Lecture 4
Hypothesis Tests

We focus on tests comparing two groups because, believe it or not, this can be generalized as a regression. And so it makes a nice segue into the main topic of the course.

Problem: We have two distinct populations, and we are interested in knowing whether a parameter (or two) differs between the two populations. The reason this isn't easy to tell is that we have to make our decision based on a (relatively) small sample of the population, and hence don't have perfect knowledge. Worse, our sample might be "clouded" by measurement error, sampling variability, or any of a number of factors which, combined with the natural variability of the populations, makes it hard to tell for certain whether there is a difference in parameters.

We'll focus on comparing two means. And at the risk of getting boring, we'll focus on the playlist data that is also on your homework.

The two populations consist of all Rock songs on my hard drive and all Classical tracks on my hard drive. The question is whether the mean lengths are the same or different.

Our data consist of random samples drawn from the two populations.

Our first goal is simply to describe the difference. As a first step this is best done graphically, and a common device is the boxplot. But a histogram of all of the data can be itneresting, to see if there is bimodality and to see if the distribution looks normal.

We also need to describe it numerically. Since Xbar and Ybar are estiamtes of the respective means of the populations (let's agree that Xbar is the average of the times in the Rock songs and Ybar is the average times of the Classical) , it makes sense that we can estimate the difference between the means with T1 = Xbar - Ybar.

For this to be a "good" estimate, we need to make some assumptions about how we achieved these values. If each was achieved via a random sample, then T1 is unbiased. This means, on average, this approach will produce an estimate of the difference in means that is "on target". More precisely, E(T1) = mean(rock) - mean(classical) where "mean(rock)" is the population mean.

We can compute the standard error (the standard deviation) of this estimator too. But this is slightly problematic. Recall that Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2ab*cov(X,Y)

So Var(T1) = Var(Xbar) + Var(Ybar) - 2Cov(Xbar, Ybar)

Now presumably we either know the variance of Xbar and of Ybar or we can estiamte from the data. But the covariance term can be tricky to compute in general. So we have

to assume that our data was collected so that (a) X and Ys are independent or (b) so that we can estiamte the covariance.

A standard assumption is to assume that the sampel of X's is independent of the sample of Y's. Then Cov() will =0.

This seems like a good assumption for this problem, since we took a sample of Rock songs and a sample of classical, and there's no way that knowledge of one can give us information about the other.

So we have two assumptions on the data collection process: observations are random, observations are independent.

So now we can say that the mean difference in our sample is 33.285 seconds. But our question is not about the sample it is about the populations.

The question is, then, is this difference this large because of the differences between the populations, or because of chance error in our sampling?

What we want to know to answer this is what's the probability of getting a difference as large or larger than this just by chance?

We want P(Xbar - Ybar > = 33.285). But to answer this, we need to know the probability distribution of T1.

Well, the probability distributioni of T1 depends on the distributions of the times of the two populations.

If both are normally distributed, then Xbar is N(mu_x, sigma_x) and Ybar is N(mu_y, sigma_y). If not, then these are approximately normally distributed, as long as the sample size is "large enough". (How large is enough is a matter of debate. But the procedure we're developing is fairly robust if the distributions are not normal.) In that case, T1 will be normal or approximately normal. And E(T1) = mu_x - mu_y
Var(T1) = sigma_x^2/n + sigma_y^2/m (no covariance, because we're assuming these are independent).

This means a natural statistic to consider is
T2 = (Xbar - Ybar)/sqrt(sigma_x^2/n + sigma_y^2/m)

And if we know sigma_x and sigma_y then this is a good statistic and T2 is N(0, sqrt(sigma_x^2/n + sigma_y^2/m) distributed. And so it's easy to find probabities.

But in this case -- and indeed in most cases, we don't know the variances. Hence they must be estimated. Here is where a particular dilemma occurs. There are two ways of estimated the standard deviation of T1. The first is to assume that both populations have

the same standard deviations. If so, then we can "pool" our data together and combine forces to estimate the sd. The pooled estimate is

s^2 _pooled = sum.....(too complicated to type here).

Now we replace this estimate into T2, and the new statistic, T_pooled follows a t-distribution with n+m-2 degrees of freedom.

As we'll later see, this is how a linear model would handle it. But this turns out to be not good for two reasons:

a) for many interesting problems, the SDs are NOT equal in the two populations

b) the statistical tests for checkign whether they are equal or not are not terribly reliable, which means this is in effect an untestable assumptions

But fortunately, there is another method that works just as well, regardless of whether the SDs are equal or not.

This method is to estimate the SDs separately for X and Y and substitute them into the formula for T2. This method is not *quite* the same distribution as T_pooled. We'll call this statistic T. It has a t distribution with degrees of freedom given by something called the Satherwaite approximation,. A "conservative" approach is to say the df = min(n-1, m-1). But the computer will easily and painlessly do the more precise approximation.

So those are the technical details. The important thing to remember are the assumptions:
Assumptions about the populations: each population is "nearly" normal, and if not, sample sizes are large.
Assumptions about the sampling: observations are random and independent.

If these conditions are met, we're in a position to do a two-sample (non-pooled) t-test. If we add another assumption about the population, that the SDs in each population are equal, then we can do a pooled t-test. But I don't recommend it.

What is the t-test? there's a formality that goes along with it. First we state two competing hypotheses. The null hypothesis is the hypothesis that the means are equal -- the skeptical position.

The alternative is that they are not.

We then compute P(T > | observed value of T|) and this is our p-value. If the p-value is small, this means that it is unlikely to get a value of T so large and we reject the null.

"small" is defined by comparing with the "significance level", alpha. Alpha needs to be established before we begin, so lets set alpha = .05.

Alpha represents the probability of a Type 1 error: the probability we reject the null hypothesis when, in fact, it is true. This means we declare that the mean times are really different, when in the population the truth is that they are not.

If we follow this rule: "reject the null hypothesis whenever the p-value is less than alpha", then we will make this mistake only 100*alpha % of the time. If 5% is too often for you, then choose a smaller alpha.

In R, we do many steps at once:

```
> t.test(classical,rock)

        Welch Two Sample t-test

data:  classical and rock
t = -0.8694, df = 25.359, p-value = 0.3928
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -112.07622   45.50633
sample estimates:
mean of x mean of y
 208.1667  241.4516
```

This is actually an "object" as you can see if you type

```
outcome <- t.test(classical, rock)
names(outcome)
```

You see our conclusion i sthat the p-value is fairly large. We do NOT reject the null and conclude that there's no evidence of a difference in the means.

If we had decided to assume the SDs were equal

```
t.test(classica., rock, var.equal=T)
```

```
> t.test(classical, rock, var.equal=T)

        Two Sample t-test

data:  classical and rock
t = -0.9271, df = 78, p-value = 0.3567
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -104.75716   38.18727
sample estimates:
mean of x mean of y
```

208.1667 241.4516

It hardly matters in this case. A slightly larger t-statistic.

What do we do if the assumptions we made bother us? We have two choices: bootstrapping, randomization tests. We won't do the details, here.