

ReCentering Psych Stats

Lynette H. Bikos, PhD, ABPP (she/her)

Last updated 15 Jul 2023

Contents

1	Introduction	27
1.1	What to expect in each chapter	27
1.2	Strategies for Accessing and Using this OER	28
1.3	If You are New to R	28
1.4	Introduction to the Data Set Used for Homeworked Examples	29
1.4.1	The Data Set	29
2	Ready_Set_R	33
2.1	Navigating this Lesson	33
2.1.1	Learning Objectives	33
2.2	downloading and installing R	33
2.2.1	So many paRts and pieces	33
2.2.2	oRienting to R Studio (focusing only on the things we will be using first and most often)	34
2.3	best pRactices	35
2.3.1	Everything is documented in the .rmd file	35
2.3.2	Setting up the file	35
2.3.3	Script in chunks and everything else in the “inline text” sections	36
2.3.4	Managing packages	36
2.3.5	Upload the data	37
2.3.5.1	To and from .csv files	38
2.3.5.2	To and from .rds files	38
2.3.5.3	From SPSS files	39
2.4	quick demonstRation	39
2.5	the knitted file	39
2.6	tRoubleshooting in R maRkdown	40

2.7 just <i>why</i> have we tRansitioned to R?	40
2.8 stRategies for success	41
2.9 Resources for getting staRted	41
2.10 Practice Problems	42
2.11 Homeworked Example	42
2.11.1 Perform a simple mathematical operation:	42
2.11.2 Install at least three packages we will commonly use:	43
2.11.3 Copy the simulation in this lesson to your .rmd file. Change the random seed and run the simulation.	43
2.11.4 Save the resulting data as a .csv or .rds file in the same file as you saved the .rmd file.	43
2.11.5 Clear your environment (broom in upper right).	44
2.11.6 Run the describe() function from the psych package with your simulated data that you imported from your local drive.	44
3 Preliminary Analyses	45
3.1 Navigating this Lesson	45
3.1.1 Learning Objectives	45
3.1.2 Planning for Practice	46
3.1.3 Readings & Resources	46
3.2 Research Vignette	46
3.3 Variable Types (Scale of Measurement)	50
3.3.1 Measurement Scale	50
3.3.2 Corresponding Variable Structure in R	51
3.4 Descriptive Statistics	53
3.4.1 Measures of Central Tendency	54
3.4.1.1 Mean	54
3.4.1.2 Median	55
3.4.1.3 Mode	55
3.4.1.4 Relationship between mean, median, and mode	56
3.5 Variability	57
3.5.1 Range	57
3.5.2 Percentiles, Quantiles, Interquartile Range	58
3.5.3 Deviations around the Mean	60
3.5.4 Variance	63

CONTENTS	5
----------	---

3.5.5 Standard Deviation	65
3.6 Are the Variables Normally Distributed?	67
3.6.1 Skew and Kurtosis	67
3.6.2 Shapiro-Wilk Test of Normality	72
3.7 Relations between Variables	73
3.8 Shortcuts to Preliminary Analyses	76
3.8.1 SPLOM	76
3.8.2 apaTables	78
3.9 An APA Style Writeup	79
3.10 Practice Problems	79
3.10.1 Problem #1: Change the Random Seed	80
3.10.2 Problem #2: Swap Variables in the Simulation	80
3.10.3 Problem #3: Use (or Simulate) Your Own Data	80
3.10.4 Grading Rubrics	80
3.11 Homeworked Example	81
3.11.1 Working the Problem with R and R Packages	81
3.11.1.1 Create a df with 3 continuously scaled variables of interest	81
3.11.1.2 Create a df with 3 continuously scaled variables of interest	82
3.11.1.3 Produce descriptive statistics	82
3.11.1.4 Produce SPLOM/pairs.panels	82
3.11.1.5 Produce an apaTables matrix	83
3.11.1.6 Produce an APA Style write-up of the preliminary analyses	83
3.11.2 Hand Calculations	83
3.11.2.1 Create a variable that represents the mean.	84
3.11.2.2 Create a variable that represents the mean deviation.	84
3.11.2.3 What is the value of the sum of mean deviations?	84
3.11.2.4 Create a variable that represents the absolute mean deviation.	84
3.11.2.5 Create a variable that represents the mean deviation squared.	84
3.11.2.6 Using the same general approach, calculate the mean deviation and standard deviation for a second, continuously scaled variable.	85
3.11.2.7 Create a variable that represents the <i>cross-product</i> (of the mean deviations). What is the sum of these cross-products?	85
3.11.2.8 Calculate the value of their covariance.	85
3.11.2.9 Calculate value of correlation coefficient.	86

4 One Sample <i>t</i>-tests	89
4.1 Navigating this Lesson	89
4.1.1 Learning Objectives	89
4.1.2 Planning for Practice	90
4.1.3 Readings & Resources	90
4.1.4 Packages	90
4.2 <i>z</i> before <i>t</i>	91
4.2.1 Simulating a Mini Research Vignette	92
4.2.2 Raw Scores, <i>z</i> -scores, and Proportions	93
4.2.3 Determining Probabilities	94
4.2.4 Percentiles	98
4.2.5 Transforming Variables to Standard Scores	98
4.2.6 The One-Sample <i>z</i> test	99
4.3 Introducing the One-Sample <i>t</i> -test	101
4.3.1 Workflow for the One-Sample <i>t</i> -test	102
4.4 Research Vignette	103
4.4.1 Data Simulation	104
4.4.2 Quick Peek at the Data	105
4.5 Working the One Sample <i>t</i> -test (by hand)	107
4.5.1 Stating the Hypothesis	107
4.5.2 Calculating the <i>t</i> -test	108
4.5.2.1 Statistical Significance	108
4.5.2.2 Confidence Intervals	109
4.5.2.3 Effect size	110
4.6 Working the One-Sample <i>t</i> -test with R Packages	111
4.6.1 Evaluating the Statistical Assumptions	111
4.6.1.1 Is the Test Variable Normally Distributed?	112
4.6.2 Computing the <i>t</i> -test	113
4.7 APA Style Results	115
4.8 Power in One-Sample <i>t</i> -tests	117
4.9 Practice Problems	119
4.9.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed	119

4.9.2	Problem #2: Rework the research vignette, but change something about the simulation	119
4.9.3	Problem #3: Use other data that is available to you	119
4.9.4	Grading Rubric	119
4.10	Homeworked Example	120
4.10.1	Working the Problem with R and R Packages	121
4.10.1.1	Narrate the research vignette, describing the variables and their role in the analysis	121
4.10.1.2	Simulate (or import) and format data	121
4.10.1.3	Evaluate statistical assumptions	121
4.10.1.4	Conduct a one sample <i>t</i> test (with an effect size)	122
4.10.1.5	APA style results with table(s) and figure	123
4.10.1.6	Conduct power analyses to determine the power of the current study and a recommended sample size	124
4.10.2	Hand Calculations	125
4.10.2.1	Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	125
4.10.2.2	Calculate the mean of your sample; identify the mean of your benchmarking sample	125
4.10.2.3	Using the steps from the previous lesson, hand-calculate the standard deviation of your sample. This should involve variables representing the mean, mean deviation, and mean deviation squared . . .	125
4.10.2.4	Calculate the one-sample <i>t</i> -test	126
4.10.2.5	Identify the degrees of freedom associated with your <i>t</i> -test	126
4.10.2.6	Locate the test critical value for your test	126
4.10.3	Is the <i>t</i> -test statistically significant? Why or why not?	126
4.10.3.1	What is the confidence interval around your sample mean?	126
4.10.3.2	Calculate the effect size (i.e., Cohen's <i>d</i> associated with your <i>t</i> -test . . .	127
5	Independent Samples <i>t</i>-test	129
5.1	Navigating this Lesson	129
5.1.1	Learning Objectives	129
5.1.2	Planning for Practice	130
5.1.3	Readings & Resources	130
5.1.4	Packages	130
5.2	Introducing the Independent Samples <i>t</i> -Test	131

5.2.1	Workflow for Independent Samples <i>t</i> -Test	131
5.3	Research Vignette	133
5.3.1	Data Simulation	133
5.3.2	Quick Peek at the Data	135
5.4	Working the Independent Samples <i>t</i> -Test (by hand)	136
5.4.1	Stating the Hypothesis	136
5.4.2	Calculating the <i>t</i> -Test	137
5.4.2.1	Statistical Significance	138
5.4.2.2	Confidence Intervals	139
5.4.2.3	Effect Size	140
5.5	Working the Independent Samples <i>t</i> -Test with R Packages	140
5.5.1	Evaluating the Statistical Assumptions	142
5.5.1.1	Is the dependent variable normally distributed at each level of the grouping variable?	142
5.5.1.2	Are the variances of the dependent variable similar across the levels of the grouping factor?	144
5.5.1.3	APA style write-up of testing the assumptions	145
5.5.2	Computing the Independent Samples <i>t</i> -Test	146
5.5.3	What if we had violated the homogeneity of variance assumption?	148
5.6	APA Style Results	148
5.7	Power in Independent Samples <i>t</i> -tests	151
5.8	Practice Problems	154
5.8.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed	154
5.8.2	Problem #2: Rework the research vignette, but change something about the simulation	154
5.8.3	Problem #3: Rework the research vignette, but swap one or more variables .	154
5.8.4	Problem #4: Use other data that is available to you	155
5.8.5	Grading Rubric	155
5.9	Homeworked Example	156
5.9.1	Working the Problem with R and R Packages	156
5.9.1.1	Narrate the research vignette, describing the variables and their role in the analysis	156
5.9.1.2	Simulate (or import) and format data	156
5.9.1.3	Evaluate statistical assumptions	157

5.9.1.4	Conduct an independent samples <i>t</i> -test (with an effect size and 95%CIs)	159
5.9.1.5	APA style results with table(s) and figure	159
5.9.1.6	Conduct power analyses to determine the power of the current study and a recommended sample size	161
5.9.2	Hand Calculations	162
5.9.2.1	Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	162
5.9.2.2	Using an R package or functions in base R, calculate the means and standard deviations for both levels of the dependent variable	162
5.9.2.3	Calulate the SE used in the denominator of the <i>t</i> -test	163
5.9.2.4	Calculate the independent samples <i>t</i> -test	163
5.9.2.5	Identify the degrees of freedom associated with your <i>t</i> -test	163
5.9.2.6	Locate the test critical value for your test	163
5.9.2.7	Is the <i>t</i> -test statistically significant? Why or why not?	163
5.9.2.8	Calculate the confidence interval around the difference in sample means	164
5.9.2.9	Calculate the effect size (i.e., Cohen's d associated with your <i>t</i> -test .	164
5.9.2.10	Assemble the results into a statistical string	164
6	Paired Samples <i>t</i>-test	165
6.1	Navigating this Lesson	165
6.1.1	Learning Objectives	165
6.1.2	Planning for Practice	166
6.1.3	Readings & Resources	166
6.1.4	Packages	166
6.2	Introducing the Paired Samples <i>t</i> -test	167
6.2.1	Workflow for Paired Samples <i>t</i> -test	168
6.3	Research Vignette	169
6.3.1	Simulating Data for the Paired Samples <i>t</i> -test	169
6.3.2	Quick Peek at the Data	172
6.4	Working the Paired Samples <i>t</i> -Test (by hand)	173
6.4.1	Stating the Hypothesis	173
6.4.2	Calculating the Paired Samples <i>t</i> -Test	173
6.4.2.1	Statistical Significance	174

6.4.2.2	Confidence Intervals	175
6.4.2.3	Effect Size	176
6.5	Working the Paired Samples <i>t</i> -Test with R Packages	177
6.5.1	Evaluating the Statistical Assumptions	177
6.5.1.1	Are the difference scores of the test variable normally distributed? .	178
6.5.1.2	APA style write-up of testing the assumptions	179
6.5.2	Computing the Paired Samples <i>t</i> -Test	179
6.6	APA Style Results	182
6.7	Power in Paired Samples <i>t</i> -Tests	185
6.8	Practice Problems	188
6.8.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed	188
6.8.2	Problem #2: Rework the research vignette, but change something about the simulation	188
6.8.3	Problem #3: Rework the research vignette, but swap one or more variables .	189
6.8.4	Problem #4: Use other data that is available to you	189
6.8.5	Grading Rubric	189
6.9	Homeworked Example	190
6.9.1	Working the Problem with R and R Packages	190
6.9.1.1	Narrate the research vignette, describing the variables and their role in the analysis	190
6.9.1.2	Simulate (or import) and format data	190
6.9.1.3	Evaluate statistical assumptions	191
6.9.1.4	Conduct a paired samples <i>t</i> -test (with an effect size & 95% CIs) .	194
6.9.1.5	APA style results with table(s) and figure	194
6.9.1.6	Conduct power analyses to determine the power of the current study and a recommended sample size	196
6.9.2	Hand Calculations	197
6.9.2.1	Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	197
6.9.2.2	Using an R package or functions in base R (and with data in the “wide” format), calculate the <i>difference</i> score between the two observations of the dependent variable	197
6.9.2.3	Obtain the mean and standard deviation of the <i>difference</i> score . .	197
6.9.2.4	Calculate the paired samples <i>t</i> -test	198

6.9.2.5	Identify the degrees of freedom associated with your paired samples <i>t</i> -test	198
6.9.2.6	Locate the test critical value for your paired samples <i>t</i> -test	198
6.9.2.7	Is the paired samples <i>t</i> -test statistically significant? Why or why not?	198
6.9.2.8	What is the confidence interval around the mean difference?	199
6.9.2.9	Calculate the effect size (i.e., Cohen's <i>d</i> associated with your paired samples <i>t</i> -test	199
6.9.2.10	Assemble the results into a statistical string.	199
7	One-way ANOVA	203
7.1	Navigating this Lesson	203
7.1.1	Learning Objectives	203
7.1.2	Planning for Practice	204
7.1.3	Readings & Resources	204
7.1.4	Packages	205
7.2	Workflow for One-Way ANOVA	205
7.3	Research Vignette	207
7.3.1	Data Simulation	207
7.3.2	Quick Peek at the Data	211
7.4	Working the Oneway ANOVA (by hand)	213
7.4.1	Sums of Squares Total	213
7.4.2	Sums of Squares for the Model (or Between)	218
7.4.3	Sums of Squares Residual (or within)	220
7.4.3.1	On the relationship between standard deviation and variance	221
7.4.4	Relationship between SS_T , SS_M , and SS_R	222
7.4.5	Mean Squares Model & Residual	222
7.4.6	Calculating the <i>F</i> Statistic	224
7.4.7	Source Table Games	224
7.5	Working the One-Way ANOVA with R Packages	225
7.5.1	Evaluating the Statistical Assumptions	227
7.5.1.1	Is the dependent variable normally distributed across levels of the factor?	227
7.5.1.2	Should we consider removing outliers?	231
7.5.1.3	Are the variances of the dependent variable similar across the levels of the grouping factor?	232

7.5.1.4	Summarizing results from the analysis of assumptions	232
7.5.2	Computing the Omnibus ANOVA	232
7.5.2.1	Effect size for the one-way ANOVA	234
7.5.3	Follow-up to the Omnibus F	235
7.5.3.1	Planning for the management of Type I Error	235
7.5.3.2	OPTION #1: Post hoc, pairwise, comparisons	236
7.5.3.3	OPTION #2: Non-orthogonal planned contrast	240
7.5.3.4	OPTION #3: Orthogonal planned contrasts	243
7.5.3.5	OPTION #4: Trend (polynomial) analysis	247
7.5.3.6	Which set of follow-up tests do we report?	249
7.5.4	What if we Violated the Homogeneity of Variance test?	250
7.6	APA Style Results	251
7.7	Power Analysis	253
7.8	A Conversation with Dr. Tran	255
7.9	Practice Problems	256
7.9.1	Problem #1: Play around with this simulation.	256
7.9.2	Problem #2: Conduct a one-way ANOVA with the <i>moreTalk</i> dependent variable.	256
7.9.3	Problem #3: Try something entirely new.	256
7.9.4	Grading Rubric	257
7.10	Homeworked Example	258
7.10.1	Working the Problem with R and R Packages	258
7.10.1.1	Narrate the research vignette, describing the IV and DV. The data you analyze should have at least 3 levels in the independent variable; at least one of the attempted problems should have a significant omnibus test so that follow-up is required).	258
7.10.1.2	Simulate (or import) and format data.	258
7.10.1.3	Evaluate statistical assumptions.	259
7.10.1.4	Conduct omnibus ANOVA (w effect size).	261
7.10.1.5	Conduct one set of follow-up tests; narrate your choice.	261
7.10.1.6	Describe approach for managing Type I error.	262
7.10.1.7	APA style results with table(s) and figure.	262
7.10.1.8	Conduct power analyses to determine the power of the current study and a recommended sample size.	264
7.10.2	Hand Calculations	265

7.10.2.1 Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses.	266
7.10.2.2 Calculate sums of squares total (SST). Steps in this calculation must include calculating a grand mean and creating variables representing the mean deviation and mean deviation squared.	266
7.10.2.3 Calculate the sums of squares for the model (SSM). A necessary step in this equation is to calculate group means.	267
7.10.2.4 Calculate the sums of squares residual (SSR). A necessary step in this equation is to calculate the variance for each group.	268
7.10.2.5 Calculate the mean square model, mean square residual, and <i>F</i> -test.	268
7.10.2.6 What are the degrees of freedom for your numerator and denominator?	269
7.10.2.7 Locate the test critical value for your one-way ANOVA.	269
7.10.2.8 Is the <i>F</i> -test statistically significant? Why or why not?	269
7.10.2.9 Calculate and interpret the η^2 effect size	269
7.10.2.10 Assemble the results into a statistical string.	269
8 Factorial (Between-Subjects) ANOVA	271
8.1 Navigating this Lesson	271
8.1.1 Learning Objectives	272
8.1.2 Planning for Practice	272
8.1.3 Readings & Resources	272
8.1.4 Packages	273
8.2 Introducing Factorial ANOVA	273
8.2.1 Workflow for Two-Way ANOVA	276
8.3 Research Vignette	276
8.3.1 Data Simulation	277
8.3.2 Quick peek at the data	278
8.4 Working the Factorial ANOVA (by hand)	283
8.4.1 Sums of Squares Total	283
8.4.2 Sums of Squares for the Model	285
8.4.3 Sums of Squares Residual (or within)	287
8.4.4 A Recap on the Relationship between SS_T , SS_M , and SS_R	288
8.4.5 Calculating SS for Each Factor and Their Products	288
8.4.5.1 Rater Main Effect	288
8.4.5.2 Photo Main Effect	289

8.4.5.3	Interaction effect	289
8.4.6	Source Table Games!	290
8.4.7	Interpreting the results	292
8.5	Working the Factorial ANOVA with R Packages	292
8.5.1	Evaluating the statistical assumptions	292
8.5.1.1	Is the dependent variable normally distributed?	295
8.5.1.2	Are the variances of the dependent variable similar across the levels of the grouping factors?	300
8.5.1.3	Summarizing results from the analysis of assumptions	300
8.5.2	Evaluating the Omnibus ANOVA	301
8.5.2.1	APA write-up of the omnibus results	303
8.5.3	Follow-up to a Significant Interaction Effect	303
8.5.3.1	Planning for the management of Type I Error	304
8.5.3.2	Option #1 the simple main effect of photo stimulus within ethnicity of the rater	304
8.5.3.3	Option #2 the simple main effect of ethnicity of rater within photo stimulus.	308
8.5.3.4	Options #3 through k	313
8.5.4	Investigating Main Effects	313
8.5.4.1	Option #1 post hoc paired comparisons	316
8.5.4.2	Option #2 planned orthogonal contrasts	319
8.5.4.3	Option #3 trend/polynomial analysis	323
8.6	APA Style Results	326
8.6.1	Comparing Our Results to Rhamdani et al. [2018]	330
8.7	Options for Violation of Statistical Assumptions	330
8.7.1	Violating the Assumption of Normality	331
8.7.2	Violating the Homogeneity of Variance Assumption	331
8.8	Power Analysis	333
8.8.1	Post Hoc Power Analysis	334
8.8.2	Estimating Sample Size Requirements	335
8.9	Practice Problems	336
8.9.1	Problem #1: Play around with this simulation.	337
8.9.2	Problem #2: Conduct a factorial ANOVA with the <i>positive evaluation</i> dependent variable.	337
8.9.3	Problem #3: Try something entirely new.	337
8.9.4	Grading Rubric	337

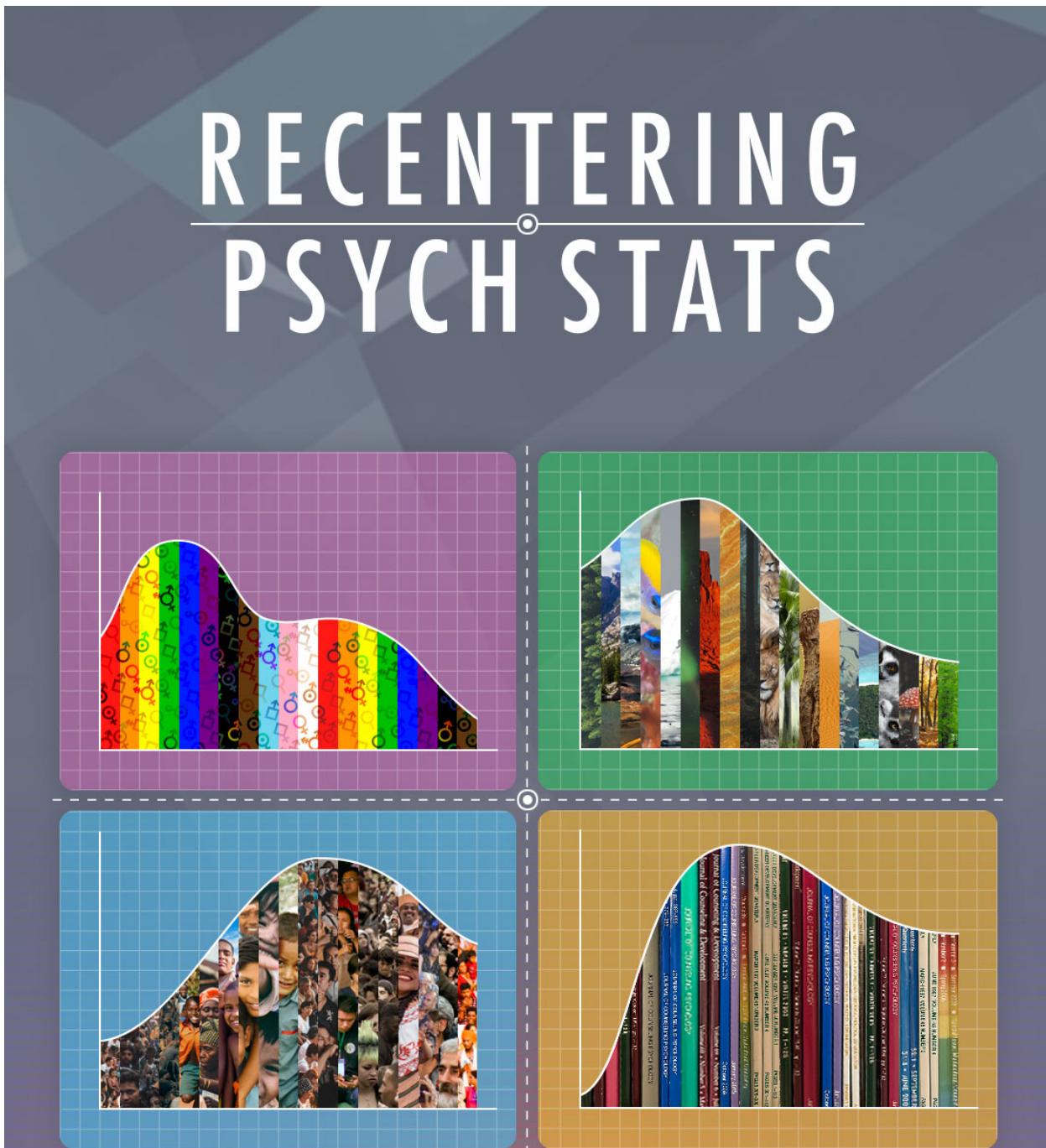
9 One-Way Repeated Measures ANOVA	339
9.1 Navigating this Lesson	339
9.1.1 Learning Objectives	339
9.1.2 Planning for Practice	340
9.1.3 Readings & Resources	340
9.1.4 Packages	341
9.2 Introducing One-way Repeated Measures ANOVA	341
9.2.1 Workflow for Oneway Repeated Measures ANOVA	342
9.3 Research Vignette	342
9.3.1 Data Simulation	344
9.3.2 Quick peek at the data	346
9.4 Working the One-Way Repeated Measures ANOVA (by hand)	348
9.4.1 Sums of Squares Total	348
9.4.2 Sums of Squares Within for Repeated Measures ANOVA	350
9.4.3 Sums of Squares Model – Effect of Time	350
9.4.4 Sums of Squares Residual	351
9.4.5 Sums of Squares Between	352
9.4.6 Mean Squares Model & Residual	353
9.4.7 <i>F</i> ratio	353
9.5 Working the One-Way Repeated Measures ANOVA with R packages	354
9.5.1 Testing the assumptions	354
9.5.1.1 Univariate assumptions for repeated measures ANOVA	354
9.5.1.2 Demonstrating sphericity	356
9.5.1.3 Is the data normally distributed?	356
9.5.1.4 Are there any outliers (and should we consider their removal)?	359
9.5.1.5 Summarizing results from the analysis of assumptions	361
9.5.1.6 Assumption of Sphericity	361
9.5.2 Computing the Test Statistic	361
9.5.3 Follow-up to Omnibus F	365
9.6 APA Style Results	366
9.6.1 Comparison with Amodeo et al.[2018]	367
9.7 Power Analysis	368
9.8 Practice Problems	370
9.8.1 Problem #1: Change the Random Seed	370

9.8.2 Problem #2: Increase N	370
9.8.3 Problem #3: Try Something Entirely New	370
9.8.4 Grading Rubric	371
10 Mixed Design ANOVA	373
10.1 Navigating this Lesson	373
10.1.1 Learning Objectives	373
10.1.2 Planning for Practice	374
10.1.3 Readings & Resources	374
10.1.4 Packages	375
10.2 Introducing Mixed Design ANOVA	375
10.2.1 Workflow for the Mixed Design ANOVA	376
10.3 Research Vignette	377
10.3.1 Data Simulation	378
10.3.2 Quick peek at the data	380
10.4 Working the Mixed Design ANOVA with R packages	385
10.4.1 Exploring data and testing assumptions	385
10.4.1.1 Is the dependent variable normally distributed?	387
10.4.1.2 Homogeneity of variance assumption	392
10.4.1.3 Assumption of homogeneity of covariance matrices	393
10.4.1.4 APA style writeup of assumptions	393
10.4.2 Omnibus ANOVA	394
10.4.2.1 Checking the sphericity assumption	395
10.4.2.2 Interpreting the omnibus results	396
10.4.3 Follow-up to Omnibus Tests	397
10.4.3.1 Planning for the management of Type I error	397
10.4.4 Simple main effect of condition within wave	397
10.4.5 Simple main effect of wave within condition	401
10.4.6 If we only had a main effect	406
10.4.7 APA Style Write-up of the Results	408
10.4.7.1 Results	408
10.4.7.2 Comparing our findings to Murrar and Brauer [2018]	410
10.5 Power in Mixed Design ANOVA	411
10.6 Practice Problems	412

10.6.1	Problem #1: Play around with this simulation.	413
10.6.2	Problem #2: Conduct a mixed design ANOVA with a different dependent variable.	413
10.6.3	Problem #3: Try something entirely new.	413
10.6.4	Grading Rubric	413
11	Analysis of Covariance	415
11.1	Navigating this Lesson	415
11.1.1	Learning Objectives	415
11.1.2	Planning for Practice	415
11.1.3	Readings & Resources	416
11.1.4	Packages	417
11.2	Introducing Analysis of Covariance (ANCOVA)	417
11.2.1	Workflow for ANCOVA	418
11.3	Research Vignette	420
11.3.1	Data Simulation	421
11.4	Working the ANCOVA – Scenario #1: Controlling for the pretest	422
11.4.1	Preparing the data	423
11.4.2	Evaluating the statistical assumptions	424
11.4.2.1	Linearity assumption	424
11.4.2.2	Homogeneity of regression slopes	427
11.4.2.3	Normality of residuals	427
11.4.2.4	Homogeneity of variances	428
11.4.2.5	Outliers	429
11.4.2.6	Summarizing results from the analysis of assumptions	429
11.4.3	Calculating the Omnibus ANOVA	431
11.4.4	Post-hoc pairwise comparisons (controlling for the covariate)	431
11.4.5	APA style results for Scenario 1	433
11.5	Working the ANCOVA – Scenario #2: Controlling for a confounding or covarying variable	436
11.5.1	Preparing the data	436
11.5.2	Evaluating the statistical assumptions	436
11.5.2.1	Linearity assumption	437
11.5.2.2	Homogeneity of regression slopes	438
11.5.2.3	Normality of residuals	439

11.5.2.4 Homogeneity of variances	440
11.5.2.5 Outliers	440
11.5.2.6 Summarizing the results from the analysis of assumptions	441
11.5.3 Calculating the Omnibus ANOVA	441
11.5.4 Post-hoc pairwise comparisons (controlling for the covariate)	443
11.5.5 APA style results for Scenario 2	444
11.6 More (and a recap) on covariates	447
11.7 Practice Problems	448
11.7.1 Problem #1: Play around with this simulation.	448
11.7.2 Problem #2: Conduct a one-way ANCOVA with the DV and covariate at post2.	448
11.7.3 Problem #3: Try something entirely new.	448
11.7.4 Grading Rubric	449
12 Type I Error	455
13 Examples for Follow-up to Factorial ANOVA	459
14 One-Way Repeated Measures with a Multivariate Approach	477
14.0.1 A Brief Commentary on Wrappers	482

BOOK COVER



LYNETTE H BIKOS, PHD, ABPP

- Formatted as an [html book](#) via GitHub Pages
- As a [PDF](#) available in the [docs](#) folder at the GitHub repository
- As an [ebook](#) available in the [docs](#) folder at the GitHub repository
- As a [Word document](#) available in the [docs](#) folder at the GitHub repository

All materials used in creating this OER are available at its [GitHub repo](#).

As a perpetually-in-progress, open education resource, feedback is always welcome. This IRB-approved (SPU IRB #202102010R, no expiration) [Qualtrics-hosted survey](#) includes formal rating scales, open-ended text boxes, and a portal for uploading attachments (e.g., marked up PDFs). You are welcome to complete only the portions that are relevant to you.

PREFACE

If you are viewing this document, you should know that this is a book-in-progress. Early drafts are released for the purpose teaching my classes and gaining formative feedback from a host of stakeholders. The document was last updated on 15 Jul 2023. Emerging volumes on other statistics are posted on the [ReCentering Psych Stats](#) page at my research team's website.

[Screencasted Lecture Link](#)

[YouTube Lecture Link](#)

To *center* a variable in regression means to set its value at zero and interpret all other values in relation to this reference point. Regarding race and gender, researchers often center male and White at zero. Further, it is typical that research vignettes in statistics textbooks are similarly seated in a White, Western (frequently U.S.), heteronormative, framework. The purpose of this project is to create a set of open educational resources (OER) appropriate for doctoral and post-doctoral training that contribute to a socially responsive pedagogy – that is, it contributes to justice, equity, diversity, and inclusion.

Statistics training in doctoral programs are frequently taught with fee-for-use programs (e.g., SPSS/AMOS, SAS, MPlus) that may not be readily available to the post-doctoral professional. In recent years, there has been an increase and improvement in R packages (e.g., *psych*, *lavaan*) used for analyses common to psychological research. Correspondingly, many graduate programs are transitioning to statistics training in R (free and open source). This is a challenge for post-doctoral psychologists who were trained with other software. This OER will offer statistics training with R and be freely available (specifically in a GitHub repository and posted through GitHub Pages) under a Creative Commons Attribution - Non Commercial - Share Alike license [CC BY-NC-SA 4.0].

Training models for doctoral programs in health service psychology are commonly scholar-practitioner, scientist-practitioner, or clinical-scientist. An emerging model, the *scientist-practitioner-advocacy* training model, incorporates social justice advocacy so that graduates are equipped to recognize and address the sociocultural context of oppression and unjust distribution of resources and opportunities [[Mallinckrodt et al., 2014](#)]. In statistics textbooks, the use of research vignettes engages the learner around a tangible scenario for identifying independent variables, dependent variables, covariates, and potential mechanisms of change. Many students recall examples in Field's [[2012](#)] popular statistics text: Viagra to teach one-way ANOVA, beer goggles for two-way ANOVA, and bushtucker for repeated measures. What if the research vignettes were more socially responsive?

In this OER, research vignettes will be from recently published articles where:

- the author's identity is from a group where scholarship is historically marginalized (e.g., BIPOC, LGBTQ+, LMIC[low-middle income countries]),
- the research is responsive to issues of justice, equity, inclusion, diversity,
- the lesson's statistic is used in the article, and
- there is sufficient information in the article to simulate the data for the chapter example(s) and practice problem(s); or it is publicly available.

In training for multicultural competence, the saying, “A fish doesn’t know that it’s wet” is often used to convey the notion that we are often unaware of our own cultural characteristics. In recent months and years, there has been an increased awakening to institutional and systemic factors that contribute to discrimination as a function of race, gender, nationality, class, and so forth. Queuing from the water metaphor, I am hopeful that a text that is recentered in the ways I have described can contribute to *changing the water* in higher education and in the profession of psychology.

Copyright with Open Access

This book is published under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. This means that this book can be reused, remixed, retained, revised and redistributed (including commercially) as long as appropriate credit is given to the authors. If you remix, or modify the original version of this open textbook, you must redistribute all versions of this open textbook under the same license: CC BY-SA 4.0.

A [GitHub open-source repository](#) contains all of the text and source code for the book, including data and images.

ACKNOWLEDGEMENTS

As a doctoral student at the University of Kansas (1992-1996), I learned that “a foreign language” was a graduation requirement. *Please note that as one who studies the intersections of global, vocational, and sustainable psychology, I regret that I do not have language skills beyond English.* This could have been met with credit from high school, but my rural, mid-Missouri high school did not offer such classes. This requirement would have typically been met with courses taken during an undergraduate program – but my non-teaching degree in the University of Missouri’s School of Education was exempt from this. The requirement could have also been met with a computer language (FORTRAN, C++) – but I did not have any of those either. There was a tiny footnote on my doctoral degree plan that indicated that a 2-credit course, “SPSS for Windows” would substitute for the language requirement. Given that it was taught by my one of my favorite professors, I readily signed up. As it turns out, Samuel B. Green, PhD, was using the course to draft chapters in the textbook [[Green and Salkind, 2017c](#)] that has been so helpful for so many. Unfortunately, Drs. Green (1947 - 2018) and Salkind (1947 - 2017) are no longer with us. I have worn out numerous versions of their text. Another favorite text of mine has been Dr. Barbara Byrne’s [[2016](#)], “Structural Equation Modeling with AMOS.” I loved the way she worked through each problem and paired it with a published journal article, so that the user could see how the statistical evaluation fit within the larger project/article. I took my tea-stained text with me to a workshop she taught at APA and was proud of the signature she added to it. Dr. Byrne created SEM texts for a number of statistical programs (e.g., LISREL, EQS, MPlus). As I was learning R, I wrote Dr. Byrne, asking if she had an edition teaching SEM/CFA with R. She promptly wrote back, saying that she did not have the bandwidth to learn a new statistics package. We lost Dr. Byrne in December 2020. I am so grateful to these role models for their contributions to my statistical training. I am also grateful for the doctoral students who have taken my courses and are continuing to provide input for how to improve the materials.

The inspiration for training materials that re*center statistics and research methods came from the [Academics for Black Survival and Wellness Initiative](#). This project, co-founded by Della V. Mosley, Ph.D., and Pearis L. Bellamy, M.S., made clear the necessity and urgency for change in higher education and the profession of psychology.

At very practical levels, I am indebted to SPU’s Library, and more specifically, SPU’s Education, Technology, and Media Department. Assistant Dean for Instructional Design and Emerging Technologies, R. John Robertson, MSc, MCS, has offered unlimited consultation, support, and connection. Senior Instructional Designer in Graphics & Illustrations, Dominic Wilkinson, designed the logo and bookcover. Psychology and Scholarly Communications Librarian, Kristin Hoffman, MLIS, has provided consultation on topics ranging from OERS to citations. I am also indebted to Associate Vice President, Teaching and Learning at Kwantlen Polytechnic University, Rajiv Jhangiani, PhD. Dr. Jhangiani’s text [[2019](#)] was the first OER I ever used and I was grateful for

his encouraging conversation.

Financial support for this project has been provided the following:

- *Call to Action on Equity, Inclusion, Diversity, Justice, and Social Responsivity Request for Proposals* grant from the Association of Psychology Postdoctoral and Internship Centers (2021-2022).
- *Diversity Seed Grant*, Office of Inclusive Excellence and Advisory Council for Diversity and Reconciliation (ACDR), Seattle Pacific University.
- *ETM Open Textbook & OER Development Funding*, Office of Education, Technology, & Media, Seattle Pacific University.

```
{r include=FALSE} # automatically create a
bib database for R packages
knitr:::write_bib(c(
  .packages(),
  'bookdown',
  'knitr',
  'rmarkdown'
), 'packages.bib')
```

Chapter 1

Introduction

[Screencasted Lecture Link](#)

1.1 What to expect in each chapter

This textbook is intended as *applied*, in that a primary goal is to help the scientist-practitioner-advocate use a variety of statistics in research problems and *writing them up* for a program evaluation, dissertation, or journal article. In support of that goal, I try to provide just enough conceptual information so that the researcher can select the appropriate statistic (i.e., distinguishing between when ANOVA is appropriate and when regression is appropriate) and assign variables to their proper role (e.g., covariate, moderator, mediator).

This conceptual approach does include occasional, step-by-step, *hand-calculations* (using R to do the math for us) to provide a *visceral feeling* of what is happening within the statistical algorithm that may be invisible to the researcher. Additionally, the conceptual review includes a review of the assumptions about the characteristics of the data and research design that are required for the statistic.

Statistics can be daunting, so I have worked hard to establish a *workflow* through each analysis. When possible, I include a flowchart that is referenced frequently in each chapter and assists the researcher keep track of their place in the many steps and choices that accompany even the simplest of analyses.

As with many statistics texts, each chapter includes a *research vignette*. Somewhat unique to this resource is that the vignettes are selected from recently published articles. Each vignette is chosen with the intent to meet as many of the following criteria as possible:

- the statistic that is the focus of the chapter was properly used in the article,
- the author's identity is from a group where scholarship is historically marginalized (e.g., BIPOC, LGBTQ+, LMIC [low middle income countries]),
- the research has a justice, equity, inclusion, diversity, and social responsibility focus and will contribute positively to a social justice pedagogy, and
- there is sufficient information in the article to simulate the data for the chapter example(s) and practice problem(s); or the data is available in a repository.

In each chapter we employ *R* packages that will efficiently calculate the statistic and the dashboard of metrics (e.g., effect sizes, confidence intervals) that are typically reported in psychological science.

1.2 Strategies for Accessing and Using this OER

There are a number of ways you can access this resource. You may wish to try several strategies and then select which works best for you. I demonstrate these in the screencast that accompanies this chapter.

1. Simply follow along in your preferred format of the book (html, PDF, or ebook) and then
 - open a fresh .rmd file of your own, copying (or retyping) the script and running it
2. Locate the original documents at the [GitHub repository](#). You can
 - open them to simply take note of the “behind the scenes” script
 - copy/download individual documents that are of interest to you
 - clone a copy of the entire project to your own GitHub site and further download it (in its entirety) to your personal workspace. The [GitHub Desktop app](#) makes this easy!
3. Listen to the accompanying lectures (I think sound best when the speed is 1.75). The lectures are being recorded in Panopto and should include the closed captioning.
4. Each time the book is updated, new .docx (Microsoft Word), PDF (Adobe Acrobat), and ebook(EPUB File) versions are also createdt. You can access these in the “docs” folder at the [GitHub repository](#).
5. Provide feedback to me! If you fork a copy to your own GitHub repository, you can
 - open up an editing tool and mark up the document with your edits,
 - start a discussion by leaving comments/questions, and then
 - sending them back to me by committing and saving. I get an e-mail notiying me of this action. I can then review (accepting or rejecting) them and, if a discussion is appropriate, reply back to you.
 - I am also seeking peer-review feedback at this [Qualtrics-hosted survey](#). You are welcome to complete only the portions that are relevant to you.

1.3 If You are New to R

R can be oveRwhelming. Jumping right into advanced statistics might not be the easiest way to start. The [Ready_Set_Rlesson](#) of this volume provides an introduction and the [waRming uplesson](#) walks through simple data preparation and descriptive statistics.

In the remaining lessons, I have attempted to provide complete code for every step of the process, starting with uploading the data. To help explain what R script is doing, I sometimes write it in the chapter text; sometimes leave hashtags-commments in the chunks; and, particularly in the accompanying screencasted lectures, try to take time to narrate what the R script is doing.

I've found that, somewhere on the internet, there's almost always a solution to what I'm trying to do. I am frequently stuck and stumped and have spent hours searching the internet for even the tiniest of tasks. When you watch my videos, you may notice that in my R studio, there is a "scRiptuRe" file. I take notes on the solutions and scripts here – using keywords that are meaningful to me so that when I need to repeat the task, I can hopefully search my own prior solutions and find a fix or a hint. You may also find it useful to create a working document of your own tips and tricks.

1.4 Introduction to the Data Set Used for Homeworked Examples

Screencast Link

This section of the Appendix is designed as a streamlined example of working the primary statistic from each lesson. This section is intended to be helpful in two ways:

- The worked example focuses on the operations and interpretations and more closely mimics “how you would work a problem in real life.”
- The grading rubric from the end of each lesson serves as the outline for the process.
- This dataset could be used for the practice problems. For homework that you submit for grading, please choose *different variables* than the ones worked in the examples.

1.4.1 The Data Set

The dataset used in the “homeworked” examples is from my own research. Along long with the pre-registration and codebooks, it is publicly available on the Open Science Framework (OSF): <https://osf.io/z84kh/> I have also provided a copy of it in the GitHub repository that hosts the ReCentering Psych Stats OER.

This data is from an IRB-approved study. The informed consent of the IRB specified that the data could be used in research as well as in teaching demonstrations and would be made available to the general public. You may notice there are student- and teacher- IDs. These numbers are *not** the institution’s identification numbers. Rather, they have been further anonymized.

The purpose of the research project was to evaluate efforts to recenter – in a socially responsive way – courses in the statistics and research methods sequence in scientist-practitioner psychology (PhD) programs. The recentering occurred in two phases: (a) a transition from SPSS to R and (b) an explicit recentering. Data were from end-of-course evaluations three courses I taught: Research Methods I: Analysis of Variance [ANOVA], Research Methods III: Multivariate Modeling [multivariate], and Research Methods IV: Psychometrics/Theory of Test Construction [psychometrics]) that were offered 2017 through 2022.

Because students could contribute up to three course evaluations, each, multilevel modeling was used for the primary analyses. The nature of the data, though, allows me to demonstrate all of the statistics utilized the OER with this data. For each analysis, I have tried to derive a sensible question that *could be* answered by the data. In-so-doing, I try to point out when the alignment of research question and statistic is less than ideal.

The data file is titled *ReC.rds* and can be retrieved with this code:

The following can serve as a codebook:

Variable	Definition or Description	Scaling
deID	Anonymized identification for each student. Each student could contribute up to three course evaluations.	Nominal/factor
CourseID	Unique number for each course taught (i.e., ANOVA has unique numbers across department and year).	Nominal/factor
Dept	CPY (Clinical Psychology), ORG (Industrial Organizational Psychology)	Nominal/factor
Course	ANOVA (analysis of variance), Multivariate (multivariate modeling), Psychometrics (psychometrics/theory of test construction), taught in that order	Nominal/factor
StatsPkg	SPSS, R	Nominal/factor
Centering	Pre (before explicit recentering), Re (included explicit recentering)	Nominal/factor
Year	Calendar year in which the course was taught	Calendrical time
Quarter	Academic term in which course was taught (fall, winter, spring)	Nominal/factor
ProgramYear	A potential confound to the study. During the changes from SPSS to R and the explicit recentering, the program was also moving the stats sequence from the second to the first year of the doctoral program. First = course taken during first year; Second = course taken during second year; Transition = course taken during the transition period.	Nominal/factor
SPFC.Decolonize	Students were given the opportunity to exclude their data from analysis. Such data was removed prior to any analysis and not included in this set.	Character

COURSE EVALUATION ITEMS; 5-point Likert scaling from 1(*strongly disagree*) to 5(*strongly agree*). Higher scores are more favorable evaluations.

Variable	Complete Item
IncrInterest	My interest in the subject matter increased over the span of the course.
IncrUnderstanding	My understanding of the subject matter increased over the span of the course.
ValObjectives	This course has objectives that are valuable in furthering my education.
ApprAssignments	This course has had an appropriate workload, given the course objectives.
EffectiveAnswers	The instructor has effectively answered student questions.
Respectful	The instructor has shown respect to students.
ClearResponsibilities	The instructor has made student responsibilities clear.
Feedback	The instructor has provided feedback to me about my learning progress.
OvInstructor	My overall rating of this instructor for this course is:

Variable	Complete Item
MultPerspectives	The instructor has helped students consider issues from multiple perspectives, where applicable.
OvCourse	My overall rating of the course content is:
InclsrvClssrm	The instructor has been intentional in fostering an inclusive classroom for students with diverse backgrounds and abilities.
DEIntegration	The instructor has, when appropriate, discussed the relationships between race/ethnicity/culture and course content.
ClearPresentation	The instructor has presented course material clearly.
ApprWorkload	This course has had an appropriate workload, given the course objectives.
MyContribution	My overall rating of my contribution in this course is:
InspiredInterest	The instructor has inspired student interest in the subject matter of this course.
Faith	The instructor has, when appropriate, discussed the relationship between the Christian faith and course content.
EquitableEval	The instructor used methods of evaluating student course work that were equitable.
ClearOrganization	This course has had a clear overall organization.
RegPrepare	I regularly read, reviewed, visited/logged on, or completed assigned readings and tasks.
EffectiveLearning	This course has consisted of course activities/tasks that were effective in helping me learn (e.g., discussions, readings, assignments, labs, or other activities).
AccessibleInstructor	The instructor has been accessible (e.g., discussion sessions, virtual office hours, phone, chat, email, online forum or conference, etc.).

From these variables, I created three scales to assess valued by the student (Valued), traditional pedagogy (TradPed), and socially responsive pedagogy (SRped). I will use these in the demonstrations.

- **Valued by the student** includes the items: ValObjectives, IncrUnderstanding, IncrInterest
- **Traditional pedagogy** includes the items: ClearResponsibilities, EffectiveAnswers, Feedback, ClearOrganization, ClearPresentation
- **Socially responsive pedagogy** includes the items: InclusvClassrm, EquitableEval, MultPerspectives, DEIntegration

In the examples where the scale scores are used, I provide code for calculating the means.

Here's how to import the data:

Chapter 2

Ready_Set_R

[Screencasted Lecture Link](#)

With the goal of creating a common, system-wide approach to using the platform, this lesson was originally created for Clinical and Industrial-Organizational doctoral students who are entering the “stats sequence.” I hope it will be useful for others (e.g., faculty, post-doctoral researchers, and practitioners) who are also making the transition to R.

2.1 Navigating this Lesson

There is about 45 minutes of lecture.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

2.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Downloading/installing R’s parts and pieces.
- Using R-Markdown as the interface for running R analyses and saving the script.
- Recognizing and adopting best practices for “R hygiene.”
- Identifying effective strategies for troubleshooting R hiccups.

2.2 downloading and installing R

2.2.1 So many paRts and pieces

Before we download R, it may be helpful to review some of R’s many parts and pieces.

The base software is free and is available [here](#)

Because R is already on my machine (and because the instructions are sufficient), I will not walk through the demo, but I will point out a few things.

- The “cran” (I think “cranium”) is the *Comprehensive R Archive Network*. In order for R to run on your computer, you have to choose a location – and it should be geographically “close to you.”
 - Follow the instructions for your operating system (Mac, Windows, Linux)
 - You will see the results of this download on your desktop (or elsewhere if you chose to not have it appear there) but you won’t ever use R through this platform.
- **R Studio** is the way in which we operate R. It’s a separate download. Choose the free, desktop, option that is appropriate for your operating system:
- *R Markdown* is the way that many analysts write *script*, conduct analyses, and even write up results. These are saved as .rmd files.
 - In R Studio, open an R Markdown document through File/New File/R Markdown
 - Specify the details of your document (title, author, desired output)
 - In a separate step, SAVE this document [File/Save] into a NEW FILE FOLDER that will contain anything else you need for your project (e.g., the data).
 - *Packages* are at the heart of working in R. Installing and activating packages require writing script.

Note If you are working on an *enterprise-owned machine* (e.g., in my specific context, if you are a faculty/staff or have a lab with institution-issued laptops) there can be complications caused by how documents are stored. In recent years we have found that letting the computer choose where to load base R, R Studio, and the packages generally works. The trick is to save R projects (i.e., folder with .rmd files and data) into the OneDrive folder that syncs to your computer. If you have difficulty knitting that is unrelated to code/script (which you can evaluate by having a classmate or colleague successfully knit on their machine), it is likely because you have saved the files to the local hard drive and not OneDrive. If you continue to have problems I recommend consulting with your computer and technology support office.

2.2.2 oRienting to R Studio (focusing only on the things we will be using first and most often)

R Studio is organized around four panes. These can be re-sized and rearranged to suit your personal preferences.

- Upper right window
 - Environment: lists the *objects* that are available to you (e.g., dataframes)
- Lower right window
 - *Files*: Displays the file structure in your computer’s environment. Make it a practice to (a) organize your work in small folders and (b) navigate to that small folder that is holding your project when you are working on it.

- *Packages*: Lists the packages that have been installed. If you navigate to a specific package, you will know if it is “on” because its box is checked. You can also access information about the package (e.g., available functions, examples of script used with the package) in this menu. This information opens in the “Help” window.
 - The *Viewer* and *Plots* tabs will be useful, later, in some advanced statistics when we can simultaneously examine output and script in windows that are side-by-side.
- Upper left window
 - If you are using R Markdown, that file lives here and is composed of open space and chunks.
 - Lower left window
 - R Studio runs in the Console (the background). Very occasionally, I can find useful troubleshooting information here.
 - More commonly, I open my R Markdown document so that it takes up the whole screen.

2.3 best pRactices

Many initial problems in R can be solved with good R hygiene. Here are some suggestions for basic practices. It can be tempting to “skip this.” However, in the first few weeks of class, these are the solutions I am presenting (and repeating, ad nauseum) to my students.

2.3.1 Everything is documented in the .rmd file

Although others do it differently, I put *everything* in my .rmd file. That is, my R script includes code for importing data and opening packages. Additionally, I make notes about the choices I am making. Relatedly, I keep a “bug log” – noting what worked and what did not work. I will also begin my APA style results section directly in the .rmd file.

Why do I do all this? Because when I return to my project hours or years later, I have a permanent record of very critical things like (a) where my data is located, (b) what version I was using, and (c) what package was associated with the functions.

2.3.2 Setting up the file

File organization is a critical key to success. In your computing environment:

- Create a project file folder.
- Put the data file in it.
- Open an R Markdown file.
- Save it in the same file folder as the data.
- When your data and .rmd files are in the same folder (not your desktop, but a specific folder) the data can be pulled into the .rmd file without creating a working directory.

2.3.3 Script in chunks and everything else in the “inline text” sections

The R Markdown document is an incredible tool for integrating text, tables, and analyses. This entire OER is written in R Markdown. A central feature of this is “chunks.”

The only thing in the chunks should be script for running R. You can also hashtag comments so they won’t run (but you can also write anything you want between chunks without using hashtags).

Syntax for simple formatting in the text (i.e., non-chunk) areas (e.g., using italics, making headings, bold) is found here: https://rmarkdown.rstudio.com/authoring_basics.html

“Chunks” start and end with three tic marks and will show up in a shaded box. Chunks have three symbols in their upper right. Those controls will disappear (and your script will not run) if you have replaced them with double or single quotation marks or one or more of the tics are missing.

The easiest way to insert a chunk is to use the INSERT/R command at the top of this editor box. You can also insert a chunk with the keyboard shortcut: CTRL/ALT/i

```
# hashtags let me write comments to remind myself what I did here I
# am simply demonstrating arithmetic (but I would normally be running
# code)
2021 - 1966
```

[1] 55

2.3.4 Managing packages

As scientist-practitioners (and not coders), we will rely on *packages* to do much of the work. At first you may feel overwhelmed about the large number of packages that are available. Soon, though, you will become accustomed to the ones most applicable to our work (e.g., psych, tidyverse, rstatix, apaTables).

Researchers treat packages differently. In these lectures, I list all the packages we will use in an opening chunk at the beginning of the lecture. When the hashtags are removed, the script will ask R to check to see if the package is installed on your machine. If it is, installation is skipped. If it is not, R installs it. Simply remove the hashtag to run the code the first time, then hashtag them back out so R is not always re-checking.

```
# will install the package if not already installed
# if(!require(psych)){install.packages('psych')}
```

To make a package operable, you need to open it. There are two primary ways to do this. The first is to use the library function.

```
#install.packages ("psych")
library (psych)
```

The second way is to place a double colon between the package and function. This second method has become my preferred practice because it helps me remember what package goes with each function. It can also prevent R hiccups when there are identical function names and R does not know which package to use. Below is an example where I might ask for descriptives from the psych package. Because I have not yet uploaded data, I have hashtags it out, making the command inoperable.

```
#psych::describe(mydata)
```

There are exceptions. One is the *tidyverse* package. Some of my script uses pipes (%>%) and pipes require *tidyverse* to be activated. This is why you will often see me call the *tidyverse* package with the *library()* function (as demonstrated above.)

2.3.5 Upload the data

When imported (or simulated) properly, data will appear as an object in the global environment.

In the context of this OER, I will be simulating data in each lesson for immediate use in the lesson. This makes this web-based OER more *portable*. This also means that when working the problems in the chapter we do not need to (a) write the data to a file or (b) import data from files. Because these are essential skills, I will demonstrate this process here – starting with simulating data.

At this point, simulating data is beyond the learning goals I have established for the chapter. I do need to include the code so that we get some data. The data I am simulating is used in the [one-way ANOVA lesson](#). The data is from the Tran and Lee [2014] random clinical trial.

In this simulation, I am simply creating an ID number, a condition (High, Low, Control), and a score on the dependent variable, “Accurate.” More information about this study is included in the [one-way ANOVA chapter](#).

```
# Note, this simulation results in a different dataset than is in the
# OnewayANOVA lesson sets a random seed so that we get the same
# results each time
set.seed(2021)
# sample size, M and SD for each group
Accurate <- c(rnorm(30, mean = 1.18, sd = 0.8), rnorm(30, mean = 1.83,
sd = 0.58), rnorm(30, mean = 1.76, sd = 0.56))
# set upper bound for DV
Accurate[Accurate > 3] <- 3
# set lower bound for DV
Accurate[Accurate < 0] <- 0
# IDs for participants
ID <- factor(seq(1, 90))
# name factors and identify how many in each group; should be in same
# order as first row of script
COND <- c(rep("High", 30), rep("Low", 30), rep("Control", 30))
# groups the 3 variables into a single df: ID, DV, condition
Acc_sim30 <- data.frame(ID, COND, Accurate)
```

At this point, this data lives only in this .rmd file after the above code is run. Although there are numerous ways to export and import data, I have a preference for two.

2.3.5.1 To and from .csv files

The first is to write the data to a .csv file. In your computer's environment (outside of R), these files are easily manipulated in Excel. I think of them as being "Excel lite" because although Excel can operate them, they lack some of the more advanced features of an Excel spreadsheet.

In the code below, I identify the R object "Acc_sim30" and give it a file name, "to_CSV.csv". This file name must have the .csv extension. I also indicate that it should preserve the column names (but ignore row names; since we don't have row names).

This file will save in the same folder as wherever you are using this .rmd file.

```
# to write it to an outfile as a .csv
write.table(Acc_sim30, file = "to_CSV.csv", sep = ",", col.names = TRUE,
            row.names = FALSE)
```

Importing this object back into the R environment can be accomplished with some simple code. For the sake of demonstration,

```
# to save the df as an .csv (think 'Excel lite') file on your
# computer; it should save in the same file as the .rmd file you are
# working with
from_CSV <- read.csv("to_CSV.csv", header = TRUE)
```

The advantage of working with .csv files is that it is then easy to inspect and manipulate them outside of the R environment. The disadvantage of .csv files is that each time they are imported they lose any formatting you may have meticulously assigned to them.

2.3.5.2 To and from .rds files

While it is easy enough to rerun the code (or copy it from data prep .rmd and paste it into an .rmd you are using for advanced analysis), there is a better way! Saving the data as an R object preserves all of its characteristics.

```
# to save the df as an .rds file on your computer; it should save in
# the same file as the .rmd file you are working with
saveRDS(Acc_sim30, "to_Robject.rds")
```

This file will save to your computer (and you can send it to colleagues). However, it is not easy to "just open it" in Excel. To open an .rds file and use it (whether you created it or it is sent to you by a colleague), use the following code:

```
from_rds <- readRDS("to_Robject.rds")
```

If you are the recipient of an R object, but want to view it as a .csv, simply import the .rds then use the above code to export it as a .csv.

2.3.5.3 From SPSS files

Your data may come to you in a variety of ways. One of the most common is SPSS. The *foreign* package is popular for importing SPSS data. Below is code which would import an SPSS file *if I had created one*. You'll see that this script is hashtags out because I rarely use SPSS and do not have a handy file to demo.

```
# opening an SPSS file requires the foreign package which I opened
# earlier from_SPSS <- foreign::read.spss ('SPSSdata.sav',
# use.value.labels = TRUE, to.data.frame = TRUE)
```

2.4 quick demonstRation

Let's run some simple descriptives. In the script below, I am using the *psych* package. Descriptive statistics will appear for all the data in the dataframe and the output will be rounded to three spaces. Note that rather than opening the psych package with the library function, I have used the double colon convention.

```
round(psych::describe(Acc_sim30), 3)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
ID*	1	90	45.50	26.12	45.50	45.50	33.36	1	90	89	0.00	-1.24
COND*	2	90	2.00	0.82	2.00	2.00	1.48	1	3	2	0.00	-1.53
Accurate	3	90	1.52	0.68	1.55	1.54	0.70	0	3	3	-0.19	-0.34
			se									
ID*			2.75									
COND*			0.09									
Accurate			0.07									

Because "ID" is the case ID and COND is the factor (high, low, control), the only variable for which this data is sensible is "Accurate." Nonetheless, this provides an example of how to apply a package's function to a dataset. As we progress through the text we will learn how to manage the data so that we get the specific output we are seeking.

2.5 the knitted file

One of the coolest things about R Markdown is its capacity to *knit* to HTML, PPT, or WORD.

- In this OER, I am writing the lessons in R markdown (.rmd files), with the package *bookdown* as a helper, and knitting the files to .html, .doc, .pdf, and .epub formats.
- The package *papaja* is designed to prepare APA manuscripts where the writing, statistics, and references are all accomplished in a single file. This process contributes to replicability and reproducibility.
- More detailed instructions for knitting to these formats are provided in the [extRas](#) mini-volume of [ReCentering Psych Stats](#).

2.6 tRoubleshooting in R maRkdown

Hiccups are normal. Here are some ideas that I have found useful in getting unstuck.

- In a given set of operations, you must run/execute each piece of code in order: every, single, time. That is, all the packages have to be in your library and activated.
 - If you open an .rmd file, you cannot just scroll down to make a boxplot. You need to run any *prerequisite* script (like loading files, putting the data in the global environment, etc.)
 - Lost? Clear your global environment (broom icon in the upper right) and start over. Fresh starts are good.
- Your .rmd file and your data need to be stored in the same file folder. Make unique folders for each project (even if each contains only a few files).
- If you have tried what seems apparent to you and cannot solve your challenge, do not wait long before typing warnings into a search engine. Odds are, you'll get some useful hints in a manner of seconds. Especially at first, these are common errors:
 - The package isn't loaded.
 - The .rmd file hasn't been saved yet, or isn't saved in the same folder as the data.
 - There are errors in punctuation or spelling.
- Restart R (it's quick – not like restarting your computer). I frequently restart and clear my output and environment so that I can better track my order of operations.
- If you receive an error indicating that a function isn't working or recognized, and you have loaded the package, type the name of the package in front of the function with two colons (e.g., psych::describe(df)). If multiple packages are loaded with functions that have the same name, R can get confused.

2.7 just *why* have we tRansitioned to R?

- It is (or at least it appears to be) the futuRe.
- SPSS individual and site licenses are increasingly expensive and limited; that is, Mplus, AMOS, HLM, or R tools may also be needed. As package development for R is exploding, we have tools to “do just about anything.”
- Most graduate psychology programs are scientist/practitioner in nature and include training in “high end” statistics. Yet, many of your employing organizations will not have SPSS. R is a free, universally accessible program, that our graduates can use anywhere.

2.8 stRategies for success

- Engage with R, but don't let it overwhelm you.
 - The *mechanical is also the conceptual*. Especially while it's *simpler*, do try to retype the script into your own .rmd file and run it. Track down the errors you are making and fix them.
 - If this stresses you out, move to simply copying the code into the .rmd file and running it. If you continue to have errors, you may have violated one of the best practices above (ask, "Is the package activated?" "Are the data and .rmd files in the same place?" "Is all the prerequisite script run?").
 - Still overwhelmed? Keep moving forward by (retrieving the original .rmd file from the GitHub repository) opening a copy of the .rmd file and just "run it along" with the lecture. Spend your mental power trying to understand what each piece does so you can translate it for any homework assignments. My suggestions for practice are intended to be parallel to the lecture with no sneaky trix.
- Copy script that works elsewhere and replace it with your datafile, variables, and so forth.
- The learning curve is steep, but not impossible. Gladwell [2008] taught us that it takes about 10,000 hours to get great at something (2,000 to get reasonably competent). Practice. Practice. Practice.
- Updates to R, R Studio, and the packages are necessary, but can also be problematic. Sometimes updates cause programs/script to fail (e.g., "X has been deprecated for version X.XX"). My personal practice is to update R, R Studio, and the packages a week or two before each academic term. I expect that
 - prior scripts may need to be updated or revised with package updates, and
 - there will be incongruencies between base R, R Studio, and the packages.
- Embrace your downward dog. And square breathing. Also, walk away, then come back.

2.9 Resources for getting staRted

R for Data Science: <https://r4ds.had.co.nz/>

R Cookbook: <http://shop.oreilly.com/product/9780596809164.do>

R Markdown homepage with tutorials: <https://rmarkdown.rstudio.com/index.html>

R has cheatsheets for everything, here's the one for R Markdown: <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>

R Markdown Reference guide: <https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>

Using R Markdown for writing reproducible scientific papers: <https://libscie.github.io/rmarkdown-workshop/handout.html>

Script for all of Field's text: <https://studysites.uk.sagepub.com/dsur/study/scriptfi.htm>

LaTeX equation editor: <https://www.codecogs.com/latex/eqneditor.php>

2.10 Practice Problems

The suggestions for practice in this lesson are foundational for starting work in R. If you struggle with any of these steps, I encourage you to get consultation from a peer, instructor, or a tutor.

Assignment Component	Points Possible	Points Earned
1. Download base R and R Studio	5	_____
2. Open and save an .rmd (R Markdown) file in a “sensible location” on your computer	5	_____
3. In the .rmd file, open a chunk and perform a simple mathematical operation of your choice (e.g., subtract your birth year from this year)	5	_____
4. Install at least three packages; we will commonly use <i>psych</i> , <i>tidyverse</i> , <i>dplyr</i> , <i>knitr</i> , <i>ggplot2</i> , <i>ggnpubr</i>)	5	_____
5. Copy the simulation in this lesson to your .rmd file. Change the random seed and run the simulation. Save the resulting data as a .csv or .rds file <i>in the same file as you saved the .rmd file</i> .	5	_____
6. Clear your environment (broom in upper right). Open the simulated file that you saved.	5	_____
7. Run the <i>describe()</i> function from the <i>psych</i> package with your simulated data that you imported from your local drive.	5	_____
8. Demonstration/discussion with a grader.	5	_____
Totals	40	_____

2.11 Homeworked Example

Screencast Link

Several elements of the practice problems (i.e., download base R and R studio, open and save an .rmd file) are not easily demonstrated and not replicated here. These are skipped.

If you wanted to use this example and dataset as a basis for a homework assignment, you could simply change the seed – again. For a greater challenge, you could adjust the simulation to have different sample sizes, means, or standard deviations.

2.11.1 Perform a simple mathematical operation:

In the .rmd file, open a chunk and perform a simple mathematical operation of your choice (e.g., subtract your birth year from this year).

2023 – 1966

[1] 57

2.11.2 Install at least three packages we will commonly use:

Below is code for installing three packages. Because continuous reinstallation can be problematic, I have hashtagged them so that they will not re-run.

```
#install.packages("tidyverse")
#install.packages("ggpubr")
#install.packages("psych")
```

2.11.3 Copy the simulation in this lesson to your .rmd file. Change the random seed and run the simulation.

```
set.seed(2023)
# sample size, M and SD for each group
Accurate <- c(rnorm(30, mean = 1.18, sd = 0.8), rnorm(30, mean = 1.83,
                 sd = 0.58), rnorm(30, mean = 1.76, sd = 0.56))
# set upper bound for DV
Accurate[Accurate > 3] <- 3
# set lower bound for DV
Accurate[Accurate < 0] <- 0
# IDs for participants
ID <- factor(seq(1, 90))
# name factors and identify how many in each group; should be in same
# order as first row of script
COND <- c(rep("High", 30), rep("Low", 30), rep("Control", 30))
# groups the 3 variables into a single df: ID, DV, condition
Acc_sim30B <- data.frame(ID, COND, Accurate)
```

2.11.4 Save the resulting data as a .csv or .rds file in the same file as you saved the .rmd file.

You only need to save it as a .csv or .rds file. I have demonstrated both.

Saving as a .csv file

```
write.table(Acc_sim30B, file = "to_CSVb.csv", sep = ",", col.names = TRUE,
            row.names = FALSE)
```

Saving as an .rds file

```
saveRDS(Acc_sim30B, "to_RobjectB.rds")
```

2.11.5 Clear your environment (broom in upper right).

You only need to import the .csv or .rds file; I have demonstrated both. Open the .csv file from my local drive.

```
from_CSV <- read.csv("to_CSVb.csv", header = TRUE)
```

Open the .rds file from my local drive.

```
from_rds <- readRDS("to_RobjectB.rds")
```

2.11.6 Run the `describe()` function from the `psych` package with your simulated data that you imported from your local drive.

You only need to retrieve descriptives from the .csv or .rds file; I have demonstrated both.

```
psych::describe(from_CSV)
```

```
psych::describe(from_rds)
```

Chapter 3

Preliminary Analyses

[Screencasted Lecture Link](#)

The beginning of any data analysis means familiarizing yourself with the data. Among other things, this includes producing and interpreting its distributional characteristics. In this lesson we mix common R operations for formatting, preparing, and analyzing the data with foundational statistical concepts in statistics.

3.1 Navigating this Lesson

There is just less than two hours of lecture. If you work through the lesson with me, I would plan for an additional three hours.

While the majority of R objects and data you will need are created within the R script that sources the lesson, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's introduction

3.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Determine the appropriate scale of measurement for variables and format them properly in R
- Produce and interpret measures of central tendency
- Analyze the distributional characteristics of data
- Describe the steps in calculating a standard deviation.
- Describe the steps in calculating a bivariate correlation coefficient (i.e., Pearson r).
- Create an APA Style table and results section that includes means, standard deviations, and correlations and addresses skew and kurtosis.

3.1.2 Planning for Practice

The practice assignment at the end of the lesson is designed as a “get (or ‘get back’) into it” assignment. You will essentially work through this very same lecture, using the same dataframe; you will simply use a different set of continuous variables.

3.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Revelle, W. (2021). An introduction to the psych package: Part I: data entry and data description. 60.
 - Revelle is the author/creator of the *psych* package. His tutorial provides both technical and interpretive information. Read pages 1-17.
- Lui, P. P. (2020). Racial microaggression, overt discrimination, and distress: (In)Direct associations with psychological adjustment. *The Counseling Psychologist*, 32.
 - This is the research vignette from which I simulate data that we can use in the lesson and practice problem.

3.2 Research Vignette

We will use data that has been simulated data from Lui [2020] as the research vignette. Controlling for overt discrimination, and neuroticism, Lui examined the degree to which racial microaggressions contributed to negative affect, alcohol consumption, and drinking problems in African American, Asian American, and Latinx American college students ($N = 713$).

Using the means, standard deviations, correlation matrix, and group sizes (n) I simulated the data. Although I provide some narration of what I did, process of simulation is beyond the learning goals of this lesson, so you are welcome to skip it. Simulating data within each chapter makes the lesson more “portable.”

```
set.seed(210807) #sets the random seed so that we consistently get the same results
# for practice, you could change (or remove) the random seed and try
# to interpret the results (they should be similar) There are
# probably more efficient ways to simulate data. Given the
# information available in the manuscript, my approach was to first
# create separate datasets for each of the racial ethnic groups and
# then bind them together.

# First, the data for the students who identified as Asian American
Asian_mu <- c(1.52, 1.72, 2.69, 1.71, 2.14, 2.35, 2.42) #creating an object containing the me
Asian_stddev <- c(2.52, 2.04, 0.47, 0.7, 0.8, 2.41, 3.36) # creating an object containing the
Asian_corMat <- matrix(c(1, 0.69, 0.19, 0.28, 0.32, 0.08, 0.23, 0.69, 1,
  0.2, 0.29, 0.33, 0.13, 0.25, 0.19, 0.2, 1, 0.5, 0.5, -0.04, 0.09, 0.28,
```



```

Black_covMat <- Black_stddev %*% t(Black_stddev) * Black_corMat
Black_dat <- MASS::mvrnorm(n = 133, mu = Black_mu, Sigma = Black_covMat,
                           empirical = TRUE)
Black_df <- as.data.frame(Black_dat)
Black_df <- rename(Black_df, OvDisc = V1, mAggr = V2, Neuro = V3, nAff = V4,
                   psyDist = V5, Alcohol = V6, drProb = V7)

# set upper and lower bound for each variable
Black_df$OvDisc[Black_df$OvDisc > 16] <- 16
Black_df$OvDisc[Black_df$OvDisc < 0] <- 0

Black_df$mAggr[Black_df$mAggr > 16] <- 16
Black_df$mAggr[Black_df$mAggr < 0] <- 0

Black_df$Neuro[Black_df$Neuro > 5] <- 5
Black_df$Neuro[Black_df$Neuro < 1] <- 1

Black_df$nAff[Black_df$nAff > 4] <- 4
Black_df$nAff[Black_df$nAff < 1] <- 1

Black_df$psyDist[Black_df$psyDist > 5] <- 5
Black_df$psyDist[Black_df$psyDist < 1] <- 1

Black_df$Alcohol[Black_df$Alcohol > 12] <- 12
Black_df$Alcohol[Black_df$Alcohol < 0] <- 0

Black_df$drProb[Black_df$drProb > 12] <- 12
Black_df$drProb[Black_df$drProb < 0] <- 0

Black_df$RacEth <- "Black"

# Third, the data for the students who identified as Latinx American
Latinx_mu <- c(1.56, 2.34, 2.69, 1.81, 2.17, 3.47, 2.69)
Latinx_stddev <- c(2.46, 2.49, 0.86, 0.71, 0.78, 2.59, 3.76)
Latinx_corMat <- matrix(c(1, 0.78, 0.27, 0.36, 0.42, -0.06, 0.08, 0.78,
                           1, 0.33, 0.26, 0.35, -0.11, -0.02, 0.27, 0.33, 1, 0.62, 0.64, -0.04,
                           0.15, 0.36, 0.26, 0.62, 1, 0.81, -0.08, 0.17, 0.42, 0.35, 0.64, 0.81,
                           1, -0.06, 0.15, -0.06, -0.11, -0.04, -0.08, -0.06, 1, 0.6, 0.08, -0.02,
                           0.15, 0.17, 0.15, 0.6, 1), ncol = 7)
Latinx_covMat <- Latinx_stddev %*% t(Latinx_stddev) * Latinx_corMat
Latinx_dat <- MASS::mvrnorm(n = 182, mu = Latinx_mu, Sigma = Latinx_covMat,
                            empirical = TRUE)
Latinx_df <- as.data.frame(Latinx_dat)
Latinx_df <- rename(Latinx_df, OvDisc = V1, mAggr = V2, Neuro = V3, nAff = V4,
                     psyDist = V5, Alcohol = V6, drProb = V7)

Latinx_df$OvDisc[Latinx_df$OvDisc > 16] <- 16

```

```

Latinx_df$OvDisc[Latinx_df$OvDisc < 0] <- 0

Latinx_df$mAggr[Latinx_df$mAggr > 16] <- 16
Latinx_df$mAggr[Latinx_df$mAggr < 0] <- 0

Latinx_df$Neuro[Latinx_df$Neuro > 5] <- 5
Latinx_df$Neuro[Latinx_df$Neuro < 1] <- 1

Latinx_df$nAff[Latinx_df$nAff > 4] <- 4
Latinx_df$nAff[Latinx_df$nAff < 1] <- 1

Latinx_df$psyDist[Latinx_df$psyDist > 5] <- 5
Latinx_df$psyDist[Latinx_df$psyDist < 1] <- 1

Latinx_df$Alcohol[Latinx_df$Alcohol > 12] <- 12
Latinx_df$Alcohol[Latinx_df$Alcohol < 0] <- 0

Latinx_df$drProb[Latinx_df$drProb > 12] <- 12
Latinx_df$drProb[Latinx_df$drProb < 0] <- 0

Latinx_df$RacEth <- "Latinx"

# binding the datasets together
Lui_sim_df <- bind_rows(Asian_df, Black_df, Latinx_df)

```

If you have simulated the data, you can continue using the the “`Lui_sim_df`” object that we created. In your own research you will likely work with a datafile stored on your computer. Although I will hashtag the code out (making it inoperable until the hashtags are removed), here is script to save the simulated data both `.csv` (think “Excel lite”) and `.rds` (it retains all the properties we specified in R) files and then bring/import them back into R. For more complete instructions see the [Ready_Set_R](#) lesson.

```

# write the simulated data as a .csv write.table(Lui_sim_df,
# file='Lui_CSV.csv', sep=',', col.names=TRUE, row.names=FALSE) bring
# back the simulated dat from a .csv file df <- read.csv
# ('Lui_CSV.csv', header = TRUE)

# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Lui_sim_df, 'Lui_RDS.rds') bring back the simulated
# dat from an .rds file df <- readRDS('Lui_RDS.rds')

```

You may have noticed a couple of things in each of these operations

- First, I named the data object to include a “`df`” (i.e., dataframe).

- It is a common (but not required) practice for researchers to simply use “df” or “dat” as the name of the object that holds their data. This practice has advantages (e.g., as making the re-use of code quite easy across datasets) and disadvantages (e.g., it is easy to get confused about what data is being used).
- Second, when you run the code, any updating *replaces* the prior object.
 - While this is irrelevant today (we are saving the same data with different names), it points out the importance of creating a sensible and systematic *order of operations* in your .rmd files and then knowing where you are in the process.

Because the data is simulated, I can simply use the data I created in the simulation, however, I will go ahead and use the convention of renaming it, “df”, which (in this case) stands for *dataframe* and is the common term for a dataset for users of R. *A quick note: in statistics “df” is also an abbreviation for “degrees of freedom.”*

```
df <- Lui_sim_df
```

3.3 Variable Types (Scale of Measurement)

When working with raw data, we begin by inspecting and preparing it for the planned analyses. The *type* of variables we have influences what statistics we will utilize. Further, the data must be formatted as that type in order for the statistic to properly execute. Variable types (or formats) are directly connected to the statistical concept of *measurement scale* (or *scale of measurement*). Researchers often think of the *categorical versus continuous* distinction, but it’s even more nuanced than that.

3.3.1 Measurement Scale

Categorical variables name *discrete* or *distinct* entities where the categorization has no inherent value or order. When there are two categories, the variable type is **binary** (e.g., pregnant or not, treatment and control conditions). When there are more than two categories, the variable type is **nominal** (e.g., teacher, student, or parent; Republican, Democrat, or Independent).

Ordinal variables are also categorical variables where the score reflects a logical order or relative rank (e.g., the order of finishing in a race). A challenge with the ordinal scale is the inability to determine the distance between rankings. The percentile rank is a (sometimes surprising) example of the ordinal scale. Technically, Likert type scaling (e.g., providing ratings on a 1-to-5 scale) is ordinal because it is uncertain that the distance between each of the anchors is equal. Practically, though, most researchers treat the Likert type scale as interval. This is facilitated, in part, because most Likert-type scales have multiple items which are averaged into a single score. Navarro[2020a] uses the term, **quasi-interval** to describe Likert-type scaling.

Continuous variables can take on any value in the measurement scale that is being used. **Interval** level data has equal distances between each unit on the scale. Two classic examples of interval level data are temperature and year. Whether using Fahrenheit or Celsius, the rating of 0 does not mean there is an absence of temperature, rather, it is simply a number along a continuum of temperature. Another interval example is calendrical time. In longitudinal research, we frequently note the date

or year (e.g., 2019) of an event. It is highly unlikely that the value zero will appear in our research and if it did, it would not represent the absence of time. A researcher can feel confident that a variable is on the interval scale if the values can be meaningfully added and subtracted.

Ratio level data also has equal distances between each unit on the scale, plus it has a true zero point where the zero indicates absence. Examples are behavioral counts (e.g., cigarettes smoked) and time-on-task (e.g., 90 seconds). Ratio data offers more manipulative power because researchers can add, subtract, multiply, and divide ratio level data.

3.3.2 Corresponding Variable Structure in R

With these definitions in mind, we will see if R is reading our variables correctly. R will provide the following designations of variables:

Abbreviation	Unabbreviated	Used for	Scale of Measurement
num	numerical	numbers that allow decimals or fractional values	quasi-interval, interval, or ratio
int	integer	whole numbers (no decimals)	quasi-interval, interval, or ratio
chr	character	sometimes termed “string” variables, these are interpreted as words	NA
Factor	factor	two or more categories; R imposes an alphabetical order; the user can re-specify the order based on the logic of the design	nominal

Looking back at the Lui [2020] article we can determine what the scale of measurement is for each variable and what the corresponding R format for that variable should be:

Name	Variable	How assessed	Scale of measurement	R format
OvDis	Overt racial discrimination	9 items, 1-to-4 Likert scaling for frequency and stressfulness assessed separately, then multiplied	quasi-interval	numerical
mAggr	Racial and ethnic microaggressions	28 items, 1-to-4 Likert scaling for frequency and stressfulness assessed separately, then multiplied	quasi-interval	numerical
Neuro	Neuroticism	4 items, 1-to-5 Likert scaling	quasi-interval	numerical
nAff	Negative affect	6 items, 1-to-4 Likert scaling	quasi-interval	numerical
psyDist	Psychological distress	6 items, 1-to-5 Likert scaling	quasi-interval	numerical
Alcohol	Hazardous alcohol use	10 items, 0-to-4 Likert scaling	quasi-interval	numerical
drProb	Drinking problems	10 items, 0-to-4 Likert scaling	quasi-interval	numerical

Name	Variable	How assessed	Scale of measurement	R format
RacEth	Race Ethnicity	3 categories	nominal	factor

We can examine the accuracy with which R interpreted the type of data with the *structure()* command.

```
str(df)
```

```
'data.frame': 713 obs. of 8 variables:
 $ OvDisc : num 1.62 0 2.08 0 0 ...
 $ mAggr : num 2.78 0 2.8 0 0 ...
 $ Neuro : num 3.24 2.59 2.79 2.53 1.34 ...
 $ nAff : num 1.11 1 1.08 1 1.03 ...
 $ psyDist: num 2.07 1 1.06 1.82 1.36 ...
 $ Alcohol: num 1.63 0 3.2 2.52 2.43 ...
 $ drProb : num 2.4073 5.3177 0.6424 1.1671 0.0774 ...
 $ RacEth : chr "Asian" "Asian" "Asian" "Asian" ...
```

Only Race/Ethnicity needs to be transformed from a character (“chr”) variable to a factor. I will use the *mutate()* function in the *dplyr* package to convert the RacEth variable to be a factor with three levels.

```
library(tidyverse)
df <- df %>%
  dplyr::mutate(RacEth = as.factor(RacEth))
```

Let's check the structure again. Below we see that the RacEth variable is now a factor. R has imposed an alphabetical order: Asian, Black, Latinx.

```
# checking the structure of the data
str(df)
```

```
'data.frame': 713 obs. of 8 variables:
 $ OvDisc : num 1.62 0 2.08 0 0 ...
 $ mAggr : num 2.78 0 2.8 0 0 ...
 $ Neuro : num 3.24 2.59 2.79 2.53 1.34 ...
 $ nAff : num 1.11 1 1.08 1 1.03 ...
 $ psyDist: num 2.07 1 1.06 1.82 1.36 ...
 $ Alcohol: num 1.63 0 3.2 2.52 2.43 ...
 $ drProb : num 2.4073 5.3177 0.6424 1.1671 0.0774 ...
 $ RacEth : Factor w/ 3 levels "Asian","Black",...: 1 1 1 1 1 1 1 1 1 ...
```

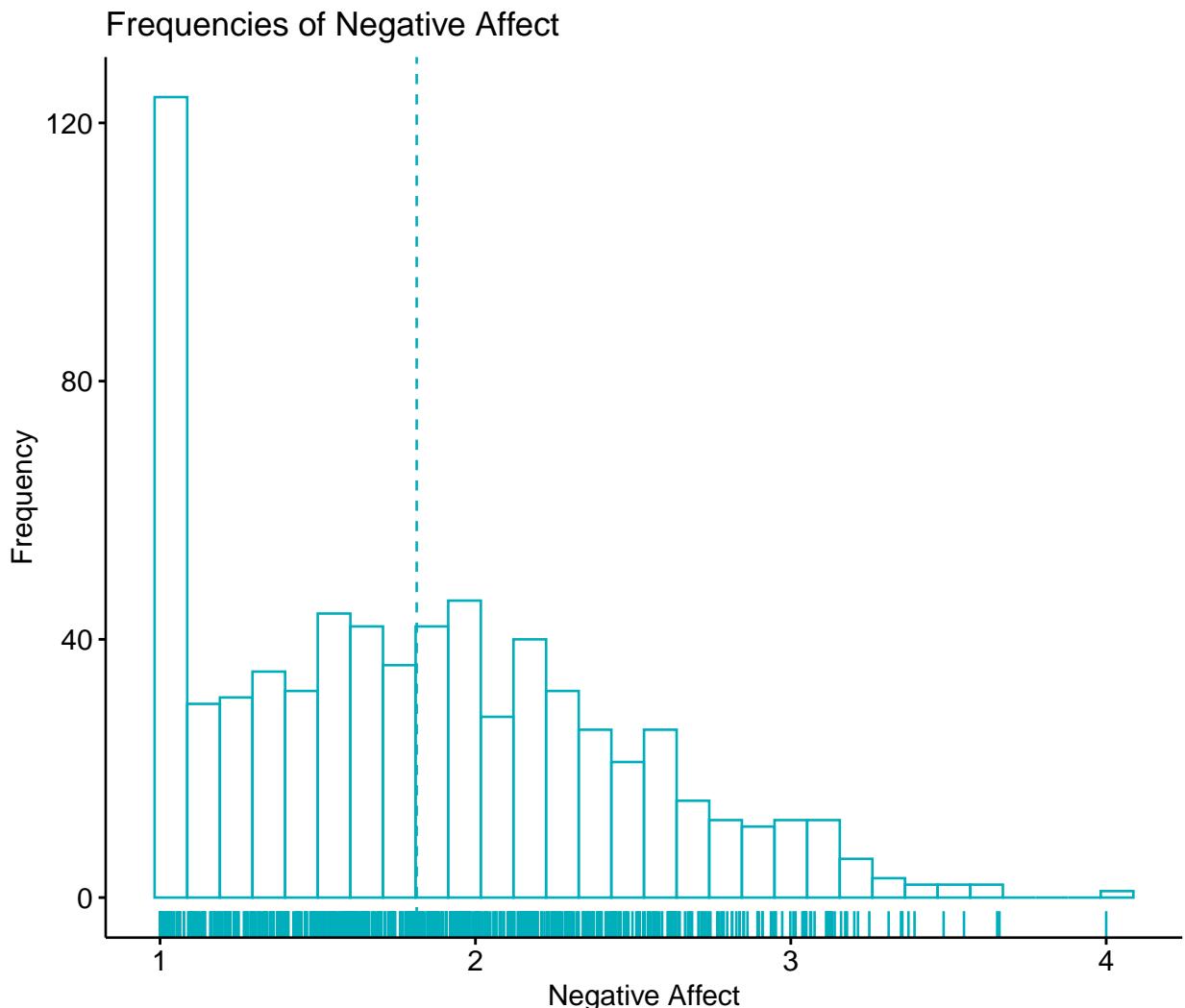
3.4 Descriptive Statistics

While the majority of this OER (and statistics training in general) concerns the ability to make predictions or inferences (hence *inferential statistics*) from data, we almost always begin data analysis by describing it (hence, *descriptive statistics*).

Our research vignette contains a number of variables. Lui [2020] was interested in predicting negative affect, alcohol consumption, and drinking problems from overt discrimination, microaggressions, neuroticism, through psychological distress. This research model is a *mediation* model (or model of indirect effects) and is beyond the learning objectives of today's instruction. In demonstrating descriptive statistics, we will focus on one of the dependent variables: negative affect.

As we begin to explore the descriptive and distributional characteristics of this variable, it may be helpful to visualize it through a histogram.

```
ggpubr::gghistogram(df$nAff, xlab = "Negative Affect", ylab = "Frequency",
add = "mean", rug = TRUE, color = "#00AFBB", title = "Frequencies of Negative Affect")
```



3.4.1 Measures of Central Tendency

Describing data almost always begins with *measures of central tendency*: the mean, median, and mode.

3.4.1.1 Mean

The **mean** is simply a mathematical average of the non-missing data. The mathematical formula is frequently expressed this way:

$$\bar{X} = \frac{X_1 + X_2 + X_3 \dots + X_N}{N}$$

Because this formula is clumsy to write, there is statistical shorthand to help us convey it more efficiently (not necessarily, more easily).

Placing information below (where to start), above (where to stop), and to the right (what data to use) of the summation operator (\sum), provides information about the nature of the data. In the formula below, we learn from the notation to the right that we use the individual data in the vector X . We start with the first piece of data ($i = 1$) and stop with the N th (or last) case.

$$\sum_{i=1}^N X_i$$

The $\frac{1}{N}$ notation to the left tells us that we are calculating the mean.

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

R is an incredible tool in that we can type out mathematical operations, use functions from base R, and use packages to do the work for us. If we had the following toy dataset (2, 3, 2, 1, 5, NA) we could calculate the mean by typing it out:

```
(2 + 3 + 2 + 1 + 5)/5
```

```
[1] 2.6
```

Alternatively we could use the built-in functions in base R to do the work for us. Let me add a little complexity by creating a single variable (a vector of data) and introducing a little missingness (i.e., the “NA”).

```
toy <- c(2, 3, 2, 1, 5, NA)
toy <- as.data.frame(toy)
```

I can use the base R function *mean()*. Inside the parentheses I point to the data. The function automatically sums the values. When there is missingness, adding *na.rm=TRUE* tells the function to exclude the missing variables from the count (i.e., the denominator would still be 5).

```
mean(toy$toy, na.rm = TRUE)
```

```
[1] 2.6
```

In my simulation of the research vignette, we have no missing values, none-the-less, it is, perhaps a good habit to include the *na.rm=TRUE* specification in our code. Because we have an entire dataframe, we just point to the dataframe and the specific variable (i.e., negative affect).

```
mean(df$nAff, na.rm = TRUE)
```

```
[1] 1.813748
```

3.4.1.2 Median

The middle value in a set of values is the **median**. The easiest way to calculate the median is to sort the numbers:

Unsorted	Sorted
2, 3, 2, 1, 5,	1, 2, 2, 3, 5

And select the middle value. Because we have an odd number of values ($N = 5$), our median is 2. If we had an even number of values, we would take the average of the middle two numbers.

We can use a base R function to calculate the median for us. Let's do it first with the toy data:

```
median(toy$toy, na.rm = TRUE)
```

```
[1] 2
```

Let's also calculate it for the negative affect variable from the research vignette.

```
median(df$nAff, na.rm = TRUE)
```

```
[1] 1.765367
```

3.4.1.3 Mode

The **mode** is the score that occurs most frequently. When a histogram is available, spotting the mode is easy because it will have the tallest bar. Determining the mode can be made complicated if there are ties for high frequencies of values. A common occurrence of this happens in the **bimodal** distribution.

Unfortunately, there is no base R function that will call a mode. In response, Navarro developed and included a function in the *lsr* package that accompanies her [2020a] textbook. Once the package is installed, you can include two colons, the function name, and then the dataset to retrieve the mode.

```
lsr::modeOf(toy$toy)
```

```
[1] 2
```

From our toy data, we see the *modeOf()* function returns a 2.

Let's retrieve the mode from the negative affect variable in our research vignette.

```
lsr::modeOf(df$nAff)
```

```
[1] 1
```

The value is a 1.0 and is likely an artifact of how I simulated the data. Specifically, to ensure that the values fell within the 1-to-4 range, I rounded up to 1.0 any negative values and rounded down to 4.0 any values that were higher than 4.0.

3.4.1.4 Relationship between mean, median, and mode

Many inferential statistics rely on manipulations of the mean. The mean, though, can be misleading when it is influenced by outliers. Therefore, as we engage in preliminary exploration, it can be quite useful to calculate all three measures of central tendency, as well as exploring other distributional characteristics.

As a bit of an advanced cognitive organizer, it may be helpful to know that in a normal distribution, the mean, median, and mode are the same number (or quite close). In a positively skewed distribution, the mean is higher than the median which is higher than the mode. In a negatively skewed distribution, the mean is lower than the median, which is lower than the mode.

```
mean(df$nAff, na.rm=TRUE)
```

```
[1] 1.813748
```

```
median(df$nAff, na.rm=TRUE)
```

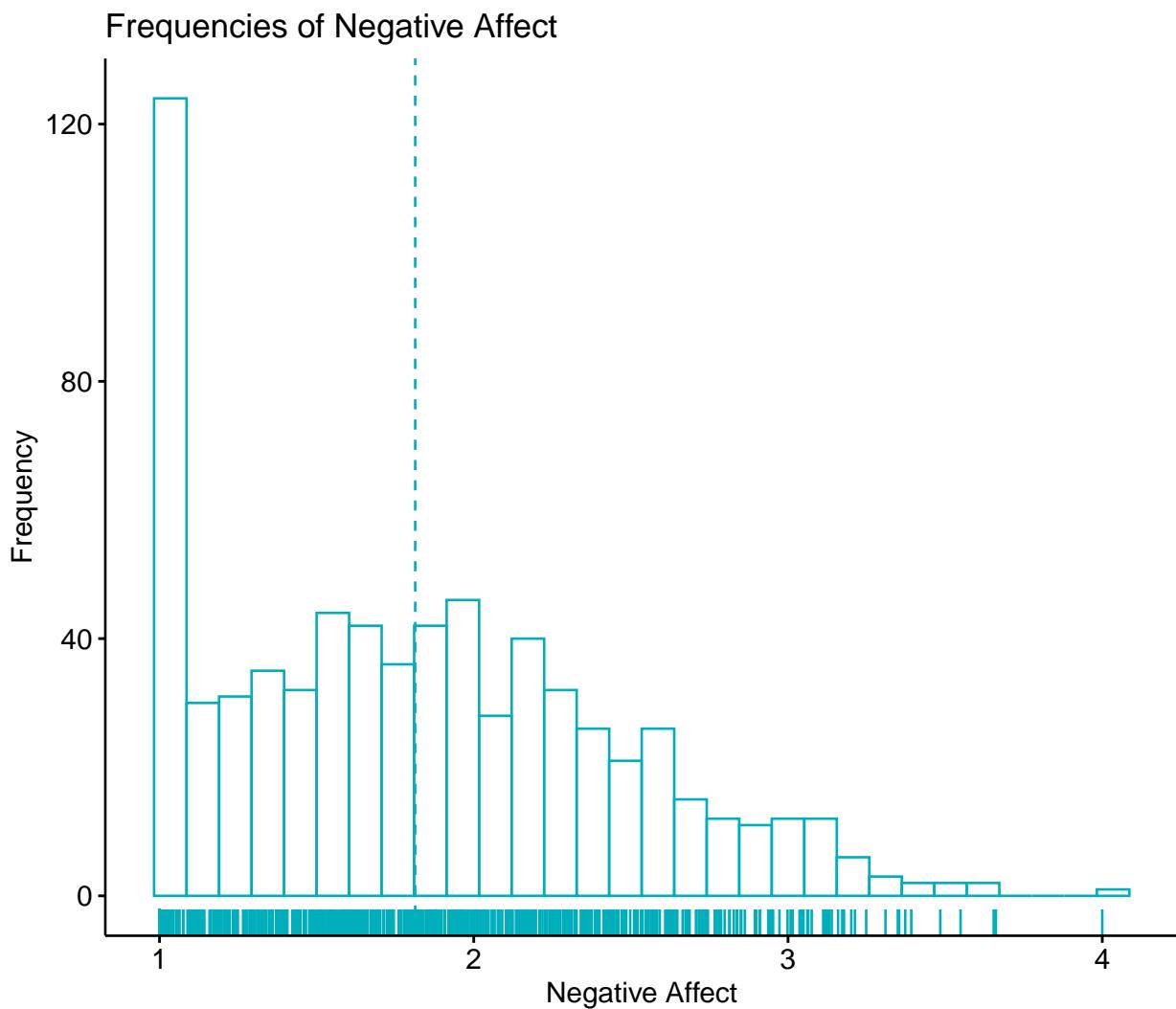
```
[1] 1.765367
```

```
lsr::modeOf(df$nAff, na.rm=TRUE)
```

```
[1] 1
```

In our research vignette, the mean (1.81) is higher than the median (1.75) is higher than the mode (1.0). This would suggest a positive skew. Here is a reminder of our histogram:

```
ggpubr::gghistogram(df$nAff, xlab = "Negative Affect", ylab = "Frequency",
  add = "mean", rug = TRUE, color = "#00AFBB", title = "Frequencies of Negative Affect")
```



3.5 Variability

Researchers are critically interested in the spread or dispersion of the scores.

3.5.1 Range

The **range** is the simplest assessment of variability and is calculated by identifying the highest and lowest scores and subtracting the lowest from the highest. In our toy dataset, arranged from low-to-high (1, 2, 2, 3, 5) we see that the low is 1 and high is 5; 4 is the range. We can retrieve this data with three base R functions that ask for the minimum score, the maximum score, or both together – the range:

```
min(toy$toy, na.rm = TRUE)
```

```
[1] 1
```

```
max(toy$toy, na.rm = TRUE)
```

```
[1] 5
```

```
range(toy$toy, na.rm = TRUE)
```

```
[1] 1 5
```

The negative affect variable from our research vignette has the following range:

```
min(df$nAff)
```

```
[1] 1
```

```
max(df$nAff)
```

```
[1] 4
```

```
range(df$nAff)
```

```
[1] 1 4
```

With a low of 1 and high of 4, the range of negative affect is 3. This is consistent with the description of the negative affect measure.

One limitation of the range is that it is easily influenced by extreme scores.

3.5.2 Percentiles, Quantiles, Interquartile Range

The **interquartile range** is middle 50% of data, or the scores that fall between 25th and 75th percentiles. Before calculating that, let's first define **quantiles** and **percentiles**. **Quantiles** are values that split a data into equal portions. **Percentiles** divide the data into 100 equal parts. Percentiles are commonly used in testing and assessment. You may have encountered them in standardized tests such as the SAT and GRE where both the score obtained and its associated percentile are reported. When graduate programs evaluate GRE scores, depending on their criteria and degree of competitiveness they may set a threshold based on percentiles (e.g., using a cut off of the 50th, 75th, or higher percentile for the verbal or quantitative GRE scores).

We have already learned the value of the median. The median is also the 50th percentile. We can now use the *quantile()* function and indicate we want the value at the 50% percentile.

Let's first examine the toy dataset:

```
median(toy$toy, na.rm = TRUE)
```

```
[1] 2
```

```
quantile(toy$toy, probs = 0.5, na.rm = TRUE)
```

```
50%
2
```

As shown by our calculation, the value at the median and the 50th percentile is 2.0. Let's look at those values for the research vignette:

```
median(df$nAff, na.rm = TRUE)
```

```
[1] 1.765367
```

```
quantile(df$nAff, probs = 0.5, na.rm = TRUE)
```

```
50%
1.765367
```

Again, we see the same result. Half of the values for negative affect are below 1.76; half are above. The *quantile()* function is extremely useful. We can retrieve the raw score at any percentile, and we could ask for as many as we desired. Here's an example.

```
quantile(df$nAff, probs = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9))
```

10%	20%	30%	40%	50%	60%	70%	80%
1.000000	1.142097	1.376633	1.582701	1.765367	1.943260	2.143177	2.360980
	90%						
		2.682303					

Quartiles divide the data into four equal parts. The **interquartile range** is the spread of data between the 25th and 75th percentiles (or quartiles). We calculate the interquartile range by first obtaining those values, and then subtracting the lower from the higher.

```
quantile(df$nAff, probs = c(0.25, 0.75))
```

25%	75%
1.271045	2.240372

We see that a score of 1.29 is at the 25th percentile and a score of 2.24 is at the 75th percentile. If we subtract 1.29 from 2.24...

```
2.24 - 1.29
```

```
[1] 0.95
```

...we learn that the interquartile range is 0.95. We could also obtain this value by using the *IQR()* function in base R.

```
IQR(df$nAff, na.rm = TRUE)
```

```
[1] 0.9693262
```

You may be asking, “When would we use the interquartile range?” When data are influenced by **outliers** (i.e., extreme scores), using a more truncated range (the middle 50%, 75%, 90%) may be an option (if the dataset it large enough). At this point, though, the goal of this lesson is simply to introduce different ways of examining the variability in a dataset. Ultimately, we are working our way to the **standard deviation**. The next logical step is the **mean deviation**.

3.5.3 Deviations around the Mean

Nearly all statistics include assessments of variability in their calculation and most are based on deviations around the mean. In fact it might be good to pause for a moment and consider as the lessons in this OER (and those that follow) continue, we will be engaged in *mathematical and statistical modeling*. In a featured article in the *American Psychologist*, Rodgers [2010] described models as a representation of reality that has two features:

- the model describes reality in some important ways, and
- the model is simpler than reality.

Albeit one of the simplest, the mean is a statistical model. Rodgers noted this when he wrote, “The mean and variance have done yeoman service to psychology and other behavioral sciences,” [2010, p. 4]. These next statistical operations will walk through the use of the mean, particularly in its role in understanding variance. In later lessons, means and variances are used in understanding relations and differences.

A first step in understanding mean deviation is to ask, “How far does each individual score deviates from the mean of scores?” We can demonstrate this with our toy dataset. I am taking more steps than necessary to (a) make clear how the mean deviation (abbreviated, mdev) is calculated and (b) practice using R.

First, I will create a variable representing the mean:

```
# Dissecting the script, each variable is referenced by
# df_nameDOLLARSIGNvariable_name
toy$mean <- mean(toy$toy, na.rm = TRUE)
head(toy) #displays the first 6 rows of the data
```

```
toy mean
1 2 2.6
2 3 2.6
3 2 2.6
4 1 2.6
5 5 2.6
6 NA 2.6
```

Next, I will subtract the mean from each individual score. The result

```
toy$mdev <- toy$toy - toy$mean
head(toy) #displays the first 6 rows of the data
```

```
toy mean mdev
1 2 2.6 -0.6
2 3 2.6 0.4
3 2 2.6 -0.6
4 1 2.6 -1.6
5 5 2.6 2.4
6 NA 2.6 NA
```

The variable, *mdev* (short for “mean deviation”) lets us know how far the individual score is from the mean. Unfortunately, it does not provide an overall estimate of variation. Further, summing and averaging these values all result in zero. Take a look:

```
# Dissecting the script, Wrapping the sum and mean script in 'round'
# and following with the desired decimal places, provides a rounde
# result.
round(sum(toy$mdev, na.rm = TRUE), 3)
```

```
[1] 0
```

```
round(mean(toy$mdev, na.rm = TRUE), 3)
```

```
[1] 0
```

One solution is to create the *mean absolute deviation*. We first transform the mean deviation score to their absolute values, and then sum them.

```
toy$abslt_m <- abs(toy$mdev)
head(toy)
```

```
toy mean mdev abslt_m
1 2 2.6 -0.6 0.6
```

2	3	2.6	0.4	0.4
3	2	2.6	-0.6	0.6
4	1	2.6	-1.6	1.6
5	5	2.6	2.4	2.4
6	NA	2.6	NA	NA

And now to average them:

```
round(mean(toy$abslt_m, na.rm = TRUE), 3)
```

```
[1] 1.12
```

This value tells how far individual observations are from the mean, “on average.” In our toy dataset, the average distance from the mean is 1.12.

So that we can keep statistical notation in our mind, this is the formula calculating the absolute mean deviation:

$$\sum_{i=1}^n |X_i - \bar{X}|$$

Let’s quickly repeat the process with the negative affect variable in our research vignette. So that we can more clearly see the relationship of the new variables to negative affect, let me create a df containing only nAff:

```
library(tidyverse)
df_nAff <- df %>%
  dplyr::select(nAff)
```

```
df_nAff$mdevNA <- df_nAff$nAff - mean(df_nAff$nAff, na.rm = TRUE)
df_nAff$abNAmdev <- abs(df_nAff$mdevNA)
head(df_nAff)
```

	nAff	mdevNA	abNAmdev
1	1.109882	-0.7038658	0.7038658
2	1.000000	-0.8137481	0.8137481
3	1.075573	-0.7381750	0.7381750
4	1.000000	-0.8137481	0.8137481
5	1.025246	-0.7885020	0.7885020
6	1.920559	0.1068111	0.1068111

```
round(mean(df_nAff$abNAmdev, na.rm = TRUE), 3)
```

```
[1] 0.523
```

Thus, the absolute mean deviation for the negative affect variable in our research vignette is 0.521.

Although relatively intuitive, the absolute mean deviation is not all that useful. Most statistics texts include it because it is one of the steps toward variance, and ultimately, the standard deviation.

3.5.4 Variance

Variance is considered to be an *average* dispersion calculated by summing the squared deviations and dividing by the number of observations (minus 1; more on that in later lessons).

Our next step is to square the mean deviations. This value is also called the *sum of squared errors*, *sum of squared deviations around the mean*, or *sums of squares* and is abbreviated as *SS*. Below are common statistical representations:

$$SS = \sum_{i=1}^n (X_i - \bar{X})^2$$

Let's do it with our toy data.

```
toy$mdev2 <- (toy$mdev) * (toy$mdev)
sum(toy$mdev2, na.rm = TRUE) #sum of squared deviations
```

```
[1] 9.2
```

```
head(toy)
```

	toy	mean	mdev	abslt_m	mdev2
1	2	2.6	-0.6	0.6	0.36
2	3	2.6	0.4	0.4	0.16
3	2	2.6	-0.6	0.6	0.36
4	1	2.6	-1.6	1.6	2.56
5	5	2.6	2.4	2.4	5.76
6	NA	2.6	NA	NA	NA

Thus, our *SS* (sums of squares or sums of squared errors) is 9.2.

To obtain the variance we divide by *N* (or *N* - 1; described in later lessons). Here are the updated formulas:

$$s^2 = \frac{SS}{N-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{N-1}$$

Let's do this with the toy data:

```
9.2/(5 - 1) #calculated with the previously obtained values
```

```
[1] 2.3
```

```
# to obtain the 'correct' calculation by using each of these
# individual R commands, we need to have non-missing data
toy <- na.omit(toy)
sum(toy$mdev2, na.rm = TRUE)/((nrow(toy) - 1)) #variance
```

```
[1] 2.3
```

Of course R also has a function that will do all the steps for us:

```
mean(toy$toy, na.rm = TRUE)
```

```
[1] 2.6
```

```
var(toy$toy, na.rm = TRUE)
```

```
[1] 2.3
```

The variance around the mean (2.6) of our toy data is 2.3.

Let's quickly repeat this process with the negative affect variable from the research vignette. In prior steps we had calculated the mean deviations by subtracting the mean from each individual score. Next we square the mean deviations....

```
df_nAff$NAmd2 <- (df_nAff$mdevNA) * (df_nAff$mdevNA)
head(df_nAff)
```

	nAff	mdevNA	abNAmdev	NAmd2
1	1.109882	-0.7038658	0.7038658	0.49542700
2	1.000000	-0.8137481	0.8137481	0.66218597
3	1.075573	-0.7381750	0.7381750	0.54490233
4	1.000000	-0.8137481	0.8137481	0.66218597
5	1.025246	-0.7885020	0.7885020	0.62173547
6	1.920559	0.1068111	0.1068111	0.01140861

... and sum them.

```
sum(df_nAff$NAmd2, na.rm = TRUE) #sum of squared deviations
```

```
[1] 283.8923
```

Our sums of squared deviations around the mean is 283.44. When we divide it by $N - 1$, we obtain the variance. We can check our work with (a) the values we calculated at each step, (b) the steps written in separate R code, and (c) the *var()* function.

```
283.44/(713 - 1) # calculating with the individual pre-calculated values
```

```
[1] 0.3980899
```

```
sum(df_nAff$NAmd2, na.rm = TRUE)/((nrow(df_nAff) - 1)) #calculated with steps from separate R
```

```
[1] 0.3987252
```

```
var(df_nAff$nAff) #calculated using the base R function
```

```
[1] 0.3987252
```

Unfortunately, because the mean deviations were squared, this doesn't interpret well. Hence, we move to the *standard deviation*.

3.5.5 Standard Deviation

The standard deviation is simply the square root of the variance. Stated another way, it is an estimate of the average spread of data, presented in the same metric as the data.

Calculating the standard deviation requires earlier steps:

1. Calculating the mean.
2. Calculating mean deviations by subtracting the mean from each individual score.
3. Squaring the mean deviations.
4. Summing the mean deviations to create the *SS*, or sums of squares.
5. Dividing the *SS* by $N - 1$; this results in the *variance* around the mean.

The 6th step is to take the square root of variance. It is represented in the formula, below:

$$s = \sqrt{\frac{SS}{N - 1}} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{N - 1}}$$

Repeated below are each of the six steps for the toy data:

```
# six steps wrapped into 1
toy$mdev <- toy$toy - mean(toy$toy, na.rm = TRUE)
toy$mdev2 <- (toy$mdev) * (toy$mdev)
# I can save the variance calculation as an object for later use
toy_var <- sum(toy$mdev2)/(nrow(toy) - 1)
# checking work with the variance function
var(toy$toy)
```

```
[1] 2.3
```

The seventh step is to take the square root of variance.

```
# grabbing the mean for quick reference
mean(toy$toy)
```

[1] 2.6

```
# below the 'toy_var' object was created in the prior step
sqrt(toy_var)
```

[1] 1.516575

```
# checking work with the R function to calculate standard deviation
sd(toy$toy)
```

[1] 1.516575

It is common to report means and standard deviations for continuous variables in our datasets. For the toy data our mean is 2.6 with a standard deviation of 1.52.

Let's repeat the process for the negative affect variable in the research vignette. First the six steps to calculate variance.

```
# six steps wrapped into 1
df_nAff$mdevNA <- df_nAff$nAff - mean(df_nAff$nAff, na.rm = TRUE)
df_nAff$NAmd2 <- (df_nAff$mdevNA) * (df_nAff$mdevNA)
# I can save the variance calculation as an object for later use
nAff_var <- sum(df_nAff$NAmd2)/(nrow(df) - 1)
# checking work with the variance function
var(df_nAff$nAff)
```

[1] 0.3987252

The seventh step is to take the square root of variance.

```
# grabbing the mean for quick reference
mean(df_nAff$nAff)
```

[1] 1.813748

```
# below the 'toy_var' object was created in the prior step
sqrt(nAff_var)
```

[1] 0.6314469

```
# checking work with the R function to calculate standard deviation
sd(df_nAff$nAff)
```

```
[1] 0.6314469
```

In APA Style we use M and SD as abbreviations for mean and standard deviation, respectively. In APA Style, non-Greek statistical symbols such as these are italicized. Thus we would write $M = 1.81 (SD = 0.63)$ in a statistical string of results.

We can examine the standard deviation in relation to its mean to understand how narrowly or broadly the data is distributed. Relative to a same-sized mean, a small standard deviation means that the mean represents the data well. A larger standard deviation, conveys that there is greater variability and the mean, alone, is a less valid representation of the score.

In later lessons we will explore the standard deviation in more detail – learning how we can use it in the determination of the significance and magnitude of relations and differences.

3.6 Are the Variables Normally Distributed?

Statistics that we use are accompanied by assumptions about the nature of variables in the dataset. A common assumption is that the data are *normally distributed*. That is, the data presumes a standard normal curve.

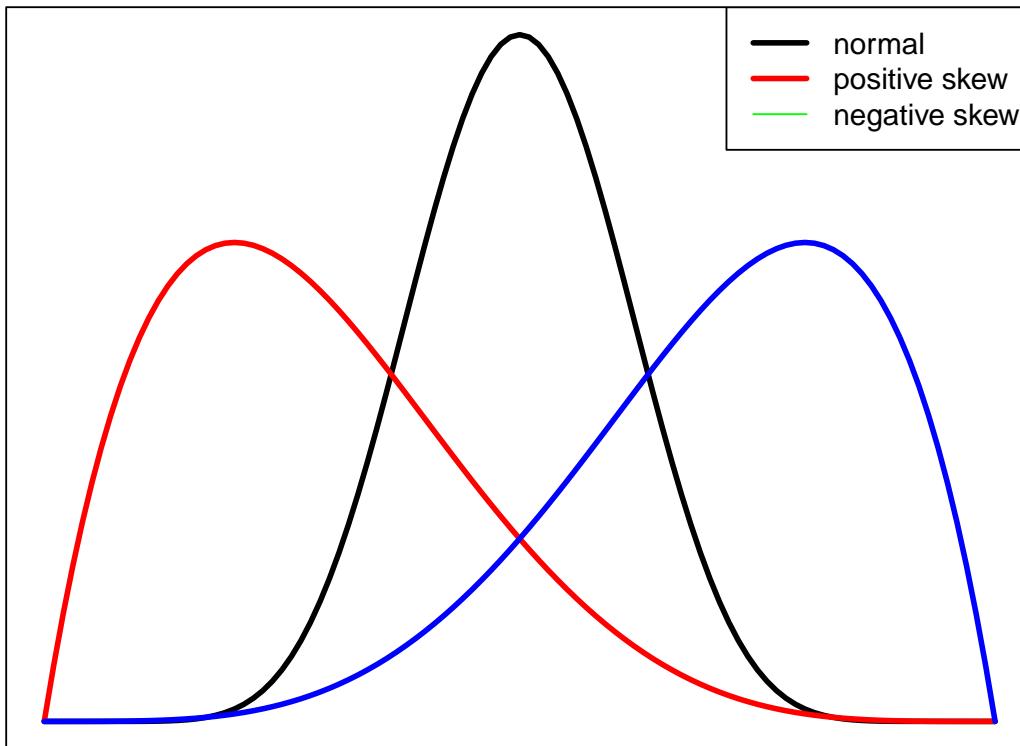
For a streamlined presentation, let me create a df with three, continuously scaled, variables of interest.

```
# I have opened the tidyverse library so that I can use the pipe
library(tidyverse)
df_3vars <- df %>%
  dplyr::select(nAff, mAggr, drProb)
```

3.6.1 Skew and Kurtosis

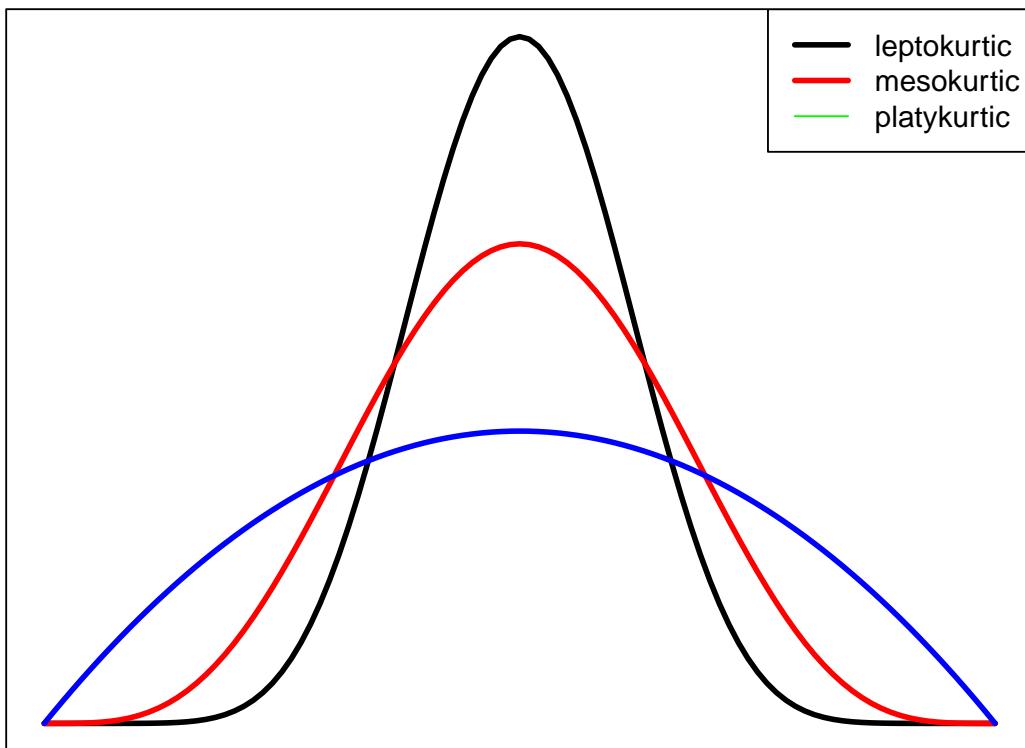
Skew and kurtosis are indicators of non-normality. Skew refers to the degree to which the data is symmetrical. In the figure below, the symmetrical distribution in the center (the black line) has no apparent evidence of skew. In contrast, the red figure whose curve (representing a majority of observations) in the left-most part of the graph (with the tail pulling to the right) is positively skewed; the blue figure whose curve (representing a majority of cases) is in the right-most part of the graph (with the tail pulling to the left) is negatively skewed.

Positive, Normal, and Negative Skew



Kurtosis refers to the degree to which the distribution of data is flat or peaked. Mesokurtic distributions are considered to be closest to normal. Leptokurtic distributions are peaked and platykurtic distributions are flat. As we will learn as we progress, visual observation of data is a legitimate component in evaluating skew and kurtosis.

Kurtosis: Platykurtic, Mesokurtic, Leptokurtic



There have been numerous approaches to calculating and interpreting skew and kurtosis. Consequently, different statistics packages calculate skew and kurtosis differently. The *psych* package (a go-to-for a variety of tasks) offers three different options for calculating skew and kurtosis. These are specified in the script as “type=#” (i.e., 1, 2, or 3 [the default]). Revelle [2021] refers readers to Joanes and Gill’s [1998] article for detailed information about each. A very helpful resource to understand skew, kurtosis, and its interpretation is found in chapter four (Data Preparation and Psychometrics Review) of Kline’s [2016a] SEM text is helpful in the interpretation of skew and kurtosis. Summarizing by simulation studies for structural equation modeling (i.e., multivariate statistics that are generally characterized as large sample studies using maximum likelihood estimation), Kline suggested that *type=1* skew values greater than the absolute value of 3.0 are “severely” skewed. Regarding *type=1* kurtosis, Kline noted the literature has suggested that values from 8.0 to 20.0 have been described as severely kurtotic. As an interpretive framework, Kline suggested that absolute values greater than 10.0 are problematic and values greater than 20 are serious. He added that this rule-of-thumb errs on the conservative side.

The *psych::describe* specification of “type=1” results in the *skew index* and *kurtosis index*. For simplicity sake, I will refer to this specific variation of skew and kutosis as “type=1.” This is a very quick way to obtain initial values.

```
psych::describe(df_3vars, type = 1)
```

Using Kline's [2016a] guidelines for evaluation, a quick review of the type=1 output indicates that no skew value exceeded the absolute value of 3.0. That is, across the nAff, mAggr, and drProb variables the highest type=1 skew value was 0.93 Regarding kurtosis, no value had a greater magnitude than .59 and all fell below the absolute value of 10. A limitation of the type=1 output and Kline's interpretative guidelines is that the simulation studies that led to the interpretive guidelines were based on structural equation modeling. These statistics are multivariate in nature, they typically use maximum likelihood estimators, and are based on large samples.

An alternative tool for identifying distributions that are severely skewed or kurtotic is the “skew.2SE” and “kurt.2SE” output from *pastecs::stat.desc*.

These values represent the type=1 skew (or kurtosis) value divided by two-times its respective standard error (i.e., the standard error of the skew or kurtosis distribution, not the *se* value associated with the variable). The result is a standardized value that, on its own, indicates statistical significance. In the case of skew.2SE and kurt2SE, values of 1 ($p < .05$), 1.29 ($p < .01$), and 1.65 ($p < .001$) represent statistically significant departures from symmetry (skew) and normal peakedness (kurtosis). Unfortunately, this tool is not without criticism.

$$skew.2SE = \frac{S-0}{2*SE_{skewness}} \text{ and } kurt.2SE = \frac{S-0}{2*SE_{kurtosis}}$$

The skew.2SE and kurt.2SE values can be obtained with `pastecs::stat.desc` by adding the “norm = TRUE” statement.

```
pastecs::stat.desc(df_3vars, norm = TRUE)
```

Statisticians have noted that these standardized values are quite sensitive to sample size [Field, 2012, Kline, 2016a]. In large samples even minor deviations from normality may appear as statistically significant. In contrast, small samples with lower power may be severely non-normal, but skew and kurtosis could go undetected. When sample sizes are smaller, using the 1.96 (or “2”) criteria is acceptable in determining a significant skew or kurtosis, however, as the sample size increases, the probability of rejecting the hypothesis that skew (or kurtosis) also zero increases. Field [2012] noted that in such cases it is appropriate to relax the standard and evaluate skew or kurtosis against the 1.29 ($p < .01$) criteria. Further, when samples are larger than 200, it may be more appropriate to abandon the interpretation of the z -values and, instead, examine the shape of the distribution rather than to interpret these standardized values.

Comparing the skewness and kurtosis type=1 values to the skew.2SE and kurt.2SE values, we can see the interpretive challenges.

Variable	skewness	skew.2SE	kurtosis	kurt.2SE
nAff	0.575	3.141***	-0.176	-0.481

Variable	skewness	skew.2SE	kurtosis	kurt.2SE
mAggr	0.926	5.059***	0.575	1.572**
drProb	0.783	4.278***	-0.168	-0.460

Values in the “skewness” column are concerning when they are exceed the absolute value of 3.0; none are. Values in the “skew.2SE” column are statistically significant at $p < .001$ when they exceed 1.65. Here, the two approaches to interpreting skew both suggest positive skew (i.e., heavy distribution in the left with a long tail to the right), but only the “skew.2SE” results suggest that the degree of skewness is significant/concerning.

Regarding kurtosis, values in the “kurtosis” column become concerning when they exceed the absolute value of 10; none are. Values in the “kurt.2SE” become statistically significant when they exceed 1.0. The mAggr variable’s value of 1.572 is statistically significant at $p < .01$.

So how do we think about skewness and kurtosis in our data? First, I simulated a dataset with more than 700 cases. This far exceeds the “large” sample size of 200. Therefore, interpreting the type 1 skewness and kurtosis values according to Kline’s [2016a] criteria of less than the absolute values of 3 and 10, respectively, is probably most appropriate. Further, skewness and kurtosis are only two dimensions of assessing whether or not a distribution is normally distributed. Thus, I will keep these results in mind as I examine additional metrics (especially when we look at histograms with superimposed curve).

In this OER, I will predominantly use the type=1 output from the *psych::describe* package and use Kline’s [2016a] interpretive criteria. I do think the “skew.2SE” and “kurt.2SE” metrics can be useful when sample sizes are smaller (perhaps $N = 100$ or less) and ordinary least squares (such as used in ANOVA and regression models) statistics will be utilized. In any case, if I have significant concerns about normality, I always return to more extensive and authoritative sources to make my decisions about preparing my data for analysis.

3.6.2 Shapiro-Wilk Test of Normality

In addition to skew and kurtosis, there are formal statistical tests that evaluate whether or not our data is statistically significantly different than a normal distribution. One of those is the Shapiro-Wilk test of normality. The output we obtained from *pastecs::stat.desc* included the Shapiro-Wilk test value and the associated p value. When $p < .05$, our data is statistically significantly different from a normal distribution.

In our simulated data, all variables were statistically significantly different than a normal distribution ($nAff : W = 0.948, p < .001$; $mAggr : W = 0.913, p < .001$; $drProb : W = 0.900, p < .001$).

Just because data is skewed, kurtotic, or non-normally distributed does not (necessarily) mean that we cannot use it. As we move through the lessons in this OER we will evaluate the quality of the data according to the statistical assumptions associated with the statistic we are using. Often there are tools that we can use (e.g., variations of the statistic that are robust to violations of assumptions, deleting univariate or multivariate outliers) in spite of our data characteristics.

3.7 Relations between Variables

Preliminary investigation of data almost always includes a report of their bivariate relations. Correlation coefficients express the magnitude of relationships on a scale ranging from -1 to +1. A correlation coefficient of

- -1.0 implies a 1:1 inverse relationship, such that for every unit of increase in variable A, there is a similar decrease in variable B,
- 0.0 implies no correspondence between two variables,
- 1.0 implies that as A increases by one unit, so does B.

Correlation coefficients are commonly represented in two formulas. In a manner that echoes the calculation of *variance*, the first part of the calculation estimates the covariation (i.e., *covariance*) of the two variables of interest.

$$\text{Cov}(X, Y) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

The problem is that the result is unstandardized and difficult to interpret. Therefore, the second part of the calculation of the correlation coefficient results in the standardization of the metric in the -1 to +1 scale.

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

Covariation and correlation matrices are central to many of our statistics therefore, those of who teach statistics believe that it is important to take a look “under the hood.” From our research vignette, let’s calculate the relationship between negative affect and psychological distress.

Examining the first formula, some parts should look familiar:

- $(X_i - \bar{X})$: We can see that we need to subtract the mean from the first(X) variable in involved in the correlation; we saw this when we calculated *mean deviations*.
- $(Y_i - \bar{Y})$: We repeat the *mean deviation* process for the second (Y) variable.

Let’s work step-by-step through the calculation of a correlation coefficient. So that we can more easily see what we are doing with the variables, I will create a super tiny dataframe with the two variables of interest (negative affect and microaggressions):

```
# just in case it turned off, I'm reopening tidyverse so that I can
# use the pipe
library(tidyverse)
# using the dplyr package to select the two variables in this tiny df
df4corr <- df %>%
  dplyr::select(nAff, mAggr)
# displaying the first 6 rows of df4corr ('dataframe for
# correlations' -- I made this up)
head(df4corr)
```

	nAff	mAggr
1	1.109882	2.779103
2	1.000000	0.000000
3	1.075573	2.798700
4	1.000000	0.000000
5	1.025246	0.000000
6	1.920559	1.857067

First we calculate the mean deviations for negative affect and microaggressions.

```
# calculating the mean deviation for negative affect
df4corr$MDnAff <- df4corr$nAff - mean(df4corr$nAff)
# calculating the mean deviation for microaggressions
df4corr$MDmAggr <- df4corr$mAggr - mean(df4corr$mAggr)
# displaying the first 6 rows of df4corr
head(df4corr)
```

	nAff	mAggr	MDnAff	MDmAggr
1	1.109882	2.779103	-0.7038658	0.2925610
2	1.000000	0.000000	-0.8137481	-2.4865424
3	1.075573	2.798700	-0.7381750	0.3121577
4	1.000000	0.000000	-0.8137481	-2.4865424
5	1.025246	0.000000	-0.7885020	-2.4865424
6	1.920559	1.857067	0.1068111	-0.6294752

The next part of the formula $\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$ suggests that we sum the cross-products of these mean deviations. Here we multiply the mean deviations to create the “cross-product.”

```
# Creating a crossproduct variabl by multiplying negative affect by
# psych distress
df4corr$crossproductXY <- df4corr$MDnAff * df4corr$MDmAggr
# displaying the first 6 rows of df4corr
head(df4corr)
```

	nAff	mAggr	MDnAff	MDmAggr	crossproductXY
1	1.109882	2.779103	-0.7038658	0.2925610	-0.20592370
2	1.000000	0.000000	-0.8137481	-2.4865424	2.02341919
3	1.075573	2.798700	-0.7381750	0.3121577	-0.23042700
4	1.000000	0.000000	-0.8137481	-2.4865424	2.02341919
5	1.025246	0.000000	-0.7885020	-2.4865424	1.96064379
6	1.920559	1.857067	0.1068111	-0.6294752	-0.06723494

Next, we sum the column of cross-products.

```
sum(df4corr$crossproductXY)
```

```
[1] 236.0952
```

To obtain the covariance, the next part of the formula suggests that we multiply the sum of cross-products by $\frac{1}{N-1}$. I will do this in one step.

```
# I have created the object 'cov' so I can use it in a calculation,
# later The 'nrow' function will count the number of rows and use
# that value
cov <- 1/(nrow(df4corr) - 1) * sum(df4corr$crossproductXY)
# Because I created an object, R markdown won't automatically display
# it; I have to request it by listing it
cov
```

```
[1] 0.3315944
```

The covariance between negative affect and psychological distress is 0.373.

We now move to the second part of the formula to create the interpretable, standardized, correlation coefficient.

$$r_{XY} = \frac{\text{Cov}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

We will use our covariance value in the numerator. The denominator involves the multiplication of the standard deviations of X and Y. Because we have already learned how to calculate standard deviation in a step-by-step manner, I will use code to simplify that process:

```
cov / (sd(df4corr$nAff) * sd(df4corr$mAggr))
```

```
[1] 0.2395157
```

Our results suggest that the relationship between negative affect and psychological distress is positive, as one increases so does the other. Is it strong? This really depends on your field of scholarship. The traditional values of .10, .30, and .50 are interpreted as small, medium, and large [Cohen et al., 2003]. Hence, when $r = 0.27$, we can say that it is (more-or-less) medium.

Is it statistically significant? Because this is an introductory chapter, we will not calculate this in a stepwise manner, but use the `cor.test()` function in base R to check our prior math and retrieve the p value associated with the correlation coefficient.

```
cor.test(df4corr$nAff, df4corr$mAggr)
```

Pearson's product-moment correlation

```
data: df4corr$nAff and df4corr$mAggr
t = 6.5781, df = 711, p-value = 0.00000000009241
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1690651 0.3075312
sample estimates:
cor
0.2395157
```

In a statistical string we would report the result of this Pearson correlation coefficient as: $r = 0.27$ ($p < .001$).

3.8 Shortcuts to Preliminary Analyses

Unless you teach statistics (or take another statistics class), you may never need to work through all those individual steps again. Rather, a number of R packages make retrieval of these values relatively simple and efficient.

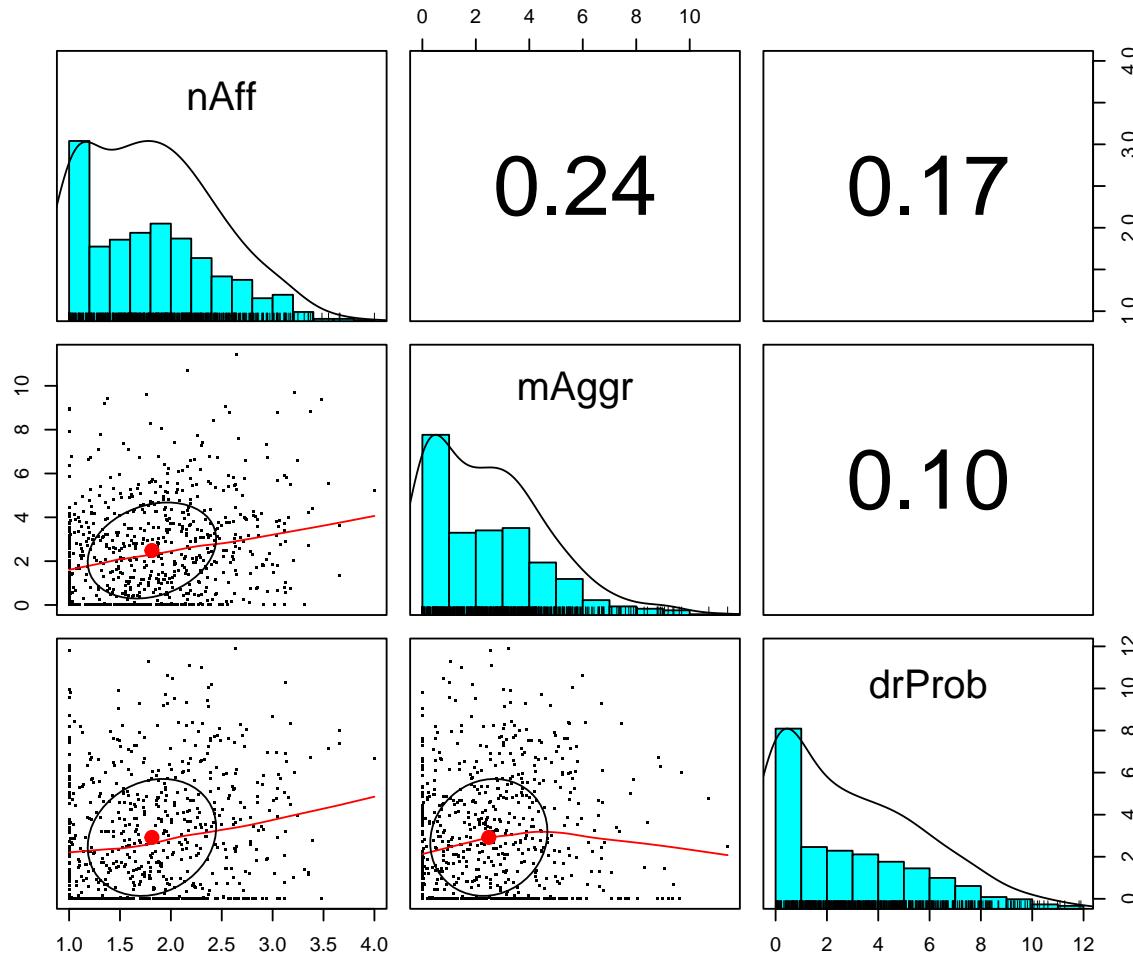
3.8.1 SPLOM

The *pairs.panels()* function in the *psych* package produces a SPLOM (i.e., scatterplot matrix) which includes:

- histograms of each individual variable within the dataframe with a curve superimposed (located on the diagonal),
- scatterplots of each bivariate combination of variables (located below the diagonal), and
- correlation coefficients of each bivariate combination of variables (located above the diagonal).

To provide a simple demonstration this, I will use our *df* with the three continuously scaled variables of interest:

```
# in the code below, psych points to the package pairs.panels points
# to the function we simply add the name of the df; if you want fewer
# variables than that are in the df, you may wish to create a smaller
# df adding the pch command is optional and produces a finer
# resolution
psych::pairs.panels(df_3vars, pch = ".")
```



What do we observe?

- There is a more-or-less moderate correlation between negative affect and microaggressions ($r = 0.27$)
- There is a small-to-moderate correlation between negative affect and drinking problems ($r = 0.18$)
- There is a small correlation between microaggressions and drinking problems ($r = 0.09$)
- All variables have a positive skew (with pile-up of scores on the lower end and tail pulling to the right); this is consistent with the values we calculated earlier
- The scatterplots can provide clues to relations that are not necessarily linear.
 - Look at the relationship between negative affect and drinking problems. As negative affect hits around 2.75, there is a change in the relationship, such that drinking problems increase.
 - Taking time to look at plots such as these can inform subsequent analyses.

3.8.2 apaTables

Writing up an APA style results section frequently involves tables. A helpful package for doing this is *apaTables*. An instructional article notes the contributions of tools like this to the *reproducibility* of science by reducing errors made when the author or analyst retypes or copies text from output to the manuscript. When the R script is shared through an open science framework, reproducibility is further enhanced [Stanley and Spence, 2018].

We pass the desired df to the *apaTables::apa.cor.table*. Commands allow us to specify what is included in the table and whether it should be displayed in the console or saved as a document to the project's folder.

```
# the apa.cor.table function removes any categorical variables that
# might be in the df
Table1_Cor <- apaTables::apa.cor.table(df_3vars, filename = "Table1_Cor.doc",
                                         table.number = 1, show.conf.interval = FALSE, landscape = TRUE)
```

The ability to suppress reporting of reporting confidence intervals has been deprecated in this version of the package. The function argument `show.conf.interval` will be removed in a later version.

```
# swap in this command to see it in the R Markdown file
print(Table1_Cor)
```

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1	2
1. nAff	1.81	0.63		
2. mAggr	2.49	2.19	.24** [.17, .31]	
3. drProb	2.92	2.78	.17** [.10, .24]	.10* [.02, .17]

Note. M and SD are used to represent mean and standard deviation, respectively.
 Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations
 that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

Because I added: `filename = "Table1_Cor.doc"`, a word version of the table will appear in the same file folder as the .rmd file and data. It is easily manipulated with tools in your word processing package.

3.9 An APA Style Writeup

The statistics used in this lesson are often presented in the preliminary results portion of an empirical manuscript. Some of the results are written in text and some are presented in tables. APA Style recommends that the narration of results not duplicate what is presented in the tables. Rather, the write-up only highlights and clarifies what is presented in the table(s).

At the outset, let me note that a primary purpose of the Lui [2020] article was to compare the relations of variables between three racial/ethnic groups in the U.S. identified as Asian American, Black, and Latinx. Because we did not run separate analyses for each of the groups, my write-up does not make these distinctions. I highly recommend that you examine the write-up of results and the accompanying tables in Lui's article. The presentation is clear and efficient (i.e., it conveys maximal information in as little space as possible).

Below is an example of how I might write up these preliminary results:

Preliminary Results

Our sample included 713 participants who self-identified as Asian American, Black/African American, and Latinx American. Inspection of the characteristics of the three variables of interest (negative affect, microaggressions, drinking problems) indicated that all variables were positively skewed, however the values of skew and kurtosis did not exceed commonly used thresholds of concern [Kline, 2016a]. In contrast, Shapiro-Wilk tests of normality suggested that the distribution of all three variables were statistically significantly different than a normal distribution ($nAff : W = 0.948, p < .001$; $mAggr : W = 0.913, p < .001$; $drProb : W = 0.900, p < .001$). Means, standard deviations, and a correlation matrix are presented in Table 1. We noted that the correlation between negative affect and microaggressions was moderate ($r = 0.27$); correlations between remaining variables were smaller.

3.10 Practice Problems

The three exercises described below are designed to “meet you where you are” and allow you to challenge your skills depending on your goals as well as your comfort with statistics and R.

Regardless which you choose, work one or more of the problems with R packages:

- Create a smaller df from a larger df selecting a minimum of three continuously scaled variables
- Calculate and interpret descriptive statistics
- Create the SPLOM (pairs.panels)
- Use the *apaTables* package to make an APA style table with means, standard deviations, and correlations

- Write an APA Style results section for these preliminary analyses

Additionally, please complete at least one set of *hand calculations*, that is using the code demonstrated in the chapter to work through the formulas that compute the descriptive statistics that are the focus of this lesson. At this stage in your learning, you may ignore any missingness in your dataset by excluding all rows with missing data in your variables of interest.

3.10.1 Problem #1: Change the Random Seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation (at the very top), then rework the lesson exactly as written. This should provide minor changes to the data (maybe in the second or third decimal point), but the results will likely be very similar.

3.10.2 Problem #2: Swap Variables in the Simulation

Use the simulated data from the Lui [2020] study. However, select three continuous variables (2 must be different from mine) and then conduct the analyses. Be sure to select from the variables that are considered to be *continuous* (and not *categorical*).

3.10.3 Problem #3: Use (or Simulate) Your Own Data

Use data for which you have permission and access. This could be IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; or data from other chapters in this OER.

3.10.4 Grading Rubrics

Regardless which option(s) you chose, use the elements in the grading rubrics to guide you through the practice. Worked examples are provided in the [Appendix](#).

Working the problem with R and R packages	Points Poss	Points Earned
1. Create a df with 3 continuously scaled variables of interest	3	
2. Produce descriptive statistics	3	
3. Produce SPLOM/pairs.panels	3	
4. Produce an apaTables matrix	3	
5. Produce an APA Style write-up of the preliminary analyses	5	
6. Explanation/discussion with a grader	5	
**Totals	22	

Hand Calculations	Points Poss	Points Earned
1. Create a variable that represents the mean.	2	

Hand Calculations	Points Poss	Points Earned
2. Create a variable that represents the mean deviation.	2	
3. What is the value of the <i>sum</i> of mean deviations?	2	
4. Create a variable that represents the absolute mean deviation. What is the <i>sum</i> of the absolute mean deviation? What is the value of the <i>mean</i> of the absolute mean deviation? What does this value tell you?	4	
5. Create a variable that represents the mean deviation squared.	2	
6. What are the values of the sum of squared deviations around the mean SS , variance s^2 , and standard deviation (s)?	3	
7. Using the same general approach, calculate the mean deviation and standard deviation for a second, continuously scaled variable.	5	
8. Create a variable that represents the <i>cross-product</i> (of the mean deviations). What is the sum of these cross-products?	2	
9. Calculate the value of their covariance.	2	
8. Calculate value of correlation coefficient.	2	
**Totals	26	

3.11 Homeworked Example

Screencast Link

Preliminary analyses often consist of means, standard deviations, and correlations. These can be helpful in determining whether or not data are normally distribution. Correlations and pairs.panels also assess the relatedness of the variables.

If you wanted to use this example and dataset as a basis for a homework assignment, you could (a) select a different course (i.e., Multivariate or Psychometrics) and/or (b) different variables.

3.11.1 Working the Problem with R and R Packages

3.11.1.1 Create a df with 3 continuously scaled variables of interest

The ReC.rds is the entire dataset. Let's first open it.

Recall that students (represented by the *deID* variable) could contribute up to three course evaluations (i.e., ANOVA, psychometrics, multivariate) each. In many statistics, repeated observations creates dependencies that need to be accounted for statistically.

To avoid this dependency and to practice an R skill, let's first filter the data, selecting only those students who took ANOVA.

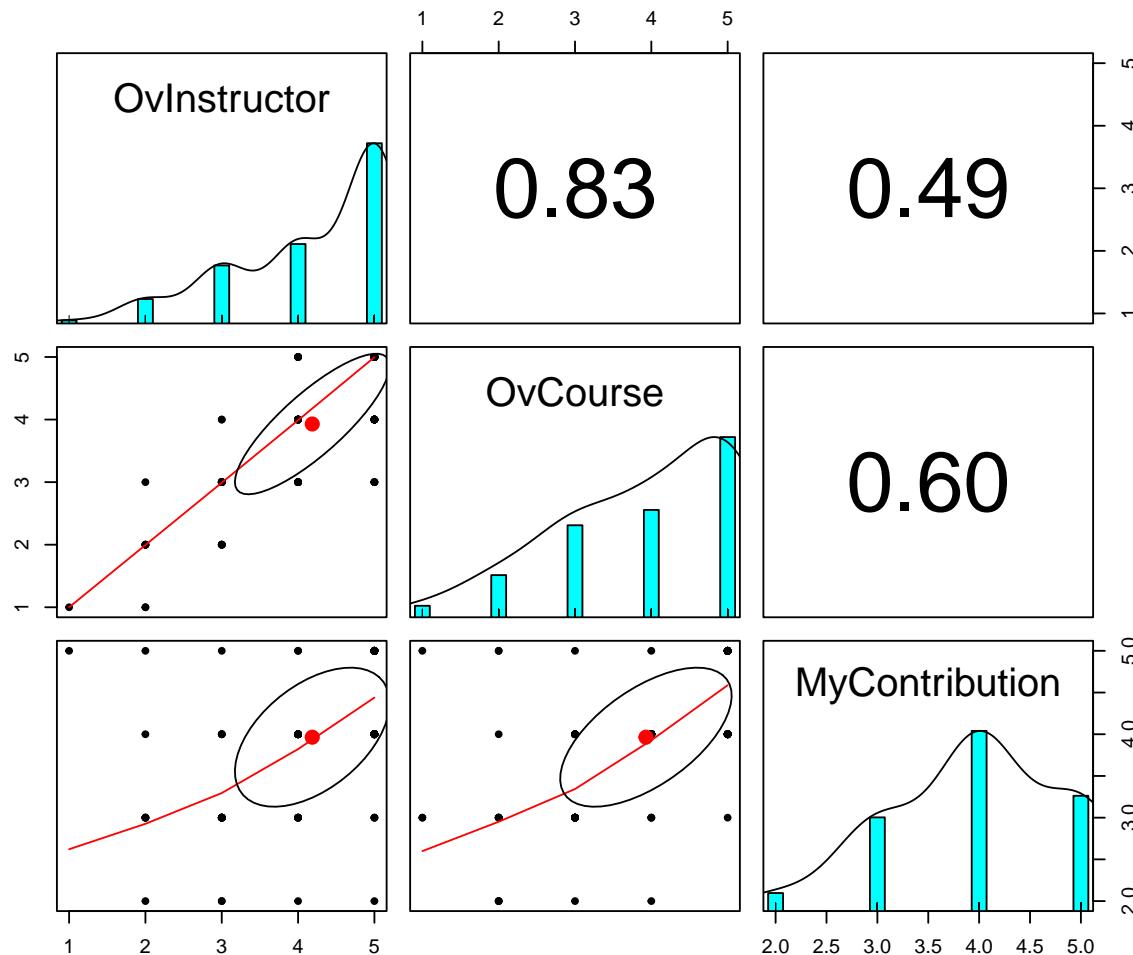
3.11.1.2 Create a df with 3 continuously scaled variables of interest

The assignment requires that we downsize to three variables. We could pick any three.

3.11.1.3 Produce descriptive statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
OvInstructor	1	113	4.19	1.01	5	4.34	0.00	1	5	4	-0.98
OvCourse	2	113	3.93	1.12	4	4.07	1.48	1	5	4	-0.72
MyContribution	3	113	3.96	0.83	4	4.01	1.48	2	5	3	-0.39
	kurtosis	se									
OvInstructor	-0.07	0.10									
OvCourse	-0.49	0.11									
MyContribution	-0.55	0.08									

3.11.1.4 Produce SPLOM/pairs.panels



3.11.1.5 Produce an apaTables matrix

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1	2
1. OvInstructor	4.19	1.01		
2. OvCourse	3.93	1.12	.83** [.76, .88]	
3. MyContribution	3.96	0.83	.49** [.34, .62]	.60** [.46, .70]

Note. M and SD are used to represent mean and standard deviation, respectively.

Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

3.11.1.6 Produce an APA Style write-up of the preliminary analyses

Our sample included 113 doctoral students in Clinical and Industrial-Organizational psychology doctoral (PhD) programs who were completing a statistics class focused on analysis of variance. Visual inspection of three dimensions of course evaluation (overall instructor, overall course, my contributions) combined with formal evaluation of skewness and kurtosis suggested that their distributions did not violate the assumption of univariate normality. That is, skew values all fell below the absolute value of 3 and kurtosis values all fell below the absolute value of 10 [Kline, 2016a]. Means, standard deviations, and a correlation matrix are presented in Table 1. All three correlations were strong and statistically significant. We noted that the correlation between the overall instructor and overall course was especially high ($r = .83$, $p < .001$)

3.11.2 Hand Calculations

Although these are termed “hand calculations,” you may use the code demonstrated in the chapter to work these problems.

I am going to continue with the *tiny3* dataset I used when I worked the problem with R and R packages. Given that this is for homework, let’s avoid problems with missingness by deleting any rows with missing data:

If you need to reimport data, here is a quick recap of the code explained earlier.

To avoid problems in the code we are used that is caused by missingness, we will eliminate any rows with missing data.

3.11.2.1 Create a variable that represents the mean.

I will start with the OvInstructor variable. Inspect the dataframe to see that this new variable exists.

3.11.2.2 Create a variable that represents the mean deviation.

Inspect the dataframe to see that this new variable exists. Note that this functions to “center” the mean around zero.

3.11.2.3 What is the value of the sum of mean deviations?

[1] 0

Yes, zero!

3.11.2.4 Create a variable that represents the absolute mean deviation.

Inspect the dataframe to see that this new variable no longer has negative values.

What is the value of the sum of the absolute mean deviation?

[1] 96.071

What is the value of the mean of the absolute mean deviation?

[1] 0.85

What does this value tell you?

Average distance of each value from the mean.

3.11.2.5 Create a variable that represents the mean deviation squared.

What is the value of the sum of squared deviations around the mean (also known as sums of squares; sometimes abbreviated as SS)?

[1] 115.0973

What is the value of the variance (s^2)?

There are at least two ways to do this with basic code (and then we can check our work).

Here’s how to do it with “more code.”

[1] 1.027655

Here's how to do it with the numbers that I calculated:

```
[1] 1.027654
```

Checking my work with the *var* function from base R. If it's wrong, I need to rework some of the previous steps.

```
[1] 1.027655
```

What is the value of the standard deviation (s)?

There are two ways to calculate it with basic code; and then we can check it with more code from base R.

```
[1] 1.013733
```

```
[1] 1.013733
```

```
[1] 1.013733
```

3.11.2.6 Using the same general approach, calculate the mean deviation and standard deviation for a second, continuously scaled variable.

My second variable is MyContribution

```
[1] 0
```

```
[1] 0
```

```
[1] 0.8337652
```

3.11.2.7 Create a variable that represents the *cross-product* (of the mean deviations). What is the sum of these cross-products?

The sum of the crossproduct is:

```
[1] 46.74336
```

3.11.2.8 Calculate the value of their covariance.

```
[1] 0.4173514
```

3.11.2.9 Calculate value of correlation coefficient.

```
[1] 0.4937606
```

And now I can check my work with a function from base R.

```
Pearson's product-moment correlation

data: tiny3$OvInstructor and tiny3$MyContribution
t = 5.9825, df = 111, p-value = 0.00000002737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3400714 0.6217934
sample estimates:
cor
0.4937812
```

The correlation between ratings of overall instructor and my contribution is 0.493, $p < .001$.

t-tests

The lessons offered in the *t*-tests section introduce *inferential statistics*. In the prior chapters, our use of measures of central tendency (i.e., mean, median, mode) and variance (i.e., range, variance, standard deviation) serve to *describe* a sample.

As we move into *inferential* statistics we evaluate data from a sample and try to determine whether or not we can use it to draw conclusions (i.e, predict or make inferences) about a larger, defined, population.

The *t*-test lessons begin with an explanation of the *z*-score and progress through one sample, independent samples, and paired samples *t*-tests. Each lesson is centered around a research vignette that was focused on physicians' communication with patients who were critically and terminally ill and in the intensive care unit at a hospital [Elliott et al., 2016].

In addition to a conceptual presentation of each statistic, each lesson includes:

- a workflow that guides researchers through decision-points in each statistic,
- the presentation of formulas and R code for “hand-calculating” each component of the formula,
- script for efficiently computing the statistic with R packages,
- an “recipe” for an APA style presentation of the results,
- a discussion of *power* in that particular statistic with R script for calculating sample sizes sufficient to reject the null hypothesis, if in fact, it is appropriate to do so, and
- suggestions for practice that vary in degree of challenge.

Chapter 4

One Sample t -tests

[Screencasted Lecture Link](#)

```
options(scipen = 999) #eliminates scientific notation
```

Researchers and program evaluators, may wish to know if their data differs from an external standard. In today's research vignette, we will ask if the time physicians spent with their patients differed from an external benchmark. The one sample t -test is an appropriate tool for this type of analysis. As we work toward the one sample t -test we take some time to explore the standard normal curve and z -scores, particularly as they relate to probability.

4.1 Navigating this Lesson

There is just over one hour of lecture. If you work through the materials with me, plan for an additional hour-and-a-half.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

4.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Convert raw scores to z -scores (and back again).
- Using the z table, determine the probability of an occurrence.
- Recognize the research questions for which utilization of a one sample t -test would be appropriate.
- Narrate the steps in conducting a one-sample t -test, beginning with testing the statistical assumptions through writing up an APA style results section.

- Calculate a one-sample t -test in R (including effect sizes).
- Interpret a 95% confidence interval around a mean difference score.
- Produce an APA style results section for a one-sample t -test .
- Determine a sample size that (given a set of parameters) would likely result in a statistically significant effect, if there was one.

4.1.2 Planning for Practice

The suggestions for homework vary in degree of complexity. The more complete descriptions at the end of the chapter follow these suggestions.

- Rework the one-sample t -test in the lesson by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- Rework the one-sample t -test in the lesson by changing something else about the simulation. For example, if you are interested in power, consider changing the sample size.
- Conduct a one sample t -test with data to which you have access and permission to use. This could include data you simulate on your own or from a published article.

4.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- How To Do a One-Sample T-test in R: Best Tutorial You Need. (n.d.). Datanovia. Retrieved May 24, 2023, from <https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/how-to-do-a-one-sample-t-test-in-r/>
 - The primary R code we use is from the rstatix/Datanovia tutorial. *Navarro, D. (2020). Chapter 13: Comparing two means. In [Learning Statistics with R - A tutorial for Psychology Students and other Beginners](#). Retrieved from [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro))
 - Navarro's OER includes a good mix of conceptual information about t -tests as well as R code. My lesson integrates her approach as well as considering information from Field's [2012] and Green and Salkind's [2017c] texts.
- Elliott, A. M., Alexander, S. C., Mescher, C. A., Mohan, D., & Barnato, A. E. (2016). Differences in Physicians' Verbal and Nonverbal Communication With Black and White Patients at the End of Life. *Journal of Pain and Symptom Management*, 51(1), 1–8. <https://doi.org/10.1016/j.jpainsymman.2015.07.008>
 - The source of our research vignette.

4.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them. Remove the hashtags for the code to work.

```
# will install the package if not already installed
# if(!require(psych)){install.packages('psych')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(dplyr)){install.packages('dplyr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(knitr)){install.packages('knitr')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(pwr)){install.packages('pwr')}
# if(!require(pastecs)){install.packages('pastecs')}
# if(!require(rstatix)){install.packages('rstatix')}
```

4.2 z before t

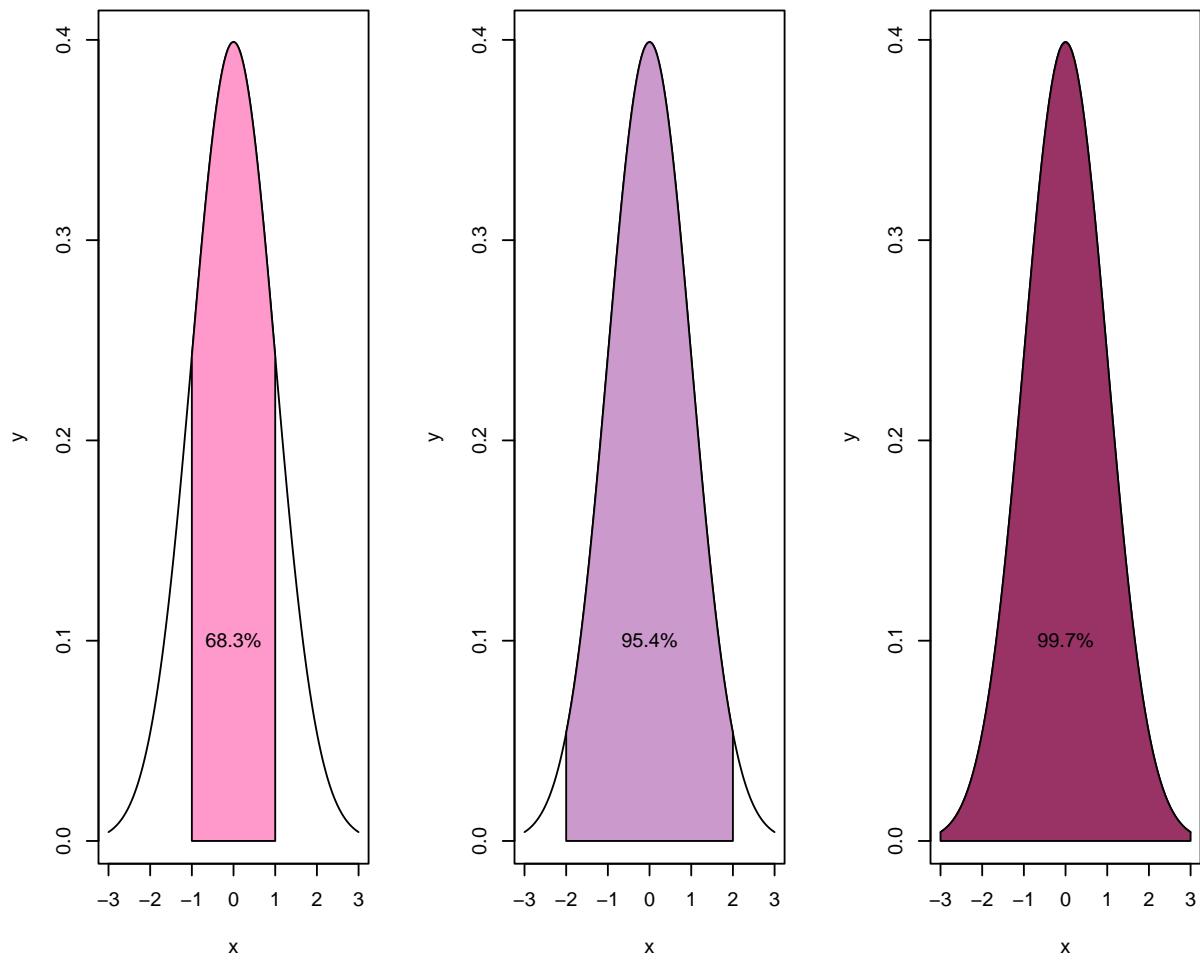
Probability density functions are mathematical formula that specifies idealized versions of known distributions. The equations that define these distributions allow us to calculate the probability of obtaining a given score. This is a powerful tool.

As students progress through statistics, they become familiar with a variety of these distributions including the t -distribution (commonly used in t -tests), F -distribution (commonly used in analysis of variance [ANOVA]), and Chi-square (X^2) distributions (used in a variety of statistics, including structural equation modeling). The z distribution is the most well-known of these distributions.

The z distribution is also known as the normal distribution, the bell curve, or the standard normal curve. Its mean is always 0.00 and its standard deviation is always 1.00. Regardless of the actual mean and standard deviation:

- 68.3% of the area falls within 1 standard deviation of the mean
- 95.4% of the distribution falls within 2 standard deviations of the mean
- 99.7% of the distribution falls within 3 standard deviations of the mean

Properties of the Normal Distribution



z-scores are transformations of raw scores, in standard deviation units. Using the following formula, so long as the mean and standard deviation are known, any set of continuously scaled scores can be transformed to a *z*-scores equivalent:

$$z = \frac{X - \bar{X}}{s}$$

We can rearrange the formula to find what raw score corresponds with the *z*-score.

$$X = \bar{X} + z(s)$$

The properties of the *z*-score and the standard normal curve allow us to make inferences about the data.

4.2.1 Simulating a Mini Research Vignette

Later in this larger section on *t*-tests we introduce a research vignette that focuses on time physicians spend with patients. Because working with the *z*-test requires a minimum sample size of 120 (and

the research vignette has a sample size of 33), I will quickly create normally distributed sample of 200 with a mean of 10 minutes and a standard deviation of 2 minutes per patient. This will allow us to ask some important questions of the data.

```
# https://r-charts.com/distribution/histogram-curves/
set.seed(220821)
PhysTime <- data.frame(minutes = rnorm(200, mean = 10, sd = 2))
```

Using the `describe()` function from the `psych` package, we can see the resulting descriptive statistics.

```
psych::describe(PhysTime$minutes)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	200	9.9	2	9.98	9.93	2	3.68	15.15	11.47	-0.2	0.03	0.14

Specifically, in this sample size of 200, our mean is 9.9 with a standard deviation of 2.0.

4.2.2 Raw Scores, *z*-scores, and Proportions

With data in hand, let's ask, "What is the range of time that physicians spend with patients that fall within 1 standard deviation of the mean?" We would answer this question by applying the raw score formula ($X = \bar{X} + z(s)$) to +1 and -1 standard deviation.

```
9.9 - 1 * (2)
```

```
[1] 7.9
```

```
9.9 + 1 * (2)
```

```
[1] 11.9
```

Because $\pm 1SD$ covers 68% of the distribution, we now know that 68% of patients have physician visits that are between 7.9 and 11.9 minutes long.

What about $\pm 2SDs$? Similarly, we would apply the raw score formula, using 2 for the standard deviation.

```
9.9 - 2 * (2)
```

```
[1] 5.9
```

```
9.9 + 2 * (2)
```

```
[1] 13.9
```

Two standard deviations around the mean captures 94.5% of patients; patients in this range receive between visits that range between 5.9 and 13.9 minutes.

And what about $\pm 3SDs$? This time we use 3 to represent the standard deviation.

```
9.9 - 3 * (2)
```

```
[1] 3.9
```

```
9.9 + 3 * (2)
```

```
[1] 15.9
```

Three standard deviations around the mean captures 99.7% of patients; patients in this range receive between visits that range between 3.9 and 15.9 minutes.

4.2.3 Determining Probabilities

We can also ask questions of **probability**. For example, what is the probability that a physician spends at least 9.9 minutes with a patient? To answer this question we first calculate the *z*-score associated with 9.9 minutes.

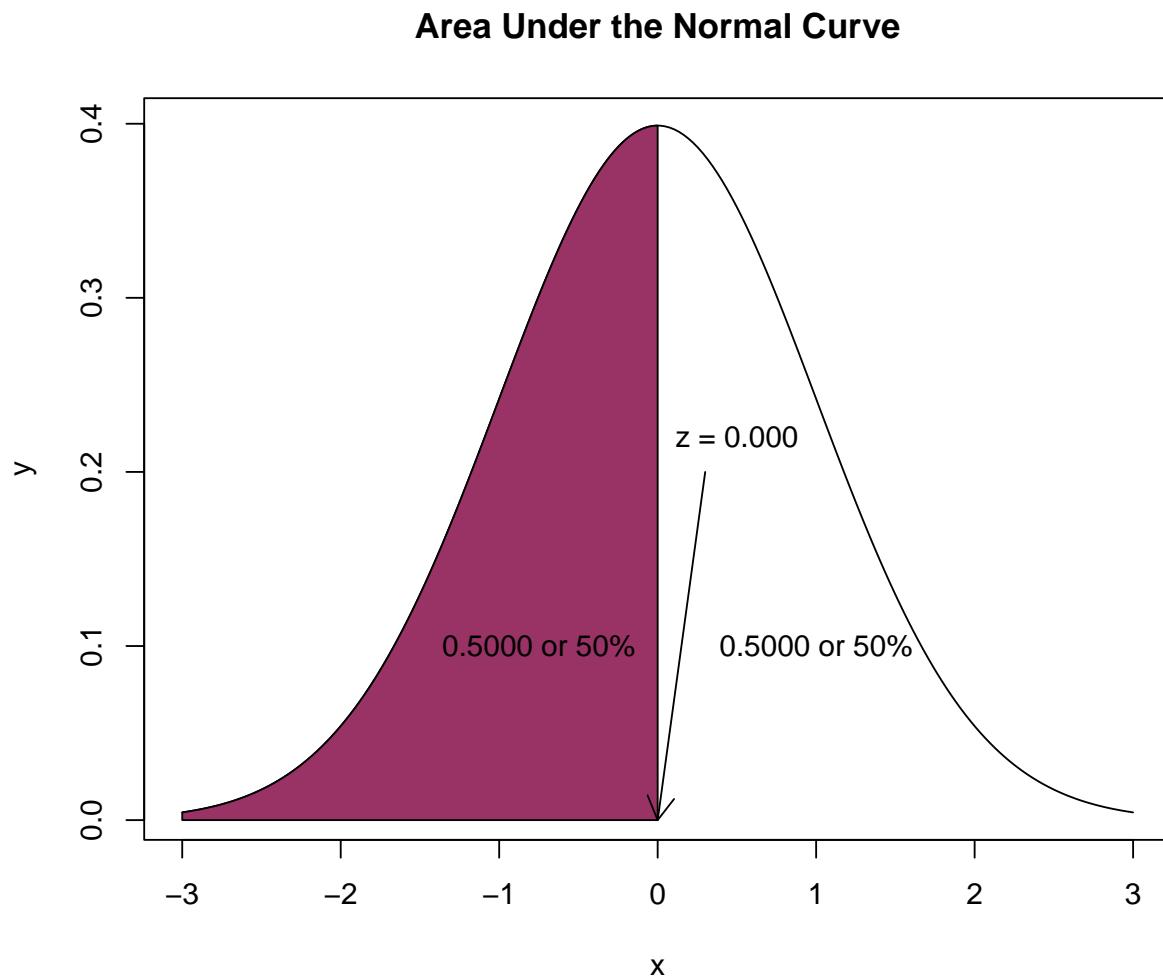
$$z = \frac{X - \bar{X}}{s}$$

```
(9.9 - 9.9)/2 #for 9.9 minutes
```

```
[1] 0
```

We learn that 9.9 minutes (the mean of the distribution of raw scores) corresponds with 0 (the mean of the distribution of *z*-scores).

Next, we examine a [table of critical *z* values](#) where we see that a score of 0.0 corresponds to an area (probability) of .50. The directionality of our table is such that fewer minutes spent with patients are represented on the left (the shaded portion) and more minutes spent with patients are represented on the right (the unshaded portion). Our question asks, what is the probability that a physician spends *at least* 9.9 minutes with a patient (i.e., 9.9 or more minutes) means that we should use the area on the right. Thus, the probability that a physician spends *at least* 9.9 minutes with a patient is 50%. In this case it is also true that the probability that a physician spends 9.9 minutes or less is also 50%. This 50/50 result helps make the point that the area under the curve is equal to 1.0.



We can also obtain the probability value with the *pnorm()* function. We enter the score, the mean, and the standard deviation. As shown below, we can enter them in *z* score formula or from the raw scores.

```
pnorm(0, mean = 0, sd = 1)
```

```
[1] 0.5
```

```
pnorm(9.9, mean = 9.9, sd = 2)
```

```
[1] 0.5
```

Next, let's ask a question that requires careful inspection of the asymmetry of the curve. What is the probability that a physician spends less than 5 minutes with a patient? First, we calculate the corresponding *z*-score:

```
# calculating the z-score
(5 - 9.9)/2 #for 5 minutes
```

```
[1] -2.45
```

Second we locate the corresponding area under the normal curve. Examining the table of critical z -values we see that a z -score of -2.45 corresponds with an area of 0.0071. We can check this with the `pnorm()` function:

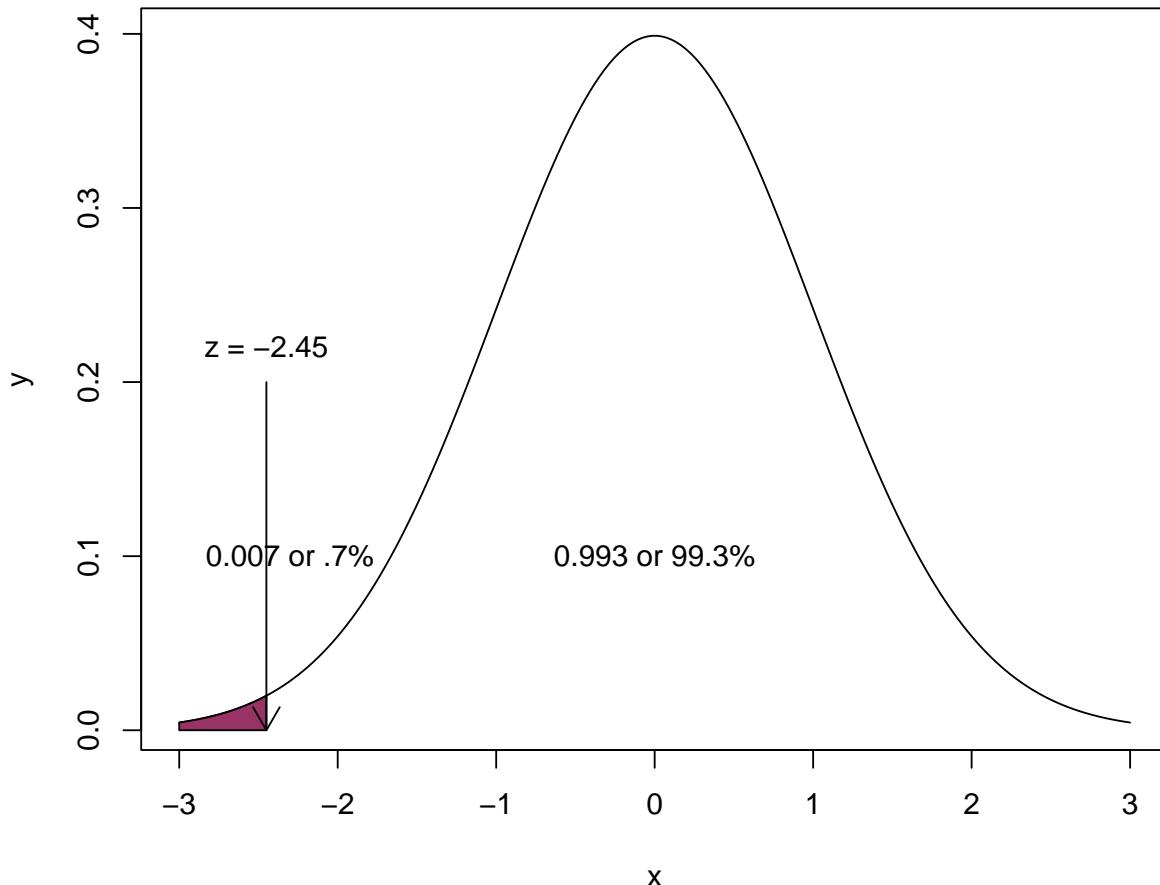
```
pnorm(-2.45, mean = 0, sd = 1) #using SD or standardized units
```

```
[1] 0.007142811
```

```
pnorm(5, mean = 9.9, sd = 2) #using raw data units
```

```
[1] 0.007142811
```

Area Under the Normal Curve



There is a .7% (that is less than 1%) probability that physicians spend less than 5 minutes with

a patient. The inverse ($1 - .7$) indicates that we can be 99% confident that patients receive 5 or more minutes with the ICU physician.

What about operations at the other end of the curve? What is the probability that a patient receives less than 12 minutes with a physician? Again, we start with the calculation of the z -score.

```
(12 - 9.9)/2 #for 12 minutes
```

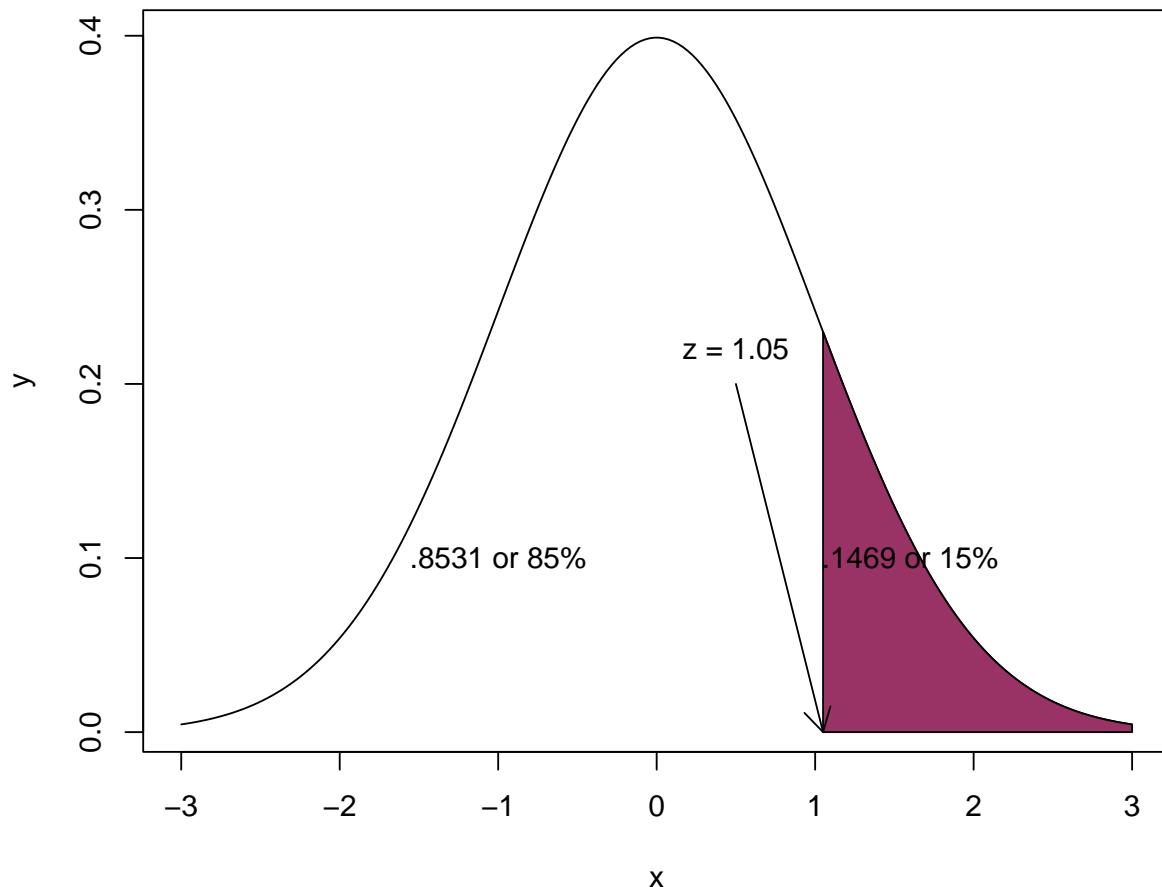
```
[1] 1.05
```

The 12 minute mark is 1.05 SD above the mean. Checking the z table lets us know that an area of 0.8531 corresponds with a z -score of 1.05.

```
1-.8531
```

```
[1] 0.1469
```

Area Under the Normal Curve



The probability of a physician spending 12 minutes *or less* with a patient is 85%; the probability of a physician spending 12 minutes *or more* with a patient is 15%.

4.2.4 Percentiles

The same values that we just collected are often interpreted as percentiles. Our prior calculations taught us that a physician/patient visit that lasted 9.9 minutes ($z = 0$), is ranked at the 50th percentile. That is, a 9.9 minute visit is longer than 50% of patient/physician visits.

A visit lasting 5 minutes ($z = -2.45$) is ranked at the .07th percentile. That is fewer than 1% of patient/physician visits are shorter than 5 minutes.

Finally, a visit lasting 12 minutes ($z = 1.05$) is ranked at the 85th percentile. That is, it is longer than 85% of patient visits.

While this seems redundant, this something of a prelude to the importance of z scores and the standard normal curve in assessment, evaluation, and psychometrics.

4.2.5 Transforming Variables to Standard Scores

At this point, we have hand-calculated each score. It is easy to transform a set of scores into a column of z -scores:

```
PhysTime$zMinutes <- (PhysTime$minutes - mean(PhysTime$minutes))/sd(PhysTime$minutes)

head(PhysTime)
```

	minutes	zMinutes
1	10.300602	0.20226980
2	10.143081	0.12370440
3	9.785452	-0.05466684
4	13.162710	1.62977447
5	6.120944	-1.88237678
6	11.793346	0.94679063

The transformation of scores is considered to be *linear*. That is, this 1:1 relationship would result in a correlation of 1.00. Further, the z -version of the variable could be used in analyses, just as the original raw score. Choices to do this are made carefully and usually done to optimize interpretation. I will demonstrate this with set of descriptive statistics produced by the *apa.cor.table()* function from the *apaTables* package.

```
apaTables::apa.cor.table(PhysTime)
```

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1
1. minutes	9.90	2.00	

```
2. zMinutes 0.00 1.00 1.00**
[1.00, 1.00]
```

Note. M and SD are used to represent mean and standard deviation, respectively.

Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

4.2.6 The One-Sample z test

The one-sample z test is a common entry point to hypothesis testing. Let's imagine that we have reason to believe that an optimal physician/patient interaction in the ICU is 10.5 minutes. We want to use this value as a contrast to our own data and ask if the physician/patient interactions in our ICU are statistically significantly different. To test this hypothesis, we first set up null (H_0) and alternative (H_A) hypotheses. Our null hypothesis states that the population mean for physician/patient visits is equal to 10.5; the alternative hypothesis states that it is unequal to 10.5.

As written, this question is *two-tailed*. That is, the external mean could be larger or smaller, we are just curious to see if it is different.

$$\begin{aligned} H_0 : \mu &= 10.5 \\ H_A : \mu &\neq 10.5 \end{aligned}$$

Alternatively, we could ask a *one-sided* question. That is, we might hypothesize that our sample mean is smaller than the external mean.

$$\begin{aligned} H_0 : \mu &= 10.5 \\ H_A : \mu &< 10.5 \end{aligned}$$

Whether the test is one- or two- sided makes a difference in the strictness with which we interpret the results and can impact whether or not the result is statistically significant. We will reject the H_0 in favor of the alternative (H_A) if the resulting test statistic (a z score) falls into the region of rejection (but that region shifts, depending on whether our test is one- or two- tailed).

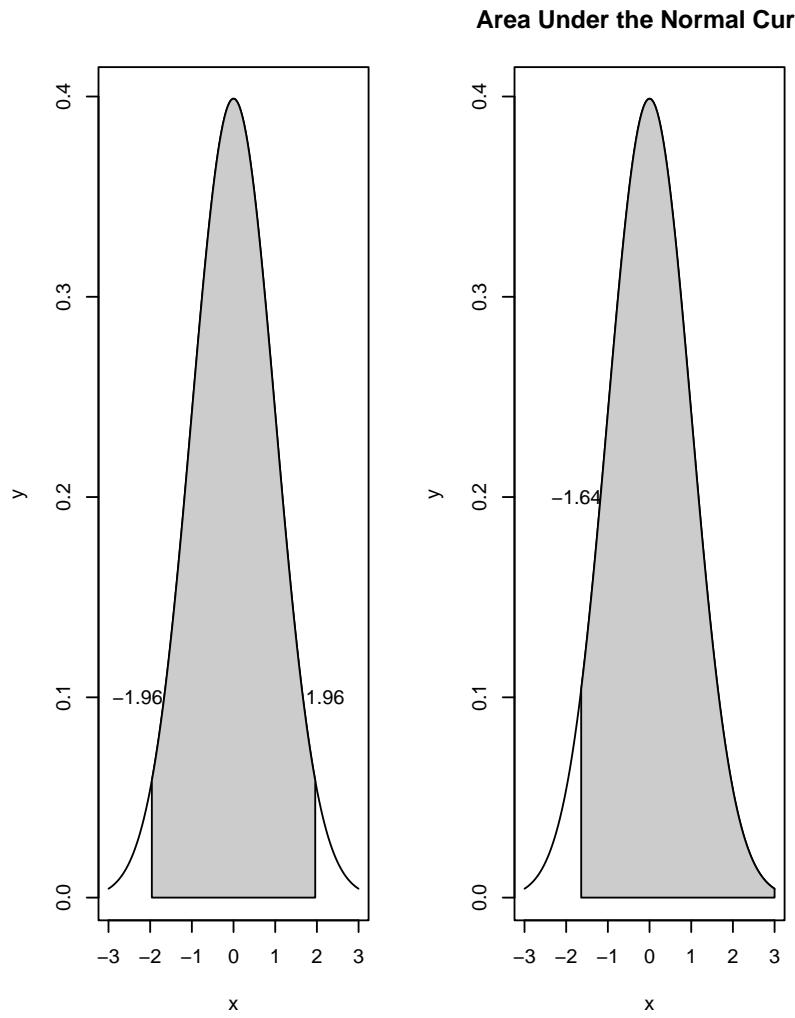
Statistician, Sir Ronald Fisher, popularized 5% as the region of rejection. Specifically, if a probability value associated with a z -score (or similar) falls into the tails of a distribution that represent 5%, then the H_0 is rejected, in favor of the H_A .

Stated another way

- p is the probability that the H_0 is true
 - $p > 0.05$ suggests that there is a 95% chance or greater that the H_0 is true
- 1 minus the p value is the probability that the alternative hypothesis is true.

- A statistically significant test result ($p < 0.05$) means that the test hypothesis is false or should be rejected.
- A p value greater than 0.05 means that no effect was observed.

If our hypothesis is two-sided, then we can spread the 5% across both tails of the test. Inspecting a table of z values shows that ± 1.96 would be the region of rejection of H_0 . In contrast, if the hypothesis is directionless (two-tailed), 1.64 would serve as the boundary for the region of rejection and the corresponding z -test would require the same sign (+ or -) as the hypothesized tail of the distribution. So long as the hypothesis is consistent with the data, a one-sided test can be more powerful, that is, there is greater probability (defined as area under the curve) for rejecting the H_0 , if it is should be rejected.



The formula for a one-sample z -test is as follows:

$$z_{\bar{X}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}}$$

We have already calculated these values. But let's calculate some of them again as a reminder:

```
psych::describe(PhysTime$minutes)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	200	9.9	2	9.98	9.93	2	3.68	15.15	11.47	-0.2	0.03	0.14

- Sample mean is 9.9
- Population mean (the one we're comparing to) is 10.5
- Standard deviation is 2
- N is 200

```
(9.9 - 10.5)/(2/sqrt(200))
```

```
[1] -4.242641
```

The resulting value, $z = -4.242$ is our test value. Because this far exceeds ± 1.96 we know (from memory) that there is a statistically significant effect. Just to be certain, let's use the *pnorm()* function to obtain the *p* value.

```
pnorm(-4.24, mean = 9.9, sd = 2)
```

```
[1] 0.000000000007746685
```

Simply with these hand-calculations, we can claim that there was a statistically significant difference between the physician/patient visit times in our simulated sample data and external benchmark criteria: $z(200) = -4.24, p < .001$.

The one sample *z*-test is rarely sighted in the published literature. However, a close inspection of a table of critical *t*-values, reveals that the very bottom row (i.e., when sample sizes are 120 or greater) is, in fact, the *z* criteria. Thus, it is time to learn about the one sample *t*-test.

4.3 Introducing the One-Sample *t*-test

The one-sample *t*-test is used to evaluate whether the mean of a sample differs from another value that, symbolically, is represented as the population mean. Green and Salkind [2017c] noted that this value is often the midpoint of set of scores, the average value of the test variable based on past research, or a test value as the chance level of performance.

Figure 4.1: An image of a row with two boxes labeled Condition A (in light blue) and the population mean (in dark blue) to which it is being compared. This represents the use of a one-sample *t*-test.

This comparison is evident in the numerator of the formula for the *t*-test that shows the population mean μ being subtracted from the sample mean \bar{X} .

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

Although this statistic is straightforward, it is quite limited. If the researcher wants to compare an outcome variable across two groups of people, they should consider the **independent samples t-test**. If the participant wants to evaluate an outcome variable with two observations from the same group of people, they should consider the **paired samples t-test**

4.3.1 Workflow for the One-Sample *t*-test

The following is a proposed workflow for conducting a one-sample *t*-test.

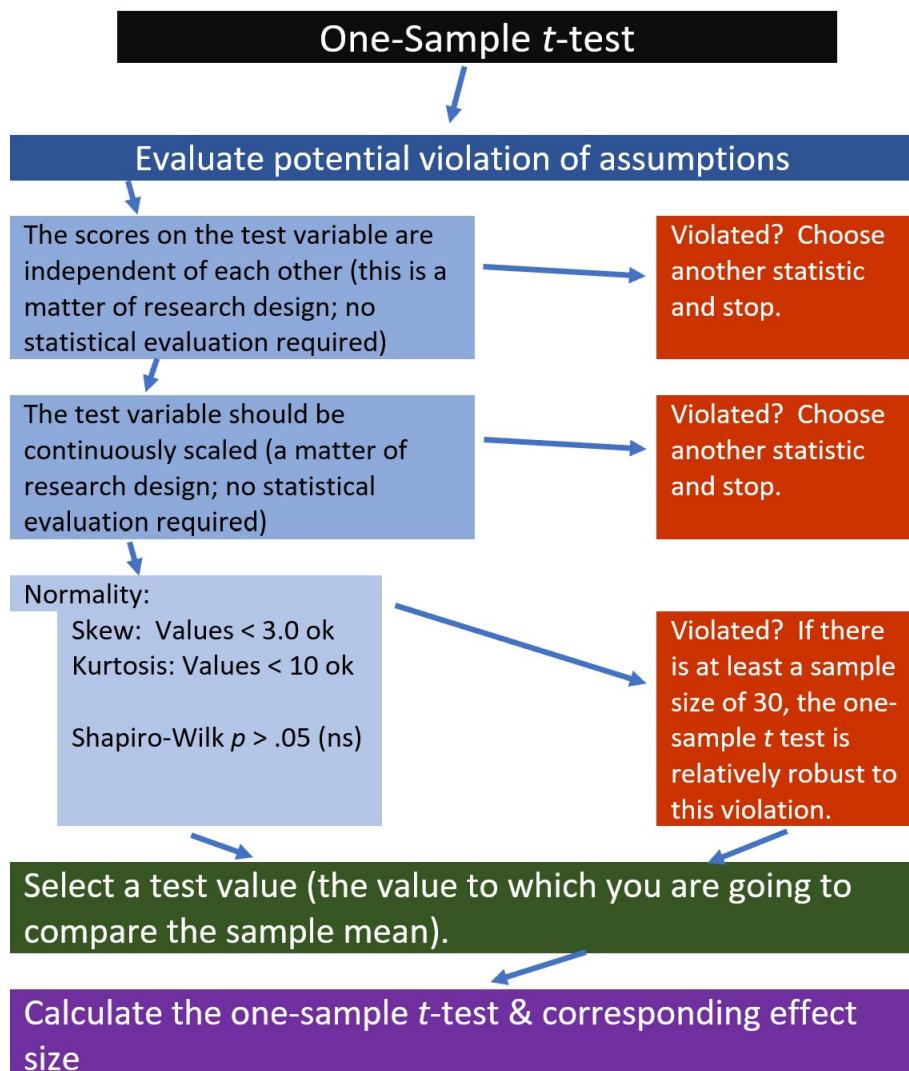


Figure 4.2: A colorful image of a workflow for the one sample *t*-test

If the data meets the assumptions associated with the research design (e.g., independence of observations and a continuously scaled metric), these are the steps for the analysis of a one-sample *t*-test:

1. Prepare (upload) data.
2. Explore data with
 - graphs
 - descriptive statistics
3. Assess normality via skew and kurtosis
4. Select the comparison (i.e., test, population) value
5. Compute the one sample *t*-test
6. Compute an effect size (frequently the *d* statistic)
7. Manage Type I error
8. Sample size/power analysis (which you should think about first, but in the context of teaching statistics, it's more pedagogically sensible, here).

4.4 Research Vignette

Empirically published articles where *t*-tests are the primary statistic are difficult to locate. Having exhausted the psychology archives, I located this article in an interdisciplinary journal focused on palliative medicine. The research vignette for this lesson examined differences in physician's verbal and nonverbal communication with Black and White patients at the end of life [Elliott et al., 2016].

Elliott and colleagues [2016] were curious to know if hospital-based physicians (56% White, 26% Asian, 7.4% each Black and Hispanic) engaged in verbal and nonverbal communication differently with Black and White patients. Black and White patient participants were matched on characteristics deemed important to the researchers (e.g., critically and terminally ill, prognostically similar, expressed similar treatment preferences). Interactions in the intensive care unit were audio and video recorded and then coded on dimensions of verbal and nonverbal communication.

Because each physician saw a pair of patients (i.e., one Black patient and one White patient), the researchers utilized a paired samples, or dependent *t*-test. This statistical choice was consistent with the element of the research design that controlled for physician effects through matching (and one we will work in a later lesson). Below are the primary findings of the study.

	Black Patients	White Patients	
Category	<i>Mean(SD)</i>	<i>Mean(SD)</i>	<i>p</i> -value
Verbal skill score (range 0 - 27)	8.37(3.36)	8.41(3.21)	0.958
Nonverbal skill score (range 0 - 5)	2.68(.84)	2.93(.77)	0.014

In the research vignette Elliott et al. [2016] indicated that physician/patient visits lasted between 3 minutes and 40 seconds to 20 minutes and 13 seconds. For the purpose of demonstrating the one sample *t*-test, we might want to ask whether the length of patient visits in this research study were statistically significantly different than patient in the ICU or in palliative care, more broadly. Elliott et al.[2016] did not indicate a measure of central tendency (i.e., mean, mode, median) therefore,

I will simulate the data by randomly generating 33 numbers with a mean of 8 and a standard deviation of 2.5. I will use *random selection with replacement*, which allows the same number to be selected more than once.

4.4.1 Data Simulation

I re-simulated (what may seem like identical data from above) to be consistent with the journal article's research sample of 33.

```
# Setting the 'random' seed ensures that everyone gets the same
# result, every time they rerun the analysis. My personal practice is
# to create a random seed that represents the day I write up the
# problem (in this case August, 15, 2022) When the Suggestions for
# Practice invite you to 'change the random seed,' simply change this
# number to anything you like (maybe your birthday or today's date)
set.seed(220822)
dfOneSample <- data.frame(PhysMins = rnorm(33, mean = 10, sd = 2.5))

head(dfOneSample)
```

	PhysMins
1	9.097343
2	11.385558
3	8.424395
4	8.640534
5	12.583856
6	8.949883

A warning: this particularly analysis (the whole lesson, in fact) is “more simulated than usual” and does not represent reality. However, this research vignette lends itself for this type of question.

With our data in hand, let's examine its structure. The variable representing physician minutes represents the ratio scale of measurement and therefore should be noted as *num* (numerical) in R.

```
str(dfOneSample)
```

```
'data.frame': 33 obs. of 1 variable:
 $ PhysMins: num 9.1 11.39 8.42 8.64 12.58 ...
```

Below is code for saving the data to your computer (and then re-importing) as .csv or .rds files. I make choices about saving data based on what I wish to do with the data. If I want to manipulate the data outside of R, I will save it as a .csv file. It is easy to open .csv files in Excel. A limitation of the .csv format is that it does not save any restructuring or reformatting of variables. For this lesson, this is not an issue.

Although you do not need to save nor re-import the data for this lesson, here is code for saving the data as a .csv and then reading it back into R. I have hashtags these out, so you will need to remove the hashtags if you wish to run any of these operations.

```
# writing the simulated data as a .csv write.table(dfOneSample, file
# = 'dfOneSample.csv', sep = ',', col.names=TRUE, row.names=FALSE) at
# this point you could clear your environment and then bring the data
# back in as a .csv reading the data back in as a .csv file
# dfOneSample<- read.csv ('dfOneSample.csv', header = TRUE)
```

The .rds form of saving variables preserves any formatting (e.g., creating ordered factors) of the data. A limitation is that these files are not easily opened in Excel. Again, you do not need to save nor re-import the data for this lesson. However, if you would like to do so, here is the hashtags code (remove hashtags if you wish to do this) for writing (and then reading) this data as an .rds file.

```
# saveRDS(dfOneSample, 'dfOneSample.rds') dfOneSample <-
# readRDS('dfOneSample.rds')
```

4.4.2 Quick Peek at the Data

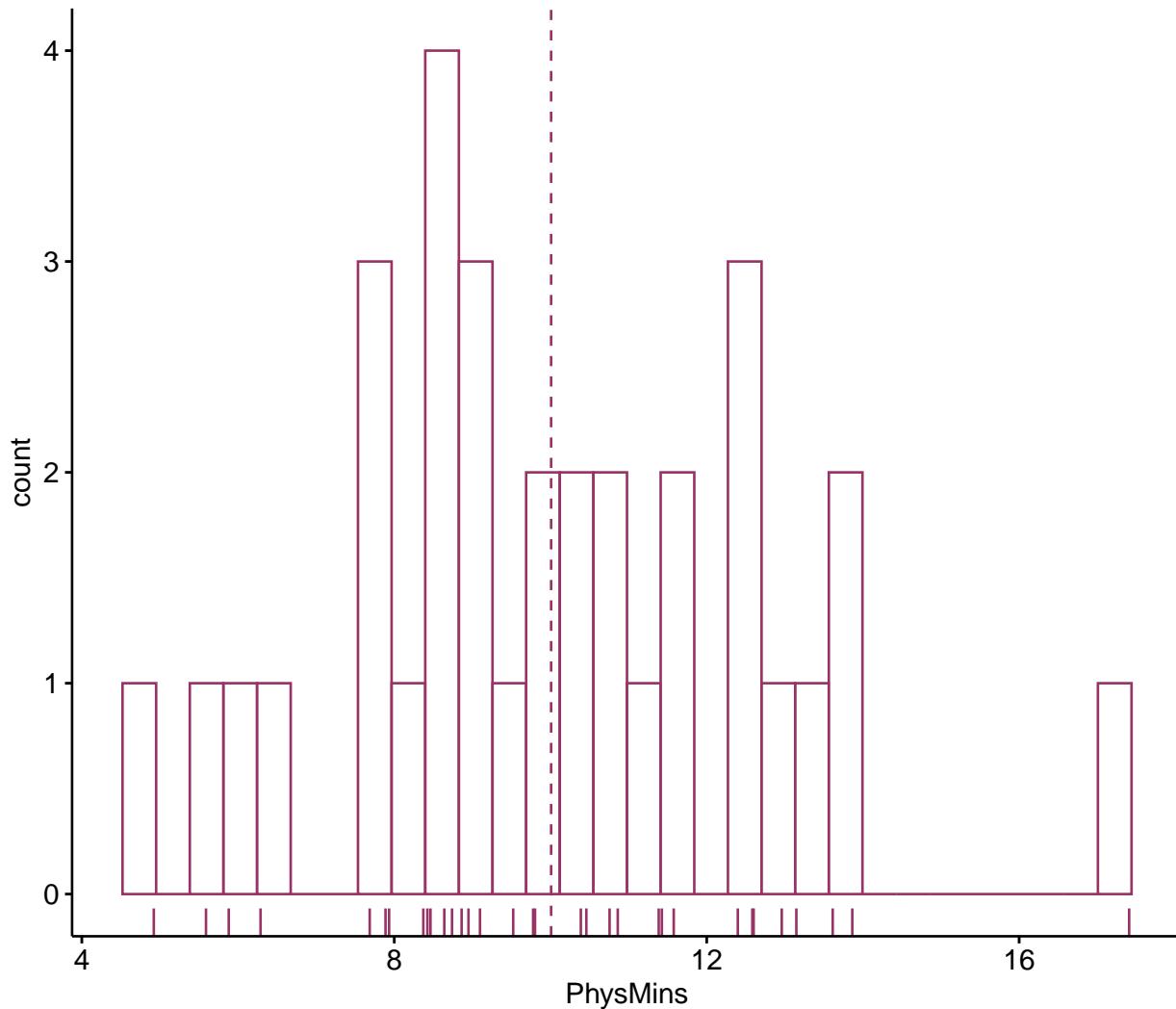
Plotting the data is best practice to any data analysis. Further, visualizing the data can help us with a conceptual notion of the statistic we are utilizing. The *ggpubr* package is one of my go-to-tools for quick and easy plots of data. Below, I have plotted the time-with-patient (Physician Seconds) variable and added the mean. As with most plotting packages, *ggpubr* will “bin” (or cluster) the data for plotting; this is especially true for data with a large number of units (a range from 220 to 1213 is quite large). The “rug = TRUE” command added a lower row of the table to identify where each of the datapoint follows.

```
ggpubr::gghistogram(dfOneSample, x = "PhysMins", add = "mean", rug = TRUE,
color = "#993366")
```

Warning: Using `bins = 30` by default. Pick better value with the argument `bins`.

Warning: `geom_vline()`¹: Ignoring `mapping` because `xintercept` was provided.

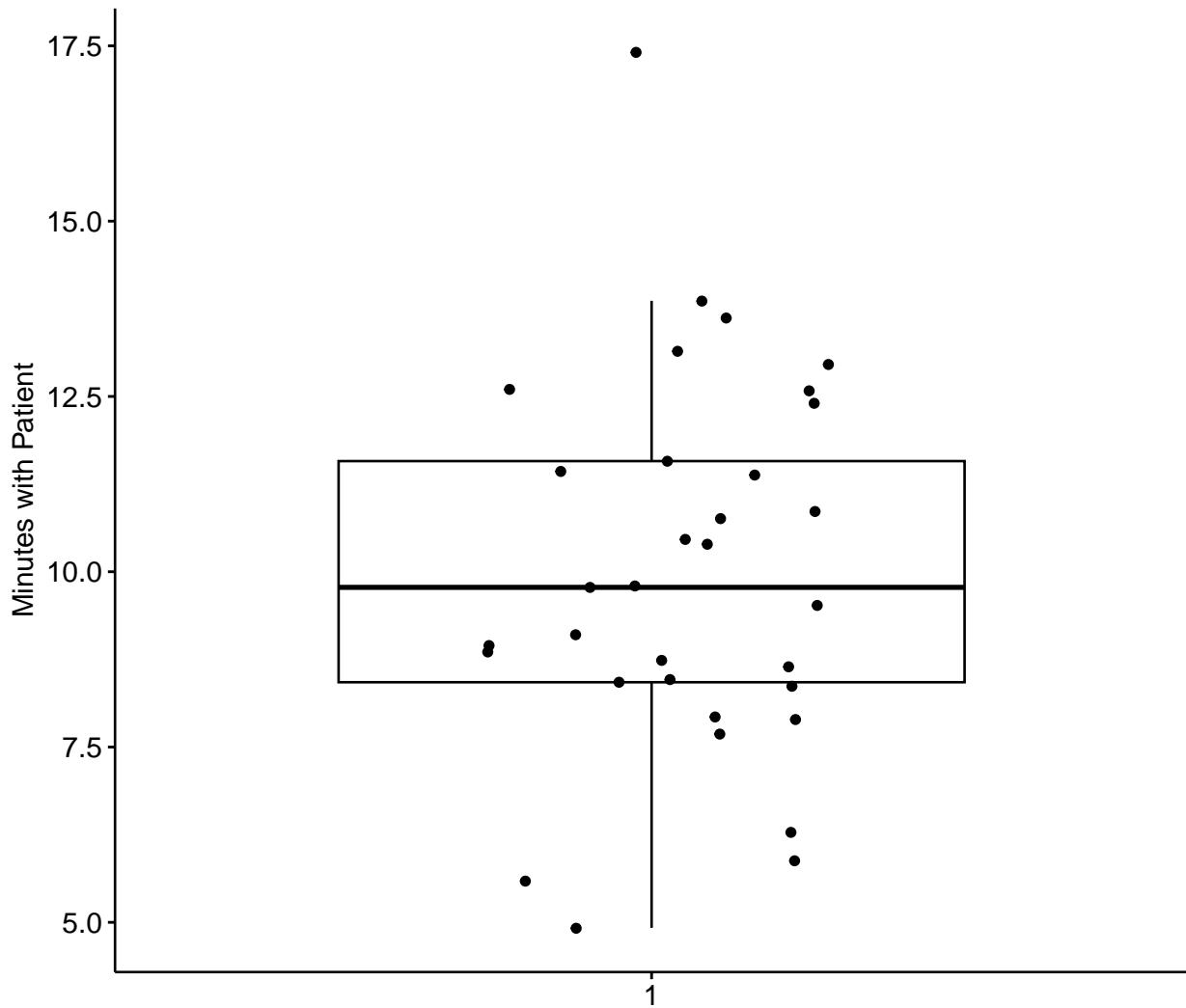
Warning: `geom_vline()`¹: Ignoring `data` because `xintercept` was provided.



Although the histogram is not perfectly normal, we can see at least the suggestion of a normal distribution. With only a sample of 33, I'm encouraged.

Another view of our data is with a boxplot. The box captures the middle 50% of data with the horizontal bar at the median. The whiskers extend three standard deviations around the mean with dots beyond the whiskers representing outliers. I personally like the `add = "jitter"` statement because it shows where each case falls.

```
ggpubr::ggboxplot(dfOneSample$PhysMins, ylab = "Minutes with Patient",
  xlab = FALSE, add = "jitter")
```



We can further evaluate normality by obtaining the descriptive statistics with the *describe()* function from the *psych* package.

```
psych::describe(dfOneSample$PhysMins)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	
X1	1	33	10.01	2.7	9.78	9.96	2.44	4.92	17.41	12.49	0.36	0.36	0.04	0.47

Here we see that our minutes range from 4.92 to 17.41 with a mean of 10.01 and a standard deviation of 2.7. We're ready to calculate the one sample *t*-test.

4.5 Working the One Sample *t*-test (by hand)

4.5.1 Stating the Hypothesis

A quick scan of the literature suggests that health care workers' visits to patients in the ICU are typically quite brief. Specifically, the average duration of a physician visit in a 2018 study was 73.5

seconds or 1.23 minutes [Butler et al., 2018]. A one-sample t -test is appropriate for comparing the visit lengths from our sample to this external metric.

As noted in the symbolic presentation below, our null hypothesis (H_0) states that our data will be equal to the test value of 1.23 minutes. In contrast, the alternative hypothesis (H_A) states that these values will not be equal.

$$\begin{aligned} H_0 &: \mu = 1.23 \\ H_A &: \mu \neq 1.23 \end{aligned}$$

4.5.2 Calculating the t -test

In learning the statistic, hand-calculations can help understand what the statistic is doing. Here's the formula again:

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

The numerator of the formula below subtracts the test value from the sample mean. The denominator involves multiplying the standard deviation by the square root of the sample size. The descriptive statistics provided the values we need to complete the analysis:

```
(10.01 - 1.23)/(2.7/sqrt(33))
```

```
[1] 18.68047
```

4.5.2.1 Statistical Significance

If we ask about *statistical significance* then we are engaged in *null hypothesis significance testing* (NHST). In the case of a one sample test, we construct our hypothesis with a null and an alternative that are relatively straightforward. Specifically, we are interested in knowing if our sample mean (10.01) is statistically, significantly different from the test value of 1.23. We can write the hypotheses in this way:

$$\begin{aligned} H_0 &: \mu = 1.23 \\ H_A &: \mu \neq 1.23 \end{aligned}$$

In two parts, our null hypothesis (H_0) states that the population mean (H_0) for physician visits with palliative care patients is 1.23; the alternative $\mu \neq$ states that it is not 1.23.

When we calculated the t -test, we obtained a t value. We can check the statistical significance by determining the test critical value from a [table of critical values](#) for the t distribution. There are many freely available on the internet. If our t value exceeds the value(s) in the table of critical values, then we can claim that our sample mean is statistically significantly different from the hypothesized value.

Heading to the table of critical values we do the following:

- For the one-sample t -test, the degrees of freedom (DF) is equal to $N - 1$ (32). The closest value in our table is 30, so we will use that row.
- A priorily, we did not specify if we thought the difference would be greater, or lower. Therefore, we will use a column that indicates *two-tails*.
- A p value of .05 is customary (but it will be split between two tails).
- Thus, if our t -value is lower than -2.042 or higher than 2.042 we know we have a statistically significant difference.

In our case, the t value of 18.68 far exceeded the test critical value of 2.042. We would write the statistical string this way: $t(32) = 18.68, p < .05$.

In base R, the $qt()$ function will look up a test critical value. For the one-sample t -test, degrees of freedom (df) is equal to $N - 1$. We “divide the p value by 2” when we want a two-tailed test. Finally, the “lower.tail” command results in positive or negative values in the tail.

```
qt(p = 0.05/2, df = 32, lower.tail = FALSE)
```

```
[1] 2.036933
```

Not surprisingly, this value is quite similar to the value we saw in the table. The $qt()$ function is more accurate because it used $df = 32$ (not rounded down to 30).

4.5.2.2 Confidence Intervals

How confident are we in our result? With the one sample t -test, it is common to report an interval in which we are 95% confident that that our sample mean exists. Below is the formula, which involves:

- \bar{X} is the sample mean; in our case this is 10.01
- t_{cv} the test critical value for a two-tailed model (even if the hypothesis was one-tailed) where $\alpha = .05$ and the degrees of freedom are $N - 1$
- $\frac{s}{\sqrt{n}}$ was the denominator of the test statistic it involves the standard deviation of our sample (2.7) and the square root of our sample size (33)

$$\bar{X} \pm t_{cv} \left(\frac{s}{\sqrt{n}} \right)$$

Let's calculate it:

First, let's calculate the proper t critical value. Even though these are identical to the one above, I am including them again. Why? Because if the original hypothesis had been one-tailed, we would need to calculate a two-tailed confidence interval; this is a placeholder to remind us.

```
qt(p = 0.05/2, df = 32, lower.tail = FALSE)
```

```
[1] 2.036933
```

Using the values from above, we can specify both the lower and upper bound of our confidence interval.

```
(10.01) - ((2.0369) * (2.7/sqrt(33)))
```

```
[1] 9.052637
```

```
(10.01) + ((2.0369) * (2.7/sqrt(33)))
```

```
[1] 10.96736
```

The resulting interval is the 95% confidence interval around our sample mean. Stated another way, we are 95% certain that the true mean of time with patients in our sample ranges between 9.05 and 10.97 minutes.

4.5.2.3 Effect size

If you have heard someone say something like, “I see there is statistical significance, but is the difference *clinically significant*,” the person is probably asking about *effect sizes*. Effect sizes provide an indication of the magnitude of the difference.

The d statistic is commonly used with t -tests; d assesses the degree that the mean on the test variable differs from the test value. Conveniently, d represents standard deviation units. A d value of 0 indicates that the mean of the sample equals the mean of the test value. As d moves away from 0 (in either direction), we can interpret the effect size to be stronger. Conventionally, the absolute values of .2, .5, and .8, represent small, medium, and large effect sizes, respectfully.

Calculating the d statistic is easy. Here are two equivalent formulas:

$$d = \frac{\text{MeanDifference}}{SD} = \frac{t}{\sqrt{N}}$$

```
# First formula
(10.01 - 1.23)/2.7
```

```
[1] 3.251852
```

```
# Second formula
18.68047/sqrt(33)
```

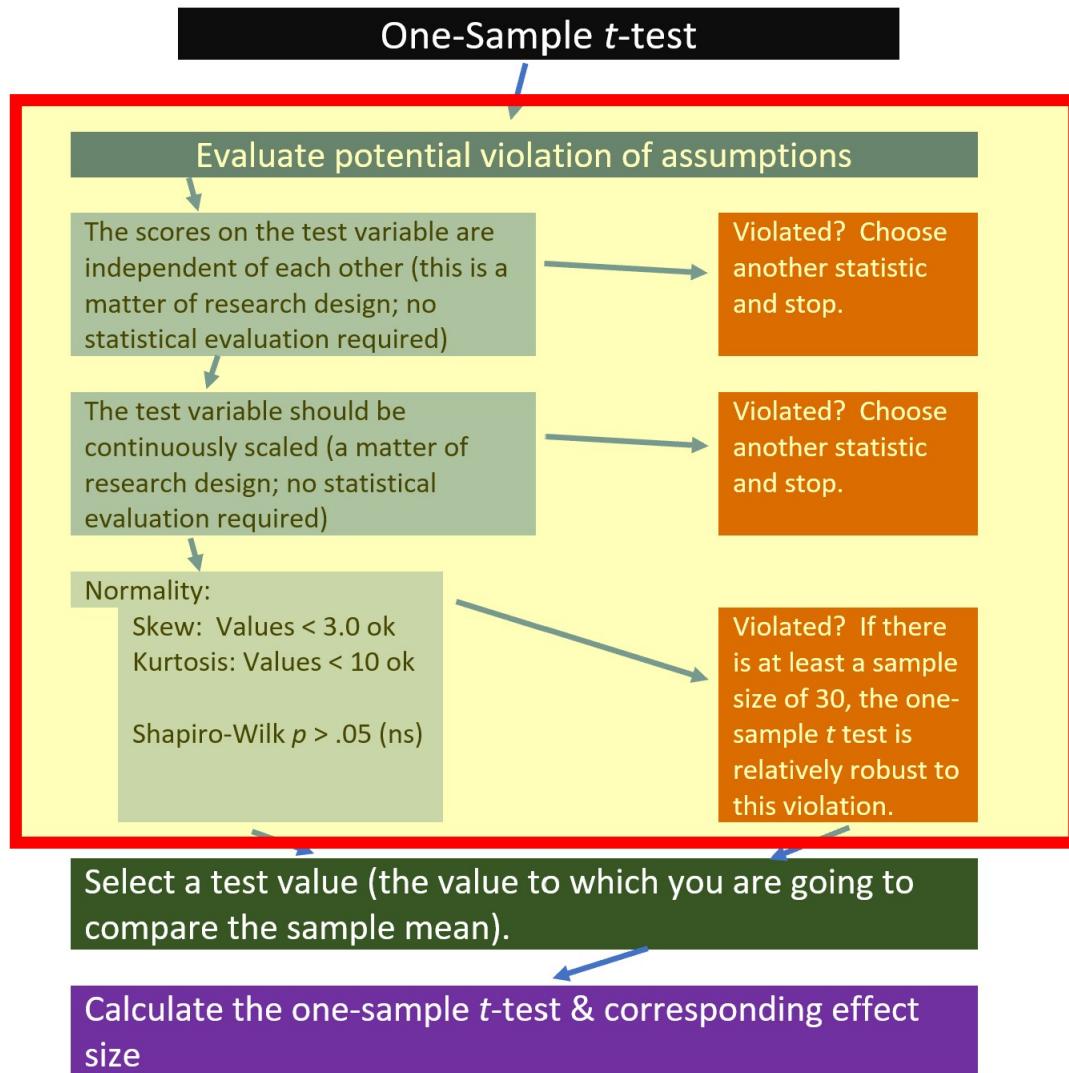
```
[1] 3.251852
```

The value of 3.25 indicates that the test value is approximately more than three standard deviations away from the sample mean. This is a very large difference.

4.6 Working the One-Sample *t*-test with R Packages

4.6.1 Evaluating the Statistical Assumptions

Let's rework the problem in R. We start at the top of the flowchart, evaluating the statistical assumptions.



All statistical tests have some assumptions about the data. The one-sample *t*-test has three.

- The scores on the test variable as independent of each other. This is a research design issue and the one-sample *t*-test is not robust to violating this assumption.
 - If physicians were contributing more than one data point, this vignette potentially violated this assumption. For the sake of simplicity, let's presume that each physician contributed visit length (minutes) for only one patient. If the research scenario was

such that physicians contributed multiple datapoints a potential analytic choice that is robust to such a violation is [multilevel modeling](#).

- The test variable should be continuously scaled. This is also a matter of research design and no statistical analysis is required.
 - Our test variable is measured in minutes; this is continuously scaled and has the properties of *ratio*-level data.
- The test variable is normally distributed. We can check this several ways:
 - visually with histograms (perhaps with superimposed curves) and boxplots,
 - calculation of skew and kurtosis values,
 - calculation of the Shapiro-Wilk test of normality

4.6.1.1 Is the Test Variable Normally Distributed?

Thus, we need only to assess whether the test variable is normally distributed. The `pastecs::stat.desc()` function will provide all of this information in one test. We need only add the specification, “norm=TRUE”.

```
# pastecs is the package, stat.desc is the function we point it to
# the data and then add the norm=TRUE command
pastecs::stat.desc(dfOneSample, norm = TRUE)
```

	PhysMins
nbr.val	33.0000000
nbr.null	0.0000000
nbr.na	0.0000000
min	4.92123791
max	17.40834882
range	12.48711091
sum	330.26971365
median	9.77737813
mean	10.00817314
SE.mean	0.47011645
CI.mean.0.95	0.95759588
var	7.29331287
std.dev	2.70061342
coef.var	0.26984080
skewness	0.35985466
skew.2SE	0.44031259
kurtosis	0.03511647
kurt.2SE	0.02199140
normtest.W	0.97666915
normtest.p	0.68198838

Recall from the lesson on [Preliminary Results](#) that there are multiple ways to assess severity of skew and kurtosis. Values greater than the absolute value of 3.0 are concerning for the “skewness”

output. The PhysMins skewness values of 0.36 is well below that threshold. Values greater than the absolute value of 10 are concerning for the “kurtosis” output. The PhysMins skewness value of 0.035 is well below that threshold. The “skew.2SE” and “kurt.2SE” values are standardized. The “2” in the “skew.2SE” is a helpful reminder that, in smaller sample sizes”, using the 1.96 (or “2”) criteria is acceptable in determining problematic skew or kurtosis. The PhysMins values of 0.44 and 0.022 fall well below those areas of concern.

Regarding a formal assessment of normality, the `pastecs::stat.descr()` output includes the Shapiro-Wilk value (`normtest.W`) and statistical significance (`normtest.p`). Non-significant results indicate that the distribution of the PhysMins variable is not statistically significantly different from a normal distribution. In the case of PhysMins, $W = 0.977, p = 0.682$.

Considering skewness, kurtosis, and normality estimates together, we are confident that we have not violated the assumption of normality.

4.6.2 Computing the *t*-test

Now we are ready to calculate the *t*-test, itself.

Calculating a one sample *t*-test is possible through base R and a number of packages. Kassambara’s [b] `rstatix` package is one we can use for the *t*-test and ANOVA problems that we will work. I like it for several reasons. First, it was designed to be “pipe-friendly” in a manner that is consistent with the `tidyverse` approach to working in R and there are numerous tutorials. Additionally, `rstatix` objects work well with `ggpubr`, one of my favorite packages for graphing data and results.

In the script below:

- the first element points to the dataframe
- the second element provides a “formula”
 - we are predicting “PhysMins” from “1” which represent an invariant/constant hypothesized mean
- the third element identifies the population/comparison mean
- specifying “`detailed = TRUE`” will produce the 95% confidence interval around the mean (i.e., in this case the average amount of time that physicians in our sample spent with their patients)

```
rstatix::t_test(dfOneSample, PhysMins ~ 1, mu = 1.23, detailed = TRUE)
```

```
# A tibble: 1 x 12
  estimate .y. group1 group2     n statistic       p    df conf.low conf.high
*   <dbl> <chr> <chr>  <chr> <int>    <dbl> <dbl> <dbl>    <dbl>
1     10.0 Phys~ 1     null ~    33     18.7 9.07e-19    32     9.05     11.0
# i 2 more variables: method <chr>, alternative <chr>
```

The results we obtained are identical to those we hand-calculated. The `rstatix` output also includes confidence intervals. In the case of the one-sample *t*-test, this represent the 95% confidence interval around the mean. That is, we are 95% confident that the true mean of the minutes that physicians

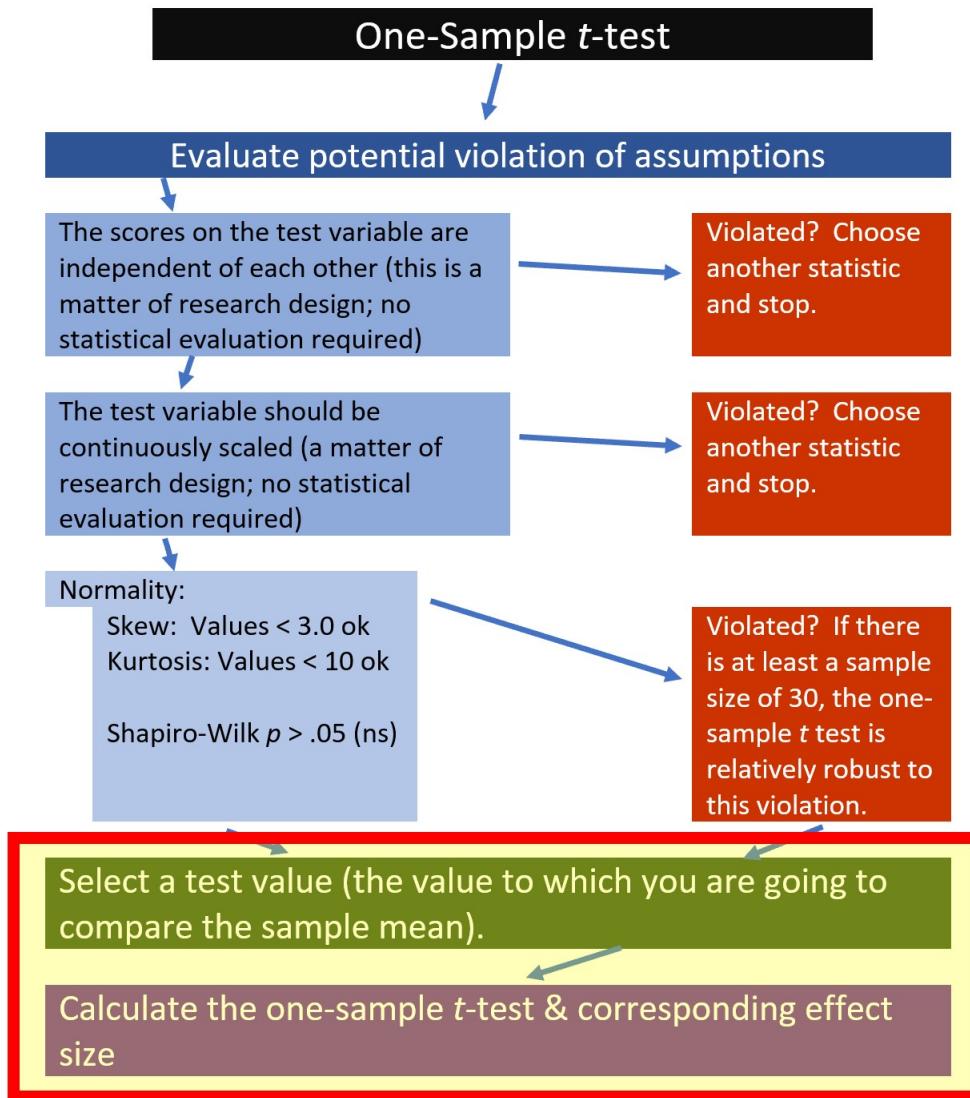


Figure 4.3: The workflow for the one sample t -test highlighting the evaluation of assumptions section

in our sample spent with patients falls between 9.05 and 10.97. I appreciate that the *rstatix* output reminds us that we are using a *t*-test and that it is a two-sided hypothesis.

Knowing what the confidence interval is “around” can be tricky. Whatever the “topic” of the confidence interval will be exactly in the middle of (most) confidence intervals. We can check ourselves by adding the two ends of the confidence interval and dividing by two.

```
(9.050577 + 10.96577)/2
```

```
[1] 10.00817
```

As we see, 10.008 is the reported as the “estimate.” We know from our earlier analysis of the descriptive statistics that this is the value of the mean. If we are uncertain, we can check:

```
mean(dfOneSample$PhysMins)
```

```
[1] 10.00817
```

From these results, we can begin to create our *t* string: $t(32) = 18.67, p < .001, CI95(9.05, 10.97)$

With a separate command, we can use the *rstatix* package to obtain the effect size, *d*. With the exception of including the “ref.group = NULL” statement, the script is quite similar.

```
rstatix::cohens_d(dfOneSample, PhysMins ~ 1, ref.group = NULL, mu = 1.23)
```

```
# A tibble: 1 x 6
  .y.      group1 group2     effsize     n magnitude
* <chr>    <chr>  <chr>     <dbl> <int> <ord>
1 PhysMins 1       null model     3.25     33 large
```

From these results, we can begin to complete our *t* string: $t(32) = 18.672, p < .001, CI95(9.05, 10.97), d = 3.25$

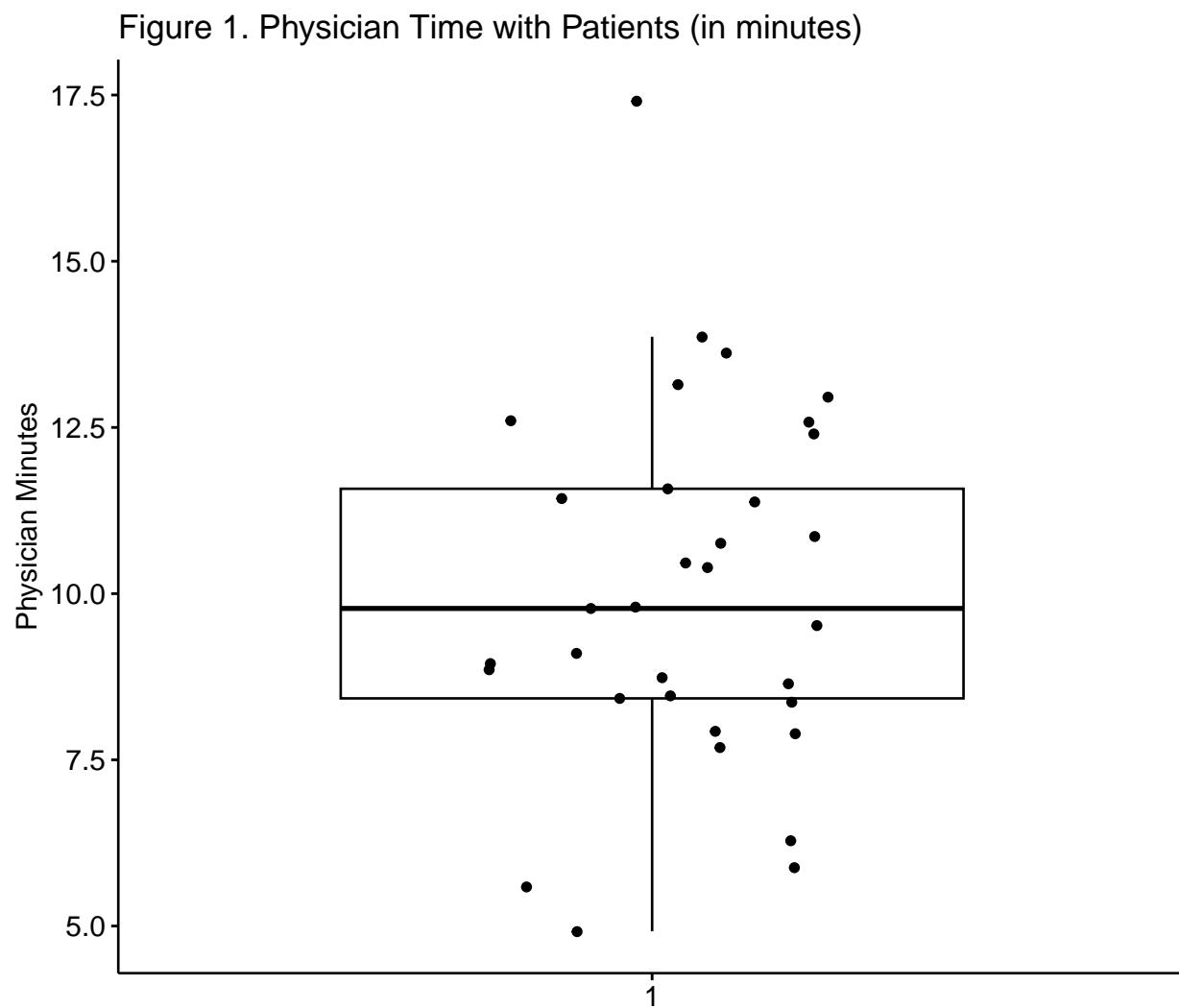
4.7 APA Style Results

Let’s write up the results. In-so-doing, I would include the boxplot we produced during our initial exploration of the data.

Preliminary inspection of the data indicated that we did not violate the assumption of normality. Specifically, our skew (0.36) and kurtosis (0.035) values fell below that absolute values (3.0, 10.0, respectively) that are concerning [Kline, 2016a]. Further, the Shapiro-Wilk test of normality suggested that the distribution of our sample data did not differ significantly from a normal distribution ($W = 0.977, p = 0.682$)

A one-sample *t*-test was used to evaluate whether average amount of time that a sample of physicians (palliative care physicians in the ICU) enrolled in a research study on patient communication was statistically significantly different from the amount of time that ICU physicians spend with their patients, in general. The sample mean 10.008 ($SD = 2.7016$) was significantly different from 1.23, $t(32) = 18.672, p < .001$, $CI_{95}(9.05, 10.97)$, $d = 3.25$. The effect size, (d) indicates a very large effect. Figure 1 illustrates the distribution of time physicians in the research study spent with their patients. The results support the conclusion that physicians in the research study spent more time with their patients than ICU physicians in general.

```
ggpubr::ggboxplot(dfOneSample$PhysMins, ylab = "Physician Minutes", xlab = FALSE,
add = "jitter", title = "Figure 1. Physician Time with Patients (in minutes)")
```



Reflecting on these results, I must remind readers that this simulated data that is even further extrapolated. Although “data” informed both the amount of time spent by the physicians in the research study and data used as the test value, there are probably many reasons that the test value was not a good choice. For example, even though both contexts were ICU, palliative physicians may have a different standard of care than ICU physicians “in general.”

4.8 Power in One-Sample *t*-tests

Researchers often use power analysis packages to estimate the sample size needed to detect a statistically significant effect, if, in fact, there is one. Utilized another way, these tools allows us to determine the probability of detecting an effect of a given size with a given level of confidence. If the probability is unacceptably low, we may want to revise or stop. A helpful overview of power as well as guidelines for how to use the *pwr* package can be found at a [Quick-R website \[Kabacoff, 2017\]](#).

In Champely's *pwr* package, we can conduct a power analysis for a variety of designs, including the one sample *t*-test that we worked in this lesson. There are a number of interrelating elements of power:

- Sample size, *n* refers to the number of observations; our vignette had 33
- *d* refers to the difference between means divided by the pooled standard deviation; ours was $(10.01 - 1.23)/2.7$; we can use the results from Cohen's *d*.
- *power* refers to the power of a statistical test; conventionally it is set at .80
- *sig.level* refers to our desired alpha level; conventionally it is set at .05
- *type* indicates the type of test we ran; this was "one.sample"
- *alternative* refers to whether the hypothesis is non-directional/two-tailed ("two.sided") or directional/one-tailed("less" or "greater")

In this script, we must specify *all-but-one* parameter; the remaining parameter must be defined as *NULL*. R will calculate the value for the missing parameter.

When we conduct a "power analysis" (i.e., the likelihood of a hypothesis test detecting an effect if there is one), we specify, "power=NULL". Using the data from our results, we learn from this first run, that our statistical power was 1.00. That is, given the value of the mean difference relative to the pooled standard deviation we had a 100% chance of detecting a statistically significant effect if there was one.

```
pwr::pwr.t.test(d = 3.25, n = 33, power = NULL, sig.level = 0.05, type = "one.sample",
                  alternative = "two.sided")
```

```
One-sample t test power calculation
```

```
n = 33
d = 3.25
sig.level = 0.05
power = 1
alternative = two.sided
```

Researchers frequently use these tools to estimate the sample size required to obtain a statistically significant effect. In these scenarios we set *n* to *NULL*.

```
pwr::pwr.t.test(d = 3.25, n = NULL, power = 0.8, sig.level = 0.05, type = "one.sample",
  alternative = "two.sided")
```

```
One-sample t test power calculation
```

```
n = 3.006908
d = 3.25
sig.level = 0.05
power = 0.8
alternative = two.sided
```

Shockingly, this suggests that a sample size of 3 could result in a statistically significant result. Let's see if this is true. Below I will re-simulate the data for the verbal scores, changing only the sample size:

```
set.seed(220822)
rdfOneSample <- data.frame(rPhysMins = rnorm(3, mean = 10, sd = 2.5))

head(rdfOneSample)
```

```
rPhysMins
1 9.097343
2 11.385558
3 8.424395
```

With the newly simulated data, I will run the one-sample *t*-test:

```
rstatix::t_test(rdfOneSample, rPhysMins ~ 1, mu = 1.23, detailed = TRUE)

# A tibble: 1 x 12
  estimate .y.    group1 group2      n statistic      p     df conf.low conf.high
*   <dbl> <chr>  <chr>  <chr>  <int>    <dbl>  <dbl> <dbl>    <dbl>    <dbl>
1     9.64 rPhysM~ 1       null ~     3      9.38 0.0112     2      5.78     13.5
# i 2 more variables: method <chr>, alternative <chr>

rstatix::cohens_d(rdfOneSample, rPhysMins ~ 1, ref.group = NULL, mu = 1.23)

# A tibble: 1 x 6
  .y.    group1 group2      effsize      n magnitude
* <chr>  <chr>  <chr>      <dbl> <int> <ord>
1 rPhysMins 1       null model     5.42      3 large
```

In this case our difference between the sample data and the external data is so huge, that a sample of three still nets a statistically significant result. This is unusual. Here's the *t* string: $t(2) = 9.379, p = 0.011, d = 5.415, CI95[5.78, 13.492]$.

4.9 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “doable” and then try at least one more problem that challenges you in some way. Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Using R packages, complete a one-sample t -test.

Additionally, please complete at least one set of *hand calculations*, that is using the code demonstrated in the chapter to work through the formulas that compute the one-sample t -test. At this stage in your learning, you may ignore any missingness in your dataset by excluding all rows with missing data in your variables of interest.

4.9.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation of the research vignette, then rework the problem. This should provide minor changes to the data but the results will likely be very similar. That said, don’t be alarmed if what was non-significant in my working of the problem becomes significant. Our selection of $p < .05$ (and the corresponding 95% confidence interval) means that 5% of the time there could be a difference in statistical significance.

4.9.2 Problem #2: Rework the research vignette, but change something about the simulation

Rework the one sample t -test in the lesson by changing something else about the simulation. Perhaps estimate another comparative number. The 1.23 was a dramatic difference from the mean of the research participants. Perhaps suggest (and, ideally, support with a reference) a different value. Alternatively, if you are interested in issues of power, specify a different sample size.

4.9.3 Problem #3: Use other data that is available to you

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete an independent samples t test.

4.9.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Working the problem with R and R packages	Points Possible	Points Earned
1. Narrate the research vignette, describing the variables and their role in the analysis	5	_____
2. Simulate (or import) and format data	5	_____

Working the problem with R and R packages	Points Possible	Points Earned
3. Evaluate statistical assumptions	5	_____
4. Conduct a one sample t -test (with an effect size)	5	_____
5. APA style results with table(s) and figure	5	_____
6. Conduct power analyses to determine the power of the current study and a recommended sample size	5	_____
7. Explanation to grader	5	_____
Totals	35	_____

Hand Calculations	Points Poss	Points Earned
1. Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	2	
2. Calculate the mean of your sample; identify the mean of your benchmarking sample	2	
3. Using the steps from the previous lesson, calculate the standard deviation of your sample. This should involve variables representing the mean, mean deviation, and mean deviation squared	6	
4. Calculate the one-sample t -test	4	
5. Identify the degrees of freedom associated with your t -test	2	
6. Locate the test critical value for your test	2	
7. Is the t -test statistically significant? Why or why not?	2	
8. Calculate the confidence interval around your sample mean	2	
9. Calculate the effect size (i.e., Cohen's d associated with your t -test	2	
Totals	24	

4.10 Homeworked Example

Screencast Link

The one-sample test comes in handy when you want to compare your dataset to an external benchmark or standard. It can be a real helper in program evaluation

If you wanted to use this example and dataset as a basis for a homework assignment, you could select a different course (i.e., Multivariate or Psychometrics) and/or compare the mean for the ORG department ($M = 4.1$).

4.10.1 Working the Problem with R and R Packages

4.10.1.1 Narrate the research vignette, describing the variables and their role in the analysis

From my course evaluation data, I want to ask the question, “Are ratings for the Overall Instructor for the ANOVA course evals statistically significantly different from the overall departmental averages for that same item?” In CPY the overall average for that specific item is 4.4.

4.10.1.2 Simulate (or import) and format data

The BIGdf is from a project that evaluated three changes to our own stats courses, over time. As a whole, this dataset violates a ton of assumptions of ANOVA, but we can create a tiny df and use it for demonstrations.

Let's first trim it to just students who took ANOVA

And further trim to our variable of interest

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2     v readr      2.1.4
v forcats   1.0.0     v stringr    1.5.0
v ggplot2   3.4.2     v tibble     3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()   masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

And further trim to non-missing data

- Is the sample variable on a continuous scale of measurement and formatted as *num* or *int* in R?
- Is the external score evaluated on the same continuous scale?

```
int [1:113] 5 4 4 3 5 3 5 4 3 5 ...
```

Yes. The format for the OvInstructor variable is integer (which is numerical); the overall course evaluation is on an equivalent (1 to 5) scale.

4.10.1.3 Evaluate statistical assumptions

- Are the skew and kurtosis values within the range expected?
- Does the distribution of the variable differ significantly from a normal distribution?

	nbr.val	nbr.null	nbr.na
113.0000000000000000000000	0.0000000000000000000000	0.0000000000000000000000	0.0000000000000000000000
min		max	range
1.0000000000000000000000	5.0000000000000000000000	4.0000000000000000000000	mean
473.0000000000000000000000	5.0000000000000000000000	4.185840707964601393	var
SE.mean	CI.mean.0.95	1.027654867256637239	skewness
0.095363991425895162	0.188951524765374329	-0.984495621273390964	kurt.2SE
std.dev	coef.var	-0.082121112321619227	normtest.W
1.013733134141642456	0.242181488706142922	-2.164227168444894378	normtest.p
skew.2SE	kurtosis	-0.074100280830601939	normtest.p
normtest.W	normtest.p	0.772806906937811733	0.00000000006195409

The skew value is -9.84 and far exceeds the absolute value of 3. The skew.2SE is -2.164 (larger than the absolute value of 2.0) is consistent. Thus, we might have some concern about skew.

The kurtosis value is -7.410 and is below the absolute value of 10. The kurt.2SE value is -8.212 which is substantially larger than the absolute value of 2.0. Thus, we are similarly concerned about kurtosis.

The Shapiro wilk test value is 7.728 ($p < 0.001$). This significant value suggests a distribution that is not normally distributed.

4.10.1.4 Conduct a one sample t test (with an effect size)

First, comparison to CPY

```
# A tibble: 1 x 12
  estimate .y. group1 group2    n statistic     p   df conf.low conf.high
*   <dbl> <chr>  <chr>  <chr>  <int>    <dbl> <dbl> <dbl> <dbl>    <dbl>
1     4.19 OvInst~ 1      null ~    113     -2.25 0.0267   112     4.00     4.37
# i 2 more variables: method <chr>, alternative <chr>
```

We can begin to create our t string:

$$t(112) = -2.246, p = 0.027, CI95(3.997, 4.374)$$

Let's interpret the results. With 112 degrees of freedom, our t value is -2.245. Because the p value is less than .05, this is statistically significant. This means that my course evaluations in ANOVA were statistically significantly lower than the average for CPY. We are 95% confident that the true course evaluation mean (for my courses) fell between 3.997 and 4.374.

Let's calculate the effect size. we will use a Cohen's d which is interpreted in standard deviation units.

```
# A tibble: 1 x 6
  .y.      group1 group2    effsize    n magnitude
* <chr>    <chr>  <chr>    <dbl> <int> <ord>
1 OvInstructor 1      null model  -0.211    113 small
```

Cohen's d was 0.211. This is a small effect. We can add it to the t string.

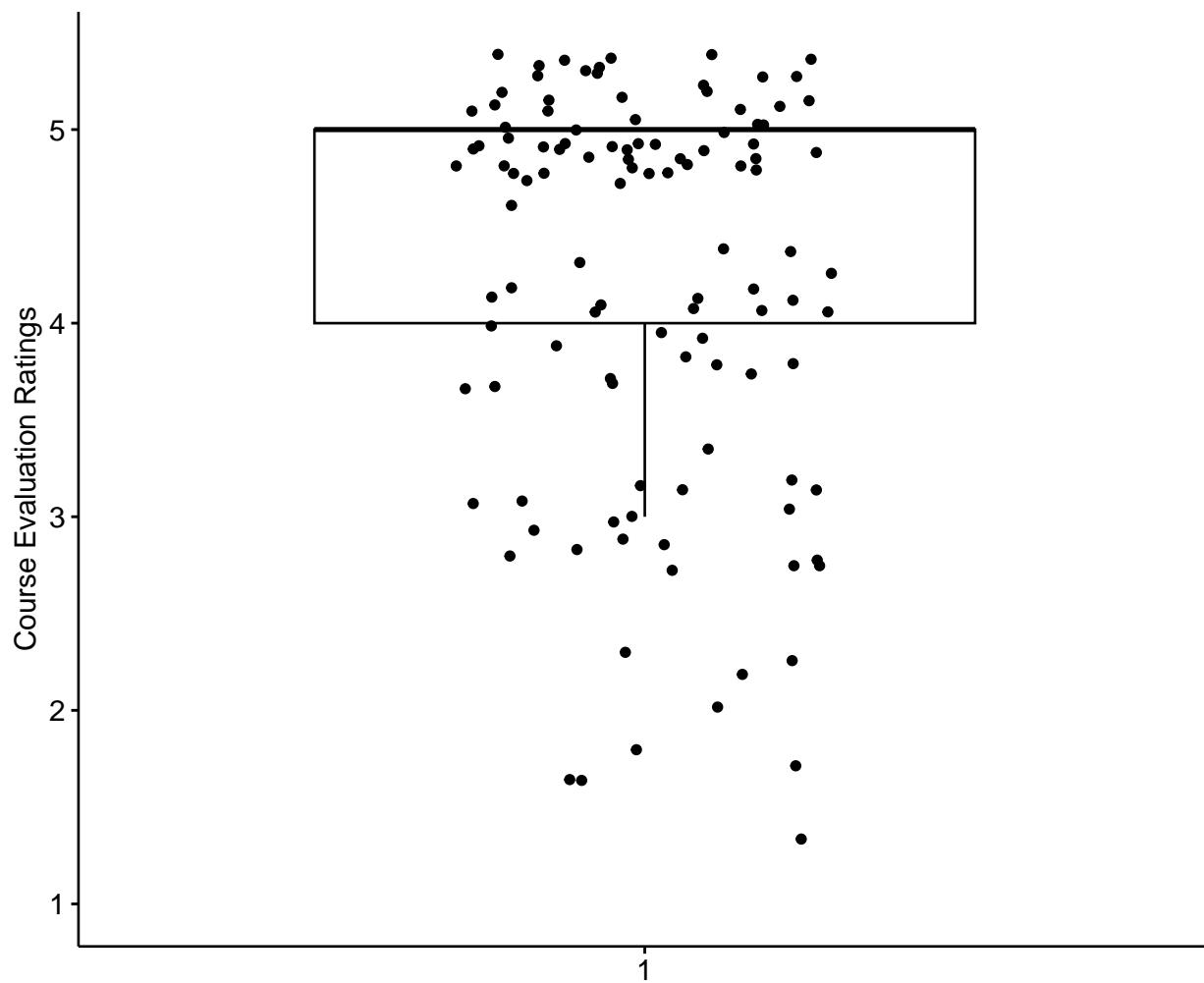
$$t(112) = -2.246, p = 0.027, CI95(3.997, 4.374), d = -0.211$$

4.10.1.5 APA style results with table(s) and figure

- t-test results should include t, df, p, d-or-eta, and CI95%
- Table
- Figure
- Grammar/style

A one-sample t -test was used to evaluate whether the *overall instructor* course evaluation ratings from the ANOVA courses were statistically significant from the departmental averages for the Clinical (CPY; $M = 4.4$) department. The sample mean for the ANOVA course evaluations was 4.186 ($SD = 1.013$). Although this mean was statistically significantly different from the average CPY course evaluation ratings of the same item, $t(112) = -2.246, p = 0.027, CI95(3.997, 4.374)$, the effect size was quite small ($d = -0.211$). A distribution of the ANOVA course ratings is found in Figure 1.

Figure 1. Overall Instructor Ratings for ANOVA



4.10.1.6 Conduct power analyses to determine the power of the current study and a recommended sample size

A quick reminder that the d in the power analysis is the difference between the means divided by the pooled standard deviation. This is the same as Cohen's d that we just calculated.

One-sample t test power calculation

```
n = 113
d = 0.211
sig.level = 0.05
power = 0.604022
alternative = two.sided
```

For the comparison to the CPY departmental average, power was 60%. That is, given the value of

the mean difference relative to the pooled standard deviation we had a 60% chance of detecting a statistically significant effect if there was one.

```
One-sample t test power calculation
```

```
n = 178.226
d = 0.211
sig.level = 0.05
power = 0.8
alternative = two.sided
```

For the CPY departmental comparison, the recommended sample size would be 178. This means there would need to be 178 individuals to find a statistically significant difference, if one existed.

4.10.2 Hand Calculations

4.10.2.1 Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses

$$\begin{aligned}H_0 : \mu &= 4.4 \\H_A : \mu &\neq 4.4\end{aligned}$$

4.10.2.2 Calculate the mean of your sample; identify the mean of your benchmarking sample

I will continue with the *tiny1* dataset and calculate the mean of the OvInstructor variable from my ANOVA course evaluations.

```
[1] 4.185841
```

The mean of my benchmarking sample is 4.4. This number is a “departmental standard” and did not need to be calculated by me for this purpose.

4.10.2.3 Using the steps from the previous lesson, hand-calculate the standard deviation of your sample. This should involve variables representing the mean, mean deviation, and mean deviation squared

```
[1] 1.027655
```

```
[1] 1.013733
```

The variance is 1.028; the standard deviation is 1.014.

```
[1] 1.013733
```

4.10.2.4 Calculate the one-sample *t*-test

Here's the formula:

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{N}}$$

```
[1] -2.245701
```

4.10.2.5 Identify the degrees of freedom associated with your *t*-test

For the one-sample *t*-test, $df = N - 1$. In our case

```
[1] 112
```

4.10.2.6 Locate the test critical value for your test

We can use a table of critical values for the one sample *t*-test: <https://www.statology.org/t-distribution-table/>

A 2-tail test, when $p = .05$, with ~ 120 individuals is 1.98

Or, this code:

```
[1] 1.98118
```

4.10.3 Is the *t*-test statistically significant? Why or why not?

Yes $t = -2.245701$ exceeds the (absolute) test critical value of 1.98.

4.10.3.1 What is the confidence interval around your sample mean?

Here is a reminder of the formula:

$$\bar{X} \pm t_{cv} \left(\frac{s}{\sqrt{n}} \right)$$

```
[1] 3.996908
```

```
[1] 4.374774
```

We are 95% confident that the sample mean for the student in the ANOVA classes is between 3.997, 4.375.

4.10.3.2 Calculate the effect size (i.e., Cohen's d associated with your t -test

A reminder of the two formula:

$$d = \frac{\text{MeanDifference}}{SD} = \frac{t}{\sqrt{N}}$$

[1] -0.2112578

[1] -0.2112578

Chapter 5

Independent Samples t -test

[Screencasted Lecture Link](#)

```
options(scipen = 999) #eliminates scientific notation
```

Researchers may wish to know if there are differences on a given outcome variable as a result of a dichotomous grouping variable. For example, during the COVID-19 pandemic, my research team asked if there were differences in the percentage of time that individuals wore facemasks as a result of 2020 Presidential voting trends (Republican or Democratic) of their county of residence. In these simple designs, the independent samples t -test could be used to test the researchers' hypotheses.

5.1 Navigating this Lesson

There is just less than one hour of lecture. If you work through the materials with me, plan for an additional hour

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

5.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Recognize the research questions for which utilization of the independent samples t -test would be appropriate.
- Narrate the steps in conducting an independent samples t -test, beginning with testing the statistical assumptions through writing up an APA style results section.
- Calculate an independent samples t -test in R (including effect sizes and 95% CIs).
- Interpret a 95% confidence interval around a mean difference score.

- Produce an APA style results section for an independent samples t -test.
- Determine a sample size that (given a set of parameters) would likely result in a statistically significant effect, if there was one.

5.1.2 Planning for Practice

The suggestions for homework vary in degree of complexity. The more complete descriptions at the end of the chapter follow these suggestions.

- Rework the independent samples t -test in the lesson by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- Rework the independent samples t -test in the lesson by changing something else about the simulation. For example, if you are interested in power, consider changing the sample size.
- Use the simulated data that is provided, but use the nonverbal variable, instead.
- Conduct an independent samples t -test with data to which you have access and permission to use. This could include data you simulate on your own or from a published article.

5.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Navarro, D. (2020). Chapter 13: Comparing two means. In [Learning Statistics with R - A tutorial for Psychology Students and other Beginners](#). Retrieved from [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro))
 - Navarro's OER includes a good mix of conceptual information about t -tests as well as R code. My lesson integrates her approach as well as considering information from Field's [2012] and Green and Salkind's [Green and Salkind, 2017c] texts (as well as searching around on the internet).
- Elliott, A. M., Alexander, S. C., Mescher, C. A., Mohan, D., & Barnato, A. E. (2016). Differences in Physicians' Verbal and Nonverbal Communication With Black and White Patients at the End of Life. *Journal of Pain and Symptom Management*, 51(1), 1–8. <https://doi.org/10.1016/j.jpainsymman.2015.07.008>
 - The source of our research vignette.

5.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(psych)){install.packages('psych')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(dplyr)){install.packages('dplyr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(pwr)){install.packages('pwr')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(knitr)){install.packages('knitr')}
# if(!require(rstatix)){install.packages('rstatix')}
```

5.2 Introducing the Independent Samples *t*-Test

The independent samples *t*-test assesses whether the population mean of the test variable for one group differs from the population mean of the test variable for a second group. This *t*-test can only accommodate two levels of a grouping variable (e.g., teachers/students, volunteers/employees, treatment/control) and the participants must be different in each group.



Figure 5.1: An image of a row with two boxes labeled Condition A (in light blue) and Condition B (in dark blue). This represents the use of an independent samples *t*-test to compare across conditions.

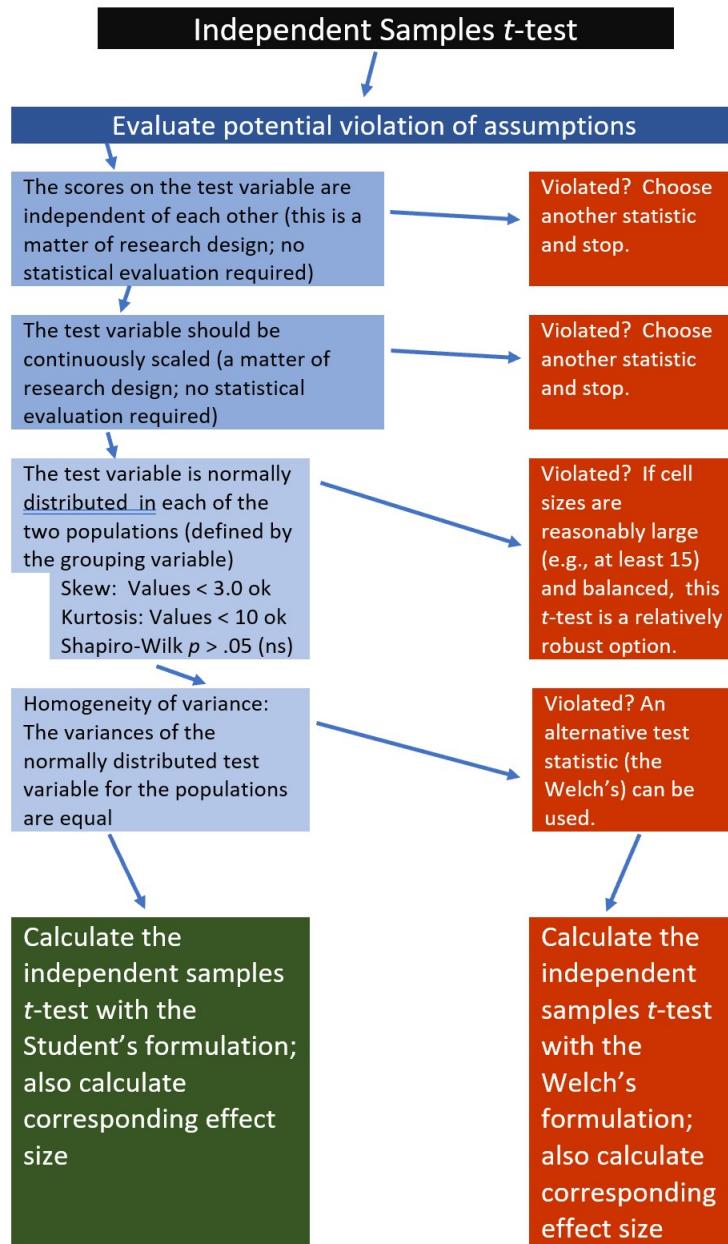
The comparison of two means is especially evident in the numerator of the formula. In the denominator we can see that the mean difference is adjusted by the standard error. At the outset, you should know that the formula in the denominator gets messy, but the formula, alone, provides an important conceptual map.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

If the researcher is interested in comparing the same participants' experiences across time or in different groups, they should consider using a **paired samples *t*-test**. Further, the independent samples *t*-test is limited to a grouping variable with only two levels. If the researcher is interested in three or more levels, they should consider using a **one-way ANOVA**.

5.2.1 Workflow for Independent Samples *t*-Test

The following is a proposed workflow for conducting a independent samples *t*-test.



If the data meets the assumptions associated with the research design (e.g., independence of observations and a continuously scaled metric), these are the steps for the analysis of an independent samples t -test:

1. Prepare (upload) data.
2. Explore data with
 - graphs
 - descriptive statistics
3. Assess normality via skew, kurtosis, and the Shapiro-Wilk test of normality
4. Consider the homogeneity of variance assumption and decide whether to use the Student's or Welch's formulation.
5. Compute the independent samples t -test

6. Compute an effect size (frequently the d or η^2 statistic)
7. Manage Type I error
8. Sample size/power analysis (which you should think about first, but in the context of teaching statistics, it's more pedagogically sensible, here).

5.3 Research Vignette

Empirically published articles where t -tests are the primary statistic are difficult to locate. Having exhausted the psychology archives, I located this article in an interdisciplinary journal focused on palliative medicine. The research vignette for this lesson examined differences in physician's verbal and nonverbal communication with Black and White patients at the end of life [Elliott et al., 2016].

Elliott and colleagues [2016] were curious to know if hospital-based physicians (56% White, 26% Asian, 7.4% each Black and Hispanic) engaged in verbal and nonverbal communication differently with Black and White patients. Black and White patient participants were matched on characteristics deemed important to the researchers (e.g., critically and terminally ill, prognostically similar, expressed similar treatment preferences). Interactions in the intensive care unit were audio and video recorded and then coded on dimensions of verbal and nonverbal communication.

Because each physician saw a pair of patients (i.e., one Black patient and one White patient), the researchers utilized a paired samples, or dependent t -test. This statistical choice was consistent with the element of the research design that controlled for physician effects through matching. Below are the primary findings of the study.

	Black Patients	White Patients	
Category	$Mean(SD)$	$Mean(SD)$	p -value
Verbal skill score (range 0 - 27)	8.37(3.36)	8.41(3.21)	0.958
Nonverbal skill score (range 0 - 5)	2.68(.84)	2.93(.77)	0.014

Although their design was more sophisticated (and, therefore, required the paired samples t -test), Elliott et al. [2016] could have simply compared the outcome variables (e.g., verbal and nonverbal communication) as a function of their dichotomous variable, patient race (Black, White).

5.3.1 Data Simulation

In the data below, I have simulated the verbal and non-verbal communication variables using the means and standard deviations listed in the article. Further, I truncated them to fit within the assigned range. I created 33 sets each and assigned them to the Black or White level of the grouping variable.

```
set.seed(220815)
# sample size, M, and SD for Black then White patients
Verbal <- c(rnorm(33, mean = 8.37, sd = 3.36), rnorm(33, mean = 8.41, sd = 3.21))
# set upper bound
Verbal[Verbal > 27] <- 27
# set lower bound
```

```

Verbal[Verbal < 0] <- 0
# sample size, M, and SD for Black then White patients
Nonverbal <- c(rnorm(33, mean = 2.68, sd = 0.84), rnorm(33, mean = 2.93,
  sd = 0.77))
# set upper bound
Nonverbal[Nonverbal > 5] <- 5
# set lower bound
Nonverbal[Nonverbal < 0] <- 0

ID <- factor(seq(1, 66))
# name factors and identify how many in each group; should be in same
# order as first row of script
PatientRace <- c(rep("Black", 33), rep("White", 33))
# groups the 3 variables into a single df: ID#, DV, condition
dfIndSamples <- data.frame(ID, PatientRace, Verbal, Nonverbal)

```

With our data in hand, let's inspect its structure (i.e., the measurement scales for the variables) to see if they are appropriate.

```
str(dfIndSamples)
```

```

'data.frame':   66 obs. of  4 variables:
 $ ID         : Factor w/ 66 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ PatientRace: chr  "Black" "Black" "Black" "Black" ...
 $ Verbal      : num  2.76 5.73 6.81 8.68 9.1 ...
 $ Nonverbal   : num  3.41 4.02 1.62 2.52 2.11 ...

```

The verbal and nonverbal variables are quasi-interval scale variables. Therefore, the numerical scale is correctly assigned by R. In contrast, patient race is a nominal variable and should be a factor. In their article, Elliot et al. [2016] assigned Black as the baseline variable and White as the comparison variable. Because R orders factors alphabetically, and “Black” precedes “White”, this would happen automatically. Because creating ordered factors is a useful skill, I will write out the full code.

```
dfIndSamples$PatientRace <- factor(dfIndSamples$PatientRace, levels = c("Black",
  "White"))
```

Let's again check the formatting of the variables:

```
str(dfIndSamples)
```

```

'data.frame':   66 obs. of  4 variables:
 $ ID         : Factor w/ 66 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ PatientRace: Factor w/ 2 levels "Black","White": 1 1 1 1 1 1 1 1 1 ...
 $ Verbal      : num  2.76 5.73 6.81 8.68 9.1 ...
 $ Nonverbal   : num  3.41 4.02 1.62 2.52 2.11 ...

```

The four variables of interest are now correctly formatted as *num* and *factor*.

Below is code for saving (and then importing) the data in .csv or .rds files. I make choices about saving data based on what I wish to do with the data. If I want to manipulate the data outside of R, I will save it as a .csv file. It is easy to open .csv files in Excel. A limitation of the .csv format is that it does not save any restructuring or reformatting of variables. For this lesson, this is not an issue.

Here is code for saving the data as a .csv and then reading it back into R. I have hashtagsged these out, so you will need to remove the hashtags if you wish to run any of these operations. If you have simulated the data (above), you do not need to save and then re-import the data.

```
# writing the simulated data as a .csv write.table(dfIndSamples, file
# = 'dfIndSamples.csv', sep = ',', col.names=TRUE, row.names=FALSE)
# at this point you could clear your environment and then bring the
# data back in as a .csv reading the data back in as a .csv file
# dfIndSamples<- read.csv ('dfIndSamples.csv', header = TRUE)
```

The .rds form of saving variables preserves any formatting (e.g., creating ordered factors) of the data. A limitation is that these files are not easily opened in Excel. Here is the hashtagsged code (remove hashtags if you wish to do this) for writing (and then reading) this data as an .rds file.

```
# saveRDS(dfIndSamples, 'dfIndSamples.rds') dfIndSamples <-
# readRDS('dfIndSamples.rds') str(dfIndSamples)
```

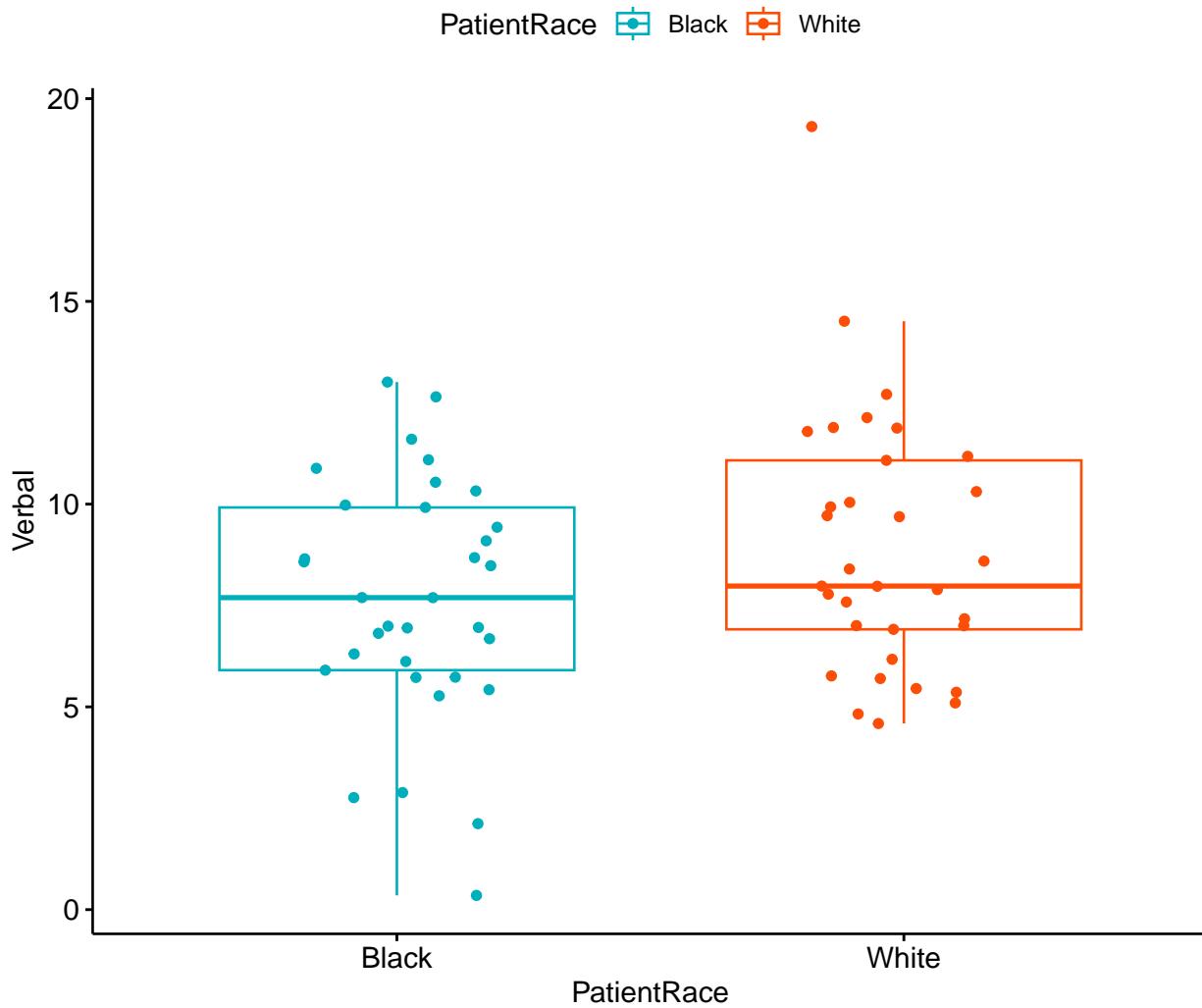
5.3.2 Quick Peek at the Data

Plotting the data is a helpful early step in any data analysis. Further, visualizing the data can help us with a conceptual notion of the statistic we are utilizing. The *ggpubr* package is one of my go-to-tools for quick and easy plots of data. Boxplots are terrific for data that is grouped. A helpful [tutorial](#) for boxplots (and related plots) can be found at datanovia.

In the code below I introduced the colors by identifying the grouping variable and assigning colors. Those color codes are the “Hex” codes you find in the custom color palette in your word processing program.

I am also fond of plotting each case with the command, *add* = “*jitter*”. To increase your comfort and confidence in creating figures (and with other tools) try deleting and adding back in different commands. This is how to distinguish between the essential and the elective elements of the code.

```
ggpubr::ggboxplot(dfIndSamples, x = "PatientRace", y = "Verbal", color = "PatientRace",
  palette = c("#00AFBB", "#FC4E07"), add = "jitter")
```



The box of the boxplot covers the middle 50% (the interquartile range). The horizontal line is the median. The whiskers represent three standard deviations above and below the mean. Any dots are outliers.

5.4 Working the Independent Samples t -Test (by hand)

5.4.1 Stating the Hypothesis

In this lesson, I will focus on differences in the verbal communication variable. Specifically, I hypothesize that physician verbal communication scores for Black and White patients will differ. In the hypotheses below, the null hypothesis (H_0) states that the two means are equal; the alternative hypothesis (H_A) states that the two means are not equal.

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_A : \mu_1 &\neq \mu_2 \end{aligned}$$

5.4.2 Calculating the *t*-Test

Earlier I presented a formula for the independent samples *t*-test.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

There are actually two formulations of the *t*-test. Student's version can be used when there is no violation of the homogeneity of variance assumption; Welch's can be used when the homogeneity of variance assumption is violated. For the hand-calculation demonstration, I will only demonstrate the formula in the most ideal of circumstances, that is: there is no violation of the homogeneity of variance assumption and sample sizes are equal.

Even so, while the formula seems straightforward enough, calculating the SE in the denominator gets a little spicy:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Let's first calculate the SE – the value of the denominator. For this, we need the standard deviations for the dependent variable (verbal) for both levels of patient race. We obtained these earlier when we used the *describeBy()* function in the *psych* package.

The standard deviation of the verbal variable for the levels in the patient race group were 2.99 for Black patients and 3.20 for White patients; the *N* in both our groups is 33. We can do the denominator math right in an R chunk:

```
sqrt((2.985^2/33) + (3.203^2/33))
```

```
[1] 0.7621627
```

Our *SE* = 0.762

With the simplification of the denominator, we can easily calculate the independent sample *t*-test.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

```
(7.615 - 8.891)/0.762
```

```
[1] -1.674541
```

Hopefully, this hand-calculation provided an indication of how the means, standard deviation, and sample sizes contribute to the estimate of this *t*-test value. Now we ask, “But it is statistically significant?”

5.4.2.1 Statistical Significance

The question of statistical significance testing invokes NHST (null hypothesis significance testing). In the case of the independent samples t -test, the null hypothesis is that the two means are equal; the alternative is that they are not equal. Our test is of the null hypothesis. When the probability (p) is less than the value we specify (usually .05), we are 95% certain that the two means are not equal. Thus, we reject the null hypothesis (the one we tested) in favor of the alternative (that the means are not equal).

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_A &: \mu_1 \neq \mu_2 \end{aligned}$$

Although still used, NHST has its critiques. Among the critiques are the layers of logic and confusing language as we interpret the results.

Our t -value was -1.675. We compare this value to the test critical value in a table of t critical values. In-so-doing we must know our degrees of freedom. In the test that involves two levels of a grouping value, we will use $N - 1$ as the value for degrees of freedom. We must also specify the p value (in our case .05) and whether-or-not our hypothesis is unidirectional or bi-directional. Our question only asked, “Are the verbal communication levels different?” In this case, the test is two-tailed, or bi-directional.

Let’s return to the [table of critical values](#) for the t distribution to compare our t -value (-1.675) to the column that is appropriate for our:

- Degrees of freedom (in this case $N - 2$ or 64)
 - We have two levels of a grouping value; for each our df is $N - 1$
- Alpha, as represented by $p < .05$
- Specification as a one-tailed or two-tailed test
 - Our alternative hypothesis made no prediction about the direction of the difference; therefore we will use a two-tailed test

In the above linked table of critical values, when the degrees of freedom reaches 30, there larger intervals. We will use the row representing degrees of freedom of 60. If our t -test value is lower than an absolute value of -2 or greater than the absolute value of 2, then our means are statistically significantly different from each other. In our case, we have not achieved statistical significance and we cannot say that the means are different. The t string would look like this: $t(64) = -1.675, p > .05$

We can also use the `qt()` function in base R. In the script below, I have indicated an alpha of .05. The “2” that follows indicates I want a two-tailed test. The 64 represents my degrees of freedom ($N - 2$). In a two-tailed test, the regions of rejection will be below the lowerbound (lower.tail=TRUE) and above the upperbound (lower.tail=FALSE).

```
qt(0.05/2, 64, lower.tail = TRUE)
```

```
[1] -1.99773
```

```
qt(0.05/2, 64, lower.tail = FALSE)
```

```
[1] 1.99773
```

Given the large intervals, it makes sense that this test critical value is slightly different than the one from the table.

5.4.2.2 Confidence Intervals

How confident are we in our result? With independent samples t -tests, it is common to report an interval in which we are 95% confident that our true mean difference exists. Below is the formula, which involves:

- $\bar{X}_1 - \bar{X}_2$ the difference in the means
- t_{cv} the test critical value for a two-tailed model (even if the hypothesis was one-tailed) where $\alpha = .05$ and the degrees of freedom are $N - 2$
- SE the standard error used in the denominator of the test statistic

$$(\bar{X}_1 - \bar{X}_2) \pm t_{cv}(SE)$$

Let's calculate it:

First, let's get the proper t critical value. Even though these are identical to the one above, I am including them again. Why? Because if the original hypothesis had been one-tailed, we would need to calculate a two-tailed confidence interval; this is a placeholder to remind us.

```
qt(0.05/2, 64, lower.tail = TRUE)
```

```
[1] -1.99773
```

```
qt(0.05/2, 64, lower.tail = FALSE)
```

```
[1] 1.99773
```

With this in hand, let's calculate the confidence intervals.

```
(7.614 - 8.891) - (1.99773 * 0.762)
```

```
[1] -2.79927
```

```
(7.614 - 8.891) + (1.99773 * 0.762)
```

```
[1] 0.2452703
```

These values indicate the range of scores in which we are 95% confident that our true mean difference ($\bar{X}_1 - \bar{X}_2$) lies. Stated another way, we are 95% confident that the true mean difference lies between -2.80 and 0.25. Because this interval crosses zero, we cannot rule out that the true mean difference is 0.00. This result is consistent with our non-significant p value. For these types of statistics, the 95% confidence interval and p value will always be yoked together.

5.4.2.3 Effect Size

Whereas p values address statistical significance, effect sizes address the magnitude of difference. There are two common effect sizes that are used with the independent samples t -test. The first is the d statistic, which measures, in standard deviation units, the distance between the two means. The simplest formula involves the t value and sample sizes:

$$d = t \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

With a t value of -1.675 and sample sizes at 33 each, we can easily calculate this. Small, medium, and large sizes for the d statistic are .2, .5, and .8, respectively (irrespective of sign).

```
-1.675 * (sqrt((33 + 33)/(33 * 33)))
```

```
[1] -0.4123565
```

Our value, -0.412 suggests a small-to-medium effect size. We might wonder why it wasn't statistically significant? Later we will discuss power and the relationship between sample size, one vs. two-tailed hypotheses, and effect sizes.

Eta square, η^2 is the proportion of variance of a test variable that is a function of the grouping variable. A value of 0 indicates that the difference in the mean scores is equal to 0, where a value of 1 indicates that the sample means differ, and the test scores do not differ within each group. The following equation can be used to compute η^2 . Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

$$\eta^2 = \frac{t^2}{t^2 + (N_1 + N_2 - 2)}$$

Let's calculate it:

```
(-1.6745 * -1.6745)/((-1.6745 * -1.6745) + (33 + 33 - 2))
```

```
[1] 0.04197282
```

Similarly, the η^2 is small-to-medium.

5.5 Working the Independent Samples t -Test with R Packages

Let's rework the problem in R. We start at the top of the flowchart, evaluating the statistical assumptions.

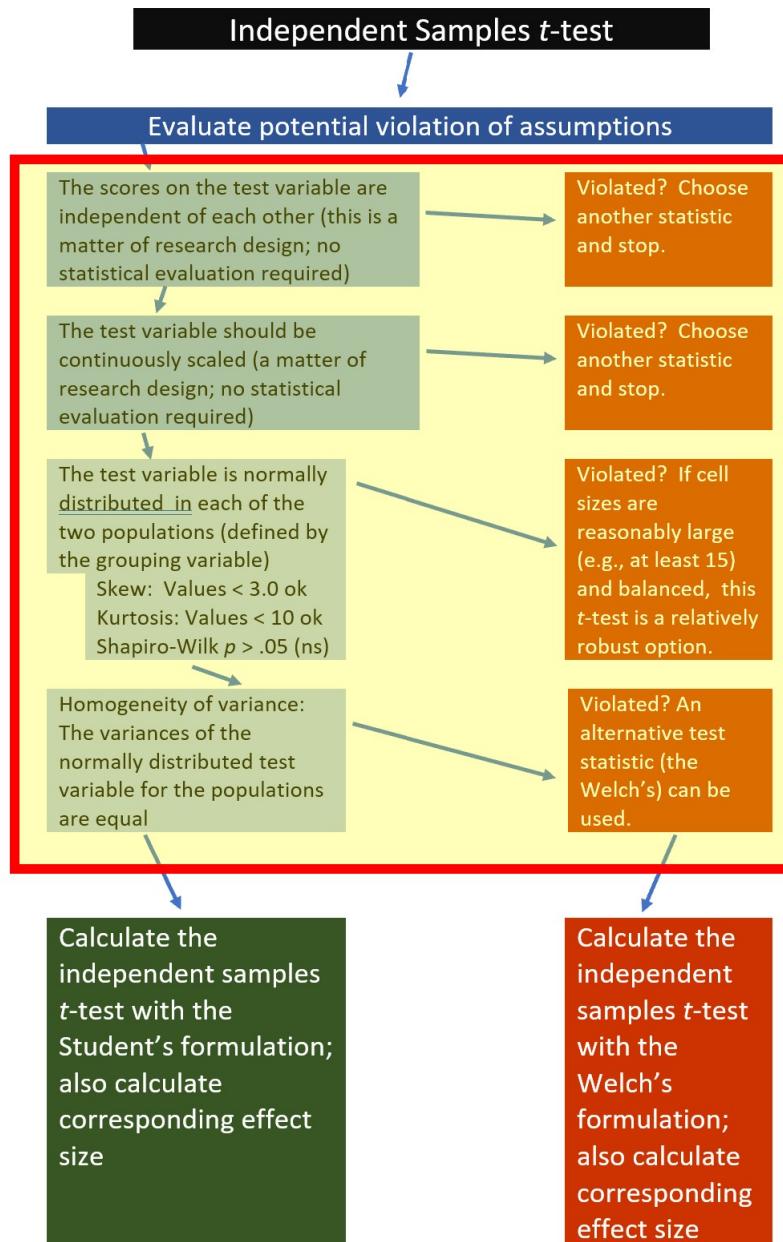


Figure 5.2: The workflow for the one sample t -test highlighting the evaluation of assumptions section

5.5.1 Evaluating the Statistical Assumptions

With an eye on our data, we can begin to explore the statistical assumptions associated with the independent samples *t*-test. Here's where we are in the workflow:

All statistical tests have some assumptions about the data. The independent-samples *t*-test has four:

- The scores on the test variable are independent of each other. This is a research design issue and the independent-samples *t*-test is not robust to violating this assumption.
 - If physicians' verbal communication was evaluated and reported for more than one patient, this vignette would violate the assumption of the independent samples *t*-test. For the sake of simplicity, let's presume that each was evaluated on verbal communication for only one patient. If the research scenario was such that physicians contributed multiple datapoints a potential analytic choice that is robust to such a violation is [multilevel modeling](#).
- The test variable should be continuously scaled. This is also a matter of research design and no statistical analysis is required.
 - Our test variable is an evaluation of verbal interactions; this is continuously scaled and has the properties of *interval*-level data.
- The test variable is normally distributed. We can check this several ways:
 - visually with histograms (perhaps with superimposed curves) and boxplots,
 - calculation of skew and kurtosis values,
 - calculation of the Shapiro-Wilk test of normality
- The variances of the normally distributed test variable for both levels of the grouping factor are equal. This is called the homogeneity of variance test and is easily calculated with a Levene's test of homogeneity of variance.

5.5.1.1 Is the dependent variable normally distributed at each level of the grouping variable?

We can begin to evaluate the assumption of normality by obtaining the descriptive statistics with the *describe()* function from the *psych* package.

```
psych::describe(dfIndSamples$Verbal, type = 1) #type=1 produces the type of skew and kurtosis
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	66	8.25	3.14	7.93	8.2	3.08	0.35	19.31	18.96	0.44	1.34	0.39

From this, we learn that the overall verbal mean is 8.25 with a standard deviation of 3.14. The values for skew (0.44) and kurtosis (1.34) fall below the areas of concern (below the absolute value of 3 for skew; below the absolute values of 10 for kurtosis) identified by Kline [2016a].

Recall that one of the assumptions for independent samples *t*-test is that the variable of interest is normally distributed within each level of the grouping variable. The *describeBy()* function in the *psych* package allows us to obtain these skew and kurtosis at both levels of the grouping variable.

If we feed the function the entire df, it will give us results for each level of PatientRace for each variable, including variables for which such disaggregation is nonsensible (i.e., ID, PatientRace). If we had a large df, we might want to create a tiny df that only includes our variable(s) of interest. For now, it is not problematic to include all the variables.

```
psych::describeBy(dfIndSamples ~ PatientRace, mat = TRUE, type = 1)
```

	item	group1	vars	n	mean	sd	median	trimmed		
ID*1		1	Black	1 33	17.000000 9.6695398	17.000000 17.000000				
ID*2		2	White	1 33	50.000000 9.6695398	50.000000 50.000000				
PatientRace*1		3	Black	2 33	1.000000 0.0000000	1.000000 1.000000				
PatientRace*2		4	White	2 33	2.000000 0.0000000	2.000000 2.000000				
Verbal1		5	Black	3 33	7.614884 2.9854116	7.693516 7.733412				
Verbal2		6	White	3 33	8.891483 3.2032222	7.979546 8.606615				
Nonverbal1		7	Black	4 33	2.943125 0.9251164	2.885724 2.931841				
Nonverbal2		8	White	4 33	2.965472 0.7001442	2.936787 2.995131				
					mad	min	max	range	skew	kurtosis
ID*1					11.8608000	1.0000000	33.000000	32.000000	0.0000000	-1.2022059
ID*2					11.8608000	34.0000000	66.000000	32.000000	0.0000000	-1.2022059
PatientRace*1					0.0000000	1.0000000	1.000000	0.000000	NaN	NaN
PatientRace*2					0.0000000	2.0000000	2.000000	0.000000	NaN	NaN
Verbal1					2.9075794	0.3507447	13.011100	12.660355	-0.3705014	-0.1377654
Verbal2					3.2861809	4.5891699	19.311207	14.722037	1.0651306	1.5382575
Nonverbal1					0.9185825	0.8333731	5.000000	4.166627	0.1204796	0.1380025
Nonverbal2					0.5560620	1.1311619	4.350886	3.219724	-0.4338806	0.3937160
					se					
ID*1					1.6832508					
ID*2					1.6832508					
PatientRace*1					0.0000000					
PatientRace*2					0.0000000					
Verbal1					0.5196935					
Verbal2					0.5576094					
Nonverbal1					0.1610421					
Nonverbal2					0.1218795					

In this analysis we are interested in the verbal variable. We see that patients who are Black received verbal interactions from physicians that were quantified by a mean score of 7.61 ($SD = 2.99$); physicians' scores for White patients were 8.89 ($SD = 3.20$). Skew and kurtosis values for the verbal ratings with Black patients were -.37 and -.14, respectively. They were 1.07 and 1.54 for White patients. As before, these fall well below the absolute values of 3 (skew) and 10 (kurtosis) that are considered to be concerning.

Beyond skew and kurtosis, we can formally test for deviations from normality with a Shapiro-Wilk. The script below first groups the data by PatientRace and then applies the *rstatix::shapiro_test()*. We want the results to be non-significant.

```
library(tidyverse) #opening this package so I can use the pipes

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr    1.3.0
v purrr    1.0.1

-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
shapiro <- dfIndSamples %>%
  group_by(PatientRace) %>%
  rstatix::shapiro_test(Verbal)
shapiro
```

	PatientRace	variable	statistic	p
<fct>	<chr>	<dbl>	<dbl>	
1	Black	Verbal	0.977	0.677
2	White	Verbal	0.922	0.0204

The Shapiro-Wilk test of normality indicated that the dependent variable, evaluation of verbal interaction with the patient was normally distributed within Black patients ($W = 0.977, p = 0.677$), but not within White patients ($W = 0.922, p = 0.020$). That is, the distribution of verbal communication scores for physicians attending to White patients was statistically significantly different from a normal distribution.

Should we be concerned? A general rule of thumb is that when cell sizes are larger than 15 the independent t -test should be relatively robust to violations of normality [Green and Salkind, 2017c].

5.5.1.2 Are the variances of the dependent variable similar across the levels of the grouping factor?

One of the assumptions of the independent samples t -test is that the variances of the dependent variable (verbal communication) are similar for both levels of the PatientRace factor. We can use the Levene's test to do this. We want this value to be non-significant ($p > .05$). If violated, we can use the Welch's test because it is robust to the violation of the homogeneity of variance.

Using `rstatix::levene_test()`, we simply need to point to the data, provide a “formula” in the form of “dependent variable by grouping variable,” and specify about how to center the data. The median is a commonly used because it provides a more robust test.

```
rstatix::levene_test(dfIndSamples, Verbal ~ PatientRace, center = median)

# A tibble: 1 x 4
  df1   df2 statistic     p
  <int> <int>    <dbl> <dbl>
1     1     64     0.0398 0.843
```

The results of the Levene's test are presented as an F statistic. We'll get to F distributions in the next chapter. For now, it is just important to know how to report and interpret them:

- Degrees of freedom are 1 and 64
- The value of the F statistic is 0.039
- The p value is 0.843 (i.e., greater than .05)

Happily, our Levene's result is ($F[1, 64] = 0.039, p = 0.843$) not significant. Because p is greater than .05, we have not violated the homogeneity of variance assumption. That is to say, the variance in the patient race groups is not statistically significantly different from each other. We can use the regular (Student's) formulation of the t -test for independent samples.

5.5.1.3 APA style write-up of testing the assumptions

My practice is to create APA style drafts of the different sections of the analysis as I work along. Here's how I might capture our evaluation of the statistical assumptions:

We began by analyzing the data to see if it met the statistical assumptions for analysis with an independent samples t -test. One assumption is that the dependent variable be normally distributed within the both levels of the grouping variable. We evaluated skew and kurtosis using Kline's [2016a] guidelines of the absolute values of 3 (skew) and 10 (kurtosis). Our results were well-within these boundary conditions. Specifically, the verbal ratings of physicians with Black patients were -.37 and -.14 for skew and kurtosis, respectively; they were 1.07 and 1.54 for White patients. The Shapiro-Wilk test of normality indicated that the dependent variable was normally distributed within Black patients ($W = 0.977, p = 0.677$), but not within White patients ($W = 0.922, p = 0.020$). That is, the distribution of verbal communication scores for physicians attending to White patients was statistically significantly different from a normal distribution. Results of Levene's homogeneity of variance test suggested that the variance in each of the patient race groups was not statistically significantly different from each other ($F[1, 64] = 0.039, p = 0.843$). Because the independent samples t -test is relatively robust to violations of normality when samples sizes have at least 15 participants per cell [Green and Salkind, 2017c] and there was no violation of the homogeneity of variance assumption we proceeded with the Student's formulation of the t -test for independent samples.

Odds are, owing to space limitations in journals, you would not provide this much detail about an independent samples t -test in an empirical manuscript. I am encouraging you to do so as you work through these chapters because it is good practice for thinking through the logic and sequencing of statistics as well as writing results in APA style.

5.5.2 Computing the Independent Samples *t*-Test

We are ready to compute the independent samples *t*-test.

Calculating an independent samples *t*-test is possible through base R and a number of packages. Kassambara's [b] *rstatix* package is one we can use for all of the *t*-test and ANOVA problems that we will work. I like it for several reasons. First, it was designed to be "pipe-friendly" in a manner that is consistent with the *tidyverse* approach to working in R and there are numerous tutorials. Additionally, *rstatix* objects work well with *ggpubr*, one of my favorite packages for graphing data and results.

In the script below:

- the first element points to the dataframe
- the second element provides a "formula"
 - we are predicting "Verbal" from "PatientRace"
- the third element, "var.equal=TRUE" means that we are using Student's formulation (because we did not violate the homogeneity of variance assumption)
- specifying "detailed = TRUE" will produce the 95% confidence interval around the difference in the two means

```
rstatix::t_test(dfIndSamples, Verbal ~ PatientRace, var.equal = TRUE, detailed = TRUE)

# A tibble: 1 x 15
  estimate estimate1 estimate2 .y.    group1 group2    n1    n2 statistic     p
*   <dbl>      <dbl>      <dbl> <chr>  <chr>  <chr>  <int> <int>      <dbl> <dbl>
1    -1.28      7.61      8.89 Verbal Black White     33     33     -1.67 0.0989
# i 5 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>
```

From this output we can start to draft our *t* string: $t(64) = -1.675, p = 0.099, CI95(-2.80, 0.25)$.

Separately, we must request the effect size. Earlier in the lesson we calculated both Cohen's *d* and eta-squared. Unfortunately, the *rstatix* package only offers the Cohen's *d* for *t*-tests. If you wanted an eta-squared, it would be easy enough to hand-calculate (or obtain from another R package).

```
rstatix::cohens_d(dfIndSamples, Verbal ~ PatientRace, var.equal = TRUE)
```

```
# A tibble: 1 x 7
  .y.    group1 group2 effsize    n1    n2 magnitude
* <chr>  <chr>  <chr>    <dbl> <int> <int> <ord>
1 Verbal Black White   -0.412     33     33 small
```

We can update our *t* string to include the effect size: $t(64) = -1.675, p = 0.099, CI95(-2.80, 0.25), d = -0.412$

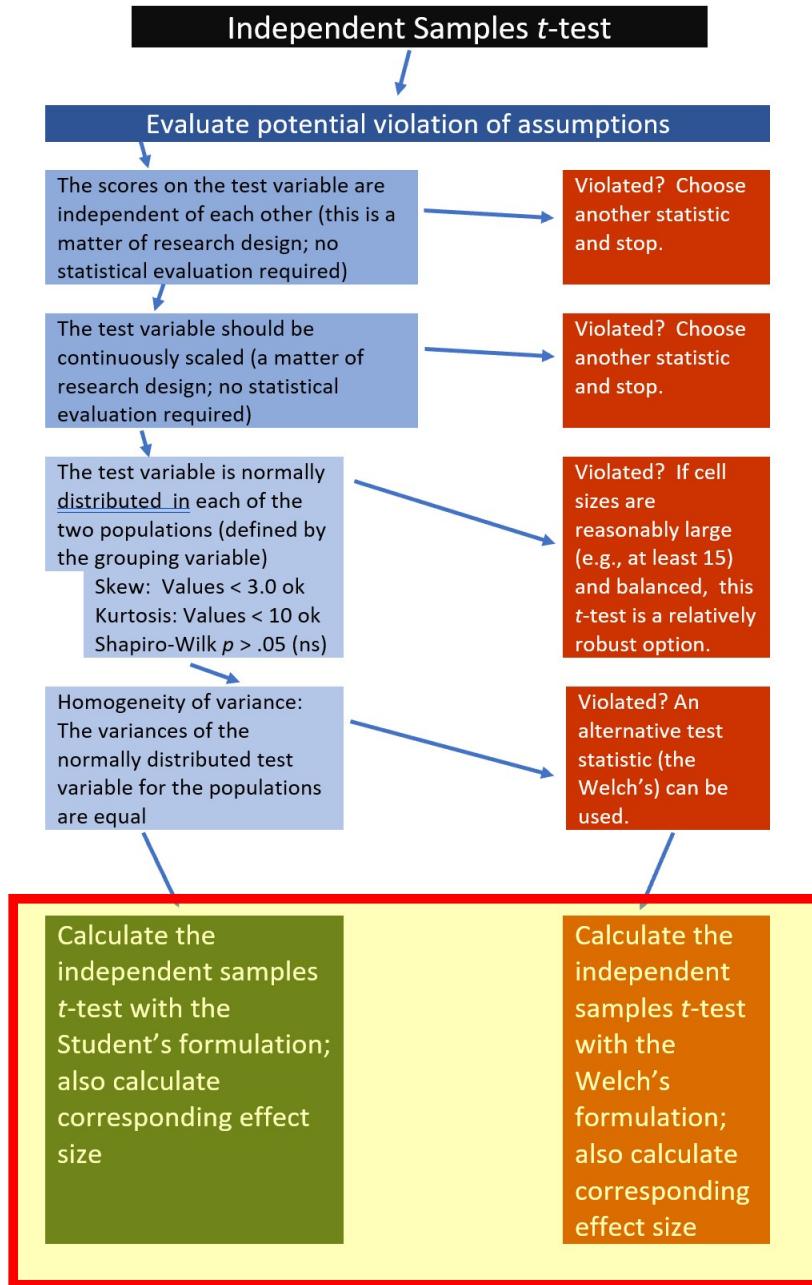


Figure 5.3: A colorful image of a workflow for the paired samples t -test focusing on the computation of the t -test

What does this mean? Our result is not-significant. Our estimate of the difference in verbal communication ratings when physicians interacted with Black and White patients was -1.675. We are 95% confident that that true mean difference is as low as -2.80 or as high as 0.25. Because the confidence interval crosses zero, we cannot be certain that the true difference is zero. This is consistent with the non-significant p value and effect size. Our output even tells us that the d of -0.41 is small.

5.5.3 What if we had violated the homogeneity of variance assumption?

Earlier we used the Levene's test to examine the homogeneity of variance assumption. If we had violated it, the Welch's formulation of the independent sample t -test is available to us. The *rstatix* package makes this easy. We simply change the *var.equal* to *FALSE*. This will produce the Welch's alternative, which takes into consideration violations of the homogeneity of variance assumption. Conveniently, "Student's" or "Welch's" will serve as the first row of the output.

```
rstatix::t_test(dfIndSamples, Verbal ~ PatientRace, var.equal = FALSE,
  detailed = TRUE)
```

```
# A tibble: 1 x 15
  estimate estimate1 estimate2 .y.   group1 group2    n1    n2 statistic      p
*   <dbl>     <dbl>     <dbl> <chr> <chr> <chr> <int> <int>    <dbl> <dbl>
1    -1.28      7.61      8.89 Verbal Black  White     33     33     -1.67 0.0989
# i 5 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
#   alternative <chr>
```

Likely because of the similarity of the standard deviations associated with each level of patient race and our equal cell sizes, this changes nothing about our conclusion. Note that the degrees of freedom in the Student's t -test analysis (the first one) was 64; in the Welch's version, the degrees of freedom is 63.685. It is this change that, when the homogeneity of variance assumption is violated, can make the Welch's results more conservative (i.e., less likely to have a statistically significant result).

5.6 APA Style Results

Putting it altogether, here is an APA Style results section:

An independent samples t -test was conducted to evaluate the hypothesis that there would be differences between the quality of physicians' verbal communication depending on whether the patient's race (Black, White).

We began by analyzing the data to see if it met the statistical assumptions for analysis with an independent samples t -test. One assumption is that the dependent variable be normally distributed within the both levels of the grouping variable. We evaluated skew and kurtosis using Kline's [2016a] guidelines of the absolute values of 3 (skew) and 10 (kurtosis). Our results were well-within these boundary conditions. Specifically,

the verbal ratings of physicians with Black patients were -.37 and -.14 for skew and kurtosis, respectively; they were 1.07 and 1.54 for White patients. The Shapiro-Wilk test of normality indicated that the dependent variable was normally distributed within Black patients ($W = 0.977, p = 0.677$), but not within White patients ($W = 0.922, p = 0.020$). That is, the distribution of verbal communication scores for physicians attending to White patients was statistically significantly different from a normal distribution. Results of Levene's homogeneity of variance test suggested that the variance in each of the patient race groups was not statistically significantly different from each other ($F[1, 64] = 0.039, p = 0.843$). Because the independent samples t -test is relatively robust to violations of normality when sample sizes have at least 15 participants per cell [Green and Salkind, 2017c] and there was no violation of the homogeneity of variance assumption we proceeded with the Student's formulation of the t -test for independent samples.

Results of the independent samples t -test was nonsignificant, $t(64) = -1.675, p = .099, d = 0.412$. The 95% confidence interval for the difference in means ranged from -2.799 to 0.246. Means and standard deviations are presented in Table 1; the results are illustrated in Figure 1.

```
apaTables::apa.1way.table(PatientRace, Verbal, dfIndSamples)
```

Descriptive statistics for Verbal as a function of PatientRace.

PatientRace	M	SD
Black	7.61	2.99
White	8.89	3.20

Note. M and SD represent mean and standard deviation, respectively.

The figure we created earlier in the lesson would be sufficient for a journal article. However, using *rstatix* in combination with *ggpubr* can be quite powerful. The result can be a figure that includes the t -test results and “significance bars.” To do this, we first need to re-run the *rstatix::t_test*, but adding to it by

- including “add_significance()” script after the pipe, and
- saving it as an object, which I’m naming “pair.test.”

We could have done this in the initial run (but I didn’t want to make the test-statistic unnecessarily confusing).

```
ind.test <- rstatix::t_test(dfIndSamples, Verbal ~ PatientRace, var.equal = TRUE,
                           detailed = TRUE) %>%
  rstatix::add_significance()
ind.test
```

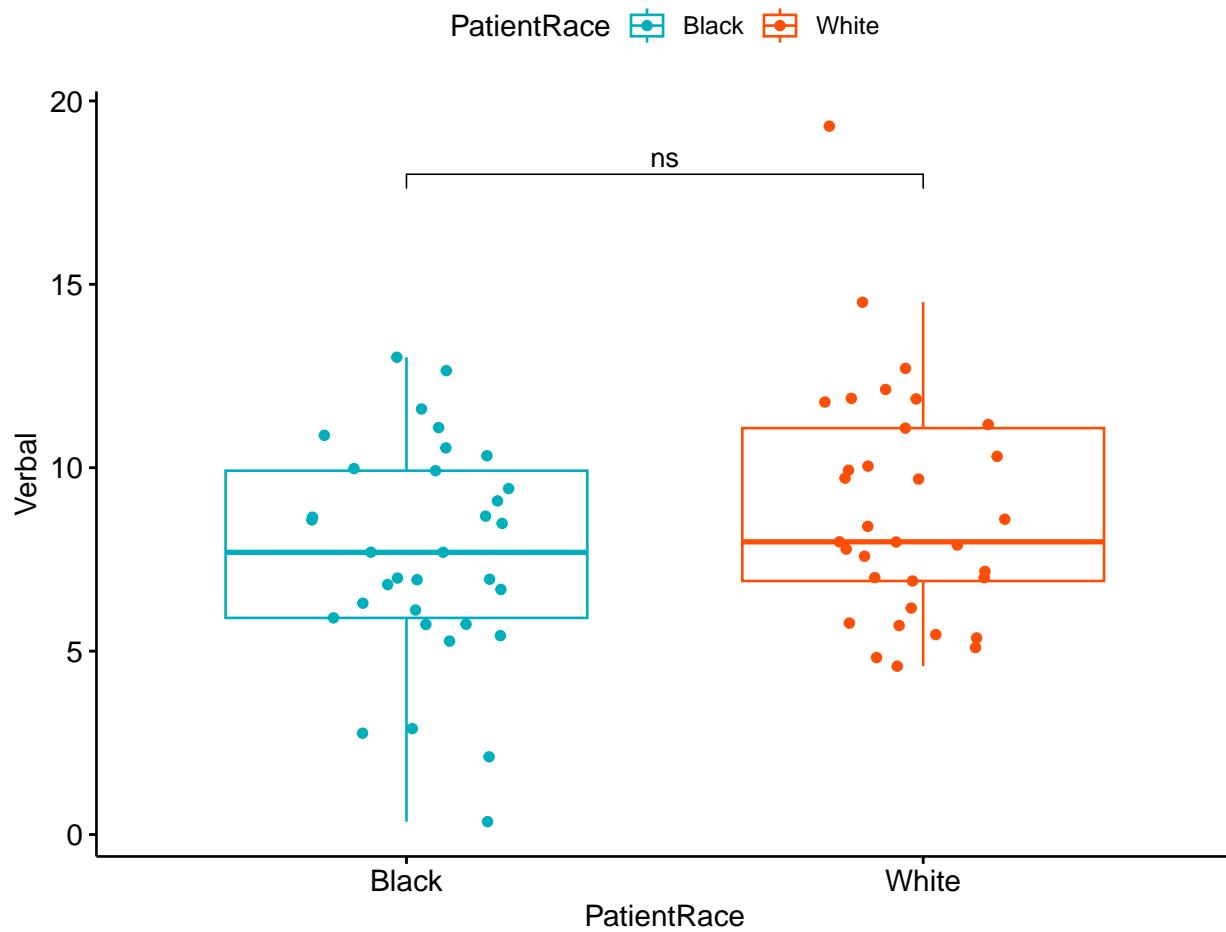
```
# A tibble: 1 x 16
  estimate estimate1 estimate2 .y.    group1 group2     n1     n2 statistic      p
  <dbl>     <dbl>     <dbl> <chr>  <chr>  <chr>  <int> <int>     <dbl>  <dbl>
1 -1.28      7.61      8.89 Verbal Black   White     33     33    -1.67 0.0989
# i 6 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>, p.signif <chr>
```

Next, we update the earlier boxplot code with the results from our statistical analyses:

```
ind.box <- ggpubr::ggbboxplot(dfIndSamples, x = "PatientRace", y = "Verbal",
  color = "PatientRace", palette = c("#00AFBB", "#FC4E07"), add = "jitter",
  title = "Figure 1. Physician Verbal Engagement as a Function of Patient Race")
ind.test <- ind.test %>%
  rstatix::add_xy_position(x = "PatientRace") #autocomputes p-value labels positions
ind.box <- ind.box + ggpubr::stat_pvalue_manual(ind.test, label = "p.signif",
  tip.length = 0.02, hide.ns = FALSE, y.position = c(18)) + labs(subtitle = rstatix::get_test_
detailed = TRUE) #adds t-test results

ind.box
```

Figure 1. Physician Verbal Engagement as a Function of Patient Race
 T test, $t(64) = -1.67$, $p = 0.099$, $n = 66$



Between the *rstatix* and *ggpubr* tools, there is a great deal of flexibility in creating figures. Determining which figure is best will likely depend on your outlet, your audience, your goals, and your personal preferences. For example, a print journal might prefer a black-and-white figure (with no fill in the boxes). This is accomplished easily enough by removing (or, hashtagging out) the “color” and “palette” arguments.

5.7 Power in Independent Samples t -tests

Researchers often use power analysis packages to estimate the sample size needed to detect a statistically significant effect, if, in fact, there is one. Utilized another way, these tools allows us to determine the probability of detecting an effect of a given size with a given level of confidence. If the probability is unacceptably low, we may want to revise or stop. A helpful overview of power as well as guidelines for how to use the *pwr* package can be found at a [Quick-R website \[Kabacoff, 2017\]](#).

In Champely’s *pwr* package, we can conduct a power analysis for a variety of designs, including the independent samples t -test that we worked in this lesson. There are a number of interrelating

elements of power:

- Sample size, n refers to the number of observations in each group; our vignette had 33
- d refers to the difference between means divided by the pooled standard deviation; we can use the value of Cohen's d for this
- $power$ refers to the power of a statistical test; conventionally it is set at .80
- $sig.level$ refers to our desired alpha level; conventionally it is set at .05
- $type$ indicates the type of test we ran; this was "two.sample"
- $alternative$ refers to whether the hypothesis is non-directional/two-tailed ("two.sided") or directional/one-tailed("less" or "greater")

In this script, we must specify *all-but-one* parameter; the remaining parameter must be defined as `NULL`. R will calculate the value for the missing parameter.

When we conduct a "power analysis" (i.e., the likelihood of a hypothesis test detecting an effect if there is one), we specify, "power=NULL". Using the data from our results, we learn from this first run, that our statistical power was 0.38. That is, given the value of the mean difference (1.276) we had a 38% chance of detecting a statistically significant effect if there was one. This is consistent with our non-significant result.

```
pwr::pwr.t.test(d = -0.412, n = 33, power = NULL, sig.level = 0.05, type = "two.sample",
  alternative = "two.sided")
```

Two-sample t test power calculation

```
n = 33
d = 0.412
sig.level = 0.05
power = 0.3778572
alternative = two.sided
```

NOTE: `n` is number in *each* group

Researchers frequently use these tools to estimate the sample size required to obtain a statistically significant effect. In these scenarios we set n to `NULL`. Using the results from the simulation of our research vignette, you can see that we would have needed 93 individuals (per group; 186 total) for the p value to be $< .05$.

```
pwr::pwr.t.test(d = -0.412, n = NULL, power = 0.8, sig.level = 0.05, type = "two.sample",
  alternative = "two.sided")
```

Two-sample t test power calculation

```
n = 93.44893
```

```
d = 0.412
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

Given that we had a non-significant result, this is not surprising. None-the-less, let's try it again. Below I will re-simulate the data for the verbal scores and change only the sample size:

```
set.seed(230525)
# sample size, M, and SD for Black then White patients
rVerbal <- c(rnorm(93, mean = 8.37, sd = 3.36), rnorm(93, mean = 8.41,
sd = 3.21))
# set upper bound
rVerbal[rVerbal > 27] <- 3
# set lower bound
rVerbal[rVerbal < 0] <- 0
# sample size, M, and SD for Black then White patients
rNonverbal <- c(rnorm(93, mean = 2.68, sd = 0.84), rnorm(93, mean = 2.93,
sd = 0.77))
# set upper bound
rNonverbal[rNonverbal > 5] <- 5
# set lower bound
rNonverbal[rNonverbal < 0] <- 0

rID <- factor(seq(1, 186))
# name factors and identify how many in each group; should be in same
# order as first row of script
rPatientRace <- c(rep("Black", 93), rep("White", 93))
# groups the 3 variables into a single df: ID#, DV, condition
rdfIndSamples <- data.frame(rID, rPatientRace, rVerbal, rNonverbal)

rdfIndSamples$rPatientRace <- factor(rdfIndSamples$rPatientRace, levels = c("Black",
"White"))

rstatix::t_test(rdfIndSamples, rVerbal ~ rPatientRace, var.equal = TRUE,
detailed = TRUE)

# A tibble: 1 x 15
  estimate estimate1 estimate2 .y.    group1 group2     n1     n2 statistic      p
*   <dbl>     <dbl>     <dbl> <chr>  <chr>  <chr>  <int>  <int>     <dbl>  <dbl>
1    0.852     9.02     8.17 rVerb~ Black  White     93     93     1.71 0.0884
# i 5 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>
```

```
rstatix::cohens_d(rdfIndSamples, rVerbal ~ rPatientRace, var.equal = TRUE)
```

```
# A tibble: 1 x 7
.y.   group1 group2 effsize    n1    n2 magnitude
* <chr>  <chr>  <chr>    <dbl> <int> <int> <ord>
1 rVerbal Black  White     0.251    93    93 small
```

Curiously, our result is still not statistically significant: $t(184) = 1.713, p = 0.088, d = 0.251, CI95[-0.129, 1.832]$. Given the closeness of our means (9.025, 8.173), this makes sense to me. Additionally, it does show us, though, how power is influenced by sample size. Holding all else equal, the larger the sample, the more likely we are to have a statistically significant result.

5.8 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “do-able” and then try at least one more problem that challenges you in some way. Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands).

Additionally, please complete at least one set of *hand calculations*, that is using the code demonstrated in the chapter to work through the formulas that compute the independent samples *t*-test. At this stage in your learning, you may ignore any missingness in your dataset by excluding all rows with missing data in your variables of interest.

5.8.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation of the research vignette, then rework the problem. This should provide minor changes to the data (maybe even in the second or third decimal point), but the results will likely be very similar. That said, don’t be alarmed if what was non-significant in my working of the problem becomes significant. Our selection of $p < .05$ (and the corresponding 95% confidence interval) means that 5% of the time there could be a difference in statistical significance.

5.8.2 Problem #2: Rework the research vignette, but change something about the simulation

Rework the independent samples *t*-test in the lesson by changing something else about the simulation. You might have noticed that my re-simulation of a smaller sample size did not produce a statistically significant result. You may wish to pick a value in between the primary lecture N and the re-simulation to see what it takes to achieve statistical significance. Alternatively, you could specify different means and/or standard deviations.

5.8.3 Problem #3: Rework the research vignette, but swap one or more variables

Use the simulated data, but select the nonverbal communication variables that were evaluated in the Elliott et al. [2016] study. Compare your results to those reported in the manuscript.

5.8.4 Problem #4: Use other data that is available to you

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete an independent samples t -test.

5.8.5 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the variables and their role in the analysis	5	_____
2. Simulate (or import) and format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct an independent samples t -test (with an effect size and 95% CIs)	5	_____
5. APA style results with table(s) and figure	5	_____
6. Conduct power analyses to determine the power of the current study and a recommended sample size	5	_____
7. Explanation to grader	5	_____
Totals	35	_____

Hand Calculations	Points Poss	Points Earned
1. Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	2	
2. Using an R package or functions in base R, calculate the means and standard deviations for both levels of the dependent variable	4	
3. Calculate the SE used in the denominator of the t -test	4	
4. Calculate the independent samples t -test	4	
5. Identify the degrees of freedom associated with your t -test	2	
6. Locate the test critical value for your test	2	

Hand Calculations	Points Poss	Points Earned
7. Is the t -test statistically significant? Why or why not?	2	
8. What is the confidence interval around the difference in sample means?	4	
9. Calculate the effect size (i.e., Cohen's d associated with your t -test)	4	
10. Assemble the results into a statistical string	4	
Totals*	32	

5.9 Homeworked Example

Screencast Link

The independent-samples t-test is useful when you want to compare means across two different groups. That is, the people in the comparison groups must be different from each other.

If you wanted to use this example and dataset as a basis for a homework assignment, you could change the course (i.e., Multivariate or Psychometrics) and/or change the dependent variable to one of the other scales.

5.9.1 Working the Problem with R and R Packages

5.9.1.1 Narrate the research vignette, describing the variables and their role in the analysis

I want to ask the question, “Do the course evaluation ratings for the traditional pedagogy subscale differ for CPY and ORG students?”

I will use the mean rating for the traditional pedagogy rating. As a mean, it retains its continuous, Likert scaling, ranging from 1 to 5 (with higher scores being more positive).

My predictor variable will be department. It has two levels: CPY and ORG.

5.9.1.2 Simulate (or import) and format data

First, bring in the dataset.

To avoid “dependency” in the data, I will just use data from the ANOVA course. Let’s first trim it to just students who took ANOVA

I will create a mean score of completed items from the traditional pedagogy scale.

To make it easier for teaching, I will make a super tiny df with just the predictor and continuous variable.

And further trim to non-missing data

Are the structures of the variables as follows: * Grouping variable: factor * Dependent variable: numerical or integer

In our case we want Department to be a factor with two levels and the SCRPed variable to be integer or numerical.

```
Classes 'data.table' and 'data.frame': 112 obs. of 2 variables:
 $ Dept    : chr "CPY" "CPY" "CPY" "CPY" ...
 $ TradPed: num 4.4 3.8 4 3 4.8 3.5 4.6 3.8 3.6 4.6 ...
 - attr(*, "na.action")= 'omit' Named int [1:2] 74 84
 ..- attr(*, "names")= chr [1:2] "202" "234"
```

Since the Department is a character variable, we need to change it to be a factor.

```
Factor w/ 2 levels "CPY","ORG": 1 1 1 1 1 1 1 1 1 1 ...
```

Without further coding, R will order the factors alphabetically. This is fine. CPY will be the base/intercept and ORG will be the comparison (this becomes more important in regression).

5.9.1.3 Evaluate statistical assumptions

- Evaluate and report skew and kurtosis
- Evaluate and correctly interpret homogeneity of variance (if Levene's $< .05$; use Welch's formulation)

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max
Dept*1	1	CPY	1	81	1.000000	0.0000000	1.0	1.000000	0.000000	1.0	1
Dept*2	2	ORG	1	31	2.000000	0.0000000	2.0	2.000000	0.000000	2.0	2
TradPed1	3	CPY	2	81	4.129630	0.7547259	4.2	4.210769	0.88956	1.8	5
TradPed2	4	ORG	2	31	3.870968	1.0948953	4.0	4.040000	1.18608	1.0	5
	range				skew	kurtosis	se				
Dept*1	0.0				NaN	NaN	0.00000000				
Dept*2	0.0				NaN	NaN	0.00000000				
TradPed1	3.2	-0.7630015	0.1555318	0.08385843							
TradPed2	4.0	-1.1832786	0.5826528	0.19664900							

Although I included Dept in the descriptives, it is a factor and therefore the values around distribution are rather senseless.

TradPed, though, is a continuously scored variable:

Skew = -0.763 (CPY) and -1.183 (ORG) falls below the $|3.0|$ threshold of concern (Klein, 2016)
 Kurtosis = 0.156 (CPY) and 0.583 (ORG) falls below the $|10.0|$ threshold of concern (Klein, 2016)

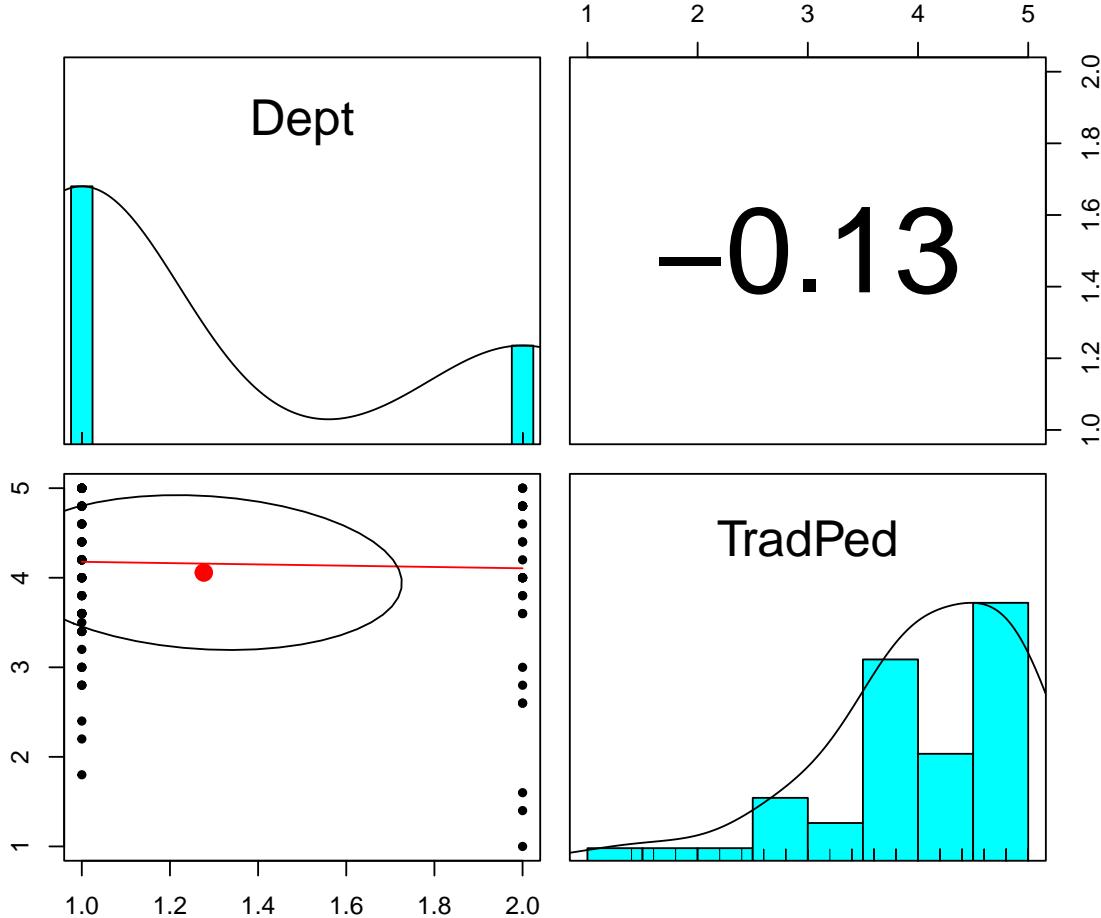
We can use the Shapiro Wilk test for a formal test of normality

```
# A tibble: 2 x 4
  Dept   variable statistic      p
  <fct> <chr>        <dbl>     <dbl>
1 CPY    TradPed     0.918 0.0000731
2 ORG    TradPed     0.851 0.000544
```

The Shapiro-Wilk test of normality indicated that the dependent variable, traditional pedagogy, differed significantly from a normal distribution for both CPY students ($W = 0.918, p < 0.001$) and ORG students($W = 0.851, p < 0.001$).

Should we be concerned? A general rule of thumb is that when cell sizes are larger than 15 the independent t -test should be relatively robust to violations of normality [Green and Salkind, 2017c]. Although there are more CPY than ORG students, we are well-powered.

For fun (not required), let's produce a pairs.panels.



We can see that we'll have more CPY students than ORG students. Although our kurtosis was below $|10|$ our distribution looks negatively skewed, with the majority of the scores being on the high end of the scale.

And now for homogeneity of variance:

```
# A tibble: 1 x 4
  df1   df2 statistic     p
  <int> <int>    <dbl> <dbl>
1     1    110      2.46 0.120
```

Levene's test for homogeneity of variance indicated that we did not violate the assumption of homogeneity of variance ($F[1, 110] = 2.460, p = 0.120$). That is to say, the variance in each of the departments is not statistically significantly different from each other. We can use the regular (Student's) formulation of the t -test for independent samples.

5.9.1.4 Conduct an independent samples t -test (with an effect size and 95%CIs)

Conduct the independent samples t -test (with an effect size)

```
# A tibble: 1 x 16
  estimate estimate1 estimate2 .y.    group1 group2   n1   n2 statistic     p
  <dbl>      <dbl>      <dbl> <chr>  <chr>  <chr> <int> <int>    <dbl> <dbl>
1 0.259       4.13       3.87 TradPed CPY      ORG      81     31      1.42 0.158
# i 6 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>, p.signif <chr>
```

From this output we learn that the value of the t -test is 1.423 and is non-significant $p = 0.148$. We are 95% confident that the mean difference falls between -0.102 and 0.618. Because this threshold crosses zero, we cannot be certain that the true difference in means is not zero. Here's how I would represent these results in a statistical string: $t(110) = 1.423, p = 0.158, CI95(0.102, 0.619)$.

Calculating the Cohen's d as the effect size.

```
# A tibble: 1 x 7
  .y.    group1 group2 effsize   n1   n2 magnitude
  * <chr>  <chr>  <chr>  <dbl> <int> <int> <ord>
1 TradPed CPY      ORG      0.300     81     31 small
```

The value of Cohen's d statistic (interpreted in standard deviation units) is 0.300 and is small. We can add this value to the statistical string: $t(110) = 1.423, p = 0.158, CI95(0.102, 0.619), d = 0.300$.

5.9.1.5 APA style results with table(s) and figure

- Complete content of results (including t , df , p , d -or- η^2 , $CI95\%$)
- Table
- Figure
- Grammar/formatting

An independent samples t -test was conducted to evaluate the hypothesis that there would be differences in course evaluation ratings of traditional pedagogy between academic departments (CPY, ORG).

We began by analyzing the data to see if it met the statistical assumptions for analysis with an independent samples t -test. One assumption is that the dependent variable be normally distributed within the both levels of the grouping variable. We evaluated skew and kurtosis using Kline's [2016a] guidelines of the absolute values of 3 (skew)

and 10 (kurtosis). Our results were well-within these boundary conditions. Specifically, the traditional pedagogy ratings for CPY were -0.763 and 0.156 for skew and kurtosis, respectively; they were -1.183 and 0.583 for ORG. The Shapiro-Wilk test of normality indicated that the dependent variable, traditional pedagogy, differed significantly from a normal distribution for both CPY students ($W = 0.918, p < 0.001$) and ORG students($W = 0.851, p < 0.001$). Levene's test for homogeneity of variance indicated that we did not violate the assumption of homogeneity of variance ($F[1, 110] = 2.460, p = 0.120$). That is to say, the variance in each of the departments is not statistically significantly different from each other. Because the independent samples t -test is relatively robust to violations of normality when samples sizes have at least 15 participants per cell [Green and Salkind, 2017c] and there was no violation of the homogeneity of variance assumption we proceeded with the Student's formulation of the t -test for independent samples.

The independent samples t -test was nonsignificant, $t(110) = 1.423, p = 0.158$, the effect size ($d = 0.300$) was small. The 95% confidence interval for the difference in means ranged from -0.102 to 0.619. Means and standard deviations are presented in Table 1; the results are illustrated in Figure 1.

We can use the *apaTables* package to create a table of means and standard deviations.

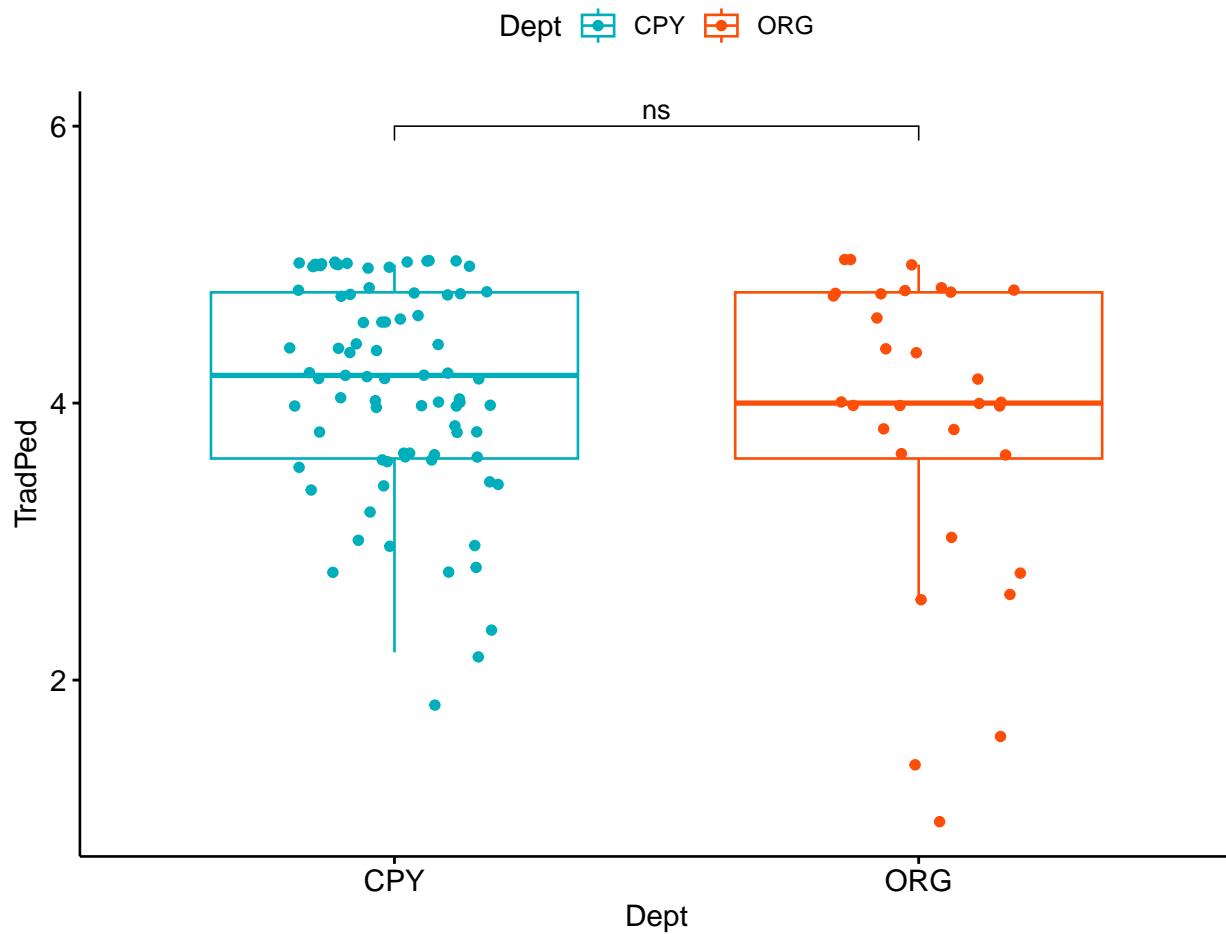
Descriptive statistics for TradPed as a function of Dept.

Dept	M	SD
CPY	4.13	0.75
ORG	3.87	1.09

Note. M and SD represent mean and standard deviation, respectively.

And now a figure.

Figure 1. Traditional Pedagogy as a Function of Academic Department
T test, $t(110) = 1.42$, $p = 0.16$, $n = 112$



5.9.1.6 Conduct power analyses to determine the power of the current study and a recommended sample size

We can use Cohen's d in this specification of d .

```
Two-sample t test power calculation
```

```

n = 112
d = 0.3
sig.level = 0.05
power = 0.6084749
alternative = two.sided
```

NOTE: n is number in *each* group

We were at 61% power. That is, given the value of the mean difference (), we had a 61% chance of detecting a statistically significant effect if there was one. How big of a sample would it take?

```
Two-sample t test power calculation
```

```
n = 175.3847
d = 0.3
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number in *each* group

To find a statistically significant difference, we would need 175 per group. This large size is consistent with the small effect – that there isn't really a difference between the two groups..

5.9.2 Hand Calculations

Note: While the values of the hand-calculations are close to those calculated with the R packages, they differ slightly.

5.9.2.1 Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

5.9.2.2 Using an R package or functions in base R, calculate the means and standard deviations for both levels of the dependent variable

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max
Dept*1	1	CPY	1	81	1.000000	0.0000000	1.0	1.000000	0.00000	1.0	1
Dept*2	2	ORG	1	31	2.000000	0.0000000	2.0	2.000000	0.00000	2.0	2
TradPed1	3	CPY	2	81	4.129630	0.7547259	4.2	4.210769	0.88956	1.8	5
TradPed2	4	ORG	2	31	3.870968	1.0948953	4.0	4.040000	1.18608	1.0	5
	range			skew	kurtosis	se					
Dept*1	0.0			NaN		NaN	0.00000000				
Dept*2	0.0			NaN		NaN	0.00000000				
TradPed1	3.2	-0.7630015	0.1555318		0.08385843						
TradPed2	4.0	-1.1832786	0.5826528		0.19664900						

CPY: $M = 4.130$, $SD = 0.755$ ORG: $M = 3.871$, $SD = 1.095$

5.9.2.3 Calculate the SE used in the denominator of the *t*-test

Just as a reminder, the SE is the denominator in the *t*-test formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

```
[1] 0.2137828
```

The *SE* = 0.214

5.9.2.4 Calculate the independent samples *t*-test

```
[1] 1.209929
```

I note that this hand calculation differs from the worked in R. I believe this is likely due to an unbalanced design with unequal cell sizes (81 and 31).

5.9.2.5 Identify the degrees of freedom associated with your *t*-test

N – 2 is the degrees of freedom: 112-2, df = 110

5.9.2.6 Locate the test critical value for your test

We can look at a [table of critical values](#)

For a two-tailed test, with alpha of 0.05, and a sample size of 120 (close enough), the *t*-statistic must be greater than 1.98.

We could also obtain a *t* critical value with this code:

```
[1] -1.981372
```

```
[1] 1.981372
```

5.9.2.7 Is the *t*-test statistically significant? Why or why not?

In a two-tailed test, if the *t*-statistic falls outside the boundaries of -1.98 and 1.98 the means of the two groups are statistically significantly different from each other.

My *t* value of 1.209929 does not exceed these boundaries and therefore is not statistically significant.

5.9.2.8 Calculate the confidence interval around the difference in sample means

Calculating a confidence interval around the difference in sample means requires the two-tailed test critical values. We can insert them into this formula:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{cv}(SE)$$

```
[1] -0.000000094212
```

```
[1] 0.517324
```

We are 95% confident that the mean difference falls between -0.165 and 0.682. Because this interval passes through zero, we cannot be certain that the difference is 0. This is consistent with the non-significant p value.

5.9.2.9 Calculate the effect size (i.e., Cohen's d associated with your t-test

Here is the formula for Cohen's d :

$$d = t \sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

```
[1] 0.2555321
```

5.9.2.10 Assemble the results into a statistical string

$t(110) = 1.210, p > 0.05, CI95(-0.000, 0.517), d = 0.256$

Chapter 6

Paired Samples *t*-test

[Screencasted Lecture Link](#)

Researchers are often interested in knowing if participants score differently on some outcome variable (like affective well-being) across two conditions. These conditions could be before and after an intervention; they could also be interventionless exposures such as scary versus funny movies. In these simple designs, the paired *t*-test can be used to test the researchers' hypotheses.

6.1 Navigating this Lesson

There is about 45 minutes of lecture. If you work through the materials with me it would be plan for an additional hour

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

6.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Recognize the research questions for which utilization of paired sample *t*-tests would be appropriate.
- Narrate the steps in conducting a paired samples *t*-test, beginning with testing the statistical assumptions through writing up an APA style results section.
- Calculate a paired samples *t*-test in R (including effect sizes).
- Interpret a 95% confidence interval around a mean difference score.
- Produce an APA style results for a paired-samples *t*-test.
- Determine a sample size that (given a set of parameters) would likely result in a statistically significant effect, if there was one.

6.1.2 Planning for Practice

The suggestions for homework vary in degree of complexity. The more complete descriptions at the end of the chapter follow these suggestions.

- Rework the paired samples *t*-test in the lesson by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- Rework the paired samples *t*-test in the lesson by changing something else about the simulation. For example, if you are interested in power, consider changing the sample size.
- Use the simulated data that is provided, but use the nonverbal variable, instead.
- Conduct paired *t*-test with data to which you have access and permission to use. This could include data you simulate on your own or from a published article.

6.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- How to Do Paired T-test in R: The Best Tutorial You Will Love. (n.d.). Datanovia. Retrieved May 25, 2023, from <https://www.datanovia.com/en/lessons/how-to-do-a-t-test-in-r-calculation-and-reporting/how-to-do-paired-t-test-in-r/>
 - This tutorial provides a demonstration of the paired samples *t*-test using the *rstatix* package.
- Navarro, D. (2020). Chapter 13: Comparing two means. In [Learning Statistics with R - A tutorial for Psychology Students and other Beginners](https://stats.libretexts.org/Bookshelves/Bookshelves/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)). Retrieved from [https://stats.libretexts.org/Bookshelves/Bookshelves/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Bookshelves/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro))
 - Navarro's OER includes a good mix of conceptual information about *t*-tests as well as R code. My lesson integrates her approach as well as considering information from Field's [2012] and Green and Salkind's [Green and Salkind, 2017c] texts (as well as searching around on the internet).
- Elliott, A. M., Alexander, S. C., Mescher, C. A., Mohan, D., & Barnato, A. E. (2016). Differences in Physicians' Verbal and Nonverbal Communication With Black and White Patients at the End of Life. *Journal of Pain and Symptom Management*, 51(1), 1–8. <https://doi.org/10.1016/j.jpainsympman.2015.07.008>
 - The source of our research vignette.

6.1.4 Packages

The script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(psych)){install.packages('psych')}
# if(!require(faux)){install.packages('faux')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(dplyr)){install.packages('dplyr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(pwr)){install.packages('pwr')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(knitr)){install.packages('knitr')}
# if(!require(rstatix)){install.packages('rstatix')}
```

6.2 Introducing the Paired Samples *t*-test

There are a couple of typical use cases for the paired samples *t*-test. Repeated measures or change-over-time is a very common use. In this case, the research participant may take a pre-test, be exposed to an intervention or other type of stimulus, then take a post-test. Owing to the limitations of the statistics, all participants must be exposed to the same intervention/stimulus.



Figure 6.1: An image of a row with three boxes: pre-test (in blue), intervention or exposure to stimulus (in light red), post-test (in blue) representing the use of a paired samples *t*-test in a repeated measures design

A second common use is the assessment of a research participant in two competing conditions. An example might be the galvanic skin response ratings when a participant's hand is submerged in ice versus the GSR ratings when the hand is not exposed in ice. A strength of this design is the within-subjects' control of the participant.



Figure 6.2: An image of a row with two boxes labeled Condition A (light blue) and Condition B (dark blue). This represents the use of a paired samples *t*-test to compare across conditions

In the formula for the paired samples *t*-test we see a \bar{D} in the numerator. This represents the *difference* between the continuously scaled scores in the two conditions. The denominator involves a standard deviation of the difference scores ($\hat{\sigma}_D$) and the square root of the sample size.

$$t = \frac{\bar{D}}{\hat{\sigma}_D / \sqrt{N}}$$

Although these types of research design and analyses are quite handy, they have some limitations. First, the paired samples *t*-test cannot establish causality because it lacks elements such as comparing conditions (e.g., treatment vs. control) and random assignment to those conditions. If a

research wants to compare pre-post change as a result of participating in more-than-one condition, a **mixed design ANOVA** would be a better option. Second, the paired samples *t*-test cannot accommodate more than two comparison conditions. If the researcher wants to compare three or more time periods or conditions, they will want to consider **repeated measures ANOVA** or **multilevel/hierarchical linear modeling**.

6.2.1 Workflow for Paired Samples *t*-test

The following is a proposed workflow for conducting the paired samples *t*-test.

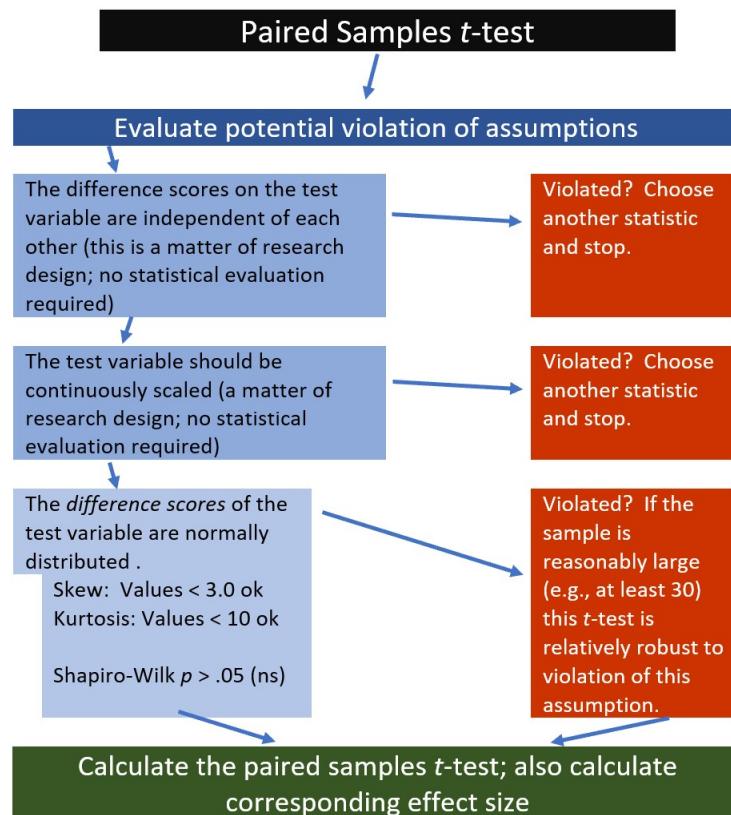


Figure 6.3: A colorful image of a workflow for the paired samples *t*-test

If the data meets the assumptions associated with the research design (e.g., independence of difference scores and a continuously scaled metric for that difference score), these are the steps for the analysis of an independent samples *t*-test:

1. Prepare (upload) data.
2. Explore data with
 - graphs
 - descriptive statistics
3. Assess normality of the difference scores via skew and kurtosis

4. Compute the paired samples t -test
5. Compute an effect size (frequently the d or η^2 statistic)
6. Manage Type I error
7. Sample size/power analysis (which you should think about first, but in the context of teaching statistics, it's more pedagogically sensible, here).

6.3 Research Vignette

Empirically published articles where t -tests are the primary statistic are difficult to locate. Having exhausted the psychology archives, I located this article in an interdisciplinary journal focused on palliative medicine. The research vignette for this lesson examined differences in physician's verbal and nonverbal communication with Black and White patients at the end of life [Elliott et al., 2016].

Elliott and colleagues [2016] were curious to know if hospital-based physicians (56% White, 26% Asian, 7.4% each Black and Hispanic) engaged in verbal and nonverbal communication differently with Black and White patients. Black and White patient participants were matched on characteristics deemed important to the researchers (e.g., critically and terminally ill, prognostically similar). Interactions in the intensive care unit were audio and video recorded and then coded on dimensions of verbal and nonverbal communication.

Because each physician saw a pair of patients (i.e., one Black patient and one White patient), the researchers utilized a paired samples, or dependent t -test. This statistical choice was consistent with the element of the research design that controlled for physician effects through matching patients on critical characteristics. Below are the primary findings of the study.

	Black Patients	White Patients	
Category	$Mean(SD)$	$Mean(SD)$	p -value
Verbal skill score (range 0 - 27)	8.37(3.36)	8.41(3.21)	0.958
Nonverbal skill score (range 0 - 5)	2.68(.84)	2.93(.77)	0.014

The primary analysis utilized by Elliott and colleagues [2016] was the paired samples t -test. We will replicate that exact analysis with simulated data.

6.3.1 Simulating Data for the Paired Samples t -test

Below is the code I used to simulate the data. The following code assumes 33 physician participants who had separate interactions with critically ill, end-of-life stage patients, who were identified as Black and White. The Elliott et al. [2016] manuscript describe the process for coding verbal and nonverbal communication for video/audio recordings of the physician/patient interactions. Using that data, I simulate verbal and nonverbal communication scores for 33 physicians who rate patients who identify as Black and White, respectively. This creates four variables.

In the lesson, we will compare verbal communication scores. The nonverbal communication score is available as an option for practice.

```

library(tidyverse)

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.2     v readr      2.1.4
v forcats   1.0.0     v stringr    1.5.0
v ggplot2   3.4.2     v tibble     3.2.1
v lubridate 1.9.2     v tidyrr     1.3.0
v purrr     1.0.1

-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become

# Setting the seed. If you choose this practice option, change the
# number below to something different.
set.seed(220817)

# These define the characteristics of the verbal variable. It is
# essential that the object names (e.g., A_mean) are not changed
# because they will be fed to the function in the faux package.
sub_n <- 33
A_mean <- 8.37
B_mean <- 8.41
A_sd <- 3.36
B_sd <- 3.21
AB_r <- 0.3

# the faux package can simulate a variety of data. This function
# within the faux package will use the objects above to simulate
# paired samples data
paired_V <- faux::rnorm_multi(n = sub_n, vars = 2, r = AB_r, mu = c(A_mean,
  B_mean), sd = c(A_sd, B_sd), varnames = c("Verbal_BL", "Verbal_WH"))

paired_V <- paired_V %>%
  dplyr::mutate(PhysID = row_number())

# Here, I repeated the process for the nonverbal variable.
sub_n <- 33
A_mean <- 2.68
B_mean <- 2.93
A_sd <- 0.84
B_sd <- 0.77
AB_r <- 0.9

paired_NV <- faux::rnorm_multi(n = sub_n, vars = 2, r = AB_r, mu = c(A_mean,
  B_mean), sd = c(A_sd, B_sd), varnames = c("NVerb_BL", "NVerb_WH"))

# This code produced an ID number for each physician

```

```

paired_NV <- paired_NV %>%
  dplyr::mutate(PhysID = row_number())

# This data joined the two sets of data. Note, I did not write any
# code that assumed tha the verbal and nonverbal data came from the
# same physician. Full confession: I'm not quite sure how to do that
# just yet.
dfPairedSamples <- dplyr::full_join(paired_V, paired_NV, by = c("PhysID"))
dfPairedSamples <- dfPairedSamples %>%
  dplyr::select(PhysID, everything())

```

Before beginning our analysis, let's check the format of the variables to see if they are consistent with the scale of measurement of the variables. In our case, we expect to see four variables representing the verbal and nonverbal communication of the physicians with the patients who are identified as Black and White. Each of the variables should be continuously scaled and, therefore, should be formatted as *num* (numerical).

```
str(dfPairedSamples)
```

```
'data.frame': 33 obs. of 5 variables:
 $ PhysID   : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Verbal_BL: num 8.19 3.3 6.18 4.85 6.91 ...
 $ Verbal_WH: num 4.63 12.85 13.47 6.49 12.27 ...
 $ NVerb_BL : num 3.099 4.234 0.429 1.835 3.704 ...
 $ NVerb_WH : num 2.74 5.02 1.34 2.38 2.91 ...
```

The four variables of interest are correctly formatted as *num*. Because PhysID (physician ID) will not be used in our analysis, its structure is irrelevant.

Below is code for saving (and then importing) the data in .csv or .rds files. I make choices about saving data based on what I wish to do with the data. If I want to manipulate the data outside of R, I will save it as a .csv file. It is easy to open .csv files in Excel. A limitation of the .csv format is that it does not save any restructuring or reformatting of variables. For this lesson, this is not an issue.

Here is code for saving the data as a .csv and then reading it back into R. I have hashtagged these out, so you will need to remove the hashtags if you wish to run any of these operations.

```
# writing the simulated data as a .csv write.table(dfPairedSamples,
# file = 'dfPairedSamples.csv', sep = ',', col.names=TRUE,
# row.names=FALSE) at this point you could clear your environment and
# then bring the data back in as a .csv reading the data back in as a
# .csv file dfPairedSamples<- read.csv ('dfPairedSamples.csv', header
# = TRUE)
```

The .rds form of saving variables preserves any formatting (e.g., creating ordered factors) of the data. A limitation is that these files are not easily opened in Excel. Here is the hashtagged code (remove hashtags if you wish to do this) for writing (and then reading) this data as an .rds file.

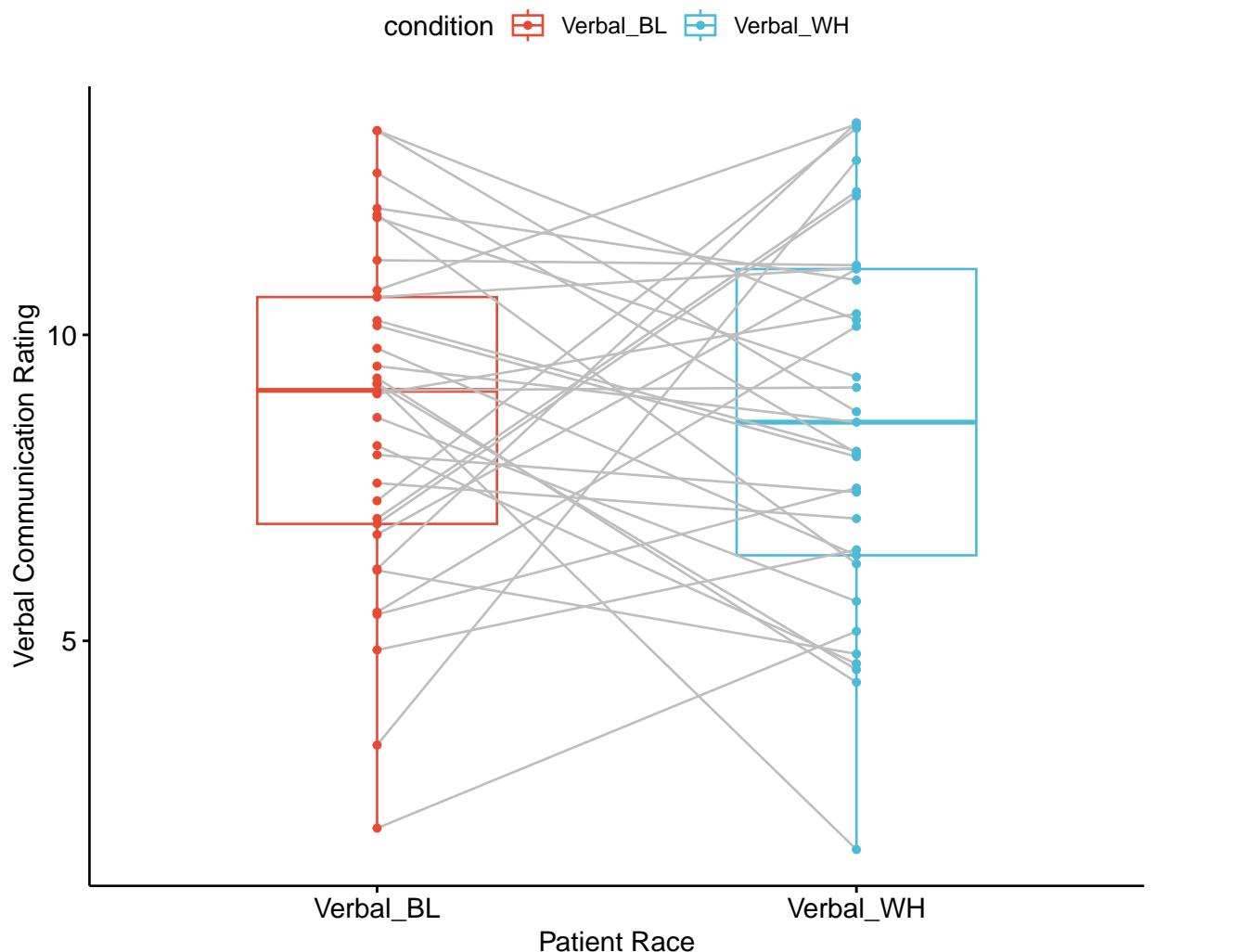
```
# saveRDS(dfPairedSamples, 'dfPairedSamples.rds') dfPairedSamples <-
# readRDS('dfPairedSamples.rds')
```

6.3.2 Quick Peek at the Data

Plotting the data is a helpful early step in any data analysis. Further, visualizing the data can help us with a conceptual notion of the statistic we are utilizing. The *ggpubr* package is one of my go-to-tools for quick and easy plots of data. The *ggpaired()* function is especially appropriate for paired data. A [tutorial](#) is available at datanovia.

Especially unique about this function is that the lines connect the scores of each person across time or conditions. In this research scenario, the lines present the amount of time the physicians spent with each of the two patients they treated.

```
ggpubr::ggpaired(dfPairedSamples, cond1 = "Verbal_BL", cond2 = "Verbal_WH",
  color = "condition", line.color = "gray", palette = c("npg"), xlab = "Patient Race",
  ylab = "Verbal Communication Rating")
```



The box of the boxplot covers the middle 50% (the interquartile range). The horizontal line is the median. The whiskers represent three standard deviations above and below the mean. Any dots beyond the whiskers are outliers.

6.4 Working the Paired Samples *t*-Test (by hand)

6.4.1 Stating the Hypothesis

In this lesson, I will focus on differences in the verbal communication variable. Specifically, I hypothesize that physician verbal communication scores for Black and White patients will differ. In the hypotheses below, the null hypothesis (μ_D) states that the difference score is zero; the alternative hypothesis (μ_D) states that the difference score is different from zero.

$$\begin{aligned} H_O : \mu_D &= 0 \\ H_A : \mu_D &\neq 0 \end{aligned}$$

Notice the focus on a *difference* score. Even though the R package we will use does not require one for calculation, creating one in our df will be useful for preliminary exploration.

```
# Creating the Verbal_D variable within the dfPairedSamples df Doing
# the 'math' that informs that variable
dfPairedSamples$Verbal_D <- (dfPairedSamples$Verbal_BL - dfPairedSamples$Verbal_WH)
# Displaying the first six rows of the df to show that the difference
# score now exists
head(dfPairedSamples)
```

	PhysID	Verbal_BL	Verbal_WH	NVerb_BL	NVerb_WH	Verbal_D
1	1	8.190342	4.625680	3.0991101	2.742055	3.564663
2	2	3.297486	12.851362	4.2338398	5.024047	-9.553876
3	3	6.176386	13.466880	0.4288566	1.337259	-7.290495
4	4	4.851426	6.488762	1.8347393	2.379431	-1.637336
5	5	6.911155	12.266646	3.7035910	2.914445	-5.355491
6	6	11.965831	6.259292	1.5369696	1.598493	5.706540

Examining this new variable, because we subtracted the verbal communication ratings of physicians with White patients from those of Black patients a negative score means that physicians had lower verbal engagement with Black patients; a positive score means that physicians had more verbal engagement with White patients.

6.4.2 Calculating the Paired Samples *t*-Test

Let's take another look at the formula for calculating paired samples *t*-test.

$$t = \frac{\bar{D}}{\hat{\sigma}_D / \sqrt{N}}$$

We can use the data from our preliminary exploration in the calculation.

- The mean difference was .08
- The standard deviation of that difference was 4.14
- The sample size is 33

```
0.08/(4.14/sqrt(33))
```

```
[1] 0.111006
```

The resultant t value is 0.111

Hopefully, this hand-calculation provided an indication of how the means, standard deviation, and sample sizes contribute to the estimate of this t -test value. Now we ask, “But it is statistically significant?”

6.4.2.1 Statistical Significance

Our t -value was 0.111. We compare this value to the test critical value in a table of t critical values. In-so-doing we must know our degrees of freedom. Because the numerator in a paired samples t -test is a single difference score \bar{D} , the associated degrees of freedom is $N - 1$. We must also specify the p value (in our case .05) and whether-or-not our hypothesis is unidirectional or bi-directional. Our question only asked, “Are the verbal communication levels different?” In this case, the test is two-tailed, or bi-directional.

Let’s return to the [table of critical values](#) for the t distribution to compare our t -value (0.111) to the column that is appropriate for our:

- Degrees of freedom (in this case $N - 1$ or 32)
- Alpha, as represented by $p < .05$
- Specification as a one-tailed or two-tailed test
 - Our alternative hypothesis made no prediction about the direction of the difference; therefore we will use a two-tailed test

In the linked table, when the degrees of freedom reaches 30, there larger intervals. We will use the row representing degrees of freedom of 30. If our t -test value is lower than an absolute value of -2.042 or greater than the absolute value of 2.042, then our means are statistically significantly different from each other. In our case, we have not achieved statistical significance and we cannot say that the means are different. The t string would look like this: $t(32) = 0.111, p > .05$

We can also use the `qt()` function in base R. In the script below, I have indicated an alpha of .05. The “2” that follows indicates I want a two-tailed test. The 32 represents my degrees of freedom ($N - 1$). In a two-tailed test, the regions of rejection will be below the lowerbound (lower.tail=TRUE) and above the upperbound (lower.tail=FALSE).

```
qt(0.05/2, 32, lower.tail = TRUE)
```

```
[1] -2.036933
```

```
qt(0.05/2, 32, lower.tail = FALSE)
```

```
[1] 2.036933
```

If our t value is below the lowerbound (-2.04) or above the upper bound (2.04), then we have rejected the null hypothesis in favor of the alternative. As we demonstrated in the hand-calculations, we have not. The ratings of physicians' verbal engagement with patients who are racially identified as Black and White are not statistically significant.

6.4.2.2 Confidence Intervals

How confident are we in our result? With paired samples t -tests, it is common to report an interval in which we are 95% confident that our true mean difference exists. Below is the formula, which involves:

- \bar{D} the mean difference score
- t_{cv} the test critical value for a two-tailed model (even if the hypothesis was one-tailed) where $\alpha = .05$ and the degrees of freedom are $N - 1$
- s_d the standard deviation of \bar{D}
- N sample size

$$\bar{D} \pm t_{cv}(s_d/\sqrt{n})$$

Let's calculate it:

First, let's get the proper t critical value:

```
qt(0.05/2, 32, lower.tail = TRUE)
```

```
[1] -2.036933
```

```
qt(0.05/2, 32, lower.tail = FALSE)
```

```
[1] 2.036933
```

```
0.08 - (2.037 * ((4.14/(sqrt(33)))))
```

```
[1] -1.388028
```

```
0.08 + (2.037 * ((4.14/sqrt(33))))
```

```
[1] 1.548028
```

These values indicate the range of scores in which we are 95% confident that our true \bar{D} lies. Stated another way, we are 95% confident that the true mean difference lies between -1.39 and 1.55. Because this interval crosses zero, we cannot rule out that the true mean difference is 0.00. This result is consistent with our non-significant p value. For these types of statistics, the 95% confidence interval and p value will always be yoked together.

6.4.2.3 Effect Size

Effect sizes address the magnitude of difference. There are two common effect sizes that are used with the paired samples t -test. The first is the d statistic, which measures, in standard deviation units, the distance between the two means. Regardless of sign, values of .2, .5, and .8 are considered to be small, medium, and large, respectively.

Because the paired samples t -test used the difference score in the numerator, there are two easy options for calculating this effect:

$$d = \frac{\bar{D}}{\hat{\sigma}_D} = \frac{t}{\sqrt{N}}$$

The first is to use the mean and standard deviation associated with the difference score:

```
0.08/4.14
```

```
[1] 0.01932367
```

The formula uses the t value and N .

```
0.111/(sqrt(33))
```

```
[1] 0.01932262
```

Within rounding error, both calculations result in a value ($d = 0.02$) that is quite small.

Eta squared, η^2 is the proportion of variance of a test variable that is a function of the grouping variable. A value of 0 indicates that mean of the difference scores is equal to 0, where a value of 1 indicates that the difference scores in the sample are all the same nonzero value, and the test scores do not differ within each group. The following equation can be used to compute η^2 . Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

$$\eta^2 = \frac{N(\bar{D}^2)}{N(\bar{D}^2 + (N-1)(\hat{\sigma}_D^2)} = \frac{t^2}{t^2 + (N_1 - 1)}$$

The first calculation option uses the N and the mean difference score:

```
(33 * (0.08^2))/((33 * (0.08^2)) + ((33 - 1) * (4.14^2)))
```

```
[1] 0.0003849249
```

The second calculation option uses the t values and sample size:

```
(0.111^2)/((0.111^2) + (33 - 1))
```

```
[1] 0.0003848831
```

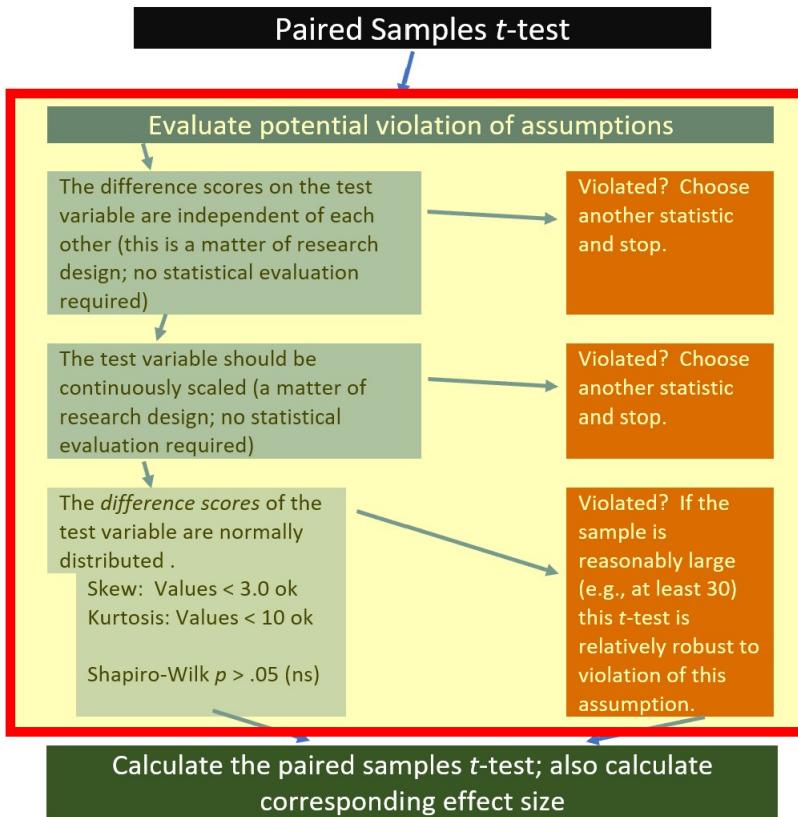
Within rounding errors, and similar to our d statistic, the η^2 value (0.0004) is quite small.

6.5 Working the Paired Samples t -Test with R Packages

Let's rework the problem in R. We start at the top of the flowchart, evaluating the statistical assumptions.

6.5.1 Evaluating the Statistical Assumptions

With an eye on our data, we can begin to explore the statistical assumptions associated with the paired samples t -test. Here's where we are in the workflow:



All statistical tests have some as-

sumptions about the data. The paired-samples *t*-test has three:

- The difference scores (i.e., the difference on the outcome across time or conditions) on the test variable are independent of each other. This is a matter of research design and no further statistical evaluation is required.
- The test variable should be continuously scaled. This is also a matter of research design and no statistical analysis is required.
 - Our test variable is measured in minutes; this is continuously scaled and has the properties of *interval*-level data.
- The *difference scores* of the test variable are normally distributed. We can check this several ways:
 - visually with histograms (perhaps with superimposed curves) and boxplots,
 - calculation of skew and kurtosis values,
 - calculation of the Shapiro-Wilk test of normality

6.5.1.1 Are the difference scores of the test variable normally distributed?

We can begin to evaluate normality by obtaining the descriptive statistics with the *describe()* function from the *psych* package.

```
psych::describe(dfPairedSamples, type = 1)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
PhysID	1	33	17.00	9.67	17.00	17.00	11.86	1.00	33.00	32.00	0.00
Verbal_BL	2	33	8.70	2.80	9.09	8.80	3.10	1.94	13.34	11.40	-0.35
Verbal_WH	3	33	8.62	3.08	8.57	8.65	3.44	1.59	13.47	11.88	-0.15
NVerb_BL	4	33	2.73	1.00	2.63	2.78	1.26	0.43	4.23	3.80	-0.37
NVerb_WH	5	33	2.89	0.85	2.94	2.87	0.64	1.34	5.02	3.69	0.25
Verbal_D	6	33	0.08	4.14	0.61	0.27	4.11	-9.55	7.61	17.17	-0.42
	kurtosis	se									
PhysID	-1.20	1.68									
Verbal_BL	-0.31	0.49									
Verbal_WH	-0.75	0.54									
NVerb_BL	-0.71	0.17									
NVerb_WH	-0.02	0.15									
Verbal_D	-0.54	0.72									

We observe that the skew and kurtosis values for Verbal_BL and Verbal_WH are well below the areas of concern (below the absolute value of 3 for skew; below the absolute values of 10 for kurtosis) identified by Kline [2016a].

Recall, though that the normality assumption for the paired samples *t*-test concerns the *difference score* (Verbal_D). We see that the mean difference is 0.08 (*SD* = 4.14). Its skew (-0.42) and kurtosis (-0.54) are also well-below the thresholds of concern.

Beyond skew and kurtosis, we can formally test for deviations from normality with a Shapiro-Wilk. We want the results to be non-significant.

```
rstatix::shapiro_test(dfPairedSamples, Verbal_D)
```

```
# A tibble: 1 x 3
  variable statistic     p
  <chr>      <dbl> <dbl>
1 Verbal_D    0.973 0.572
```

Results of the Shapiro-Wilk test of normality are not statistically significant ($W = 0.97, p = 0.57$). This means that the distribution of difference scores is not statistically significantly different from a normal distribution.

6.5.1.2 APA style write-up of testing the assumptions

My practice is to create APA style drafts of the different sections of the analysis as I work along. Here's how I might capture our evaluation of the statistical assumptions:

We began by analyzing the data to see if it met the statistical assumptions for analysis with a paired samples t -test. One assumption is that the difference scores of dependent variable are normally distributed. We evaluated skew and kurtosis using Kline's [2016a] guidelines of the absolute values of 3 (skew) and 10 (kurtosis). Our results were well-within these boundary conditions. Further, a non-significant Shapiro-Wilk test of normality suggested that the distribution of difference scores was not statistically significant from a normal distribution ($W = 0.97, p = 0.57$).

6.5.2 Computing the Paired Samples t -Test

We are ready to compute the paired samples t -test.

Calculating a paired samples t -test is possible through base R and a number of packages. Kassambara's [b] *rstatix* package is one we can use for all of the t -test and ANOVA problems that we will work.

A challenge in evaluating within-persons data is the *shape* of the data. The simulation resulted in a *wide* (also termed person-level or multivariate) format, where each of the 33 physicians has the verbal and non-verbal communication scores with Black and White patients. We need to reshape the data to a long (also termed person-period or univariate) format. Although it may seem a bit tricky at first, this is a skill you will return to in many longitudinal or repeated measures analyses.

In the script below we are using the *melt()* and *setDT* functions from the *data.table* package. We put stable (i.e., time-invariant, "one-per-person") variables in a concatenated variable list of "id.vars." We create separate lists of the variables that change over time. In this case, each physician saw one Black patient and one White patient. Therefore, every physician will have two rows of data. For each variable collected at both points, we create concatenated lists.

```
df_long <- data.table::melt(data.table::setDT(dfPairedSamples), id.vars = c("PhysID"),
  measure.vars = list(c("Verbal_BL", "Verbal_WH"), c("NVerb_BL", "NVerb_WH")))
head(df_long)
```

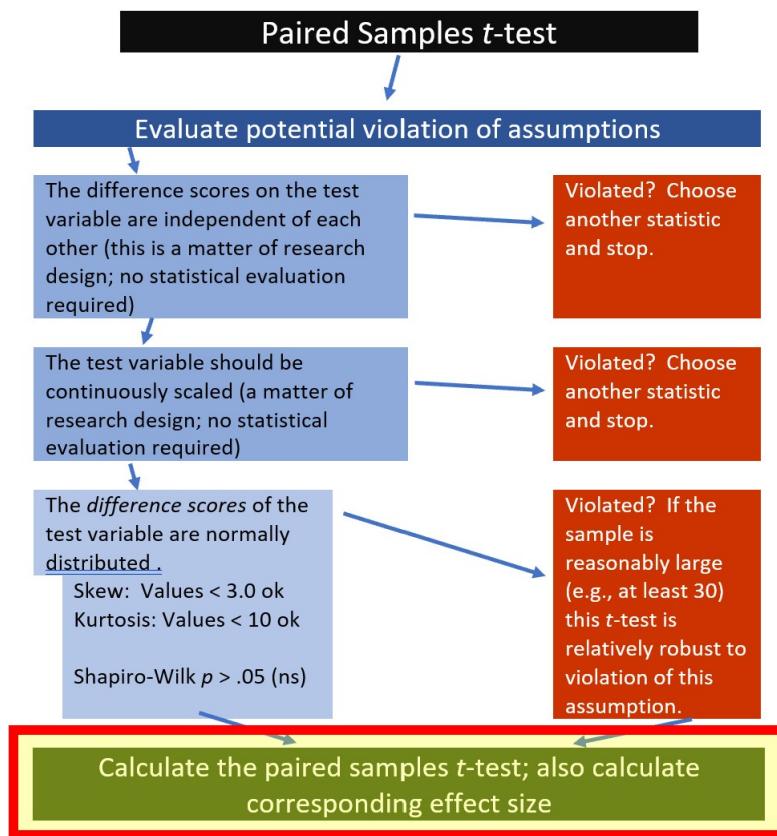


Figure 6.4: A colorful image of a workflow for the paired samples *t*-test

```

PhysID variable    value1    value2
1:      1      1  8.190342 3.0991101
2:      2      1  3.297486 4.2338398
3:      3      1  6.176386 0.4288566
4:      4      1  4.851426 1.8347393
5:      5      1  6.911155 3.7035910
6:      6      1 11.965831 1.5369696

```

While that code performed the magic, it did not name the variables. We must provide that in separate code.

```
df_long <- rename(df_long, PatientRace = variable, Verbal = value1, Nonverbal = value2)
```

After the reshaping, let's recheck the structure of our data:

```
str(df_long)
```

```

Classes 'data.table' and 'data.frame': 66 obs. of 4 variables:
 $ PhysID     : int 1 2 3 4 5 6 7 8 9 10 ...
 $ PatientRace: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Verbal      : num 8.19 3.3 6.18 4.85 6.91 ...
 $ Nonverbal   : num 3.099 4.234 0.429 1.835 3.704 ...
 - attr(*, ".internal.selfref")=<externalptr>

```

The dependent variables Verbal and Nonverbal are continuously scaled, so the *num* designation is appropriate. Similarly, PatientRace is categorical, so *Factor* is appropriate. Because labels (instead of numbers) can minimize misinterpretation (or forgetting), I would prefer to use “Black” and “White” as opposed to “1” and “2”. To further reduce the possibility of error, it is easy enough to create a second, parallel, variable.

```
df_long$PtRace <- plyr::mapvalues(df_long$PatientRace, from = c(1, 2),
                                     to = c("Black", "White"))
```

We are now ready to perform the paired samples *t*-test. In the script below:

- the first element points to the dataframe
- the second element provides a “formula”
 - we are predicting “Verbal” from “PtRace”
- specifying “detailed = TRUE” will produce the 95% confidence interval around the difference

```
rstatix::t_test(df_long, Verbal ~ PtRace, paired = TRUE, detailed = TRUE)
```

```
# A tibble: 1 x 13
  estimate .y. group1 group2   n1   n2 statistic     p    df conf.low
* <dbl> <chr> <chr> <chr> <int> <int>    <dbl> <dbl> <dbl> <dbl>
1 0.0813 Verbal Black White     33    33     0.113 0.911    32   -1.39
# i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

This output provides information to get us started in drafting the APA style results. Identical to all the information we hand-calculated, we would write the *t* string this way: $t(32) = 0.113, p = .911, CI95(-1.39, 1.55)$. Our results show that the mean difference in physician verbal communication scores with Black and White patients was 0.081. Taking a look at the confidence interval, we are 95% confident that the true difference in means falls between the values of -1.386 and 1.549. What is critically important is that this confidence interval crosses zero. There is an important link between the CI95% and statistical significance. When the CI95% includes zero, *p* will not be lower than 0.05.

We still need to calculate the effect size.

```
rstatix::cohens_d(df_long, Verbal ~ PtRace, paired = TRUE)
```

```
# A tibble: 1 x 7
  .y.   group1 group2 effsize   n1   n2 magnitude
* <chr> <chr> <chr>    <dbl> <int> <int> <ord>
1 Verbal Black White    0.0196    33    33 negligible
```

Keeping in mind the interpretive criteria of .2, .5, and .8, as small, medium, and large effect sizes, we see that $d = 0.020$ is quite small. We can add it to our *t*-string and draft the results: $t(32) = 0.113, p = .911, d = 0.020, CI95(-1.39, 1.55)$.

6.6 APA Style Results

Putting it altogether we can assemble an APA style results section. Code for a table of means, standard deviations, and correlation follow the write-up of results. For inclusion in a manuscript, I would rework the export of the table to delete the difference score (i.e., Verbal_D). I might also exclude the rows of confidence intervals around the correlations.

We began by analyzing the data to see if it met the statistical assumptions for analysis with a paired samples *t*-test. One assumption is that the difference scores of dependent variable are normally distributed. We evaluated skew and kurtosis using Kline's [2016a] guidelines of the absolute values of 3 (skew) and 10 (kurtosis). Our results were well-within these boundary conditions. Further, a non-significant Shapiro-Wilk test of normality suggested that the distribution of difference scores was not statistically significant from a normal distribution ($W = 0.97, p = 0.57$).

A paired samples *t*-test was conducted to evaluate the hypothesis that there would be differences in the degree of physicians' verbal engagement as a function of the patient's

race (Black, White). The paired samples t -test was nonsignificant, $t(32) = 0.133$, $p = .911$. The small magnitude of the effect size ($d = 0.02$) was consistent with the nonsignificant result. The 95% confidence interval for the difference in means was quite wide and included the value of zero (95%CI[-1.386, 1.549]). Means and standard deviations are presented in Table 1; the results are illustrated in Figure 1.

```
library(tidyverse) #needed to use the pipe
# Creating a smaller df to include only the variables I want in the
# table
PairedDescriptors <- dfPairedSamples %>%
  select(Verbal_BL, Verbal_WH, Verbal_D)
# using the apa.cor.table function for means, standard deviations,
# and correlations the filename command will write the table as a
# word document to your file
apaTables::apa.cor.table(PairedDescriptors, table.number = 1, filename = "Tab1_PairedV.doc")
```

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1	2
1. Verbal_BL	8.70	2.80		
2. Verbal_WH	8.62	3.08	.01 [-.33, .35]	
3. Verbal_D	0.08	4.14	.67** [.42, .82]	-.74** [-.86, -.53]

Note. M and SD are used to represent mean and standard deviation, respectively.

Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

The figure we created earlier in the lesson would be sufficient for a journal article. However, using *rstatix* in combination with *ggbubbr* can be quite powerful. The result can be a figure that includes the t -test results and “significance bars.” To do this, we first need to re-run the *rstatix::t_test*, but adding to it by

- including “add_significance()” script after the pipe, and
- saving it as an object, which I’m naming “pair.test.”

We could have done this in the initial run (but I didn't want to make the test-statistic unnecessarily confusing).

```
library(tidyverse)
pair.test <- rstatix::t_test(df_long, Verbal ~ PtRace, paired = TRUE, detailed = TRUE) %>%
  rstatix::add_significance()
pair.test

# A tibble: 1 x 14
  estimate .y.   group1 group2     n1     n2 statistic      p     df conf.low
  <dbl> <chr> <chr> <chr> <int> <int>    <dbl> <dbl> <dbl> <dbl>
1 0.0813 Verbal Black White     33     33     0.113 0.911    32    -1.39
# i 4 more variables: conf.high <dbl>, method <chr>, alternative <chr>,
#   p.signif <chr>
```

Next, we create boxplot code with the long form of our data:

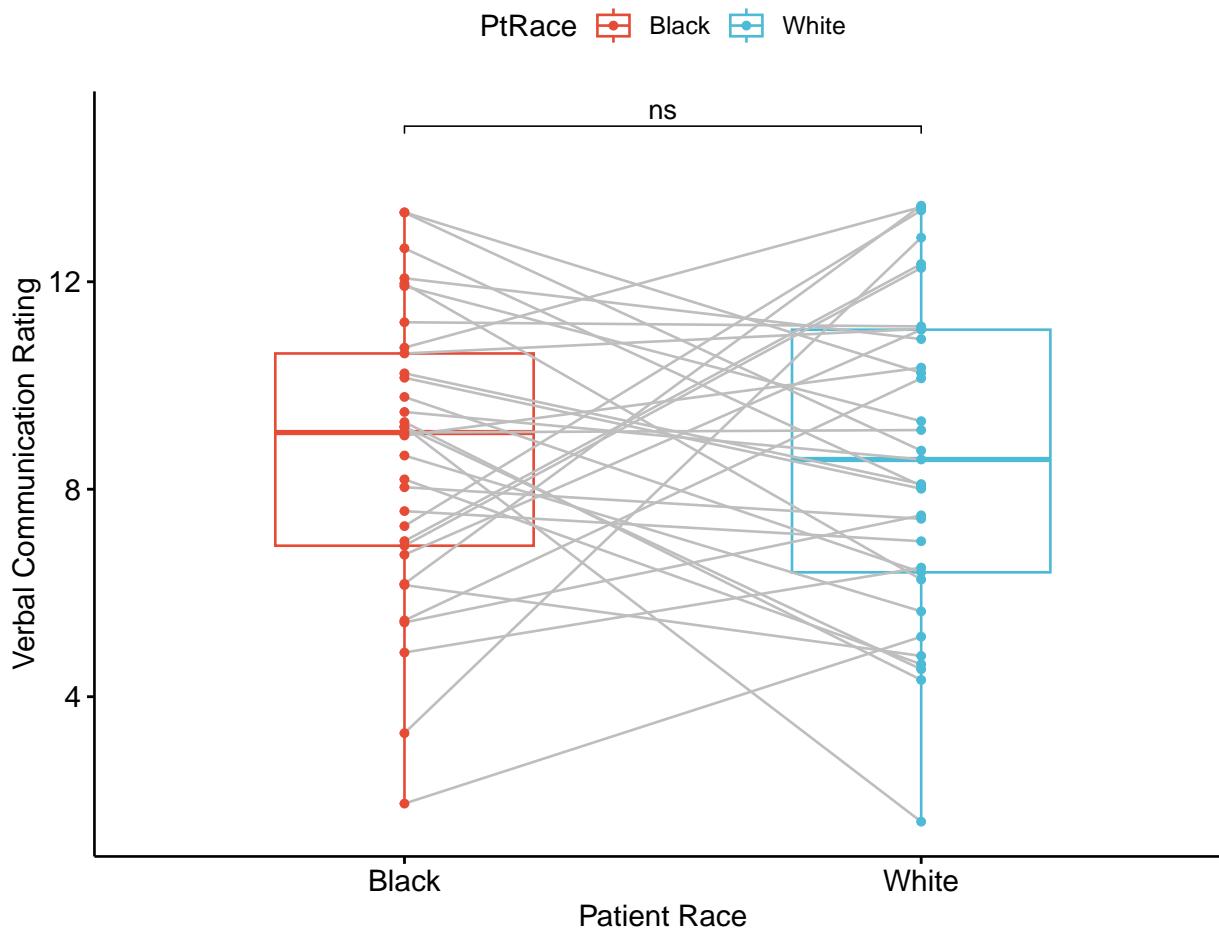
```
pair.box <- ggpubr::ggpaired(df_long, x = "PtRace", y = "Verbal", order = c("Black",
  "White"), line.color = "gray", palette = c("npg"), color = "PtRace",
  ylab = "Verbal Communication Rating", xlab = "Patient Race", title = "Figure 1. Physician

pair.test <- pair.test %>%
  rstatix::add_xy_position(x = "PtRace") #autocomputes p-value labels positions

pair.box <- pair.box + ggpubr::stat_pvalue_manual(pair.test, tip.length = 0.01,
  y.position = c(15)) + labs(subtitle = rstatix::get_test_label(pair.test,
  detailed = TRUE))

pair.box
```

Figure 1. Physician Verbal Engagement as a Function of Patient Race
 T test, $t(32) = 0.11$, $p = 0.91$, $n = 33$



The tools available offer a great deal of flexibility. Determining which figure is best will likely depend on your outlet, your audience, and your personal preferences. For example, a print journal might prefer a black-and-white figure (with no fill in the boxes). This is accomplished easily enough by removing (or, hashtagging out) the “fill = PtRace” argument.

6.7 Power in Paired Samples t -Tests

Researchers often use power analysis packages to estimate the sample size needed to detect a statistically significant effect, if, in fact, there is one. Utilized another way, these tools allows us to determine the probability of detecting an effect of a given size with a given level of confidence. If the probability is unacceptably low, we may want to revise or stop. A helpful overview of power as well as guidelines for how to use the *pwr* package can be found at a [Quick-R website](#) [Kabacoff, 2017].

In Champely’s *pwr* package, we can conduct a power analysis for a variety of designs, including the paired t -test that we worked in this chapter. There are a number of interrelating elements of power:

- Sample size, n refers to the number of pairs; our vignette had 33
- d refers to the difference between means divided by the pooled standard deviation; we can use the value of Cohen's d for this
- $power$ refers to the power of a statistical test; conventionally it is set at .80
- $sig.level$ refers to our desired alpha level; conventionally it is set at .05
- $type$ indicates the type of test we ran; ours was "paired"
- $alternative$ refers to whether the hypothesis is non-directional/two-tailed ("two.sided") or directional/one-tailed("less" or "greater")

In this script, we must specify *all-but-one* parameter; the remaining parameter must be defined as `NULL`. R will calculate the value for the missing parameter.

When we conduct a "power analysis" (i.e., the likelihood of a hypothesis test detecting an effect if there is one), we specify, "power=NULL". Using the data from our results, we learn from this first run, that our statistical power was at 5%. That is, given the low value of the mean difference (.08) and the relatively large standard deviation (4.14), we had only a 5% chance of detecting a statistically significant effect if there was one.

```
pwr::pwr.t.test(d = 0.02, n = 33, power = NULL, sig.level = 0.05, type = "paired",
  alternative = "two.sided")
```

Paired t test power calculation

```
n = 33
d = 0.02
sig.level = 0.05
power = 0.05142498
alternative = two.sided
```

NOTE: n is number of *pairs*

The results indicate that we were powered at 5%. That is, we had a 5% chance of finding a statistically significant difference, if in fact there was one.

Researchers frequently use these tools to estimate the sample size required to obtain a statistically significant effect. In these scenarios we set n to `NULL`.

```
pwr::pwr.t.test(d = 0.02, n = NULL, power = 0.8, sig.level = 0.05, type = "paired",
  alternative = "two.sided")
```

Paired t test power calculation

```
n = 19624.07
d = 0.02
sig.level = 0.05
```

```
power = 0.8
alternative = two.sided
```

NOTE: n is number of *pairs*

Using the results from the simulation of our research vignette, you can see that we would have needed 19624 individuals for the p value to be $< .05$, if, in fact there were a significant difference.

Let's see if this is true. Below I will re-simulate the data for the verbal scores, changing only the sample size:

```
set.seed(220820)
# These define the characteristics of the verbal variable. It is
# essential that the object names (e.g., A_mean) are not changed
# because they will be fed to the function in the faux package.
sub_n <- 19624
A_mean <- 8.37
B_mean <- 8.41
A_sd <- 3.36
B_sd <- 3.21
AB_r <- 0.3

# the faux package can simulate a variety of data. This function
# within the faux package will use the objects above to simulate
# paired samples data
paired_V2 <- faux::rnorm_multi(n = sub_n, vars = 2, r = AB_r, mu = c(A_mean,
  B_mean), sd = c(A_sd, B_sd), varnames = c("Verbal_BL", "Verbal_WH"))

paired_V2 <- paired_V2 %>%
  dplyr::mutate(PhysID = row_number())

# restructuring data to the long form
df_longV2 <- data.table::melt(data.table::setDT(paired_V2), id.vars = c("PhysID"),
  measure.vars = list(c("Verbal_BL", "Verbal_WH")))
df_longV2 <- rename(df_longV2, PatientRace = variable, Verbal = value)
df_longV2$PtRace <- plyr::mapvalues(df_longV2$PatientRace, from = c("Verbal_BL",
  "Verbal_WH"), to = c("Black", "White"))
```

Now I will conduct the paired samples t -test and corresponding effect size.

```
rstatix::t_test(df_longV2, Verbal ~ PtRace, paired = TRUE, detailed = TRUE)
```

```
# A tibble: 1 x 13
  estimate .y. group1 group2    n1    n2 statistic      p    df conf.low
*   <dbl> <chr> <chr> <chr> <int> <int>     <dbl> <dbl> <dbl>    <dbl>
1 -0.0343 Verbal Black White  19624 19624     -1.24 0.214 19623  -0.0885
# i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

```
rstatix::cohens_d(df_longV2, Verbal ~ PtRace, paired = TRUE)
```

```
# A tibble: 1 x 7
  .y.    group1 group2  effsize     n1     n2 magnitude
* <chr> <chr>   <chr>    <dbl> <int> <int> <ord>
1 Verbal Black   White   -0.00887 19624 19624 negligible
```

The new results remain non-significant: $t(19623) = -1.243, p = 0.241, d = -0.009, 95CI(-0.088, 0.020)$. This tells me these means are quite similar and this is not a function of being under powered.

Conducting power analyses requires that researchers speculate about their values. In this case, in order to estimate sample size, the researcher would need to make some guesses about the difference scores means and standard deviations. These values could be estimated from prior literature or a pilot study.

6.8 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “do-able” and then try at least one more problem that challenges you in some way. Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands).

Additionally, please complete at least one set of *hand calculations*, that is use the code demonstrated in the chapter to work through the formulas that compute the paired samples *t*-test. At this stage in your learning, you may ignore any missingness in your dataset by excluding all rows with missing data in your variables of interest.

6.8.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed

If this topic feels a bit overwhelming, simply change the random seed in the data simulation of the research vignette, then rework the problem. This should provide minor changes to the data (maybe even in the second or third decimal point), but the results will likely be very similar. That said, don’t be alarmed if what was non-significant in my working of the problem becomes significant. Our selection of $p < .05$ (and the corresponding 95% confidence interval) means that 5% of the time there could be a difference in statistical significance.

6.8.2 Problem #2: Rework the research vignette, but change something about the simulation

Rework the paired samples *t*-test in the lesson by changing something else about the simulation. For example, if you are interested in understanding more about power, consider changing the sample size. Alternatively, you could specify different means and/or standard deviations.

6.8.3 Problem #3: Rework the research vignette, but swap one or more variables

Use the simulated data, but select the nonverbal communication variables that were evaluated in the Elliott et al. [2016] study. Compare your results to those reported in the manuscript.

6.8.4 Problem #4: Use other data that is available to you

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a paired samples *t*-test.

6.8.5 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the variables and their role in the analysis	5	_____
2. Simulate (or import) and format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct a paired samples <i>t</i> -test (with an effect size & 95%CIs)	5	_____
5. APA style results with table(s) and figure	5	_____
6. Conduct power analyses to determine the power of the current study and a recommended sample size	5	_____
7. Explanation to grader	5	_____
Totals	35	_____

Hand Calculations	Points Poss	Points Earned
1. Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses	2	
2. Using an R package or functions in base R (and with data in the “wide” format), calculate the <i>difference</i> score between the two observations of the dependent variable	2	
3. Obtain the mean and standard deviation of the <i>difference</i> score	2	
4. Calculate the paired samples <i>t</i> -test	4	
5. Identify the degrees of freedom associated with your paired samples <i>t</i> -test	2	

Hand Calculations	Points Poss	Points Earned
6. Locate the test critical value for your paired samples t -test	2	
7. Is the paired samples t -test statistically significant? Why or why not?	2	
8. What is the confidence interval around the mean difference?	4	
9. Calculate the effect size (i.e., Cohen's d associated with your paired samples t -test	4	
10. Assemble the results into a statistical string	4	
Totals*	28	

6.9 Homeworked Example

Screencast Link

If you wanted to use this example and dataset as a basis for a homework assignment, you could compare a differenc combination of courses and/or score one of the other course evaluation subscales (e.g., socially responsive pedagogy or valued-by-me).

6.9.1 Working the Problem with R and R Packages

6.9.1.1 Narrate the research vignette, describing the variables and their role in the analysis

I want to ask the question, “Do students’ evaluations of traditional pedagogy (TradPed) change from ANOVA (the first course in the series) to Multivariate (the second course in the series)?” Unlike the independent samples t -test where we compared students in two different departments, we are comparing *the same* students across two different conditions. In this particular analysis, there is also an element of time. That is the ANOVA class always precedes the multivariate class (with a regression class, taught by a different instructor) in the intervening academic quarter.

This research design has some clear limitations. Threats to internal validity are caused by issues like history and maturation. None-the-less, for the purpose of a statistical demonstration, this dataset works.

Like most data, some manipulation is required before we can begin the analyses.

6.9.1.2 Simulate (or import) and format data

Let’s import the larger dataset.

The TradPed (traditional pedagogy) variable is an average of the items on that scale. I will first create that variable.

From the “larger” data, let’s select only the variable we will use in the analysis. I have included “long” in the filename because the structure of the dataset is that course evaluation by each student is in its own row. That is, each student could have up to three rows of data.

We need both “long” and “wide” forms to conduct the analyses required for both testing the statistical assumptions and performing the paired samples t -test.

From that reduced variable set, let’s create a subset with students only from those two courses.

Regarding the structure of the data, we want the conditions (ANOVA, multivariate) to be factors and the TradPed variable to be continuously scaled. The format of the deID variable can be any numerical or categorical format – just not a “chr” (character) variable.

```
Classes 'data.table' and 'data.frame': 198 obs. of 3 variables:
 $ deID    : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Course  : Factor w/ 3 levels "Psychometrics",...: 2 2 2 2 2 2 2 2 2 ...
 $ TradPed: num 4.4 3.8 4 3 4.8 3.5 4.6 3.8 3.6 4.6 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

R correctly interpreted our variables.

For analyzing the assumptions associated with the paired-samples t -test, the format needs to be “wide” form (where each student has both observations on one row). Our data is presently in “long” form (where each observation is listed in each row). Here’s how to reshape the data.

Let’s recheck the structure.

```
'data.frame': 119 obs. of 3 variables:
 $ deID      : int 1 2 3 4 5 6 7 8 9 10 ...
 $ ANOVA     : num 4.4 3.8 4 3 4.8 3.5 4.6 3.8 3.6 4.6 ...
 $ Multivariate: num NA NA NA NA NA NA NA NA NA ...
```

You will notice that there is a good deal of missingness in the Multivariate condition. This is caused because the most recent cohort of students had not yet taken the course. While managing missingness is more complex than this, for the sake of simplicity, I will create a dataframe with non-missing data.

Doing so should also help with the hand-calculations later in the worked example.

6.9.1.3 Evaluate statistical assumptions

We need to evaluate the *distribution of the difference score* in terms of skew and kurtosis. We want this distribution of difference scores to be normally distributed.

This means we need to create a difference score:

We can use the *psych::describe()* function to obtain skew and kurtosis.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew
deID		1	77	62.27	35.43	60.0	59.54	34.10	11.0	142.0	131.0
ANOVA		2	77	4.21	0.73	4.2	4.29	0.89	2.2	5.0	2.8
Multivariate		3	77	4.33	0.73	4.4	4.45	0.59	1.2	5.0	3.8
DIFF		4	77	-0.12	0.80	-0.2	-0.13	0.59	-2.4	3.2	5.6
					kurtosis	se					

deID	-0.40	4.04
ANOVA	-0.19	0.08
Multivariate	5.00	0.08
DIFF	3.15	0.09

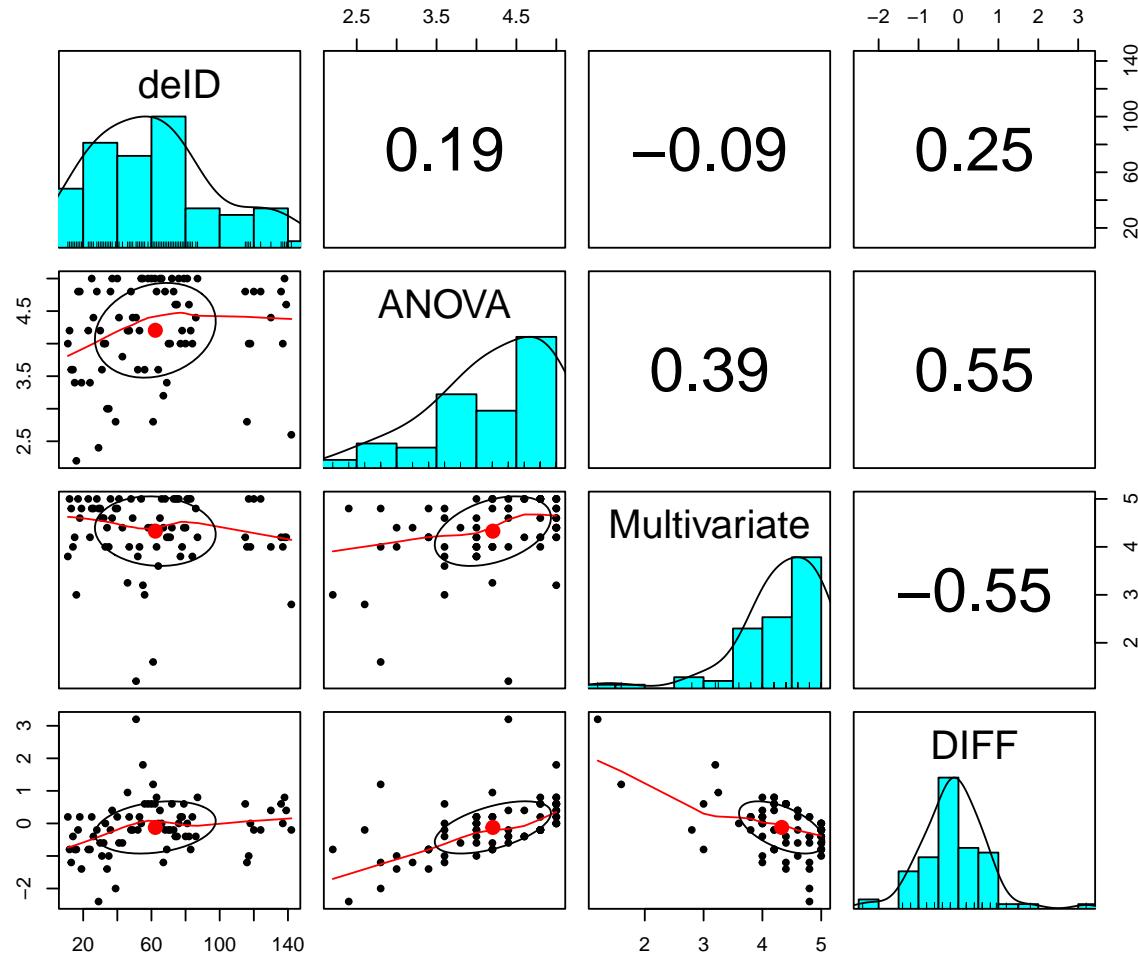
Regarding the DIFF score, the skew (0.56) and kurtosis (3.15) values were well below the thresholds of concern identified by Klein (2016).

We can formally test for deviations from normality with a Shapiro-Wilk. We want the results to be non-significant.

```
# A tibble: 1 x 3
  variable statistic      p
  <chr>       <dbl>    <dbl>
1 DIFF        0.943  0.00187
```

Results of the Shapiro-Wilk test of normality are statistically significant ($W = 0.943, p = 0.002$). This means that the distribution of difference scores are statistically significantly different from a normal distribution.

although not required in the formal test of instructions, a *pairs panel* of correlations and distributions can be useful in undersatnding our data.



Visual inspection of the distributions of the specific course variables were negatively skewed, with values clustered at the high end of the course evaluation ratings. However, the distribution for the DIFF variable seems relatively normal (although maybe a bit leptokurtic). This is consistent with the statistically significant Shapiro-Wilk test.

Before moving forward, I want to capture my analysis of assumptions:

We began by analyzing the data to see if it met the statistical assumptions for analysis with a paired samples t-test. Regarding the assumption of normality, the skew (0.56) and kurtosis (3.15) values associated with the difference between conditions (ANOVA and multivariate) were below the thresholds of concern identified by Klein (2016). In contrast, results of the Shapiro-Wilk test of normality suggested that the distribution of difference scores was statistically significantly different than a normal distribution ($W = 0.943, p = 0.002$).

6.9.1.4 Conduct a paired samples t-test (with an effect size & 95% CIs)

So this may be a bit tricky, but our original “long” form of the data has more ANOVA evaluations (students who had taken ANOVA had not yet taken multivariate) than multivariate. The paired samples t test requires the design to be balanced. When we used the `na.omit()` function with the wide case, we effectively balanced the design, eliminating students who lacked observations across both courses. Let’s restructure that wide format back to long format so that the design will be balanced.

```

deID Course TradPed
1:   11 ANOVA    4.0
2:   12 ANOVA    4.2
3:   13 ANOVA    3.6
4:   14 ANOVA    3.6
5:   15 ANOVA    3.4
6:   16 ANOVA    2.2

# A tibble: 1 x 13
  estimate .y. group1 group2      n1      n2 statistic     p     df conf.low
* <dbl> <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <dbl>
1 -0.123 TradPed ANOVA Multivaria~    77     77    -1.34 0.184    76   -0.305
# i 3 more variables: conf.high <dbl>, method <chr>, alternative <chr>
```

I'll begin the t string with this output: $t(76) = -1.341, p = 0.184, CI95(-0.305, 0.069)$. The difference in course evaluations is not statistically significantly difference. We are 955 confident that the true difference in means is as low as -0.301 or as high as 0.060.

we calculate the Cohen's d (the effect size) this way:

```

# A tibble: 1 x 7
  .y. group1 group2      effsize      n1      n2 magnitude
* <chr> <chr> <chr> <dbl> <int> <int> <ord>
1 TradPed ANOVA Multivariate -0.153     77     77 negligible
```

The value of -0.153 is quite small. We can add this value to our statistical string: $t(76) = -1.341, p = 0.184, CI95(-0.305, 0.069), d = -0.153$

6.9.1.5 APA style results with table(s) and figure

A paired samples t -test was conducted to evaluate the hypothesis that there would be statistically significant differences in students' course evaluations of ANOVA and multivariate statistics classses.

We began by analyzing the data to see if it met the statistical assumptions for analysis with a paired samples t-test. Regarding the assumption of normality, the skew (0.56) and kurtosis (3.15) values associated with the difference between conditions (ANOVA

and multivariate) were below the thresholds of concern identified by Klein (2016). In contrast, results of the Shapiro-Wilk test of normality suggested that the distribution of difference scores was statistically significantly different than a normal distribution \$(W=0.943, p = 0.002)\$

Results of the paired samples *t*-test suggested nonsignificant differences $t(76) = -1.341, p = 0.184, d = -0.153$. The 95% confidence interval crossed zero, ranging from -0.305 to 0.069. Means and standard deviations are presented in Table 1 and illustrated in Figure 1.

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1	2
1. ANOVA	4.21	0.73		
2. Multivariate	4.33	0.73	.39**	
			[.18, .56]	
3. DIFF	-0.12	0.80	.55**	-.55**
			[.38, .69]	[-.69, -.37]

Note. M and SD are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

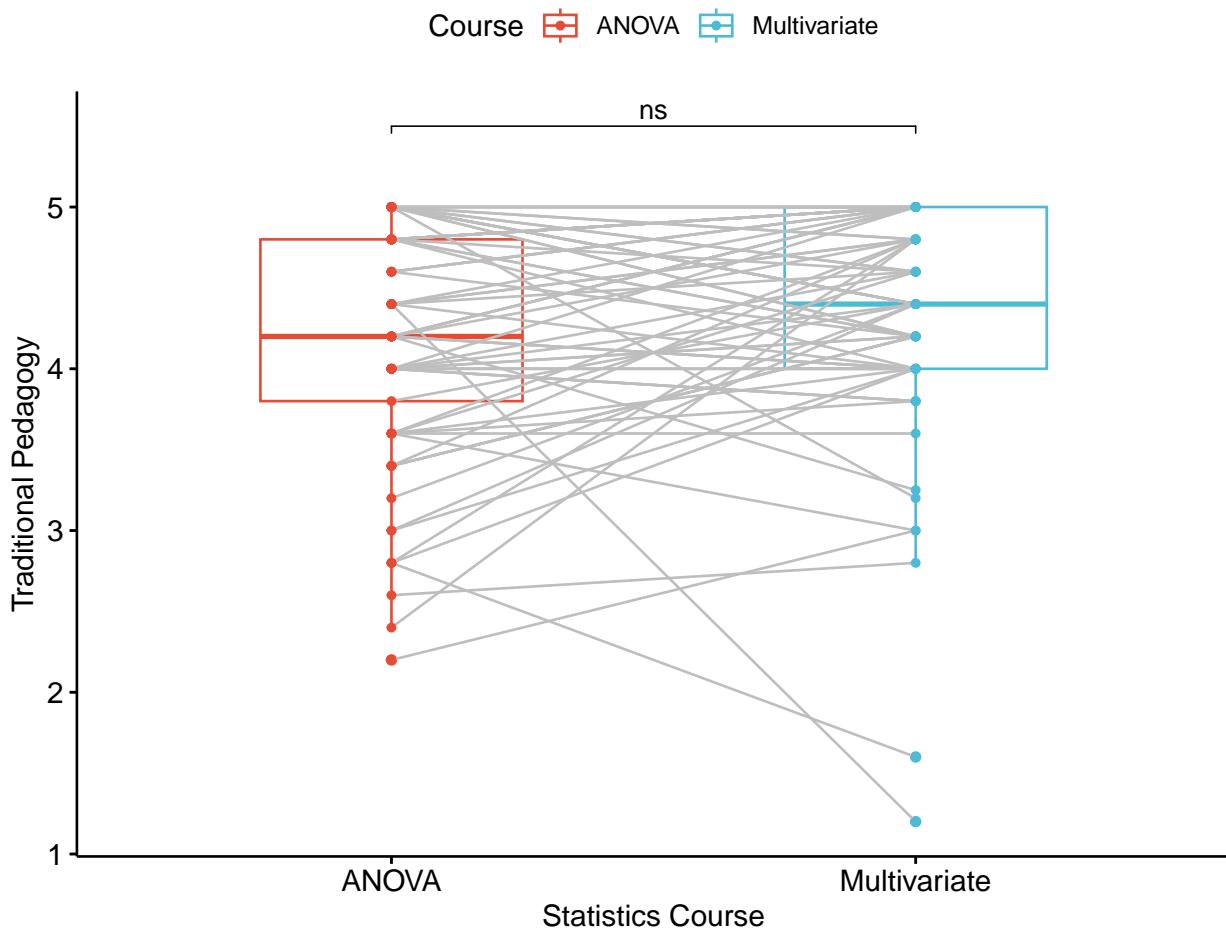
* indicates $p < .05$. ** indicates $p < .01$.

For the figure, let's re-run the paired samples *t* test, save it as an object, and use the "add_significance" function so that we can add it to our figure.

```
# A tibble: 1 x 14
  estimate .y.    group1 group2      n1      n2 statistic     p     df conf.low
  <dbl> <chr>  <chr>  <chr>    <int>    <int>    <dbl> <dbl> <dbl>    <dbl>
1   -0.123 TradPed ANOVA Multivaria~    77      77     -1.34 0.184    76   -0.305
# i 4 more variables: conf.high <dbl>, method <chr>, alternative <chr>,
# p.signif <chr>
```

Next, we create boxplot:

Figure 1. Evaluation of Traditional Pedagogy as a Function of Course
 T test, $t(76) = -1.34$, $p = 0.18$, $n = 77$



6.9.1.6 Conduct power analyses to determine the power of the current study and a recommended sample size

Script for estimating current power:

- d is Cohen's d
- n is number of pairs, but set to NULL if we want to estimate sample size
- power is conventionally set at .80, but left at NULL when we want to estimate power
- sig.level is conventionlaly set at 0.05
- type indicates the type of t -test; in this example it is "paired"
- alternative indicates one or two.sided

Paired t test power calculation

n = 77

```
d = 0.153
sig.level = 0.05
power = 0.2634404
alternative = two.sided
```

NOTE: n is number of *pairs*

We had a 26% chance of finding a statistically significant result if, in fact, one existed.

Paired t test power calculation

```
n = 337.2182
d = 0.153
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: n is number of *pairs*

If we presumed power were at 80%, we would need a sample size of 337.

6.9.2 Hand Calculations

For these hand calculations I will used the “paired_wide” dataframe that we had prepared for the homework assignment intended for R and R packages.

6.9.2.1 Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses

The null hypotheses states that the true difference in means is zero. $H_O : \mu_D = 0$

The alternative hypothesis states that the true difference in means is not zero. $H_A : \mu_D \neq 0$

6.9.2.2 Using an R package or functions in base R (and with data in the “wide” format), calculate the *difference* score between the two observations of the dependent variable

We had already calculated a difference score in the earlier assignment. Here it is again.

6.9.2.3 Obtain the mean and standard deviation of the *difference* score

We can obtain the mean and standard deviation for the difference score with this script.

```
vars n mean sd median trimmed mad min max range skew kurtosis se
x1   1 77 -0.12 0.8   -0.2   -0.13 0.59 -2.4 3.2   5.6 0.56     3.15 0.09
```

The mean difference (\bar{D}) is -0.12; the standard deviation ($\hat{\sigma}_D$) of the difference score is 0.8.

6.9.2.4 Calculate the paired samples t -test

Here is the formula for the paired samples t -test:

$$t = \frac{\bar{D}}{\hat{\sigma}_D / \sqrt{N}}$$

Using the values we located we can calculate the value of the t statistic.

```
[1] -1.316245
```

The value we calculated with the `rstatix::t_test()` function was -1.34. Considering rounding error, I think we got it!

6.9.2.5 Identify the degrees of freedom associated with your paired samples t -test

We have 77 pairs. The df for the paired samples t -test is $N - 1$. Therefore, $df = 76$.

6.9.2.6 Locate the test critical value for your paired samples t -test

I could look at the [table of critical values](#) for the t -distribution. Because I have non-directional hypotheses, I would use the column for a p -value of .05 for a two-tailed test. I roll down to the closest sample size (I'll pick 60). This suggests that my t -test statistic would need to be greater than 2.0 in order to be statistically significant.

I can also use the `qt()` function in base R. This function requires that I specify the alpha level (0.05), whether the test is one- or two-tailed (2), and my degrees of freedom (76). Specifying “TRUE” and “FALSE” after the lower.tail command gives the positive and negative regions of rejection.

```
[1] -1.991673
```

```
[1] 1.991673
```

It is not surprising that these values are a smidge lower than 2.0. Why? Because in the table we stopped at df of 60, when it is actually 76.

6.9.2.7 Is the paired samples t -test statistically significant? Why or why not?

The paired samples t -test is not statistically significant because the t -value of -1.316245 does not exceed -1.992.

6.9.2.8 What is the confidence interval around the mean difference?

Here is the formula for hand-calculating the confidence interval.

$$\bar{D} \pm t_{cv}(s_d/\sqrt{n})$$

- \bar{D} the mean difference score
- t_{cv} the test critical value for a two-tailed model (even if the hypothesis was one-tailed) where $\alpha = .05$ and the degrees of freedom are $N - 1$
- s_d the standard deviation of \bar{D}
- N sample size

Let's calculate it:

```
[1] 0.06157776
```

```
[1] -0.3015778
```

These values indicate the range of scores in which we are 95% confident that our true \bar{D} lies. Stated another way, we are 95% confident that the true mean difference lies between -0.302 and 0.062. Because this interval crosses zero, we cannot rule out that the true mean difference is 0.00. This result is consistent with our non-significant p value. For these types of statistics, the 95% confidence interval and p value will always be yoked together.

6.9.2.9 Calculate the effect size (i.e., Cohen's d associated with your paired samples t -test

Cohen's d measures, in standard deviation units, the distance between the two means. Regardless of sign, values of .2, .5, and .8 are considered to be small, medium, and large, respectively.

Because the paired samples t -test used the difference score in the numerator, there are two easy options for calculating this effect:

$$d = \frac{\bar{D}}{\hat{\sigma}_D} = \frac{t}{\sqrt{N}}$$

Here's a demonstration of both:

```
[1] -0.15
```

```
[1] -0.15
```

6.9.2.10 Assemble the results into a statistical string.

$t(76) = -1.316, p > .05, CI95(-0.302, 0.062), d = -0.15$

Analysis of Variance

Chapter 7

One-way ANOVA

[Screencasted Lecture Link](#)

One-way ANOVA allows the researcher to analyze mean differences between two or more groups on a between-subjects factor. For the one-way ANOVA, each case (i.e., individual, participant) must have scores on two variables: a factor and a dependent variable.

The factor must be categorical in nature, dividing the cases into two or more groups or levels. These levels could be ordered (e.g., placebo, low dose, high dose) or unordered (e.g., cognitive-behavioral, existential, psychodynamic). The dependent variable must be assessed on a quantitative, continuous dimension. The ANOVA F test evaluates whether population means on the dependent variable differ across the levels of the factor.

One-way ANOVA can be used in experimental, quasi-experimental, and field studies. As we work through the chapter, we will examine some of the requirements (assumptions) of the statistic in greater detail.

7.1 Navigating this Lesson

There is about 2 hours of lecture. If you work through the materials with me, plan for another two hours of study.

7.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Evaluate the statistical assumptions associated with one-way analysis of variance (ANOVA).
- Describe the relationship between model/between-subjects and residual/within-subjects variance.
- Narrate the steps in conducting a formal one-way ANOVA beginning with testing the statistical assumptions through writing up an APA style results section.
- Conduct a one-way ANOVA in R (including calculation of effect sizes and follow-up to the omnibus).
- Conduct a power analysis for a one-way ANOVA.
- Produce an APA style results section for one-way ANOVA.

7.1.2 Planning for Practice

In each of these lessons I provide suggestions for practice that allow you to select one or more problems that are graded in difficulty. The least complex is to change the random seed and rework the problem demonstrated in the lesson. The results *should* map onto the ones obtained in the lecture.

The second option comes from the research vignette. The Tran et al. [2014] vignette has two variables where the authors have conducted one-way ANOVAs. I will demonstrate one (*Accurate*) in this lecture; the second is available as one of the homework options.

As a third option, you are welcome to use data to which you have access and is suitable for two-way ANOVA. In either case the practice options suggest that you:

- test the statistical assumptions
- conduct a one-way ANOVA, including
 - omnibus test and effect size
 - follow-up (pairwise, planned comparisons, polynomial trends)
- write a results section to include a figure and tables

7.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s) that are freely available on the internet. Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Crump, M. J. C. (2018). Chapter 5.5.2, Simulating data for one-way between subjects design with 3 levels. In [Programming for Psychologists: Data Creation and Analysis](#). Retrieved from <https://crumplab.github.io/programmingforpsych/simulating-and-analyzing-data-in-r.html#single-factor-anovas-data-simulation-and-analysis>
 - Although this reference is on simulating data, the process of simulation can provide another perspective on one-way ANOVA.
- Kassambara, A. (n.d.). ANOVA in R: The Ultimate Guide. Datanovia. Retrieved December 28, 2022, from <https://www.datanovia.com/en/lessons/anova-in-r/>
 - In order to streamline the learning process, I have chosen to use *rstatix* package for the majority of ANOVA lessons. There are a number of tutorials about this package as well as its integration with *ggpubr* for creating relatively easy creation of attractive and informative figures. This tutorial is especially helpful.
- Navarro, D. (2020). Chapter 14: Comparing Several Means (one-Way ANOVA). In [Learning Statistics with R - A tutorial for Psychology Students and other Beginners](#). Retrieved from [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro))

- Navarro's OER includes a good mix of conceptual information about one-way ANOVA as well as R code. My code/approach is a mix of Green and Salkind's [2017c], Field's [2012], Navarro's [2020b] chapters as well as other techniques I have found on the internet and learned from my students.
- Tran, A. G. T. T., & Lee, R. M. (2014). You speak English well! Asian Americans' reactions to an exceptionalizing stereotype. *Journal of Counseling Psychology*, 61(3), 484–490. <https://doi.org/10.1037/cou0000034>
 - The source of our research vignette.

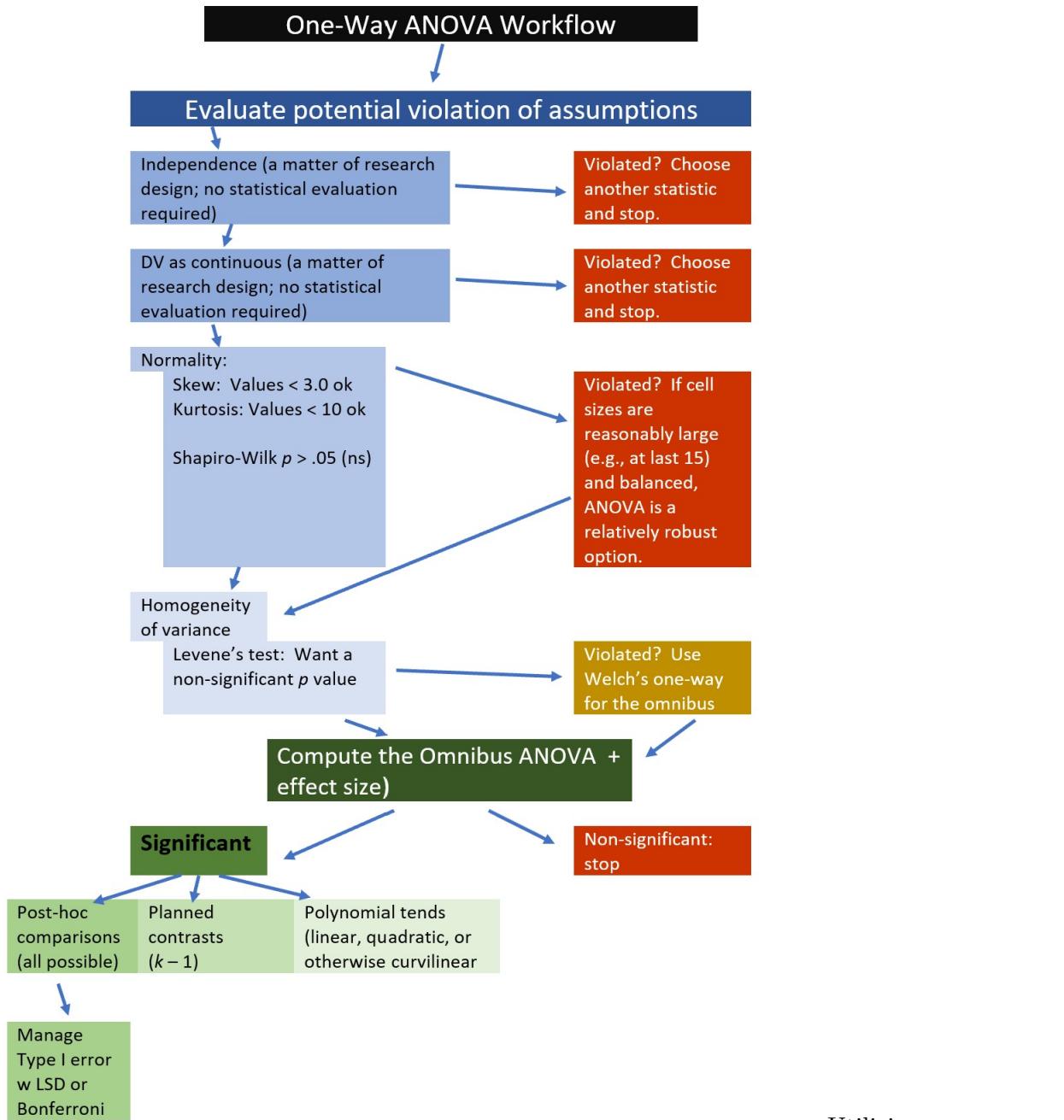
7.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed #easy plotting
# for simple ANOVA if(!require(knitr)){install.packages('knitr')}
# #not needed for conducting the statistics, but necessary for
# knitting the document (if desired)
# if(!require(tidyverse)){install.packages('tidyverse')} #a specific
# part of the tidyverse with useful tools for manipulating data
# if(!require(dplyr)){install.packages('dplyr')} #for descriptive
# statistics and writing them as csv files
# if(!require(psych)){install.packages('psych')} #a number of
# wrappers for ANOVA models; today for evaluating the Shapiro
# if(!require(ggpubr)){install.packages('ggpubr')} #the package we
# will use to create figures
# if(!require(rstatix)){install.packages('rstatix')} #the package we
# will use for the majority of the ANOVA computations
# if(!require(apaTables)){install.packages('apaTables')} #helps with
# formats like decimals and percentages for inline code
# if(!require(effectsize)){install.packages('effectsize')}
# if(!require(pwr)){install.packages('pwr')} #produces an APA style
# table for ANOVAs and other models
# if(!require(car)){install.packages('car')}#although we don't call
# this package directly, there are rstatix functions that are a
# wrapper for it and therefore it needs to be installed
```

7.2 Workflow for One-Way ANOVA

The following is a proposed workflow for conducting a one-way ANOVA.



ANOVA involves the following steps:

1. Prepare (upload) data.
2. Explore data
 - graphs
 - descriptive statistics
3. Checking distributional assumptions
 - assessing normality via skew, kurtosis, Shapiro-Wilks

Utilizing one-way

- checking for violation of homogeneity of variance assumption with Levene's test; if we violate this we can use Welch's omnibus ANOVA
4. Compute the omnibus ANOVA (remember to use Welch's if Levene's $p < .05$)
 5. Compute post hoc comparisons, planned contrasts, or polynomial trends
 6. Managing Type I error
 7. Sample size/power analysis (which you should think about first – but in the context of teaching ANOVA, it's more pedagogically sensible, here)

7.3 Research Vignette

The *exceptionalizing racial stereotype* is microaggression framed as interpersonally complimentary, but perpetuates negative stereotypical views of a racial/ethnic group. We are using data that is simulated from a random clinical trial (RCT) conducted by Tran and Lee [2014].

The one-way ANOVA examples we are simulating represent the post-only design which investigated three levels of the exceptionalizing stereotype in a sample of Asian American participants. This experimental design involved a confederate (posing as a peer) whose parting comment fell into the low racial loading, high racial loading, or control conditions.

COND	Assignment	Manipulation	Post-test Observation
Low racial loading condition ($n = 22$)	Random	Yes: "Nice talking to you. You speak English well."	Accurate
High racial loading ($n = 23$)	Random	Yes: "Nice talking to you. You speak English well for an Asian."	Accurate
Control ($n = 23$)	Random	No: "Nice talking to you."	Accurate

In the article, the one-way ANOVA is a relatively smaller focus. In fact, Tran and Lee [2014] reported results from two ANOVAs and 4 ANCOVAs, using a pre-test as a covariate. A preprint of their article is available [here](#). If you are interested in this topic, I highly encourage you to review the more complex analyses and their results.

- **Accurate** is the DV we will be exploring in this lesson. Participants rated how *accurate* they believed their partner's impression of them was ($0 = \text{very inaccurate}$, $3 = \text{very accurate}$).
- **moreTalk** is the DV suggested as a practice problem. Participants rated how much longer they would continue the interaction with their partner compared to their interactions in general ($-2 = \text{much less than average}$, $0 = \text{average}$, $2 = \text{much more than average}$).

7.3.1 Data Simulation

Simulating data for a one-way ANOVA requires the sample size (listed first), mean (mean=), and standard deviation (sd=) for each of the groups [Crump, 2018]. In creating this simulation, I used the data from Table 1 in the Tran and Lee [2014] article. Having worked the problem several times,

I made one change. The group sizes in the original study were 23, 22, and 23. To increase the probability that we would have statistically significant results in our worked example, I increased the sample sizes to 30 for each group. In this way we have a perfectly *balanced* (equal cell sizes) design.

```
# Note, this script results in a different simulation than is in the
# ReadySetR lesson sets a random seed so that we get the same results
# each time
set.seed(210820)
# sample size, M and SD for each group
Accurate <- c(rnorm(30, mean = 1.18, sd = 0.8), rnorm(30, mean = 1.83,
  sd = 0.58), rnorm(30, mean = 1.76, sd = 0.56))
# set upper bound for DV
Accurate[Accurate > 3] <- 3
# set lower bound for DV
Accurate[Accurate < 0] <- 0
# sample size, M and SD for each group
moreTalk <- c(rnorm(30, mean = -0.82, sd = 0.91), rnorm(30, mean = -0.39,
  sd = 0.66), rnorm(30, mean = -0.04, sd = 0.71))
# set upper bound for DV
moreTalk[moreTalk > 2] <- 2
# set lower bound for DV
moreTalk[moreTalk < -2] <- -2
# IDs for participants
ID <- factor(seq(1, 90))
# name factors and identify how many in each group; should be in same
# order as first row of script
COND <- c(rep("High", 30), rep("Low", 30), rep("Control", 30))
# groups the 3 variables into a single df: ID#, DV, condition
accSIM30 <- data.frame(ID, COND, Accurate, moreTalk)
```

Examining the data is important for several reasons. First, we can begin our inspection for anomalies. Second, if we are confused about what statistic we wish to apply, understanding the characteristics of the data can provide clues.

We can see the entire dataframe by clicking open the dataframe object found in the Environment window of R studio. This will open a tab that allows scrolling up/down and left/right through the entire dataframe. It is also possible to sort by variables.

Alternatively the *head()* function from base R displays a static view of the first six rows of data.

```
head(accSIM30)
```

	ID	COND	Accurate	moreTalk
1	1	High	0.4203896	-0.6398265
2	2	High	1.1226505	-2.0000000
3	3	High	0.8852238	-0.2497750
4	4	High	1.5689439	0.1455637

```
5 5 High 1.8307196 -0.9960413
6 6 High 1.8874431 -1.0692978
```

Yet another option is to use the *str()* function from base R. This provides a list of variables and provides detail about their formats.

```
str(accSIM30)
```

```
'data.frame': 90 obs. of 4 variables:
 $ ID      : Factor w/ 90 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ COND    : chr "High" "High" "High" "High" ...
 $ Accurate: num 0.42 1.123 0.885 1.569 1.831 ...
 $ moreTalk: num -0.64 -2 -0.25 0.146 -0.996 ...
```

If we look at this simple dataset, we see that we see that

- **COND** is a grouping variable) with 3 levels (high, low, control)
 - it is presently in “chr” (character) format, it needs to be changed to be a factor.
- **Accurate** is a continuous variable
 - it is presently in “num” (numerical) format, this is an appropriate format.
- **moreTalk** is a continuous variable
 - it is presently in “num” (numerical) format, this is an appropriate format

There are many ways to convert variables to factors; here is one of the simplest.

```
#convert variable to factor
accSIM30$COND <- factor(accSIM30$COND)
```

Let's recheck the structure

```
str(accSIM30)
```

```
'data.frame': 90 obs. of 4 variables:
 $ ID      : Factor w/ 90 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ COND    : Factor w/ 3 levels "Control","High",...: 2 2 2 2 2 2 2 2 2 2 ...
 $ Accurate: num 0.42 1.123 0.885 1.569 1.831 ...
 $ moreTalk: num -0.64 -2 -0.25 0.146 -0.996 ...
```

By default, R orders factors alphabetically. This means, analyses will assume that “Control” (C) is the lowest condition, then “High,” then “Low.” Since these have theoretically ordered values, we want them in the order of “Control,” “Low,” “High.”

Here is the script to create an ordered factor. The order in which the variables are entered in the concatenated list (“c”) establishes the order (e.g., levels).

```
# ordering the factor
accSIM30$COND <- factor(accSIM30$COND, levels = c("Control", "Low", "High"))
```

Again, we can check our work.

```
#another structure check
str(accSIM30)
```

```
'data.frame': 90 obs. of 4 variables:
 $ ID      : Factor w/ 90 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ COND    : Factor w/ 3 levels "Control","Low",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ Accurate: num  0.42 1.123 0.885 1.569 1.831 ...
 $ moreTalk: num  -0.64 -2 -0.25 0.146 -0.996 ...
```

Now our variables are suitable for analysis.

Although you may continue working with the simulated data, at this point, you may wish to export and/or import the data as a .csv (think “Excel lite”) or .rds (R object that preserves the information about the variables – such changing COND to an ordered factor). Here is the code to do so. The data should save in the same folder as the .rmd file. Therefore, it is really important (think, “good R hygiene”) to have organized your folders so that your .rmd and data files are co-located.

I have hashtags out the code. If you wish to use it, delete the hashtags. Although I show the .csv code first, my personal preference is to save R data as .rds files. While they aren’t easy to “see” as an independent file, they retain the formatting of the variables. For a demonstration, refer back to the [Ready_Set_R](#) lesson.

```
# write the simulated data as a .csv write.table(accSIM30,
# file='accSIM.csv', sep=',', col.names=TRUE, row.names=FALSE) bring
# back the simulated dat from a .csv file acc_csv <-
# read.csv('accSIM.csv', header = TRUE)
```

If you have cleared the environment and then imported the .csv file, examining the structure of the .csv file shows that the prior formatting is lost. This is demonstrated in the accompanying screencast.

```
# a quick demo to show that the .csv format loses the variable
# formatting str(acc_csv)
```

Below is the code to write and then import the data as an .rds file.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(accSIM30, 'accSIM.rds') bring back the simulated dat
# from an .rds file acc_RDS <- readRDS('accSIM.rds')
```

By examining the structure of the .rds file we can see that the .rds file preserves the variable formatting. This is demonstrated in the accompanying screencast.

```
# a quick demo to show that the .rds format preserves the variable
# formatting str(acc_RDS)
```

Note that I renamed each of these data objects to reflect the form in which I saved them (i.e., “acc_csv”, “acc_RDS”). If you have followed this step, you will want to rename the file before continuing with the rest of the chapter. Alternatively, you can start from scratch, re-run the code to simulate the data, and skip this portion on importing/exporting data.

```
#accSIM30 <- acc_RDS
#or
#accSIM30 <- acc_csv
```

7.3.2 Quick Peek at the Data

This lesson’s exploration of the data is designed to introduce multiple tools for doing so. In this first demonstration I will quickly produce a mean and standard deviation using functions from base R.

The *aggregate()* function lets R know we want output by a grouping variable. We then list the variable of interest, a tilda (I think of the word “by”), and then the grouping variable (I think “Accurate by COND”). Finally we list the dataframe and the statistic (e.g., mean or standard deviation). R is case sensitive – so check your capitalization if your code fails to execute.

```
aggregate(Accurate ~ COND, accSIM30, mean)
```

```
COND Accurate
1 Control 1.756195
2 Low 1.900116
3 High 1.152815
```

```
aggregate(Accurate ~ COND, accSIM30, sd)
```

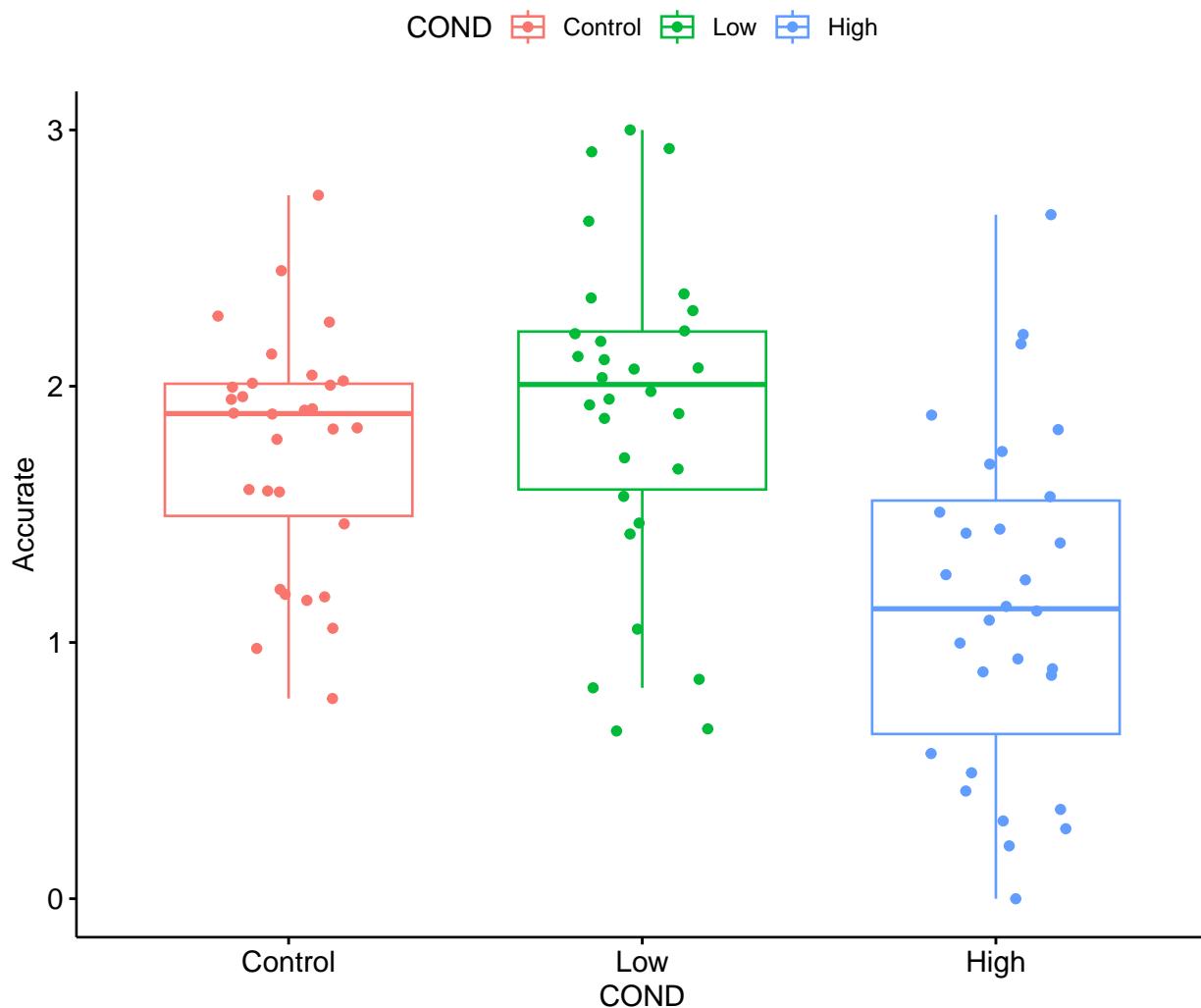
```
COND Accurate
1 Control 0.4603964
2 Low 0.6301138
3 High 0.6587486
```

Inspection of the means and standard deviations shows that the racially loaded *high* condition has the lowest accuracy score ($M = 1.153$) and the largest variability ($SD = 0.659$).

Graphing data is a best practice for early exploration and inspection of the data. In ANOVA models the boxplot is especially useful. The *ggpubr* package offers terrific options. After calling *ggpubr::ggbboxplot()*, we list the data frame and name the x and y variables. That would be sufficient

to produce a simple boxplot. The “add = jitter” command will plot each individual case, but “jitter” them to the right and left such that they are not overlapping and we can see all scores. For fun I added some color.

```
ggpubr::ggboxplot(accSIM30, x = "COND", y = "Accurate", add = "jitter",
color = "COND", )
```



In boxplots the center value is the median. The box spans the *interquartile range* and ranges from the 25th to the 75th percentile. The whiskers cover 1.5 times the interquartile range. When this does not capture the entire range, outliers are represented with dots.

From both the boxplot and the linegraph with error bars, we can see that participants in the low racial loading condition have the highest accuracy ratings. This is followed by the control and then high racial loading conditions. Are these differences statistically significant? This is why we need the one-way ANOVA.

7.4 Working the Oneway ANOVA (by hand)

ANOVA was developed by Sir Ronald Fisher in the early 20th century. The name is a bit of a misnomer – rather than analyzing *variances*, we are investigating differences in *means* (but the formula does take variances into consideration...stay tuned).

ANOVA falls squarely within the tradition of **null hypothesis significance testing** (NHST). As such, a formal, traditional, ANOVA begins with statements of the null and alternate hypotheses. *Note. In their article, Tran and Lee [2014] do not list such. This is fairly common in present-day journal articles.*

In our example, we would hypothesize that the population means (i.e., Asian or Asian American individuals in the U.S.) are equal:

$$H_O : \mu_1 = \mu_2 = \mu_3$$

There are an number of ways that the H_O could be false. Here are a few:

$$H_{a1} : \mu_1 \neq \mu_2 \neq \mu_3$$

$$H_{a2} : \mu_1 = \mu_2 > \mu_3$$

$$H_{a3} : \mu_1 > \mu_2 > \mu_3$$

The bottom line is that if we have a statistically significant omnibus ANOVA (i.e., the test of the overall significance of the model) and the H_O is false, somewhere between the three levels of the grouping factor, the means are statistically significantly different from each other.

In evaluating the differences between means, one-way ANOVA compares:

- systematic variance to unsystematic variance
- explained to unexplained variation
- experimental effect to the individual differences
- model variance to residual variance
- between group variance to within group variance

The ratio of these variances is the *F*-ratio.

Navarro [2020a] offers a set of useful figures to compare between- and within-group variation.

When between-group variance (i.e., model variance) is greater than within-group variance (i.e., residual variance) there may be support to suggest that there are statistically significant differences between groups.

Let's examine how variance is partitioned by hand-calculating sums of squares total, model, and residual. Along the way we will use some basic R skills to manipulate the data.

7.4.1 Sums of Squares Total

Sums of squares total represents the total amount of variance within our data. Examining the formula(s; there are variants of each) can help us gain a conceptual understanding of this.

In this first version of the formula we can see that the grand (or overall) mean is subtracted from each individual score, squared, and then summed. This makes sense: *sums of squares, total*.

Between-group variation
(i.e., differences among group means)

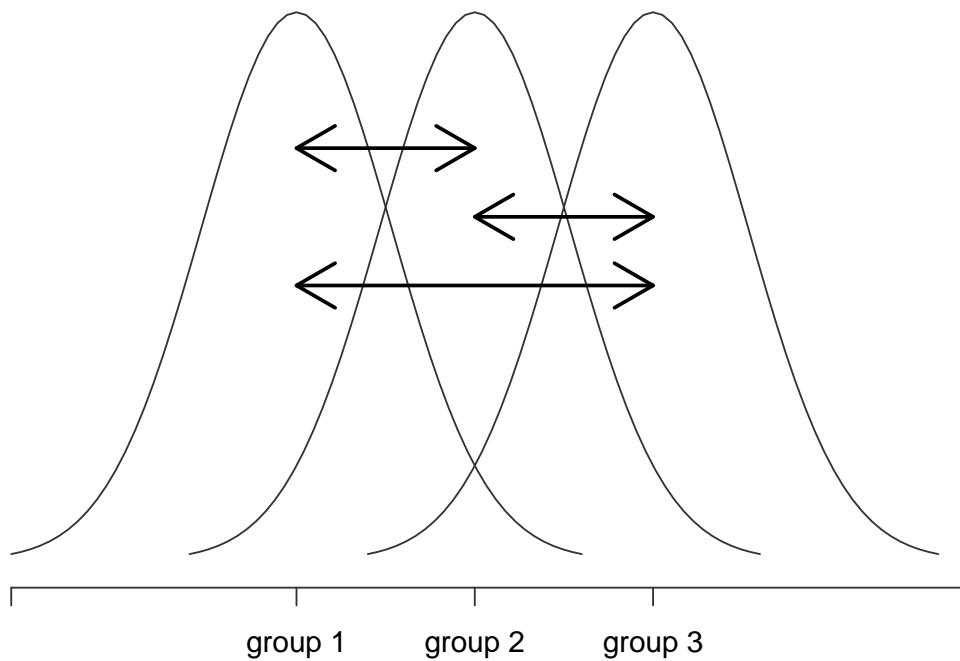


Figure 7.1: Graphical illustration of “between groups” variation

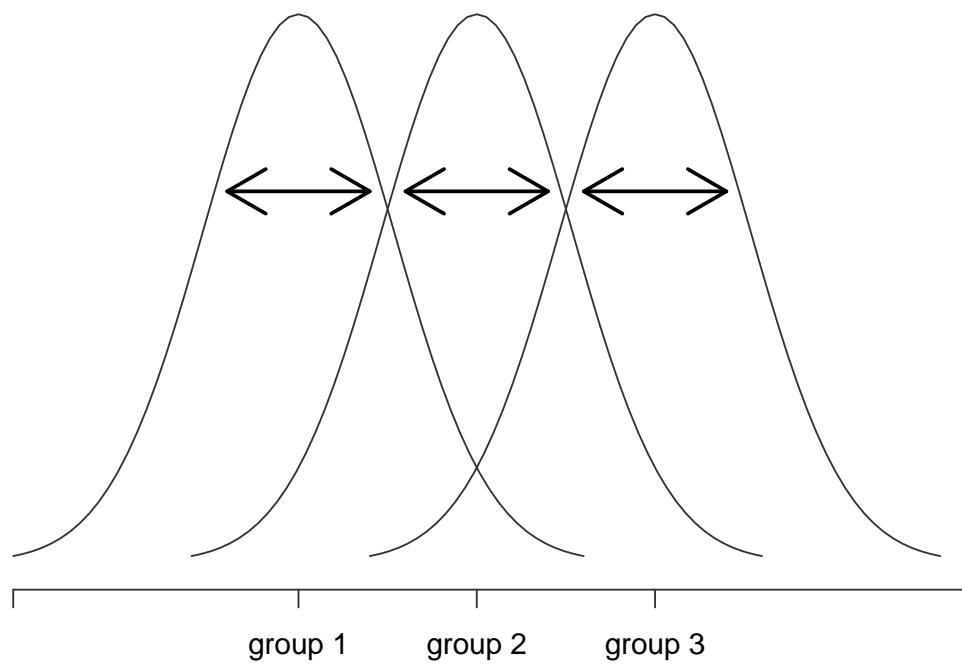


Figure 7.2: Graphical illustration of “within groups” variation

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

In the next version of the formula we see that the sums of square total is the addition of the sums of squares model and residual.

$$SS_T = SS_M + SS_R$$

“Between” and “within” are another way to understand “model” and “residual.” This is reflected in the next formula.

$$SS_T = SS_B + SS_W$$

Finally, think of the sums of squares total as the grand variance multiplied by the overall degrees of freedom ($N - 1$).

$$SS_T = s_{grand}^2(n - 1)$$

Let’s take a moment to *hand-calculate* SS_T . Not to worry – we’ll get R to do the math for us!

Our grand (i.e., overall) mean is

```
GrandMean <- mean(accSIM30$Accurate)
GrandMean
```

```
[1] 1.603042
```

Subtracting the grand mean from each Accurate rating yields a mean difference. In the script below I have used the *mutate()* function from the *dplyr* package (a part of the *tidyverse*) to created a new variable (“m_dev”) in the dataframe. The *tidyverse* package is one of the few exceptions that I will open via the library. This is because we need it if we are going to use the pipe (%>%) to string parts of our script together.

```
library(tidyverse)

accSIM30 <- accSIM30 %>%
  dplyr::mutate(m_dev = Accurate - mean(Accurate))

head(accSIM30)
```

	ID	COND	Accurate	moreTalk	m_dev
1	1	High	0.4203896	-0.6398265	-1.18265259
2	2	High	1.1226505	-2.0000000	-0.48039170
3	3	High	0.8852238	-0.2497750	-0.71781837
4	4	High	1.5689439	0.1455637	-0.03409829
5	5	High	1.8307196	-0.9960413	0.22767748
6	6	High	1.8874431	-1.0692978	0.28440098

Pop quiz: What's the sum of our new *m_dev* variable? Let's check.

```
mean(accSIM30$m_dev)
```

```
[1] 0.0000000000000003830065
```

Unless you run the script at the top of this document (“options(scipen=999)”), R will (seemingly selectively) use **scientific e notation** to report your results. The proper value is one where the base number (before the “e”) is multiplied by 10, raised to the power shown: 3.830065×10^{17} . Another way to think of it is to move the decimal 17 places to the left. In any case, this number is essentially zero.

Back to the point of sums of squares total, the sum of deviations around the grand mean will always be zero. To make them useful, we must square them:

```
accSIM30 <- accSIM30 %>%
  dplyr::mutate(m_devSQ = m_dev^2)

head(accSIM30)
```

ID	COND	Accurate	moreTalk	<i>m_dev</i>	<i>m_devSQ</i>
1	High	0.4203896	-0.6398265	-1.18265259	1.398667144
2	High	1.1226505	-2.0000000	-0.48039170	0.230776185
3	High	0.8852238	-0.2497750	-0.71781837	0.515263216
4	High	1.5689439	0.1455637	-0.03409829	0.001162694
5	High	1.8307196	-0.9960413	0.22767748	0.051837034
6	High	1.8874431	-1.0692978	0.28440098	0.080883915

If we sum the squared mean deviations we will obtain the total variance (sums of squares total):

```
SST <- sum(accSIM30$m_devSQ)
SST
```

```
[1] 39.67818
```

This value, the sum of squared deviations around the grand mean, is our SS_T . The associated *degrees of freedom* is $N - 1$; in our case this is $90 - 1 = 89$.

In one-way ANOVA, we divide SS_T into **model/between sums of squares** and **residual/within sums of squares**.

The *model* generally represents the notion that the means are different than each other. We want the variation between our means to be greater than the variation within each of the groups from which our means are calculated.

7.4.2 Sums of Squares for the Model (or Between)

We just determined that the total amount of variation within the data is 39.678 units. From this we can estimate how much of this variation our model can explain. SS_M tells us how much of the total variation can be explained by the fact that different data points come from different groups.

We see this reflected in the formula below, where

- the grand mean is subtracted from each group mean
- this value is squared and multiplied by the number of cases in each group
- these values are summed

$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$$

To calculate this, we start with the grand mean (previously calculated): 1.603.

We also estimate the group means. The script below provides the formula, the dataset, and the particular statistic (mean) that we want calculated.

```
GroupMeans <- aggregate(Accurate ~ COND, accSIM30, mean)
GroupMeans
```

COND	Accurate
1 Control	1.756195
2 Low	1.900116
3 High	1.152815

This script is used to extract the specific means so that I can demonstrate the formulas with the words/terms as well as the numbers.

```
ControlMean <- (GroupMeans$Accurate[1])
ControlMean
```

[1] 1.756195

```
LowMean <- (GroupMeans$Accurate[2])
LowMean
```

[1] 1.900116

```
HighMean <- (GroupMeans$Accurate[3])
HighMean
```

[1] 1.152815

```
nGroup <- accSIM30 %>%
  count(COND)
nGroup
```

	COND	n
1	Control	30
2	Low	30
3	High	30

```
nControl <- nGroup$n[1]
nControl
```

[1] 30

```
nLow <- nGroup$n[2]
nLow
```

[1] 30

```
nHigh <- nGroup$n[3]
nHigh
```

[1] 30

This formula occurs in three chunks, representing the control, low, and high racial loading conditions. In each of the chunks we have the n , group mean, and grand mean.

```
# Calculated by using object names from our calculations
SSM <- nControl * (ControlMean - GrandMean)^2 + nLow * (LowMean - GrandMean)^2 +
  nHigh * (HighMean - GrandMean)^2
SSM
```

[1] 9.432402

```
# calculated by specifying the actual values from our calculations
30 * (1.756 - 1.603)^2 + 30 * (1.9 - 1.603)^2 + 30 * (1.153 - 1.603)^2
```

[1] 9.42354

```
# Both result in the same
```

This value, SS_M is the amount of variance accounted for by the model; that is, the the amount of variance accounted for by the grouping variable/factor, COND. Degrees of freedom for SS_M is always one less than the number of elements (e.g., groups) used in its calculation ($k - 1$). Because we have three groups, our degrees of freedom for the model is two.

7.4.3 Sums of Squares Residual (or within)

To recap, we know there are 39.678 units of variation to be explained in our data. Our model explains 9.432 of these units. Sums of squares residual tells us how much of the variation cannot be explained by the model. This value is influenced by extraneous factors; some will refer to it as “noise.”

Looking at the formula can assist us in with a conceptual formula. In SS_R we subtract the group mean from each individual member of the group and then square it.

$$SS_R = \sum (x_{ik} - \bar{x}_k)^2$$

Below is another approach to calculating SS_R . In this one the variance for each group is multiplied by its respective degrees of freedom, then summed.

$$SS_R = s_{group1}^2(n-1) + s_{group2}^2(n-1) + s_{group3}^2(n-1)$$

Again, the formula is in three chunks – but this time the calculations are *within-group*. We need the variance (the standard deviation squared) for the calculation.

```
SDs <- aggregate(Accurate ~ COND, accSIM30, sd)
SDs
```

COND	Accurate
1 Control	0.4603964
2 Low	0.6301138
3 High	0.6587486

This script is used to create objects for each of the SDs associated with the grouping level. I created this so that I could demonstrate the formulas with words/terms as well as with numbers.

```
sdControl <- (SDs$Accurate[1])
sdControl
```

```
[1] 0.4603964
```

```
sdLow <- (SDs$Accurate[2])
sdLow
```

```
[1] 0.6301138
```

```
sdHigh <- (SDs$Accurate[3])
sdHigh
```

```
[1] 0.6587486
```

7.4.3.1 On the relationship between standard deviation and variance

Early in statistics training the difference between standard deviation (s or σ_{n-1}) and variance(s^2 or σ^2) can be confusing. This calculation demonstrates the relationship between standard deviation and variance. Variance is the standard deviation, squared.

```
#when squared, the standard deviation of the control group,
#should equal the variance reported in the next chunk
sdControl^2
```

```
[1] 0.2119648
```

```
VARs <- aggregate(Accurate ~ COND, accSIM30, var)
VARs
```

COND	Accurate
1 Control	0.2119648
2 Low	0.3970434
3 High	0.4339497

This script is used to extract the variances for each level of the grouping variable. I created them to be able to demonstrate the later formulas with words/terms as well as numbers.

```
varControl <- (VARs$Accurate[1])
varControl
```

```
[1] 0.2119648
```

```
varLow <- (VARs$Accurate[2])
varLow
```

```
[1] 0.3970434
```

```
varHigh <- (VARs$Accurate[3])
varHigh
```

```
[1] 0.4339497
```

We will use the second formula to calculate SS_R . For each of the groups, we multiply the variance by the respective degrees of freedom for the group ($n - 1$).

```
# Calculated by using object names from our calculations
SSR <- varControl * (nControl - 1) + varLow * (nLow - 1) + varHigh * (nHigh -
1)
```

```
# Re-calculated by specifying the actual values from our calculations
SSR
```

```
[1] 30.24578
```

```
0.212 * (30 - 1) + 0.397 * (30 - 1) + 0.434 * (30 - 1)
```

```
[1] 30.247
```

```
# Both result in the same
```

The value for our SS_R is 30.246. Degrees of freedom for the residual is $df_T - df_M$.

- df_T was $N - 1$: $90 - 1 = 89$
- df_M was $k - 1$: $3 - 1 = 2$
- Therefore, df_R : is $89 - 2 = 87$

7.4.4 Relationship between SS_T , SS_M , and SS_R .

In case it is not clear:

$$SS_T = 9.432 + 30.246$$

```
#calculated with object names
SSM + SSR
```

```
[1] 39.67818
```

```
#Re-calculated with the actual values
9.432 + 30.247
```

```
[1] 39.679
```

```
#Both result in the same
```

Our SST, calculated from above was 39.678.

7.4.5 Mean Squares Model & Residual

Our estimates of variation were *sums of squares* and are influenced by the number of scores that were summed. We can correct this bias by calculating their average – the *mean squares* or *MS*. We will use these in the calculation of the *F* ratio – the statistic that tests if there are significant differences between groups.

Like the constellation of sums of squares, we calculate mean squares for the model (MS_M) and residual(MS_R). Each formula simply divides the corresponding sums of squares by their respective degrees of freedom.

$$MS_M = \frac{SS_M}{df_M}$$

Regarding the calculation of our model mean squares:

- SS_M was 9.432
- df_M was 2
- Therefore, MS_M is:

```
#mean squares for the model
#calculated with object names
MSM <- SSM/dfM
MSM
```

[1] 4.716201

```
#Re-calculated with actual values
9.432/2
```

[1] 4.716

```
#Both result in the same
```

$$MS_R = \frac{SS_R}{df_R}$$

Regarding the calculation of our model residual squares:

- SS_R was 30.247
- df_R was 87
- Therefore, MS_R is:

```
#mean squares for the residual
#calculated with object names
MSR <- SSR/ dfR
MSR
```

[1] 0.3476526

```
#calculated with actual values
30.247/87
```

```
[1] 0.3476667
```

```
#Both result in the same
```

7.4.6 Calculating the F Statistic

The F statistic (or F ratio) assesses the ratio (as its name implies) of variation explained by the model to unsystematic factors (i.e., the residual). Earlier we used “between” and “within” language. Especially when we think of our example – where the model is composed of three groups, we can think of the F statistic as assessing the ratio of variation explained by between-subjects differences to within-subjects differences. Navarro’s [2020b] figures (earlier in the chapter) illustrate this well.

$$F = \frac{MS_M}{MS_R}$$

Regarding the calculation of our F -ratio:

- MS_M was 4.716
- MS_R was 0.348
- Therefore, F is:

```
#calculated with object names
Fratio <- MSM / MSR
Fratio
```

```
[1] 13.56584
```

```
#calculated with actual values
#Both result in the same
4.716/0.348
```

```
[1] 13.55172
```

7.4.7 Source Table Games

These last few calculations are actually less complicated than this presentation makes them seem. To better understand the relation between sums of squares, degrees of freedom, and mean squares, let’s play a couple of rounds of *Source Table Games*!

Rules of the game:

- In each case, mean squares are determined by dividing the sums of squares by its respective degrees of freedom.

- The F statistic is determined by dividing MS_M by MS_R

Knowing only two of the values, challenge yourself to complete the rest of the table. Before looking at the answers (below), try to fill in the blanks based in the table based on what we have learned so far.

Game	Total (df, $N - 1$)	Model (df, $k - 1$)	Residual (df, $df_T - df_M$)
SS	39.678(89)	9.432(2)	_____
MS	NA	_____	_____

$$F = MS_M/MS_R = \underline{\hspace{2cm}}$$

DON'T PEEK! TRY TO DO THE CALCULATIONS IN THE “SOURCE TABLE GAMES” EXERCISE BEFORE LOOKING AT THESE ANSWERS

Answers	Total (df, $N - 1$)	Model (df, $k - 1$)	Residual (df, $df_T - df_M$)
SS	39.678(89)	9.432(2)	30.247(87)
MS	NA	4.716 ^c	0.348

$$F = MS_M/MS_R = 13.566$$

To determine whether or not it is statistically significant, we can check a [table of critical values \[Zach, 2019\]](#) for the F test.

Our example has 2 (numerator) and 87 (denominator) degrees of freedom. Rolling down to the table where $\alpha = .05$, we can see that any F value > 3.11 (a value somewhere between 3.07 and 3.15) will be statistically significant. Our $F = 13.566$, so we have clearly exceeded the threshold. This is our *omnibus F test*.

We can also use a look-up function, which follows this general form: `qf(p, df1, df2, lower.tail=FALSE)`

```
qf(0.05, 2, 87, lower.tail = FALSE)
```

```
[1] 3.101296
```

Significance at this level lets us know that there is at least 1 statistically significant difference between our control, low, and high racially loaded conditions. While it is important to follow-up to see where these significant differences lie, we will not do these by hand. Rather, let's rework the problem in R.

7.5 Working the One-Way ANOVA with R Packages

Let's rework the problem in R. We start at the top of the flowchart, evaluating the statistical assumptions.

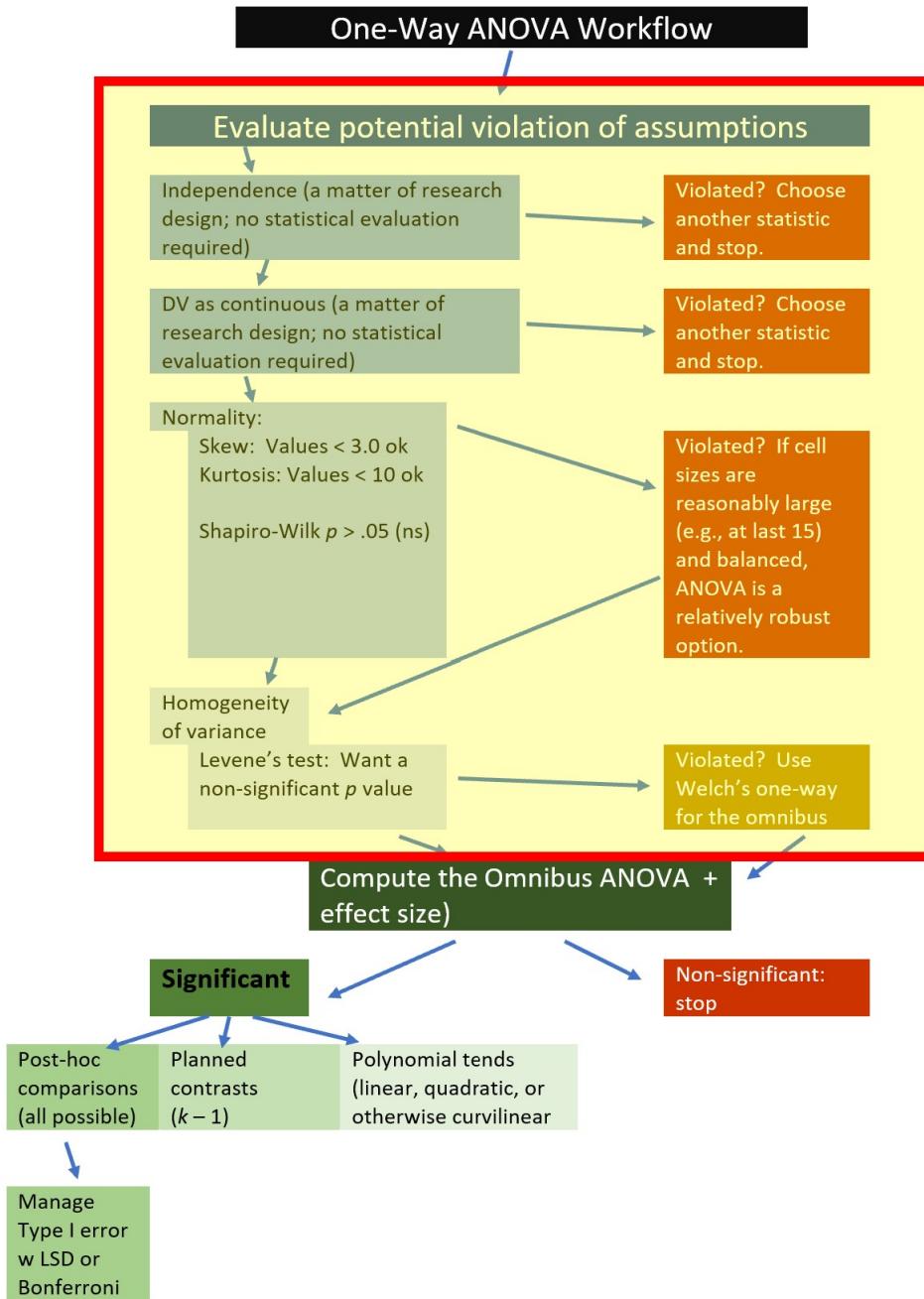


Figure 7.3: An image of the workflow for one-way ANOVA, showing that we are at the beginning: evaluating the potential violation of the assumptions.

7.5.1 Evaluating the Statistical Assumptions

All statistical tests have some assumptions about the data. The one-way ANOVA has four assumptions:

- The dependent variable is normally distributed for each of the populations as defined by the different levels of the factor. We will examine this by
 - evaluating skew and kurtosis
 - visually inspecting the distribution
 - conducting Shapiro-Wilk tests of normality
 - examining a QQ plot
- The variances of the dependent variable are the same for all populations. This is often termed the *homogeneity of variance* assumption. We will examine this with
 - Levene's Test
- The cases represent *random* samples from the populations and scores on the test variable are *independent* of each other. That is, comparing related cases (e.g., parent/child, manager/employee, time1/time2) violates this assumption and this question would need to be evaluated by a different statistic such as **repeated measures ANOVA** or dyadic data analysis.
 - *Independence* in observations is a research design issue. ANOVA is not robust to violating this assumption. When observations are correlated/dependent there is a dramatic increase in Type I error.
- The dependent variable is measured on an interval scale.
 - If the dependent variable is categorical, another statistic (such as logistic regression) should be chosen.

7.5.1.1 Is the dependent variable normally distributed across levels of the factor?

From the *psych* package, the *describe()* function can be used to provide descriptive statistics (or, “descriptives”) of continuously scaled variables (i.e., variables measured on the interval or ratio scale). In this simple example, we can specify the specific continuous, DV.

```
# we name the function in parentheses we list data source
psych::describe(accSIM30$Accurate, type = 1) #the type=1 argument provides the specific skew
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	90	1.6	0.67	1.73	1.62	0.68	0	3	3	-0.29	-0.43	0.07

If we want descriptives for each level of the grouping variable (factor), we can use the *describeBy()* function of the *psych* package. The order of entry within the script is the DV followed by the grouping variable (IV). In our research vignette below, I mentally interpret the *Accurate ~ COND* formula as, “Accurate by condition.”

```

# It is unnecessary to create an object, but an object allows you to
# do cool stuff, like write it to a .csv file and use that as a basis
# for APA style tables In this script we can think 'Accurate by COND'
# meaning that the descriptives for accuracy will be grouped by COND
# which is a categorical variable mat = TRUE presents the output in
# matrix (table) form digits = 3 rounds the output to 3 decimal
# places data = accSIM30 is a different (I think easier) way to
# identify the object that holds the dataframe
des.mat <- psych::describeBy(Accurate ~ COND, mat = TRUE, digits = 3, data = accSIM30,
    type = 1)
# Note. Recently my students and I have been having intermittent
# struggles with the describeBy function in the psych package. We
# have noticed that it is problematic when using .rds files and when
# using data directly imported from Qualtrics. If you are having
# similar difficulties, try uploading the .csv file and making the
# appropriate formatting changes. displays the matrix object that we
# just created
des.mat

```

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max
Accurate1	1	Control	1	30	1.756	0.460	1.893	1.767	0.392	0.781	2.745
Accurate2	2	Low	1	30	1.900	0.630	2.007	1.918	0.458	0.655	3.000
Accurate3	3	High	1	30	1.153	0.659	1.131	1.128	0.743	0.000	2.669
	range	skew	kurtosis	se							
Accurate1	1.964	-0.289	-0.364	0.084							
Accurate2	2.345	-0.398	-0.288	0.115							
Accurate3	2.669	0.218	-0.528	0.120							

```

# optional to write it to a .csv file for further manipulation and
# formatting for a paper or presentation
write.csv(des.mat, file = "Table1.csv")

```

Skew and kurtosis are one way to evaluate whether or not data are normally distributed. When we use the “type=1” argument, the skew and kurtosis indices in the *psych* package can be interpreted according to Kline’s [2016a] guidelines. Regarding skew, values greater than the absolute value of 3.0 are generally considered “severely skewed.” Regarding kurtosis, “severely kurtotic” is argued to be anywhere greater 8 to 20. Kline recommended using a conservative threshold of the absolute value of 10.

The *Shapiro-Wilk* test evaluates the hypothesis that the distribution of the data deviates from a comparable normal distribution. If the test is non-significant ($p > .05$) the distribution of the sample is not significantly different from a normal distribution. If, however, the test is significant ($p < .05$), then the sample distribution is significantly different from a normal distribution. The *rstatix* package can conduct this test for us.

```
library(tidyverse)
shapiro <- accSIM30 %>%
  group_by(COND) %>%
  rstatix::shapiro_test(Accurate)
shapiro
```

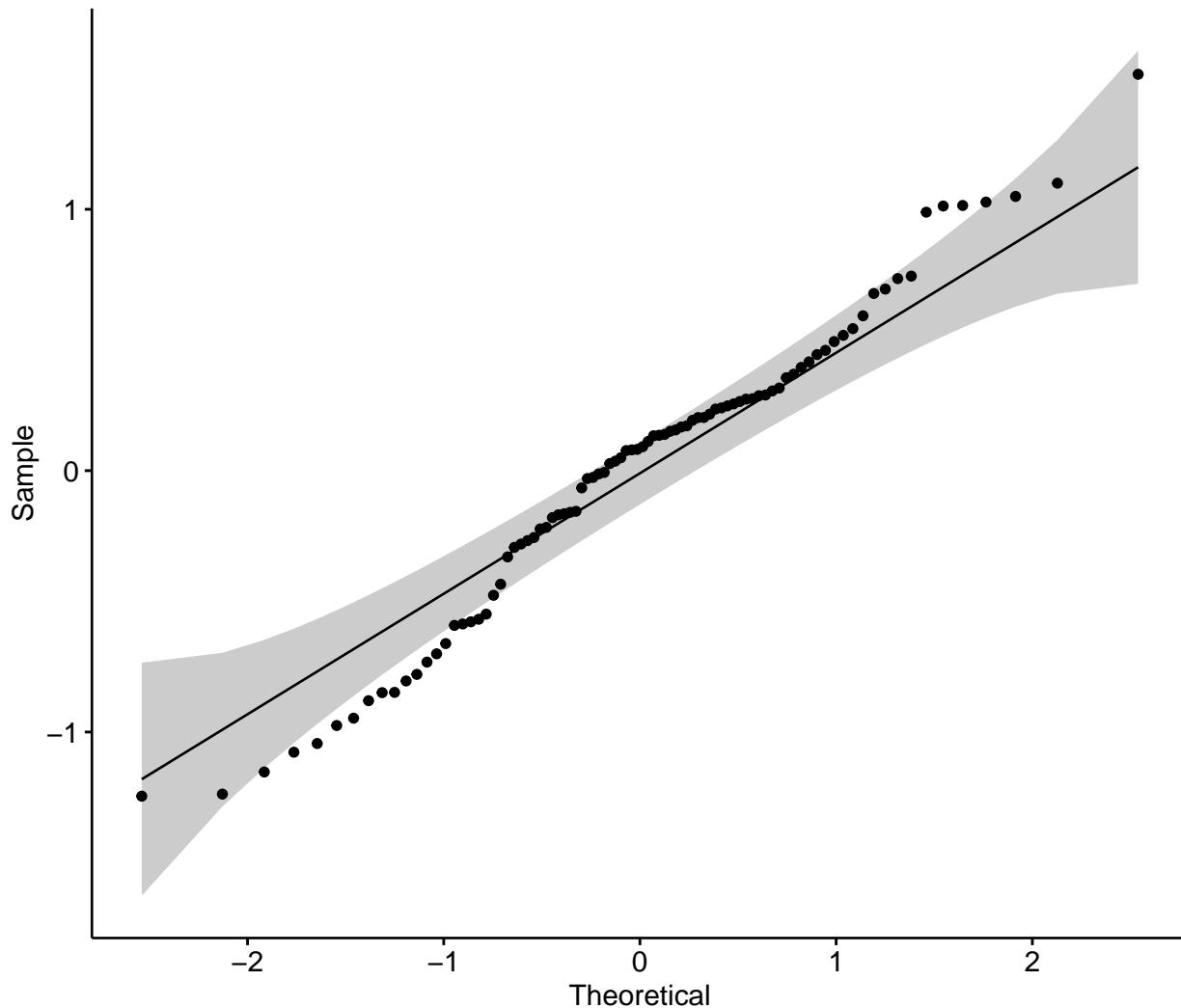
```
# A tibble: 3 x 4
  COND    variable statistic     p
  <fct>   <chr>      <dbl> <dbl>
1 Control Accurate    0.954 0.215
2 Low     Accurate    0.944 0.115
3 High    Accurate    0.980 0.831
```

The p values for the distributions of the dependent variable (accurate) in each of the three conditions are all well above .05. This tells us that the Accurate variable does not deviate from a statistically significant distribution at any level (Control, $W = 0.954$, $p = 0.215$; Low, $W = 0.944$, $p = 0.115$; High, $W = 0.980$, $p = 0.831$).

Especially in the more simple “ANOVA’s” I like this form of the Shapiro-Wilk test because it makes it clear that we expect normality within each of the grouping levels. This approach, however, is only appropriate when there are a low number of levels/groupings and there are many data points per group. As models become more complex, researchers should use the model-based option for assessing normality. To do this, we first create an object that tests our research model.

Although that model (a regression model) has information about our primary statistic, we are using it to carefully investigate the assumption of normality. One product of the analysis is *residuals*. Residuals are the unexplained variance in the outcome (or dependent) variable after accounting for the predictor (or independent) variable. When we plot these “leftovers” against the values of x, we can visualize the fit of the model in a QQ plot. The dots represent the residuals. When they are relatively close to the line they not only suggest good fit of the model, but we know they are small and evenly distributed around zero (i.e., normally distributed).

```
res_model <- lm(Accurate ~ COND, data = accSIM30)
ggnpubr::ggqqplot(residuals(res_model))
```



We can also use the model in a Shapiro-Wilk test. As before, we want a non-significant result.

```
rstatix::shapiro_test(residuals(res_model))
```

```
# A tibble: 1 x 3
  variable      statistic p.value
  <chr>          <dbl>     <dbl>
1 residuals(res_model) 0.979    0.150
```

These results parallel what we have already learned. That is, the non-significant p value associated with the model-based Shapiro-Wilk test of normality indicates that our distribution of residuals does not differ from a normal distribution ($W = 0.979, p = 0.15$). Given the space restrictions in journal articles and the greater interest in results of the primary analyses, I am more likely to report model-level results than the results from the cell-based Shapiro-Wilk tests.

There are limitations to the Shapiro-Wilk test. As the dataset being evaluated gets larger, the Shapiro-Wilk test becomes more sensitive to small deviations; this leads to a greater probability of rejecting the null hypothesis (null hypothesis being the values come from a normal distribution).

Green and Salkind [2017c] advised that ANOVA is relatively robust to violations of normality if there are at least 15 cases per cell and the design is reasonably balanced (i.e., equal cell sizes).

7.5.1.2 Should we consider removing outliers?

If our data pointed to significant violations of normality, we could consider identifying and removing outliers. Removing data is a serious consideration that should not be made lightly. If needed, though, here is a tool to inspect the data and then, if necessary, remove it.

We can think of outlier identification in a couple of ways. First, we might look at dependent variable across the entire dataset. That is, without regard to the levels of the grouping variable. We can point `rstatix::identify_outliers()` to the data.

```
accSIM30 %>%
  rstatix::identify_outliers(Accurate)

[1] ID      COND      Accurate moreTalk m_dev      m_devSQ      is.outlier
[8] is.extreme
<0 rows> (or 0-length row.names)
```

The output “0 rows” is not an error. It means if we consider the distribution of the Accurate variable as a whole, there are no outliers. Let’s re-run the code, this time requiring it to look within each of the grouping levels of the condition variable.

```
accSIM30 %>%
  group_by(COND) %>%
  rstatix::identify_outliers(Accurate)

# A tibble: 2 x 8
  COND   ID   Accurate moreTalk  m_dev m_devSQ is.outlier is.extreme
  <fct> <fct>    <dbl>    <dbl>  <dbl>  <dbl> <lgl>     <lgl>
1 Low    31     0.663   -1.10  -0.940  0.884 TRUE      FALSE
2 Low    39     0.655   -0.201 -0.948  0.899 TRUE      FALSE
```

This output tells us that in the low-racial loading condition there are two cases that are identified as outliers (denoted as TRUE) but not as extreme outliers (denoted as FALSE). Handily, the function returns information (i.e., the values of the Accurate and moreTalk variables, the ID number) that would help us delete it.

Let’s say that, after very careful consideration, we decided to remove the case with ID = 31. We could use `dplyr::filter()` to do so. In this code, the `filter()` function locates all the cases where ID = 31. The exclamation point that precedes the equal sign indicates that the purpose is to remove the case.

```
# accSIM30 <- dplyr::filter (accSIM30, ID != '31')
```

Once executed, we can see that this case is no longer in the dataframe. Although I demonstrated this in the accompanying lecture, I have hashtags out the command because I would not delete the case. If you already deleted the case, you can return the hashtag and re-run all the code up to this point.

7.5.1.3 Are the variances of the dependent variable similar across the levels of the grouping factor?

The Levene's test evaluates the ANOVA assumption that variances of the dependent variable for each level of the independent variable are similarly distributed. We want this to be non-significant ($p > .05$). If violated, we need to use an ANOVA test that is “robust to the violation of the homogeneity of variance” (e.g., Welch's oneway).

```
rstatix::levene_test(accSIM30, Accurate ~ COND)
```

```
# A tibble: 1 x 4
  df1    df2 statistic     p
  <int> <int>    <dbl> <dbl>
1     2     87      1.70  0.190
```

We write the result of the Levene's as $F(2, 87) = 1.695, p = 0.190$. Because $p > .05$, we know that the result is nonsignificant – that the variances of the three groups are not statistically significantly different from each other. If the results had been statistically significantly different, we would have needed to use a Welch's F or robust version of ANOVA.

7.5.1.4 Summarizing results from the analysis of assumptions

It is common for an APA style results section to begin with a review of the evaluation of the statistical assumptions. As we have just finished these analyses, I will document what we have learned so far:

Regarding the assumption of normality, skew and kurtosis values at each of the levels of the condition value fell well below the thresholds that Kline [2016a] identified as concerning (i.e., below |3| for skew and |10| for kurtosis). Similarly, no extreme outliers were identified and results of a model-based Shapiro-Wilk test of normality, indicated that the model residuals did not differ from a normal distribution ($W = 0.979, p = 0.15$). Finally, Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption ($F[2, 87] = 1.695, p = 0.190$).

7.5.2 Computing the Omnibus ANOVA

Having met all the assumptions, we are now ready to calculate the omnibus F test. *Omnibus* is the term applied to the first F test that evaluates if all groups have the same mean [Chen et al., 2018]. If this test is not significant there is no evidence in the data to reject the null; that is, there

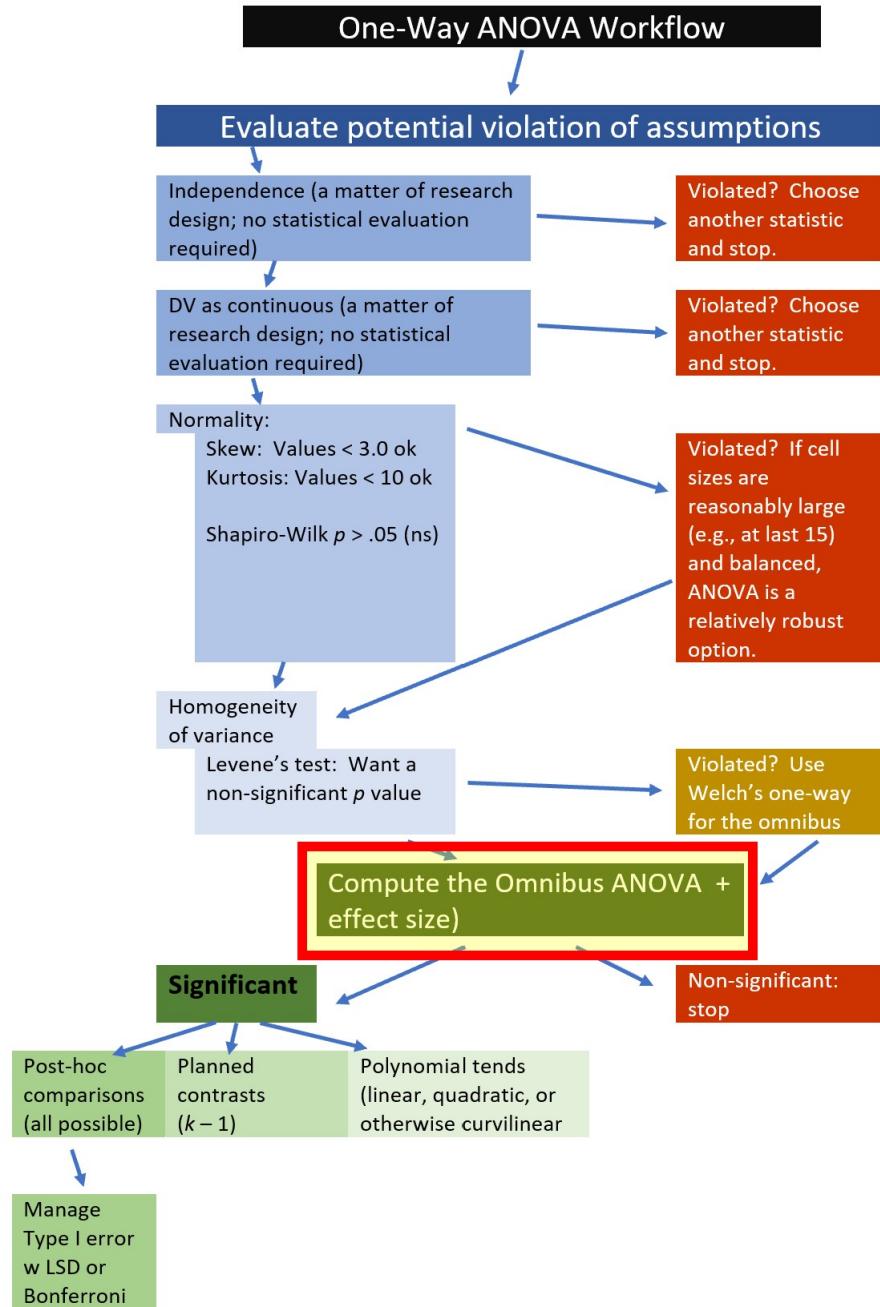


Figure 7.4: An image of the workflow for one-way ANOVA, showing that we are at the stage of computing the omnibus ANOVA.

is no evidence to suggest that group means are different. If it is significant – and there are three or more groups – follow-up testing will be needed to determine where the differences lie.

We will use `rstatix::anova_test` to calculate the omnibus. In script we must point to the data and provide the formula (`Accurate ~ COND`). By specifying “`detailed=TRUE`” we get can view our sums of squares values. When we run this test, we will save all of the results in an object. We can name this object anything – I will call it `omnibus1w`. When we create objects, we have to re-type the name of the object below our formula in order for the results to display. Objects are incredibly useful because we can later use them in follow-up tests, in creating figures, and in exporting results that we can use outside of R (e.g., to create tables for papers or presentations).

```
omnibus1w <- rstatix::anova_test(accSIM30, Accurate ~ COND, detailed = FALSE)
omnibus1w
```

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	COND	2	87	13.566	0.00000745	*	0.238

The values we see map onto those we calculated by hand. Our SS_M (9.432) plus SS_R (30.246) sum to equal the SS_T (39.678). Dividing the two sums of squares by their respective degrees of freedom produces the means squared. Then, dividing the MS_M (COND) by MS_R (4.716/0.348) provides the F ratio. By using a table of F critical values, we already knew that our F value exceeded the value in the table of critical values. Here we see that $p < .001$.

The “ F string” for an APA style results section should be written like this: $F(2, 87) = 13.566, p < .001$.

7.5.2.1 Effect size for the one-way ANOVA

Eta squared is one of the most commonly used measures of effect. It refers to the proportion of variability in the dependent variable/outcome that can be explained in terms of the independent variable/predictor. Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

You may see different values (.02, .13, .26) suggested as small, medium, and large effects – these values are used when multiple regression is used. A useful summary of effect sizes, guide to interpreting their magnitudes, and common usage can be found [here \[Watson, 2020\]](#).

The formula for η^2 is straightforward. If we think back to our hand-calculations of all the sums of squares, we can see that this is the proportion of variance that is accounted for by the model.

$$\eta^2 = \frac{SS_M}{SS_T}$$

Hand calculation, then, is straightforward.:

```
9.432/(9.432 + 30.246)
```

```
[1] 0.2377136
```

Luckily, `rstatix::anova_test()` has provided the η^2 for us; it is found in the column, `ges`. Using the interpretive criteria suggests that our effect is rather large. We can update our F string this way: $F(2, 87) = 13.566, p < .001, \eta^2 = 0.238$. An APA style write-up of the omnibus might read like this:

Results of the omnibus ANOVA indicated a significant effect of COND on accuracy perception ($F[2, 87] = 13.566, p < .001, \eta^2 = 0.238$).

7.5.3 Follow-up to the Omnibus F

The F -test associated with the one-way ANOVA is the *omnibus* – giving the result for the overall test. Looking at the workflow for the one-way ANOVA we see that if we had had we had a non-significant F , we would have stopped our analysis.

However, if the omnibus F is significant, we know that there is at least one pair of cells where there is a statistically significant difference. We have several ways (each with its own strengths/limitations) to figure out where these differences lie.

7.5.3.1 Planning for the management of Type I Error

Type I error is the concern about false positives – that we would incorrectly reject a true null hypothesis (i.e., claiming a statistically significant difference when there is not one). In ANOVA, we become increasingly concerned about Type I error as the number of pairwise or post hoc comparison increases. In ANOVA, we generally begin controlling for Type I error when follow-up to a significant omnibus test.

The *traditional Bonferroni* is, perhaps, the most well-known approach to managing Type I error. Although the lessons in this OER will frequently suggest alternative approaches to managing Type I error, I will quickly review it now because it is relatively straightforward and intuitive. We start by establishing the α_{family} ; this is traditionally $p = .05$.

Next, we determine how many pairwise comparisons that we are going to conduct. If we want to conduct all possible comparisons, we could use this formula to determining the number: $N_{pc} = \frac{N_g(N_g - 1)}{2}$, where

- N_{pc} is the number of pairwise comparisons, and
- N_g is the number of groups.

In the current research vignette, the COND factor had three levels: control, low, high. Thus, if we wanted to conduct all possible comparisons we would determine N_{pc} this way:

```
3*(3-1)/2
```

```
[1] 3
```

Subsequently, we would compute a new alpha that would be used for each comparison with this formula: $\alpha_{pc} = \frac{\alpha_{family}}{N_{pc}}$.

In the current research vignette we would calculate it this way:

```
.05/3
```

```
[1] 0.01666667
```

If we were to use the traditional Bonferroni to manage Type I error, the resultant p value would need to be $< .017$ in order for statistical significance to be claimed.

Luckily, the traditional Bonferroni (and other approaches to managing Type I error) has been reverse-engineered so that we do not have to determine the more conservative alpha levels. Rather, when we specify these options in the R script, the p value is adjusted and we can continue to use the customary criteria of $p < .05$, $p < .01$, and $p < .001$. In the case of the traditional Bonferroni, the p value has been adjusted upward by multiplying it (i.e., the raw p values) by the number of comparisons being completed. This holds the *total* Type I error rate across these tests to be $\alpha = 0.05$.

Although the traditional Bonferroni is easy-to-understand and compute, it has been criticized as being too restrictive. That is, it increases the risk of making a Type II error (i.e., failing to reject the null hypothesis when it is false). Therefore, as we work through each option for follow-up testing for the ANOVA models, I will introduce one or more methods for managing Type I error that are commonly used with that follow-up. Descriptions of all the methods for managing Type I error that are used in this OER are described in an [appendix](#)

7.5.3.2 OPTION #1: Post hoc, pairwise, comparisons

A very common follow-up to the omnibus test from a one-way ANOVA is to conduct post hoc, pairwise comparisons, of all possible combinations of pairs.

Post hoc, pairwise comparisons are:

- used for exploratory work when no firm hypotheses were articulated a priori,
- used to compare the means of all combinations of pairs of an experimental condition,
- less powerful than planned comparisons because more strict criterion for significance should be used.

By specifying the *formula* of the ANOVA, the *rstatix::t_test()* function will provide comparisons of all possible combinations. The arguments in the code mirror those we used for the omnibus. Note that I am saving the results as an object. We will use this object (“ttest”) later when we create an accompanying figure.

We will request the traditional Bonferroni using the *p.adjust.method*. The *rstatix::t_test()* offers multiple options for adjusting the p values.

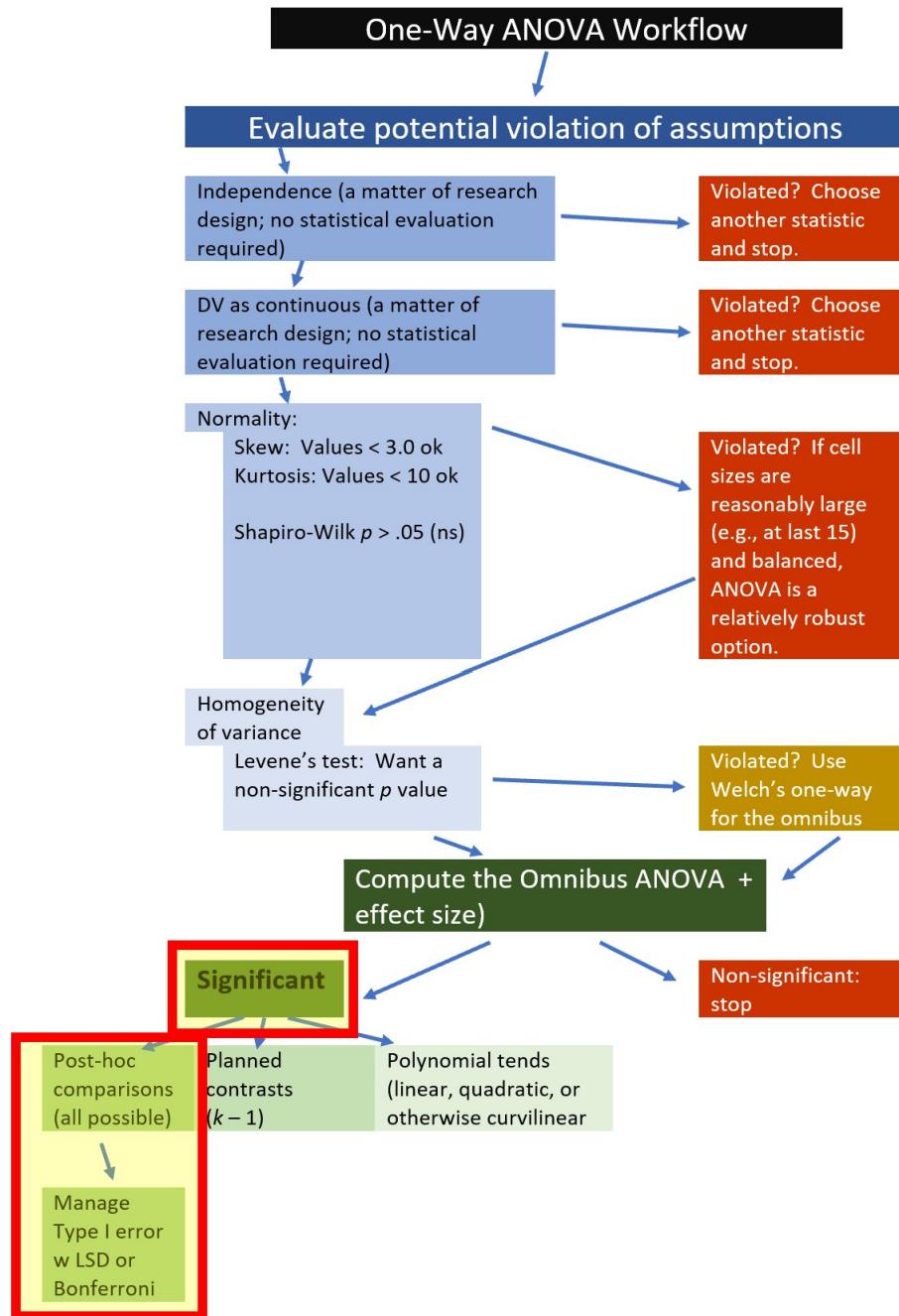


Figure 7.5: An image of the workflow for one-way ANOVA, showing that we are at the stage of following a statistically significant omnibus F test and are now conducting post hoc comparisons.

```
ttest <- rstatix::t_test(accSIM30, Accurate ~ COND, p.adjust.method = "bonferroni",
  detailed = TRUE)
ttest

# A tibble: 3 x 17
  estimate estimate1 estimate2 .y. group1 group2     n1     n2 statistic      p
*   <dbl>     <dbl>     <dbl> <chr> <chr> <chr> <int> <int>    <dbl>    <dbl>
1   -0.144     1.76     1.90 Accu~ Contr~ Low       30     30    -1.01 3.17e-1
2    0.603     1.76     1.15 Accu~ Contr~ High      30     30     4.11 1.4e-4
3    0.747     1.90     1.15 Accu~ Low     High      30     30     4.49 3.45e-5
# i 7 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>, p.adj <dbl>, p.adj.signif <chr>
```

The *estimate* column provide the mean difference between the two levels of the independent different. The *estimate1/group1* and *estimate2/group2* columns provide those means and identify the group levels. The *statistic* column provides the value of the *t*-test.

The *p* value is the unadjusted *p*-value, it will usually be “more significant” (i.e., a lower value) than the *p.adj* value that we specified in our code. The column *p.adj.signif* provides symbolic notation associated with the “*p.adj*” value. In this specific case we specified the traditional Bonferroni as the adjusted *p* value.

An APA style results section of this follow-up might read like this:

We followed up the significant omnibus with a series of post hoc, pairwise comparisons. We controlled for Type I error with the traditional Bonferroni adjustment. Results suggested that there were statistically significant differences between the control and high ($M_{diff} = 0.601, p < .001$) and low and high ($M_{diff} = 0.75, p < 0.001$) conditions, but not control and low conditions ($M_{diff} = -.14, p = 0.951$). Consequently, it appeared that only the highest degree of racial loading (e.g., “You speak English well for an Asian”) resulted in the decreased perceptions of accuracy of impressions from the confederate. Means and standard deviations are presented in Table 1 and complete ANOVA results are presented in Table 2. Figure 1 provides an illustration of the results.

Below is an augmentation of the figure that appeared at the beginning of the chapter. We can use the objects from the omnibus tests (named, “omnibus1w”) and post hoc pairwise comparisons (“*ttest*”) to add the ANOVA string and significance bars to the figure. Although they may not be appropriate in every circumstance, such detail can assist the figure in conveying maximal amounts of information.

```
# updates the ttest object so that it will auto-compute p-value
# positions in the graph
ttest <- ttest %>%
  rstatix::add_xy_position(x = "COND")

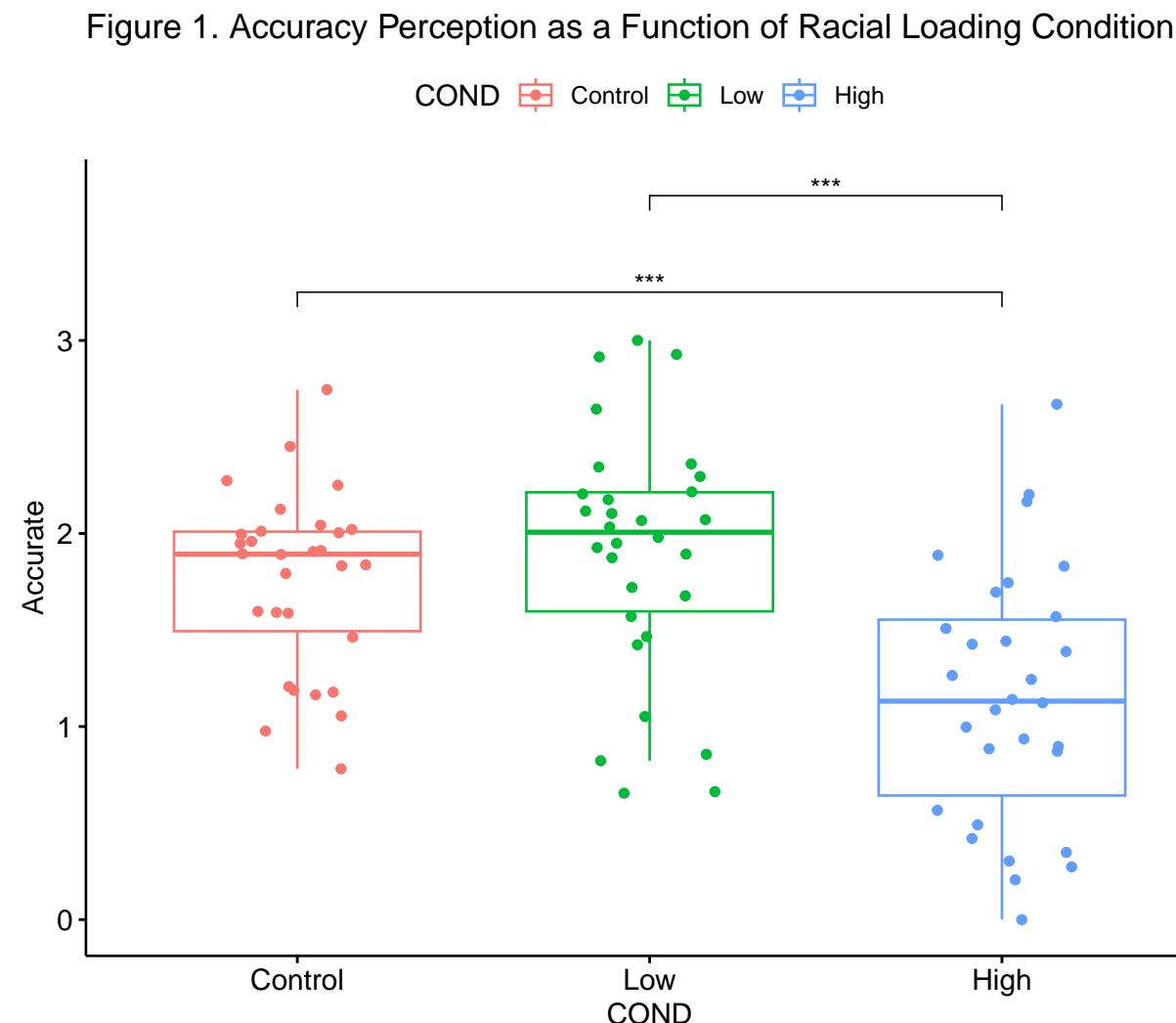
# our original plot
Fig1 <- ggpubr::ggboxplot(accSIM30, x = "COND", y = "Accurate", add = "jitter",
```

```

color = "COND", title = "Figure 1. Accuracy Perception as a Function of Racial Loading Condition"
ggpubr::stat_pvalue_manual(ttest, label = "p.adj.signif", tip.length = 0.02,
  hide.ns = TRUE, y.position = c(3.25, 3.75))
# tip.length instructs how long to make the dropped edges of the
# significance bar; hide.ns will suppress or display non-significant
# bars; step.increase will separate the bars from each other so that
# they do not overlap

```

Fig1



Although it would not make a difference in this research vignette, the LSD (least significant differences) method is commonly used for controlling Type I error in the follow-up to a one-way ANOVA. The LSD method is appropriate in the one-way ANOVA scenario when there are only three levels in the factor. In this case, Green and Salkind [2017c] have suggested that alpha could be retained at the alpha level for the “family” (α_{family}), which is conventionally $p = .05$ and used both to evaluate the omnibus and, so long as they don’t exceed three in number, the planned or pairwise comparisons that follow.

7.5.3.3 OPTION #2: Non-orthogonal planned contrast

Another option for follow-up to a significant omnibus test is to evaluate *planned* comparisons. These can either be orthogonal (i.e., a complete partitioning of variance) or *non-orthogonal* (i.e., allowing for overlapping variance). We will start with a non-orthogonal example.

Planned comparisons are

- theory-driven comparisons constructed prior to data collection,
- based on the idea of partitioning the variance created by the overall effect of group differences into gradually smaller portions of variance, and
- more powerful than post hoc tests.

Planned contrasts involve further considerations regarding the *partitioning of variance*.

- There will always be $k - 1$ contrasts; in our case this means we will have two contrasts
- Each contrast must involve only two *chunks* of variance.

If the researcher chooses this route, they must decide which two comparisons will best tell the story of the data as it relates to the hypotheses and a priori theory. I will compare differences between the no and low racial loading conditions, and then between low and high racial loading conditions. I have chosen to not adjust the p values. In the results write-up, I will reference the LSD method as my rationale for this approach.

```
contr2 <- rstatix::pairwise_t_test(accSIM30, Accurate ~ COND, comparison = list(c("Control",
  "Low"), c("Low", "High")), p.adjust.method = "none", detailed = TRUE)
contr2

# A tibble: 2 x 10
  .y.    group1 group2   n1   n2      p method  p.adj p.signif p.adj.signif
* <chr>  <chr>  <chr> <int> <int>  <dbl> <chr>  <dbl> <chr>    <chr>
1 Accura~ Contr~ Low     30    30  3.47e-1 T-test 3.47e-1 ns        ns
2 Accura~ Low     High    30    30  4.25e-6 T-test 4.25e-6 ****     ****
```

The format of the output is quite similar to the preceding examples. One difference is that this function does not provide mean differences nor confidence intervals. If I wanted them, I would need to calculate them.

An APA style results section of this follow-up might read like this:

We followed up the significant omnibus with two, non-orthogonal, planned comparisons. Because we had fewer than three comparisons, we chose to retain alpha at .05. This is consistent with the LSD method for control of Type I error [Green and Salkind, 2017c]. Results suggested a statistically significant difference between the low and high ($M_{diff} = 0.75, p < .001$) conditions, but not between the control and low conditions ($M_{diff} = -.14, p = 0.347$). Consequently, it appeared that only the highest degree of racial loading (e.g., “You speak English well for an Asian”) resulted in the decreased perceptions of accuracy of impressions from the confederate. Means and standard deviations are presented in Table 1 and complete ANOVA results are presented in Table 2. Figure 2 provides an illustration of the results.

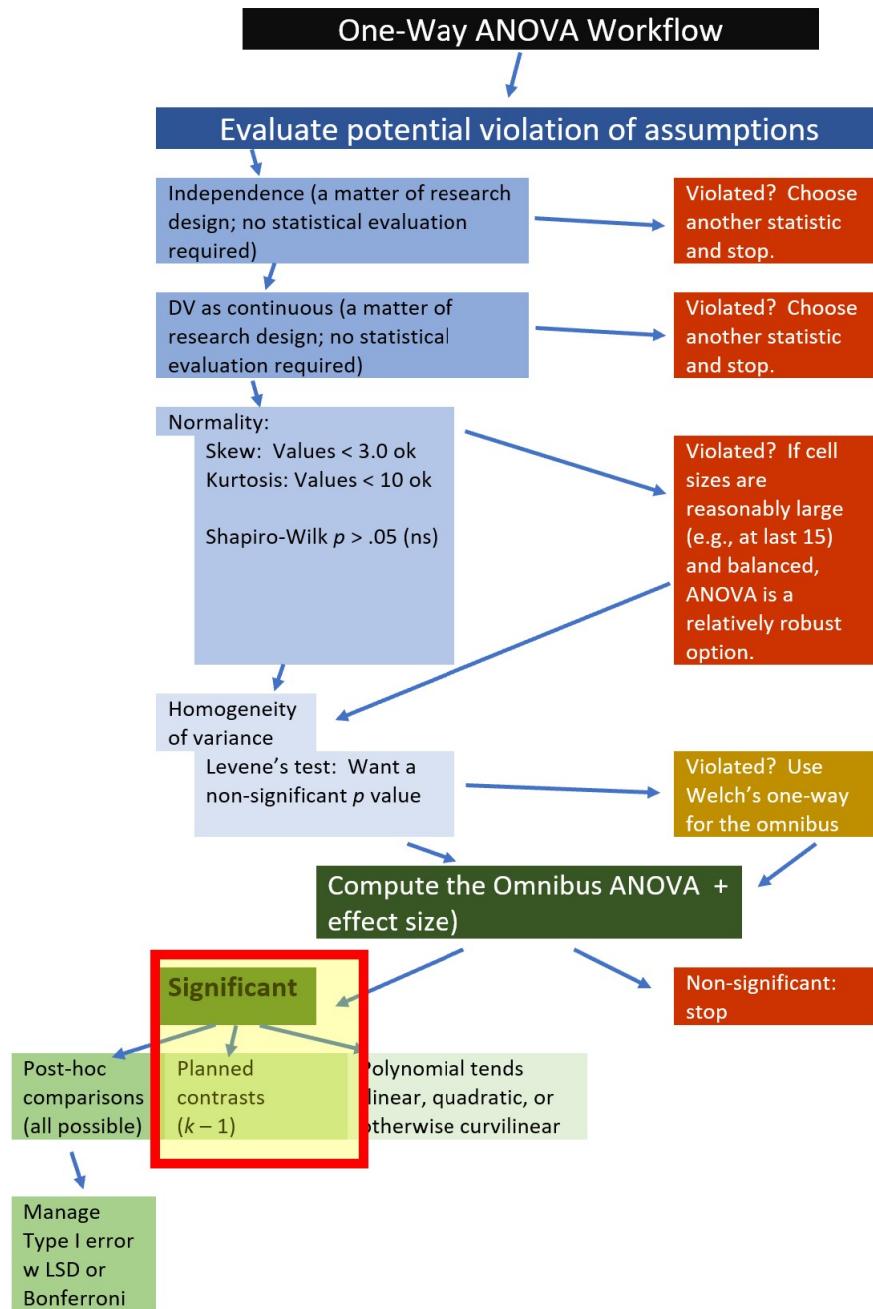


Figure 7.6: An image of the workflow for one-way ANOVA, showing that we are at the following up to a significant omnibus F by conducting planned comparisons

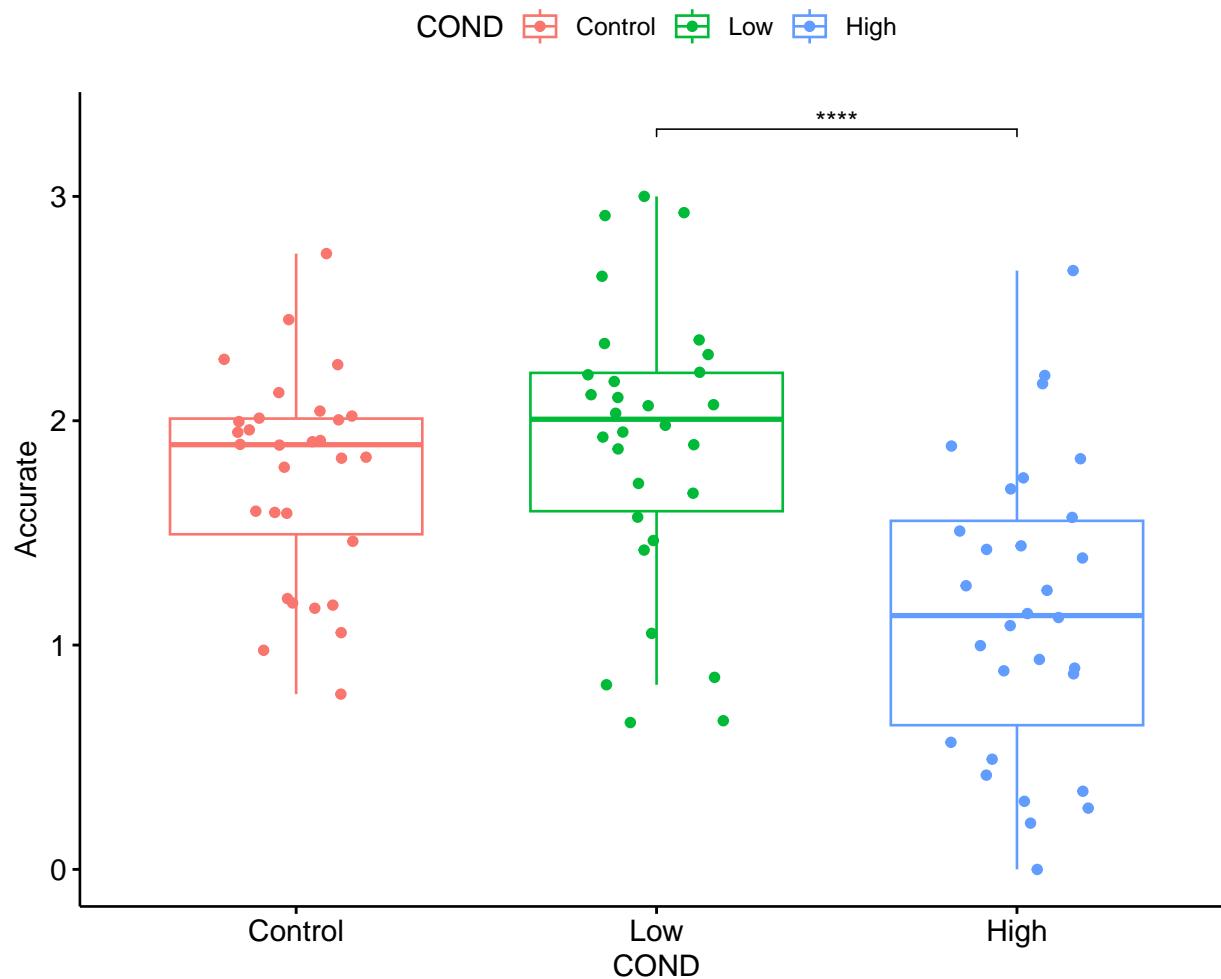
Below is an augmentation of the figure that appeared at the beginning of the chapter. We can use the previously created objects from the omnibus test (“omnibus1w”) and post hoc pairwise comparisons (“ttest”) to add the ANOVA string and significance bars to the figure. Although they may not be appropriate in every circumstance, such detail can assist the figure in conveying maximal amounts of information.

```
# updates the ttest object so that it will autocompute p-value
# positions in the graph
contr2 <- contr2 %>%
  rstatix::add_xy_position(x = "COND")

# our original plot
ggpubr::ggboxplot(accSIM30, x = "COND", y = "Accurate", add = "jitter",
  color = "COND", title = "Figure 2. Accuracy Perception as a Function of Racial Loading Condition")

# retrieves information from the contr2 object; label tells the
# figure to use the 'p.adj.signif' column in the contr2 output
ggpubr::stat_pvalue_manual(contr2, label = "p.adj.signif", tip.length = 0.01,
  hide.ns = TRUE, y.position = c(3.3)) #tip.length instructs how long to make the drop edges
```

Figure 2. Accuracy Perception as a Function of Racial Loading Condition



In this particular research vignette, I probably would not compute nor report the non-orthogonal option. The statistically significant difference pattern from the post hoc pairwise comparisons (Option 1) was straightforward. Using the non-orthogonal planned comparisons to aid in the control of Type I error way (a) does not change the result (i.e., does not increase the power) and (b) provides a less complete picture of the results.

7.5.3.4 OPTION #3: Orthogonal planned contrasts

Orthogonal contrasts are even more sophisticated. Essential to conducting an orthogonal contrast is the requirement that if a group is singled out in one comparison it should be excluded from subsequent contrasts. The typical, orthogonal scenario with three ordered groups has only two contrasts:

1. Control versus the combined low and high conditions
 - because control was excluded, it should not reappear in the next contrast

2. Low versus high

Especially in scenarios where there are no, low and high dose (or exposure) conditions, this is an elegant comparison. Unfortunately, at the time of this writing, the *rstatix* package does not offer a function to make these computations. We can, however, use functions from base R. Given that *rstatix* is a wrapper for the *aov()* function in base R, the code should feel somewhat familiar.

To work toward our orthogonal contrasts, we first need to create an object (“*omnibus1w_b*”) from a one-way ANOVA test using the base R, *aov()* function. You can see that the script involves the same elements as in *rstatix*. We can view the results by using the *summary()* function.

```
omnibus1w_b <- aov(Accurate ~ COND, data = accSIM30)
summary(omnibus1w_b)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
COND	2	9.432	4.716	13.57	0.00000745 ***						
Residuals	87	30.246	0.348								

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'..'	0.1	' '	1

This foray into orthogonal contrasts gives us a peek into multiple regression. Let’s take a peek at “regression results” from our ANOVA model.

```
summary.lm(omnibus1w_b)
```

```
Call:
aov(formula = Accurate ~ COND, data = accSIM30)

Residuals:
    Min      1Q      Median      3Q      Max 
-1.24533 -0.32092  0.08642  0.30101  1.51646 

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)    
(Intercept)  1.7562     0.1076 16.314 < 0.000000000000002 *** 
CONDLow      0.1439     0.1522  0.945     0.347095    
CONDHigh     -0.6034    0.1522 -3.963     0.000151 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.5896 on 87 degrees of freedom
Multiple R-squared:  0.2377,    Adjusted R-squared:  0.2202 
F-statistic: 13.57 on 2 and 87 DF,  p-value: 0.000007446
```

The values on the row labeled *intercept* are the values of the baseline or comparison group. Since CONDLow and CONDHIGH follow, we know the three groups are sensibly ordered as control (0), low (1), and high (2).

While we are here, we observe that the control mean is 1.76 and that this value is statistically significantly different than zero. The *CONDLow* row represents the low level of the condition variable. The mean for “low” is 0.14 units bigger than the control group and this is not a statistically significant difference ($p = 0.347$). The third row is the high level of the condition variable. This value is 0.60 units lower than the control condition and is statistically significantly different than zero ($p < .001$). This information is consistent with what we have already learned.

To move forward with the orthogonal contrasts we must first specify our contrasts.

- Specifying the contrasts means you know their order within the factor
- Early in the data preparation, we created an ordered factor with Control, Low, High as the order.
- We want orthogonal contrasts, this means there will be
 - $k - 1$ contrasts; with three groups we will have two contrasts
 - once we single out a condition for comparison, we cannot use it again.

In *contrast1* we compare the control condition to the combined low and high conditions. In *contrast2* we discard the control condition (it was already singled out) and we compare the low and high conditions.

This is sensible because we likely hypothesize that any degree of racially loaded stereotypes may have a deleterious outcome, so we first compare control to the two conditions with any degree of racial loading. Subsequently, we compare the low and high levels of the factor.

In the second step we must bind the contrasts together and check the output to ensure that we've mapped them correctly.

```
# Contrast1 compares Control against the combined effects of Low and
# High.
contrast1 <- c(-2, 1, 1)

# Contrast2 excludes Control; compares Low to High.
contrast2 <- c(0, -1, 1)

# binding the contrasts together
contrasts(accSIM30$COND) <- cbind(contrast1, contrast2)
accSIM30$COND
```

```
[1] High   High   High   High   High   High   High   High   High
[10] High  High   High   High   High   High   High   High   High
[19] High  High   High   High   High   High   High   High   High
[28] High  High   High   Low    Low    Low    Low    Low    Low
[37] Low   Low    Low    Low    Low    Low    Low    Low    Low
[46] Low   Low    Low    Low    Low    Low    Low    Low    Low
[55] Low   Low    Low    Low    Low    Low    Control Control Control
[64] Control Control Control Control Control Control Control Control
[73] Control Control Control Control Control Control Control Control
[82] Control Control Control Control Control Control Control Control
```

```
attr(", "contrasts")
  contrast1 contrast2
Control      -2      0
Low          1     -1
High          1      1
Levels: Control Low High
```

Thinking back to the hand-calculations and contrast mapping, the table of weights that R just produced confirms that

- Contrast 1 compares the Control condition against the levels with any racial loading.
- Contrast 2 compares the Low and High loadings.

Finally, we create a new *aov()* model and apply the contrasts.

```
# create a new objects
accPlanned <- aov(Accurate ~ COND, data = accSIM30)
summary.lm(accPlanned)
```

Call:
`aov(formula = Accurate ~ COND, data = accSIM30)`

Residuals:

Min	1Q	Median	3Q	Max
-1.24533	-0.32092	0.08642	0.30101	1.51646

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.60304	0.06215	25.793	< 0.0000000000000002 ***
CONDcontrast1	-0.07658	0.04395	-1.742	0.085 .
CONDcontrast2	-0.37365	0.07612	-4.909	0.00000425 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5896 on 87 degrees of freedom
Multiple R-squared: 0.2377, Adjusted R-squared: 0.2202
F-statistic: 13.57 on 2 and 87 DF, p-value: 0.000007446

These planned contrasts show that when the control condition is compared to the combined low and high racial loading conditions, there is not a statistically significant difference, $t(87) = -1.742$, $p = 0.085$. However, when the low and high racial loading conditions are compared, there is a statistically significant difference, $t(87) = -4.909$, $p < 0.001$. An APA style results write-up might look like this:

We followed the significant omnibus test with a pair of orthogonal contrasts. The first compared the control condition to the combined low and high racial loading conditions. The result was non-significant ($t[87] = -1.742, p = 0.085$). The second contrast compared the low and high racial loading conditions. In this contrast, accuracy ratings were statistically significantly lower for the high racial loading condition ($t[87] = -4.909, p < 0.001$).

At this point, I do not have script that would update the *ggpubr* graph with these results. If I were to use this follow-up in my APA style results, I would likely use the boxplot we produced at the beginning of the lesson.

7.5.3.5 OPTION #4: Trend (polynomial) analysis

Polynomial contrasts let us see if there is a linear (or curvilinear) pattern to the data. To detect a trend, the data must be coded in an ascending order...and it needs to be a sensible comparison. Here's where this would fall in our workflow.

To detect a trend, the data must be coded in an ascending order and the comparison needs to be sensible and theoretically defensible. Our data has a theoretically ordered effect (control/none, low, and high racially loaded conditions). Recall that we created an ordered factor when we imported the data. However, we can use the *contrasts()* function from base R to verify the order.

```
contrasts(accSIM30$COND)
```

	contrast1	contrast2
Control	-2	0
Low	1	-1
High	1	1

In a polynomial analysis, the statistical analysis looks across the ordered means to see if they fit a linear or curvilinear shape that is one fewer than the number of levels (i.e., $k - 1$). Because the COND factor has three levels, the polynomial contrast checks for linear (.L) and quadratic (one change in direction) trends (.Q). If we had four levels, *contr.poly()* could also check for cubic change (two changes in direction). Conventionally, when more than one trend is significant, we interpret the most complex one (i.e., quadratic over linear).

To the best of my knowledge, *rstatix* does not offer these contrasts. We can fairly easily make these calculations in base R by creating a set of polynomial contrasts. In the prior example we specified our contrasts through coding. Here we can the *contr.poly(3)* function. The “3” lets R know that there are three levels in COND. The *aov()* function will automatically test for quadratic (one hump) and linear (straight line) trends.

```
contrasts(accSIM30$COND) <- contr.poly(3)
accTrend <- aov(Accurate ~ COND, data = accSIM30)
summary.lm(accTrend)
```

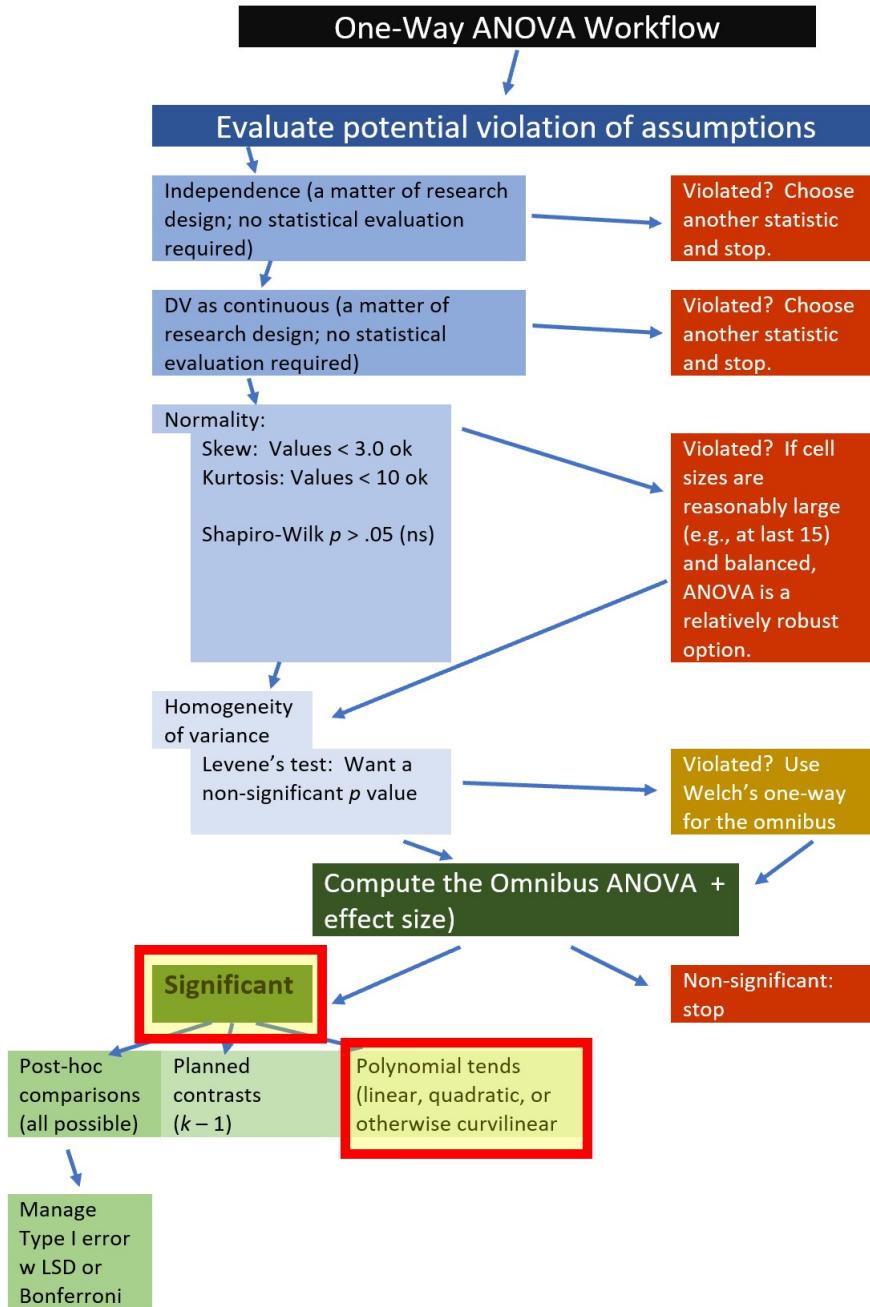


Figure 7.7: An image of the workflow for one-way ANOVA, showing that we are at the following up to a significant omnibus F by assessing for a polynomial trend

```

Call:
aov(formula = Accurate ~ COND, data = accSIM30)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.24533 -0.32092  0.08642  0.30101  1.51646 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.60304   0.06215 25.793 < 0.000000000000002 *** 
COND.L       -0.42665   0.10765 -3.963      0.000151 ***  
COND.Q       -0.36384   0.10765 -3.380      0.001087 **   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5896 on 87 degrees of freedom
Multiple R-squared:  0.2377,    Adjusted R-squared:  0.2202 
F-statistic: 13.57 on 2 and 87 DF,  p-value: 0.000007446

```

Results of our polynomial contrast suggested statistically significant results for both linear $t(87) = -3.963, p < .001$ and quadratic $t(87) = -3.380, p = .001$ trends. A quick peek back at any of our boxplots illustrates the quadratic trend (an small increase in accuracy from control to low; a larger decrease in accuracy from low to high) that was supported by this analysis.

Given that our earlier analyses did not support statistically significant differences between control and low racial loading conditions, I am disinclined to include this information. That said, there are times when I will include results of a polynomial trend along with the results of posthoc or planned pairwise comparisons. I will do this when the overall trend in the data helps clarify the results. For example, if in a circumstance where there was a clear linear trend between no, low, and high dose conditions and the pairwise results were consistent with that (i.e., statistically significant differences between no and low, no and high, low and high), I would likely add the results of the polynomial, after presenting the results of the posthoc or planned comparisons:

Additionally, results of a polynomial constrained suggested a statistically significant linear trend across the three conditions, $t(87) = -3.963, p < .001$.

At this point, I do not have script that would update the *ggpubr* figure in a manner that would clearly convey these results. If I were to use this follow-up in my APA style results, I would likely use the simple boxplot we produced at the beginning of the lesson.

7.5.3.6 Which set of follow-up tests do we report?

It depends! Here are some things to consider.

- If the post hoc comparisons are robustly statistically significant (and controlling Type I error is not going to change that significance), I would lean toward reporting those.

- If p values are hovering around 0.05, an orthogonal contrast will offer more power because
 - a $k - 1$ comparison will be more powerful and
 - (when the research design allows) the contrast of no dose/exposure to any exposure followed by a contrast between low and high doses/exposures is compelling.
- The polynomial can be a useful descriptive addition if there is a linear or quadratic relationship that is sensible or interesting.

Although I would report either the post hoc or planned contrasts, I will sometimes add a polynomial if it clarifies the result (i.e., there is a meaningful linear or curvilinear pattern essential to understanding the data).

7.5.4 What if we Violated the Homogeneity of Variance test?

The `rstatix::welch_anova_test` produces Welch's F – a test that is robust to violation of the homogeneity of variance assumption. The Welch's approach adjusts the residual degrees of freedom used to produce the Welch's F -ratio. The format of the argument is quite similar to what we have been doing all along.

```
omnibus_w <- rstatix::welch_anova_test(accSIM30, Accurate ~ COND)
omnibus_w
```

```
# A tibble: 1 x 7
  .y.       n statistic   DFn   DFd      p method
* <chr> <int>    <dbl> <dbl> <dbl>    <dbl> <chr>
1 Accurate     90     11.6    2  56.3 0.0000617 Welch ANOVA
```

Note that the denominator df is now 56.34 (not 87) and p value is a little larger (it was 0.00000745). With its design intended to avoid making a Type I error, the Welch's F is more restrictive. While it wouldn't alter the conclusions in our research vignette, it could if the p value was closer to 0.05. These are some of the tradeoffs we must consider in order to have confidence in the results. At this time the `rstatix::welch_anova_test()` function does not offer an effect size. The omega squared is an effect size that is commonly reported with the Welch's F . It would either need to be calculated by hand or with another R package.

In terms of follow-up to the omnibus test, `rstatix` includes Games-Howell pairwise comparisons and pairwise t-tests. Neither of these follow-up options requires the assumption of equal variance. Consequently either could be used as a follow-up. Here's an example from the Games-Howell test.

```
gw_pwc <- rstatix::games_howell_test(accSIM30, Accurate ~ COND, conf.level = 0.95,
                                       detailed = TRUE)
gw_pwc
```

```
# A tibble: 3 x 14
  .y.    group1 group2   n1   n2 estimate conf.low conf.high    se statistic
* <chr> <chr>  <chr> <int> <int>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
```

```

1 Accurate Contr~ Low      30     30    0.144   -0.200    0.487 0.101    1.01
2 Accurate Contr~ High    30     30   -0.603   -0.957   -0.249 0.104    4.11
3 Accurate Low    High    30     30   -0.747   -1.15    -0.347 0.118    4.49
# i 4 more variables: df <dbl>, p.adj <dbl>, p.adj.signif <chr>, method <chr>

```

Another common correction for evaluating the omnibus test when there is a violation of the homogeneity of variance assumption is the Brown and Forsythe F -ratio. The *rstatix* package does not include this option (but other packages do).

7.6 APA Style Results

All that's left to do to decide which set of follow-up tests to report and assemble the write-up. APA style results sections in empirical manuscripts are typically accompanied by tables and figures. APA style discourages redundancy in information (i.e., if information is clearly presented in a table, do not repeat it verbatim in written text) and encourages reducing the cognitive load of the reader. For this example, I suggest two tables – (a) one with means and standard deviations the dependent variable (disaggregated by level)and (b) a second that reports the output from the one-way ANOVA.

The package *apaTables* can produce journal-ready tables. Deciding what to report in text and table is important. First, I create Table 1 with means and standard deviations (plus a 95% confidence interval around each mean).

```

# table.number = 1 assigns a table number to the top of the table
# filename = 'Table1.doc' writes the table to Microsoft Word and puts
# it in your project folder
apaTables::apa.1way.table(iv = COND, dv = Accurate, show.conf.interval = TRUE,
                           data = accSIM30, table.number = 1, filename = "Table1.doc")

```

Table 1

Descriptive statistics for Accurate as a function of COND.

COND	M	M_95%_CI	SD
Control	1.76	[1.58, 1.93]	0.46
Low	1.90	[1.66, 2.14]	0.63
High	1.15	[0.91, 1.40]	0.66

Note. M and SD represent mean and standard deviation, respectively.

LL and UL indicate the lower and upper limits of the 95% confidence interval for the mean, respectively.

The confidence interval is a plausible range of population means that could have caused a sample mean (Cumming, 2014).

Next, I create Table 2 with source table for the one-way ANOVA. The result can be edited in Microsoft Word for the paper or presentation (e.g., I would replace the partial-eta squared with

η^2). One trick about *apaTables::aov* is that it requires an object from the base R's *aov* function. Recall that we used this in our contrasts. None-the-less, I will repeat it in this code.

```
omnibus1w_b <- aov(Accurate ~ COND, data = accSIM30)
apaTables::apa.aov.table(omnibus1w_b, table.number = 2, filename = "Table2.doc")
```

Table 2

ANOVA results using Accurate as the dependent variable

Predictor	SS	df	MS	F	p	partial_eta2	CI_90_partial_eta2
(Intercept)	231.28	1	231.28	665.25	.000		
COND	9.43	2	4.71	13.57	.000	.24	[.11, .34]
Error	30.25	87	0.35				

Note: Values in square brackets indicate the bounds of the 90% confidence interval for partial

Regarding figures, I would use the one I created with the set of follow-up results.

With table and figure at hand, here is how I would write up these results:

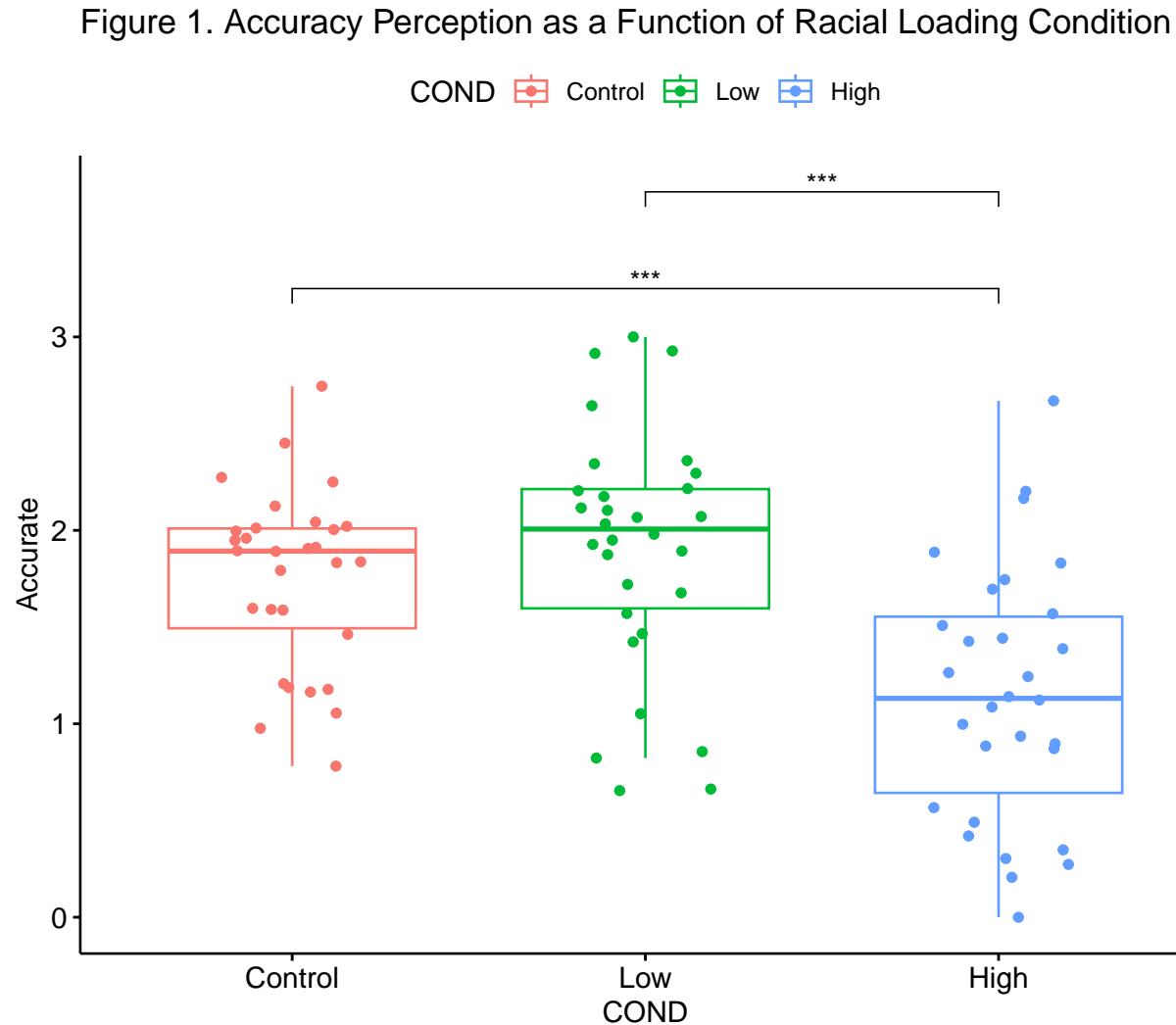
A one-way analysis of variance was conducted to evaluate the relationship between degree of racial loading of an exceptionalizing microaggression and the perceived accuracy of a research confederate's impression of the Asian or Asian American participant. The independent variable, condition, included three levels: control/none, low, and high levels of racial loading.

Regarding the assumption of normality, skew and kurtosis values at each of the levels of the condition value fell well below the thresholds that Kline [2016a] identified as concerning (i.e., below |3| for skew and |10| for kurtosis). Similarly, no extreme outliers were identified and results of a model-based Shapiro-Wilk test of normality, indicated that the model residuals did not differ from a normal distribution ($W = 0.979, p = 0.15$). Finally, Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption ($F[2, 87] = 1.695, p = 0.190$).

Results of the omnibus ANOVA indicated a significant effect of COND on accuracy perception ($F[2, 87] = 13.566, p < .001, \eta^2 = 0.238$). We followed up the significant omnibus with a series of post hoc, pairwise comparisons. We controlled for Type I error with the traditional Bonferroni adjustment. Results suggested that there were statistically significant differences between the control and high ($M_{diff} = 0.601, p < .001$) and low and high ($M_{diff} = 0.75, p < 0.001$) conditions, but not control and low conditions ($M_{diff} = -0.14, p = 0.951$). Consequently, it appeared that only the highest degree of racial loading (e.g., "You speak English well for an Asian") resulted in the decreased perceptions of accuracy of impressions from the confederate. Means

and standard deviations are presented in Table 1 and complete ANOVA results are presented in Table 2. Figure 1 provides an illustration of the results.

Fig1



7.7 Power Analysis

Power analysis allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence. Utilized another way, it allows us to determine the probability of detecting an effect of a given size with a given level of confidence. If the probability is unacceptably low, we may want to revise or stop. A helpful overview of power as well as guidelines for how to use the *pwr* package can be found at a [Quick-R website \[Kabacoff, 2017\]](#).

There are four interrelating elements of power:

1. Sample size, N

2. Effect size,

- For one-way ANOVAs, Cohen suggests that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes, respectively.

3. Significance level = $P(\text{Type I error})$,

- Recall that Type I error is the rejection of a true null hypothesis (a false positive).
- Stated another way, Type I error is the probability of finding an effect that is not there.

4. Power = $1 - P(\text{Type II error})$,

- Recall that Type II error is the non-rejection of a false null hypothesis (a false negative).
- Power is the probability of finding an effect that is there.

If we have any three of these values, we can calculate the fourth.

In Champely's *pwr* package, we can conduct a power analysis for a variety of designs, including the balanced one-way ANOVA (i.e., roughly equal cell sizes) design that we worked in this chapter.

The *pwr.anova.test()* has five parameters:

- k = # groups
- n = sample size
- f = effect sizes, where 0.1/small, 0.25/medium, and 0.4/large
 - In the absence from an estimate from our own data, we make a guess about the expected effect size value based on our knowledge of the literature
- sig.level = p value that you will use
- $\text{power} = .80$ is the standard value

In the script below, we simply add our values. So long as we have four values, the fifth will be calculated for us.

Because this calculator requires the effect size in the metric of Cohen's f (this is not the same as the F ratio), we need to convert it. The *effectsize* package has a series of converters. We can use the *eta2_to_f()* function.

```
effectsize::eta2_to_f(0.238)
```

```
[1] 0.5588703
```

We simply plug this value into the "f =".

```
pwr::pwr.anova.test(k = 3, f = 0.5589, sig.level = 0.05, power = 0.8)
```

```
Balanced one-way analysis of variance power calculation
```

```
k = 3
```

```
n = 11.3421
f = 0.5589
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

This result suggested that we would need 11 people per group.

If we were unsure about what to expect in terms of our results, we could take a guess. I like to be on the safe(r) side and go with a smaller effect. What would happen if we had a Cohen's f that represented a small effect?

```
pwr::pwr.anova.test(k = 3, f = 0.1, sig.level = 0.05, power = 0.8)
```

Balanced one-way analysis of variance power calculation

```
k = 3
n = 322.157
f = 0.1
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

Yikes! We would need over 300 per group!

If effect sizes are new to you, play around with this effect size converter hosted at [Psychometrica.de](#). For examples like this one, use the option labeled, “Transformation of the effect sizes d , r , f , Odds Ratio, η^2 , and Common Language Effect Size (CLES).”

7.8 A Conversation with Dr. Tran

Doctoral student (and student in one of my classes) Emi Ichimura and I were able to interview the first author (Alisia Tran, PhD) about the article and what it means. Here’s a direct [link](#) to that interview.

Among others, we asked:

- What were unexpected challenges to the research method or statistical analysis?
- What were the experiences of the confederates as they offered the statements in the racial loading conditions? And in the debriefings, did the research participants share anything more anecdotally in their experiences as research participants?
- What are your current ideas about interventions or methods for mitigating the harm caused by racial microaggressions?
- How do you expect the article to change science, practice, and/or advocacy?

7.9 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “doable” and then try at least one more problem that challenges you in some way. The data for each vignette should have at least three levels in the independent variable. Further, at least one of the problems you work should have a significant omnibus test so that follow-up is required.

Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Whichever you choose, you will focus on these larger steps in one-way ANOVA, including:

- testing the statistical assumptions
- conducting a one-way ANOVA, including
 - omnibus test and effect size
 - follow-up (pairwise, planned comparisons, polynomial trends)
- writing a results section to include a figure and tables

7.9.1 Problem #1: Play around with this simulation.

If one-way ANOVA is new to you, perhaps you just change the number in “set.seed(2021)” from 2021 to something else. Your results should parallel those obtained in the lecture, making it easier for you to check your work as you go.

There are other ways to change the dataset. For example, if you are interested in power, change the sample size to something larger or smaller. If you are interested in variability (i.e., the homogeneity of variance assumption), perhaps you change the standard deviations in a way that violates the assumption.

7.9.2 Problem #2: Conduct a one-way ANOVA with the *moreTalk* dependent variable.

In their study, Tran and Lee [2014] included an outcome variable where participants rated how much longer they would continue the interaction with their partner compared to their interactions in general. The scale ranged from -2 (*much less than average*) through 0 (*average*) to 2 (*much more than average*). This variable is available in the original simulation and is an option for a slightly more challenging analysis.

7.9.3 Problem #3: Try something entirely new.

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a one-way ANOVA. Please have at least 3 levels for the predictor variable.

7.9.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the IV and DV. The data you analyze should have at least 3 levels in the independent variable; at least one of the attempted problems should have a significant omnibus test so that follow-up is required).	5	_____
2. Simulate (or import) and format data.	5	_____
3. Evaluate statistical assumptions.	5	_____
4. Conduct omnibus ANOVA (w effect size).	5	_____
5. Conduct one set of follow-up tests; narrate your choice.	5	_____
6. Describe approach for managing Type I error.	5	_____
7. APA style results with table(s) and figure.	5	_____
8. Conduct power analyses to determine the power of the current study and a recommended sample size.	5	_____
9. Explanation to grader.	5	_____
Totals	40	_____

Hand Calculations	Points Poss	Points Earned
1. Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses.	2	
2. Calculate sums of squares total (SST). Steps in this calculation must include calculating a grand mean and creating variables representing the mean deviation and mean deviation squared.	4	
3. Calculate the sums of squares for the model (SSM). A necessary step in this equation is to calculate group means.	4	
4. Calculate the sums of squares residual (SSR). A necessary step in this equation is to calculate the variance for each group.	4	
5. Calculate the mean square model, mean square residual, and <i>F</i> -test.	2	
6. What are the degrees of freedom for your numerator and denominator?	2	
7. Locate the test critical value for your one-way ANOVA.	2	
8. Is the <i>F</i> -test statistically significant? Why or why not?	2	

Hand Calculations	Points Poss	Points Earned
9. Calculate and interpret the η^2 effect size	2	

|10. Assemble the results into a statistical string. |4 || |Totals* | 28 ||

7.10 Homeworked Example

Screencast Link

If you wanted to use this example and dataset as a basis for a homework assignment, you could...

7.10.1 Working the Problem with R and R Packages

7.10.1.1 Narrate the research vignette, describing the IV and DV. The data you analyze should have at least 3 levels in the independent variable; at least one of the attempted problems should have a significant omnibus test so that follow-up is required).

I want to ask the question, do course evaluation ratings for traditional pedagogy differ for students as we enacted a substantive revision to our statistics series. The evaluative focus is on the ANOVA course and we will compare ratings from the stable, transition, and resettled stages of the transitional period. The variable (Stage) of interest will have three levels:

- STABLE: 2017 represents the last year of “stability during the old way” when we taught with SPSS and during the 2nd year of the doctoral programs.
- TRANSITION: 2018 & 2019 represent the transition to R, when the classes were 30% larger because each of the IOP and CPY departments were transitioning to the 1st year (they did it separately, so as not to double the classes)
- RESETTLED: 2020 & 2021 represent the “resettled” phase where the transition to R was fairly complete and the class size returned to normal because the classes were offered in the first year.

This is not a variable that was included in the dataset posted to the OSF repository, so we will need to create it.

7.10.1.2 Simulate (or import) and format data.

This df includes course evaluations from ANOVA, multivariate, and psychometrics. To include up to three evaluations per student would violate the assumption of independence, therefore, I will only select the students in ANOVA course.

Let's first create the “Stage” variable that represents the three levels of transition.

First I will map the years to the three levels (factors).

Then check the structure.

```
chr [1:114] "Resettled" "Resettled" "Resettled" "Resettled" "Resettled" ...
```

R is reading the variable as a character, so I need to make it to be an ordered factor.

Let's check the structure again:

```
Factor w/ 3 levels "Stable","Transition",...: 3 3 3 3 3 3 3 3 3 ...
```

The TradPed (traditional pedagogy) variable is an average of the items on that scale. I will first create that variable.

With our variables properly formatted, let's trim it to just the variables we need.

Although we would handle missing data more carefully in a “real study,” I will delete all cases with any missingness. This will prevent problems in the hand-calculations section, later (and keep the two sets of results more similar).

7.10.1.3 Evaluate statistical assumptions.

Is the dependent variable normally distributed across levels of the factor?

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max	range
TradPed1	1	Stable	1	21	4.419	0.544	4.6	4.482	0.593	3.2	5	1.8
TradPed2	2	Transition	1	44	4.045	1.029	4.3	4.206	1.038	1.0	5	4.0
TradPed3	3	Resettled	1	47	3.909	0.778	4.0	3.967	0.890	1.8	5	3.2
					skew	kurtosis	se					
TradPed1	-0.623		-0.492	0.119								
TradPed2	-1.312		1.177	0.155								
TradPed3	-0.601		-0.041	0.113								

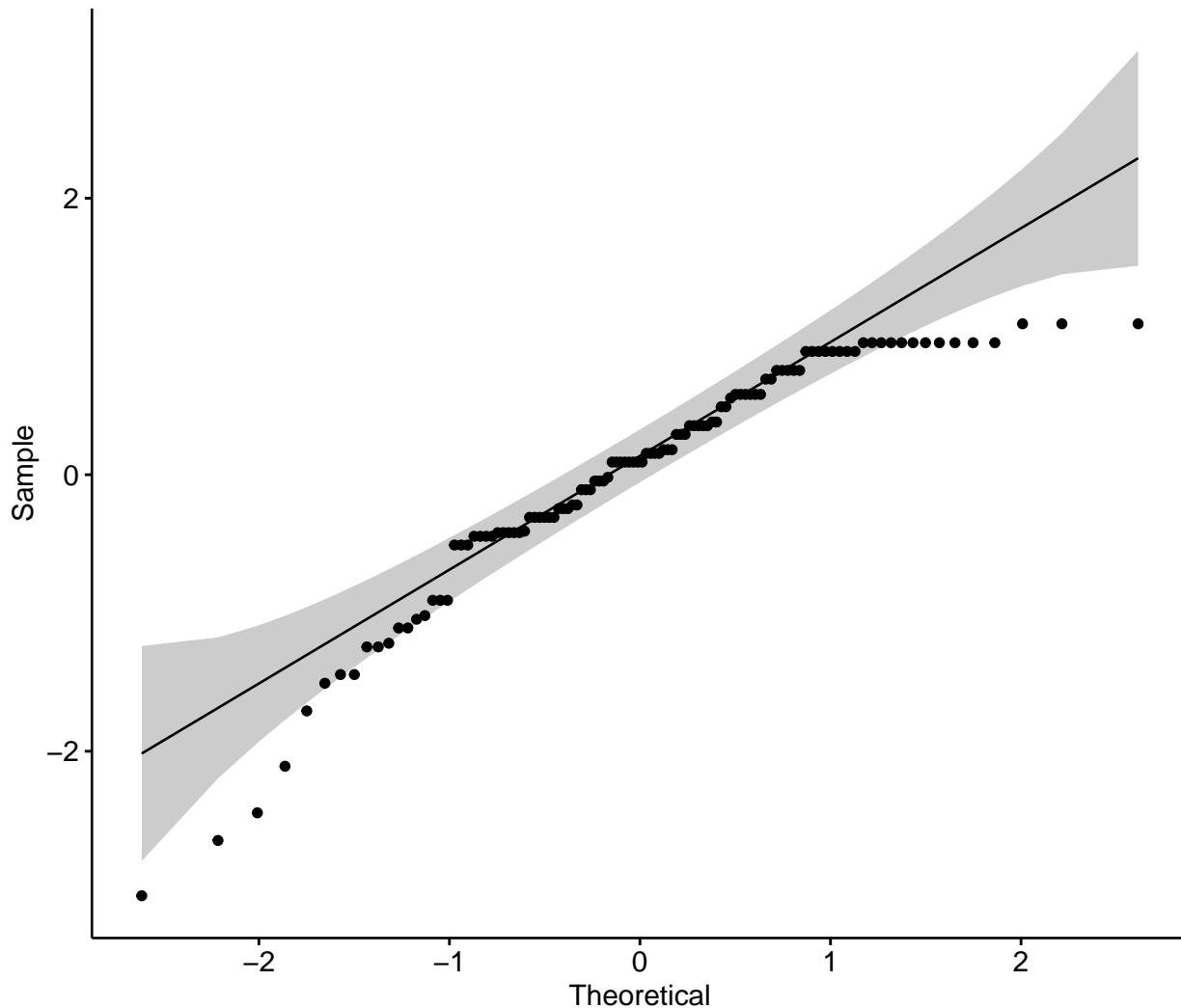
We'll use Kline's (2016) threshholds of the absolute values of 3 (skew) and 10 (kurtosis). The highest absolute value of skew is -1.31; the highest absolute value of kurtosis is 1.18. These are well below the areas of concern.

the Shapiro-wilk test is a formal assessment of normality. It is a 2-part test that begins with creating an ANOVA model from which we can extract residuals, then testing the residuals.

```
# A tibble: 1 x 3
  variable           statistic    p.value
  <chr>              <dbl>      <dbl>
1 residuals(TradPed_res) 0.910 0.00000130
```

The Shapiro-Wilk test suggests that the our distribution of residuals is statistically significantly different from a normal distribution ($W = 0.910, p < .001$).

It is possible to plot the residuals to see how and where they deviate from the line.



Oof! at the ends of the distribution they really deviate.

Should we remove outliers?

The `rstatix::identify_outliers()` function identifies outliers and extreme outliers.

	Stage	TradPed	is.outlier	is.extreme
1	Resettled	1.8	TRUE	FALSE
2	Transition	1.0	TRUE	FALSE
3	Transition	1.4	TRUE	FALSE
4	Transition	1.6	TRUE	FALSE

There are 4 cases identified with outliers; none of those is extreme. I also notice that these outliers are low course evaluations. It seems only fair to retain the data from individuals who were not satisfied with the course.

Are the variances of the dependent variable similar across the levels of the grouping factor?

We want the results of the Levene's homogeneity of variance test to be non-significant. This would support the notion that the TradPed variance is equivalent across the three stages of the transition.

```
# A tibble: 1 x 4
  df1    df2 statistic     p
  <int> <int>     <dbl> <dbl>
1     2    109     2.09 0.128
```

The non-significant p value suggests that the variances across the three stages are not statistically significantly different: $F(2, 109) = 2.094, p = 0.128$.

Before moving on, I will capture our findings in an APA style write-up of the testing of assumptions:

Regarding the assumption of normality, skew and kurtosis values at each of the levels of program year fell well below the thresholds that Kline (2016a) identified as concerning (i.e., below $|3|$ for skew and $|10|$ for kurtosis). In contrast, results of a model-based Shapiro-Wilk test of normality, indicated that the model residuals differed from a normal distribution ($W = 0.910, p < .001$). Although 4 outliers were identified none were extreme, thus we retained all cases. Finally, Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption $F(2, 109) = 2.094, p = 0.128$.

7.10.1.4 Conduct omnibus ANOVA (w effect size).

The `rstatix::anova_test()` function calculates the one-way ANOVA and includes the effect size, η^2 in the column, *ges*. Values of .01, .07, and .14 are considered to be small, medium, and large. The value of .05 would be small-to-medium.

```
Warning: NA detected in rows: 74,84.
Removing this rows before the analysis.
```

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	Stage	2	109	2.61	0.078		0.046

The one-way ANOVA is not statistically significant. This means there should not be differences between any combination of variables in the dependent variable. Before moving on, I will capture the F string: $F(2, 109) = 2.61, p = 0.078, \eta^2 = 0.046$.

Normally, the researcher would stop here. However, because the homework requires follow-up, I will continue.

7.10.1.5 Conduct one set of follow-up tests; narrate your choice.

I will simply calculate post-hoc comparisons. That is, all possible pairwise comparisons. I will specify the traditional Bonferroni as the approach to managing Type I error.

```
# A tibble: 3 x 17
  estimate estimate1 estimate2 .y.    group1 group2   n1   n2 statistic     p
* <dbl>     <dbl>     <dbl> <chr>  <chr>  <int> <int>     <dbl> <dbl>
1 0.374      4.42      4.05 TradPed Stable Trans~    21    44     1.91  0.06
2 0.511      4.42      3.91 TradPed Stable Reset~    21    47     3.11  0.003
3 0.137      4.05      3.91 TradPed Trans~ Reset~    44    47     0.713 0.478
# i 7 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>, p.adj <dbl>, p.adj.signif <chr>
```

Curiously, the post hoc tests suggested statistically significant differences between the stable and resettled stages, favoring the stable period of time (i.e., using SPSS and taught in the second year).

7.10.1.6 Describe approach for managing Type I error.

We used the Bonferroni. The Bonferroni divides the overall alpha (.05) by the number of comparisons (3). In this case, a p value would have to be lower than 0.017 to be statistically significant. The calculation reverse-engineers this so that we can interpret the * p values by the traditional 0.05. In the output, it is possible to see the higher thresholds necessary to claim statistical significance.

7.10.1.7 APA style results with table(s) and figure.

A one-way analysis of variance was conducted to evaluate the effects significant transitions (e.g., from SPSS to R; to the second to the first year in a doctoral program) on students ratings of traditional pedagogy. The independent variable, stage, included three levels: stable (with SPSS and taught in the second year of a doctoral program), transitioning (with R and students moving from second to first year), and resettled (with R and in the first year of the program).

We began by testing the statistical assumptions associated with one-way ANOVA. Regarding the assumption of normality, skew and kurtosis values at each of the levels of program year fell well below the thresholds that Kline (2016a) identified as concerning (i.e., below $|3|$ for skew and $|10|$ for kurtosis). In contrast, results of a model-based Shapiro-Wilk test of normality, indicated that the model residuals differed from a normal distribution ($W = 0.910, p < .001$). Although 4 outliers were identified none were extreme, thus we retained all cases. Finally, Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption $F(2, 109) = 2.094, p = 0.128$.

Results of the omnibus ANOVA indicated a non-significant effect of stage on students assessments of traditional pedagogy, $F(2, 109) = 2.61, p = 0.078, \eta^2 = 0.046$. The effect size was small-to-medium. We followed up the non-significant omnibus with all possible pairwise comparisons. We controlled for Type I error with the traditional Bonferroni adjustment. Curiously, results suggested that there were statistically significant differences between the transition and resettled ($M_{diff} = 0.511, p = 0.009$) stages, but not between stable and transition ($M_{diff} = 0.374, p = 0.181$) or transition and resettled

($M_{diff} = -.137, p = 1.000$). Given that the doctoral programs are unlikely to transition back to SPSS or into the second year, the instructor(s) are advised to consider ways that could result in greater student satisfaction. Means and standard deviations are presented in Table 1 and complete ANOVA results are presented in Table 2. Figure 1 provides an illustration of the results.

Table 1

Descriptive statistics for TradPed as a function of Stage.

Stage	M	M_95%_CI	SD
Stable	4.42	[4.17, 4.67]	0.54
Transition	4.05	[3.73, 4.36]	1.03
Resettled	3.91	[3.68, 4.14]	0.78

Note. M and SD represent mean and standard deviation, respectively.

LL and UL indicate the lower and upper limits of the 95% confidence interval for the mean, respectively.

The confidence interval is a plausible range of population means that could have caused a sample mean (Cumming, 2014).

Table 2

ANOVA results using TradPed as the dependent variable

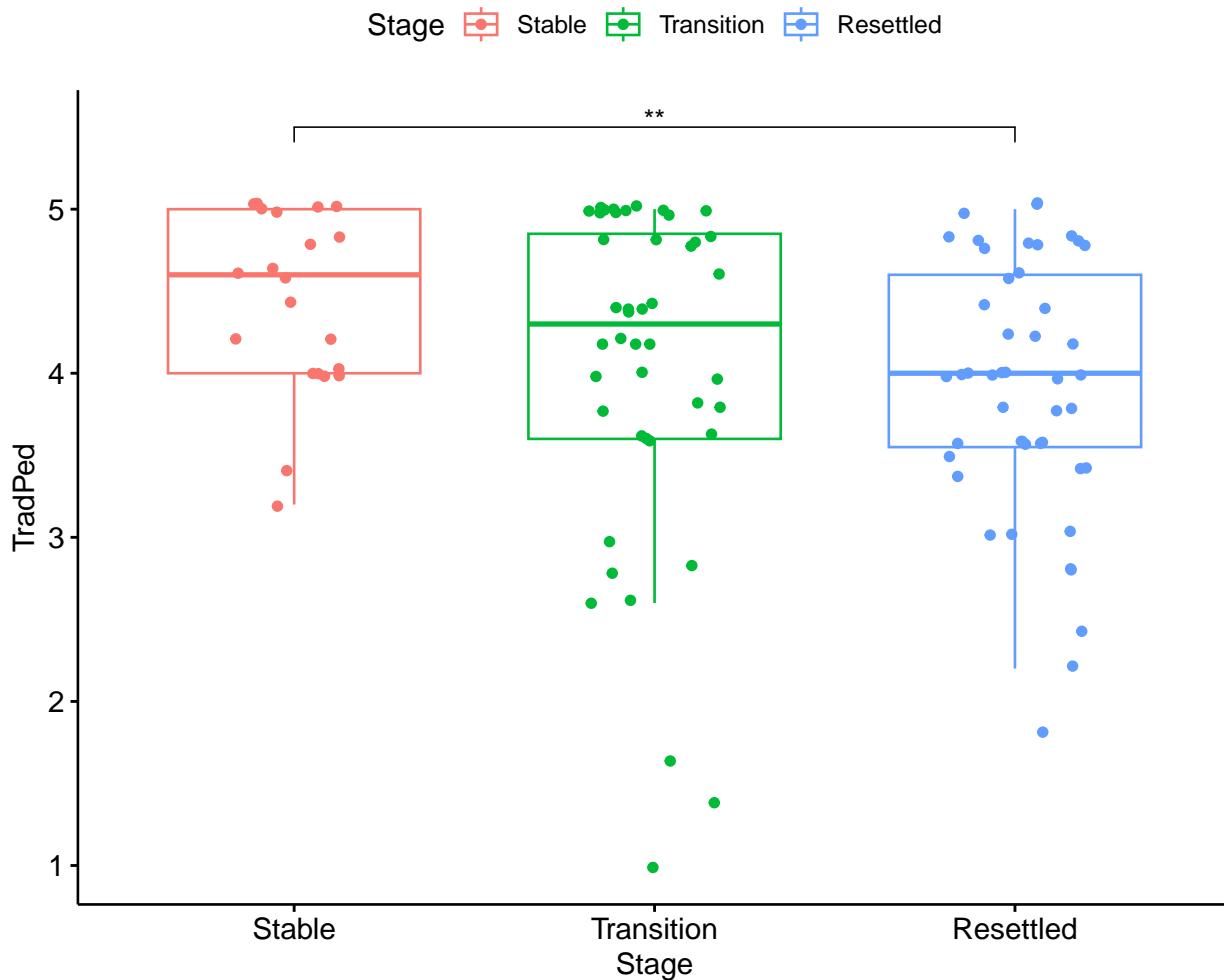
Predictor	SS	df	MS	F	p	partial_eta2	CI_90_partial_eta2
(Intercept)	410.09	1	410.09	564.12	.000		
Stage	3.79	2	1.90	2.61	.078	.05	[.00, .11]
Error	79.24	109	0.73				

Note: Values in square brackets indicate the bounds of the 90% confidence interval for partial

Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).

Warning: Removed 2 rows containing missing values (`geom_point()`).

Figure 1. Evaluations of Traditional Pedagogy as a Result of Transition



7.10.1.8 Conduct power analyses to determine the power of the current study and a recommended sample size.

The `pwr.anova.test()` has five parameters:

- $k = \#$ groups
- $n =$ sample size per group
- $f =$ effect sizes, where 0.1/small, 0.25/medium, and 0.4/large
 - In the absence from an estimate from our own data, we make a guess about the expected effect size value based on our knowledge of the literature
- $sig.level = p$ value that you will use
- $power = .80$ is the standard value

In the script below, we simply add our values. So long as we have four values, the fifth will be calculated for us.

Because this calculator requires the effect size in the metric of Cohen's f (this is not the same as the F ratio), we need to convert it. The *effectsize* package has a series of converters. We can use the *eta2_to_f()* function.

```
[1] 0.219586
```

We simply plug this value into the "f =".

First let's ask what our level of power was? Our goal would be 80%.

Given that our design was unbalanced (21, 44, 47 across the three stages), I used 38 (114/3).

```
Balanced one-way analysis of variance power calculation
```

```
k = 3
n = 38
f = 0.219586
sig.level = 0.05
power = 0.5327864
```

NOTE: n is number in each group

Our power was 0.53. That is, we had 53% chance to find a statistically significant result if one existed. In the next power analysis, let's see what sample size is recommended.

```
Balanced one-way analysis of variance power calculation
```

```
k = 3
n = 67.61369
f = 0.219586
sig.level = 0.05
power = 0.8
```

NOTE: n is number in each group

In order to be at 80% power to find a statistically significant result if there is one, we would need 68 people per group. We currently had an unbalanced design of 21, 44, and 47.

7.10.2 Hand Calculations

I will use the same example (and same dataset) for hand calculations. Because of the unbalanced design (e.g., unequal cell sizes across stages), my hand calculations will likely be different from the results from the *rstatix::anova_test()* function.

7.10.2.1 Using traditional NHST (null hypothesis testing language), state your null and alternative hypotheses.

Regarding the evaluation of traditional pedagogy across three stages of transitions to a doctoral ANOVA course, the null hypothesis predicts no differences between the three levels of the dependent variable:

$$H_O : \mu_1 = \mu_2 = \mu_3$$

In contrast, the alternative hypothesis suggests there will be differences. Apriorily, I did not make any specific predictions.

$$H_{a1} : \mu_1 \neq \mu_2 \neq \mu_3$$

7.10.2.2 Calculate sums of squares total (SST). Steps in this calculation must include calculating a grand mean and creating variables representing the mean deviation and mean deviation squared.

I will use this approach to calculating sums of squares total:

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

I will use the *psych::describe()* function to obtain the overall mean:

```
vars   n  mean    sd median trimmed  mad min max range skew kurtosis
Stage*    1 114 2.23  0.75     2.0    2.28 1.48    1    3      2 -0.39    -1.17
TradPed    2 112 4.06  0.86     4.2    4.17 0.89    1    5      4 -1.14    1.25
          se
Stage*  0.07
TradPed 0.08
```

Next, I will subtract this value from each person's TradPed value. This will create a mean deviation.

```
num [1:114] 4.4 3.8 4 3 4.8 3.5 4.6 3.8 3.6 4.6 ...
```

	Stage	TradPed	mdevTP	mdevTPb
1	Resettled	4.4	0.34	0.34196429
2	Resettled	3.8	-0.26	-0.25803571
3	Resettled	4.0	-0.06	-0.05803571
4	Resettled	3.0	-1.06	-1.05803571
5	Resettled	4.8	0.74	0.74196429
6	Resettled	3.5	-0.56	-0.55803571

	Stage	TradPed	mdevTP	mdevTPb	m_devSQTP
1	Resettled	4.4	0.34	0.34196429	0.1156
2	Resettled	3.8	-0.26	-0.25803571	0.0676
3	Resettled	4.0	-0.06	-0.05803571	0.0036
4	Resettled	3.0	-1.06	-1.05803571	1.1236
5	Resettled	4.8	0.74	0.74196429	0.5476
6	Resettled	3.5	-0.56	-0.55803571	0.3136

I will ask for a sum of the mean deviation squared column. The function was not running, sometimes this occurs when there is missing data. While I didn't think that was true, adding "na.rm = TRUE" solved the problem.

```
[1] 83.0332
```

$SST = 83.0332$

7.10.2.3 Calculate the sums of squares for the model (SSM). A necessary step in this equation is to calculate group means.

The formula for SSM is

$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$$

We will need:

- n for each group,
- Grand mean (earlier we learned it was 4.06),
- Group means

We can obtain the group means several ways. I think the *psych::describeBy()* function is one of the easiest.

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max	range
TradPed1	1	Stable	1	21	4.419	0.544	4.6	4.482	0.593	3.2	5	1.8
TradPed2	2	Transition	1	44	4.045	1.029	4.3	4.206	1.038	1.0	5	4.0
TradPed3	3	Resettled	1	47	3.909	0.778	4.0	3.967	0.890	1.8	5	3.2
			skew		kurtosis	se						
TradPed1	-0.623		-0.492		0.119							
TradPed2	-1.312		1.177		0.155							
TradPed3	-0.601		-0.041		0.113							

Now we can pop these values into the formula.

```
[1] 3.400431
```

$SSM = 3.400$

7.10.2.4 Calculate the sums of squares residual (SSR). A necessary step in this equation is to calculate the variance for each group.

The formula for I will use to calculate SSR is

$$SS_R = s_{group1}^2(n - 1) + s_{group2}^2(n - 1) + s_{group3}^2(n - 1)$$

We will need:

- n for each group,
- variance (standard deviation, squared) for each group

We can obtain these values from the previous run of the `psych::describeBy()` function.

[1] 79.29195

$SSR = 79.29$

7.10.2.5 Calculate the mean square model, mean square residual, and F -test.

The formula for mean square model is

$$MS_M = \frac{SS_M}{df_M}$$

- SS_M was 3.400
- df_M is $k - 1$ (where k is number of groups/levels)

[1] 1.7

The formula for mean square residual is

$$MS_R = \frac{SS_R}{df_R}$$

- SS_R was 79.292
- df_R is $N - k$ ($114 - 3 = 111$)

[1] 0.7143423

The formula for the F ratio is

$$F = \frac{MS_M}{MS_R}$$

[1] 2.380952

$F = 2.381$

This “isn’t exactly” what we found for the same data using R and R packages. However, the algorithms for those packages would take into consideration the unbalanced design (i.e., unequal cell sizes). Such a characteristic is a limitation, but is beyond this lesson.

7.10.2.6 What are the degrees of freedom for your numerator an denominator?

Numerator or df_M : 2 Denominator or df_R : 111

7.10.2.7 Locate the test critical value for your one-way ANOVA.

We could use use a [table of critical values](#) for the F distribution.

The closest N in the table I am using is 120. If we set alpha at 0.05, our test value would need to exceed the absolute value of 3.0718.

We can also use a look-up function, which follows this general form: `qf(p, df1, df2, lower.tail=FALSE)`

```
[1] 3.078057
```

Not surprisingly the values are quite similar.

7.10.2.8 Is the F -test statistically significant? Why or why not?

Because the value of the F test (2.381) did not exceed the absolute value of the critical value (3.078), the F test is not statistically significant.

7.10.2.9 Calculate and interpret the η^2 effect size

The formula to calculate the effect size is

$$\eta^2 = \frac{SS_M}{SS_T}$$

- SS_M was 3.400
- SS_R was 79.292

```
[1] 0.04287948
```

Eta square is 0.043. Values of .01, .06, and .14 are interpreted as small, medium, and large. Our value of 0.043 is small-to-medium.

7.10.2.10 Assemble the results into a statistical string.

$$F(2, 111) = 2.381, p > .05, \eta^2 = 0.43$$

Chapter 8

Factorial (Between-Subjects) ANOVA

[Screencasted Lecture Link](#)

In this (somewhat long and complex) lesson we conduct a 3X2 ANOVA. We will

- Work an actual example from the literature.
 - “by hand”, and
 - with R packages
- I will also demonstrate
 - several options for exploring interaction effects, and
 - several options for exploring main effects.
- Exploring these options will allow us to:
 - Gain familiarity with the concepts central to multi-factor ANOVAs.
 - Explore tools for analyzing the complexity in designs.

The complexity is that not all of these things need to be conducted for every analysis. The two-way ANOVA Workflow is provided to help you map a way through your own analyses. I will periodically refer to this map so that we can more easily keep track of where we are in the process.

8.1 Navigating this Lesson

There is about 1 hour and 30 minutes hours of lecture. If you work through the materials with me plan for another two hours of study.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

8.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Define, locate, and interpret all the effects associated with two-way ANOVA:
 - main
 - interaction (introducing the concept, *moderator*)
 - simple main effects
- Identify which means belong with which effects. Then compare and interpret them.
 - marginal means
 - individual cell means
 - comparing them
- Map a process/workflow for investigating a factorial ANOVA
- Manage Type I error
- Conduct a power analysis to determine sample size

8.1.2 Planning for Practice

In each of these lessons I provide suggestions for practice that allow you to select from options that vary in degree of difficulty. The least complex is to change the random seed and rework the problem demonstrated in the lesson. The results *should* map onto the ones obtained in the lecture.

The second option comes from the research vignette. The Ramdhani et al. [2018] article has two dependent variables (DVs; negative and positive evaluation) which are suitable for two-way ANOVAs. I will demonstrate a simulation of one of their 3X2 ANOVAs (negative) in this lecturette. The second dependent variable (positive) is suggested for the moderate level of difficulty.

As a third option, you are welcome to use data to which you have access and is suitable for two-way ANOVA. In either case the practice options suggest that you:

- test the statistical assumptions
- conduct a two-way ANOVA, including
 - omnibus test and effect size
 - report main and interaction effects
 - conduct follow-up testing of simple main effects
- write a results section to include a figure and tables

8.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s) that are freely available on the internet. Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Kassambara, A. (n.d.). ANOVA in R: The Ultimate Guide. Datanovia. Retrieved December 28, 2022, from <https://www.datanovia.com/en/lessons/anova-in-r/>

- In order to streamline the learning process, I have chosen to use *rstatix* package for the majority of ANOVA lessons. There are a number of tutorials about this package as well as its integration with *ggpubr* for creating relatively easy creation of attractive and informative figures. This tutorial is especially helpful.
- Navarro, D. (2020). Chapter 16: Factorial ANOVA. In [Learning Statistics with R - A tutorial for Psychology Students and other Beginners](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)). Retrieved from [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro))
 - Navarro's OER includes a good mix of conceptual information about one-way ANOVA as well as R code. My code/approach is a mix of Green and Salkind's [2017c], Field's [2012], Navarro's [2020b], and other techniques I have found on the internet and learned from my students.
- Ramdhani, N., Thontowi, H. B., & Ancok, D. (2018). Affective Reactions Among Students Belonging to Ethnic Groups Engaged in Prior Conflict. *Journal of Pacific Rim Psychology*, 12, e2. <https://doi.org/10.1017/prp.2017.22>
 - The source of our research vignette.

8.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(knitr)){install.packages('knitr')}
# if(!require(psych)){install.packages('psych')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(dplyr)){install.packages('dplyr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(rstatix)){install.packages('rstatix')}
# if(!require(effectsize)){install.packages('effectsize')}
# if(!require(pwr2)){install.packages('pwr2')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(emmeans)){install.packages('emmeans')}#although we
# don't call this package directly, there are rstatix functions that
# are a wrapper for it and therefore it needs to be installed
# if(!require(car)){install.packages('car')}#although we don't call
# this package directly, there are rstatix functions that are a
# wrapper for it and therefore it needs to be installed
```

8.2 Introducing Factorial ANOVA

My approach to teaching is to address the conceptual as we work problems. That said, there are some critical ideas we should address first.

ANOVA is for experiments (or arguably closely related designs). As we learn about the assumptions you'll see that ANOVA has some rather restrictive ones (e.g., there should be an equal/equivalent number of cases per cell). To the degree that we violate these assumptions, we should locate alternative statistical approaches where these assumptions are relaxed.

Factorial: a term used when there are two or more independent variables (IVs; the factors). The factors could be between-groups, within-groups, repeated measures, or a combination of between and within.

- **Independent factorial design:** several IVs (predictors/factors) and each has been measured using different participants (between groups).
- **Related factorial design:** several IVs (factors/predictors) have been measured, but the same participants have been used in all conditions (repeated measures or within-subjects).
- **Mixed design:** several IVs (factors/predictors) have been measured. One or more factors uses different participants (between-subjects) and one or more factors uses the same participants (within-subjects). Thus, there is a combination of independent (between) and related (within or repeated) designs.

“Naming” the ANOVA model follows a number/levels convention. The example in this lesson is a 3X2 ANOVA. We know there are two factors that have three and two levels, respectively:

- *rater ethnicity* has three levels representing the two ethnic groups that were in prior conflict (Marudese, Dayaknese) and a third group who was uninvolved in the conflict (Javanese);
- *photo stimulus* has two levels representing members of the two ethnic groups that were in prior conflict (Madurese, Dayaknese);

Moderator is what creates an interaction. Below are traditional representations of the *statistical* and *conceptual* figures of interaction effects. We will say that Factor B, *moderates* the relationship between Factor A (the IV) and the DV.

In a later lesson we work an ANCOVA – where we will distinguish between a *moderator* and a *covariate*. In lessons on regression models, you will likely be introduced to the notion of *mediator*.

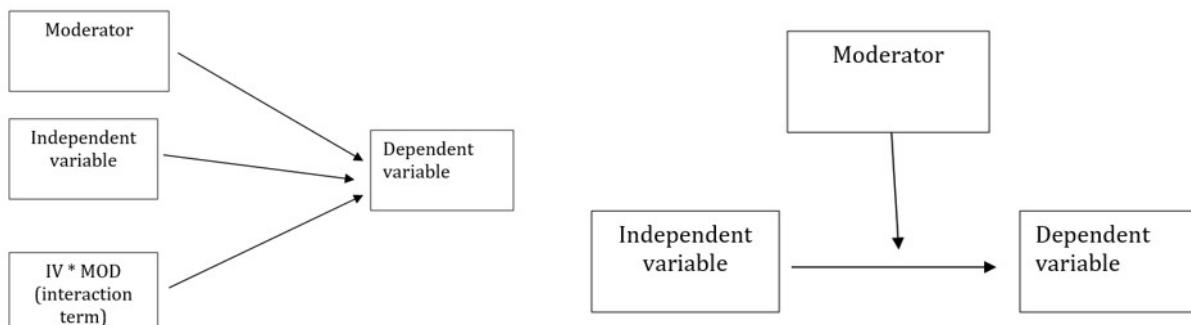


Figure 8.1: Graphic representations of a moderated relationship?

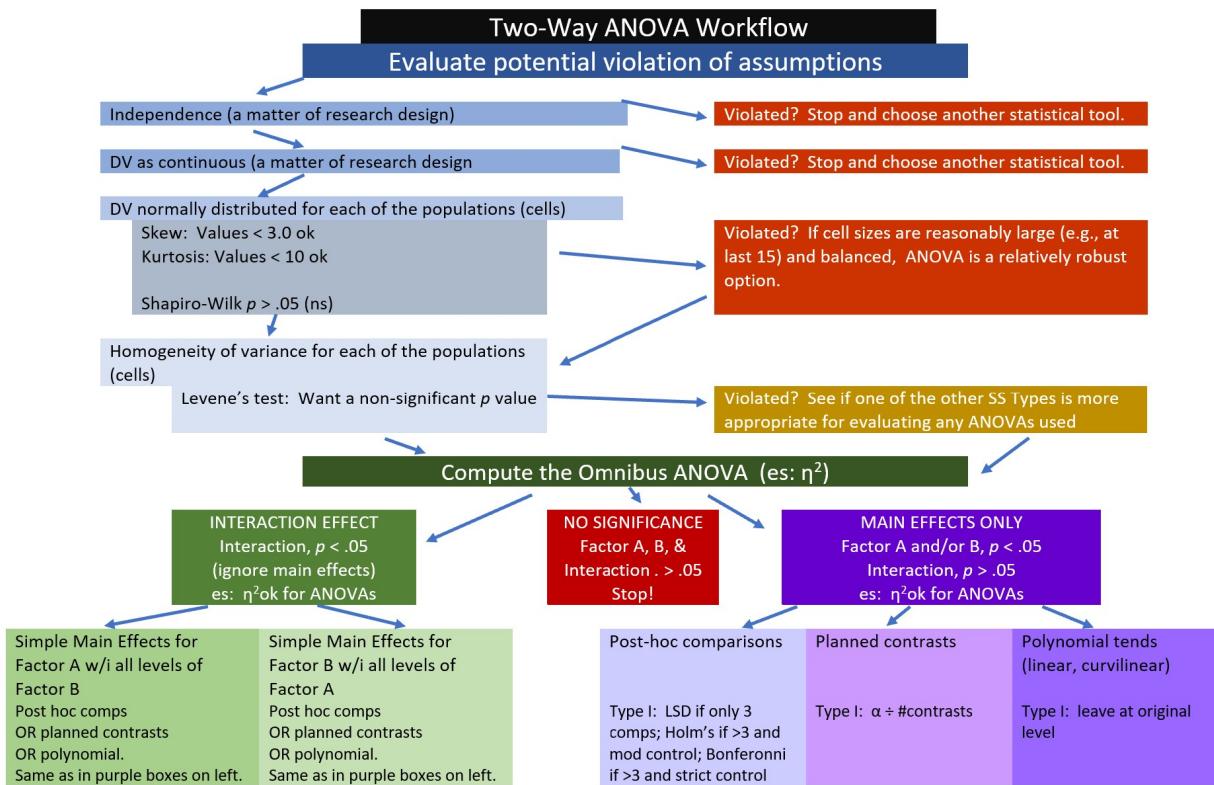


Figure 8.2: An image of a workflow for the two-way ANOVA

8.2.1 Workflow for Two-Way ANOVA

The following is a proposed workflow for conducting a two-way ANOVA.

Steps of the workflow include:

1. Enter data
 - predictors should be formatted as factors (ordered or unordered); the dependent variable should be continuously scaled
 - understanding the format of data can often provide clues as to which ANOVA/statistic to use
2. Explore data
 - graph the data
 - compute descriptive statistics
 - evaluate distributional assumptions
 - assess the homogeneity of variance assumption with Levene's test
 - assess the normality assumption with the Shapiro Wilk test
 - determine if there are outliers; if appropriate, delete
3. Compute the omnibus ANOVA
 - *depending on what you found in the data exploration phase, you may need to run a robust version of the test*
4. Follow-up testing based on significant main or interaction effects
 - significant interactions require tests of simple main effects which could be further explored with contrasts, posthoc comparisons, and/or polynomials
 - *the exact methods you choose will depend upon the tests of assumptions during data exploration*
5. Managing Type I error

8.3 Research Vignette

The research vignette for this example was located in Kalimantan, Indonesia and focused on bias in young people from three ethnic groups. The Madurese and Dayaknese groups were engaged in ethnic conflict that spanned 1996 to 2001. The last incidence of mass violence was in 2001 where approximately 500 people (mostly from the Madurese ethnic group) were expelled from the province. Ramdhani et al.'s [2018] research hypotheses were based on the roles of the three ethnic groups in the study. According to the author, the Madurese were viewed as the transgressors when they occupied lands and took employment and business opportunities from the Dayaknese. Ramdhani et al. also included a third group who were not involved in the conflict (Javanese). The research participants were students studying in Yogyakarta who were not involved in the conflict. They included 39 Madurese, 35 Dyaknese, and 37 Javanese; 83 were male and 28 were female.

In the study [Ramdhani et al., 2018], participants viewed facial pictures of three men and three women (in traditional dress) from each ethnic group (6 photos per ethnic group). Participants were

asked, “How do you feel when you see this photo? Please indicate your answers based on your actual feelings.” Participants responded on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicated ratings of higher intensity on that scale. The two scales included the following words:

- Positive: friendly, kind, helpful, happy
- Negative: disgusting, suspicious, hateful, angry

8.3.1 Data Simulation

Below is script to simulate data for the negative reactions variable from the information available from the manuscript [Ramdhani et al., 2018].

```
library(tidyverse)
set.seed(210731)
# sample size, M and SD for each cell; this will put it in a long
# file
Negative <- round(c(rnorm(17, mean = 1.91, sd = 0.73), rnorm(18, mean = 3.16,
  sd = 0.19), rnorm(19, mean = 3.3, sd = 1.05), rnorm(20, mean = 3, sd = 1.07),
  rnorm(18, mean = 2.64, sd = 0.95), rnorm(19, mean = 2.99, sd = 0.8)), 3)
# sample size, M and SD for each cell; this will put it in a long
# file
Positive <- round(c(rnorm(17, mean = 4.99, sd = 1.38), rnorm(18, mean = 3.83,
  sd = 1.13), rnorm(19, mean = 4.2, sd = 0.82), rnorm(20, mean = 4.19,
  sd = 0.91), rnorm(18, mean = 4.17, sd = 0.6), rnorm(19, mean = 3.26,
  sd = 0.94)), 3)
ID <- factor(seq(1, 111))
Rater <- c(rep("Dayaknese", 35), rep("Madurese", 39), rep("Javanese", 37))
Photo <- c(rep("Dayaknese", 17), rep("Madurese", 18), rep("Dayaknese",
  19), rep("Madurese", 20), rep("Dayaknese", 18), rep("Madurese", 19))
# groups the 3 variables into a single df: ID#, DV, condition
Ramdhani_df <- data.frame(ID, Negative, Positive, Rater, Photo)
```

For two-way ANOVA our variables need to be properly formatted. In our case:

- Negative is a continuously scaled DV and should be *num*
- Positive is a continuously scaled DV and should be *num*
- Rater should be an unordered factor
- Photo should be an unordered factor

```
str(Ramdhani_df)
```

```
'data.frame': 111 obs. of 5 variables:
 $ ID      : Factor w/ 111 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
```

```
$ Negative: num  2.768 1.811 0.869 1.857 2.087 ...
$ Positive: num  5.91 5.23 3.54 5.63 5.44 ...
$ Rater    : chr  "Dayaknese" "Dayaknese" "Dayaknese" "Dayaknese" ...
$ Photo    : chr  "Dayaknese" "Dayaknese" "Dayaknese" "Dayaknese" ...
```

Our Negative variable is correctly formatted. Let's reformat Rater and Photo to be factors and re-evaluate the structure. R's default is to order the factors alphabetically. In this case this is fine. If we had ordered factors such as dosage (placebo, lo, hi) we would want to respecify the order.

```
Ramdhani_df[, "Rater"] <- as.factor(Ramdhani_df[, "Rater"])
Ramdhani_df[, "Photo"] <- as.factor(Ramdhani_df[, "Photo"])
str(Ramdhani_df)
```

```
'data.frame': 111 obs. of 5 variables:
 $ ID      : Factor w/ 111 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ Negative: num  2.768 1.811 0.869 1.857 2.087 ...
 $ Positive: num  5.91 5.23 3.54 5.63 5.44 ...
 $ Rater   : Factor w/ 3 levels "Dayaknese","Javanese",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ Photo   : Factor w/ 2 levels "Dayaknese","Madurese": 1 1 1 1 1 1 1 1 1 1 ...
```

If you want to export this data as a file to your computer, remove the hashtags to save it (and re-import it) as a .csv ("Excel lite") or .rds (R object) file. This is not a necessary step.

The code for .csv will likely lose the formatting (i.e., making the Rater and Photo variables factors), but it is easy to view in Excel.

```
# write the simulated data as a .csv write.table(Ramdhani_df,
# file='RamdhaniCSV.csv', sep=',', col.names=TRUE, row.names=FALSE)
# bring back the simulated dat from a .csv file Ramdhani_df <-
# read.csv ('RamdhaniCSV.csv', header = TRUE) str(Ramdhani_df)
```

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Ramdhani_df, 'Ramdhani_RDS.rds') bring back the
# simulated dat from an .rds file Ramdhani_df <-
# readRDS('Ramdhani_RDS.rds') str(Ramdhani_RDS)
```

8.3.2 Quick peek at the data

Let's first examine the descriptive statistics (e.g., means of the variable, Negative) by group. We can use the *describeBy()* function from the *psych* package.

```
negative.descripts <- psych::describeBy(Negative ~ Rater + Photo, mat = TRUE,
  data = Ramdhani_df, digits = 3) #digits allows us to round the output
negative.descripts
```

	item	group1	group2	vars	n	mean	sd	median	trimmed	mad
Negative1	1	Dayaknese	Dayaknese		1 17	1.818 0.768	1.692	1.783	0.694	
Negative2	2	Javanese	Dayaknese		1 18	2.524 0.742	2.391	2.460	0.569	
Negative3	3	Madurese	Dayaknese		1 19	3.301 1.030	3.314	3.321	1.294	
Negative4	4	Dayaknese	Madurese		1 18	3.129 0.156	3.160	3.136	0.104	
Negative5	5	Javanese	Madurese		1 19	3.465 0.637	3.430	3.456	0.767	
Negative6	6	Madurese	Madurese		1 20	3.297 1.332	2.958	3.254	1.615	
		min	max	range	skew	kurtosis	se			
Negative1	0.706	3.453	2.747	0.513	-0.881	0.186				
Negative2	1.406	4.664	3.258	1.205	1.475	0.175				
Negative3	1.406	4.854	3.448	-0.126	-1.267	0.236				
Negative4	2.732	3.423	0.691	-0.623	0.481	0.037				
Negative5	2.456	4.631	2.175	-0.010	-1.307	0.146				
Negative6	1.211	5.641	4.430	0.215	-1.238	0.298				

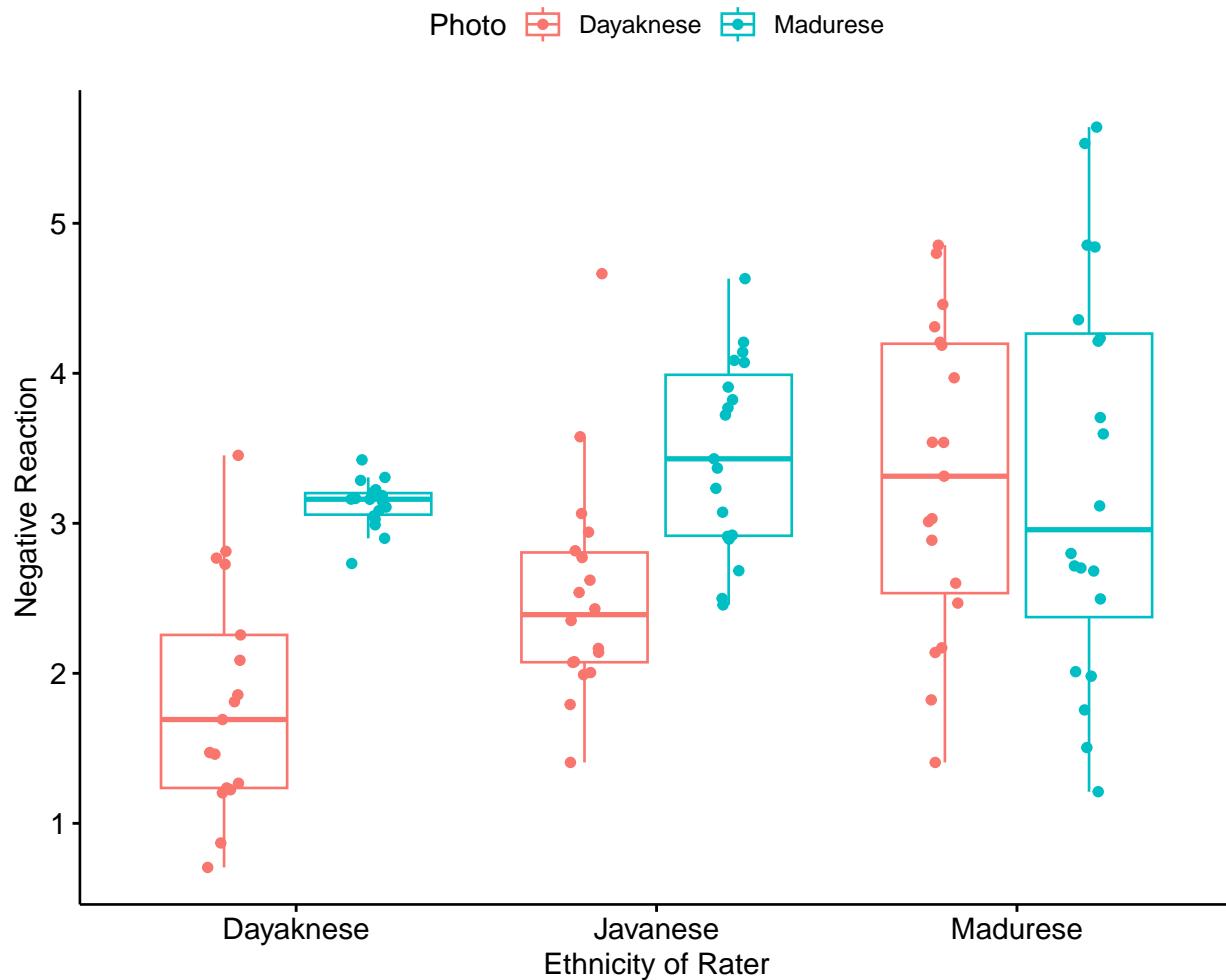
The `write.table()` function can be a helpful way to export output to .csv files so that you can manipulate it into tables.

```
write.table(negative.descripts, file = "NegativeDescriptions.csv", sep = ",",
  col.names = TRUE, row.names = FALSE)
```

At this stage, it would be useful to plot our data. Figures can assist in the conceptualization of the analysis.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction", add = "jitter",
  title = "Boxplots Clustered by Rater Ethnicity")
```

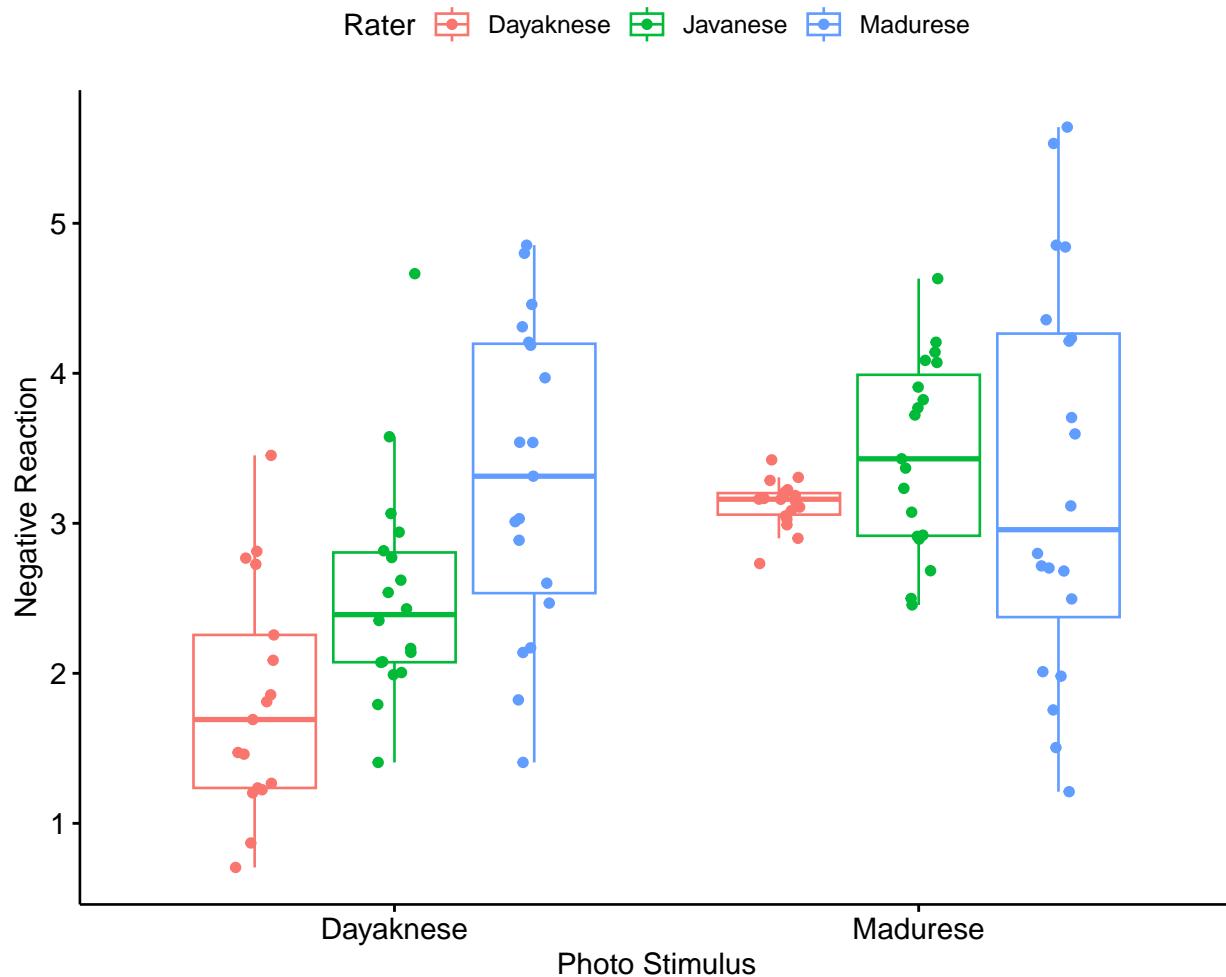
Boxplots Clustered by Rater Ethnicity



Narrating results is sometimes made easier if variables are switched. There is usually not a right or wrong answer. Here is another view, switching the Rater and Photo predictors.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Photo Stimulus", ylab = "Negative Reaction", add = "jitter",
  title = "Boxplots Clustered by Ethnicity Represented in Photo Stimulus")
```

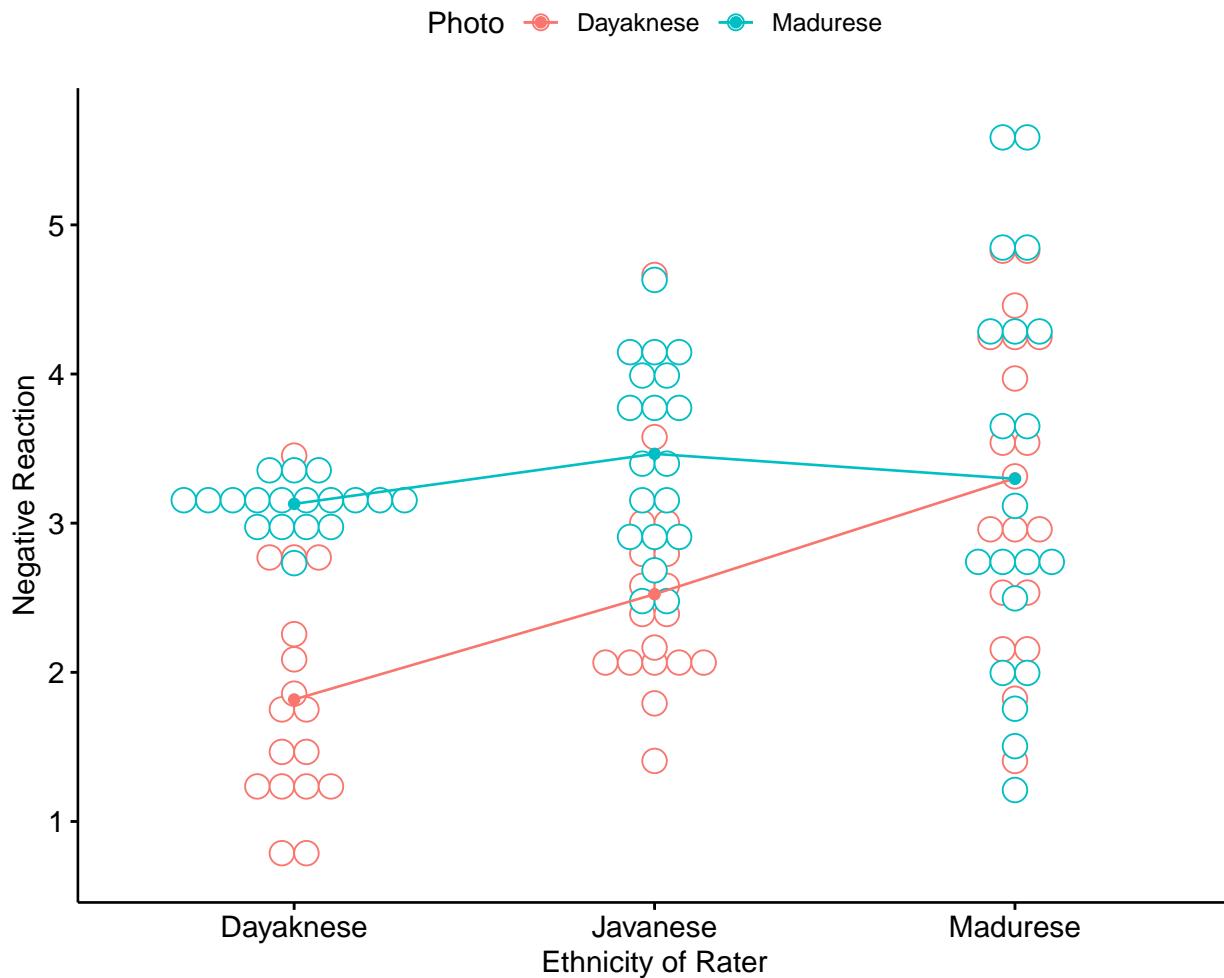
Boxplots Clustered by Ethnicity Represented in Photo Stimulus



Yet another option plots the raw data as bubbles, the means as lines, and denotes differences in the moderator with color.

```
ggpubr::gglime(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction", add = c("mean_se",
  "dotplot"), title = "Lineplot Clustered by Rater Ethnicity")
```

Lineplot Clustered by Rater Ethnicity



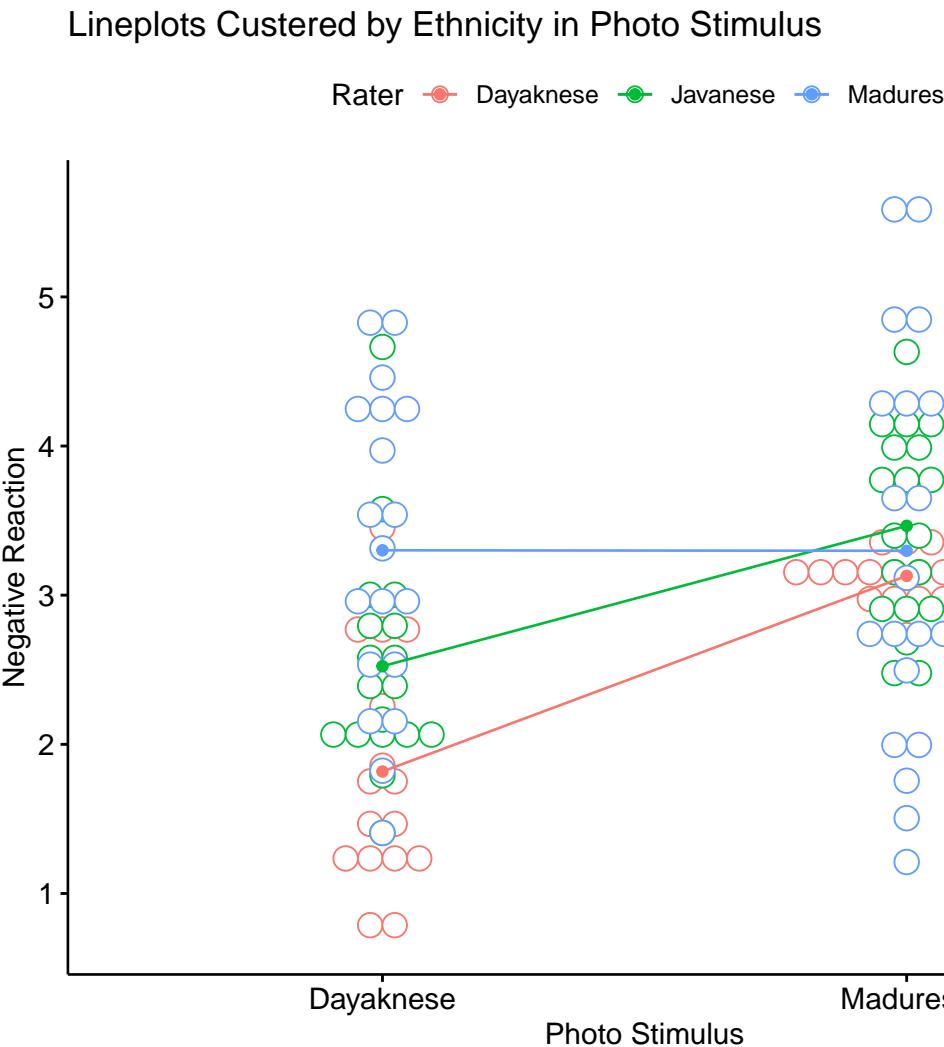
```
# add this for a different color palette: palette = c('#00AFBB',
# '#E7B800')
```

We can reverse this to see if it assists with our conceptualization.

```
ggpubr::gglue(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Photo Stimulus", ylab = "Negative Reaction", add = c("mean_se",
  "dotplot"), title = "Lineplots Custered by Ethnicity in Photo Stimulus")
```

Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.

```
Warning: Computation failed in `stat_summary()`
Caused by error in `get()`:
! object 'mean_se_' of mode 'function' was not found
```



8.4 Working the Factorial ANOVA (by hand)

Before we work an ANOVA let's take a moment to consider what we are doing and how it informs our decision-making. This figure (which already contains "the answers") may help conceptualize how variance is partitioned.

As in one-way ANOVA, we partition variance into **total**, **model**, and **residual**. However, we now further divide the SS_M into its respective factors A(column), B(row,) and their a x b product.

In this, we begin to talk about main effects and interactions.

8.4.1 Sums of Squares Total

Our formula is the same as it was for one-way ANOVA:

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

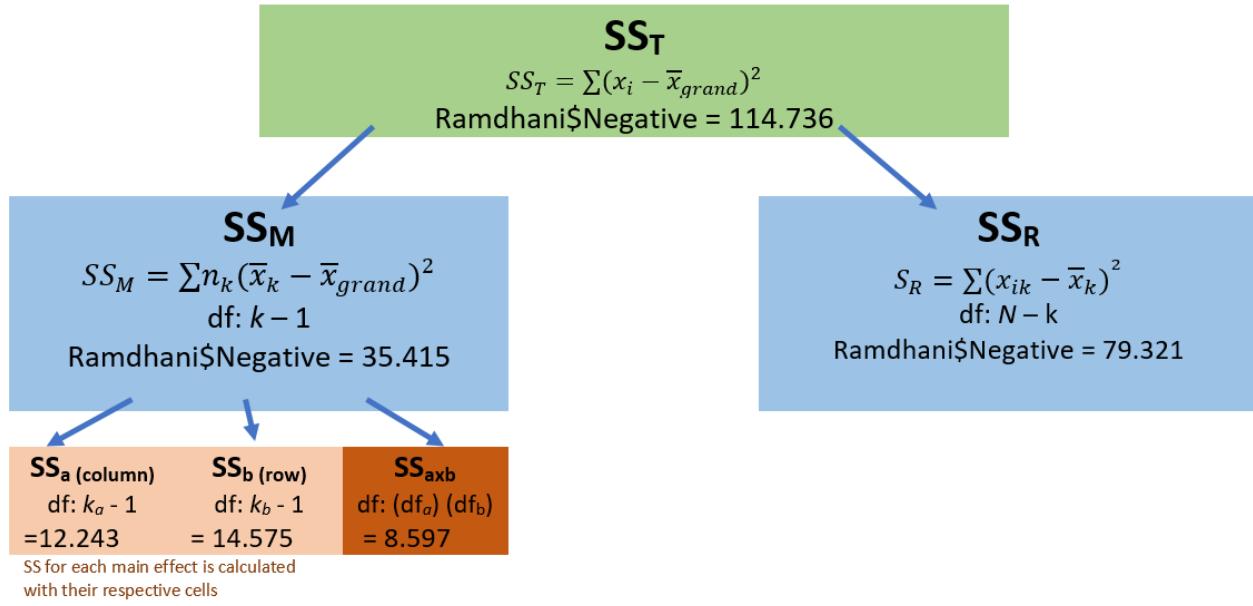


Figure 8.3: Image of a flowchart that partitions variance from sums of squares totals to its component pieces

Let's calculate it for the Ramdhani et al. [2018] data. Our grand (i.e., overall) mean is

```
mean(Ramdhani_df$Negative)
```

```
[1] 2.947369
```

Subtracting the grand mean from each Negative rating yields a mean difference.

```
library(tidyverse)
Ramdhani_df <- Ramdhani_df %>%
  mutate(m_dev = Negative - mean(Negative))
head(Ramdhani_df)
```

ID	Negative	Positive	Rater	Photo	m_dev
1	2.768	5.907	Dayaknese	Dayaknese	-0.1793694
2	1.811	5.234	Dayaknese	Dayaknese	-1.1363694
3	0.869	3.544	Dayaknese	Dayaknese	-2.0783694
4	1.857	5.628	Dayaknese	Dayaknese	-1.0903694
5	2.087	5.438	Dayaknese	Dayaknese	-0.8603694
6	0.706	5.833	Dayaknese	Dayaknese	-2.2413694

Pop quiz: What's the sum of our new *m_dev* variable?

Let's find out!

```
sum(Ramdhani_df$m_dev)
```

```
[1] -0.000000000000007549517
```

Of course! The sum of squared deviations around the mean is zero. Next we square those mean deviations.

```
Ramdhani_df <- Ramdhani_df %>%
  mutate(m_devSQ = m_dev^2)
head(Ramdhani_df)
```

	ID	Negative	Positive	Rater	Photo	m_dev	m_devSQ
1	1	2.768	5.907	Dayaknese	Dayaknese	-0.1793694	0.03217337
2	2	1.811	5.234	Dayaknese	Dayaknese	-1.1363694	1.29133534
3	3	0.869	3.544	Dayaknese	Dayaknese	-2.0783694	4.31961924
4	4	1.857	5.628	Dayaknese	Dayaknese	-1.0903694	1.18890536
5	5	2.087	5.438	Dayaknese	Dayaknese	-0.8603694	0.74023545
6	6	0.706	5.833	Dayaknese	Dayaknese	-2.2413694	5.02373665

Then we sum the squared mean deviations.

```
sum(Ramdhani_df$m_devSQ)
```

```
[1] 114.7746
```

This value, 114.775, the sum of squared deviations around the grand mean, is our SS_T ; the associated *degrees of freedom* is $N - 1$.

In factorial ANOVA, we divide SS_T into **model/between** sums of squares and **residual/within** sums of squares.

8.4.2 Sums of Squares for the Model

$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$$

The *model* generally represents the notion that the means are different than each other. We want the variation between our means to be greater than the variation within each of the groups from which our means are calculated.

In factorial ANOVA, we need means for each of the combinations of the factors. We have a 3 x 2 model:

- Rater with three levels: Dayaknese, Madurese, Javanese
- Photo with two levels: Dayaknese, Madurese

Let's repeat some code we used before to obtain the cell-level means and cell sizes.

```
psych::describeBy(Negative ~ Rater + Photo, mat = TRUE, data = Ramdhani_df,
  digits = 3)
```

	item	group1	group2	vars	n	mean	sd	median	trimmed	mad
Negative1	1	Dayaknese	Dayaknese		1 17	1.818 0.768	1.692	1.783	0.694	
Negative2	2	Javanese	Dayaknese		1 18	2.524 0.742	2.391	2.460	0.569	
Negative3	3	Madurese	Dayaknese		1 19	3.301 1.030	3.314	3.321	1.294	
Negative4	4	Dayaknese	Madurese		1 18	3.129 0.156	3.160	3.136	0.104	
Negative5	5	Javanese	Madurese		1 19	3.465 0.637	3.430	3.456	0.767	
Negative6	6	Madurese	Madurese		1 20	3.297 1.332	2.958	3.254	1.615	
					min	max	range	skew	kurtosis	se
Negative1		0.706	3.453	2.747	0.513		-0.881	0.186		
Negative2		1.406	4.664	3.258	1.205		1.475	0.175		
Negative3		1.406	4.854	3.448	-0.126		-1.267	0.236		
Negative4		2.732	3.423	0.691	-0.623		0.481	0.037		
Negative5		2.456	4.631	2.175	-0.010		-1.307	0.146		
Negative6		1.211	5.641	4.430	0.215		-1.238	0.298		

```
# Note. Recently my students and I have been having intermittent
# struggles with the describeBy function in the psych package. We
# have noticed that it is problematic when using .rds files and when
# using data directly imported from Qualtrics. If you are having
# similar difficulties, try uploading the .csv file and making the
# appropriate formatting changes.
```

We also need the grand mean (i.e., the mean that disregards [or “collapses across”] the factors).

```
mean(Ramdhani_df$Negative)
```

```
[1] 2.947369
```

This formula occurs in six chunks, representing the six cells of our designed. In each of the chunks we have the n , group mean, and grand mean.

```
17 * (1.818 - 2.947)^2 + 18 * (2.524 - 2.947)^2 + 19 * (3.301 - 2.947)^2 +
  18 * (3.129 - 2.947)^2 + 19 * (3.465 - 2.947)^2 + 20 * (3.297 - 2.947)^2
```

```
[1] 35.41501
```

This value, 35.415, SS_M is the value accounted for by the model. That is, the amount of variance accounted for by the grouping variable/factors, Rater and Photo.

8.4.3 Sums of Squares Residual (or within)

SS_R is error associated with within group variability. If people are randomly assigned to conditions there should be no other confounding variable. Thus, all SS_R variability is *uninteresting* for the research and treated as noise.

$$SS_R = \sum (x_{ik} - \bar{x}_k)^2$$

Here's another configuration of the same:

$$SS_R = s_{group1}^2(n-1) + s_{group2}^2(n-1) + s_{group3}^2(n-1) + s_{group4}^2(n-1) + s_{group5}^2(n-1) + s_{group6}^2(n-1)$$

Again, the formula is in six chunks – but this time the calculations are *within-group*. We need the variance (the standard deviation squared) for the calculation. We can retrieve these from the descriptive statistics.

```
psych::describeBy(Negative ~ Rater + Photo, mat = TRUE, data = Ramdhani_df,
  digits = 3)
```

	item	group1	group2	vars	n	mean	sd	median	trimmed	mad
Negative1	1	Dayaknese	Dayaknese		17	1.818	0.768	1.692	1.783	0.694
Negative2	2	Javanese	Dayaknese		18	2.524	0.742	2.391	2.460	0.569
Negative3	3	Madurese	Dayaknese		19	3.301	1.030	3.314	3.321	1.294
Negative4	4	Dayaknese	Madurese		18	3.129	0.156	3.160	3.136	0.104
Negative5	5	Javanese	Madurese		19	3.465	0.637	3.430	3.456	0.767
Negative6	6	Madurese	Madurese		20	3.297	1.332	2.958	3.254	1.615
		min	max	range	skew	kurtosis	se			
Negative1		0.706	3.453	2.747	0.513	-0.881	0.186			
Negative2		1.406	4.664	3.258	1.205	1.475	0.175			
Negative3		1.406	4.854	3.448	-0.126	-1.267	0.236			
Negative4		2.732	3.423	0.691	-0.623	0.481	0.037			
Negative5		2.456	4.631	2.175	-0.010	-1.307	0.146			
Negative6		1.211	5.641	4.430	0.215	-1.238	0.298			

Calculating SS_R

```
((0.768^2) * (17 - 1)) + ((0.742^2) * (18 - 1)) + ((1.03^2) * (19 - 1)) +
  ((0.156^2) * (18 - 1)) + ((0.637^2) * (19 - 1)) + ((1.332^2) * (20 -
  1))
```

[1] 79.32078

The value for our SS_R is 79.321. Its degrees of freedom is $N - k$. That is, the total N minus the number of groups:

```
111 - 6
```

```
[1] 105
```

8.4.4 A Recap on the Relationship between SS_T , SS_M , and SS_R

$SS_T = SS_M + SS_R$ In our case:

- SS_T was 114.775
- SS_M was 35.415
- SS_R was 79.321

Considering rounding error, we were successful!

```
35.415 + 79.321
```

```
[1] 114.736
```

8.4.5 Calculating SS for Each Factor and Their Products

8.4.5.1 Rater Main Effect

$SS_a : Rater$ is calculated the same way as SS_M for one-way ANOVA. Simply collapse across Photo and calculate the *marginal means* for Negative as a function of the Rater's ethnicity.

Reminder of the formula: $SS_{a:Rater} = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$

There are three cells involved in the calculation of $SS_a : Rater$.

```
psych::describeBy(Negative ~ Rater, mat = TRUE, data = Ramdhani_df, digits = 3)
```

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max	
Negative1	1	Dayaknese		1	35	2.492	0.856	2.900	2.561	0.480	0.706	3.453
Negative2	2	Javanese		1	37	3.007	0.831	2.913	2.986	0.984	1.406	4.664
Negative3	3	Madurese		1	39	3.299	1.179	3.116	3.288	1.588	1.211	5.641
	range	skew	kurtosis	se								
Negative1	2.747	-0.682	-1.132	0.145								
Negative2	3.258	0.239	-0.923	0.137								
Negative3	4.430	0.117	-1.036	0.189								

Again, we need the grand mean.

```
mean(Ramdhani_df$Negative)
```

```
[1] 2.947369
```

Now to calculate the Rater main effect.

```
35 * (2.491 - 2.947)^2 + 37 * (3.007 - 2.947)^2 + 39 * (3.299 - 2.947)^2
```

```
[1] 12.24322
```

8.4.5.2 Photo Main Effect

SS_b : *Photo* is calculated the same way as SS_M for one-way ANOVA. Simply collapse across Rater and calculate the *marginal means* for Negative as a function of the ethnicity reflected in the Photo stimulus:

Reminder of the formula: $SS_{a:Photo} = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$.

With Photo, we have only two cells.

```
psych::describeBy(Negative ~ Photo, mat = TRUE, data = Ramdhani_df, digits = 3)
```

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max
Negative1	1	Dayaknese	1	54	2.575	1.043	2.449	2.516	0.921	0.706	4.854
Negative2	2	Madurese	1	57	3.300	0.871	3.166	3.280	0.667	1.211	5.641
		range	skew	kurtosis	se						
Negative1	4.148	0.47	-0.555	0.142							
Negative2	4.430	0.35	0.581	0.115							

Again, we need the grand mean.

```
mean(Ramdhani_df$Negative)
```

```
[1] 2.947369
```

```
54 * (2.575 - 2.947)^2 + 57 * (3.3 - 2.947)^2
```

```
[1] 14.57545
```

8.4.5.3 Interaction effect

The interaction term is simply the SS_M remaining after subtracting the SS from the main effects.

$$SS_{axb} = SS_M - (SS_a + SS_b)$$

```
35.415 - (12.243 + 14.575)
```

```
[1] 8.597
```

Let's revisit the figure I showed at the beginning of this section to see, again, how variance is partitioned.

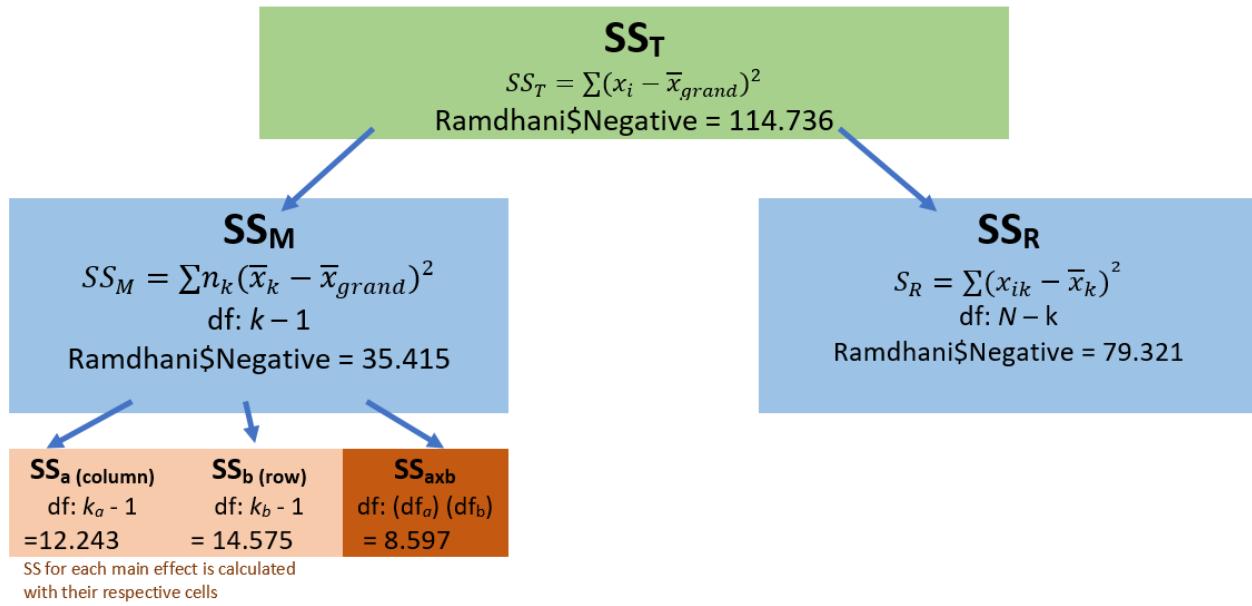


Figure 8.4: Image of a flowchart that partitions variance from sums of squares totals to its component pieces

8.4.6 Source Table Games!

As in the lesson for one-way ANOVA, we can use the information in this source table to determine if we have statistically significance in the model. There is enough information in the source table to be able to calculate all the elements. The formulas in the table provide some hints. Before scrolling onto the answers, try to complete it yourself.

Summary ANOVA for Negative Reaction

Source	SS	df	$MS = \frac{SS}{df}$	$F = \frac{MS_{source}}{MS_{resid}}$	F_{CV}
Model		$k - 1$			
a		$k_a - 1$			
b		$k_b - 1$			
aXb		$(df_a)(df_b)$			
Residual		$n - k$			
Total					

```
# hand-calculating the MS values
35.415/5 #Model
```

[1] 7.083

```
12.243/2 #a: Rater
```

```
[1] 6.1215
```

```
14.575/1 #b: Photo
```

```
[1] 14.575
```

```
8.597/2 #axb interaction term
```

```
[1] 4.2985
```

```
79.321/105 #residual
```

```
[1] 0.7554381
```

```
# hand-calculating the F values
```

```
7.083/0.755 #Model
```

```
[1] 9.381457
```

```
6.122/0.755 #a: Rater
```

```
[1] 8.108609
```

```
14.575/0.755 #b: Photo
```

```
[1] 19.30464
```

```
4.299/0.755 #axb interaction term
```

```
[1] 5.69404
```

To find the F_{CV} we can use an [F distribution table](#).

Or use a look-up function, which follows this general form: `qf(p, df1, df2, lower.tail=FALSE)`

```
# looking up the F critical values
qf(0.05, 5, 105, lower.tail = FALSE) #Model F critical value
```

```
[1] 2.300888
```

```
qf(0.05, 2, 105, lower.tail = FALSE) #a and axb F critical value
```

```
[1] 3.082852
```

```
qf(0.05, 1, 105, lower.tail = FALSE) #b F critical value
```

```
[1] 3.931556
```

When the F value exceeds the F_{CV} , the effect is statistically significant.

Summary ANOVA for Negative Reaction

Source	SS	df	$MS = \frac{SS}{df}$	$F = \frac{MS_{source}}{MS_{resid}}$	F_{CV}
Model	35.415	5	7.083	9.381	2.301
a	12.243	2	6.122	8.109	3.083
b	14.575	1	14.575	19.305	3.932
aXb	8.597	2	4.299	5.694	3.083
Residual	79.321	105	0.755		
Total	114.775				

8.4.7 Interpreting the results

What have we learned?

- there is a main effect for Rater
- there is a main effect for Photo
- there is a significant interaction effect

In the face of this significant interaction effect, we would follow-up by investigating the interaction effect. Why? The significant interaction effect means that findings (e.g., the story of the results) are more complex than group identity or photo stimulus, alone, can explain.

8.5 Working the Factorial ANOVA with R Packages

8.5.1 Evaluating the statistical assumptions

All statistical tests have some assumptions about the data. I have marked our Two-Way ANOVA Workflow with a yellow box outlined in red to let us know that we are just beginning the process of analyzing our data with an evaluation of the statistical assumptions.

The are four critical assumptions in factorial ANOVA:

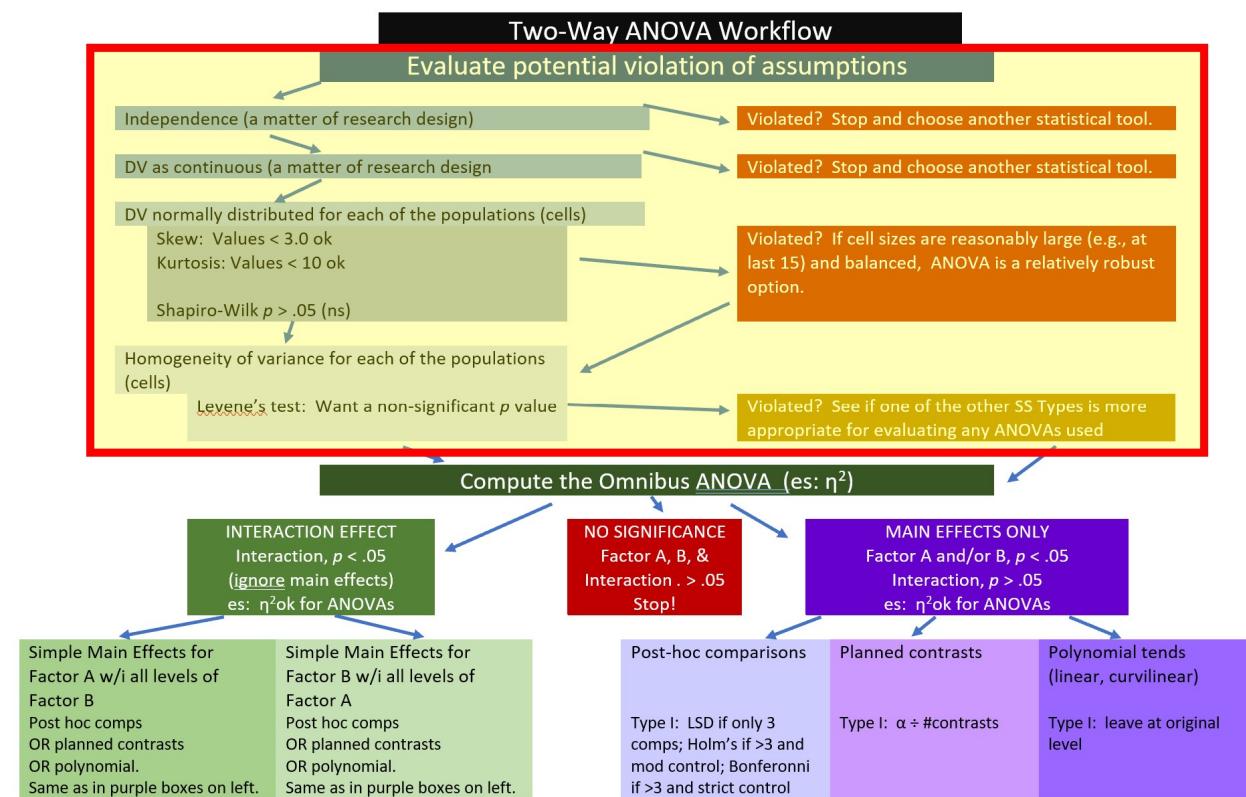


Figure 8.5: Image of a flowchart showing that we are on the “Evaluating assumptions” portion of the workflow

- Cases represent random samples from the populations
 - This is an issue of research design
 - Although we see ANOVA used (often incorrectly) in other settings, ANOVA was really designed for the random clinical trial (RCT).
- Scores on the DV are independent of each other.
 - This is an issue of research design
 - With correlated observations, there is a dramatic increase of Type I error
 - There are alternative statistics designed for analyzing data that has dependencies (e.g., repeated measures ANOVA, dyadic data analysis, multilevel modeling)
- The DV is normally distributed for each of the populations
 - that is, data for each cell (representing the combinations of each factor) is normally distributed
- Population variances of the DV are the same for all cells
 - When cell sizes are not equal, ANOVA not robust to this violation and cannot trust F ratio

Even though we position the evaluation of assumptions first – some of the best tests of the assumptions use the resulting ANOVA model. Because of this, I will quickly run the model now. I will not explain the results until after we evaluate the assumptions.

```
TwoWay_neg <- aov(Negative ~ Rater * Photo, Ramdhani_df)
summary(TwoWay_neg)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
Rater	2	12.21	6.103	8.077	0.000546 ***						
Photo	1	14.62	14.619	19.346	0.0000262 ***						
Rater:Photo	2	8.61	4.304	5.696	0.004480 **						
Residuals	105	79.34	0.756								

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

```
model.tables(TwoWay_neg, "means")
```

Tables of means

Grand mean

2.947369

Rater

	Dayaknese	Javanese	Madurese
rep	2.492	3.007	3.299
rep	35.000	37.000	39.000

Photo

	Dayaknese	Madurese
	2.575	3.301
rep	54.000	57.000

Rater:Photo

	Photo	
Rater	Dayaknese	Madurese
Dayaknese	1.818	3.129
rep	17.000	18.000
Javanese	2.524	3.465
rep	18.000	19.000
Madurese	3.301	3.298
rep	19.000	20.000

8.5.1.1 Is the dependent variable normally distributed?

8.5.1.1.1 Is there evidence of skew or kurtosis? Let's start by analyzing `skew` and `kurtosis`. Skew and kurtosis are one way to evaluate whether or not data are normally distributed. When we use the “`type=1`” argument, the skew and kurtosis indices in `psych::describe` (or `psych::describeBy`) can be interpreted according to Kline's [2016a] guidelines. Regarding skew, values greater than the absolute value of 3.0 are generally considered “severely skewed.” Regarding kurtosis, “severely kurtotic” is argued to be anywhere greater the absolute values of 8 to 20. Kline recommended using a conservative threshold of the absolute value of 10.

```
psych::describeBy(Negative ~ Rater + Photo, mat = TRUE, data = Ramdhani_df,
  digits = 3, type = 1)
```

	item	group1	group2	vars	n	mean	sd	median	trimmed	mad
Negative1	1	Dayaknese	Dayaknese		17	1.818	0.768	1.692	1.783	0.694
Negative2	2	Javanese	Dayaknese		18	2.524	0.742	2.391	2.460	0.569
Negative3	3	Madurese	Dayaknese		19	3.301	1.030	3.314	3.321	1.294
Negative4	4	Dayaknese	Madurese		18	3.129	0.156	3.160	3.136	0.104
Negative5	5	Javanese	Madurese		19	3.465	0.637	3.430	3.456	0.767
Negative6	6	Madurese	Madurese		20	3.297	1.332	2.958	3.254	1.615
		min	max	range	skew	kurtosis	se			
Negative1	0.706	3.453	2.747	0.562	-0.608	0.186				
Negative2	1.406	4.664	3.258	1.313	2.017	0.175				
Negative3	1.406	4.854	3.448	-0.137	-1.069	0.236				
Negative4	2.732	3.423	0.691	-0.679	0.903	0.037				
Negative5	2.456	4.631	2.175	-0.010	-1.114	0.146				
Negative6	1.211	5.641	4.430	0.232	-1.048	0.298				

Using guidelines from Kline [2016b] our values for skewness fall below $|3.0|$ and our values for kurtosis fall below $|10|$.

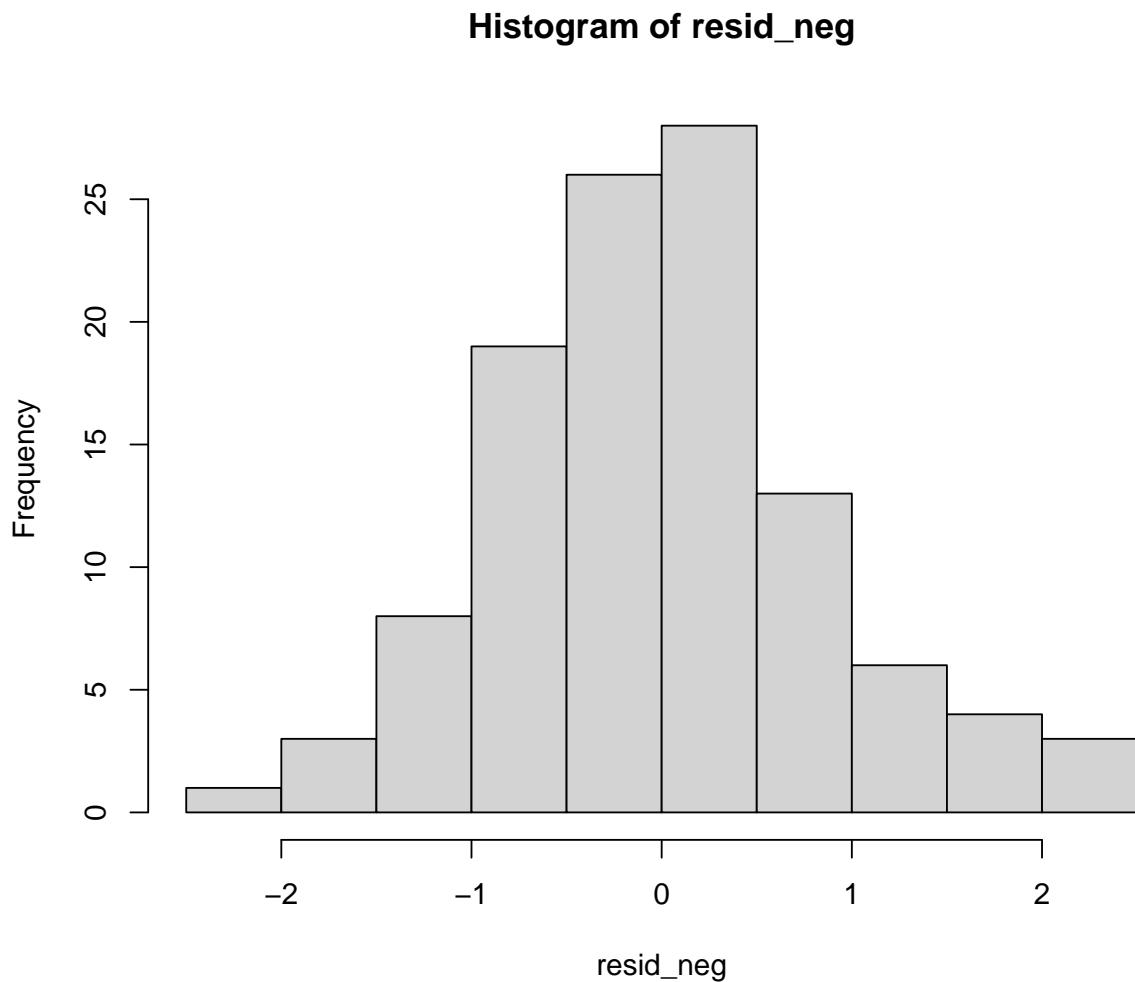
8.5.1.1.2 Are the model residuals normally distributed? We can further investigate normality with the Shapiro-Wilk test. The assumption requires that the distribution be normal in each of the levels of each factor. In the case of multiple factors (such as is the case in factorial ANOVA), the assumption requires a normal distribution in each combination of these levels (e.g., Javanese rater of Dyaknese photo). In this lesson's 3 x 2 ANOVA, there are six such combinations. This cell-level analysis has been demonstrated in [one-way ANOVA](#) and independent [t-test](#) lessons. To the degree that there are many factorial combinations (and therefore, cells), this approach becomes unwieldy to calculate, interpret, and report. The cell-level analysis of normality is also only appropriate when there are a low number of levels/groupings and there are many data points per group. Thus, as models become more complex, researchers turn to the model-based option for assessing normality. To do this, we first create an object that tests our research model.

Just a paragraph or two earlier, I ran the factorial ANOVA and saved the results in an object. Among the information contained in that object are *residuals*. Residuals are the unexplained variance in the outcome (or dependent) variable after accounting for the predictor (or independent) variable. In the code below we extract the residuals (i.e., that which is left-over/unexplained) from the model. We can examine their distribution with a plot.

```
# creates object of residuals
resid_neg <- residuals(TwoWay_neg)
```

Next, we can take a “look” them with a couple of plots.

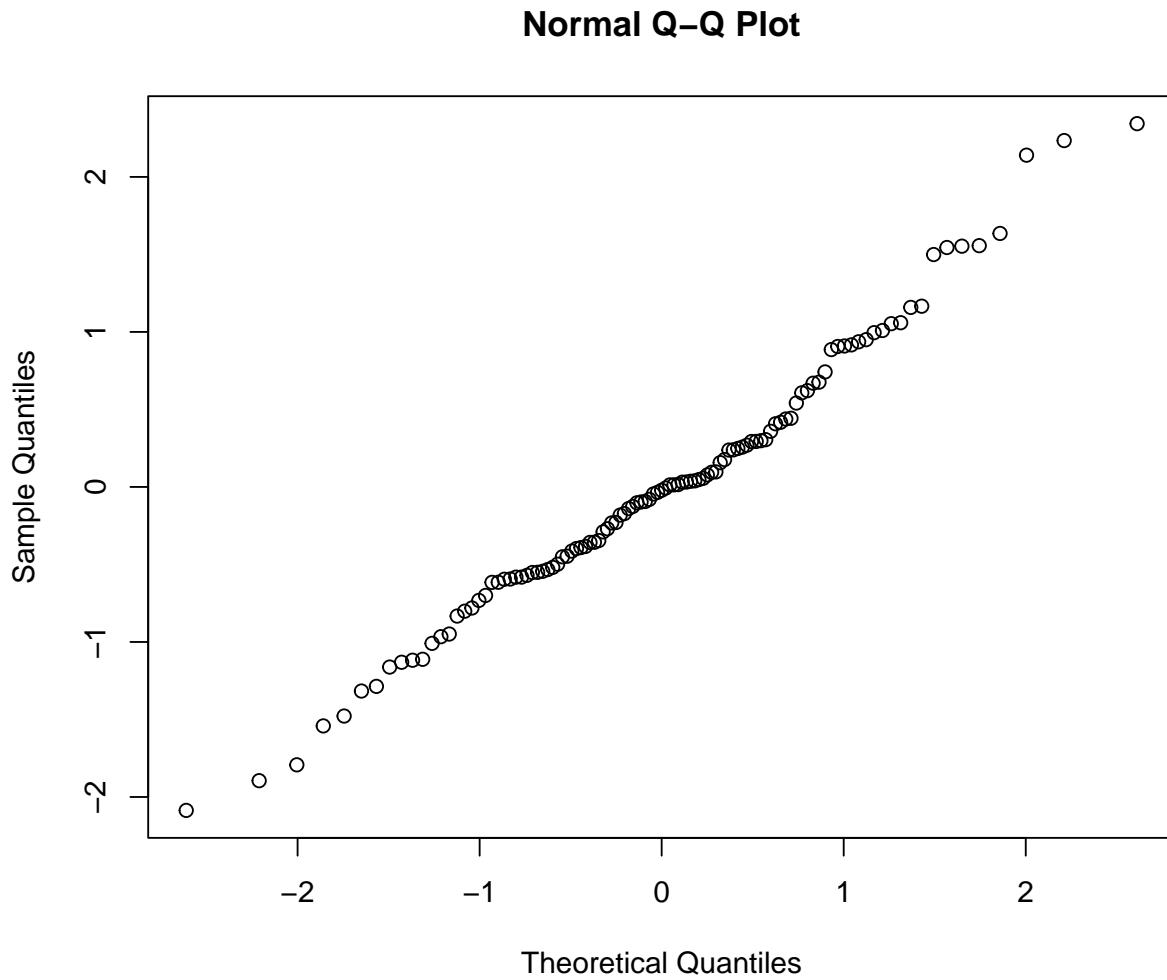
```
hist(resid_neg)
```



So far so good – our distribution of *residuals* (i.e., what is leftover after the model is applied) resembles a normal distribution.

The Q-Q plot provides another view. The dots represent the residuals. When they are relatively close to the line they not only suggest good fit of the model, but we know they are small and evenly distributed around zero (i.e., normally distributed).

```
qqnorm(resid_neg)
```



Additionally, we can formally test the distribution of the residuals with a Shapiro test. We want the associated p value to be greater than 0.05.

```
shapiro.test(resid_neg)
```

```
Shapiro-Wilk normality test

data: resid_neg
W = 0.98464, p-value = 0.2344
```

Whooo hoo! $p > 0.05$. This means that our distribution of residuals is not statistically significantly different from a normal distribution ($W = 0.985, p = 0.234$).

8.5.1.1.3 Are there outliers? If our data pointed to significant violations of normality, we could consider identifying and removing outliers. Removing data is a serious consideration that

should not be made lightly. If needed, though, here is a tool to inspect the data and then, if necessary, remove it.

We can think of outlier identification in a couple of ways. First, we might look at dependent variable across the entire dataset. That is, without regard to the levels of the grouping variable. We can point `rstatix::identify_outliers()` to the data.

```
Ramdhani_df %>%
  rstatix::identify_outliers(Negative)
```

	ID	Negative	Positive	Rater	Photo	m_dev	m_devSQ	is.outlier
1	73	5.641	4.813	Madurese	Madurese	2.693631	7.255646	TRUE
		is.extreme						
1		FALSE						

Our results indicate that one case (ID = 73) had an outlier (TRUE), but it was not extreme (FALSE).

Let's re-run the code, this time requiring it to look within each of the grouping levels of the condition variable.

```
Ramdhani_df %>%
  group_by(Rater, Photo) %>%
  rstatix::identify_outliers(Negative)
```

	Rater	Photo	ID	Negative	Positive	m_dev	m_devSQ	is.outlier	is.extreme
1	Dayaknese	Madure~	18	2.73	5.22	-0.215	0.0464	TRUE	FALSE
2	Dayaknese	Madure~	19	3.42	3.17	0.476	0.226	TRUE	FALSE
3	Javanese	Dayakn~	87	4.66	3.54	1.72	2.95	TRUE	FALSE

This time there are three cases where there are outliers (TRUE), but they are not extreme (FALSE). Handily, the function returns information about each row of data. We can use such information to help us delete it.

Let's say that, after very careful consideration, we decided to remove the case with ID = 18. We could use `dplyr::filter()` to do so. In this code, the `filter()` function locates all the cases where ID = 18. The exclamation point that precedes the equal sign indicates that the purpose is to remove the case.

```
# Ramdhani_df <- dplyr::filter(Ramdhani_df, ID != '18')
```

Once executed, we can see that this case is no longer in the dataframe. Although I demonstrated this in the accompanying lecture, I have hashtags out the command because I would not delete the case. If you already deleted the case, you can return the hashtag and re-run all the code up to this point.

Here's how I would summarize our data in terms of normality:

Factorial ANOVA assumes that the dependent variable is normally distributed for all cells in the design. Skew and kurtosis values for each factorial combinations fell below the guidelines recommended by Kline [2016a]. That is, they were below the absolute values of 3 for skew and 10 for kurtosis. Similarly, no extreme outliers were identified and results of the Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) suggested that model residuals did not differ significantly from a normal distribution ($W = 0.9846, p = 0.234$).

8.5.1.2 Are the variances of the dependent variable similar across the levels of the grouping factors?

We can evaluate the homogeneity of variance test with the Levene's test for the equality of error variances. Levene's requires a *fully saturated model*. This means that the prediction model requires an interaction effect (not just two, non-interacting predictors). We can use the `rstatix::levene_test()`. Within the function we point to the dataset, then specify the formula of the factorial ANOVA. That is, predicting Negative from the Rater and Photo factors. The asterisk indicates that they will also be added as an interaction term.

```
rstatix::levene_test(Ramdhani_df, Negative ~ Rater * Photo)
```

```
# A tibble: 1 x 4
  df1   df2 statistic      p
  <int> <int>    <dbl>    <dbl>
1     5    105     8.63 0.000000700
```

Levene's test, itself, is an F -test. Thus, its reporting assumes the form of an F -string. Our result has indicated a violation of the homogeneity of variance assumption ($F[5, 105] = 8.634, p < .001$). This is not surprising as the boxplots displayed some widely varying variances.

Should we be concerned? Addressing violations of homogeneity of variance in factorial ANOVA is complex. The following have been suggested:

- One approach is to use different error variances in follow-up to the omnibus. Kassambara [a] suggested that separate one-way ANOVAs for the analysis of simple main effects will provide these separate error terms.
- Green and Salkind [2017c] indicated that we should become more concerned about the trustworthiness of the p values from the omnibus two-way ANOVA when this assumption is violated and the cell sizes are unequal. In today's research vignette, our design is balanced (i.e., the cell sizes are quite similar).

8.5.1.3 Summarizing results from the analysis of assumptions

It is common for an APA style results section to begin with a review of the evaluation of the statistical assumptions. As we have just finished these analyses, I will document what we have learned so far:

Factorial ANOVA assumes that the dependent variable is normally distributed for all cells in the design. Skew and kurtosis values for each factorial combinations fell below the guidelines recommended by Kline [2016a]. That is, they were below the absolute values of 3 for skew and 10 for kurtosis. Similarly, no extreme outliers were identified and results of the Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) suggested that model residuals did not differ significantly from a normal distribution ($W = 0.9846, p = 0.234$). Results of Levene's test for equality of error variances indicated a violation of the homogeneity of variance assumption, ($F[5, 105] = 8.834, p < .001$). Given that cell sample sizes were roughly equal and greater than 15, each [Green and Salkind, 2017c] we proceeded with the two-way ANOVA.

8.5.2 Evaluating the Omnibus ANOVA

The F -tests associated with the two-way ANOVA are the *omnibus* – providing the result for the main and interaction effects.

Here's where we are in the workflow.

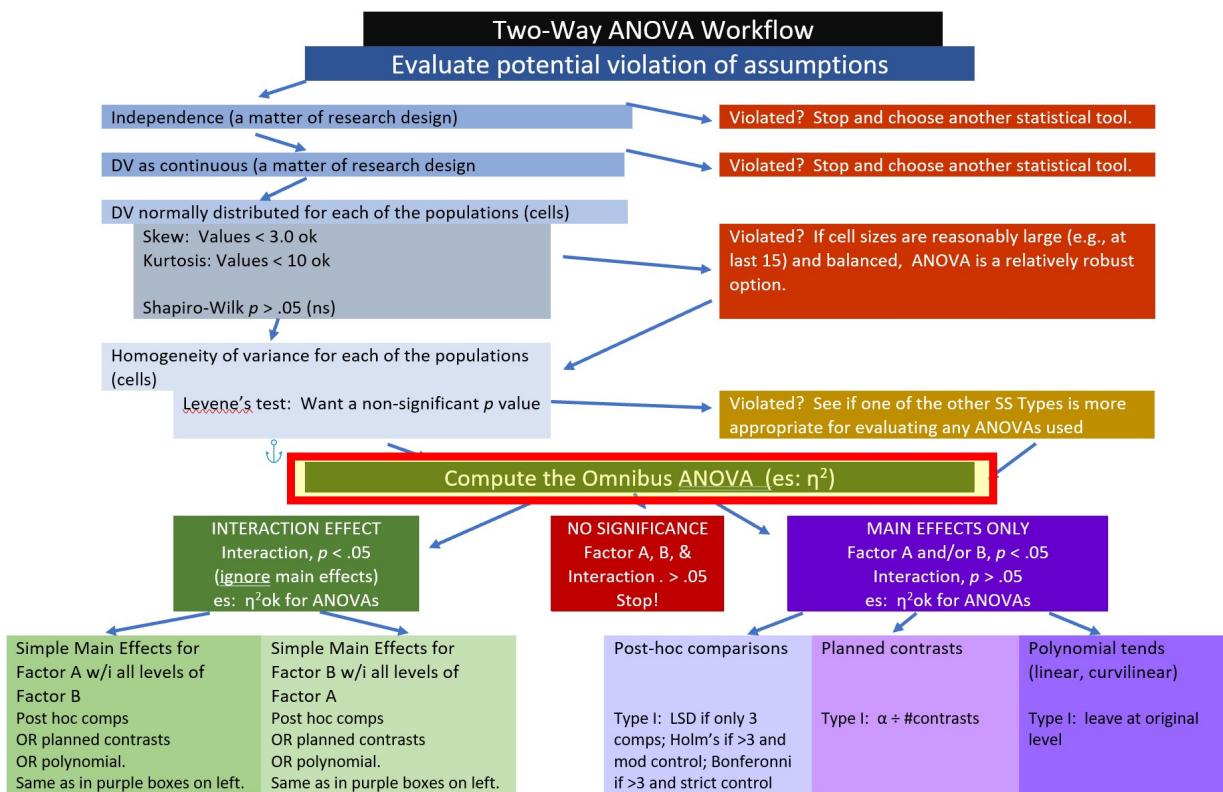


Figure 8.6: Image of our place in the Two-Way ANOVA Workflow.

When we run the two-way ANOVA we will be looking for several effects:

- main effects for each predictor, and
- the interaction effect.

It is possible that all effects will be significant, none will be significant, or some will be significant. The interaction effect always takes precedence over the main effect because it lets us know there is a more nuanced/complex result.

In the code below, the *type* argument is used to specify the type of sums of squares that are used. Type II is the *rstatix::anova_test()*'s default and is what I will use in this demonstration. It will yield identical results as *type=1* when data are balanced (i.e., cell sizes are equal). In specifying the ANOVA, order of entry matters if you choose *type=1*. In that case, if there are distinctions between independent variable and moderator, enter the independent variable first because it will claim the most variance. I provide more information on these options related to types of sums of squares calculations near the end of the chapter.

```
omnibus2w <- rstatix::anova_test(Ramdhani_df, Negative ~ Rater * Photo,
  type = "2", detailed = TRUE)
omnibus2w
```

ANOVA Table (type II tests)

	Effect	SSn	SSd	DFn	DFd	F	p	p<.05	ges
1	Rater	12.238	79.341	2	105	8.098	0.0005360	*	0.134
2	Photo	14.619	79.341	1	105	19.346	0.0000262	*	0.156
3	Rater:Photo	8.609	79.341	2	105	5.696	0.0040000	*	0.098

Let's write the *F strings* from the above table.

- Rater main effect: $F[2, 105] = 8.098, p < 0.001, \eta^2 = 0.134$
- Photo stimulus main effect: $F[1, 105] = 19.346, p < 0.001, \eta^2 = 0.156$
- Interaction effect: $F[2, 105] = 5.696, p = 0.004, \eta^2 = 0.098$

Eta squared (represented in the “ges” column of ouput) is one of the most commonly used measures of effect. It refers to the proportion of variability in the DV/outcome variable that can be explained in terms of the IVs/predictors. Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

You may see different values (.02, .13, .26) offered as small, medium, and large – these values are used when multiple regression is used. A useful summary of effect sizes, guide to interpreting their magnitudes, and common usage can be found [here \[Watson, 2020\]](#).

The formula for η^2 is straightforward:

$$\eta^2 = \frac{SS_M}{SS_T}$$

Before moving to follow-up, an APA style write-up of the omnibus might read like this:

8.5.2.1 APA write-up of the omnibus results

A 3 X 2 ANOVA was conducted to evaluate the effects of rater ethnicity (3 levels, Dayaknese, Madurese, Javanese) and photo stimulus (2 levels, Dayaknese on Madurese,) on negative reactions to the photo stimuli.

Computing sums of squares with a Type II approach, the results for the ANOVA indicated a significant main effect for ethnicity of the rater ($F[2, 105] = 8.098, p < 0.001, \eta^2 = 0.134$), a significant main effect for photo stimulus, ($F[1, 105] = 19.346, p < 0.001, \eta^2 = 0.156$), and a significant interaction effect ($F[2, 105] = 5.696, p = 0.004, \eta^2 = 0.098$).

8.5.3 Follow-up to a Significant Interaction Effect

In factorial ANOVA we are interested in main effects and interaction effects. When the result is explained by a main effect, then there is a consistent trend as a function of a factor (e.g., Madurese raters had consistently higher Negative evaluations, irrespective of stimulus). In an interaction effect, the results are more complex (e.g., the ratings across the stimulus differed for the three groups of raters).

There are a variety of strategies to follow-up a significant interaction effect. In this lesson, I demonstrate the two I believe to be the most useful in the context of psychologists operating within the scientist-practitioner-advocacy context. I provide additional examples in the [appendix](#).

When an interaction effect is significant (irrespective of the significance of one or more main effects), examination of **simple main effects** is a common statistical/explanatory approached that is used. The Two-Way ANOVA Workflow shows where we are in this process. Our research vignette is a 3 x 2 ANOVA. The first factor, ethnicity, has three levels (Dayaknese, Javanese, Madurese) and the second factor, photo stimulus, has two levels (Dayaknese, Madurese). When we conduct simple main effects, we evaluate one factor within the levels of the other factor. The number of levels in each factor changes the number of steps (i.e., the complexity) in the analysis.

When I am analyzing the simple main effect of photo stimulus (two levels) within ethnicity of the rater (three levels), I only need a one-step procedure that will conduct pairwise comparisons of the negative evaluating of the photo stimulus for the Dayaknese, Javanese, and Madurese raters, separately (while controlling for Type I error). Traditionally, researchers will follow with three, separate, one-way ANOVAs. However, any procedure (e.g., t-tests, pairwise comparisons) that will make these pairwise comparisons is sufficient.

When I am analyzing the simple main effect of ethnicity of the rater (three levels) within photo stimulus (two levels), I will need a two-step process. The first step will require the one-way ANOVA to determine, first, if there were statistically significant differences within the photo stimulus (e.g., Were there differences between Dayaknese, Javanese, and Madurese raters when viewing the Dayaknese photos?). If there were statistically significant differences, we follow up with an analysis of pairwise comparisons.

Although I will demonstrate both rater ethnicity within photo stimulus and photo stimulus within rater ethnicity in this lesson, we will choose only one for the write-up of results.

8.5.3.1 Planning for the management of Type I Error

Controlling for Type I error can depend, in part, on the design of the follow-up tests that are planned, and the number of pairwise comparisons that follow.

In the first option, the examination of the simple main effect of photo stimulus within ethnicity of rater results in only three pairwise comparisons. In this case, I will use the traditional Bonferroni. Why? Because there are only three post omnibus analyses, its more restrictive control is less likely to be problematic.

In the second option, the examination of the simple main effect of ethnicity of the rater within photo stimulus results in the potential comparison of six pairwise comparisons. If we used a traditional Bonferroni and divided $.05/6$, the p value for each comparison would need to be less than 0.008. Most would agree that this is too restrictive.

```
.05/6
```

```
[1] 0.008333333
```

The Holm's sequential Bonferroni [Green and Salkind, 2017c] offers a middle-of-the-road approach (not as strict as $.05/6$ with the traditional Bonferroni; not as lenient as “none”) to managing Type I error.

If we were to hand-calculate the Holms, we would rank order the p values associated with the six comparisons in order from lowest (e.g., 0.000001448891) to highest (e.g., 1.000). The first p value is evaluated with the most strict criterion ($.05/6$; the traditional Bonferonni approach). Then, each successive comparison calculates the p value by using the number of *remaining* comparisons as the denominator (e.g., $.05/5$, $.05/4$, $.05/3$). As the p values increase and the alpha levels relax, there will be a cut-point where remaining comparisons are not statistically significant. Luckily, most R packages offer the Holm's sequential Bonferroni as an option. The algorithm in the package rearranges the mathematical formula and produces a p value that we can interpret according to the traditional values of $p < .05$, $p < .01$ and $p < .001$. I will demonstrate use of Holm's in the examination of the simple main effect of ethnicity of rater within photo stimulus.

8.5.3.2 Option #1 the simple main effect of photo stimulus within ethnicity of the rater

In the examination of the simple main effect of photo stimulus within ethnicity of the rater our goal is to compare the:

- Dayaknese raters' negative evaluation of the Dayaknese and Madurese photos,
- Javanese raters' negative evaluation fo the Dayaknese and Madurese photos, and
- Madurese raters' negative evaluation of the Dayaknese and Madurese photos.

Thus, we only need three, pairwise comparisons. I will demonstrate two ways to conduct these analyses. Here's where we are in the two-way ANOVA workflow

Separate one-way ANOVAs are a traditional option for this evaluation. Using `dplyr::group_by()` we can efficiently calculate the three ANOVAs by the grouping variable, Rater. One advantage of

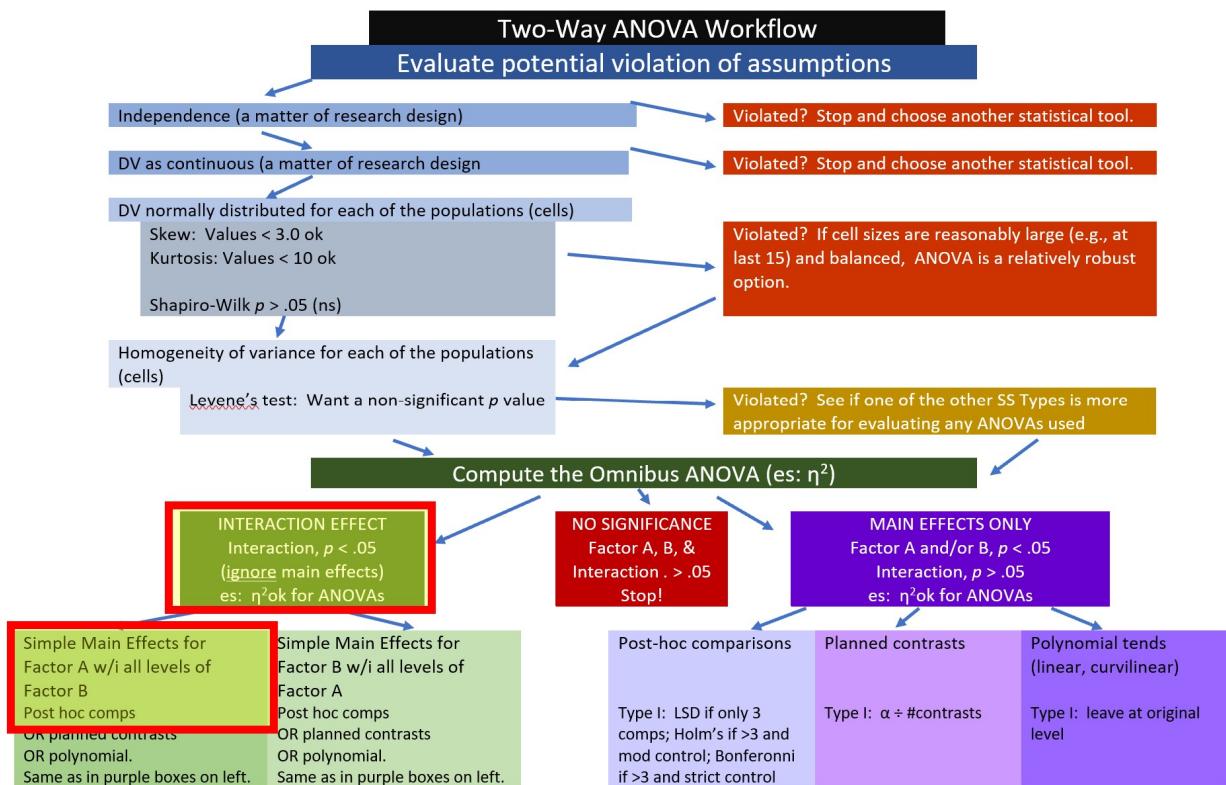


Figure 8.7: Image our place in the Two-Way ANOVA Workflow – analysis of simple main effects of factor A within levels of factor B.

separate one-way ANOVAs is that they each have their own error term and that this can help mitigate problems associated with violation of the homogeneity of variance assumption [Kassambara, a].

Note that in this method there is no option for controlling Type I error. Thus, we would need to do it manually. The traditional Bonferroni involves dividing family-wise error (traditionally $p < .05$) by the number of follow-up comparisons. In our case $.05/3 = .017$.

```
Ramdhani_df %>%
  dplyr::group_by(Rater) %>%
  rstatix::anova_test(Negative ~ Photo)

# A tibble: 3 x 8
  Rater   Effect   DFn   DFd       F      p `p<.05`     ges
* <fct>   <chr>   <dbl> <dbl>    <dbl>    <dbl> <chr>     <dbl>
1 Dayaknese Photo     1     33  50.4     0.0000000395 "*"     0.604
2 Javanese   Photo     1     35  17.2     0.000205      "*"     0.329
3 Madurese   Photo     1     37  0.0000762 0.993      ""     0.00000206
```

The APA style write-up will convey what we have found using this traditional approach:

To explore the interaction effect, we followed with a test of the simple main effect of photo stimulus within the ethnicity of the rater. That is, with separate one-way ANOVAs (chosen, in part, to mitigate violation of the homogeneity of variance assumption [Kassambara, a]) we examined the effect of the photo stimulus within the Dayaknese, Madurese, and Javanese groups. To control for Type I error across the three simple main effects, we set alpha at .017 (.05/3). Results indicated significant differences for Dayaknese ($F[1, 33] = 50.404, p < 0.001, \eta^2 = 0.604$) and Javanese ethnic groups ($F[1, 35] = 17.183, p < 0.001, \eta^2 = 0.329$), but not for the Madurese ethnic group ($F[1, 37] < 0.001, p = .993, \eta^2 < .001$). As illustrated in Figure 1, the Dayaknese and Javanese raters both reported stronger negative reactions to the Madurese. The differences in ratings for the Madurese were not statistically significantly different. In this way, the rater's ethnic group moderated the relationship between the photo stimulus and negative reactions.

The `rstatix::emmeans_test()` offers an efficient alternative to this pairwise analysis that will (a) automatically control for Type I error and (b) integrate well into a figure. Note that this function is a wrapper to functions in the `emmeans` package. If you haven't already, you will need to install the `emmeans` package. For each, the resulting test statistic is a *t.ratio*. The result of this *t*-test will be slightly different than an independent sample *t*-test because it is based on *estimated marginal means* (i.e., means based on the model, not directly on the data). We will spend more time with estimated marginal means in the ANCOVA lesson.

In the script below, we will group the dependent variable by Rater and then conduct pairwise comparisons. Note that I have requested that that the traditional Bonferroni be used to manage Type I error. We can see these adjusted *p* values in the output.

```

library(tidyverse)
pwPHwiETH <- Ramdhani_df%>%
  group_by(Rater)%>%
  rstatix::emmeans_test(Negative ~ Photo, detailed = TRUE, p.adjust.method = "bonferroni")
pwPHwiETH

# A tibble: 3 x 15
  Rater    term   .y.    group1 group2 null.value estimate     se     df conf.low
* <fct>   <chr> <chr>   <chr>   <chr>      <dbl>     <dbl> <dbl> <dbl>    <dbl>
1 Dayaknese Photo Negati~ Dayak~ Madur~       0 -1.31    0.294  105  -1.89
2 Javanese   Photo Negati~ Dayak~ Madur~       0 -0.941   0.286  105  -1.51
3 Madurese   Photo Negati~ Dayak~ Madur~       0  0.00334 0.278  105  -0.549
# i 5 more variables: conf.high <dbl>, statistic <dbl>, p <dbl>, p.adj <dbl>,
#   p.adj.signif <chr>

```

Not surprisingly, our results are quite similar. I would report them this way:

To explore the interaction effect, we followed with a test of the simple main effect of photo stimulus within the ethnicity of the rater. Specifically, we conducted pairwise comparisons between the groups using the estimated marginal means. We specified the Bonferroni method for managing Type I error. Results suggested statistically significant differences for the Dayaknese ($M_{diff} = -1.312, t[105] = -4.461, p < 0.001$) and Javanese ethnic groups ($M_{diff} = -0.941, t[105] = -3.291, p < 0.001$) but not for the Madurese ethnic group ($M_{diff} = 0.003, t[105] = 0.0121, p = 0.990$). As illustrated in Figure 1, the Dayaknese and Javanese raters both reported stronger negative reactions to the Madurese. The differences in ratings for the Madurese were not statistically significantly different. In this way, the rater's ethnic group moderated the relationship between the photo stimulus and negative reactions.

Because we used the *rstatix* functions, we can easily integrate them into our *ggpubr::ggboxplot()*. Let's first re-run the version of the boxplot where "Rater" is on the x-axis (and, is therefore our grouping variable). Because I want the data to be as true-to-scale as possible, I have added the full range of the y axis through the *ylim* argument. In order to update the ggboxplot, we will need to save it as an option. My object name represents the "PHoto within Ethnicity" simple main effect.

```

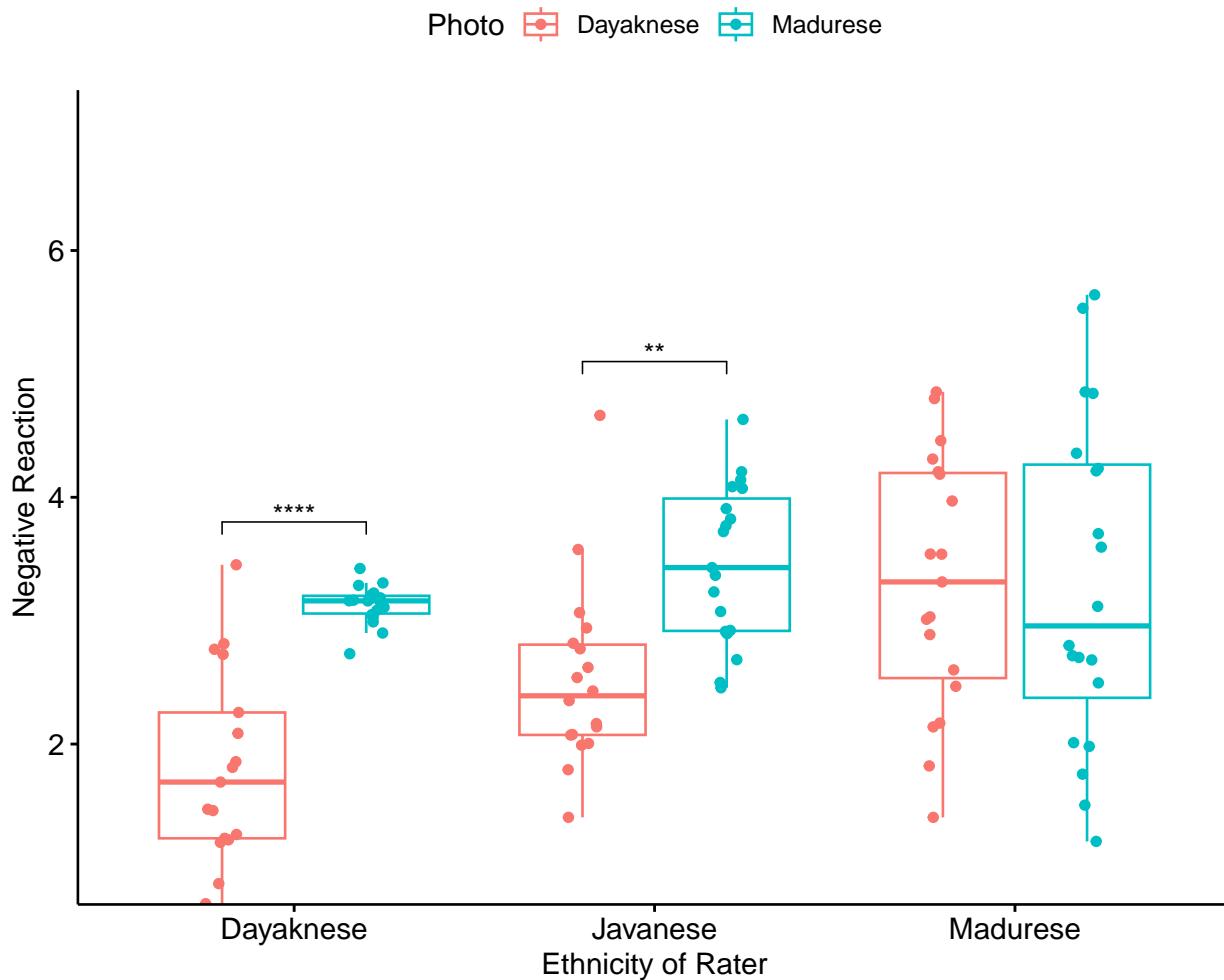
boxPHwiETH <- ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative",
  color = "Photo", xlab = "Ethnicity of Rater", ylab = "Negative Reaction",
  add = "jitter", title = "Simple Main Effect of Photo Stimulus within Rater",
  ylim = c(1, 7))

pwPHwiETH <- pwPHwiETH %>%
  rstatix::add_xy_position(x = "Rater") #x should be whatever the variable was used in the
boxPHwiETH <- boxPHwiETH + ggpubr::stat_pvalue_manual(pwPHwiETH, label = "p.adj.signif",
  tip.length = 0.02, hide.ns = TRUE, y.position = c(3.8, 5.1))

boxPHwiETH

```

Simple Main Effect of Photo Stimulus within Rater



8.5.3.3 Option #2 the simple main effect of ethnicity of rater within photo stimulus.

In the examination of the simple main effect of rater ethnicity photo stimulus our goal is to compare:

- Dayaknese, Javanese, and Madurese negative evaluations of the Dayaknese photos, and
- Dayaknese, Javanese, and Madurese negative evaluations of the Maudurese photos.

Consequently, we will need a two-staged evaluation. First, we will conduct separate one-way ANOVAs. Second, we will follow-up with pairwise comparisons.

Let's start with the one-way ANOVAs. Using `dplyr::group_by()` we can efficiently calculate the three ANOVAs by the grouping variable, Photo. One advantage of separate one-way ANOVAs is that they each have their own error term and that this can help mitigate problems associated with violation of the homogeneity of variance assumption [Kassambara, a].

Note that in this method there is no option for controlling Type I error. Thus, we would need to do it manually. The traditional Bonferroni involves dividing family-wise error (traditionally $p < .05$) by the number of follow-up comparisons. In our case $.05/2 = .025$.

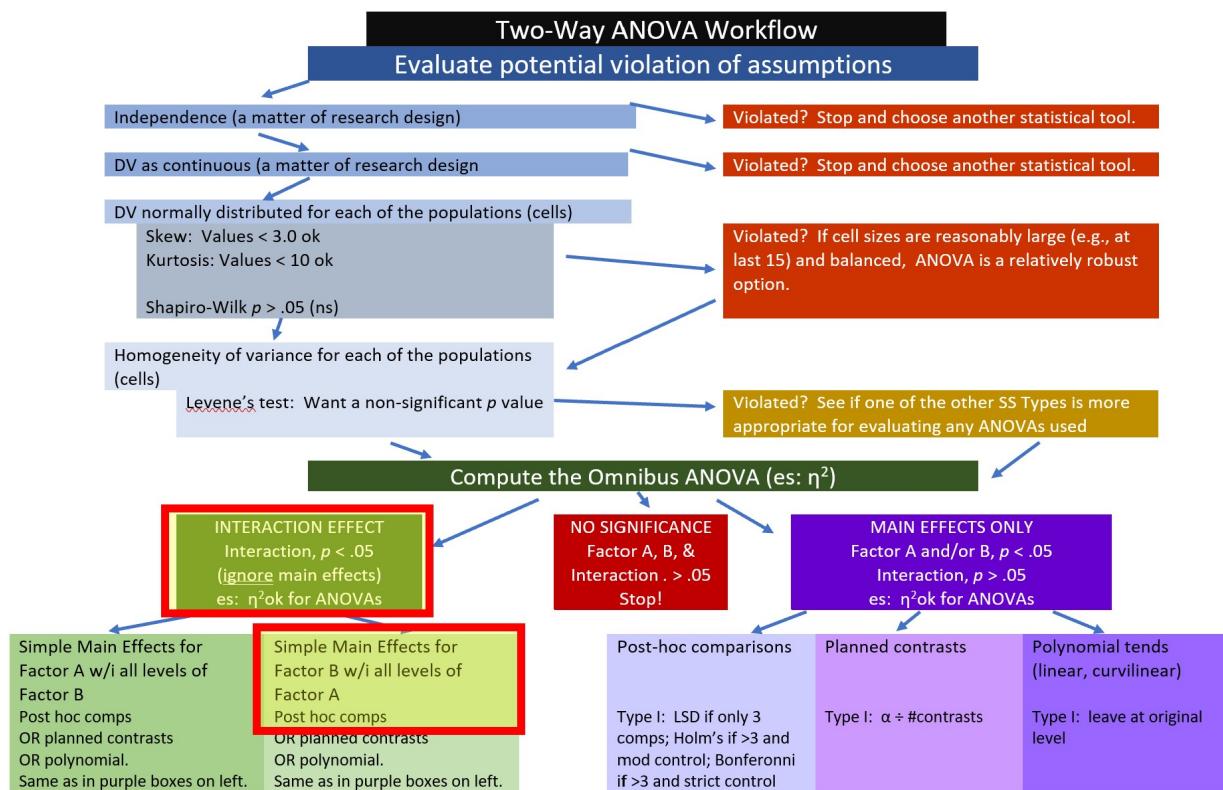


Figure 8.8: Image our place in the Two-Way ANOVA Workflow – analysis of simple main effects of factor B within levels of factor A.

```
Ramdhani_df %>%
  dplyr::group_by(Photo) %>%
  rstatix::anova_test(Negative ~ Rater)
```

	Photo	Effect	DFn	DFd	F	p `p<.05`	ges	
*	<fct>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	
1	Dayaknese	Rater	2	51	13.3	0.0000221	"*"	0.343
2	Madurese	Rater	2	54	0.679	0.512	""	0.025

The APA style write-up will convey what we have found (so far) using this approach:

To explore the interaction effect, we followed with a test of the simple main effect of ethnicity of the rater within the photo stimulus. We began with separate one-way ANOVAs (chosen, in part, to mitigate violation of the homogeneity of variance assumption [Kassambara, a]). To control for Type I error across the two simple main effects, we set alpha at .025 (.05/2). Results indicated significant differences for Dayaknese photo ($F[2, 51] = 13.325, p < 0.001, \eta^2 = 0.343$) but not for the Madurese photo ($F[2, 54] = 0.679, p = 0.512, \eta^2 = 0.025$).

Results suggest that there are differences within the Dayaknese group, yet because there are three groups, we cannot know with certainty where there are statistically significant difference. As before, we can use the `rstatix::emmeans_test()` to conduct the pairwise analysis. This function will (a) automatically control for Type I error and (b) integrate well into a figure. For each comparison, the resulting test statistic is a *t.ratios*. The result of this *t*-test will be slightly different than an independent sample *t*-test because it is based on *estimated marginal means* (i.e., means based on the model, not directly on the data). We will spend more time with estimated marginal means in the ANCOVA lesson.

In the script below, we will group the dependent variable by Photo and then conduct pairwise comparisons. Note that I have requested that that the Holm's sequential Bonferroni be used to manage Type I error. We can see these adjusted *p* values in the output.

```
pwETHwiPH <- Ramdhani_df %>%
  dplyr::group_by(Photo) %>%
  rstatix::emmeans_test(Negative ~ Rater, p.adjust.method = "holm")
pwETHwiPH
```

	Photo	term	.y.	group1	group2	df	statistic	p	p.adj	p.adj.signif
*	<fct>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Dayakn~	Rater	Nega~	Dayak~	Javan~	105	-2.40	1.81e-2	1.81e-2	*
2	Dayakn~	Rater	Nega~	Dayak~	Madur~	105	-5.11	1.45e-6	4.35e-6	****
3	Dayakn~	Rater	Nega~	Javan~	Madur~	105	-2.72	7.69e-3	1.54e-2	*
4	Madure~	Rater	Nega~	Dayak~	Javan~	105	-1.17	2.43e-1	7.29e-1	ns
5	Madure~	Rater	Nega~	Dayak~	Madur~	105	-0.595	5.53e-1	1	e+0 ns
6	Madure~	Rater	Nega~	Javan~	Madur~	105	0.601	5.49e-1	1	e+0 ns

Very consistent with the one-way ANOVAs, we see that there were significant rater differences in the evaluation of the Dayaknese photo, but not for the Madurese photo. Further, in the rating of the Dayaknese photo, there were statistically significant differences between all three comparisons of ethnic groups. The *p*-values remained statistically significant with the adjustment of the Holm's.

For a quick demonstration of differences in managing Type I error, I will replace “holm” with “bonferroni.” Here, we will see the more restrictive result, where one of the previously significant comparisons drops out. Note that I am not saving this results as an object – I don't want it to interfere with our subsequent analyses

```
# demonstration of the more restrictive bonferroni approach to
# managing Type I error
Ramdhani_df %>%
  dplyr::group_by(Photo) %>%
  rstatix::emmeans_test(Negative ~ Rater, p.adjust.method = "bonferroni")
```

	Photo	term	.y.	group1	group2	df	statistic	p	p.adj	p.adj.signif
*	<fct>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	Dayakn~	Rater	Nega~	Dayak~	Javan~	105	-2.40	1.81e-2	5.43e-2	ns
2	Dayakn~	Rater	Nega~	Dayak~	Madur~	105	-5.11	1.45e-6	4.35e-6	****
3	Dayakn~	Rater	Nega~	Javan~	Madur~	105	-2.72	7.69e-3	2.31e-2	*
4	Madure~	Rater	Nega~	Dayak~	Javan~	105	-1.17	2.43e-1	7.29e-1	ns
5	Madure~	Rater	Nega~	Dayak~	Madur~	105	-0.595	5.53e-1	1	e+0 ns
6	Madure~	Rater	Nega~	Javan~	Madur~	105	0.601	5.49e-1	1	e+0 ns

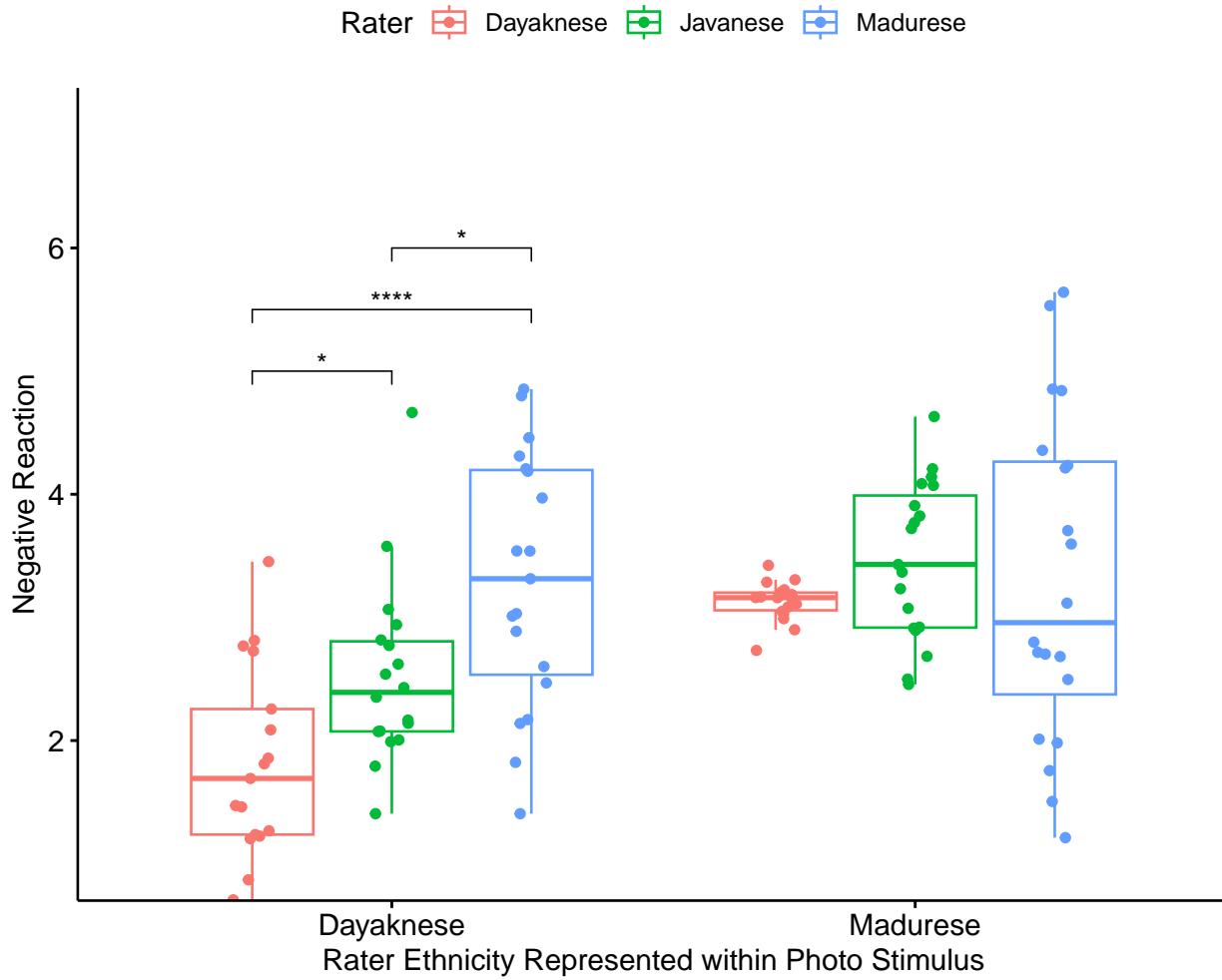
Let's create a figure that reflects the results of this simple main effect of rater ethnicity within photo stimulus. As before, we start with the corresponding figure where “Photo” is on the x-axis.

```
boxETHwiPH <- ggpubr::ggboxplot(Ramdhani_df, x = "Photo", y = "Negative",
  color = "Rater", xlab = "Rater Ethnicity Represented within Photo Stimulus",
  ylab = "Negative Reaction", add = "jitter", title = "Simple Main Effect of Rater within Photo Stimulus",
  ylim = c(1, 7))

pwETHwiPH <- pwETHwiPH %>%
  rstatix::add_xy_position(x = "Photo") #x should be whatever the variable was used in the boxETHwiPH <- boxETHwiPH + ggpubr::stat_pvalue_manual(pwETHwiPH, label = "p.adj.signif",
  tip.length = 0.02, hide.ns = TRUE, y.position = c(5, 5.5, 6))

boxETHwiPH
```

Simple Main Effect of Rater within Photo Stimulus



Here's how I would update the APA style reporting of results:

To explore the interaction effect, we followed with a test of the simple main effect of ethnicity of the rater within the photo stimulus. We began with separate one-way ANOVAs (chosen, in part, to mitigate violation of the homogeneity of variance assumption [Kassambara, a]). To control for Type I error across the two simple main effects, we set alpha at .025 (.05/2). Results indicated significant differences for Dayaknese photo ($F[2, 51] = 13.325, p < 0.001, \eta^2 = 0.343$) but not for the Madurese photo ($F[2, 54] = 0.679, p = 0.512, \eta^2 = 0.025$). We followed up significant one-way ANOVA with pairwise comparisons between the groups using the estimated marginal means. We specified the Holm's sequential Bonferroni for managing Type I error. Regarding evaluation of the Dayaknese photo, results suggested statistically significant differences in all combinations of raters. As shown in Figure 1, the Dayaknese raters had the lowest ratings, followed by Javanese raters, and then Madurese raters. Consistent with the non-significant one-way ANOVA evaluating ratings of the Madurese photo, there were no statistically significant differences for raters. Results of these tests are presented in Table 1.

8.5.3.4 Options #3 through k

There are seemingly infinite approaches to analyzing significant interaction effects. I am frequently asked, “But what about _____?” And “Do you have an example of _____?” In prior versions of this lesson, I included a few more examples in this section of follow-up to a significant interaction effect. However, in an effort to reduce the cognitive load of the chapter and stay focused on the primary learning goals I have relocated some of these to the [appendix](#). At the time of this update, there are worked examples that highlight:

- Orthogonal contrast-coding
- All possible post hoc comparisons
- Polynomial trends

If, as a reader, you have recommendations for more specific examples, please suggest them using the contact information provided at the beginning of the OER.

8.5.4 Investigating Main Effects

We now focus on the possibility that there might be significant main effects, but a non-significant interaction effect. We only interpret main effects when there is a non-significant interaction effect. Why? Because in the presence of a significant interaction effect, the main effect will not tell a complete story. If we didn’t specify a correct model, we still might have an incomplete story. But that’s another issue.

Here’s where we are on the workflow.

Recall that main effects are the *marginal means* – that is the effects of factor A *collapsed across* all the levels of factor B.

If the main effect has only two levels (e.g., the ratings of the Dayaknese and Madurese photos):

- the comparison was already ignoring/including all levels of the rater ethnicity factor (Dayaknese, Madurese, Javanese),
- it was only a comparison of two cells (Dayaknese rater, Madurese rater), therefore
- there is no need for further follow-up.

In the case of our specific research vignette, we learned from the omnibus test that the Photo main effect was statistically significant ($F[1, 105] = 19.346, p < 0.001, \eta^2 = 0.156$). This means that we know there are statistically significant differences between ratings of Dayaknese and Madurese photos overall.

```
psych::describeBy(Negative ~ Photo, data = Ramdhani_df, mat = TRUE)
```

	item	group1	vars	n	mean	sd	median	trimmed	mad
Negative1	1	Dayaknese	1	54	2.574926	1.0434646	2.449	2.516386	0.9206946
Negative2	2	Madurese	1	57	3.300211	0.8709631	3.166	3.279745	0.6671700
	min	max	range		skew	kurtosis	se		
Negative1	0.706	4.854	4.148	0.4699817	-0.5548515	0.1419975			
Negative2	1.211	5.641	4.430	0.3501228	0.5814430	0.1153619			

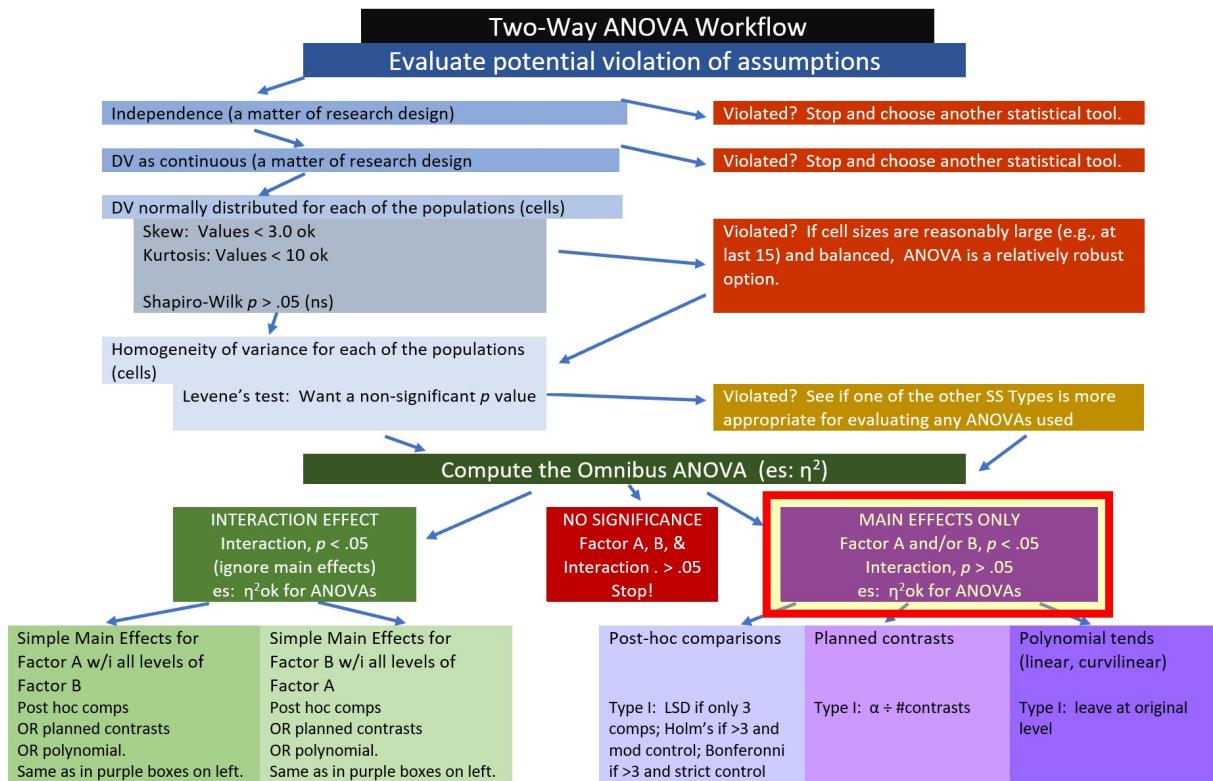


Figure 8.9: Image our place in the Two-Way ANOVA Workflow.

A quick review of the descriptive statistics, aggregated by photo stimulus indicates that, overall, Madurese photos were evaluated more negatively.

If the main effect has three or more levels (e.g., ethnicity of rater with Dayaknese, Madurese, Javanese levels), then we follow-up with one or more of the myriad of options. I tend to focus on three:

- planned contrasts
- posthoc comparisons (all possible cells)
- polynomial

From our omnibus evaluation, our rater main effect was $F[2, 105] = 8.098, p < .001, \eta^2 = 0.134$. I will demonstrate how to do each as follow-up to a *pretend* scenario where a main effect (but not the interaction effect) had been significant. In fact, our follow-up of Rater main effects will be quite similar to the manner in which we followed up the significant omnibus in the [one-way ANOVA lesson](#).

Here's what would happen if we simply ran a one-way ANOVA.

```
rater_main <- rstatix::anova_test(Ramdhani_df, Negative ~ Rater, detailed = FALSE)
rater_main
```

ANOVA Table (type II tests)

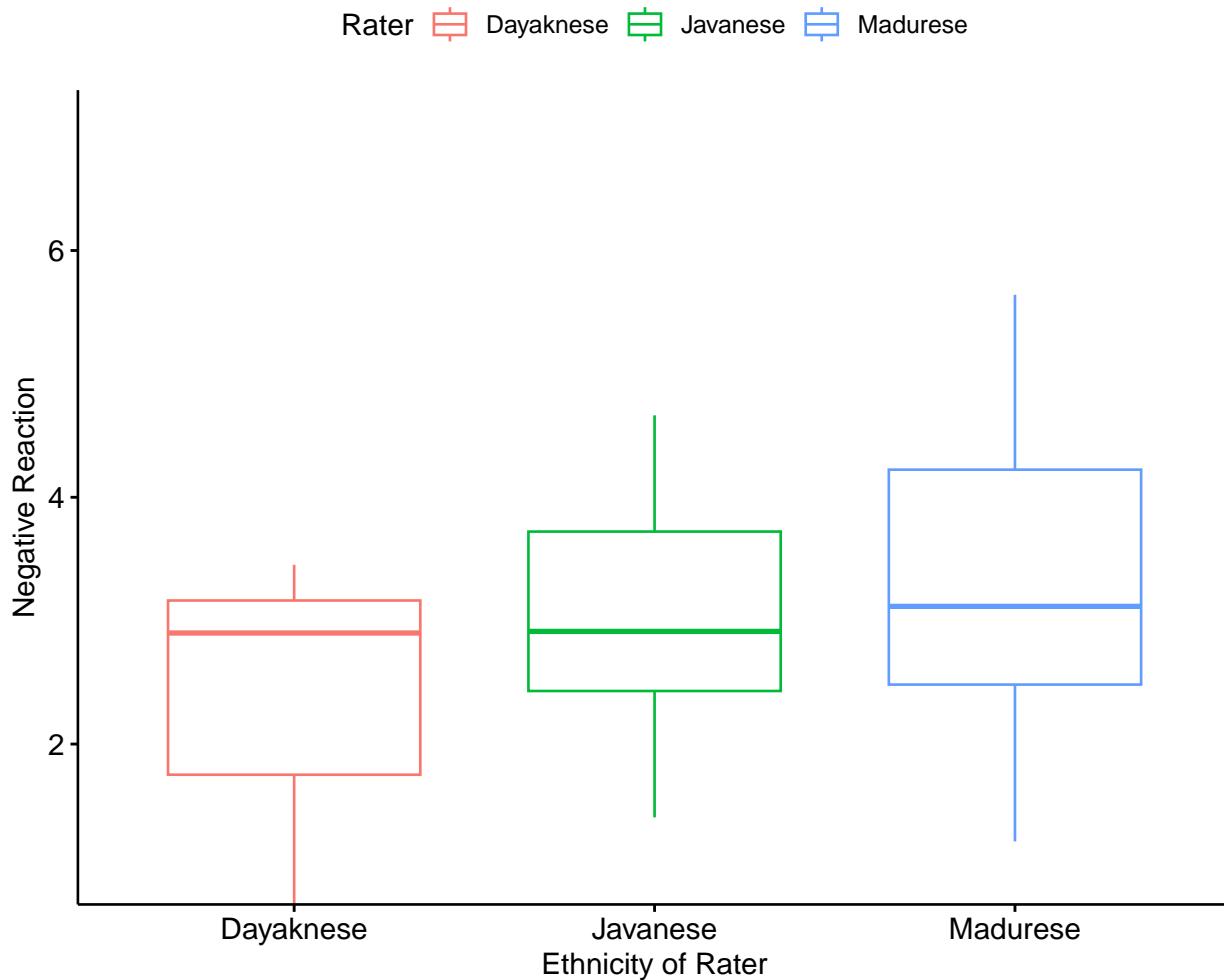
Effect	DFn	DFd	F	p	p<.05	ges
1 Rater	2	108	6.426	0.002	*	0.106

Results of a one-way ANOVA evaluating negative reaction to photos of members of Dayaknese and Madurese ethnic groups indicate a statistically differences as a function of the ethnicity of the rater ($F[2, 108] = 6.426, p = 0.002, \eta^2 = 0.106$)

A boxplot representing this main effect may help convey how the main effect of Rater (collapsed across Photo) is different than an interaction effect.

```
box_RaterMain <- ggpubr::ggbboxplot(Ramdhani_df, x = "Rater", y = "Negative",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction", color = "Rater",
  ylim = c(1, 7), title = "Boxplots of Rater Main Effect")
box_RaterMain
```

Boxplots of Rater Main Effect



8.5.4.1 Option #1 post hoc paired comparisons

An easy possibility is to follow-up with all possible post hoc pairwise comparisons. Here is a reminder of our location on the workflow.

Post hoc, pairwise comparisons are:

- used for exploratory work when no firm hypotheses were articulated a priori,
- used to compare the means of all combinations of pairs of an experimental condition, and
- less powerful than planned comparisons because more strict criterion for significance should be used.

By specifying the *formula* of the ANOVA, the *rstatix::t_test()* function will provide comparisons of all possible combinations. The arguments in the code mirror those we used for the omnibus. Note that I am saving the results as an object. We will use this object ("ttest") later when we create an accompanying figure.

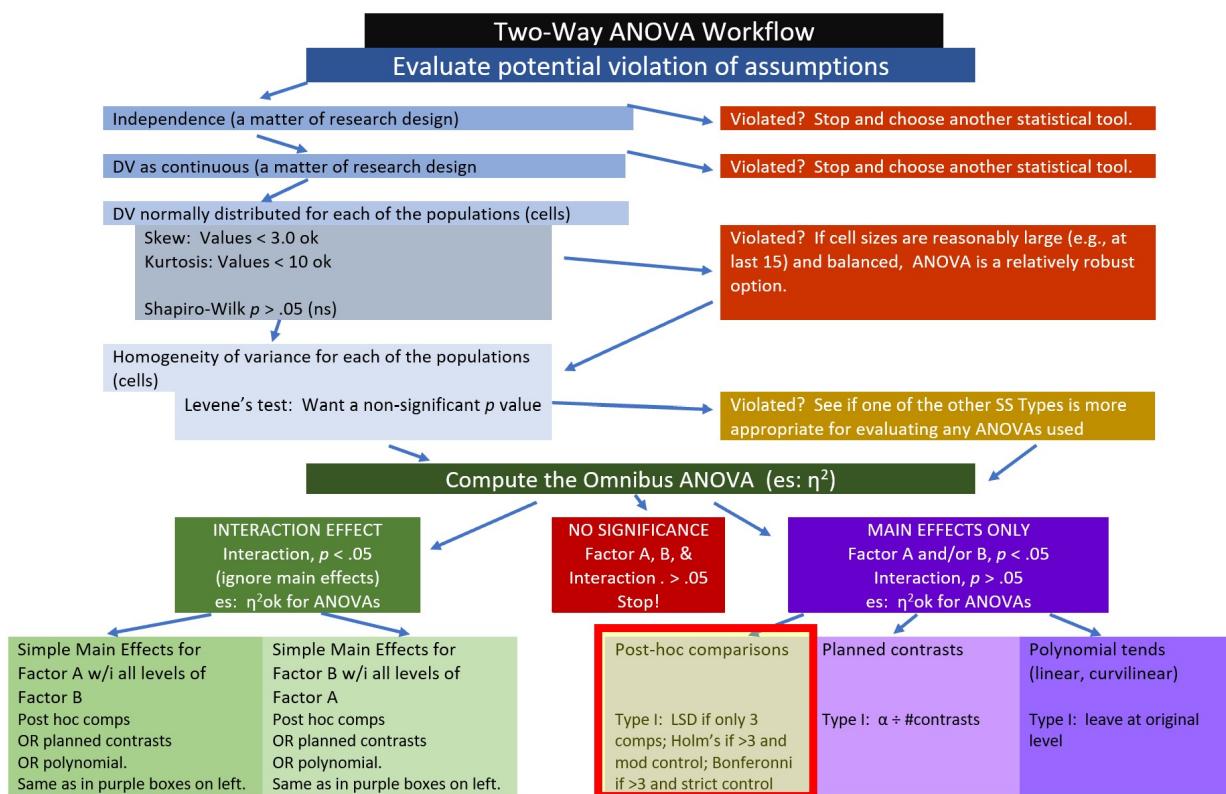


Figure 8.10: Image our place in the Two-Way ANOVA Workflow: Following up a significant main effect with post hoc comparisons.

We will request the traditional Bonferroni using the *p.adjust.method*. The *rstatix::t_test()* offers multiple options for adjusting the *p* values.

```
RaterMain_ttest <- rstatix::t_test(Ramdhani_df, Negative ~ Rater, p.adjust.method = "bonferroni",
  detailed = TRUE)
RaterMain_ttest

# A tibble: 3 x 17
  estimate estimate1 estimate2 .y. group1 group2   n1   n2 statistic     p
*    <dbl>     <dbl>     <dbl> <chr>  <chr>  <chr> <int> <int>    <dbl> <dbl>
1    -0.515      2.49      3.01 Negati~ Dayak~ Javan~    35    37    -2.59 0.012
2    -0.807      2.49      3.30 Negati~ Dayak~ Madur~    35    39    -3.39 0.001
3    -0.292      3.01      3.30 Negati~ Javan~ Madur~    37    39    -1.25 0.214
# i 7 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
# alternative <chr>, p.adj <dbl>, p.adj.signif <chr>
```

The *estimate* column provide the mean difference between the two levels of the independent different. The *estimate1/group1* and *estimate2/group2* columns provide those means and identify the group levels. The *statistic* column provides the value of the *t*-test.

The *p* value is the unadjusted *p*-value, it will usually be “more significant” (i.e., a lower value) than the *p.adj* value associated with the strategy for managing Type I error that we specified in our code. The column *p.adj.signif* provides symbolic notation associated with the *p.adj* value. In this specific case we specified the traditional Bonferroni as the adjusted *p* value.

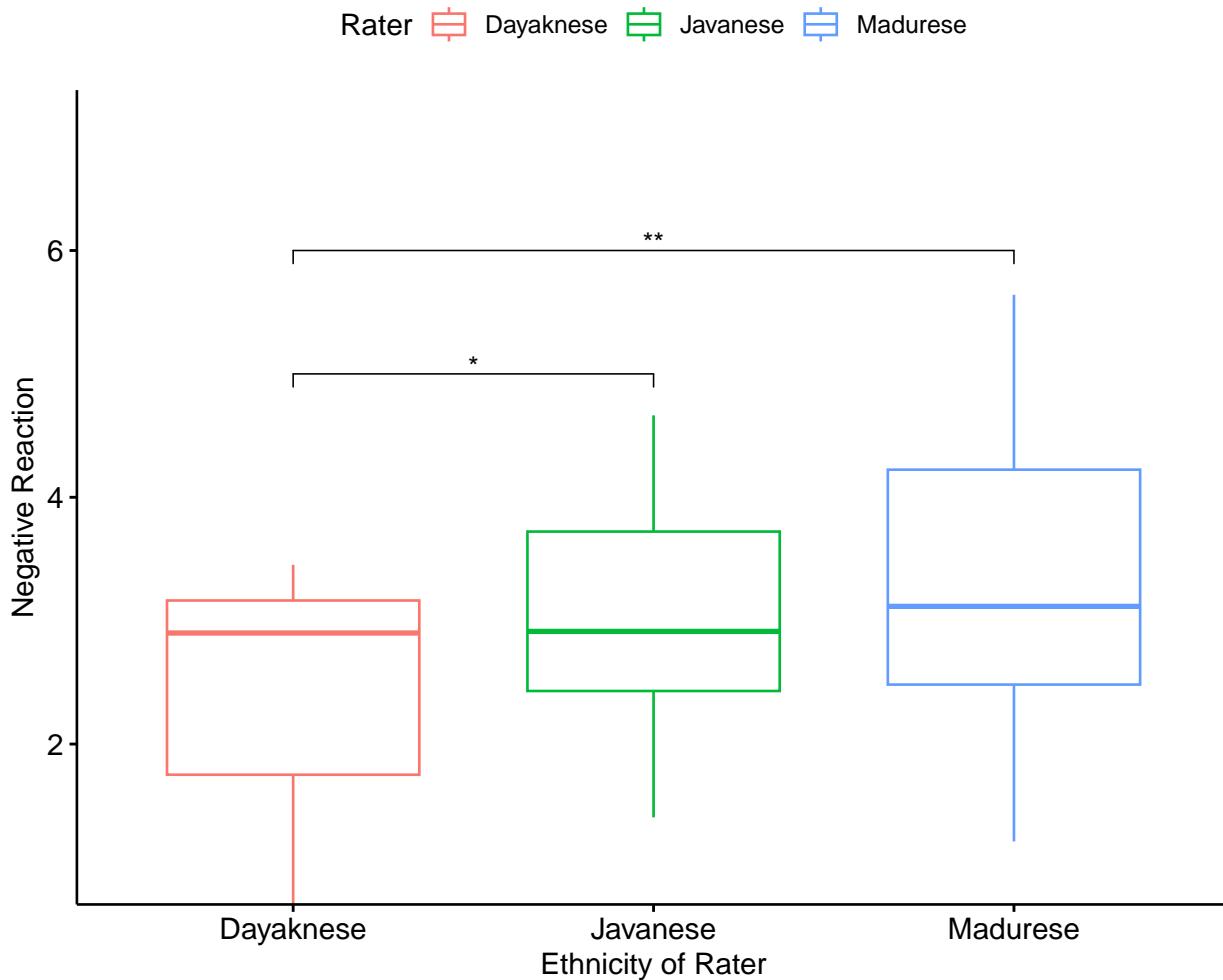
An APA style results section of this portion of follow-up might read like this:

We followed significant the rater main effect with a series of post hoc, pairwise comparisons. We controlled for Type I error with the traditional Bonferroni adjustment. Results suggested that there were statistically significant differences between the Dayaknese and Javanese ($M_{diff} = -0.515, p = 0.035$) and Dayaknese and Madurese ($M_{diff} = -0.807, p = < 0.003$) raters, but not Javanese and Madurese rater ($M_{diff} = -0.292, p = 0.642$). This analysis disregards the ethnic identity displayed on the photo.

Below is an augmentation of the figure that appeared at the beginning of the chapter. We can use the objects from the omnibus tests (named, “omnibus2w”) and post hoc pairwise comparisons (“RaterMain_ttest”) to add the ANOVA string and significance bars to the figure. Although they may not be appropriate in every circumstance, such detail can assist the figure in conveying maximal amounts of information.

```
RaterMain_ttest <- RaterMain_ttest %>%
  rstatix::add_xy_position(x = "Rater")
box_RaterMain + ggpubr::stat_pvalue_manual(RaterMain_ttest, label = "p.adj.signif",
  tip.length = 0.02, hide.ns = TRUE, y.position = c(5, 6))
```

Boxplots of Rater Main Effect



8.5.4.2 Option #2 planned orthogonal contrasts

We generally try for *orthogonal* contrasts so that the partitioning of variance is independent (clean, not overlapping). Planned contrasts are a great way to do this. Here's where we are in the workflow.

If you aren't extremely careful about your order-of-operations in R, it can confuse objects, so I have named these contrasts *c1* and *c2* to remind myself that they refer to the main effect of ethnicity of the rater.

In this hypothetical scenario (remember we are pretending we are in the circumstance of a non-significant interaction effect but a significant main effect), I am:

- comparing the DV for the Javanese rater to the combined Dayaknese and Madurese raters (c1).
- comparing the DV for the Dayaknese and Madurese raters (c2).

These are orthogonal because:

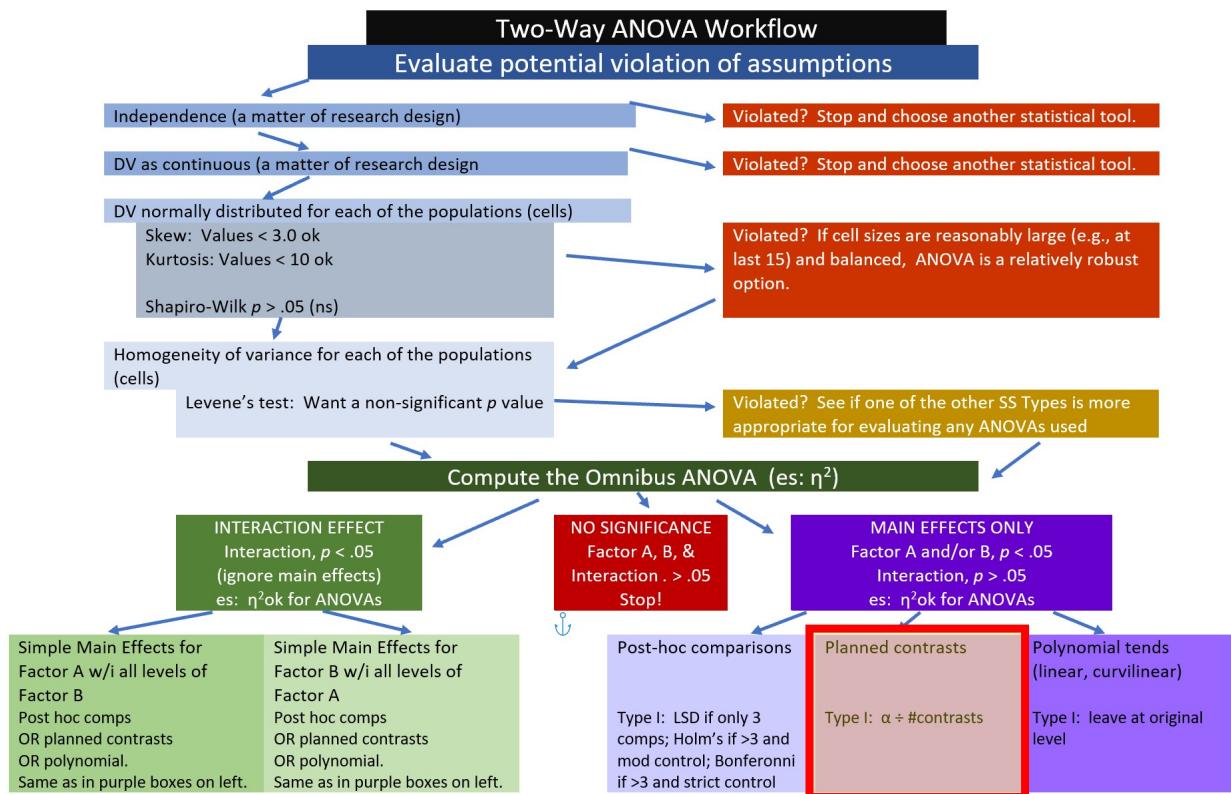


Figure 8.11: Image our place in the Two-Way ANOVA Workflow: Followup to a significant main effect with planned comparisons.

- there are $k - 1$ comparisons, and
- once a contrast is isolated (i.e., the Javanese rater in contrast #1) it cannot be used again
 - The “cake” analogy can be a useful mnemonic: once you take out a piece of the cake, you really can’t put it back in

I am not aware of *rstatix* functions or arguments that can complete these analyses. Therefore, we will use functions from base R. It helps to know what the default contrast codes are; we can get that information with the *contrasts()* function.

```
contrasts(Ramdhani_df$Rater)
```

	Javanese	Madurese
Dayaknese	0	0
Javanese	1	0
Madurese	0	1

Next, we set up the contrast conditions. In the code below,

- c1 indicates that the Javanese (noted as -2) are compared to the combined ratings from the Dayaknese (1) and Madurese (1)
- c2 indicates that the Dayaknese (-1) and Madurese (1) are compared; Javanese (0) is removed from the contrast.

```
# tell R which groups to compare
c1 <- c(1, -2, 1)
c2 <- c(-1, 0, 1)
mat <- cbind(c1, c2) #combine the above bits
contrasts(Ramdhani_df$Rater) <- mat # attach the contrasts to the variable
```

This allows us to recheck the contrasts.

```
contrasts(Ramdhani_df$Rater)
```

	c1	c2
Dayaknese	1	-1
Javanese	-2	0
Madurese	1	1

With this output we can confirm that, in contrast 1 (the first column) we are comparing the Javanese to the combined Dayaknese and Madurese. In contrast 2 (the second column) we are comparing the Dayaknese to the Madurese.

Then we run the contrast and extract the output.

```
mainPlanned <- aov(Negative ~ Rater, data = Ramdhani_df)
summary.lm(mainPlanned)
```

Call:

```
aov(formula = Negative ~ Rater, data = Ramdhani_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.08813	-0.74921	0.05792	0.71482	2.34187

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.93283	0.09259	31.676	< 0.0000000000000002 ***
Raterc1	-0.03712	0.06544	-0.567	0.571670
Raterc2	0.40342	0.11345	3.556	0.000561 ***

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	'	'	'	'
	'	'	'	'

Residual standard error: 0.9745 on 108 degrees of freedom

Multiple R-squared: 0.1063, Adjusted R-squared: 0.0898

F-statistic: 6.426 on 2 and 108 DF, p-value: 0.002307

```
contrasts(Ramdhani_df$Rater) <- cbind(c(1, -2, 1), c(-1, 0, 1))
```

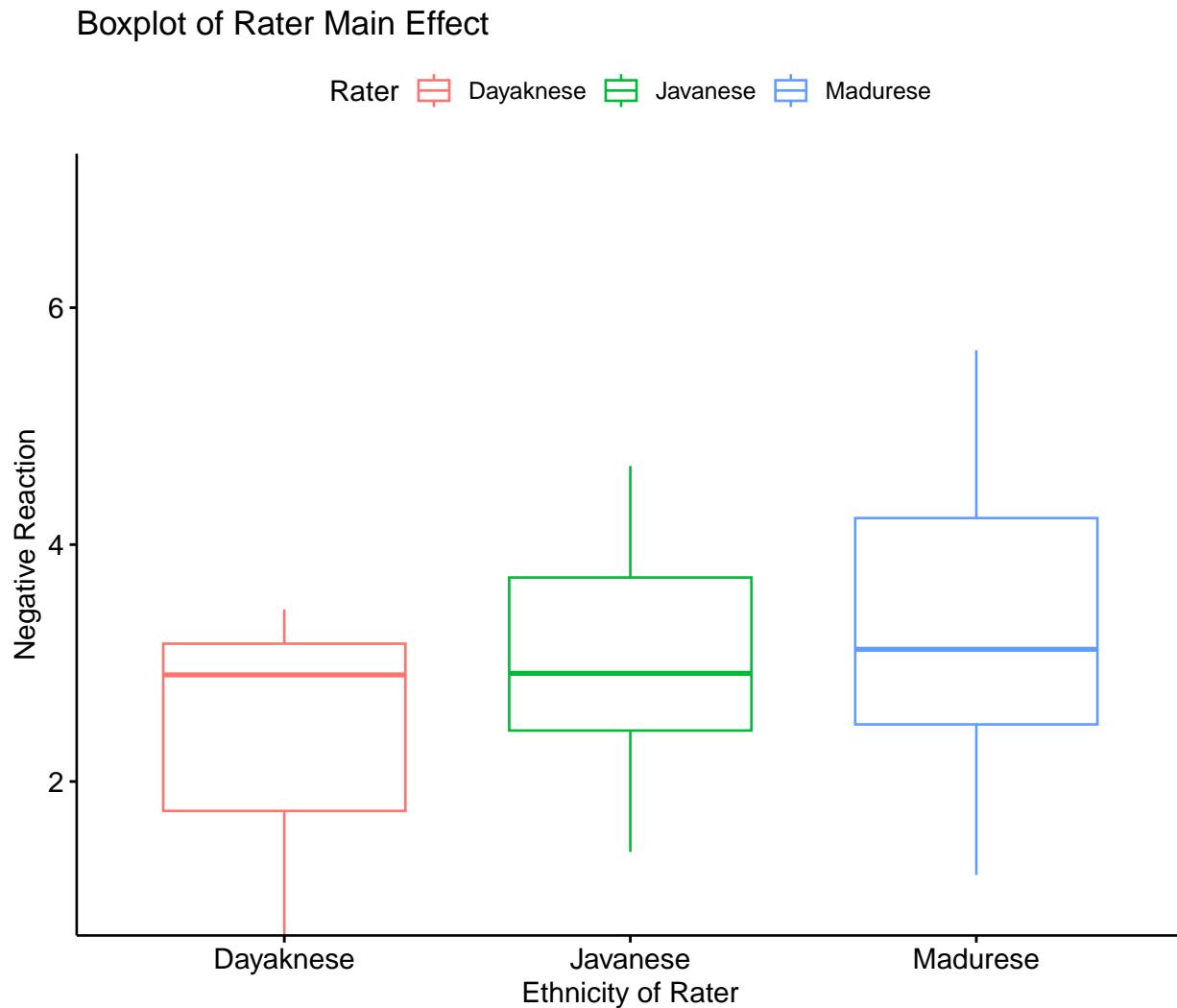
These planned contrasts show that when the Javanese raters are compared to the combined Dayaknese and Madurese raters, there was a non significant difference, $t(108) = -0.567, p = 0.572$. However, there were significant differences between Dayaknese and Javanese raters, $t(108) = 3.556, p < 0.001$.

An mini APA style reporting of these results might look like this:

We followed the significant rater main effect with a pair of planned, orthogonal, contrasts. The first compared Javanese raters to the combined Dayakneses and Madurese raters; there was a nonsignificant difference ($t[108] = -0.567, p = 0.572$). There was significant differences between Dayaknese and Javanese raters, $t(108) = 3.556, p < 0.001$.

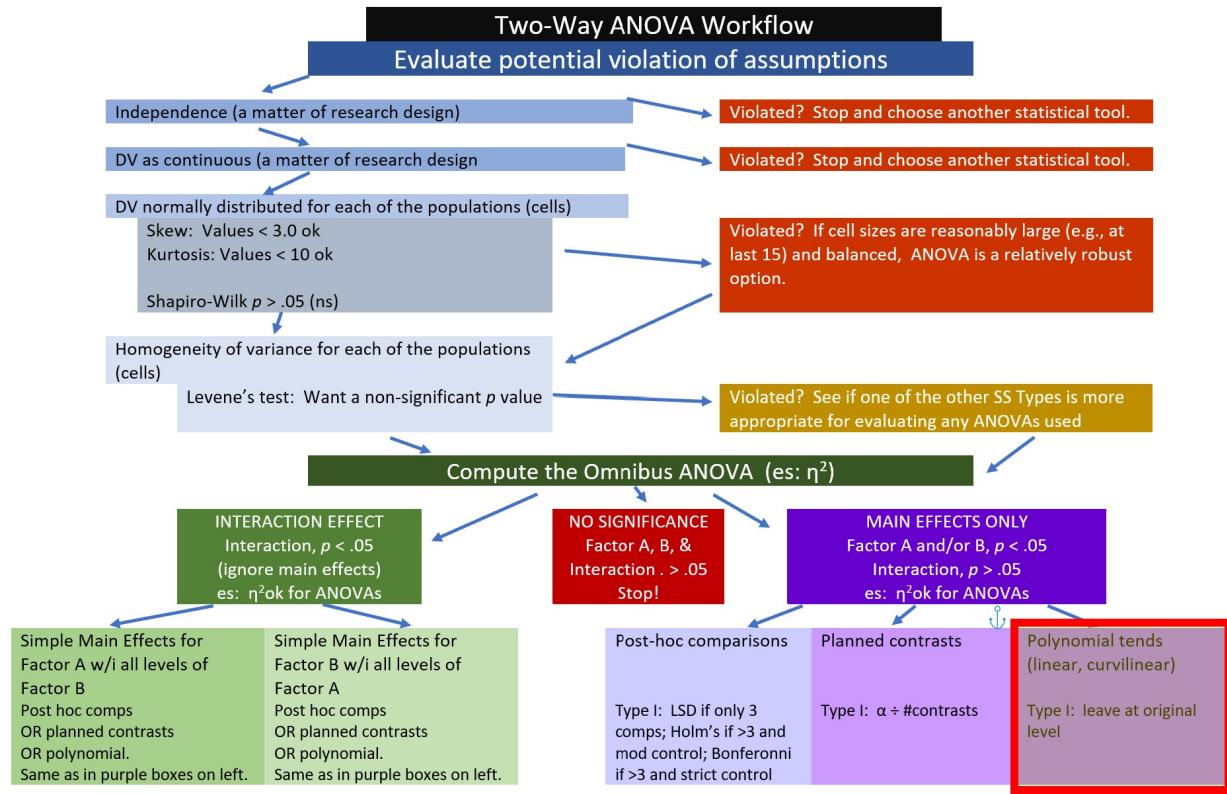
I am not aware of script that would effectively display this in a figure. Therefore, I would use a simple boxplot for the rater main effect.

```
box_RaterMain <- ggpubr::ggbboxplot(Ramdhani_df, x = "Rater", y = "Negative",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction", color = "Rater",
  ylim = c(1, 7), title = "Boxplot of Rater Main Effect")
box_RaterMain
```



8.5.4.3 Option #3 trend/polynomial analysis

Polynomial contrasts let us see if there is a linear (or curvilinear) pattern to the data. To detect a trend, the data must be coded in an ascending order...and it needs to be a sensible comparison. Here's where this would fall in our workflow.



Because these three ethnic groups are not *ordered* in the same way as would an experiment involving dosage (e.g., placebo, lo dose, hi dose), evaluation of the polynomial trend is not really justified (even though it is statistically possible). None-the-less, I will demonstrate how it is conducted.

The polynomial fits linear and curvilinear trends across levels of a factor based on how the variable is coded in R. The *contrasts()* function from base R will reveal this ordering. Not surprisingly, this is the same order seen in our boxplots. In terms of the “story” of the vignette, the authors suggest that the Dayaknese are typically viewed as the ones who were victimized, the Javanese were not involved, and the Madurese have been viewed as aggressors.

```
contrasts(Ramdhani_df$Rater)
```

	[,1]	[,2]
Dayaknese	1	-1
Javanese	-2	0
Madurese	1	1

Viewing the *contrasts()* output, we see that the trends (linear, quadratic) in our contrast coding will be fit across Dayaknese, Javanese, and Madurese.

In a polynomial analysis, the statistical analysis looks across the ordered means to see if they fit a linear or curvilinear shape that is one fewer than the number of levels (i.e., $k - 1$). Because the Rater factor has three levels, the polynomial contrast checks for linear (.L) and quadratic (one change in direction) trends (.Q). If we had four levels, *contr.poly()* could also check for cubic change (two changes in direction). Conventionally, when more than one trend is significant, we interpret the most complex one (i.e., quadratic over linear).

To the best of my knowledge, *rstatix* does not offer these contrasts. We can fairly easily make these calculations in base R by creating a set of polynomial contrasts. In the prior example we specified our contrasts through coding. Here we can the *contr.poly(3)* function. The “3” lets R know that there are three levels in Rater. The *aov()* function will automatically test for quadratic (one hump) and linear (straight line) trends.

```
contrasts(Ramdhani_df$Rater) <- contr.poly(3)
mainTrend <- aov(Negative ~ Rater, data = Ramdhani_df)
summary.lm(mainTrend)
```

Call:

```
aov(formula = Negative ~ Rater, data = Ramdhani_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.08813	-0.74921	0.05792	0.71482	2.34187

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.93283	0.09259	31.676 < 0.0000000000000002	***
Rater.L	0.57052	0.16045	3.556	0.000561 ***
Rater.Q	-0.09094	0.16029	-0.567	0.571670

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9745 on 108 degrees of freedom

Multiple R-squared: 0.1063, Adjusted R-squared: 0.0898

F-statistic: 6.426 on 2 and 108 DF, p-value: 0.002307

Rater.L tests the data to see if there is a significant linear trend. There is: $t(108) = 3.556, < 0.001$.

Rater.Q tests to see if there is a significant quadratic (curvilinear, one hump) trend. There is not: $t(108) = -0.567, p = .572$.

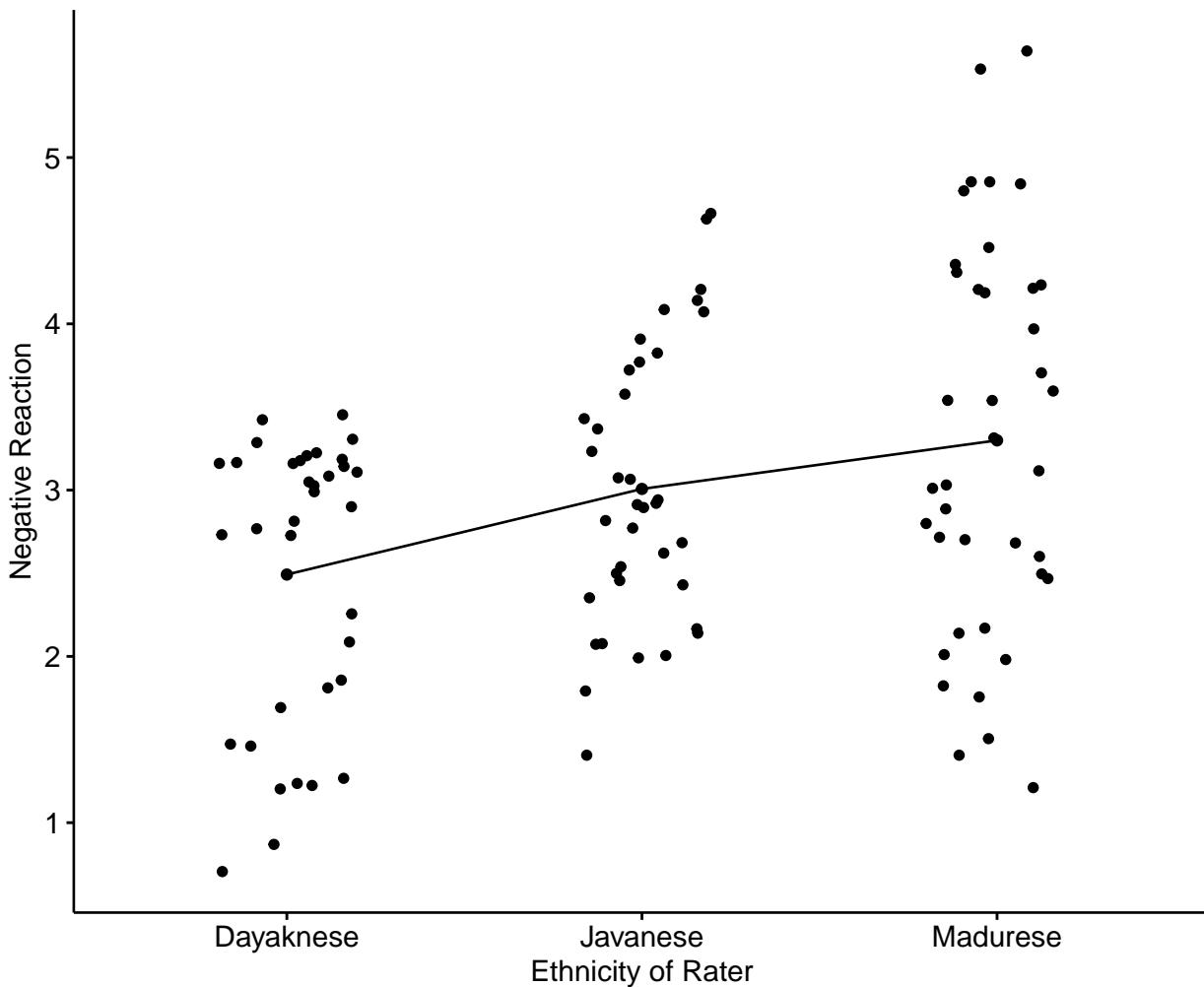
Here's how I might prepare a statement for inclusion in an write-up of APA style results:

Our follow-up to a significant main effect for Rater included a polynomial contrast. Results supported a significant linear trend ($t[108] = 3.556, p < .001$) such that negative reactions increased linearly across Dayaknese, Javanese, and Madurese raters.

A line plot might be a useful choice in conveying the linear trend.

```
ggpubr::gglne(Ramdhani_df, x = "Rater", y = "Negative", xlab = "Ethnicity of Rater",
  linetype = "solid", ylab = "Negative Reaction", add = c("mean_sd",
  "jitter"), title = "Linear Trend for Rater Main Effect")
```

Linear Trend for Rater Main Effect



```
# add this for a different color palette: palette = c('#00AFBB',
# '#E7B800')
```

8.6 APA Style Results

First, I am loathe to term anything “final.” In academia, there is *always* the possibility or revision. Given that I demonstrated a number of options in the workflow (with more in the [appendix](#)), let me first show the workflow with the particular path I took:

That is, I first tested the statistical assumptions and computed the omnibus ANOVA. Because there was a significant interaction effect, I followed with examination of the simple main effect of photo stimulus within ethnicity of the rater. It made sense to me to conduct the all post hoc pairwise comparisons within this simple main effect. In light of that, here’s how I might write it up:

A 3 X 2 ANOVA was conducted to evaluate the effects of rater ethnicity (3 levels, Dayaknese, Madurese, Javanese) and photo stimulus (2 levels, Dayaknese on Madurese,) on

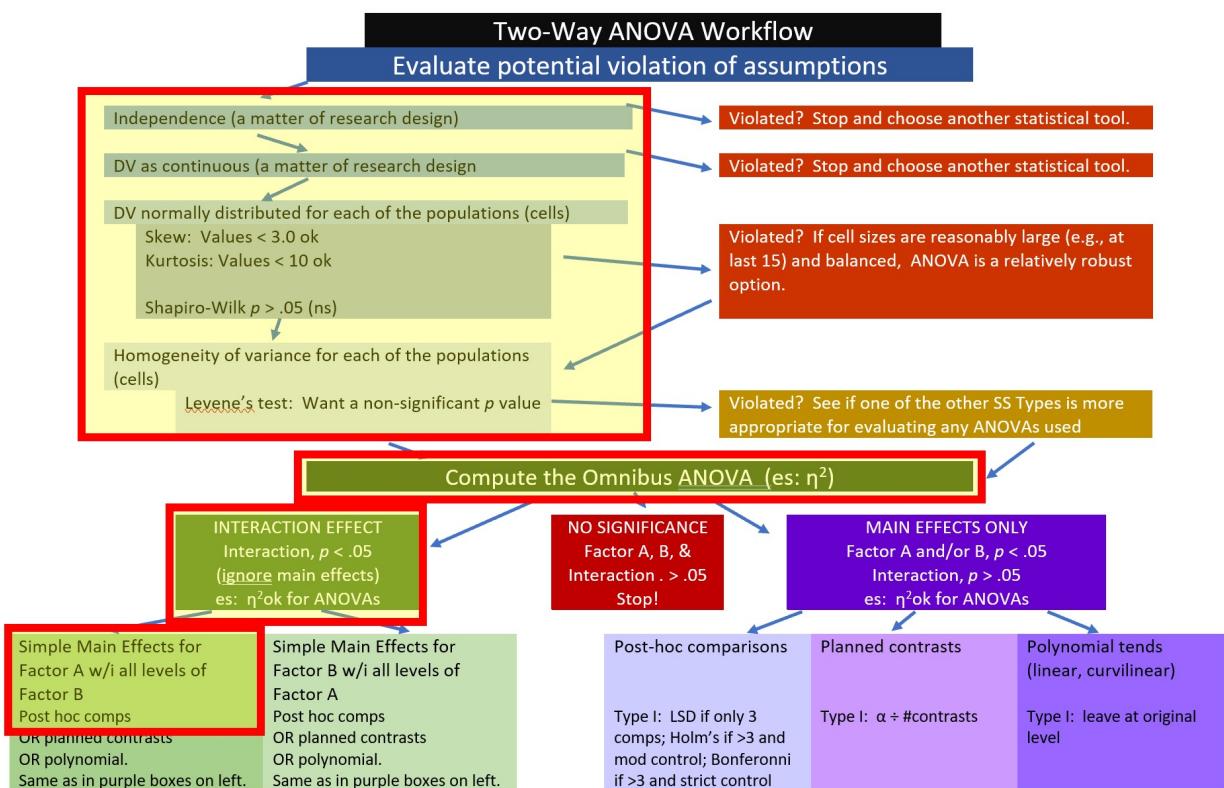


Figure 8.12: Image our place in the Two-Way ANOVA Workflow illustrating the path taken for this analysis.

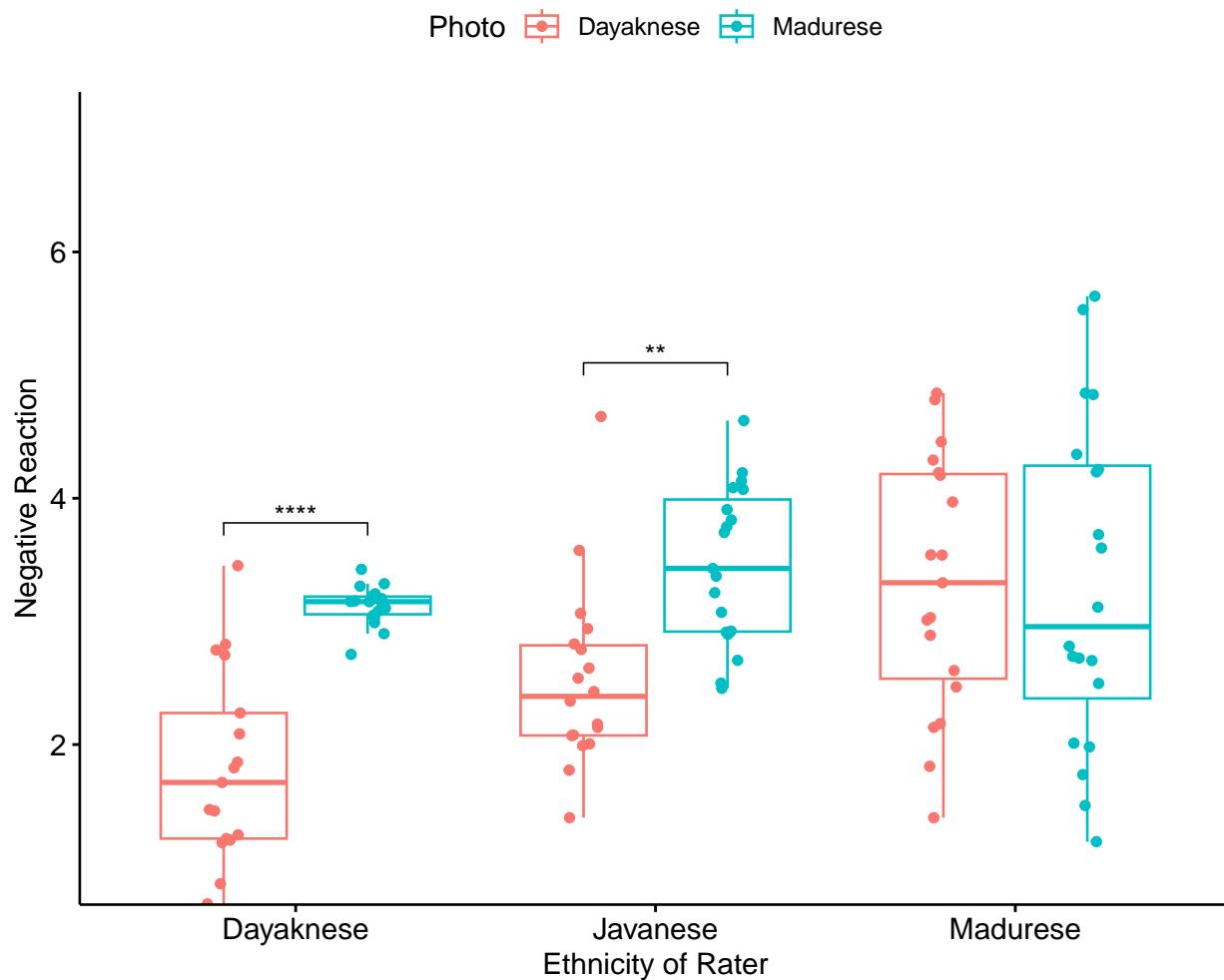
negative reactions to the photo stimuli. Factorial ANOVA assumes that the dependent variable is normally distributed for all cells in the design. Skew and kurtosis values for each factorial combinations fell below the guidelines recommended by Kline [2016a]. That is, they were below the absolute values of 3 for skew and 10 for kurtosis. Similarly, no extreme outliers were identified and results of the Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) suggested that model residuals did not differ significantly from a normal distribution ($W = 0.9846, p = 0.234$). Results of Levene's test for equality of error variances indicated a violation of the homogeneity of variance assumption, ($F[5, 105] = 8.834, p < 0.001$). Given that cell sample sizes were roughly equal and greater than 15, each; [Green and Salkind, 2017c] we proceeded with the two-way ANOVA.

Computing sums of squares with a Type II approach, the results for the ANOVA indicated a significant main effect for ethnicity of the rater ($F[2, 105] = 8.098, p < 0.001, \eta^2 = 0.134$), a significant main effect for photo stimulus, ($F[1, 105] = 19.346, p < 0.001, \eta^2 = 0.156$), and a significant interaction effect ($F[2, 105] = 5.696, p = .004, \eta^2 = 0.098$).

To explore the interaction effect, we followed with a test of the simple main effect of photo stimulus within the ethnicity of the rater. That is, with separate one-way ANOVAs (chosen, in part, to mitigate violation of the homogeneity of variance assumption [Kassambara, a]) we examined the effect of the photo stimulus within the Dayaknese, Madurese, and Javanese groups. To control for Type I error across the three simple main effects, we set alpha at .017 (.05/3). Results indicated significant differences for Dayaknese ($F[1, 33] = 50.404, p < 0.001, \eta^2 = 0.604$) and Javanese ethnic groups ($F[1, 35] = 17.183, p < 0.001, \eta^2 = 0.329$), but not for the Madurese ethnic group ($F[1, 37] < 0.001, p = .993, \eta^2 < .001$). As illustrated in Figure 1, the Dayaknese and Javanese raters both reported stronger negative reactions to the Madurese. The differences in ratings for the Madurese were not statistically significantly different. In this way, the rater's ethnic group moderated the relationship between the photo stimulus and negative reactions.

We can simply call the Figure we created before:

Simple Main Effect of Photo Stimulus within Rater



```
apaTables::apa.2way.table(iv1 = Rater, iv2 = Photo, dv = Negative, data = Ramdhani_df,
  landscape = TRUE, table.number = 1, filename = "Table_1_MeansSDs.doc")
```

Table 1

Means and standard deviations for Negative as a function of a 3(Rater) X 2(Photo) design

Rater	Photo			
	Dayaknese		Madurese	
	M	SD	M	SD
Dayaknese	1.82	0.77	3.13	0.16
Javanese	2.52	0.74	3.46	0.64
Madurese	3.30	1.03	3.30	1.33

Note. M and SD represent mean and standard deviation, respectively.

```
apaTables::apa.aov.table(TwoWay_neg, filename = "Table_2_effects.doc",
  table.number = 2, type = "II")
```

Table 2

ANOVA results using Negative as the dependent variable

Predictor	SS	df	MS	F	p	partial_eta2	CI_90_partial_eta2
Rater	12.24	2	6.12	8.10	.001	.13	
Photo	14.62	1	14.62	19.35	.000	.16	[.06, .26]
Rater x Photo	8.61	2	4.30	5.70	.004	.10	[.02, .18]
Error	79.34	105	0.76				

Note: Values in square brackets indicate the bounds of the 90% confidence interval for partial

While I have not located a package that will take *rstatix* output to make an APA style table with our pairwise comparisons, we can use the *MASS* package to write the pwc object to a .csv file, then manually make our own table.

```
MASS::write.matrix(pwPHwiETH, sep = ",", file = "pwPHwiETH.csv")
```

8.6.1 Comparing Our Results to Rhamdani et al. [2018]

As is common in simulations, our results approximate the findings reported in the manuscript, but does not replicate them exactly. Our main and interaction effects map on very closely. However, in the follow-up tests, while our findings that Dayaknese rated the Madurese photos more negatively, the findings related to the Javanese' and Madurese' ratings wiggled around some. Given the varying variability around each of the group means (i.e., and violation of the homogeneity of variance assumption) this makes sense to me. I find it to be a useful lesson in “what it takes” to get stable, meaningful results.

8.7 Options for Violation of Statistical Assumptions

Statistical assumptions are conditions that we should meet in order for the results of a particular statistical test to be valid. They are frequently focused on the trustworthiness of the *p* value. Some assumptions (e.g., dependency, random sampling) are specific to the research design. Others (e.g., normal distribution, homogeneity of variance) are ones that we evaluate with statistical tests. Thus, we are often asking, “What do we do if we violate one of these statistical assumptions?”

8.7.1 Violating the Assumption of Normality

Regarding the assumption of normal distribution within each of the cells (i.e., the combinations of the levels of the two factors in the design), Green and Salkind [2017c] provide some assurance that ANOVA is robust to violation of the normality assumption when there are at least 15 cases per cell and that the design is balanced (i.e., the number of cases per cell are roughly equal).

In the case of low power caused by low sample size or severely unbalanced designs, the research team may wish to consider extending the study to collect more data. Alternatively, although it is always difficult to remove cases, the researchers may inspect the data for outliers and see what happens if extreme outliers are removed or if the data is truncated at the extreme ends. Further, Kline [2016a] has helpful coverage for options regarding transforming data. A substantial concern about transformations relates to the interpretability of the results.

8.7.2 Violating the Homogeneity of Variance Assumption

Addressing violations of the homogeneity of variance assumption feel more tricky. In one-way ANOVA, the Welch's alternative was an easily accessible alternative that is robust to the homogeneity of variance assumption. The [WRS2 package](#) has been identified as a resource for ANOVA designs with statistical methods based on Wilcox' WRS functions that are robust to these statistical violations.

One potential alternative is to change the sums of squares type used in the ANOVA calculations. In ANOVA models sums of squares can be calculated four different ways: Types I, II, III, and IV.

SS Type II is the `aov()` default. Because `rstatix()` is a wrapper for `aov()` it is similarly the default for `rstatix()`. I find it to be a best practice to include the `type =` argument in my code so that I am reminded of the need to make this choice.

Type I sums of squares is similar to hierarchical linear regression in that the first predictor in the model claims as much variance as it can and the leftovers are claimed by the variable entered next – each claiming as much as possible leaving the leftovers for what follows. Unless the variables are completely independent of each other (unlikely), Type I sums of squares cannot evaluate the true main effect of each variable. Type I should not be used to evaluate main effects and interactions because the order of predictors will affect the results.

Type II (the default in the package we used) is appropriate if you are interested main effects because it ignores the effect of any interactions involving the main effect. Thus, variance from a main effect is not “lost” to any interaction terms containing that effect. Type II is appropriate for main effects analyses only, but should not be used when evaluating interaction effects. Type II sums of squares is not affected by the type of contrast coding used to specify the predictor variables.

Type III is the default in many stats packages – but not the R packages we used. In Type III all effects (main effects and interactions) are evaluated (simultaneously) taking into consideration all other effects in the model (not just the ones entered before). Type III is more robust to unequal samples sizes (e.g., unbalanced designs). Type III is best when predictors are encoded with orthogonal contrasts.

Type IV is identical to Type III except it requires no missing cells. In `rstatix::anova_test*`, this type is not available.

Researchers appear to disagree about which sums of squares type to use. Certainly, when package and program developers specify a default. The `rstatix::anova_test` that we used (which is a wrapper for the `aov()` in base R) has set Type II as the default. In contrast, Field [2012] suggested that it is safest to stick with Type III sums of squares. For more information, check out this explanation on [r-bloggers](#).

In this lesson I stuck with the `rstatix::anova_test()` default because

- Type II sums of squares was used in hand-calculations,
 - Our example was reasonably balanced (equal cell sizes), and
 - We had only violated the homogeneity of variance assumption.

For demonstration purposes, let's run the Type III alternative to see the differences:

```
rstatix::anova_test(Ramdhani_df, Negative ~ Rater * Photo, type = "3",  
detailed = TRUE)
```

ANOVA Table (type III tests)

For comparison, this was our earlier analysis:

```
rstatix::anova_test(Ramdhani_df, Negative ~ Rater * Photo, type = "2",  
  detailed = TRUE)
```

ANOVA Table (type II tests)

	Effect	SSn	SSd	DFn	DFd	F	p	p<.05	ges
1	Rater	12.238	79.341	2	105	8.098	0.0005360	*	0.134
2	Photo	14.619	79.341	1	105	19.346	0.0000262	*	0.156
3	Rater:Photo	8.609	79.341	2	105	5.696	0.0040000	*	0.098

Note that the sums of squares are somewhat different between models – and that the Type III results includes an intercept. In today’s example, the statistical significance remains the same across the models.

Unfortunately, violations of the assumption of homogeneity variance impact choices at both the omnibus and follow-up levels of analysis [Green and Salkind, 2017c]. Kassambara [a] noted that when the homogeneity of variance assumption has been violated, it is better to follow-up a significant omnibus with separate one-way ANOVAs because these offer separate and unique error terms. One operational advantage to this is option is that researchers can return to procedures for one-way ANOVA, assessing for violations at that level and using the Welch's alternative for the follow-up of simple main effects or main effects.

8.8 Power Analysis

Asking about *power* can be a euphemistic way of asking, “How large should my sample size be?”

Power is defined as the ability of the test to detect statistical significance when there is such. It's represented formulaically as $(1 - P)$ (Type II error). Power is traditionally set at 80% (or .8)

We will do both – evaluate the power of our current example and then work backwards to estimate the sample size needed.

We'll use the *pwr.2way()* function from the *pwr2* package. Helpful resources are found here:

- <https://cran.r-project.org/web/packages/pwr2/pwr2.pdf>
- <https://rdrr.io/cran/pwr2/man/ss.2way.html>

The *pwr.2way()* and *ss.2way()* functions require the following:

- **a** number of groups in Factor A
- **b** number of groups in Factor B
- **alpha** significant level (Type I error probability)
- **beta** Type II error probability (Power = 1 - beta; traditionally set at .1 or .2)
- **f.A** the *f* effect size of Factor A
- **f.B** the *f* effect size of Factor B
- **B** Iteration times, default is 100

Hints for calculating the *f.A* and *f.B* values:

- In this case, we will rerun the statistic, grab both effect sizes, and convert them to the *f* (not the *f²*)
 - calculation can be straightforward, either use an online calculator, a hand-calculated formula, or the *eta2_to_f* function from the *effectsize*
- When an effect size is unknown, you can substitute what you expect using Cohen's guidelines of .10, .25, and .40 as small, medium, and large (for the *f*, not *F²*)

Let's quickly rerun our model to get both the df and calculate the *f* effect value

```
rstatix::anova_test(Ramdhani_df, Negative ~ Rater * Photo, type = "2",
detailed = TRUE)
```

ANOVA Table (type II tests)

	Effect	SSn	SSd	DFn	DFd	F	p	p<.05	ges
1	Rater	12.238	79.341	2	105	8.098	0.0005360	*	0.134
2	Photo	14.619	79.341	1	105	19.346	0.0000262	*	0.156
3	Rater:Photo	8.609	79.341	2	105	5.696	0.0040000	*	0.098

If we want to understand power in our analysis, we need to convert our effect size (η^2) for the *interaction* to *f* effect size (this is not the same as the *F* test). The *effectsize* package has a series of converters. We can use the *eta2_to_f()* function.

```
effectsize::eta2_to_f(0.134) #FactorA -- Rater
```

[1] 0.393363

```
effectsize::eta2_to_f(0.156) #Factor B -- Photo
```

[1] 0.4299234

8.8.1 Post Hoc Power Analysis

Now we calculate power for our existing model. We'll use the package *pwr2* and the function *pwr.2way()*. To specify this we identify:

- a: number of groups for Factor A (Rater)
- b: number of groups for Factor B (Photo)
- size.A: sample size per group in Factor A (because ours differ slightly, I divided the N by the number of groups)
- size.B: sample size per group in Factor B (because ours differ slightly, I divided the N by the number of groups)
- f.A: Effect size of Factor A
- f.A.: Effect size of Factor B

```
pwr2::pwr.2way(a = 3, b = 2, alpha = 0.05, size.A = 37, size.B = 55, f.A = 0.3935,
f.B = 0.43)
```

Balanced two-way analysis of variance power calculation

```
a = 3
b = 2
```

```

n.A = 37
n.B = 55
sig.level = 0.05
power.A = 0.9997716
power.B = 1
power = 0.9997716

```

NOTE: power is the minimum power among two factors

At 0.9998 (Rater), 1.0000 (Photo), and 0.9998 (interaction), our power to detect a significant effect for Factor A/Rater and Factor B/Photo was huge!

8.8.2 Estimating Sample Size Requirements

If we want to replicate this study we could use its results to estimate what would be needed for the replication.

In this specification:

- a: number of groups for Factor A (Rater)
- b: number of groups for Factor B (Photo)
- alpha: significance level (Type I error probability); usually .05
- beta: Type II error probability (Power = 1-beta); usually .80
- f.A: Effect size (*f*) of Factor A (this time we know; other times we can guess from previously published literature)
- f.A.: Effect size (*f*) of Factor B
- B: iteration times, default number is 100

```
pwr2::ss.2way(a = 3, b = 2, alpha = 0.05, beta = 0.8, f.A = 0.3935, f.B = 0.43,
B = 100)
```

Balanced two-way analysis of variance sample size adjustment

```

a = 3
b = 2
sig.level = 0.05
power = 0.2
n = 3

```

NOTE: n is number in each group, total sample = 18

Curiously, 18 was just about the number that was in each of the six cells!

Often times researchers will play around with the *f* values. Remember Cohen's indication of small (.10), medium (.25), and large (.40). Let's see what happens when we enter different values. Specifically, what if we only had a medium effect?

```
pwr2::ss.2way(a = 3, b = 2, alpha = 0.05, beta = 0.8, f.A = 0.25, f.B = 0.25,
B = 100) #if we expected a medium effect
```

Balanced two-way analysis of variance sample size adjustment

```
a = 3
b = 2
sig.level = 0.05
power = 0.2
n = 6
```

NOTE: n is number in each group, total sample = 36

And what would happen if we only had a small effect?

```
pwr2::ss.2way(a = 3, b = 2, alpha = 0.05, beta = 0.8, f.A = 0.1, f.B = 0.1,
B = 100) #if we expected a small effect
```

Balanced two-way analysis of variance sample size adjustment

```
a = 3
b = 2
sig.level = 0.05
power = 0.2
n = 30
```

NOTE: n is number in each group, total sample = 180

8.9 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “doable” and then try at least one more problem that challenges you in some way. At a minimum your data should allow for a 2 X 3 (or 3 X 2) design. At least one of the problems you work should have a statistically significant interaction effect that you work all the way through.

Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Whichever you choose, you will focus on these larger steps in factorial-way ANOVA, including:

- test the statistical assumptions
- conduct a two-way ANOVA, including
 - omnibus test and effect size

- report main and interaction effects
- conduct follow-up testing of simple main effects
- write a results section to include a figure and tables

8.9.1 Problem #1: Play around with this simulation.

Copy the script for the simulation and then change (at least) one thing in the simulation to see how it impacts the results.

- If two-way ANOVA is new to you, perhaps you just change the number in “set.seed(210731)” from 210731 to something else. Your results should parallel those obtained in the lecture, making it easier for you to check your work as you go.
- If you are interested in power, change the sample size to something larger or smaller.
- If you are interested in variability (i.e., the homogeneity of variance assumption), perhaps you change the standard deviations in a way that violates the assumption.

8.9.2 Problem #2: Conduct a factorial ANOVA with the *positive evaluation* dependent variable.

The Ramdhani et al. [2018] article has two dependent variables (negative and positive evaluation). Each is suitable for two-way ANOVA. I used *negative evaluation* as the dependent variable; you are welcome to conduct the analysis with *positive evaluation* as the dependent variable.

8.9.3 Problem #3: Try something entirely new.

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a two-way, factorial ANOVA. Please have at least 3 levels for one predictor and at least 2 levels for the second predictor.

8.9.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the IV and DV. Minimally, the data should allow the analysis of a 2 x 3 (or 3 X 2) design. At least one of the problems you work should have a significant interaction effect so that follow-up is required.	5	_____
2. Simulate (or import) and format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct omnibus ANOVA (w effect size)	5	_____

Assignment Component	Points Possible	Points Earned
5. Conduct one set of follow-up tests; narrate your choice	5	_____
6. Describe approach for managing Type I error	5	_____
7. APA style results with table(s) and figure	5	_____
8 Explanation to grader	5	_____
Totals	40	_____

Chapter 9

One-Way Repeated Measures ANOVA

[Screencasted Lecture Link](#)

In the prior lessons, a critical assumption is that the observations must be “independent.” That is, related people (partners, parent/child, manager/employee) cannot comprise the data and there cannot be multiple waves of data for the same person. Repeated measures ANOVA is created specifically for this *dependent* purpose. This lesson focuses on the one-way repeated measures ANOVA, where we measure changes across time.

9.1 Navigating this Lesson

There is just over one hour of lecture. If you work through the materials with me plan for an additional two hours

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER’s [introduction](#)

9.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Evaluate the suitability of a research design/question and dataset for conducting a one-way repeated measures ANOVA; identify alternatives if the data is not suitable.
- Hand-calculate a one-way repeated measures ANOVAs
 - describing the partitioning of variance as it relates to model/residual; within/between.
- Test the assumptions for one-way repeated measures ANOVA.
- Conduct a one-way repeated measures ANOVA (omnibus and follow-up) in R.

- Interpret output from the one-way repeated measures ANOVA (and follow-up).
- Prepare an APA style results section of the one-way repeated measures ANOVA output.
- Demonstrate how an increased sample size increases the power of a statistical test.

9.1.2 Planning for Practice

The suggestions for homework vary in degree of challenge with more complete descriptions at the end of the chapter follow these suggestions.

- Rework the problem in the chapter by changing the random seed in the code that simulates the data. This should provide minor changes to the data, but the results will likely be very similar.
- There were no additional variables in this example. However, you'll see we do have an issue with statistical power. Perhaps change the sample size to see if it changes (maybe stabilizes?) the results.
- Conduct a one-way repeated measures ANOVA with data to which you have access. This could include data you simulate on your own or from a published article.

9.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- *Repeated Measures ANOVA in R: The Ultimate Guide.* (n.d.). Datanovia. Retrieved October 19, 2020, from <https://www.datanovia.com/en/lessons/repeated-measures-anova-in-r>
 - This website is an excellent guide for both one-way repeated measures and mixed design ANOVA. A great resource for both the conceptual and procedural. This is the guide I have used for the basis of the lecture. Working through their example would be great additional practice.
- Green, S. B., & Salkind, N. J. (2017). One-Way Repeated Measures Analysis of Variance (Lesson 29). In *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (Eighth edition., pp. 209–217). Pearson.
 - For years I taught from the Green and Salkind text. Even though it was written for SPSS, the authors do a terrific job of walking the reader through the one-way repeated measures logic and process.
- Amodeo, A. L., Picariello, S., Valerio, P., & Scandurra, C. (2018). Empowering transgender youths: Promoting resilience through a group training program. *Journal of Gay & Lesbian Mental Health*, 22(1), 3–19.
 - This mixed methods (qualitative and quantitative) includes a one-way repeated measures example.

9.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(knitr)){install.packages('knitr')}
# if(!require(tidyverse)){install.packages('tidyverse')} #manipulate
# data if(!require(psych)){install.packages('psych')}
# if(!require(ggpubr)){install.packages('ggpubr')} #easy plots
# if(!require(rstatix)){install.packages('rstatix')} #pipe-friendly R
# functions if(!require(data.table)){install.packages('data.table')}
# #pipe-friendly R functions
# if(!require(reshape2)){install.packages('reshape2')} #pipe-friendly
# R functions
# if(!require(effectsize)){install.packages('effectsize')} #converts
# effect sizes for use in power analysis
# if(!require(WebPower)){install.packages('WebPower')} #power
# analysis tools for repeated measures
# if(!require(MASS)){install.packages('MASS')} #power analysis tools
# for repeated measures
```

9.2 Introducing One-way Repeated Measures ANOVA

There are a couple of typical use cases for one-way repeated measures ANOVA. In the first, the research participant is assessed in multiple conditions – with no interested in change-over-time.

An example of a research design using this approach occurred in the Green and Salkind [2017b] statistics text, the one-way repeated measures vignette compared teachers' perception of stress when responding to parents, teachers, and school administrators.



Figure 9.1: Illustration of a research design appropriate for one-way repeated measures ANOVA

Another common use case is about time. The classic design is a pre-test, an intervention, a post-test, and a follow up. In designs like these researchers often hope that there is a positive change from pre-to-post and that that change either stays constant (from post-to-follow-up) or, perhaps, increases even further. The research vignette for this lesson is interested in change-over-time.



Figure 9.2: Illustration of a research design appropriate for one-way repeated measures ANOVA

9.2.1 Workflow for Oneway Repeated Measures ANOVA

The following is a proposed workflow for conducting a one-way repeated measures ANOVA.

Steps involved in analyzing the data include:

1. Preparing and importing the data.
2. Exploring the data
 - graphs
 - descriptive statistics
3. Checking distributional assumptions
 - assessing normality via skew, kurtosis, Shapiro Wilks
 - checking or violation of the *sphericity* assumption with Mauchly's test; if violated interpret the corrected output or use a multivariate approach for the analysis
4. Computing the omnibus ANOVA
5. Computing post hoc comparisons, planned contrasts, or polynomial trends
6. Managing Type I error
7. Sample size/power analysis (which you should think about first – but in the context of teaching ANOVA, it's more pedagogically sensible, here)

9.3 Research Vignette

Amodeo [Amodeo et al., 2018] and colleagues conducted a mixed methods study (qualitative and quantitative) to evaluate the effectiveness of an empowerment, peer-group-based, intervention with participants ($N = 8$) who experienced transphobic episodes. Focus groups used qualitative methods to summarize emergent themes from the program (identity affirmation, self-acceptance, group as support) and a one-way, repeated measures ANOVA provided evidence of increased resilience from pre to three-month followup.

Eight participants (seven transgender women and one genderqueer person) participated in the intervention. The mean age was 28.5 ($SD = 5.85$). All participants were located in Italy.

The within-subjects condition was wave, represented by T1, T2, and T3:

- T1, beginning of training
- Training, three 8-hour days,
 - content included identity and heterosexism, sociopolitical issues and minority stress, resilience, and empowerment
- T2, at the conclusion of the 3-day training
- Follow-up session 3 months later
- T3, at the conclusion of the +3 month follow-up session

The dependent variable (assessed at each wave) was a 14-item resilience scale [Wagnild and Young, 1993]. Items were assessed on a 7-point scale ranging from *strongly disagree* to *strongly agree* with higher scores indicating higher levels of resilience. An example items was, “I usually manage one way or another.”

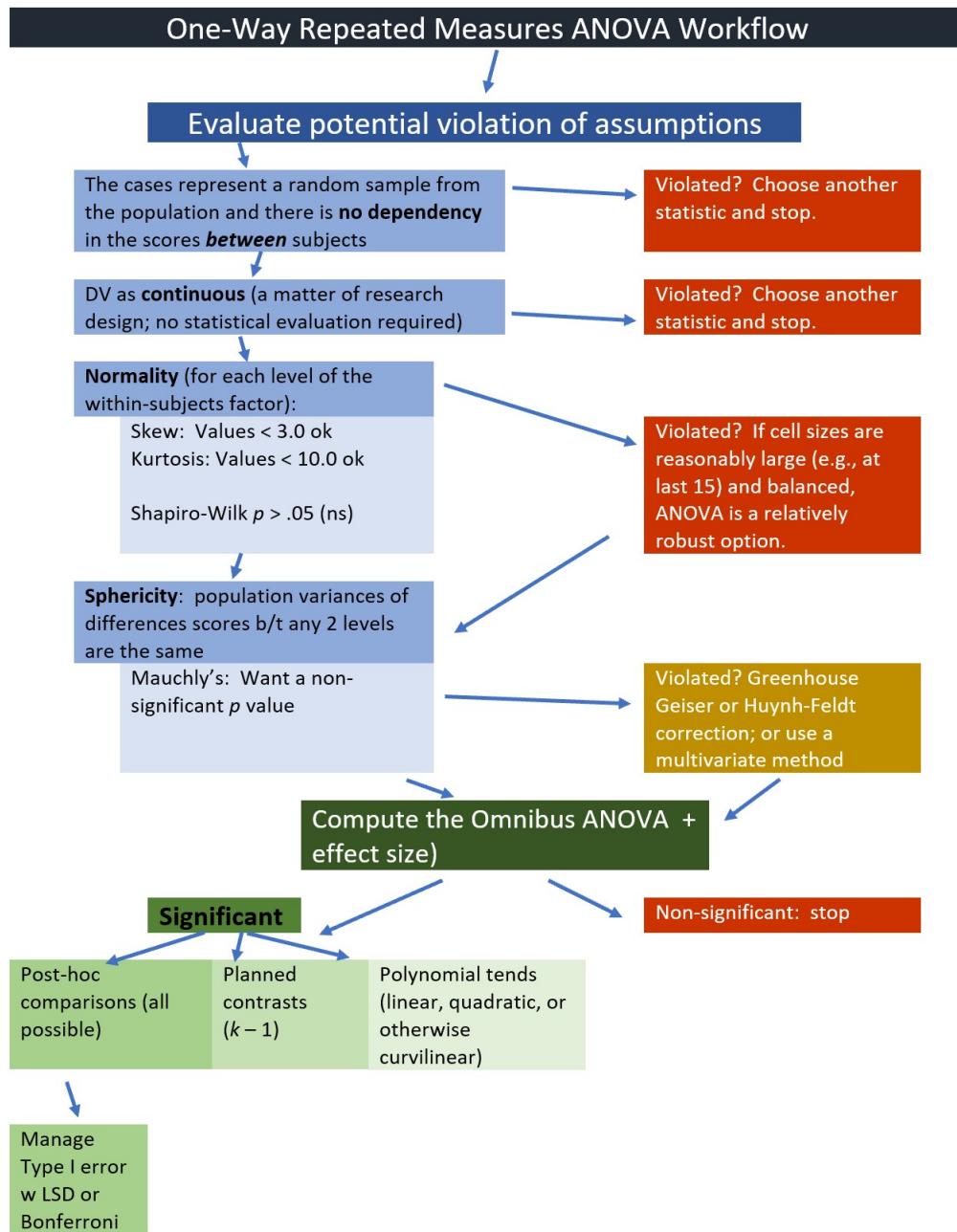


Figure 9.3: Image of a workflow for the one-way repeated measures ANOVA



Figure 9.4: Diagram of the research design for the Amodeo et al study

9.3.1 Data Simulation

Below is the code I used to simulate data. The following code assumes 8 participants who each participated in 3 waves (pre, post, followup).

```
set.seed(2022)
# gives me 8 numbers, assigning each number 3 consecutive spots, in
# sequence
ID <- factor(rep(seq(1, 8), each = 3))
# gives me a column of 24 numbers with the specified Ms and SD
Resilience <- rnorm(24, mean = c(5.7, 6.21, 6.26), sd = c(0.88, 0.79, 0.37))
# repeats pre, post, follow-up once each, 8 times
Wave <- rep(c("Pre", "Post", "FollowUp"), each = 1, 8)
Amodeo_long <- data.frame(ID, Wave, Resilience)
```

Let's take a look at the structure of our variables. We want ID to be a factor, Resilience to be numeric, and Wave to be an ordered factor (Pre, Post, FollowUp).

```
str(Amodeo_long)
```

```
'data.frame': 24 obs. of 3 variables:
 $ ID       : Factor w/ 8 levels "1","2","3","4",...: 1 1 1 2 2 2 3 3 3 4 ...
 $ Wave     : chr  "Pre" "Post" "FollowUp" "Pre" ...
 $ Resilience: num  6.49 5.28 5.93 4.43 5.95 ...
```

We need to update Wave to be an ordered factor. Because R's default is to order factors alphabetically, we can use the levels command and identify our preferred order.

```
Amodeo_long$Wave <- factor(Amodeo_long$Wave, levels = c("Pre", "Post",
"FollowUp"))
```

We check the structure again.

```
str(Amodeo_long)
```

```
'data.frame': 24 obs. of 3 variables:
 $ ID       : Factor w/ 8 levels "1","2","3","4",...: 1 1 1 2 2 2 3 3 3 4 ...
 $ Wave     : Factor w/ 3 levels "Pre","Post","FollowUp": 1 2 3 1 2 3 1 2 3 1 ...
 $ Resilience: num  6.49 5.28 5.93 4.43 5.95 ...
```

9.3.1.0.1 Shape Shifters The form of our data matters. The simulation created a *long* form (formally called the *person-period* form) of data. That is, each observation for each person is listed in its own row. In this dataset where we have 8 people with 3 observation (pre, post, follow-up) each, we have 24 rows. This is convenient, because this is the form we need for repeated measures ANOVA.

However, for some of the calculations (particularly those we will do by hand), we need the data to be in its more familiar wide form (formally called the *person level* form). We can do this with the *data.table* and *reshape2*()* packages.

```
# Create a new df (Amodeo_wide) Identify the original df We are
# telling it to connect the values of the Resilience variable its
# respective Wave designation
Amodeo_wide <- reshape2::dcast(data = Amodeo_long, formula = ID ~ Wave,
  value.var = "Resilience")
# doublecheck to see if they did what you think
str(Amodeo_wide)
```

```
'data.frame': 8 obs. of 4 variables:
$ ID      : Factor w/ 8 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8
$ Pre     : num  6.49 4.43 4.77 5.91 4.84 ...
$ Post    : num  5.28 5.95 6.43 7 6.28 ...
$ FollowUp: num  5.93 5.19 6.54 6.19 6.24 ...
```

```
Amodeo_wide$ID <- factor(Amodeo_wide$ID)
```

In this reshape script, I asked for a quick structure check. The format of the variables looks correct. If you want to export these data as files to your computer, remove the hashtags to save (and re-import) them as .rds (R object) or .csv (“Excel lite”) files. This is not a necessary step to continue working the problem in this lesson.

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R. I would choose this option.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Amodeo_long, 'Amodeo_longRDS.rds')
# saveRDS(Amodeo_wide, 'Amodeo_wideRDS.rds') bring back the simulated
# dat from an .rds file Amodeo_long <- readRDS('Amodeo_longRDS.rds')
# Amodeo_wide <- readRDS('Amodeo_wideRDS.rds')
```

Another option is to write them as .csv files. The code for .csv will likely lose any variable formatting, but the .csv file is easy to view and manipulate in Excel. If you choose this option, you will probably need to re-run the prior code to reformat Wave as an ordered factor

```
# write the simulated data as a .csv write.table(Amodeo_long,
# file='Amodeo_longCSV.csv', sep=',', col.names=TRUE,
# row.names=FALSE) write.table(Amodeo_wide,
# file='Amodeo_wideCSV.csv', sep=',', col.names=TRUE,
# row.names=FALSE) bring back the simulated dat from a .csv file
# Amodeo_long <- read.csv ('Amodeo_longCSV.csv', header = TRUE)
# Amodeo_wide <- read.csv ('Amodeo_wideCSV.csv', header = TRUE)
```

9.3.2 Quick peek at the data

Before we get into the statistic let's inspect our data. As we work the problem we will switch between long and wide formats. We can start with the long form.

```
str(Amdeo_long)
```

```
'data.frame': 24 obs. of 3 variables:  
 $ ID       : Factor w/ 8 levels "1","2","3","4",...: 1 1 1 2 2 2 3 3 3 4 ...  
 $ Wave     : Factor w/ 3 levels "Pre","Post","FollowUp": 1 2 3 1 2 3 1 2 3 1 ...  
 $ Resilience: num 6.49 5.28 5.93 4.43 5.95 ...
```

In the following output, note the order of presentation of the grouping variable (i.e., FollowUp, Post, Pre). Even though we have ordered our factor so that “Pre” is first, the *describeBy()* function seems to be ordering them alphabetically.

```
psych::describeBy(Amodeo_long$Resilience, Wave, mat = TRUE, data = Amodeo_long,  
    digits = 3)
```

```
# Note. Recently my students and I have been having intermittent  
# struggles with the describeBy function in the psych package. We  
# have noticed that it is problematic when using .rds files and when  
# using data directly imported from Qualtrics. If you are having  
# similar difficulties, try uploading the .csv file and making the  
# appropriate formatting changes.
```

Another view (if we use the wide file).

```
psych::describe(Amodeo_wide)
```

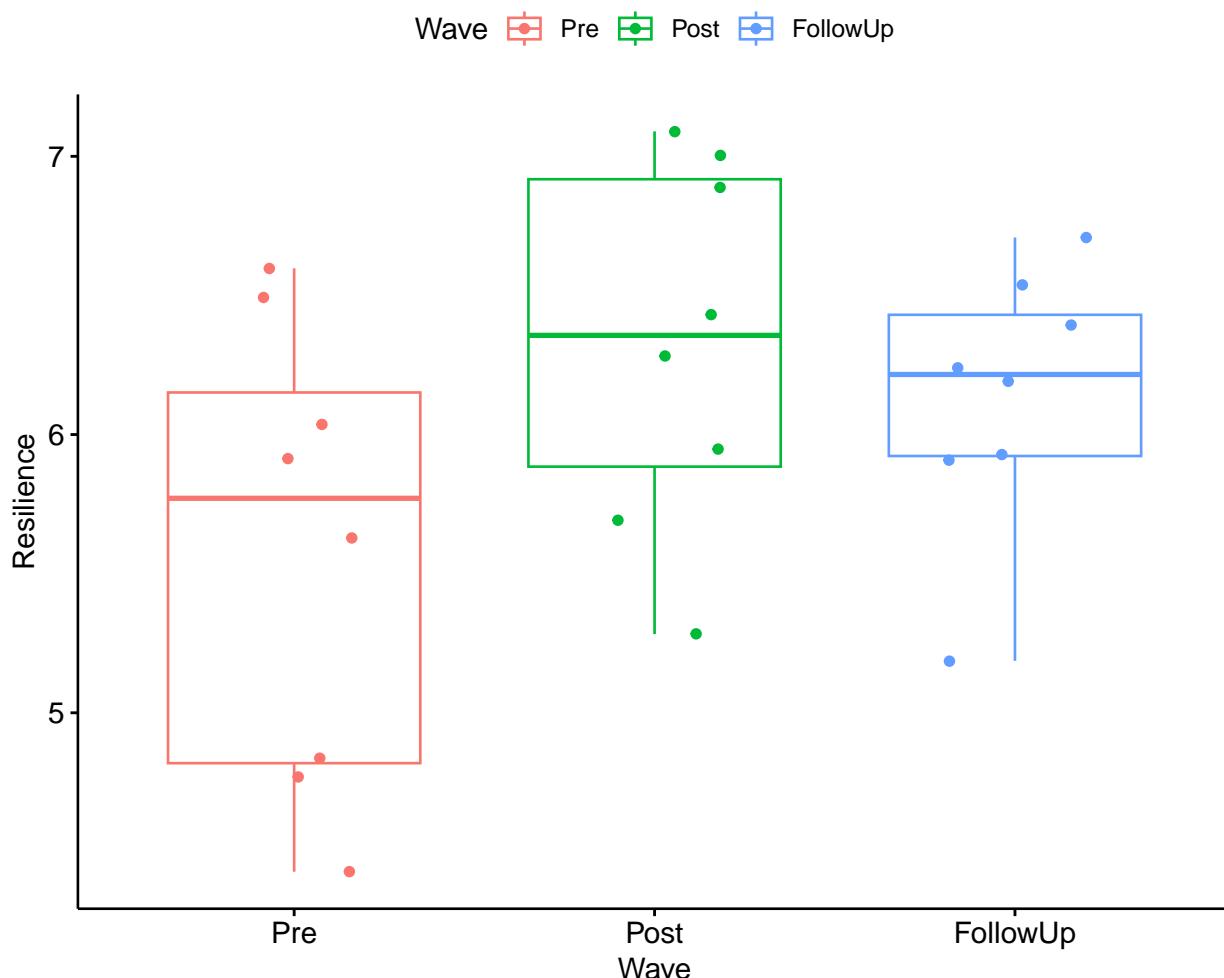
ID*	0.87
Pre	0.29
Post	0.23
FollowUp	0.17

Our means suggest that resilience increases from pre to post, then declines a bit. We use one-way repeated measures ANOVA to learn if there are statistically significant differences between the pairs of means and over time.

Let's also take a quick look at a boxplot of our data.

```
ggpubr::ggbboxplot(Amodeo_long, x = "Wave", y = "Resilience", add = "jitter",
  color = "Wave", title = "Figure 9.1 Boxplots of Resilience Over Time")
```

Figure 9.1 Boxplots of Resilience Over Time



9.4 Working the One-Way Repeated Measures ANOVA (by hand)

Before working our problem in R, let's gain a conceptual understanding by partitioning the variance by hand.

In repeated measures ANOVA: $SS_T = SS_B + SS_W$, where

- B = between-subjects variance
 - W = within-subjects variance
- $SS_W = SS_M + SS_R$

What differs is that SS_M and SS_R (model and residual) are located in SS_W

- $SS_T = SS_B + (SS_M + SS_R)$

		SS Within = 6.64 df (N - k) = 16		
Total	Model df	Residual	Between	
df formula	#cells-1	#levels-1	df _W - df _M	#people-1
SS	11.66	2.36	4.27	5.03
df	23	2	14	7
MS = SS/df		1.18	0.305	
F = MS _M /MS _R =		3.87		
*where N is number of cells				
F critical value = 3.73				
Because F > F _{cv} , we can reject the null hypothesis: F (2, 14) = 3.87, p < .05				

Figure 9.5: Demonstration of partitioning variance

9.4.1 Sums of Squares Total

Our formulas for SS_T are the same as they were for one-way and factorial ANOVA:

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

$$SS_T = s_{grand}^2(n - 1)$$

Degrees of freedom for SS_T is $N - 1$, where N is the total number of cells.

Let's take a moment to *hand-calculate* SS_T (but using R).

Our grand (i.e., overall) mean is

```
mean(Amdeo_long$Resilience)
```

```
[1] 6.017408
```

Subtracting the grand mean from each resilience score yields a mean difference.

```
library(tidyverse)

Amodeo_long <- Amodeo_long %>%
  mutate(m_dev = Resilience - mean(Resilience))

head(Amodeo_long)
```

	ID	Wave	Resilience	m_dev
1	1	Pre	6.492125	0.47471697
2	1	Post	5.283057	-0.73435114
3	1	FollowUp	5.927930	-0.08947756
4	2	Pre	4.428839	-1.58856921
5	2	Post	5.948499	-0.06890871
6	2	FollowUp	5.186767	-0.83064071

Pop quiz: What's the sum of our new *m_dev* variable?

```
sum(Amodeo_long$m_dev)
```

```
[1] 0.00000000000007993606
```

If we square those mean deviations:

```
Amodeo_long <- Amodeo_long %>%
  mutate(m_devSQ = m_dev^2)

head(Amodeo_long)
```

	ID	Wave	Resilience	m_dev	m_devSQ
1	1	Pre	6.492125	0.47471697	0.225356199
2	1	Post	5.283057	-0.73435114	0.539271599
3	1	FollowUp	5.927930	-0.08947756	0.008006235
4	2	Pre	4.428839	-1.58856921	2.523552145
5	2	Post	5.948499	-0.06890871	0.004748410
6	2	FollowUp	5.186767	-0.83064071	0.689963983

If we sum the squared mean deviations:

```
sum(Amodeo_long$m_devSQ)
```

```
[1] 11.65769
```

This value, the sum of squared deviations around the grand mean, is our SS_T ; the associated *degrees of freedom* is 23 ($24 - 1$; $N - 1$).

9.4.2 Sums of Squares Within for Repeated Measures ANOVA

The format of the formula is parallel to the formulae for SS we have seen before. In this case each person serves as its own grouping factor.

$$SS_W = s_{person1}^2(n_1 - 1) + s_{person2}^2(n_2 - 1) + s_{person3}^2(n_3 - 1) + \dots + s_{personk}^2(n_k - 1)$$

The degrees of freedom (df) within is $N - k$; or the summation of the df for each of the persons.

```
psych::describeBy(Resilience ~ ID, data = Amodeo_long, mat = TRUE, digits = 3)
```

	item	group1	vars	n	mean	sd	median	trimmed	mad	min	max
Resilience1	1	1	1 3	5.901	0.605	5.928	5.901	0.836	5.283	6.492	
Resilience2	2	2	1 3	5.188	0.760	5.187	5.188	1.124	4.429	5.948	
Resilience3	3	3	1 3	5.912	0.992	6.430	5.912	0.160	4.768	6.537	
Resilience4	4	4	1 3	6.370	0.568	6.191	6.370	0.414	5.913	7.005	
Resilience5	5	5	1 3	5.787	0.824	6.240	5.787	0.064	4.836	6.283	
Resilience6	6	6	1 3	5.744	0.146	5.693	5.744	0.095	5.629	5.908	
Resilience7	7	7	1 3	6.627	0.248	6.597	6.627	0.300	6.395	6.889	
Resilience8	8	8	1 3	6.612	0.533	6.708	6.612	0.565	6.038	7.090	
	range	skew	kurtosis	se							
Resilience1	1.209	-0.044	-2.333	0.349							
Resilience2	1.520	0.002	-2.333	0.439							
Resilience3	1.769	-0.380	-2.333	0.573							
Resilience4	1.092	0.283	-2.333	0.328							
Resilience5	1.447	-0.384	-2.333	0.475							
Resilience6	0.279	0.304	-2.333	0.084							
Resilience7	0.494	0.118	-2.333	0.143							
Resilience8	1.052	-0.175	-2.333	0.307							

With 8 people, there will be 8 chunks of the analysis, in each:

- SD squared (to get the variance)
- multiplied by the number of observations less 1

```
(0.605^2 * (3 - 1)) + (0.76^2 * (3 - 1)) + (0.992^2 * (3 - 1)) + (0.568^2 *
(3 - 1)) + (0.824^2 * (3 - 1)) + (0.146^2 * (3 - 1)) + (0.248^2 * (3 -
1)) + (0.553^2 * (3 - 1))
```

```
[1] 6.635836
```

9.4.3 Sums of Squares Model – Effect of Time

The SS_M conceptualizes the within-persons (or repeated measures) element as the grouping factor. In our case these are the pre, post, and follow-up clusters.

$$SS_M = \sum n_k (\bar{x}_k - \bar{x}_{grand})^2$$

The degrees of freedom will be $k - 1$ (number of levels, minus one).

```
psych::describe(Amdeo_wide)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
ID*		1	8	4.50	2.45	4.50	4.50	2.97	1.00	8.00	7.00	0.00
Pre		2	8	5.59	0.82	5.77	5.59	1.15	4.43	6.60	2.17	-0.14
Post		3	8	6.33	0.66	6.36	6.33	0.88	5.28	7.09	1.81	-0.23
FollowUp		4	8	6.14	0.47	6.22	6.14	0.44	5.19	6.71	1.52	-0.72
	se											
ID*			0.87									
Pre			0.29									
Post			0.23									
FollowUp			0.17									

In this case, we are interested in change in resilience over time. Hence, *time* is our factor. In our equation, we have three chunks representing the pre, post, and follow-up *conditions* (waves). From each, we subtract the grand mean, square it, and multiply by the n observed in each wave.

The degrees of freedom (df) for SS_M is $k - 1$

Let's calculate grand mean; that is the resilience score for all participants across all waves.

```
mean(Amdeo_long$Resilience)
```

```
[1] 6.017408
```

Now we can calculate the SS_M .

```
(8 * (6.14 - 6.017)^2) + (8 * (6.33 - 6.017)^2) + (8 * (5.59 - 6.017)^2)
```

```
[1] 2.363416
```

```
# df is 3-1 = 2
```

9.4.4 Sums of Squares Residual

Because $SS_W = SS_M + SS_R$ we can calculate SS_R with simple subtraction:

- $SS_w = 6.636$
- $SS_M = 2.363$

```
6.636 - 2.363
```

```
[1] 4.273
```

Correspondingly, the degrees of freedom (also taking the easy way out) is calculated by subtracting (the associated degrees of freedom) SS_M from SS_W .

```
16-2
```

```
[1] 14
```

9.4.5 Sums of Squares Between

The SS_B is not used in our calculations today, but also calculated easily. Given that $SS_T = SS_W + SS_B$:

- $SS_T = 11.66; df = 23$
- $SS_W = 6.64; df = 16$

```
11.66 - 6.64
```

```
[1] 5.02
```

```
23-16
```

```
[1] 7
```

$SS_B = 5.02, df = 7$

		SS Within = 6.64 $df (N-k) = 16$		
	Total	Model df	Residual	Between
df formula	#cells-1	#levels-1	$df_W - df_M$	#people-1
SS	11.66	2.36	4.27	5.03
df	23	2	14	7
$MS = SS/df$		1.18	0.305	
$F = MS_M/MS_R =$		3.87		
*where N is number of cells				
F critical value = 3.73				
Because $F > F_{cv}$, we can reject the null hypothesis: $F (2, 14) = 3.87, p < .05$				

Figure 9.6: Screenshot of the ANOVA source Table

Looking again at our sourcetable, we can move through the steps to calculate our F statistic.

9.4.6 Mean Squares Model & Residual

Now that we have the Sums of Squares, we can calculate the mean squares (we need these for our F statistic). Here is the formula for the mean square model.

$$MS_M = \frac{SS_M}{df_M}$$

```
#mean squares for the model  
2.36/2
```

[1] 1.18

Here is the formula for mean square residual.

And $MS_R =$

$$MS_R = \frac{SS_R}{df_R}$$

Recall, degrees of freedom for the residual is $N - k$. In our case that is 90 - 3.

```
#mean squares for the residual  
4.27 / 14
```

[1] 0.305

9.4.7 F ratio

The F ratio is calculated with MS_M and $MS_R =$.

$$F = \frac{MS_M}{MS_R}$$

```
1.18 / .305
```

[1] 3.868852

To find the F_{CV} we can use an [F distribution table](#). Or use a look-up function in R, which follows this general form: `qf(p, df1, df2, lower.tail=FALSE)`

```
qf(.05, 2, 14, lower.tail=FALSE)
```

[1] 3.738892

Our example has 2 (numerator) and 14 (denominator) degrees of freedom. If we use a table we find the corresponding degrees of freedom combinations for the column where $\alpha = .05$. We observe that any F value > 3.73 will be statistically significant. Our $F = 3.87$, so we have (just barely) exceeded the threshold. This is our *omnibus F*. We know there is at least 1 statistically significant difference between our pre, post, and follow-up conditions.

9.5 Working the One-Way Repeated Measures ANOVA with R packages

As usual, we will work through the testing of statistical assumptions, calculating the omnibus, and then (if the omnibus is significant), conducting follow-up tests.

9.5.1 Testing the assumptions

We will start by testing the assumptions. Highlighting in the figure notes our place in the one-way ANOVA workflow:

There are several different ways to conduct a repeated measures ANOVA. Each has different assumptions/requirements. These include:

- univariate statistics
 - This is what we will use today.
- multivariate statistics
 - Functionally similar to univariate, except the underlying algorithm does not require the sphericity assumption.
 - An example of using a multivariate approach to working the problem (using the *car* package) is in the [appendix](#).
- multi-level modeling/hierarchical linear modeling
 - This a different statistic altogether and is addressed in the [multilevel modeling OER](#).

9.5.1.1 Univariate assumptions for repeated measures ANOVA

- The cases represent a random sample from the population.
- There is no dependency in the scores *between* participants.
 - Of course there is intentional dependency in the repeated measures (or within-subjects) factor.
- There are no significant outliers in any cell of the design
 - Check by visualizing the data using box plots. The *identify_outliers()* function in the *rstatix* package identifies extreme outliers.
- The dependent variable is normally distributed in the population for each level of the within-subjects factor.
 - Conduct a Shapiro-Wilk test of normality for each of the levels of the DV.
 - Visually examine Q-Q plots.
- The population variance of difference scores computed between any two levels of a within-subjects factor is the same value regardless of which two levels are chosen; termed the **sphericity assumption**. This assumption is

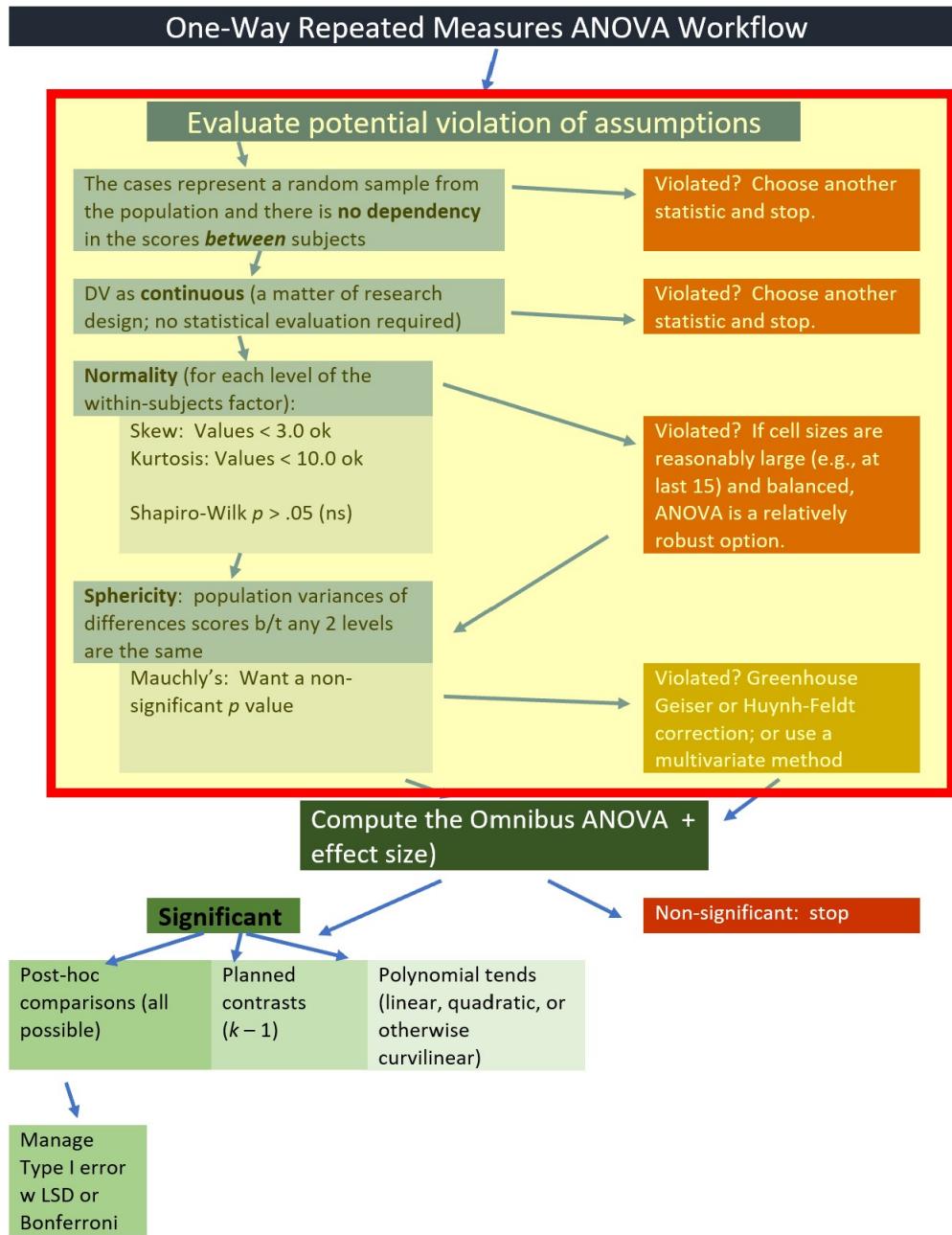


Figure 9.7: Image of our position in the workflow for the one-way repeated measures ANOVA

- akin to compound symmetry (both variances across conditions are equal).
- akin to the homogeneity of variance assumption in between-group designs.
- sometimes called the homogeneity-of-variance-of-differences assumption.
- statistically evaluated with *Mauchly's test*. If Mauchly's $p < .05$, there are statistically significant differences. The *anova_test()* function in the *rstatix* package reports Mauchly's and two alternatives to the traditional F that correct the values by the degree to which the sphericity assumption is violated.

9.5.1.2 Demonstrating sphericity

Using the data from our motivating example, I calculated differences for each of the time variables. These are the three columns (in green shading) on the right. The variance for each is reported at the bottom of the column.

When we get into the analysis, we will use *Mauchly's test* in hopes that there are non-significant differences in variances between all three of the comparisons.

We are only concerned with the sphericity assumption if there are three or more groups.

$$\text{variance}_{A-B} \approx \text{variance}_{A-C} \approx \text{variance}_{B-C}$$

ID	Pre	Post	FollowUp	Pre-Pos	Pre-Fup	Pos-Fup
1	6.49	5.28	5.93	1.21	0.56	-0.64
2	4.43	5.95	5.19	-1.52	-0.76	0.76
3	4.77	6.43	6.54	-1.66	-1.77	-0.11
4	5.91	7.00	6.19	-1.09	-0.28	0.81
5	4.84	6.28	6.24	-1.45	-1.40	0.04
6	5.63	5.69	5.91	-0.06	-0.28	-0.21
7	6.60	6.89	6.39	-0.29	0.20	0.49
8	6.04	7.09	6.71	-1.05	-0.67	0.38
			Variance	0.95	0.60	0.26

Figure 9.8: Demonstration of unequal variances

9.5.1.3 Is the data normally distributed?

We can obtain skew and kurtosis values for each cell in our model with the *psych::describeBy()* function.

```
psych::describeBy(Resilience ~ Wave, mat = TRUE, type = 1, data = Amodeo_long)
```

	item	group1	vars	n	mean	sd	median	trimmed	mad
Resilience1	1	Pre	1 8	5.587693	0.8217561	5.770952	5.587693	1.1471137	
Resilience2	2	Post	1 8	6.327615	0.6550520	6.356491	6.327615	0.8751431	

```
Resilience3    3 FollowUp    1 8 6.136916 0.4729432 6.215983 6.136916 0.4416578
               min      max   range     skew kurtosis       se
Resilience1 4.428839 6.597214 2.168376 -0.1755752 -1.448137 0.2905347
Resilience2 5.283057 7.089591 1.806534 -0.2819094 -1.209000 0.2315959
Resilience3 5.186767 6.708259 1.521491 -0.8802629  0.121247 0.1672107
```

Our skew and kurtosis values fall below the thresholds of concern [Kline, 2016a]:

- $< |3|$ for skew
- $< |10|$ for kurtosis

The *Shapiro-Wilk* test evaluates the hypothesis that the distribution of the data deviates from a comparable normal distribution. If the test is non-significant ($p >.05$) the distribution of the sample is not significantly different from a normal distribution. If, however, the test is significant ($p < .05$), then the sample distribution is significantly different from a normal distribution. The *rstatix* package can conduct this test for us.

```
Amodeo_long %>%
  group_by(Wave) %>%
  rstatix::shapiro_test(Resilience)
```

```
# A tibble: 3 x 4
  Wave     variable   statistic     p
  <fct>   <chr>        <dbl> <dbl>
1 Pre      Resilience  0.919  0.419
2 Post     Resilience  0.941  0.617
3 FollowUp Resilience  0.926  0.480
```

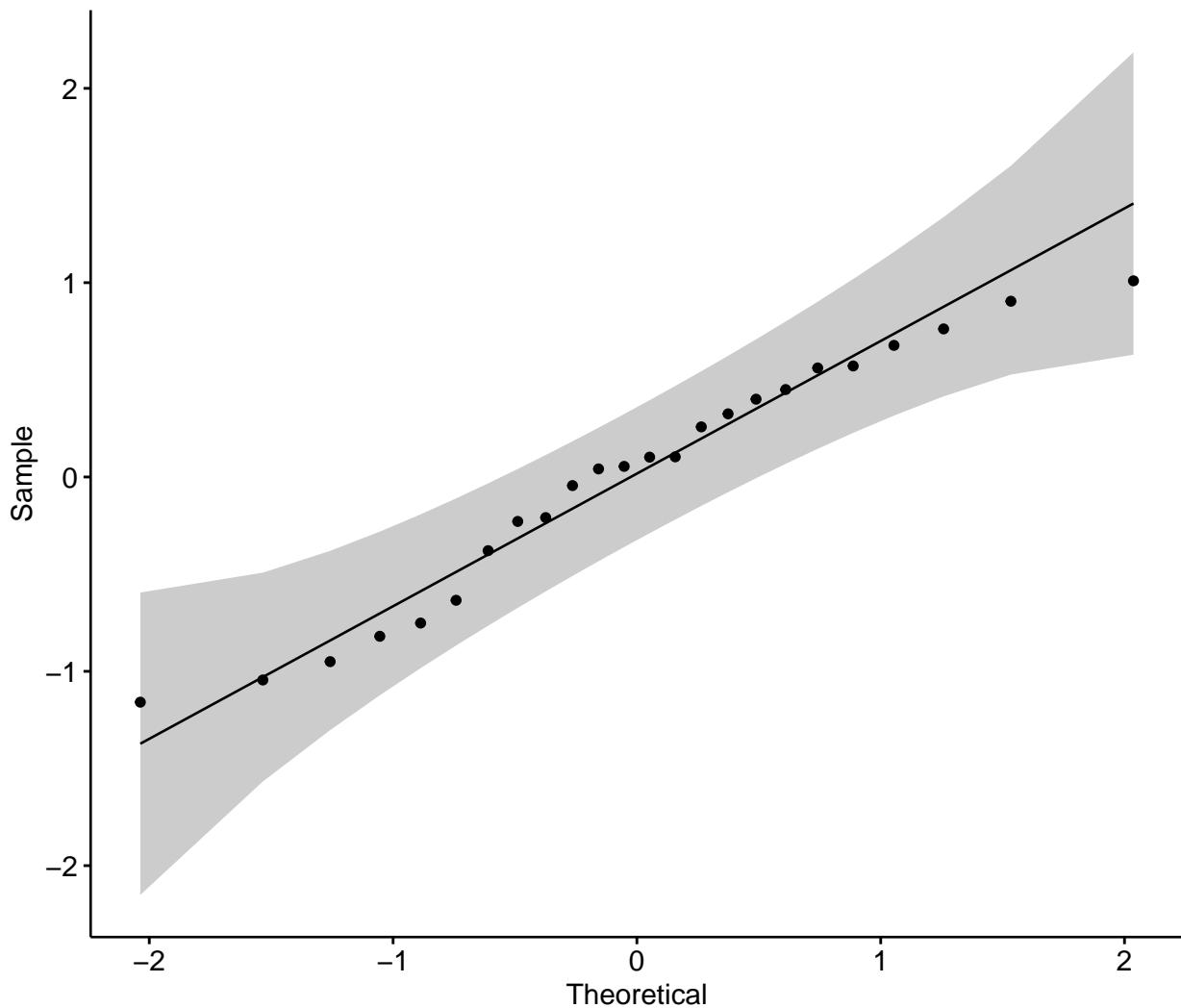
The p value is $> .05$ for each of the cells. This provides some assurance that we have not violated the assumption of normality at any level of the design.

The p values for the distributions of the dependent variable (Resilience) in each wave of the study are all well above .05. This tells us that the Resilience variable does not deviate from a statistically significant distribution at any level (Pre, $W = 0.929, p = 0.418$; Post, $p = 0.941, p = 0.617$; FellowUp, $W = 0.926, p = 0.430$).

Especially in the more simple “ANOVA’s” I like this form of the Shapiro-Wilk test because it makes it clear that we expect normality within each of the grouping levels. This approach, however, is only appropriate when there are a low number of levels/groupings and there are many data points per group. As models become more complex, researchers should use the model-based option for assessing normality. To do this, we first create an object that tests our research model.

Although that model (a regression model) has information about the results of our primary analysis, at this point we are only using it to investigate the assumption of normality. One product of the analysis is *residuals*. Residuals are the unexplained variance in the outcome (or dependent) variable after accounting for the predictor (or independent) variable. When we plot these “leftovers” against the values of x , we can visualize the fit of the model in a QQ plot. The dots represent the residuals. When they are relatively close to the line they not only suggest good fit of the model, but we know they are small and evenly distributed around zero (i.e., normally distributed).

```
RMres_model <- lm(Resilience ~ Wave, data = Amodeo_long)
ggpubr::ggqqplot(residuals(RMres_model))
```



We can also use the model in a Shapiro-Wilk test. As before, we want a non-significant result.

```
rstatix::shapiro_test(residuals(RMres_model))
```

```
# A tibble: 1 x 3
  variable      statistic p.value
  <chr>          <dbl>    <dbl>
1 residuals(RMres_model) 0.957    0.385
```

These results are consistent with what we have already learned. That is, the non-significant p value associated with the model-based Shapiro-Wilk test of normality indicates that our distribution of residuals does not differ from a normal distribution ($W = 0.957, p = 0.385$). Given the space restrictions in journal articles and the higher priority of describing the results of the primary

analyses, I am more likely to report model-level results than the results from the cell-based Shapiro-Wilk tests.

There are limitations to the Shapiro-Wilk test. As the dataset being evaluated gets larger, the Shapiro-Wilk test becomes more sensitive to small deviations; this leads to a greater probability of rejecting the null hypothesis (null hypothesis being the values come from a normal distribution). Green and Salkind [2017c] advised that ANOVA is relatively robust to violations of normality if there are at least 15 cases per cell and the design is reasonably balanced (i.e., equal cell sizes).

9.5.1.4 Are there any outliers (and should we consider their removal)?

The boxplot is one common way for identifying outliers. The boxplot uses the median and the lower (25th percentile) and upper (75th percentile) quartiles. The difference between Q3 and Q1 is the *interquartile range* (IQR). Outliers are generally identified when values fall outside these lower and upper boundaries:

- $Q1 - 1.5 \times IQR$
- $Q3 + 1.5 \times IQR$

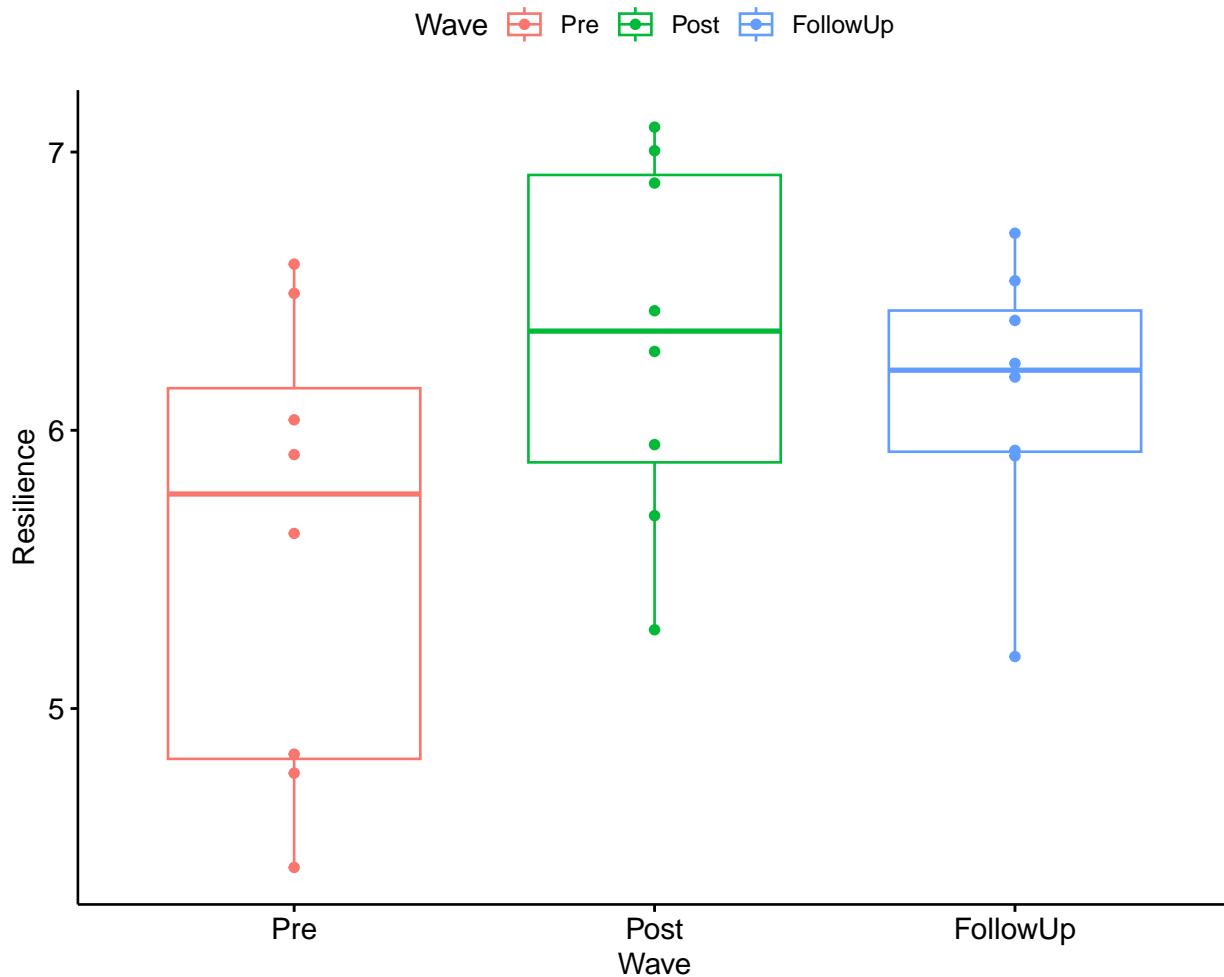
Extreme values occur when values fall outside these boundaries:

- $Q1 - 3 \times IQR$
- $Q3 + 3 \times IQR$

Let's take another look at the boxplot. Swapping “jitter” for “point” may help with the visual inspection.

```
ggpubr::ggboxplot(Amodeo_long, x = "Wave", y = "Resilience", add = "point",
  color = "Wave", title = "Figure 9.2 Identifying Outliers with Boxplots")
```

Figure 9.2 Identifying Outliers with Boxplots



The package *rstatix* has features that allow us to identify outliers.

```
Amodeo_long %>%
  group_by(Wave)%>%
  rstatix::identify_outliers(Resilience)
```

```
[1] Wave      ID       Resilience m_dev      m_devSQ    is.outlier is.extreme
<0 rows> (or 0-length row.names)
```

```
#?identify_outliers
```

The output, “0 rows” indicates there are no outliers.

This is consistent with the visual inspection of boxplots (above), where all observed scores fell within the 1.5x the interquartile range. If there were outliers and you chose to delete them, instructions for doing so are found in the parallel sections of the [one-way ANOVA](#) and [factorial ANOVA](#) lessons.

9.5.1.5 Summarizing results from the analysis of assumptions

Here's how I would write up the assumptions we have tested thus far:

Similarly, no extreme outliers were identified and results of a model-based Shapiro-Wilk test of normality, indicated that the model residuals did not differ from a normal distribution ($W = 0.979, p = 0.15$).

Repeated measures ANOVA has several assumptions regarding normality, outliers, and sphericity. Regarding normality, no values of skew and kurtosis (at each wave of assessment) fell within cautionary ranges for skew and kurtosis [Kline, 2016a]. Additionally, results of a model-based Shapiro-Wilk test of normality indicated that the model residuals did not differ from a normal distribution ($W = 0.957, p = 0.385$). Visual inspection of boxplots for each wave of the design, assisted by the `identify_outliers()` function in the `rstatix` package (which reports values above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$, where IQR is the interquartile range) indicated no outliers.

9.5.1.6 Assumption of Sphericity

The sphericity assumption is automatically checked with Mauchley's test during the computation of the ANOVA when the `rstatix::anova_test()` function is used. When the `rstatix::get_anova_table()` function is used, the Greenhouse-Geisser sphericity correction is automatically applied to factors violating the sphericity assumption.

The effect size, η^2 is reported in the column labeled "ges." Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

You may see different values (.02, .13, .26) offered as small, medium, and large – these values are used when multiple regression is used. A useful summary of effect sizes, guide to interpreting their magnitudes, and common usage can be found [here](#) [Watson, 2020].

Earlier in the lesson I noted that the evaluation of the sphericity assumption occurs at the same time that we evaluate the omnibus ANOVA. We are at that point in the analyses. The workflow points to our stage in the process.

9.5.2 Computing the Test Statistic

As we prepare to run the omnibus ANOVA it may be helpful to think again about our variables. Our DV, Resilience, should be a continuous variable. In R, its structure should be "num." Our predictor, Wave, should be categorical. In R case, Wave should be an ordered factor that is consistent with its timing: pre, post, follow-up.

The repeated measures ANOVA must be run with a long form of the data. This means that there needs to be a within-subjects identifier. In our case, it is the "ID" variable which is also formatted as a factor.

We can verify the format of our variables by examining the structure of our dataframe. Recall that we created the "m_dev" and "m_devSQ" variables earlier in the demonstration. We will not use them in this analysis; it does not harm anything for them to "ride along" in the dataframe.

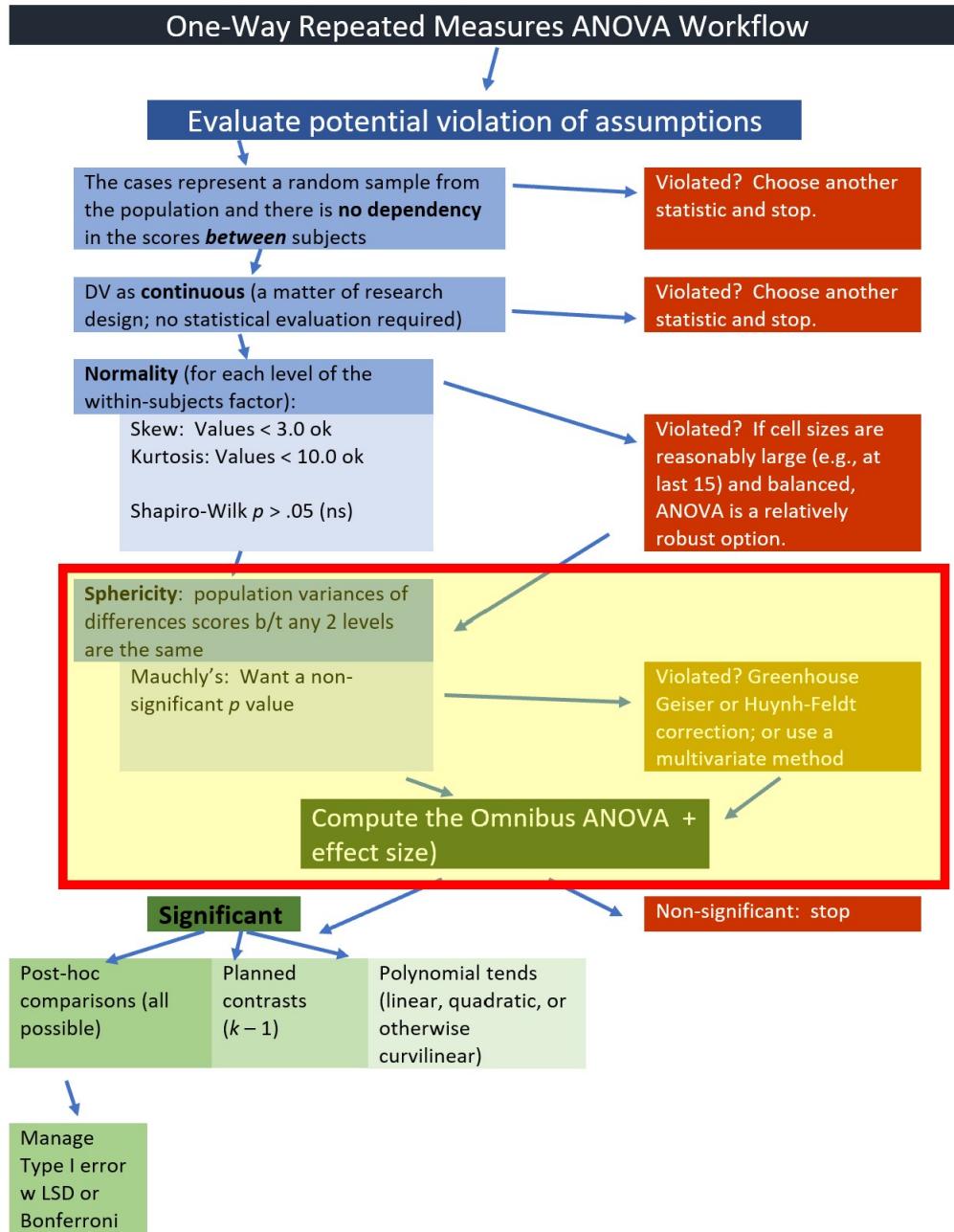


Figure 9.9: Image of our position in the workflow for the one-way repeated measures ANOVA

```
str(Amodeo_long)
```

```
'data.frame': 24 obs. of 5 variables:
 $ ID       : Factor w/ 8 levels "1","2","3","4",...: 1 1 1 2 2 2 3 3 3 4 ...
 $ Wave     : Factor w/ 3 levels "Pre","Post","FollowUp": 1 2 3 1 2 3 1 2 3 1 ...
 $ Resilience: num  6.49 5.28 5.93 4.43 5.95 ...
 $ m_dev    : num  0.4747 -0.7344 -0.0895 -1.5886 -0.0689 ...
 $ m_devSQ  : num  0.22536 0.53927 0.00801 2.52355 0.00475 ...
```

We can use the `rstatix::anova_test()` function to calculate the omnibus ANOVA. Notice where our variables are included in the script:

- Resilience is the dv
- ID is the wid
- Wave is assigned to within

```
RM_AOV <- rstatix::anova_test(data = Amodeo_long, dv = Resilience, wid = ID,
                               within = Wave)
RM_AOV
```

ANOVA Table (type III tests)

```
$ANOVA
  Effect DFn DFd      F      p p<.05    ges
1   Wave    2  14 3.91 0.045      * 0.203

$`Mauchly's Test for Sphericity`
  Effect      W      p p<.05
1   Wave 0.566 0.182

$`Sphericity Corrections`
  Effect    GGe      DF[GG] p[GG] p[GG]<.05    HFe      DF[HF] p[HF] p[HF]<.05
1   Wave 0.698 1.4, 9.77 0.068          0.817 1.63, 11.44 0.057
```

We can assemble our F string from the ANOVA object: $F(2, 14) = 3.91, p = 0.045, \eta^2 = 0.203$

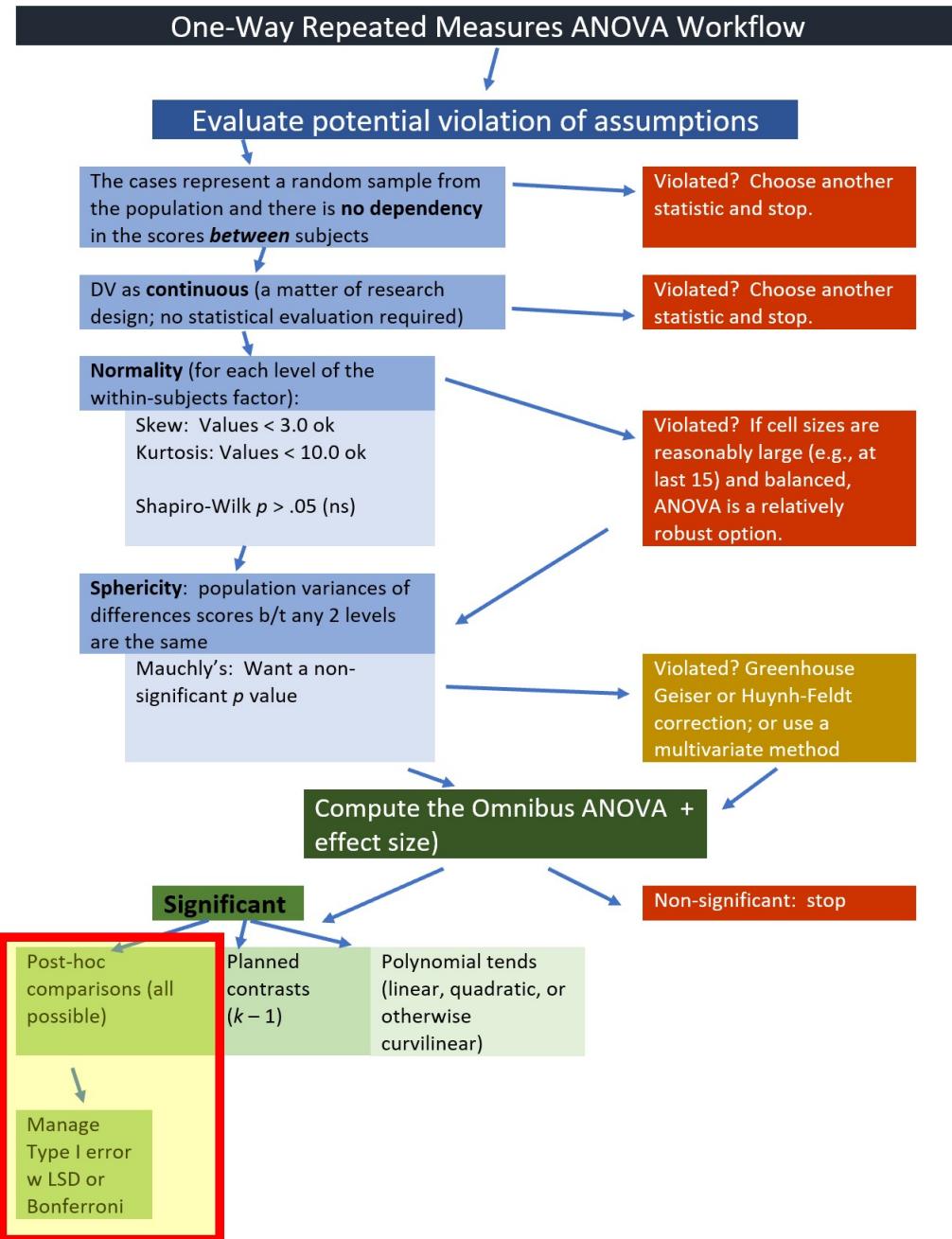
From the Mauchly's Test for Sphericity object we learn that we did not violate the sphericity assumption: $W = 0.566, p = .182$

From the Sphericity Corrections object are output for two alternative corrections to the F statistic, the Greenhouse-Geiser epsilon (GGe), and Huynh-Feldt epsilon (HFe). Because we did not violate the sphericity assumption we do not need to use them. Notice that these two tests adjust our df (both numerator and denominator) to have a more conservative p value.

If we needed to write an F string that is corrected for violation of the sphericity assumption, it might look like this:

The Greenhouse Geiser estimate was 0.698 the correct omnibus was $F(1.4, 9.77) = 3.91$, $p = .068$. The Huyhn Feldt estimate was 0.817 and the corrected omnibus was $F(1.63, 11.44) = 3.91 p = .057$.

You might be surprised that we are at follow-up already.



for the management of Type I Error

Planning

In a one-way repeated measures ANOVA, managing Type I error can be relatively straightforward.

The LSD (least significant differences) method is especially appropriate in the one-way ANOVA scenario when there are only three levels in the factor. In this case, Green and Salkind [2017c]

have suggested that alpha can be retained at the alpha level for the “family” (α_{family}), which is conventionally $p = .05$ and used both to evaluate the omnibus and, so long as they don’t exceed three in number, the planned or pairwise comparisons that follow. Because there are only three levels (i.e., pre, post, follow-up) in this one-way repeated measures design this is what I will do.

More information about options for managing Type I error is included in the [appendix](#).

9.5.3 Follow-up to Omnibus F

Given the simplicity of our design, it makes sense to me to follow-up with post hoc, pairwise, comparisons. Note that when I am calculating these pairwise t tests, I am creating an object (named “pwc”). The object will be a helpful tool in creating a Figure and an APA Style table.

Note that the script used to produced the figure will pull the symbols from the column labeled, “p.adj.signif.” The `rstatix::pairwise_t_test` default is the traditional Bonferroni. Therefore, if we want to use the LSD approach, we must “p.adjust.method” as “none.”

```
pwc <- Amodeo_long %>%
  rstatix::pairwise_t_test(Resilience ~ Wave, paired = TRUE, p.adjust.method = "none")
pwc
```

```
# A tibble: 3 x 10
  .y.      group1 group2     n1     n2 statistic     df     p p.adj p.adj.signif
* <chr>    <chr>  <chr>   <int>   <int>     <dbl>   <dbl> <dbl> <dbl> <chr>
1 Resilience Pre     Post       8       8     -2.15     7 0.069 0.069 ns
2 Resilience Pre     Follow~     8       8     -2.00     7 0.086 0.086 ns
3 Resilience Post    Follow~     8       8      1.06     7 0.325 0.325 ns
```

Although omnibus test had a statistically significant result, we did not obtain statistically significant results in an of the posthoc pairwise comparisons. Why not?

- Our omnibus F was right at the margins
 - a larger sample size (assuming that the effects would hold) would have been more powerful.
 - there could be significance if we compared pre to the combined effects of post and follow-up.

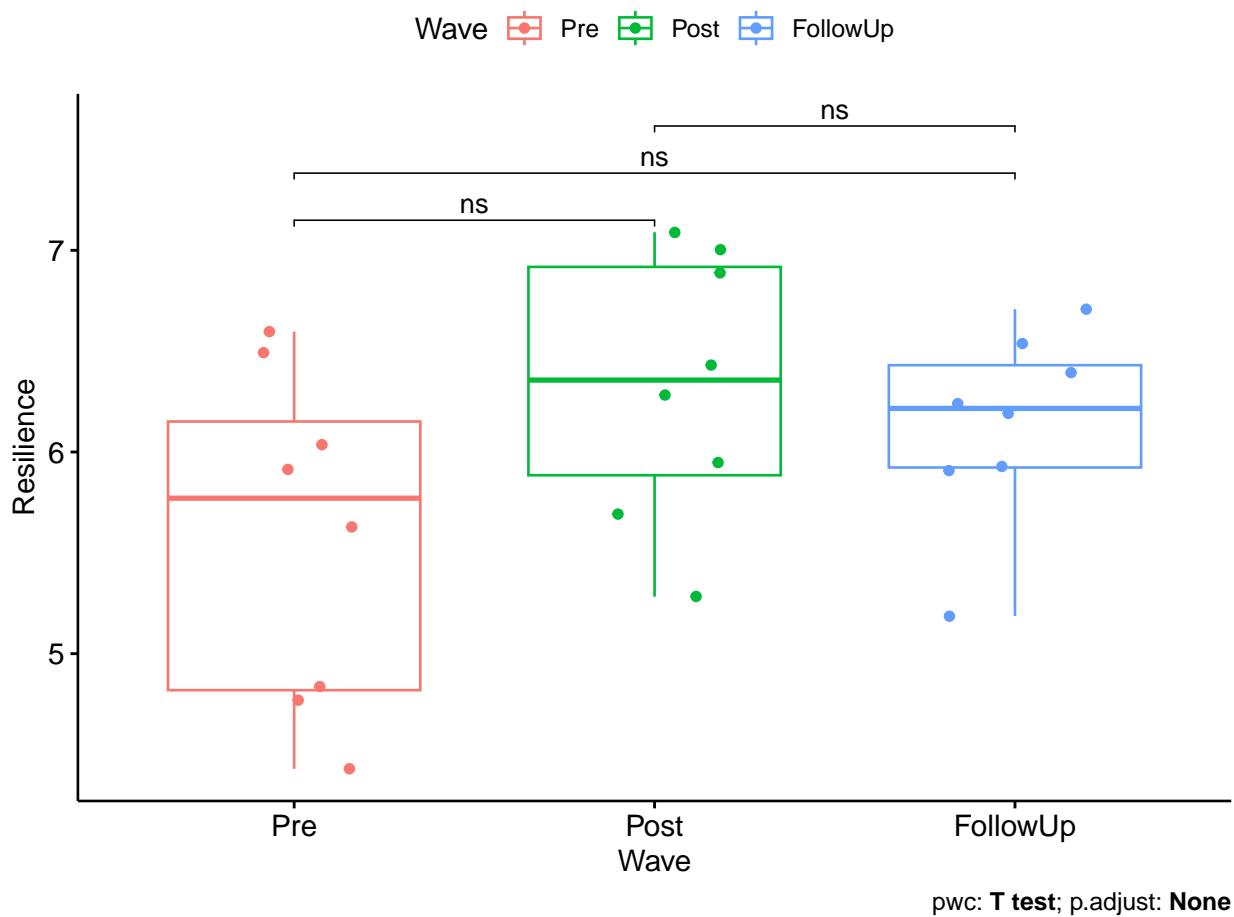
How would we manage Type I error? With only three possible post-omnibus comparisons, I would cite the Tukey LSD approach and not adjust the alpha to a more conservative level [[Green and Salkind, 2017c](#)].

We can combine information from the object we created (“bxp”) from an earlier boxplot with the object we saved from the posthoc pairwise comparisons (“pwc”) to enhance our boxplot with the F string and indications of pairwise significant (or, in our case, non-significance).

```
RMbox <- ggpubr::ggboxplot(Amodeo_long, x = "Wave", y = "Resilience", add = "jitter",
  color = "Wave", title = "Figure 9.3 Resilience as a Function of Wave")
pwc <- pwc %>%
  rstatix::add_xy_position(x = "Wave")
RMbox <- RMbox + ggpubr::stat_pvalue_manual(pwc, label = "p.adj.signif",
  tip.length = 0.01, hide.ns = FALSE, step.increase = 0.05) + labs(subtitle = rstatix::get_tukey_detailed = TRUE), caption = rstatix::get_pwc_label(pwc))
RMbox
```

Figure 9.3 Resilience as a Function of Wave

Anova, $F(2,14) = 3.91, p = 0.045, \eta^2_g = 0.2$



Unfortunately, the *apaTables* package does not work with the *rstatix* package, so a table would need to be crafted by hand.

9.6 APA Style Results

Repeated measures ANOVA has several assumptions regarding normality, outliers, and sphericity. Regarding normality, no values of skew and kurtosis (at each wave of assessment) fell within cautionary ranges for skew and kurtosis [Kline, 2016a]. Additionally,

results of a model-based Shapiro-Wilk test of normality indicated that the model residuals did not differ from a normal distribution ($W = 0.957, p = 0.385$). Visual inspection of boxplots for each wave of the design, assisted by the `identify_outliers()` function in the `rstatix` package (which reports values above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$, where IQR is the interquartile range) indicated no outliers. A non-significant Mauchley's test ($W = 0.566, p = .182$) indicated that the sphericity assumption was not violated.

The omnibus ANOVA was significant: $F(2, 14) = 3.91, p = 0.045, \eta^2 = 0.203$. We followed up with all pairwise comparisons. Curiously, and in spite of a significant omnibus test, there were not statistically significant differences between any of the pairs. Regarding pre versus post, $t = -2.15, p = .069$. Regarding pre versus follow-up, $t = -2.00, p = .068$. Regarding post versus follow-up, $t = 1.059, p = .325$. Because there were only three pairwise comparisons subsequent to the omnibus test, we used the LSD (least significant differences) approach to managing Type I error and alpha was retained at .05 [Green and Salkind, 2017c]. While the trajectories from pre-to-post and pre-to-follow-up were in the expected direction, the small sample size likely contributed to a Type II error. Descriptive statistics are reported in Table 1 and the differences are illustrated in Figure 1.

While I have not located a package that will take `rstatix` output to make an APA style table, we can use the `MASS` package to write the `pwc` object to a `.csv` file, then manually make our own table.

```
MASS:::write.matrix(pwc, sep = ", ", file = "PWC.csv")
```

9.6.1 Comparison with Amodeo et al.[2018]

How do our findings and our write-up from the simulated data compare with the original article?

In the published manuscript, the F string is presented in the Table 1 note ($F[1.612, 11.283] = 6.390, p = 0.18, \eta_p^2$). We can tell from the fractional degrees of freedom that the p value has been had a correction for violation of the sphericity assumption.

Table 1 also reports all of the post hoc, pairwise comparisons. There is no mention of control for Type I error. Had they used a traditional Bonferroni, they would have needed to reduce the alpha to $(k * (k-1)/2)$ and then divide .05 by that number.

```
(3 * (3-1))/2
```

```
[1] 3
```

```
.05/3
```

```
[1] 0.01666667
```

Although Amodeo et al. report six comparisons; three are repeated because they are merely in reverse. Thus, the revised alpha would be .016 and the one, lone, comparison would not have been statistically significant. That said, the Tukey LSD is appropriate because there are only 3 comparisons and holding alpha at .05 can be defended [Green and Salkind, 2017c].

- Regarding the presentation of the results
 - there is no figure
 - there is no data presented in the text; all data is presented in Table 1
- Regarding the research design and its limitations
 - the authors note that a control condition would have better supported the conclusions
 - the authors note the limited sample size and argue that this is a difficult group to recruit for intervention and evaluation
 - the article is centered around the qualitative aspect of the design; the quantitative portion is, appropriately, secondary



Figure 9.10: Another peek at the research design for the Amodeo et al study

9.7 Power Analysis

Power analysis allows us to determine the probability of detecting an effect of a given size with a given level of confidence. The package *wp.rmanova* was designed for power analysis in repeated measures ANOVA.

In the *WebPower* package, we specify 6 of 7 interrelated elements; the package computes the missing one.

- n = sample size (number of individuals in the whole study).
- ng = number of groups.
- nm = number of measurements/conditions/waves.
- f = Cohen's f (an effect size; we can use an effect size converter to obtain this value)
 - Cohen suggests that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes, respectively.
- $nscor$ = the Greenhouse Geiser correction from our output; 1.0 means no correction was needed and is the package's default; < 1 means some correction was applied.
- $alpha$ = is the probability of Type I error; we traditionally set this at .05
- $power$ = $1 - P(\text{Type II error})$ we traditionally set this at .80 (so anything less is less than what we want).

- *type* = 0 is for between-subjects, 1 is for repeated measures, 2 is for interaction effect.

I used *effectsize::eta2_to_f* packages convert our η^2 to Cohen's *f*.

```
effectsize::eta2_to_f(.203)
```

```
[1] 0.5046832
```

Retrieving the information about our study, we add it to all the arguments except the one we wish to calculate. For power analysis, we write “power = *NULL*.”

```
WebPower::wp.ranova(n = 8, ng = 1, nm = 3, f = 0.5047, nscor = 0.689,
alpha = 0.05, power = NULL, type = 1)
```

Repeated-measures ANOVA analysis

n	f	ng	nm	nscor	alpha	power
8	0.5047	1	3	0.689	0.05	0.1619613

NOTE: Power analysis for within-effect test

URL: <http://psychstat.org/ranova>

Here we learned that we were only powered at .16. That is, we had a 16% chance of finding a statistically significant effect if, in fact, it existed. This is low!

In reverse, setting *power* at .80 (the traditional value) and changing *n* to *NULL* yields a recommended sample size.

In many cases we won't know some of the values in advance. We can make best guesses based on our review of the literature.

```
WebPower::wp.ranova(n = NULL, ng = 1, nm = 3, f = 0.5047, nscor = 0.689,
alpha = 0.05, power = 0.8, type = 1)
```

Repeated-measures ANOVA analysis

n	f	ng	nm	nscor	alpha	power
50.87736	0.5047	1	3	0.689	0.05	0.8

NOTE: Power analysis for within-effect test

URL: <http://psychstat.org/ranova>

With these new values, we learn that we would need 50 individuals in order to feel confident in our ability to get a statistically significant result if, in fact, it existed.

9.8 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “do-able” and then try at least one more problem that challenges you in some way. All problems attempted should have at least three levels in the independent variable. At least one problem should have a significant omnibus test and require follow-up.

Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Whichever you choose, you will focus on these larger steps in one-way repeated measures/within-subjects ANOVA, including:

- test the statistical assumptions
- conduct a one-way, including
 - omnibus test and effect size
 - conduct follow-up testing
- write a results section to include a figure and tables

9.8.1 Problem #1: Change the Random Seed

If repeated measures ANOVA is new to you, perhaps change the random seed and follow-along with the lesson.

9.8.2 Problem #2: Increase N

Our analysis of the Amodeo et al. [Amodeo et al., 2018] data failed to find significant increases in resilience from pre-to-post through follow-up. Our power analysis suggested that a sample size of 50 would be sufficient to garner statistical significance. The script below re-simulates the data by increasing the sample size to 50 (from 8). All else remains the same.

```
set.seed(2022)
ID <- factor(c(rep(seq(1, 50), each = 3))) #gives me 8 numbers, assigning each number 3 consecutive
Resilience <- rnorm(150, mean = c(5.7, 6.21, 6.26), sd = c(0.88, 0.79,
  0.37)) #gives me a column of 24 numbers with the specified Ms and SD
Wave <- rep(c("Pre", "Post", "FollowUp"), each = 1, 50) #repeats pre, post, follow-up once each
Amodeo50_long <- data.frame(ID, Wave, Resilience)
```

9.8.3 Problem #3: Try Something Entirely New

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a one-way repeated measures ANOVA. Please have at least 3 levels for the predictor variable.

9.8.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the IV and DV. The data you analyze should have at least 3 levels in the independent variable; at least one of the attempted problems should have a significant omnibus test so that follow-up is required)	5	_____
2. Check and, if needed, format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct omnibus ANOVA (w effect size)	5	_____
5. Conduct all possible pairwise comparisons (like in the lecture)	5	_____
6. Describe approach for managing Type I error	5	_____
7. APA style results with table(s) and figure	5	_____
8. Explanation to grader	5	_____
Totals	35	_____

Chapter 10

Mixed Design ANOVA

[Screencasted Lecture Link](#)

The focus of this lecture is mixed design ANOVA. That is, we are conducting a two-way ANOVA where one of the factors is repeated measures and one of the factors is between groups. The mixed design ANOVA is often associated with the random clinical trial (RCT) where the researcher hopes for a significant interaction effect. Specifically, the researcher hopes that the individuals who were randomly assigned to the treatment condition improve from pre-test to post-test and maintain (or continue to improve) after post-test, while the people assigned to the no-treatment control are not statistically significantly different from treatment group at pre-test, and do not improve over time.

10.1 Navigating this Lesson

There is just over one hour of lecture. If you work through the materials with me it would be plan for an additional two hours.

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's [introduction](#)

10.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Evaluate the suitability of a research design/question and dataset for conducting a mixed design ANOVA; identify alternatives if the data is not suitable.
- Test the assumptions for mixed design ANOVA.
- Conduct a mixed design ANOVA (omnibus and follow-up) in R.
- Interpret output from the mixed design ANOVA (and follow-up).
- Prepare an APA style results section of the mixed design ANOVA output.
- Conduct a power analysis for mixed design ANOVA.

10.1.2 Planning for Practice

In each of these lessons I provide suggestions for practice that allow you to select from problems that vary in degree of difficulty. The least complex is to change the random seed and rework the problem demonstrated in the lesson. The results *should* map onto the ones obtained in the lecture.

The second option comes from the research vignette. The Murrar and Brauer [2018] article has three variables (attitudes toward Arabs, attitudes toward Whites, and a difference score) which are suitable for mixed design ANOVAs. I will demonstrate a mixed design ANOVA with the difference score. I'll leave the other two variables for opportunities for practice.

As a third option, you are welcome to use data to which you have access and is suitable for two-way ANOVA. In either case the practice options suggest that you:

- test the statistical assumptions
- conduct a mixed design ANOVA, including
 - omnibus test and effect size
 - report main and interaction effects
 - conduct follow-up testing of simple main effects
- write a results section to include a figure and tables

10.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Mixed ANOVA in R. (n.d.). Datanovia. Retrieved October 19, 2020, from <https://www.datanovia.com/en/lessons/mixed-anova-in-r/>
 - This website is an excellent guide for mixed design ANOVA and providing explanatory figures of the results. It is a great resource for both the conceptual and procedural. This is the guide I have used for the basis of the lecture. Working through their example would be provide an additional, excellent, opportunity for practice.
- Murrar, S., & Brauer, M. (2018). Entertainment-education effectively reduces prejudice. *Group Processes & Intergroup Relations*, 21(7), 1053–1077. <https://doi.org/10.1177/1368430216682350>
 - This article is the source of our research vignette. Our vignette is simulated from the first of their two experiments. The authors did not conduct mixed design ANOVA. Instead, they ran independent-samples *t* tests to test the differences between the sitcom conditions for each of the three waves. This is comparable to conducting the simple-main effect analysis of condition within wave subsequent to a significant interaction.
 - Full-text of the article is available at the [authors' ResearchGate](#).

10.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# will install the package if not already installed
# if(!require(knitr)){install.packages('knitr')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(psych)){install.packages('psych')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(rstatix)){install.packages('rstatix')}
# if(!require(MASS)){install.packages('MASS')}
# if(!require(effectsize)){install.packages('effectsize')}
# if(!require(WebPower)){install.packages('WebPower')}
```

10.2 Introducing Mixed Design ANOVA

Mixed design ANOVA is characterized by the following:

- at least two independent variables.
- Termed “mixed” because
 - one is a between-subjects factor, and
 - one is a repeated-measures (i.e., within-subjects) factor.
- In essence, we are simultaneously conducting
 - a one-way independent ANOVA and a
 - a one-way repeated-measures ANOVA.

The illustration below represents the simplest of the mixed ANOVA designs. This 2x2 ANOVA includes random assignment to the between-participants factor. In this case it represents a control (or comparison) condition to a treatment condition. The within-persons factor is pre-test and post-test.

By increasing the number of between-persons conditions or follow-up assessments, the mixed design ANOVA can quickly become more complex.

Random assignment	Pre-Test	Control or Comparison Condition	Post-Test
	Pre-Test	Treatment Condition	Post-Test

Figure 10.1: Illustration of a research design appropriate for a mixed design ANOVA

Especially when there is a significant interaction there can be numerous ways to follow up. We will work one set of analyses: simple main effects (condition within wave; wave within condition)

and, when needed, conduct posthoc pairwise comparisons as follow-up. Other good options include identifying a priori contrasts and conducting polynomials (not demonstrated in this lecture).

10.2.1 Workflow for the Mixed Design ANOVA

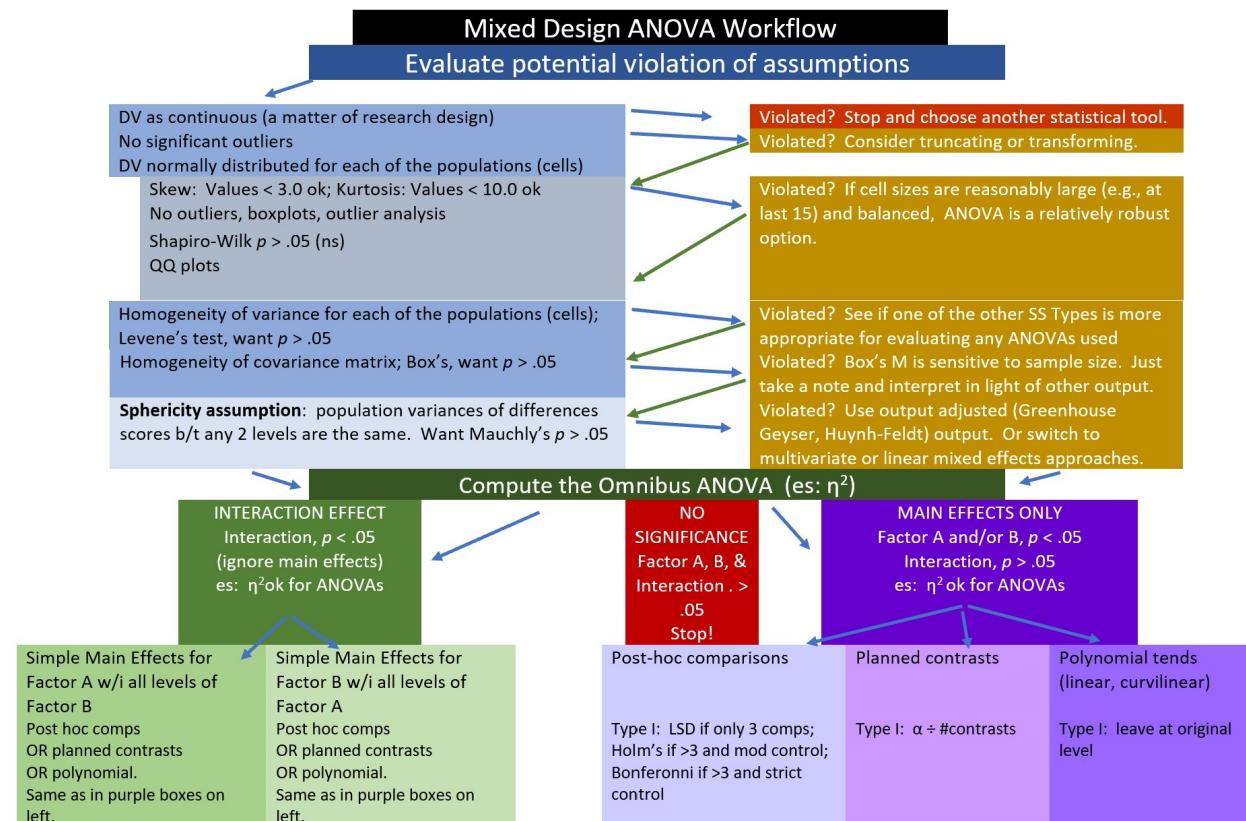


Figure 10.2: Image of a workflow for mixed design ANOVA

The steps in working the mixed design generally include,

- Exploring the data/evaluating the assumptions
- Evaluating the omnibus test
- Follow-up to the omnibus
 - if significant interaction effect: simple main effects and further follow-up to those
 - if significant main effect (but no significant interaction effect), identify source of significance in the main effect
 - if no significance, stop
- Write it up with tables, figure(s)

Assumptions for the mixed design ANOVA include the following:

- The dependent variable should be continuous with no significant outliers in any cell of the design

- Check by visualizing the data using box plots and by using the `rstatix::identify_outliers()` function
- The DV should be approximately normally distributed in each cell of the design
 - Check with Shapiro-Wilk normality test `rstatix::shapiro_test()` function and with visual inspection by creating Q-Q plots. The `ggpubr::ggqqplot()` function is a great tool.
- The variances of the differences between groups should be equal. This is termed the **sphericity assumption**. This can be checked with Mauchly's test of sphericity, which is reported automatically in the `rstatix::anova_test()` output.

The best way to address violations of these assumptions is not always clear. Possible solutions include:

- For 2- and 3- way ANOVAs, violations of the normality assumption might be addressed by removing extreme outliers or considering transformations of the data. Transformations, though, introduce their own complexities regarding interpretation. Kline's text [Kline, 2016a] provides excellent coverage of options.
- A robust ANOVA option is available in the `WRS2` package
- If there are three or more waves/conditions and the sample is large, it may be possible to run a multilevel, model.
- In the absence of alternatives, it may be necessary to run the mixed design with the violated assumptions, but report them.
-and more. Internet searches continue to offer new approaches and alternatives.

10.3 Research Vignette

This lesson's research vignette is from Murrar and Brauer's [2018] article that describes the results of two studies that evaluated interventions designed to reduce prejudice against Arabs/Muslims. We are working only a portion of the first study reported in the article. Participants ($N = 193$), all who were White, were randomly assigned to one of two conditions where they watched six episodes of the sitcom *Friends* or *Little Mosque on the Prairie*. The sitcoms and specific episodes were selected after significant pilot testing. The researchers wanted stimuli that were as similar as possible while ensuring that the intervention-oriented sitcom would be capable of reducing prejudice. The authors felt that both series had characters that were likable and relatable and were engaged in regular activities of daily living. The Friends series featured characters who were predominantly White, cisgender, and straight. The Little Mosque series portrayed the experience of Western Muslims and Arabs as they lived in a small Canadian town. This study involved assessment across three waves: baseline (before watching the assigned episodes), post1 (immediately after watching the episodes), and post2 (completed 4-6 weeks after watching the episodes).

The study used *feelings and liking thermometers*, rating their feelings and liking toward 10 different groups of people on a 0 to 100 sliding scale (with higher scores reflecting greater liking and positive feelings). For the purpose of this analysis, the ratings of attitudes toward White people and attitudes toward Arabs/Muslims were used. A third metric was introduced by subtracting the attitudes towards Arabs/Muslims from the attitudes toward Whites. Higher scores indicated more positive attitudes toward Whites where as low scores indicated no difference in attitudes. To recap, there were three potential dependent variables, all continuously scaled:

- *AttWhite*: attitudes toward White people; higher scores reflect greater liking
- *AttArab*: attitudes toward Arab people; higher scores reflect greater liking
- *Diff*: the difference between AttWhite and AttArab; higher scores reflect a greater liking for White people

With random assignment, nearly equal cell sizes, a condition with two levels (Friends, Little Mosque), and three waves (baseline, post1, post2), this is perfect for mixed design ANOVA.

	COND	Baseline At start of study (prior to viewing sitcoms)	Intervention 6 episodes of the sitcom	Post1 Toward end of viewing the sitcoms	Post2 4-6 weeks after viewing the final sitcom
Random Assignment	Friends	X		X	X
	Little Mosque on the Prairie	X	Selected for potential for prejudice reduction	X	X

Figure 10.3: Image of the design for the Murrar and Brauer (2018) study

10.3.1 Data Simulation

Below is the code used to simulate the data. The simulation includes two dependent variables (AttWhite, AttArab), Wave (baseline, post1, post2), and COND (condition; Friends, Little_Mosque). There is also a caseID (repeated three times across the three waves) and rowID (giving each observation within each case an ID). This creates the long-file, where each person has 3 rows of data representing baseline, post1, and post2. You can use this simulation for two of the three practice suggestions.

```
library(tidyverse)
# change this to any different number (and rerun the simulation) to
# rework the chapter problem
set.seed(210813)
AttWhite <- round(c(rnorm(98, mean = 76.79, sd = 18.55), rnorm(95, mean = 75.37,
  sd = 18.99), rnorm(98, mean = 77.47, sd = 18.95), rnorm(95, mean = 75.81,
  sd = 19.29), rnorm(98, mean = 77.79, sd = 17.25), rnorm(95, mean = 75.89,
  sd = 19.44)), 3) #sample size, M and SD for each cell; this will put it in a long file
# set upper bound for variable
AttWhite[AttWhite > 100] <- 100
# set lower bound for variable
AttWhite[AttWhite < 0] <- 0
AttArab <- round(c(rnorm(98, mean = 64.11, sd = 20.97), rnorm(95, mean = 64.37,
  sd = 20.03), rnorm(98, mean = 64.16, sd = 21.64), rnorm(95, mean = 70.52,
  sd = 18.55), rnorm(98, mean = 65.29, sd = 19.76), rnorm(95, mean = 70.3,
  sd = 17.98)), 3)
# set upper bound for variable
AttArab[AttArab > 100] <- 100
# set lower bound for variable
AttArab[AttArab < 0] <- 0
```

```

rowID <- factor(seq(1, 579))
caseID <- rep((1:193), 3)
Wave <- c(rep("Baseline", 193), rep("Post1", 193), rep("Post2", 193))
COND <- c(rep("Friends", 98), rep("LittleMosque", 95), rep("Friends", 98),
      rep("LittleMosque", 95), rep("Friends", 98), rep("LittleMosque", 95))
# groups the 3 variables into a single df: ID#, DV, condition
Murrar_df <- data.frame(rowID, caseID, Wave, COND, AttArab, AttWhite)

```

Let's check the structure. We want

- rowID and caseID to be unordered factors
- Wave and COND to be ordered factors
- AttArab and AttWhite to be numerical

```
str(Murrar_df)
```

```

'data.frame': 579 obs. of 6 variables:
$ rowID   : Factor w/ 579 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
$ caseID  : int  1 2 3 4 5 6 7 8 9 10 ...
$ Wave    : chr  "Baseline" "Baseline" "Baseline" "Baseline" ...
$ COND    : chr  "Friends" "Friends" "Friends" "Friends" ...
$ AttArab : num  74.3 55.8 33.3 66.3 71 ...
$ AttWhite: num  100 79 75.9 68.2 100 ...

```

The script below changes

- caseID from integer to factor
- Wave and COND from factor to ordered factors
 - It makes sense to order Friends and LittleMosque, since we believe that LittleMosque contains prejudice-reducing properties

```

# make caseID a factor
Murrar_df[, "caseID"] <- as.factor(Murrar_df[, "caseID"])
# make Wave an ordered factor
Murrar_df$Wave <- factor(Murrar_df$Wave, levels = c("Baseline", "Post1",
                                                       "Post2"))
# make COND an ordered factor
Murrar_df$COND <- factor(Murrar_df$COND, levels = c("Friends", "LittleMosque"))

```

Let's check the structure again.

```
str(Murrar_df)
```

```
'data.frame': 579 obs. of 6 variables:
 $ rowID   : Factor w/ 579 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ caseID   : Factor w/ 193 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ Wave     : Factor w/ 3 levels "Baseline","Post1",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ COND     : Factor w/ 2 levels "Friends","LittleMosque": 1 1 1 1 1 1 1 1 1 1 ...
 $ AttArab  : num  74.3 55.8 33.3 66.3 71 ...
 $ AttWhite : num  100 79 75.9 68.2 100 ...
```

A key dependent variable in the Murrar and Brauer [Murrar and Brauer, 2018] article is *attitude difference*. Specifically, the attitudes toward Arabs score was subtracted from the attitudes toward Whites scores. Higher attitude difference indicate a greater preference for Whites. Let's create that variable, here.

```
Murrar_df$Diff <- Murrar_df$AttWhite - Murrar_df$AttArab
head(Murrar_df)
```

	rowID	caseID	Wave	COND	AttArab	AttWhite	Diff
1	1	1	Baseline	Friends	74.291	100.000	25.709
2	2	2	Baseline	Friends	55.796	78.977	23.181
3	3	3	Baseline	Friends	33.267	75.938	42.671
4	4	4	Baseline	Friends	66.315	68.232	1.917
5	5	5	Baseline	Friends	70.992	100.000	29.008
6	6	6	Baseline	Friends	94.297	77.961	-16.336

If you want to export this data as a file to your computer, remove the hashtags to save it (and re-import it) as a .csv (“Excel lite”) or .rds (R object) file. This is not a necessary step.

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R. This is what I would do.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Murrar_df, 'Murrar_RDS.rds') bring back the simulated
# dat from an .rds file Murrar_df <- readRDS('Murrar_RDS.rds')
```

The code for .csv will likely lose the formatting (i.e., stripping Wave and COND of their ordered factors), but it is easy to view in Excel.

```
# write the simulated data as a .csv write.table(Murrar_df,
# file='DiffCSV.csv', sep=',', col.names=TRUE, row.names=FALSE) bring
# back the simulated dat from a .csv file Murrar_df <- read.csv
# ('DiffCSV.csv', header = TRUE)
```

10.3.2 Quick peek at the data

Let's first examine the descriptive statistics (e.g., means of the variable, Negative) by group. We can use the *describeBy()* function from the *psych* package.

```
Diff.descripts <- psych::describeBy(Diff ~ COND + Wave, mat = TRUE, data = Murrar_df,
  digits = 3) #digits allows us to round the output
Diff.descripts
```

	item	group1	group2	vars	n	mean	sd	median	trimmed	mad
Diff1	1	Friends	Baseline		1 98	9.306	23.909	8.804	8.991	24.481
Diff2	2	LittleMosque	Baseline		1 95	9.733	30.519	10.797	10.554	30.905
Diff3	3	Friends	Post1		1 98	15.926	26.418	16.191	16.231	29.771
Diff4	4	LittleMosque	Post1		1 95	-0.149	26.969	-1.280	-0.940	23.932
Diff5	5	Friends	Post2		1 98	11.954	23.336	10.882	11.834	24.592
Diff6	6	LittleMosque	Post2		1 95	3.670	23.665	1.860	3.786	25.261
		min	max	range	skew	kurtosis	se			
Diff1		-47.342	72.565	119.907	0.176	-0.356	2.415			
Diff2		-71.510	90.737	162.247	-0.201	-0.030	3.131			
Diff3		-42.598	82.288	124.886	-0.046	-0.575	2.669			
Diff4		-65.259	83.367	148.626	0.328	0.549	2.767			
Diff5		-46.528	75.014	121.542	0.132	0.065	2.357			
Diff6		-53.856	55.264	109.120	-0.065	-0.424	2.428			

```
# Note. Recently my students and I have been having intermittent
# struggles with the describeBy function in the psych package. We
# have noticed that it is problematic when using .rds files and when
# using data directly imported from Qualtrics. If you are having
# similar difficulties, try uploading the .csv file and making the
# appropriate formatting changes.
```

First we inspect the means. We see that the baseline scores for the Friends and Little Mosque conditions are similar. However, the post1 and post2 difference scores (i.e., difference in attitudes toward White and Arab individuals, where higher scores indicate more favorable ratings of White individuals) are higher in the Friends condition than in the Little Mosque condition.

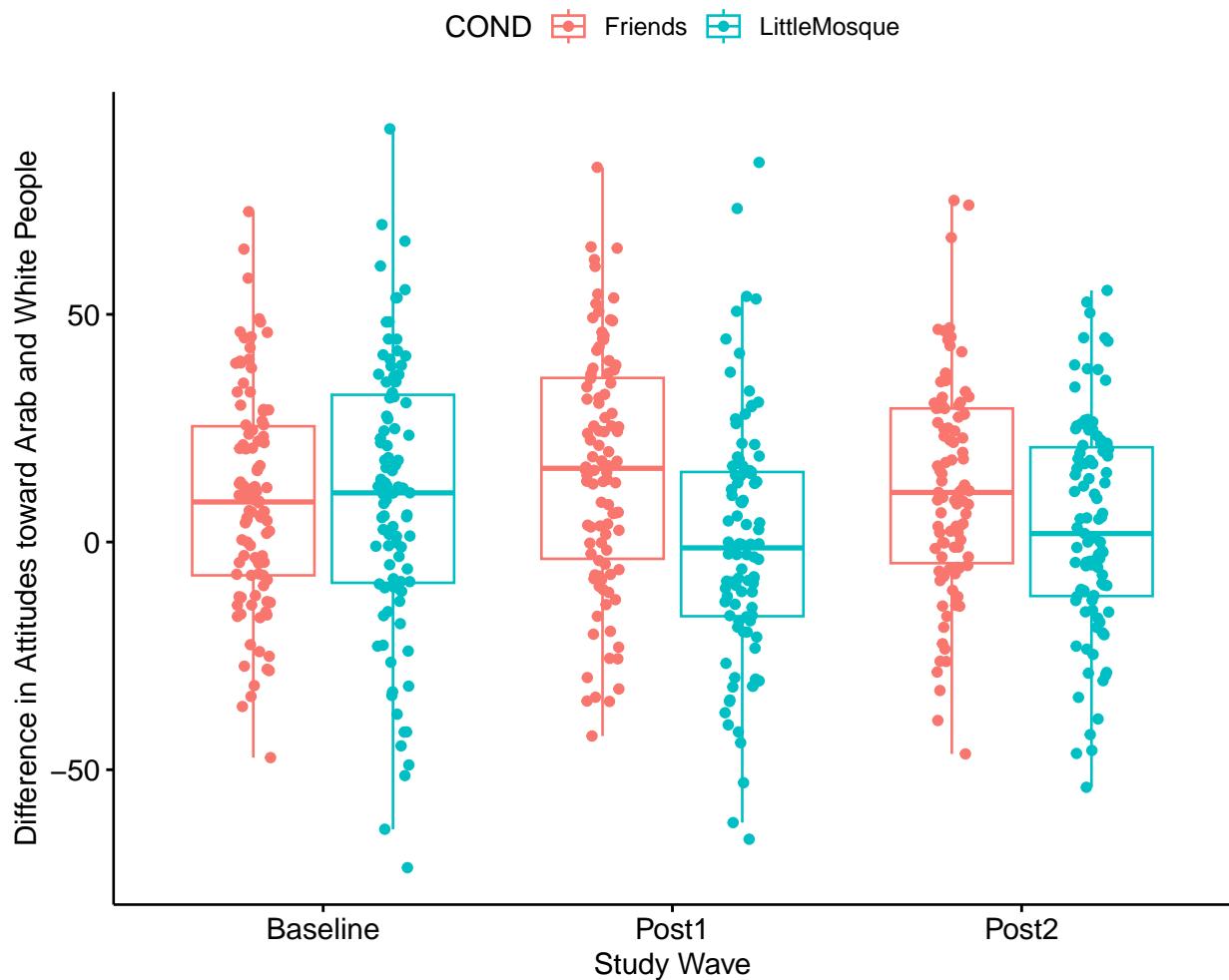
The `write.table()` function can be a helpful way to export output to .csv files so that you can manipulate it into tables.

```
write.table(Diff.descripts, file = "DiffDescripts.csv", sep = ", ", col.names = TRUE,
  row.names = FALSE)
```

At this stage, it would be useful to plot our data. Figures can assist in the conceptualization of the analysis.

```
ggpubr::ggboxplot(Murrar_df, x = "Wave", y = "Diff", color = "COND", xlab = "Study Wave",
  ylab = "Difference in Attitudes toward Arab and White People", add = "jitter",
  title = "Difference Scores: Condition within Wave")
```

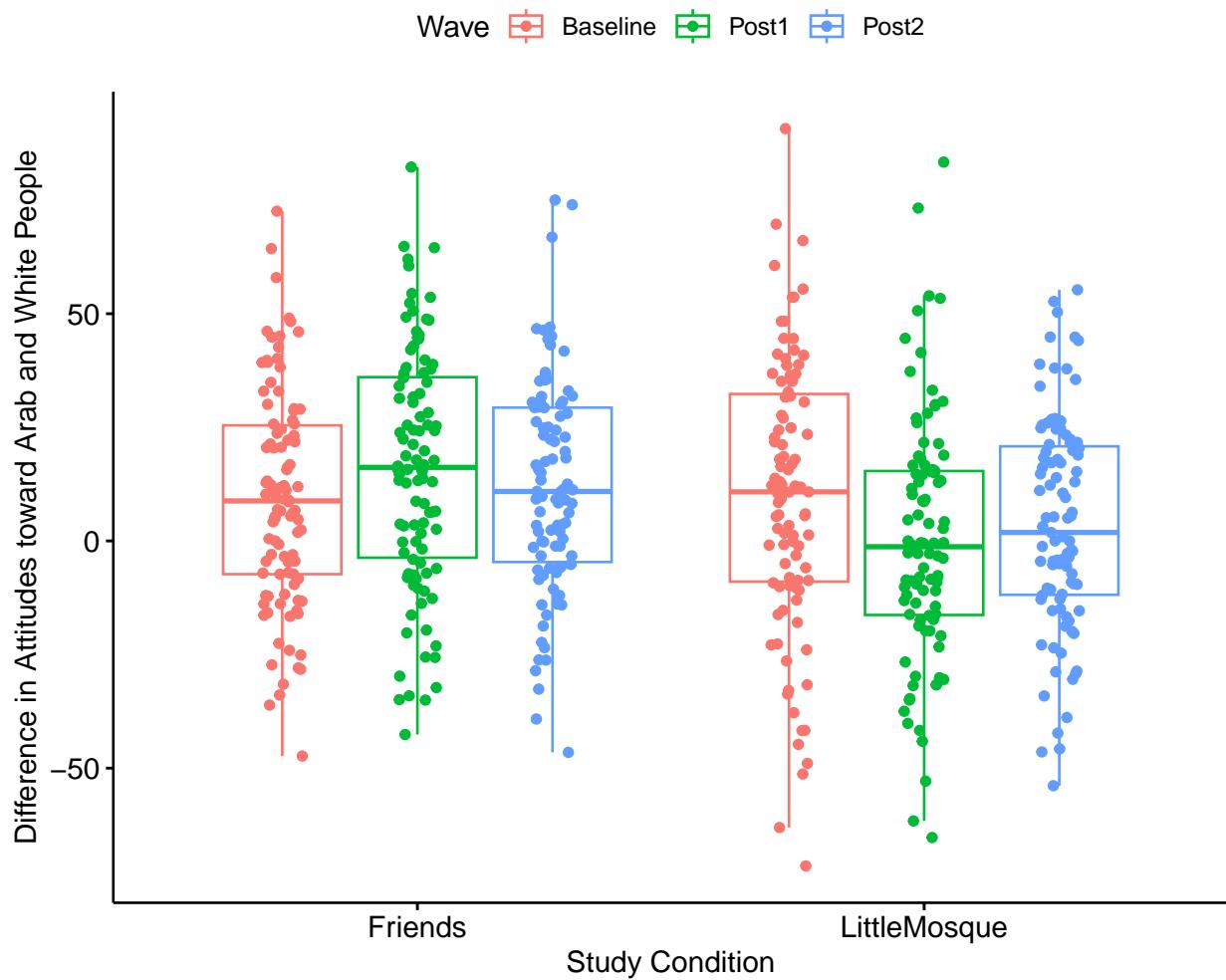
Difference Scores: Condition within Wave



Narrating results is sometimes made easier if variables are switched. There is usually not a right or wrong answer. Here is another view, switching the Rater and Photo predictors.

```
ggpubr::ggbboxplot(Murrar_df, x = "COND", y = "Diff", color = "Wave", xlab = "Study Condition",
  ylab = "Difference in Attitudes toward Arab and White People", add = "jitter",
  title = "Difference Scores: Wave within Condition")
```

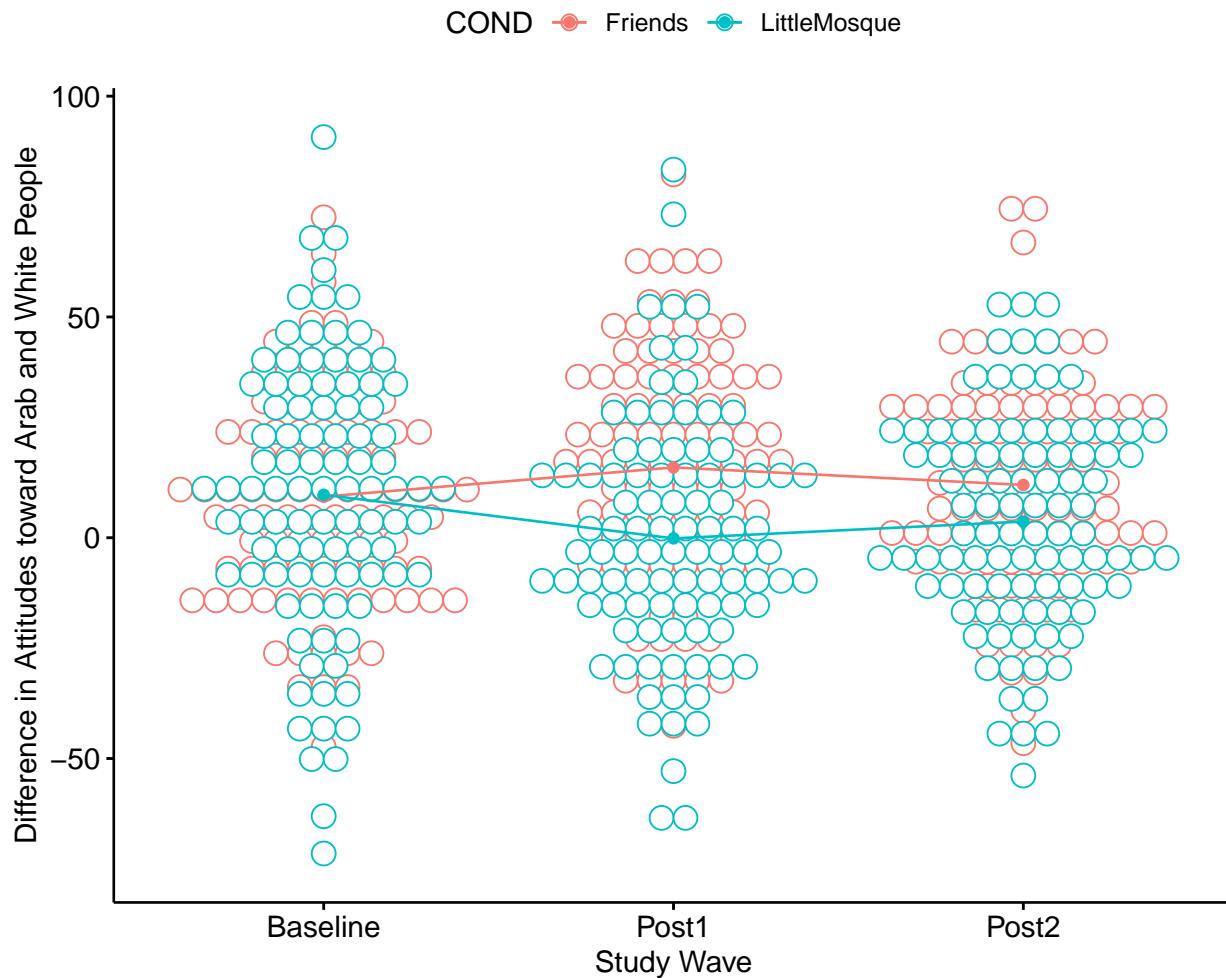
Difference Scores: Wave within Condition



Yet another option plots the raw data as bubbles, the means as lines, and denotes differences in the moderator with color.

```
ggpubr::gglime(Murrar_df, x = "Wave", y = "Diff", color = "COND", xlab = "Study Wave",
  ylab = "Difference in Attitudes toward Arab and White People", add = c("mean_se",
  "dotplot"), title = "Lineplots: Condition within Wave")
```

Lineplots: Condition within Wave

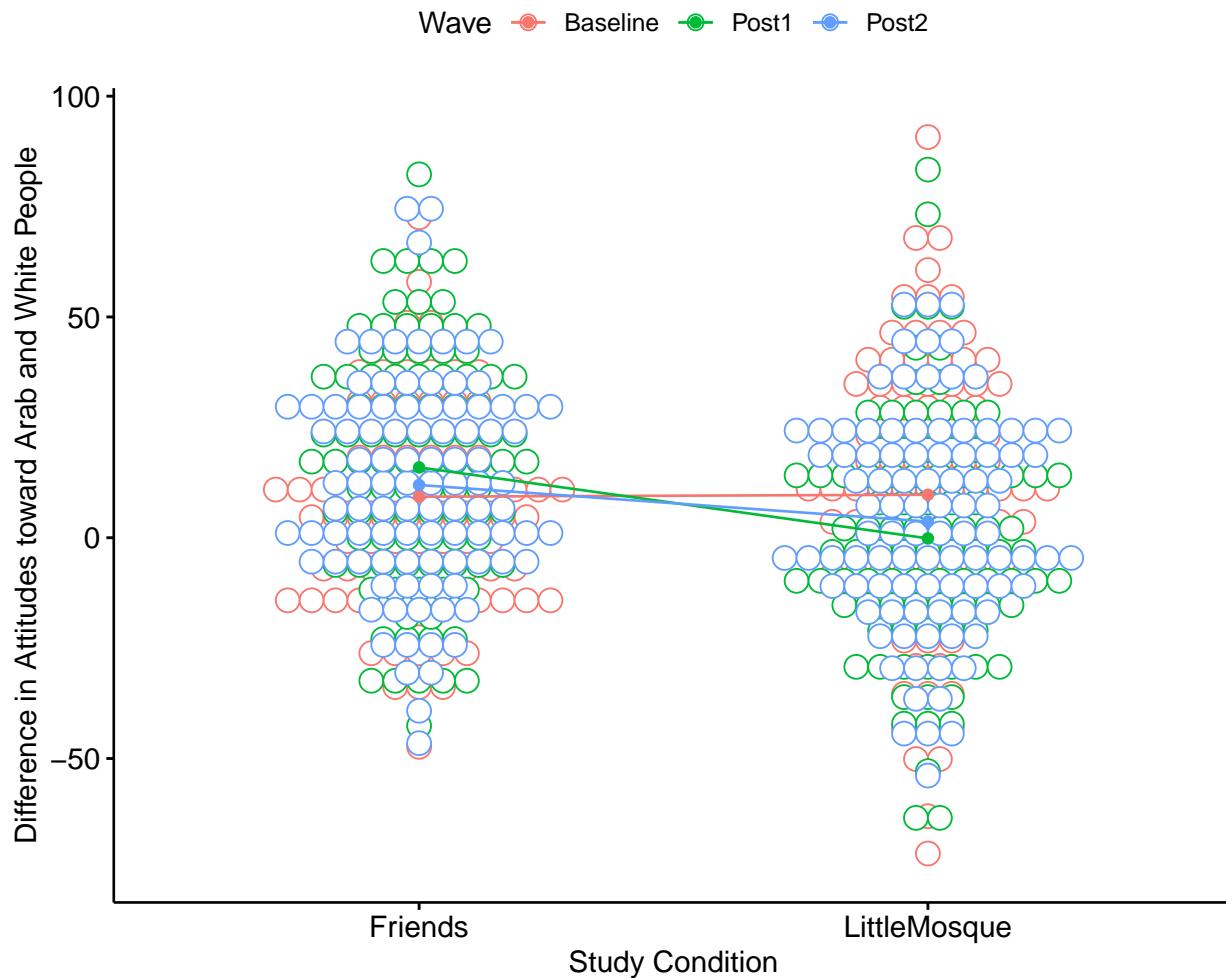


```
# add this for a different color palette: palette = c('#00AFBB',
# '#E7B800')
```

We can reverse this to see if it assists with our conceptualization.

```
ggpubr::gglue(Murrar_df, x = "COND", y = "Diff", color = "Wave", xlab = "Study Condition",
ylab = "Difference in Attitudes toward Arab and White People", add = c("mean_se",
"dotplot"), title = "Lineplots: Wave within Condition")
```

Lineplots: Wave within Condition



10.4 Working the Mixed Design ANOVA with R packages

10.4.1 Exploring data and testing assumptions

We begin the 2x3 mixed design ANOVA with a preliminary exploration of the data and testing of the assumptions. Here's where we are on the workflow:

There are several critical assumptions in factorial ANOVA:

- Cases represent random samples from the populations
 - This is an issue of research design
 - Although we see ANOVA used (often incorrectly) in other settings, ANOVA was really designed for the random clinical trial (RCT).
- The DV is continuously scaled (i.e., assessed on an interval or ratio scale); this is a matter of research design.

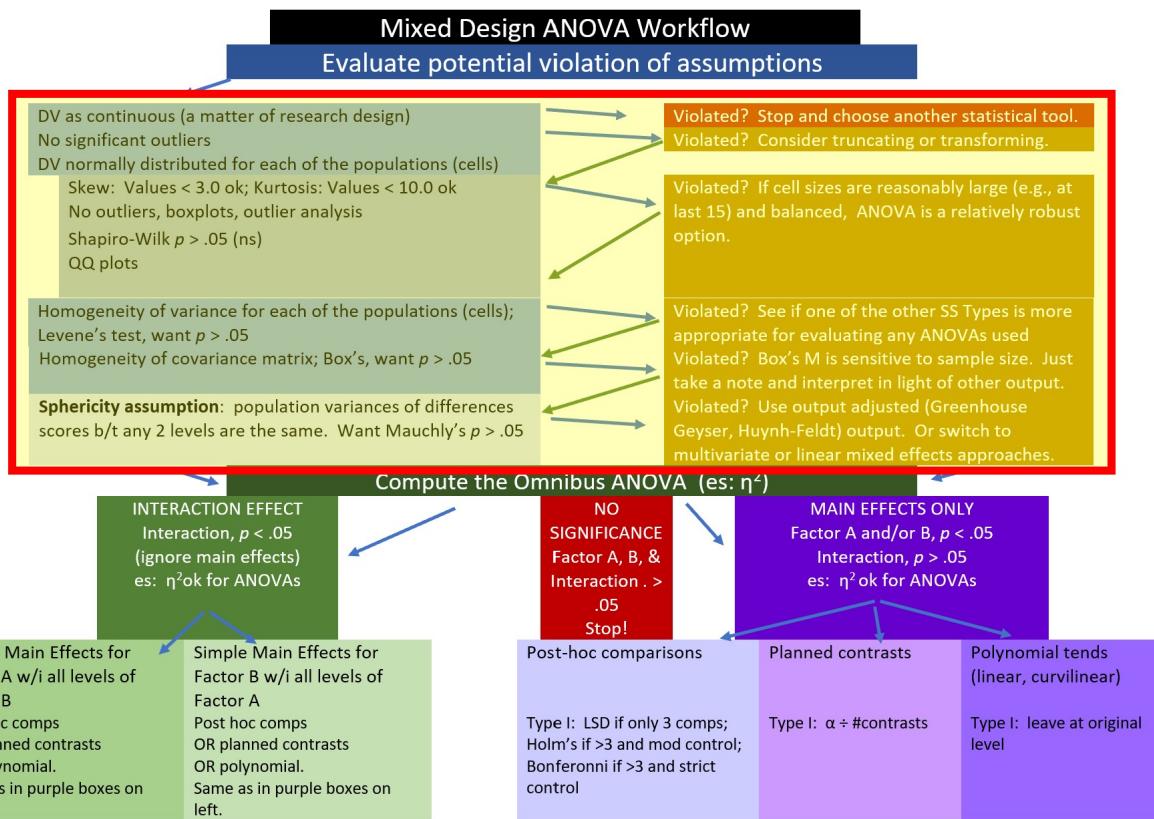


Figure 10.4: Image of the workflow showing that we are on the “Evaluating assumptions” portion

- The DV is normally distributed for each of the populations
 - that is, data for each cell (representing the combinations of each factor) is normally distributed.
- Population variances of the DV are the same for all cells (i.e., homogeneity of variance assumption)
 - When cell sizes are not equal, ANOVA not robust to this violation and we cannot trust F ratio.
- If the repeated measures factor has three or more levels the population variance of difference scores computed between any two levels of a within-subjects factor is the same value regardless of which two levels are chosen; termed the **sphericity assumption**. This assumption is
 - akin to compound symmetry (both variances across conditions are equal).
 - akin to the homogeneity of variance assumption in between-group designs.
 - sometimes called the homogeneity-of-variance-of-differences assumption.
 - statistically evaluated with *Mauchly's test*. If Mauchly's $p < .05$, there are statistically significant differences. The *anova_test()* function in the *rstatix* package reports Mauchly's and two alternatives to the traditional F that correct the values by the degree to which the sphericity assumption is violated.
- The covariance matrix of the DV is the same for all levels of the between-subjects factors (i.e., homogeneity of covariance matrix).

10.4.1.1 Is the dependent variable normally distributed?

10.4.1.1.1 Are skew and kurtosis levels concerning? Our analysis will use the difference score (Diff) as the dependent variable. Let's inspect values of skew and kurtosis for this variable in its combinations of wave and condition.

```
psych::describeBy(Diff ~ Wave + COND, data = Murrar_df, type = 1, mat = TRUE)
```

	item	group1	group2	vars	n	mean	sd	median	trimmed
Diff1	1	Baseline	Friends	1	98	9.3064898	23.90867	8.804	8.9906625
Diff2	2	Post1	Friends	1	98	15.9261327	26.41789	16.191	16.2309375
Diff3	3	Post2	Friends	1	98	11.9540102	23.33602	10.882	11.8340000
Diff4	4	Baseline	LittleMosque	1	95	9.7331158	30.51895	10.797	10.5544156
Diff5	5	Post1	LittleMosque	1	95	-0.1486632	26.96858	-1.280	-0.9402727
Diff6	6	Post2	LittleMosque	1	95	3.6704737	23.66524	1.860	3.7857403
				mad	min	max	range	skew	kurtosis
Diff1	24.48143	-47.342	72.565	119.907	0.17873477	-0.30140294	2.415140		
Diff2	29.77135	-42.598	82.288	124.886	-0.04684904	-0.52430090	2.668609		
Diff3	24.59189	-46.528	75.014	121.542	0.13445276	0.12837311	2.357294		
Diff4	30.90480	-71.510	90.737	162.247	-0.20457520	0.03318234	3.131178		
Diff5	23.93213	-65.259	83.367	148.626	0.33344683	0.62510747	2.766918		
Diff6	25.26054	-53.856	55.264	109.120	-0.06578811	-0.36855661	2.428002		
				se					

```
# Note. Recently my students and I have been having intermittent
# struggles with the describeBy function in the psych package. We
# have noticed that it is problematic when using .rds files and when
# using data directly imported from Qualtrics. If you are having
# similar difficulties, try uploading the .csv file and making the
# appropriate formatting changes.
```

Our values of skew and kurtosis are well within the limits [Kline, 2016a] of a normal distribution.

- skew: $< |3|$; the highest skew value in our data is 0.32
- kurtosis: $< |10|$; the highest kurtosis value in our data is $|.57|$

10.4.1.1.2 Are the model residuals normally distributed? We can formally investigate the normality assumption with the Shapiro-Wilk test. In the case of multiple factors (such as is the case in mixed design ANOVA), the assumption requires a normal distribution in each combination of these levels (e.g., difference scores at baseline for participants in the Friends condition). In this lesson's 3 x 2 ANOVA, there are six such combinations. A cell-level analysis (i.e., testing for normal distributions within each combination of factor levels) has been demonstrated in [one-way ANOVA](#) and independent [t-test](#) lessons. To the degree that there are many factorial combinations (and therefore, cells), this approach becomes unwieldy to calculate, interpret, and report. Further, the cell-level analysis of normality is only appropriate when there are a low number of levels/groupings and there are many data points per group. As designs increase in complexity, researchers turn to the model-based option for assessing normality.

To do this, we first create an object that tests our research model. Because the model-based approach to calculating the Shapiro-Wilk test of normality requires an object created by the `aov()` function from base R, I will quickly run this. For the moment, we will only peek at the ANOVA results to make sure it ran, but will save the interpretation until later.

```
Mixed_diff <- aov(Diff ~ COND * Wave, Murrar_df)
summary(Mixed_diff)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
COND	1	9209	9209	13.723	0.000232 ***						
Wave	2	320	160	0.238	0.788240						
COND:Wave	2	6574	3287	4.898	0.007774 **						
Residuals	573	384530	671								

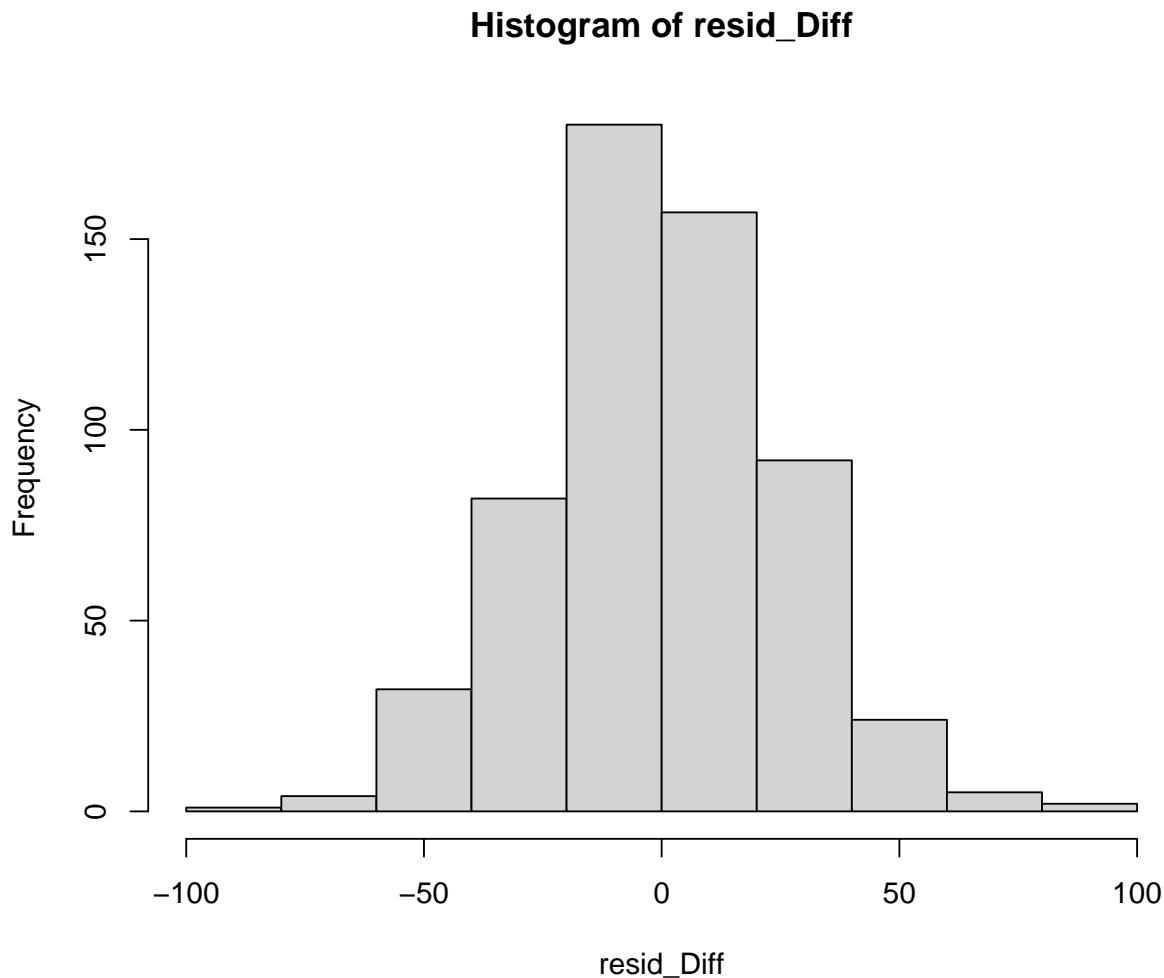
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

From this object we can extract the residuals.

```
# creates object of residuals
resid_Diff <- residuals(Mixed_diff)
```

We can visually inspect the distribution of the residuals with a couple of plots.

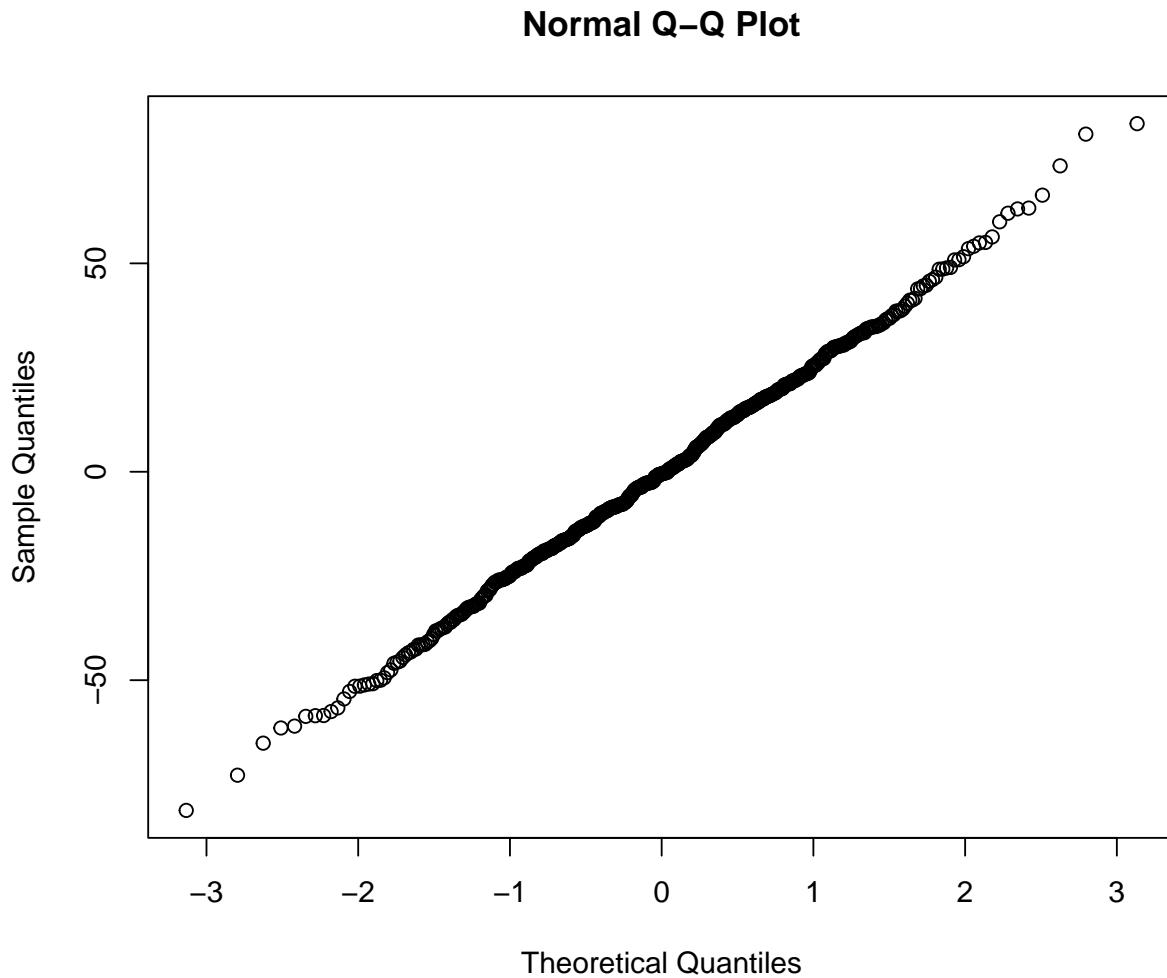
```
hist(resid_Diff)
```



So far so good – our distribution of *residuals* (i.e., what is leftover after the model is applied) resembles a normal distribution.

The Q-Q plot provides another view. The dots represent the residuals. When they are relatively close to the line they not only suggest good fit of the model, but we know they are small and evenly distributed around zero (i.e., normally distributed).

```
qqnorm(resid_Diff)
```



Finally, we can formally evaluate whether or not the distribution of residuals is statistically significantly different from a normal distribution with a Shapiro test. We want the associated p value to be greater than 0.05.

```
shapiro.test(resid_Diff)
```

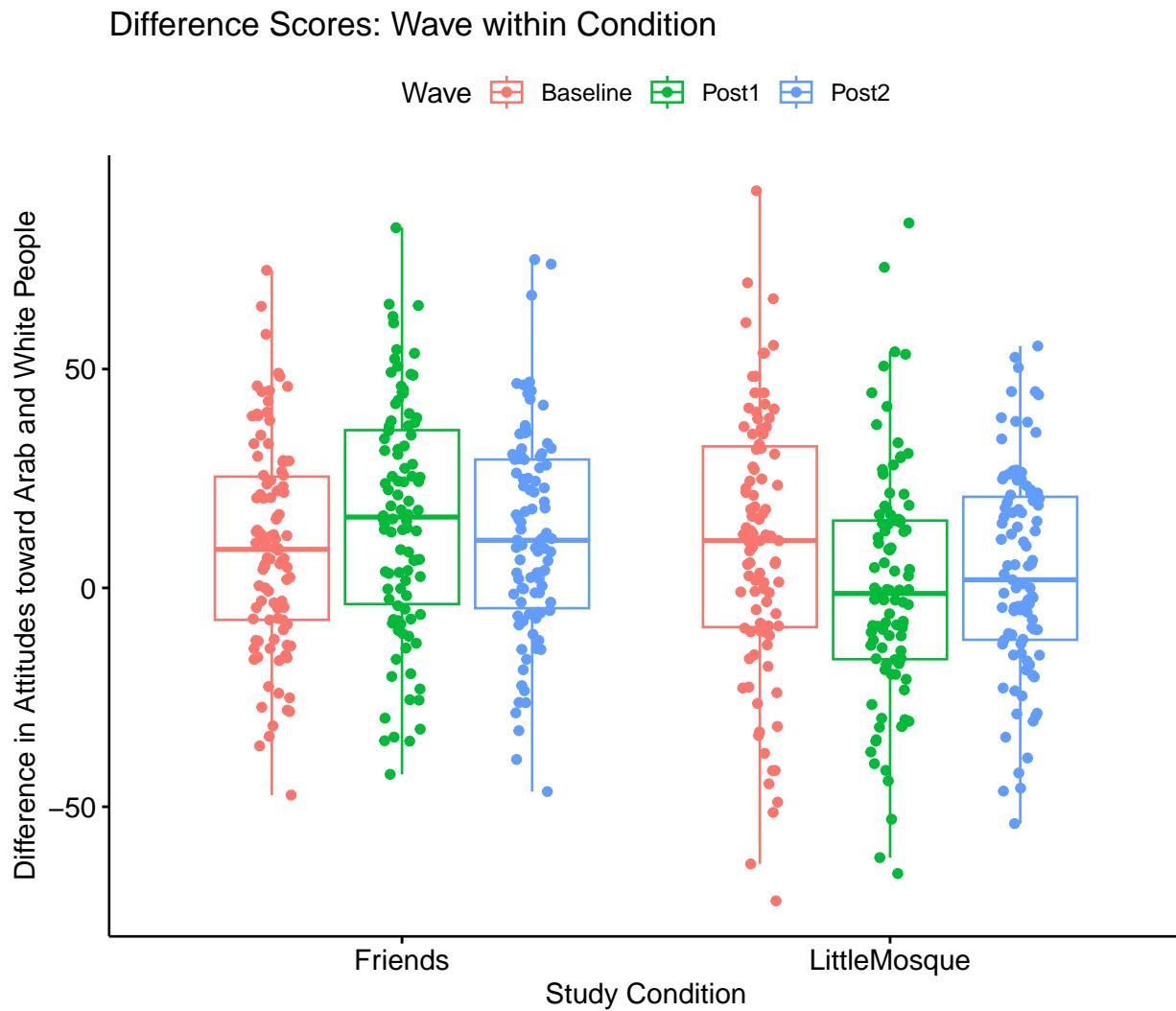
```
Shapiro-Wilk normality test
data: resid_Diff
W = 0.99869, p-value = 0.9526
```

The non-significant p value indicates that the distribution of our model residuals are not statistically significantly different from a normal distribution ($W = 0.999, p = 0.953$).

10.4.1.1.3 Is there evidence of outliers? The boxplot is one common way for identifying outliers. The boxplot uses the median and the lower (25th percentile) and upper (75th percentile)

quartiles. The difference between Q3 and Q1 is the *interquartile range* (IQR). Let's revisit one of our boxplots to see if there are any dots above the whiskers.

```
ggpubr::ggboxplot(Murrar_df, x = "COND", y = "Diff", color = "Wave", xlab = "Study Condition",
  ylab = "Difference in Attitudes toward Arab and White People", add = "jitter",
  title = "Difference Scores: Wave within Condition")
```



The distributions look relatively normal with the mean well-centered. Given that we simulated the data from means and standard deviations, this is somewhat expected.

Outliers are generally identified when values fall outside these lower and upper boundaries. In the short formulas below, IQR is the *interquartile range* (i.e., the middle 50%, the distance of the box):

- $Q1 - 1.5 \times IQR$
- $Q3 + 1.5 \times IQR$

Extreme values occur when values fall outside these boundaries:

- Q1 - 3xIQR
- Q3 + 3xIQR

Using the `rstatix::identify_outliers` function we can identifier outliers, doubly grouped by our predictor variables.

```
Murrar_df %>%
  group_by(Wave, COND) %>%
  rstatix::identify_outliers(Diff)
```

	Wave	COND	rowID	caseID	AttArab	AttWhite	Diff	is.outlier	is.extreme
	<fct>	<fct>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<lgl>	<lgl>
1	Baseline	LittleMosq~	107	107	100	28.5	-71.5	TRUE	FALSE
2	Post1	LittleMosq~	297	104	16.6	100	83.4	TRUE	FALSE
3	Post1	LittleMosq~	315	122	26.8	100	73.2	TRUE	FALSE
4	Post1	LittleMosq~	337	144	97.4	32.2	-65.3	TRUE	FALSE

While we have some outliers (where “is.outlier” = “TRUE”), none are extreme (where “is.outlier” = “FALSE”). We’ll keep these in mind as we continue to evaluate the data.

If I had extreme outliers, I would individually inspect them. Especially if something looked awry (e.g., erratic responding extreme scores across variables) I might consider deleting them. In this mixed design ANOVA, a case will only be included in the final analysis if all three waves of data are present. Therefore, if we decided to remove case at rowID = 337, we would need to remove all three cases. That is, we would need to remove it by caseID = 144. This would decrease the number of rows/observations by 3.

Let’s say that, after very careful consideration, we decided to remove the caseID = 144. We could use `dplyr::filter()` to do so. In this code, the `filter()` function locates all the cases where caseID = 144. The exclamation point that precedes the equal sign indicates that the purpose is to remove the case.

```
# Murrar_df <- dplyr::filter (Murrar_df, caseID != '144')
```

Once executed, we can see that this case is no longer in the dataframe. Because each person had three observations, this reduces the number of observations by 3. Although I demonstrated this in the accompanying lecture, I have hashtagged out the command because I would not delete the case. If you already deleted the case, you can return the hashtag and re-run all the code up to this point.

10.4.1.2 Homogeneity of variance assumption

Because there is a between-subjects variable, we need need to evaluate the homogeneity of variance assumption. As before, we can use the Levene’s test with the `rstatix::levene_test()` function. Considering each of the comparisons of condition within wave, there is no instance where we violate the assumption.

```
Murrar_df %>%
  group_by(Wave) %>%
  rstatix::levene_test(Diff ~ COND)
```

```
# A tibble: 3 x 5
  Wave      df1    df2 statistic     p
  <fct>    <int> <int>    <dbl>   <dbl>
1 Baseline     1    191     3.97  0.0477
2 Post1        1    191     0.141  0.708
3 Post2        1    191     0.107  0.744
```

Levene's test indicated a violation of this assumption between the Friends and Little Mosque conditions at baseline ($F[1, 191] = 3.973, p = .047$). However, there was no indication of assumption violation at post1 ($F[1, 191] = 0.141, p = .708$), and post2 ($F[1, 191] = 0.107, p = .743$) waves of the design.

10.4.1.3 Assumption of homogeneity of covariance matrices

In this multivariate sample, the Box's M test evaluates if two or more covariance matrices are homogeneous. Like other tests of assumptions, we want a non-significant test result (i.e., where $p > .05$). Box's M has some disadvantages. Box's M has low power in small sample sizes and is overly sensitive in large sample sizes. We would unlikely make a decision about our data with Box's M alone. Rather, we consider it along with our dashboard of diagnostic screeners.

```
rstatix::box_m(Murrar_df[, "Diff", drop = FALSE], Murrar_df$COND)
```

```
# A tibble: 1 x 4
  statistic p.value parameter method
  <dbl>     <dbl>     <dbl> <chr>
1       3.21  0.0732          1 Box's M-test for Homogeneity of Covariance Matric~
```

Box's M indicated no violation of the homogeneity of covariance matrices assumption ($M = 3.209, p = .073$).

10.4.1.4 APA style writeup of assumptions

At this stage we are ready to draft the portion of the APA style writeup that evaluates the assumptions.

Mixed design ANOVA has a number of assumptions related to both the within-subjects and between-subjects elements. Data are expected to be normally distributed at each level of design. There was no evidence of skew (all values were at or below the absolute value of 0.32) or kurtosis (all values were below the absolute value of .55; [Kline, 2016a]). Similarly, results of the Shapiro-Wilk normality test (applied to the residuals from the

factorial ANOVA model) suggested that model residuals did not differ significantly from a normal distribution ($W = 0.999, p = 0.953$). Visual inspection of boxplots for each wave of the design, assisted by the `rstatix::identify_outliers()` function (which reports values above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$, where IQR is the interquartile range) indicated some outliers, but none at the extreme level. Because of the between-subjects aspect of the design, the homogeneity of variance assumption was evaluated. Levene's test indicated a violation of this assumption between the Friends and Little Mosque conditions at baseline ($F[1, 191] = 3.973, p = .047$). However, there was no indication of assumption violation at post1 ($F[1, 191] = 0.141, p = .708$), and post2 ($F[1, 191] = 0.107, p = .743$) waves of the design. Further, Box's M-test ($M = 3.209, p = .073$) indicated no violation of the homogeneity of covariance matrices. *LATER WE WILL ADD INFORMATION ABOUT THE SPHERICITY ASSUMPTION.*

10.4.2 Omnibus ANOVA

Having evaluated the assumptions (excepting sphericity) we are ready to move to the evaluation of the omnibus ANOVA. This next step produces both the omnibus test as well as testing the sphericity assumption. Conceptually, evaluating the sphericity assumption precedes the omnibus; procedurally these are evaluated simultaneously. The figure also reflects that decisions related to follow-up are dependent upon the significance of the main and omnibus effects.

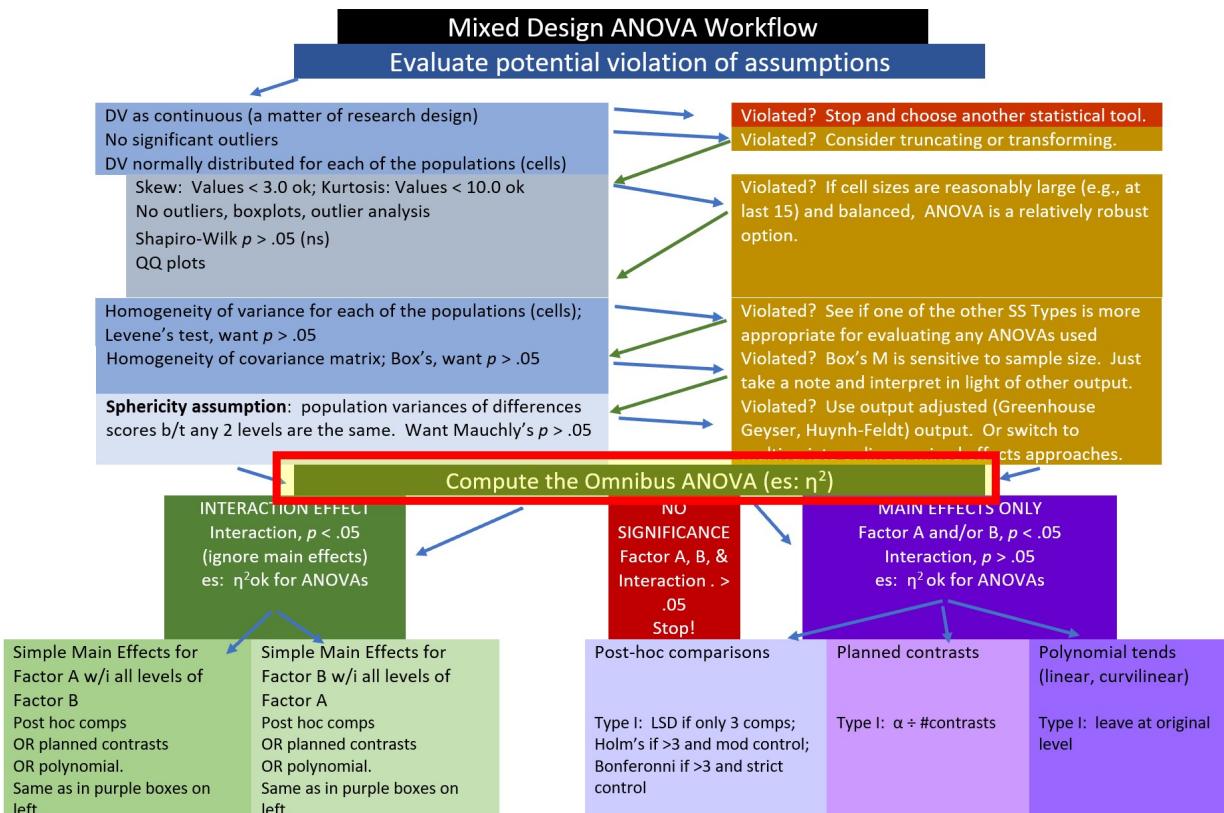


Figure 10.5: Image of the workflow showing that we are at the “Compute the Omnibus ANOVA” step

The `rstatix` package is a wrapper for the `car` package. Authors of *wrappers* attempt to streamline

a more complex program to simplify the input needed and maximize the output produced for the typical use-cases.

If we are ever confused about a function, we can place a question mark in front of it. It will summons information and, if the package is in our library, let us know to which package it belongs and open the instructions that are embedded in R/R Studio.

```
#?anova_test
```

In the code below the identification of the data, DV, between, and within variables are likely to be intuitive. The within-subjects identifier (*wid*) is the person-level ID that assists the statistic in controlling for the dependency introduced by the repeated-measures factor.

```
# Murrar_df is our df, Diff is our df, wid is the caseID between is
# the between-subjects variable, within is the within subjects
# variable
Diff_2way <- rstatix::anova_test(data = Murrar_df, dv = Diff, wid = caseID,
    between = COND, within = Wave, detailed = TRUE)
Diff_2way
```

ANOVA Table (type III tests)

\$ANOVA

	Effect	DFn	DFd	SSn	SSd	F	p	p<.05	ges
1	(Intercept)	1	191	40911.756	133769.4	58.415	0.000000000001	*	0.096000
2	COND	1	191	9209.127	133769.4	13.149	0.000369000000	*	0.023000
3	Wave	2	382	359.022	250761.1	0.273	0.761000000000		0.000933
4	COND:Wave	2	382	6574.365	250761.1	5.008	0.007000000000	*	0.017000

\$`Mauchly's Test for Sphericity`

	Effect	W	p	p<.05
1	Wave	0.99	0.369	
2	COND:Wave	0.99	0.369	

\$`Sphericity Corrections`

	Effect	GGe	DF[GG]	p[GG]	p[GG]<.05	HFe	DF[HF]	p[HF]	p[HF]<.05
1	Wave	0.99	1.98, 378.06	0.759		1	2, 382	0.761	
2	COND:Wave	0.99	1.98, 378.06	0.007	*	1	2, 382	0.007	*

10.4.2.1 Checking the sphericity assumption

We continue our evaluation of statistical assumptions by examining the Mauchly's test for sphericity. Fortunately, the Mauchly's is included with the results of the omnibus ANOVA. The sphericity assumption becomes important when there are three or more levels in the repeated measures factor. The assumption requires that the variance of difference scores computed between any two levels of a within-subjects factor is the same value regardless of which two levels are chosen. As with so many of our statistical tests of assumptions, we want the *p* value to be $> .05$.

Mauchly's test (in the middle of the output) reports the result. Because our Wave variable has three levels and appears in the main and interaction effects. The results are identical and we didn't violate the sphericity assumption.

- main effect for Wave: $W = .99, p = .369$
- interaction effect for COND:Wave: $W = .99, p = .369$

We will be able to add this statement to our assumptions write-up:

Mauchly's test indicated no violation of the sphericity assumption for the main ($W = .99, p = .369$) and interaction ($W = .99, p = .369$) effects.

If the p value associated with Mauchly's test had been less than .05, we could have used one of the two options (Greenhouse Geyser/GGe or Huynh-Feldt/HFe). In each of these an epsilon value provides an adjustment to the degrees of freedom used in the estimation of the p value.

When there are concerns about sphericity violations, there is also a multivariate approach that does not require the assumption of sphericity. The *rstatix* package does include this analysis. There is also an option to use a multivariate approach when ANOVA designs include a repeated measures factor. In the [appendix](#), I included an example of the multivariate approach with a one-way repeated measures design using with the *car* package.

10.4.2.2 Interpreting the omnibus results

We are interested in the output reported in rows 2, 3, 4. These include the:

- main effect for condition: $F(1, 191) = 13.149, p < .001, \eta^2 = 0.023$
- main effect for wave: $F(2, 382) = 0.273, p = .761, \eta^2 = 0.001$
- condition:wave interaction effect: $F(2, 382) = 5.008, p = 0.007, \eta^2 = 0.017$

Results of the omnibus ANOVA indicated a significant main effect for condition ($F[1, 191] = 13.149, p < .001, \eta^2 = 0.023$), a non-significant main effect for wave ($F[2, 382] = 0.273, p = .761, \eta^2 = 0.001$), and a significant interaction effect ($F[2, 382] = 5.008, p = 0.007, \eta^2 = 0.017$).

In the output, the column labeled “ges” provides the value for the effect size, η^2 . Recall that *eta-squared* is one of the most commonly used measures of effect. It refers to the proportion of variability in the dependent variable/outcome that can be explained in terms of the independent variable/predictor. Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

You may see different values (.02, .13, .26) offered as small, medium, and large – these values are used when multiple regression is used. A useful summary of effect sizes, guide to interpreting their magnitudes, and common usage can be found [here \[Watson, 2020\]](#).

With a significant interaction effect, we would focus on interpreting one or both of the simple main effects. Let's first look at the simple main effect of condition within wave option.

10.4.3 Follow-up to Omnibus Tests

10.4.3.1 Planning for the management of Type I error

Controlling for Type I error can depend, in part, on the design of the follow-up tests that are planned, and the number of pairwise comparisons that follow.

In the first option, the examination of the simple main effect of condition within wave results in only three pairwise comparisons. In this case, I will use the traditional Bonferroni. Why? Because there are only three post omnibus analyses and the traditional Bonferroni's more restrictive control is less likely to be problematic.

In the second option, the examination of the simple main effect of wave within condition results in the potential comparison of nine pairwise comparisons. If we used a traditional Bonferroni and divided .05/6, the p value for each comparison would need to be less than 0.008. Most would agree that this is too restrictive.

```
.05/6
```

```
[1] 0.008333333
```

The Holm's sequential Bonferroni [Green and Salkind, 2017c] offers a middle-of-the-road approach (not as strict as .05/6 with the traditional Bonferroni; not as lenient as “none”) to managing Type I error.

If we were to hand-calculate the Holm's, we would rank order the p values associated with the 6 comparisons in order from lowest (e.g., 0.016) to highest (e.g., 1.000). The first p value is evaluated with the most strict criterion (.05/6; the traditional Bonferroni approach). Then, each successive comparison calculates the p value by using the number of *remaining* comparisons as the denominator (e.g., .05/5, .05/4, .05/3). As the p values increase and the alpha levels relax, there will be a cut-point where remaining comparisons are not statistically significant.

Luckily, most R packages offer the Holm's sequential Bonferroni as an option. The algorithm in the package rearranges the mathematical formula and produces a p value that we can interpret according to the traditional values of $p < .05$, $p < .01$ and $p < .001$. I will demonstrate use of Holm's in the examination of the simple main effect of ethnicity of rater within photo stimulus.

10.4.4 Simple main effect of condition within wave

The figure reflects our path in the workflow. In the presence of a significant interaction effect we could choose from a variety of follow-up tests.

If we take this option we follow up with three t -tests. Why? Because within each wave (the repeated measures factor), there are two conditions. Statistically, we could use one-way ANOVAs, however, using the `rstatix::t_test` function will allow us to map statistically significant differences on a subsequent figure. Here are the comparisons:

- comparison of Friends and Little Mosque within the baseline wave
- comparison of Friends and Little Mosque within the post1 wave

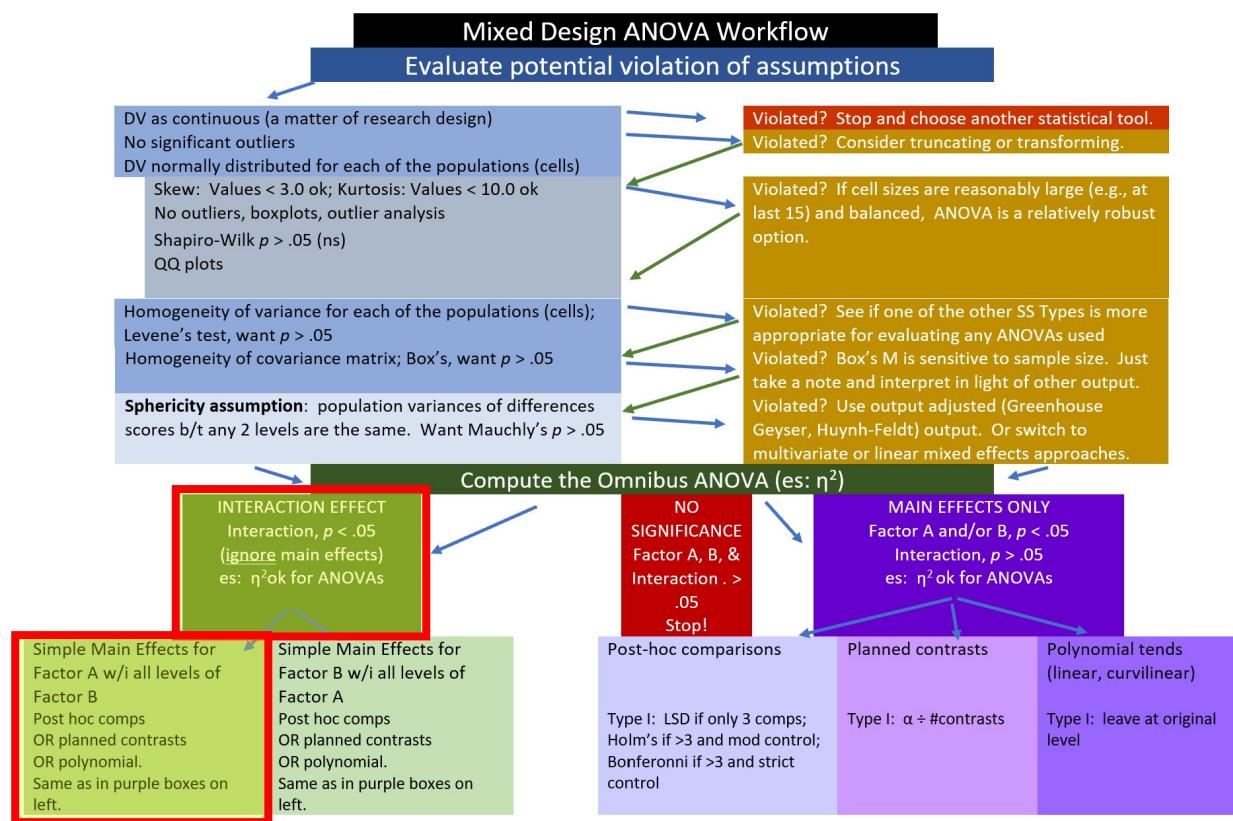


Figure 10.6: Image of the workflow showing that we are at the “Simple Main Effects for Factor A within all levels of Factor B” step

- comparison of Friends and Little Mosque within the post2 wave

We will control for Type I error by requesting the traditional Bonferroni.

Note that the function is *rstatix::t_test*. Recall that *t*-tests can be for independent samples or paired samples. The *rstatix::t_test* is for independent samples and therefore appropriate for comparing the Friends to LittleMosque conditions (from the between-groups factor, COND).

Finally, I have also specified *detailed=TRUE*. This permits me to see the means (i.e., estimate and estimate 1) for the Friends and Little Mosque conditions, the sample sizes for each (i.e., n1, n2), the confidence interval around the true difference between the means (i.e, conf.low, conf.high), and that we used a two-sided *t*-test.

```
SimpleWave <- Murrar_df %>%
  group_by(Wave) %>%
  rstatix::t_test(Diff ~ COND, detailed = TRUE, p.adjust.method = "bonferroni") %>%
  rstatix::add_significance()
# rstatix::adjust_pvalue(method = 'bonferroni') #this displays the
# adjusted Bonferroni values
SimpleWave
```

	Wave	estimate	estimate1	estimate2	.y.	group1	group2	n1	n2	statistic
	<fct>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<int>	<int>	<dbl>
1	Baseli~	-0.427	9.31	9.73	Diff	Frien~	Littl~	98	95	-0.108
2	Post1	16.1	15.9	-0.149	Diff	Frien~	Littl~	98	95	4.18
3	Post2	8.28	12.0	3.67	Diff	Frien~	Littl~	98	95	2.45

i 7 more variables: p <dbl>, df <dbl>, conf.low <dbl>, conf.high <dbl>,
method <chr>, alternative <chr>, p.signif <chr>

We can begin to assemble the results. At each wave we are comparing the Friends and Little Mosque conditions.

Baseline: $t(178.04) = -0.108, p = 0.914$ Post1: $t(190.49) = 4.182, p < 0.001$ Post2: $t(190.61) = 2.448, p = 0.015$

You might wonder about effect sizes. The *rstatix::pairwise_t_test* does not produce any. A commonly used effect size for *t* tests is the Cohen's *d*. This gives the degree to which means differ in the metric of standard deviations. Values of .02, .05, and .08 are interpreted as small, moderate, and large, respectively. We can use *rstatix::cohens_d*. A helpful feature of this function is that interpretive language is provided.

```
SimpleWave_d <- Murrar_df %>%
  group_by(Wave) %>%
  rstatix::cohens_d(Diff ~ COND)
SimpleWave_d
```

```
# A tibble: 3 x 8
```

.y.	group1	group2	effsize	Wave	n1	n2	magnitude	
*	<chr>	<chr>	<dbl>	<fct>	<int>	<int>	<ord>	
1	Diff	Friends	LittleMosque	-0.0156	Baseline	98	95	negligible
2	Diff	Friends	LittleMosque	0.602	Post1	98	95	moderate
3	Diff	Friends	LittleMosque	0.352	Post2	98	95	small

We can update our t strings with this information.

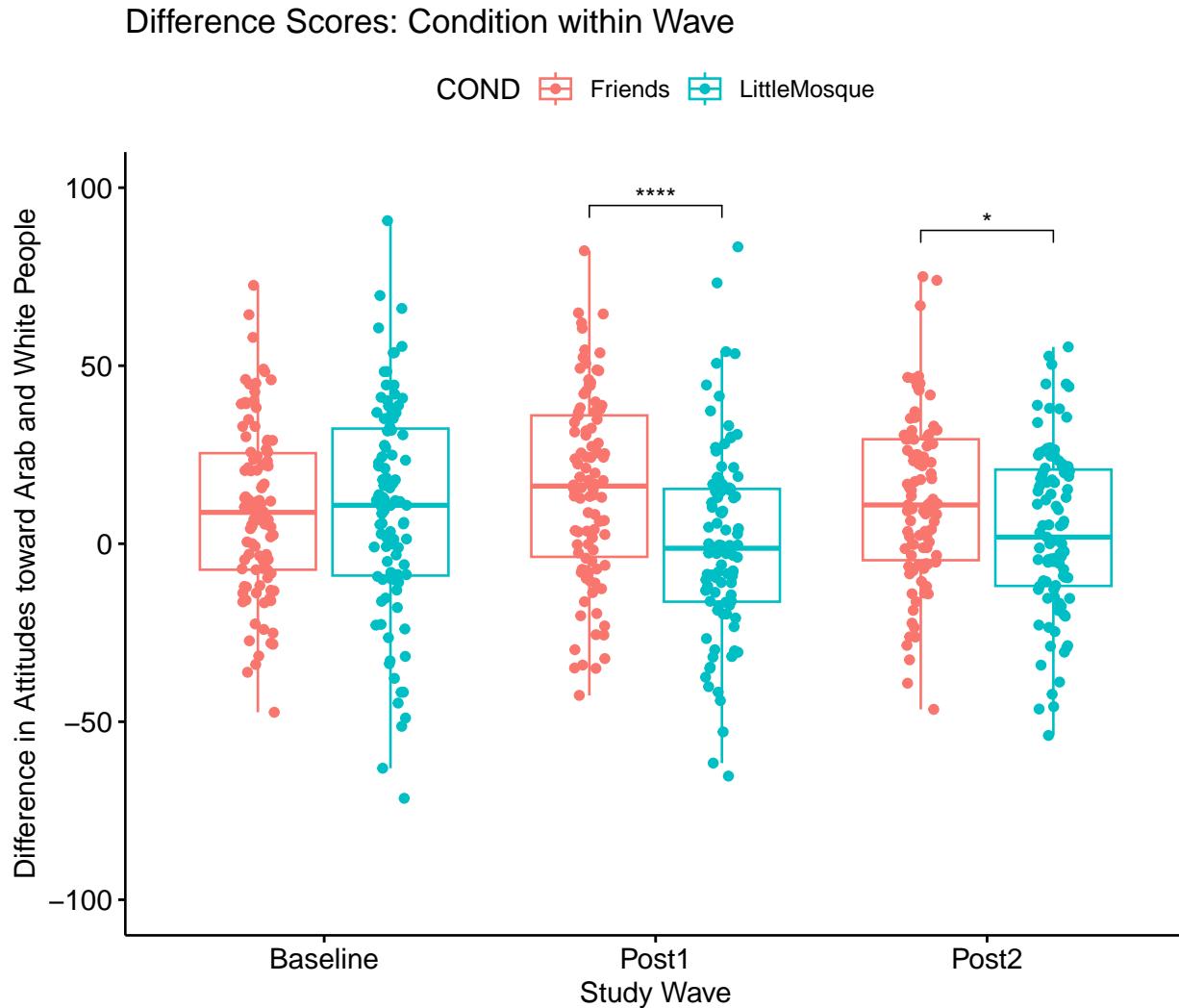
Baseline: $t(178.04) = -0.108, p = 0.914, d = -0.016$ Post1: $t(190.49) = 4.182, p < 0.001, d = 0.602$
Post2: $t(190.61) = 2.448, p = 0.015, d = 0.352$

Producing a figure can be helpful in conceptualizing what we have first done. Because we used the *rstatix* functions, we can easily integrate them into our *ggpubr::ggbboxplot()*. Let's re-run the version of the boxplot where "Wave" is on the x-axis (and, is therefore our grouping variable). Because I want the data to be as true-to-scale as possible, I have added the full, potential, range of the y axis through the *ylim* argument. In order to update the ggbboxplot, we will need to save it as an option. My object name represents the "Condition within Wave" simple main effect.

```
# Although we have used this code this before, I respecified the
# basic figure here.
CNDwiWV <- ggpubr::ggbboxplot(Murrar_df, x = "Wave", y = "Diff", color = "COND",
  xlab = "Study Wave", ylab = "Difference in Attitudes toward Arab and White People",
  add = "jitter", ylim = c(-100, 100), title = "Difference Scores: Condition within Wave")

# This updates the SimpleWave object (which holds the t-tests) to
# include plotting information about the xy positions
SimpleWave <- SimpleWave %>%
  rstatix::add_xy_position(x = "Wave")
# SimpleWave #unhashtag if you want to see the plotting information

# Now we update the figure to include the significance bars and stars
# label = 'p.adj.signif' points to the values in the rstatix output
# from the pairwise_t_test tip.length is the amount of downward
# pointing on the lines that hold the p-values hide.ns=TRUE
# suppresses a bar over non-significant comparisons y.position
# adjusts the significance bars up and down, I pushed them up
CNDwiWV <- CNDwiWV + ggpubr::stat_pvalue_manual(SimpleWave, label = "p.signif",
  tip.length = 0.02, hide.ns = TRUE, y.position = c(95, 88))
CNDwiWV
```



If we were to write up this simple main effect of condition within wave:

We followed the significant interaction effect with an evaluation of simple main effects of condition within wave. A traditional Bonferroni was used to manage Type I error [Green and Salkind, 2017c]. There was a non-statistically significant difference between conditions at baseline ($t[178.04] = -0.108, p = 0.914, d = -0.016$) However there were statistically significant differences at post1 ($t[190.49] = 4.182, p < 0.001, d = 0.602$) and post2 $t[190.61] = 2.448, p = 0.015, d = 0.352$. We note that the effect size at post1 approached a moderate size; the effect size at post2 was small.

10.4.5 Simple main effect of wave within condition

Alternatively, we could evaluate the simple main effect of wave within condition. The figure reflects our path along the workflow.

Because there are three waves within each condition, would start with two one-way ANOVAs and then follow each of those with pairwise comparisons. First, the one-way repeated measures ANOVAs:

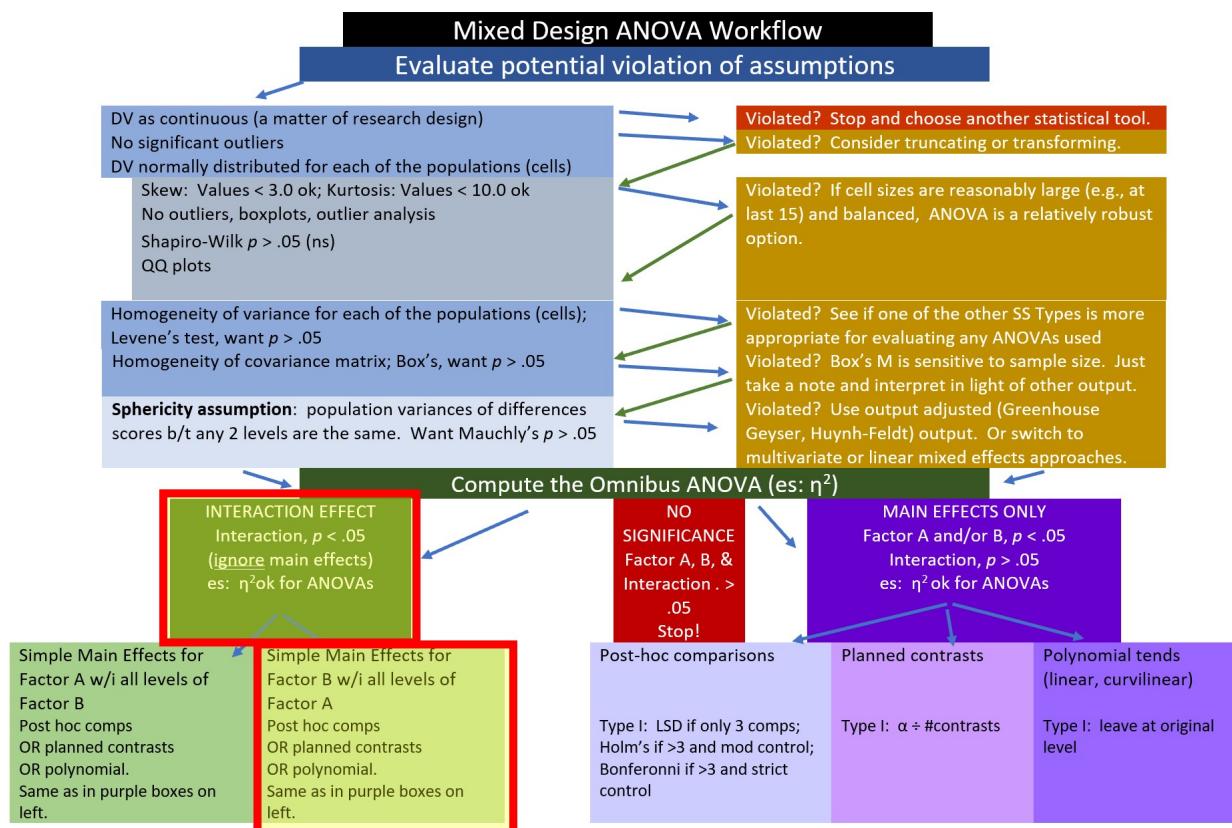


Figure 10.7: Image of the workflow showing that we are at the “Simple Main Effects for Factor B within all levels of Factor A” step

- comparison of baseline, post1, and post2 within the Friends condition
- comparison of baseline, post1, and post2 within the Little Mosque condition

```
SimpleCond <- Murrar_df %>%
  group_by(COND) %>%
  rstatix::anova_test(dv = Diff, wid = caseID, within = Wave) %>%
  rstatix::get_anova_table() %>%
  rstatix::adjust_pvalue(method = "none")
SimpleCond
```

```
# A tibble: 2 x 9
  COND      Effect   DFn   DFD      F      p `p<.05`    ges p.adj
  <fct>     <chr>   <dbl> <dbl> <dbl> <dbl> <chr>   <dbl> <dbl>
1 Friends   Wave     2     194  1.76  0.175   ""     0.012  0.175
2 LittleMosque Wave     2     188  3.39  0.036   "*"    0.022  0.036
```

Below are the F strings for the one-way ANOVAs the followed the omnibus, mixed design, ANOVA:

- Friends: $F(2, 194) = 1.759, p = 0.175, \eta^2 = 0.012$ (effect size indicates a small effect)
- Little Mosque: $F(2, 188) = 3.392, p = 0.036, \eta^2 = 0.072$ (a moderate effect size)

Because each of these one-way ANOVAs has three levels, we need to follow with pairwise comparisons. However, because the oneway repeated measures ANOVA was non-significant for the Friends condition, we only need to report them for the Little Mosque condition. As you can see we generally work our way down to comparing chunks to each other to find the source(s) of significant differences.

You will notice that we are saving the results of the pairwise comparisons as an object. This allows us to update the object in combination with the boxplot we created earlier.

Note that in this follow-up, once we have arrived at the level of paired comparisons, we are using `rstatix::pairwise_t_test`. Because wave (within condition) is repeated measures, we including the command, `paired = TRUE`.

In order to manage Type I error, I have specified “holm.” The Holm’s sequential Bonferroni offers a middle-of-the-road approach (not as strict as .05/6 with the traditional Bonferroni; not as lenient as “none”) to managing Type I error.

```
pwcWVwiGP <- Murrar_df %>%
  group_by(COND) %>%
  rstatix::pairwise_t_test(Diff ~ Wave, paired = TRUE, detailed = TRUE,
                          p.adjust.method = "holm") # %>%
# select(-df, -statistic, -p) # Remove details
pwcWVwiGP
```

```
# A tibble: 6 x 16
  COND   estimate .y.   group1 group2     n1     n2 statistic      p     df conf.low
  <fct>     <dbl> <chr> <fct> <fct> <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
* <fct>    <dbl> <chr> <chr> <chr> <int> <int>    <dbl> <dbl> <dbl>    <dbl>
1 Frien~ -6.62 Diff  Basel~ Post1    98    98    -1.80  0.075    97   -13.9
2 Frien~ -2.65 Diff  Basel~ Post2    98    98    -0.793 0.43     97   -9.27
3 Frien~  3.97 Diff  Post1  Post2    98    98     1.09  0.276    97   -3.23
4 Littl~  9.88 Diff  Basel~ Post1    95    95     2.45  0.016    94   1.86
5 Littl~  6.06 Diff  Basel~ Post2    95    95     1.62  0.108    94   -1.36
6 Littl~ -3.82 Diff  Post1  Post2    95    95    -1.03  0.304    94   -11.2
# i 5 more variables: conf.high <dbl>, method <chr>, alternative <chr>,
# p.adj <dbl>, p.adj.signif <chr>
```

Consistent with the non-significant one-way repeated measures ANOVA, there were non-significant pairwise comparisons for the Friends condition.

Within the Little Mosque condition, we find a significant difference between baseline and post1 ($t[94] = 2.447, p = .049$), but non-significant differences between baseline and post2 ($t[94] = 1.621, p = .216$) and post1 and post2 ($t[94] = -1.034, p = .304$)

We can use `rstatix::cohens_d` to calculate effect sizes. In the metric of standard deviation units, values of .02, .05, and .08 are interpreted as small, moderate, and large, respectively.

```
pwcWVwiGP_d <- Murrar_df %>%
  group_by(COND) %>%
  rstatix::cohens_d(Diff~Wave)
pwcWVwiGP_d
```

```
# A tibble: 6 x 8
  .y.   group1  group2 effsize COND          n1      n2 magnitude
* <chr> <chr>    <chr> <dbl> <fct>      <int> <int> <ord>
1 Diff  Baseline Post1  -0.263 Friends      98    98 small
2 Diff  Baseline Post2  -0.112 Friends      98    98 negligible
3 Diff  Post1    Post2   0.159 Friends      98    98 negligible
4 Diff  Baseline Post1   0.343 LittleMosque  95    95 small
5 Diff  Baseline Post2   0.222 LittleMosque  95    95 small
6 Diff  Post1    Post2  -0.151 LittleMosque  95    95 negligible
```

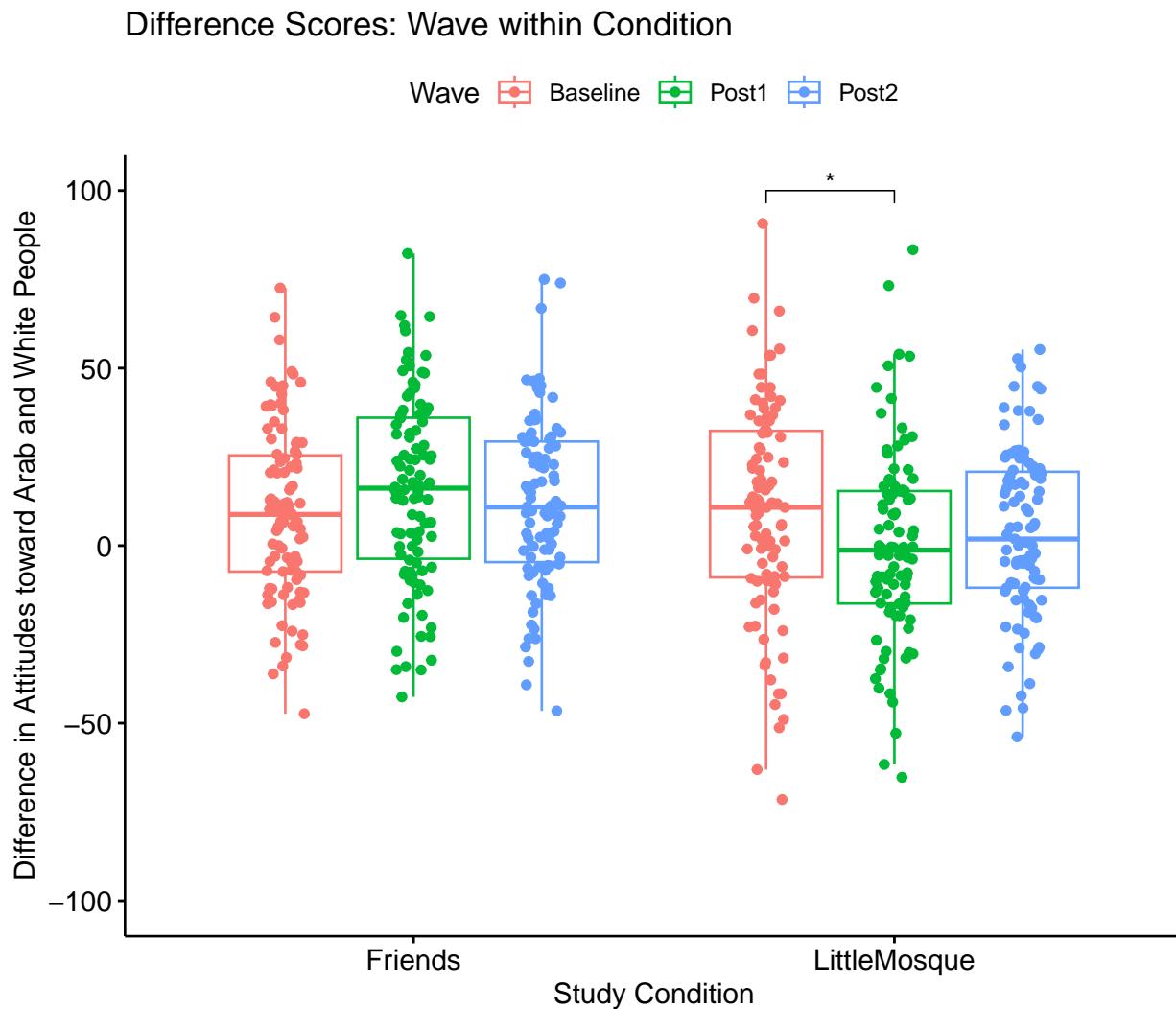
We can update our t strings with this information. That is, within the Little Mosque condition: .

Baseline versus post1: $t(94) = 2.447, p = .049, d = 0.343$, Baseline versus post2: $t(94) = 1.621, p = .216, d = 0.222$, and Post1 versus post2: ($t(94) = -1.034, p = .304, d = -0.150$)

Let's create a figure to illustrate what we've just learned.

```
# We ran this before -- grabbing again to make it clear how creating
# and updating the boxplot works
WVwiCND <- ggpubr::ggboxplot(Murrar_df, x = "COND", y = "Diff", color = "Wave",
  xlab = "Study Condition", ylim = c(-100, 100), ylab = "Difference in Attitudes toward Arab
  add = "jitter", title = "Difference Scores: Wave within Condition")
# WVwiCND
```

```
# This updates the pwcWVwiGP object (which holds the t-tests) to
# include plotting information about the xy positions
pwcWVwiGP <- pwcWVwiGP %>%
  rstatix::add_xy_position(x = "COND")
# pwcWVwiGP Diff_2way was my omnibus ANOVA object
WVwiCND <- WVwiCND + ggpubr::stat_pvalue_manual(pwcWVwiGP, label = "p.adj.signif",
  tip.length = 0.02, hide.ns = TRUE, y.position = c(100))
WVwiCND
```



If we were to write up this result:

We followed the significant interaction effect with an evaluation of simple main effects of wave within condition. There were non-significant difference within the Friends condition ($F[2, 194] = 1.759, p = 0.175, \eta^2 = 0.012$). There were significant differences with an effect size indicating a moderate effect in the Little Mosque condition ($F[2, 188] = 3.392, p = 0.036, \eta^2 = 0.072$). We followed up the significant simple main effect for with pairwise comparisons. At this level we controlled for Type I

error with the Holm's sequential Bonferroni [Green and Salkind, 2017c]. Within the Little Mosque condition, we found a significant difference between baseline and post1 ($t[94] = 2.447, p = .049, d = 0.343$), but non-significant differences between baseline and post2 ($t[94] = 1.621, p = .216, d = 0.222$) and post1 and post2 ($t[94] = 1.034, p = .304, d = -0.150$).

10.4.6 If we only had a main effect

When there is an interaction effect, we do not interpret main effects. This is because the solution is more complicated than a main effect could explain. It is important, though, to know how to interpret a main effect. We would do this if we had one or more significant main effects and no interaction effect.

The figure shows our place on the workflow.

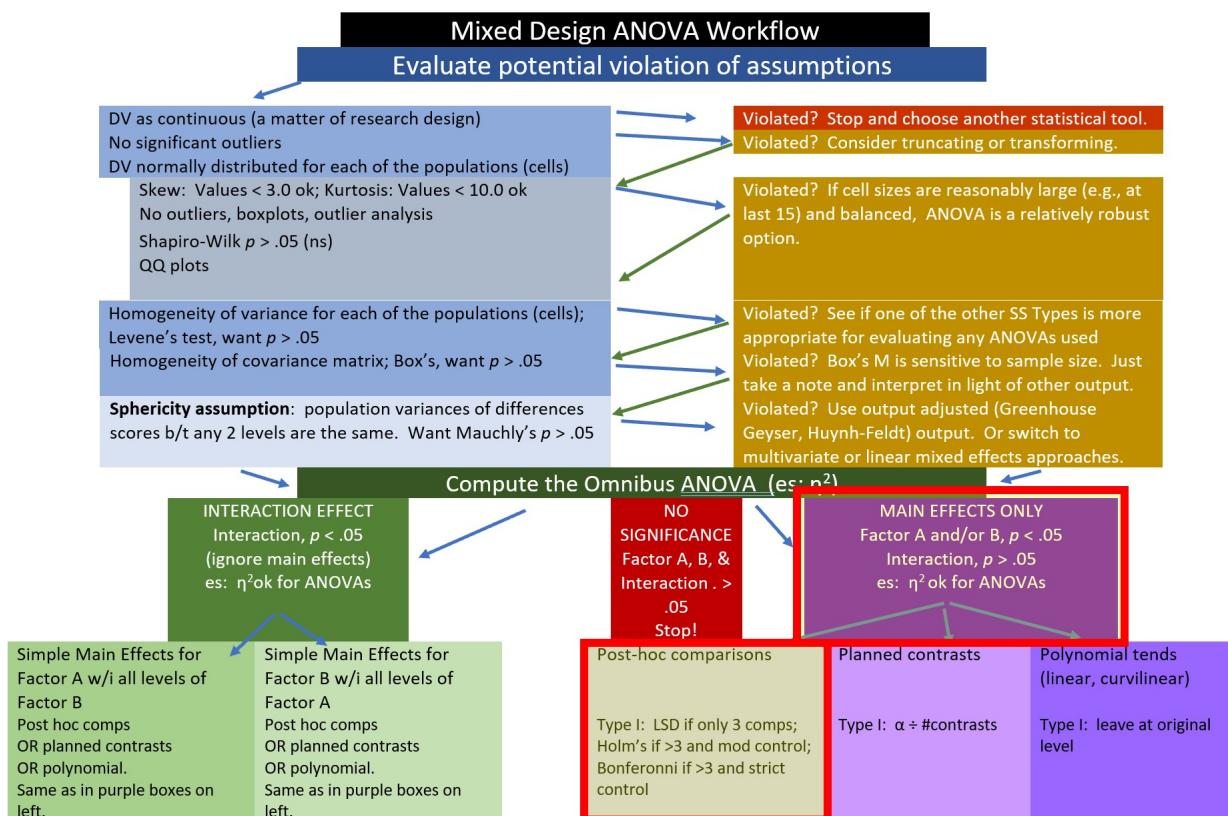


Figure 10.8: Image of a workflow showing that we are at the “Main effects only” step

If we had not had a significant interaction, but did have a significant main effect for wave, we could have conducted pairwise comparisons for pre, post1, and post2 – collapsing across condition.

```
pwcMain <- Murrar_df %>%
  rstatix::pairwise_t_test(Diff ~ Wave, paired = TRUE, p.adjust.method = "bonferroni")
pwcMain
```

```
# A tibble: 3 x 10
  .y.   group1   group2     n1     n2 statistic     df      p p.adj p.adj.signif
* <chr> <chr>    <chr> <int> <int>     <dbl> <dbl> <dbl> <dbl> <chr>
1 Diff  Baseline Post1     193     193     0.539     192  0.59      1 ns
2 Diff  Baseline Post2     193     193     0.652     192  0.515     1 ns
3 Diff  Post1    Post2     193     193     0.0528    192  0.958     1 ns
```

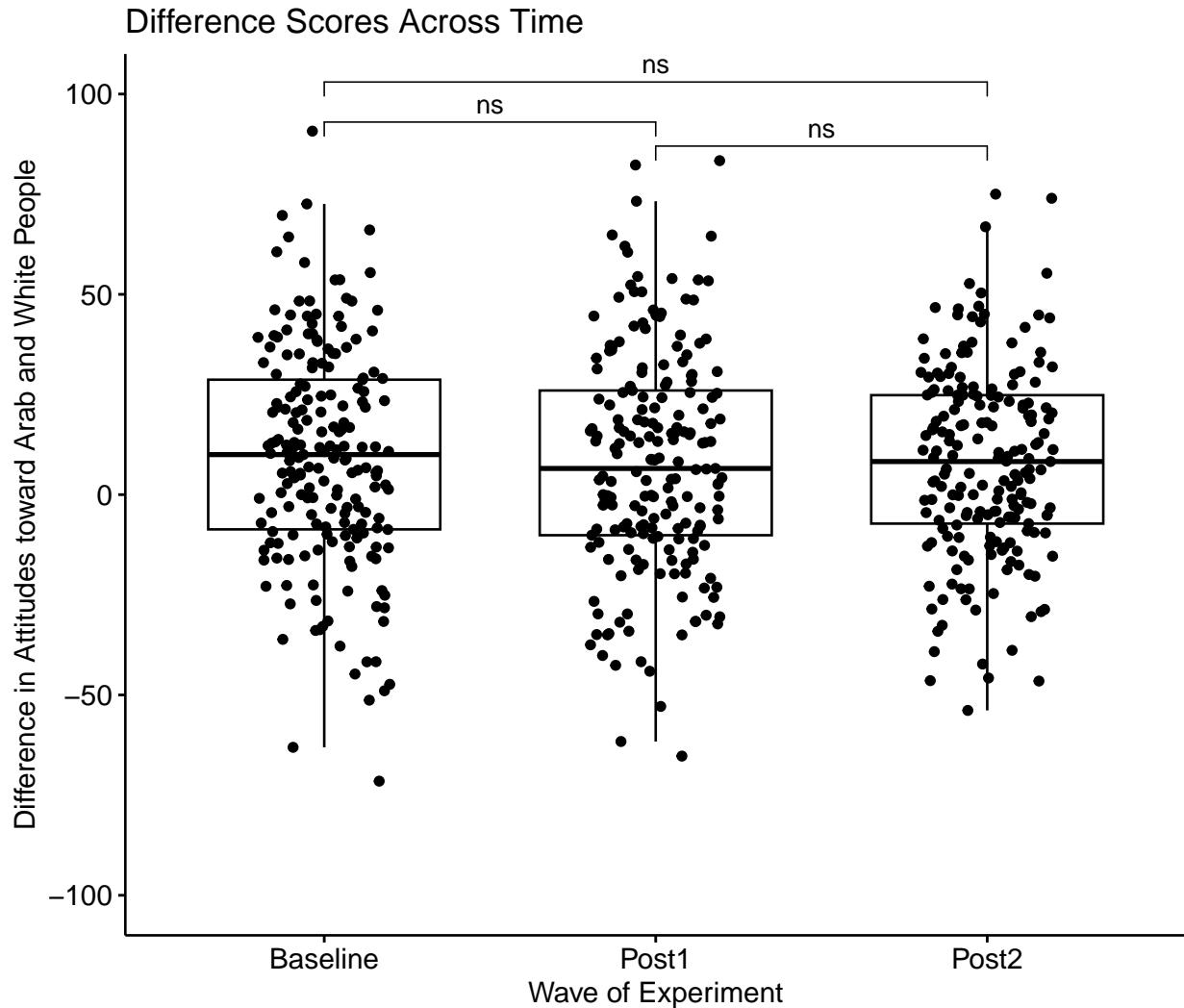
Ignoring condition (Friends, Little Mosque), we do not see changes across time. This is not surprising since the F test for the main effect was also non-significant ($F[2, 382] = 0.273, p = .761, \eta^2 = 0.0014$),

If we had had a non-significant interaction effect but a significant main effect for condition, there would have been no need for further follow-up. Why? Because there were only two levels the significant main effect already tells us there were statistically significant differences between Friends and Little Mosque ($F[1, 191] = 13.149, p < .001, \eta^2 = 0.023$).

Here is a figure to represent this analysis.

```
# We ran this before -- grabbing again to make it clear how creating
# and updating the boxplot works
WaveMain <- ggpubr::ggboxplot(Murrar_df, x = "Wave", y = "Diff", xlab = "Wave of Experiment",
                                ylim = c(-100, 100), ylab = "Difference in Attitudes toward Arab and White People",
                                add = "jitter", title = "Difference Scores Across Time")

# This updates the pwcWVwiGP object (which holds the t-tests) to
# include plotting information about the xy positions
pwcMain <- pwcMain %>%
  rstatix::add_xy_position(x = "Wave")
# pwcWVwiGP Diff_2way was my omnibus ANOVA object
WaveMain <- WaveMain + ggpubr::stat_pvalue_manual(pwcMain, label = "p.adj.signif",
                                                    tip.length = 0.02, hide.ns = FALSE, y.position = c(93, 103, 87))
WaveMain
```



10.4.7 APA Style Write-up of the Results

As I looked across the different approaches to describing the results, I felt that the simple main effect of condition within wave best explained the findings.

10.4.7.1 Results

We conducted a 2 X 3 mixed design ANOVA to evaluate the combined effects of condition (Friends and Little Mosque) and wave (baseline, post1, post2) on a difference score that compared attitudes toward White and Arab people.

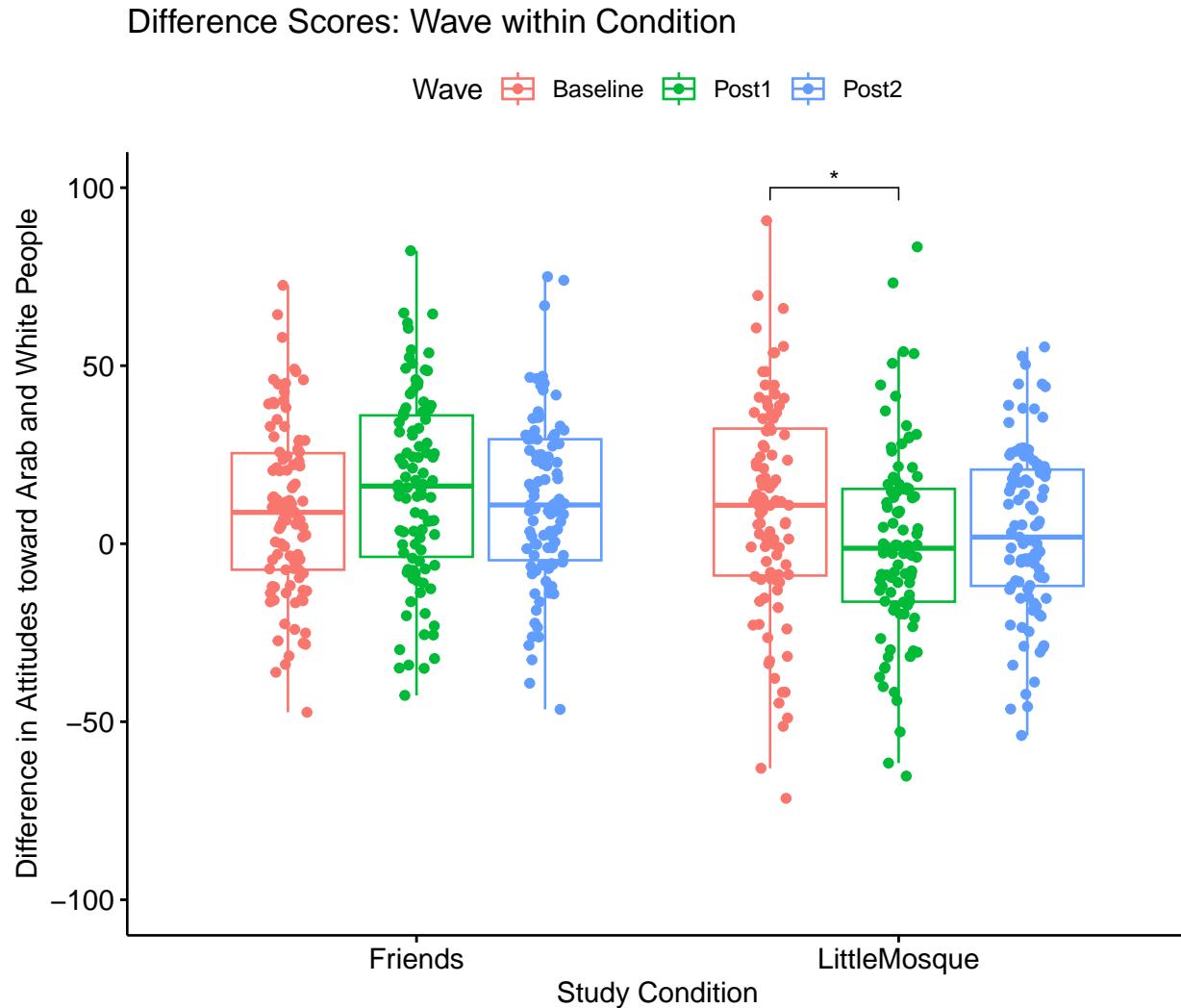
Mixed design ANOVA has a number of assumptions related to both the within-subjects and between-subjects elements. Data are expected to be normally distributed at each level of design. There was no evidence of skew (all values were at or below the absolute value of 0.32) or kurtosis (all values were below the absolute value of .57 [Kline, 2016a]). Similarly, results of the Shapiro-Wilk normality test (applied to the residuals from the

factorial ANOVA model) suggested that model residuals did not differ significantly from a normal distribution ($W = 0.999, p = 0.953$). Visual inspection of boxplots for each wave of the design, assisted by the `rstatix::identify_outliers()` function (which reports values above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$, where IQR is the interquartile range) indicated some outliers, but none at the extreme level. Because of the between-subjects aspect of the design, the homogeneity of variance assumption was evaluated. Levene's test indicated a violation of this assumption between the Friends and Little Mosque conditions at baseline ($F[1, 191] = 3.973, p = .047$). However, there was no indication of assumption violation at post1 ($F[1, 191] = 0.141, p = .708$) and post2 ($F[1, 191] = 0.107, p = .743$) waves of the design. Further, Box's M-test ($M = 3.21, p = .073$) indicated no violation of the homogeneity of covariance matrices. Mauchly's test indicated no violation of the sphericity assumption for the main ($W = .99, p = .369$) and interaction ($W = .99, p = .369$) effects.

Results of the omnibus ANOVA indicated a significant main effect for condition ($F[1, 191] = 13.149, p < .001, \eta^2 = 0.023$), a non-significant main effect for wave ($F[2, 382] = 0.273, p = .761, \eta^2 = 0.001$), and a significant interaction effect ($F[2, 382] = 5.008, p = 0.007, \eta^2 = 0.017$).

We followed the significant interaction effect with an evaluation of simple main effects of wave within condition. A one-way ANOVA indicated non-significant differences within the Friends condition ($F[2, 194] = 1.759, p = 0.175, \eta^2 = 0.012$). In contrast, there were significant differences in the Little Mosque condition ($F[2, 188] = 3.392, p = 0.036, \eta^2 = 0.072$). We followed up this significant simple main effect with pairwisse comparisons. At this level we controlled for Type I error with the Holm's sequential Bonferroni [Green and Salkind, 2017c]. Within the Little Mosque condition, we found a significant difference between baseline and post1 ($t[94] = 2.447, p = .049, d = 0.343$), but non-significant differences between baseline and post2 ($t[94] = 1.621, p = .216, d = 0.222$) and post1 and post2 ($t[94] = 1.034, p = .304, d = -0.150$).

As illustrated in Figure 1 difference scores were comparable at baseline. After the intervention, difference scores increased for those in the Friends condition – indicating more favorable attitudes toward White people. In contrast, those exposed to the Little Mosque condition had difference scores that were lower from baseline to post1 and stayed at that same level at post2. Means and standard deviations are reported in Table 1.



The following code can be used to write output to .csv files. From there it is easy(er) to manipulate them into tables for use in an empirical manuscript.

```
MASS::write.matrix(pwcWVwiGP, sep = ",", file = "pwcWVwiGP.csv")
# this command can also be used to export other output
MASS::write.matrix(Diff_2way$ANOVA, sep = ",", file = "Diff_2way.csv")
MASS::write.matrix(SimpleWave, sep = ",", file = "SimpleWave.csv")
MASS::write.matrix(SimpleCond, sep = ",", file = "SimpleCond.csv")
```

10.4.7.2 Comparing our findings to Murrar and Brauer [2018]

In general, the results of our simulation mapped onto the findings. If you have access to the article I encourage you to examine it as you consider my observations.

- The authors started their primary analyses of Experiment 1 with independent t tests comparing the Friends and Little Mosque conditions within each of the baseline, post1, and post2

waves. This is equivalent to our simple main effects of condition within wave that we conducted as follow-up to the significant interaction effect. It is not clear to me why they did not precede this with a mixed design ANOVA.

- The results of the article are presented in their Table 1
- Our results were comparable in that we found no attitude difference at baseline
- Similar to the results in the article we found statistically significant differences (with comparable p values and effect sizes) at post1 and post2
- With two experiments (each with a number of associated hypotheses) in a single paper there were a large number of analyses conducted by the authors. I think they designed tables and figures that provided an efficient and clear review of the study design and their findings.
- This finding is exciting to me. Anti-racism education frequently encourages individuals to expose themselves to content authored/created by individuals from groups with marginalized identities. This finding supports that approach to prejudice reduction.

10.5 Power in Mixed Design ANOVA

The package `wp.rmanova` was designed for power analysis in repeated measures ANOVA.

Power analysis allows us to determine the probability of detecting an effect of a given size with a given level of confidence. Especially when we don't achieve significance, we may want to stop.

In the `WebPower` package, we specify 6 of 7 interrelated elements; the package computes the missing element

- n = sample size (number of individuals in the whole study)
- ng = number of groups
- nm = number of repeated measurements (i.e., waves)
- f = Cohen's f (an effect size; we can use a conversion calculator); Cohen suggests that f values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes, respectively
- $nscor$ = the Greenhouse Geiser correction from our output; 1.0 means no correction was needed and is the package's default; < 1 means some correction was applied
- $alpha$ = is the probability of Type I error; we traditionally set this at .05
- $power$ = $1 - P(\text{Type II error})$ we traditionally set this at .80 (so anything less is less than what we want)
- $type$ = 0 is for between-subjects, 1 is for repeated measures, 2 is for interaction effect; in a mixed design ANOVA we will select "2"

As in the prior lessons, we need to convert our effect size for the *interaction* to f effect size (this is not the same as the F test). The `effectsize` package has a series of converters. We can use the `eta2_to_f()` function to translate the η^2 associated with the interaction effect to Cohen's f .

```
#interaction effect
effectsize::eta2_to_f(0.017)
```

[1] 0.1315066

We can now retrieve information from our study (including the Cohen's f value we just calculated) and insert it into the script for the power analysis.

```
WebPower::wp.rmanova(n=193, ng=2, nm=3, f = .1315, nscor = .99, alpha = .05, power = NULL, type = 2)
```

Repeated-measures ANOVA analysis

n	f	ng	nm	nscor	alpha	power
193	0.1315	2	3	0.99	0.05	0.3493183

NOTE: Power analysis for interaction-effect test

URL: <http://psychstat.org/rmanova>

We are powered at .349 (we have a 35% of rejecting the null hypothesis, if it is true)

In reverse, setting *power* at .80 (the traditional value) and changing *n* to *NULL* yields a recommended sample size.

```
WebPower::wp.rmanova(n = NULL, ng = 2, nm = 3, f = 0.1315, nscor = 0.99, alpha = 0.05, power = 0.8, type = 2)
```

Repeated-measures ANOVA analysis

n	f	ng	nm	nscor	alpha	power
562.608	0.1315	2	3	0.99	0.05	0.8

NOTE: Power analysis for interaction-effect test

URL: <http://psychstat.org/rmanova>

Given our desire for strong power and our weak effect size, this power analysis suggests a sample size of 562 participants to detect a significant interaction effect.

10.6 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “do-able” and then try at least one more problem that challenges you in some way. At a minimum your data should allow for a 2 X 3 (or 3 X 2) design. At least one of the problems you work should have a statistically significant interaction effect that you work all the way through.

Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Whichever you choose, you will focus on these larger steps in one-way ANOVA, including:

- test the statistical assumptions
- conduct a two-way (minimally a 2x3), mixed design, ANOVA, including

- omnibus test and effect size
 - report main and interaction effects
 - conduct follow-up testing of simple main effects
- write a results section to include a figure and tables

10.6.1 Problem #1: Play around with this simulation.

Copy the script for the simulation and then change (at least) one thing in the simulation to see how it impacts the results.

- If mixed design ANOVA is new to you, perhaps you just change the number in “set.seed(210813)” from 210813 to something else. Your results should parallel those obtained in the lecture, making it easier for you to check your work as you go.
- If you are interested in power, change the sample size to something larger or smaller.
- If you are interested in variability (i.e., the homogeneity of variance assumption), perhaps you change the standard deviations in a way that violates the assumption.

10.6.2 Problem #2: Conduct a mixed design ANOVA with a different dependent variable.

The Murrar et al. [2018] article has three dependent variables (attitudes toward people who are Arab, attitudes toward people who are White, and the difference score). I analyzed the difference score. Select one of the other dependent variables. If you do not get a significant interaction, play around with the simulation (changing the sample size, standard deviations, or both) until you get a significant interaction effect.

10.6.3 Problem #3: Try something entirely new.

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete a mixed design ANOVA. Please have at least 3 levels for one predictor and at least 2 levels for the second predictor.

10.6.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the IV and DV. Minimally, the data should allow the analysis of a 2 x 3 (or 3 X 2) design. At least one of the problems you work should have a significant interaction effect so that follow-up is required.	5	_____

Assignment Component	Points Possible	Points Earned
2. Simulate (or import) and format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct omnibus ANOVA (w effect size)	5	_____
5. Conduct one set of follow-up tests; narrate your choice	5	_____
6. Describe approach for managing Type I error	5	_____
7. APA style results with table(s) and figure	5	_____
8 Explanation to grader	5	_____
Totals	40	_____

Chapter 11

Analysis of Covariance

[Screencasted Lecture Link](#)

The focus of this lecture is analysis of covariance. Sticking with the same research vignette as we used for the mixed design ANOVA, we rearrange the variables a bit to see how they work in an ANCOVA design. The results help clarify the distinction between *moderator* and *covariate*.

11.1 Navigating this Lesson

There is about just about an hour of lecture. If you work through the materials with me, plan for an additional hour or two

While the majority of R objects and data you will need are created within the R script that sources the chapter, occasionally there are some that cannot be created from within the R framework. Additionally, sometimes links fail. All original materials are provided at the [Github site](#) that hosts the book. More detailed guidelines for ways to access all these materials are provided in the OER's introduction

11.1.1 Learning Objectives

Learning objectives from this lecture include the following:

- Define a *covariate* and distinguish it from a *moderator*.
- Recognize the case where ANCOVA is a defensible statistical approach for analyzing the data.
- Name and test the assumptions underlying ANCOVA.
- Analyze, interpret, and write up results for ANCOVA.
- List the conditions that are prerequisite for the appropriate use of a covariate or control variable.

11.1.2 Planning for Practice

In each of these lessons I provide suggestions for practice that allow you to select from problems that vary in degree of difficulty. The least complex is to change the random seed and rework the problem demonstrated in the lesson. The results *should* map onto the ones obtained in the lecture.

The second option comes from the research vignette. For this ANCOVA article, I take a lot of liberties with the variables and research design. You could further mix and match for a different ANCOVA constellation.

As a third option, you are welcome to use data to which you have access and is suitable for ANCOVA. In either case the practice options suggest that you:

- test the statistical assumptions
- conduct an ANCOVA, including
 - omnibus test and effect size
 - report main effects and engage in any follow-up testing
 - interpret results in light of the role of the second predictor variable as a *covariate* (as opposed to the moderating role in the prior lessons)
- write a results section to include a figure and tables

11.1.3 Readings & Resources

In preparing this chapter, I drew heavily from the following resource(s). Other resources are cited (when possible, linked) in the text with complete citations in the reference list.

- Green, S. B., & Salkind, N. J. (2017). One-Way Analysis of Covariance (Lesson 27). In *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (Eighth edition., pp. 151–160). Boston: Pearson. OR
 - This lesson provides an excellent review of ANCOVA with examples of APA style write-ups. The downside is that it is written for use in SPSS.
- ANCOVA in R: The Ultimate Practical Guide. (n.d.). Retrieved from <https://www.datanovia.com/en/lessons/ancova-in-r/>
 - This is the workflow we are using for the lecture and written specifically for R.
- Bernerth, J. B., & Aguinis, H. (2016). A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1), 229–283. <https://doi.org/10.1111/peps.12103>
 - An article from the industrial-organizational psychology world. Especially relevant for this lesson is the flowchart on page 273 and the discussion (pp. 270 to the end).
- Murrar, S., & Brauer, M. (2018). Entertainment-education effectively reduces prejudice. *Group Processes & Intergroup Relations*, 21(7), 1053–1077. <https://doi.org/10.1177/1368430216682350>
- This article is the source of our research vignette. I used this same article in the lesson on **mixed design ANOVA**. Swapping variable roles can be useful in demonstrating how ANCOVA is different than mixed design ANOVA.

11.1.4 Packages

The packages used in this lesson are embedded in this code. When the hashtags are removed, the script below will (a) check to see if the following packages are installed on your computer and, if not (b) install them.

```
# used to convert data from long to wide
# if(!require(reshape2)){install.packages('reshape2')}
# if(!require(broom)){install.packages('broom')}
# if(!require(tidyverse)){install.packages('tidyverse')}
# if(!require(psych)){install.packages('psych')} easy plots
# if(!require(ggpubr)){install.packages('ggpubr')} pipe-friendly R
# functions if(!require(rstatix)){install.packages('rstatix')} export
# objects for table making
# if(!require(MASS)){install.packages('MASS')}
# if(!require(knitr)){install.packages('knitr')}
# if(!require(dplyr)){install.packages('dplyr')}
# if(!require(apaTables)){install.packages('apaTables')}
```

11.2 Introducing Analysis of Covariance (ANCOVA)

Analysis of covariance (ANCOVA) evaluates the null hypothesis that

- population means on a dependent variable are equal across levels of a factor(s) adjusting for differences on a covariate(s); stated differently -
- the population adjusted means are equal across groups

This lecture introduces a distinction between **moderators** and **covariates**.

Moderator: a variable that changes the strength or direction of an effect between two variables X (predictor, independent variable) and Y (criterion, dependent variable).

Covariate: an observed, continuous variable, that (when used properly) has a relationship with the dependent variable. It is included in the analysis, as a predictor, so that the predictive relationship between the independent (IV) and dependent (DV) are adjusted.

Understanding this difference may be facilitated by understanding one of the assumptions of ANCOVA – that the slopes relating the covariate to the dependent variable are the same for all groups (i.e., the homogeneity-of-slopes assumption). If this assumption is violated then the between-group differences in adjusted means are not interpretable and the covariate should be treated as a moderator and analyses that assess the simple main effects (i.e., follow-up to a significant interaction) should be conducted.

A one-way ANCOVA requires three variables:

- IV/factor – categorical (2 or more)
- DV – continuous
- covariate – continuous

Green and Salkind [2017a] identified common uses of ANCOVA:

- Studies with a pretest and random assignment of subjects to factor levels. Variations on this research design include:
 - assignment to factor levels based on that pretest,
 - matching based on the pretest, and random assignment to factor levels,
 - simply using the pretest as a covariate for the posttest DV.
- Studies with a potentially confounding variable (best when there is theoretical justification and prior empirical evidence for such) over which the researcher wants “control”

Although it is possible to have multi-way (e.g., 2-way, 3-way) ANCOVA, in this lecture we will only work two, one-way ANCOVAs representing these common use cases.

11.2.1 Workflow for ANCOVA

Our analytic process will be similar to others in the ANOVA series. An ANCOVA workflow maps this in further detail.

1. Prepare the data
2. Evaluate potential violation of the assumptions
3. Compute the omnibus ANCOVA, and follow-up accordingly
 - If significant: follow-up with post-hoc comparisons, planned contrasts, and/or polynomial
 - If non-significant: stopping.

ANCOVA has four primary assumptions:

Linearity: The covariate is linearly related to the dependent variable within all levels of the factor (IV).

Homogeneity of regression slopes: The weights or slopes relating the covariate to the DV are equal across all levels of the factor.

Normally distributed: The DV is normally distributed in the population for any specific value of the covariate and for any one level of a factor. This assumption applies to every combination of the values of the covariate and levels ohttps://www.datanovia.com/en/lessons/ancova-in-r/f the factor and requires them all to be normally distributed. To the degree that population distributions are not normal and sample sizes are small, p values may not be trustworthy and power reduced. Evaluating this is frequently operationalized by inspecting the residuals and identifying outliers.

Homogeneity of variances: The variances of the DV for the conditional distributions (i.e., every combination of the values of the covariate and levels of the factor) are equal.

We are following the approach to analyzing ANCOVA identified in the Datanovia lesson on ANCOVA [Datanovia].

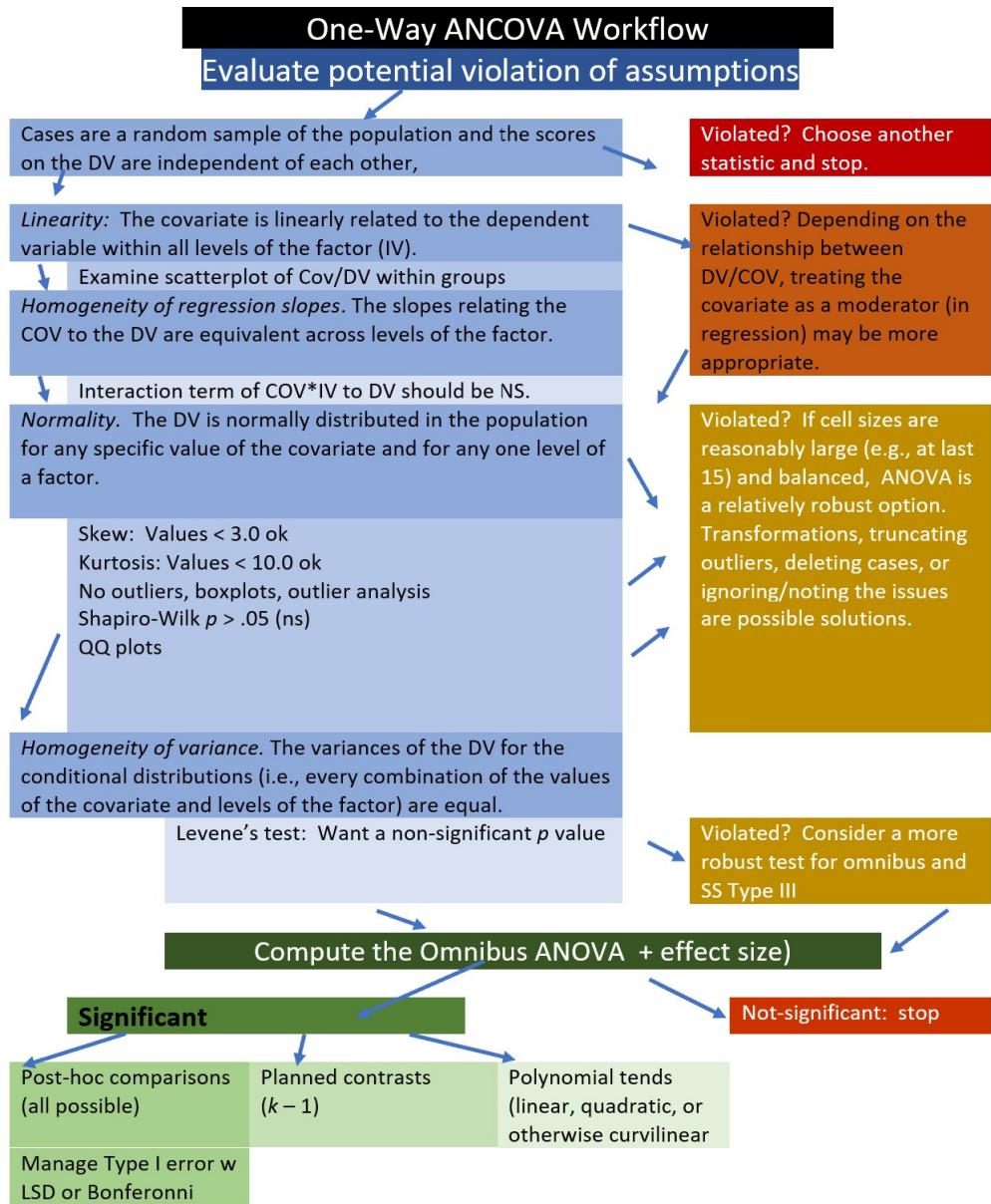


Figure 11.1: Image of the ANCOVA workflow

11.3 Research Vignette

We will continue with the example used in the [mixed design ANOVA lesson](#). The article does not contain any ANCOVA analyses, but there is enough data that I can demonstrate the two general ways (i.e., controlling for the pretest, controlling for a potentially confounding variable) that ANCOVA is used.

Here is a quick reminder of the research vignette.

Murrar and Brauer's [2018] article described the results of two studies designed to reduce prejudice against Arabs/Muslims. In the lesson on mixed design ANOVA, we only worked the first of two experiments reported in the study. Participants ($N = 193$), all who were White, were randomly assigned to one of two conditions where they watched six episodes of the sitcom *Friends* or *Little Mosque on the Prairie*. The sitcoms and specific episodes were selected after significant pilot testing. The selection was based on the tension selecting stimuli that were as similar as possible, yet the intervention-oriented sitcom needed to invoke psychological processes known to reduce prejudice. The authors felt that both series had characters that were likable and relatable who were engaged in activities of daily living. The Friends series featured characters who were predominantly White, cis-gendered, and straight. The Little Mosque series portrays the experience Western Muslims and Arabs as they live in a small Canadian town. This study involved assessment across three waves: baseline (before watching the assigned episodes), post1 (immediately after watching the episodes), and post2 (completed 4-6 weeks after watching the episodes).

The study used *feelings and liking thermometers*, rating their feelings and liking toward 10 different groups of people on a 0 to 100 sliding scale (with higher scores reflecting greater liking and positive feelings). For the purpose of this analysis, the ratings of attitudes toward White people and attitudes toward Arabs/Muslims were used. A third metric was introduced by subtracting the attitudes towards Arabs/Muslims from the attitudes toward Whites. Higher scores indicated more positive attitudes toward Whites while low scores indicated no difference in attitudes. To recap, there were three potential dependent variables, all continuously scaled:

- AttWhite: attitudes toward White people; higher scores reflect greater liking
- AttArab: attitudes toward Arab people; higher scores reflect greater liking
- Diff: the difference between AttWhite and AttArab; higher scores reflect a greater liking for White people

With random assignment, nearly equal cell sizes, a condition with two levels (Friends, Little Mosque), and three waves (baseline, post1, post2), this is perfect for mixed design ANOVA but suitable for an ANCOVA demonstration.

	COND	Baseline At start of study (prior to viewing sitcoms)	Intervention 6 episodes of the sitcom	Post1 Toward end of viewing the sitcoms	Post2 4-6 weeks after viewing the final sitcom
Random Assignment	Friends	X		X	X
	Little Mosque on the Prairie	X	Selected for potential for prejudice reduction	X	X

Figure 11.2: Image of the design for the Murrar and Brauer (2018) study

11.3.1 Data Simulation

Below is the code I have used to simulate the data. The simulation includes two dependent variables (AttWhite, AttArab), Wave (baseline, post1, post2), and COND (condition; Friends, Little_Mosque). There is also a caseID (repeated three times across the three waves) and rowID (giving each observation within each case an ID). You can use this simulation for two of the three practice suggestions.

```
library(tidyverse)
# change this to any different number (and rerun the simulation) to
# rework the chapter problem
set.seed(210813)
# sample size, M and SD for each cell; this will put it in a long
# file
AttWhite <- round(c(rnorm(98, mean = 76.79, sd = 18.55), rnorm(95, mean = 75.37,
  sd = 18.99), rnorm(98, mean = 77.47, sd = 18.95), rnorm(95, mean = 75.81,
  sd = 19.29), rnorm(98, mean = 77.79, sd = 17.25), rnorm(95, mean = 75.89,
  sd = 19.44)), 3)
# set upper bound for variable
AttWhite[AttWhite > 100] <- 100
# set lower bound for variable
AttWhite[AttWhite < 0] <- 0
AttArab <- round(c(rnorm(98, mean = 64.11, sd = 20.97), rnorm(95, mean = 64.37,
  sd = 20.03), rnorm(98, mean = 64.16, sd = 21.64), rnorm(95, mean = 70.52,
  sd = 18.55), rnorm(98, mean = 65.29, sd = 19.76), rnorm(95, mean = 70.3,
  sd = 17.98)), 3)
# set upper bound for variable
AttArab[AttArab > 100] <- 100
# set lower bound for variable
AttArab[AttArab < 0] <- 0
rowID <- factor(seq(1, 579))
caseID <- rep((1:193), 3)
Wave <- c(rep("Baseline", 193), rep("Post1", 193), rep("Post2", 193))
COND <- c(rep("Friends", 98), rep("LittleMosque", 95), rep("Friends", 98),
  rep("LittleMosque", 95), rep("Friends", 98), rep("LittleMosque", 95))
# groups the 3 variables into a single df: ID#, DV, condition
Murrar_df <- data.frame(rowID, caseID, Wave, COND, AttArab, AttWhite)
# make caseID a factor
Murrar_df[, "caseID"] <- as.factor(Murrar_df[, "caseID"])
# make Wave an ordered factor
Murrar_df$Wave <- factor(Murrar_df$Wave, levels = c("Baseline", "Post1",
  "Post2"))
# make COND an ordered factor
Murrar_df$COND <- factor(Murrar_df$COND, levels = c("Friends", "LittleMosque"))
# creates the difference score
Murrar_df$Diff <- Murrar_df$AttWhite - Murrar_df$AttArab
```

Let's check the structure. We want

- rowID and caseID to be unordered factors,
- Wave and COND to be ordered factors,
- AttArab, AttWhite, and Diff to be numerical

```
str(Murrar_df)
```

```
'data.frame': 579 obs. of 7 variables:
 $ rowID   : Factor w/ 579 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ caseID   : Factor w/ 193 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ Wave     : Factor w/ 3 levels "Baseline","Post1",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ COND     : Factor w/ 2 levels "Friends","LittleMosque": 1 1 1 1 1 1 1 1 1 1 ...
 $ AttArab  : num  74.3 55.8 33.3 66.3 71 ...
 $ AttWhite : num  100 79 75.9 68.2 100 ...
 $ Diff     : num  25.71 23.18 42.67 1.92 29.01 ...
```

The structure looks satisfactory. R will automatically “order” factors alphabetically or numerically. In this lesson’s example the alphabetical ordering (i.e., Baseline, Post1, Post2; Friends, LittleMosque) is consistent with the logic in our study.

If you want to export this data as a file to your computer, remove the hashtags to save it (and re-import it) as a .csv (“Excel lite”) or .rds (R object) file. This is not a necessary step.

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R. This is what I would do. *Note: My students and I have discovered that the psych::describeBy() function seems to not work with files in the .rds format, but does work when the data are imported with .csv.*

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Murrar_df, 'Murrar_RDS.rds') bring back the simulated
# dat from an .rds file Murrar_df <- readRDS('Murrar_RDS.rds')
```

The code for .csv will likely lose the formatting (i.e., stripping Wave and COND of their ordered factors), but it is easy to view in Excel.

```
# write the simulated data as a .csv write.table(Murrar_df,
# file='DiffCSV.csv', sep=',', col.names=TRUE, row.names=FALSE) bring
# back the simulated dat from a .csv file Murrar_df <- read.csv
# ('DiffCSV.csv', header = TRUE)
```

11.4 Working the ANCOVA – Scenario #1: Controlling for the pretest

So that we can begin to understand how the covariate operates, we are going to predict attitudes towards Arabs at post-test (AttArabP1) by condition (COND), controlling for attitudes toward Arabs at baseline (AttArabB). You may notice that in this analysis we are ignoring the second post-test. This is because I am simply demonstrating ANCOVA. To ignore the second post test would be a significant loss of information.

11.4.1 Preparing the data

When the covariate in ANCOVA is a pretest, we need three variables:

- IV that has two or more levels; in our case it is the Friends and Little Mosque conditions
- DV that is continuous; in our case it is the attitudes toward Arabs at post1
- Covariate that is continuous; in our case it is the attitudes toward Arabs at baseline

The form of our data matters. The simulation created a *long* form (formally called the *person-period* form) of data. That is, each observation for each person is listed in its own row. In this dataset where we have 193 people with 3 observations (baseline, post1, post2) each, we have 579 rows. In ANCOVA where we use the pre-test as a covariate, we need all the data to be on a single row. This is termed the *person level* form of data. We can restructure the data with the *data.table* and *reshape2()** packages.

```
# Create a new df (Murrar_wide) Identify the original df In the
# transition from long-to-wide it seems like you can only do one
# time-varying variable at a time When there are multiple
# time-varying and time-static variables, put all the time-static
# variables on the left side of the tilde Put the name of the single
# time-varying variable in the concatenated list
Murrar1 <- reshape2::dcast(data = Murrar_df, formula = caseID + COND ~
    Wave, value.var = "AttArab")
# before restructuring a second variable, rename the first variable
Murrar1 <- rename(Murrar1, AttArabB = "Baseline", AttArabP1 = "Post1",
    AttArabP2 = "Post2")
# repeat the process for additional variables; but give the new df
# new names -- otherwise you'll overwrite your work
Murrar2 <- reshape2::dcast(data = Murrar_df, formula = caseID ~ Wave, value.var = "AttWhite")
Murrar2 <- rename(Murrar2, AttWhiteB = "Baseline", AttWhiteP1 = "Post1",
    AttWhiteP2 = "Post2")
# Now we join them
Murrar_wide <- dplyr::full_join(Murrar1, Murrar2, by = c("caseID"))

str(Murrar_wide)
```

```
'data.frame': 193 obs. of 8 variables:
 $ caseID     : Factor w/ 193 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ COND       : Factor w/ 2 levels "Friends","LittleMosque": 1 1 1 1 1 1 1 1 1 1 ...
 $ AttArabB   : num  74.3 55.8 33.3 66.3 71 ...
 $ AttArabP1  : num  80.3 76.6 92 96.5 59.1 ...
 $ AttArabP2  : num  64.8 43.3 40.3 69.1 74.9 ...
 $ AttWhiteB  : num  100 79 75.9 68.2 100 ...
 $ AttWhiteP1 : num  95.6 51 91.9 86.7 75.8 ...
 $ AttWhiteP2: num  100 89.7 49.5 99.4 83.1 ...
```

If you want to export this data as a file to your computer, remove the hashtags to save it (and re-import it) as a .csv (“Excel lite”) or .rds (R object) file. This is not a necessary step.

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R. This is what I would do.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Murrar_wide, 'MurrarW_RDS.rds') bring back the
# simulated dat from an .rds file Murrar_wide <-
# readRDS('MurrarW_RDS.rds')
```

The code for .csv will likely lose the formatting (i.e., stripping Wave and COND of their ordered factors), but it is easy to view in Excel.

```
# write the simulated data as a .csv write.table(Murrar_wide,
# file='MurrarW_CCSV.csv', sep=',', col.names=TRUE, row.names=FALSE)
# bring back the simulated dat from a .csv file Murrar_wide <-
# read.csv ('MurrarW_CCSV.csv', header = TRUE)
```

11.4.2 Evaluating the statistical assumptions

There are a number of assumptions in ANCOVA. These include:

- random sampling
- independence in the scores representing the dependent variable
 - there is, of course, intentional dependence in any repeated measures or within-subjects variable
- linearity of the relationship between the covariate and DV within all levels of the independent variable
- homogeneity of the regression slopes
- a normally distributed DV for any specific value of the covariate and for any one level of a factor
- homogeneity of variance

These are depicted in the flowchart, below.

11.4.2.1 Linearity assumption

ANCOVA assumes that there is linearity between the covariate and outcome variable at each level of the grouping variable. In our case this means that there is linearity between the pre-test (covariate) and post-test (outcome variable) at each level of the intervention (Friends, Little Mosque).

We can create a scatterplot (with regression lines) between covariate (our pretest) and the outcome (post-test1).

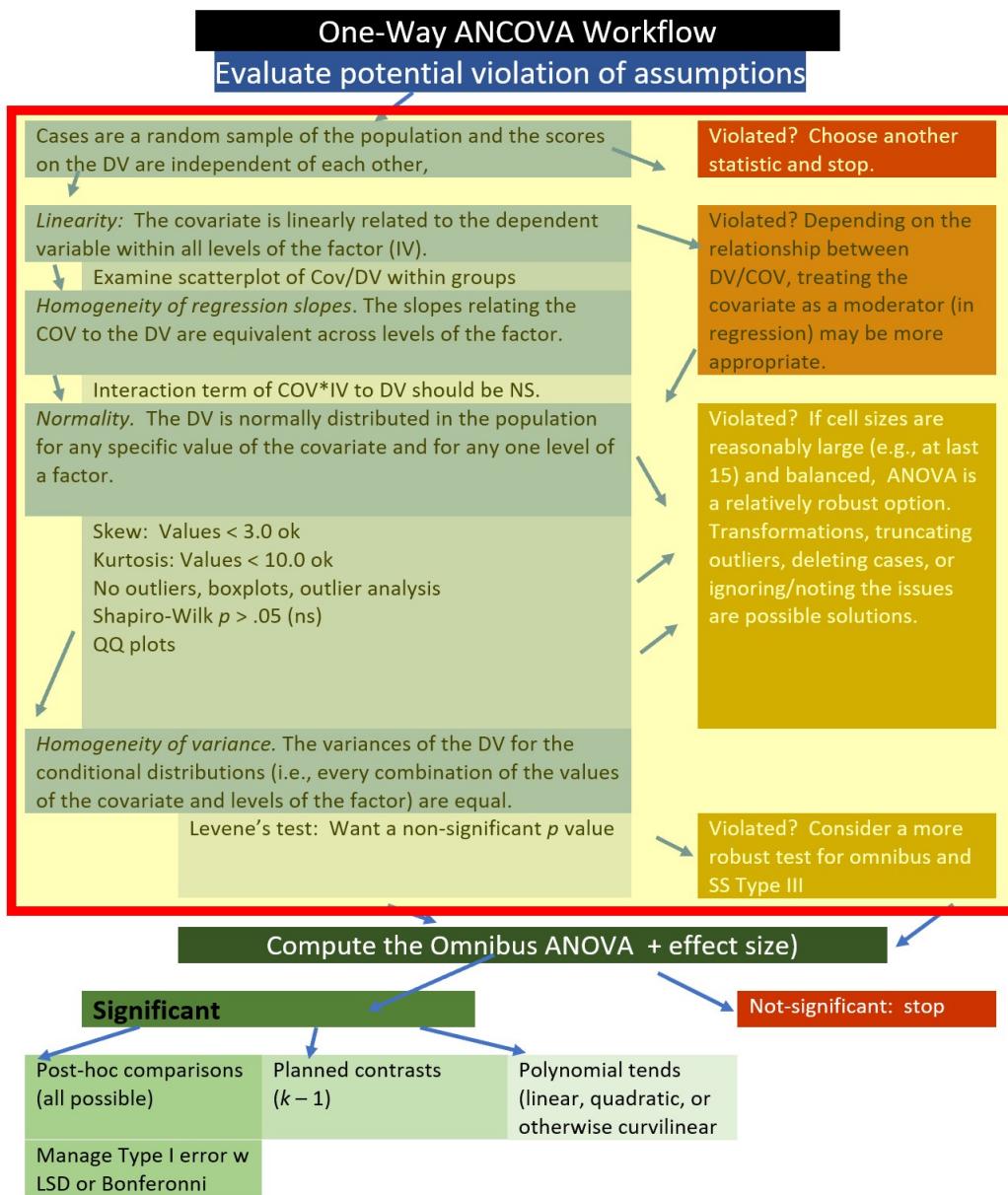


Figure 11.3: Image of the ANCOVA workflow, showing our current place in the process

```
ggpubr::ggscatter(Murrar_wide, x = "AttArabB", y = "AttArabP1", color = "COND",
  add = "reg.line") + ggpubr::stat_regrline_equation(aes(label = paste(..eq.label..,
  ..rr.label.., sep = "~~~~"), color = COND))
```

Warning: The dot-dot notation (`..eq.label..`) was deprecated in ggplot2 3.4.0.

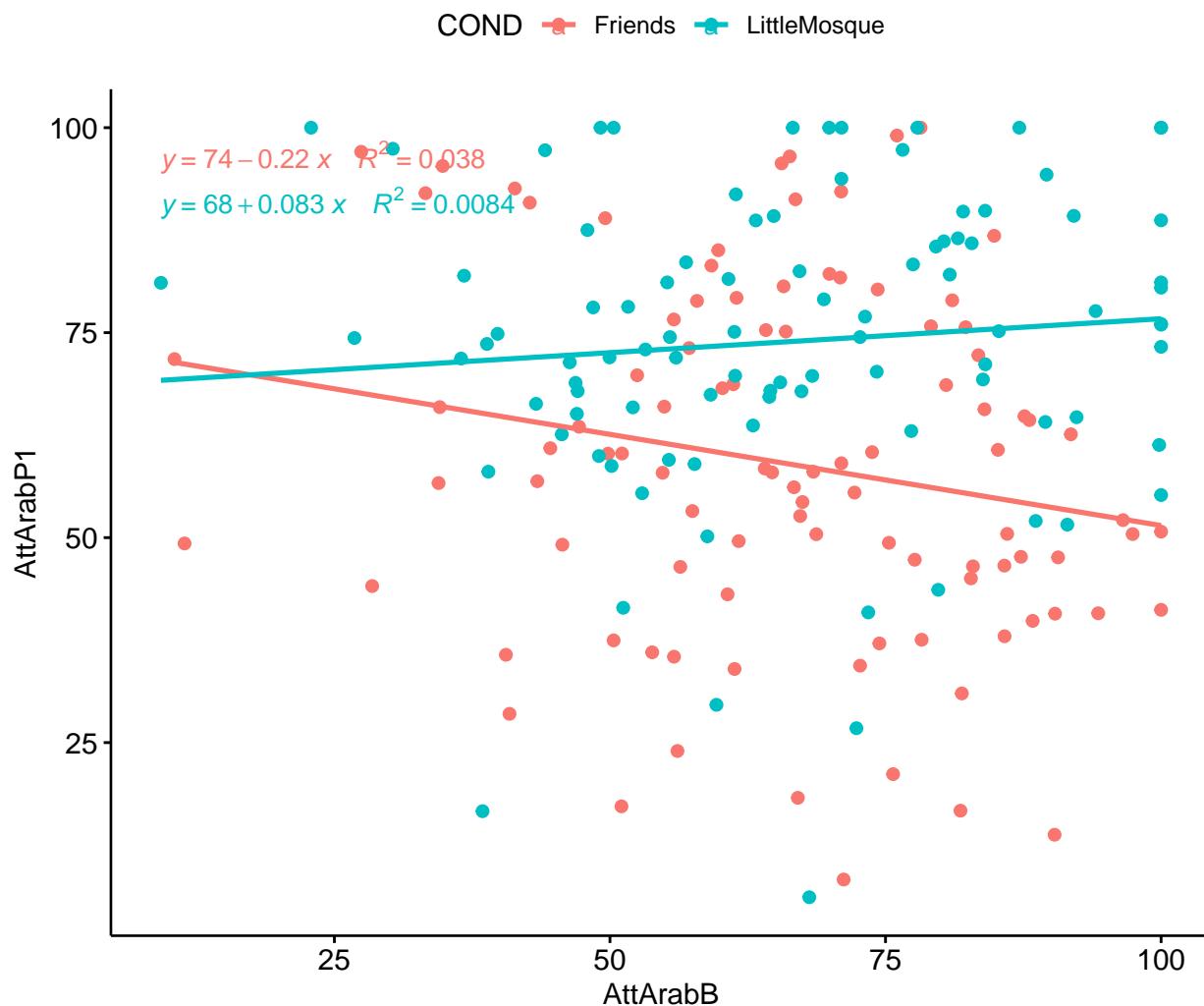
i Please use `after_stat(eq.label)` instead.

i The deprecated feature was likely used in the ggpublisher package.

Please report the issue at <<https://github.com/kassambara/ggpublisher/issues>>.

This warning is displayed once every 8 hours.

Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.



As is not surprising (because we tested a similar set of variables in the mixed design chapter), this relationship look like an interaction effect. Let's continue our exploration.

11.4.2.2 Homogeneity of regression slopes

This assumption requires that the slopes of the regression lines formed by the covariate and the outcome variable are the same for each group. The assumption evaluates that there is no interaction between the outcome and covariate. The plotted regression lines should be parallel.

```
Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ COND * AttArabB)
```

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	COND	1	189	26.819	0.000000569	*	0.124
2	AttArabB	1	189	0.676	0.412000000		0.004
3	COND:AttArabB	1	189	4.297	0.040000000	*	0.022

Because the statistically significant interaction term is violation of homogeneity of regression slopes ($F[1, 189] = 4.297, p = .040, \eta^2 = 0.022$) we should not proceed with ANCOVA as a statistical option. However, for the sake of demonstration, I will continue. One of the reasons I wanted to work this example as ANCOVA is to demonstrate that covariates and moderators each have their role. We can already see how this data is best analyzed with mixed design ANOVA.

11.4.2.3 Normality of residuals

Our goal here is to specify a model and extract *residuals*: the difference between the observed value of the DV and its predicted value. Each data point has one residual. The sum and mean of residuals are equal to 0.

Once we have saved the residuals, we can treat them as data and evaluate the shape of their distribution. We hope that the distribution is not statistically significantly different from a normal one. We first compute the model with *lm()* (*lm* stands for “linear model”). This is a linear regression.

```
# Create a linear regression model predicting DV from COV & IV
AttArabB_Mod <- lm(AttArabP1 ~ AttArabB + COND, data = Murrar_wide)
AttArabB_Mod
```

Call:

```
lm(formula = AttArabP1 ~ AttArabB + COND, data = Murrar_wide)
```

Coefficients:

(Intercept)	AttArabB	CONDLittleMosque
63.01428	-0.06042	14.92165

With the *broom::augment()* function we can augment our *lm()* model object to add fitted values and residuals.

```
# new model by augmenting the lm model
AttArabB_Mod.metrics <- broom::augment(AttArabB_Mod)
# shows the first three rows of the UEmodel.metrics
head(AttArabB_Mod.metrics, 3)

# A tibble: 3 x 9
  AttArabP1 AttArabB COND     .fitted .resid   .hat .sigma .cooksdi .std.resid
  <dbl>      <dbl> <fct>      <dbl>   <dbl>  <dbl>   <dbl>    <dbl>      <dbl>
1     80.3     74.3 Friends    58.5   21.7  0.0111  20.2  0.00440   1.08
2     76.6     55.8 Friends    59.6   17.0  0.0116  20.2  0.00280   0.845
3     92.0     33.3 Friends    61.0   31.0  0.0247  20.1  0.0204   1.56
```

From this, we can assess the normality of the residuals using the Shapiro Wilk test

```
# apply shapiro_test to that augmented model
rstatix::shapiro_test(AttArabB_Mod.metrics$.resid)
```

```
# A tibble: 1 x 3
  variable           statistic p.value
  <chr>              <dbl>    <dbl>
1 AttArabB_Mod.metrics$.resid  0.984  0.0261
```

The statistically significant Shapiro Wilk test has indicated a violation of the normality assumption ($W = 0.984$, $p = .026$).

11.4.2.4 Homogeneity of variances

ANCOVA presumes that the variance of the residuals is equal for all groups. We can check this with the Levene's test.

```
AttArabB_Mod.metrics %>%
  rstatix::levene_test(.resid ~ COND)
```

```
# A tibble: 1 x 4
  df1   df2 statistic     p
  <int> <int>    <dbl> <dbl>
1     1    191     3.52 0.0623
```

A non-significant Levene's test indicated no violation of the homogeneity of the residual variances for all groups ($F[1, 191] = 3.515$, $p = .062$).

11.4.2.5 Outliers

We can identify outliers by examining the standardized (or studentized) residuals. This is the residual divided by its estimated standard error. Standardized residuals are interpreted as the number of standard errors away from the regression line.

```
# from our model metrics show us any standardized residuals that are
# >3
AttArabB_Mod.metrics %>%
  filter(abs(.std.resid) > 3) %>%
  as.data.frame()
```

	AttArabP1	AttArabB	COND	.fitted	.resid	.hat	.sigma
1	6.137	68.085	LittleMosque	73.82234	-67.68534	0.01056251	19.62279
	.cooksdi	.std.resid					
1	0.04044273	-3.371254					

We do have one outlier with a standardized residual that has an absolute value greater than 3. At this point I am making a mental note of this. If this were “for real” I might more closely inspect these data. I would look at the whole response. If any response seemed invalid (e.g., random, extreme, or erratic responding) I would delete it. If the responses seemed valid, I *could* truncate them to exactly 3 SEs or. I could also ignore it. Kline [2016a] has a great section on some of these options.

Code for deleting outliers can be found in earlier chapters, including [Mixed Design ANOVA](#).

As noted by the suggestion of an interaction effect, our preliminary analyses suggests that ANCOVA is not the best option. We know from the prior lesson that a mixed design ANOVA worked well. In the spirit of an example, here’s a preliminary write-up so far:

11.4.2.6 Summarizing results from the analysis of assumptions

A one-way analysis of covariance (ANCOVA) was conducted. The independent variable, condition, had two levels: Friends, Little Mosque. The dependent variable was attitudes towards Arabs expressed by the participant at post-test and covariate was the pre-test assessment of the same variable. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate and the dependent variable differed significantly as a function of the independent variable, $F(1, 189) = 4.297, p = .040, \eta^2 = 0.022$. Regarding the assumption that the dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor, results of the Shapiro-Wilk test of normality on the model residuals was also significant, $W = 0.984, p = .026$. Only one datapoint (in the Little Mosque condition) had a standardized residual (-3.37) that exceeded an absolute value of 3.0. A non-significant Levene’s test indicated no violation of the homogeneity of the residual variances for all groups, $F(1, 191) = 3.515, p = .062$.

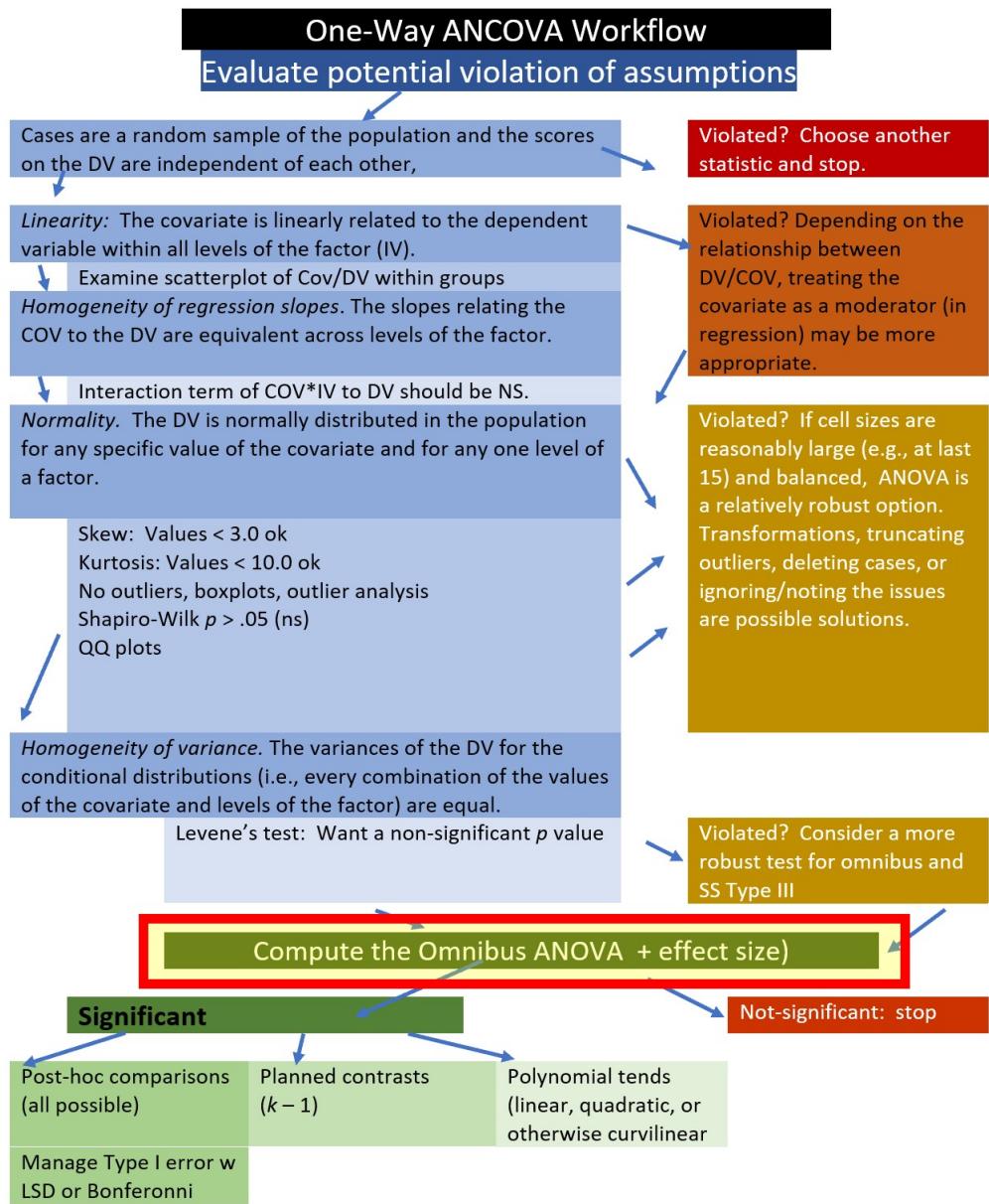


Figure 11.4: Image of the ANCOVA workflow, showing our current place in the process.

11.4.3 Calculating the Omnibus ANOVA

We are ready to conduct the omnibus ANOVA.

Order of variable entry matters in ANCOVA. Thinking of the *controlling for* language associate with covariates, we want to remove the effect of the covariate before we run the one-way ANOVA. With this ANCOVA we are asking the question, “Does the condition (Friends or Little Mosque) contribute to more positive attitudes toward Arabs, when controlling for the pre-test score?”

In repeated measures projects, we expect there to be dependency in the data. That is, in most cases prior waves will have significant prediction on later waves. When ANCOVA uses a prior assessment or wave as a covariate, that variable “claims” as much variance as possible and the subsequent variable can capture what is left over.

In the code below, we are predicting attitudes toward Arabs at post1 from the condition (Friends or Little Mosque), controlling for attitudes toward Arabs at baseline.

The *ges* column provides the effect size, η^2 . Conventionally, values of .01, .06, and .14 are considered to be small, medium, and large effect sizes, respectively.

You may see different values (.02, .13, .26) offered as small, medium, and large – these values are used when multiple regression is used. A useful summary of effect sizes, guide to interpreting their magnitudes, and common usage can be found [here \[Watson, 2020\]](#).

```
MurrarB_ANCOVA <- Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ AttArabB + COND)
rstatix::get_anova_table(MurrarB_ANCOVA)
```

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	AttArabB	1	190	0.665	0.416	0.000000000	0.003
2	COND	1	190	26.361	0.000000698	*	0.122

There was a non-significant effect of the baseline covariate on the post-test ($F[1, 190] = 0.665, p = .416, \eta^2 = 0.003$). After controlling for the baseline attitudes toward Arabs, there was a statistically significant effect of condition on post-test attitudes toward Arabs, $F(1, 190) = 26.361, p < .001, \eta^2 = 0.122$. This effect appears to be moderate-to-large in size.

11.4.4 Post-hoc pairwise comparisons (controlling for the covariate)

Just like in one-way ANOVA, we follow-up the significant effect of condition. We'll use all-possible pairwise comparisons. In our case, we only have two levels of the categorical factor, so this run wouldn't be necessary. I included it to provide the code for doing so. If there were three or more variables, we would see all possible comparisons.

```
pwc_B <- Murrar_wide %>%
  rstatix::emmeans_test(AttArabP1 ~ COND, covariate = AttArabB, p.adjust.method = "none")
pwc_B
```

```
# A tibble: 1 x 9
  term          .y. group1 group2    df statistic      p   p.adj p.adj.signif
* <chr>        <chr> <chr>  <dbl>     <dbl>    <dbl>    <dbl> <chr>
1 AttArabB*COND AttA~ Frien~ Littl~    190     -5.13 6.98e-7 6.98e-7 ****
```

Not surprisingly (since this single pairwise comparison is redundant with the omnibus ANCOVA), results suggest a statistically significant difference between Friends and Little Mosque at Post1.

With the script below we can obtain the covariate-adjusted marginal means. These are termed *estimated marginal means*. Take a look at these and compare them to what we would see in the regular descriptives. It is helpful to see the grand mean (AttArabB) and then the marginal means (emmmean).

```
emmeans_B <- rstatix::get_emmeans(pwc_B)
emmeans_B
```

```
# A tibble: 2 x 8
  AttArabB COND       emmean     se     df conf.low conf.high method
  <dbl> <fct>      <dbl> <dbl> <dbl>    <dbl>    <dbl> <chr>
1     66.2 Friends    59.0  2.04  190     55.0     63.0 Emmeans test
2     66.2 LittleMosque 73.9  2.07  190     69.8     78.0 Emmeans test
```

Note that the *emmeans* process produces slightly different means than the raw means produced with the *psych* package's *describeBy()* function. Why? Because the *get_emmeans()* function uses the model that included the covariate. That is, the *estimated* means are covariate-adjusted.

```
descripts_P1 <- psych::describeBy(AttArabP1 ~ COND, data = Murrar_wide,
                                    mat = TRUE)
descripts_P1
```

	item	group1	vars	n	mean	sd	median	trimmed			
AttArabP11	1	Friends	1	98	59.02351	21.65024	57.9955	59.31306			
AttArabP12	2	LittleMosque	1	95	73.92134	18.51082	74.4600	75.52858			
					mad	min	max	range	skew	kurtosis	se
AttArabP11	23.67045	8.297	100	91.703	-0.0518848	-0.6252126	2.187005				
AttArabP12	15.98984	6.137	100	93.863	-0.9798189	1.6335325	1.899170				

```
# Note. Recently my students and I have been having intermittent
# struggles with the describeBy function in the psych package. We
# have noticed that it is problematic when using .rds files and when
# using data directly imported from Qualtrics. If you are having
# similar difficulties, try uploading the .csv file and making the
# appropriate formatting changes.
```

$(M = 59.02, SD = 21.65)$ $(M = 73.92, SD = 18.51)$

In our case the adjustments are very minor. Why? The effect of the attitudes toward Arabs baseline test on the attitudes toward Arabs post test was nonsignificant. We can see this in the bivariate correlations, below.

```
MurP1_Rmat <- psych::corr.test(Murrar_wide[c("AttArabB", "AttArabP1")])
MurP1_Rmat
```

```
Call:psych::corr.test(x = Murrar_wide[c("AttArabB", "AttArabP1")])
Correlation matrix
  AttArabB AttArabP1
AttArabB    1.00   -0.05
AttArabP1   -0.05    1.00
Sample Size
[1] 193
Probability values (Entries above the diagonal are adjusted for multiple tests.)
  AttArabB AttArabP1
AttArabB    0.00    0.47
AttArabP1   0.47    0.00
```

To see confidence intervals of the correlations, print with the `short=FALSE` option

The correlation between attitudes toward Arabs at baseline and post test are nearly negligible ($r = -0.05, p = .47$).

11.4.5 APA style results for Scenario 1

As we assemble the elements for an APA style result sections, a table with the means, adjusted means, and correlations may be helpful.

```
apaTables::apa.cor.table(Murrar_wide[c("AttArabB", "AttArabP1")], table.number = 1)
```

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1
1. AttArabB	66.25	19.66	
2. AttArabP1	66.36	21.46	-.05 [-.19, .09]

Note. M and SD are used to represent mean and standard deviation, respectively.
Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

```
# You can save this as a Microsoft word document by adding this
# statement into the command: filename = 'your_filename.doc'
```

Additionally, writing this output to excel files helped create the two tables that follow. The *MASS* package is useful to export the model objects into .csv files. They are easily opened in Excel where they can be manipulated into tables for presentations and manuscripts.

```
MASS::write.matrix(pwc_B, sep = ",", file = "pwc_B.csv")
MASS::write.matrix(emmeans_B, sep = ",", file = "emmeans_B.csv")
MASS::write.matrix(descripts_P1, sep = ",", file = "descripts_P1.csv")
```

Ultimately, I would want a table that included this information. Please refer to the APA style manual for more proper formatting for a manuscript that requires APA style.

Table 1

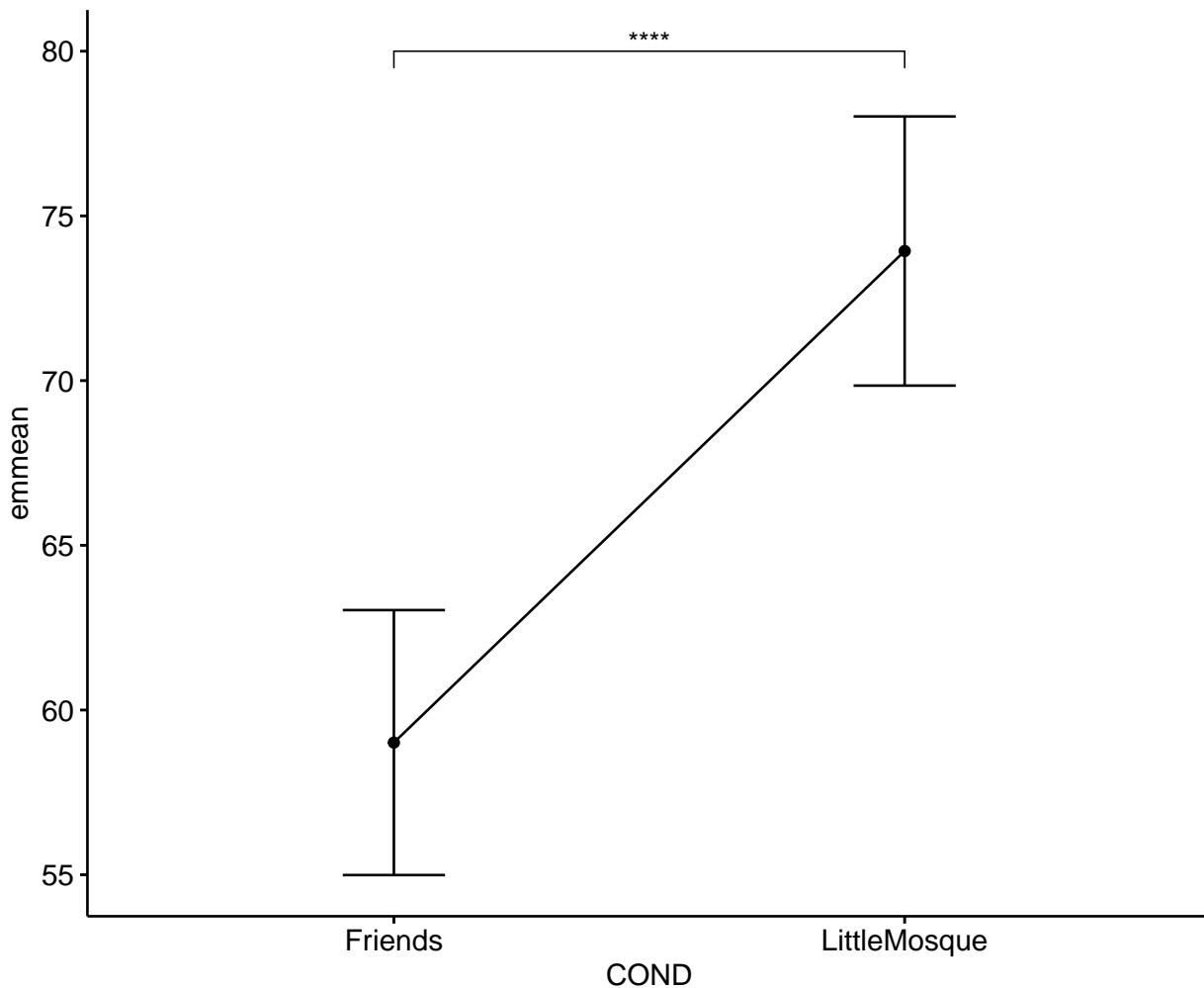
Unadjusted and Covariate-Adjusted Descriptive Statistics

Condition	Unadjusted	Covariate-Adjusted		
	<i>M</i>	<i>SD</i>	<i>EMM</i>	<i>SE</i>
Friends	59.02	21.65	59.01	2.04
Little Mosque	73.92	18.51	73.93	2.07

Unlike the figure we created when we were testing assumptions, this script creates a plot from the model (which identifies AttArabB in its role as covariate). Thus, the relationship between condition and AttArabP1 controls for the effect of the AttArabB covariate.

```
pwc_B <- pwc_B %>%
  rstatix::add_xy_position(x = "COND", fun = "mean_se")
ggpubr::ggline(rstatix::get_emmeans(pwc_B), x = "COND", y = "emmean", title = "Figure 10.14 Pos
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) +
  ggpubr::stat_pvalue_manual(pwc_B, hide.ns = TRUE, tip.length = 0.02,
  y.position = c(80))
```

Figure 10.14 Post-test Attitudes by Condition, Controlling for Pre-test Attitude



Results

A one-way analysis of covariance (ANCOVA) was conducted. The independent variable, condition, had two levels: Friends, Little Mosque. The dependent variable was attitudes towards Arabs expressed by the participant at post-test and covariate was the baseline assessment of the same variable. Descriptive statistics are presented in Table 1. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate and the dependent variable differed significantly as a function of the independent variable, $F(1, 189) = 4.297, p = .040, \eta^2 = 0.022$. Regarding the assumption that the dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor, results of the Shapiro-Wilk test of normality on the model residuals was also significant, $W = 0.984, p = .026$. Only one datapoint (in the Little Mosque condition) had a standardized residual (-3.37) that exceeded an absolute value of 3.0. A non-significant Levene's test indicated no violation of the homogeneity of the residual variances for all groups, $F(1, 191) = 3.515, p = .062$.

There was a non-significant effect of the baseline covariate on the post-test ($F[1, 190] =$

$0.665, p = .416, \eta^2 = 0.003$). After controlling for the baseline attitudes toward Arabs, there was a statistically significant effect of condition on post-test attitudes toward Arabs, $F(1, 190) = 26.361, p < .001, \eta^2 = 0.122$. This effect appears to be moderate-to-large. Given there were only two conditions, no further follow-up was required. As illustrated in Figure 1, results suggest that those in the Little Mosque condition ($M = 73.92, SD = 18.51$) had more favorable attitudes toward Arabs than those in the Friends condition ($M = 59.02, SD = 21.65$). Means and covariate-adjusted means are presented in Table 1b.

11.5 Working the ANCOVA – Scenario #2: Controlling for a confounding or covarying variable

In the scenario below, I am simulating a one-way ANCOVA, predicting attitudes toward Arabs at post1 as a function of sitcom condition (Friends, Little Mosque), controlling for the participants' attitudes toward Whites. That is, the ANCOVA will compare the the means of the two groups (at post1, only), adjusted for level of attitudes toward Whites

TO BE CLEAR: This is not the best way to analyze this data. With such a strong, balanced design, the multi-way, mixed design ANOVAs were an excellent choice that provided much fuller information than this demonstration, below. The purpose of this over-simplified demonstration is merely to give another example of using a variable as a *covariate* rather than a *moderator*.

11.5.1 Preparing the data

When the covariate in ANCOVA is a potentially confounding variable, we need three variables:

- IV that has two or more levels; in our case it is the Friends and Littls Mosque sitcom conditions.
 - DV that is continuous; in our case it attitudes toward Arabs at post1 (AttArabP1).
 - Covariate that is continuous; in our case it attitudes toward Whites at post1 (AttWhiteP1).
- Note* We could have also chosen attitudes toward Whites at baseline.

We can continue using the Murrar_wide df.

11.5.2 Evaluating the statistical assumptions

There are a number of assumptions in ANCOVA. These include:

- random sampling
- independence in the scores representing the dependent variable
- linearity of the relationship between the covariate and DV within all levels of the independent variable
- homogeneity of the regression slopes
- a normally distributed DV for any specific value of the covariate and for any one level of a factor

11.5. WORKING THE ANCOVA – SCENARIO #2: CONTROLLING FOR A CONFOUNDING OR COVARYING VARIABLE

- homogeneity of variance

These are depicted in the flowchart, below.

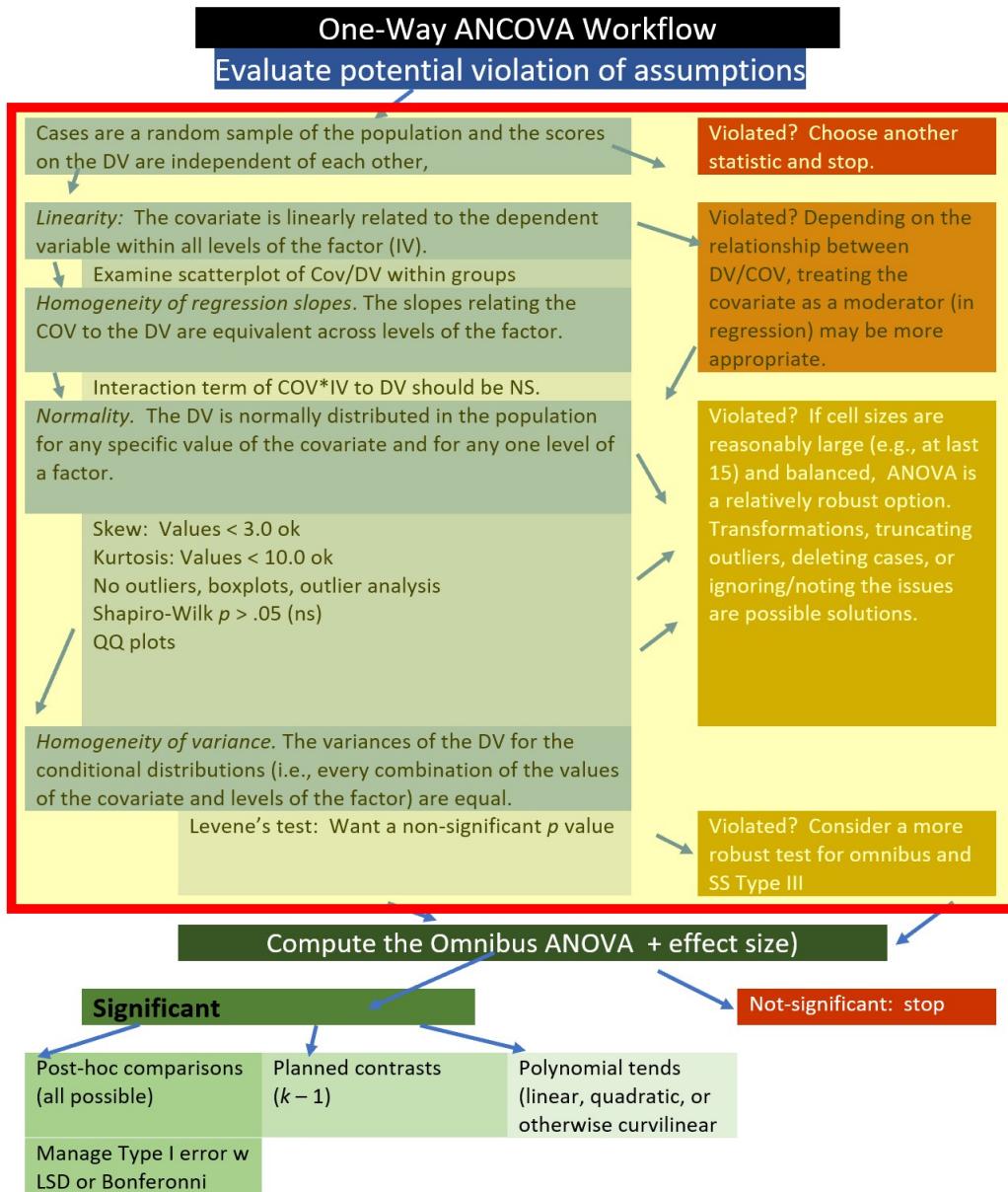


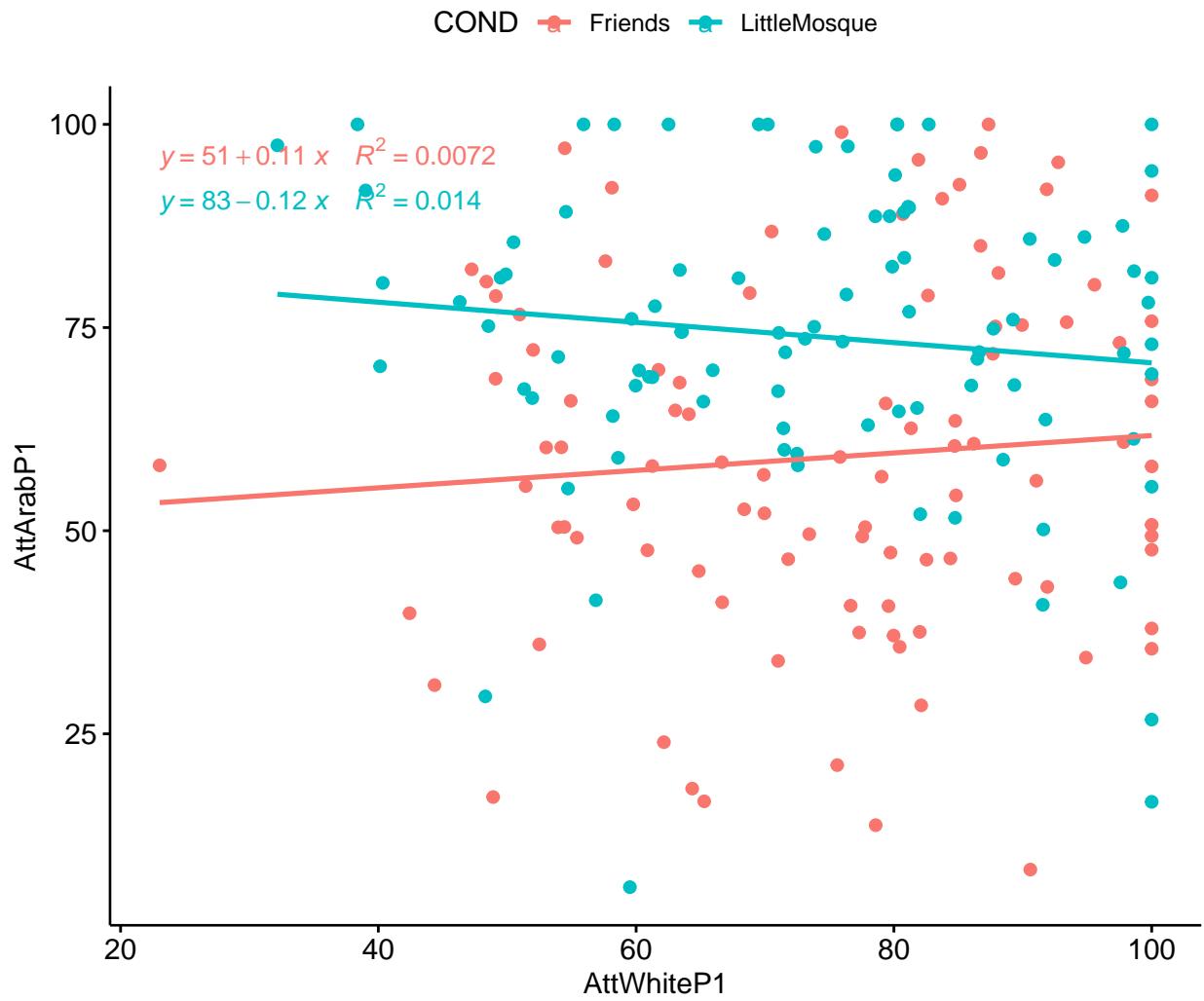
Figure 11.5: Image of the ANCOVA workflow, showing our current place in the process

11.5.2.1 Linearity assumption

ANCOVA assumes that there is linearity between the covariate and outcome variable at each level of the grouping variable. In our case this means that there is linearity between the attitudes toward Whites (covariate) and attitudes toward Arabs (outcome variable) at each level of the intervention (Friends, Little Mosque).

We can create a scatterplot (with regression lines) between the covariate (attitudes toward Whites) and the outcome (attitudes toward Arabs).

```
ggpubr::ggscatter(Murrar_wide, x = "AttWhiteP1", y = "AttArabP1", color = "COND",
  add = "reg.line") + ggpubr::stat_regrline_equation(aes(label = paste(..eq.label..,
  ..rr.label.., sep = "~~~~"), color = COND))
```



As we look at this scatterplot, we are trying to determine if there is an interaction effect (rather than a covarying effect). The linearity here looks reasonable and not terribly “interacting” (to help us decide whether empathy should be a covariate or a moderator). More testing can help us make this distinction.

11.5.2.2 Homogeneity of regression slopes

This assumption requires that the slopes of the regression lines formed by the covariate and the outcome variable are the same for each group. The assumption evaluates that there is no interaction between the outcome and covariate. The plotted regression lines should be parallel.

11.5. WORKING THE ANCOVA – SCENARIO #2: CONTROLLING FOR A CONFOUNDING OR COVARYING VARIABLE

```
Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ COND * AttWhiteP1)
```

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	COND	1	189	26.240	0.00000074	*	0.1220000
2	AttWhiteP1	1	189	0.014	0.90700000		0.0000729
3	COND:AttWhiteP1	1	189	1.886	0.17100000		0.0100000

Preliminary analysis supported ANCOVA as a statistical option in that there was no violation of the homogeneity of regression slopes as the interaction term was not statistically significant, $F(1, 189) = 1.886, p = .171, \eta^2 = 0.010$.

11.5.2.3 Normality of residuals

Assessing the normality of residuals means running the model, capturing the unexplained portion of the model (i.e., the *residuals*), and then seeing if they are normally distributed. Proper use of ANCOVA is predicated on normally distributed residuals.

We first compute the model with *lm()*. The *lm()* function is actually testing what we want to test. However, at this early stage, we are just doing a “quick run and interpretation” to see if we are within the assumptions of ANCOVA.

```
# Create a linear regression model predicting DV from COV & IV
WhCov_mod <- lm(AttArabP1 ~ AttWhiteP1 + COND, data = Murrar_wide)
WhCov_mod
```

Call:

```
lm(formula = AttArabP1 ~ AttWhiteP1 + COND, data = Murrar_wide)
```

Coefficients:

(Intercept)	AttWhiteP1	CONDLittleMosque
59.765300	-0.009897	14.886178

We can use the *augment(model)* function from the *broom* package to add fitted values and residuals.

```
WhCov_mod.metrics <- broom::augment(WhCov_mod)
# shows the first three rows of the UEcon_model.metrics
head(WhCov_mod.metrics, 3)
```

```
# A tibble: 3 x 9
  AttArabP1 AttWhiteP1 COND    .fitted .resid   .hat .sigma .cooksdi .std.resid
  <dbl>      <dbl> <fct>     <dbl>  <dbl>  <dbl>  <dbl>    <dbl>      <dbl>
```

1	80.3	95.6	Friends	58.8	21.4	0.0176	20.2	0.00685	1.07
2	76.6	51.0	Friends	59.3	17.3	0.0203	20.2	0.00518	0.867
3	92.0	91.9	Friends	58.9	33.2	0.0152	20.1	0.0140	1.65

Now we assess the normality of residuals using the Shapiro Wilk test. The script below captures the “.resid” column from the model.

```
rstatix::shapiro_test(WhCov_mod.metrics$.resid)
```

```
# A tibble: 1 x 3
  variable      statistic p.value
  <chr>          <dbl>     <dbl>
1 WhCov_mod.metrics$.resid 0.984   0.0294
```

The statistically significant Shapiro Wilk test indicate a violation of the normality assumption ($W = 0.984, p = .029$). As I mentioned before, there are better ways to analyze this research vignette. None-the-less, we will continue with this demonstration so that you will have the procedural and conceptual framework for conducting ANCOVA.

11.5.2.4 Homogeneity of variances

ANCOVA presumes that the variance of the residuals is equal for all groups. We can check this with the Levene’s test.

```
WhCov_mod.metrics %>%
  rstatix::levene_test(.resid ~ COND)
```

```
# A tibble: 1 x 4
  df1   df2 statistic    p
  <int> <int>    <dbl> <dbl>
1     1    191     4.54 0.0344
```

Contributing more evidence that ANCOVA is not the best way to analyze this data, a statistically significant Levene’s test indicates a violation of the homogeneity of the residual variances ($F[1, 191] = 4.539, p = .034$).

11.5.2.5 Outliers

We can identify outliers by examining the standardized (or studentized) residual. This is the residual divided by its estimated standard error. Standardized residuals are interpreted as the number of standard errors away from the regression line.

```
WhCov_mod.metrics %>%
  filter(abs(.std.resid) > 3) %>%
  as.data.frame()
```

11.5. WORKING THE ANCOVA – SCENARIO #2: CONTROLLING FOR A CONFOUNDING OR COVARYING VARIABLE

```
AttArabP1 AttWhiteP1           COND .fitted   .resid      .hat   .sigma
1       6.137     59.518 LittleMosque 74.06242 -67.92542 0.01407535 19.65185
       .cooksdi .std.resid
1 0.05447684 -3.383443
```

There is one outlier with a standardized residual with an absolute value greater than 3. At this point I am making a mental note of this. If this were “for real” I might more closely inspect these data. I would look at the whole response. If any response seems invalid (e.g., random, erratic, or extreme responding) I would delete it. If the response seem valid, I *could* truncate them to within 3 SEs. I could also ignore it. Kline [2016a] has a great section on some of these options.

11.5.2.6 Summarizing the results from the analysis of assumptions

A one-way analysis of covariance (ANCOVA) was conducted. The independent variable, sitcom condition, had two levels: Friends, Little Mosque. The dependent variable was attitudes towards Arabs at pre-test. Preliminary analyses which tested the assumptions of ANCOVA were mixed. Results suggesting that the relationship between the covariate and the dependent variable did not differ significantly as a function of the independent variable ($F[1, 189] = 1.886, p = .171, \eta^2 = 0.010$) provided evidence that we did not violate the homogeneity-of-slopes assumption. In contrast, the Shapiro-Wilk test of normality on the model residuals was statistically significant ($W = 0.984, p = .029$). This means that we likely violated the assumption that the dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor. Regarding outliers, one datapoint (-3.38) had a standardized residual that exceeded an absolute value of 3.0. Further, a statistically significant Levene’s test indicated a violation of the homogeneity of the residual variances for all groups, ($F[1, 191] = 4.539, p = .034$).

Because the intent of this analysis was to demonstrate how ANCOVA differs from mixed design ANOVA we proceeded with the analysis. Were this for “real research” we would have chosen a different analysis.

11.5.3 Calculating the Omnibus ANOVA

We are ready to conduct the omnibus ANOVA.

Order of variable entry matters in ANCOVA. Thinking of the *controlling for* language associated with covariates, we firstly want to remove the effect of the covariate.

In the code below we are predicting attitudes toward Arabs at post1 from attitudes toward Whites at post1 (the covariate) and sitcom condition (Friends, Little Mosque).

The *ges* column provides the effect size, η^2 where a general rule-of-thumb for interpretation is .01 (small), .06 (medium), and .14 (large) [Lakens, 2013].

```
WhCov_ANCOVA <- Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ AttWhiteP1 + COND)
rstatix::get_anova_table(WhCov_ANCOVA)
```

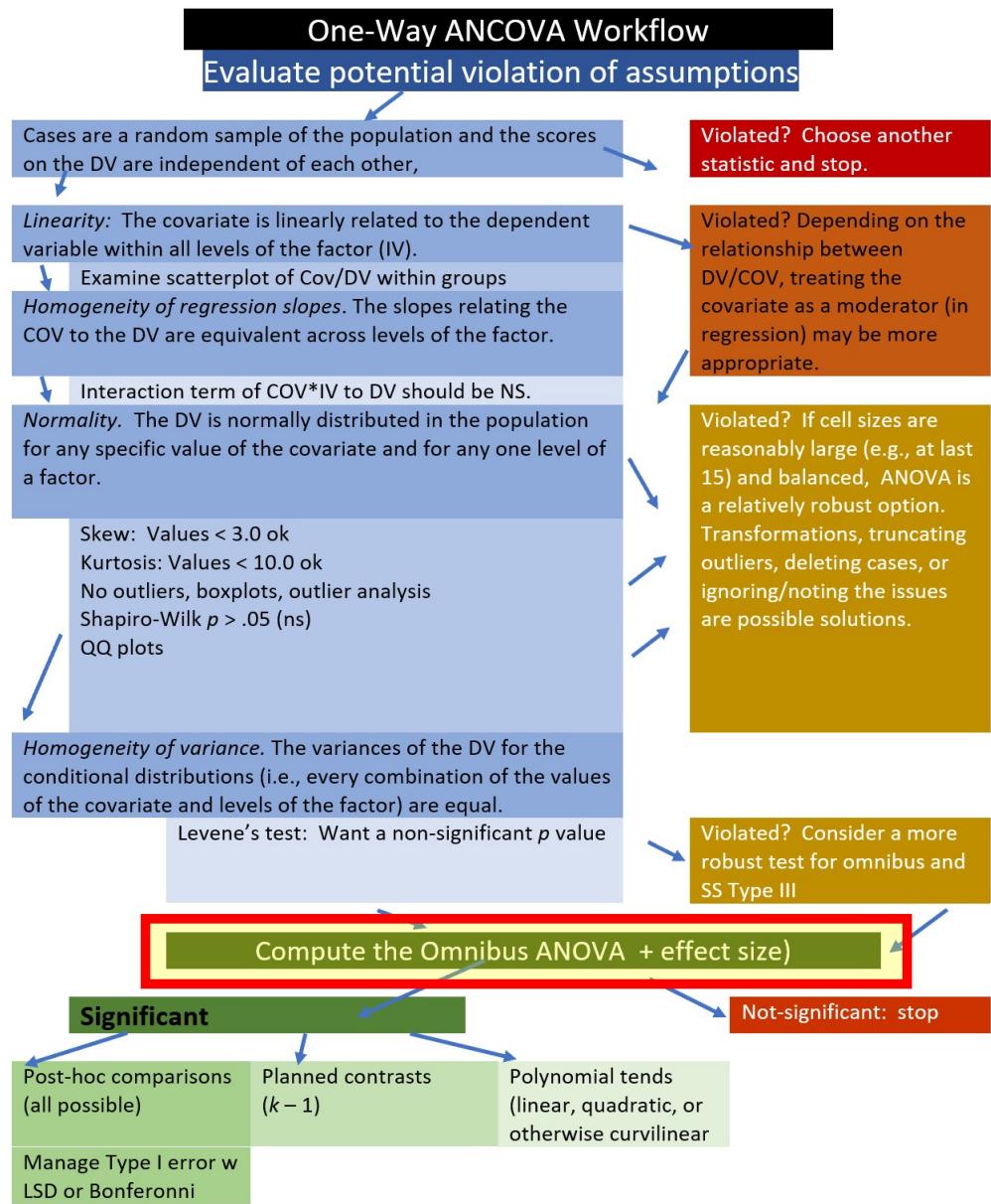


Figure 11.6: Image of the ANCOVA workflow, showing our current place in the process.

11.5. WORKING THE ANCOVA – SCENARIO #2: CONTROLLING FOR A CONFOUNDING OR COVARYING VARIABLE

ANOVA Table (type II tests)

	Effect	DFn	DFd	F	p	p<.05	ges
1	AttWhiteP1	1	190	0.014	0.907000000		0.0000722
2	COND	1	190	26.119	0.000000779	*	0.1210000

There was a non-significant effect of the attitudes toward Whites covariate on the attitudes toward Arabs at post-test, $F(1, 190) = 0.014, p = .907, \eta^2 < .001$. After controlling for attitudes toward Whites, there was a statistically significant effect in attitudes toward Arabs at post-test between the conditions, $F(1, 190) = 26.119, p < .001, \eta^2 = 0.121$. The effect size was moderate-to-large.

11.5.4 Post-hoc pairwise comparisons (controlling for the covariate)

With only two levels of sitcom condition (Friends, Little Mosque), we do not need to conduct post-hoc pairwise comparisons. However, because many research designs involve three or more levels, I will use code that evaluates them here.

```
pwc_cond <- Murrar_wide %>%
  rstatix::emmeans_test(AttArabP1 ~ COND, covariate = AttWhiteP1, p.adjust.method = "none")
pwc_cond
```

```
# A tibble: 1 x 9
  term      .y. group1 group2   df statistic      p  p.adj p.adj.signif
* <chr>    <chr> <chr>  <dbl>    <dbl>  <dbl> <dbl> <chr>
1 AttWhiteP1*C~ AttA~ Frien~ Littl~    190     -5.11 7.79e-7 7.79e-7 ****
```

Results suggest a statistically significant post-test difference between the Friends and Little Mosque sitcom conditions. With the script below we can obtain the covariate-adjusted marginal means. These are termed *estimated marginal means*.

```
emmeans_cond <- rstatix::get_emmeans(pwc_cond)
emmeans_cond
```

```
# A tibble: 2 x 8
  AttWhiteP1 COND       emmean     se   df conf.low conf.high method
  <dbl> <fct>      <dbl> <dbl> <dbl>    <dbl>  <dbl> <chr>
1     74.4 Friends     59.0  2.04   190     55.0    63.1 Emmeans test
2     74.4 LittleMosque 73.9  2.08   190     69.8    78.0 Emmeans test
```

As before, these means are usually different (even if only ever-so-slightly) than the raw means you would obtain from the descriptives.

```
descripts_cond <- psych::describeBy(AttArabP1 ~ COND, data = Murrar_wide,
  mat = TRUE)
descripts_cond
```

	item	group1	vars	n	mean	sd	median	trimmed	
	AttArabP11	1	Friends	1	98	59.02351	21.65024	57.9955	59.31306
	AttArabP12	2	LittleMosque	1	95	73.92134	18.51082	74.4600	75.52858
				mad	min	max	range	skew	kurtosis
	AttArabP11			23.67045	8.297	100	91.703	-0.0518848	-0.6252126
	AttArabP12			15.98984	6.137	100	93.863	-0.9798189	1.6335325
								se	
								2.187005	
									1.899170

11.5.5 APA style results for Scenario 2

Tables with the means, adjusted means, and pairwise comparison output may be helpful. The *apa.cor.table()* function in the *apaTables* package is helpful for providing means, standarddeviations, and correlations.

```
apaTables::apa.cor.table(Murrar_wide[c("AttArabP1", "AttWhiteP1")], table.number = 2)
```

Table 2

Means, standard deviations, and correlations with confidence intervals

Variable	M	SD	1
1. AttArabP1	66.36	21.46	
2. AttWhiteP1	74.37	17.28	-.02 [-.16, .12]

Note. M and SD are used to represent mean and standard deviation, respectively.
Values in square brackets indicate the 95% confidence interval.

The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

* indicates $p < .05$. ** indicates $p < .01$.

```
# You can save this as a Microsoft word document by adding this
# statement into the command: filename = 'your_filename.doc'
```

Writing this output to excel files helped create the two tables that follow.

```
MASS::write.matrix(pwc_cond, sep = ",", file = "pwc_con.csv")
MASS::write.matrix(emmeans_cond, sep = ",", file = "emmeans_con.csv")
MASS::write.matrix(descripts_cond, sep = ",", file = "descripts_con.csv")
```

11.5. WORKING THE ANCOVA – SCENARIO #2: CONTROLLING FOR A CONFOUNDING OR COVARYING VARIABLE

Ultimately, I would want a table that included this information. Please refer to the APA style manual for more proper formatting for a manuscript that requires APA style.

Table 1b

Unadjusted and Covariate-Adjusted Descriptive Statistics

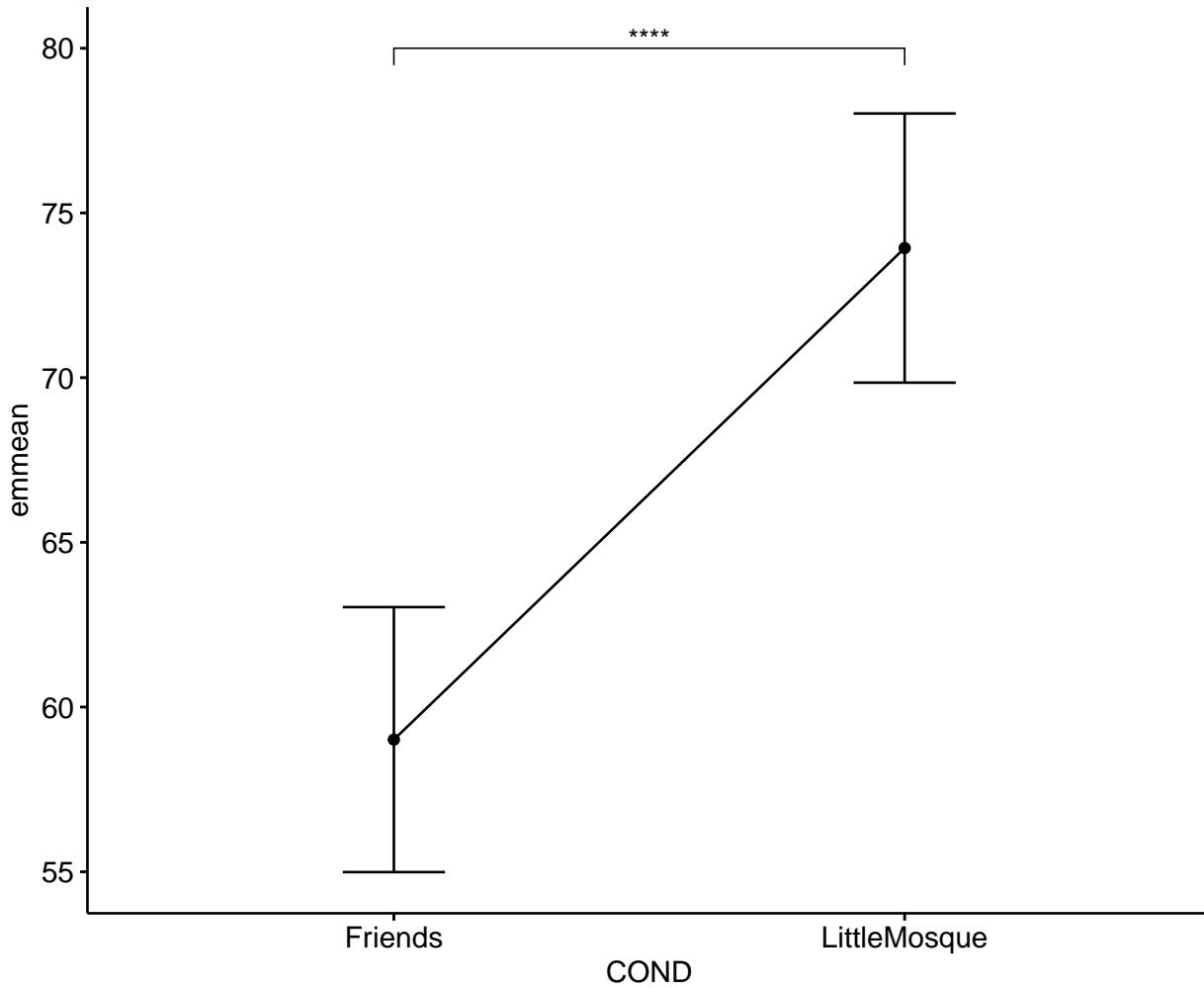
Condition	Unadjusted	Covariate-Adjusted
-----------	------------	--------------------

	<i>M</i>	<i>SD</i>	<i>EMM</i>	<i>SE</i>
Friends	59.02	21.65	59.03	2.04
Little Mosque	73.92	18.51	73.92	2.08

Unlike the figure we created when we were testing assumptions, this script creates a plot from the model (which identifies AttWhiteP1 in its role as covariate). Thus, the relationship between condition and AttArabP1 controls for the effect of the AttArabB covariate.

```
pwc_cond <- pwc_cond %>%
  rstatix::add_xy_position(x = "COND", fun = "mean_se")
ggpubr::ggline(rstatix::get_emmeans(pwc_B), x = "COND", y = "emmean", title = "Figure 10.5 AttarabB vs Condition", y.position = 80)
```

Figure 10.5 Attitudes toward Arabs by Condition, Controlling for Attitudes to



Results

A one-way analysis of covariance (ANCOVA) was conducted. The independent variable, sitcom condition, had two levels: Friends, Little Mosque. The dependent variable was attitudes towards Arabs at pre-test. We controlled for attitudes toward Whites. Preliminary analyses which tested the assumptions of ANCOVA were mixed. Results suggesting that the relationship between the covariate and the dependent variable did not differ significantly as a function of the independent variable ($F[1, 189] = 1.886, p = .171, \eta^2 = 0.010$) provided evidence that we did not violate the homogeneity-of-slopes assumption. In contrast, the Shapiro-Wilk test of normality on the model residuals was statistically significant ($W = 0.984, p = .029$). This means that we likely violated the assumption that the dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor. Regarding outliers, one datapoint (-3.38) had a standardized residual that exceeded an absolute value of 3.0. Further, a statistically significant Levene's test indicated a violation of the homogeneity of the residual variances for all groups, ($F[1, 191] = 4.539, p = .034$).

Because the intent of this analysis was to demonstrate how ANCOVA differs from mixed design

ANOVA we proceeded with the analysis. Were this for “real research” we would have chosen a different analysis.

There was a non-significant effect of the attitudes toward Whites covariate on the attitudes toward Arabs post-test, $F(1, 190) = 0.014, p = .907, \eta^2 < .001$. After controlling for attitudes toward Whites, there was a statistically significant effect in attitudes toward Arabs at post-test between the conditions, $F(1, 190) = 26.119, p < .001, \eta^2 = 0.121$. The effect size was moderately large. Means and covariate-adjusted means are presented in Table 1b.

11.6 More (and a recap) on covariates

Covariates, sometimes termed *controls* are often used to gain statistical control over variables that are difficult to control in a research design. That is, it may be impractical for polychotomize an otherwise continuous variable and/or it is impractical to have multiple factors and so a covariate is a more manageable approach. Common reasons for including covariates include [Bernerth and Aguinis, 2016]:

- they mathematically remove variance associated with nonfocal variables,
- the *purification principle* – removing unwanted or confusing variance,
- they remove the *noise* in the analysis to clear up the relationship between IV and DVs.

Perhaps it is an oversimplification, but we can think of three categories of variables: moderators, covariates, and mediators. Through ANOVA and ANCOVA, we distinguish between moderator and covariate.

Moderator: a variable that changes the strength or direction of an effect between two variables X (predictor, independent variable) and Y (criterion, dependent variable).

Covariate: an observed, continuous variable, that (when used properly) has a relationship with the dependent variable. It is included in the analysis, as a predictor, so that the predictive relationship between the independent (IV) and dependent (DV) are adjusted.

Bernerth and Aguinis [2016] conducted a review of how and when control variables were used in nearly 600 articles published between 2003 and 2012. Concurrently with their analysis, they provided guidance for when to use control variables (covariates). The flowchart that accompanies their article is quite helpful. Control variables (covariates) should only be used when:

1. Theory suggests that the potential covariate(s) relate(s) to variable(s) in the currrent study.
2. There is empirical justification for including the covariate in the study.
3. The covariate can be measured reliably.

Want more? Instructions for calculating a two-way ANCOVA are here: <https://www.datanovia.com/en/lessons/ancova-in-r/>

11.7 Practice Problems

The suggestions for homework differ in degree of complexity. I encourage you to start with a problem that feels “do-able” and then try at least one more problem that challenges you in some way. At a minimum your data should have three levels in the independent variable. At least one of the problems you work should have a statistically significant interaction effect that you work all the way through.

Regardless, your choices should meet you where you are (e.g., in terms of your self-efficacy for statistics, your learning goals, and competing life demands). Whichever you choose, you will focus on these larger steps in one-way ANCOVA, including:

- test the statistical assumptions
- conduct an ANCOVA
- if the predictor variable has more than three or more levels, conduct follow-up testing
- present both means and covariate-adjusted means
- write a results section to include a figure and tables

11.7.1 Problem #1: Play around with this simulation.

Copy the script for the simulation and then change (at least) one thing in the simulation to see how it impacts the results.

- If ANCOVA is new to you, perhaps you just change the number in “set.seed(210813)” from 210813 to something else. Then rework Scenario#1, Scenario#2, or both. Your results should parallel those obtained in the lecture, making it easier for you to check your work as you go.
- If you are interested in power, change the sample size to something larger or smaller.
- If you are interested in variability (i.e., the homogeneity of variance assumption), perhaps you change the standard deviations in a way that violates the assumption.

11.7.2 Problem #2: Conduct a one-way ANCOVA with the DV and covariate at post2.

The Murrar et al. [2018] article has three waves: baseline, post1, post2. In this lesson, I focused on the post1 waves. Rerun this analysis using the post2 wave data.

11.7.3 Problem #3: Try something entirely new.

Using data for which you have permission and access (e.g., IRB approved data you have collected or from your lab; data you simulate from a published article; data from an open science repository; data from other chapters in this OER), complete an ANCOVA.

11.7.4 Grading Rubric

Regardless which option(s) you chose, use the elements in the grading rubric to guide you through the practice. Using the lecture and workflow (chart) as a guide, please work through all the steps listed in the proposed assignment/grading rubric.

Assignment Component	Points Possible	Points Earned
1. Narrate the research vignette, describing the IV and DV. Minimally, the data at least three levels in the independent variable. At least one of the problems you work should have a significant omnibus test so that follow-up is required.	5	_____
2. Check and, if needed, format data	5	_____
3. Evaluate statistical assumptions	5	_____
4. Conduct omnibus ANCOVA (w effect size)	5	_____
5. If the IV has three or more levels, conduct follow-up tests	5	_____
6. Present means and covariate-adjusted means; interpret them	5	_____
7. APA style results with table(s) and figure	5	_____
8. Explanation to grader	5	_____
Totals	35	_____

References

APPENDICES

Chapter 12

Type I Error

Type I Error Defined

Type I error is the concern about false positives – that we would incorrectly reject a true null hypothesis (that we would say that there is a statistically significant difference when there is not one). This concern is increased when there are multiple hypothesis tests. This concern increases when we have a large number of pairwise comparisons.

Throughout the chapters, I noted the importance and relative risk of Type I error with each statistic and options for follow-up testing. Because there are so many options, I have provided a review and summary of each option in this appendix. For each, I provide a definition, a review of the steps and options for utilizing the statistic, and suggest the types of follow-up for which this approach is indicated.

Methods for Managing Type I Error

LSD (Least Significant Difference) Method

The LSD method is especially appropriate in the one-way ANOVA scenario when there are only three levels in the factor. In this case, Green and Salkind [2017c] have suggested that alpha can be retained at the alpha level for the “family” (α_{family}), which is conventionally $p = .05$ and used both to evaluate the omnibus and, so long as they don’t exceed three in number, the planned or pairwise comparisons that follow.

Traditional Bonferroni

The *traditional Bonferroni* is, perhaps, the most well-known approach to managing Type I error. Although the lessons in this OER will frequently suggest another approach to managing Type I error, I will quickly review it now because, conceptually it is easy to understand. We start by establishing the α_{family} ; this is traditionally $p = .05$.

Next, we determine how many pairwise comparisons that we are going to conduct. If we are going to conduct all possible comparisons, we could use this formula: $N_{pc} = \frac{N_g(N-1)}{2}$, where

- N_{pc} is the number of pairwise comparisons, and
- N_g is the number of groups.

In the one-way ANOVA research vignette, the COND factor had three levels: control, low, high. Thus, if we wanted to conduct all possible comparisons we would determine N_{pc} this way:

```
3*(3-1)/2
```

```
## [1] 3
```

Subsequently, we would compute a new alpha that would be used for each comparison with this formula: $\alpha_{pc} = \frac{\alpha_{family}}{N_{pc}}$.

In the one-way ANOVA research vignette we would calculate it this way:

```
.05/3
```

```
## [1] 0.01666667
```

If we were to use the traditional Bonferroni to manage Type I error, the resultant p value would need to be $< .017$ in order for statistical significance to be claimed.

Luckily, each of these options has been reverse-engineered so that we do not have to determine the more conservative alpha levels. Instead, when we specify these options (and, as you will see, more) in the script, the p value is adjusted and we can continue to use the customary $p < .05$, $p < .01$ and $p < .001$ levels of interpretation. In the case of the traditional Bonferroni, the p value can be adjusted upward by multiplying it (i.e., the raw p values) by the number of comparisons being completed. This holds the *total* Type I error rate across these tests to be α (usually 0.05). Further, most *R* packages allow specification of one or more types of p values in the script. The result is the Type I error-adjusted p values.

Although the traditional Bonferroni is easy-to-understand and computer, it has been criticized as being too restrictive. That is, it increases the risk of making a Type II error (i.e., failing to reject the null hypothesis when it is false). This is why the majority of follow-up options to ANOVA did not use the traditional Bonferroni.

Tukey HSD

The Tukey HSD (honestly significant difference test) is a multiple comparison procedure used to identify significant differences between means of multiple groups. In the ANOVA context, it examines which specific pairs of groups differ from one another. The Tukey HSD was designed to control for Type I error. It does so by calculating the difference between the largest and smallest group means, then dividing this mean difference by the standard error of the same mean difference. The resulting statistic, q has an associated Studentized Range Distribution. Critical values for this distribution come from a Studentized Range q Table and are based on the alpha level, the number of groups, and the denominator degrees of freedom (i.e., df_W).

The Tukey HSD (“Tukey’s honestly significantly different”) test automatically controls for Type I error (i.e., false positives) by using the studentized range distribution to calculate a critical value. Subsequently, it compares the difference between pairs of means to this critical value. In the *rstatix* package, the *tukey_hsd()* function will perform the t-tests of all possible pairwise combinations. The Tukey HSD *p* value is automatically adjusted. In fact, there is nothing additional that can be specified about *p* values (i.e., there are no other choice options).

I had intended to demonstrate this with the one-way ANOVA chapter, but could not get the results to render a figure with the significance bars and results. An online search suggested that I am not the only one to have experienced this glitch.

Holms Sequential Bonferroni

The Holm’s sequential Bonferroni [[Green and Salkind, 2017c](#)] offers a middle-of-the-road approach (not as strict as .05/9 with the traditional Bonferroni; not as lenient as “none”) to managing Type I error.

If we were to hand-calculate the Holms, we would rank order the *p* values associated with the 9 comparisons in order from lowest (e.g., 0.000001448891) to highest (e.g., 1.000). The first *p* value is evaluated with the most strict criterion (.05/9; the traditional Bonferonni approach). Then, each successive comparison calculates the *p* value by using the number of *remaining* comparisons as the denominator (e.g., .05/8, .05/7, .05/6). As the *p* values increase and the alpha levels relax, there will be a cut-point where remaining comparisons are not statistically significant. Luckily, most R packages offer the Holm’s sequential Bonferroni as an option. The algorithm behind the output rearranges the mathematical formula and produces a *p* value that we can interpret according to the traditional values of $p < .05$, $p < .01$ and $p < .001$. [[Green and Salkind, 2017c](#)]

Chapter 13

Examples for Follow-up to Factorial ANOVA

As noted in the lesson on [factorial ANOVA](#), the options for follow-up to a significant interaction effect are infinite. In order to maintain a streamlined chapter with minimal distractions to student learning (through numerous examples and changes in R packages), I have moved examples of some these variations to this section.

As a quick reminder, I will describe and re-simulate the data. The narration will presume familiarity with the [factorial ANOVA](#) lesson.

Research Vignette

The research vignette for this example was located in Kalimantan, Indonesia and focused on bias in young people from three ethnic groups. The Madurese and Dayaknese groups were engaged in ethnic conflict that spanned 1996 to 2001. The last incidence of mass violence was in 2001 where approximately 500 people (mostly from the Madurese ethnic group) were expelled from the province. Ramdhani et al.'s [2018] research hypotheses were based on the roles of the three ethnic groups in the study. The Madurese appear to be viewed as the transgressors when they occupied lands and took employment and business opportunities from the Dayaknese. Ramdhani et al. also included a third group who were not involved in the conflict (Javanese). The research participants were students studying in Yogyakarta who were not involved in the conflict. They included 39 Madurese, 35 Dyaknese, and 37 Javanese; 83 were male and 28 were female.

In the study [Ramdhani et al., 2018], participants viewed facial pictures of three men and three women (in traditional dress) from each ethnic group (6 photos per ethnic group). Participant were asked, "How do you feel when you see this photo? Please indicate your answers based on your actual feelings." Participants responded on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Higher scores indicated ratings of higher intensity on that scale. The two scales included the following words:

- Positive: friendly, kind, helpful, happy
- Negative: disgusting, suspicious, hateful, angry

Quick Resimulating of the Data

Below is script to simulate data for the negative reactions variable from the information available from the manuscript [Ramdhani et al., 2018]. If you would like more information about the details of this simulation, please visit the lesson on [factorial ANOVA](#).

```
library(tidyverse)
set.seed(210731)
# sample size, M and SD for each cell; this will put it in a long
# file
Negative <- round(c(rnorm(17, mean = 1.91, sd = 0.73), rnorm(18, mean = 3.16,
  sd = 0.19), rnorm(19, mean = 3.3, sd = 1.05), rnorm(20, mean = 3, sd = 1.07),
  rnorm(18, mean = 2.64, sd = 0.95), rnorm(19, mean = 2.99, sd = 0.8)), 3)
# sample size, M and SD for each cell; this will put it in a long
# file
Positive <- round(c(rnorm(17, mean = 4.99, sd = 1.38), rnorm(18, mean = 3.83,
  sd = 1.13), rnorm(19, mean = 4.2, sd = 0.82), rnorm(20, mean = 4.19,
  sd = 0.91), rnorm(18, mean = 4.17, sd = 0.6), rnorm(19, mean = 3.26,
  sd = 0.94)), 3)
ID <- factor(seq(1, 111))
Rater <- c(rep("Dayaknese", 35), rep("Madurese", 39), rep("Javanese", 37))
Photo <- c(rep("Dayaknese", 17), rep("Madurese", 18), rep("Dayaknese",
  19), rep("Madurese", 20), rep("Dayaknese", 18), rep("Madurese", 19))
# groups the 3 variables into a single df: ID#, DV, condition
Ramdhani_df <- data.frame(ID, Negative, Positive, Rater, Photo)

Ramdhani_df[, "Rater"] <- as.factor(Ramdhani_df[, "Rater"])
Ramdhani_df[, "Photo"] <- as.factor(Ramdhani_df[, "Photo"])
```

If you want to export this data as a file to your computer, remove the hashtags to save it (and re-import it) as a .csv (“Excel lite”) or .rds (R object) file. This is not a necessary step.

The code for .csv will likely lose the formatting (i.e., making the Rater and Photo variables factors), but it is easy to view in Excel.

```
# write the simulated data as a .csv write.table(Ramdhani_df,
# file='RamdhaniCSV.csv', sep=',', col.names=TRUE, row.names=FALSE)
# bring back the simulated dat from a .csv file Ramdhani_df <-
# read.csv ('RamdhaniCSV.csv', header = TRUE) str(Ramdhani_df)
```

The code for the .rds file will retain the formatting of the variables, but is not easy to view outside of R.

```
# to save the df as an .rds (think 'R object') file on your computer;
# it should save in the same file as the .rmd file you are working
# with saveRDS(Ramdhani_df, 'Ramdhani_RDS.rds') bring back the
```

```
# simulated dat from an .rds file Ramdhani_df <-
# readRDS('Ramdhani_RDS.rds') str(Ramdhani_RDS)
```

Analysis of Simple Main Effects with Orthogonal Contrasts

This example follows a significant interaction effect. Specifically, we will analyze the effects of ethnicity of rater (three levels) within photo stimulus (two levels). We will conduct two one-way ANOVAs for the Dayaknese and Madurese photos, separately. In this example, we will utilize orthogonal contrast-coding for rater ethnicity.

In the lesson on **factorial ANOVA** I used the *rstatix* package. I am not aware of a way to do this type of analysis in *rstatix*, therefore this worked example will use functions from base R and _____.

This is our place on the ANOVA workflow.

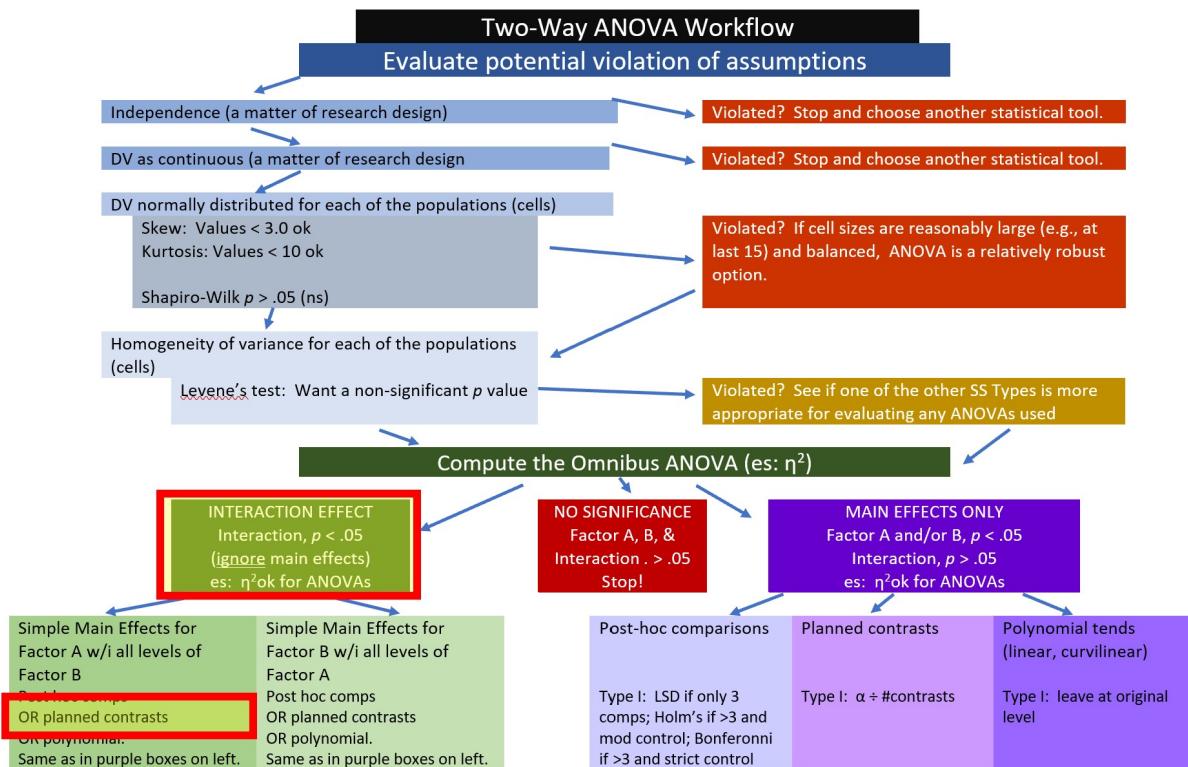


Figure 13.1: Image our place in the Two-Way ANOVA Workflow – analysis of simple main effects of factor A within levels of factor B with orthogonal contrasts

Among the requirements for orthogonal contrasts are these critical ones:

- there be one fewer contrast than the number of groups, (i.e., $k - 1$), and
- once a group is singled out, it cannot be compared again.

Thus, with a limit of two contrasts I want to compare the

- Javanese to the Dayaknese and Madurese combined (asking, “Do the Javanese evaluations of the photo differ from the combined Dyaknese/Madurese evaluations?”), then
- Dayaknese to Madurese (asking, “Do the Dayknese and Madurese evaluations of the photos differ from each other?”)

Such contrasts should be theoretically or rationally defensible. In the case of none, low, and high dose/intervention/exposure designs this is an easy requirement to meet. Typically, the no-dose is compared to the combined low and high dosage conditions. Then the low and high dosage conditions are compared. I would argue that because the Javanese were observers to the conflict, we can single them out in the first contrast, then compare the two groups who were directly involved in the conflict.

It helps to know what the default contrast codes are; we can get that information with the *contrasts()* function.

```
contrasts(Ramdhani_df$Rater)
```

	Javanese	Madurese
Dayaknese	0	0
Javanese	1	0
Madurese	0	1

Next, we set up the contrast conditions. In the code below,

- c1 indicates that the Javanese (noted as -2) are compared to the Dayaknese (1) and Madurese (1)
- c2 indicates that the Dayaknese (-1) and Madurese (1) are compared; Javanese (0) is removed from the contrast.

```
# tell R which groups to compare
c1 <- c(1, -2, 1)
c2 <- c(-1, 0, 1)
mat <- cbind(c1, c2) #combine the above bits
contrasts(Ramdhani_df$Rater) <- mat # attach the contrasts to the variable
```

This allows us to recheck the contrasts.

```
contrasts(Ramdhani_df$Rater)
```

	c1	c2
Dayaknese	1	-1
Javanese	-2	0
Madurese	1	1

With this output we can confirm that, in contrast 1 (the first column) we are comparing the Javanese to the combined Dayaknese and Madurese. In contrast 2 (the second column) we are comparing the Dayaknese to the Madurese.

We will conduct these contrasts with one group at a time. First, we must create a subset of all observations of the Dayaknese photo:

```
# subset data
Dayaknese_Ph <- subset(Ramdhani_df, Photo == "Dayaknese")
```

Next we use the *aov()* function from base R for the one-way ANOVA. Like magic, the contrast that we specified is assigned to the Rater variable. We can apply the *summary()* function to the *aov()* object that we created to see the results.

```
Dykn_simple <- aov(Negative ~ Rater, data = Dayaknese_Ph)
summary(Dykn_simple)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
Rater	2	19.81	9.903	13.32	0.0000221 ***						
Residuals	51	37.90	0.743								
<hr/>											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

We can apply the *etaSquared()* function from the *lsr* package to retrieve an η^2 .

```
# effect size for simple main effect can add 'type = 1,2,3,4' to
# correspond with the ANOVA that was run
lsr::etaSquared(Dykn_simple, anova = FALSE)
```

```
eta.sq eta.sq.part
Rater 0.3432006 0.3432006
```

We can capture the *F* string from this output: $F [2, 51] = 13.32, p < .001, \eta^2 = 0.343$.

This code produces the contrasts we specified. Note that in our code we can improve the interpretability of the output by adding labels. We know the specific contrasts from our prior work.

```
summary.aov(Dykn_simple, split = list(Rater = list(`Javanese v Dayaknese and Madurese` = 1,
`Dayaknese Madurese` = 2)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Rater	2	19.81	9.903	13.325	0.00002211
Rater: Javanese v Dayaknese and Madurese	1	0.07	0.071	0.095	0.759
Rater: Dayaknese Madurese	1	19.73	19.735	26.554	0.00000419
Residuals	51	37.90	0.743		

```
Rater                               ***
Rater: Javanese v Dayaknese and Madurese
Rater: Dayaknese Madurese          ***
Residuals
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

An APA style reporting of results-so-far might look like this

The simple main effect evaluating differences between rater ethnicity when evaluating photos of Dayaknese ethnic group was statistically significant: $F(2, 5) = 13.32, p < .001, \eta^2 = 0.343$. Follow-up testing indicated non-significant differences when the ratings from members of the Javanese ethnic group were compared to the Dayaknese and Madurese, combined ($F[1, 51] = 0.095, p = .759$). There was a statistically significant difference when Dayaknese and Madurese raters were compared ($F[1, 51] = 26.554, p < .001$).

We repeat the simple main effect process for evaluation of the Madurese photos.

```
# subset data
Madurese_Ph <- subset(Ramdhani_df, Photo == "Madurese")
# change df to subset, new model name
Mdrs_simple <- aov(Negative ~ Rater, data = Madurese_Ph)
# output for simple main effect
summary(Mdrs_simple)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Rater	2	1.04	0.5207	0.679	0.512
Residuals	54	41.44	0.7674		

```
# effect size for simple main effect can add 'type = 1,2,3,4' to
# correspond with the ANOVA that was run
lsr::etaSquared(Mdrs_simple, anova = FALSE)
```

```
eta.sq eta.sq.part
Rater 0.02451385 0.02451385
```

Let's capture the F string for ratings of the Madurese photos: $F(2, 54) = 0.679, p = .512, \eta^2 = 0.024$.

We can use the procedure described above to obtain our orthogonal contrasts.

```
summary.aov(Mdrs_simple, split = list(Rater = list(`Javanese v Dayaknese and Madurese` = 1,
`Dayaknese Madurese` = 2)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Rater	2	1.04	0.5207	0.679	0.512
Rater: Javanese v Dayaknese and Madurese	1	0.77	0.7734	1.008	0.320
Rater: Dayaknese Madurese	1	0.27	0.2679	0.349	0.557
Residuals	54	41.44	0.7674		

Here's a write-up of this portion of the result.

The simple main effect evaluating differences between rater ethnicity when evaluating photos of Madurese ethnic group was not statistically significant: $F(2, 54) = 0.679, p = .512, \eta^2 = 0.024$. Correspondingly, follow-up testing indicated non-significant differences when the ratings of the Javanese were compared to Dayaknese and Madurese, combined ($F[1, 54] = 1.008, p = .320$) and when the ratings of the Dayaknese and Madurese were compared ($F[1, 54] = 0.349, p = .557$)

In this series of analyses we did not have an opportunity to “let R manage Type I error for us.” Therefore, we will need to do it manually. We had 4 follow-up contrasts (2 for Dayaknese, 2 for Madurese). Using a traditional Bonferroni we could control Type I error with $.05/4 = .0125$

0.05/4

[1] 0.0125

APA Write-up of the simple main effect of photo stimulus within rater ethnicity.

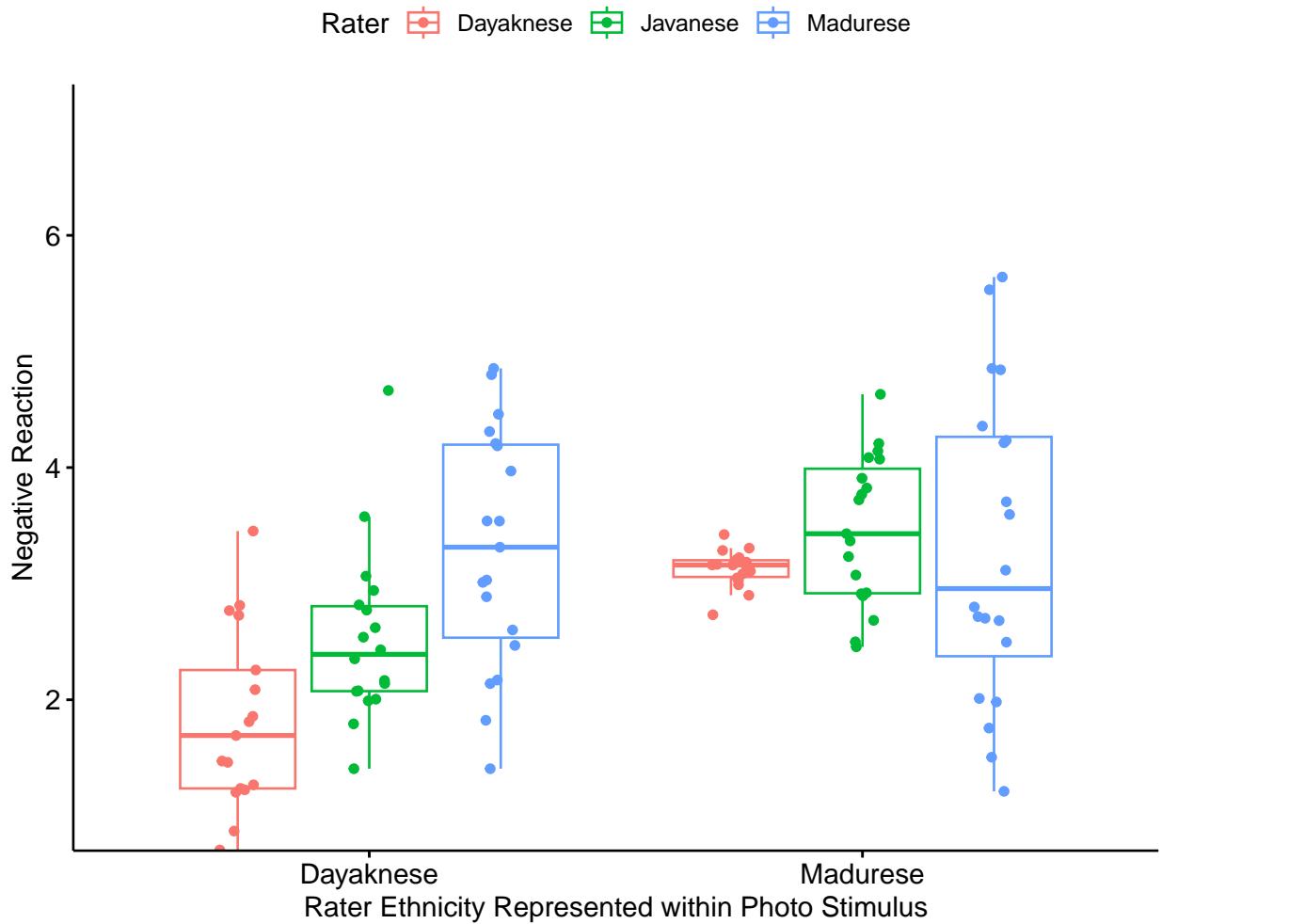
This would be added to the write-up of the omnibus two-way ANOVA test.

To explore the interaction effect, we followed with tests of simple effect of rater ethnicity within the photo stimulus. That is, we examined the effect of each each rater's ethnicity within the Madurese and Dayaknese photo stimulus, separately. Our first analysis evaluated the effect of the rater's ethnicity when evaluating the Dayaknese photo; our second analysis evaluated effect of the rater's ethnicity when evaluating the Madurese photo. To control for Type I error across the two simple main effects, we set alpha at .0125 (.05/4). The simple main effect evaluating differences between rater ethnicity when evaluating photos of Dayaknese ethnic group was statistically significant: $F(2, 51) = 13.32, p < .001, \eta^2 = 0.343$. Follow-up testing indicated non-significant differences when the ratings from members of the Javanese ethnic group were compared to the Dayaknese and Madurese, combined ($F[1, 51] = 0.095, p = .759$). There was a statistically significant difference when Dayaknese and Madurese raters were compared ($F[1, 51] = 26.554, p < .001$). The simple main effect evaluating differences between rater ethnicity when evaluating photos of Madurese ethnic group was not statistically significant: $F(2, 54) = 0.679, p = .512, \eta^2 = 0.024$. Correspondingly, follow-up testing indicated non-significant differences when the ratings of the Javanese were compared to Dayaknese and Madurese, combined ($F[1, 54] = 1.008, p = .320$) and when the ratings of the Dayaknese and Madurese were compared ($F[1, 54] = 0.349, p = .557$). This moderating effect of ethnicity of the rater on the negative reaction to the photo stimulus is illustrated in Figure 1.

I am not aware of an integration of packages that would represent this type of orthogonal contrast in a figure. Therefore, I would simply present the boxplots clustered by photo stimulus.

```
ggpubr::ggbboxplot(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Rater Ethnicity Represented within Photo Stimulus", ylab = "Negative Reaction",
  add = "jitter", title = "Figure 1. Simple Main Effect of Rater within Photo Stimulus",
  ylim = c(1, 7))
```

Figure 1. Simple Main Effect of Rater within Photo Stimulus



Analysis of Simple Main Effects with a Polynomial Trend

In the context of the significant interaction effect, we might be interested in polynomial trends for any simple main effects where three or more cells are compared.

Why? If there are only two cells being compared, then the significance of that has already been tested and if significant, it is also a significant linear effect (because the shape between any two points is a line).

Here is where we are in the workflow:

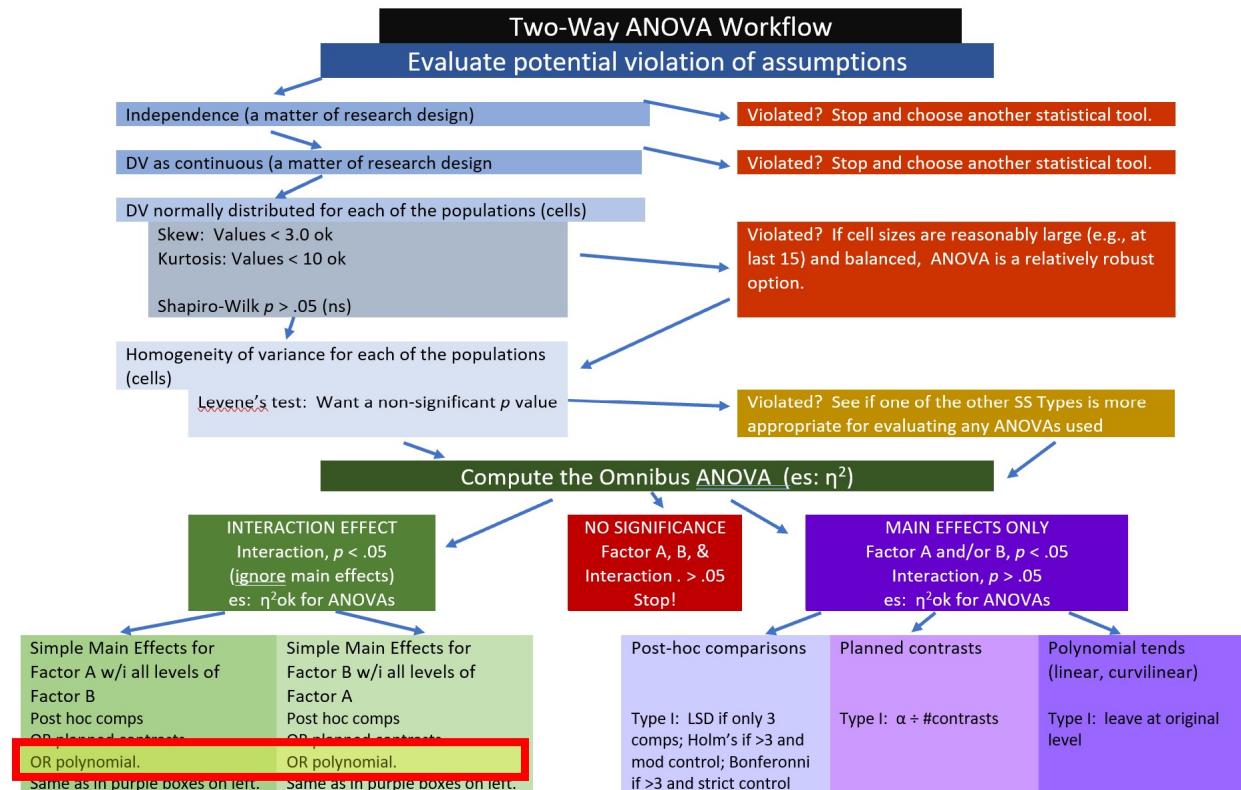


Figure 13.2: Image our place in the Two-Way ANOVA Workflow.

At the outset, let me acknowledge that this is not the best example to demonstrate a polynomial trend. Why? We do not necessarily have an ordered prediction across categories for this vignette. Other research scenarios (e.g., when dosage, intervention, or exposure is none, low, high) are more readily suited for this analytic strategy.

In our example, Rater has three groups. Thus, we could evaluate a polynomial for the simple main effect of ethnicity of the rater within photo stimulus. That is, we conduct polynomial analyses separately for the Dayaknese and Madurese photo stimuli.

If you haven't already, we need to subset the data, creating separate datasets for the evaluations of the Dayaknese photos and Madurese photos:

```
Dayaknese_Ph <- subset(Ramdhani_df, Photo == "Dayaknese")
Madurese_Ph <- subset(Ramdhani_df, Photo == "Madurese")
```

We will work the entire contrast for each of the datasets, separately.

First, we assign the polynomial contrast to the Rater variable. This is easily accomplished because the *contr.poly(#)* argument is built into base R. We simply indicate the number of levels in the variable. With Javanese, Dayaknese, and Madurese ethnic groups, we have three.

Second, we calculate the one-way ANOVA. Because we are using the *aov()* function in base R, we will need to extract the results. This time we need to use the *summary.lm()* function.

```
contrasts(Dayaknese_Ph$Rater) <- contr.poly(3)
poly_Dy <- aov(Negative ~ Rater, data = Dayaknese_Ph)
summary.lm(poly_Dy)
```

Call:

```
aov(formula = Negative ~ Rater, data = Dayaknese_Ph)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.8948	-0.5463	-0.1098	0.5155	2.1402

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	2.54746	0.11744	21.693	< 0.0000000000000002 ***							
Rater.L	1.04869	0.20351	5.153	0.00000419 ***							
Rater.Q	0.02901	0.20330	0.143	0.887							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

Residual standard error: 0.8621 on 51 degrees of freedom

Multiple R-squared: 0.3432, Adjusted R-squared: 0.3174

F-statistic: 13.32 on 2 and 51 DF, p-value: 0.00002211

We are interested in the regression output that end in the extensions “L” (for linear trend) and “.Q” (for quadratic trend). In the event that more than one polynomial trend is significant, select the higher one. For example, if both linear and quadratic are selected, interpret the quadratic trend

Results of polynomial trend analysis indicated a statistically significant linear trend for evaluation of the Dayaknese photos across the three raters $t(51) = 5.153, p < .001$.

Let's repeat the process for the Madurese photos.

```
contrasts(Madurese_Ph$Rater) <- contr.poly(3)
poly_Md <- aov(Negative ~ Rater, data = Madurese_Ph)
summary.lm(poly_Md)
```

Call:

```
aov(formula = Negative ~ Rater, data = Madurese_Ph)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.08650	-0.54395	0.01367	0.35905	2.34350

Coefficients:

Estimate	Std. Error	t value	Pr(> t)
----------	------------	---------	----------

```
(Intercept) 3.2973     0.1161   28.391 <0.0000000000000002 ***
Rater.L      0.1189     0.2012    0.591                   0.557
Rater.Q     -0.2054     0.2011   -1.021                   0.312
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.876 on 54 degrees of freedom
Multiple R-squared:  0.02451, Adjusted R-squared:  -0.01162
F-statistic: 0.6785 on 2 and 54 DF,  p-value: 0.5116
```

Results of a polynomial trend analyses were non-significant when ethnicity of the rater was evaluated when rating Madurese photos. Compared to the significant linear trend for the Dayaknese photos, results for ratings of the Madurese photos were non-significant ($t[54] = 0.591, p = 0.557$).

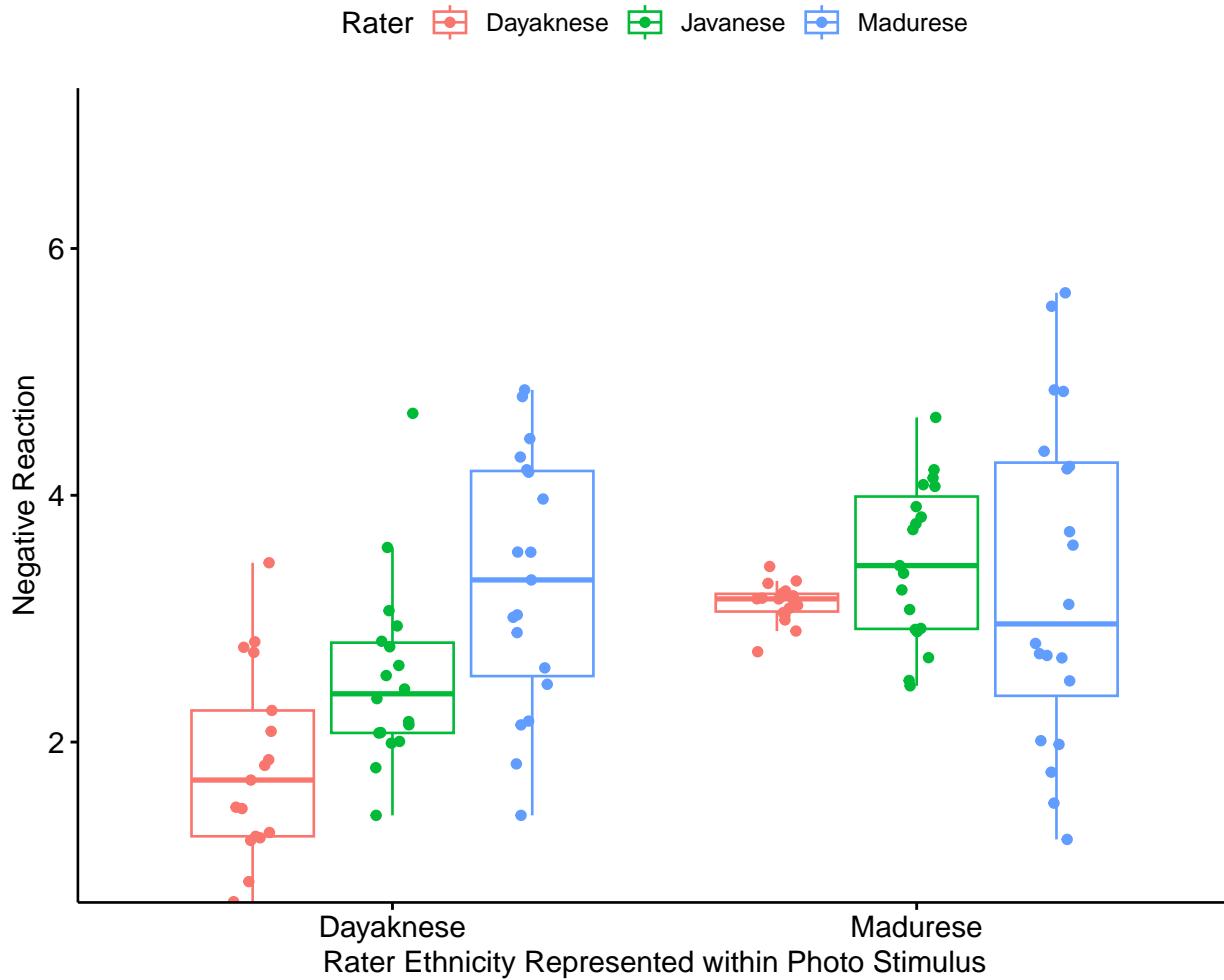
Here's how I might write up the results. In the case of polynomials, I will sometimes add them to an analysis that uses post hoc comparisons, particularly if the polynomial is helpful in conveying meaningful information about the result.

We followed up a significant interaction effect with a simple main effect of rater ethnicity within photo stimulus. Specifically, we were curious to see if there was a polynomial trend across rater ethnicity (ordered as Dayaknese, Javanese, and Madurese). Results indicated a statistically significant linear trend for evaluation of the Dayaknese photos $t(51) = 5.153, p < .001$, but not for the Madurese photos ($t[54] = 0.591, p = 0.557$).

The figure we have been using would be appropriate to illustrate the significant linear trend.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Rater Ethnicity Represented within Photo Stimulus", ylab = "Negative Reaction",
  add = "jitter", title = "Figure 1. Simple Main Effect of Rater within Photo Stimulus",
  ylim = c(1, 7))
```

Figure 1. Simple Main Effect of Rater within Photo Stimulus



All Possible Post Hoc Comparisons

Another option is the comparison possible cells. These are termed *post hoc comparisons*. They are an alternative to simple main effects; you would not report both. A potential criticism of this approach is that it is atheoretical. Without compelling justification, reviewers may criticize this approach as “fishing,” “p-hacking,” or “HARKing” (hypothesizing after results are known). None-the-less, particularly when our results are not as expected, I do think having these tools available can be a helpful resource.

The figure shows our place on the Two-Way ANOVA Workflow.

As the numbers of levels increase, post hoc comparisons become somewhat unwieldy. Even though this procedure produces them all, you can select which sensible number you want to compare and control for Type I error according to the number in that set.

With rater ethnicity (3 levels) and photo stimulus (2 levels), we have 6 groupings. When k is the number of groups, the total number of paired comparisons is: $k(k-1)/2$

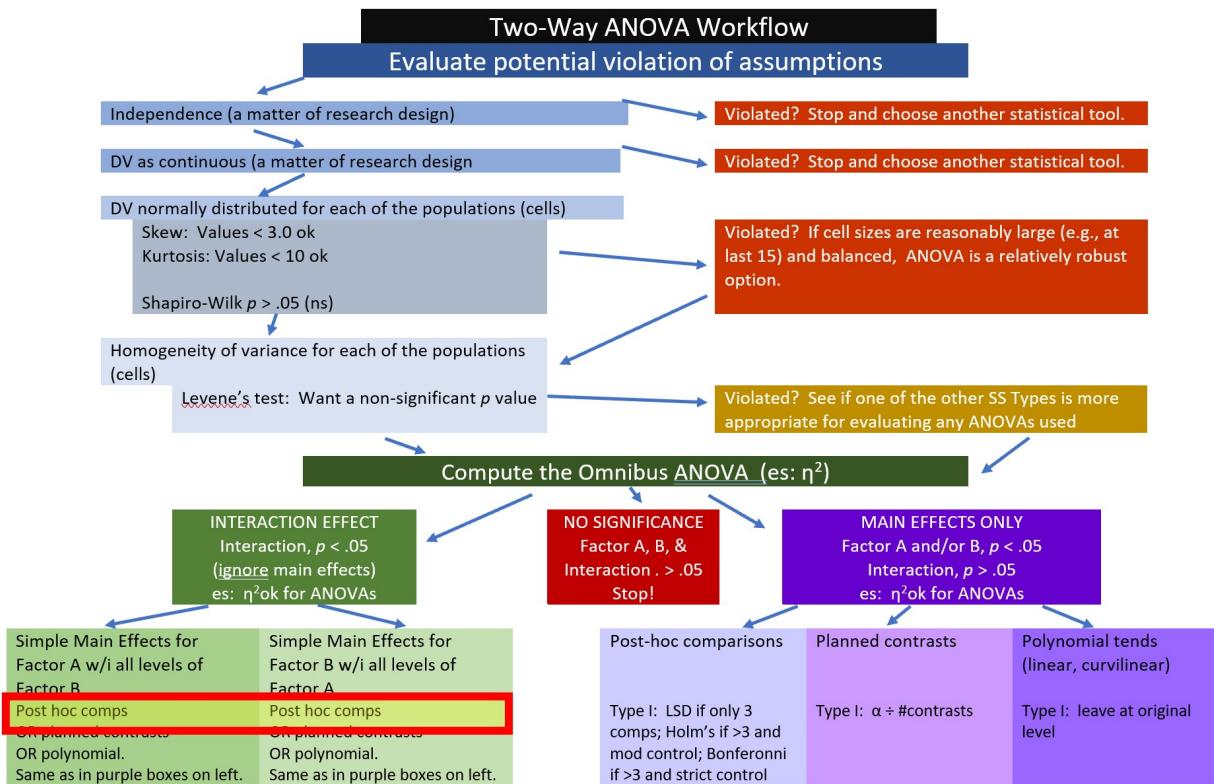


Figure 13.3: Image our place in the Two-Way ANOVA Workflow.

```
6 * (6 - 1)/2
```

```
[1] 15
```

Before running this analysis, we must calculate the omnibus ANOVA with the `aov()` function in base R and save the result as an object.

```
TwoWay_neg <- aov(Negative ~ Rater * Photo, Ramdhani_df)
summary(TwoWay_neg)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
Rater	2	12.21	6.103	8.077	0.000546 ***						
Photo	1	14.62	14.619	19.346	0.0000262 ***						
Rater:Photo	2	8.61	4.304	5.696	0.004480 **						
Residuals	105	79.34	0.756								

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

We can calculate the 15 post-hoc paired comparisons with the `TukeyHSD()` function from base R.

```
posthocs <- TukeyHSD(TwoWay_neg, ordered = TRUE)
posthocs
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
factor levels have been ordered

Fit: aov(formula = Negative ~ Rater * Photo, data = Ramdhani_df)

$Rater
      diff      lwr      upr      p adj
Javanese-Dayaknese 0.5147954  0.02750358 1.0020872 0.0358235
Madurese-Dayaknese 0.8068425  0.32566283 1.2880222 0.0003629
Madurese-Javanese  0.2920471 -0.18222911 0.7663234 0.3124227

$Photo
      diff      lwr      upr      p adj
Madurese-Dayaknese 0.726071  0.3987575 1.053385 0.0000262

$`Rater:Photo`
      diff      lwr      upr      p adj
Javanese:Dayaknese-Dayaknese:Dayaknese 0.706013072 -0.14743916 1.5594653
Dayaknese:Madurese-Dayaknese:Dayaknese 1.311568627  0.45811640 2.1650209
Madurese:Madurese-Dayaknese:Dayaknese  1.479735294  0.64726775 2.3122028
Madurese:Dayaknese-Dayaknese:Dayaknese 1.483077399  0.64060458 2.3255502
```

Javanese:Madurese-Dayaknese:Dayaknese	1.647182663	0.80470985	2.4896555
Dayaknese:Madurese-Javanese:Dayaknese	0.605555556	-0.23561614	1.4467273
Madurese:Madurese-Javanese:Dayaknese	0.773722222	-0.04615053	1.5935950
Madurese:Dayaknese-Javanese:Dayaknese	0.777064327	-0.05296553	1.6070942
Javanese:Madurese-Javanese:Dayaknese	0.941169591	0.11113973	1.7711995
Madurese:Madurese-Dayaknese:Madurese	0.168166667	-0.65170609	0.9880394
Madurese:Dayaknese-Dayaknese:Madurese	0.171508772	-0.65852109	1.0015386
Javanese:Madurese-Dayaknese:Madurese	0.335614035	-0.49441582	1.1656439
Madurese:Dayaknese-Madurese:Madurese	0.003342105	-0.80509532	0.8117795
Javanese:Madurese-Madurese:Madurese	0.167447368	-0.64099006	0.9758848
Javanese:Madurese-Madurese:Dayaknese	0.164105263	-0.65463115	0.9828417
	p adj		
Javanese:Dayaknese-Dayaknese:Dayaknese	0.1652148		
Dayaknese:Madurese-Dayaknese:Dayaknese	0.0002907		
Madurese:Madurese-Dayaknese:Dayaknese	0.0000171		
Madurese:Dayaknese-Dayaknese:Dayaknese	0.0000211		
Javanese:Madurese-Dayaknese:Dayaknese	0.0000018		
Dayaknese:Madurese-Javanese:Dayaknese	0.3005963		
Madurese:Madurese-Javanese:Dayaknese	0.0760131		
Madurese:Dayaknese-Javanese:Dayaknese	0.0802217		
Javanese:Madurese-Javanese:Dayaknese	0.0166363		
Madurese:Madurese-Dayaknese:Madurese	0.9911395		
Madurese:Dayaknese-Dayaknese:Madurese	0.9908344		
Javanese:Madurese-Dayaknese:Madurese	0.8482970		
Madurese:Dayaknese-Madurese:Madurese	1.0000000		
Javanese:Madurese-Madurese:Madurese	0.9907331		
Javanese:Madurese-Madurese:Dayaknese	0.9920328		

If we want to consider all 15 pairwise comparisons and also control for Type I error, a Holm's sequential Bonferroni [Green and Salkind, 2017c] will help us take a middle-of-the-road approach (not as strict as .05/15 with the traditional Bonferroni; not as lenient as "none") to managing Type I error.

With the Holms, we rank order the p values associated with the 15 comparisons in order from lowest (e.g., .0000018) to highest (e.g., 1.000). The first p value is evaluated with the most strict criterion (.05/15; the traditional Bonferonni approach). Then, each successive comparison calculates the p value by using the number of *remaining* comparisons as the denominator (e.g., .05/14, .05/13, .05/12). As the p values rise and the alpha levels relax, there will be a cut-point where remaining comparisons are not statistically significant.

0.05/15

[1] 0.003333333

0.05/14

[1] 0.003571429

To facilitate this contrast, let's extract the 15 TukeyHSD tests and work with them in Excel.

First, obtain the structure of the *posthoc* object

```
str(posthocs)
```

```
List of 3
$ Rater      : num [1:3, 1:4] 0.5148 0.8068 0.292 0.0275 0.3257 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:3] "Javanese-Dayaknese" "Madurese-Dayaknese" "Madurese-Javanese"
.. ..$ : chr [1:4] "diff" "lwr" "upr" "p adj"
$ Photo       : num [1, 1:4] 0.726071 0.3987575 1.0533845 0.0000262
..- attr(*, "dimnames")=List of 2
.. ..$ : chr "Madurese-Dayaknese"
.. ..$ : chr [1:4] "diff" "lwr" "upr" "p adj"
$ Rater:Photo: num [1:15, 1:4] 0.706 1.312 1.48 1.483 1.647 ...
..- attr(*, "dimnames")=List of 2
.. ..$ : chr [1:15] "Javanese:Dayaknese-Dayaknese:Dayaknese" "Dayaknese:Madurese-Dayaknese:Dayaknese"
.. ..$ : chr [1:4] "diff" "lwr" "upr" "p adj"
- attr(*, "class")= chr [1:2] "TukeyHSD" "multicomp"
- attr(*, "orig.call")= language aov(formula = Negative ~ Rater * Photo, data = Ramdhani_df)
- attr(*, "conf.level")= num 0.95
- attr(*, "ordered")= logi TRUE
```

```
write.csv(posthocs$"Rater:Photo", "posthocsOUT.csv")
```

In Excel, I would sort my results by their *p* values (low to high) and consider my threshold ($p < .0033$) to determine which effects were statistically significant. Using the strictest criteria of $p < .0033$, we would have four statistically significant values.

I would ask, “Is this what we want?” Similar to the simple main effects we just tested, I am interested in two sets of comparisons:

First, how are the two sets of photos (Madurese and Dayaknese) rated within each set of raters.

- Javanese:Madurese - Javanese:Dayaknese
- Dayaknese:Madurese - Dayaknese:Dayaknese
- Madurese:Madurese - Madurese:Dayaknese

Second, focused on each photo, what are the relative ratings.

- Javanese:Madurese - Dayaknese:Madurese
- Madurese: Madurese - Dayaknese:Madurese
- Javanese:Dayaknese - Dayaknese:Dayaknese
- Madurese: Dayaknese - Dayaknese:Dayaknese

This is only seven sets of comparisons and would considerably reduce the alpha:

	diff	lwr	upr	p adj
Javanese:Madurese-Dayaknese:Dayaknese	1.647182663	0.80470985	2.489655478	0.00000182
Madurese:Madurese-Dayaknese:Dayaknese	1.479735294	0.64726775	2.312202839	0.00001714
Madurese:Dayaknese-Dayaknese:Dayaknese	1.483077399	0.64060458	2.325550215	0.00002114
Dayaknese:Madurese-Dayaknese:Dayaknese	1.311568627	0.4581164	2.165020855	0.00029074
Javanese:Madurese-Javanese:Dayaknese	0.941169591	0.11113973	1.77119945	0.01663633
Madurese:Madurese-Javanese:Dayaknese	0.773722222	-0.0461505	1.593594978	0.07601309
Madurese:Dayaknese-Javanese:Dayaknese	0.777064327	-0.0529655	1.607094187	0.08022174
Javanese:Dayaknese-Dayaknese:Dayaknese	0.706013072	-0.1474392	1.559465299	0.16521479
Dayaknese:Madurese-Javanese:Dayaknese	0.605555556	-0.2356161	1.446727254	0.30059630
Javanese:Madurese-Dayaknese:Madurese	0.335614035	-0.4944158	1.165643895	0.84829701
Javanese:Madurese-Madurese:Madurese	0.167447368	-0.6409901	0.975884798	0.99073306
Madurese:Dayaknese-Dayaknese:Madurese	0.171508772	-0.6585211	1.001538632	0.99083435
Madurese:Madurese-Dayaknese:Madurese	0.168166667	-0.6517061	0.988039422	0.99113951
Javanese:Madurese-Madurese:Dayaknese	0.164105263	-0.6546311	0.982841674	0.99203282
Madurese:Dayaknese-Madurese:Madurese	0.003342105	-0.8050953	0.811779534	1.000000000

Figure 13.4: Image of the results of the Holms sequential Bonferroni.

0.05/7

[1] 0.007142857

Below I have greyed-out the comparisons that are less interesting to me and left the seven that are my focal interest. I have highlighted in green the two comparisons that are statistically significant based on the Holms' sequential criteria. In this case, it does not make any difference in our interpretation of these focal predictors.

	diff	lwr	upr	p adj
Javanese:Madurese-Dayaknese:Dayaknese	1.647182663	0.80470985	2.489655478	0.00000182
Madurese:Madurese-Dayaknese:Dayaknese	1.479735294	0.64726775	2.312202839	0.00001714
Madurese:Dayaknese-Dayaknese:Dayaknese	1.483077399	0.64060458	2.325550215	0.00002114
Dayaknese:Madurese-Dayaknese:Dayaknese	1.311568627	0.4581164	2.165020855	0.00029074
Javanese:Madurese-Javanese:Dayaknese	0.941169591	0.11113973	1.77119945	0.01663633
Madurese:Madurese-Javanese:Dayaknese	0.773722222	-0.0461505	1.593594978	0.07601309
Madurese:Dayaknese-Javanese:Dayaknese	0.777064327	-0.0529655	1.607094187	0.08022174
Javanese:Dayaknese-Dayaknese:Dayaknese	0.706013072	-0.1474392	1.559465299	0.16521479
Dayaknese:Madurese-Javanese:Dayaknese	0.605555556	-0.2356161	1.446727254	0.30059630
Javanese:Madurese-Dayaknese:Madurese	0.335614035	-0.4944158	1.165643895	0.84829701
Javanese:Madurese-Madurese:Madurese	0.167447368	-0.6409901	0.975884798	0.99073306
Madurese:Dayaknese-Dayaknese:Madurese	0.171508772	-0.6585211	1.001538632	0.99083435
Madurese:Madurese-Dayaknese:Madurese	0.168166667	-0.6517061	0.988039422	0.99113951
Javanese:Madurese-Madurese:Dayaknese	0.164105263	-0.6546311	0.982841674	0.99203282
Madurese:Dayaknese-Madurese:Madurese	0.003342105	-0.8050953	0.811779534	1.000000000

Given that my “tinkering around” analysis resembles the results of the simple main effects analyses

in the [factorial lessonn](#), I will not write this up as an APA style results section, but rather offer this is as a set of tools when you would like to explore the data in an atheoretical manner.

Chapter 14

One-Way Repeated Measures with a Multivariate Approach

As noted in the lesson on [one-way repeated measures ANOVA](#), the researcher can use a univariate or multivariate approach to analyzing the data. The `rstatix::anova_test()` is limited to the univariate approach. In order to maintain a streamlined chapter with minimal distractions to student learning I have chosen to provide a quick and separate demonstration of the multivariate approach in this appendix. In-so-doing, I will use the `car` package.

As a quick reminder, I will describe and resimulate the data. The narration will presume familiarity with the [one-way repeated measures ANOVA](#) lesson.

Research Vignette

Amodeo [Amodeo et al., 2018] and colleagues conducted a mixed methods study (qualitative and quantitative) to evaluate the effectiveness of an empowerment, peer-group-based, intervention with participants ($N = 8$) who experienced transphobic episodes. Focus groups used qualitative methods to summarize emergent themes from the program (identity affirmation, self-acceptance, group as support) and a one-way, repeated measures ANOVA provided evidence of increased resilience from pre to three-month followup.

Eight participants (seven transgender women and one genderqueer person) participated in the intervention. The mean age was 28.5 ($SD = 5.85$). All participants were located in Italy.

The within-subjects condition was wave, represented by T1, T2, and T3:

- T1, beginning of training
- Training, three 8-hour days,
 - content included identity and heterosexism, sociopolitical issues and minority stress, resilience, and empowerment
- T2, at the conclusion of the 3-day training
- Follow-up session 3 months later
- T3, at the conclusion of the +3 month follow-up session

The dependent variable (assessed at each wave) was a 14-item resilience scale [Wagnild and Young, 1993]. Items were assessed on a 7-point scale ranging from *strongly disagree* to *strongly agree* with higher scores indicating higher levels of resilience. An example items was, “I usually manage one way or another.”

Data Simulation

Below is the code I used to simulate data. The following code assumes 8 participants who each participated in 3 waves (pre, post, followup). The script produces “long” and “wide” forms are created.

```
set.seed(2022)
# gives me 8 numbers, assigning each number 3 consecutive spots, in
# sequence
ID <- factor(c(rep(seq(1, 8), each = 3)))
# gives me a column of 24 numbers with the specified Ms and SD
Resilience <- rnorm(24, mean = c(5.7, 6.21, 6.26), sd = c(0.88, 0.79, 0.37))
# repeats pre, post, follow-up once each, 8 times
Wave <- rep(c("Pre", "Post", "FollowUp"), each = 1, 8)
Amodeo_long <- data.frame(ID, Wave, Resilience)

Amodeo_long$Wave <- factor(Amodeo_long$Wave, levels = c("Pre", "Post",
"FollowUp"))

# Create a new df (Amodeo_wide) Identify the original df We are
# telling it to connect the values of the Resilience variable its
# respective Wave designation
Amodeo_wide <- reshape2::dcast(data = Amodeo_long, formula = ID ~ Wave,
    value.var = "Resilience")
# doublecheck to see if they did what you think
str(Amodeo_wide)

'data.frame':   8 obs. of  4 variables:
 $ ID      : Factor w/ 8 levels "1","2","3","4",...: 1 2 3 4 5 6 7 8
 $ Pre     : num  6.49 4.43 4.77 5.91 4.84 ...
 $ Post    : num  5.28 5.95 6.43 7 6.28 ...
 $ FollowUp: num  5.93 5.19 6.54 6.19 6.24 ...
```

```
Amodeo_wide$ID <- factor(Amodeo_wide$ID)
```

Computing the Omnibus F

Without the *rstatix* helper package, here is how the analysis would be run in the package, *car*. Although this package is less intuitive to use, it results in both univariate output (both sphericity

assumed and sphericity violated) and multivariate output (which does not require the sphericity assumption).

Evaluating the data requires that we create some objects that will be fed into function. We can name these objects anything we like.

In this script below I define the objects that are required.

- waveLevels is an object that will specify three levels of the independent variable (pre, post, follow-up),
- waveFactor simply makes “waveLevels” a factor
- waveBind column-binds (i.e., cbind) the pre, post, and follow-up variables from the wide form of the Amodeo dataset
- waveModel calculates the intercept (i.e., the means) of the pre, post, and follow-up levels

```
# library(car)
waveLevels <- c(1, 2, 3)
waveFactor <- as.factor(waveLevels)
waveFrame <- data.frame(waveFactor)
waveBind <- cbind(Amdeo_wide$Pre, Amdeo_wide$Post, Amdeo_wide$FollowUp)
waveModel <- lm(waveBind ~ 1)
waveModel
```

Call:

```
lm(formula = waveBind ~ 1)
```

Coefficients:

	[,1]	[,2]	[,3]
(Intercept)	5.588	6.328	6.137

To run the analysis, we insert these objects into arguments:

- waveModel is the first argument,
- waveFrame is assigned to the *idata* command,
- waveFactor is assigned to the *idata* command

```
analysis <- car:::Anova(waveModel, idata = waveFrame, idesign = ~waveFactor)
```

Note: model has only an intercept; equivalent type-III tests substituted.

```
summary(analysis)
```

Type III Repeated Measures MANOVA Tests:

Term: (Intercept)

Response transformation matrix:

(Intercept)

```
[1,]      1
[2,]      1
[3,]      1
```

Sum of squares and products for the hypothesis:

(Intercept)

```
(Intercept) 2607.062
```

Multivariate Tests: (Intercept)

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.9942	1200.028	1	7	0.0000000043326 ***
Wilks	1	0.0058	1200.028	1	7	0.0000000043326 ***
Hotelling-Lawley	1	171.4325	1200.028	1	7	0.0000000043326 ***
Roy	1	171.4325	1200.028	1	7	0.0000000043326 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Term: waveFactor

Response transformation matrix:

waveFactor1 waveFactor2

```
[1,]      1      0
[2,]      0      1
[3,]     -1     -1
```

Sum of squares and products for the hypothesis:

waveFactor1 waveFactor2

```
waveFactor1  2.4131705 -0.8378898
waveFactor2 -0.8378898  0.2909282
```

Multivariate Tests: waveFactor

	Df	test stat	approx F	num Df	den Df	Pr(>F)
Pillai	1	0.4026101	2.021846	2	6	0.21319
Wilks	1	0.5973899	2.021846	2	6	0.21319
Hotelling-Lawley	1	0.6739486	2.021846	2	6	0.21319
Roy	1	0.6739486	2.021846	2	6	0.21319

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

Sum Sq	num Df	Error SS	den Df	F value	Pr(>F)
--------	--------	----------	--------	---------	--------

```
(Intercept) 869.02      1   5.0692      7 1200.0279 0.000000004333 ***
waveFactor    2.36      2   4.2272     14   3.9102      0.04476 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mauchly Tests for Sphericity

Test	statistic	p-value
waveFactor	0.56648	0.18179

Greenhouse-Geisser and Huynh-Feldt Corrections for Departure from Sphericity

GG	eps	Pr(>F[GG])
waveFactor	0.69759	0.06754 .

HF	eps	Pr(>F[HF])
waveFactor	0.8172743	0.05734876

The `car::Anova()` function produces both univariate and multivariate results. To begin to understand this data, let's start with what we learned in the [one-way repeated measures ANOVA lesson](#).

Univariate Results

When we ran the univariate approach in the lesson, we first checked the sphericity assumption. Our results here are identical to those from `rstatix::anova_test`. That is, we did not violate the sphericity assumption: Mauchley's test = $.566p = 0.182$. Although I do not see the complete F string, we learn that, if we had violated the sphericity assumption, that:

The Greenhouse Geiser estimate was 0.698 the corrected $p = .068$. The Huyhn Feldt estimate was 0.817 and the corrected $p = .057$.

The univariate ANOVA results are under the “Univariate Type III REpeated-Measures ANOVA Assuming Sphericity” heading. We find the ANOVA output on the row titled, “waveFactor.” The results are identical to what we found in teh lesson: $F(2, 14) = 3.91, p = 0.045$. I do not see that an effect size is reported.

Multivariate Results

Researchers may prefer the multivariate approach because it does not require the sphericity assumption. Stated another way, if the sphericity assumption is violated, researchers can report the results of the multivariate analysis.

We find the multivariate results in the middle of the output, under the heading, “Multivariate Tests: waveFactor.” There are four choices: Pillai, Wilks, Hotelling-Lawley, and Roy. Green and Salkind [2017b] have noted that in the one-way within-subjects ANOVA, all four will yield the same F and p values. They recommended reporting Wilks’ lambda because researchers will have greatest familiarity with it. Thus, I would write up the result of this omnibus test like this:

Results of the one-way repeated measures ANOVA indicated a significant wave effect, $\text{Wilks}'\lambda = .597, F(2, 6) = 2.022, p = 0.213$.

Because follow-up testing is *pairwise* (i.e., there are only two levels being compared), the sphericity assumption is not required and those could proceed in the manner demonstrated in the [one-way repeated measures ANOVA lesson](#).

14.0.1 A Brief Commentary on Wrappers

As noted several times, because of its relative ease-of-use, the relevance of information included in the results, and its integration with the *ggpubr* package, I chose to use *rstatix* package in all of the ANOVA lessons. As I worked through this example, I spent several hours creating and interpreting the code. For me, there was value in this exercise:

- I am encouraged and reassured with the consistency of results between the two approaches,
- I am in awe of the power of these programs and a little intimidated by all the options that are available within a given package, and
- I am deeply grateful to package developers who take the time to create packages for discipline-specific use-cases and then freely share their work with others. Thank you [Alboukadel Kas-sambara!](#)

Bibliography

- Anna Lisa Amodeo, Simona Picariello, Paolo Valerio, and Cristiano Scandurra. Empowering trans-gender youths: Promoting resilience through a group training program. *Journal of Gay & Lesbian Mental Health*, 22(1):3–19, 2018. URL <https://alliance-primo.hosted.exlibrisgroup.com>.
- Jeremy B. Bernerth and Herman Aguinis. A critical review and best-practice recommendations for control variable usage. *Personnel Psychology*, 69(1):229–283, 2016. ISSN 0031-5826. doi: 10.1111/peps.12103. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2015-25446-001&site=ehost-live>. Publisher: Wiley-Blackwell Publishing Ltd.
- Rachel Butler, Mauricio Monsalve, Geb W. Thomas, Ted Herman, Alberto M. Segre, Philip M. Polgreen, and Manish Suneja. Estimating Time Physicians and Other Health Care Workers Spend with Patients in an Intensive Care Unit Using a Sensor Network. *The American Journal of Medicine*, 131(8):972.e9–972.e15, August 2018. ISSN 00029343. doi: 10.1016/j.amjmed.2018.03.015. URL <https://linkinghub.elsevier.com/retrieve/pii/S0002934318302961>.
- Barbara M. Byrne. Structural Equation Modeling: The basics (Chapter 1). In *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming, Third Edition*. Taylor & Francis Group, London, UNITED KINGDOM, 2016. ISBN 978-1-317-63313-6. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4556523>.
- Tian Chen, Manfei Xu, Justin Tu, Hongyue Wang, and Xiaohui Niu. Relationship between Omnibus and Post-hoc Tests: An Investigation of performance of the F test in ANOVA. *Shanghai Archives of Psychiatry*, 30(1):60–64, 2018. ISSN 1002-0829. doi: 10.11919/j.issn.1002-0829.218014. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5925602/>.
- Jacob Cohen, P. Cohen, Stephen G. West, and Leona S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Erlbaum Associates, Mahwah, N.J., 3rd ed. edition, 2003. ISBN 978-0-8058-2223-6.
- Matthew J. C. Crump. Simulating and analyzing data in R (Chapter 5). In *Programming for Psychologists: Data Creation and Analysis*. 2018. URL <https://crumplab.github.io/programmingforpsych/index.html>.
- Datanovia. ANCOVA in R: The Ultimate Practical Guide. URL <https://www.datanovia.com/en/lessons/ancova-in-r/>.
- Andrea M. Elliott, Stewart C. Alexander, Craig A. Mescher, Deepika Mohan, and Amber E. Barnato. Differences in Physicians’ Verbal and Nonverbal Communication With Black and

- White Patients at the End of Life. *Journal of Pain and Symptom Management*, 51(1):1–8, January 2016. ISSN 0885-3924. doi: 10.1016/j.jpainsymman.2015.07.008. URL <https://www.sciencedirect.com/science/article/pii/S0885392415004029>.
- Andy P. Field. *Discovering statistics using R*. Sage, Thousand Oaks, California, 2012. ISBN 978-1-4462-0046-9.
- Malcolm Gladwell. *Outliers: the story of success*. New York Times best sellers. Little, Brown and Company, New York, first edition. edition, 2008. ISBN 978-0-316-01792-3.
- Samuel B. Green and Neil J. Salkind. One-Way Analysis of Covariance (Lesson 27). In *Using SPSS for Windows and Macintosh: analyzing and understanding data*, pages 151–160. Pearson, Boston, eighth edition. edition, 2017a. ISBN 978-0-13-431988-9.
- Samuel B. Green and Neil J. Salkind. One-Way Repeated Measures Analysis of Variance (Lesson 29). In *Using SPSS for Windows and Macintosh: analyzing and understanding data*, pages 209–217. Pearson, Boston, eighth edition. edition, 2017b. ISBN 978-0-13-431988-9.
- Samuel B. Green and Neil J. Salkind. *Using SPSS for Windows and Macintosh: analyzing and understanding data*. Pearson, Boston, eighth edition. edition, 2017c. ISBN 978-0-13-431988-9.
- Rajiv S. Jhangiani, I.-Chant A. Chiang, Carrie Cuttler, and Dana C. Leighton. *Research Methods in Psychology*. August 2019. ISBN 978-1-9991981-0-7. doi: 10.17605/OSF.IO/HF7DQ. URL <https://kpu.pressbooks.pub/psychmethods4e/>.
- D. N. Joanes and C. A. Gill. Comparing Measures of Sample Skewness and Kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):183–189, 1998. ISSN 0039-0526. URL <https://www.jstor.org/stable/2988433>. Publisher: [Royal Statistical Society, Wiley].
- Robert I. Kabacoff. Power Analysis, 2017. URL <https://www.statmethods.net/stats/power.html>.
- Alboukadel Kassambara. ANOVA in R: The Ultimate Guide, a. URL <https://www.datanovia.com/en/lessons/anova-in-r/>.
- Alboukadel Kassambara. Pipe-Friendly Framework for Basic Statistical Tests, b. URL <https://rpkgs.datanovia.com/rstatix/>.
- Rex B. Kline. Data Preparation and Psychometrics Review (Chapter 4). In *Principles and practice of structural equation modeling*, pages 64–96. Guilford Publications, New York, UNITED STATES, 4th edition, 2016a. ISBN 978-1-4625-2336-8. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>.
- Rex B. Kline. *Principles and practice of structural equation modeling*. Guilford Publications, New York, UNITED STATES, 4th edition, 2016b. ISBN 978-1-4625-2336-8. URL <http://ebookcentral.proquest.com/lib/spu/detail.action?docID=4000663>.
- Daniel Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00863. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863/full>. Publisher: Frontiers.
- P Priscilla Lui. Racial Microaggression, Overt Discrimination, and Distress: (In)Direct Associations With Psychological Adjustment. *The Counseling Psychologist*, page 32, 2020.

- Brent Mallinckrodt, Joseph R. Miles, and Jacob J. Levy. The scientist-practitioner-advocate model: Addressing contemporary training needs for social justice advocacy. *Training and Education in Professional Psychology*, 8(4):303–311, November 2014. ISSN 1931-3918. doi: 10.1037/tep0000045. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2014-25072-001&site=ehost-live>. Publisher: Educational Publishing Foundation.
- Sohad Murrar and Markus Brauer. Entertainment-education effectively reduces prejudice. *Group Processes & Intergroup Relations*, 21(7):1053–1077, October 2018. ISSN 1368-4302, 1461-7188. doi: 10.1177/1368430216682350. URL <http://journals.sagepub.com/doi/10.1177/1368430216682350>.
- Danielle Navarro. Book: *Learning Statistics with R - A tutorial for Psychology Students and other Beginners*. Open Education Resource (OER) LibreTexts Project, July 2020a. URL [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)).
- Danielle Navarro. Chapter 14: Comparing Several Means (One-Way ANOVA). In Book: *Learning Statistics with R - A tutorial for Psychology Students and other Beginners*. Open Education Resource (OER) LibreTexts Project, July 2020b. URL [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_\(Navarro\)](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)).
- Neila Ramdhani, Haidar Buldan Thontowi, and Djamarudin Ancok. Affective Reactions Among Students Belonging to Ethnic Groups Engaged in Prior Conflict. *Journal of Pacific Rim Psychology*, 12:e2, January 2018. ISSN 1834-4909, 1834-4909. doi: 10.1017/prp.2017.22. URL <http://journals.sagepub.com/doi/10.1017/prp.2017.22>.
- William Revelle. An introduction to the psych package: Part I: data entry and data description. page 60, 2021. URL <https://rdrr.io/cran/psych/f/inst/doc/intro.pdf>.
- Joseph Lee Rodgers. The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, 65(1):1–12, January 2010. ISSN 0003-066X. doi: 10.1037/a0018326. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2009-24989-001&site=ehost-live>.
- David J. Stanley and Jeffrey R. Spence. Reproducible Tables in Psychology Using the apaTables Package. *Advances in Methods and Practices in Psychological Science*, 1(3):415–431, September 2018. ISSN 2515-2459. doi: 10.1177/2515245918773743. URL <https://doi.org/10.1177/2515245918773743>. Publisher: SAGE Publications Inc.
- Alisia G. T. T. Tran and Richard M. Lee. You speak English well! Asian Americans' reactions to an exceptionalizing stereotype. *Journal of Counseling Psychology*, 61(3):484–490, July 2014. ISSN 0022-0167. doi: 10.1037/cou0000034. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=2014-28261-016&site=ehost-live>.
- Gail M. Wagnild and Heather M. Young. Development and psychometric evaluation of the Resilience Scale. *Journal of Nursing Measurement*, 1(2):165–178, 1993. ISSN 1061-3749. URL <https://ezproxy.spu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip&db=psyh&AN=1996-05738-006&site=ehost-live>. Publisher: Springer Publishing.

Peter Watson. Rules of thumb on magnitudes of effect sizes, 2020. URL <https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize>.

Zach. How to Read the F-Distribution Table, May 2019. URL <https://www.statology.org/how-to-read-the-f-distribution-table/>.