**Inference for Numerical Variables**

YOUNGJIN LEE

**T - distribution**

**Review: What purpose does a large sample serve?**

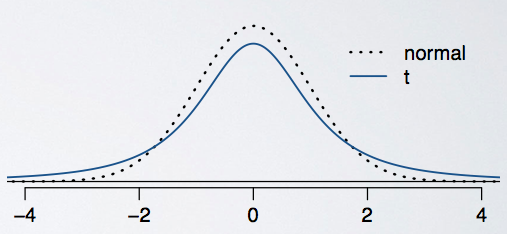
With independent observations and not extremely skewed population distribution, a large sample would ensure…

* A nearly normal sampling distribution of the mean
* A better s estimate for the population SD, sigma
* A reliable estimate of the Standard Error (s / route(n))
* Standard Error is the standard deviation of the sampling distribution of a statistic.
* The uncertainty of Standard Error estimate is addressed by using the T-distribution

**T - distribution (1)**

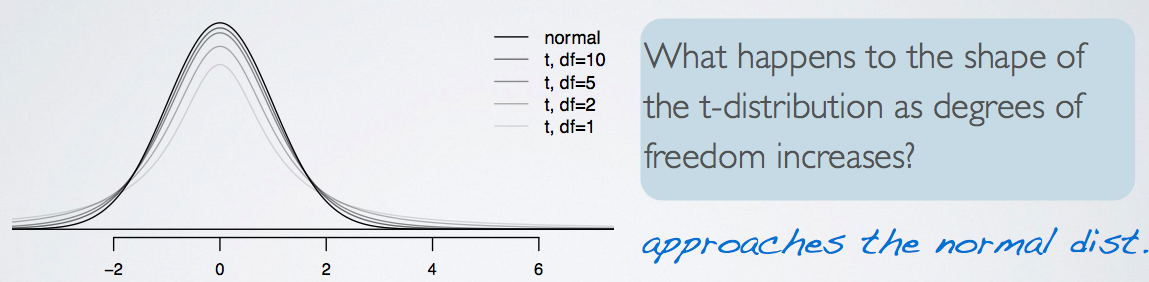
* Used for inference on a mean where sigma is unknown. (almost always)
* To address the uncertainty of the Standard Error estimate
* Unimodal, symmetric
* Bell shaped but thicker tails than the normal distribution
* As s (sample standard deviation) is used instead of sigma (population standard deviation)

**Normal distribution vs T – distribution**



* Peek of the t-distribution is lower than that of the normal distribution, somewhat squashed in the middle and the additional area is added to the tails.
* Observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.
* Confidence intervals constructed using the t distribution will be wider, more conservative than those constructed with the normal distribution.
* Thicker tails are helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution

**T - distribution (2)**

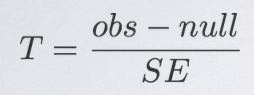
****

* Always centered at 0 (like the standard normal)
* Has one parameter: degrees of freedom(df) – determines the thickness of tails

(the normal distribution has two parameters: mean and SD)

* degrees of freedom is tied to sample size

**T Score for a Hypothesis Test**

****

**Example) Effect of degrees of freedom in t-distribution, significance level = 0.05**

P(|Z| > 2) = 0.0455 - reject

P(|tdf = 50| > 2) = 0.0509 – fail to reject ?

P(|tdf = 10| > 2) = 0.0734 – fail to reject

* More conservative with a t distribution with lower degrees of freedom, less likely to be able to reject the null hypothesis.
* Not as easy to reject the null hypothesis with a small sample size and stronger evidence is needed in order to be able to do so.

**Inference of a Mean**

2 Steps: Construct a confidence Interval, Conduct a hypothesis test

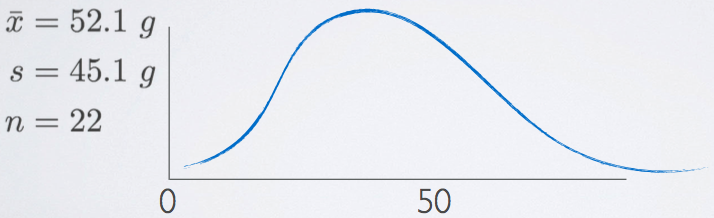
**Conditions**

1. Check conditions
2. Independence

* Sampled observations must be independent
  + Random sample / assignment
  + If sample without replacement, n < 10% of population

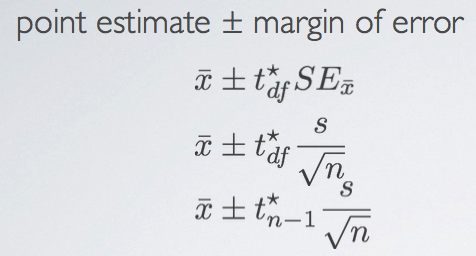
1. Sample size / skew

* The more skew in the population distributions, the higher the sample size is needed

****

* Given the sample statistics, we can kind of sketch it out. The sample mean is 52 and there's a natural boundary at 0 since one cannot eat less than 0 grams of biscuit. So the 68, 95, 99.7 rule is just not going to apply here. Because if we go more than one standard deviation below the mean, we're going to hit that natural boundary of 0 grams. Therefore, that data are likely right skewed. The t distribution is pretty robustious units but ideally, we would like to see a visualization of this distribution and the size this sample size queue distribution accordingly, especially given the low sample size.

**Estimate the mean**



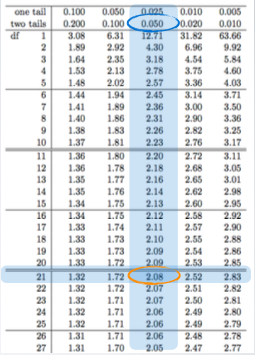
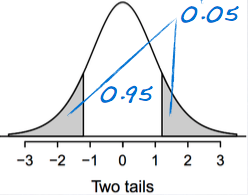
* T\*df : the critical t score to be found

**Degrees of freedom** for t statistic for inference on one sample mean



* We lose one degree of freedom because we're estimating the standard error of the sample mean using the sample standard deviation.

**Finding the Critical T Score (table)**

****

* Find corresponding tail area for desired confidence level

**INSTRUCTIONS**

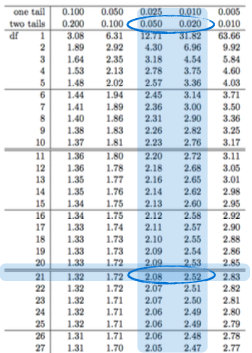
1. Check the conditions
2. Construct a confidence interval by estimating the mean
3. Make a decision, and interpret it in context of the research question:

* We are 95% confident that IT happens between the lower bound and the upper bound

**Calculating T Score**

t = (x̄ - μ) / SE

**Finding the P-value (table)**



* We have df and t value. P values are on the top.
* t values are absolute so the sign is ignored.

**INSTRUCTIONS**

1. Set the hypothesis

* H0: μ = suggested value
* H1: μ != suggested value

1. Check conditions
2. Calculate the point estimate x̄
3. Draw sampling distribution, calculate test statistic, shade p-value,
4. Make a decision, and interpret it in context of the research question:

*See ‘Inference\_of\_a\_mean.pdf’ for an example*

**Inference for Comparing Two Independent Means**

2 Steps: Construct a confidence Interval, Conduct a hypothesis test

**Conditions**

1. Independence

* Within groups: sampled observations must be independent
  + Random sample / assignment
  + If sample without replacement, n < 10% of population
* Between groups: the two groups must be independent of each other (non-paired)

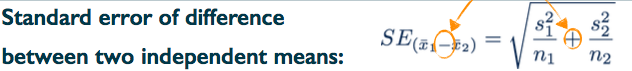
1. Sample size / skew

* The more skew in the population distributions, the higher the sample size is needed

**Point Estimate**



**Standard Error**



* This is new. Note that we add the two variances even though we're looking for the standard error of the difference of the two means. This is not simple to prove mathematically with only the tools we've learned in this course so far, but conceptually, you can think about it as bringing together two measures with an inherent variability around them. These are two sample means. When you bring two unknowns together, the result should always be more variable, regardless of whether you're adding them or subtracting them.

**INSTRUCTIONS**

1. Check the conditions
2. Construct a confidence interval by estimating the difference between means
3. Make a decision, and interpret it in context of the research question:

* We are 95% confident that IT happens between the lower bound and the upper bound

**Degrees of Freedom**



* You might come across a few other estimates for the degrees of freedom for comparing two means, perhaps in another textbook or so, but what we noted here is actually not the exact degrees of freedom, which is quite tedious to compute by hand. The estimate of degrees of freedom we noted here, is a conservative estimate, since it relies on the lower of the two sample sizes.

**INSTRUCTIONS**

1. Set the hypothesis

* H0: μx – μy = 0
* H1: μx – μy != 0

1. Check conditions
2. Calculate the point estimate x̄1 - x̄2
3. Draw sampling distribution, calculate test statistic, shade p-value,
4. Make a decision, and interpret it in context of the research question:

*See ‘Inference\_for\_comparing\_two\_independent\_means.pdf’ for an example*

**Inference for Comparing Two Paired Means**

**(much like the ‘inference of a mean’ with a new column of difference)**

**Paired Data**

When two sets of observations have this special correspondence (not independent), they are said to be paired.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations

* The parameter of interest is the average difference between the two events of the population(μdiff)
* The point estimate would be the average difference between the two events of the sample(x̄diff)

**INSTRUCTIONS**

1. Create a new column named ‘difference’ by subtracting event A from B or vice versa.

Difference = event A – event B

1. Set the hypothesis

* H0 : μdiff  = 0 ; There is no difference between the average of the two events
* H1 : μdiff  != 0 ; There is difference between the average of the two events

1. Check conditions
2. Calculate the point estimate x̄, standard deviation for this new column. Then we have ; x̄diff, sdiff, ndiff.
3. Draw sampling distribution, shade p-value,
4. Make a decision, and interpret it in context of the research question:

*See ‘Inference\_for\_comparing\_two\_paired\_means.pdf’ for an example*

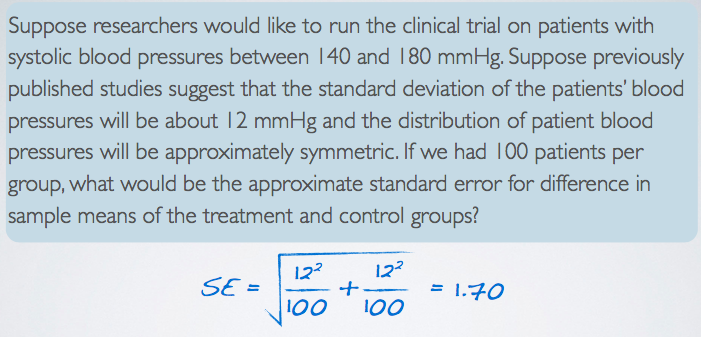
**Power of The Test**

****

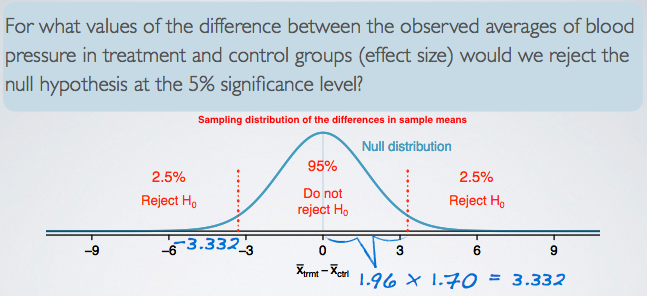
* Type 1 error is rejecting H0 when you shouldn’t have, and the probability of doing so is alpha (significance level)
* Type 2 error is failing to reject H0 when you should have, and the probability of doing so is Beta
* Power of a test is the probability of correctly rejecting H0, and the probability of doing so is 1 – Beta.
* In a hypothesis test, we obviously want to keep our error rates low, both alpha and beta. However, decreasing one increases the other, and one solution for this problem is getting a larger sample size. Hence, it's important to think about the sample size when designing an experiment. And making sure that resources are invested to recruit a sufficiently large number of subjects to obtain the desired power of the test.

**Example)**





Then, according to the central limit theorem, the distribution of the differences in sample means will be nearly normal, with mean 0, because remember, that's what our null value was. And the standard error is the standard error that we calculated, 1.70. Using this information, we can find out what values of the sample statistic we would need to reject the null hypothesis.



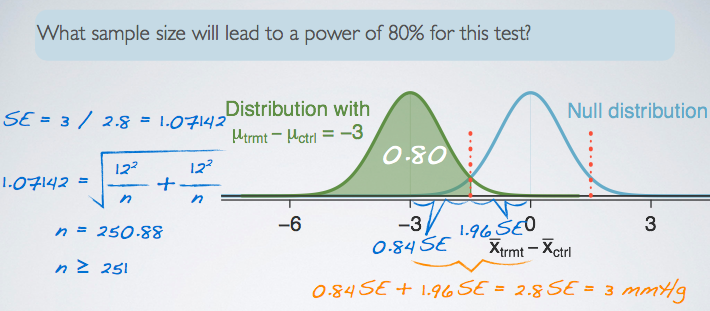
Under the normal model, 95% of the observations fall within 1.96 standard deviations of the mean.



3 millimeters of mercury is the minimum effect size of interest and we want to know how likely we are to detect this size of an effect in this study. If the treatment is indeed effective enough to result in a 3 millimeters of mercury drop in blood pressure on average, then it means the observed distribution of differences in average blood pressures between the two groups will be shifted from the null by 3 millimeters of mercury, as shown in this plot here. We also know that we can only reject the null hypothesis if the observed difference is less than negative 3.332 millimeters of mercury. Putting all of these together, the probability of being able to reject the null hypothesis if the true effect size is negative 3, is equal to the green shaded area under this curve.

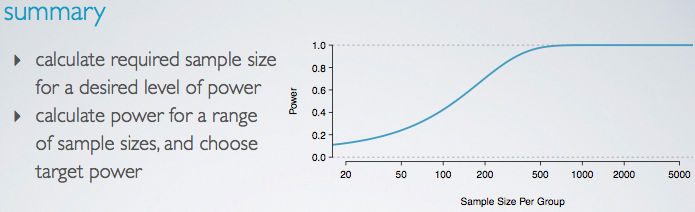
The power of the test is 0.4207 when the effect size is negative 3 and each group has a sample size of 100.

It highlights how important it is to not just arbitrarily select a sample size and risk being left with an under powered study. How can we fix things ?



The distance from 0(‘no difference’) to the value of interest(‘-3’) is equal to 3 and in terms of Standard Error it is 2.8SE. We can solve for SE with this equation and finally find out the value of ‘n’ in the SE formula.

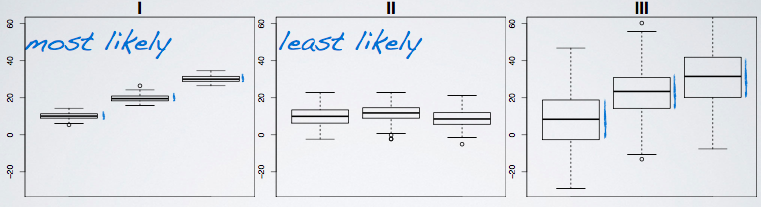
The required sample size for 80% of power of the test is at least 251.



When are these calculations actually used in practice? We can use them when designing a study to calculate a required sample size for a desired level of power. Or we can calculate the power for a range of sample sizes, and choose the target level of the power based on the resources available for collecting the required sample size.

We can see that as the sample size increases so does power but only up to a point, there seems to be no good reason to recruit more than 500 patients or so for each group since the power plateaus at that point. This is important to know when designing a study in order to avoid wasting resources on a sample size that is larger than needed for the maximum power desired.

**Comparing Many Means(ANOVA)**



* These plots show how much groups with means are likely to be significant from each other.
* There are more than two means to compare
* We use a new test called Analysis of Variance(ANOVA) and a new statistic called F for comparing means of 3+ groups

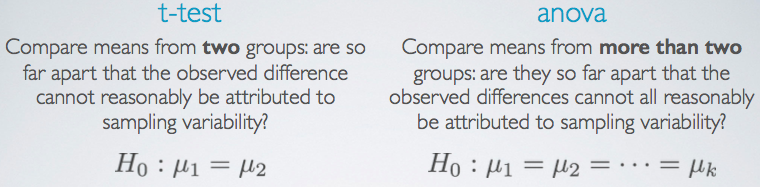
In ANOVA, defining the hypothesis is somewhat different from t-test

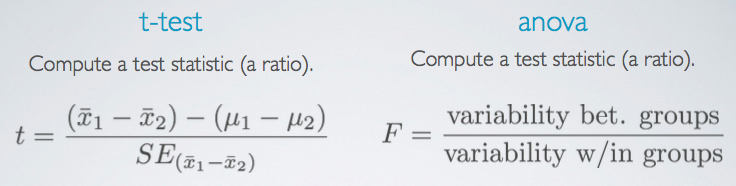
Where μi : mean of the outcome for observations in category i and k : number of groups,

H0 : The mean outcome is the same across all categories. μ1 = μ2 = … = μk

H1 : At least one pair of means are different from each other.

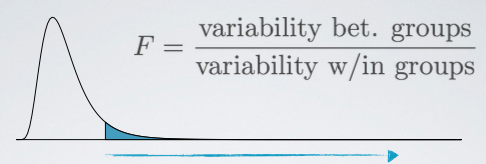
**T – test vs ANOVA**





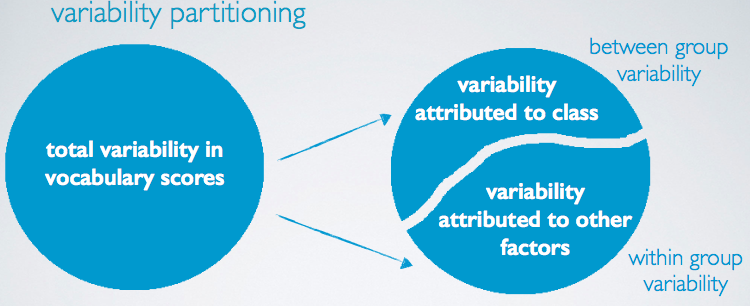
* Large test statistics lead to small p-values
* If the p-value is small enough H0 is rejected, and we conclude that the data provide evidence of a difference in the population means.

**F – distribution**

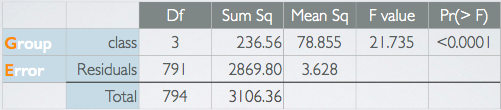


* Right skewed, always positive (since it's a ratio of two measures of variability, which can never be negative)
* In order to be able to reject H0, we need a small p-value, which requires a large F statistic
* Obtaining a large F statistic requires that the variability between sample means is greater than the variability within the samples.

**Variability Partitioning**



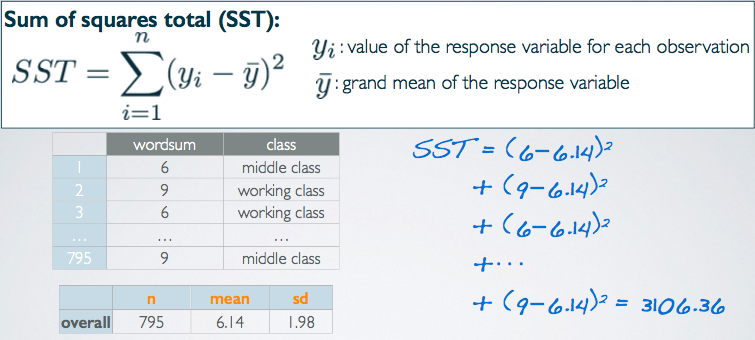
**Interpreting ANOVA Output**



The between-group variability is what we are interested in since it is the variability attributed to the class we are considering atm.

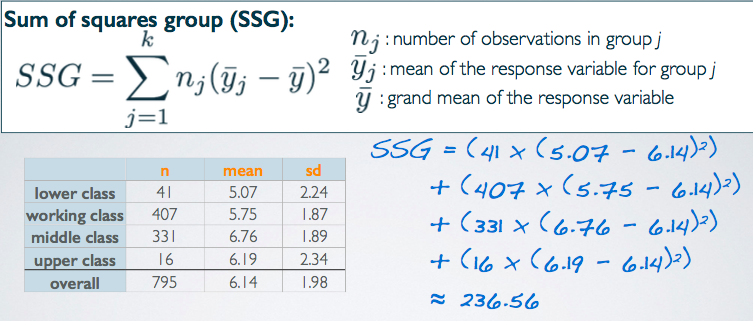
* ‘Group’ row: between-group variability
* ‘Error’ row: within-group variability

1. Sum of Squares Total(=SST), (Total, Sum Sq), 3106.36



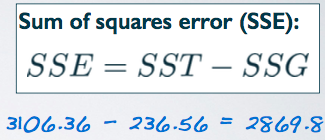
* measures the total variability in the response variable
* calculated very similarly to variance (except not scaled by the sample size)

1. Sum of Squares Groups(SSG), (Group, Sum Sq), 236.56



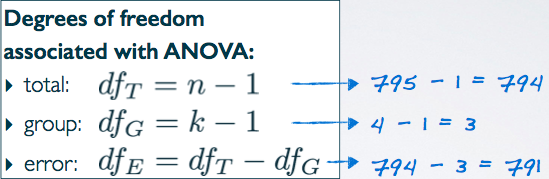
* Measures the variability between groups
* Explained variability: squared deviation of group means from overall mean, weighted by sample size
* Its own is not a meaningful number but it's interesting how it compares to the total sum of squares we calculated earlier. For example, this value is roughly 7.6% of SST. Meaning that 7.6% of the variability in vocabulary scores is explained by social class and the remainder is not explained by the explanatory variable we're considering in this analysis.

1. Sum of Squares Error(SSE), (Error, Sum Sq), 2869.8



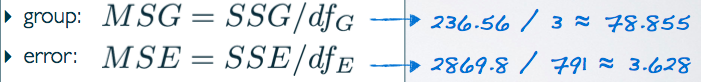
* Measures the variability within groups
* Unexplained variability: unexplained by the group variable, due to other reasons

1. Degrees of Freedom associated with ANOVA



* Used to calculate the average variability from the measures of total variability(Sum Sq) as a scaling measure that incorporates sample sizes and number of groups.

1. Mean Square Error(Mean Sq)



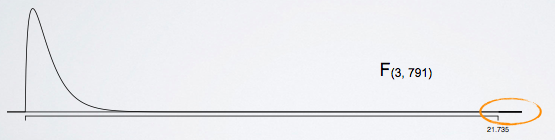
* Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom

1. F statistic



* Ratio of the average between group and within group variabilities

1. P-value



* The probability of at least as large a ratio between the “between” and “within” group variabilities if in fact the means of all groups are equal
* Area under the F curve, with degrees of freedom dfG and dfE, above the observed F statistic

**ANOVA Conclusion**

* If p-value is small (less than alpha), reject H0: The data provide convincing evidence that at least one pair of population means are different from each other (no tell which one though)
* If p-value is large, fail to reject H0: The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance)

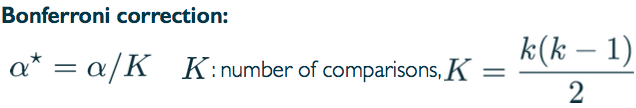
*See ‘conditions\_for\_anova.pdf’*

**Multiple Comparisons(사후검정)**

**Which Means Differ**

* Conduct two sample t tests for differences in each possible pair of groups
* And with each test that you do, you incur a probability of doing a Type I error.
* The probability of committing a Type I error is the significance level of the test, which is often set to 5%.
* So when you do multiple tests, you're going to be inflating your Type I error rate, which is an undesirable outcome.
* Thankfully, there is a simple solution.
* Use a modified significance level that is lower than the original significance level for these pairwise tests, so that the overall Type I error rate for the series of tests you have to do can still be held at the original low rate.

**Multiple Comparisons**

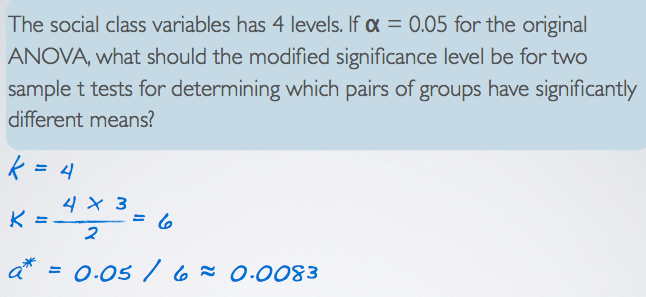


* Use Bonferroni correction for testing many pairs of groups, which suggests a more stringent significance level is more appropriate for these tests
* Adjust alpha by the number of comparisons being considered
* For example, if you have four groups in your ANOVA, and it does yield a significant result, then you need to compare group one to group two, two to three, three to four, so on and so forth. But counting these out is somewhat tedious and error prone, so we usually use a shortcut formula for determining this value and then use this value to adjust the significance level.

**2 Steps for Bonferroni correction**

1. Find the number of comparisons as k times k minus one divided by two,
2. Correct your original alpha by this level, as alpha divided by the number of comparisons.

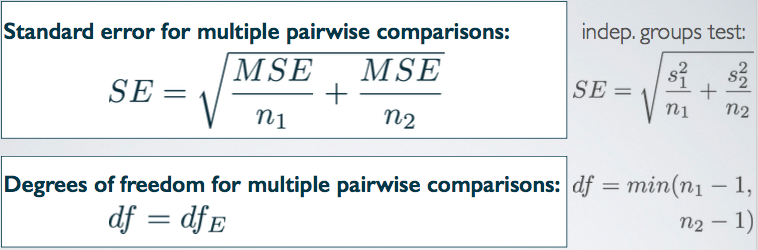
**Example)**



**Other considerations when doing multiple comparisons after ANOVA.**

1. Constant variance

* Need to re-think standard error and degrees of freedom for all tests

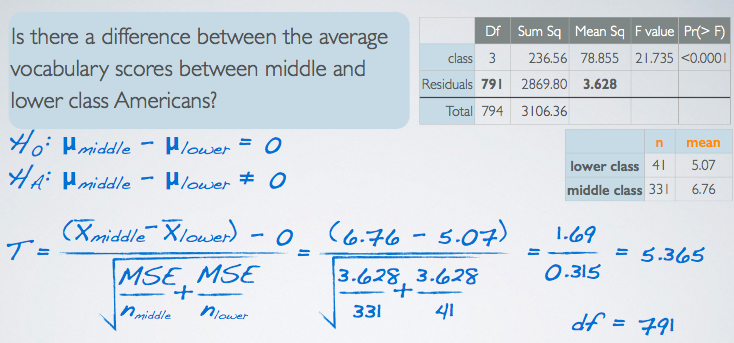


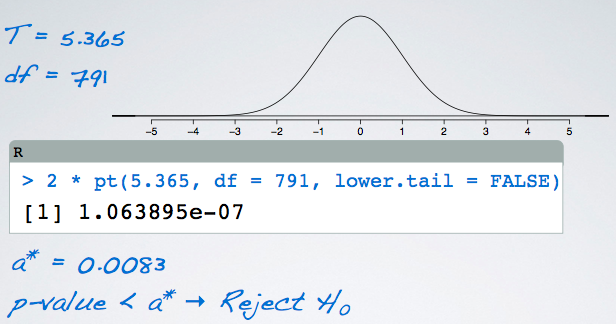
The mean's squared error is actually the average within group variance, so we're still getting at the same thing, the individual group variances, but now, we have a consistent measure that we can use for all of the tests. If, indeed, the constant variance condition is satisfied, this value should be very close to your group variances anyway.

The consistent degrees of freedom is going to be the DF error from the ANOVA output, as opposed to the minimum of the sample sizes minus one, from the two groups that we're comparing.

1. Compare the p-values from each test to the modified significance level

**Example)**



****

**Boot-Strapping**

*See BootStrapping.pdf*

*Test yourself:*

*1. What is the t⋆ for a 95% confidence interval for a mean, where the sample size is 13.*

*2. What is the p-value for a hypothesis test where the alternative hypothesis is two-sided, the sample size is 20, and the test statistic, T, is calculated to be 1.75?*

*3. 20 cardiac patients’ blood pressure is measured before taking a medication, and after. For a given patient, are the before and after blood pressure measurements dependent (paired) or independent?*

*4. A random sample of 100 students were obtained and then randomly assigned into two equal sized groups. One group went on a roller coaster while the other in a simulator at an amusement park. Afterwards their blood pressure measurements were taken. Are the measurements dependent (paired) or independent?*

*5. Describe how the two sample means test is different from the paired means test, both conceptually and in terms of the calculation of the standard error.*

*6. A 95% confidence interval for the difference between the number of calories consumed by mature and juvenile cats (μmat−μjuv) is (80 calories, 100 calories). Interpret this interval, and determine if it suggests a significant difference between the two means.*

*7. We would like to compare the average income of Americans who live in the Northeast, Midwest, South, and West. What are the appropriate hypotheses?*

*8. Suppose the sample in the question above has 1000 observations, what are the degrees of freedom associated with the F-statistic?*

*9. Suppose the null hypothesis is rejected. Describe how we would discover which regions’ averages are different from each other. Make sure to discuss how many pairwise comparisons we would need to make, and what the corrected significance level would be.*

*10. What visualizations are useful for checking each of the conditions required for performing ANOVA?*

*11. How is a bootstrap distribution different from a sampling distribution?*

*12. If a bootstrap distribution is constructed using 200 simulations, how would we find the 95% bootstrap confidence interval? Hint: Draw a sketch.*