Hypothesis Test Summary

I. General Framework

Hypothesis testing is used to make decisions about the values of parameters. Parameters, you'll recall, are factors that determine the shape of a probability distribution. The Normal probability distribution, for example, has two parameters. The mean determines the center, and the standard deviation determines the spread. The binomial distribution also has two. The sample size, n, and the probability of success on a single trial, p. In our discussion we focused on hypotheses on just two parameters: the mean (of a normal) and p (of a binomial). Later you'll see hypotheses tests concerning the standard deviation, as well as parameters regarding mathematical models.

1) State hypotheses

The null hypothesis will always be of this form:

the parameter is equal to *this value*

where this value is an actual number, say 0. Sometimes, if we don't want to specify which number it is, we write a greek character and subscript it with a 0.

The alternative hypothesis will take one of three forms:
a) one-sided
the parameter is greater than what the null hypothesis said.
b) one-sided
the parameter is less than what the null hypothesis said.
c) two-sided
the parameter is not equal to what the null hypothesis said.

NOTE: Sometimes, the hypotheses make statements about functions of parameters. For example, in words we might say "the means of the two groups are the same". But the null hypothesis would state this as "mean1 - mean2 = 0" and the alternative might be, for example, "mean1 - mean2 <> 0 ".

2) Choose a Test Statistic
The test statistic is based upon an estimator of the parameter mentioned in the null hypothesis. For example, if we're trying to understand something about the mean, we would use the average of our sample. If we're trying to make a statement about the parameter p, then we might use the proportion of successes in our sample.

Eventually, we will have to be able to calculate probabilities associated with these estimators. For this reason, certain test statistics are used again and again because their distributions are known (if certain assumptions hold true.)

3) Calculate the p-value.

The null hypothesis tells us what value the parameter has, and therefore tells us something about what value we should see for our test statistic. Of course, because the test statistic is a random number, we probably won't see exactly the value the null hypothesis says we should, even if the null hypothesis is right. Therefore, we need to understand how far off our observed value is from what we expected.

For example, the null hypothesis says that a coin is fair, and therefore p = .5. And therefore, in 100 flips we should see 50 heads. But we see 55. Is this too many?

To answer a question like that, we need to know what a "typical" result is. And we need to know what an extreme result is. If I told you that quite often, if you flip a fair coin you'll get between 45 and 55 heads, then you wouldn't find 55 remarkable. But if I told you that in my entire career of flipping fair coins, only once have I seen a result so far out of line, then you might be suspicious of that coin.

Let's do some math here. If X represents the number of heads in 100 flips of a fair coin, then E(X) = 50 and SD(X) = sqrt(np(1-p)) = sqrt(100*.5*.5) = 5. This tells us that, very roughly speaking, 68% of the time we should get 50 +_ 5 heads (45 to 55), and 95% of the time 50+- 10 heads (40 to 60). So probably a reasonable person would be suspicious if he or she saw fewer than 40 or more than 60 heads in 100 flips. But that person shouldn't be alarmed by, say, 45 heads.

The p-value is the probability of getting an outcome as extreme or more extreme than the observed outcome, ASSUMING THE NULL HYPOTHESIS IS TRUE. If the p-value is small, this weighs against the null hypothesis, because it says that the observed outcome is quite rare, and therefore unlikely. A large value for the p-value weights in favor of the null hypothesis, because it says that the observed outcome is pretty much what the null hypothesis said you would see.

4) Making a decision

Before you can make a decision, you need to set a value for alpha, a.k.a. the significance level. The significance level is defined to be the probability that you reject the null hypothesis even though it's true. This is, obviously, a bad thing to do, and so you want alpha to be small. Typicall, alpha = .05 (or 5%) is taken to be an acceptable level.

The "decision rule" then, is to reject the null hypothesis if the pvalue < alpha. Following this rule ensures that you will mistakenly reject the null hypothesis at most alpha*100 % of the time.

5) Semantics

Never say "We *accept* the null hypothesis". Should your p-value turn out to be greater than alpha, then your conclusion is that "there's no evidence to reject the null hypothesis" or "we fail to reject the null hypothesis". This is different from concluding that the null hypothesis is true.

A hypothesis test will not tell you if the null hypothesis is true. It is quite possible that if you had a larger sample size, you might reach a different conclusion. The best you can say is that that the null hypothesis provides a sufficient explanation of the outcome, and there's no evidence for rejecting it.

II. One-Sample Tests

By "one sample" I mean that we're concerned with a single variable from a single population. An example of this is the body temperature example discussed in class. Another example is the coin-flipping example discussed above.

There are only a few tests that we covered in this situation.
First Decision: The first question you ask is : are we testing the mean? Or are we testing a proportion or probability of success? If "the mean", then tests (1) and (2) are for you. If you're testing a proportion or probability, then see (3a) and (3b).

Tests of means

To decide between tests (1) and (2) ask yourself: do I know the SD of the population?

1) Z-test
Use this when:

a) You're examining the mean of a population.
b) The SD of the population is a **known** value.
c) The population is known to follow a normal distribution.

If (c) turns out to not be true, then the Z-test is still an acceptable approximation in many situations. Certainly, though, it pays to have a large sample size.

The test statistic is: Z = (Xbar - mu0)/ (sigma/sqrt(n))

where

Xbar is the average of the sample.
mu0 is the value that the null hypothesis says the mean has
sigma is the known SD of the population
n is the sample size

The sampling distribution of Z is N(0,1). Why? Because each of the observations, X1, ..., Xn is N(mu0, sigma). Therefore Xbar is also normally distributed. The mean of Xbar is E(Xbar) = mu0 (you should be able to derive this using the rules for means of linear combinations) and the SD is sigma/sqrt(n) (and you should be able to derive this, too.) Therefore, Xbar is distributed like a N(mu0, sigma/sqrt(n)) random variable. Which means that if we subtract its mean and divide by its standard deviation, we have standardized it and Z is therefore N(0,1).

Example: The heights of adult women in the U.S. varies according to a normal distribution with standard deviation of 3 inches. The mean height is 63.5 inches. An anthropologist claims that women in Los Angeles are taller than in the rest of the country. She collects a random sample of

100 women. She finds that the average height of the women in her sample is 63.7 inches. Does this support her hypothesis?

H0: mean = 63.5
Ha: mean > 63.5

Our observed value for Z is (63.7 - 63.5)/(3/sqrt(100)) = .6667. That is, the observed value is about 2/3rds of a standard deviation above where we expected it to be.

The null hypothesis, note, tells us that we expected to see a test statistic of 0, but instead we saw .6667. Is this a typical outcome?

The p-value is P(Z > .6667). Because we know that Z is N(0,1), we can find this probability from Table A.3 in the book: P(Z > .6667) = 1 - P(Z <= .6667) = 1 - .7157 (approximately)
so this is about .28. This means that such values occur fairly often: about 28% of the time. Apparently this outcome is in line with the null hypothesis.

You might not think 28% is all that often. But remember we agreed that anything more than 5% was “often”. Since 28% > alpha (which we set at 5%), we do not reject H0 and conclude that there is no evidence that women in LA are taller than in the rest of the US.

Question: Why do we need Z? Why not just use Xbar?
Answer: You can if you want. But using Z saves you a step. Here’s why. Let’s use Xbar as a test statistic. We know the distribution of Xbar is N(mu0, sigma/sqrt(n)) = N(63.5, .3). Our observed value was 63.7. So our p-value is P(Xbar > 63.7). But to use the table A.3, we must first standardize this, which means
P(Xbar > 63.7) = P((Xbar - 63.5)/(3/sqrt(100) > (63.7 - 63.5)/(3/10))
= P(Z > .6667)
So you see, we’re right back where we started.

2) T-test
Use this when:

a) You’re examining the mean of a population.
b) The SD of the population is **unknown**.
c) The population is known to follow a normal distribution.

If (c) turns out to not be true, then the T-test is still an acceptable approximation in many situations. Certainly, though, it pays to have a large sample size.

The test statistic is T = (Xbar - mu0)/(s/sqrt(n))
where

Xbar is the average of the sample
mu0 is the value that the null hypothesis says the mean has
s is the standard deviation of the sample

The sampling distribution of T is a t-distribution with n-1 degrees of freedom.

Example: bodytemperature
H0: mean = 98.6
Ha: mean <> 98.6

Our data show:
sample size = 130
sample average: 98.3
sample SD: 0.76

Observed test stat = (98.3 - 98.6)/(.76/sqrt(130)) = -4.5

Is this extreme? The null hypothesis says to expect values close to 0. Is -4.5 close to 0? In particular, if the null hypothesis is true, how often can we expect such extreme values?

The p-value is $P(T > 4.5) + P(T < 4.5)$. We can use R to get an exact value:

```
> 2*pt(-4.5,129)
[1] 1.499910e-05
```

Or we can look up possible values in Table A.4.

But notice there's no row for df = 129. There is for 100 and infinity. You'll see that it doesn't matter much. We reject our null hypothesis if p-value < .05. If df = 100, according to the table, this would be the case for any test statistic that was bigger than 1.98 or less than -1.98. (Draw a picture of the distribution to see why.) If df = infinity, then we would reject if the test statistic were bigger than 1.96 or less than -1.96. Now our degrees of freedom are 129 which is (much) closer to 100 than to infinity. But it doesn't really matter, because -4.5 is much smaller than the most conservative of the choices. So we know, no matter how many degrees of freedom, that the p-value of minus 4.5 is less than .05. Therefore, reject the null hypothesis.

Tests of Proportions

Ask yourself: do I have a small sample size? Or large? If small, use (3a), if large (3b).

3a) Binomimal Test: exact
Use this if

a) You're in a binomial situation (e.g. counting the number of successes in a fixed number,n, of indpt. trials)

b) you're conjecturing about the value of p where p is the probability of a sucess on a single trial.

c) The sample size is “small”. How small depends on your patience and access to a computer.

Test Stat: X, where X is the number of successes.

Sampling distribution: binomial

This is a very straight-forward test. You calculate the p-value using the binomial density formula. This is given for n ranging from 2 to 25 on Table A.1. But you can also calculate it using R. It can be tedious, because it means adding up lots of values. But on a computer this is easy.

Example: Someone claims that if a coin is spun, rather than flipped, than the probability of Heads is no longer .50. I did this 10 times and got 4 heads. Does this support the claim?

This is a binomial situation because

a) number of trials fixed: 10
b) trials indpt.
c) each trial success (heads) or failure
d) we’re counting the number of successes

H0: p = .5
Ha: p <> .5

Our observed value is x = 4. Because it’s a two-sided alternative, our p-value is
P(X <=4 OR X >=6) = sum(from 0 to 4) of (10 choose i) .5^{10} + sum(6 to 10) (10 choose i) .5
10

If using A.1, then find the page for n = 10, go to the column for p = .5. P(X <=4) = .3770. And similarly,
P(X >=6) = 1 - P(X < 6)= 1 - P(X <=5) = 1 - .6230 = .3770

So the p-value is .754. In other words, this outcome is quite consistent with the null hypothesis and we do not reject.

To get the same result from R:

```
> sum(dbinom(c(0:4, 6:10),10,.5))
[1] 0.7539062
```

Note: you can also use X/n as your test statistic and get the same results.

3b) Binomial for large n
If n > 25, say, then the table will do you no good. You can still use the computer, but often this approach works just as well:

Use the normal approximation if:

a) the first two conditions of 3a hold
b) $np_0 >= 10$ AND $n(1-p_0) >= 10$

Test Stat: $Z = ((X/n) - p_0) / \text{sqrt}(p_0 (1 - p_0) /n)$

Here's our reasoning:
If H0 is true:
E(X/n) = p0
SD(X/n) = sqrt(p0 (1 - p0)/ n)

And since X is a sum of bernoulli random variables, it is therefore approximately (not exactly) normal. If you want to be exact, then X is binomial. But if you can live with an approximation, then "pretend" that its normal. Therefore X/n is normal, with mean and SD as given above. And therefore

Sampling Distribution of Z is N(0,1).

The rest follows just as it did with the z-test (case (1)).

III. Comparing Two Groups

Again, you need to determine if you're comparing the means of two groups (the mean blood pressure of men compared with the mean blood pressure of women) or proportions (the percent of men who vote Democrat vs. the percent of women who vote Democrat.)

1. Comparing Means
a) Are the data "paired"? That is, one individual provides two observations? In that case, compute D = X - Y and treat D like a one-group hypothesis test. (Note: if the two groups have different sample sizes, then it's NOT a paired data set.)

b) If data are not paired, then you are comparing two independent groups. The primary test here is the "unpooled" t-test:
H0: mean1 - mean2 = 0.

The formulas here are too complicated to type. But the unpooled test is preferred because it does not assume that the SDs of both populations are the same.

The test statistic (look it up in the book) follows a T distribution, and the degrees of freedom for this distribution are calculated using a rather awkward formula.

If you KNOW that the SDs of the two populations are the same, then you can get away with the "pooled" t-test. The test statistic is slightly different, but the sampling distribution is again the T distribution, but now with (n + m - 2) degrees of freedom. (n is the sample size of one group, m of the other.)

If the data are not normal, than the distribution of the test statistic (either pooled or unpooled) is still approximately the T distribution, but the accuracy of the approximation depends on the sample size.

2. Comparing two proportions
This is complicated, and the best way is to use the normal approximation. So let X be the number of successes in group 1, Y the number of successes in group 2. Let p1 be the proportion of successes in Population 1, p2 in Population 2. We estimate these with p1-hat = X/n and p2-hat = Y/m.

H0: p1 - p2 = 0

To do this approximation, we need np1 >= 10 and n(1-p1) >= 10 and mp2 >=10 and m(1 - p2) >= 10.

The formula is again complicated, so look it up in the book. But the bottom line is that the test statistic is approximately N(0,1).