Week 5: Two-way ANOVA and Repeated Measures ANOVA

Time stamper!!

Quizzes and canvas

Midterm/final

Two-way ANOVA

- When you have two categorical predictor variables (each with more than one level) and one quantitative response variable
  - Can keep going three-way, four-way, everything generally the same, but interpretation gets more complicated (88% correct)
- Didn't have you read the beginning of the chapter, because the first example they use is more of a 'repeated' measures ANOVA
  - Where each subject participates in each level
    - Book called this a randomized complete block design
  - Analyzed differently in SPSS
  - Will talk about later

Two-way ANOVA

- Variability between group still compared to variability with the groups, but now this is looked at separately for each grouping variable
- The assumptions are the same as for one-way ANOVA

TOTAL variation in response, *Y* = Variability between groups + Variability within group

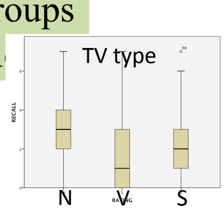
Variability between groups:
- Variability between groups of variable A
- Variability between groups of variable B
- Variability due to interaction of A and B

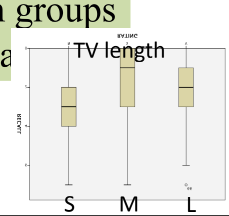
Two-way ANOVA

- If we think of tv example from before, maybe we want to measure commercial recall for not just tv type (neutral, violent, sexual), but also length of program (short, medium, long)

TOTAL variation in response, *Y* = Variability between groups + Variability within group

Variability between groups:
- Variability between groups of variabl
- Variability between groups of varia





Some terminology

- TV Type
  - 3 levels: Neutral, Violent, Sexual
- TV Length
  - 3 levels: Short, Medium, Long
- TV Type and TV Length are called:
  - Factors
  - Each factor tested separately (versus grand mean for that factor)
  - When test each factor, its called a main effect
  - Main effect of TV Type, would investigate whether recall differs for different TV types, regardless of length

Example: Do students at all levels of academic ability benefit from a practice exam?

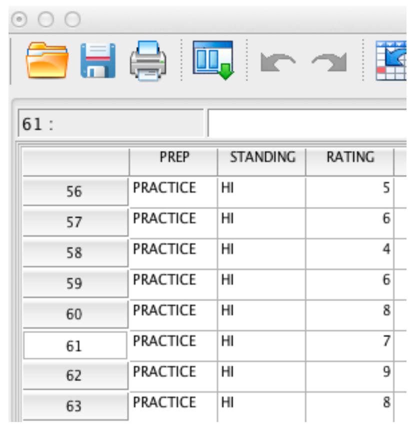
- 132 students in an introduction to psychology class divided into three groups based on class standing (hi, medium, low), and into two groups based on whether they attended a review section or took a practice exam prior to the final exam. After completing the final exam, they rated their exam preparation on an 11 point scale.
- What are the factors?
  - Class standing, type of preparation
- How many levels does each one have?
  - 3 for class standing, 2 for type of prep
- What is the response variable?
  - Rating of exam preparation
- What are the treatments/cells?
  - Hi-review, med-review, low-review, hi-prac, med-prac, low-prac
- The ANOVA is often referred to by the number of levels, so this would be a 3x2 ANOVA

What does the data look like?

- 22 students in each 'cell'
- Each experimental unit is a line, has a value for each factor
  - Each subject has a standing and a preparation value

Class Standing / Type of preparation

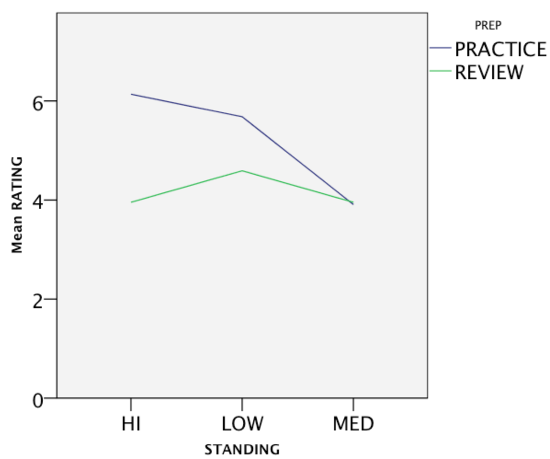
	Review Section	Practice Exam
Hi	Students in hi standing that went to a review section	Students in hi standing that took a practice exam
Medium	Students in medium standing that went to a review section	Students in medium standing that took a practice exam
Low	Students in low standing that went to a review section	Students in low standing that took a practice exam


61 :

	PREP	STANDING	RATING
56	PRACTICE	HI	5
57	PRACTICE	HI	6
58	PRACTICE	HI	4
59	PRACTICE	HI	6
60	PRACTICE	HI	8
61	PRACTICE	HI	7
62	PRACTICE	HI	9
63	PRACTICE	HI	8

Choose

- Saw already that we have two categorical and quantitative variable, lets make sure to look at our data.
- Use a line graph, where we use the factor with the smallest number of levels to create multiple lines
- Called an interaction plot, made through the ANOVA



Fit

- Looks a little different….. Can peek at results

Tests of Between-Subjects Effects

Dependent Variable: RATING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.068[a]	5	21.014	5.677	.000
Intercept	2921.523	1	2921.523	789.247	.000
PREP	38.189	1	38.189	10.317	.002
STANDING	39.591	2	19.795	5.348	.006
PREP * STANDING	27.288	2	13.644	3.686	.028
Error	466.409	126	3.702		
Total	3493.000	132			
Corrected Total	571.477	131			

a. R Squared = .184 (Adjusted R Squared = .151)

Assess: verify assumptions

- Have a mean of zero
- Equal variance: have the same standard deviation **for each group – each cell/treatment**
  - Residual plot
  - Rule of thumb: gets a little crazy would have 6 stdev
  - Levene's test of equal variance
- Follow a normal distribution
  - Normal probability plot
- Be independent
  - usually achieved by random assignment or random sampling
  - But all from one psychology class……

Assess the model

Equal variance:

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$

H_a: Not all variances are equal

- – Fail to reject null hypothesis that all variances are equal, this along with the residual plot suggests we meet the assumption of equal variance
- But only has 5 'groups' not 6

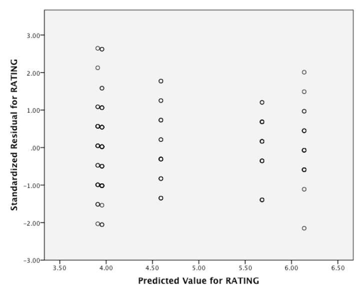
Levene's Test of Equality of Error Variances[a]

Dependent Variable: RATING

F	df1	df2	Sig.
1.314	5	126	.262

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

a. Design: Intercept + PREP + STANDING + PREP * STANDING



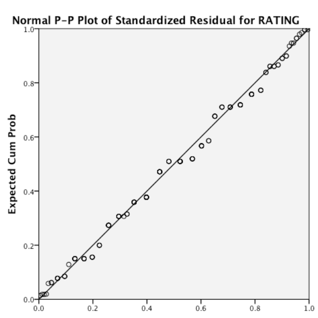
Descriptive Statistics

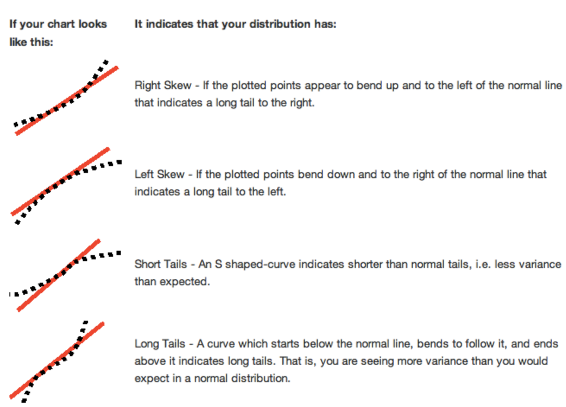
Dependent Variable: RATING

PREP	STANDING	Mean	Std. Deviation	N
PRACTICE	HI	6.14	1.726	22
	LOW	5.68	1.585	22
	MED	3.91	2.245	22
	Total	5.24	2.083	66
REVIEW	HI	3.95	1.618	22
	LOW	4.59	1.843	22
	MED	3.95	2.380	22
	Total	4.17	1.966	66
Total	HI	5.05	1.988	44
	LOW	5.14	1.786	44
	MED	3.93	2.286	44
	Total	4.70	2.089	132

Assess the model: normality

- Normality assumption met?
  – Yes!





Use

- Going to answer 3 questions
  - Are there significant differences in the mean preparation score for students with different class standings?
  - Are there significant differences in the mean preparation score for students who attend a review section compared to those who take a practice exam
  - Is there a significant interaction in the mean preparation scores between class standing and preparation method?

Tests of Between-Subjects Effects

Dependent Variable: RATING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.068[a]	5	21.014	5.677	.000
Intercept	2921.523	1	2921.523	789.247	.000
PREP	38.189	1	38.189	10.317	.002
STANDING	39.591	2	19.795	5.348	.006
PREP * STANDING	27.288	2	13.644	3.686	.028
Error	466.409	126	3.702		
Total	3493.000	132			
Corrected Total	571.477	131			

a. R Squared = .184 (Adjusted R Squared = .151)

Use

- Going to answer 3 questions
  - Are there significant differences in the mean preparation score for students with different class standings?
  - Are there significant differences in the mean preparation score for students who attend a review section compared to those who take a practice exam
  - Is there a significant interaction on preparation scores between class standing and preparation method?
- The first two are tests for 'main effects', and the hypotheses are the same as in a 1-way ANOVA

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

H_a: the means are not all equal

- The third test is for the interaction
  - For now, lets ignore this one

Use

- Test the main effect of type of preparation:

H_0: $\mu_{review} = \mu_{pracexam}$

H_a: the means are not all equal

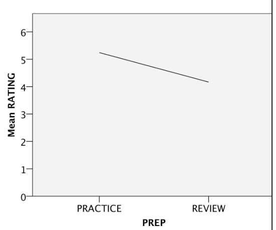
- Decision:
  - Since p<.01 and F=10.32, we can reject the null hypothesis.
- Conclusion:
  - This suggests that students in psychology feel more prepared on average when they take a practice exam compared to if they attend a review section
- When you investigate the main effect of one factor, you average across the other – the main effect of type or preparation is thought to be the same for ALL LEVELS of the other factor

Tests of Between-Subjects Effects

Dependent Variable: RATING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.068[a]	5	21.014	5.677	.000
Intercept	2921.523	1	2921.523	789.247	.000
PREP	38.189	1	38.189	10.317	.002
STANDING	39.591	2	19.795	5.348	.006
PREP * STANDING	27.288	2	13.644	3.686	.028
Error	466.409	126	3.702		
Total	3493.000	132			
Corrected Total	571.477	131			

a. R Squared = .184 (Adjusted R Squared = .151)



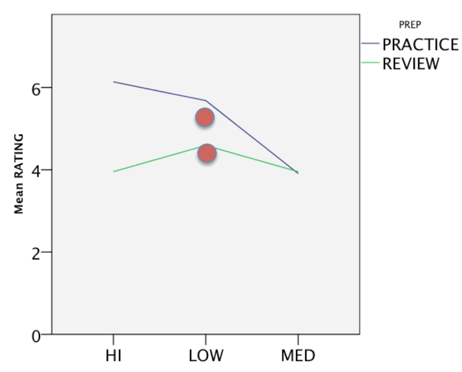
Main effect of preparation

- Average across the levels of the other factor

Type of preparation

Class Standing

	Review Section	Practice Exam
Hi	Students in hi standing that went to a review section	Students in hi standing that took a practice exam
Medium	Students in medium standing that went to a review section	Students in medium standing that took a practice exam
Low	Students in low standing that went to a review section	Students in low standing that took a practice exam

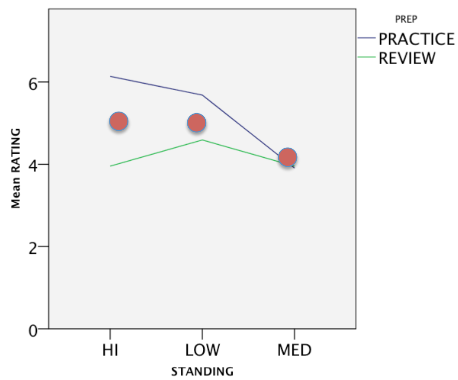


Main effect of Class Standing

- Average across the levels of the other main effect
- Can do this because we think each factor is independent of each other

Type of preparation

Class Standing	Review Section	Practice Exam
Hi	Students in hi standing that went to a review section	Students in hi standing that took a practice exam
Medium	Students in medium standing that went to a review section	Students in medium standing that took a practice exam
Low	Students in low standing that went to a review section	Students in low standing that took a practice exam



Use

- Test the main effect of class standing:

H_0: $\mu_{hi} = \mu_{med} = \mu_{low}$

H_a: the means are not all equal

Decision: Since p<.01, we can reject the null hypothesis.

Conclusion:

  – This suggests that students mean rating of preparedness differs for different class standings.

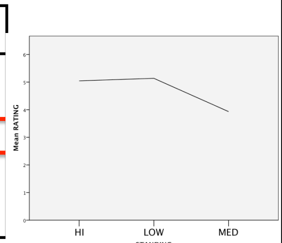
- What would we do to find out which class standings differ?
  – Posthoc tests: Tukey's
- Going to hold off for now and investigate interaction term

Tests of Between-Subjects Effects

Dependent Variable: RATING

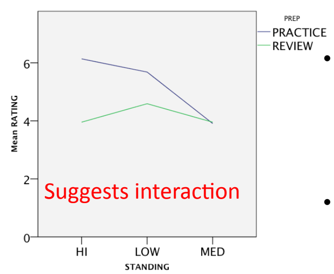
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.068[a]	5	21.014	5.677	.000
Intercept	2921.523	1	2921.523	789.247	.000
PREP	38.189	1	38.189	10.317	.002
STANDING	39.591	2	19.795	5.348	.006
PREP * STANDING	27.288	2	13.644	3.686	.028
Error	466.409	126	3.702		
Total	3493.000	132			
Corrected Total	571.477	131			

a. R Squared = .184 (Adjusted R Squared = .151)

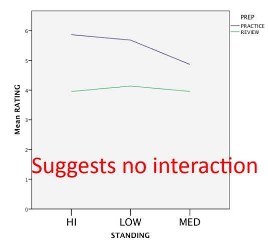


What about interactions?

- Just like in regression, these higher order terms mess things up!!!!
- If there is an interaction, it implies you can't just simply collapse across, or average across the other factor, because the two factors interact with each other!
  - Interpreting the main effects makes no sense!!
  - Makes F-tests of main effects less trustworthy
  - Like in regression, an interaction implies the 'lines' aren't parallel.



- Quick test for whether there might be an interaction is, whether the lines are parallel
- Some say if the lines cross, but really there are two ways the graph can be plotted, so you would need to plot both to know if cross (32% correct)
- So whether they deviate from parallel is best



Formal test for interaction

- Hypothesis

H_0: the main effect of each factor is the same for each level of the other factor

H_a: the two factors interact

- Decision:
  - Since p<.05, we can reject the null hypothesis that the two factors are independent of each other
- Conclusion:
  - The data suggests that students' average rating of preparedness depends on a combination of the class standing and type of preparation.
- Not very satisfying.

Tests of Between-Subjects Effects

Dependent Variable: RATING

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.068[a]	5	21.014	5.677	.000
Intercept	2921.523	1	2921.523	789.247	.000
PREP	38.189	1	38.189	10.317	.002
STANDING	39.591	2	19.795	5.348	.006
$\boxed{\text{PREP * STANDING}}$	27.288	2	13.644	3.686	.028
Error	466.409	126	3.702		
Total	3493.000	132			
Corrected Total	571.477	131			

a. R Squared = .184 (Adjusted R Squared = .151)

What if there is an interaction?

- The data suggests that students average rating of preparedness depends on a combination of the class standing and type of preparation.
- This usually isn't very useful in helping to answer our question of interest:
  - Do students at all levels of academic ability benefit from a practice exam
- Depends on field, but might follow

up with a series of t-tests, or one-way ANOVAs.

- Follow up by testing simple main effects
  - Pulling apart two factors
  - Investigate separately for each level of other factor

H_0: $\mu_{hireview} = \mu_{hipracexam}$ and
H_0: $\mu_{medreview} = \mu_{medpracexam}$ and
H_0: $\mu_{lowreview} = \mu_{lowpracexam}$

- But have to be careful of multiple comparisons

Class Standing	Type of preparation: Review Section	Type of preparation: Practice Exam
Hi	Students in hi standing that went to a review section	Students in hi standing that took a practice exam
Medium	Students in medium standing that went to a review section	Students in medium standing that took a practice exam
Low	Students in low standing that went to a review section	Students in low standing that took a practice exam

What if significant interaction

- This implies we should first look at whether the interaction is significant
- Because if it is, don't want to interpret main effects (82% correct).

Summary

- Learned about two-way ANOVA
- Same assumptions as one-way
- Learned about interactions – higher order terms that mess things up
  – Can't simply interpret main effects in same way as you would if not an interaction.

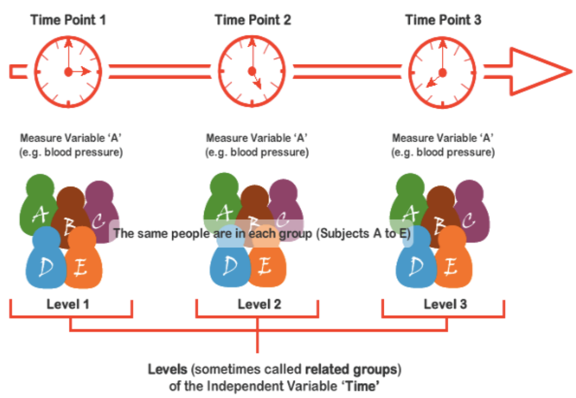
Last kind of ANOVA: Repeated measures

- This is your friend – very common in lots of fields of psychology.
- Easiest way to get a significant effect!

Repeated measures ANOVA

- Ideal for experimental designs
- Most commonly used when you are investigating an effect over time (longitudinal study)
  - Like weight loss
- But really used anytime the same people participate in all levels of a factor (94% correct)

Called within-subject factors
vs
Between subject factors:
- different people in different levels



Why RM ANOVA so cool ☺

- Remember idea of ANOVA is to compare the between group variability (what we are manipulating) to the within group variability (error due to randomness across subjects)
- Now RM ANOVA can further subdivide the error term – can predict subject variability (Because have more than one measure per subj),
  - So reduces the error term, increasing significance!!!!

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

SS(total) = Variability between groups + Variability within group

Variability within group → Subject variability + Error variability

Assumptions similar: Sphericity

- Still have normality and randomness
- Except now instead of equal variance for each group, you have the assumption of sphericity
- Sphericity: is the condition where the variances of the differences between all combinations of related levels are equal
  - Quiz question: In repeated measures ANOVA, one of the assumptions is that the variances of each treatment must be equal -> 37% correct
- This is a big deal in RM ANOVA – but easily testable and fixable
  - Mauchly's test of spehricity

  H_0: the variances of the differences are equal

  H_a: The variances of the differences are not equal
- If reject the null hypothesis, and fail the assumption of sphericity, can use the Greenhouse-Geisser correction
  - Changes degrees of freedom to correct for the increased risk of Type I error

Dataset: Diets and exercise

- Participants were randomly assigned to 2 different diet plans
- Their pulse was measured during 3 different types of exercise, low impact, med, and high impact
- What are the factors? And how many levels does each have?
  - Diet: 2, exercise: 3
- What type of factor are they?
  - Diet: between subject, exercise: within subject

Data organized differently

- Now each subject participated in multiple levels of a factor, and need to link them from level to level

	pulse1	pulse2	pulse3	diet
1	112.00	166.00	215.00	1.00
2	111.00	166.00	225.00	1.00
3	89.00	132.00	189.00	1.00
4	95.00	134.00	186.00	2.00
5	66.00	109.00	150.00	2.00
6	69.00	119.00	177.00	2.00
7	125.00	177.00	241.00	1.00

Repeated Measures: Sphericity

- Hypothesis:
  H_0: the variances of the differences are equal
  H_a: The variances of the differences are not equal
- Decision/conclusion:
- Because p<.05, we reject the null hypothesis that the variances of the differences are equal. This suggests that we fail to meet the assumption of sphericity, and should use the Greenhouse-Geisser corrected significance.

Mauchly's Test of Sphericity[a]

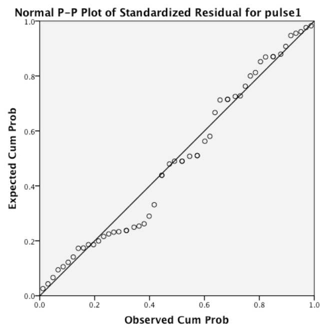
Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
exerlevel	.502	10.328	2	.006	.668	.751	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.
a. Design: Intercept + diet
Within Subjects Design: exerlevel
b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Normality

- Need to make the plot again just like in other ANOVA's, but saves the residuals out in multiple columns, one for each level of your repeated measures. Need to copy and paste them all into one column before plotting.



Repeated Measures: within subj effects

- Main effect of exercise intensity

H_0: $\mu_{low} = \mu_{med} = \mu_{high}$ H_a: the means are not all equal

- Decision:
  - p<.001, reject null
- Conclusion:
  - Average pulse rate differed significantly (p<.001, Greenhouse-Geisser corrected) across the three exercise intensity levels, regardless of diet plan (interaction p=.118, Greenhouse Geisser corrected)

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
exerlevel	Sphericity Assumed	93972.111	2	46986.056	690.666	.000
	Greenhouse-Geisser	93972.111	1.335	70370.245	690.666	.000
	Huynh-Feldt	93972.111	1.503	62533.755	690.666	.000
	Lower-bound	93972.111	1.000	93972.111	690.666	.000
exerlevel * diet	Sphericity Assumed	344.926	2	172.463	2.535	.095
	Greenhouse-Geisser	344.926	1.335	258.295	2.535	.118
	Huynh-Feldt	344.926	1.503	229.531	2.535	.112
	Lower-bound	344.926	1.000	344.926	2.535	.131
Error(exerlevel)	Sphericity Assumed	2176.963	32	68.030		
	Greenhouse-Geisser	2176.963	21.366	101.888		
	Huynh-Feldt	2176.963	24.044	90.541		
	Lower-bound	2176.963	16.000	136.060		

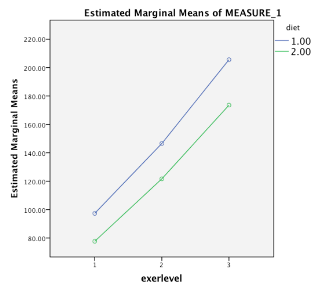
Between subject effects

- Main effect of diet plan

H_0: $\mu_{diet1} = \mu_{diet2}$

H_a: the means are not all equal

- Decision:
  - p<.01, reject null
- Conclusion:
  - Average pulse rate differed significantly between the diet plans (p<.001), such that on average people in diet plan 1 had a higher pulse rate than those on diet2.



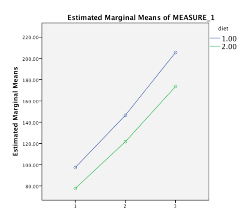
Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1014348.167	1	1014348.167	1113.315	.000
diet	8791.130	1	8791.130	9.649	.007
Error	14577.704	16	911.106		

But which exercise levels differed?



- Can't do traditional posthoc tests….

But spss will do something similar:

- Choices are LSD, Bonferroni, or Sidak
  - Sidak is the 'just right' test this time
- Conclusion:
  - Average pulse rate differed for all three exercise levels (all p<.01), with pulse rate increasing from level1 to level3.

Pairwise Comparisons

Measure: MEASURE_1

(I) exerlevel	(J) exerlevel	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
1	2	-46.611*	2.155	.000	-52.354	-40.868
	3	-102.056*	3.589	.000	-111.620	-92.491
2	1	46.611*	2.155	.000	40.868	52.354
	3	-55.444*	2.269	.000	-61.491	-49.398
3	1	102.056*	3.589	.000	92.491	111.620
	2	55.444*	2.269	.000	49.398	61.491

Based on estimated marginal means

*. The mean difference is significant at the

b. Adjustment for multiple comparisons: Sidak.