ANOVA: comparing means and Research methods

Timestamper!
Projects
spss instructions

Week 4

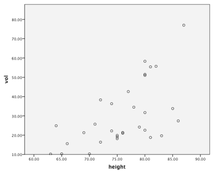
Projects

- Groups of 4-5 (March 11)
- Data from internet or observational study
- In-class project proposal presentations (March 25)
- Project proposal document (April 15)
  - 2 pages
- Final project document (May 6)
  - 10 pages

A switch: ANOVA

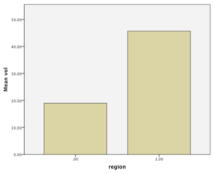
- Haven't we learned about this already?
  - Yes, as part of the 'puzzle' of regression.
  - Same idea here, partitioning variability
  - But now not just a table, but the main model that we will use
- Used to ask a different question than regression!!

Regression
Investigate a relationship



Quantitative predictors

ANOVA
Investigate difference between means



Categorical predictors

One-way ANOVA

- ANOVA is an extension of the two-sample ttest, and is used to compare the means of several populations.
  - Does the average time spent on studying differ for students that get an A, vs B, vs C.
  - Are people faster to respond when identifying the gender of family members, friends, or strangers
  - Does the average home price differ between Cambridge, Newton, and Arlington.
- Another way to say that is, are we better able to predict the response value if we know which group the observation came from?

The basics

- Data requirements:
  - 1 categorical predictor/explanatory variable with any number of ‘levels’ or groups
    - Predictor often called factor
    - Groups sometimes called levels or treatments
  - 1 response
    - Quantitative just like in regression
- Still assume the response variable is related to the explanatory variable through a model with random errors, and still looks at sum of square deviations.

Dataset

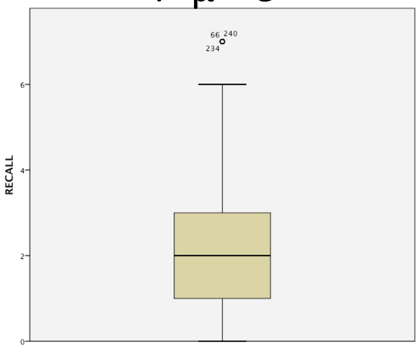
- Advertisers want to get the most for their money, so they wondered: Do TV shows with violence and sex impair memory for commercials? Iowa State professors conducted an experiment where 324 adults were randomly assigned to one of three viewer groups. One watched one high in violence, one high in sex content, and one a neutral program. Each program had several commercials embedded. Afterwards, they were scored on their recall of the brand names in the commercial messages.
- What is the explanatory variable?
  - Type of tv program
- How many levels does it have?
  - three
- What is the response variable?
  - Recall score

Model: does the mean differ by group?

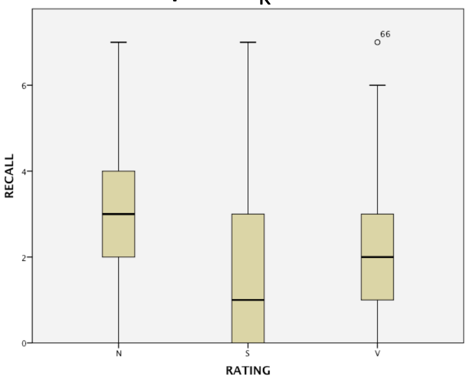
- Is the grand mean the best predictor? Or does knowing what group the value belongs to help?

ANOVA model tested

$Y=\mu + \varepsilon$



$Y=\mu + \alpha_k + \varepsilon$

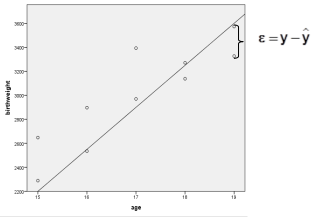


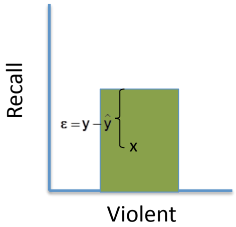
ANOVA assumptions

- All about the errors/residuals again
- Have a mean of zero
- Equal variance: have the same standard deviation for each group
- Errors follow a normal distribution
  - Not so bad to violate if no outliers
- Errors be independent: usually achieved by random assignment
- But what is a residual in ANOVA?

Residual

- It is still observed – predicted
- Here predicted is the mean of the group the observation is part of.
- So if participant 10 watched a violent tv show, it would be:
  - Their recall score – the average recall score of everyone in the violent tv show group
- Book uses a bunch of symbols, don't stress too much





Going to use the residuals to create SSE: just like in regression ANOVA table

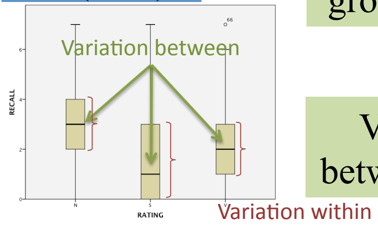
TOTAL variation in response, *Y* = Variation explained by MODEL + Unexplained variation in RESIDUALS

SS(total) = SS(Model) + SS(Error)

SS(total) = Variation from group to group + Variation from one individual to the next

Variability between groups

Variability within group



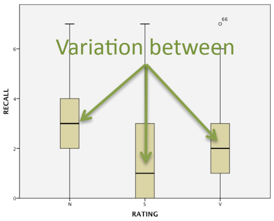
Test statistic: a ratio

$$F = \frac{\text{Variability between groups}}{\text{Variability within group}}$$

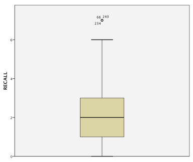
To test the hypothesis:

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

H_a: the means are not all equal



Grand mean



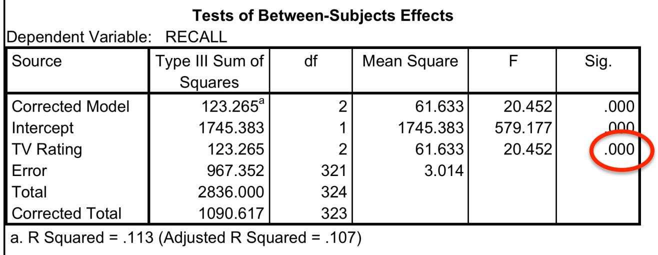
- When the numerator is large compared to the denominator, we will reject the null hypothesis.
- When the variability between groups is larger than the variability within groups, it suggests that the grand mean is not the best predictor of the response variable. (66%)

ANOVA output

H_0: $\mu_{\text{violent}} = \mu_{\text{sexual}} = \mu_{\text{neutral}}$

H_a: the means are not all equal

- Decision:
  - Since p<.05, we can reject the null hypothesis.
- Conclusion:
  - The data suggests that the mean recall for commercials presented during violent, sexual, and neutral TV shows differs.



Tests of Between-Subjects Effects

Dependent Variable: RECALL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	123.265[a]	2	61.633	20.452	.000
Intercept	1745.383	1	1745.383	579.177	.000
TV Rating	123.265	2	61.633	20.452	.000
Error	967.352	321	3.014		
Total	2836.000	324			
Corrected Total	1090.617	323			

a. R Squared = .113 (Adjusted R Squared = .107)

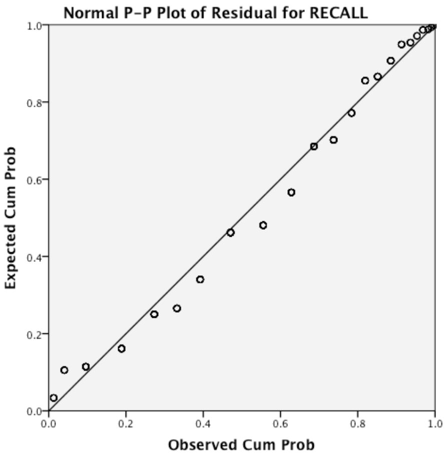
But can't get too excited about pvalue until we check the conditions of application – assumptions!

ANOVA assumptions

- Have a mean of zero
  - again automatically true due to analysis
- Equal variance: have the same standard deviation for each group
  - Residual plot
  - Rule of thumb
  - Levene's test of equal variance
- Follow a normal distribution
  - Normal probability plot
- Be independent: usually achieved by random assignment

Normality: NPP

- Same as in regression, want points to fall on the line – but have to make by hand, via saving the residuals (directions in spss document)

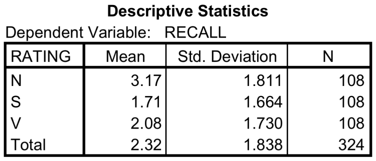


ANOVA assumptions

- Have a mean of zero
  - again automatically true due to analysis
- Equal variance: have the same standard deviation **for each group**
  - Residual plot
  - Rule of thumb
  - Levene's test of equal variance
- Follow a normal distribution
  - Normal probability plot
- Be independent: usually achieved by random assignment

Residual plot

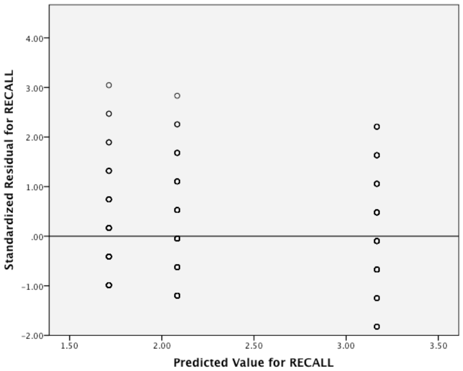
- Looks different than in regression, because we have only a categorical variable
  - Best predictor of a categorical variable is the mean
  - So get a number of column equal to the number/levels of our group
  - Each one is at the mean value of the group
- Can still look at the spread



Descriptive Statistics

Dependent Variable: RECALL

RATING	Mean	Std. Deviation	N
N	3.17	1.811	108
S	1.71	1.664	108
V	2.08	1.730	108
Total	2.32	1.838	324



What do you think?

Yes!

Equal variance: stdev rule of thumb

- Compute ratio: Find the largest and smallest standard deviation, looking across all groups

$$\frac{\text{Largest stdev}}{\text{Smallest stdev}}$$

Rule of thumb:
< 2 = OK
> 2 = concern

- Even willing to accept a ratio somewhat larger than 2 if sample size per group is small, or if designed is 'balanced'
  - When the sample sizes are the same cross all groups/ levels

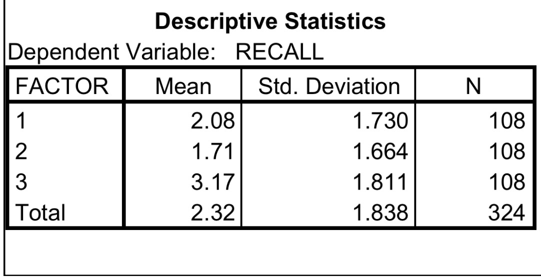
Equal variance: stdev rule of thumb

- Compute ratio: Find the largest and smallest standard deviation, looking across all groups

$$\frac{\text{Largest stdev}}{\text{Smallest stdev}}$$

Rule of thumb:
< 2 = OK
> 2 = concern

Good!



Descriptive Statistics

Dependent Variable: RECALL

FACTOR	Mean	Std. Deviation	N
1	2.08	1.730	108
2	1.71	1.664	108
3	3.17	1.811	108
Total	2.32	1.838	324

1.811/1.664 = 1.088

1 would mean perfectly equal

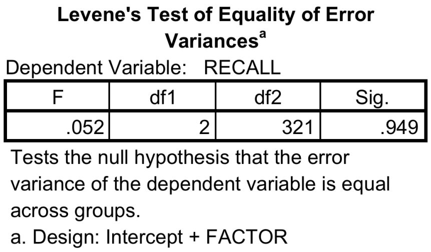
Equal variance: Levene's test

- Levene's test for homogeneity of variances
- Hypotheses:

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2$

H_a: Not all variances are equal

- For once we want the null hypothesis to be true!!!! (71%)
- Decision:
  - Since p is greater than .05, we fail to reject the null hypothesis
- Conclusion:
  - The data suggests that the condition of equal variance is not violated.



Levene's Test of Equality of Error Variances[a]

Dependent Variable: RECALL

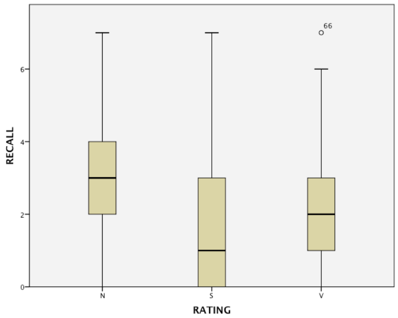
F	df1	df2	Sig.
.052	2	321	.949

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

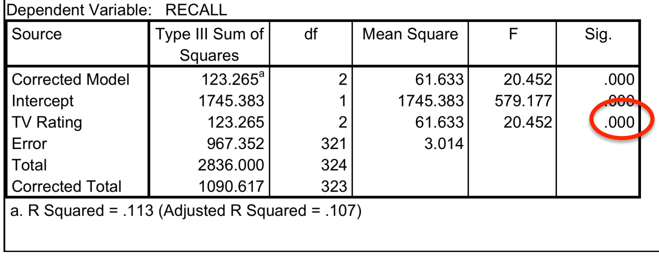
a. Design: Intercept + FACTOR

A significant difference across groups

- But it doesn't tell us which groups differ



Tests of Between-Subjects Effects



Dependent Variable: RECALL

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	123.265[a]	2	61.633	20.452	.000
Intercept	1745.383	1	1745.383	579.177	.000
TV Rating	123.265	2	61.633	20.452	.000
Error	967.352	321	3.014		
Total	2836.000	324			
Corrected Total	1090.617	323			

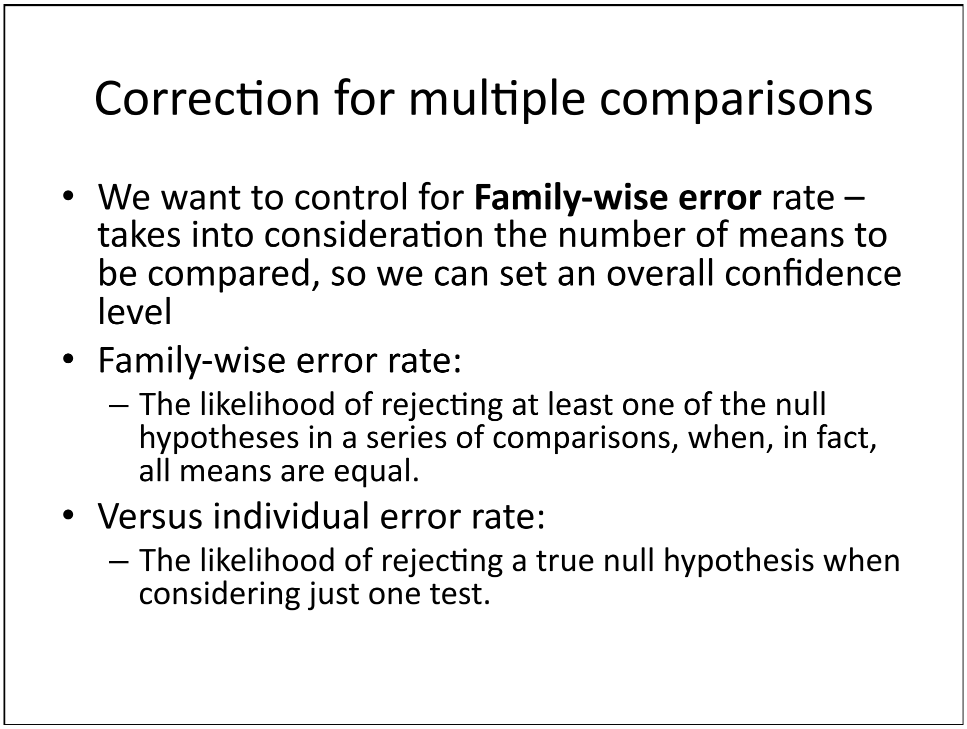
a. R Squared = .113 (Adjusted R Squared = .107)

But what groups differ?

- If we find a significant effect of our grouping variable, or a significant main effect, then we want to follow this up and find out which specific groups differ.
  - Often called posthoc tests, because run after the main analysis
  - Should not interpret if main effect is nonsignificant (52%)
- Essentially we want to run a series of t-tests to compare the groups
  - Only need when there are more than two groups, so called multiple comparisons

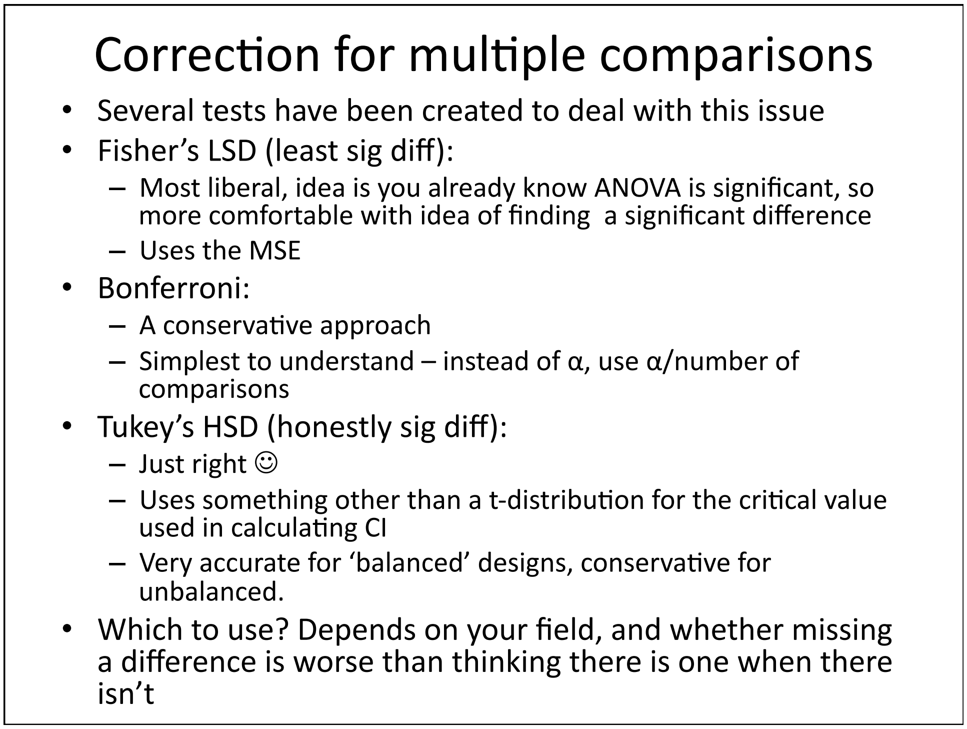
So just run a bunch of t-tests?

- But there is an issue if we just run several individual t-tests
  - If we run enough of them, we are likely to end up with something showing up significant just by chance - a Type I error
- Can think of the idea of a confidence interval, if we compare two means and 0 is not in the interval, we can be 95% confident that the true mean is in that interval
  - Which also means 5% of the time it isn't – and this is what we are worried about.
- We would really like to be '95% confident' regardless of the number of comparisons we make.
- Or if think of the p-value, it's the probability of getting a difference as large as you did if there was no difference. The probability is less than 5% (alpha=.05), but not zero.



Correction for multiple comparisons

- We want to control for **Family-wise error** rate – takes into consideration the number of means to be compared, so we can set an overall confidence level
- Family-wise error rate:
  - The likelihood of rejecting at least one of the null hypotheses in a series of comparisons, when, in fact, all means are equal.
- Versus individual error rate:
  - The likelihood of rejecting a true null hypothesis when considering just one test.



Correction for multiple comparisons

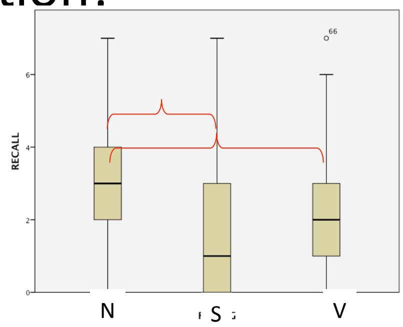
- Several tests have been created to deal with this issue
- Fisher's LSD (least sig diff):
  - Most liberal, idea is you already know ANOVA is significant, so more comfortable with idea of finding a significant difference
  - Uses the MSE
- Bonferroni:
  - A conservative approach
  - Simplest to understand – instead of α, use α/number of comparisons
- Tukey's HSD (honestly sig diff):
  - Just right ☺
  - Uses something other than a t-distribution for the critical value used in calculating CI
  - Very accurate for 'balanced' designs, conservative for unbalanced.
- Which to use? Depends on your field, and whether missing a difference is worse than thinking there is one when there isn't

See it in action!

- Hypothesis being tested is

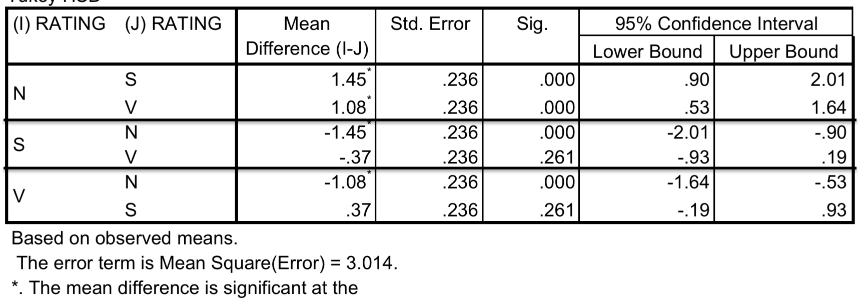
H_0: $\mu_i = \mu_j$

H_a: $\mu_i \neq \mu_j$



Multiple Comparisc

Dependent Variable: RECALL
Tukey HSD



(I) RATING	(J) RATING	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
N	S	1.45*	.236	.000	.90	2.01
	V	1.08*	.236	.000	.53	1.64
S	N	-1.45*	.236	.000	-2.01	-.90
	V	-.37	.236	.261	-.93	.19
V	N	-1.08*	.236	.000	-1.64	-.53
	S	.37	.236	.261	-.19	.93

Based on observed means.
The error term is Mean Square(Error) = 3.014.
*. The mean difference is significant at the

Compare the different tests

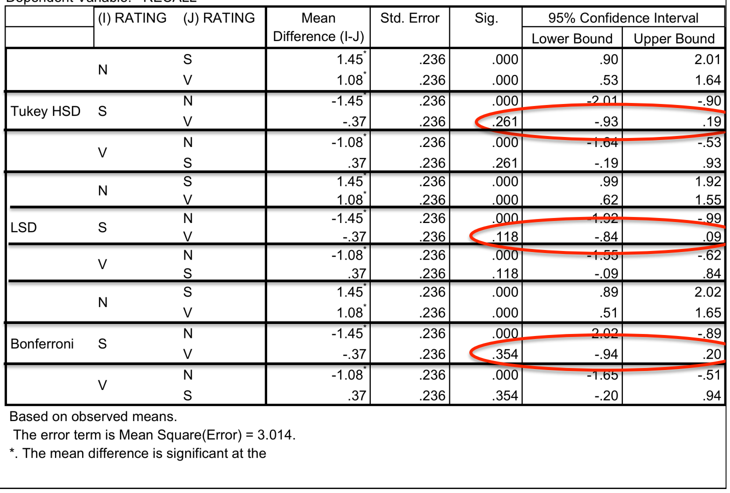
- Hypothesis being tested is essentially a ttest:

H_0: $\mu_i = \mu_j$

H_a: $\mu_i \neq \mu_j$

Multiple Comparisons

Dependent Variable: RECALL



	(I) RATING	(J) RATING	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	N	S	1.45*	.236	.000	.90	2.01
		V	1.08*	.236	.000	.53	1.64
	S	N	-1.45*	.236	.000	-2.01	-.90
		V	-.37	.236	.261	-.93	.19
	V	N	-1.08*	.236	.000	-1.64	-.53
		S	.37	.236	.261	-.19	.93
LSD	N	S	1.45*	.236	.000	.99	1.92
		V	1.08*	.236	.000	.62	1.55
	S	N	-1.45*	.236	.000	-1.92	-.99
		V	-.37	.236	.118	-.84	.09
	V	N	-1.08*	.236	.000	-1.55	-.62
		S	.37	.236	.118	-.09	.84
Bonferroni	N	S	1.45*	.236	.000	.89	2.02
		V	1.08*	.236	.000	.51	1.65
	S	N	-1.45*	.236	.000	-2.02	-.89
		V	-.37	.236	.354	-.94	.20
	V	N	-1.08*	.236	.000	-1.65	-.51
		S	.37	.236	.354	-.20	.94

Based on observed means.
The error term is Mean Square(Error) = 3.014.
*. The mean difference is significant at the

A complete example

Restoring self-control when intoxicated.

- Does coffee or anything else really allow a person suffering from alcohol intoxication to 'sober-up'? A sample of 44 male college students participated in an experiment. Each student was asked to memorize a list of 40 words, then they were given two drinks, rested for 25 min, and then were given a word completion task where they received a score based on the number of correct responses. The sample was randomly divided into four groups.
  - Group A: received only alcohol
  - Group AC: Alcohol mixed with caffeine
  - Group AR: only alcohol, but received a monetary reward for correct responses on post test
  - Group P: the placebo group, were told they were getting alcohol, but received a carbonated beverage instead.

The Four-Step Process for statistical modeling:

1. **Choose** a form for the model
   Identify the variables and their types:
   Examine graphs to help identify the appropriate model

2. **Fit** the model to the data
   Use the sample data to estimate the values of the model parameters

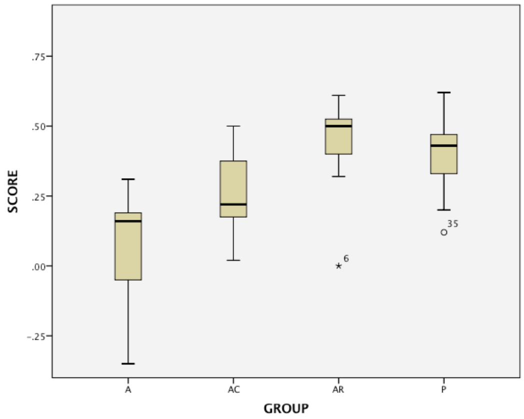
3. **Assess** how well the model fits the data
   Verify assumptions
   Examine the residuals
   Investigate significance, refine model

4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

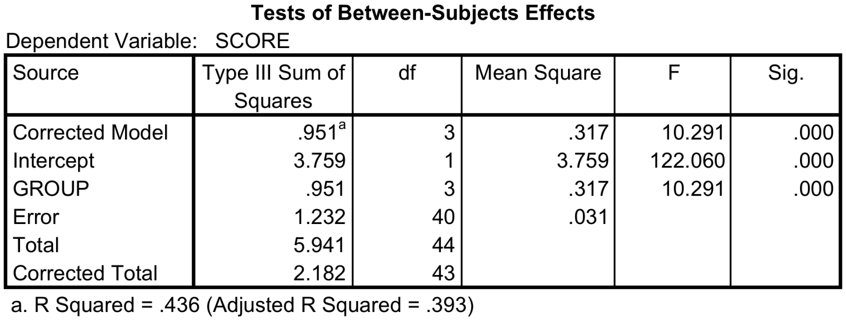
Choose

- Identify variables:
  - Which group/what kind of drink: categorical
  - Word complete score: quantitative
- Look at your data
- Is it appropriate for a one-way ANOVA?



Fit

- Can peek, but need to check assumptions first.



Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.951[a]	3	.317	10.291	.000
Intercept	3.759	1	3.759	122.060	.000
GROUP	.951	3	.317	10.291	.000
Error	1.232	40	.031		
Total	5.941	44			
Corrected Total	2.182	43			

a. R Squared = .436 (Adjusted R Squared = .393)

Assess: verify assumptions

- Have a mean of zero
- Equal variance: have the same standard deviation **for each group**
  - Residual plot
  - Rule of thumb
  - Levene's test of equal variance
- Follow a normal distribution
  - Normal probability plot
- Be independent
  - usually achieved by random assignment or random sampling

Equal variance

- Residual plot:
  - Might have an issue
- Rule of thumb
  - Max stdev/min stdev < 2

.21805/.15251 = 1.43

- Levene's test

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$

H_a: Not all variances are equal

Decision:

fail to reject null hypothesis

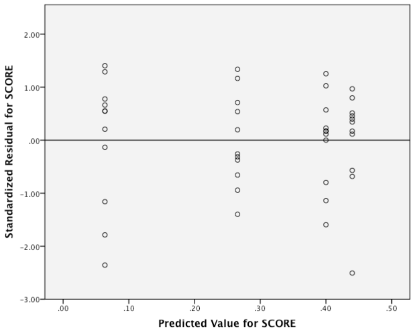
Conclusion:

Data suggests that we meet the assumption of equal variance

Descriptive Statistics

Dependent Variable: SCORE

GROUP	Mean	Std. Deviation	N
A	.0636	.21805	11
AC	.2655	.15260	11
AR	.4400	.17053	11
P	.4000	.15251	11
Total	.2923	.22528	44



Levene's Test of Equality of Error Variances[a]

Dependent Variable: SCORE

F	df1	df2	Sig.
.780	3	40	.512

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

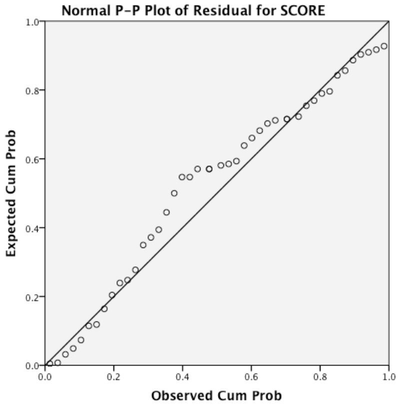
a. Design: Intercept + GROUP

What if tests don't agree?

- If one fails, fine to go ahead
- If two fails, we can go ahead
  - should be cautious in interpreting the results – though if it is a balanced design we are probably fine
- If all three fail, usually not a good sign
  - can try transforming the response variable, sqrt
  - There are posthoc tests designed for unequal variances
- What if fail the assumption of normality?
  - Should be cautious in interpreting the results – though if there are no outliers, we are probably fine
  - Log transformations of response variable

Normality

- How does it look?
  - Looks pretty good



Use: Can people sober up?

- Hypothesis

H_0: $\mu_P = \mu_A = \mu_{AC} = \mu_{AR}$

H_a: the means are not all equal

Decision:

Since p<.001, we can reject the null hypothesis that all the means are equal.

Conclusion:

The data suggests that average memory performance differs among the four groups

Tests of Between-Subjects Effects

Dependent Variable: SCORE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	.951[a]	3	.317	10.291	.000
Intercept	3.759	1	3.759	122.060	.000
GROUP	.951	3	.317	10.291	.000
Error	1.232	40	.031		
Total	5.941	44			
Corrected Total	2.182	43			

a. R Squared = .436 (Adjusted R Squared = .393)

But our question of interest is still not answered, so need to keep going – which groups differ?

- Hypothesis:

H_0: $\mu_i = \mu_j$

H_a: $\mu_i \neq \mu_j$

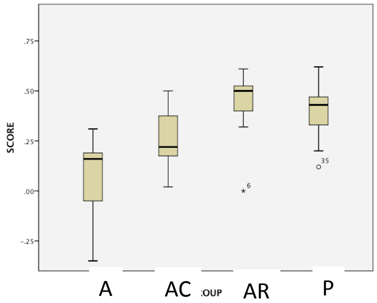
- What groups differ? Think about our question

The data suggests that the group that received only alcohol is impaired on average, compared to all other groups (A vs P, p<.001; A vs AC, p<.05; A vs AR, p<.001). This suggests that people can 'sober-up' with the help of caffeine or when there is a monetary incentive to do so

Dependent Variable: SCORE
Tukey HSD

(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	AC	-.2018*	.07482	.048	-.4024	-.0013
	AR	-.3764*	.07482	.000	-.5769	-.1758
	P	-.3364*	.07482	.000		
AC	A	.2018*	.07482	.048		
	AR	-.1745	.07482	.108		
	P	-.1345	.07482	.289		
AR	A	.3764*	.07482	.000		
	AC	.1745	.07482	.108		
	P	.0400	.07482	.950		
P	A	.3364*	.07482	.000		
	AC	.1345	.07482	.289		
	AR	-.0400	.07482	.950		

Based on observed means.
The error term is Mean Square(Error) = .031.
*. The mean difference is significant at the



Summary

- Learned a new tool for answer the question of ‘do the means of several groups differ?’
- Learned to test the assumptions of ANOVA
- Learned to interpret the output
- Learned to follow up a ‘main effect’ with posthoc tests
- Seems to have a different feel than regression
  - easier to randomly assign people to the different levels of our factor of interest
    - since a limited number of levels
    - Categorical not quantitative
  - Does this buy us anything?!

Research methods: goals

- Learn the importance of randomization
- Learn what kind of conclusions can be drawn from your study

Comparative studies

- Often want to compare more than one condition or treatment
  - Different encoding strategies and their affect on test performance
  - How do different treatments for lung cancer affect the survival rate?
- What type of analyses would you use for these studies?
  - ANOVA

Need a control group

- How do you know if something is 'better' if you have nothing to compare it to?
  - Very important in medical field
  - Lots of good examples of bad studies
- Placebo effect:
  - Simply believing you are receiving treatment can improve the situation
- Practice effects:
  - People might score better on a test after learning a new strategy just because they are more familiar with the testing situation
- Time effects:
  - People might get better after treatment because they would have gotten better with the passage of time
    - Studies of depression

Need for randomization

- People are biased!!!!
  - And we often don't know it.
  - Can't be trusted so we need a computer to keep us honest
- If your control and experimental groups don't come from the same population of subjects than you can't conclude anything about how they compare.
- Quiz question: If you wanted to study the affect of a new anti-depressant, a good control group would consist of people who haven't been depressed in the last 6 months. (85%)

Randomization is our hero!

- Prevents bias
- Permits conclusions about cause!!!
  - Which we love to draw!
- Justify using a probability model

Randomization allows for causation

- If you randomize your subjects and find a significant difference between groups
  - It has to be due to your manipulation
  - You caused the difference!
- Because the small p-value implies that you are very unlikely to observe a difference as large as you found by chance.
- Quiz question: If they find that the students who attended the group study sessions did better than students in the other two groups, they can conclude that attending the group study sessions directly caused the increase in their grades. (40%)

Experiments: randomization

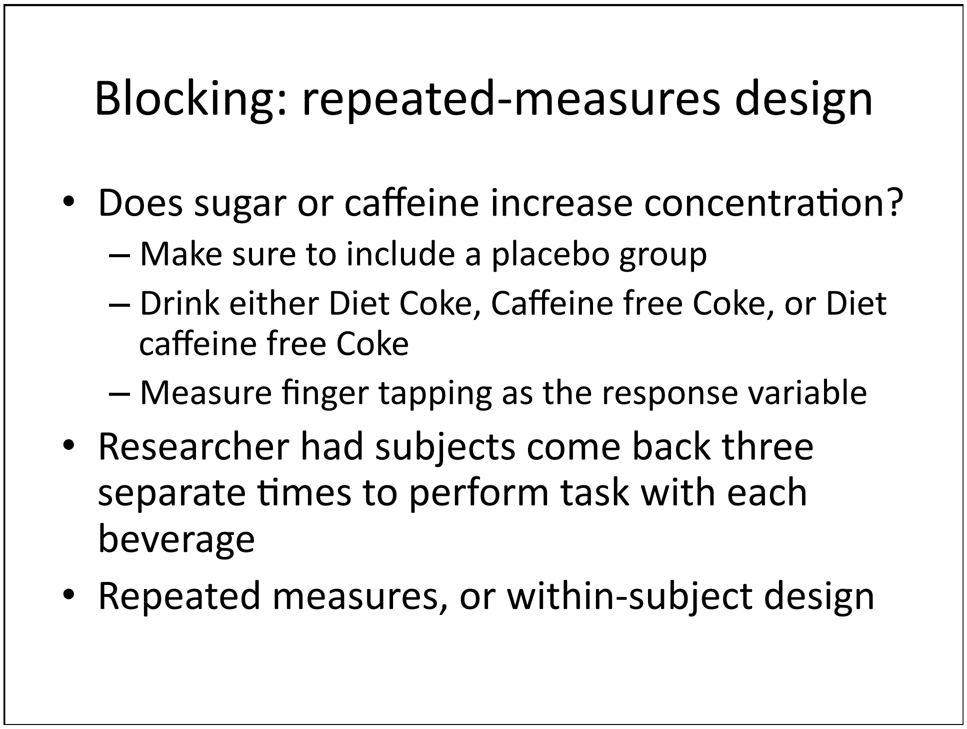
- The idea is when you randomly assign people or units to the groups, you are 'randomly' distributing other sources of error.
  - TV example: some people remember more than others, some people more or less exposed to violent tv
  - hopefully randomly got a similar number of good and bad remembers and high/low exposure in each group.
- If we find a difference, allows us to attribute it to the experimental manipulation
  - Cause and effect!!
- But still need to think about population, if it isn't a random selection, then we can't extend to the larger population.

Don’t confuse with random selection

- Random assignment to groups is separate from the random sampling from the population
  - If all the subjects in this experiment were people who watched more than 6 hours of TV a day, we can say that among heavy tv watchers, TV program content impairs recall for advertising, but that might not generalize to all tv watchers
- You can have random selection from your population without having random assignment to treatments/ groups
- Want random selection so that you can extend your findings from your sample to the larger population
  - Important for assumptions

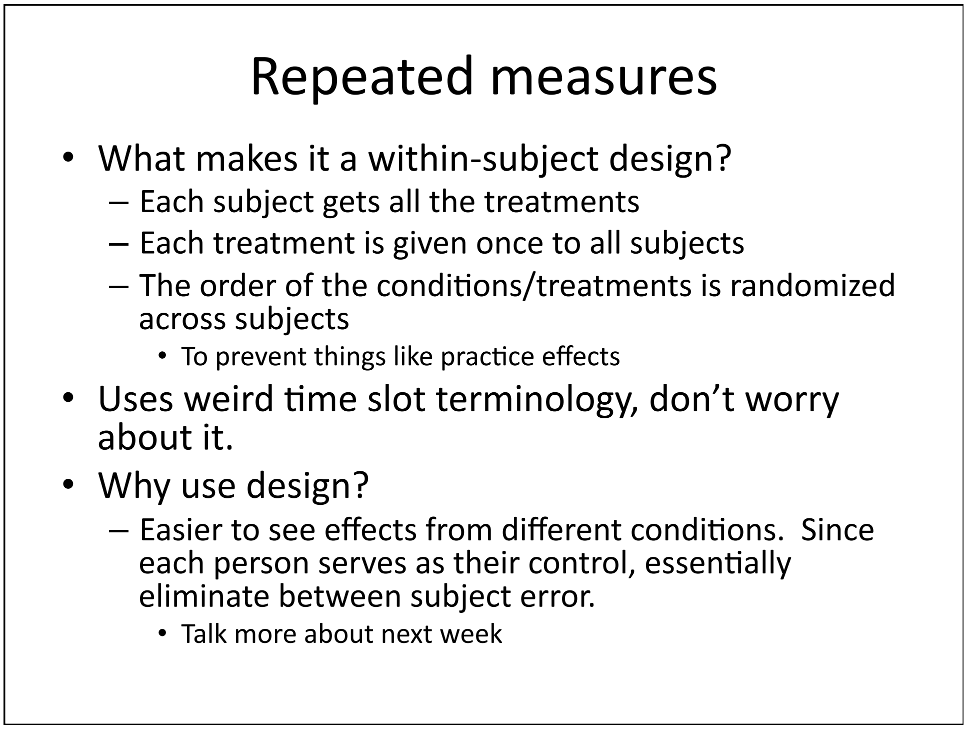
So if experiments get us cause and effect, what do observational studies get us?

- Inference in observational studies depends on random selection from the population of interest
- Random selection is also thought to try and get rid of biases, or factors that would prevent from generalizing to the greater population.
  - What if selected all drinkers from Harvard
- Can’t say cause and effect, but does ‘generalize’
  - Can’t say that drinking a lot impairs memory function, but that on average, heavy drinkers have a worse memory than non drinkers.



Blocking: repeated-measures design

- Does sugar or caffeine increase concentration?
  - – Make sure to include a placebo group
  - – Drink either Diet Coke, Caffeine free Coke, or Diet caffeine free Coke
  - – Measure finger tapping as the response variable
- Researcher had subjects come back three separate times to perform task with each beverage
- Repeated measures, or within-subject design



Repeated measures

- What makes it a within-subject design?
  - – Each subject gets all the treatments
  - – Each treatment is given once to all subjects
  - – The order of the conditions/treatments is randomized across subjects
    - To prevent things like practice effects
- Uses weird time slot terminology, don't worry about it.
- Why use design?
  - – Easier to see effects from different conditions. Since each person serves as their control, essentially eliminate between subject error.
    - Talk more about next week

Other types of blocking

- Can’t always reuse subjects/experimental units
- Farmland example: Blocks by subdividing
  - Study different kinds of corn. If wanted to grow them on the same land, would take multiple years to conduct.
- Twin example: Blocks by grouping
  - Study the effect of parents education on children’s success in school by looking at twins raised in different households

Factorial crossing

- In one-way ANOVA, we have one factor with multiple levels
  - type of TV program (neutral, violent, sexual)
- Design is ‘balanced’ if we have the same number of observations for each level
  - 10 people watch neutral, 10 watch violent, 10 watch sexual

Factorial crossing

- In two-way ANOVA we have two factors, each with multiple levels
- Two factors are crossed if all combinations of levels of two factors appear in the design
  - If you crossed type of TV program (neutral, violent, sexual) with Animation type (Animated and not animated)
    - Would have Animated neutral, animated violent, animated sexual, nonanimated neutral, nonanimated violent, nonanimated sexual
  - Each combination of factor levels is called a cell or a treatment
- A design is balanced if there are equal numbers of observations per cell (equal #s of experimental units)
- In ANOVA: in order to test for the presence of interaction effects, you must have more than one observation per cell.
  - Say you want to investigate whether math and verbal skills differ by school across boston– but only have the average score for each school.

Factorial crossing: advantages

- Usually can investigate two factors instead of one for very little additional ‘cost’
- Allows you to see interactions.

Example

- Researcher wants to investigate what treatment works best for depression. They randomly assign depressed people to three groups, a control group who receives no treatment, a group that receives Cognitive Behavioral therapy, and a group that receives antidepressants.

Factor A: Depression treatments

Control	CBT	Meds
30	30	30

Example

- But what if therapy effectiveness differs for men and women. Can easily investigate if just make sure some of the people are men and women
- Can we investigate the interaction of treatment and gender?
  - Yes, there are multiple experimental units in each cell

Factor A: Depression treatments

		Control	CBT	Meds
Factor B: Gender	men	15	15	15
	women	15	15	15

- Factors?
  - Type of depression treatment and gender
- Levels of each factor?
  - Type of depression Treatment: 3 types of treatment
  - Gender: men or women
- Whether observation or experimental?
  - Type ofdepression treatment: experimental
  - Gender: observational
- What are experimental units?
  - subjects
- What are the treatments/cells?
  - Cognitive behavioral therapy/men, cognitive behavioral therapy/women, anti-depressants/men, anti-depressants/women, control/men, control/women
- Is it a balanced design?
  - Yes!

		Factor A: Depression treatments		
		Control	CBT	Meds
Factor B: Gender	men	15	15	15
	women	15	15	15

Summary

- Randomize!
  - Experiments:
    - Randomly assigning subjects to groups/levels allows you to infer cause and effect
  - In both observational and experiments, if your sample is randomly selected from population
    - Allows you to generalize results to larger population
    - In general, or on average, the results will be true.
- If want to compare a new technique, test, or treatment, need a control group
- Factorial crossing in ANOVA is easy and powerful