 9 is high, and we certainly don't think it's the "average" litter size.

But now suppose we're really worried about making a Type I error, so we choose $\alpha = 0.00000005$. Note that none of the math changes, but the t-table value *does* change. We now need to use $t_{17,0.00000005}$. This value is so absurd that it's not in our t-tables (nor in anyone else's t-tables!). But we know how to use R, and R doesn't necessarily stop us from doing something stupid. We get the following:

$$t_{17,0.00000005} = 8.36$$

So now what happens? Our comparison becomes:

Because $|t^*| = 6.73 < t_{17,0.00000005} = 8.36$, *we fail to reject* H_0.

We obviously made a type II error. Seriously - the average litter size for lions is *not* 9.

As we decrease the probability of a type I error (by picking a small value of α), we increase the probability of a type II error and vice versa. This is a fundamental problem in hypothesis testing: how do we balance these two errors?

If possible, we look at the *cost* of making a mistake. Let's suppose we develop a new medicine for AIDS. We need new medicines as the current medicines can be complicated to take and often are very expensive. One possible way of treating AIDS might be to develop a medicine that increases T-cell production. That gives us the following:

Let μ = change in T-cell production.

Then:

$H_0 : \mu = 0$

$H_1 : \mu \neq 0$

Notice that if $\mu = 0$ this implies that the medicine does *not* work since there is no change in T-cell production.

> This example is a bit simplistic and does not completely reflect the biology of HIV/AIDS. Also, we should mention that we're obviously interested in an *increase* in T-cell production, not a decrease. This would give a *one sided hypothesis test*, a concept we will cover later.

In hypothesis testing we almost always want to *reject* H_0 since that means there is a change, or there is a difference, etc. (the status quo, implied by H_0, is not interesting).

> In other words, we have:
>
> H_0 : medicine does not work (the status quo)
>
> H_1 : medicine works

So what does all this mean for our "new" AIDS medication?

Let's first choose an value of α that is too high and pick $\alpha = 0.3$ (you should *never* use a value of α that high!). What happens?

> The probability of making a type I error is now quite high (30%). In other words, we have a 30% chance of deciding the medicine works, when *it really doesn't*.
>
> What happens if we make this medicine and give it to people (but the medicine doesn't work)?
>
> A lot of people might start dying since the medicine is not effective (despite the drawbacks, we do have medicine that works).

Now let's choose a value or α that is too low and pick $\alpha = 0.0001$. What happens?

> The probability of making a type II error is now quite high (we don't know exactly how high since we don't know β).
>
> What happens if we don't make the medicine (even though it really works)?
>
> Potentially we keep a cheaper and more effective medicine off the market *but* we are not killing people.

So in this scenario, we really should choose a small value of α. We don't want people to die as a result of our medicine not working.

In an introductory type class like this, we usually don't try to analyze the cost of making a mistake like this, but you should be aware of the consequences. Instead, we usually just pick a value of α. Commonly used values for α are:

0.1 0.05 0.01

We usually pick amongst these depending on how worried we are about making a type I (or occasionally, a type II) error. There is nothing wrong with choosing other values of α:

0.025 0.005 0.001

However, you really shouldn't pick a value of α that's absurdly small, and you really shouldn't go higher than 0.1 (there is nothing wrong with 0.1, but that's also a good limit).

You should always decide on your level of α ahead of time. Although it doesn't happen often, it is possible that not deciding ahead of time will let you do what you want. For example, let's suppose you have $n = 20$ and $|t^*| = 2.3$. You haven't decided on α, so you look in the table and find:

For $\alpha = 0.05, t_{table} = t_{.05,19} = 2.093$.

For $\alpha = 0.01, t_{table} = t_{.01,19} = 2.861$.

Your value ($|t^*| = 2.3$) is right between the two. If you haven't picked α ahead of time, you can how decide what you *want* to do.

If you want to *reject*, use $\alpha = 0.05$.

If you want to *fail to reject*, use $\alpha = 0.01$.

It's entirely up to you and now you can do whatever you want(?!?). This is *WRONG*. You don't get to pick the answer you want. This is one reason for always deciding on α ahead of time.

P-values

A very important concept in statistics is the idea of *p-values*. This idea ties in closely with hypothesis testing and is intimately connected with α. In brief, a *p-value* is the probability that you would have gotten the results you did or worse. Let's define this a bit more carefully:

> We assume our null hypothesis (H_0) is true. We then calculate the probability that we got the result we did. This is our *p-value*.
>
> Suppose, for example, that we had gotten $\bar{y} = 7.5$cm and $s = 0.5$cm for our finger length experiment with a sample size of $n = 25$.

We assume the null hypothesis ($H_0 = 8.4$cm) is true. What is the probability of getting $\bar{y} = 7.5$cm, $s = 0.25$cm with $n = 25$? This would be our *p-value* (in this case R tells us that $p = 3.69\text{e} - 9$.).

> If this probability (= *p-value*) is small, we don't believe H_0 is true.
>
> In one sense, all that α is, is our *cut-off* probability. If the probability of our result is less than α, we reject our H_0 because we don't believe it is true (it's too improbable).

We now have two ways of making a decision about whether or not to reject H_0. *They are identical in the sense that they will* ***ALWAYS*** *give you the same decision.* If you calculate a test statistic and reject, you will reject if you compare the *p-value* to α:

1) Compare your test statistic to the tabulated value (e.g., compare $|t^*|$ with t_{table}, and reject if $|t^*| \geq t_{table}$).

   OR

2) Compare your *p-value* with α and reject if *p-value* $\leq \alpha$.

One big advantage of *p-values* is that most software will always print out *p-values*. This saves you the trouble of having to use tables; all you need to do is compare the printed *p-value* with the value of α that you picked. This is true not just for *t*-tests, but for any statistical hypothesis test. Tables are usually only used in introductory statistics classes these days as software has made most of them obsolete (you're in an introductory statistics class, so you're still stuck with them!).

Let's try to illustrate *p-values* using another approach. We will introduce another hypothesis test called the *sign* test. To illustrate this, let's talk about lemurs (primates from

Madagascar) for a bit.

> In some species of lemur, females are dominant and lead the troop. Males will occasionally challenge females, but will usually loose. A researcher wants to verify that this is true. He goes out, and observes 7 aggressive encounters between males and females and finds the following results:
>
> | Encounter # | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
> |---|---|---|---|---|---|---|---|
> | Winner: | F | F | F | F | F | F | F |
>
> Where: **F** = female wins, and **M** = male wins.

So how do we analyze this? You shouldn't be surprised if we say it's With something called the sign test. Let's just do the test and then explain things as we go along.

First, what are our hypotheses?

> H_0: neither sex is dominant.
>
> H_1: females are dominant.
>
> > Note that this alternative hypothesis is actually one sided. We haven't looked at one sided alternative hypotheses yet; we'll learn about them later. For now, notice that we expect females to win (a two sided alternative here would say females OR males are dominant). A one sided alternative will also make our calculations just a bit easier.
>
> Let's pick $\alpha = 0.05$.

Now we assign a sign to each encounter:

> If the female wins, we'll use a $(+)$ sign.
>
> If the male wins, we'll use a $(-)$ sign.

So let's fill in the signs below the results:

Encounter #	1	2	3	4	5	6	7
Winner:	F	F	F	F	F	F	F
Sign	+	+	+	+	+	+	+

And we see that we have seven (+) signs, and not a single (−) sign. So the question becomes:

> What is the probability of getting seven (+) signs in seven encounters?
>
> For each encounter, we have two possibilities: either the male or the female wins.
>
> Our null hypothesis implies that *males and females are equally dominant* so the probability of a female (or male) winning is $p = 0.5$.
>
> *This is exactly the same as asking*: What is the probability of getting seven heads in seven tosses?
>
> We know how to do this! We use the binomial:

$$Pr\{7\ wins\ by\ females\} = \binom{7}{7} 0.5^7 0.5^0 = 0.0078125$$

Are you willing to believe that this is due to chance? Could females have won all 7 encounters due to chance? Do you believe the coin is fair if you get 7 heads in 7 tosses?

No! Because the probability of this outcome is absurdly small ($\approx 0.8\%$). In other words, the probability of females winning all seven times if both sexes are equally matched is 0.0078125.

> We say that the ***p-value*** $= 0.0078125$.
>
> Also notice that because $p\text{-}value = 0.0078125 \leq \alpha = 0.05$, we reject H_0.

So we can conclude that the sexes are not evenly matched and that females are dominant. We did this entirely by using *p-values*. We calculated our *p-value* directly and then compared this to α.

Calculating *p-values* for other types of tests, like our t-test is very difficult (but see below) so we use tables. There are also tables for the sign test, but we didn't need them in this example.

To reiterate, when we set $\alpha = 0.05$ what we are saying is that we are willing to make a mistake (type I error) 5% of the time. Therefore, if our *p-value* is less than 5%, we say that we don't believe our results are due to chance.

A comment on two sided tests (this will be discussed more thoroughly later):

As mentioned, the alternative hypothesis (H_1) was one sided. If we had used a two sided hypothesis (e.g., H_1: one of the sexes is dominant.) then calculating *p-values* becomes a bit more complicated. We need to calculate the probability of females winning seven times out of seven trials and *add* the probability of males winning seven times out of seven trials.

Why? Because the alternative hypothesis says *either males or females win.* It no longer says *females win.* In other words, males or females could have won seven times. So we need to do the following:

$$Pr\{7\ wins\ by\ females\} = Pr\{7\ wins\ by\ males\}$$

so we get:

$$Pr\{7\ wins\ by\ females\ OR\ 7\ wins\ by\ males\}$$

$$= 0.0078125 + 0.0078125 = 0.015625.$$

Notice that our *p-value* has gotten bigger (which is always true for two sided tests), but that it is still less than α.

p-value $= 0.015625 \leq \alpha = 0.05$ so we still reject H_0.

Suppose you don't have R handy, but really do want to get a *p-value* for a t^* that you calculate. As mentioned, this is difficult, but you can get an approximate *p-value* with just the *t*-tables. Let's use the example from before with $n = 20$ and $|t^*| = 2.3$. How do we get an approximate *p-value*?

Go into the *t*-table and find the row for $d.f. = 19$. now go across. You will find that our value of 2.3 between 2.093 and 2.539 (using the tables on our web page).

Now go to the top of the table and use the row for two sided. you will find that 2.093 corresponds to a probability of 0.05, and 2.539 corresponds to a probability of 0.02.

This means that our value of 2.3 must have a probability between 0.02 and 0.05, or:

$$0.02 < our\ p\text{-}value < 0.05$$

So we know our *p-value* is between 0.02 and 0.05.

If we use R, we can confirm that the actual *p-value* = 0.03295, which is between 0.02 and 0.05.

So while you can't get exact *p-values* using your tables, you can bracket the actual *p-value* with just the tables.

Some textbooks will use this bracketing approach for all hypothesis tests. In our example (with $n = 20$ and $|t^*| = 2.3$) we note that the *p-value* < 0.05. If we had picked $\alpha = 0.05$ this implies that since the *p-value* $\leq \alpha = 0.05$ we *reject* H_0.

Instead of comparing $|t^*| = 2.3$ with $t_{table} = t_{0.05,19} = 2.093$ and rejecting, this approach always uses a *p-value* (or approximate *p-value*) to make a decision about rejecting.

There is nothing wrong with this approach, but the approach outlined here (comparing t^* with t_{table}) is generally a bit easier.

You should, however, understand why these two approaches are the same.

Let's now suppose you reject H_0. The *p-value* can also be used to let you know how confident you are in having made the right decision.

You reject H_0 and get a *p-value* of 0.000067. This means you're really happy with your decision since the probability of getting the results you got by chance are absurdly small. You should feel very confident about your decision to reject.

You reject H_0 and get a *p-value* of 0.03. This means that yes, you get to reject (if $\alpha = 0.05$). But it does mean that the probability you got the results you did by chance is 3%. That's much higher than the previous example. You don't feel as confident this time (but you still get to reject!).

So let's summarize what *p-values* can be used for:

1. They can be used to make a decision about H_0. If *p-value* $\leq \alpha$ you reject.
2. They let you know how confident you are about your decision to reject. The smaller the *p-value* the more comfortable you are with your decision.

You should get in the habit of *always* reporting *p-values* if you can. Most journals and other scientific publications will require a *p-value* for the results of any statistical hypothesis tests.

Power

Another topic that needs to be addressed when discussing hypothesis tests is the concept of power. So far, we've mostly talked about α. What about β? Remember:

$$\beta = Pr\{do\ not\ reject\ H_0\ if\ H_0\ is\ false\}.$$

Which is the probability of making a type II error.

Let's now subtract this from 1. So we want 1-β:

$$1 - \beta = Pr\{reject\ H_0\ if\ H_0\ is\ false\}.$$

This is good! We want this probability to be as high as possible.

This is also called the *power* of a test:

$$Power = 1 - \beta = Pr\{reject\ H_0\ if\ H_0\ is\ false\}.$$

A test with more power will be better able to detect a false H_0. As we will learn, sometimes you will have a choice of two (or more) tests that test for the same thing. How do you pick which test to use?

Suppose you have two tests, A, and B. Both can test $H_0 : \mu = 8.4$ vs. $H_1 : \mu \neq 8.4$.

How do you pick between A and B?

You pick the test with the most power!

So how do you know which test has the most power? It turns out that calculating which test has the most power can be rather difficult and is beyond an introductory class like this. However, in most instances, some simple rules will let you know which test to use. We'll get back to this idea soon (particularly when we start discussing the Wilcoxon-Mann-Whitney-U test).

Why don't we "accept" H_0?

There are several reasons for this, in including some that are more mathematical in nature. But despite this, the main reason is actually fairly simple to understand:

We don't know if H_0 is true.

Just because we *fail to reject* H_0 doesn't mean we suddenly know the truth.

Let's use the example of finger lengths:

$H_0 : \mu = 8.4$

$H_1 : \mu \neq 8.4$

Suppose we now fail to reject. Does this mean we suddenly know that $\mu = 8.4$?? *Of course not!*

(We'd have to measure the finger lengths of every living person to know this).

So we never "accept" H_0 because acceptance implies we know the truth. The best we can say is we *fail to reject*.

If you wanted to make a stronger statement in support of the H_0 you could say something like *the data are consistent with* H_0 or something similar, but you still can't accept the null hypothesis (or worse, say it's true).

Summary of hypothesis testing

Finally, to finish this section, let's summarize how to do a hypothesis test. We'll follow this basic outline for the rest of the semester:

Decide on H_0 and H_1.

Decide on α.

Verify your assumptions (We have not discussed this at all yet).

Calculate a test statistic from the data.

Do a comparison:

Compare the test statistic to the tabulated value.

-OR -

Compare the *p-value* with α.

Based on this, you *reject* H_0 or *fail to reject* H_0.

So far the only test we've learned is the one sample t-test (see if you can fit what we learned into the above outline). But this outline will apply to all the tests we will be learning.

Yes, the details will vary (e.g., how to calculate the test statistic or which tables to use), but this basic outline will stick with us throughout the rest of the course.