10. Hypothesis Testing

In this chapter, we continue to study methods for making inference about a population using our sample. In chapter 9, we used CIs, essentially, to answer questions of the type, "What is the population mean?" In this chapter, we discuss hypothesis testing for a single population parameter (e.g. pop. mean). Hypothesis testing answers questions of the type, "Is the population mean equal to (greater than, less than) 20?" The last few chapters prepared us for dealing with the t-distribution (and Z distribution), the basis of CIs and hypothesis testing for a pop. mean. You can conduct hypothesis tests for other pop. parameters, but would need to study other distributions. We'll stick to CIs and testing for pop. means unless otherwise indicated.

Let's "86" the subsection numbering for this chapter.

We'll start off with an example just to get some ideas, then we'll discuss some details.

E.g. As an archeologist, you're studying the femur bone lengths of fossils unearthed from several archeological sites. Archeologists use femur bone lengths to help distinguish between dinosaur species. Suppose you calculate a sample mean femur bone length, $\bar{x} = 8.72$cm (n=25). You're wondering if the bones are from a familiar species, which you've coded as Species A, or if you've found a new species. Species A, well known to archeologists, is known to have a pop. mean femur bone length, $\mu = 8$cm. If $\bar{x} = 8.72$ is "close" to 8, you tend to think the bones you've measured are those of Species A. If $\bar{x} = 8.72$ is "far" from 8, you tend to think that the

bones are not those of Species A. Suppose that we also know the distribution of femur bone lengths to be reasonably unimodal and symmetric (we'll assume normality).

We'll use a t-distribution (perhaps a Z-distribution) to help us quantify "close" and "far".

Some pictures (assuming, for now, pop. s.d. = σ= 1):

Femur bone lengths

Sample mean femur bone lengths

What does the above picture suggest about the sample mean value, $\bar{x} = 8.72$? Let's calculate some probabilities. Assume for the sake of argument, that your dino bones do come from the same population as those of Species A (we don't know this, we're just assuming this so we can proceed in a reasonable manner). What's the prob. of having observed a sample mean value at least as large as 8.72? (Note: use z-tables if we really know $\sigma = 1$ (unlikely) or use the t-tables if we're estimating σ by the sample standard deviation, s=1).

- $P(\overline{X} \geq 8.72 \mid \mu = 8)$

What's the probability of observing a sample mean as extreme as 8.72 (i.e., either 8.72 or greater, or 7.28 or smaller)?

- $\boldsymbol{P}(\left|\overline{\boldsymbol{X}} - 8\right| \geq 0.72 \mid m = 8)$

A bit more formally now...

- Define a hypothesis to be a statement about a population parameter.

E.g. Some hypothesis in terms of pop. parameters.

- A hypothesis test (or significance test) is simply a formalized procedure for deciding the veracity of 2 complimentary hypotheses about a population parameter. These 2 complimentary hypotheses are called the null hypothesis and alternative hypothesis.

E.g. Hypothesis notation:

The null hypothesis is assumed to be true (for the sake of argument). Usually (not always), the alternative hypothesis is "what you want to 'prove'" and the null is its complement . The idea here is to be conservative: don't start off assuming the thing you want to prove; assume it's not true (the null) and evaluate the evidence (using probabilities) that leads to rejecting the null (in favor of your alternative) or not rejecting the null. It's like "proof" by contradiction. This approach is

sometimes compared to a court case where the defendant is assumed innocent (null), unless evidence supports the rejection of that assumption in favor of guilty (alternative).

E.g., Let's do a hypothesis test using the Dino Bones example (We'll perform a 2-sided test (Section 10.2)).

- State the null and alternative hypotheses

- Calculate a test statistic (we'll use a t, although you could use a Z if you know σ (not likely) or your sample size is "large")

- Calculate a p-value

- State your conclusions. Based on the evidence (p-value), reject or do not reject the null. Or, just use the p-value as a measure of evidence against the null hypothesis (i.e., you may not actually have to make a decision between the two alternatives)

If we must make a decision, for what p-values do we reject/not reject? That is, what's our "cut-off" p-value? Does it make sense to reject the null for small p-values or for large p-values? In this case, we compare the p-value to the significance level of the test, denoted α.

- The significance level, α, of the test is a pre-determined (i.e., before you see your data) value below which your p-value will cause you to reject the null.
- "Significant", "highly significant", "significant at the α-level"

When actually making a decision to reject or not to reject the null hypothesis (rather than just use the p-value as a measure of evidence against the null), we might possibly make a wrong decision. That is, we may reject the null hypothesis when we shouldn't, or we my not reject the null hypothesis when we should. Or, we may make a correct decision. Since we don't know the population parameter about which we've constructed

our hypotheses (here, the pop. mean), we never know if we're right or wrong when we make our decision. But, we can calculate probabilities of being right or wrong if we make some assumptions about the parameter value.

The table and picture below illustrate everything you want to know about hypothesis tests (well, okay, not everything, but they go a long way to explain things). We use a 1-sided test in the picture just for illustration.

Decision	Reality	
	H_o True	H_o False
Do not reject H_o	Correct Decision P(not reject H_o | H_o True)=1-α	Type II Error P(Type II Error)= P(not reject H_o | H_o False)=β β is also called the Type II error rate
Reject H_o	Type I Error P(Type I Error)= P(reject H_o | H_o True)=α α is also called the Type I error rate	Correct Decision P(reject H_o | H_o False)=1-β 1-β is called the power of the test

Note on error rates and power: When we calculate error rates and power, we need to assume a distribution. Here, that means we need to assume a pop. mean (if using a Student's t distribution) or pop. mean and pop. s.d. (if using a standard normal (Z) distribution). Most of the illustrations you see for error rates and power (including POB) use a Z distribution for the sake of simplicity, but, in reality, we typically use a Student's t distribution for testing. No matter, the concepts or errors and power are the same whether using a t or Z.

Notice that the Type I error rate, α, is the significance level of the test. Also, since it's the probability of an error, we'd like to make this small. But, we can't make it too small because as α get smaller, then, all else being equal, β (another error probability) gets large (or, equivalently, we lose power)—the power applet (URL below) is a good way to see this. α=0.05 is by far the most commonly used value. The reason for this is simply due to established convention, and it should not be taken as "the" α value. Other "typical" α values are 0.1 and 0.01. In practice, often α is set at some reasonable level (e.g. 0.05) and β is ignored. This is not good practice! Why? Ideally, we would set α at some value and then collect a large enough sample size, n, so that β is low (or, equivalently power = 1-β is high).

Recall that we compare our p-value to α and reject H_o when the p-value is smaller than α. Also, remember that the p-value is the probability of observing a test statistic (sample mean or (after standardizing) a t-score or z-score) at least as extreme (high or low or both depending on 1-sided test or 2-sided test) as the one we calculated (assuming H_o is true or "under H_o" or

"given H_o"). Thus, they're both probabilities calculated under H_o. But, the p-value is not α. α is determined before you see your data, the p-value is based on your test statistic.

Check out the applet on power:
http://www.stat.sc.edu/%7Eogden/javahtml/power/power.html

The distributions in the applet are normal distributions (pop. s.d., σ, is known). You can think of them as distributions of the sample mean under different pop. mean values and with the same variability (s.e.), but each distribution has been "partially standardized" in some sense. That is, the value of the pop. mean under the null hypothesis has been subtracted from both distributions but division by the standard error has not been performed. Some questions to answer when looking at the applet:

- Which corresponds to the null and which to the alternative?
- What's the α-level illustrated in the applet?
- How many ways can we affect a change in power (I can think of four ways, 3 of which can be illustrated by the applet)
- All else the same, which is more powerful, a one-sided test or a two-sided test?
- Given that a one-sided test makes sense, which would you perform, a one-sided test or a two-sided test?

Some summarizing:

It's always nice to be organized when we perform a hypothesis test. Let's use the following steps whenever we report a hypothesis test. We'll assume we've chosen an α-value if we must make a decision

1.) State the null and alternative hypotheses using notation and in words
2.) Calculate a test statistic
3.) Determine a p-value
4.) Formally reject or do not reject if you are making a decision, and state your conclusions in terms of the problem.

E.g., Dino Bones, continued. (State null/alt, calc. test stat, calc. p-value, use p-value directly or compare to α and reject/not reject, state conclusions.)

Critical values. We've already seen these, but here's a synopsis

When faced with a decision to reject or not, we compared the p-value with a pre-determined α-value. But, for any α-value (a tail(s) probability) we can find the associated t-score (or z-score) and sample mean value, and vice versa. These values associated with the α-values are called critical values, because they're the values on which your decision hinges. The p-value (also a tail(s) probability) is associated with our test statistic. That is, for every p-value we have a test statistic, and vice versa. So, we could use test statistics and critical values or p-values and α-values to do our tests; either way, the result is the same.

Sample Size Estimation

We talked about sample size estimation when doing CIs. There we only talked about using α (yes, it's related to the α that we're using now) to help us in estimating sample size. Now we have β (equivalently, power) to consider in our estimation of sample size.

Here's the basic idea:

- choose α and β
- decide a null mean value, μ_o, and an alternative mean value, call it μ_1,
- estimate the common pop. s.d., σ, using previous study
- get two expressions for the sample mean: one expression satisfies the type I error rate, α, and is derived using the null distribution; the other expression for the sample mean

satisfies the type II error rate, β (or power=1-β), and is derived using the alternative distribution

- set the two expression equal to each other and solve for sample size, n
- see Section 10.6, POB

Other Stuff:

Power curve--see page 245, POB

Can you see the correspondence between CIs and Tests? For example, consider a 2-side test and a 2-side CI. If you constructed a 95% CI for the pop. mean and, for a reject/not reject test situation, your null hypothesized pop. mean value, μ_o, falls outside the interval, then you would reject at the 5% level. A similar relationship exists for 1-sided tests.

One last example (a 1-sided test):