Distributions for means and statistical tests

The last set of notes covered a fair amount of material related to probability and distributions.

For example, we assumed we had a variable (y), and we want to know some probability:

$\Pr\{Y < 6 \text{ feet}\}$ for a basketball player

or

$\Pr\{Y < 1500 \text{ gm}\}$ for brain weight of Swedish men.

In both cases, Y was assumed to have a normal distribution (we say $Y \sim \mathrm{N}$, which means "Y has a normal distribution")

We then proceeded to use the normal tables (or R) to look up the probabilities we were interested in.

Now we want to discuss the distributions of means and how to answer probability questions for them.

For example, we want to know:

$\Pr\{\bar{Y} < 6 \text{ feet}\}$ for basketball players (i.e., what is the probability that the average of a group of basketball players is less than 6 feet)

or

$\Pr\{\bar{Y} < 1320\}$ or, the probability that the average brain weight for a group of Swedish men is less than 1320 gm.

To do this, we need to know three things:

1) What is the distribution of $\bar{Y}$?

2) What is the mean of $\bar{Y}$?

3) What is the standard deviation of $\bar{Y}$?

To jump right into the answers:

1) a) If $Y \sim \mathrm{N}$, then $\bar{Y} \sim \mathrm{N}$
   b) If Y is not normally distributed, then $\bar{Y} \sim \mathrm{N}$, ***if n is big enough.***

   (b) is a direct result of the CLT

2) The mean of $\bar{Y}$ is μ

3) The standard deviation of $\bar{Y}$ is $\frac{\sigma}{\sqrt{n}}$.

(2) hopefully is fairly obvious, (3) requires a bit of math to explain; let's just take it as a given.

Why did we need to know the mean and standard deviation of $\bar{Y}$?

Because we need them to calculate our probabilities. We need to get our z values, and to do that we need to convert $\bar{y}$ into z:

$$z = \frac{\bar{y}-\mu}{\sigma/\sqrt{n}}$$

So now we can answer our questions from above:

Let's try the brain weight example first. Suppose we had 15 men. We know that μ = 1,400gm, and σ = 100 gm:

$$z = \frac{1{,}325 - 1400}{100/\sqrt{15}} = -2.905$$

so we go into our tables and look up -2.91 and get a probability of 0.0018.

Let's try the other one:

What is our μ? σ?

For the brain weights above, do we really know μ or σ?

Of course not! So we use $\bar{y}$ and s for μ and σ.

The only problem with this is that if we use s instead of σ we are no longer calculating z:

$$z \neq \frac{\bar{y}-\mu}{s/\sqrt{n}}$$

Why not?

Because there are two random variables on the right side of this equation, $\bar{y}$ and s. It is important to realize that σ is NOT a variable. It is a constant (we don't know what it is, but it is a constant).

Because s does not have a normal distribution, "z" no longer has a normal distribution, so we need to do something else.

(Yes, if n gets big enough, we can use the CLT, but that doesn't solve the problem for small n).

So, since we almost never know μ and σ we proceed as follows:

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

Instead of using z, which has a normal distribution, we use t, which has a t-distribution.

Now, suppose we took a sample of 10 basketball players, and we calculate:

$\bar{y} = 76$ inches and $s = 3$ inches.

We can substitute $\bar{y}$ for μ in the equation above without having to change anything, so we get:

$$t = \frac{72 - 76}{3/\sqrt{10}} = -4.22$$

Now we just look up -4.22 in the t-tables (not the normal tables) and get a probability of less than .0025.

> Note that the t-table is arranged the opposite of the normal table. Probabilities are across the top, and the actual t values are in the table.
>
> We also need to know our sample size.
>
>> There are an infinite number of t-distributions. We pick the one we want based on the size of our sample:
>>
>> d.f. = $v = n - 1$

We can also use R to get probabilities (since we can't get the exact probability using our table):

```
pt(-4.22,9,lower.tail = TRUE)
```

This says to give us the probability for a variable with a t-distribution and a value of -4.22. The “9” tells R how many degrees of freedom our distribution has, and the lower.tail = TRUE statement tells R that we want the probability in the lower tail (remember, we want the probability that the average height for 10 basketball players is less than 72 inches).

Now we can use this to lead us directly into hypothesis tests:

Let's consider the following idea.

Suppose I thought the true average height of people at GMU was 72 inches.

How could I test this based on what we've learned so far?

Let's take a sample:

$\bar{y}$ = s = $s/\sqrt{n}$ =

Incidentally, that last quantity is also called the "standard error", abbreviated SE, or, in this case, $SE_{\bar{y}}$.

Now does anyone have any ideas?

Let's do:

H_0: $\mu = 72$
H_1: $\mu \neq 72$

Now let's calculate the "probability" that our null hypothesis (H_0) is true:

do on board using equation for t.

(Get probability with `pt(t,9,lower.tail = TRUE)*2`*)*

Now we know the probability that H_0 is true. What do we decide???

Introducing α:

We select α, and based on this value we make our decision:

if probability $\leq \alpha$, we say the null hypothesis is ***not*** true.

otherwise we say that the null hypothesis ***might*** be true (we ***never*** say that it is true).

In statistical language, we ***reject H_0*** or ***fail to reject H_0***.

α is the probability of making a mistake. In other words, it's the probability that we reject the null hypothesis if the null hypothesis is actually true:

α = Pr {reject H_0} if H_0 is true

Further discussion of α:

	H_0 is true	H_1 is true
we decide H_0 is true	we're right	we're wrong Type II error
we decide H_1 is true	we're wrong Type I error	we're right

So, as mentioned, α is the probability of making a type I error.

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

Notice: if we try to minimize type I by choosing a really small α (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.

Notice our rejection rule this time:

Instead of getting the probability, we just compare our value of t (t^*) with t_{table} at a specified probability (i.e. at α).

This is much easier (usually) if you don't have a statistical package handy.

Our rule becomes:

if $|t^*| \geq t_{table}$, reject

otherwise fail to reject

The conclusions of our test seem 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 R):

$$t_{17,.00000005} = 922$$

(note: try "`qt(0.000000025,17)`" from the R command line)

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 α 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 α 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:

0.1, 0.05, and 0.01

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.).

A little bit more about α

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 α ahead of time.

P-values:

We've already discussed this. It's simply the probability that we (can) calculate when we look up the probability associated with our t^*.

It's what we did above.

Specifically:

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

p-values can be used when we make our decision (reject or fail to reject), but they can also be used to tell us how comfortable we our with our decision:

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.

Iif 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)

note that your book points out that if possible, you might want to consider doing another study (maybe with a larger sample size) to see what happens if your p-value is really close to α.

You should get in the habit of always reporting p-values if you can.

Power of a test

Remember:

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

which is the probability of making a type II error.

Now, what is $1\text{-}\beta$??

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

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.

See section 6.3(d), p. 82, in your text for more on power and what affects power.

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 α.

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 α.

(we will repeat this list many, many times)

Why don’t we “accept” H_0 (or why can't we prove H_0 is true)??

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, μ 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 μ 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 H_0”, 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. It'll also be particularly important when we do ANOVA (and multiple comparisons).