C2 - Week 1

v1

\*The role of probability in inference

You've identified a dataset, and then used exploratory data analysis to organize and summarize the raw data in a meaningful and informative way.

The tools of exploratory data analysis, including:

1. Examination of frequency distribution,

2. Graphical representations of your variables of interest,

3. Calculations of center and spread,

help us to discover important features and patterns in the data and any striking deviations from those patterns.

This all falls under the rubric of Descriptive Statistics. Put simply Descriptive Statistics aims to quantitatively describe or summarize a sample of data.

\*Inferential Statistics

Inferential Statistics allow us to Directly Test Our Hypothesis by evaluating, based on a sample, a research question with the goal of generalizing the results to the larger population from which the sample was drawn.

Hypothesis Testing is one of the most important inferential tools in the application of statistics to real life problems. It's used when we need to make decisions concerning populations, on the basis of only a sample.

Statistical Hypothesis Testing is defined as assessing evidence provided by the data in favor of or against each hypothesis about the population.

We use probability to quantify how much we expect random samples to vary. This gives us a way to draw conclusions about the population in the face of the uncertainty that is generated by the use of a random sample.

Infere - in another words is to derive conclusion.

we can use probability to describe the likelihood that our sample is within a desired level of accuracy.

Probability is defined as the likelihood of something occurring; of an event occurring; the chance of something happening.

Probability is a mathematical description of randomness and uncertainty. It's a way to measure or quantify uncertainty. Another way to think about probability is that it's the official name for chance.

v2

\*From sample to population

The idea that sample results change from sample to sample, is called sampling variability.

Statistics are computed from samples, and each sample of a population is going to have different statistics. The statistics of different samples of a population vary. This is due to sampling variability.

As we draw more and more samples, the distribution of the sample statistics will become more and more normally distributed.

Central Limit Theorem states that as long as adequately large samples and an adequately large number of samples are drawn from a population, the distribution of the statistics of the samples, whether of mean, proportion, standard deviation, or any other statistic, will be normally distributed.

v3

\*Steps in hypothesis testing

Hypothesis testing is one of the most important inferential tools when it comes to the application of statistics to real life problems. Hypothesis testing is used when we need to make decisions concerning populations on the basis of only sample information.

A variety of statistical tests are used to help arrive at these decisions. For example, the analysis of variance test, ANOVA. And the Chi- Square Test of Independence, to name a couple. But they all include the same basic steps.

Steps involved in hypothesis testing, include:

1. specifying the null hypothesis H subscript 0, and the alternate hypothesis, H subscript a.

2. Choosing a sample

3. assessing the evidence

4. drawing conclusions

Statistical hypothesis testing is defined as assessing evidence provided by the data in favor of or against each hypothesis about the population.

Null Hypothesis - There is no relationship

Alternate Hypothesis - There is relationship

v4

\*What is P value?

The reason for using an Inferential Test is to get a probability value, commonly called p-value.

The p-value provides an estimate of how often we would get the obtained result by chance if in fact, the null hypothesis is true.

In statistics a result is called statistically significant if it's unlikely to have occurred by chance alone. The most commonly used standard or cutoff is 0.05 or 5%. Because this standard, or cutoff is so important it has a special name. It's called the significance level of a test, and is usually denoted by the Greek letter alpha, so alpha equals 0.05. If the p-value is small, less than 0.05, this suggests that it is more than 95% likely that the association of interest would be present following repeated samples drawn from the population. AKA, a sampling distribution.

Rejection of NULL Hypothesis:

If the p-value is less than alpha, which is usually 0.05, then the data we got is considered to be rare or surprising enough when the null hypothesis, H subscript 0 is true. And we say, that the data provides significant evidence against the null hypothesis. So, we reject the null hypothesis and accept the alternate hypothesis, H subscript a.

Accepting of Null Hypothesis:

If the p-value is greater than alpha, then the data is not considered to be surprising enough when the null hypothesis is true. And we say, that our data do not provide enough evidence to reject the null hypothesis. Or equivalently, that the data do not provide enough evidence to accept the alternate hypothesis.

This p-value is also known as the Type One Error Rate, since it denotes the number of times we would be wrong to reject the null hypothesis when it was true.

v5

\*How to choose a statistical test

Bivariate Statistical Tools:

1. ANOVA - Analysis of Variance

2. X2 - Chi-Square Test of independence

3. r - Correlation Coefficient

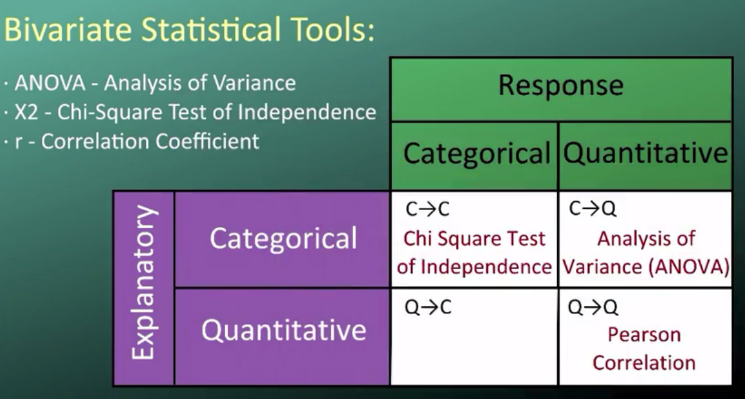
Inferential test types:

C(e)---> C(r) - Chi-Square Test of independence

C(e)---> Q(r) - ANOVA, analysis of Variance

Q(e)---> Q(r) - r, Pearson correlation

Q(e)---> C(r) - categorize your explanatory variable with only two levels and then use the Chi-Square Test of Independence



v6

\*Ideas behind ANOVA

**ANOVA F Test**

The question we need to answer with the ANOVA F Test is, are the differences among the sample means due to true differences among the population means, or merely due to sampling variability?

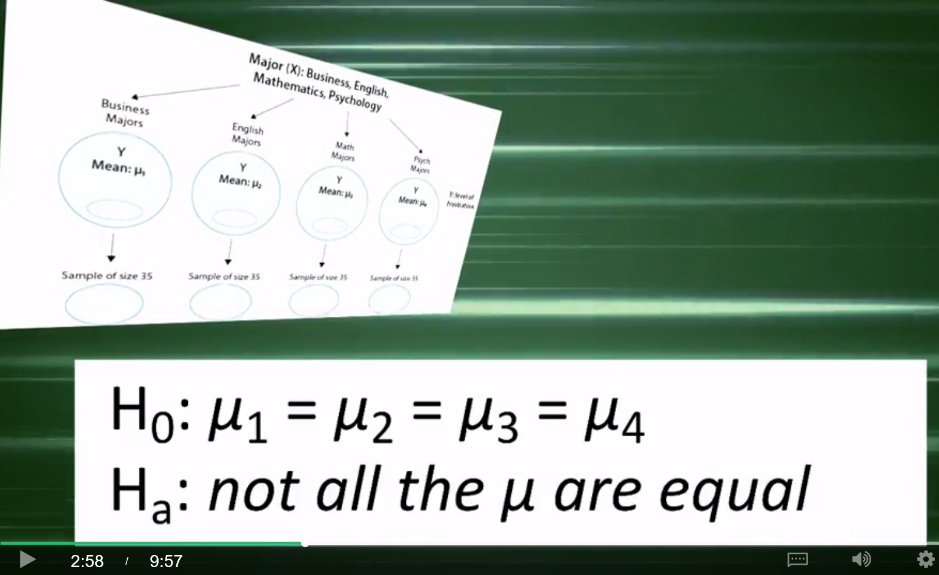
Comparison of means of population.

The Null Hypothesis (H0):

The null hypothesis claims that there's no relationship between the explanatory and response variables, x and y. Since the relationship is examined by comparing the means of y in the populations, defined by the values of x, no relationship would signify that all the means are equal. Therefore, the null hypothesis of the f test is population mean 1 equals population mean 2 equals population mean 3 equals population mean 4.

The Alternate Hypothesis (Ha):

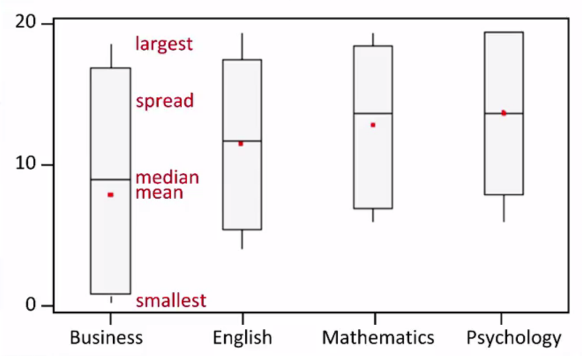
There is a relationship between x (Explanatory variable, C) and y (Response variable, Q). In terms of the means, it simply says the opposite, that not all of the means are equal and we simply write h subscript a, not all of the population means are equal.

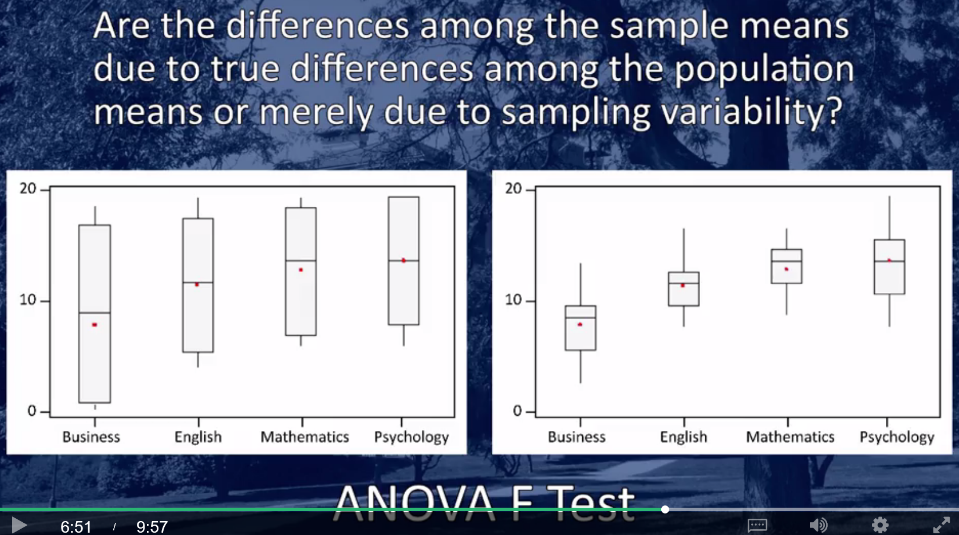


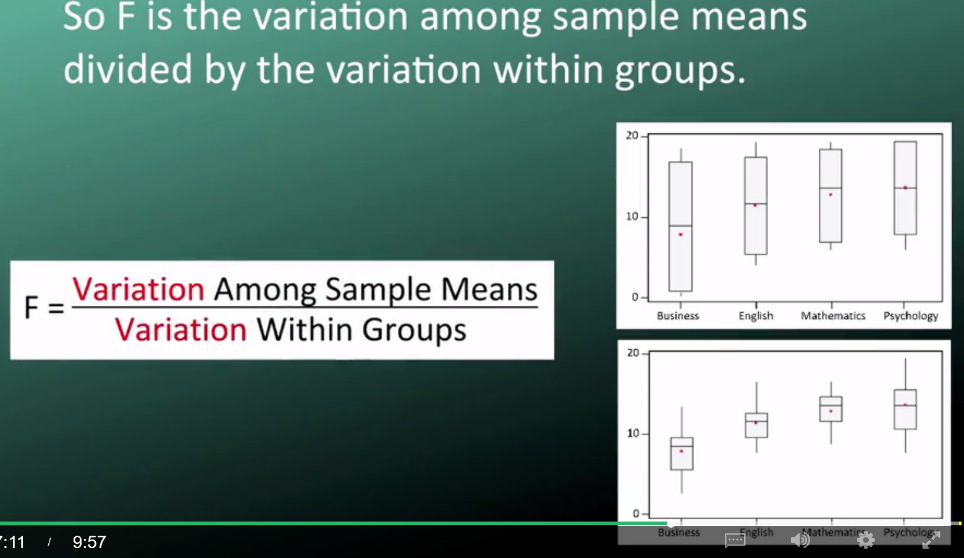
There are many ways for the population means not to be equal. Let's think about how we would go about testing whether the population means are equal. We could calculate the mean frustration level for each major and see how far apart those sample means are. Or, in other words, **measure the variation between the sample means**.

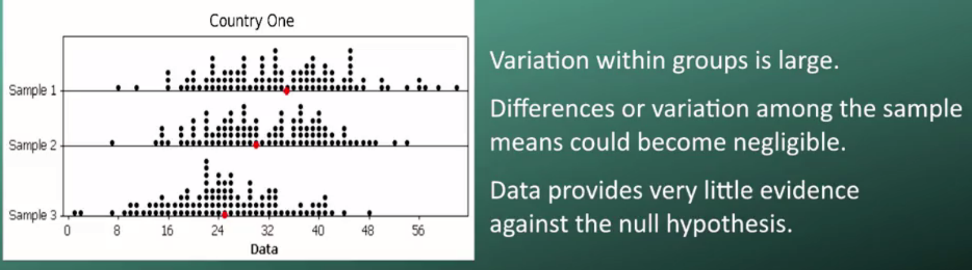
If we find that the four sample means are not all close together, we'll say that we have evidence against the null hypothesis. And otherwise, if they are close together, we'll say that we do not have evidence against the null hypothesis.

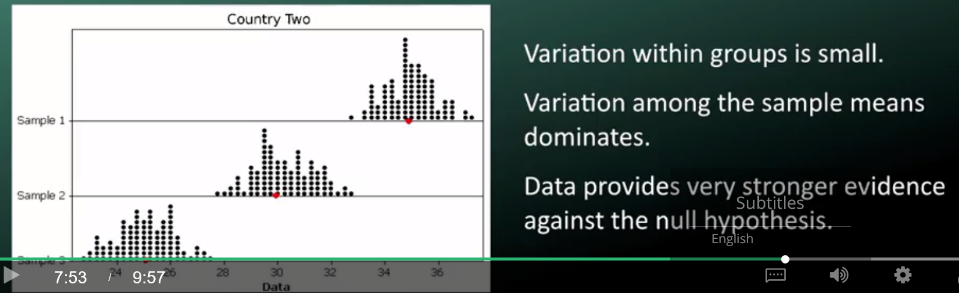
A boxplot is a convenient way of graphically depicting groups of numerical data including such descriptive information as the smallest observation of the group, the mean and median, the largest observation, and the spread or variability of the values.

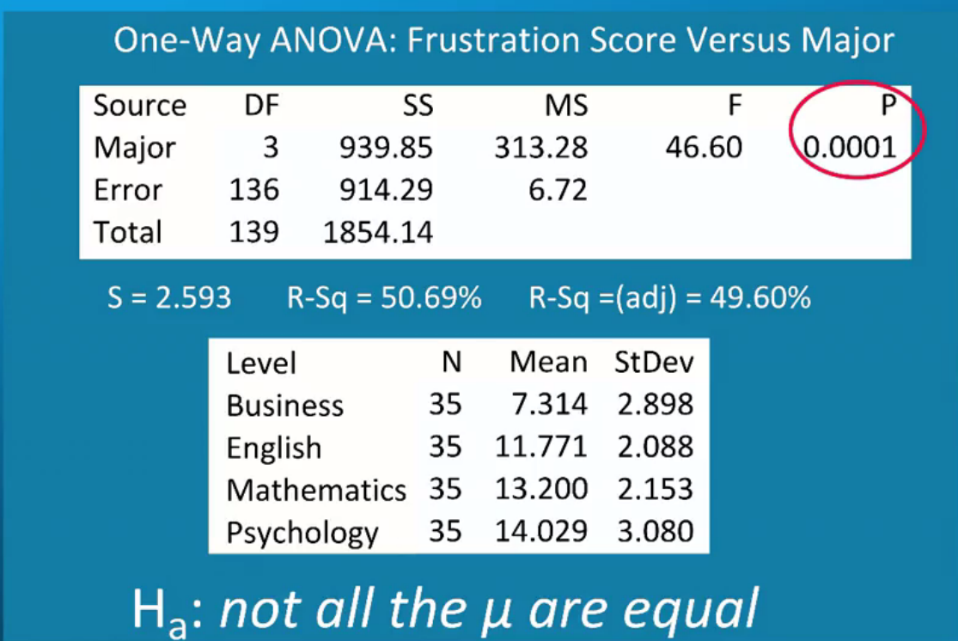












Here are the results of the analysis of variance for Country Two. Testing the relationship between major and frustration score. **The F statistic circled in red is 46.60. Since we know this is the variability among sample means divided by the variability within groups, this large number suggests that the variability among sample means is much greater than that within sample groups.**

The P value of the ANOVA F Test is the probability of getting an F statistic as largest we got or even larger had the null hypothesis been true. That is, had the population means been equal. In other words, it tells us how surprising it is to find data like those observed, assuming that there is no difference among the population means.

This P value is practically 0, telling us that it would be next to impossible to get data like those observed had the mean frustration level of the four majors been the same as the null hypothesis claims.

**The P value 0.0001 suggests that we will incorrectly reject the null hypothesis one in ten thousand times. And we will be correct in accepting the alternate hypothesis 9999 times out of 10,000 times.**

So, we can confidently conclude that the frustration level means of the four majors are not all the same. Or in other words, there's a significant association between frustration level and major. So, we accept the alternate hypothesis and reject the null hypothesis.

V7

# \*ANOVA: Explanatory variable with 2 levels

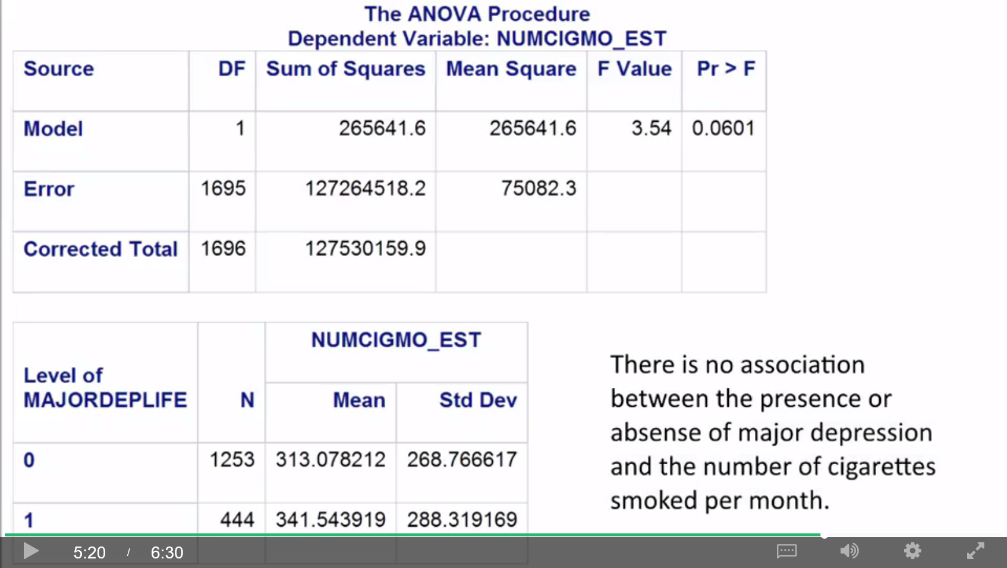
# Sample ANOVA procedure code with two levels explanatory variable:

# C:\Users\martc\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Screenshot-2017-10-3 SAS Lesson 7 - ANOVA Explanatory variable with 2 levels - Wesleyan University Coursera.png

# CLASS statement identifies the categorical explanatory variable.

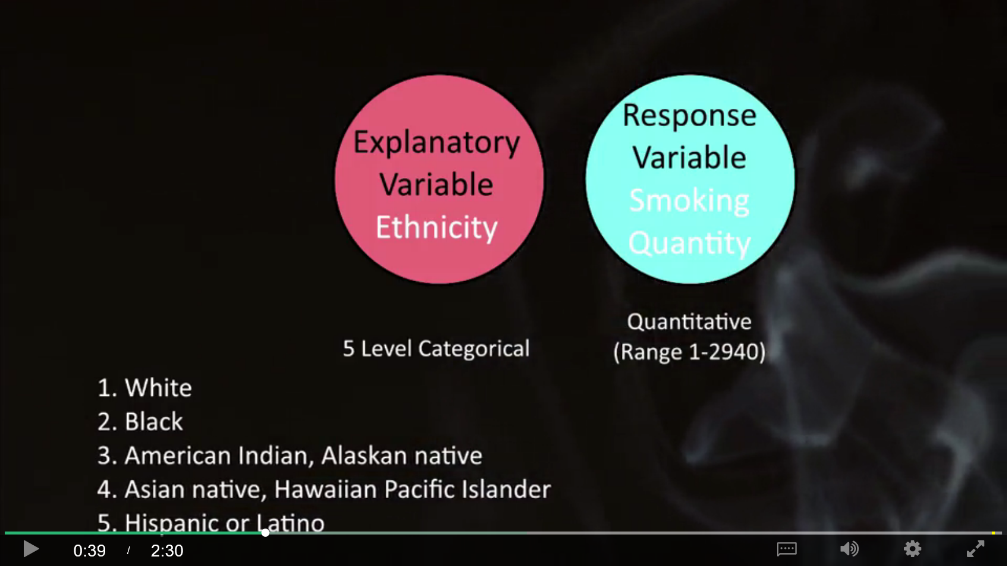
# MODEL statement naming the quantitative response variable.

MEANS statement tells us which groups you would like to compare mean number of cigarettes smoked per month.



In this case, our response or dependent variable was NUMCIGMO\_EST. Our calculated F statistic, called the F Value in this output, is 3.54. The significance, probability, or P value associated with this F statistic, is labeled Pr > F. And as you can see, the P value is .0601, just over our P value .05 cut point. If we look at the means table, we see that young adult smokers without major depression, as indicated by a value of 0, smoke an average of 312 cigarettes per month. And that those with major depression, indicated by a value of one, smoke on average 341.5 cigarettes per month. **Because the P value is greater than 0.05, actually 0.06, we must accept the null hypothesis and say that these means are statistically equal.** And that there's no association between the presence or absence of major depression in the number of cigarettes smoked per month among young adult smokers. >> If I chose to reject the null hypothesis, I would be wrong six out of 100 times. And again, by normal scientific standards, this is not adequate certainty to reject the null hypothesis and say that there is an association.

V8

\*ANOVA: Explanatory variable with more than two levels

Submitting ANOVA the way we submitted for 2 level explanatory variable, the F-test and the p-value do not provide insight into why the null hypothesis can be rejected, because there are multiple levels to categorical explanatory variable. They do not tell us in what way the population means are not statistically equal.

V9

\*Post-Hoc test for ANOVA

In the case where the explanatory variable represents more than two groups, a significant ANOVA does not tell us which groups are different from the others. **To determine which groups are different from the others, we would need to perform a post hoc test.**

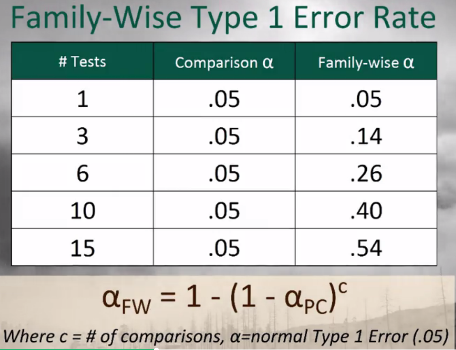
A post hoc test conducts post hoc paired comparisons. Post hoc means after the fact, and these post hoc paired comparisons must be conducted in a particular way in order to prevent excessive Type 1 error.

**Type 1 error**, as you'll recall, **occurs when you make an incorrect decision about the null hypothesis, that is you reject the null hypothesis when the null hypothesis is true.**

Why we cannot perform pair-wise ANOVA comparing each group to the other?

There's actually a 5% chance of making a Type 1 error for each analysis of variance that we conduct on this question. Therefore, performing multiple tests means that our overall chance of committing Type 1 error could be far greater than 5%. Here's how it works out.

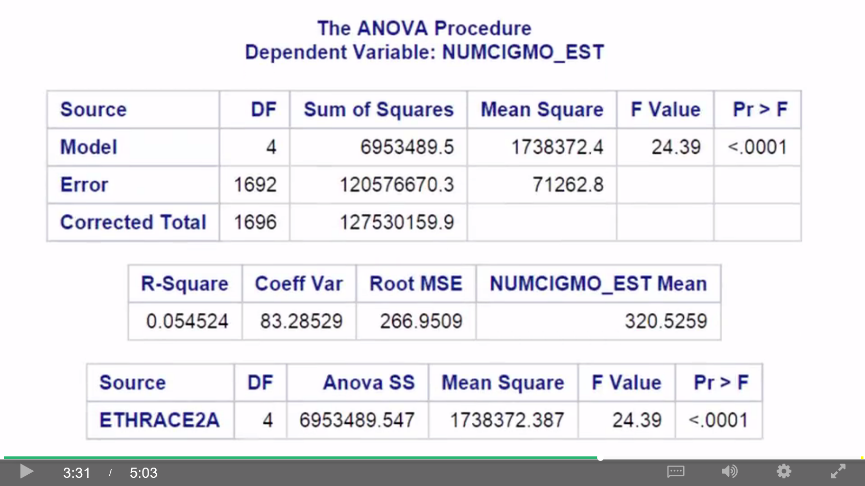
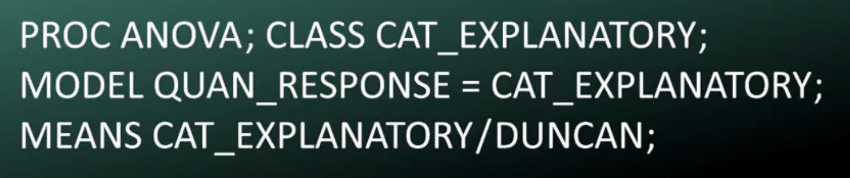
Using the formula displayed under this table, you can see that while one test has a Type 1 Error Rate of 0.05, by the time we've conducted 10 tests on this question, our chance of rejecting the null hypothesis, when the null hypothesis is true, is up to 40%.



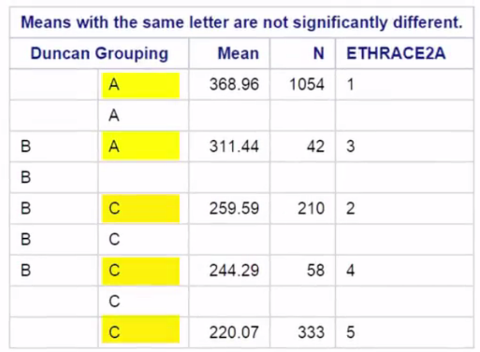
This increase in the Type 1 error rate is called the **family-wise error rate and is the error rate for pair comparison.**

Post hoc tests are designed to evaluate the difference between pairs of means while protecting against inflation of Type 1 errors.

There's the **Sidak**, and the **Holm T test**, and **Fisher's Least Significant Difference test**, **Tukey's Honestly Significant Difference test**, the **Scheffee test**, the **Newman-Keuls** test, **Dunnett's Multiple Comparison test**, the **Duncan Multiple Range test**, and the **Bonferroni Procedure**



Here are the Proc ANOVA results with Duncan Post hoc Test. The top of our results looks the same as in our original test. The F value or F statistic is 24.4 and it's significant at the P < 0.0001 level. However, if we scroll down, we see a new table displaying the results of the paired comparisons conducted by the Duncan Multiple Range Test.



Basically, means with the same capitol letter next to them are not significantly different. So, you can see that ethnic group 1 and 3 are not significantly different because they both have A's. Groups 2, 4, and 5 are not significantly different because each has a C. Groups 3, 2, and 4, again are not significantly different from one another, each with a B next to the group number. So where are the significant differences? Group 1, which indicates White ethnicity, smoked significantly more cigarettes per month than ethnic groups 2, 4, and 5. That is Black, Asian, and Hispanic or Latino. Group 3, American Indian and Alaskan Native, smoked significantly more per month than group 5, which is Hispanic or Latino. So you need to be careful, and follow the rule that means that share even one letter in common are not significantly different from one another.