# Hypothesis testing I

I. What is hypothesis testing?

[Note we're temporarily bouncing around in the book a lot! Things will settle down again in a week or so]

- Exactly what it says. We develop a hypothesis, then test it, and either reject it or not.

- In particular, we are talking about statistical hypotheses.

- For instance, we have some idea about index finger length in humans. We can test to see if our idea is correct.

Suppose I think that the true average index finger length is 8.4 cm. I want to test to see if this is true.

- Let's take a sample:

[get some people's index finger length!] n =

- mean = std. dev = std. error =

- Now let's come up with a hypothesis:

$H_0$: the average population finger length is 8.4 cm, or ($\mu = 8.4$ cm)

- the answer to this will be either yes (sort of) or no. If the answer is no, what are the alternatives?

- we need to put the 'no' answers into an "alternative" hypothesis. We have three possibilities:

First, a two-sided alternative (or non-directional)

$H_1$: $\mu$ is not 8.4 cm (or $\mu \neq 8.4$)

Second (& third), a one sided alternative (or directional)

$H_1$: $\mu > 8.4$ cm

$H_1$: $\mu < 8.4$ cm

- Important: we only pick one of these. For now, let's pick the first of these (two-sided).

- Now what? Well, does it make sense that if our 95% CI includes 8.4 then we could say:

    - We have no evidence to doubt $H_0$?

    - In other words, our $H_0$ seems reasonable, and the only reason $\mu$ might not be exactly 8.4 might be due to random chance. It's k8ind of like saying you got 7 heads purely by chance.

- That's sort of what we do, except that we don't directly calculate a CI, instead we proceed as follows (this is not given in your textbook):

    - we calculate the t-value:

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

    - then we look up the t value for whatever "confidence" we want using d.f. = n-1.

    - Finally, if our $|t^*|$ > than the t from the tables, we **REJECT** our $H_0$ and accept our $H_1$.

        - notice the absolute value symbol around t

    - On the other hand, if our $|t|$ < than the t from the tables, we "**FAIL TO REJECT**" our $H_0$.

- Pay close attention to the way things are phrased above - we'll get back to that soon.

- So what about our little finger experiment?

    - go through example.

- We just conducted our first hypothesis test. We skipped a LOT of the details. This was an example of a one-sample two sided t-test

    - a one sample t-test isn't terribly useful, which is why it isn't covered in your text.

    - why is it called a one-sample t-test?

        - because we have one sample (e.g., our classes' finger lengths)

        - because it uses the t-distribution.

- Note how this is sort of backwards from a CI. For a CI we would calculate:

$$\bar{y} \pm t_{\nu,\alpha} \times \frac{s}{\sqrt{n}}$$

- for the t-test, we calculate "our" t from the data, and then compare this to a critical value from the t-distribution.

- if a 95% confidence interval includes the value we're testing (in the finger example, if it includes 8.4 cm), then a t-test (at $\alpha = .05$) would "fail to reject" the $H_0$, and vice versa. This is always true.

II. Some details.

1) More on notation & error types (see also p. 257-258 **[252-254]** *{238 - 239}* in your text). Notice in the equation just above that the subscript for t includes both ν and α. ν we already know is our degrees of freedom (*n-1*, in this case). We already discussed the other value, but now we're ready to give it a more precise name: α. (alpha).

- In a simple way, α corresponds to the critical value that we look up in our t-tables. But more specifically:

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

- What does this mean?

- Suppose you decide to reject $H_0$:

- Why do you reject $H_0$??

- because you don't "think" it's true.

- but it might be!!!

- you can never be certain, all you can do is assign probabilities. You think the probability is less than 5% that $H_0$ is true, so you're willing to believe it's not true. But 5% is not 0%.

- In other words, α is the probability of $H_0$ being true when we think it isn't true.

- Here's another way of looking at it:

| | $H_0$ is true | $H_1$ is true |
|---|---|---|
| we decide $H_0$ is true | we're right | we're wrong<br>Type II error |
| we decide $H_1$ is true | we're wrong<br>Type I error | we're right |

- So $\alpha$ is the probability of making a type I error.

- What about type II? As it turns out, the probability of making a type II error is called $\beta$, but generally not possible to know what $\beta$ is.

  - but notice: if we try to minimize type I by choosing a really small $\alpha$ (say, 0.00000001), what happens to type II? It goes way up. Here's an example:

  Suppose we had information on litter sizes in lions (from a real study):

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

  In this case, $\bar{y} = 4.4$, s = 2.9 and n = 18.

  Now suppose our hypothesis for litter sizes in lions is that $\mu = 9$ (ignoring that this really isn't normally distributed).

  First we figure out our $H_0$: $\mu = 9$

  Then our $H_1$: $\mu \neq 9$

  We calculate our t-value:

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

  Now we look up t with 17 degrees of freedom and $\alpha = .05$ (a very standard thing to do):

$$t_{17,.05} = 2.110$$

  Because the $|-6.73| > t_{17,.05} = 2.110$, we reject our $H_0$.

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

But now suppose we're really worried about Type I error, so we choose $\alpha = .00000005$ (this is so silly it's not in your tables; I got this one using a calculator):

$$t_{17,.00000005} = 8.36$$

So now what? We don't reject $H_0$, and conclude that we have no evidence to show $H_0$ is wrong.

What did we probably do??? We probably committed a type II error. Seriously - don't you think that average lion cub size is probably not 9???

So as we decrease the probability of a type I error, we increase the probability of a type II error, and vice versa.

- this a fundamental problem in statistics. How do we balance these two errors??

- if we can, we look at the "cost" of making a mistake. For example:

We try out a new medicine for AIDS:

- Note that generally if we want to show a medicine works we need to reject $H_0$. Almost always we want to be able to reject the $H_0$.

- The reasons for this will become a bit more obvious as we learn about other tests. In any case, usually we do:

$H_0$: medicine does not work (the status quo)
$H_1$: medicine works

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

- if we set $\alpha$ too high (e.g., at 0.3), we run the risk of deciding the medicine works when it really doesn't (type I error).

- if we set $\alpha$ too low (e.g. at 0.000001), we run the risk of deciding the medicine does not work when it really does (type II error).

Which is worse?? Probably the first option. We want to be pretty sure our medicine works before we start giving it to people - because otherwise a lot of folks might start dying, {particularly since we already have a medicine for AIDS that is working}.

In case of the second option, the worst that would have happened is that we would not have a new medication for AIDS (which admittedly, we do need!).

- Kind of standard levels for α are:

  - .1, .05, and .01

  - we usually pick amongst these depending on how worried we are about making a type I error. BUT there's nothing wrong with picking other values (e.g., .025, .001, etc.).

  - Incidentally, if you pick something weird (say, .0368), that's also alright, but you're going to have to explain why you picked this (see particularly, the next section)!

- a little bit more about α (yes, you're probably sick of α by now, but we're almost done).

  - in general, you should decide on your level of α before doing the test.

    - Why??? Suppose you get a t-value that's between the cut offs for .05 and .01. What are you going to do?? Unless you decided on α ahead of time, you might decide to do WHAT YOU WANT TO DO, which is cheap, sleazy, crummy and just bad statistics. This is an example of why some statisticians have a bad reputation.

      - As an extreme example, you might calculate your t-value, and then pick a tabulated t from the table in such a way so that you always get the result you want, and never mind what the probabilities are!

  - Let's do a simple example. Suppose you have n = 20 and you get a $|t^*| = 2.3$.

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

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

  - Your value is right in between the two.

    - If you haven't made picked α ahead of time, you can now decide what you "want" to do. If you want to reject, use $\alpha = .05$. If you want to "fail to reject", use $\alpha = .01$.

    - It's entirely up to you, and you can do whatever you want!?!?

- This is *WRONG* and is *CHEATING*. You don't get to "pick" the answer you want!!

- You must decide on $\alpha$ ahead of time.

III) p-values

- a p-value is the *probability* that you would have gotten the result you did or worse.

- it all comes down to probability. What is the probability that we got the result we did ***if $H_0$ is true***?

- if the probability (= p-value) is small, then we say we don't believe $H_0$ is true.

- in one sense, all that $\alpha$ is, is our "cut-off" probability. If the probability of our result is less than $\alpha$, we reject the $H_0$ because we don't believe it (it's too improbable).

- if the p-value $\leq \alpha$ then you reject the $H_0$.

- so you actually have two ways of making the decision about $H_0$ (to reject or fail to reject).

- compare test statistic to tabulated value

OR

- compare p-value to $\alpha$.

- they are equivalent! It's irrelevant which method you choose. If you calculate a test statistic and reject, you will reject if you compare the p-value to $\alpha$.

- before we get confused, let's try something else. We'll introduce the sign test.

The sign test can be used when you can't use a t-test. For example, when you don't have measurements for your data.

Let's go through an example from your text (see problem 9.20, p. 370 **[9.22, p. 371]** *{8.4.6, p. 321}*):

A researcher is trying to figure out if one or the other of two subspecies of Junco (a little slate-gray bird seen around here in the winter time) is dominant.

How to do this?

Observe Juncos (obviously!), and see which one is dominant, and then "test" a hypothesis.

- which hypothesis? Let's set it up as follows:

$H_0$: neither subspecies is dominant.
$H_1$: one of them is dominant over the other.

let's also pick $\alpha = .05$.

Our researcher looked at 45 minute periods spread out over 8 days and wrote down how often each subspecies was dominant over the other. Here are the results:

# of times dominant:

| northern | southern |
|---|---|
| 0 | 9 |
| 0 | 6 |
| 0 | 22 |
| 2 | 16 |
| 0 | 17 |
| 2 | 33 |
| 1 | 24 |
| 0 | 40 |

Instead of doing anything else, let's just use probability to see if we can answer the question of whether or not one species is dominant over the other.

Suppose we decide for each day which species was dominant. If the northern species was dominant, we'll give that 45 minute period a (+), if the southern was dominant, we'll give that period a (-).

So let's fill in the above table:

| northern | southern | sign |
|---|---|---|
| 0 | 9 | - |
| 0 | 6 | - |
| 0 | 22 | - |
| 2 | 16 | - |
| 0 | 17 | - |
| 2 | 33 | - |
| 1 | 24 | - |
| 0 | 40 | - |

We have eight (-) signs, not a single (+) sign.

So the question is:

What is the probability of getting eight (-) signs?

*This is exactly the same as asking:*

What is the probability of getting eight heads in eight tosses?

Each time we have one of two possibilities: either the northern species wins or the southern species wins (if we assume they're equally dominant (*our null hypothesis above implies this!*), we get p = ½).

Well, we know how to calculate the probability that the southern bird won all eight times:

$$Pr\{8\,wins\,by\,southern\,subspecies\} = \binom{8}{8}.5^8.5^0 = .003906$$

Remember, that this is the probability if either bird has the same chance of winning (or that the coin is fair).

*Are you willing to believe that either bird has the same chance of winning?*

**No!** Because the probability of this outcome is absurdly small (0.4%). In other words, the probability of the southern bird winning all eight times if both birds are actually equally matched is 0.4%.

**This probability is our *p-value.***

If birds are not evenly matched, the only other possibility is that one of them does better. The southern bird obviously does better!

So we conclude that that "the birds are not evenly matched; the southern bird does better".

Note:

Technically we have to calculate the probability of the northern bird doing better as well and add that to the above probability:

.003906 + .003906 = .007812

Why? Because our alternative hypothesis doesn't say “the southern will win”, it says “either one or the other will win”, so we need both probabilities to do it right.

This should become more obvious when we do one sided tests.

So how does this compare to $\alpha$? Here's how:

$p = .007812 < \alpha = .05$, so we reject.

We set $\alpha = .05$. What we are saying is that we are willing to make a mistake (type I error) 5% of the time.

The probability of getting our result is .7812%, which is less than 5%, so we reject our $H_0$, and say that we don't believe our results are due to chance.

Sometimes it's hard to see where the probability is coming from (here it's easy).

- Now that we're done with this example, let's summarize:

In all statistical tests we can calculate:

The probability of the observed outcome if the outcome is due entirely to chance.

- this will be our p-value.

- We need to use a computer (or calculator) to get exact p-values.

If this probability is really really small, we say the results are not due to chance. Something “real” is happening (for example, the southern bird is stronger!).

So how can we calculate a p-value?

- if all you have is tables, all you can do is an approximation.

- if you have R, or a more sophisticated calculator, you can get p-values that way.

- lets do another example. Let's assume that $n = 20$ and $|t^*| = 2.3$ (a bit similar to what we did on one of the previous problems above).

Using the table, we see that the value of 2.3 (and 19 d.f.) is between the columns for $\alpha = .04$ and $\alpha = .02$.

What, exactly, does the t-value of 2.205 mean (for $\alpha = .04$ and 19 d.f.)?

- it means that the probability of getting a $t^*$ of 2.205 (or worse) is exactly .04.

- remember, $\alpha$ is the probability that we're willing to live with. If $\alpha = .05$, we're saying that if the probability of our result is less than .05, we reject.

So what is our probability if $t^* = 2.3$?

- since 2.3 is between the columns for $\alpha = .04$ and $\alpha = .02$, our probability = p-value is between .04 and .02. Or we could say:

$$.02 < \text{p-value} < .04$$

- note that we're almost never interested in what our p-value is greater than, only what it is less than.

- a lot of books make a big fuss about bracketing p-values (which is what we did above), but almost always we're just interested in p-value < .04 (or whatever).

Let's now use a calculator (or R) to get an exact p-value:

p-value = probability = .03295

Which is between .02 and .04.

Remember - this means that the probability of getting $t^* = 2.3$ is 0.03295 (if $\alpha = .05$ we reject).

The nice thing about p-values is that all statistical software (including, obviously, R) will give you p-values anytime you do any test.

This means you never have to look up anything in tables, or even remember what kind of test it is (well, you really should remember that!).

You just look at the printout, and compare the p-value to your value of $\alpha$, and you're done (if p-value $\leq \alpha$ reject)!

- Another comment on p-values:

Suppose you reject. You can also use a p-value to tell you how confident you are that you made the right decision.

For example:

- if our p-value is .00001, then we're very happy. The probability of getting what we got by chance is absurdly small, so we're very confident we made the right decision.

- if our p-value is .03, then we might not be that happy, even if we decided on $\alpha = .05$. We still get to reject, but .03 is not as small as the example above.

(but we're still happy enough)

- You should get in the habit of always reporting p-values if you can (though we won't do much of this in class except with R).

See sections 7.4 and 7.5 **[same in 3rd]** *{7.2 and 7.3}* in your book for more on p-values if you're confused.

IV) Power of a test (see also p. 259 **[254]** *{240}*).

Remember:

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

which is the probability of making a type II error.

Now, what is 1-β??

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

This is good! Obviously, we want to reject a false $H_0$ as "much as possible". This is also called the "power" of a test. A test with more power will be better able to detect a false $H_0$. Your text has a good example - more power is analogous to better resolution in a microscope. It's better able to detect true differences.

- different tests might have different "powers". We'll discuss this a little more when we do the Wilcoxon-Mann-Whitney test, and actually be able to see how power might work.

- if you're really interested, check out Appendix 7.1, p. 600 **[623]** *{574}*.

V) Concluding remarks.

- A summary of doing a hypothesis test (a basic outline that we'll follow all the way to the end of the semester):

  - Decide on $H_0$ and $H_1$.

  - Decide on $\alpha$.

  - Verify your assumptions (we haven't discussed this step yet!).

  - Calculate a test statistics from the data (t in our example).

  - Compare this test statistic to tabulated values (t-tables) and reject or "fail to reject" $H_0$.

    - Or compare the p-value with $\alpha$.

  - (we will repeat this list many, many times)

- Why don't we "accept" $H_0$??

  - Several reasons. Some are mathematical in nature, but here basically it comes down to:

    - We don't know if $H_0$ is true. $H_1$ might be true, even if we decide not to go with $H_1$ (If, for example, $\mu$ is actually close to, but not exactly, our hypothesized value, then it might be very difficult to figure out that $H_0$ is actually not true).

      [e.g., a more concrete example: suppose we test $H_0$: $\mu = 10$, $H_1$: $\mu \neq 10$, and $\mu$ is actually 10.1. Would we be able to detect this?? We'd need a small variance and/or a very large sample size].

    - So all we say is "we fail to reject $H_0$, and have no reason to doubt $H_0$ is true".

      - if you want to say something a bit stronger, you could try "the data are consistent with the H0", but that's about as much as you really can say.

      - you can never "prove" the $H_0$ is true.

      - we'll get back to this when we do two-sample tests.