

ReCentering Psych Stats

Lynette H. Bikos, PhD, ABPP

10 Oct 2022

Contents

BOOK COVER	16
PREFACE	19
Copyright with Open Access	20
ACKNOWLEDGEMENTS	21
1 Introduction	23
1.1 What to expect in each chapter	23
1.2 Strategies for Accessing and Using this OER	24
1.3 If You are New to R	24
2 Ready_Set_R	27
2.1 Navigating this Lesson	27
2.1.1 Learning Objectives	27
2.2 downloading and installing R	27
2.2.1 So many paRts and pieces	27
2.2.2 oRienting to R Studio (focusing only on the things we will be using first and most often)	28
2.3 best pRactices	29
2.3.1 Everything is documented in the .rmd file	29
2.3.2 Setting up the file	29
2.3.3 Script in chunks and everything else in the “inline text” sections	30
2.3.4 Managing packages	30
2.3.5 Upload the data	31
2.3.5.1 To and from .csv files	32
2.3.5.2 To and from .rds files	32
2.3.5.3 From SPSS files	33

2.4 quick demonstRation	33
2.5 the knitted file	33
2.6 tRoubleshooting in R maRkdown	34
2.7 just <i>why</i> have we tRansitioned to R?	34
2.8 stRategies for success	35
2.9 Resources for getting staRted	35
2.10 Practice Problems	36
Preliminary Analyses	37
3 Preliminary Results	39
3.1 Navigating this Lesson	39
3.1.1 Learning Objectives	39
3.1.2 Planning for Practice	40
3.1.3 Readings & Resources	40
3.2 Research Vignette	40
3.3 Variable Types (Scale of Measurement)	44
3.3.1 Measurement Scale	44
3.3.2 Corresponding Variable Structure in R	45
3.4 Descriptive Statistics	46
3.4.1 Measures of Central Tendency	48
3.4.1.1 Mean	48
3.4.1.2 Median	49
3.4.1.3 Mode	49
3.4.1.4 Relationship between mean, median, and mode	50
3.4.2 Range	51
3.4.3 Percentiles, Quantiles, Interquartile Range	52
3.4.4 Deviations around the Mean	54
3.4.5 Variance	57
3.4.6 Standard Deviation	59
3.5 Are the Variables Normally Distributed?	61
3.6 Relations between Variables	64
3.7 Shortcuts to Preliminary Analyses	67
3.7.1 SPLOM	68

CONTENTS	5
3.7.2 apaTables	70
3.8 An APA Style Writeup	71
3.9 Practice Problems	71
3.9.1 Problem #1: Change the Random Seed	72
3.9.2 Problem #2: Swap Variables in the Simulation	72
3.9.3 Problem #3: Use (or Simulate) Your Own Data	72
3.9.4 Grading Rubric	72
<i>t</i>-tests	73
4 One Sample <i>t</i>-tests	75
4.1 Navigating this Lesson	75
4.1.1 Learning Objectives	75
4.1.2 Planning for Practice	76
4.1.3 Readings & Resources	76
4.1.4 Packages	76
4.2 <i>z</i> before <i>t</i>	77
4.2.1 Simulating a Mini Research Vignette	78
4.2.2 Raw Scores, <i>z</i> -scores, and Proportions	79
4.2.3 Determining Probabilities	80
4.2.4 Percentiles	84
4.2.5 Transforming Variables to Standard Scores	84
4.2.6 The One-Sample <i>z</i> test	85
4.3 Introducing the One Sample <i>t</i> -test	87
4.3.1 Workflow for the One Sample <i>t</i> -test	88
4.4 Research Vignette	88
4.5 Working the Problem	91
4.5.1 Stating the Hypothesis	91
4.5.2 Preliminary Exploration	92
4.5.3 Hand-Calculations	94
4.5.3.1 Statistical Significance	94
4.5.3.2 Confidence Intervals	95
4.5.3.3 Effect size	96
4.6 Computation in R	97

4.7	APA Style Results	97
4.8	Power in Independent Samples t tests	99
4.9	Practice Problems	101
4.9.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed	101
4.9.2	Problem #2: Rework the research vignette, but change something about the simulation	101
4.9.3	Problem #3: Use other data that is available to you	101
5	Independent Samples t test	103
5.1	Navigating this Lesson	103
5.1.1	Learning Objectives	103
5.1.2	Planning for Practice	104
5.1.3	Readings & Resources	104
5.1.4	Packages	104
5.2	Introducing the Independent Samples t Test	105
5.2.1	Workflow for Independent Samples t test	105
5.3	Research Vignette	107
5.4	Working the Problem	109
5.4.1	Stating the Hypothesis	109
5.4.2	Preliminary Exploration	109
5.4.3	Hand-Calculations	112
5.4.3.1	Statistical Significance	113
5.4.3.2	Confidence Intervals	114
5.4.3.3	Effect Size	115
5.5	Computation in R	116
5.5.1	What if we had violated the homogeneity of variance assumption?	117
5.6	APA Style Results	118
5.7	Power in Independent Samples t tests	119
5.8	Practice Problems	122
5.8.1	Problem #1: Rework the research vignette as demonstrated, but change the random seed	122
5.8.2	Problem #2: Rework the research vignette, but change something about the simulation	122
5.8.3	Problem #3: Rework the research vignette, but swap one or more variables .	122
5.8.4	Problem #4: Use other data that is available to you	123

6 Paired Samples <i>t</i>-test	125
6.1 Navigating this Lesson	125
6.1.1 Learning Objectives	125
6.1.2 Planning for Practice	126
6.1.3 Readings & Resources	126
6.1.4 Packages	126
6.2 Introducing the Paired Samples <i>t</i> -test	127
6.3 Workflow for Paired Samples <i>t</i> -test	127
6.4 Research Vignette	128
6.4.1 Simulating Data for the Paired Samples <i>t</i> test	129
6.5 Working the Problem	131
6.5.1 Stating the Hypothesis	131
6.5.2 Preliminary Exploration	132
6.5.3 Hand Calculations	134
6.5.3.1 Statistical Significance	134
6.5.3.2 Confidence Intervals	135
6.5.3.3 Effect Size	136
6.6 Computation in R	137
6.7 APA Style Results	138
6.8 Power in Paired Samples <i>t</i> tests	140
6.9 Practice Problems	143
6.9.1 Problem #1: Rework the research vignette as demonstrated, but change the random seed	143
6.9.2 Problem #2: Rework the research vignette, but change something about the simulation	143
6.9.3 Problem #3: Rework the research vignette, but swap one or more variables .	143
6.9.4 Problem #4: Use other data that is available to you	143
Analysis of Variance	145
7 One-way ANOVA	147
7.1 Navigating this Lesson	147
7.1.1 Learning Objectives	147
7.1.2 Planning for Practice	148
7.1.3 Readings & Resources	148

7.1.4	Packages	149
7.2	Workflow for One-Way ANOVA	149
7.3	Research Vignette	151
7.3.1	Data Simulation	151
7.4	Working the Problem	152
7.4.1	Preparing the Data	152
7.4.2	Exploring the Distributional Characteristics Numerically	154
7.4.3	Exploring the Distributional Characteristics Graphically	155
7.5	Understanding ANOVA with <i>Hand Calculations</i>	157
7.5.1	Sums of Squares Total	158
7.5.2	Sums of Squares for the Model (or Between)	162
7.5.3	Sums of Squares Residual (or within)	163
7.5.3.1	On the relationship between standard deviation and variance	164
7.5.4	Relationship between SS_T , SS_M , and SS_R	165
7.5.5	Mean Squares Model & Residual	165
7.5.6	Calculating the F Statistic	167
7.5.7	Source Table Games	167
7.6	Working the One-Way ANOVA in R	168
7.6.1	Evaluating the Statistical Assumptions	168
7.6.1.1	Is the dependent variable normally distributed across levels of the factor?	170
7.6.1.2	Are the variances of the dependent variable similar across the levels of the grouping factor?	172
7.6.1.3	Summarizing results from the analysis of assumptions	172
7.6.2	Computing the Omnibus ANOVA	173
7.6.2.1	Effect size for the one-way ANOVA	179
7.6.2.2	Summarizing results from the omnibus ANOVA	180
7.6.3	Follow-up to the Omnibus F	181
7.6.3.1	OPTION 1: Post-hoc, pairwise, comparisons	181
7.6.3.2	OPTION 2: Planned contrasts (non-orthogonal)	183
7.6.3.3	OPTION 3: Planned contrasts (orthogonal)	185
7.6.3.4	OPTION 4: Trend (polynomial) analysis	187
7.6.3.5	Which set of follow-up tests do we report?	189
7.6.4	What if we Violated the Homogeneity of Variance test?	190

7.7	Power Analysis	190
7.8	APA Style Results	192
7.9	A Conversation with Dr. Tran	196
7.10	Practice Problems	196
7.10.1	Problem #1: Play around with this simulation.	196
7.10.2	Problem #2: Conduct a one-way ANOVA with the <i>moreTalk</i> dependent variable.	197
7.10.3	Problem #3: Try something entirely new.	197
7.10.4	Grading Rubric	197
7.11	Bonus Reel:	197
7.11.1	What's with the inline code?	197
8	Factorial (Between-Subjects) ANOVA	199
8.1	Navigating this Lesson	199
8.1.1	Learning Objectives	200
8.1.2	Planning for Practice	200
8.1.3	Readings & Resources	200
8.1.4	Packages	201
8.2	Introducing Factorial ANOVA	201
8.2.1	Workflow for Two-Way ANOVA	202
8.3	Research Vignette	204
8.3.1	Preliminary exploration of our research vignette	206
8.4	Working the Factorial ANOVA (by hand)	210
8.4.1	Sums of Squares Total	210
8.4.2	Sums of Squares for the Model	212
8.4.3	Sums of Squares Residual (or within)	213
8.4.4	A Recap on the Relationship between SS_T , SS_M , and SS_R	214
8.4.5	Calculating SS for Each Factor and Their Products	215
8.4.5.1	Rater Main Effect	215
8.4.5.2	Photo Main Effect	215
8.4.5.3	Interaction effect	216
8.4.6	Source Table Games!	216
8.4.7	Interpreting the results	219
8.5	Working the Factorial ANOVA with R packages	219

8.5.1	Evaluating the statistical assumptions	219
8.5.1.1	DV is normally distributed	221
8.5.1.2	Homogeneity of variance	225
8.5.2	Evaluating the Omnibus ANOVA	225
8.5.2.1	Effect sizes	231
8.5.2.2	APA Write-up of the omnibus results	232
8.5.3	Follow-up a significant interaction effect	232
8.5.3.1	Option #1 the simple main effect of photo stimulus within ethnicity of the rater	234
8.5.3.2	Option #2 the simple main effect of ethnicity of rater within photo stimulus.	237
8.5.3.3	Option #3 post hoc comparisons	241
8.5.3.4	Option #4 polynomial trends	246
8.6	Investigating Main Effects	249
8.6.1	Follow-up with all Post-Hocs	251
8.6.2	Follow-up with planned contrasts	253
8.6.3	Polynomial Trends	255
8.7	My APA Style Results Section	258
8.7.1	Comparing Our Results to Rhamdani et al. [2018]	260
8.8	Options for Assumption Violations	261
8.9	Power	263
8.9.1	Post Hoc Power Analysis	264
8.9.2	Estimating Sample Size Requirements	265
8.10	Practice Problems	266
8.10.1	Problem #1: Play around with this simulation.	266
8.10.2	Problem #2: Conduct a factorial ANOVA with the <i>positive evaluation</i> dependent variable.	267
8.10.3	Problem #3: Try something entirely new.	267
8.10.4	Grading Rubric	267
9	One-Way Repeated Measures ANOVA	269
9.1	Navigating this Lesson	269
9.1.1	Learning Objectives	269
9.1.2	Planning for Practice	270
9.1.3	Readings & Resources	270

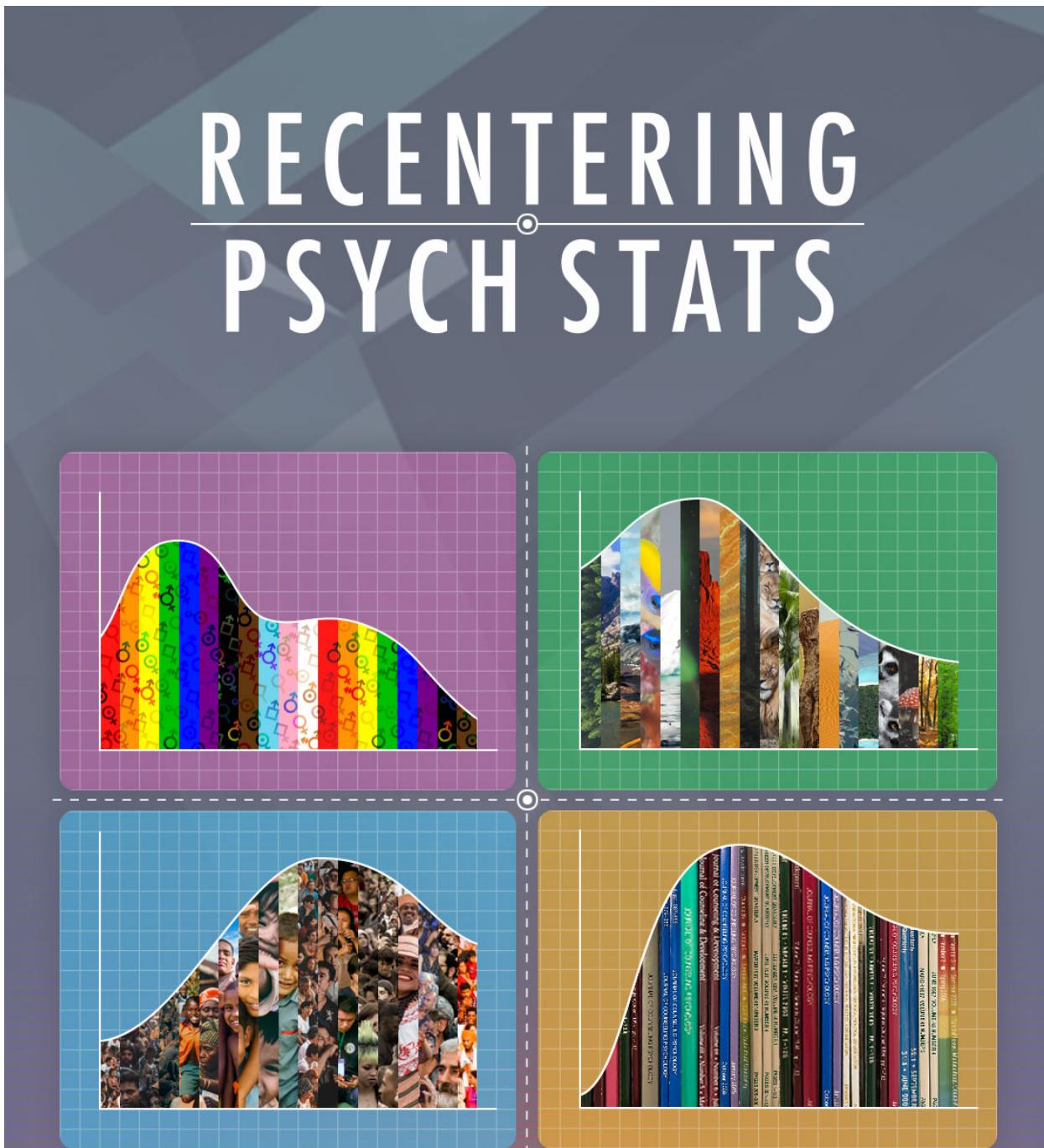
9.1.4 Packages	271
9.2 Introducing One-way Repeated Measures ANOVA	271
9.2.1 Workflow for Oneway Repeated Measures ANOVA	272
9.3 Research Vignette	272
9.3.1 Code for simulating the data used today.	274
9.3.2 Quick peek at the data	276
9.4 Working the One-Way Repeated Measures ANOVA (by hand)	278
9.4.1 Sums of Squares Total	278
9.4.2 Sums of Squares Within for Repeated Measures ANOVA	280
9.4.3 Sums of Squares Model – Effect of Time	280
9.4.4 Sums of Squares Residual	281
9.4.5 Sums of Squares Between	282
9.4.6 Mean Squares Model & Residual	283
9.4.7 <i>F</i> ratio	283
9.5 Working the One-Way ANOVA with R packages	284
9.5.1 Testing the assumptions	284
9.5.1.1 Univariate assumptions for repeated measures ANOVA	284
9.5.1.2 Demonstrating sphericity	286
9.5.1.3 Any outliers?	286
9.5.1.4 Assessing normality	288
9.5.1.5 Assumption of Sphericity	290
9.5.2 Omnibus Repeated Measures ANOVA	290
9.5.3 Follow-up	292
9.5.4 Results Section	295
9.5.4.1 Creating an APA Style Table**	296
9.5.4.2 Comparison with Amodeo et al.[2018]	296
9.6 Power in Repeated Measures ANOVA	297
9.7 Practice Problems	298
9.7.1 Problem #1: Change the Random Seed	299
9.7.2 Problem #2: Increase <i>N</i>	299
9.7.3 Problem #3: Try Something Entirely New	299
9.7.4 Grading Rubric	299
9.8 Bonus Reel:	300

10 Mixed Design ANOVA	303
10.1 Navigating this Lesson	303
10.1.1 Learning Objectives	303
10.1.2 Planning for Practice	304
10.1.3 Readings & Resources	304
10.1.4 Packages	305
10.2 Introducing Mixed Design ANOVA	305
10.3 Research Vignette	307
10.3.1 Simulating the data from the journal article	308
10.4 Working the Mixed Design ANOVA with R packages	311
10.4.1 Exploring data and testing assumptions	311
10.4.1.1 Assumption of Normality	312
10.4.1.2 Homogeneity of variance assumption	316
10.4.1.3 Assumption of homogeneity of covariance matrices	317
10.4.1.4 APA style writeup of assumptions	317
10.4.2 Omnibus ANOVA	318
10.4.2.1 Checking the sphericity assumption	319
10.4.3 Simple main effect of condition within wave	320
10.4.4 Simple main effect of wave within condition	323
10.4.5 If we only had a main effect	325
10.4.6 APA Style Write-up of the Results	326
10.4.6.1 Results	328
10.4.6.2 Comparing our findings to Murrar and Brauer [2018]	329
10.5 Power in Mixed Design ANOVA	330
10.6 Practice Problems	331
10.6.1 Problem #1: Play around with this simulation.	332
10.6.2 Problem #2: Conduct a one-way ANOVA with a different dependent variable.	332
10.6.3 Problem #3: Try something entirely new.	332
10.6.4 Grading Rubric	332
11 Analysis of Covariance	333
11.1 Navigating this Lesson	333
11.1.1 Learning Objectives	333
11.1.2 Planning for Practice	333

11.1.3 Readings & Resources	334
11.1.4 Packages	335
11.2 Introducing Analysis of Covariance (ANCOVA)	335
11.3 Research Vignette	338
11.3.1 Simulating the data from the journal article	338
11.4 Scenario #1: Controlling for the pretest	341
11.4.1 Preparing the data	341
11.4.2 Checking the assumptions	342
11.4.2.1 Linearity assumption	344
11.4.2.2 Homogeneity of regression slopes	345
11.4.2.3 Normality of residuals	345
11.4.2.4 Homogeneity of variances	346
11.4.2.5 Outliers	347
11.4.2.6 Write-up of Assumptions	347
11.4.3 Calculating the Omnibus ANOVA	347
11.4.4 Post-hoc pairwise comparisons (controlling for the covariate)	349
11.4.5 Toward an APA style results section	351
11.5 Scenario #2: Controlling for a confounding or covarying variable	354
11.5.1 Preparing the data	354
11.5.2 Checking the assumptions	354
11.5.2.1 Linearity assumption	355
11.5.2.2 Homogeneity of regression slopes	356
11.5.2.3 Normality of residuals	357
11.5.2.4 Homogeneity of variances	358
11.5.2.5 Outliers	358
11.5.2.6 Write-up of Assumptions	359
11.5.3 Calculating the Omnibus ANOVA	359
11.5.4 Post-hoc pairwise comparisons (controlling for the covariate)	361
11.5.5 Toward an APA style results section	362
11.6 More (and a recap) on covariates	365
11.7 Practice Problems	366
11.7.1 Problem #1: Play around with this simulation.	366
11.7.2 Problem #2: Conduct a one-way ANCOVA with the DV and covariate at post2.	366

11.7.3 Problem #3: Try something entirely new.	366
11.7.4 Grading Rubric	366
References	369

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 10 Oct 2022. Emerging volumes on other statistics are posted on the [ReCentering Psych Stats](#) page at my research team's website.

[Screencasted 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 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, 2014b](#)] 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.

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 created. 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 notifying 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_R](#)lesson of this volume provides an introduction and the [waRming up](#)lesson 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.

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 have 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, though, 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 (verified by getting a classmate or colleague to successfully knit it), 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 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) navigating 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 it, you can see if it is “on.” 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 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, for uploading data I write the code in my .rmd file. For opening packages, I include the package in my script. I also use the .rmd file to make notes about what I was thinking and why I made the choices I did. I also 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:

- Create a project file folder.
- Put the data file in it.
- Open an R Markdown file.
- Save it in the same file folder.
- When your data and .rmd files are in the same folder (not your desktop, but a shared 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-out comments so they won’t run.

You can put almost anything you want in the “inline text with simple formatting.” Syntax for simple formatting in the text 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, lavaan, 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. 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 they 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 right in each lesson for use in the lesson. This makes the web-based platform more *portable*. This means that when working the problems in the chapter we do not (a) write the data to a file and (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 the .html format. In prior years, I knitted these documents to .doc formats. There are numerous possibilities!
- 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.
- If you are using the PDF or ebook version of this OER you will realize that it is also possible to render to these formats. Albeit slightly more complicated, this is possible, too! More detailed instructions for this 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 like to 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 (or at least it appears to be) is the futuRe.
- SPSS site (and individual) licenses are increasingly expensive and limited (e.g., we need Mplus, AMOS, HLM, or R). 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 aspire 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 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 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	_____

Preliminary Analyses

Chapter 3

Preliminary Results

[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. While the process of simulation is beyond the learning goals of this lesson (you can skip that part), I include it here so that it is easy to work the rest of the script.

```
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 quite similar) There are
# probably more efficient ways to simulate data. Given the
# information available in the manuscript, my approach was to first
# create the datasets for each of the racial ethnic groups that were
# provided and then binding 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)
Asian_stddev <- c(2.52, 2.04, 0.47, 0.7, 0.8, 2.41, 3.36)
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,
  0.29, 0.5, 1, 0.76, 0.04, 0.18, 0.32, 0.33, 0.5, 0.76, 1, 0.1, 0.21,
```

```

  0.08, 0.13, -0.04, 0.04, 0.1, 1, 0.42, 0.23, 0.25, 0.09, 0.18, 0.21,
  0.42, 1), ncol = 7)
Asian_covMat <- Asian_stddev %*% t(Asian_stddev) * Asian_corMat

Asian_dat <- MASS::mvrnorm(n = 398, mu = Asian_mu, Sigma = Asian_covMat,
  empirical = TRUE)
Asian_df <- as.data.frame(Asian_dat)

library(tidyverse)
Asian_df <- rename(Asian_df, OvDisc = V1, mAggr = V2, Neuro = V3, nAff = V4,
  psyDist = V5, Alcohol = V6, drProb = V7)

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

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

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

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

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

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

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

Asian_df$RacEth <- "Asian"

# Second, the data for the students who identified as Black/African American
Black_mu <- c(4.45, 3.84, 2.6, 1.84, 2.1, 2.81, 2.14)
Black_stddev <- c(4.22, 3.08, 0.89, 0.8, 0.81, 2.49, 3.24)
Black_corMat <- matrix(c(1, 0.81, 0.17, 0.15, 0.09, 0.05, -0.16, 0.81,
  1, 0.17, 0.21, 0.11, 0.09, -0.01, 0.17, 0.17, 1, 0.59, 0.54, 0.05,
  0.24, 0.15, 0.21, 0.59, 1, 0.72, 0.12, 0.22, 0.09, 0.11, 0.54, 0.72,
  1, 0.21, 0.4, 0.05, 0.09, 0.05, 0.12, 0.21, 1, 0.65, -0.16, -0.01,
  0.24, 0.22, 0.4, 0.65, 1), ncol = 7)
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"

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 may wish to save data as a file. 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 stands for *dataframe* and is the common term for a dataset for users of R.

```
df <- Lui_sim_df
```

3.3 Variable Types (Scale of Measurement)

Starting with raw data always begins with inspecting the data preparing it for the planned analyses. The *type* of variables we have influences what statistics we choose to analyze our data. 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 technically 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 considered to be ordinal because it is uncertain that the distance between each of the numbers 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] terms the Likert a **quasi-interval scale**.

Continuous variables can take on any value in the measurement scale that is being used. **Interval** level data have 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 an 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
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 1.45 4.9 0.45 2.85 2.09 ...
 $ mAggr : num 0.682 4.383 0.225 2.235 1.977 ...
 $ Neuro : num 3.11 3.69 3.5 2.68 2.08 ...
 $ nAff : num 2.32 2.59 2.27 2.28 2.01 ...
 $ psyDist: num 1.83 3.41 2.75 2.11 3.12 ...
 $ Alcohol: num 3.125 4.388 0.999 0.137 0 ...
 $ drProb : num 2.5 0 3.04 0 0 ...
 $ 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.

```
# A .csv file is uninformed -- it just holds data (and R guesses what
# it is); respecifying the type of variable will likely need to be
# completed each time the file is used.
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 1.45 4.9 0.45 2.85 2.09 ...
 $ mAggr : num 0.682 4.383 0.225 2.235 1.977 ...
 $ Neuro : num 3.11 3.69 3.5 2.68 2.08 ...
 $ nAff : num 2.32 2.59 2.27 2.28 2.01 ...
 $ psyDist: num 1.83 3.41 2.75 2.11 3.12 ...
 $ Alcohol: num 3.125 4.388 0.999 0.137 0 ...
 $ drProb : num 2.5 0 3.04 0 0 ...
 $ RacEth : Factor w/ 3 levels "Asian","Black",...: 1 1 1 1 1 1 1 1 1 1 ...
```

It is possible to work with this data without restructuring them into a “tiny df.” However, this function is often one of the first skills we want to use so I will demonstrate it here.

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.

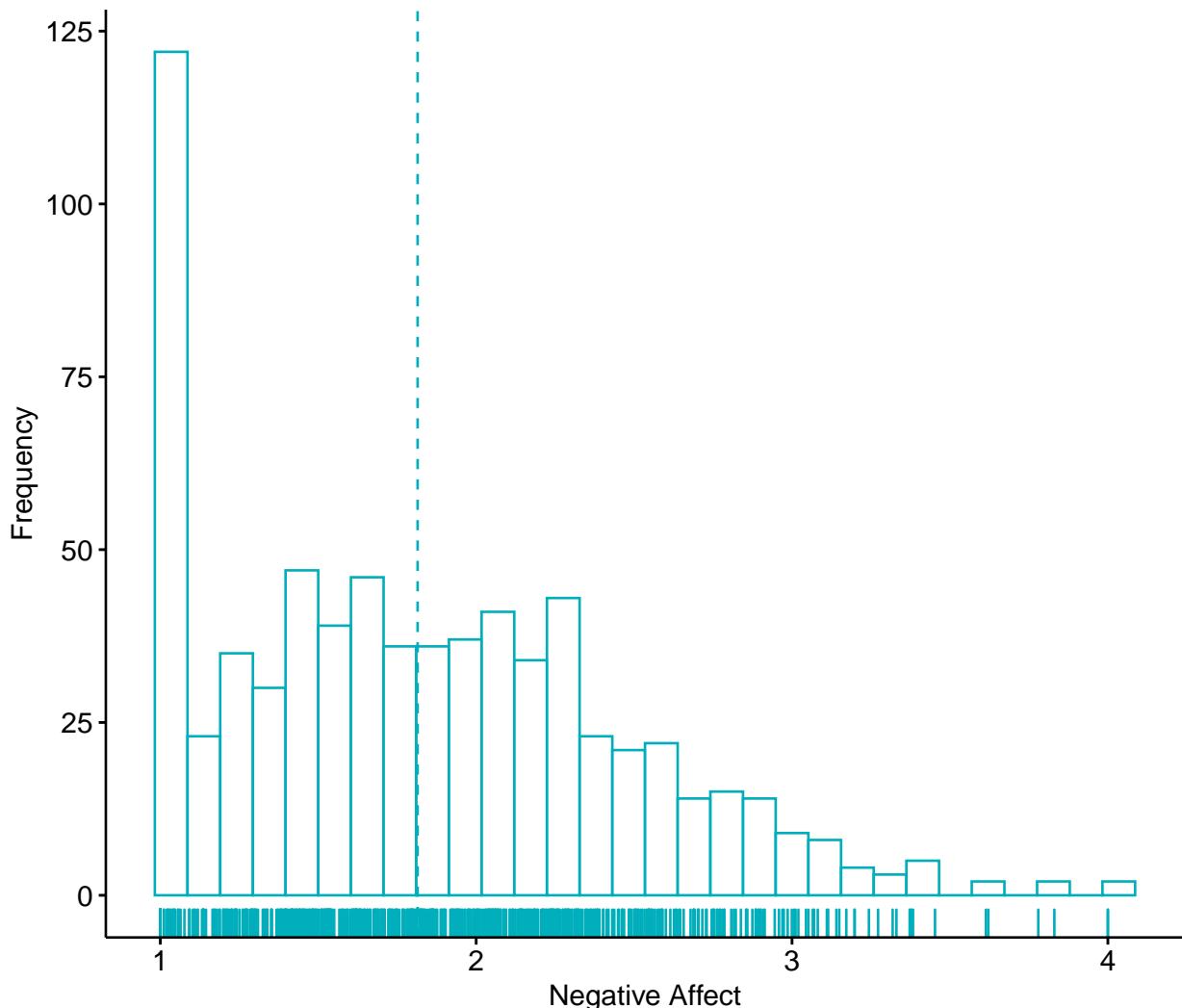
As we begin to explore the descriptive and distributional characteristics of this variable, it may be helpful to visualize it through a histogram.

```
ggpubr::gghistogram(df$nAff, xlab="Negative Affect", ylab = "Frequency", add = "mean", rug=TRUE)
```

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

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

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



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 such a 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 instructions 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
```

Because of my simulation, 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.

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

```
[1] 1.814666
```

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:

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

```
[1] 2
```

Let's also calculate it for our negative affect variable.

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

```
[1] 1.750954
```

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] 1.814666
```

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

```
[1] 1.750954
```

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

```
[1] 1
```

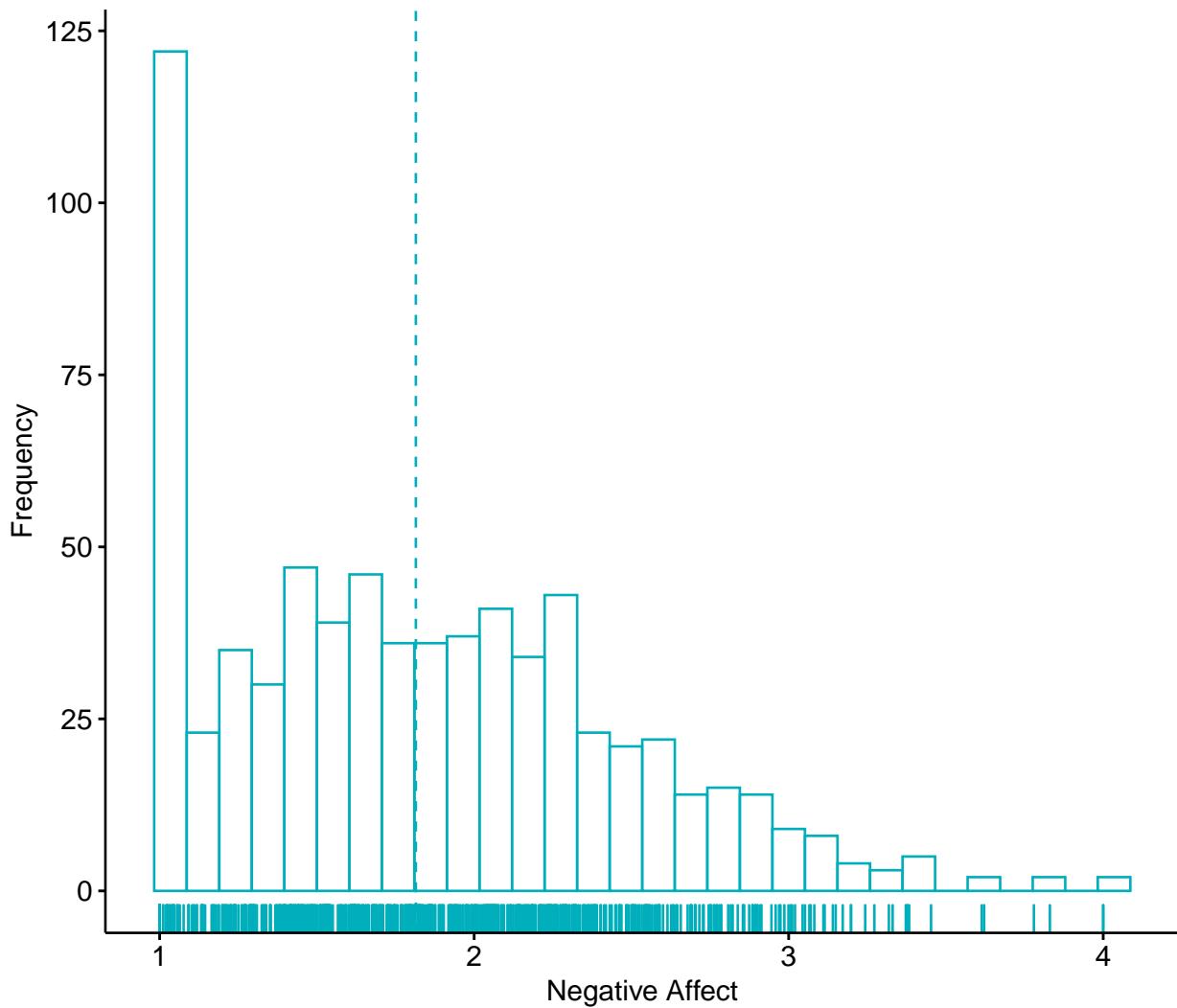
In our research vignette, the mean (1.81) is higher than the median (1.75) is higher than the mode (1.0). This would suggest a positive skew. Here is a reminder of our histogram:

```
ggpubr::gghistogram(df$nAff, xlab="Negative Affect", ylab = "Frequency", add = "mean", rug=TRUE)
```

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

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

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



Variability

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

3.4.2 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.

3.4.3 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. **Percentile** 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] 1.750954
```

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

```
50%
1.750954
```

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	1.182154	1.402778	1.549423	1.750954	1.946013	2.142151	2.334597
	90%						
		2.688985					

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%
1.289916 2.241442
```

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] 0.9515256
```

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 is 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**.

3.4.4 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_name$variable_name
toy$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	2.316454	0.5017878	0.5017878
2	2.585344	0.7706780	0.7706780
3	2.274760	0.4600937	0.4600937
4	2.281637	0.4669707	0.4669707
5	2.005462	0.1907964	0.1907964
6	1.174359	-0.6403072	0.6403072

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

```
[1] 0.521
```

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.

3.4.5 Variance

Variance is considered to be an *average* dispersion calculated by summing the squared deviations and dividing by the number of observations (less 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	2.316454	0.5017878	0.5017878	0.25179095
2	2.585344	0.7706780	0.7706780	0.59394458
3	2.274760	0.4600937	0.4600937	0.21168625
4	2.281637	0.4669707	0.4669707	0.21806160
5	2.005462	0.1907964	0.1907964	0.03640325
6	1.174359	-0.6403072	0.6403072	0.40999325

... and sum them.

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

```
[1] 283.4418
```

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)
```

```
[1] 0.3980899
```

```
sum(df_nAff$NAmd2, na.rm = TRUE)/((nrow(df_nAff) - 1)) #variance
```

```
[1] 0.3980924
```

```
var(df_nAff$nAff)
```

```
[1] 0.3980924
```

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

3.4.6 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] 0.3980924
```

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

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

```
[1] 1.814666
```

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

[1] 0.6309456

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

[1] 0.6309456

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 could include $M = 1.81 (SD = 0.63)$ in a statistical string of results.

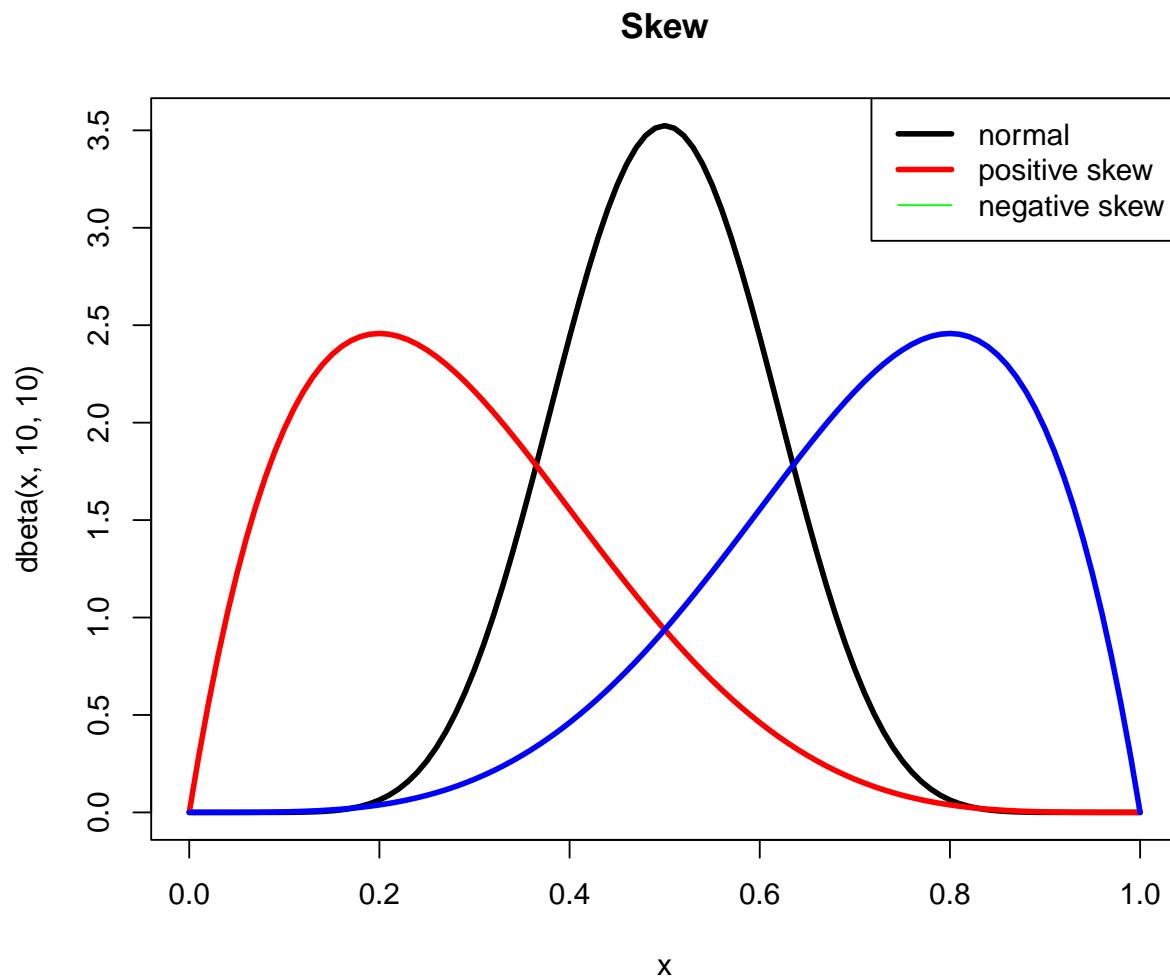
We can examine the standard deviation in relation to its mean to understand how narrowly or broadly the data is distributed. Relative to a same-sized mean, a small standard deviation means that the mean represents the data well. A larger standard deviation, means there is more variability and the mean,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

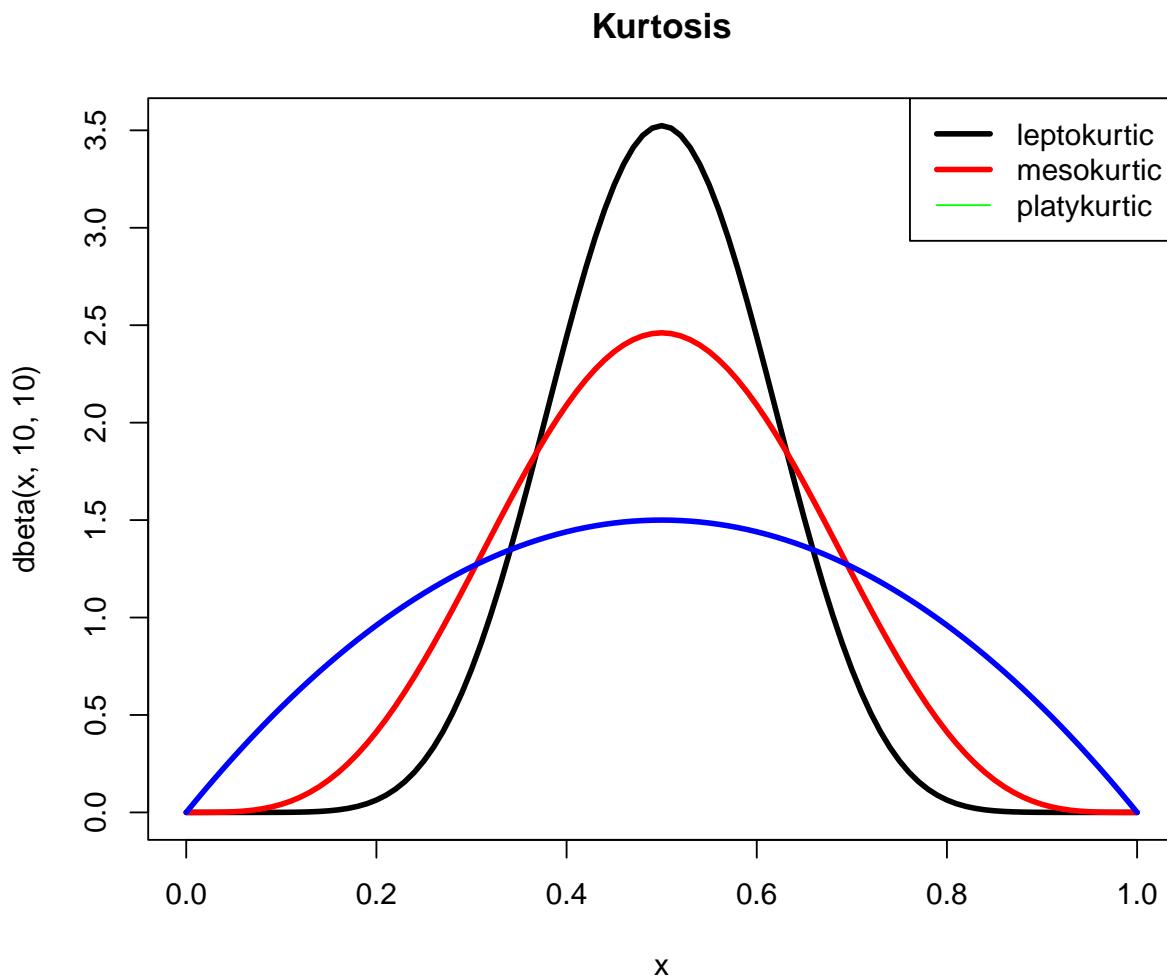
3.5 Are the Variables Normally Distributed?

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.

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) would have no skewness. 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) would be 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) would be negatively skewed.



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.



The *describe()* function in the *psych* package is one of the most common ways to obtain information on skew and kurtosis, as well as other descriptive information.

For just a quick view, call the *psych* package, add two colons, use the *describe()* function, and indicate the df.

For a simplified presentation, let me create a df with three 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)
```

```
psych::describe(df_3vars)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
nAff	1	713	1.81	0.63	1.75	1.76	0.71	1	4.00	3.00	0.58	-0.18
mAggr	2	713	2.49	2.20	2.17	2.22	2.42	0	10.42	10.42	0.93	0.58

	drProb	3	713	2.92	2.77	2.40	2.58	3.49	0	11.69	11.69	0.78	-0.17
se													
nAff		0.02											
mAggr		0.08											
drProb		0.10											

If you will be needing to create a table of information in another application such as Word or Excel, save your descriptives as an object and then write them to a .csv file.

```
LuiDescripts <- psych::describe(df_3vars)
write.csv(LuiDescripts, file = "LuiDescripts.csv")
# Once you create an object, the result won't automatically display;
# to see it simply type the name of the object
LuiDescripts
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis
nAff	1	713	1.81	0.63	1.75	1.76	0.71	1	4.00	3.00	0.58	-0.18
mAggr	2	713	2.49	2.20	2.17	2.22	2.42	0	10.42	10.42	0.93	0.58
drProb	3	713	2.92	2.77	2.40	2.58	3.49	0	11.69	11.69	0.78	-0.17
se												
nAff		0.02										
mAggr		0.08										
drProb		0.10										

To understand whether our data is normally distributed, we can look at skew and kurtosis. The skew and kurtosis indices in the *psych* package are reported as *z* scores. Positive skew values indicate positive skew; negative skew values represent negative skew. Skew values greater than 3.0 are generally considered to be “severely skewed” [Kline, 2016].

Regarding kurtosis, positive values indicate the distribution is somewhat more peaked than a normal distribution (i.e., more leptokurtic); negative values indicate the distribution is flatter than a normal distribution (i.e., more platykurtic). Regarding kurtosis, “severely kurtotic” is argued anywhere > 8 to 20 [Kline, 2016].

The values for all the variables in the research vignette are well below the regions of concern indicated by Kline [2016]

3.6 Relations between Variables

Preliminary investigation of data almost always includes a report of their bivariate relations, or correlations. 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. The problem is that the result is unstandardized and difficult to interpret.

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

The second significant calculation 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 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 this much of the problem. 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	2.316454	0.6822586
2	2.585344	4.3834353
3	2.274760	0.2251289
4	2.281637	2.2351541
5	2.005462	1.9765313
6	1.174359	0.0000000

```
# 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	2.316454	0.6822586	0.5017878	-1.8033045
2	2.585344	4.3834353	0.7706780	1.8978722
3	2.274760	0.2251289	0.4600937	-2.2604342
4	2.281637	2.2351541	0.4669707	-0.2504090
5	2.005462	1.9765313	0.1907964	-0.5090318
6	1.174359	0.0000000	-0.6403072	-2.4855631

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. That is, first we multiply the mean deviations and then sum them.

```
# Creating a crossproduct variable 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	2.316454	0.6822586	0.5017878	-1.8033045	-0.90487609
2	2.585344	4.3834353	0.7706780	1.8978722	1.46264835
3	2.274760	0.2251289	0.4600937	-2.2604342	-1.04001163
4	2.281637	2.2351541	0.4669707	-0.2504090	-0.11693366
5	2.005462	1.9765313	0.1907964	-0.5090318	-0.09712141
6	1.174359	0.0000000	-0.6403072	-2.4855631	1.59152382

Next, we sum the column of cross-products.

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

```
[1] 265.9915
```

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

```
# I have created the object 'cov' so I can use it in a calculation,
# later
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] 0.3735836
```

The covariance between negative affect and psychological distress is 0.373.

We now engage the second part of the formula to standardize it.

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

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] 0.2690291
```

Our results suggest that the relationship between negative affect and psychological distress is positive, as one increases so does the other. Is it strong? That really depends on what you are studying. The traditional values of .10, .30, and .50 are used 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 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 = 7.4481, df = 711, p-value = 0.0000000000002749
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1995470 0.3358194
sample estimates:
cor
0.2690291
```

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

3.7 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 simple and efficient.

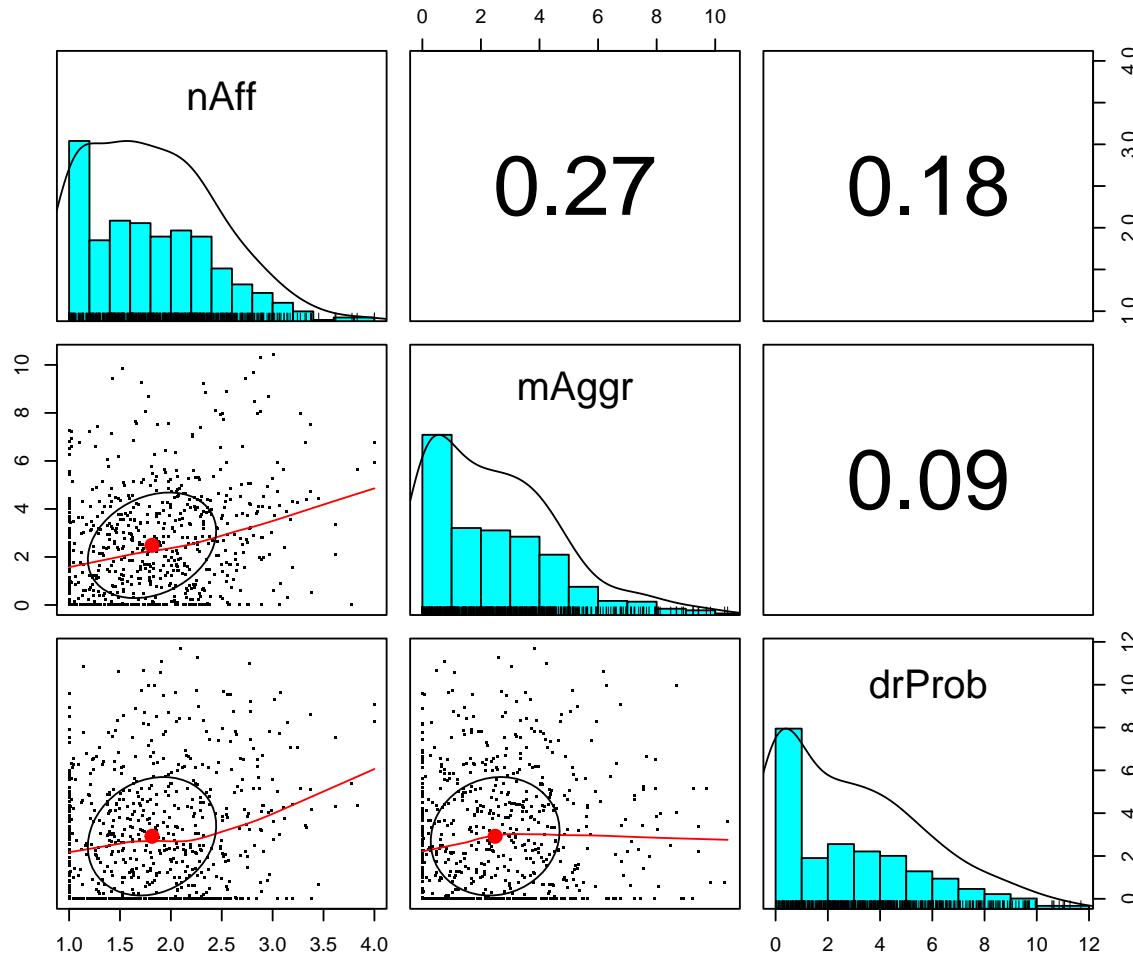
3.7.1 SPLOM

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 three 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)
- 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.

3.7.2 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 contributing 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 *apa.cor.table()* function of the *apaTables* package. 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	2.20	.27** [.20, .34]	
3. drProb	2.92	2.77	.18** [.11, .25]	.09* [.01, .16]

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.

3.8 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., takes up 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. Visual inspection of the three variables of interest (negative affect, microaggressions, drinking problems) combined with formal evaluation of skewness and kurtosis suggested that their distributions did not violate the assumption of univariate normality. 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.

3.9 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, you should:

- Create a smaller df from a larger df selecting only continuously scaled variables
- Calculate and interpret descriptives
- Create the SPLOM (pairs.panels)
- Use the *apaTables* package to make an APA style table with means, standard deviations, and correlations
- Write up an APA Style results section for these preliminary analyses

3.9.1 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.

3.9.2 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*).

3.9.3 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.

3.9.4 Grading Rubric

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

Element	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	

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 it would be 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 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.

- 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 [2014b] 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.

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(lsr)){install.packages('lsr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(knitr)){install.packages('knitr')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(pwr)){install.packages('pwr')}
```

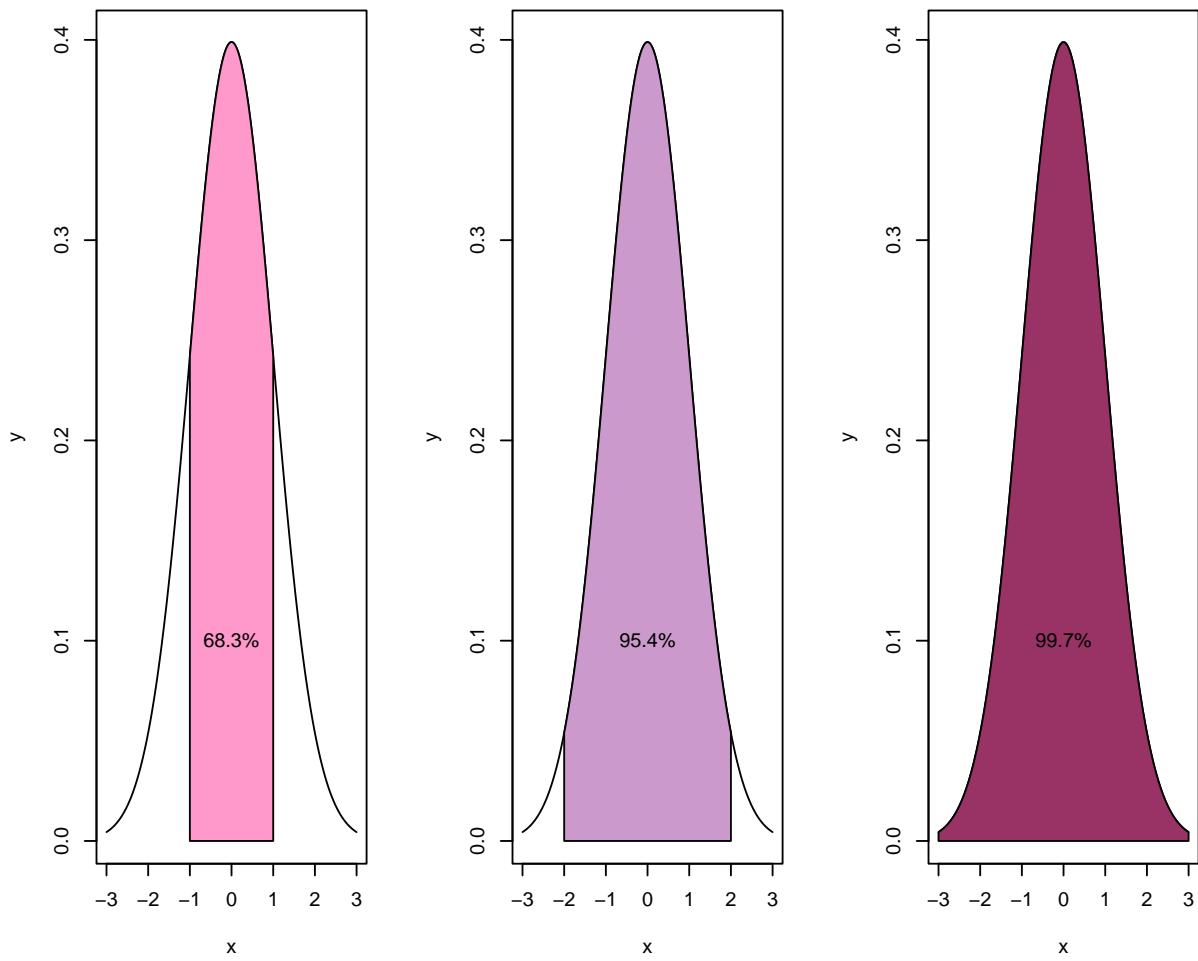
4.2 *z* before *t*

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 what the actual mean and standard deviation are

- 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



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.

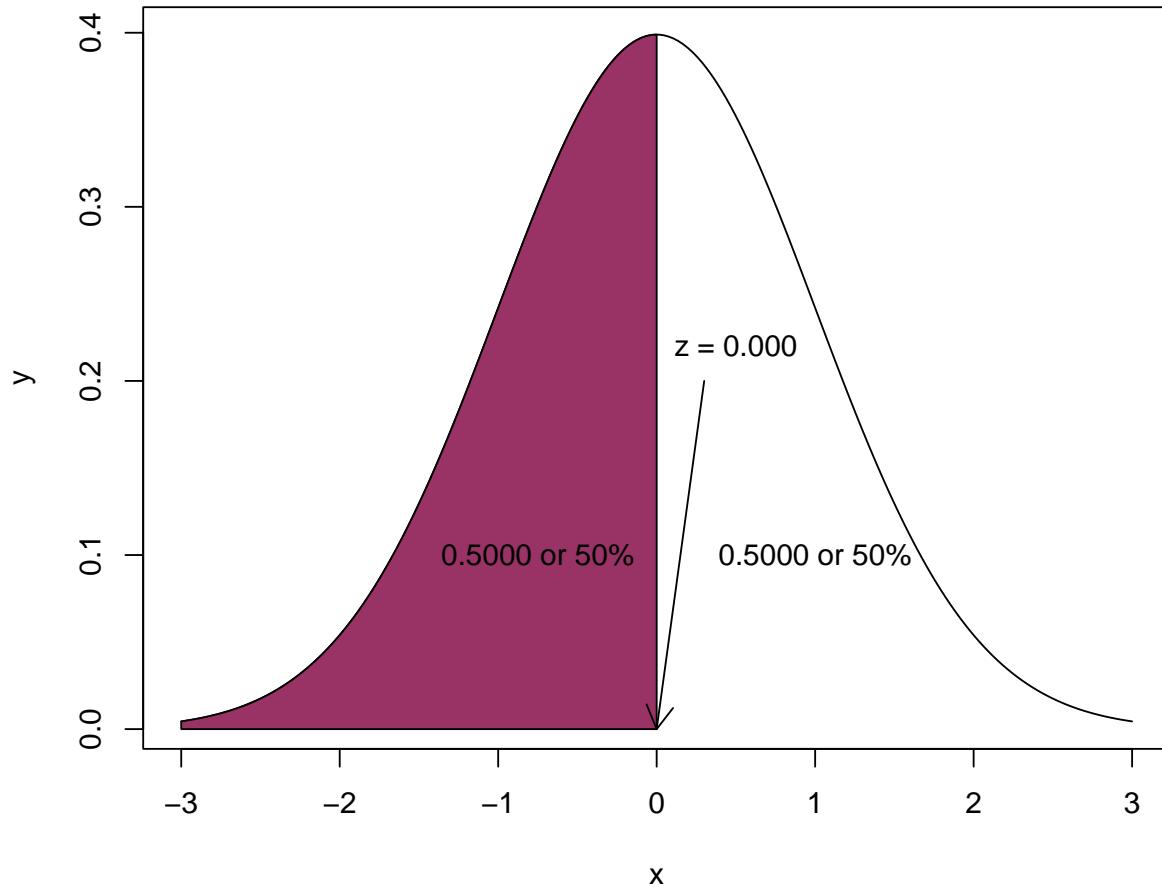
$$z = \frac{X - \bar{X}}{s}$$

```
(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
```

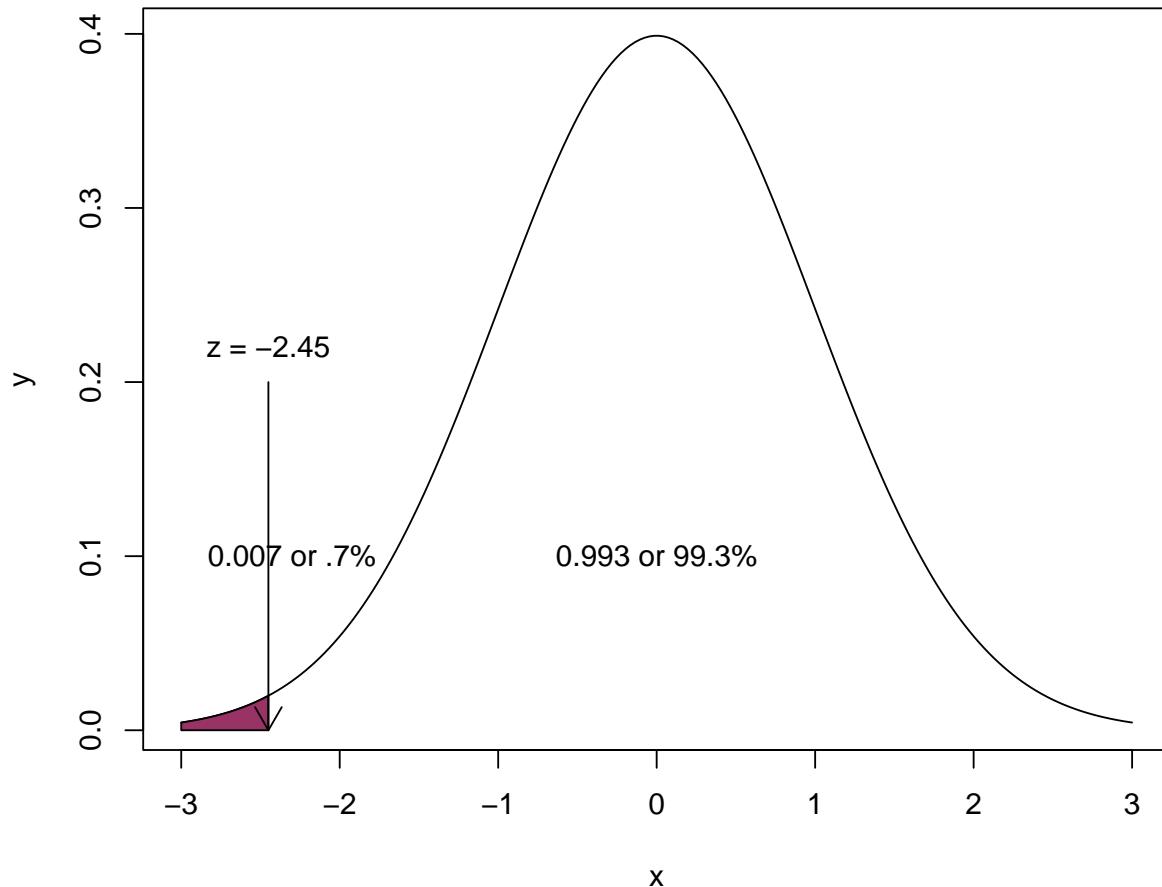
Second we locate the corresponding area under the normal curve. Examining the table of critical z -values we see that a z -score of -2.45 corresponds with an area of 0.0071. We can check this with the `pnorm()` function:

```
pnorm(-2.45, mean = 0, sd = 1)
```

```
[1] 0.007142811
```

```
pnorm(5, mean = 9.9, sd = 2)
```

```
[1] 0.007142811
```



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.

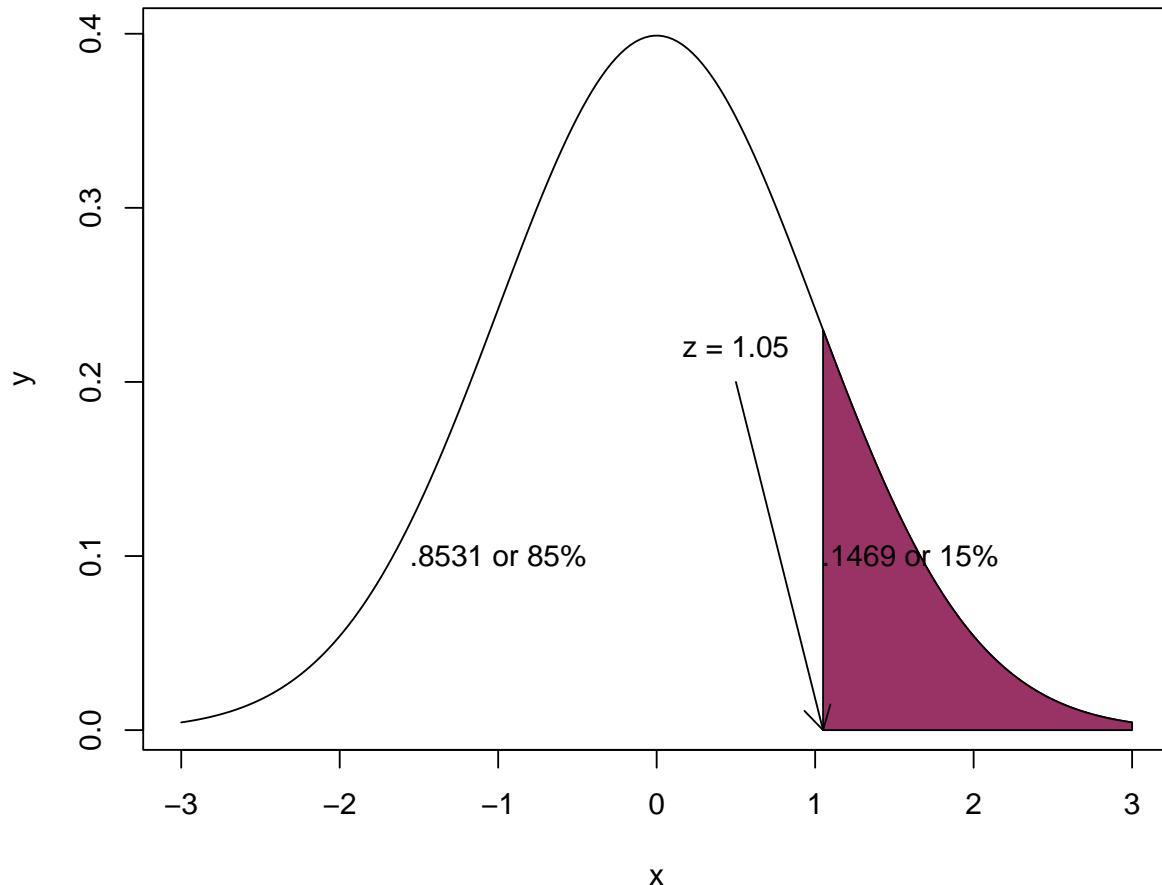
```
(12 - 9.9)/2 #for 12 minutes
```

```
[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
```



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 zversion 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..

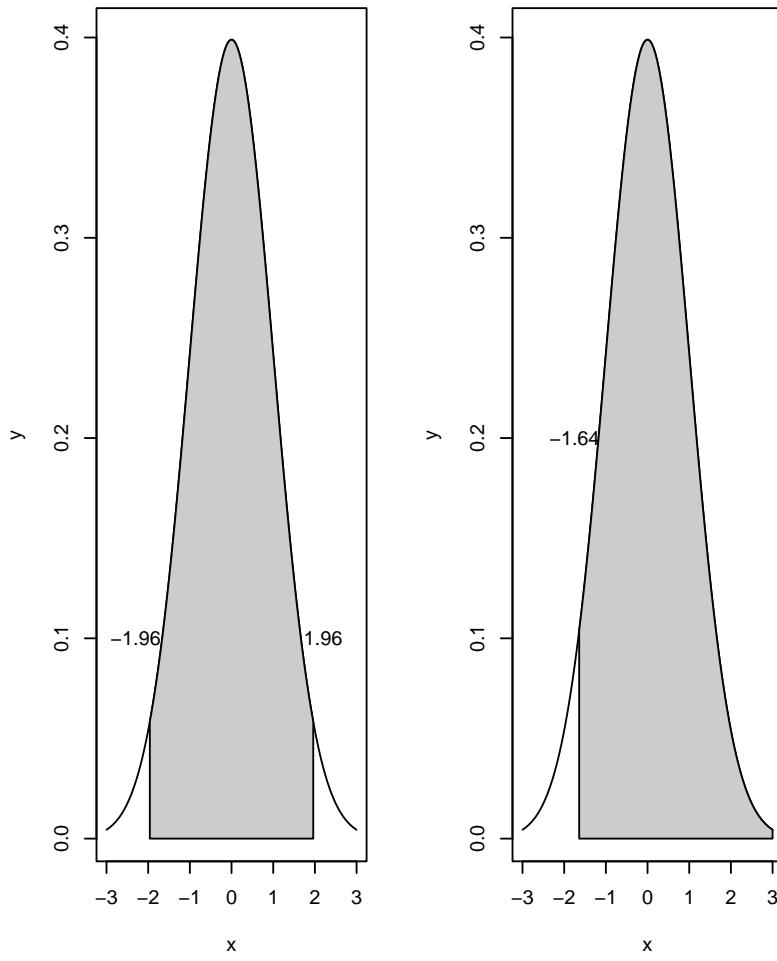
Statistician, Sir Ronald Fisher, popularized 5% as the region of rejection. Specifically, if a probability value associated with a z -score (or similar) falls into the tails of a distribution that represent 5%, then the H_0 is rejected, in favor of the H_A .

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 divide 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 have to have 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 grab 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 faaaaaar exceeds ± 1.96 we know that there is a statistically significant effect. Just to be sure, 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 seen in the published literature. However, a close inspection of tables that contain the critical values for *t*-tests reveals that the very bottom row (i.e., when sample sizes are 120 or greater) is, in fact, the *z* criteria. That must mean it's 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 [Green and Salkind, 2014b] 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.

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:

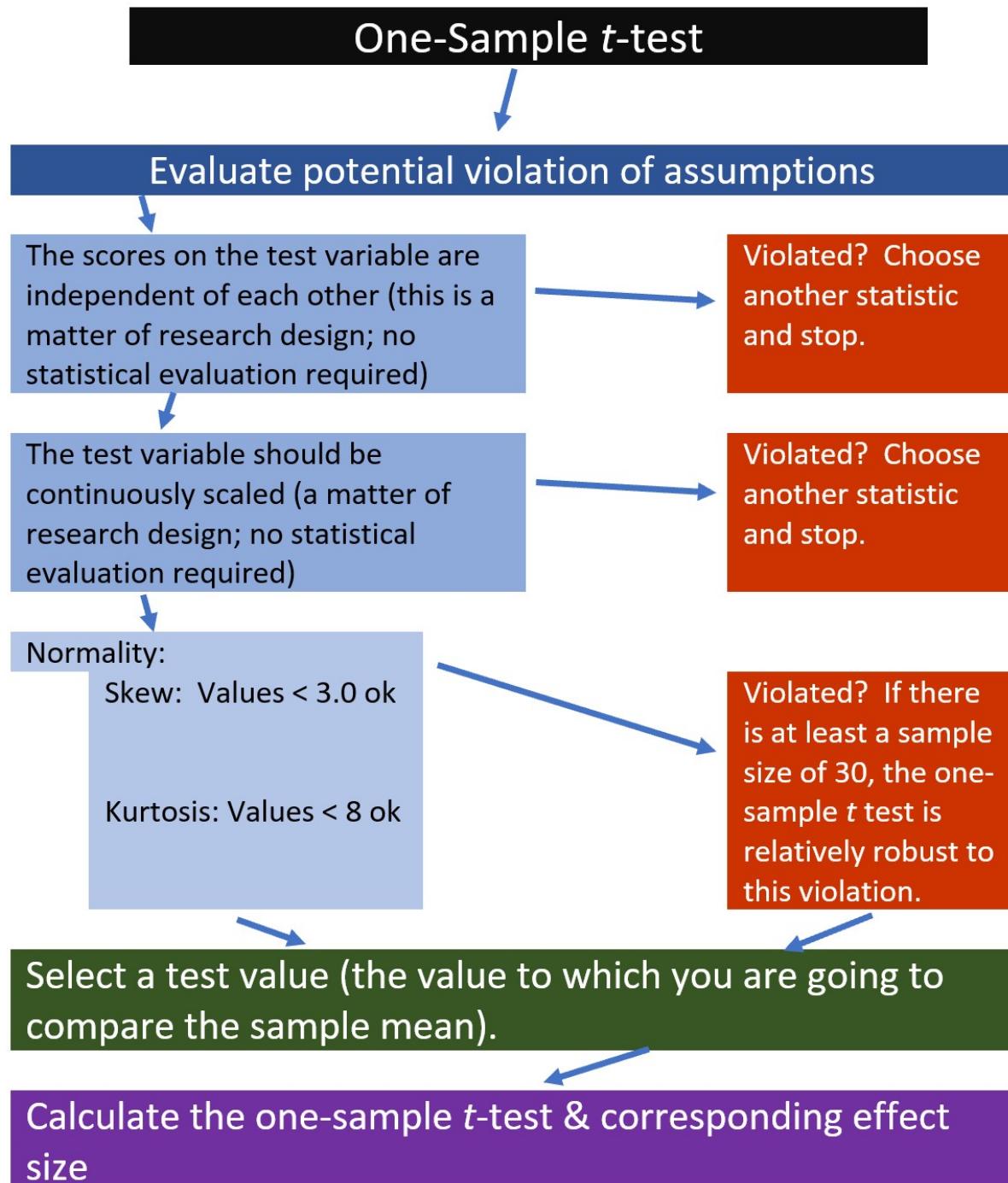
1. Prepare (upload) data.
2. Explore data with
 - graphs
 - descriptive statistics
3. Assess normality via skew and kurtosis
4. Select the comparison (i.e., test) value
5. Compute the one sample t -test
6. Compute an effect size (frequently the d statistic)
7. Manage Type I error
8. Sample size/power 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

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

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	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

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.

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

4.5 Working the Problem

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 Preliminary Exploration

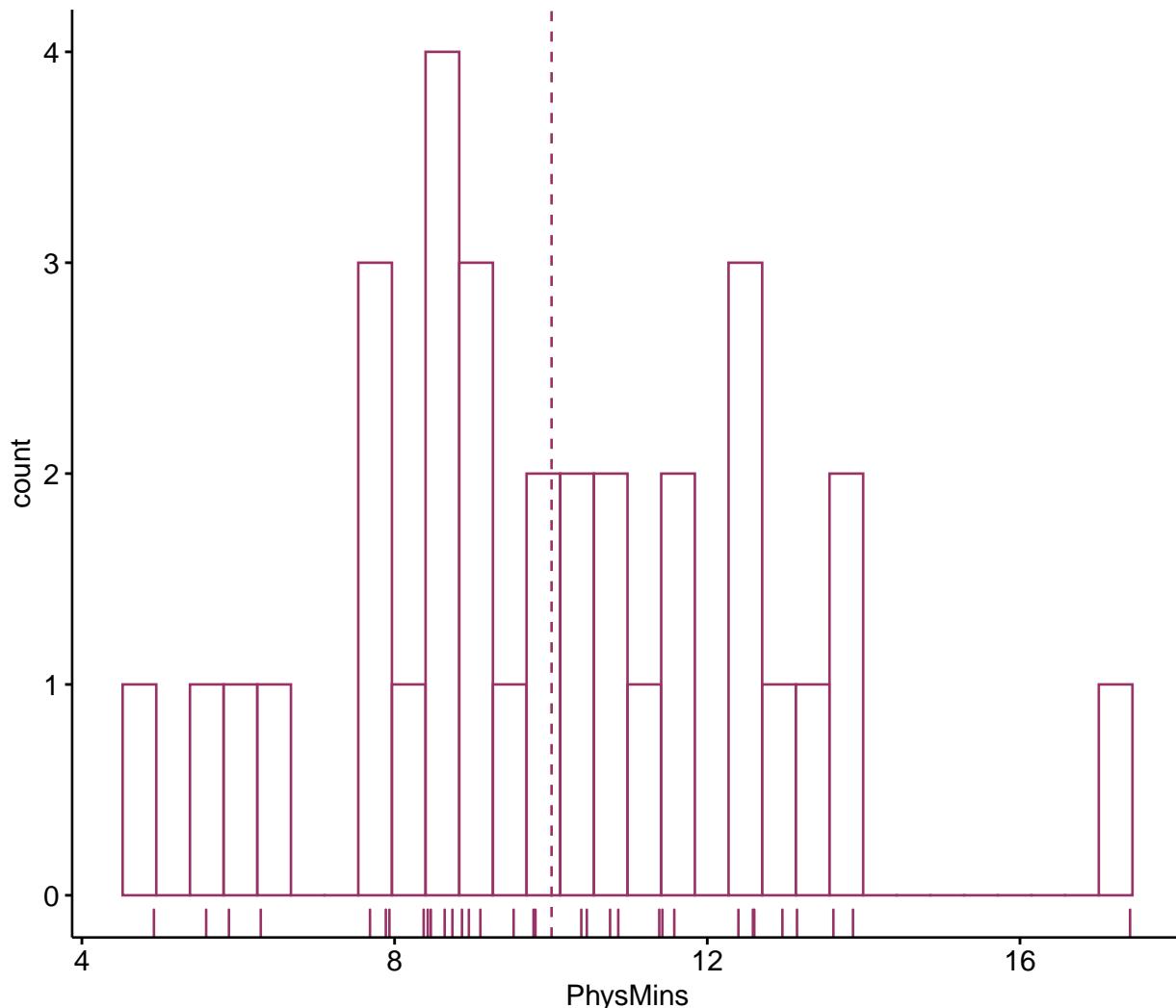
Plotting the data is best practice to any data analysis. The *ggpubr* package is one of my go-to-tools for quick and easy plots of data. Below, I have plotted the time-with-patient (Physician Seconds) variable and added the mean. As with most plotting packages, *ggpubr* will “bin” (or cluster) the data 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(): Ignoring `mapping` because `xintercept` was provided.

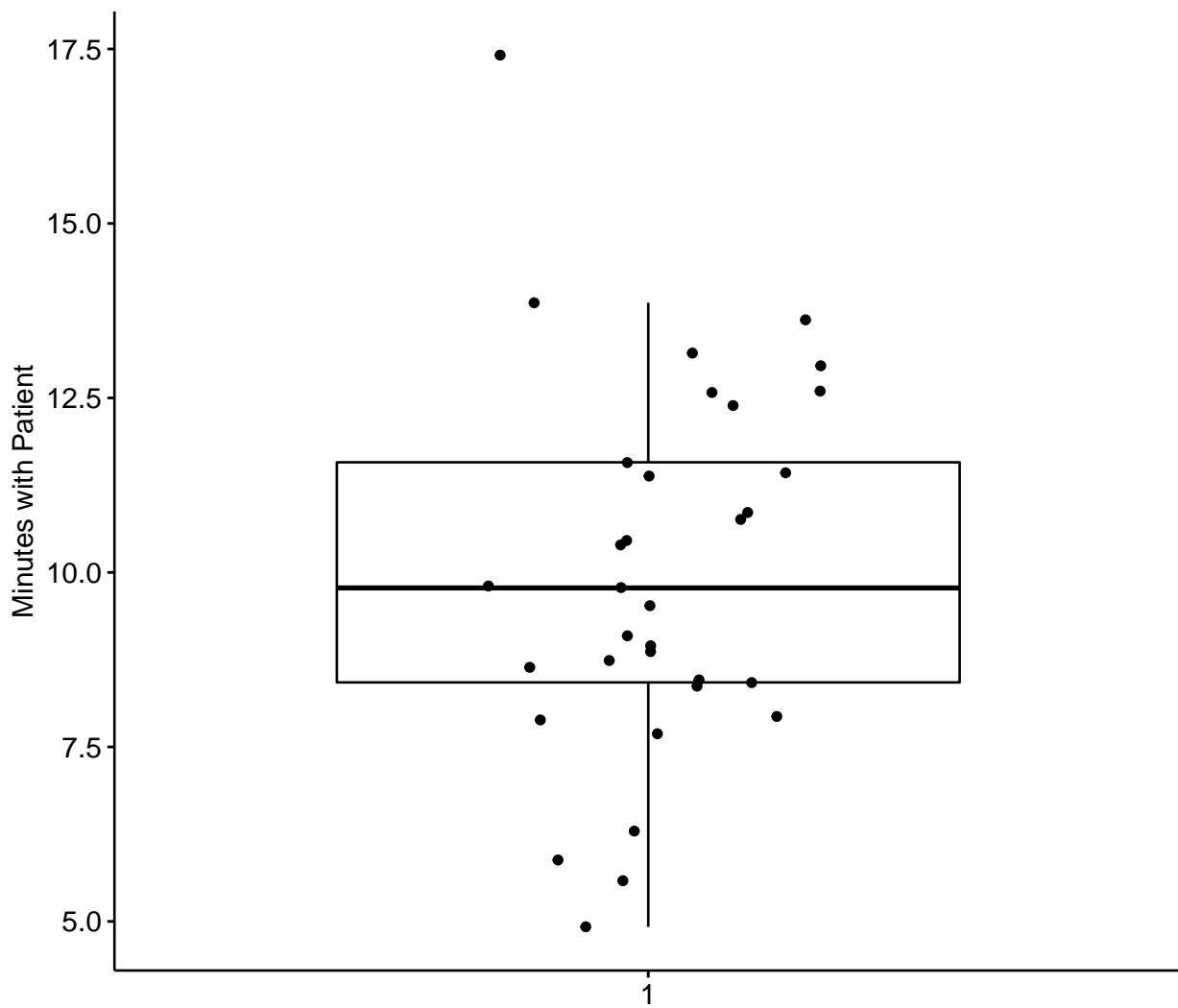
Warning: geom_vline(): Ignoring `data` because `xintercept` was provided.



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	0.04	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.

4.5.3 Hand-Calculations

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 denominator 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
```

4.5.3.1 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 of 32 (not rounded down to 30).

4.5.3.2 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 our true mean difference 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 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(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.

4.5.3.3 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 Computation in R

Calculating a one sample t -test is possible through base R and a number of packages. Navarro's [2020a] *lsr* package provides output that is commonly used in psychology.

```
lsr::oneSampleTTest(x = dfOneSample$PhysMins, mu = 1.23)
```

```
One sample t-test

Data variable: dfOneSample$PhysMins

Descriptive statistics:
  PhysMins
  mean      10.008
  std dev.   2.701

Hypotheses:
  null: population mean equals 1.23
  alternative: population mean not equal to 1.23

Test results:
  t-statistic: 18.672
  degrees of freedom: 32
  p-value: <.001

Other information:
  two-sided 95% confidence interval: [9.051, 10.966]
  estimated effect size (Cohen's d): 3.25
```

This well-organized output has everything we need for an APA style presentation of results. Identical to all the information we hand-calculated, we would write the t string this way: $t(32) = 11.68, p < .001, d = 3.25$. The *lsr* output also includes confidence intervals. These represent the 95% confidence interval of the true difference between the means. That is, we are 95% confident that the true mean of the minutes that physicians in our sample spent with patients falls between 9.05 and 10.97.

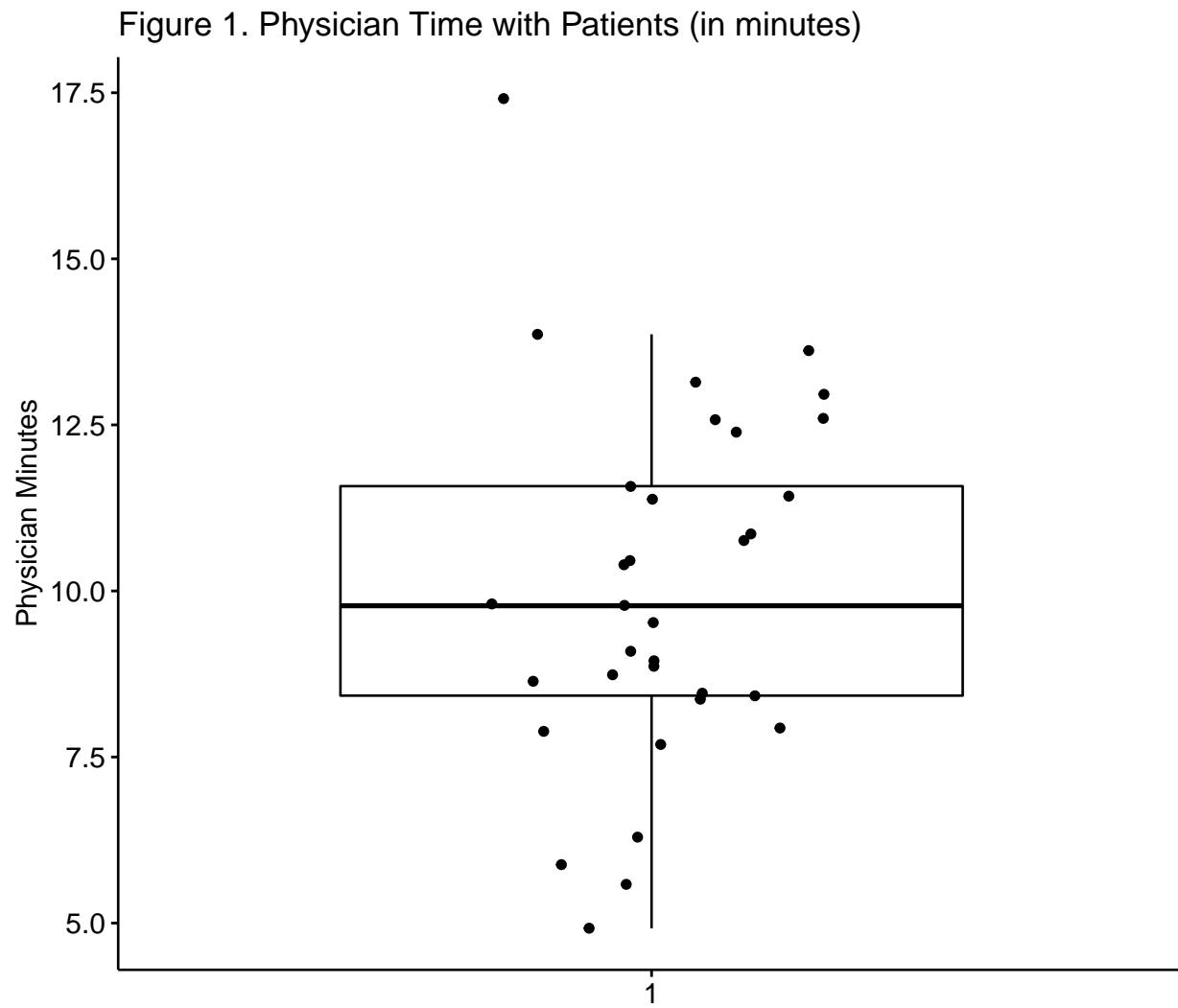
4.7 APA Style Results

Let's write up the results. I would probably choose to include the boxplot produced in the initial exploration of the data.

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, $t(32) = 18.672, p < .001.$, 95. 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::ggbboxplot(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 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 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 $(10.01 - 1.23)/2.7$
- *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 = (10.01 - 1.23)/2.7, n = 33, power = NULL, sig.level = 0.05,
                  type = "one.sample", alternative = "two.sided")
```

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

```
n = 33
d = 3.251852
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 = (10.01 - 1.23)/2.7, n = NULL, power = 0.8, sig.level = 0.05,
  type = "one.sample", alternative = "two.sided")
```

One-sample t test power calculation

```
n = 3.005993
d = 3.251852
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
```

```
lsr::oneSampleTTest(x = rdfOneSample$rPhysMins, mu = 1.23)
```

One sample t-test

Data variable: rdfOneSample\$rPhysMins

Descriptive statistics:

rPhysMins	
mean	9.636
std dev.	1.552

Hypotheses:

null:	population mean equals 1.23
alternative:	population mean not equal to 1.23

Test results:

t-statistic:	9.379
degrees of freedom:	2

p-value: 0.011

Other information:

two-sided 95% confidence interval: [5.78, 13.492]
estimated effect size (Cohen's d): 5.415

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: $t(2) = 9.379, p = 0.011, d = 5.415, 95$.

4.9 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).

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) another number. 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.

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 one sample t test (with an effect size)	5	_____
5. APA style results with table(s) and figure	5	_____
6. Explanation to grader	5	_____
Totals	30	_____

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 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 [2014b] 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(lsr)){install.packages('lsr')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(pwr)){install.packages('pwr')}
# if(!require(car)){install.packages('car')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(knitr)){install.packages('knitr')}
```

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.

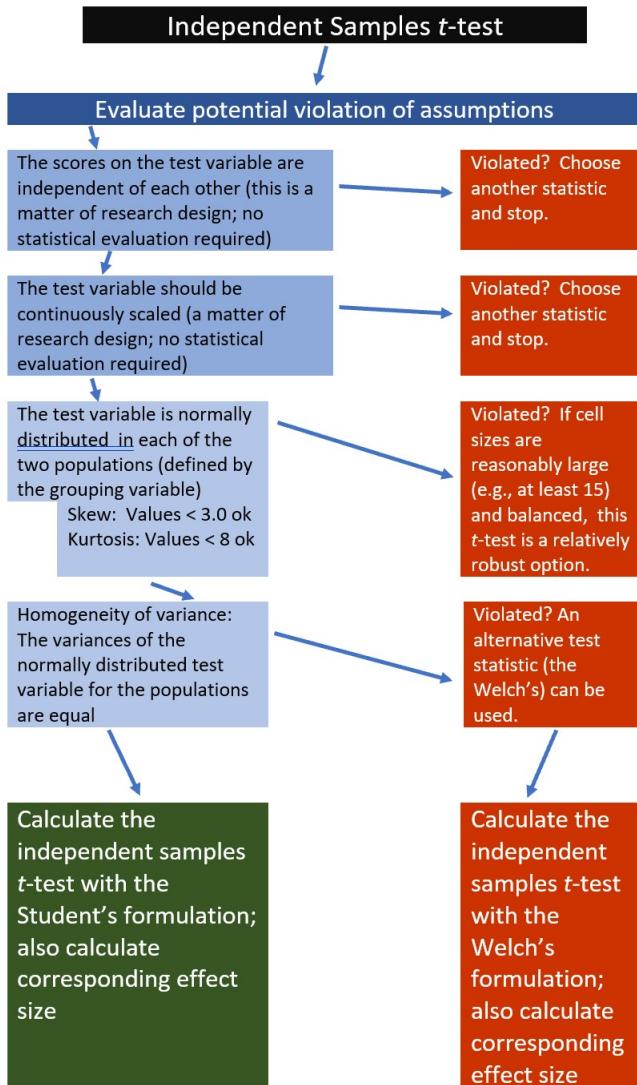
The comparison of two means is especially evident in the numerator of the formula. In the denominator we can see that the mean difference is adjusted by the standard error. At the outset, you should know that the formula in the denominator gets messy, but the formula, alone, provides an important conceptual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

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 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.

```
# 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 hashtagged 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)
```

5.4 Working the Problem

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.

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_A : \mu_1 &\neq \mu_2 \end{aligned}$$

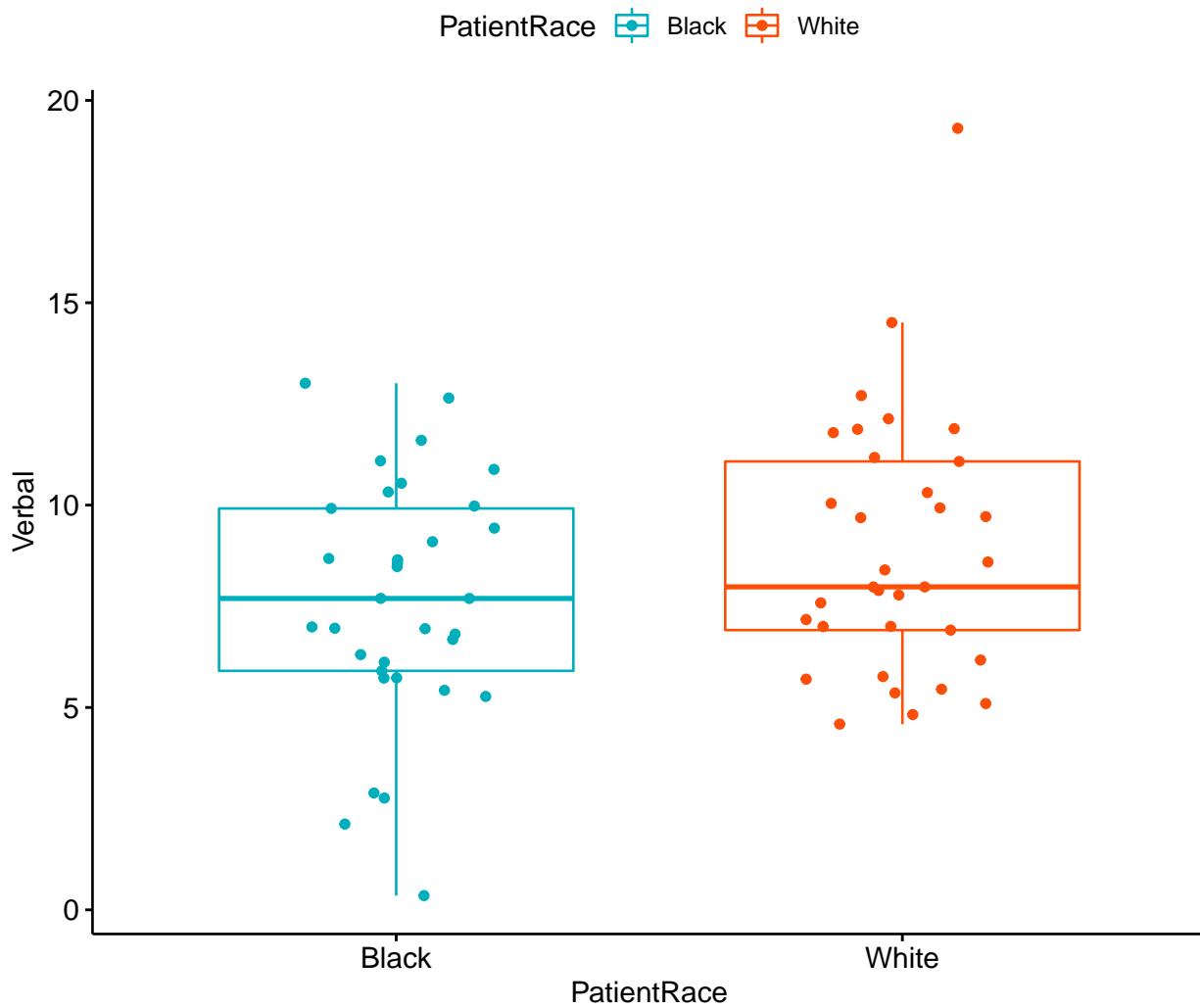
5.4.2 Preliminary Exploration

Plotting the data is a helpful early step in any data analysis. 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.

```
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.

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)
```

	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	0.43	1.21	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 (0.43) and kurtosis (1.21) fall below the areas of concern (below the absolute value of 3 for skew; below the absolute values of 8 for kurtosis) identified by Kline [2016].

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 values for each level 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., physID, 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, group = "PatientRace", mat = TRUE)
```

	item	group1	vars	n	mean	sd	median	trimmed		
ID*1		1	Black	1 33	17.000000 9.6695398	17.000000 9.6695398	17.000000 17.000000			
ID*2		2	White	1 33	50.000000 9.6695398	50.000000 9.6695398	50.000000 50.000000			
PatientRace*1		3	Black	2 33	1.000000 0.0000000	1.000000 0.0000000	1.000000 1.000000			
PatientRace*2		4	White	2 33	2.000000 0.0000000	2.000000 0.0000000	2.000000 2.000000			
Verbal1		5	Black	3 33	7.614884 2.9854116	7.693516 2.9854116	7.733412 7.693516			
Verbal2		6	White	3 33	8.891483 3.2032222	7.979546 3.2032222	8.606615 7.979546			
Nonverbal1		7	Black	4 33	2.943125 0.9251164	2.885724 0.9251164	2.931841 2.885724			
Nonverbal2		8	White	4 33	2.965472 0.7001442	2.936787 0.7001442	2.995131 2.936787			
					mad	min	max	range	skew	kurtosis
ID*1					11.8608000	1.0000000	33.000000	32.000000	0.0000000	-1.30951223
ID*2					11.8608000	34.0000000	66.000000	32.000000	0.0000000	-1.30951223
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	-0.3537887	-0.30860583
Verbal2					3.2861809	4.5891699	19.311207	14.722037	1.0170842	1.26737896
Nonverbal1					0.9185825	0.8333731	5.000000	4.166627	0.1150450	-0.04929788
Nonverbal2					0.5560620	1.1311619	4.350886	3.219724	-0.4143090	0.19115265
					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 -.35 and -.31, respectively. They were 1.02 and 1.27 for White patients. As before, these fall well below the absolute values of 3 (skew) and 8 (kurtosis) that are considered to be concerning.

One of the assumptions of the independent samples t -test is that the variances of the dependent variable are similar for both levels of the grouping 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."

In R, Levene's test is found in the *car* package.

```
car::leveneTest(Verbal ~ PatientRace, dfIndSamples, center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
Df F value Pr(>F)
group 1 0.1422 0.7074
64
```

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 0.142
- We want our p value to be $> .05$

Happily, our Levene's result is ($F[1, 64] = 0.142, p = .707$) not significant. Because p is greater than 0.05, we have not violated the homogeneity of variance assumption. That is to say, the variance in each of the patient race groups is not statistically significantly different. we can use the regular (Student's) formulation of the t test for independent samples.

5.4.3 Hand-Calculations

Earlier I presented a formula for the independent samples t -test.

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

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?”

5.4.3.1 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 linked table, 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.

5.4.3.2 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.

5.4.3.3 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 Computation in R

Navarro's *lsr* package makes the computation of the independent *t*-test easy and produces output that is commonly used in psychological science.

```
lsr::independentSamplesTTest(formula = Verbal ~ PatientRace, data = dfIndSamples,
  var.equal = TRUE)
```

Student's independent samples t-test

Outcome variable: Verbal
Grouping variable: PatientRace

Descriptive statistics:
Black White
mean 7.615 8.891
std dev. 2.985 3.203

Hypotheses:
null: population means equal for both groups
alternative: different population means in each group

Test results:
t-statistic: -1.675

```
degrees of freedom: 64
p-value: 0.099
```

Other information:

```
two-sided 95% confidence interval: [-2.799, 0.246]
estimated effect size (Cohen's d): 0.412
```

This well-organized output has everything we need for an APA style presentation of results. Identical to all the information we hand-calculated, we would write the *t* string this way: $t(64) = -1.675, p = .099, d = 0.412$. The *lsr* output also includes confidence intervals. These represent the 95% confidence interval of the true difference between the means. That is, we are 95% confident that the true difference between means could be as large as -2.799 or as (small/medium/large) as 0.246. 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.

5.5.1 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. Navarro's *lsr* 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.

```
lsr::independentSamplesTTest(formula = Verbal ~ PatientRace, data = dfIndSamples,
  var.equal = FALSE)
```

Welch's independent samples t-test

Outcome variable: Verbal
Grouping variable: PatientRace

Descriptive statistics:

	Black	White
mean	7.615	8.891
std dev.	2.985	3.203

Hypotheses:

```
null: population means equal for both groups
alternative: different population means in each group
```

Test results:

```
t-statistic: -1.675
degrees of freedom: 63.685
p-value: 0.099
```

Other information:

two-sided 95% confidence interval: [-2.799, 0.246]
 estimated effect size (Cohen's d): 0.412

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

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). Although the independent samples *t*-test was nonsignificant, $t(64) = -1.675$, $p = .099$, the effect size ($d = 0.412$) was somewhat moderate in size. 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)
```

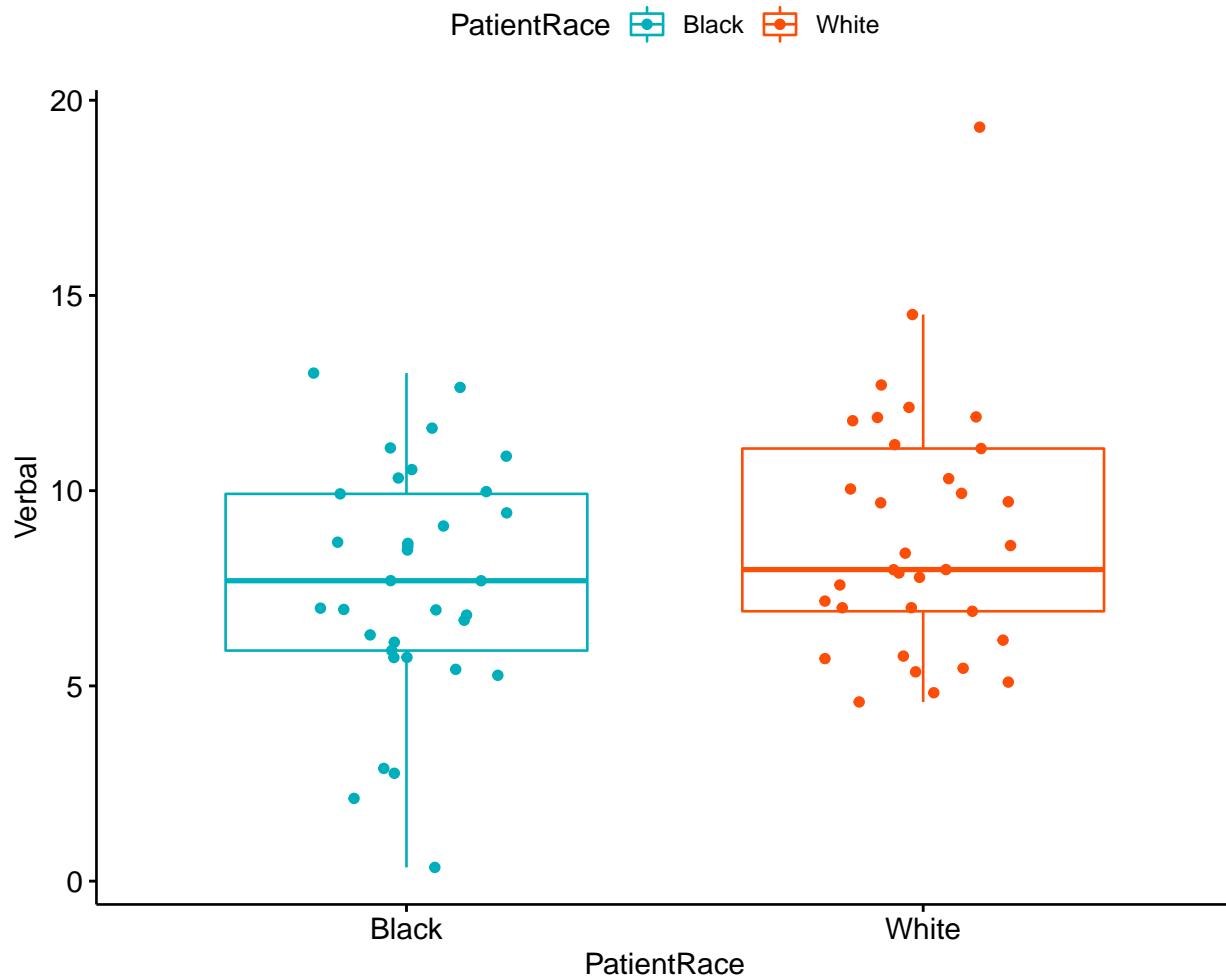
Descriptive statistics for Verbal as a function of PatientRace.

PatientRace	M	SD
Black	7.61	2.99
White	8.89	3.20

Note. M and SD represent mean and standard deviation, respectively.

```
ggpubr::ggboxplot(dfIndSamples, x = "PatientRace", y = "Verbal", color = "PatientRace",
  palette = c("#00AFBB", "#FC4E07"), add = "jitter", title = "Figure 1. Physician Verbal Eng
```

Figure 1. Physician Verbal Engagement as a Function of Patient Race



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 add the final number or the elements required (and have R do the math for us)

- *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 0.99. That is, given the value of the mean difference (1.276) we had a 99% chance of detecting a statistically significant effect if there was one.

```
pwr::pwr.t.test(d = (7.615 - 8.891)/0.762, n = 33, power = NULL, sig.level = 0.05,
  type = "two.sample", alternative = "two.sided")
```

Two-sample t test power calculation

```
n = 33
d = 1.674541
sig.level = 0.05
power = 0.9999989
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 7 individuals (per group; 14 total) for the *p* value to be < .05.

```
pwr::pwr.t.test(d = ((7.615 - 8.891)/0.762), n = NULL, power = 0.8, sig.level = 0.05,
  type = "two.sample", alternative = "two.sided")
```

Two-sample t test power calculation

```
n = 6.709177
d = 1.674541
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 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(220821)
# sample size, M, and SD for Black then White patients
rVerbal <- c(rnorm(7, mean = 8.37, sd = 3.36), rnorm(7, 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(7, mean = 2.68, sd = 0.84), rnorm(7, mean = 2.93,
sd = 0.77))
# set upper bound
rNonverbal[rNonverbal > 5] <- 5
# set lower bound
rNonverbal[rNonverbal < 0] <- 0

rID <- factor(seq(1, 14))
# name factors and identify how many in each group; should be in same
# order as first row of script
rPatientRace <- c(rep("Black", 7), rep("White", 7))
# 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"))

lsr::independentSamplesTTest(formula = rVerbal ~ rPatientRace, data = rdfIndSamples,
var.equal = TRUE)
```

Student's independent samples t-test

Outcome variable: rVerbal
 Grouping variable: rPatientRace

Descriptive statistics:

	Black	White
mean	8.821	8.608
std dev.	3.645	4.718

Hypotheses:

null: population means equal for both groups
 alternative: different population means in each group

Test results:

```
t-statistic: 0.095
degrees of freedom: 12
p-value: 0.926
```

Other information:

```
two-sided 95% confidence interval: [-4.696, 5.123]
estimated effect size (Cohen's d): 0.051
```

Not surprisingly, this did not result in a statistically significant result: $t(12) = 0.095, p = 0.9267, d = 0.051, CI_{95}$. 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 “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).

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.

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 Explanation to grader	5	_____
Totals	30	_____

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.

- 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 [2014b] 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.

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(lsrr)){install.packages('lsr')}
```

```
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(pwr)){install.packages('pwr')}
# if(!require(apaTables)){install.packages('apaTables')}
# if(!require(knitr)){install.packages('knitr')}
```

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.

Condition A	Condition B	In the formula for the paired samples <i>t</i> test we see a \bar{D} in the numerator. This represents the <i>difference</i> 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**.

6.3 Workflow for Paired Samples *t*-test

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

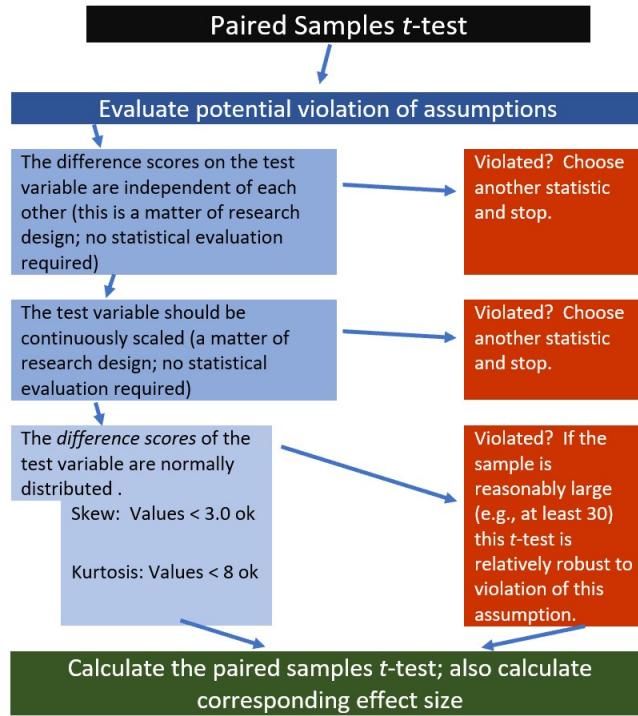


Figure 6.2: A colorful image of a workflow for the paired samples t test

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).

6.4 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). 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(<i>SD</i>)	Mean(<i>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

The primary analysis utilized by Elliott and colleagues [2016] was the paired samples *t*-test. We will replicate that exact analysis with simulated data.

6.4.1 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)
# 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')
```

6.5 Working the Problem

6.5.1 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.

$$\begin{aligned} H_0 : \mu_D &= 0 \\ H_A : \mu_D &\neq 0 \end{aligned}$$

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

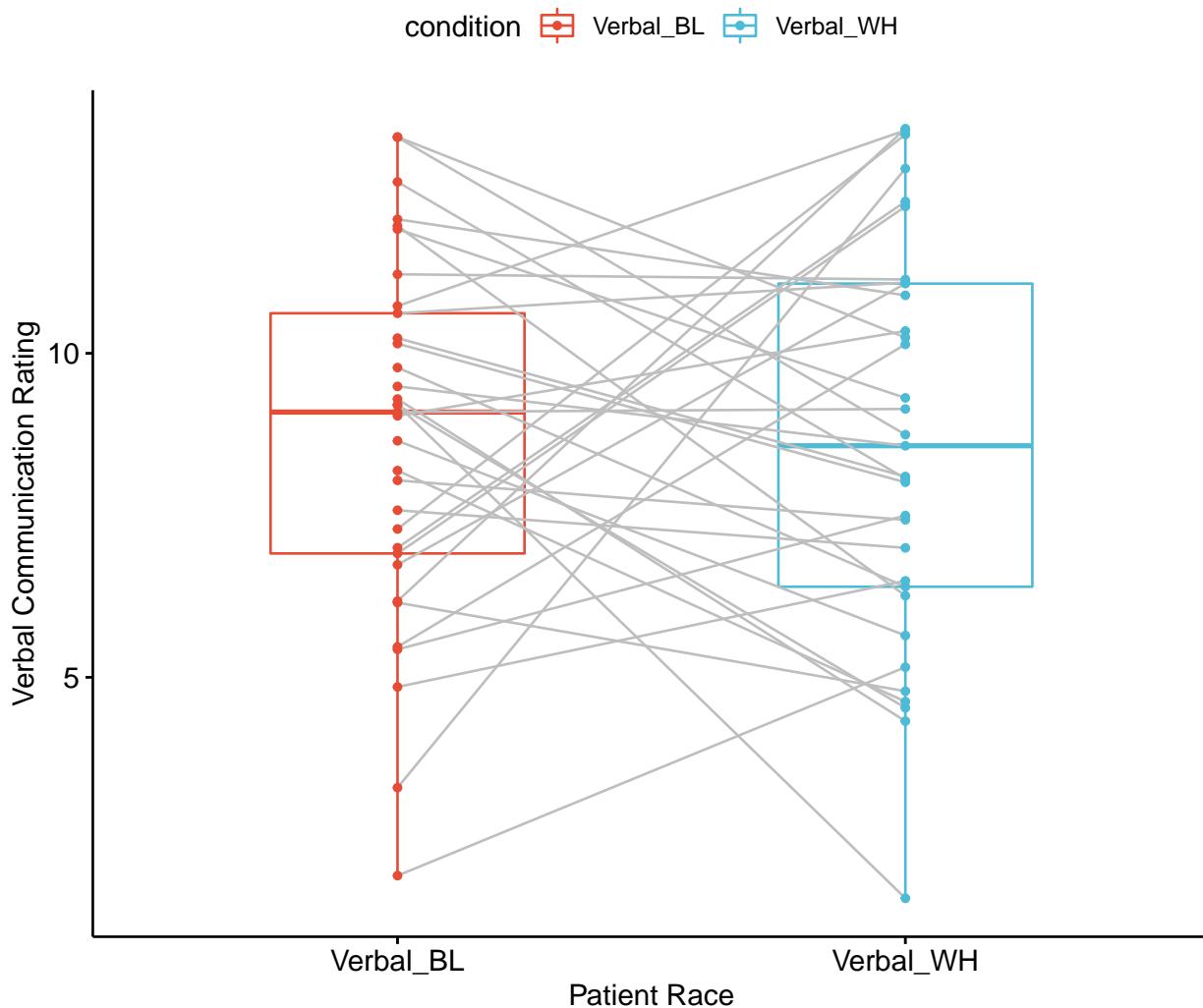
```

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.

6.5.2 Preliminary Exploration

Let's plot the data. 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.

```
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.

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

```
psych::describe(dfPairedSamples)
```

	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	-0.33
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	-0.36
NVerb_WH	5	33	2.89	0.85	2.94	2.87	0.64	1.34	5.02	3.69	0.24
Verbal_D	6	33	0.08	4.14	0.61	0.27	4.11	-9.55	7.61	17.17	-0.41
			kurtosis	se							
PhysID			-1.31	1.68							

Verbal_BL	-0.47	0.49
Verbal_WH	-0.89	0.54
NVerb_BL	-0.85	0.17
NVerb_WH	-0.19	0.15
Verbal_D	-0.69	0.72

From this, we can see the statistics for all of our variables, including the difference variable. Focused on the verbal communication scores, we see that the means are slightly higher for patients who are Black ($M = 8.70$, $SD = 2.80$), than patients who are White ($M = 8.62$, $SD = 3.08$). Further, the skew and kurtosis values are well below the areas of concern (below the absolute value of 3 for skew; below the absolute values of 8 for kurtosis) identified by Kline [2016].

Recall, though that the normality assumption for the paired samples t -test concerns the difference score. We see that the mean difference is .08($SD = 4.14$). Its skew (-0.41) and kurtosis (-0.69) are also well-below the thresholds of concern.

6.5.3 Hand Calculations

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?”

6.5.3.1 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.

6.5.3.2 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.

6.5.3.3 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 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 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.

6.6 Computation in R

Navarro's *lsr* package makes the computation of the paired samples t -test easy and produces output that is commonly used in psychological science.

```
lsr::pairedSamplesTTest(formula = ~Verbal_BL + Verbal_WH, data = dfPairedSamples)
```

Paired samples t-test

Variables: Verbal_BL , Verbal_WH

Descriptive statistics:

	Verbal_BL	Verbal_WH	difference
mean	8.698	8.617	0.081
std dev.	2.795	3.083	4.139

Hypotheses:

null: population means equal for both measurements
 alternative: different population means for each measurement

Test results:

t-statistic: 0.113
 degrees of freedom: 32
 p-value: 0.911

Other information:

two-sided 95% confidence interval: [-1.386, 1.549]
 estimated effect size (Cohen's d): 0.02

This well-organized output has everything we need for an APA style presentation of results. Identical to all the information we hand-calculated, we would write the *t* string this way: $t(32) = 0.113$, $p = .911$, $d = 0.02$. The *lsr* output also includes confidence intervals. These represent the 95% confidence interval of the true difference between the means. That is, we are 95% confident that the true difference between 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.

6.7 APA Style Results

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
PairedDescripts <- 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(PairedDescripts, 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.

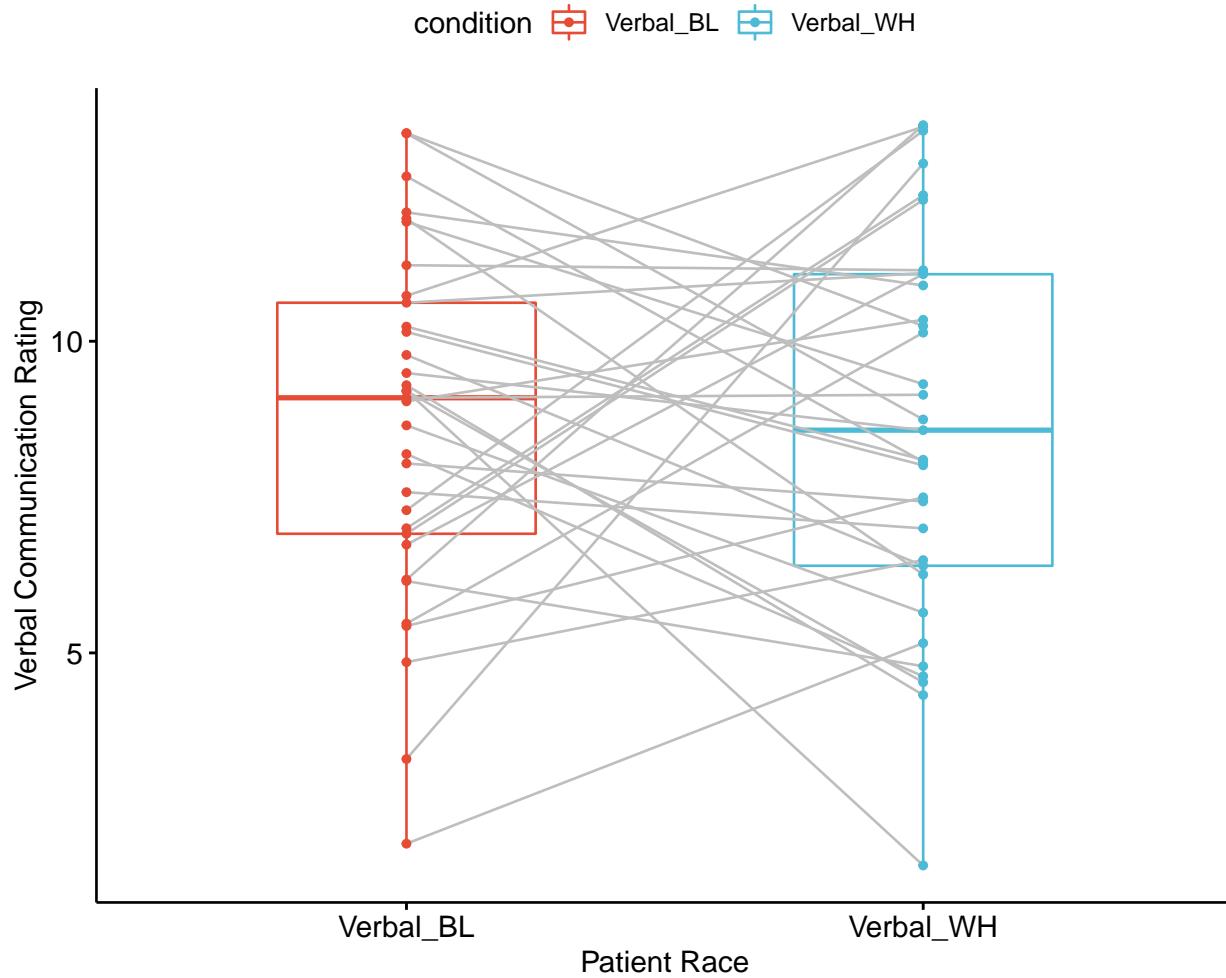
Values in square brackets indicate the 95% confidence interval.

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

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

```
ggbetween::ggpaired(dfPairedSamples, cond1 = "Verbal_BL", cond2 = "Verbal_WH",
  color = "condition", line.color = "gray", palette = c("npg"), xlab = "Patient Race",
  ylab = "Verbal Communication Rating", title = "Figure 1. Physician Verbal Engagement as a Function of Patient Race and Condition")
```

Figure 1. Physician Verbal Engagement as a Function of Patient Race



6.8 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; using data from our vignette it would be $(0-.08/4.14)$

- *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 = (0.08)/4.14, n = 33, power = NULL, sig.level = 0.05,
  type = "paired", alternative = "two.sided")
```

Paired t test power calculation

```
n = 33
d = 0.01932367
sig.level = 0.05
power = 0.05133016
alternative = two.sided
```

NOTE: *n* is number of *pairs*

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 21,022 individuals for the *p* value to be $< .05$.

```
pwr::pwr.t.test(d = (0.08)/4.14, n = NULL, power = 0.8, sig.level = 0.05,
  type = "paired", alternative = "two.sided")
```

Paired t test power calculation

```
n = 21021.66
d = 0.01932367
sig.level = 0.05
power = 0.8
alternative = two.sided
```

NOTE: *n* is number of *pairs*

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 <- 21022
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"))
```

Now I will conduct the paired samples *t* test.

```
lsr::pairedSamplesTTest(formula = ~Verbal_BL + Verbal_WH, data = paired_V2)
```

Paired samples t-test

Variables: Verbal_BL , Verbal_WH

Descriptive statistics:

	Verbal_BL	Verbal_WH	difference
mean	8.367	8.419	-0.052
std dev.	3.359	3.213	3.867

Hypotheses:

null: population means equal for both measurements
 alternative: different population means for each measurement

Test results:

t-statistic: -1.936
 degrees of freedom: 21021
 p-value: 0.053

Other information:

two-sided 95% confidence interval: [-0.104, 0.001]
 estimated effect size (Cohen's d): 0.013

The new results is much closer to statistical significance, but, wisely, remains non-significant: $t(21021) = -1.936, p = .053, d = 0.013$.

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.

6.9 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).

6.9.1 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 (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.

6.9.2 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.

6.9.3 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.

6.9.4 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.

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 Explanation to grader	5	_____
Totals	30	_____

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.

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.

- 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 [2014b], Field's [2012], Navarro's [2020b], and other techniques I have found on the internet and learned from my students.
- 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.
- 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(gplots)){install.packages('gplots')}
# creating new variables and other handy functions
# 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(rstatix)){install.packages('rstatix')} produces effect
# sizes if(!require(lsrr)){install.packages('lsrr')} estimating sample
# sizes and power analysis if(!require(pwr)){install.packages('pwr')}
# produces an APA style table for ANOVAs and other models
# if(!require(apaTables)){install.packages('apaTables')} helps with
# formats like decimals and percentages for inline code
# if(!require(formattable)){install.packages('formattable')} more
# effect size options
# if(!require(effectsize)){install.packages('effectsize')}
```

7.2 Workflow for One-Way ANOVA

The following is a proposed workflow for conducting a one-way ANOVA.

1. Prepare (upload) data.
2. Explore data
 - graphs
 - descriptive statistics
3. Checking distributional assumptions
 - assessing normality via skew, kurtosis, Shapiro Wilks
 - 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)

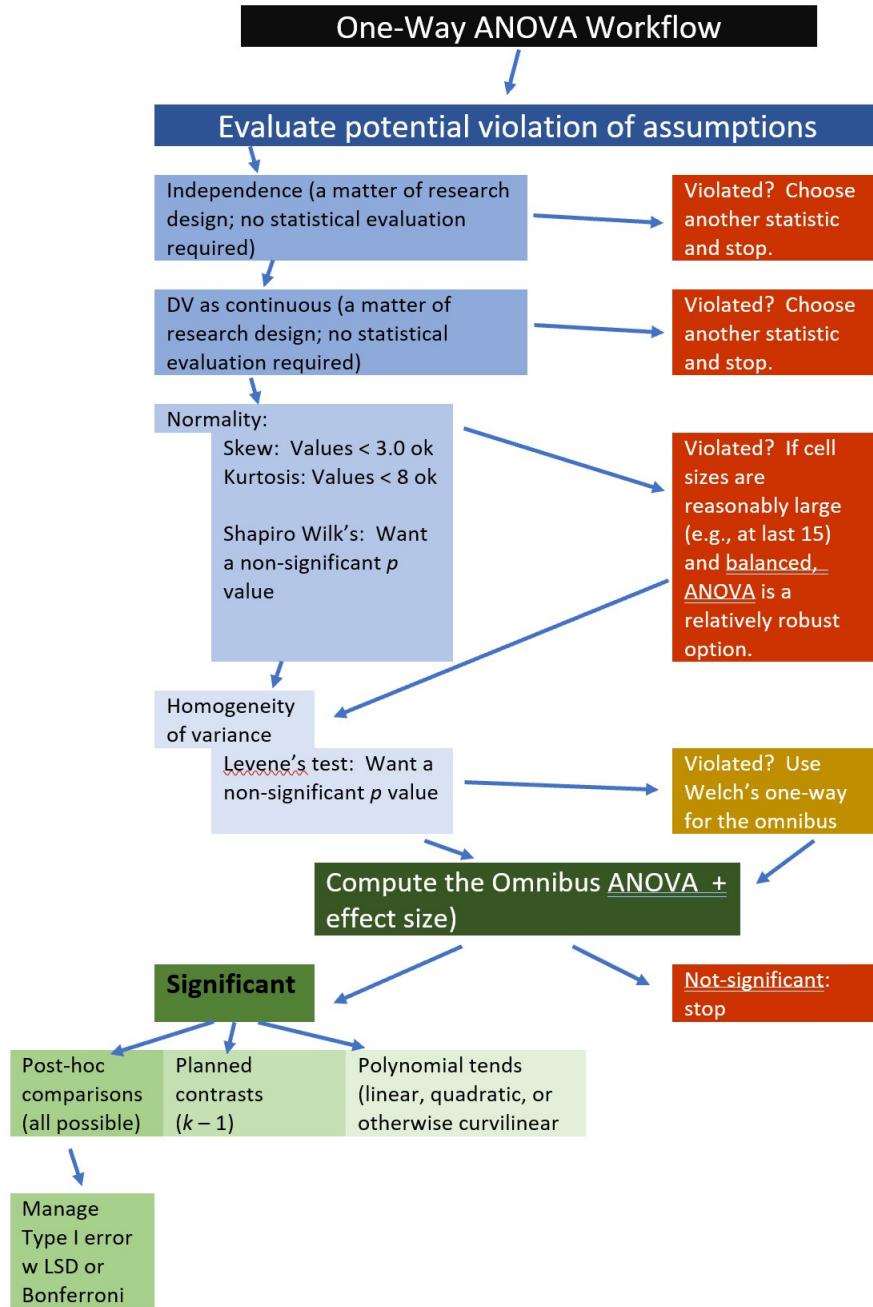


Figure 7.1: A colorful image of a workflow for the one-way ANOVA

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

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](#).

- **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 (rnorm), 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)

```

7.4 Working the Problem

7.4.1 Preparing the Data

Examining the data is important for several reasons. First, we can begin to inspect for any anomalies. Second, if we are confused about what statistic we wish to apply, understanding the characteristics of the data can provide clues.

In R markdown we can

- look at the data by clicking on it, and
- examine its structure with the `str()` function.

Let's do both.

```
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.

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)
```

```
# a quick demo to show that the .csv format loses the variable
# formatting str(acc_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(accSIM30, 'accSIM.rds') bring back the simulated dat
# from an .rds file acc_RDS <- readRDS('accSIM.rds')
```

```
# 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
```

7.4.2 Exploring the Distributional Characteristics Numerically

We will explore the data such that you will have several tools for future exploration. In this first demonstration I will quickly produce a mean and standard deviation.

These functions are in 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 to 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)
```

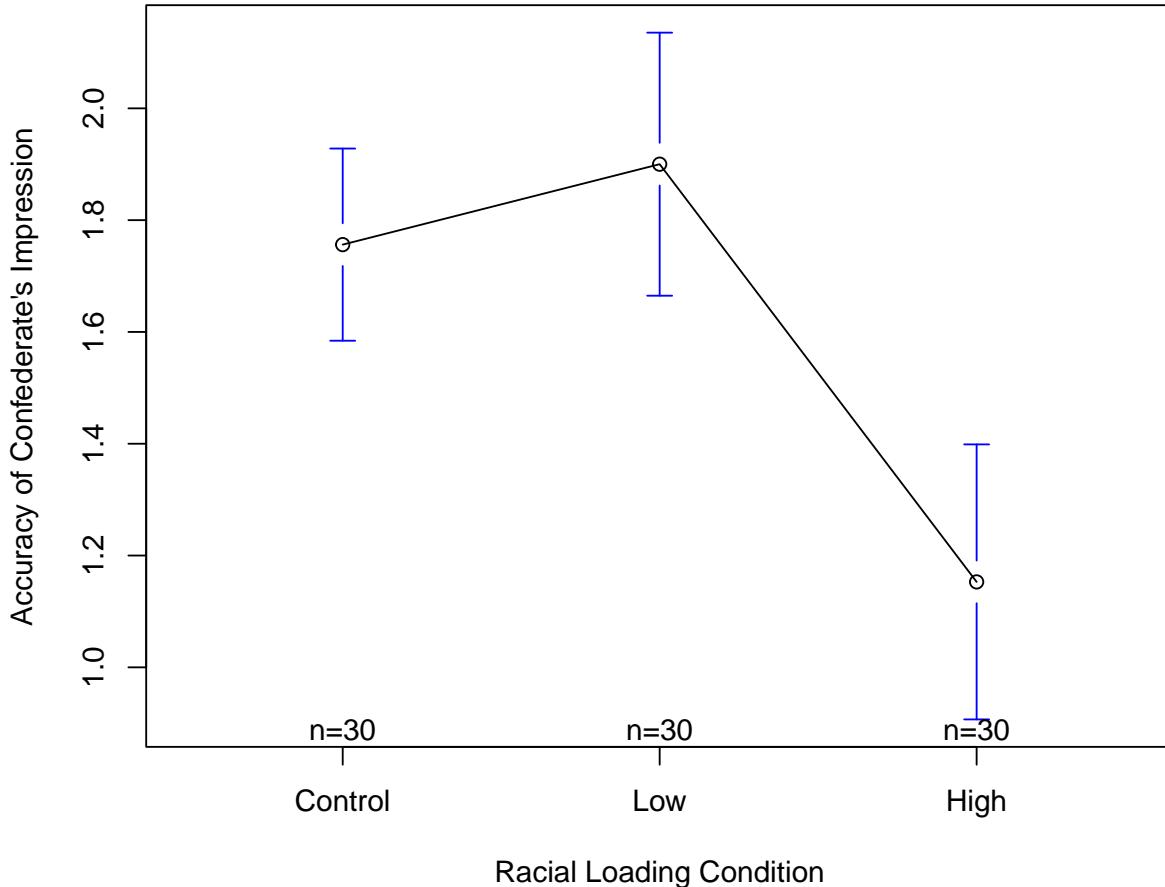
```
COND Accurate
1 Control 0.4603964
2 Low 0.6301138
3 High 0.6587486
```

Before looking at graphs, we can see that racially loaded *high* condition has the lowest accuracy score and the largest variability. Let's produce some helpful figures so that we can visualize this.

7.4.3 Exploring the Distributional Characteristics Graphically

The package *gplots* produces a simple line graph and the script is fairly intuitive. The *plotmeans()* function plots the means with error bars (95% confidence intervals) around the mean. Regarding the confidence intervals, we can think, “How confident are we that the mean is this particular value?” Earlier we noted that the “high racial loading condition” had the lowest mean and the widest variability. Is this apparent from the graph?

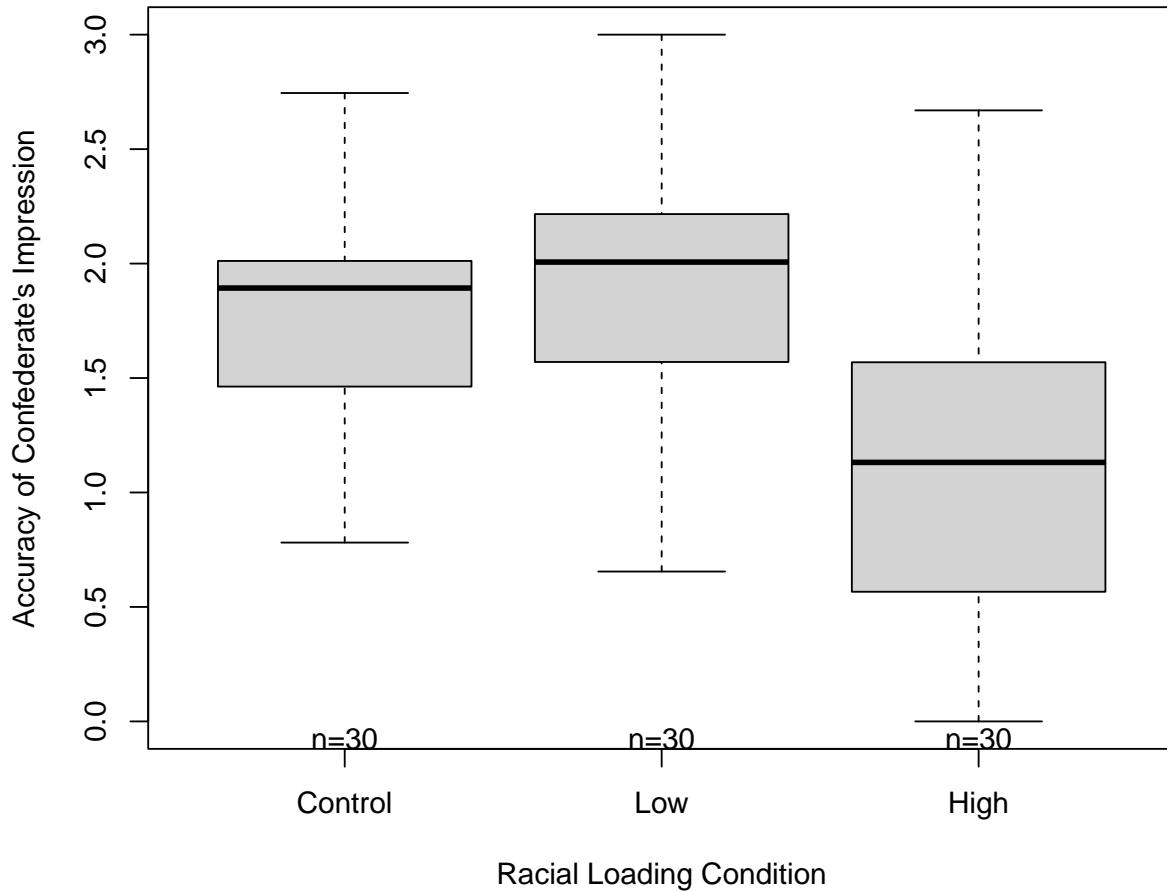
```
# plots DV by IV
gplots::plotmeans(formula = Accurate ~ COND, data = accSIM30, xlab = "Racial Loading Condition",
                  ylab = "Accuracy of Confederate's Impression", n.label = TRUE)
```



```
# this code could be more elegantly written in one row
# (formula = Accurate ~ COND, data = accSIM30, xlab = 'Racial Loading
# Condition', ylab = 'Accuracy of Confederate's Impression', n.label
# = TRUE)
```

Boxplots, with the *boxplot2()* function provide another view of our data. 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.

```
gplots::boxplot2(Accurate ~ COND, data = accSIM30, xlab = "Racial Loading Condition",
                  ylab = "Accuracy of Confederate's Impression", n.label = TRUE)
```



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.

7.5 Understanding ANOVA with *Hand Calculations*

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.*

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.

7.5.1 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*.

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

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.

$$SS_T = SS_B + SS_W$$

Between-group variation
(i.e., differences among group means)

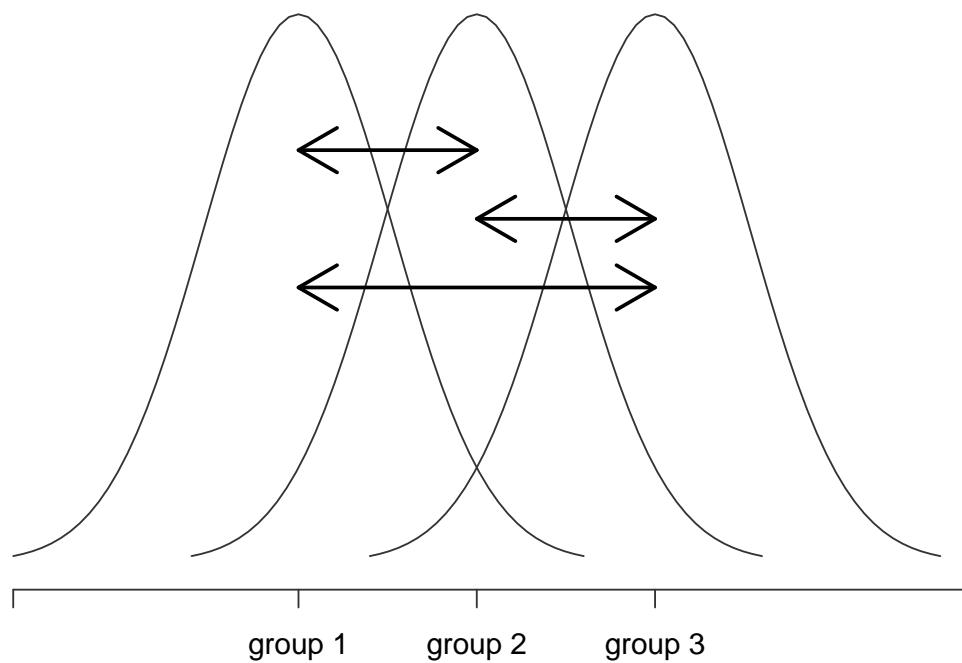


Figure 7.2: Graphical illustration of “between groups” variation

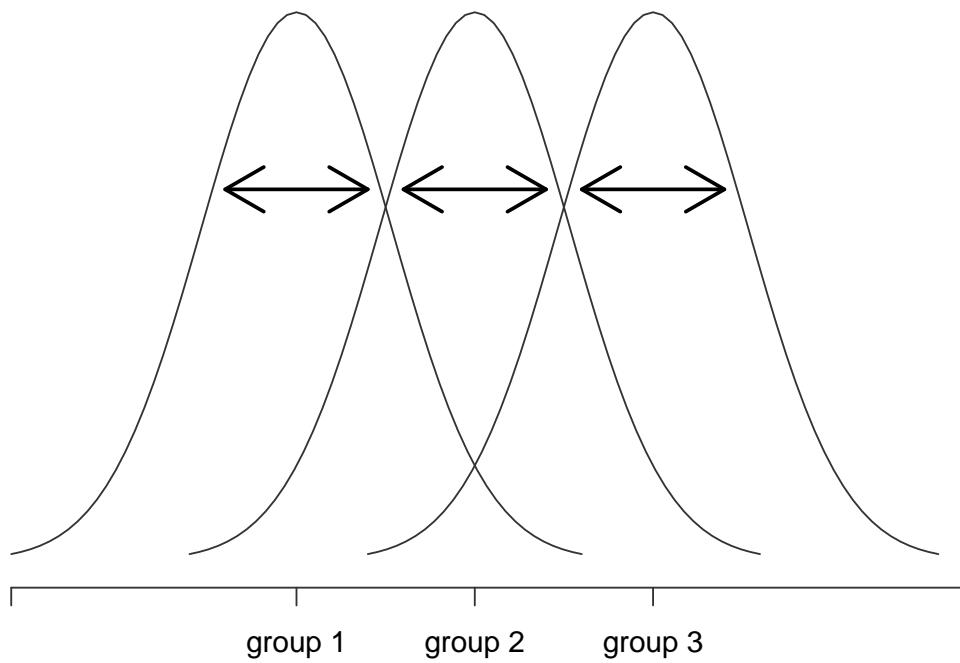


Figure 7.3: Graphical illustration of “within groups” variation

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      m_dev      m_devSQ
1  1 High 0.4203896 -0.6398265 -1.18265259 1.398667144
2  2 High 1.1226505 -2.0000000 -0.48039170 0.230776185
3  3 High 0.8852238 -0.2497750 -0.71781837 0.515263216
4  4 High 1.5689439  0.1455637 -0.03409829 0.001162694
5  5 High 1.8307196 -0.9960413  0.22767748 0.051837034
6  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.

7.5.2 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.

```
GroupMeans <- aggregate(Accurate ~ COND, accSIM30, mean)
GroupMeans
```

COND	Accurate
1 Control	1.756195
2 Low	1.900116
3 High	1.152815

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] 9.432
```

```
# 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.

7.5.3 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

7.5.3.1 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] 0.212
```

```
VARs <- aggregate (Accurate ~ COND, accSIM30, var)
VARs
```

	COND	Accurate
1	Control	0.2119648
2	Low	0.3970434
3	High	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] 30.246
```

```
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 $89 - 2 = 87$

7.5.4 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] 39.678

```
#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.

7.5.5 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] 4.716

```
#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 30.246
- df_R was 87
- Therefore, MS_R is:

```
#mean squares for the residual
#calculated with object names
MSR <- SSR/ dfR
MSR
```

[1] 0.348

```
#calculated with actual values
30.247/87
```

[1] 0.3476667

```
#Both result in the same
```

7.5.6 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] 13.566

```
#calculated with actual values
#Both result in the same
4.716/0.348
```

[1] 13.55172

7.5.7 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 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)	30.246(87)
MS	NA	4.716	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(.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.

7.6 Working the One-Way ANOVA in R

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

7.6.1 Evaluating the Statistical Assumptions

All statistical tests have some assumptions about the data. The one-way ANOVA has four assumptions:

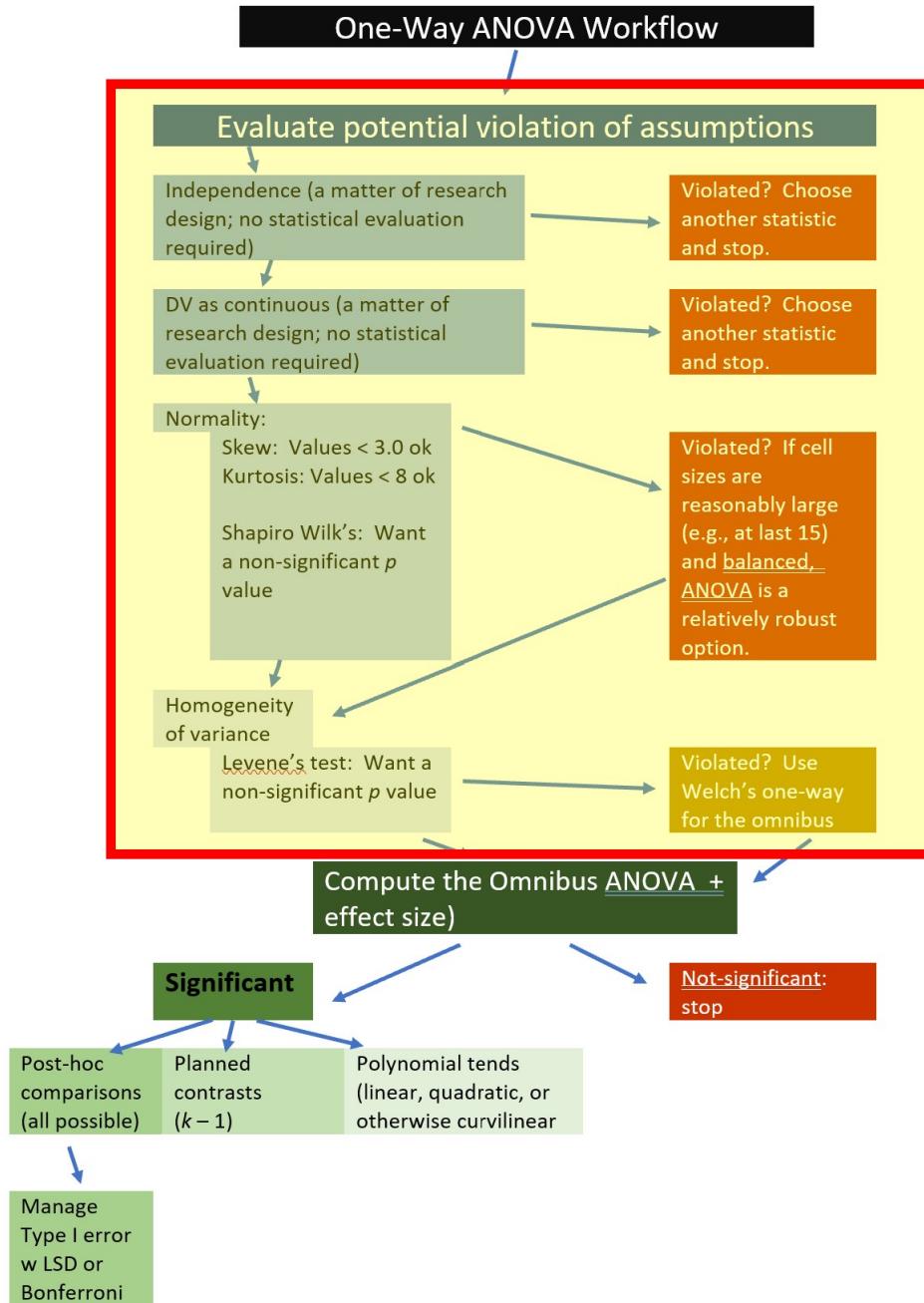


Figure 7.4: An image of the workflow for one-way ANOVA, showing that we are at the beginning: evaluating the potential violation of the assumptions.

- 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
 - conduct a Shapiro Wilks test
 - examine 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.

7.6.1.1 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)
```

	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	-0.28	-0.48	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).

```
# 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)
# 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	-0.275	-0.537	0.084							
Accurate2	2.345	-0.378	-0.465	0.115							
Accurate3	2.669	0.208	-0.690	0.120							

```
# optional to write it to a .csv file
write.csv(des.mat, file = "Table1.csv")
```

Skew and kurtosis speaks to normal distributions. The skew and kurtosis indices in the *psych* package are reported as *z* scores. Regarding skew, values greater than 3.0 are generally considered “severely skewed.” Regarding kurtosis, “severely kurtotic” is argued to be anywhere greater 8 to 20 [Kline, 2016].

The *Shapiro-Wilks* test evaluates the hypothesis that the distribution of the data as a whole 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 has a wrapper that can conduct the Shapiro-Wilks test for us.

```
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 (Control, $W = 0.954$, $p = 0.215$; Low, $W = 0.944$, $p = 0.115$; High, $W = 0.980$, $p = 0.831$).

There are limitations to the Shapiro-Wilks test. As the dataset being evaluated gets larger, the Shapiro-Wilks 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 [2014b] 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).

7.6.1.2 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).

In R, Levene's test is found in the *car* package.

```
# Our set up is similar: Accurate by condition, followed by the
# object that holds the data frame, followed by the instruction to
# center the analysis around the mean
levene <- car::leveneTest(Accurate ~ COND, accSIM30, center = mean)
levene
```

```
Levene's Test for Homogeneity of Variance (center = mean)
    Df F value Pr(>F)
group  2  1.5315  0.222
      87
```

We write the result of the Levene's as $F(2,87) = 1.532$, $p = 0.222$. Because $p > .05$, we know that the result is nonsignificant – that the variances of the three groups are not statistically significantly different than 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.

7.6.1.3 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:

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. Results of Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption ($F[2,87] = 1.532, p = 0.222$). Similarly, results of the Shapiro Wilk's test indicated no violation of the normality assumption in each of the cells (control, $W = 0.954, p = 0.215$; low, $W = 0.944, p = 0.115$; high, $W = 0.980, p = 0.831$).

Now we can move onto computing the omnibus ANOVA. *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 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.

7.6.2 Computing the Omnibus ANOVA

Having met all the assumptions, we are now ready to calculate the omnibus F test.

ANOVA is a special case of the general linear model (regression is a “not so special case” of the general linear model), therefore we use the linear model function, *aov()* to run the analysis.

In the code below, we predict Accuracy from COND (3 levels: control, low, high).

By assigning the results of the *aov()* function to an object (omnibus) we can then use that object (think *model*) in other functions to get details about our analysis.

```
# the script looks familiar, 'Accurate by Condition' DV ~ IV I say,
# 'DV by IV'
omnibus <- aov(Accurate ~ COND, data = accSIM30)
# prints the ANOVA output
summary(omnibus)
```

	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

Inserting the *aov()* object (omnibus) into the summary command produces the ANOVA Source Table that we manually created above.

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 $p = 0.000$.

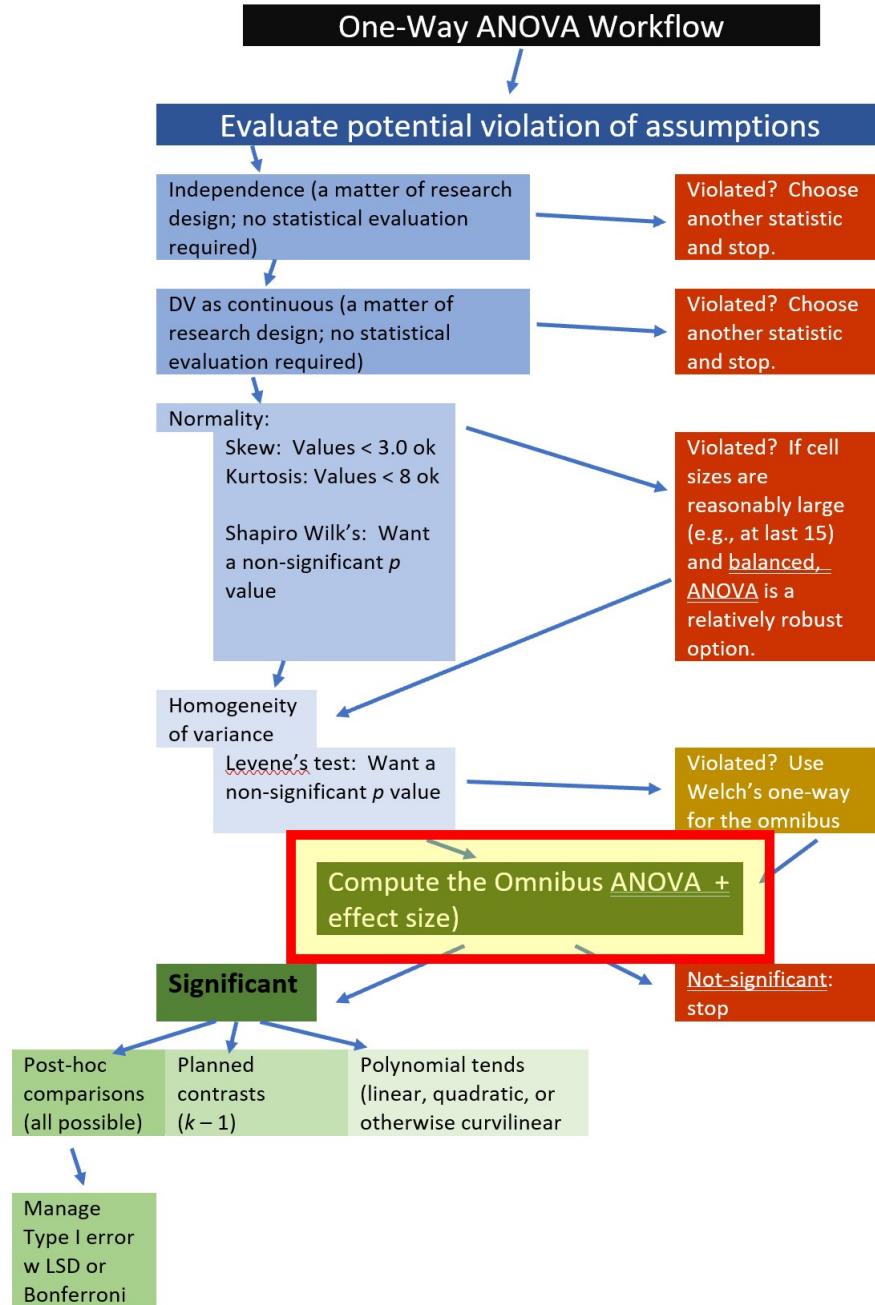


Figure 7.5: An image of the workflow for one-way ANOVA, showing that we are at the stage of computing the omnibus ANOVA.

The “*F* string” for an APA style results section should be written like this: $F(2, 87) = 13.566, p < .001$.

The object we created with the *aov()* function is capable of producing much information. Applying the *names()* function to the object can give us a list of values within it.

```
names(omnibus)
```

```
[1] "coefficients"   "residuals"      "effects"       "rank"
[5] "fitted.values"  "assign"        "qr"           "df.residual"
[9] "contrasts"      "xlevels"       "call"          "terms"
[13] "model"
```

One of the most commonly used functions to be applied to the *aov()* objects is *summary()*; but it’s not the only one. Let’s try some other options.

```
model.tables (omnibus, "means")
```

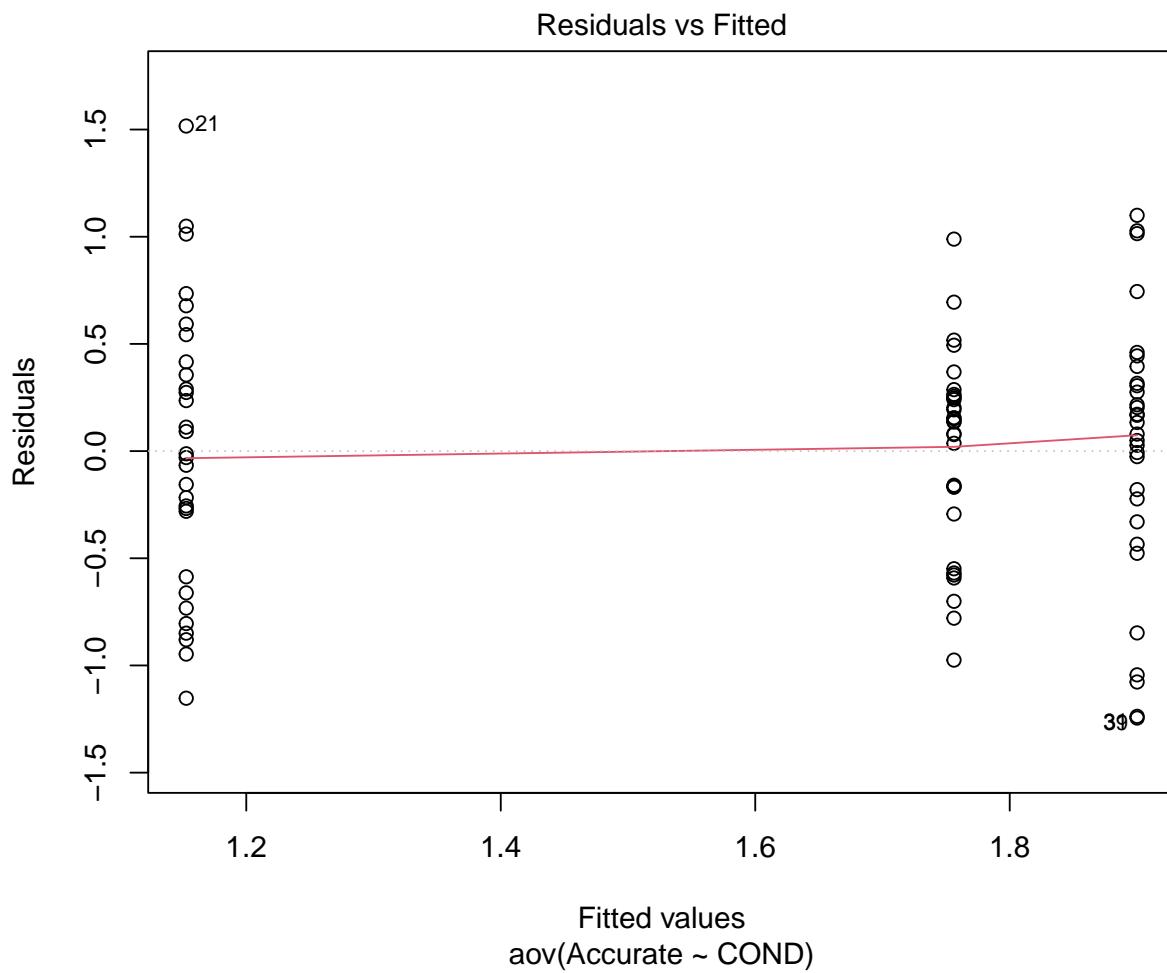
```
Tables of means
Grand mean
```

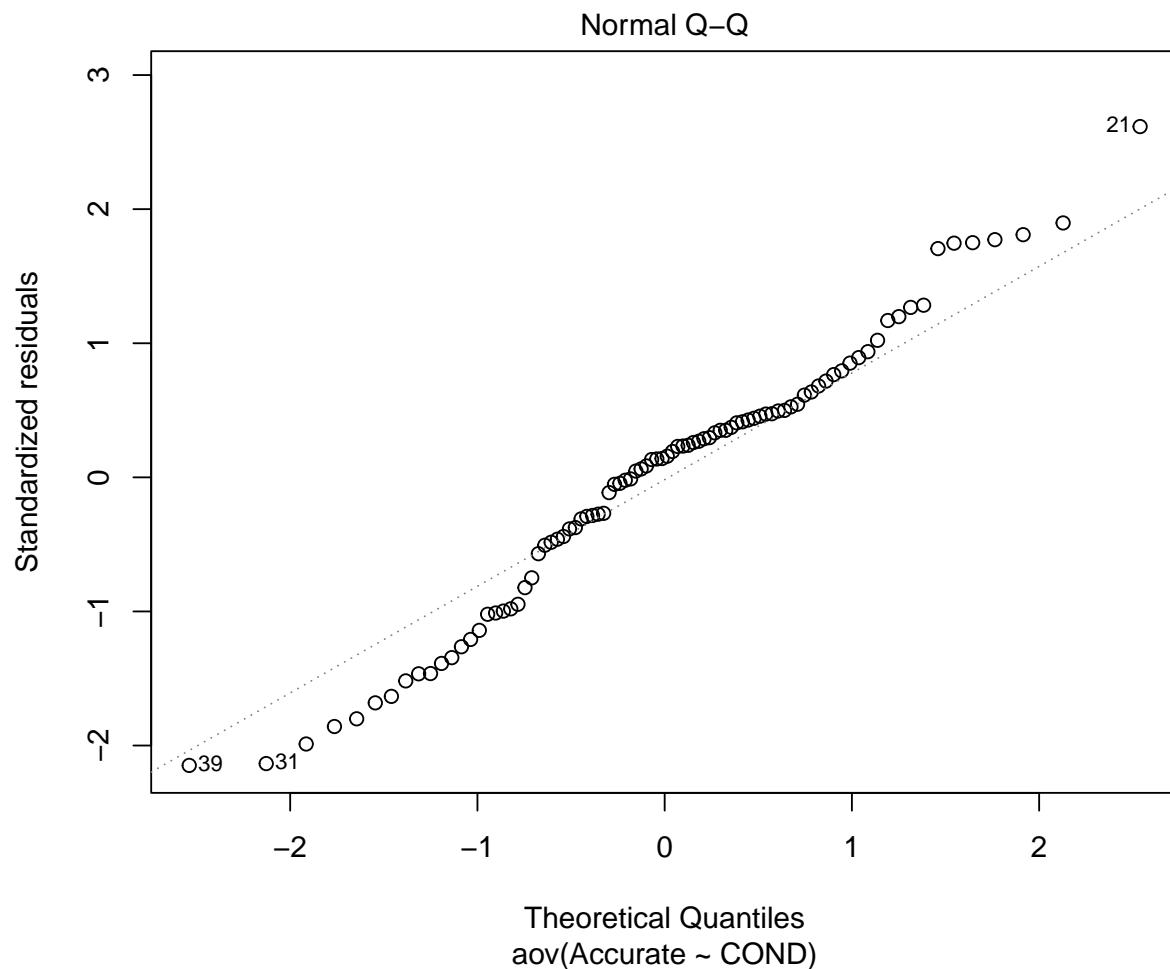
```
1.603042
```

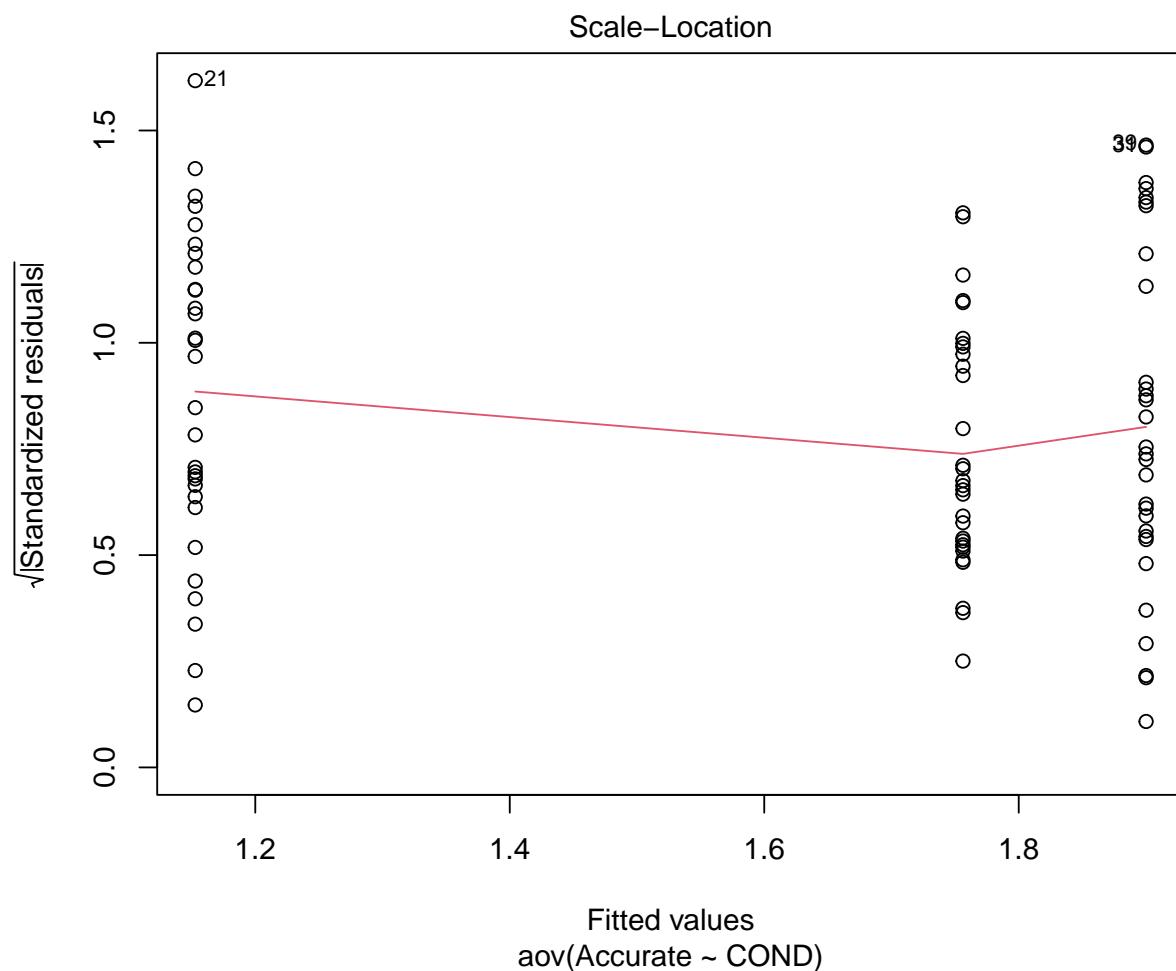
```
COND
COND
Control    Low     High
1.7562  1.9001  1.1528
```

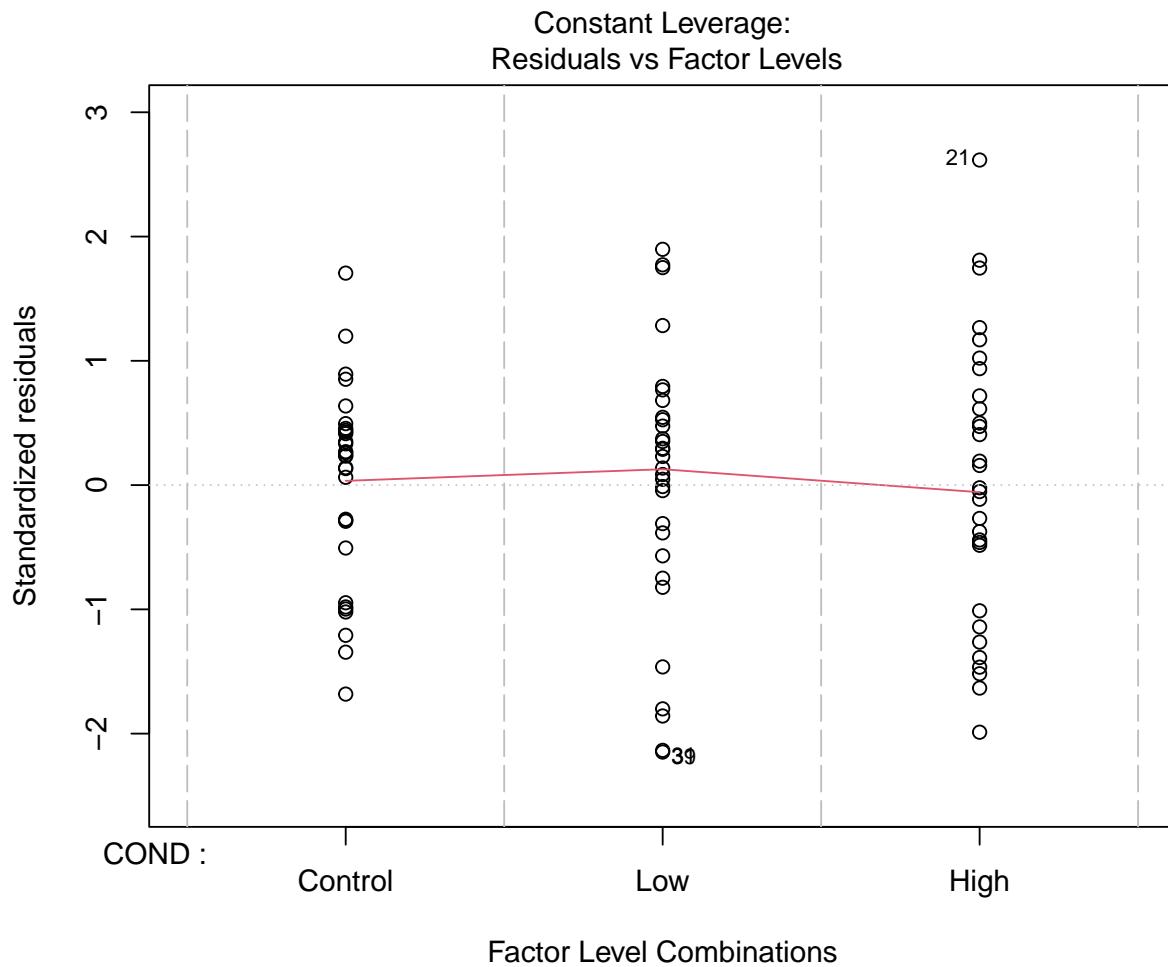
The *aov()* command has a quickplot feature.

```
plot(omnibus)
```









The first of the four plots fits the residuals. We already know from Levene's that we did not violate the homogeneity of variance test. With its straight line, this plot shows an equal spread across the three groups.

When the dots of the Q-Q plot map onto the diagonal, we have some indication of normality of the residuals (we want residuals to be normally distributed).

7.6.2.1 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) 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. 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.:

```
#Calculated using the object names
SSM omnibus / (SSM omnibus + SSMR omnibus)
```

```
[1] 0.238
```

```
#Re-calculated by using the numeric values
9.432/(9.432 + 30.246)
```

```
[1] 0.2377136
```

```
#Both options should produce the same result
```

The *lsr* package includes an eta-squared calculator. To use it, we simply insert the model/object we created with the *aov()* function to *lsr*'s *etaSquared()* function.

```
lsr::etaSquared(omnibus)
```

```
eta.sq eta.sq.part
COND 0.2377226 0.2377226
```

Notice that there are two effect sizes. We described eta-squared. Partial eta-squared is the default effect size reported in SPSS. There is a long history of debate about which to use. In certain circumstances (especially in more complicated analyses), partial-eta squared can be a bit more generous (i.e., larger than η^2). Thus, many prefer the reporting of the more cautious η^2 .

In our case, we see no difference between the two values. Differences begin to appear in datasets that are more complicated, such as when sample sizes across the levels of a factor differ.

7.6.2.2 Summarizing results from the omnibus ANOVA

Presenting the APA style results of the omnibus test is very straightforward:

Results indicated a significant effect of COND on accuracy perception $F (2, 87) = 13.566, p < .001, \eta^2 = 0.238$.

7.6.3 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.

A very common option is to conduct post-hoc, pairwise comparisons.

7.6.3.1 OPTION 1: Post-hoc, pairwise, comparisons

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 b/c strict criterion for significance must be used.

Helpful information about how to conduct post-hoc pairwise comparisons in R can be found at the [UCLA Institute for Digital Research and Education site \[noa\]](#).

```
pairwise.t.test(accSIM30$Accurate, accSIM30$COND, p.adj = "none")
```

```
Pairwise comparisons using t tests with pooled SD

data: accSIM30$Accurate and accSIM30$COND

Control Low
Low 0.34709 -
High 0.00015 0.0000042

P value adjustment method: none

#can swap "bonf" or "holm" for p.adj
```

The output only provides the p values associated with the mean differences in each of the conditions. We see that $p < .05$ when high is compared to control (0.00015) and high is compared to low (0.0000042). An APA style reporting results of these typically involves referencing the means (often reported in a table of means and standard deviations) or mean differences (hand calculated) with their p values.

Should we be concerned about Type I error?

Recall that *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 increases when we have a large number of pairwise comparisons.

Green and Salkind [2014b] reviewed three options for managing Type I error.

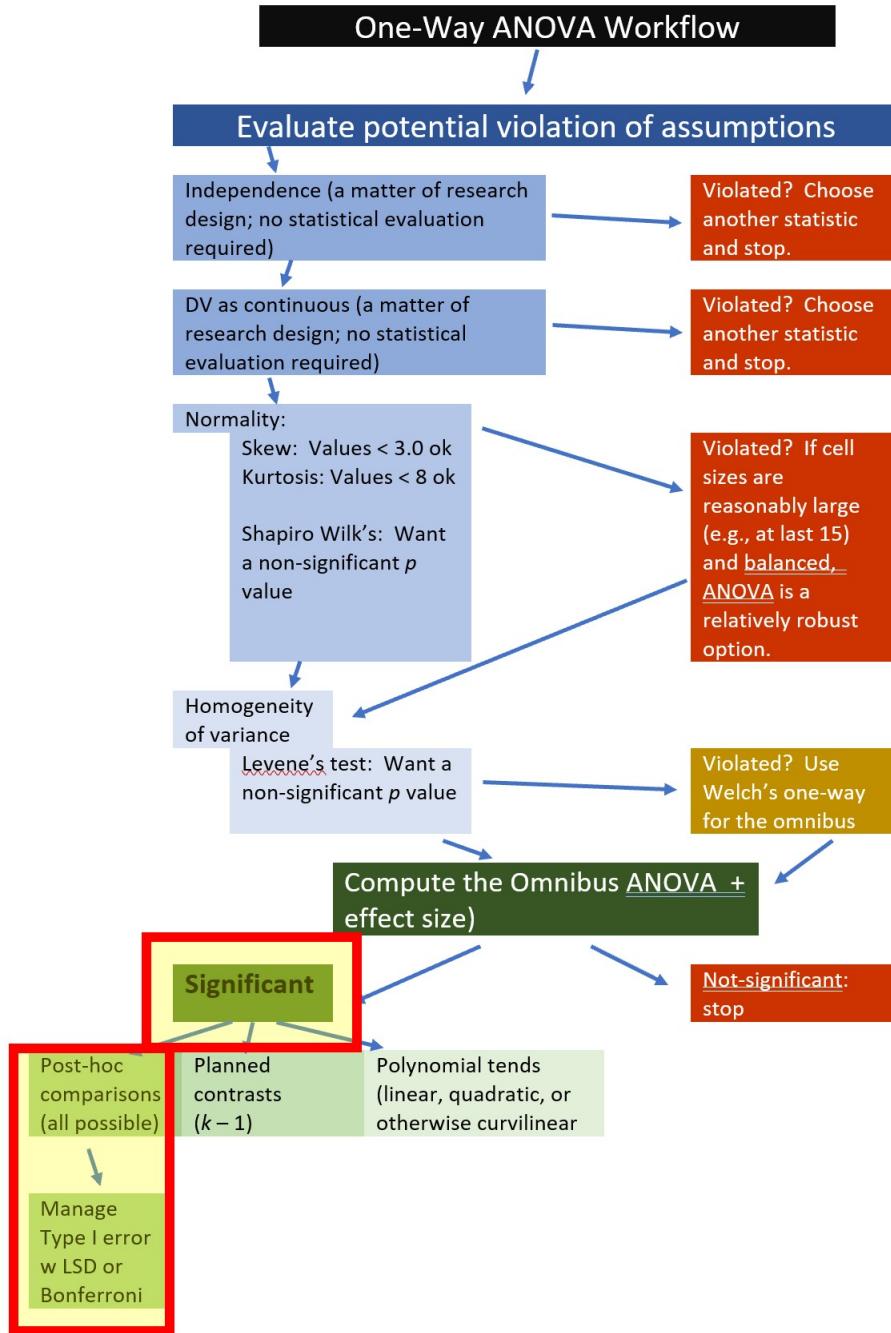


Figure 7.6: An image of the workflow for one-way ANOVA, showing that we are at the stage of following a statistically significant omnibus F test and are now conducting posthoc comparisons.

- Traditional Bonferroni:
 - Adjusts the p value upward by multiplying it (the raw p values) by the number of comparisons being completed. This holds the *total* Type I error rate across these tests to α (usually .05). The traditional Bonferroni is simple and therefore attractive, but when p values hover around .05, it can be too restrictive.
- Holms Sequential Bonferroni:
 - We'll describe this in more detail later. Briefly, it allows us to rank order the comparisons by their p value (smallest to largest). We determine the significance of each p value *sequentially* where the criteria for significance is incrementally relaxed.
- LSD method:
 - Permitted when there are only three pairwise comparisons among three groups, researchers can leave the p values as they are. Since the Tran and Lee [2014] research vignette is one of those circumstances, I will not make adjustments for Type I error. That is, I would claim the LSD and cite Green and Salkind [2014b] as justification for that decision.

There is, though, an even more powerful approach...

7.6.3.2 OPTION 2: Planned contrasts (non-orthogonal)

Another option is to evaluate planned comparisons.

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.
- Each contrast must involve only two chunks of variance.

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 low and high (because control was excluded, it should not reappear in the next contrast)
2. Low versus high

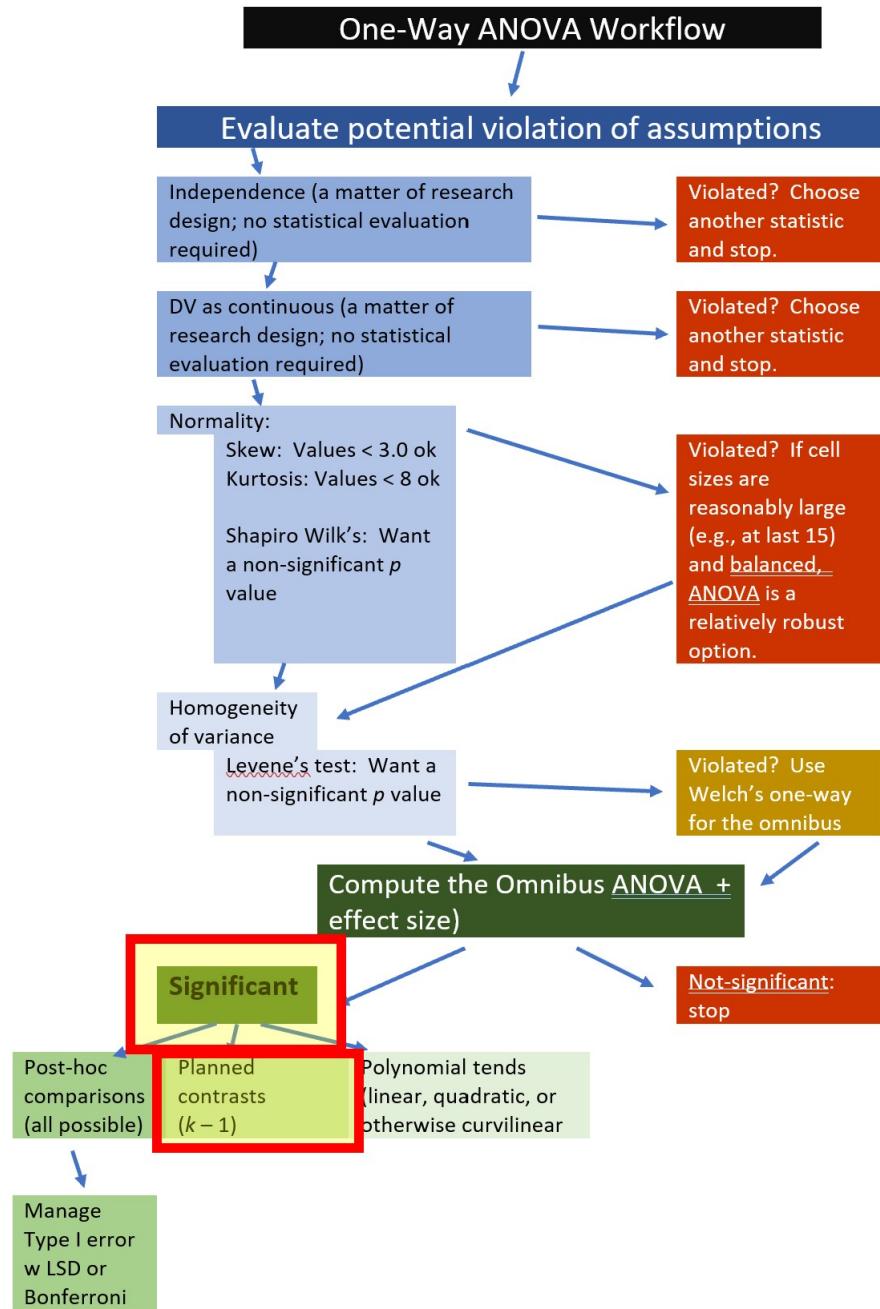


Figure 7.7: An image of the workflow for one-way ANOVA, showing that we are at the following up to a significant omnibus F by conducting planned comparisons

Underlying the `aov()` program is the linear model (“lm”). We could have used it to calculate the omnibus ANOVA, but it has clunky output.

We use it now to retrieve some contrast information. The code below is a planned comparison that uses the coding in the database (created when we formatted COND as an ordered factor) to compare the lowest coded group (control was 1, low was 2, high was 3) to the other two groups.

```
summary.lm(omnibus)
```

```
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 < 0.0000000000000002 *** 
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
```

From the above, regression output, we see that there was not a statistically significant difference between CONDControl and CONDLow ($p = 0.347$). There were statistically significant difference when CONDControl (the intercept) was compared to CONDHIGH ($p < .001$).

Note that these tests are conducted as t tests. Why? We are comparing only two groups and can use the t distribution.

Also note that these involved the comparison of the control group to low; then the control to high. This is consistent with an orthogonal contrast in that there are $k - 1$ contrasts (two contrasts with three original groups). However it is inconsistent with the requirement that once a group is singled out, it should not be used in a subsequent contrast. Therefore, it is quite possible we want something different.

7.6.3.3 OPTION 3: Planned contrasts (orthogonal)

Step 1: 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.

Step 2: Bind them 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.

Step 3: Create a new *aov()* model

```
# create a new object, the ANOVA looks the same, but it will now
# consider the contrasts (this is where order-of-operations matters)
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

```
contrasts(accSIM30$COND) <- cbind(c(-2, 1, 1), c(0, -1, 1))
```

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, $t(87) = -4.909$, $p < 0.001$.

7.6.3.4 OPTION 4: Trend (polynomial) analysis

Polynomial contrasts evaluate for the presence of a linear (or curvilinear) pattern in the data. To detect a trend, the data must be coded in an ascending order and it needs to be a sensible, theoretically defensible, comparison. Our data has a theoretically ordered effect (control, low racially loaded condition, high racially loaded condition). Recall that we created an ordered factor when we imported the data.

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 less than the number of levels. Our factor has three levels,

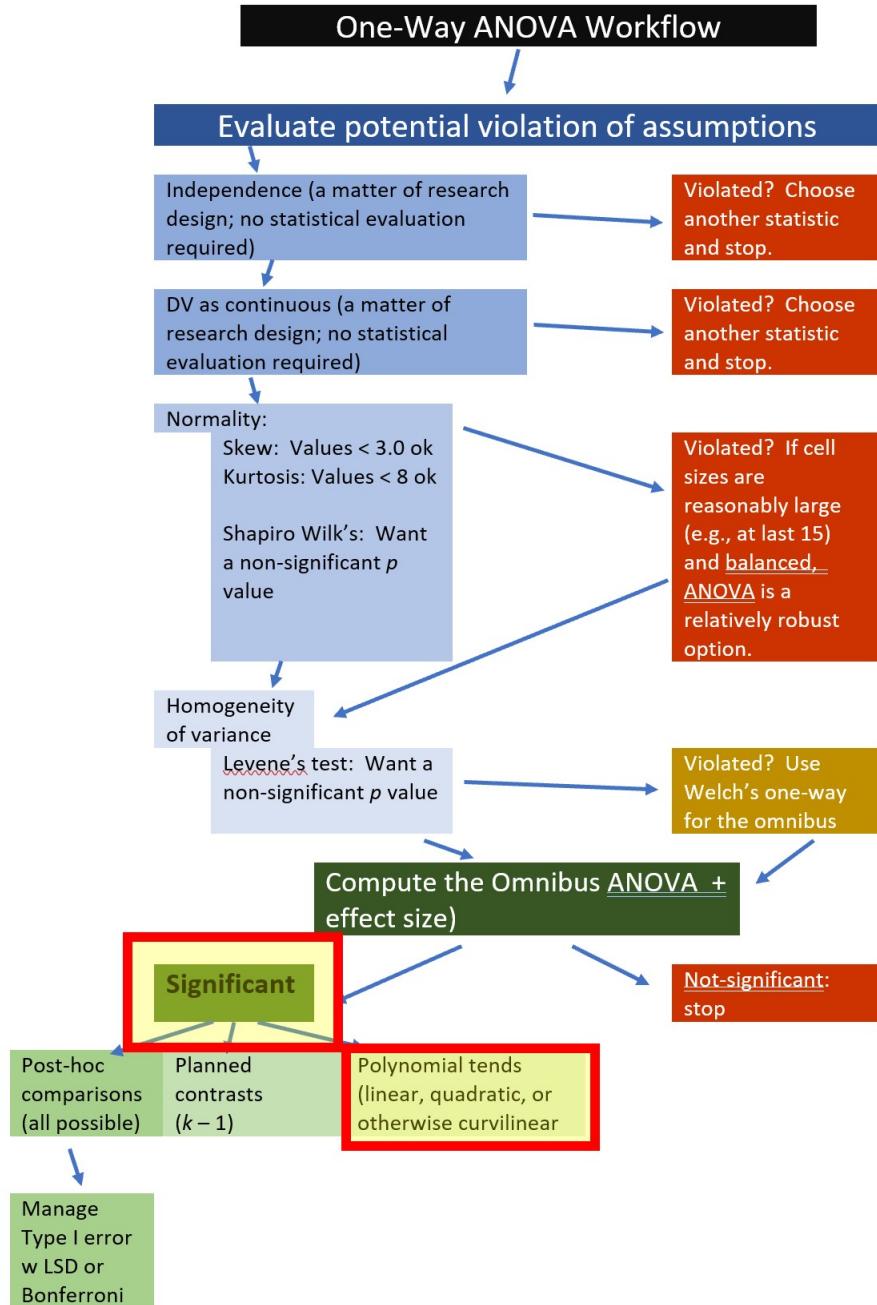


Figure 7.8: An image of the workflow for one-way ANOVA, showing that we are up to a significant omnibus F by assessing for a polynomial trend

therefore the polynomial contrast can check for a linear shape (.L) or a quadratic (one change in direction) shape (.Q). If we had four levels, the `contr.poly` could 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).

```
contrasts(accSIM30$COND) <- contr.poly(3)
accTrend <- aov(Accurate ~ COND, data = accSIM30)
summary.lm(accTrend)
```

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 ***
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

A quick peek back at an early plot shows illustrates the quadratic trend that was supported by the analysis.

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.

7.6.3.5 Which set of follow-up tests do we report?

It depends! What best tells the story of your data? 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 think this provides good information and 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
 - a priori theory 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).

7.6.4 What if we Violated the Homogeneity of Variance test?

The *oneway.test* produces Welch's F – a test more robust to violation of the homogeneity of variance assumption. The Welch's approach to attenuating error or erroneous conclusions caused by violations of the homogeneity of variance assumption is to adjust the residual degrees of freedom used to produce the Welch's F -ratio.

Another common correction is the Brown and Forsythe F -ratio. At this time I have not located and tried an R package that produces this.

```
oneway.test(Accurate ~ COND, data = accSIM30)
```

```
One-way analysis of means (not assuming equal variances)
```

```
data: Accurate and COND
F = 11.569, num df = 2.000, denom df = 56.343, p-value = 0.00006174
```

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 "conservative." While it doesn't matter in this case, if it were a study where the p value was closer to .05, it could make a difference. These are some of the tradeoffs made in order to have confidence in the results.

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
- *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)
```

```
Registered S3 method overwritten by 'parameters':
method                  from
format.parameters_distribution datawizard
```

```
[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 APA Style Results

All that's left to do is *write it up!* APA style results sections in empirical manuscripts are typically accompanied by tables and figures. APA style discourages repeated material and encourages reducing the cognitive load of the reader. For this example, I suggest two tables – one with means and standard deviations of the groups and a second that reports the output from the one-way ANOVA. In an article with multiple statistics, the authors might wish to combine or delete these.

The package *apaTables* can produce journal-ready tables by using the object produced by the *aov()* function. Deciding what to report in text and table is important.

Here 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).

I will want to give the values of mean differences (M_{diff}) in the results. I can quickly use R to calculate them here.

```
# calculating mean difference between control and high
1.76 - 1.15
```

[1] 0.61

```
# calculating mean difference between low and high
1.9 - 1.15
```

[1] 0.75

```
# calculating mean difference between control and low
1.76 - 1.9
```

[1] -0.14

Here I create Table 2 with results of the one-way ANOVA. The result in Microsoft Word can be edited (e.g., I would replace the partial-eta squared with η^2) for the journal article.

```
apaTables::apa.aov.table(omnibus, table.number = 2, filename = "Table2.doc")
```

Table 2

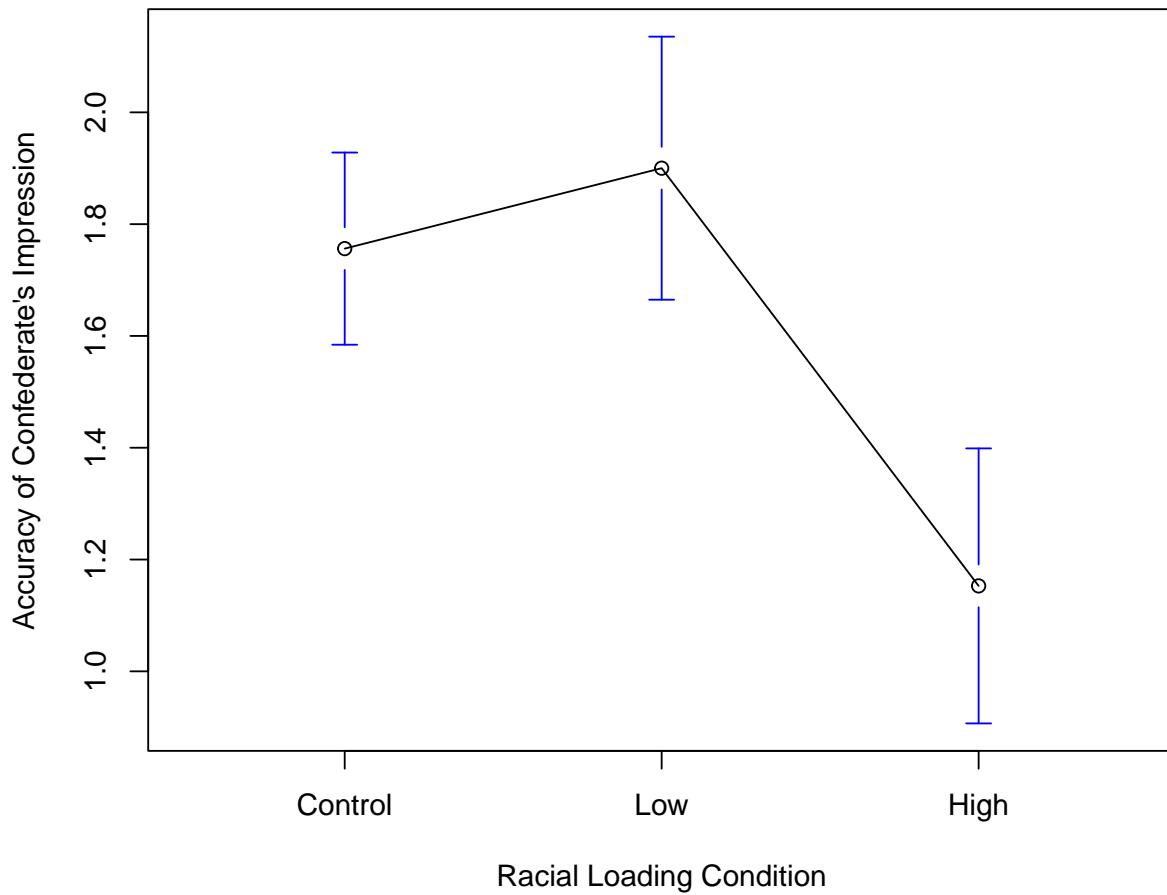
ANOVA results using Accurate as the dependent variable

Predictor	SS	df	MS	F	p	partial_eta2	CI_90_partial_eta2
(Intercept)	92.53	1	92.53	266.15	.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, there are a number of options. I find the *plotmeans()* function (used earlier) to be adequate for the purpose of one-way ANOVA. As we progress through this textbook, I will demonstrate options that offer more and less complexity.

```
gplots::plotmeans(formula = Accurate ~ COND, data = accSIM30, xlab = "Racial Loading Condition",
                  ylab = "Accuracy of Confederate's Impression", n.label = FALSE)
```



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, COND, included three levels: control/none, low, and high levels of racial loading. Results of Levene's homogeneity of variance test indicated no violation of the homogeneity of variance assumption ($F[2,87] = 1.532, p = 0.222$). Similarly, results of the Shapiro Wilk's test indicated no violation of the normality assumption in each of the cells (Control, $W = 0.954, p = 0.215$; Low, $W = 0.944, p = 0.115$; High, $W = 0.980, p = 0.831$).

Results indicated a significant effect of COND on accuracy perception $F(2, 87) = 13.566, p < .001, \eta^2 = 0.238$. In a series of post-hoc comparisons, there were statistically significant differences between the control and high ($M_{diff} = 0.61, p < .001$) and low and high ($M_{diff} = 0.75, p < 0.001$) conditions, but not the control and low conditions ($M_{diff} = -.14, p = 0.347$). A statistically significant polynomial contrast suggested a quadratic

trend across the three, ordered, levels of the conditions ($t[87] = -0.364, p = .001$) such that perception of accuracy increased slightly from the control to low conditions, but decreased more dramatically from low to high. 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.

7.9 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 teh 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?

7.10 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). 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

7.10.1 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.

7.10.2 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.

7.10.3 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.

7.10.4 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	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 Explanation to grader	5	_____
Totals	40	_____

7.11 Bonus Reel:

7.11.1 What's with the inline code?

If you are working in the .rmd file you may notice sections that look like this:

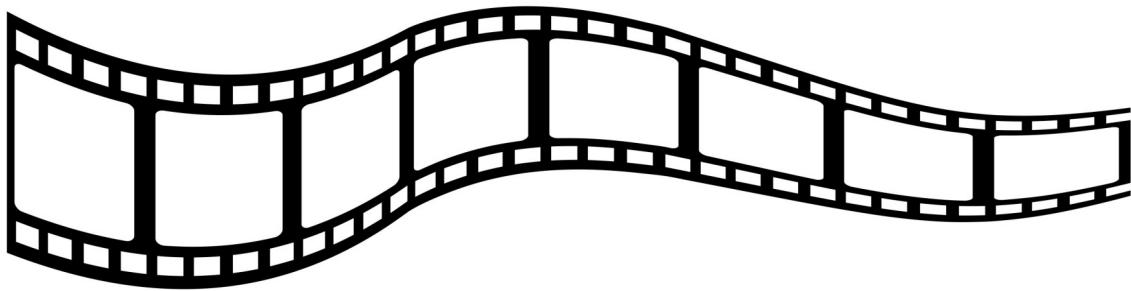


Figure 7.9: Image of a filmstrip signifying that the what follows is considered to be supplemental

```
# this script is used for the inline coding in the lesson and,
# although you will want to run it so you can 'feed' the objects into
# later script, there is no 'lesson' relative to this lecture
df1omnibus <- formattable::digits(summary(omnibus)[[1]][1, "Df"], 0)
df1omnibus
df2omnibus <- formattable::digits(summary(omnibus)[[1]][2, "Df"], 0)
df2omnibus
SSMomnibus <- formattable::digits(summary(omnibus)[[1]][1, "Sum Sq"], 3)
SSMomnibus
SSRomnibus <- formattable::digits(summary(omnibus)[[1]][2, "Sum Sq"], 3)
SSRomnibus
SSTomnibus <- formattable::digits((SSMomnibus + SSRomnibus), 3)
SSTomnibus
MSMomnibus <- formattable::digits(summary(omnibus)[[1]][1, "Mean Sq"],
  3)
MSMomnibus
MSRomnibus <- formattable::digits(summary(omnibus)[[1]][2, "Mean Sq"],
  3)
MSRomnibus
Fomnibus <- formattable::digits(summary(omnibus)[[1]][1, "F value"], 3)
Fomnibus
pomnibus <- formattable::digits(summary(omnibus)[[1]][1, "Pr(>F)"], 3)
pomnibus
```

I'm still learning about this and am using the lessons to practice. In these hidden (from the rendered view) boxes of script, I am capturing the output values as R objects. Later I can use inline code (the object's name, preceded with an "r", surrounded by tics [unfortunately, I cannot demo it in the .rmd file without getting an error]) to insert the value into the lecture notes and results. I'm working up to writing an entire journal article in R. This way, if something changes (e.g., more participants in the Qualtrics survey, a different approach to cleaning data) when I re-run the script everything auto-updates and I'm at less of a risk (or maybe a different kind of risk) for making typographical errors.

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.

- Navarro, D. (2020). Chapter 16: Factorial 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 [2014b], 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(lsr)){install.packages('lsr')}
# if(!require(car)){install.packages('car')}
# if(!require(ggpubr)){install.packages('ggpubr')}
# if(!require(effectsize)){install.packages('effectsize')}
# if(!require(pwr2)){install.packages('pwr2')}
# if(!require(apaTables)){install.packages('apaTables')}
```

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 get into the assumptions you'll see that it 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.
- Factor naming follows a number/levels convention.
- Today's example is a 3X2 ANOVA. We know there are two factors that have three and two levels, respectively: *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/IV and the DV.

In a later lesson we work an ANCOVA – where we will distinguish between a *moderator* and a *covariate*. In regression models, you will likely be introduced to the *mediator*.

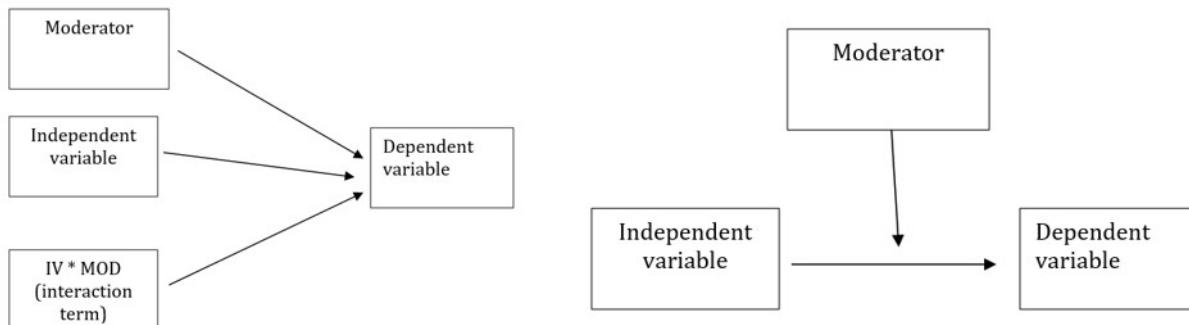


Figure 8.1: Graphic representations of a moderated relationship?

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
 - graphing the data

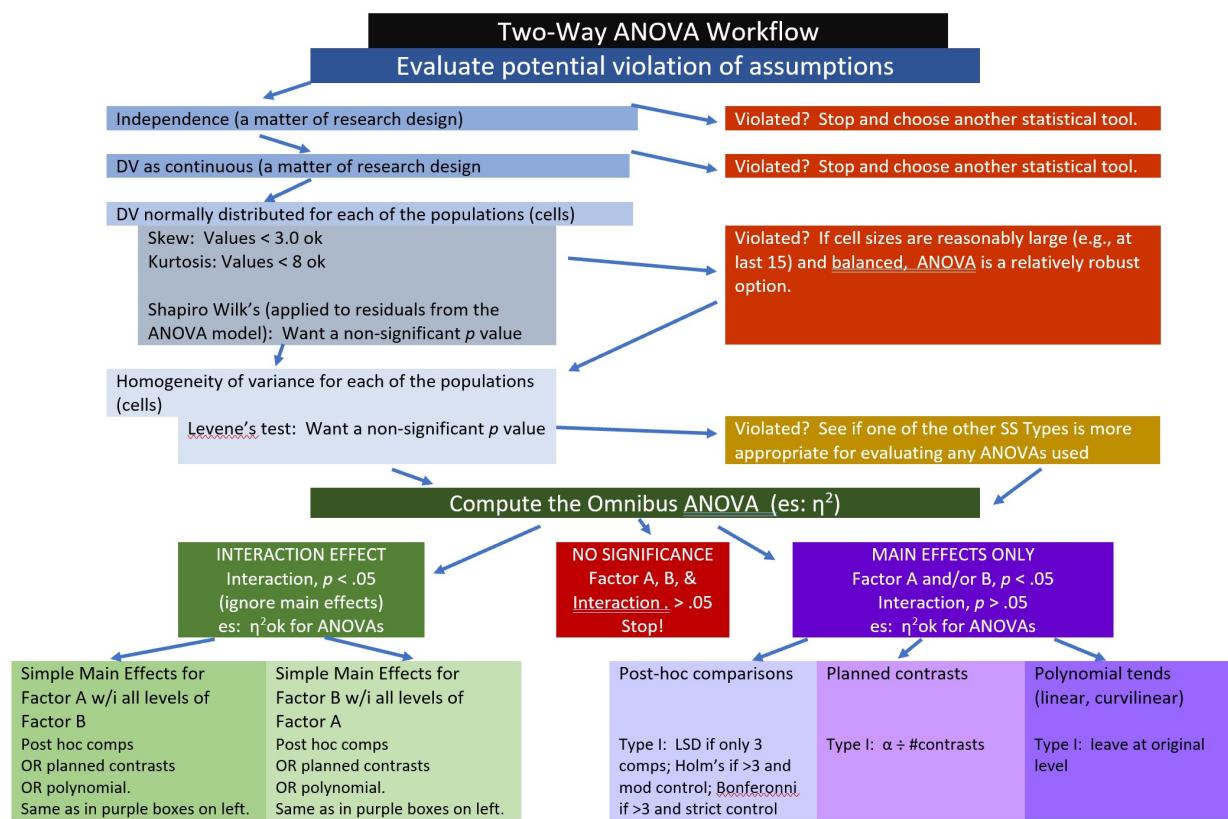


Figure 8.2: An image of a workflow for the two-way ANOVA

- computing descriptive statistics
 - check distributional assumptions
 - use Levene's test to check for homogeneity of variance
 - use Shapiro Wilks to check for normality
3. Construct or choose contrasts
 - select contrasts and specify for all of the IVs in the analysis
 - if you want to use Type III sums of squares, contrasts must be orthogonal
 4. 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*
 5. Follow-up testing based on significant main or interaction effects
 - significant interactions require test 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*
 6. 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. 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

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 ask for the structure again. 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 ...
 $ Photo   : Factor w/ 2 levels "Dayaknese","Madurese": 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_RDS <-
# readRDS('Ramdhani_RDS.rds') str(Ramdhani_RDS)
```

8.3.1 Preliminary exploration of our research vignette

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

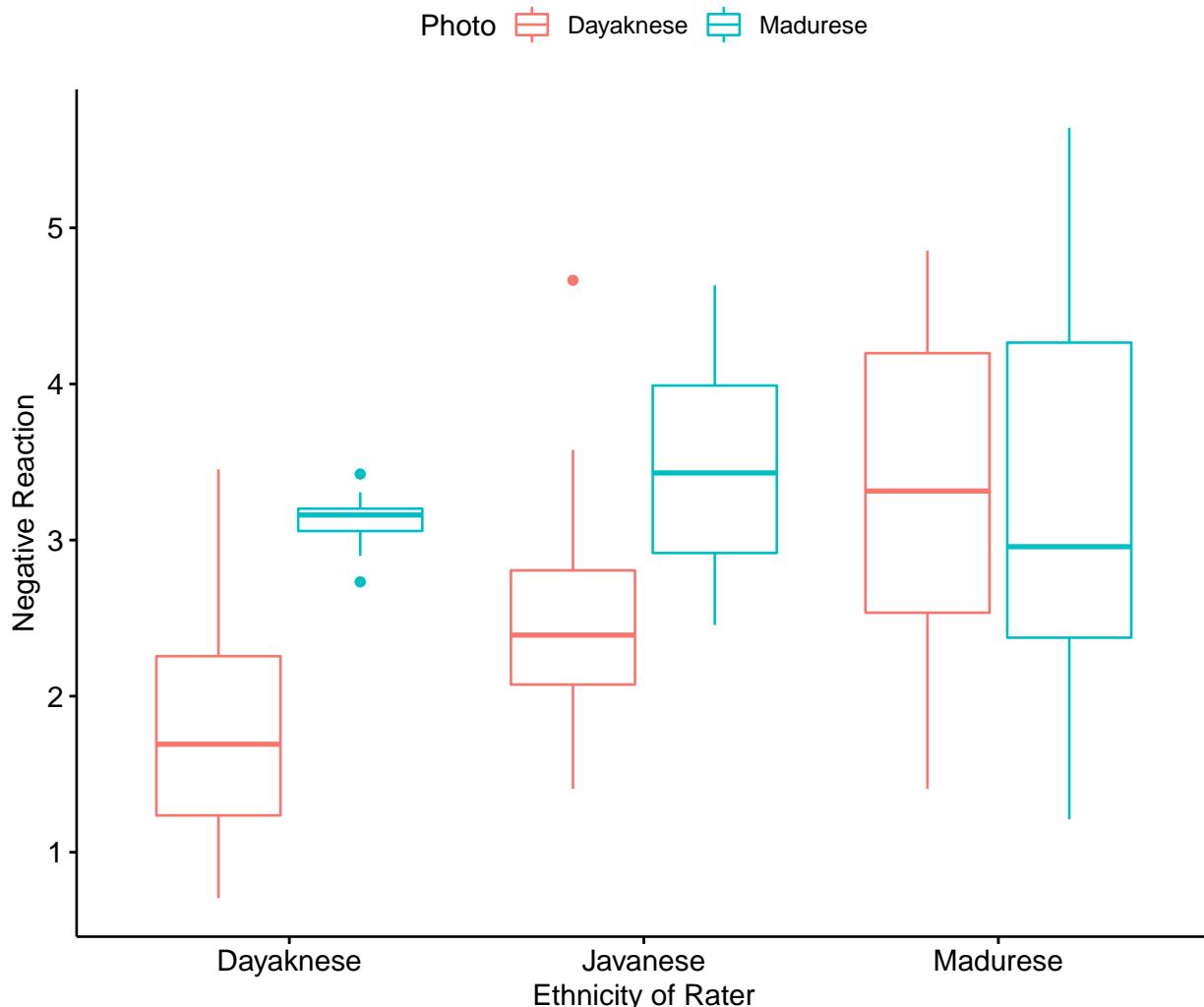
Negative3	1.406	4.854	3.448	-0.126	-1.267	0.236
Negative4	2.732	3.423	0.691	-0.623	0.481	0.037
Negative5	2.456	4.631	2.175	-0.010	-1.307	0.146
Negative6	1.211	5.641	4.430	0.215	-1.238	0.298

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)
```

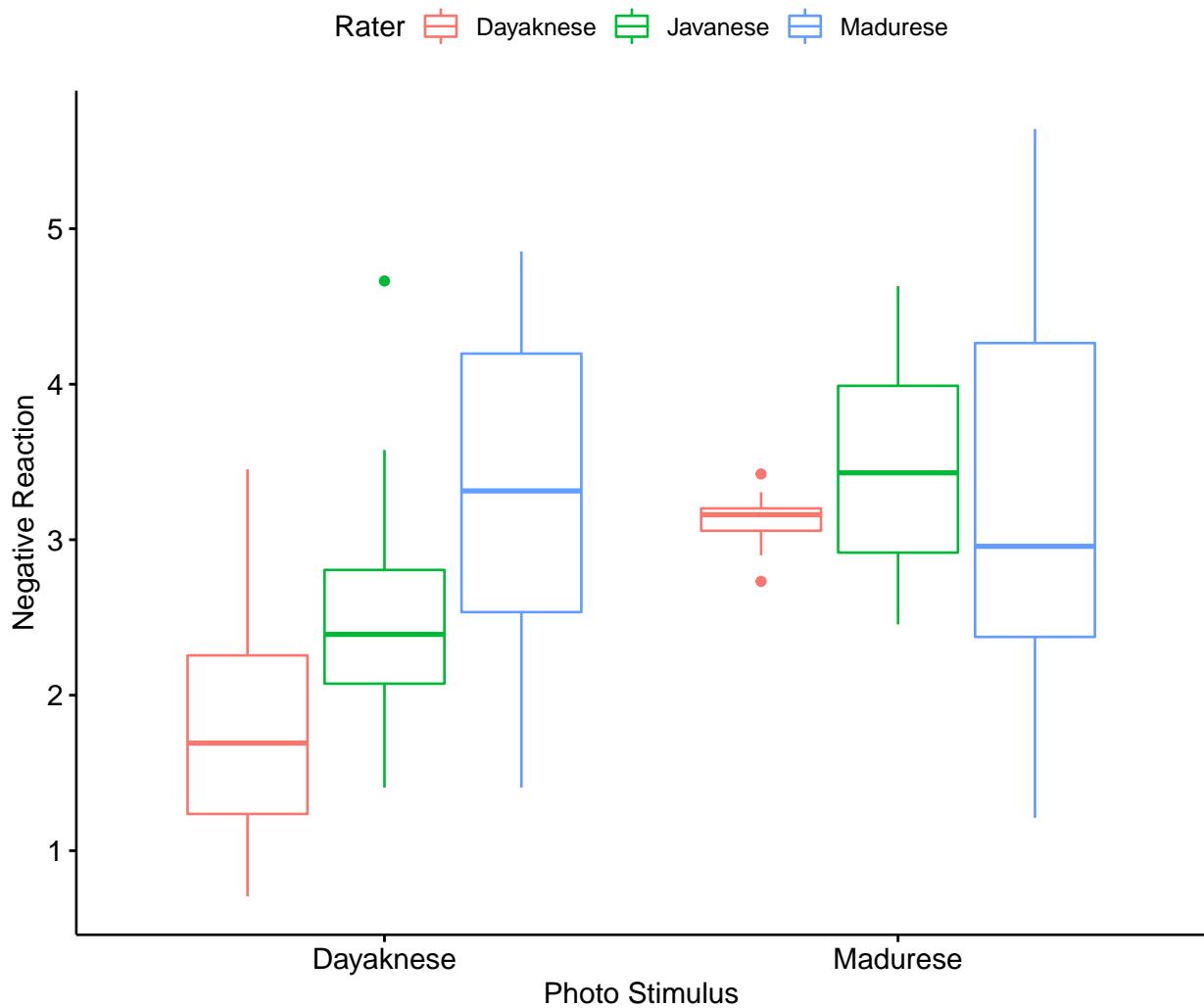
At this stage, it would be useful to plot our data. Figures can assist in the conceptualization of the analysis.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
                   xlab = "Ethnicity of Rater", ylab = "Negative Reaction")
```



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")
```

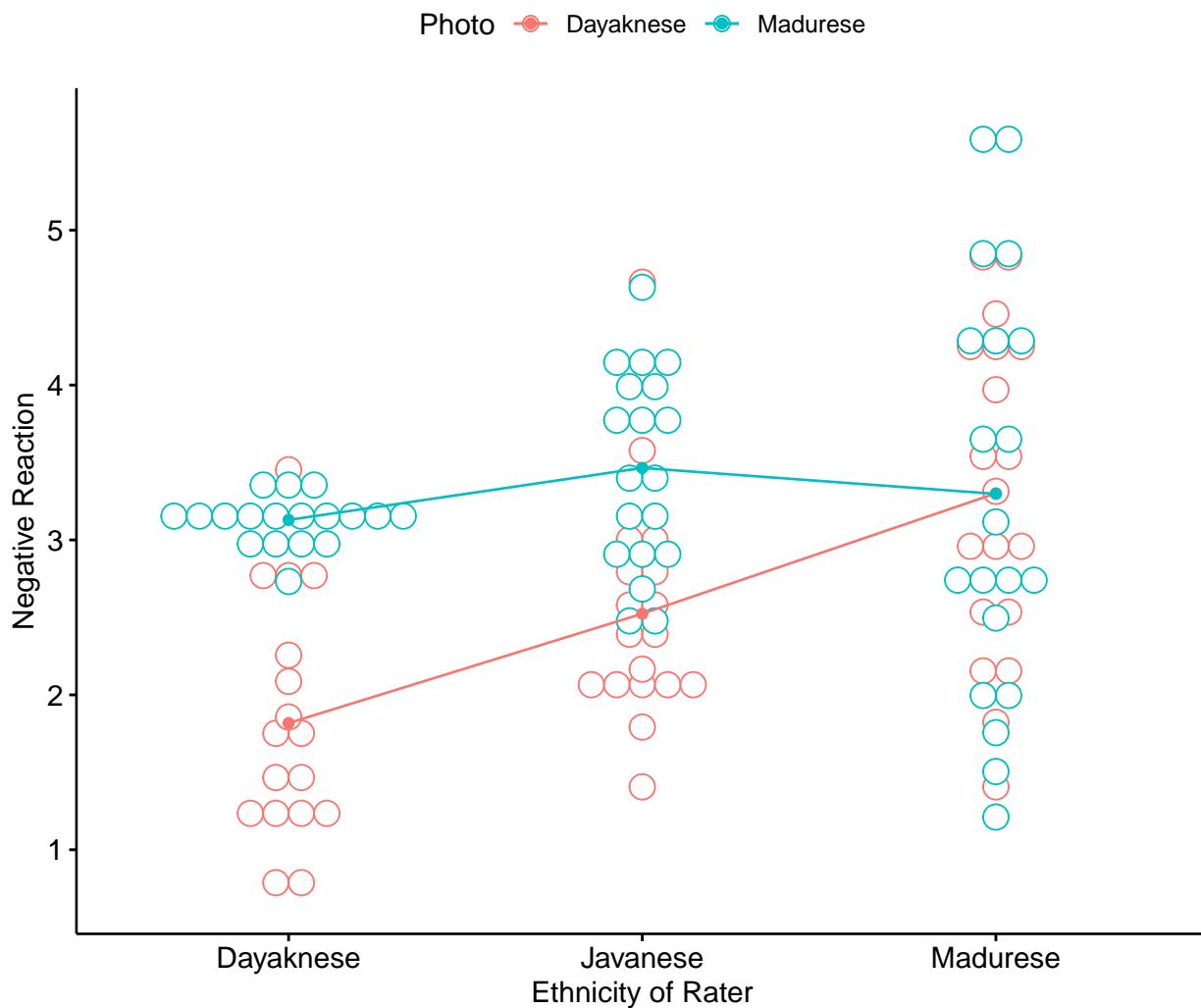


Yet another option plots the raw data as bubbles, the means as lines, and denotes differences in the moderator with color.

```
ggpubr::ggline(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction", add = c("mean_se",
  "dotplot"))
```

Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.

Warning: Computation failed in `stat_summary()`:
object 'mean_se_' of mode 'function' was not found



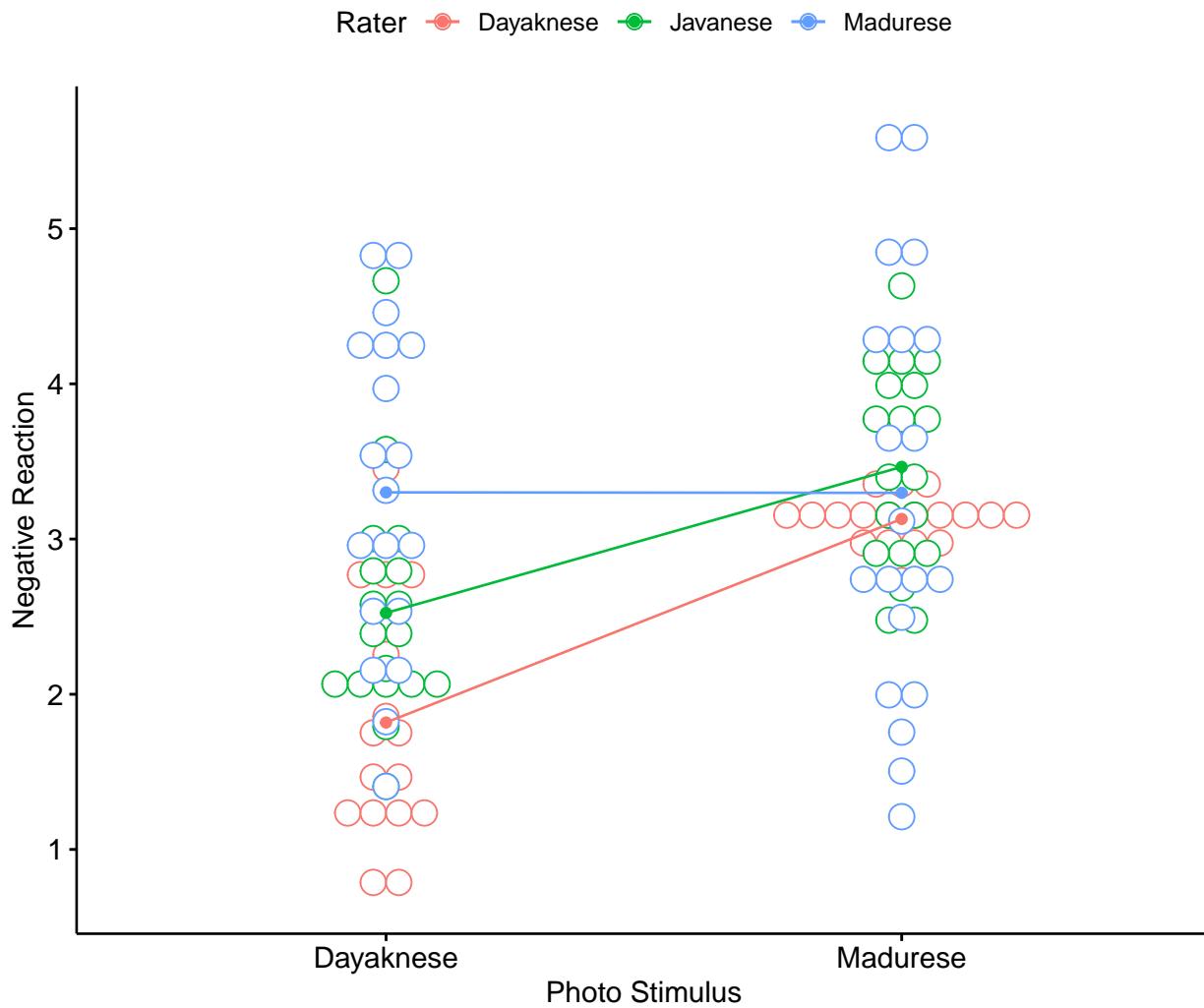
```
# add this for a different color palette: palette = c('#00AFBB',
# '#E7B800')
```

We can reverse this to see if it assists with our conceptualization.

```
ggpubr::gglime(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Photo Stimulus", ylab = "Negative Reaction", add = c("mean_se",
  "dotplot"))
```

Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.

```
Warning: Computation failed in `stat_summary()`:
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 becomes 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:

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

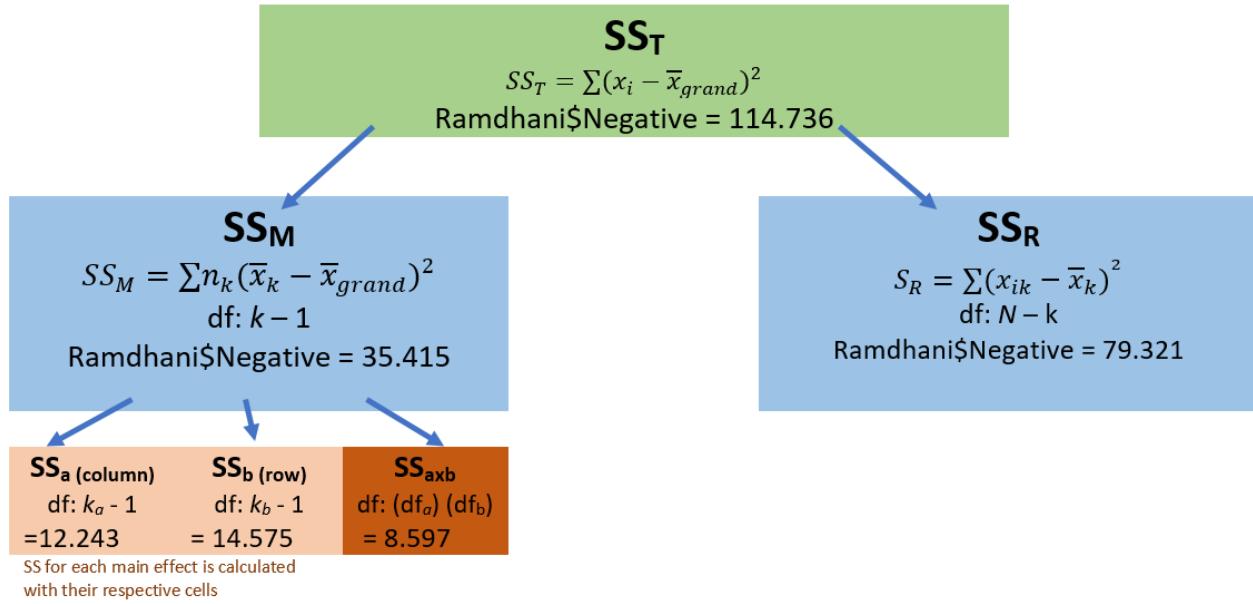


Figure 8.3: Image of a flowchart that partitions variance from sums of squares totals to its component pieces

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))
```

Pop quiz: What's the sum of our new *m_dev* variable?

Let's find out!

```
sum(Ramdhani_df$m_dev)
```

```
[1] -0.0000000000000007549517
```

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)
```

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, we need to obtain means for each of the combinations of the factors. We have a 3 x 2:

- 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 treatment group there should be no other covariate (confounding variable) so that 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. Let's take another look at our descriptives.

```
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
Negative2	2	Javanese	Dayaknese		1	18	2.524	0.742	2.391	2.460
Negative3	3	Madurese	Dayaknese		1	19	3.301	1.030	3.314	3.321
Negative4	4	Dayaknese	Madurese		1	18	3.129	0.156	3.160	3.136
Negative5	5	Javanese	Madurese		1	19	3.465	0.637	3.430	3.456
Negative6	6	Madurese	Madurese		1	20	3.297	1.332	2.958	3.254
		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	2.492	0.856	2.900	2.561	0.480	0.706	3.453
Negative2	2	Javanese		1	3.007	0.831	2.913	2.986	0.984	1.406	4.664
Negative3	3	Madurese		1	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.

8.4.6 Source Table Games!

As in the lesson for one-way ANOVA, we can use the hints in this source table to determine if we have statistically significance in the model. The formulas in the table provide some hints.

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		$k_b - 1$			

Source	SS	df	$MS = \frac{SS}{df}$	$F = \frac{MS_{source}}{MS_{resid}}$	F_{CV}
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

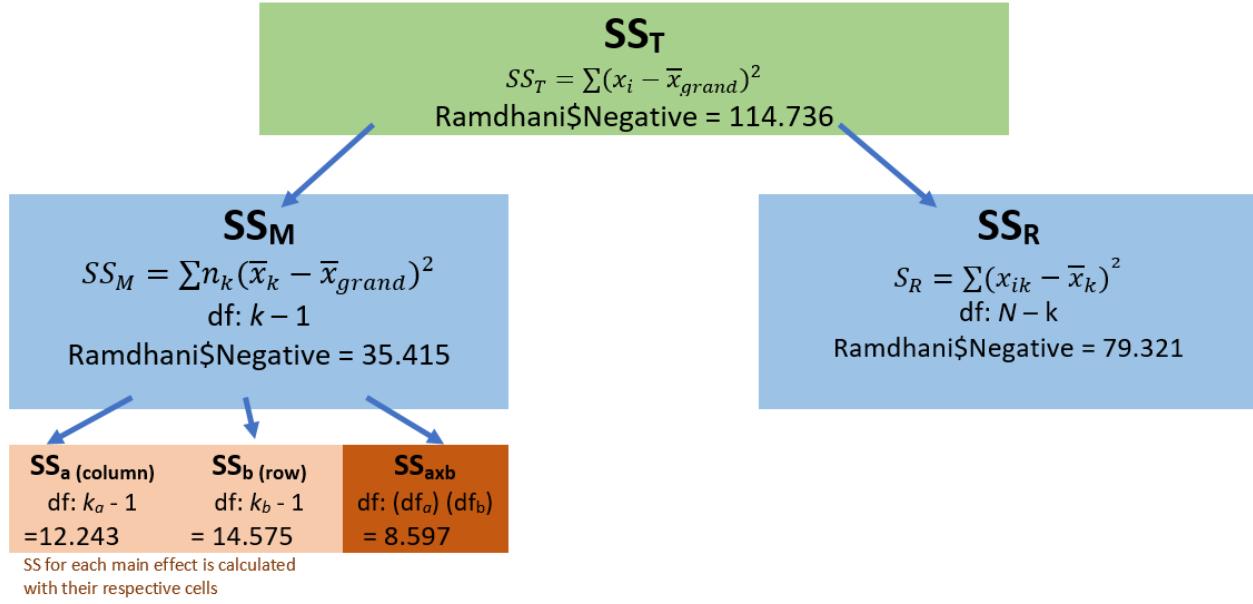


Figure 8.4: Image of a flowchart that partitions variance from sums of squares totals to its component pieces

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. This particular ANOVA has four:

Assumptions

- 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.
 - With correlated observations, there is a dramatic increase of Type I error
 - There are options 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.

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.

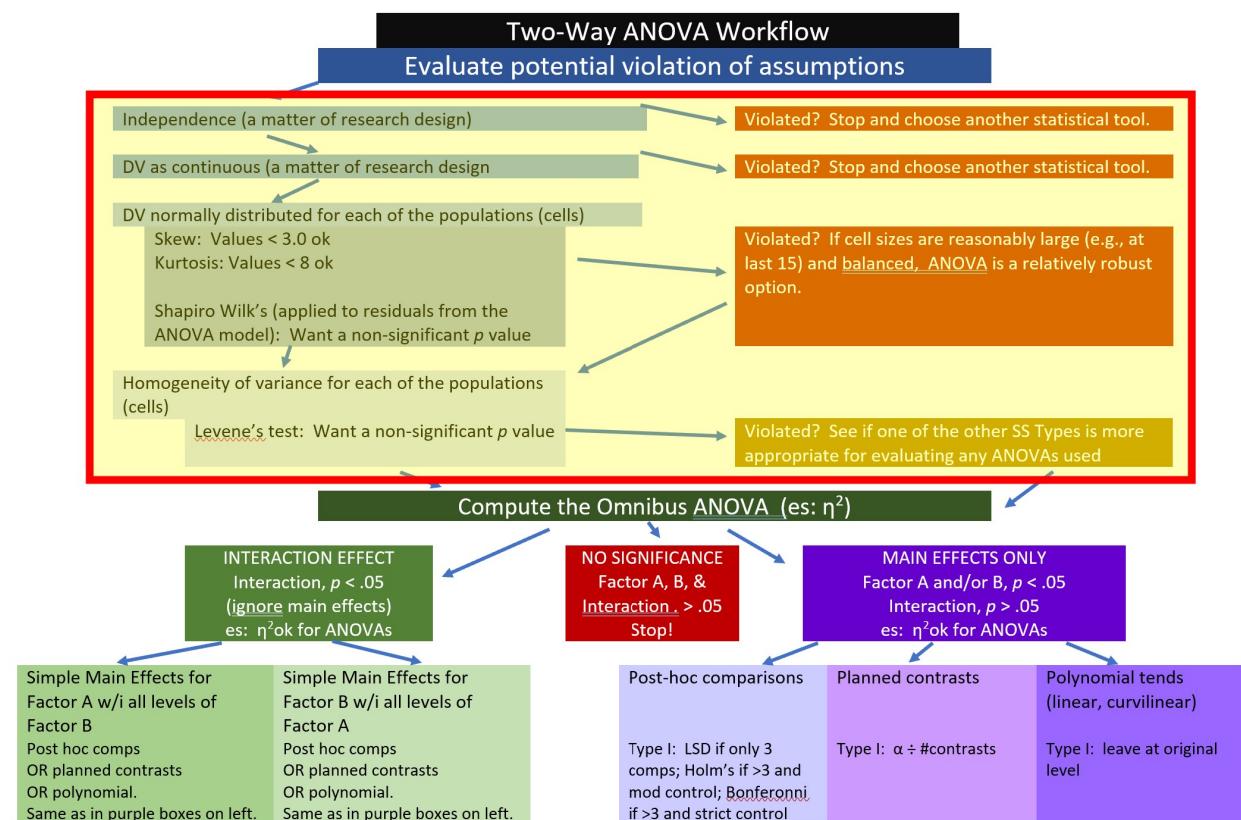


Figure 8.5: Image of a flowchart showing that we are on the “Evaluating assumptions” portion of the workflow

```
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
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 DV is normally distributed

Let's start by analyzing **skew** and **kurtosis**.

```
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

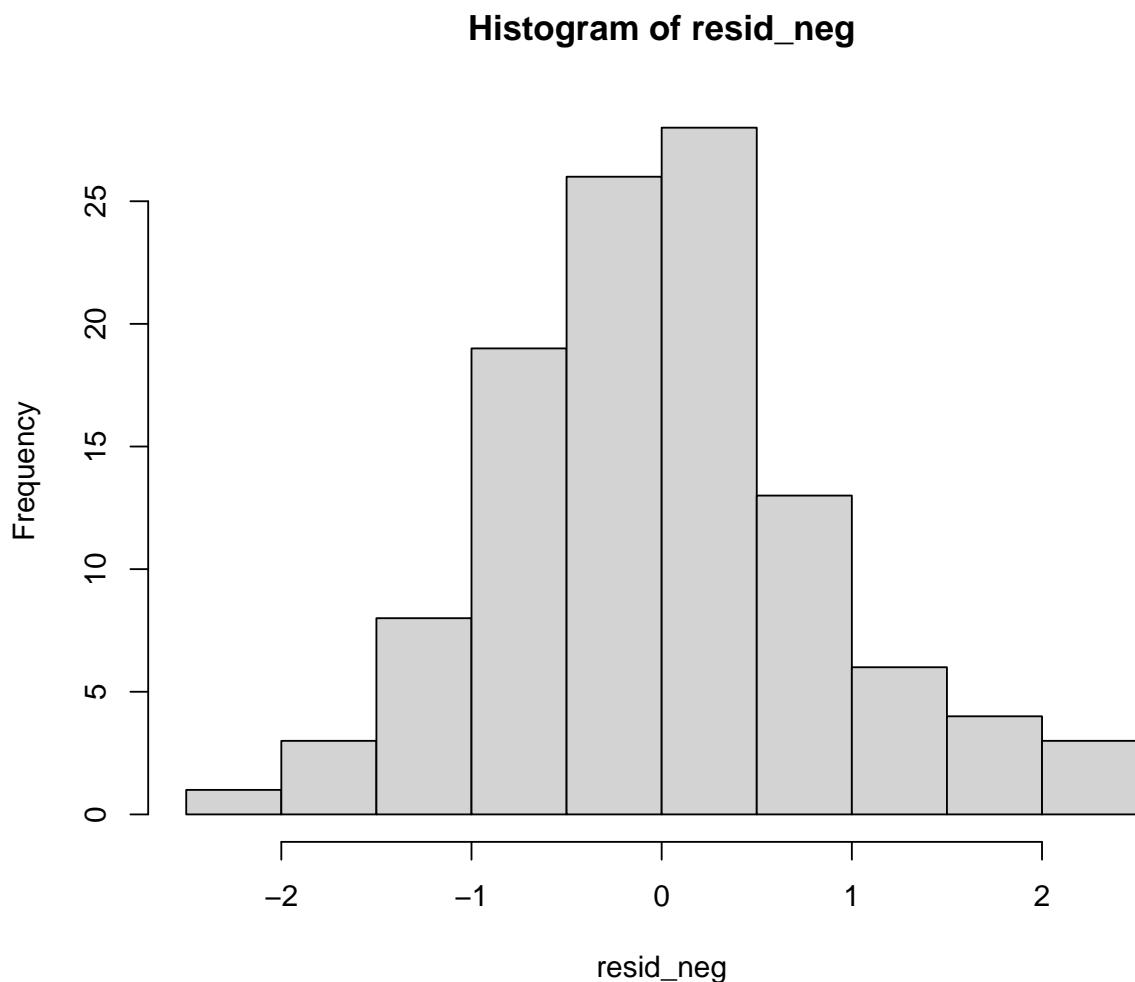
	min	max	range	skew	kurtosis	se
Negative1	0.706	3.453	2.747	0.513	-0.881	0.186
Negative2	1.406	4.664	3.258	1.205	1.475	0.175
Negative3	1.406	4.854	3.448	-0.126	-1.267	0.236
Negative4	2.732	3.423	0.691	-0.623	0.481	0.037
Negative5	2.456	4.631	2.175	-0.010	-1.307	0.146
Negative6	1.211	5.641	4.430	0.215	-1.238	0.298

Using guidelines from Kline [2016] our values for skewness should fall below 3.0 (they do) and all values for kurtosis should fall below 8 to 20 (ours do).

In a factorial design, the Shapiro-Wilk test is applied to residuals from the model itself. Examination of those residuals can give us a good indication of normality.

First, we extract the residuals (i.e., that which is left-over/unexplained) from the model. We can examine their distribution with a plot.

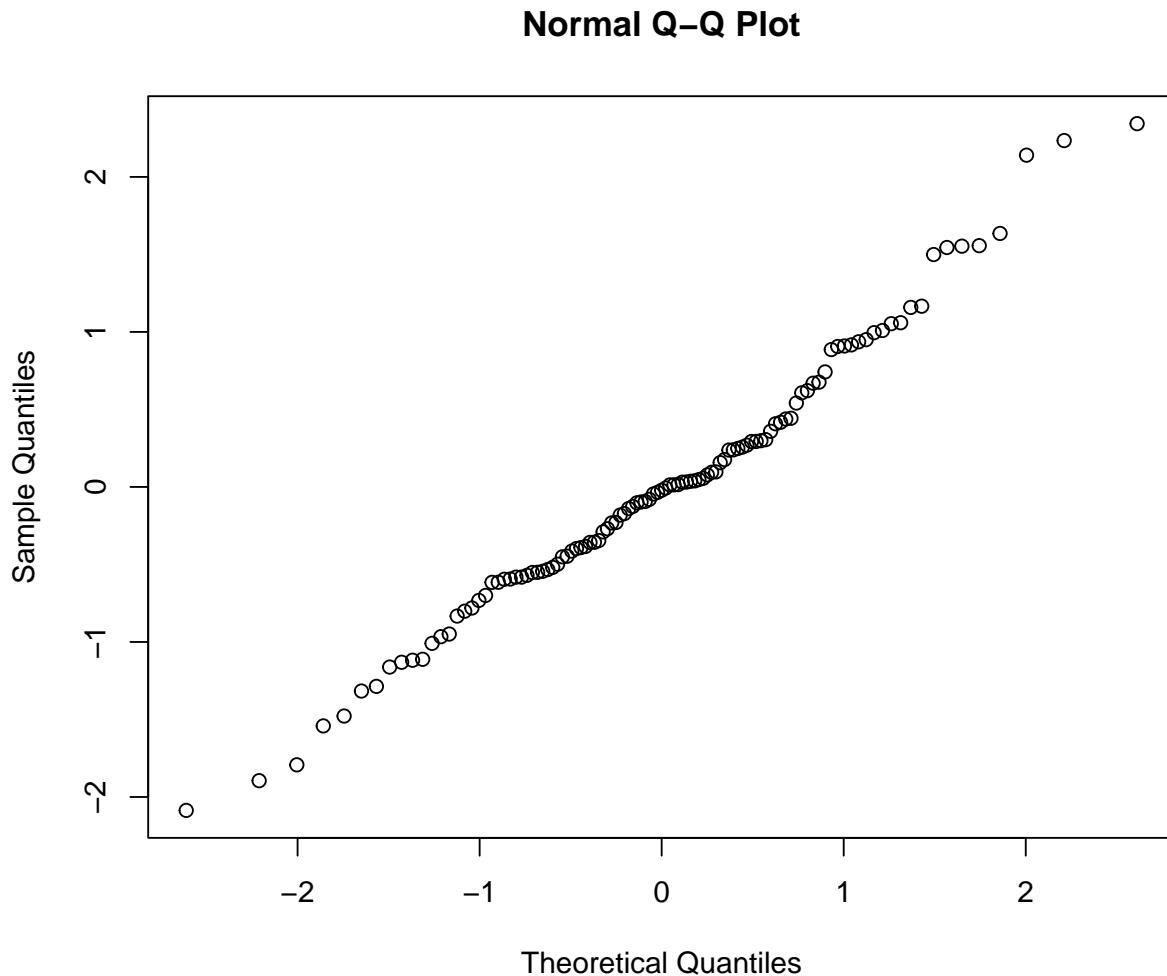
```
# creates object of residuals
resid_neg <- residuals(TwoWay_neg)
hist(resid_neg)
```



So far so good – these look normal. Let's examine a QQ plot. We want the dots (the 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. If the residuals are normally distributed, they will line up on the diagonal line (within reason).

```
qqnorm(resid_neg)
```



We can formally test the distribution of the residuals with a Shapiro test. We want the 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! $p > 0.05$

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. Our analysis suggested skew and kurtosis were within the bounds considered to be

normally distributed. Further, the Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) suggested that the plotting of the residuals did not differ significantly from a normal distribution ($W = 0.9846$, $p = 0.234$).

8.5.1.2 Homogeneity of variance

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 *leveneTest()* function from the *car* package. Within the function we specify the model we will be testing in 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.

```
car:::leveneTest(Negative ~ Rater * Photo, data = Ramdhani_df)
```

```
Levene's Test for Homogeneity of Variance (center = median)
Df F value    Pr(>F)
group   5 8.6342 0.0000007002 ***
105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's test has indicated a violation of the homogeneity of variance assumption ($F[5, 105] = 8.634$, $p < .001$). This is not surprising. The boxplots shows some widely varying variances.

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.

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 let's us know there is a more nuanced/complex story.

In specifying the ANOVA, order of entry matters. If you have a distinction between IV and moderator, put the IV first.

```
TwoWay_neg <- aov(Negative ~ Rater * Photo, Ramdhani_df)
summary(TwoWay_neg)
```

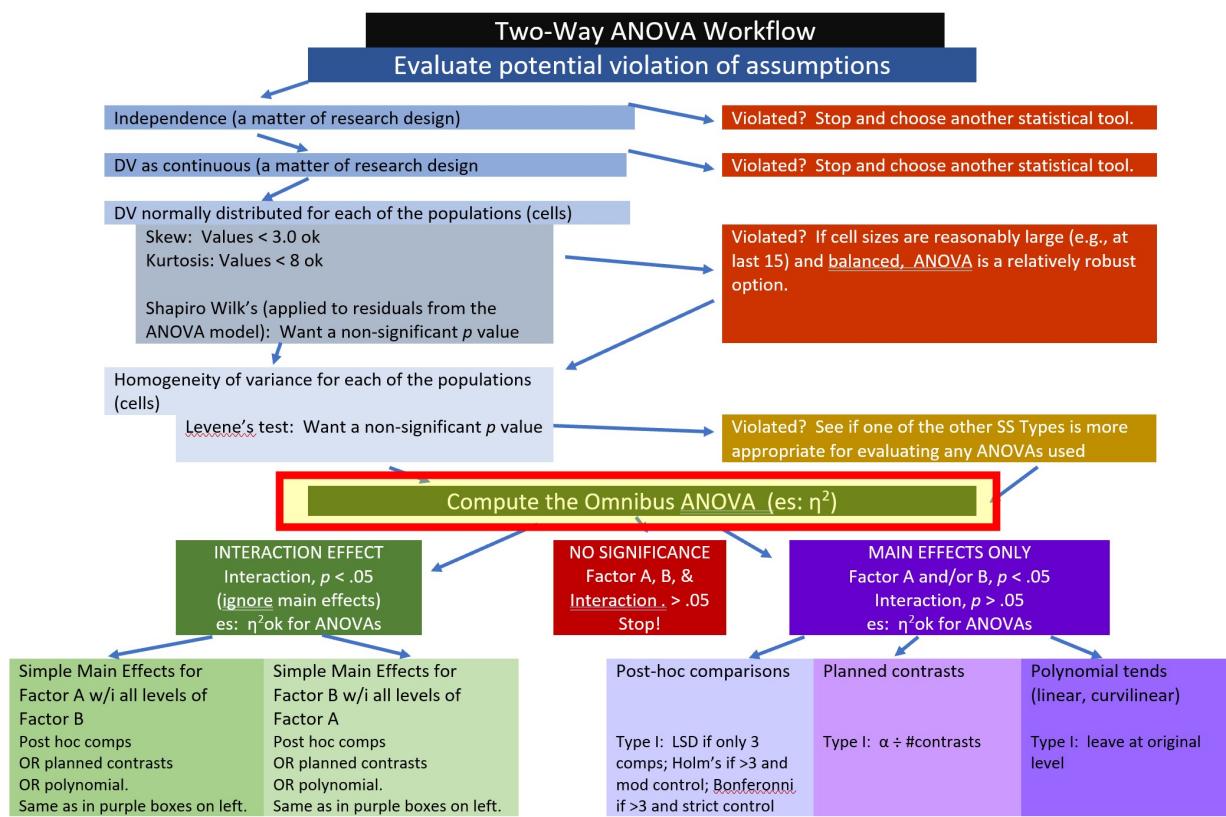


Figure 8.6: Image our place in the Two-Way ANOVA Workflow.

```
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
Dayaknese	2.492	3.007	3.299
rep	35.000	37.000	39.000

Photo

	Dayaknese	Madurese
Dayaknese	2.575	3.301
rep	54.000	57.000

Rater:Photo

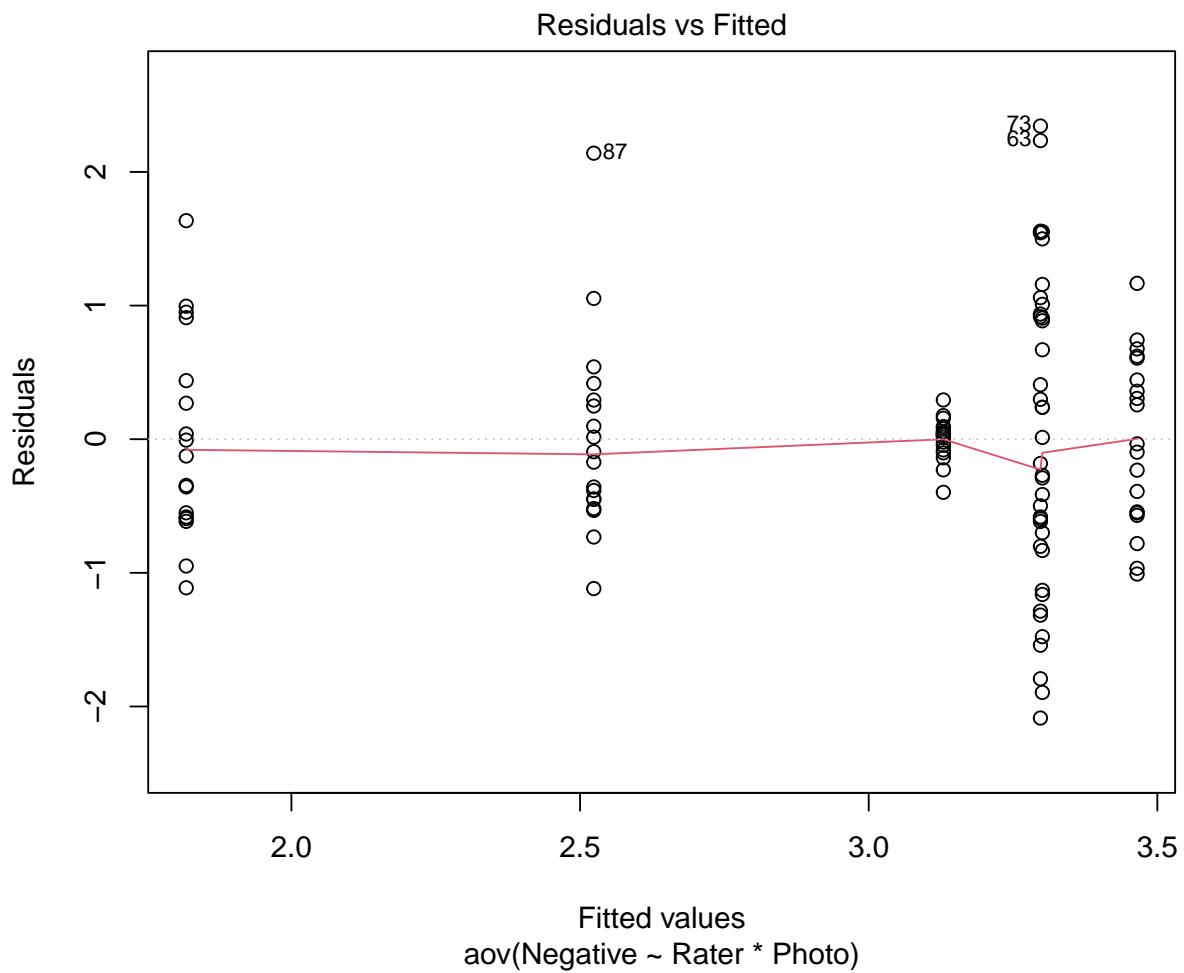
	Photo	
Rater	Dayaknese	Madurese
Dayaknese	1.818	3.129
rep	17.000	18.000
Javanese	2.524	3.465
rep	18.000	19.000
Madurese	3.301	3.298
rep	19.000	20.000

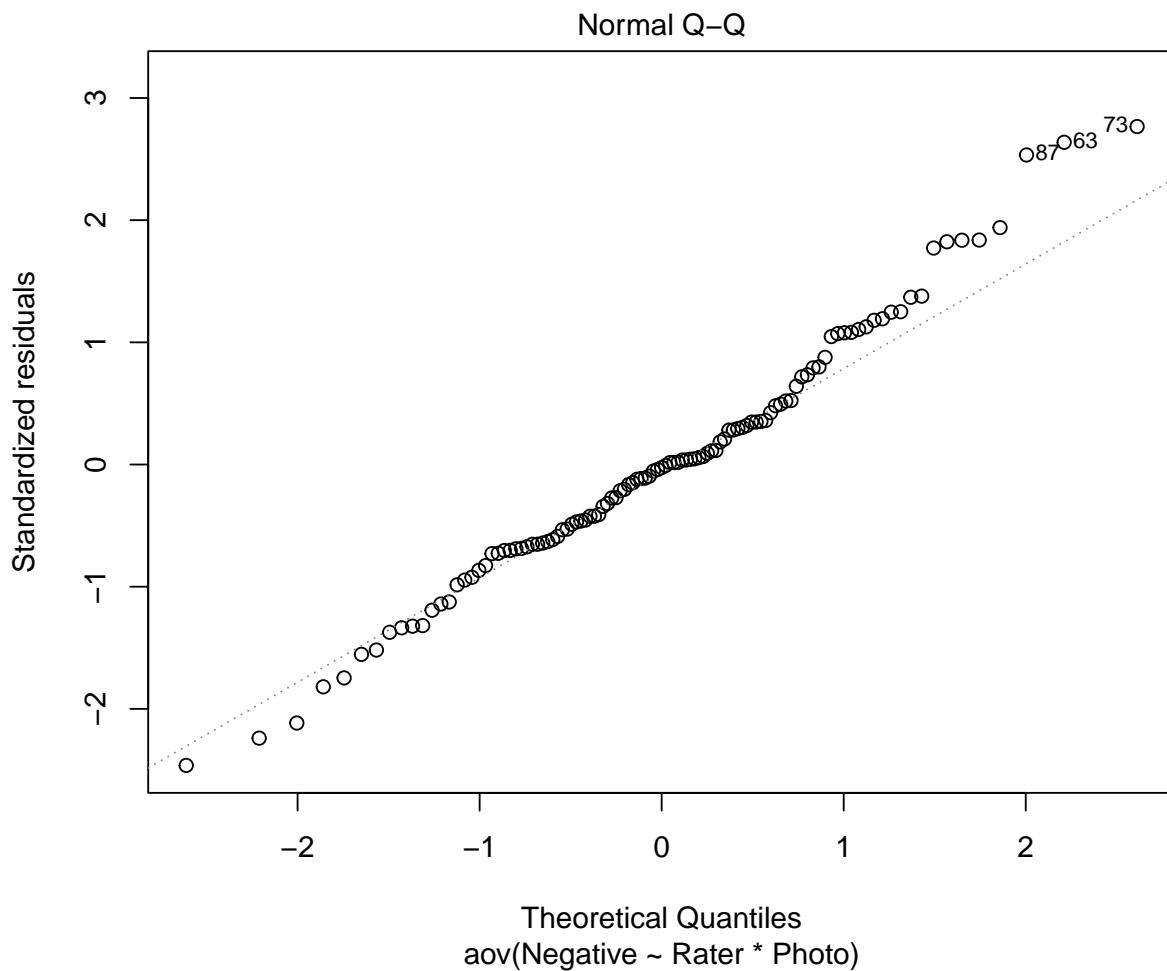
Let's write the *F strings** from the above table.

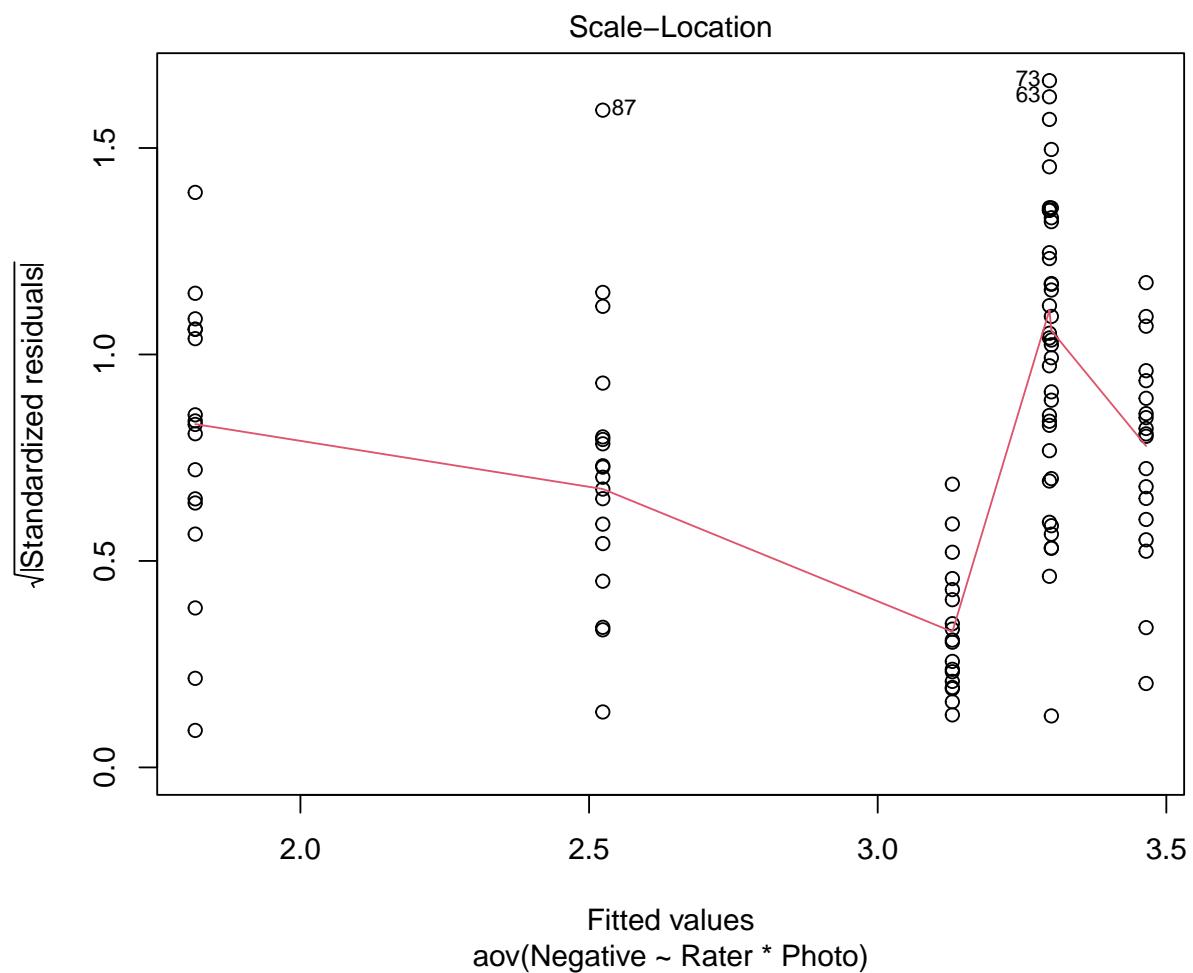
- Rater main effect: ($F[2, 105] = 8.077, p < .001$).
- Photo stimulus main effect: ($F[1, 105] = 19.346, p < .001$).
- Interaction effect: ($F[2, 105] = 5.696, p = .004$).

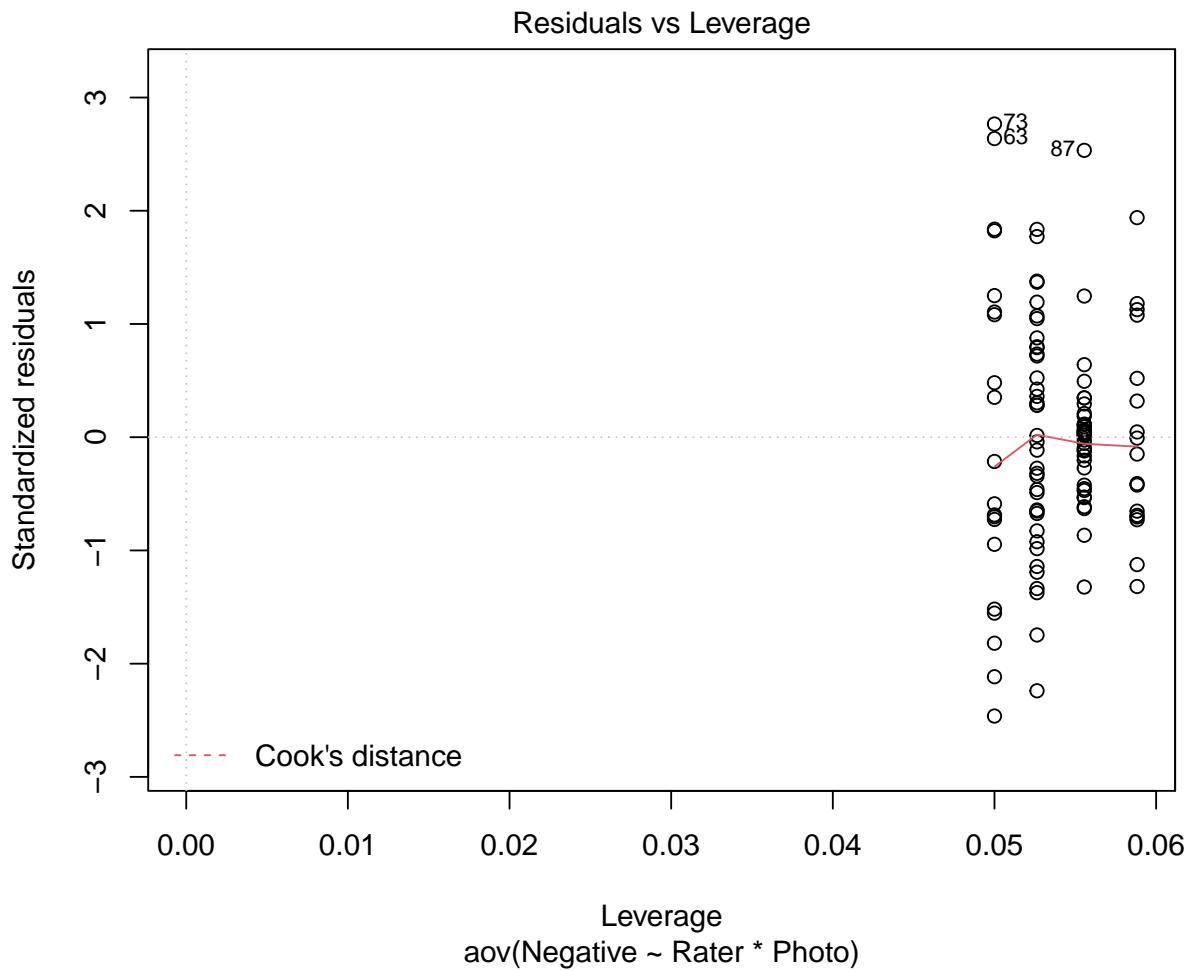
The *plot()* function provides some quick plots from the object created from the ANOVA.

```
plot(TwoWay_neg)
```









8.5.2.1 Effect sizes

Eta squared 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:

$$\eta^2 = \frac{SS_M}{SS_T}$$

We can apply the *etaSquared()* function from the *lsr* package to our ANOVA object to retrieve η^2 .

```
lsr::etaSquared(TwoWay_neg)
```

	eta.sq	eta.sq.part
Rater	0.10662441	0.13363091
Photo	0.12736755	0.15558329
Rater:Photo	0.07500609	0.09788289

We can update our F strings to include the effect size:

- Rater main effect: ($F[2, 105] = 8.077, p < .001, \eta^2 = 0.107$).
- Photo stimulus main effect: ($F[1, 105] = 19.346, p < .001, \eta^2 = 0.127$).
- Interaction effect: ($F[2, 105] = 5.696, p = .004, \eta^2 = 0.075$).

Before moving to follow-up, let's capture an APA style write-up so far.

8.5.2.2 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. Results of Levene's Test for Equality of Error Variances indicated violation of the homogeneity of variance assumption, ($F[5, 105] = 8.834, p < .001$). Our analysis of the individual cell means (see Table 1 for means and standard deviations) suggested skew and kurtosis were within the bounds considered to be normally distributed [Kline, 2016]. A non-significant Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) provided further evidence that the assumption of normality was not violated ($W = 0.9846, p = 0.234$).

Computing sums of squares with a Type II approach, the results for the ANOVA indicated a significant main effect for ethnicity of the rater ($F[2, 105] = 8.077, p < .001, \eta^2 = 0.107$), a significant main effect for photo stimulus, ($F[1, 105] = 19.346, p < .001, \eta^2 = 0.127$), and a significant interaction effect ($F[2, 105] = 5.696, p = .004, \eta^2 = 0.075$).

Note. The next paragraph will have one of the follow-up options. We will add it later in the lesson

8.5.3 Follow-up 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. I will demonstrate the four I believe to be the most useful in the context of psychologists operating within the scientist-practitioner-advocacy context.

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.

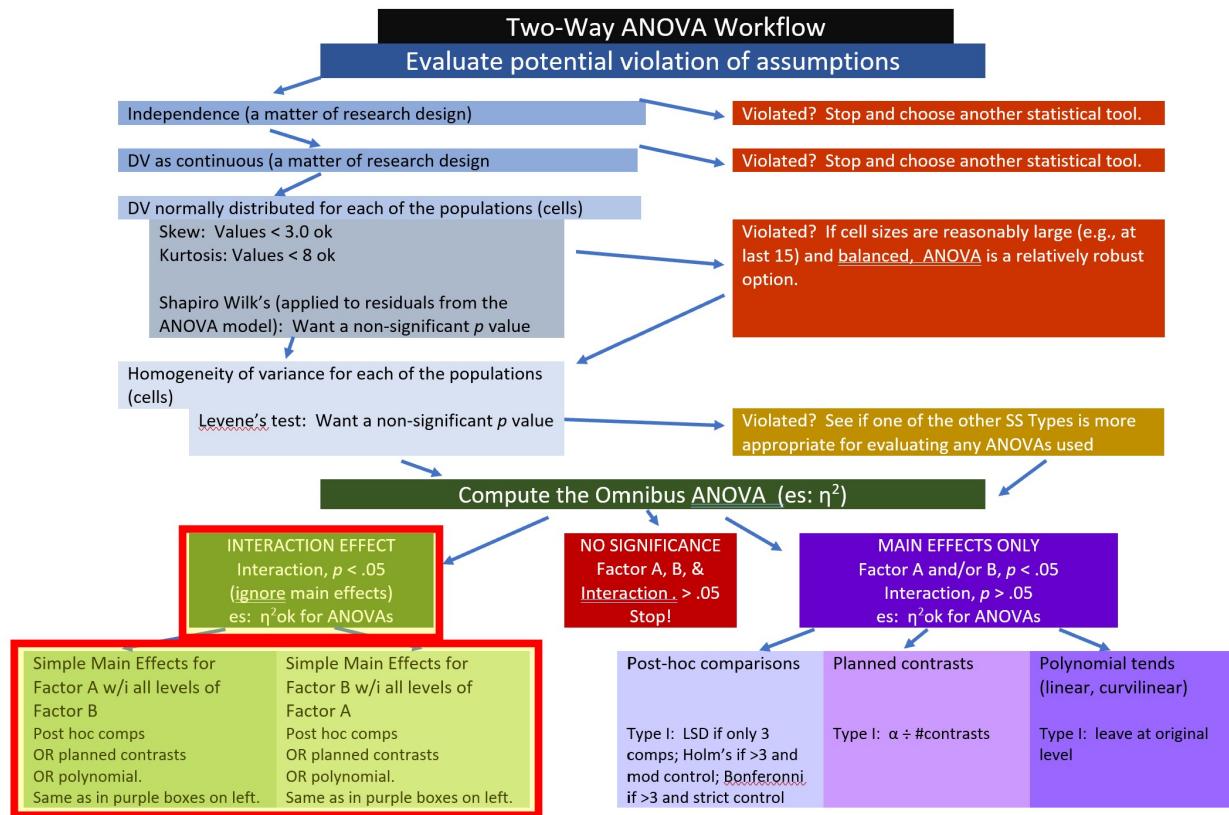


Figure 8.7: Image our place in the Two-Way ANOVA Workflow.

8.5.3.1 Option #1 the simple main effect of photo stimulus within ethnicity of the rater

Here we subset each of the three ethnic groups and then compare their ratings of the two photos. Essentially, we are conducting a one-way ANOVA for the Dyaknese ratings of the Dayaknese and Madurese photos.

```
# subset data
Dayaknese <- subset(Ramdhani_df, Rater == "Dayaknese")
# change df to subset, new model name
Dayaknese_simple <- aov(Negative ~ Photo, data = Dayaknese)
# output for simple main effect
summary(Dayaknese_simple)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Photo	1	15.040	15.040	50.4	0.0000000395 ***
Residuals	33	9.847	0.298		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

```
# effect size for simple main effect can add 'type = 1,2,3,4' to
# correspond with the ANOVA that was run
lsr::etaSquared(Dayaknese_simple, anova = FALSE)
```

```
eta.sq eta.sq.part
Photo 0.6043362 0.6043362
```

Within the Dayaknese ethnic group, there is a statistically significant difference in negative reactions to Dayaknese and Madurese photos: $F(1, 33) = 50.4$, $p < .001$, $\eta^2 = 0.60$.

Next we evaluate photo rating within the Madurese ethnic group.

```
# subset data
Madurese <- subset(Ramdhani_df, Rater == "Madurese")
# change df to subset, new model name
Madurese_simple <- aov(Negative ~ Photo, data = Madurese)
# output for simple main effect
summary(Madurese_simple)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Photo	1	0.00	0.0001	0	0.993
Residuals	37	52.82	1.4275		

```
# effect size for simple main effect can add 'type = 1,2,3,4' to
# correspond with the ANOVA that was run
lsr::etaSquared(Madurese_simple, anova = FALSE)
```

```
eta.sq      eta.sq.part
Photo 0.000002060568 0.000002060568
```

Within the Madurese ethnic group, there was a nonsignificant difference in negative reactions to Dayaknese and Madurese photos: $F(1, 37) = 0.00$, $p = .993$, $\eta^2 < .001$.

```
# subset data
Javanese <- subset(Ramdhani_df, Rater == "Javanese")
# change df to subset, new model name
Javanese_simple <- aov(Negative ~ Photo, data = Javanese)
# output for simple main effect
summary(Javanese_simple)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
Photo	1	8.188	8.188	17.18	0.000205 ***						
Residuals	35	16.678	0.477								

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

```
# effect size for simple main effect can add 'type = 1,2,3,4' to
# correspond with the ANOVA that was run
lsr::etaSquared(Javanese_simple, anova = FALSE)
```

```
eta.sq eta.sq.part
Photo 0.3292776 0.3292776
```

Within the Javanese ethnic group, there was a significant difference in negative reactions to Dayaknese and Madurese photos: $F(1, 35) = 17.18$, $p < .001$, $\eta^2 = 0.33$.

If I were using this approach in a 3 X 2 ANOVA, I would probably not control for Type I error. Why? I only conducted follow-up comparisons to evaluate the simple main effect of photo stimulus within rater ethnicity; that is, I would hold it at alpha = 0.05.

- Photo stimulus (Dayaknese or Madurese) within the Dayaknese ethnic group.
- Photo stimulus (Dayaknese or Madurese) within the Madurese ethnic group.
- Photo stimulus (Dayaknese or Madurese) within the Javanese ethnic group.

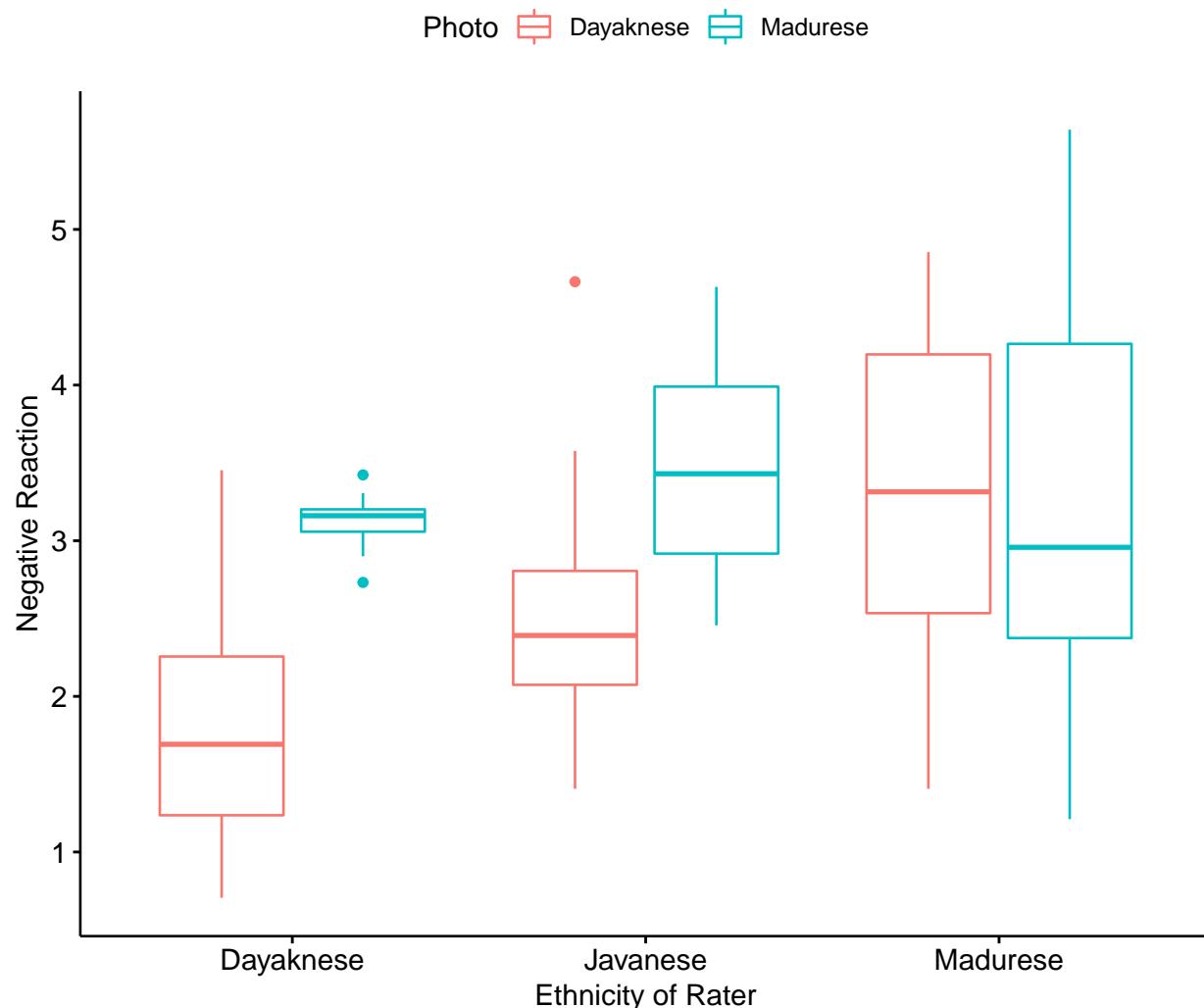
However, because it is good for instruction, it would be equally fine to use a traditional Bonferroni, dividing $.05/3 = 0.017$ and testing each at 0.017. I will use this approach in the write-up.

FAQ: Could we do the reverse simple effect, ethnicity of rater within the photo stimulus? Absolutely! The choice is yours (and sometimes the results will differ). I usually run both and then report ONE – the one that conveys the story the data has to tell. You *could* report both sets, but then you would really want to control Type I error and your repetitive contrasts are far from independent/orthogonal.

APA Style Results for Option #1 follow-up. This would be added to the results of the omnibus two-way ANOVA.

Option #1: 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, we looked at the effect of the photo stimulus within the Dayaknese, Madurese, and Javanese groups, separately. 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 ($F [1, 33] = 50.4, p < .001, \eta^2 = 0.60.$) and Javanese ethnic groups ($F [1, 35] = 17.18, p < .001, \eta^2 = 0.33$), but not for the Madurese ethnic group ($F [1, 37] = 0.000, p = .993, \eta^2 < .001$). 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.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction")
```



8.5.3.2 Option #2 the simple main effect of ethnicity of rater within photo stimulus.

In this simple main effect of ethnicity of rater (3 levels) within photo stimulus (2 levels), we will conduct two one-way ANOVAs for the Dayaknese and Madurese photos, separately. However, we will want to do orthogonal contrast-coding for rater ethnicity for the follow-up (to the follow-up).

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

Let's create custom contrasts. Recall that an orthogonal contrast requires that there be one less contrast than the number of groups and that once a group is singled out, it cannot be compared again.

Thus, I want to compare the

- Javanese to the Dayaknese and Madurese combined, then
- Dayaknese to Madurese

```
# 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

Yes, in contrast 1 we are comparing the Javanese to the combined Dayaknese and Madurese. In contrast 2 we are comparing the Dayaknese to the Madureses.

```
# subset data
Dayaknese_Ph <- subset(Ramdhani_df, Photo == "Dayaknese")
# change df to subset, new model name
Dykn_simple <- aov(Negative ~ Rater, data = Dayaknese_Ph)
# output for simple main effect
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
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# 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
  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
```

The simple main effect of ethnicity of the rater within the reaction to the photos of members of the 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 ($F [1, 51] = 26.554, p < .001$)

We repeat the simple main effect process when the Madurese photos were the stimulus.

```
# 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 of ethnicity of the rater within rating the photos of Madurese people 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 ($F [1, 54] = 0.349, p = .557$)

To control for Type I error, we have 4 follow-up contrasts (2 for Dayaknese, 2 for Madurese). We'll control Type I error with $.05/4 = .0125$

0.05/4

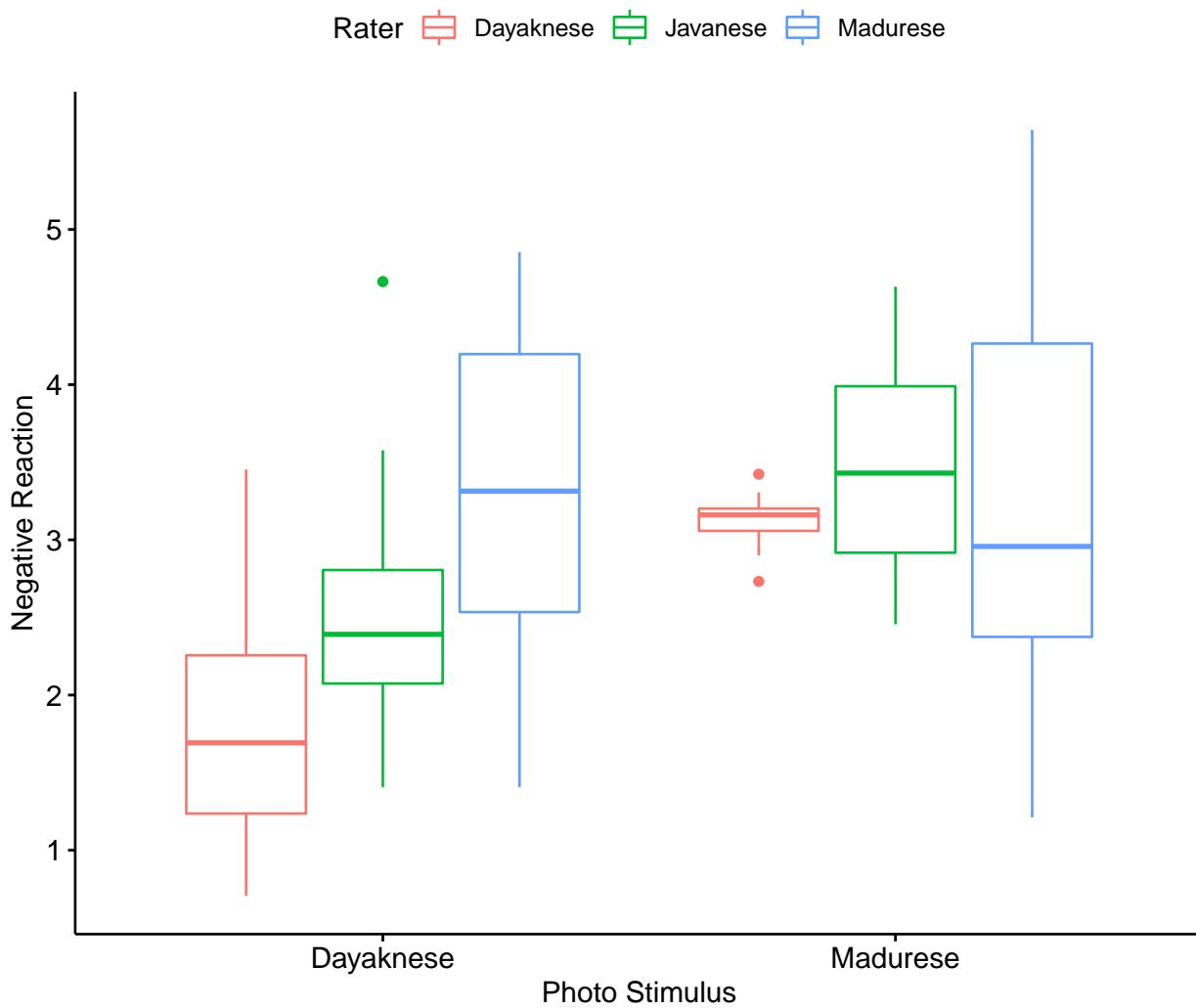
[1] 0.0125

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.

Option #2: To explore the interaction effect, we followed with tests of simple effect of rater ethnicity within the photo stimulus. That is, we looked at the effect of 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 of ethnicity of the rater within the reaction to the photos of members of the 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 of ethnicity of the rater within when rating the photos of Madurese people 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.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Photo", y = "Negative", color = "Rater",
  xlab = "Photo Stimulus", ylab = "Negative Reaction")
```



8.5.3.3 Option #3 post hoc comparisons

Another option is compare all possible cells. These are termed *post hoc comparisons*. They are an alternative to simple main effects; you would not report both. 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: $k(k-1)/2$

```
6 * (6 - 1)/2
```

```
[1] 15
```

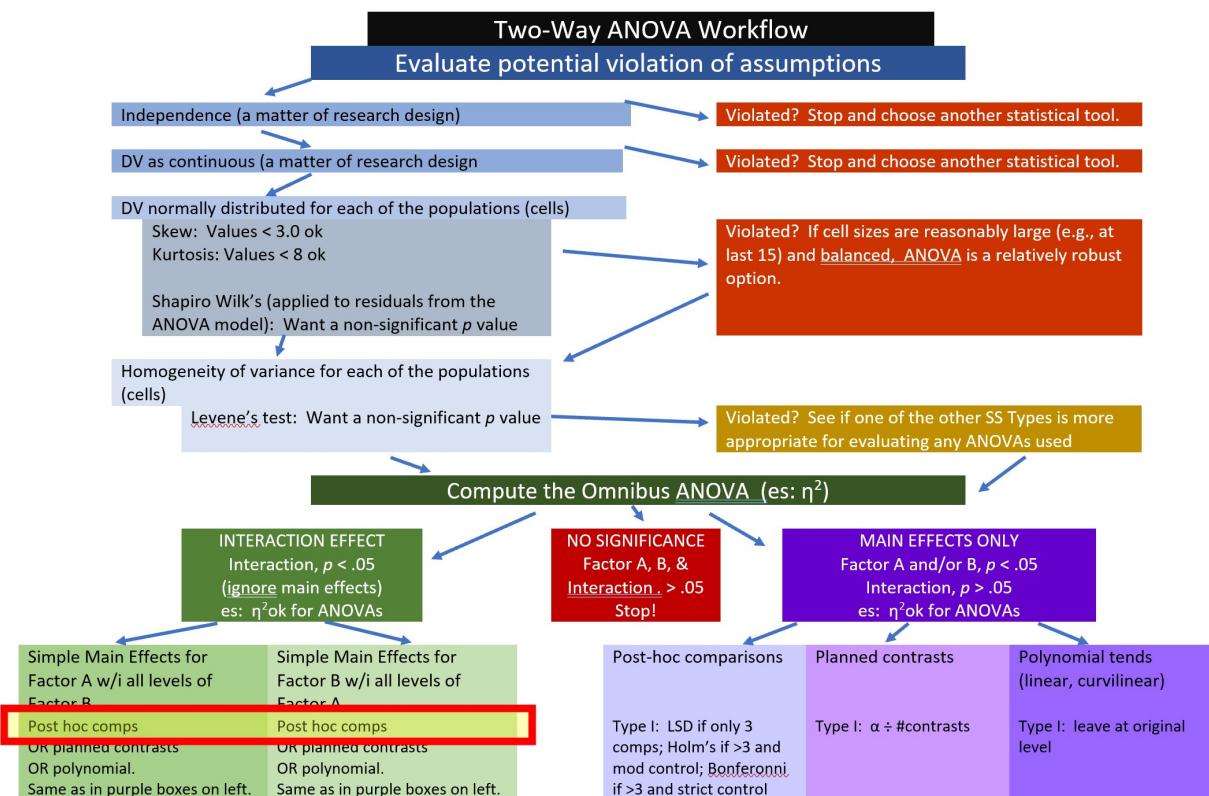


Figure 8.8: Image our place in the Two-Way ANOVA Workflow.

Obtain the 15 post-hoc paired comparisons with the *TukeyHSD()* function.

```
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, 2014b] 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" "Dayaknese:Madurese-Dayaknese:Javanese" "Javanese:Dayaknese-Dayaknese:Javanese" "Javanese:Dayaknese:Dayaknese-Dayaknese" "Javanese:Dayaknese:Dayaknese:Javanese" "Javanese:Dayaknese:Javanese-Dayaknese" "Javanese:Dayaknese:Javanese:Dayaknese" "Javanese:Madurese-Dayaknese-Dayaknese" "Javanese:Madurese-Dayaknese:Javanese" "Javanese:Madurese:Dayaknese-Dayaknese" "Javanese:Madurese:Dayaknese:Javanese" "Javanese:Madurese:Javanese-Dayaknese" "Javanese:Madurese:Javanese:Dayaknese" "Javanese:Javanese-Dayaknese-Dayaknese" "Javanese:Javanese-Dayaknese:Javanese" "Javanese:Javanese:Dayaknese-Dayaknese" "Javanese:Javanese:Dayaknese:Javanese"
```

```

... .$. : 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.

	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	1.00000000

Figure 8.9: Image of the results of the Holms sequential Bonferroni.

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:

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	1.00000000

Figure 8.10: Image of the results of the Holms sequential Bonferroni.

8.5.3.4 Option #4 polynomial trends

In the context of the significant interaction effect, we might also be interested in polynomial trends for any simple main effects where 3 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). Below is a figure of where the polynomial test of an interaction effect may fall in the process.

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 is none, low, high) are more readily suited for this approach.

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 (separately for the photos of the Dayaknese and Madurese). We conduct these separately for Dayaknese, Madurese, and Javanese groups.

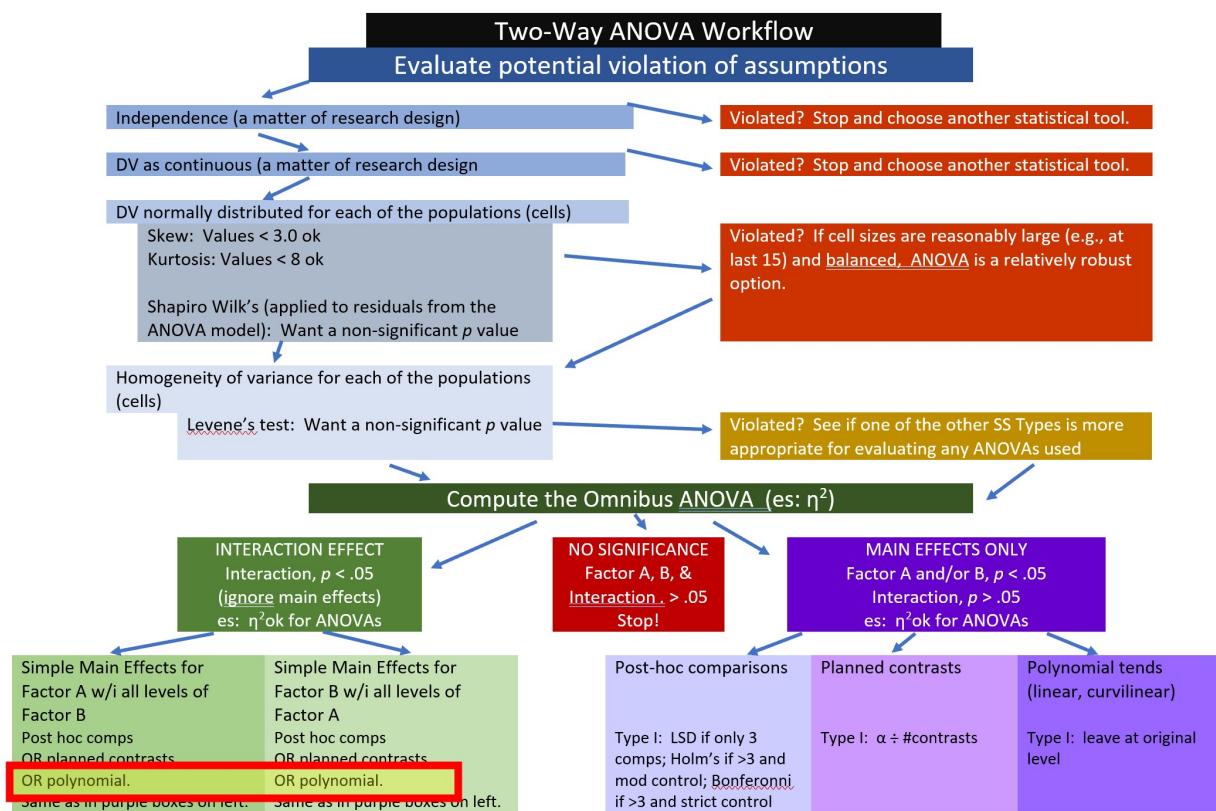


Figure 8.11: Image our place in the Two-Way ANOVA Workflow.

In the event that more than one polynomial trend select the higher one. For example, if both linear and quadratic are selected, interpret the quadratic trend

```
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 < 0.000000000000002 *** 
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
```

Results of a polynomial trend analysis indicated a statistically significant linear trend for evaluation of the Dayaknese photos across the three raters $t(51) = 5.153$, $p < .001$.

```
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 < 0.000000000000002 *** 
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 analysis were non-significant when ethnicity of the rater was evaluated when rating Madurese photos.

8.6 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. (*And, 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.

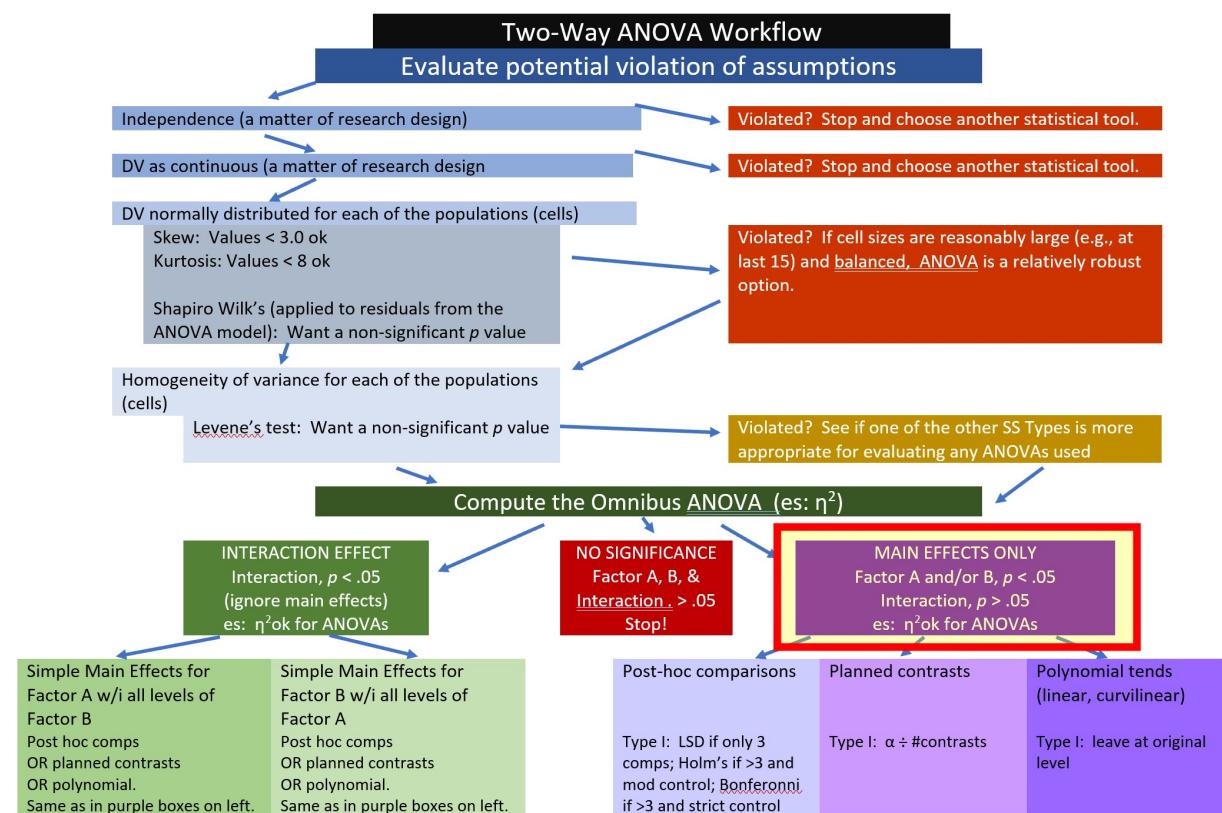


Figure 8.12: Image our place in the Two-Way ANOVA Workflow.

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.

If the main effect has three or more levels (e.g., ethnicity of rater with Dayaknese, Madurese, Javanese levels), then you would follow-up with one or more of the myriad of options. In this class we have focused on three:

- planned contrasts
- posthoc comparisons (all possible cells)
- polynomial

I will demonstrate how to do each as follow-up to a *pretend* scenario where a main effect (but not an interaction) had been significant. I will write up the portion that would be inserted in an APA style results section.

Essentially, we treat these main effect analyses as the follow-up to a significant one-way ANOVA evaluating, in our case, the ethnicity of the Rater.

```
RaterMain <- aov(Negative ~ Rater, data = Ramdhani_df) #DV ~ IV I say, 'DV by IV'
model.tables(RaterMain) #ANOVA output
```

Tables of effects

Rater	Dayaknese	Javanese	Madurese
	-0.4551	0.05971	0.3518
rep	35.0000	37.00000	39.0000

```
summary(RaterMain)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
Rater	2	12.21	6.103	6.426	0.00231 **						
Residuals	108	102.57	0.950								

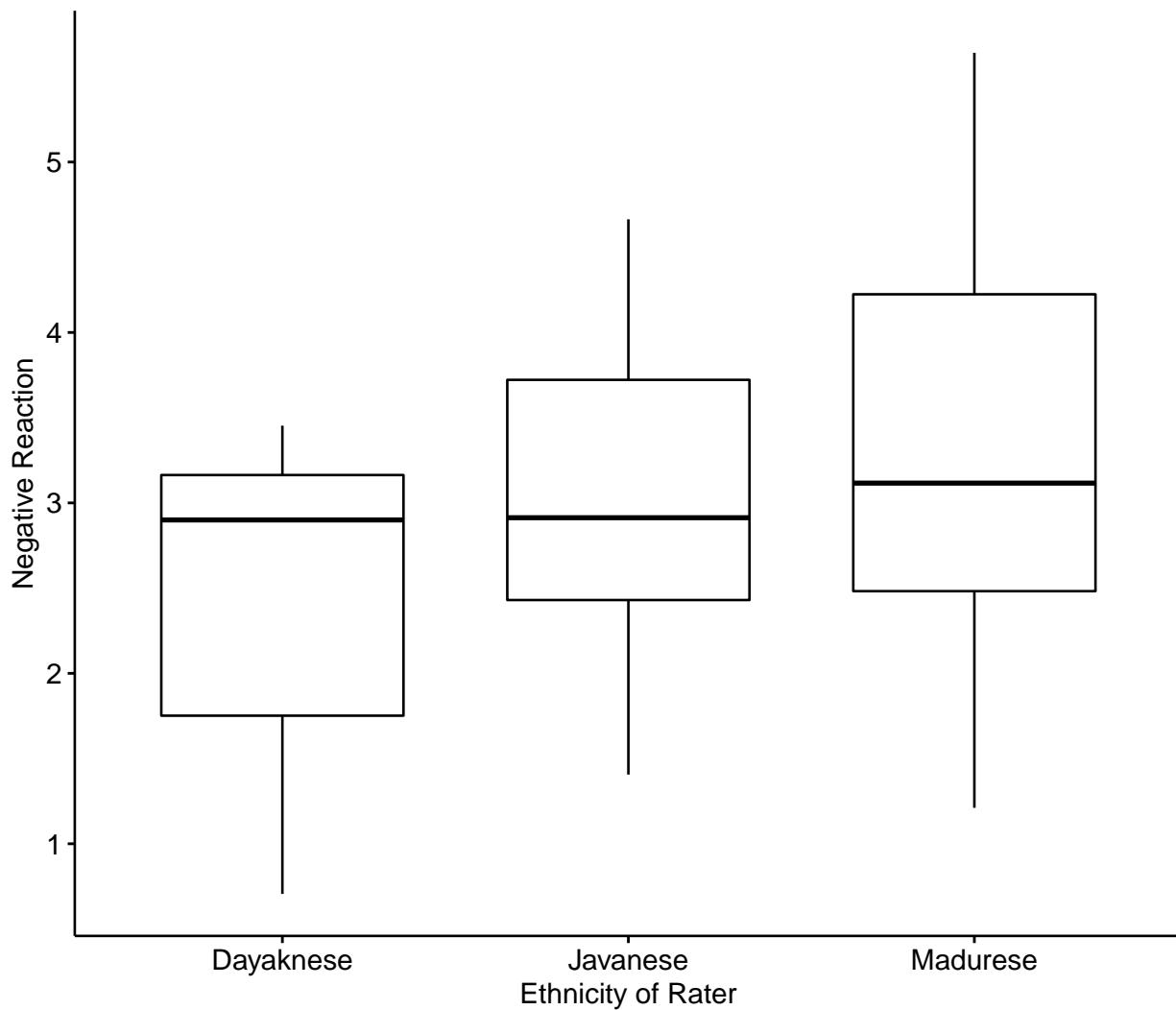
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

```
lsr::etaSquared(RaterMain)
```

	eta.sq	eta.sq.part
Rater	0.1063485	0.1063485

A boxplot representing this main effect may help convey how the main effect of Rater (collapsed across Photo) is different than an interaction effect.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative", xlab = "Ethnicity of Rater",
  ylab = "Negative Reaction")
```



8.6.1 Follow-up with all Post-Hocs

An easy possibility is to follow-up with all possible post-hocs. In the main effect case, these are far simpler than where we conducted all possible posthocts for the interaction effect (remember the Holms sequential Bonferroni?).

Here is a reminder of our location on the workflow.

The *TukeyHSD()* function produces posthoc comparisons by providing the mean difference, a 95% confidence interval of those differences, and the associated *p* value.

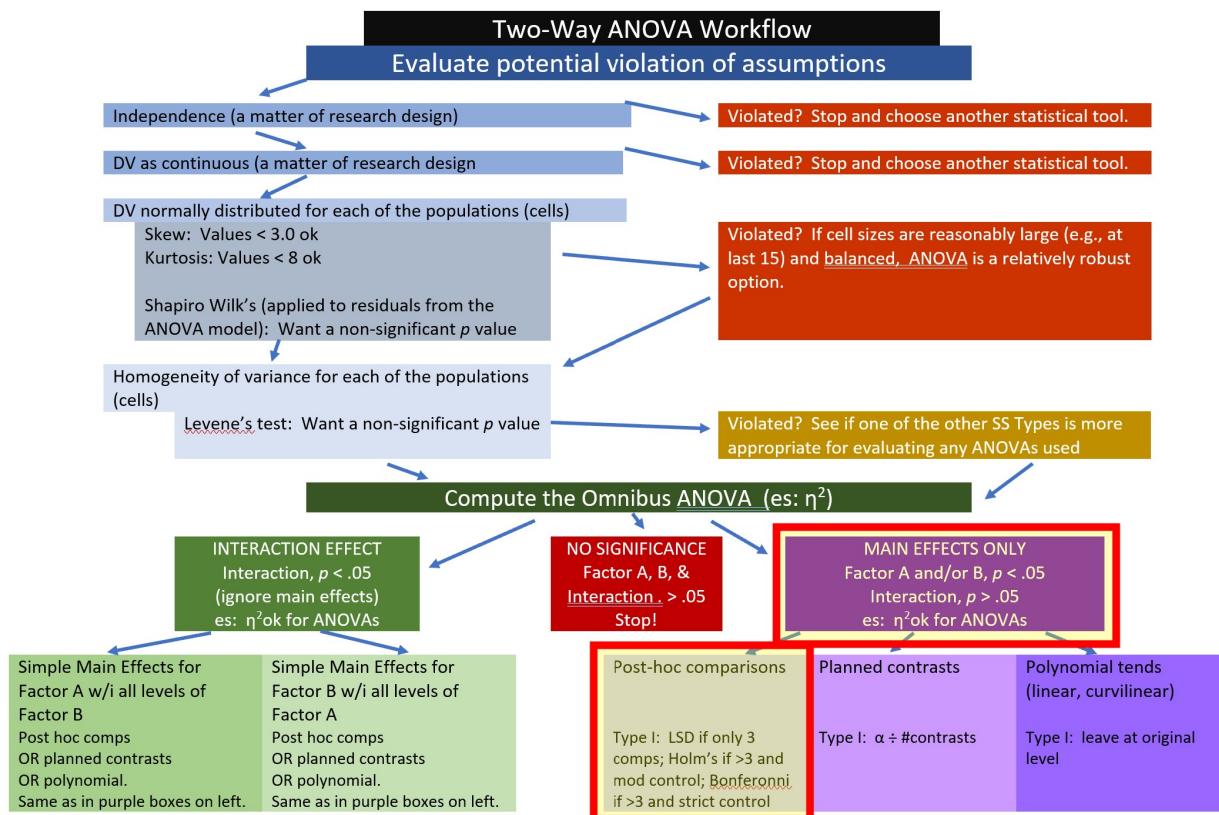


Figure 8.13: Image our place in the Two-Way ANOVA Workflow.

```
TukeyHSD(RaterMain, ordered = TRUE)
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
factor levels have been ordered
```

```
Fit: aov(formula = Negative ~ Rater, data = Ramdhani_df)
```

```
$Rater
```

	diff	lwr	upr	p	adj
Javanese-Dayaknese	0.5147954	-0.03128503	1.0608758	0.0690316	
Madurese-Dayaknese	0.8068425	0.26761161	1.3460734	0.0016135	
Madurese-Javanese	0.2920471	-0.23944747	0.8235417	0.3950430	

Results suggest there were statistically significant differences ($p < .05$) between the Madurese and Dayaknese. These differences, though, would have been when rating *all* photos. This analysis disregards the ethnicity portrayed in the photo.

8.6.2 Follow-up with planned 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 *main_c1* and *main_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:

- Contrast #1: comparing the DV for the Javanese rater to the combined Dayaknese and Madurese raters.
- Contrast #2: comparing the DV for the Dayaknese and Madurese raters.

These are orthogonal because:

- there are $k - 1$ comparisons, and
- once a contrast is isolated (i.e., the Javanese rater in contrast #1) it cannot be used again
 - Recall the piece of cake analogy: once you take out a piece of the cake, you really can't put it back in

```
# Contrast1 compares Control against the combined effects of Low and
# High.
main_c1 <- c(1, -2, 1)
```

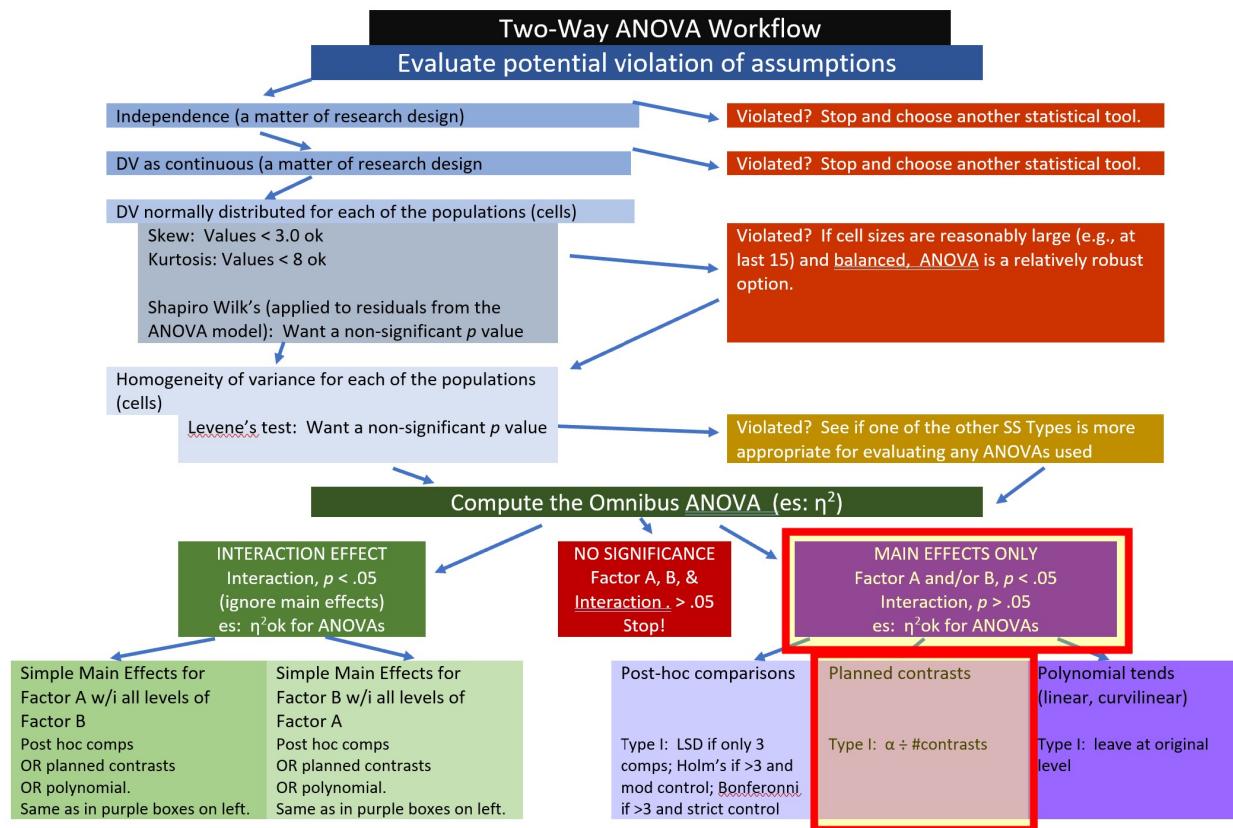


Figure 8.14: Image our place in the Two-Way ANOVA Workflow.

```
# Contrast2 excludes Control; compares Low to High.
main_c2 <- c(-1, 0, 1)
contrasts(Ramdhani_df$Rater) <- cbind(main_c1, main_c2)
contrasts(Ramdhani_df$Rater)
```

	main_c1	main_c2
Dayaknese	1	-1
Javanese	-2	0
Madurese	1	1

Then we run the contrast

```
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 ***
Ratermain_c1	-0.03712	0.06544	-0.567	0.571670
Ratermain_c2	0.40342	0.11345	3.556	0.000561 ***

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

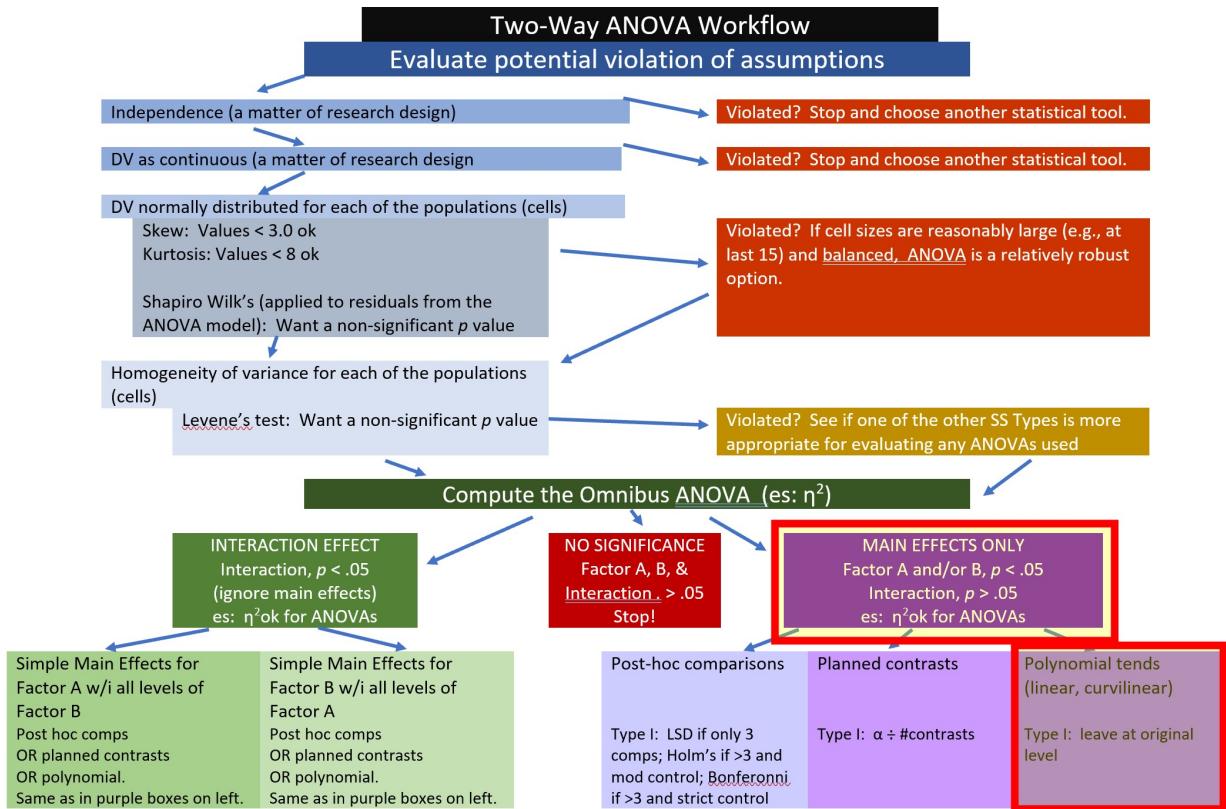
F-statistic: 6.426 on 2 and 108 DF, p-value: 0.002307

```
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, $t(108) = -0.567$, $p = .572$. However, there were significant differences between Dayaknese and Javanese raters, $t(108) = 3.556$, $p < .001$.

8.6.3 Polynomial Trends

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.

```

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		

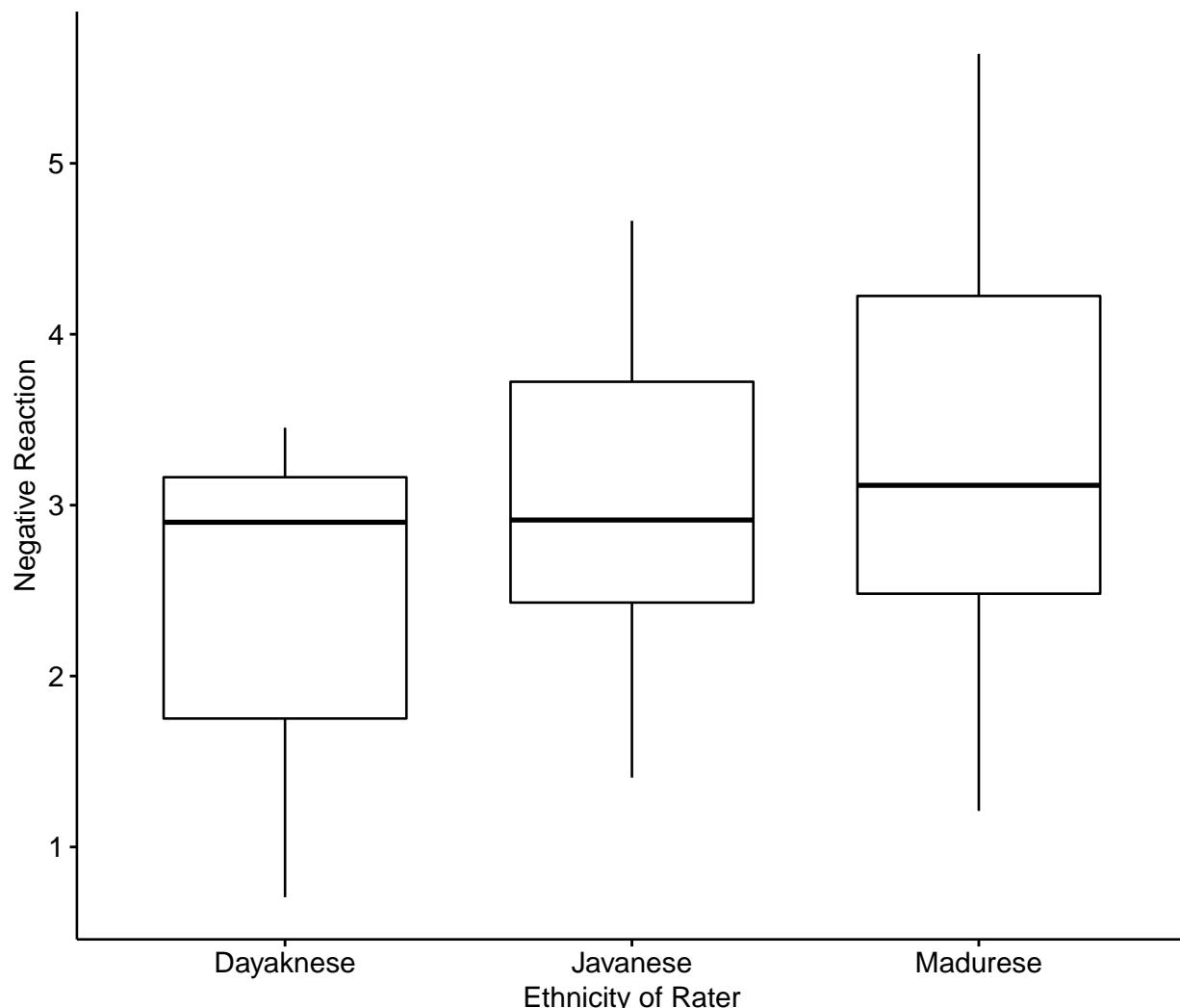
Residual standard error: 0.9745 on 108 degrees of freedom
 Multiple R-squared: 0.1063, Adjusted R-squared: 0.0898
 F-statistic: 6.426 on 2 and 108 DF, p-value: 0.002307

Rater.L tests the data to see if there is a significant linear trend. There is: $t = 3.556, p < .001$.

Rater.Q tests to see if there is a significant quadratic (curvilinear, one hump) trend. There is not: $t = -0.567, p = .572$.

Results supported a significant linear trend ($t = 3.556, p < .001$) such that negative reactions increased in a linear reaction across the three rating groups.

```
ggpubr::ggboxplot(Ramdhani_df, x = "Rater", y = "Negative", xlab = "Ethnicity of Rater",
  ylab = "Negative Reaction")
```



8.7 My APA Style Results Section

First, I am reluctant to term anything “final.” It seems like there is always the possibility or revision. Given that I demonstrated a number of options, let me first show the workflow with the particular path I took:

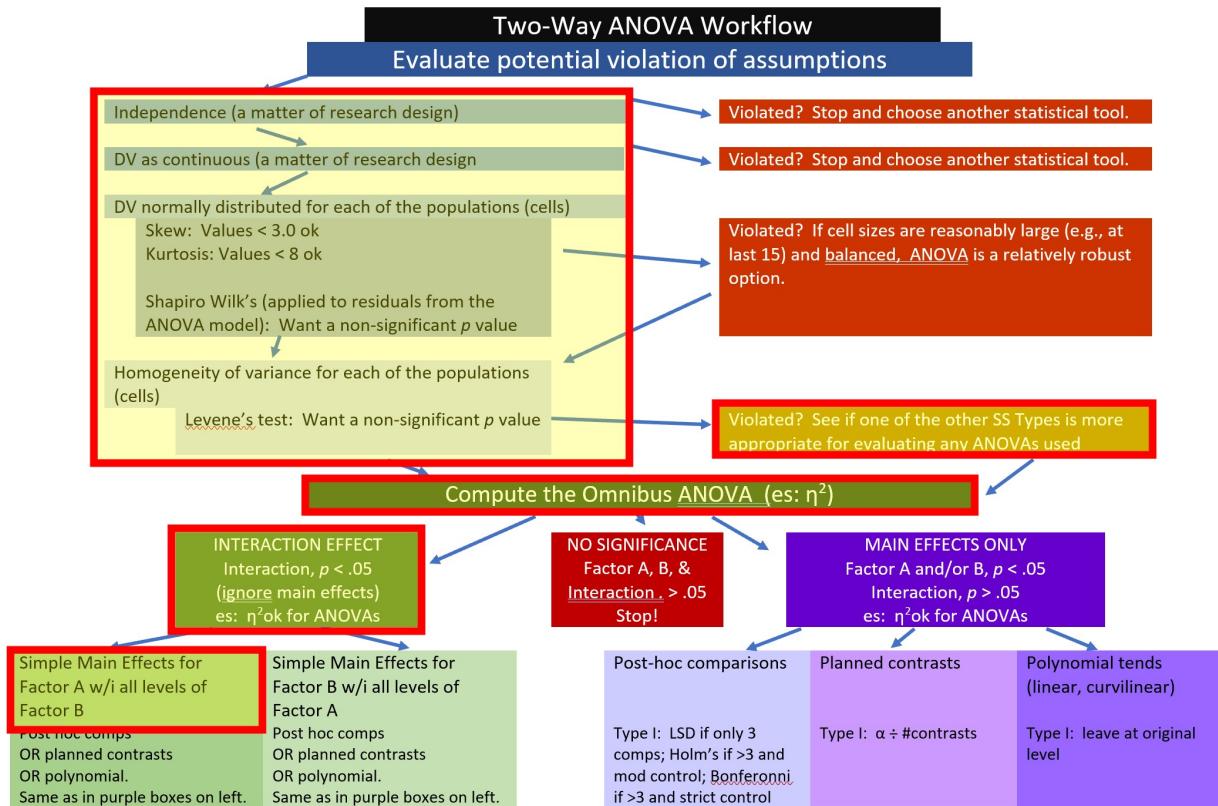


Figure 8.15: Image our place in the Two-Way ANOVA Workflow.

In light of that, here's the final write-up:

Results

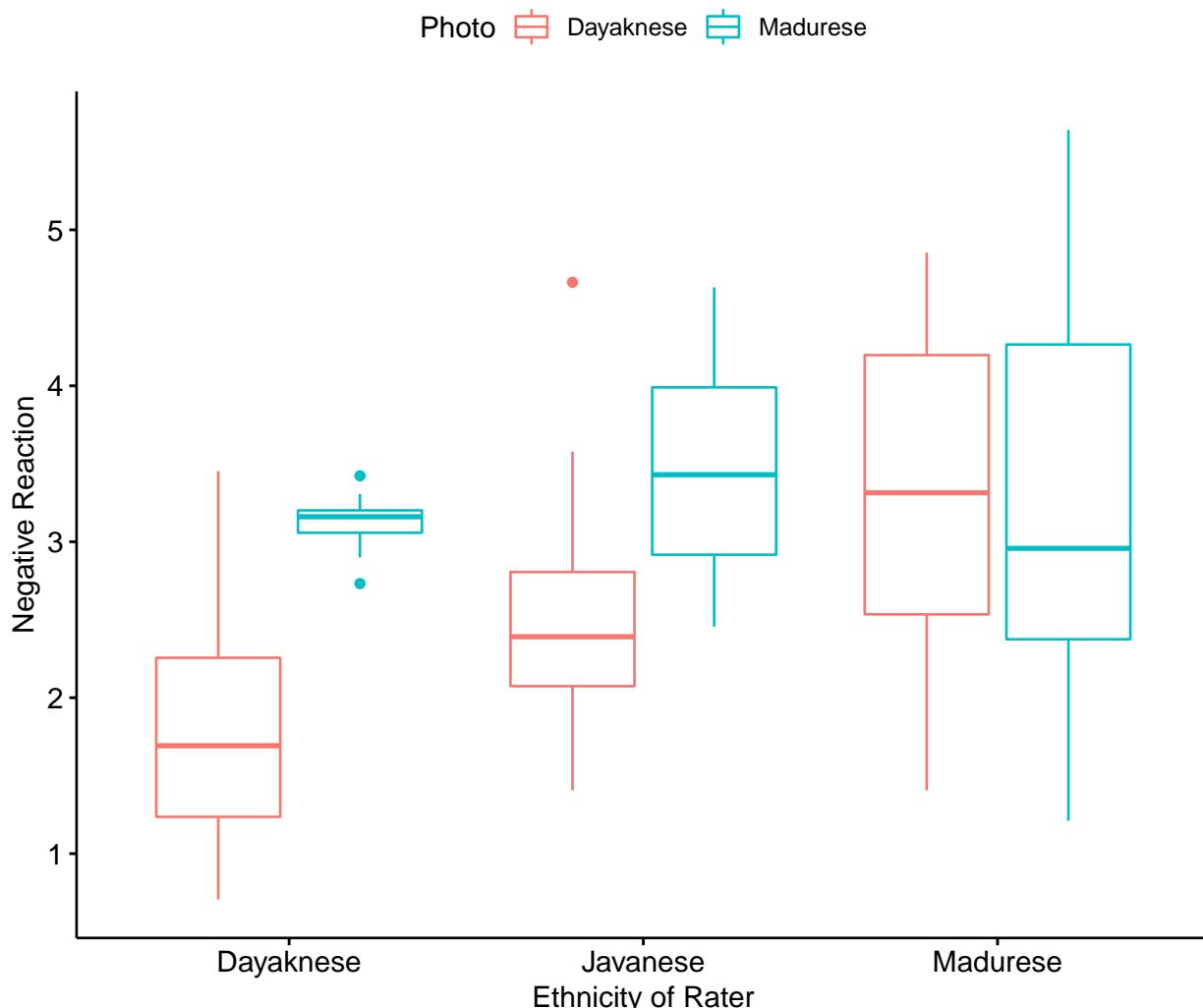
A 3 X 2 ANOVA was conducted to evaluate the effects of rater ethnicity (3 levels, Dayaknese, Madurese, Javanese) and photo stimulus (2 levels, Dayaknese on Madurese,) on negative reactions to the photo stimuli. Results of Levene's test for equality of error variances indicated violation of the assumption, ($F[5, 105] = 8.834, p < .001$). Our analysis of the individual cell means (see Table 1 for means and standard deviations) suggested skew and kurtosis were within the bounds considered to be normally distributed [Kline, 2016]. A non-significant Shapiro-Wilk normality test (applied to the residuals from the factorial ANOVA model) provided further evidence that the assumption of normality was not violated ($W = 0.9846, p = 0.234$).

Computing sums of squares with a Type II approach, the results for the ANOVA indicated a significant main effect for ethnicity of the rater ($F[2, 105] = 8.077, p < .001, \eta^2$

$\eta^2 = 0.107$), a significant main effect for photo stimulus, ($F[1, 105] = 19.346, p < .001, \eta^2 = 0.127$), and a significant interaction effect ($F[2, 105] = 5.696, p = .004, \eta^2 = 0.075$).

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, we looked at the effect of the photo stimulus within the Dayaknese, Madurese, and Javanese groups, separately. 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 ($F [1, 33] = 50.4, p < .001, \eta^2 = 0.60.$) and Javanese ethnic groups ($F [1, 35] = 17.18, p < .001, \eta^2 = 0.33$), but not for the Madurese ethnic group ($F [1, 37] = 0.000, p = .993, \eta^2 < .001$). 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.

```
ggpubr::ggbboxplot(Ramdhani_df, x = "Rater", y = "Negative", color = "Photo",
  xlab = "Ethnicity of Rater", ylab = "Negative Reaction")
```



```
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

		Photo			
		Dayaknese		Madurese	
Rater		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

8.7.1 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. A close look at the figures can explain that with varying variability and means with similar values, this is probable. I find it to be a useful lesson in “what it takes” to get stable, meaningful results.

8.8 Options for Assumption Violations

In one-way ANOVA we could simply apply the Welch's alternative. It's not so easy in factorial ANOVA. One alternative, though, 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: Type I, II, III, and IV. This matters.

SS Type II is the *aov()* default. It may be a best practice to go ahead and specify the SS Type in both the *aov()*, eta-squared, and apaTables script so that they are consistent.

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 R default) 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 R. 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.

Field [2012] suggested that it is safest to stick with Type III sums of squares. We apply the type to the model we create in the initial run. For more information, check out this explanation on [r-bloggers](#).

Many researchers automatically use Type III as the SS type. Today I went with the R 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:

```
# this is what we did
car::Anova(TwoWay_neg)
```

Anova Table (Type II tests)

Response: Negative

```

      Sum Sq Df F value    Pr(>F)
Rater      12.238  2  8.0977  0.0005363 ***
Photo       14.619  1 19.3462  0.00002623 ***
Rater:Photo  8.609  2  5.6964  0.0044803 **
Residuals   79.341 105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
# We change the SS type by applying it to our model.
car::Anova(TwoWay_neg, type = "III")
```

Anova Table (Type III tests)

```

Response: Negative
      Sum Sq Df F value    Pr(>F)
(Intercept) 56.173  1 74.3388 0.00000000000007422 ***
Rater       19.805  2 13.1051 0.00000830116650983 ***
Photo        15.040  1 19.9034 0.00002051829053731 ***
Rater:Photo  8.609  2  5.6964           0.00448 **
Residuals   79.341 105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Note that the sums of squares are somewhat different between models – and that the Type III tests includes an intercept. In today's example, the statistical significance remains the same across the models.

Now let's compare the effect sizes across models.

```
lsr::etaSquared(TwoWay_neg)
```

	eta.sq	eta.sq.part
Rater	0.10662441	0.13363091
Photo	0.12736755	0.15558329
Rater:Photo	0.07500609	0.09788289

```
lsr::etaSquared(TwoWay_neg, type = 3)
```

	eta.sq	eta.sq.part
Rater	0.17255745	0.19975734
Photo	0.13103575	0.15935009
Rater:Photo	0.07500609	0.09788289

The Type III effect size for Rater is higher; the others are quite similar.

8.9 Power

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 (which is our usual question for MRPs and dissertations).

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 f^2)
 - 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 F^2)

Let’s quickly rerun our model to get both the df and calculate the *f* effect value

```
lsr::etaSquared(TwoWay_neg, anova = TRUE)
```

	eta.sq	eta.sq.part	SS	df	MS	F
Rater	0.10662441	0.13363091	12.237770	2	6.1188848	8.097730
Photo	0.12736755	0.15558329	14.618555	1	14.6185546	19.346190
Rater:Photo	0.07500609	0.09788289	8.608791	2	4.3043955	5.696435
Residuals	0.69127788		NA	79.341113	105	0.7556296
		p				
Rater	0.00053629628					
Photo	0.00002622821					
Rater:Photo	0.00448026007					
Residuals	NA					

```
# get the partial eta-square (second number) and dfs
```

If we want to understand power in our analysis, 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.

```
effectsize::eta2_to_f(0.1066) #FactorA -- Rater
```

```
[1] 0.3454265
```

```
effectsize::eta2_to_f(0.1274) #Factor B -- Photo
```

```
[1] 0.3821001
```

8.9.1 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 = 0.345,
f.B = 0.382)
```

```
Balanced two-way analysis of variance power calculation
```

```
a = 3
b = 2
n.A = 37
n.B = 55
sig.level = 0.05
power.A = 0.9974259
power.B = 0.9999996
power = 0.9974259
```

NOTE: power is the minimum power among two factors

At 0.997 (Rater), 0.999 (Photo), and 0.997 (interaction), our power to detect a significant effect for Factor A/Rater and Factor B/Photo was huge!

8.9.2 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.B.: 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 = 0.345, f.B = 0.382,
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, that's 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

8.10 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). 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

8.10.1 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.

8.10.2 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.

8.10.3 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.

8.10.4 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	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 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. (2014). One-Way Repeated Measures Analysis of Variance (Lesson 29). In *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (Seventh 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 [2014b] 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.

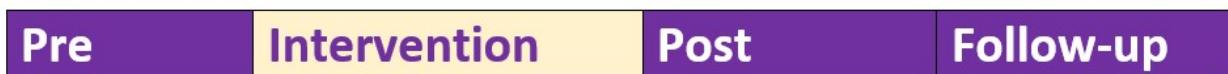


Figure 9.2: Illustration of a research design appropriate for one-way repeated measures ANOVA

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.”

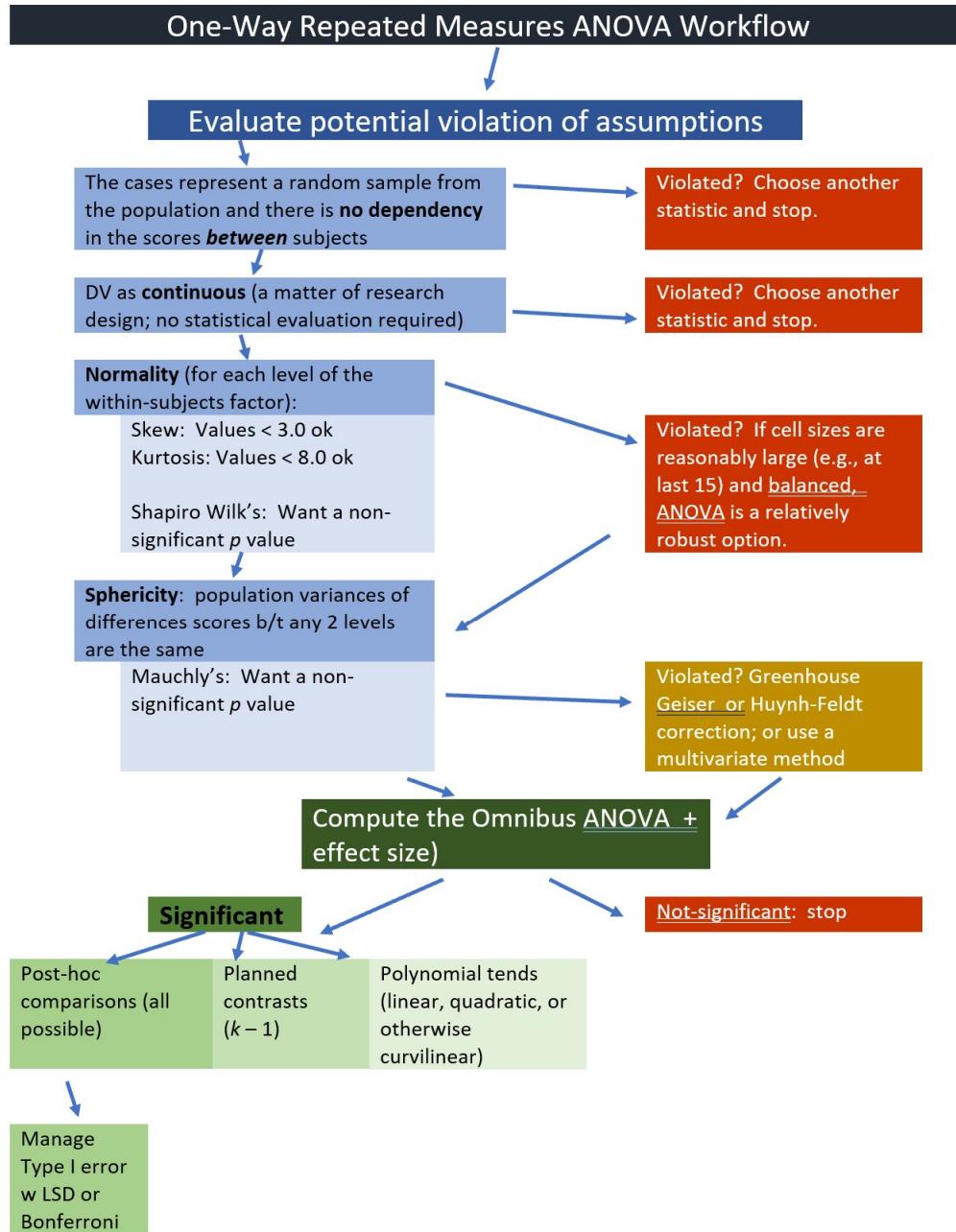


Figure 9.3: Image of a workflow for the one-way repeated measures ANOVA



Figure 9.4: Diagram of the research design for the Amodeo et al study

9.3.1 Code for simulating the data used today.

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 just need to change 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 ...
```

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

As we work the problem we will switch between long and wide formats. Before we get into the statistic let's inspect our data. 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).

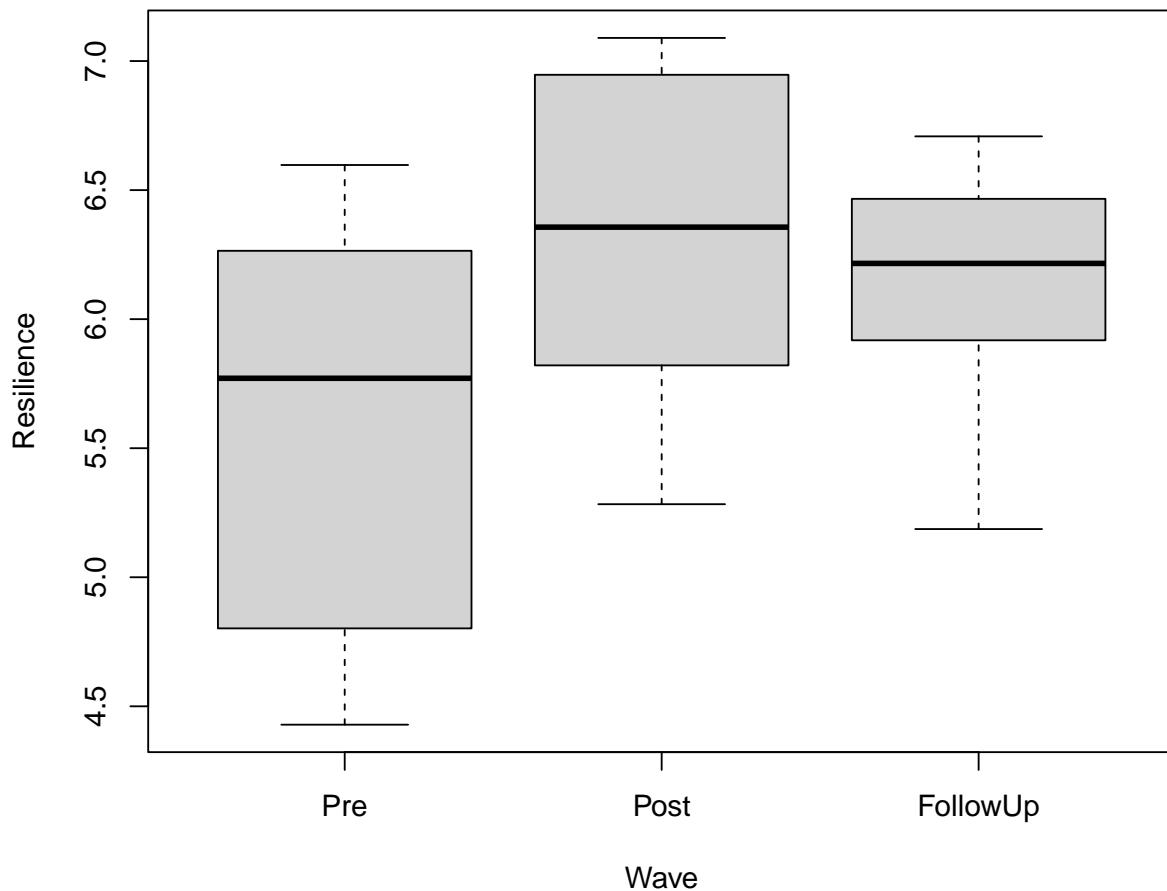
```
psych::describe(Amdeo_wide)
```

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.

```
boxplot(Resilience ~ Wave, data = Amodeo_long, xlab = "Wave", ylab = "Resilience",
       n.label = TRUE)
```



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(Amodeo_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											-0.61
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(Amodeo_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		
<i>df formula</i>	Total <i>#cells-1</i>	Model df <i>#levels-1</i>	Residual <i>df_w - df_M</i>	Between <i>#people-1</i>
SS	11.66		2.36	4.27
df	23		2	14
$MS = SS/df$			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: $F (2, 14) = 3.87, p < .05$

sourcetable, we can move through the steps to calculate our F statistic.

Looking again at our

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 ANOVA with R packages

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.
- multi-level modeling/hierarchical linear modeling
 - This is 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
 - 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.

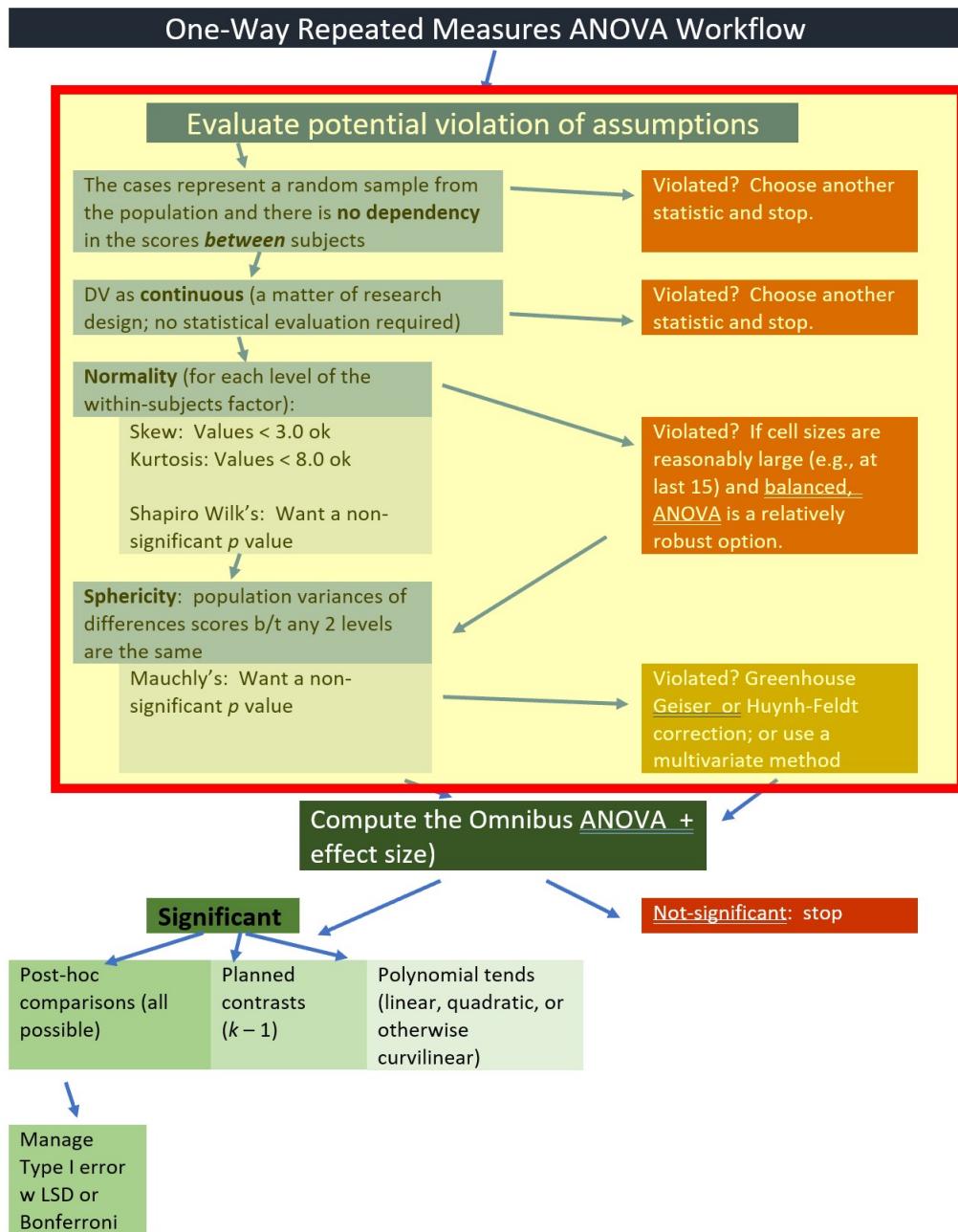


Figure 9.6: Image of our position in the workflow for the one-way repeated measures ANOVA

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 9.7: Demonstration of unequal variances

9.5.1.3 Any 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 bewteen Q3 and Q1 is the *interquartile range* (IQR). Outliers are generally identified when values fall outside these lower and upper boundaries:

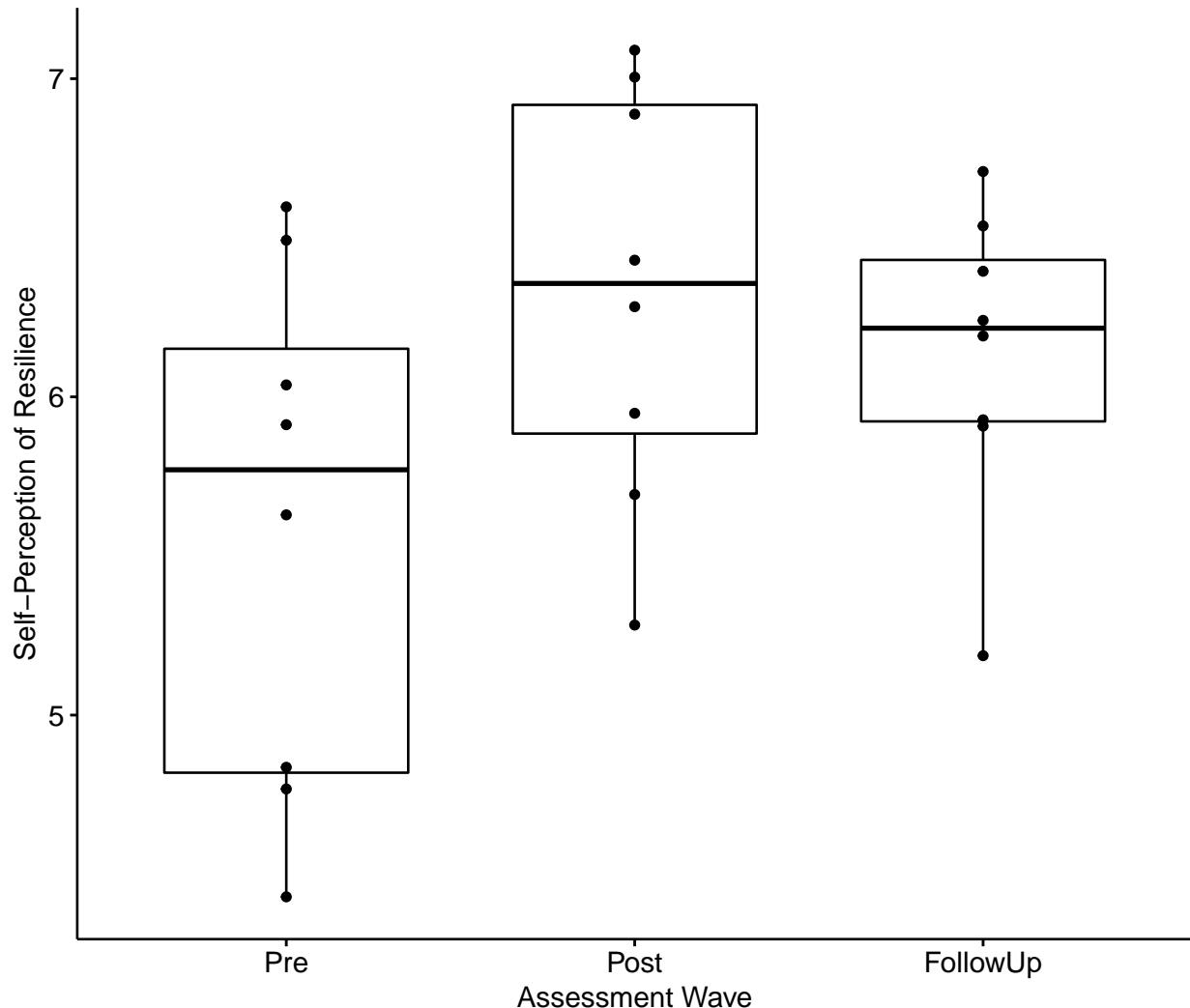
- $\text{Q1} - 1.5 \times \text{IQR}$
- $\text{Q3} + 1.5 \times \text{IQR}$

Extreme values occur when values fall outside these boundaries:

- $\text{Q1} - 3 \times \text{IQR}$
- $\text{Q3} + 3 \times \text{IQR}$

Let's take a look at a boxplot.

```
# Note that we are creating an object (bxp) from our work. This
# script creates the basic boxplot, we will add to it (by using the
# object) later.
bxp <- ggpubr::ggboxplot(Amdeo_long, x = "Wave", y = "Resilience", add = "point",
                           xlab = "Assessment Wave", ylab = "Self-Perception of Resilience")
bxp
```



The package *rstatix* has features that allow us to identify outliers.

```
Amdeo_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.

9.5.1.4 Assessing normality

We can obtain skew and kurtosis values for each cell in our model with the *psych::describeBy()* function.

```
psych::describeBy(Resilience ~ Wave, mat = TRUE, 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	-0.1437061	-1.8118551	0.2905347
Resilience2					5.283057	7.089591	1.806534	-0.2307393	-1.6287654	0.2315959
Resilience3					5.186767	6.708259	1.521491	-0.7204842	-0.6102953	0.1672107

Our skew and kurtosis values fall below the thresholds of concern [Kline, 2016]:

- < 3 for skew
- 8 - 20 indicates extreme skew for kurtosis

We can use the Shapiro-Wilk test for a formal detection of normality. When $p < .05$, it indicates that the distribution is statistically significantly different than a normal one. Therefore, $p > .05$ indicates we did not violate the normal distribution assumption. The code below groups the DV by wave so that we can test normality for each cell in the model.

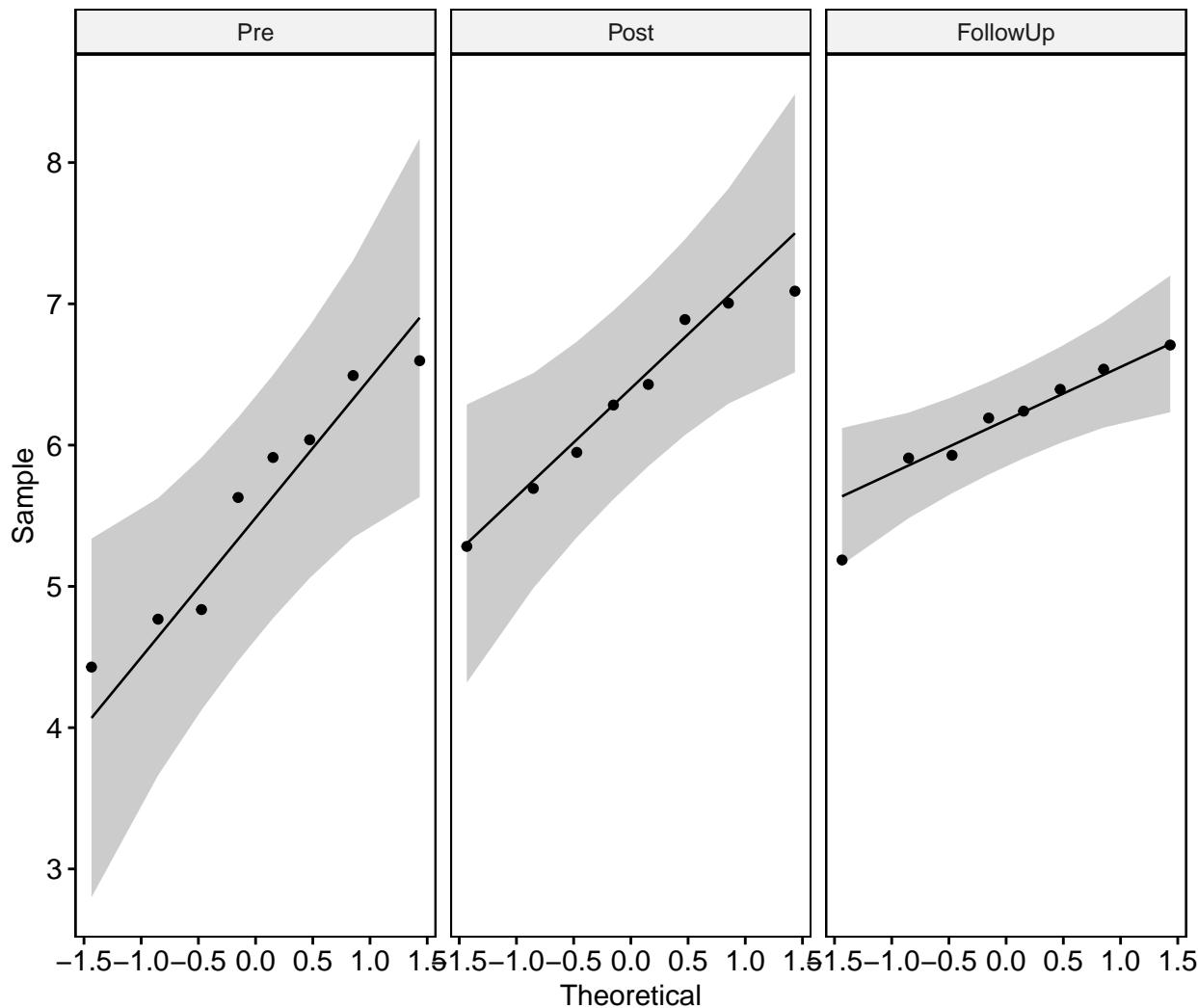
```
Amodeo_long %>%
  group_by(Wave) %>%
  rstatix::shapiro_test(Resilience)
```

```
# 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 Shapiro-Wilk test is sensitive to sample size [Datanovia, b]. Samples > 50 may lead to p values that are $< .05$, even when non-normality is not problematic. Therefore a second check with a Q-Q plot can be helpful. In a normal distribution the residuals will align along the line for each of the cells in the model.

```
ggpubr::ggqqplot(Amodeo_long, "Resilience", facet.by = "Wave")
```



APA Assumption Write-up So Far

Repeated measures ANOVA has several assumptions regarding outliers, normality, and sphericity. 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. Regarding normality, no values of skew and kurtosis (at each wave of assessment) fell within cautionary ranges for skew and kurtosis [Kline, 2016]. Additionally, the Shapiro-Wilk tests applied at each wave of the design were non-significant.

9.5.1.5 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 Omnibus Repeated Measures ANOVA

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.

```
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

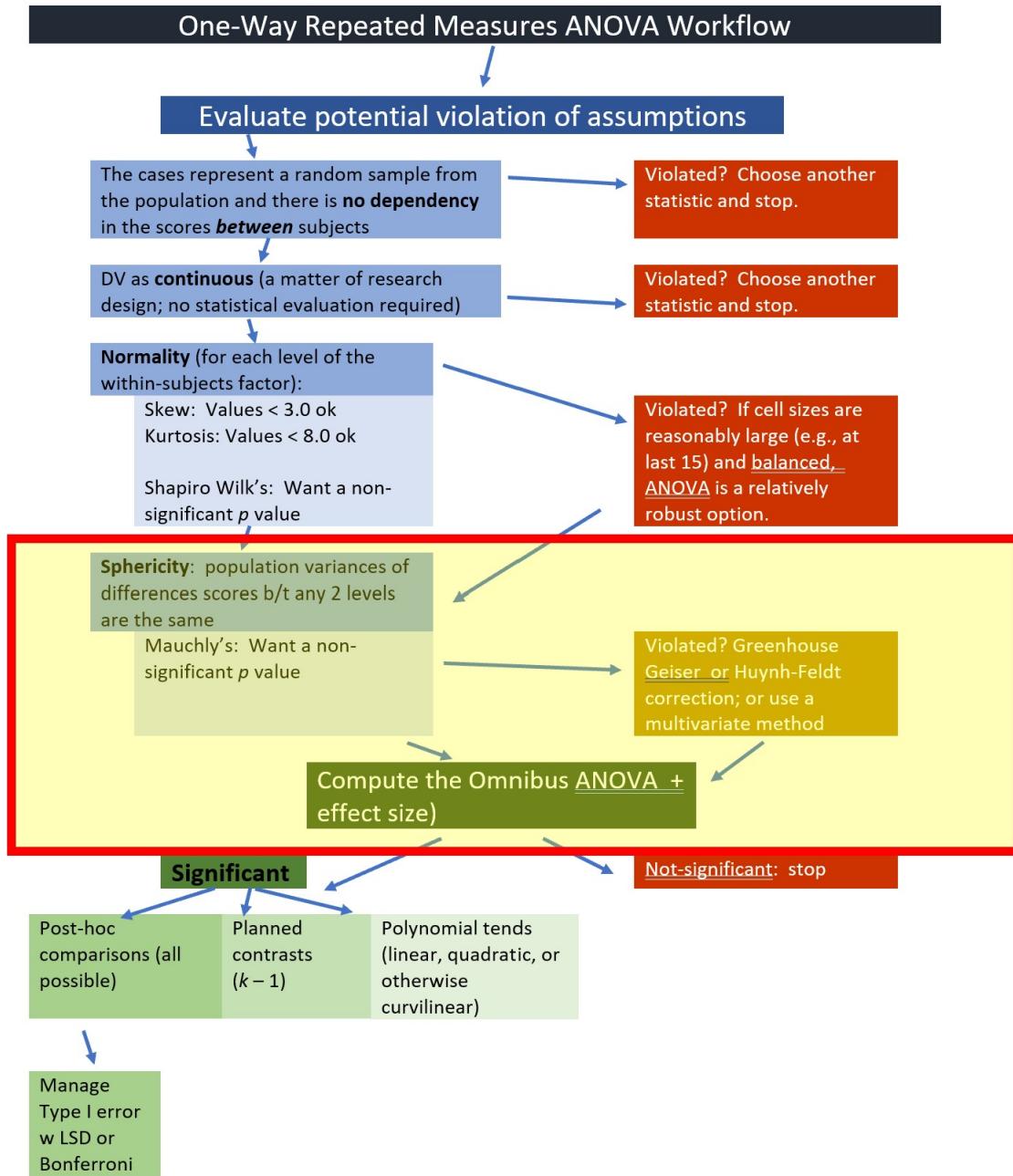


Figure 9.8: Image of our position in the workflow for the one-way repeated measures ANOVA

```
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.

9.5.3 Follow-up

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.

```
pwc <- Amodeo_long %>%
  rstatix::pairwise_t_test(Resilience ~ Wave, paired = TRUE, p.adjust.method = "bonferroni")
pwc
```

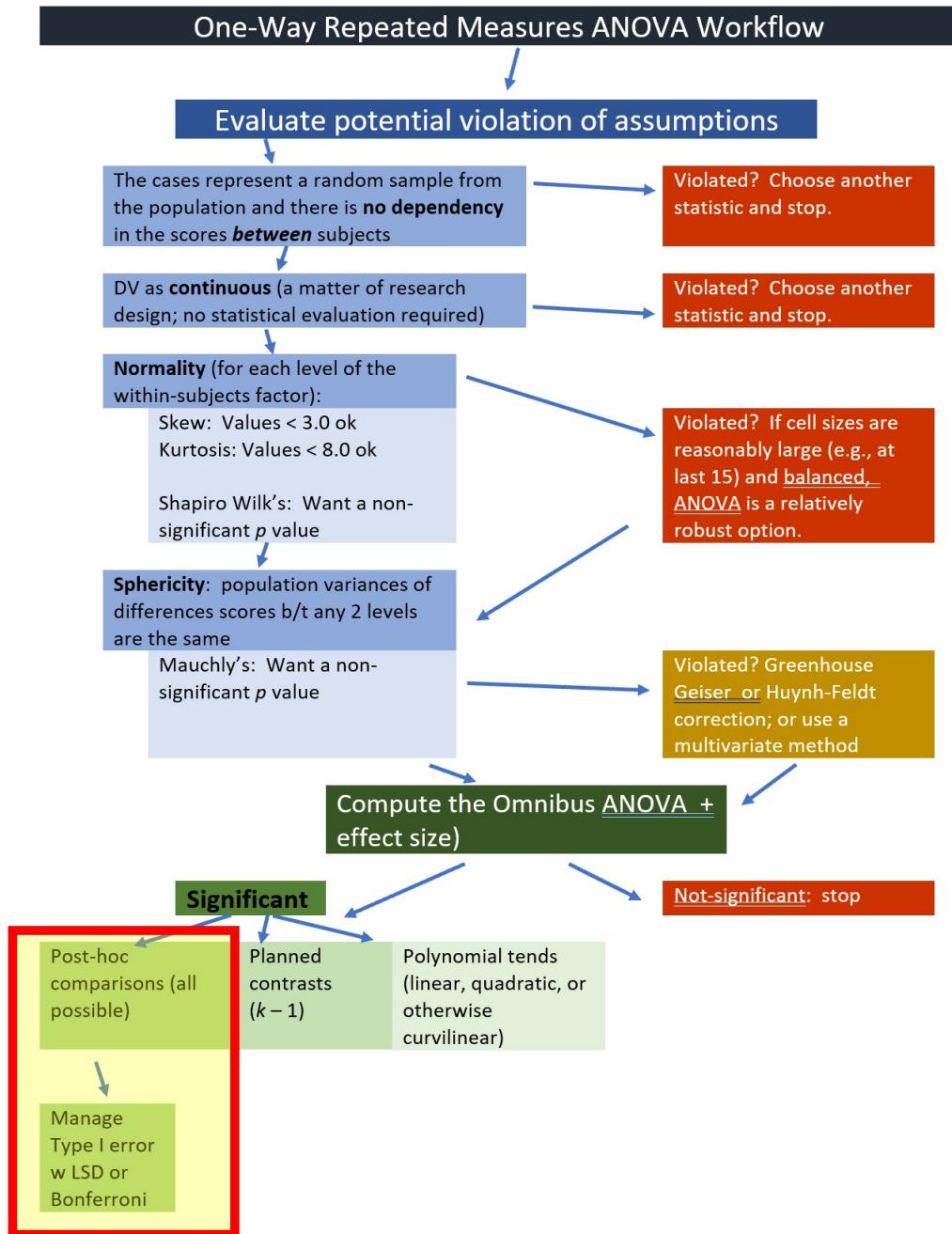


Figure 9.9: Image of our position in the workflow for the one-way repeated measures ANOVA

```
# 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 0.206 ns
2 Resilience Pre    Follow~    8     8     -2.00     7 0.086 0.257 ns
3 Resilience Post   Follow~    8     8      1.06     7 0.325 0.975 ns
```

Although we had a statistically significant omnibus test, we did not obtain statistically significant results in any of the posthoc pairwise comparisons. Why not?

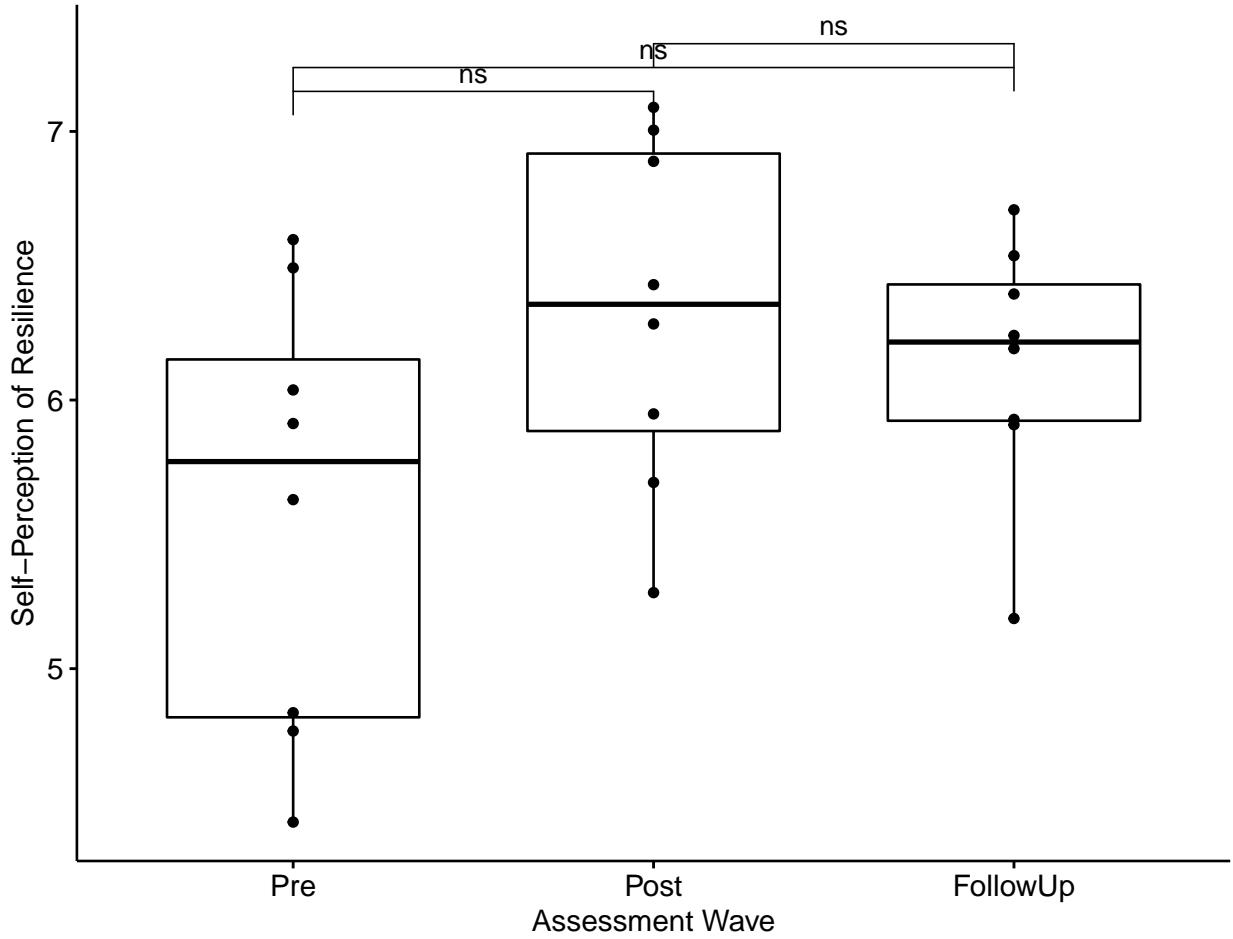
- Our omnibus F was right at the margins
 - a larger sample size (assuming that the effects would hold) would have been more powerful.
 - there could be significance if we compared pre to the combined effects of post and follow-up.

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, 2014b].

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).

```
pwc <- pwc %>%
  rstatix::add_xy_position(x = "Wave")
bxp + ggpubr::stat_pvalue_manual(pwc) + labs(subtitle = rstatix::get_test_label(RM_AOV,
  detailed = TRUE), caption = rstatix::get_pwc_label(pwc))
```

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.

9.5.4 Results Section

Repeated measures ANOVA has several assumptions regarding outliers, normality, and sphericity. 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 no outliers. Regarding normality, no values of skew and kurtosis (at each wave of assessment) fell within cautionary ranges for skew and kurtosis [Kline, 2016]. Additionally, the Shapiro-Wilk tests applied at each wave of the design were non-significant. 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, alpha was retained at .05 [Green and Salkind, 2014b]. 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.

9.5.4.1 Creating an APA Style Table**

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")
```

9.5.4.2 Comparison with Amodeo et al.[2018]

How do our findings and our write-up from the simulated data compare with the original article?

The F string is presented in the Table 1 note ($F[1.612, 11.283] = 6.390$, $p = 0.18$, η_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 they report 6 comparisons; 3 are repeated because they are merely in reverse. Yet, the revised alpha would be .016 and the one, lone, comparison would not have been statistically significant. That said, we can use the Tukey LSD because there are only 3 comparisons and holding alpha at .05 can be defended [Green and Salkind, 2014b].

- 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

9.6 Power in Repeated Measures 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 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 a conversion calculator)
- $nscor$ = 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* packages converter to transform our η^2 to Cohen's f .

```
effectsize::eta2_to_f(.203)
```

```
[1] 0.5046832
```

```
WebPower::wp.rmanova(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/rmanova>

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. In the script below:

- *nscor* is the degree of violation of the sphericity assumption. If we think we won't violate it, we can enter 1.0 or leave it out (the *wp.ranova* default is 1.0)
- *f* is the effect size estimate; Cohen suggests that *f* values of 0.1, 0.25, and 0.4 represent small, medium, and large effect sizes, respectively.

```
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/rmanova>

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.

9.7 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). 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

9.7.1 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.

9.7.2 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 resimulates 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)
```

9.7.3 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 a one-way repeated measures ANOVA. Please have at least 3 levels for the predictor variable.

9.7.4 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. Check and, if needed, format data	5	_____
2. Evaluate statistical assumptions	5	_____
3. Conduct omnibus ANOVA (w effect size)	5	_____
4. Conduct all possible pairwise comparisons (like in the lecture)	5	_____
5. Describe approach for managing Type I error	5	_____
6. APA style results with table(s) and figure	5	_____
7. Explanation to grader	5	_____

Assignment Component	Points Possible	Points Earned
Totals	35	_____

9.8 Bonus Reel:

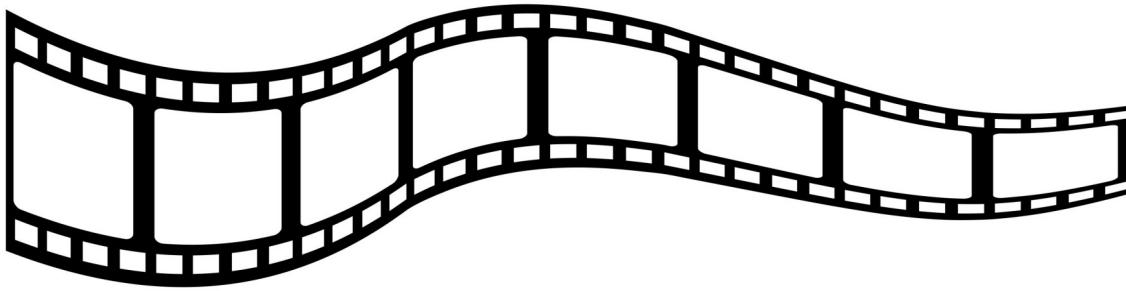


Figure 9.11: Image of a filmstrip

Without the *rstatix* helper package, here is how the analysis would be run in the package, *car*. Note that this package results in the multivariate output. The *p* value of the omnibus *F* was non-significant from the start (*p* = .213).

```
library(car)

waveLevels <- c(1, 2, 3)
waveFactor <- as.factor(waveLevels)
waveFrame <- data.frame(waveFactor)
waveBind <- cbind(Amodeo_wide$Pre, Amodeo_wide$Post, Amodeo_wide$FollowUp)
waveModel <- lm(waveBind ~ 1)
analysis <- Anova(waveModel, idata = waveFrame, idesign = ~waveFactor)
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 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
HF eps Pr(>F[HF])
waveFactor 0.8172743 0.05734876
```

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.

- 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. 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 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.

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).

The steps in working the mixed design generally include,

1. Exploring the data/evaluating the assumptions
2. Evaluating the omnibus test
3. 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
4. Write it up with tables, figure(s)

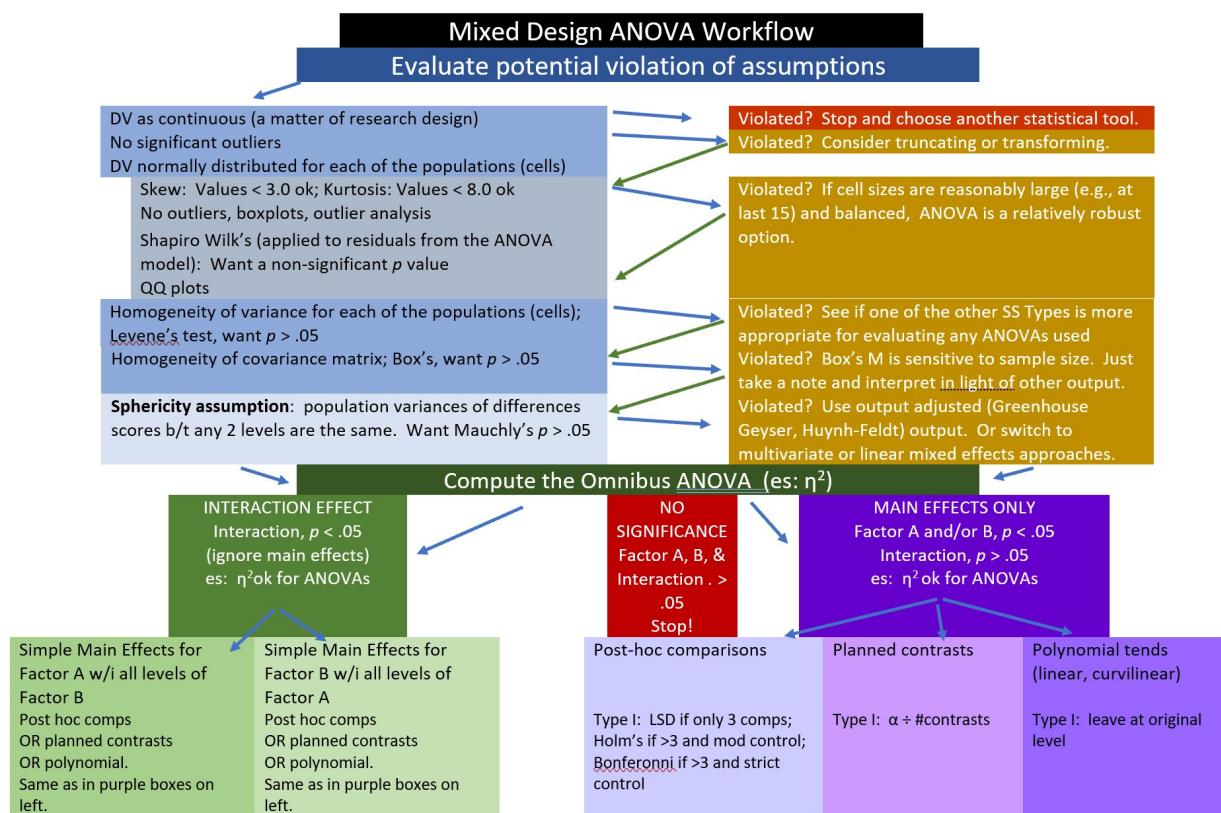


Figure 10.1: Image of a workflow for mixed design ANOVA

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 [2016] 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 multi-level, 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 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 and were engaged in regular activities of daily living. The Friends series featured characters who were predominantly White, cis-gendered, 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
Friends	X		X	X
Little Mosque on the Prairie	X	Selected for potential for prejudice reduction	X	X

Figure 10.2: Image of the design for the Murrar and Brauer (2018) study

10.3.1 Simulating the data from the journal article

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). 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
```

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')
```

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., 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)
```

10.4 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:

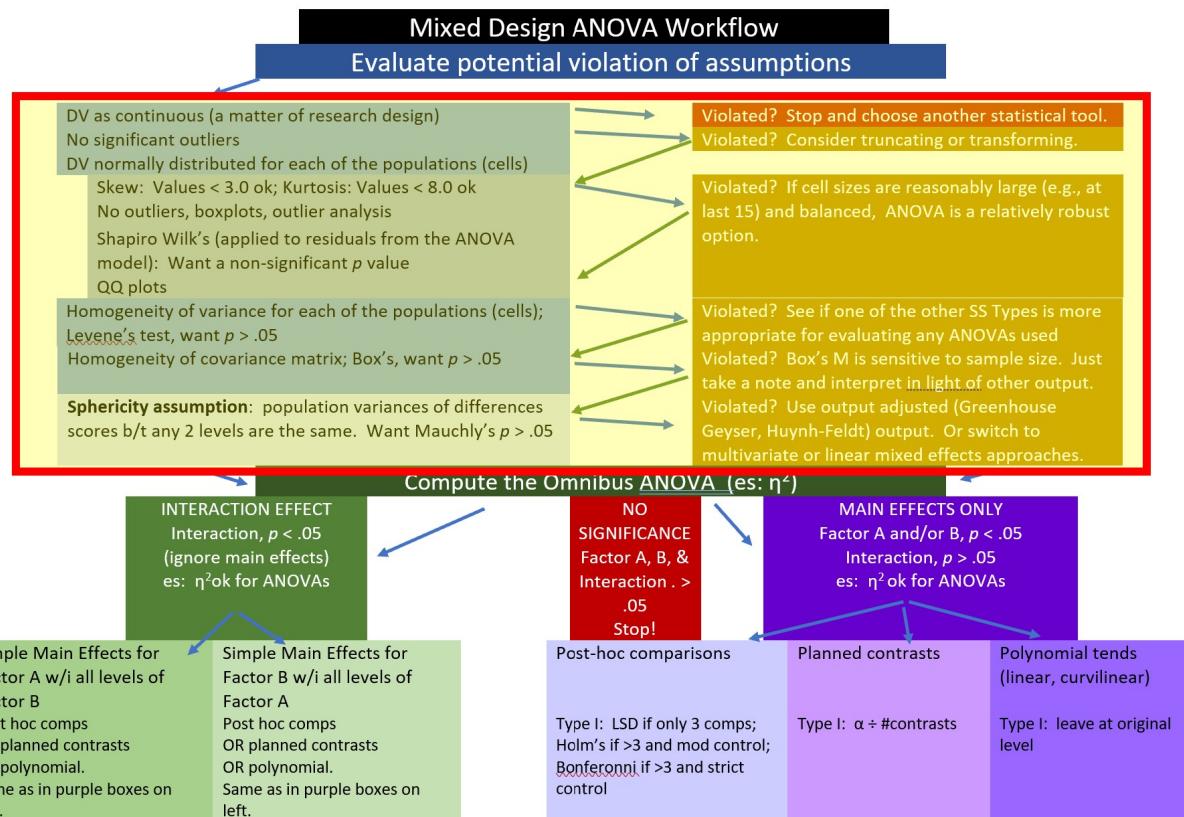


Figure 10.3: Image of the workflow showing that we are on the “Evaluating assumptions” portion

First, let's examine the overall descriptive statistics.

```
psych::describe(Murrar_df)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range
rowID*	1	579	290.00	167.29	290.00	290.00	214.98	1.00	579.00	578.00
caseID*	2	579	97.00	55.76	97.00	97.00	71.16	1.00	193.00	192.00
Wave*	3	579	2.00	0.82	2.00	2.00	1.48	1.00	3.00	2.00
COND*	4	579	1.49	0.50	1.00	1.49	0.00	1.00	2.00	1.00
AttArab	5	579	66.84	19.75	68.04	67.64	20.38	6.14	100.00	93.86
AttWhite	6	579	75.31	17.02	76.72	76.19	18.50	23.03	100.00	76.97
Diff	7	579	8.47	26.33	8.65	8.50	25.73	-71.51	90.74	162.25
			skew	kurtosis	se					
rowID*		0.00	-1.21	6.95						

caseID*	0.00	-1.21	2.32
Wave*	0.00	-1.51	0.03
COND*	0.03	-2.00	0.02
AttArab	-0.39	-0.16	0.82
AttWhite	-0.37	-0.51	0.71
Diff	0.03	0.00	1.09

Our analysis will use the difference score (Diff) as the dependent variable. Let's look at this variable in its combinations of wave and condition.

```
psych::describeBy(Diff ~ Wave + COND, data = Murrar_df, mat = TRUE)
```

item	group1	group2	vars	n	mean		sd	median	trimmed
					Friends				
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	
Diff1	24.48143	-47.342	72.565	119.907	0.17600603	-0.35619536	2.415140		
Diff2	29.77135	-42.598	82.288	124.886	-0.04613379	-0.57456759	2.668609		
Diff3	24.59189	-46.528	75.014	121.542	0.13240007	0.06485450	2.357294		
Diff4	30.90480	-71.510	90.737	162.247	-0.20135358	-0.03033805	3.131178		
Diff5	23.93213	-65.259	83.367	148.626	0.32819576	0.54919109	2.766918		
Diff6	25.26054	-53.856	55.264	109.120	-0.06475209	-0.42366384	2.428002		

```
# 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.

10.4.1.1 Assumption of Normality

We can use this output to evaluate the distributional characteristics of the dependent variable. Recall that mixed design ANOVA assumes a normal distribution.

Our values of skew and kurtosis are well within the limits [Kline, 2016] of a normal distribution.

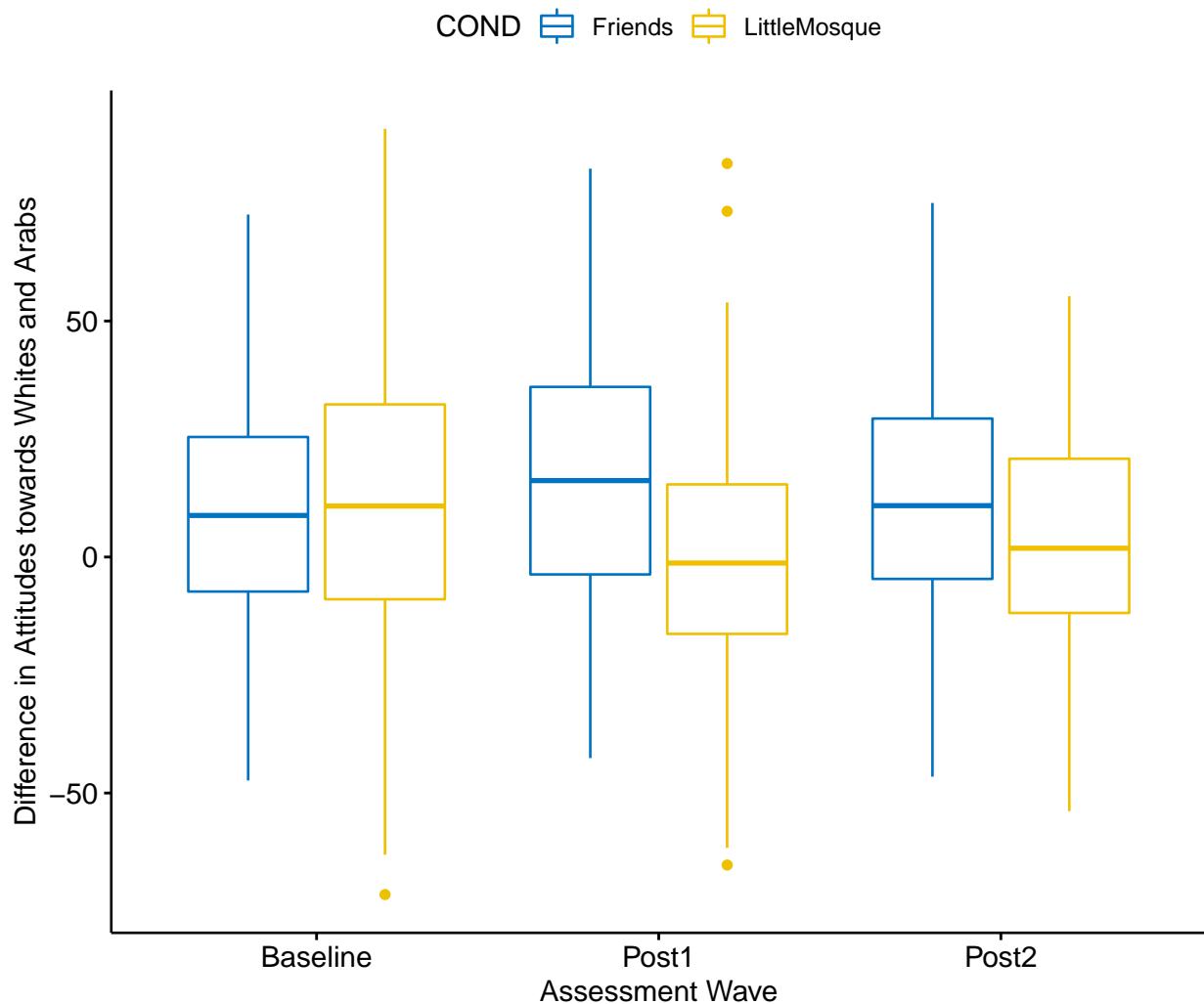
- skew: < 3; the highest skew value in our data is 0.32

- kurtosis: extreme values are between 8 and 20; the highest kurtosis value in our data is .55

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).

You'll notice that as we are creating these boxplots we are saving them as objects. This is not necessary to produce the graph. However, we will combine the object with other data, later, to embed results if the analysis in our figures.

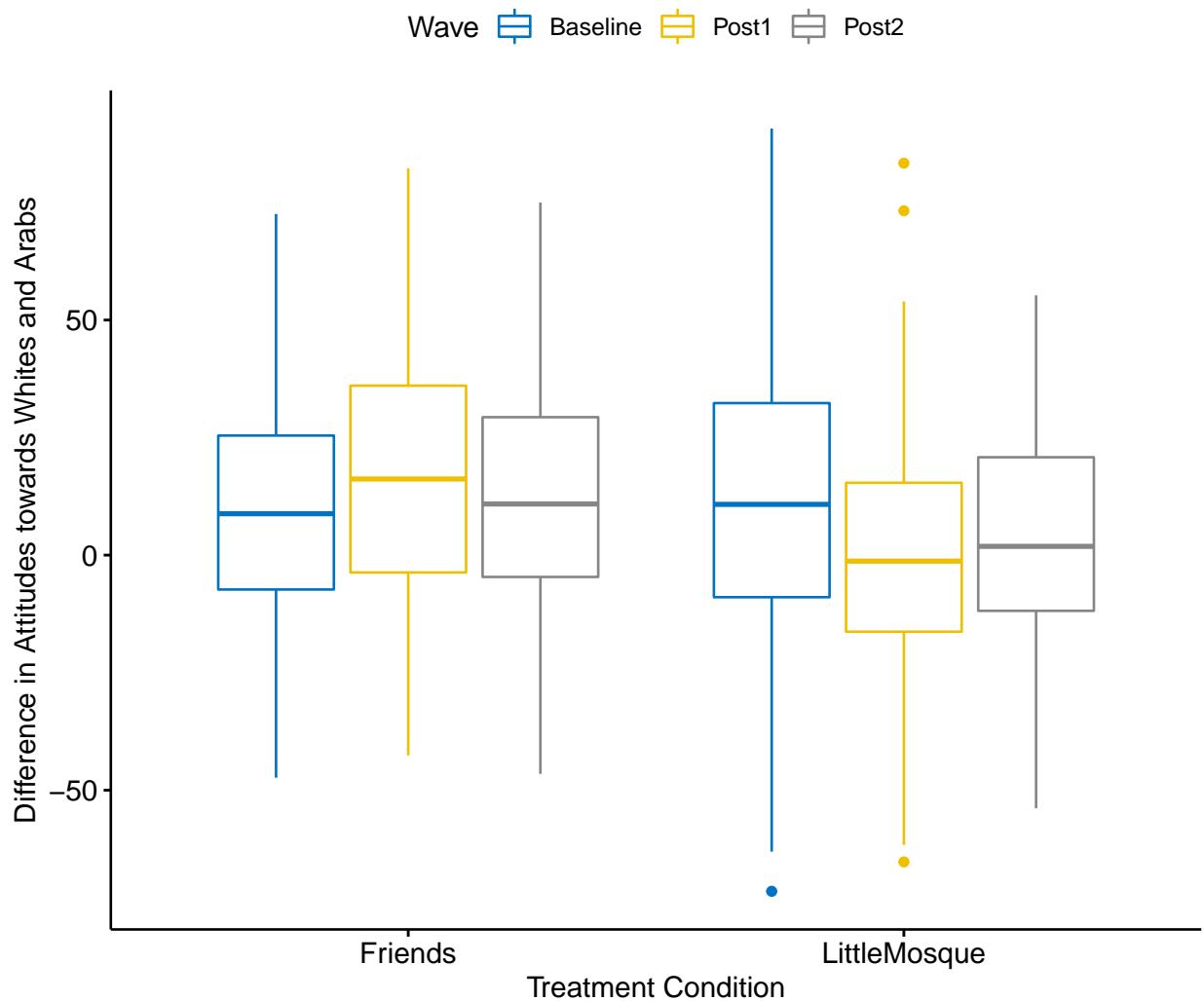
```
CNDwiWV <- ggpubr::ggboxplot(Murrar_df, x = "Wave", y = "Diff", color = "COND",
  palette = "jco", xlab = "Assessment Wave", ylab = "Difference in Attitudes towards Whites and Arabs")
CNDwiWV
```



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. This boxplot also provides a glimpse of the patterns in our data. That is, the means are quite similar at baseline; in the post intervention waves we see greater difference scores for the Friends condition.

Let's reconfigure the data by putting the wave on the X axis. Plotting it both ways (i.e., swapping roles of predictor and moderator) can help us get a sense of what is happening.

```
WVwiCND <- ggpubr::ggboxplot(Murrar_df, x = "COND", y = "Diff", color = "Wave",
  palette = "jco", xlab = "Treatment Condition", ylab = "Difference in Attitudes towards White and Arabs")
```



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:

- $Q1 - 3 \times IQR$
- $Q3 + 3 \times IQR$

Using the `rstatix::identify_outliers` function we can look for outliers in the dependent variable, 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.

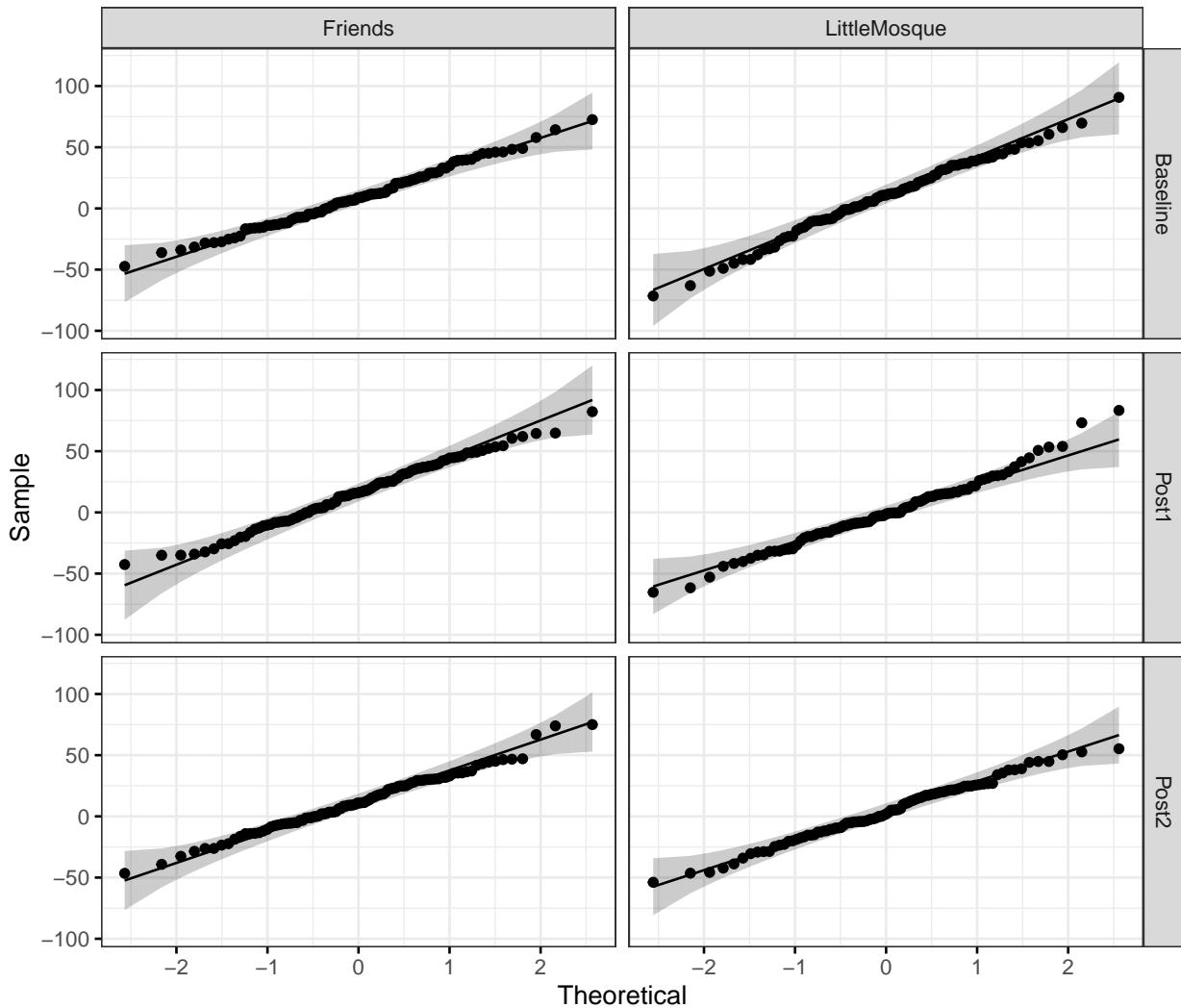
Next we can use the `rstatix::shapiro_test()` to see if any of the distributions of the dependent variable (Diff) within each wave-by-condition combinations differs significantly from a normal distribution.

```
Murrar_df %>%
  group_by(Wave, COND) %>%
  rstatix::shapiro_test(Diff)
```

	Wave	COND	variable	statistic	p
	<fct>	<fct>	<chr>	<dbl>	<dbl>
1	Baseline	Friends	Diff	0.993	0.915
2	Baseline	LittleMosque	Diff	0.993	0.923
3	Post1	Friends	Diff	0.992	0.798
4	Post1	LittleMosque	Diff	0.986	0.437
5	Post2	Friends	Diff	0.990	0.708
6	Post2	LittleMosque	Diff	0.991	0.762

The Shapiro Wilks test suggests that distribution in each of our cells is not significantly different than normal. We can further visualize this with QQ plots.

```
ggpubr::ggqqplot(Murrar_df, "Diff", ggtheme = theme_bw()) + facet_grid(Wave ~ COND)
```



10.4.1.2 Homogeneity of variance assumption

Because there is a between-subjects variable, we 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~
```

None-the-less, 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. 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. 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, 2016]). Additionally, the Shapiro-Wilk tests applied at each level of the design were non-significant. 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.

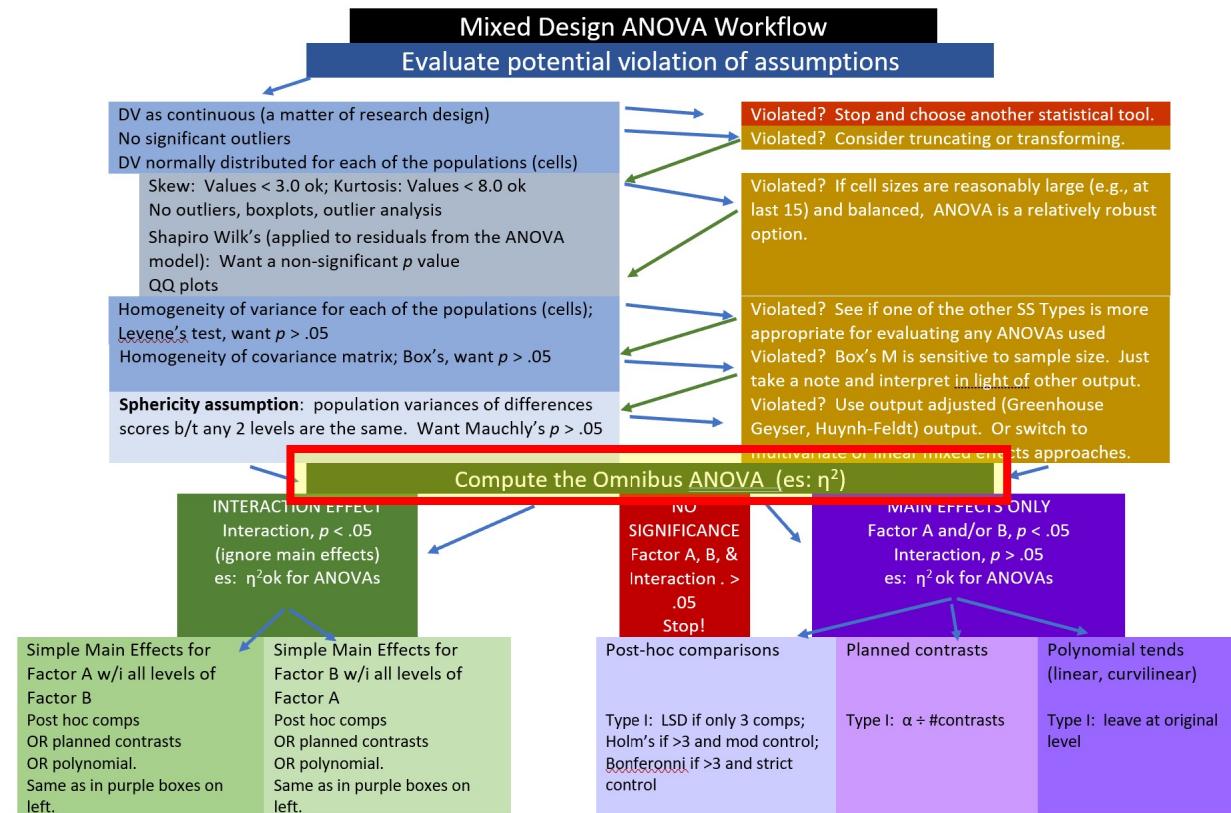


Figure 10.4: Image of the workflow showing that we are at the “Compute the Omnibus ANOVA” step

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 summon 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)
Diff_2way
```

ANOVA Table (type III tests)

\$ANOVA

	Effect	DFn	DFd	F	p	p<.05	ges
1	COND	1	191	13.149	0.000369	*	0.023000
2	Wave	2	382	0.273	0.761000		0.000933
3	COND:Wave	2	382	5.008	0.007000	*	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

First, we check Mauchly's test for the main and interaction effects that involve the repeated measures variable.

- main effect for Wave: $W = .99$, $p = .369$
- main effect for 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 effect ($W = 0.99$, $p = .369$) and interaction effect ($W = 0.99$, $p = .369$).

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. There is also an option to use a multivariate approach when ANOVA designs include a repeated measures factor.

Omnibus Results

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 note that according to Cohen et al.'s [Cohen et al., 2003] guidelines, the effect size for the interaction term is small.

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 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 3 one-way ANOVAs. When we look at condition within wave, our ANOVAs will look like this:

- comparison of Friends and Little Mosque within the baseline wave
- comparison of Friends and Little Mosque within the post1 wave
- comparison of Friends and Little Mosque within the post2 wave

```
# create an object to hold the output the group_by function is what
# results in three, one-way ANOVAs for each of the waves, separately
# the between = Cond means that each level of cond will be compared
# method - 'bonferroni' gets us both the standard and adjusted p
# values
SimpleWave <- Murrar_df %>%
  group_by(Wave) %>%
  rstatix::anova_test(dv = Diff, wid = caseID, between = COND) %>%
  rstatix::get_anova_table() %>%
  rstatix::adjust_pvalue(method = "bonferroni")
```

```
Coefficient covariances computed by hccm()
Coefficient covariances computed by hccm()
Coefficient covariances computed by hccm()
```

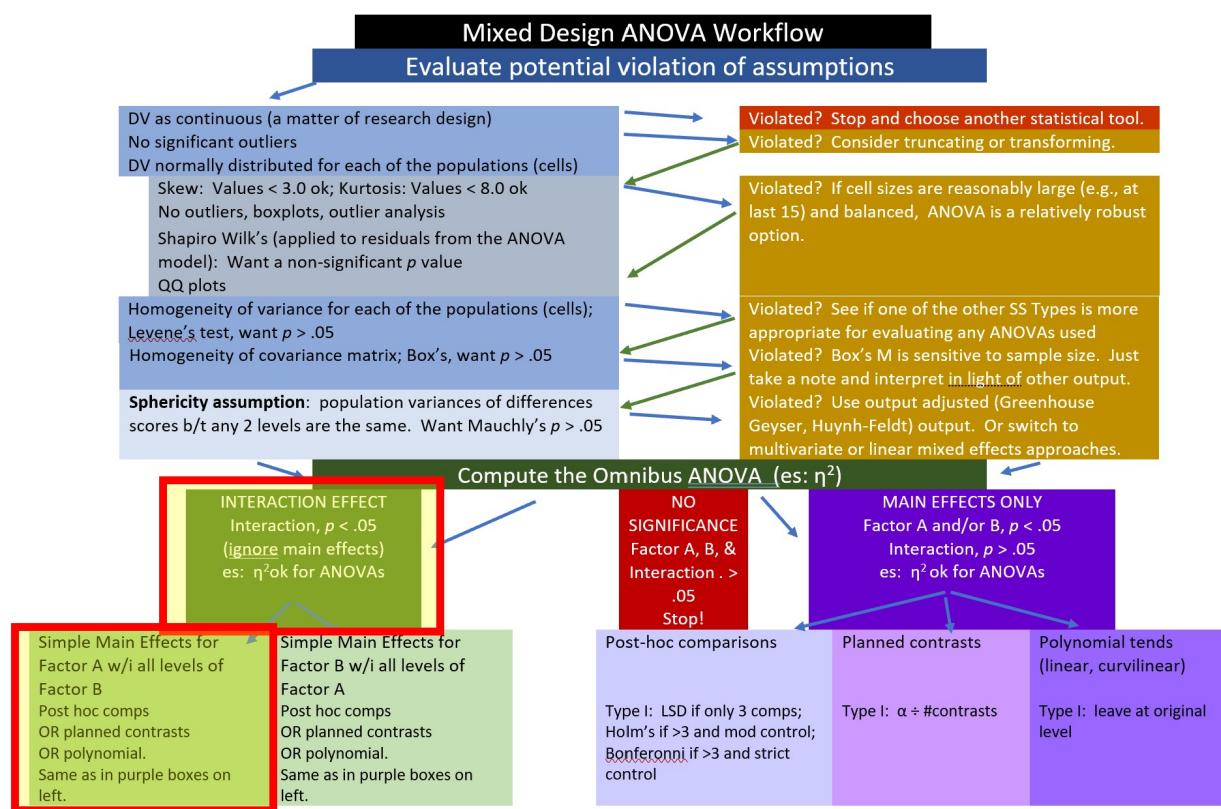


Figure 10.5: Image of the workflow showing that we are at the “Simple Main Effects for Factor A within all levels of Factor B” step

SimpleWave

	Wave	Effect	DFn	DFd	F	p `p < .05`	ges	p.adj
*	<fct>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>
1	Baseline	COND	1	191	0.012	0.914	""	0.0000614
2	Post1	COND	1	191	17.5	0.0000438	"*"	0.084
3	Post2	COND	1	191	5.99	0.015	"*"	0.03

In prior lectures we have adjusted the p value against which we compare the resulting p value. When we specify “bonferroni” on the `adjust_pvalue()` command, the algorithm adjusts the reported p value for us. We can see the unadjusted p value in the “ $p < .05$ ” column and the Bonferroni adjustment in the “ $p.adj$ ” column.

I think that it will be easiest for us to interpret this simple main effect as the traditional $p < .05$ and then apply the restrictions to the alpha at the next level of analysis. In this particular instance, we would have statistically significant differences somewhere between the Friends and Little Mosque conditions for both the Post ($p = .027$) and Post2 ($p = .010$) waves.

F strings:

- Pre: $F(1, 191) = 0.012, p = .914, \eta^2 = 0.000$ (the effect size is zero)
- Post: $F(1, 191) = 17.497, p < .001, \eta^2 = 0.084$ (a moderate effect size)
- Post2: $F(1, 191) = 5.994, p = .015, \eta^2 = 0.030$ (a small effect size)

Recall, interpretation for the eta-squared are $.01 \sim \text{small}, .06 \sim \text{medium}$, and $>.14 \sim \text{large}$

Because there are only two levels (Friends, Little Mosque) within each wave (baseline, post1, post2), this simple effects analysis is complete with the three pairwise comparisons.

As always, we have several choices about how to manage Type I error. In a circumstance when the analysis of simple main effects (condition within wave) includes only three pairwise comparisons, we can use the LSD method [Green and Salkind, 2014b]. This means that we can leave the alpha at 0.05. If we were to use a traditional Bonferroni, we would use $\alpha = .017 (.05/3)$. Although the more restrictive Bonferroni criteria comes close, in both cases we would still have one non-significant(baseline) and two significant (post1, post2) simple main effects.

.05/3

[1] 0.01666667

If we were to write up this result:

We followed the significant interaction effect with an evaluation of simple main effects of condition within wave. Because there were only three comparisons following the omnibus evaluation, we used the LSD method to control for Type I error and left the alpha at .05 [Green and Salkind, 2014b]. There was a non-statistically significant difference between conditions at baseline: $F(1, 191) = 0.012, p = .914, \eta^2 = 0.000$. However other were statistically significant differences at post1 ($F[1, 191] = 17.497, p < .001, \eta^2 = 0.084$) and post2 ($F[1, 191] = 5.994, p = .015, \eta^2 = 0.030$). We note that the effect size at post1 approached a moderate size; the effect size at post2 was small.

10.4.4 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.

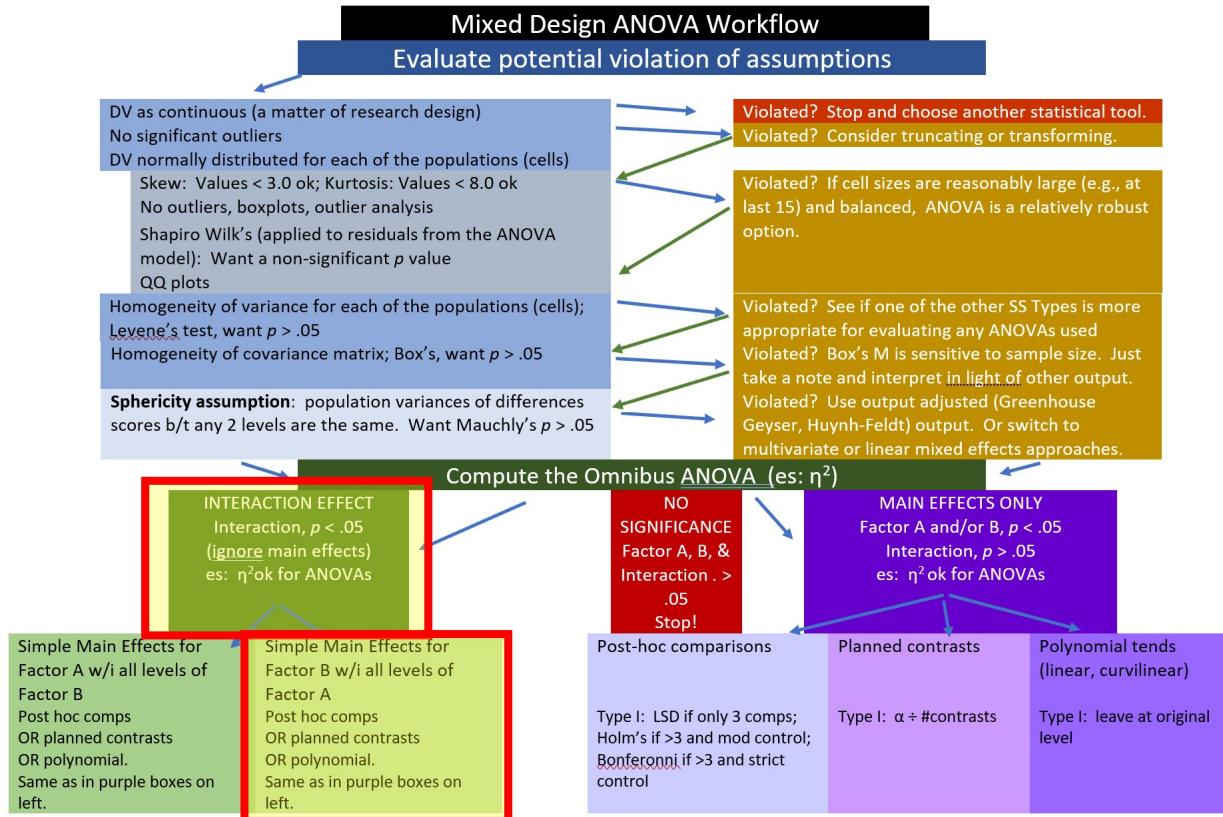


Figure 10.6: Image of the workflow showing that we are at the “Simple Main Effects for Factor B within all levels of Factor A” step

If we conducted this alternative we would start with three one-way ANOVAs and then follow each of those with pairwise comparisons. First, the one-way repeated measures ANOVAs:

- 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 = "bonferroni")
SimpleCond
```

```
# A tibble: 2 x 9
  COND     Effect    DFn    DFd      F      p `p<.05`    ges p.adj
  <fct>   <fct>    <dbl>   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
```

1 Friends	Wave	2	194	1.76	0.175	""	0.012	0.35		
2 LittleMosque	Wave	2	188	3.39	0.036	"*"	0.022	0.072		

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 no relationship)
- 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, we only need to conduct 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 is not necessary, however we can use this object in combination with the boxplot we created earlier to embed results of our analysis in the resulting figure.

```
pwcWVwiGP <- Murrar_df %>%
  group_by(COND) %>%
  rstatix::pairwise_t_test(Diff ~ Wave, paired = TRUE, detailed = TRUE,
                           p.adjust.method = "bonferroni") # %>%
# 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> <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
# ... with 5 more variables: conf.high <dbl>, method <chr>, alternative <chr>,
#   p.adj <dbl>, p.adj.signif <chr>
```

At this point, we likely need to control for Type I error. Why? We have already conducted two one-way ANOVAs after the omnibus. Now we will conduct three more pairwise comparisons in the Little Mosque condition. I would divide $.05/3$ and interpret these pairwise comparisons with an alpha of $.017$.

We find a significant difference between baseline and post1 ($t[95] = 2.447, p = .016$), but non-significant differences between baseline and post2 ($t[95] = 1.621, p = .108$) and post1 and post2 ($t[95] = -1.034, p = .304$)

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 small-to-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 the Little Mosque condition with pairwise comparisons. At this level we controlled for Type I error by dividing alpha (.05) by the number of paired comparisons (3). We found a significant difference between baseline and post1 ($t[95] = 2.447, p = .016$), but non-significant differences between baseline and post2 ($t[95] = 1.621, p = .108$) and post1 and post2 ($t[95] = -1.034, p = .304$).

10.4.5 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.

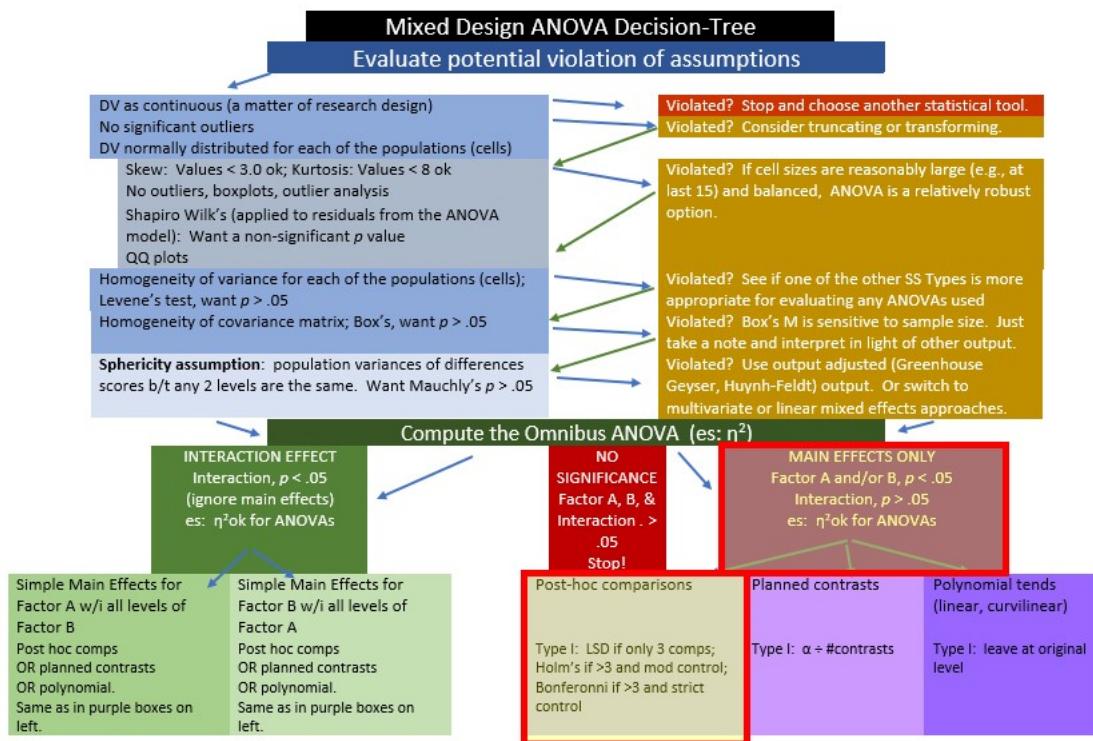


Figure 10.7: Image of a workflow showing that we are at the “Main effects only” step

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.

```
Murrar_df %>%
  rstatix::pairwise_t_test(Diff ~ Wave, paired = TRUE, p.adjust.method = "bonferroni")

# 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 ($F[1, 191] = 13.149, p < .001, \eta^2 = 0.023$).

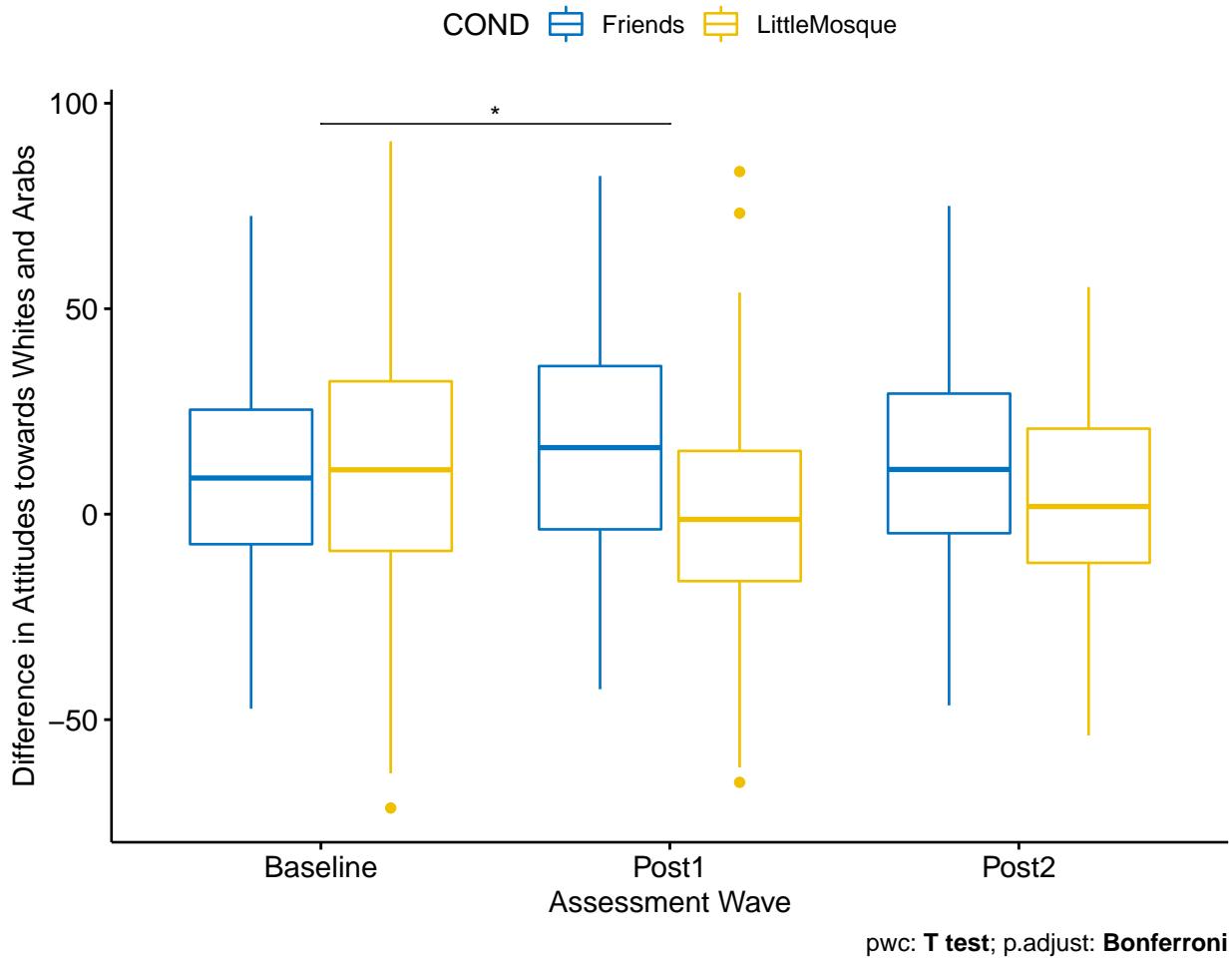
10.4.6 APA Style Write-up of the Results

Recall that earlier in this lesson we save objects for the boxplots (e.g., CNDwiWV) and the pairwise comparisons (e.g., pwcVWwiGP). The script below use the objects created from omnibus ANOVA and the pairwise comparisons to add results to the figure. Depending on where you are presenting your results, these may be useful.

This first figure would pair well if you report the simple main effect of condition within wave.

```
pwcVWwiGP <- pwcVWwiGP %>%
  rstatix::add_xy_position(x = "Wave")
CNDwiWV + ggpubr::stat_pvalue_manual(pwcVWwiGP, tip.length = 0, hide.ns = TRUE) +
  labs(subtitle = rstatix::get_test_label(Diff_2way, detailed = TRUE),
       caption = rstatix::get_pwc_label(pwcVWwiGP))
```

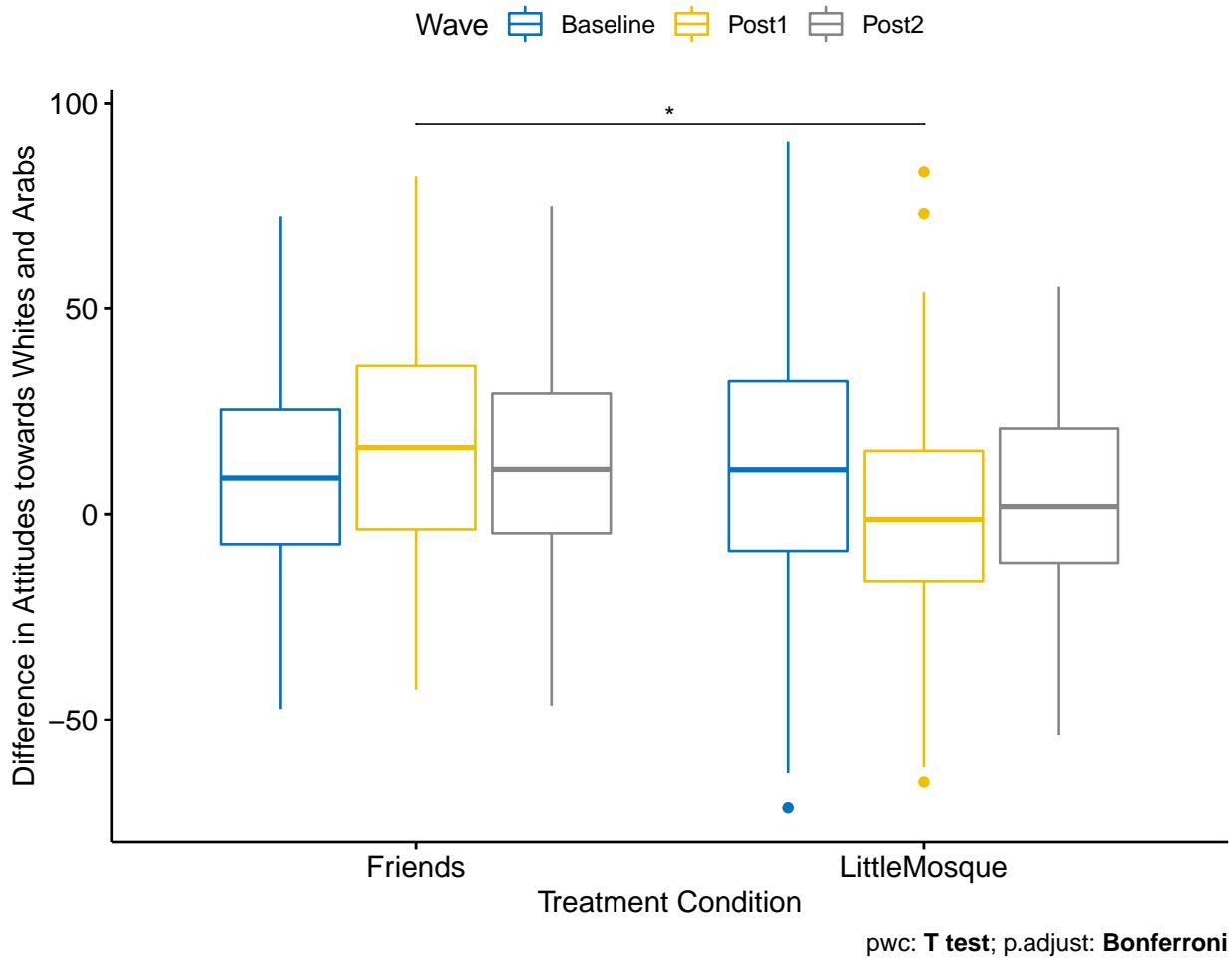
Anova, $F(2,382) = 5.01, p = 0.007, \eta^2_g = 0.02$



This second figure would pair well with the results that reported the simple main effect of wave within condition.

```
# pwcWVwiGP were my pairwise comparisons for the simple effect
# UE_2way was my omnibus ANOVA object WVwiCND was the boxplot before
# I did the ANOVA
pwcWVwiGP <- pwcWVwiGP %>%
  rstatix::add_xy_position(x = "Wave")
WVwiCND + ggpubr::stat_pvalue_manual(pwcWVwiGP, tip.length = 0, hide.ns = TRUE) +
  labs(subtitle = rstatix::get_test_label(Diff_2way, detailed = TRUE),
       caption = rstatix::get_pwc_label(pwcWVwiGP))
```

Anova, $F(2,382) = 5.01, p = 0.007, \eta^2_g = 0.02$



10.4.6.1 Results

We conducted a 2×3 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. 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. 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, 2016]). Additionally, the Shapiro-Wilk tests applied at each level of the design were non-significant. 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. Mauchly's test indicated no violation of the sphericity assumption for the main effect ($W = 0.99, p = .369$) and interaction effect ($W = 0.99, p = .369$).

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. 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 small-to-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 the Little Mosque condition with pairwise comparisons. At this level we controlled for Type I error by dividing alpha (.05) by the number of paired comparisons (3). We found a significant difference between baseline and post1 ($t[95] = 2.447, p = .016$), but non-significant differences between baseline and post2 ($t[95] = 1.621, p = .108$) and post1 and post2 ($t[95] = -1.034, p = .304$).

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. 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")
```

10.4.6.2 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 “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). 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 one-way 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	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 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. (2014). One-Way Analysis of Covariance (Lesson 27). In *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (Seventh 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 [Green and Salkind, 2014a] 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.

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, a](#)].

Our analytic process will be similar to others in the ANOVA series:

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.

An ANCOVA workflow maps this in further detail.

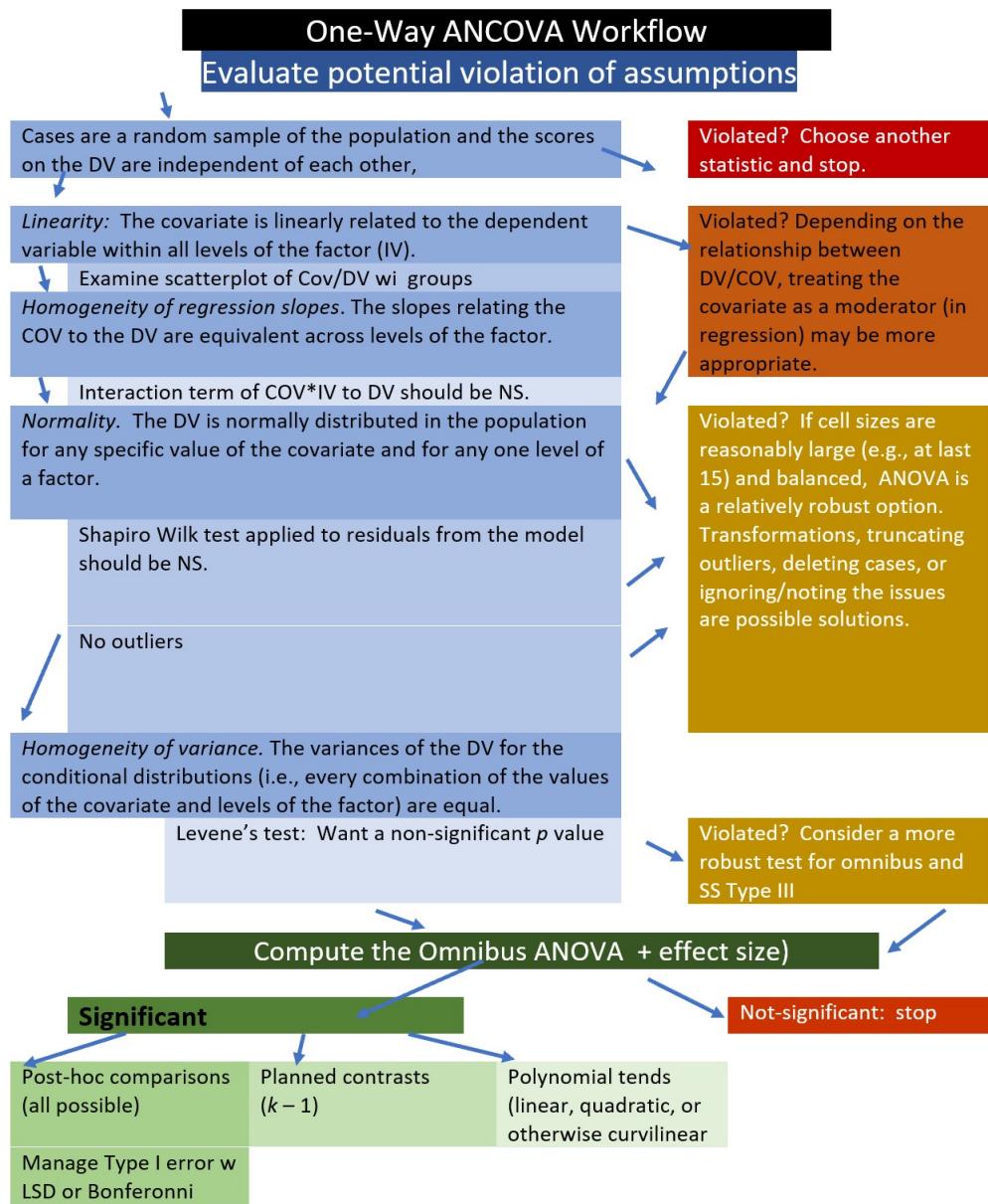


Figure 11.1: Image of the ANCOVA workflow

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 relateable 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 whereas 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.

11.3.1 Simulating the data from the journal article

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.

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
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

```

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 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_CSV.csv', sep=',', col.names=TRUE, row.names=FALSE)
# bring back the simulated dat from a .csv file Murrar_wide <-
# read.csv ('MurrarW_CSV.csv', header = TRUE)
```

11.4.2 Checking the 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.

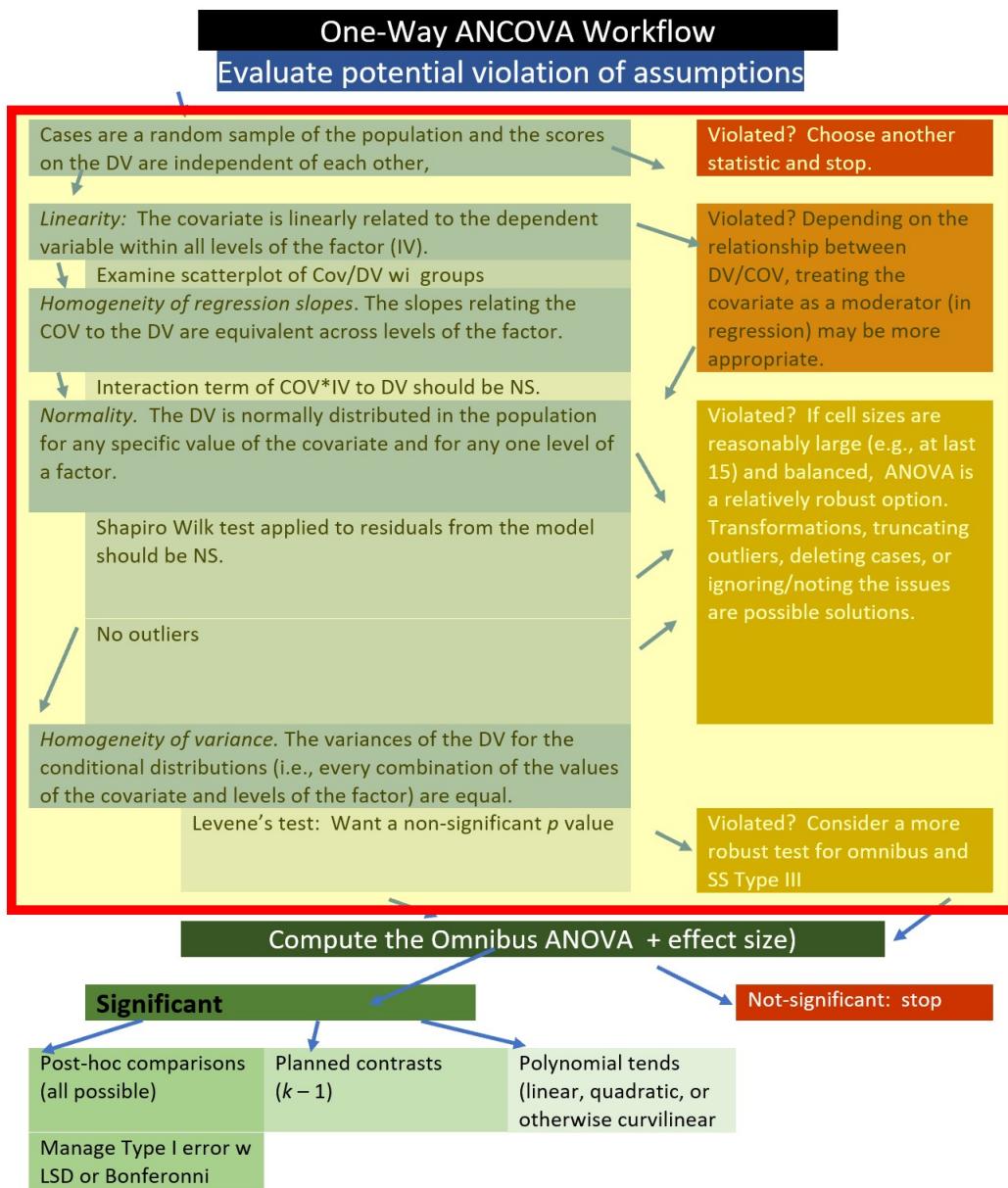


Figure 11.3: Image of the ANCOVA workflow, showing our current place in the process

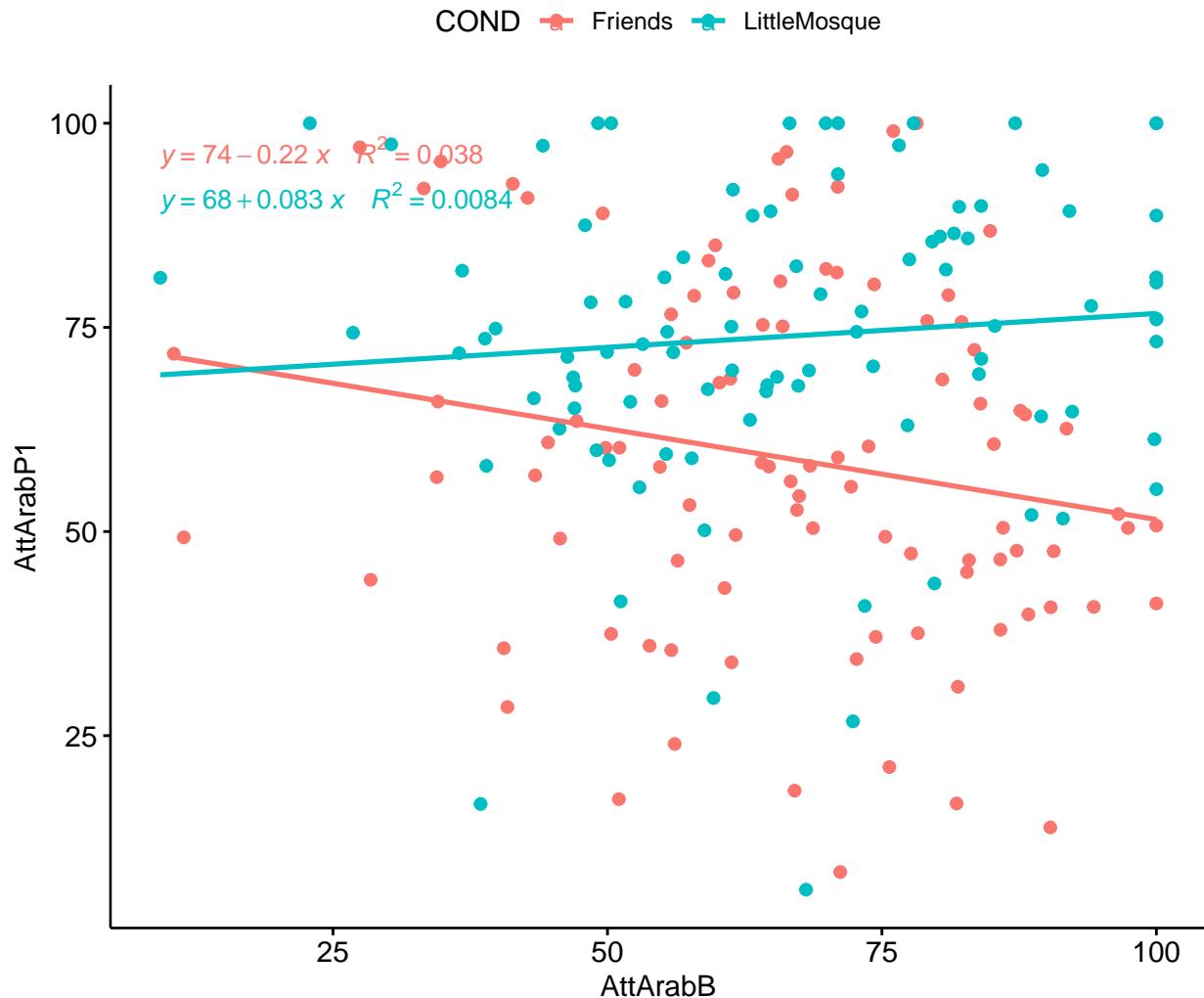
11.4.2.1 Linearity assumption

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).

```
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))
```

```
`geom_smooth()` using formula 'y ~ x'
```



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)
```

Coefficient covariances computed by hccm()

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>
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 [2016] has a great section on some of these options.

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 Write-up 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, $F(1, 191) = 3.515, p = .062$.

11.4.3 Calculating the Omnibus ANOVA

We are ready to conduct the omnibus ANOVA.

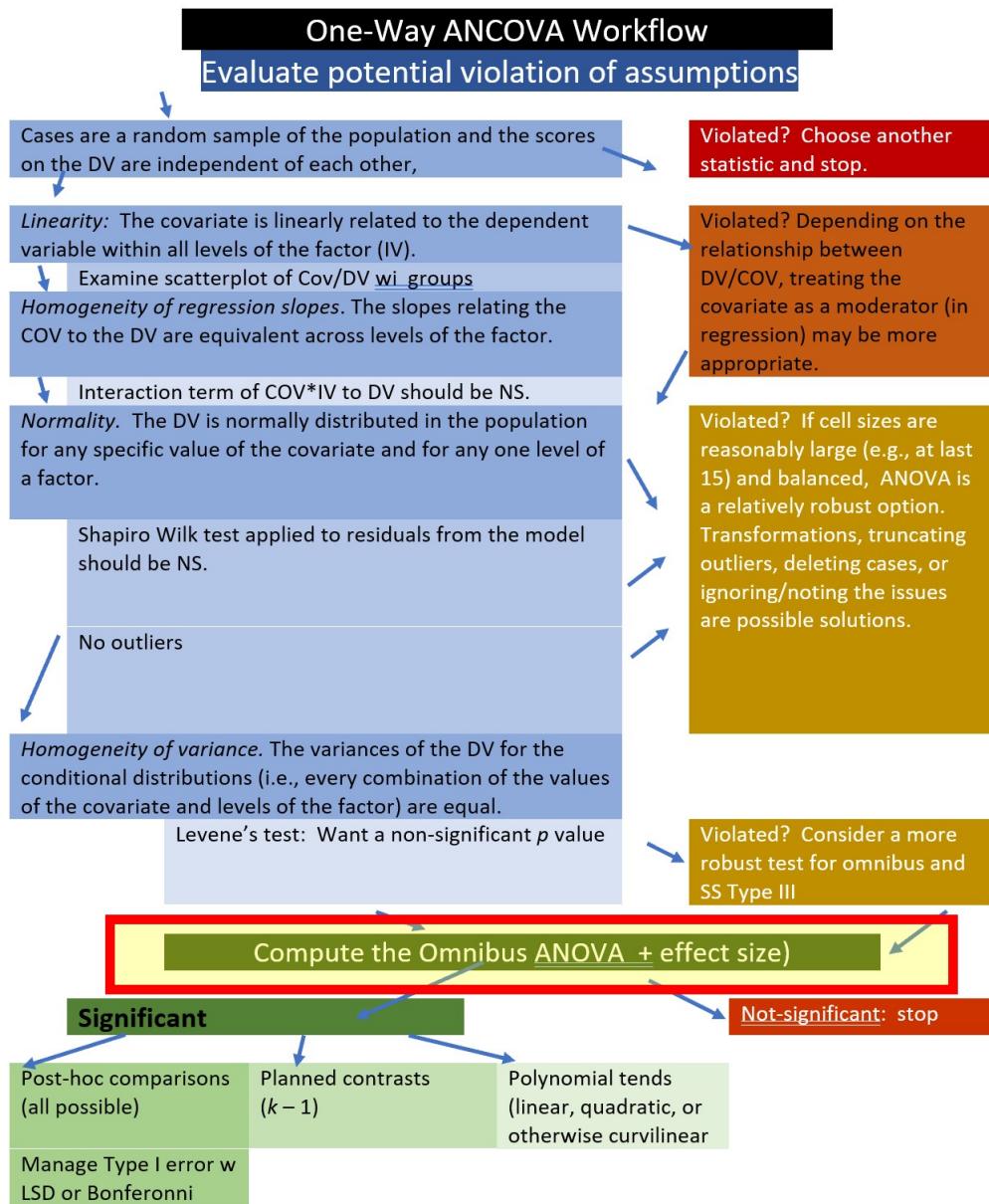


Figure 11.4: Image of the ANCOVA workflow, showing our current place in the process.

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)
```

Coefficient covariances computed by hccm()

```
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	0.416000000		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 appears to be a moderately sized effect.

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 Toward an APA style results section

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

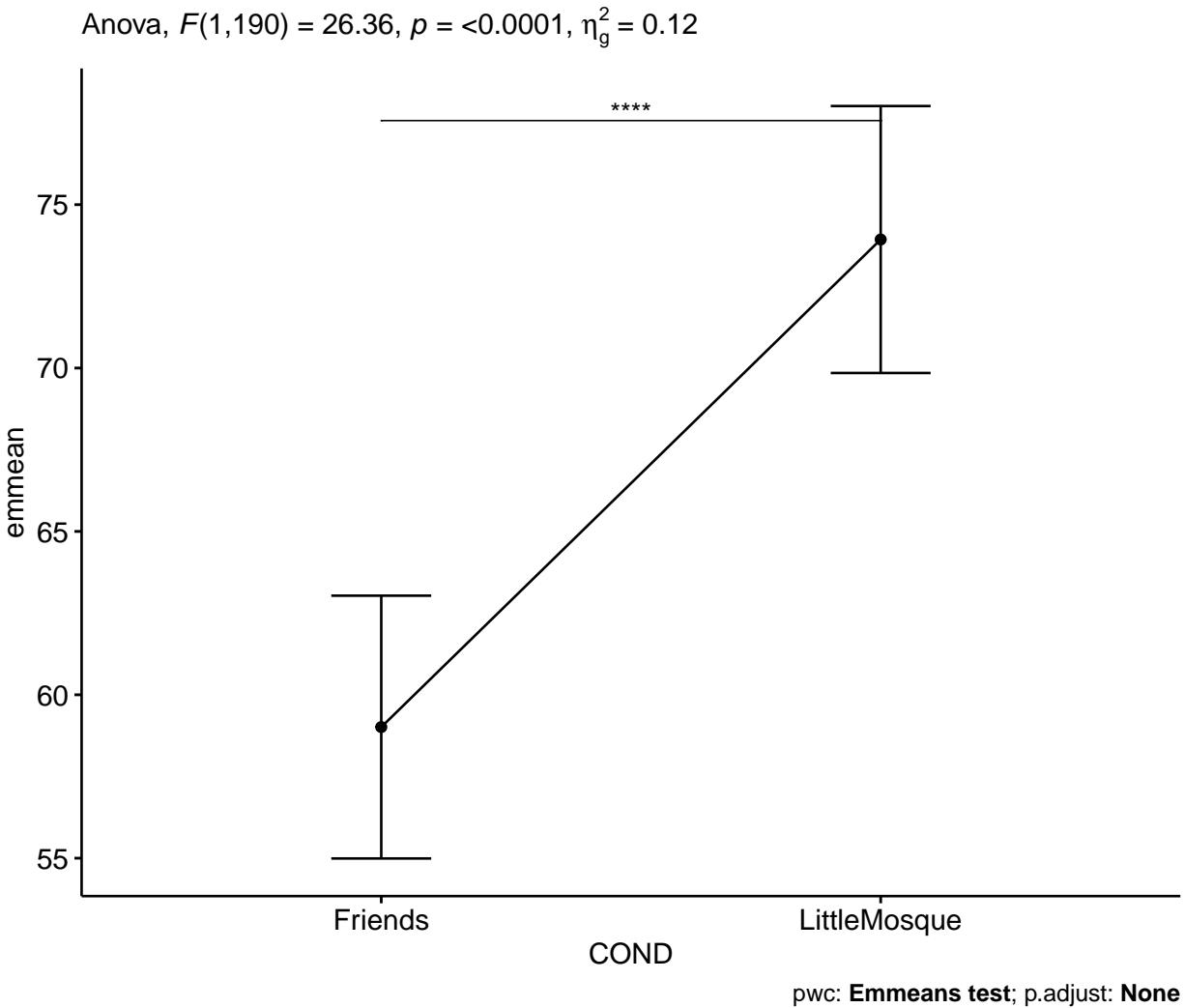
Unadjusted and Covariate-Adjusted Descriptive Statistics

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") +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) +
  ggpubr::stat_pvalue_manual(pwc_B, hide.ns = TRUE, tip.length = FALSE) +
  labs(subtitle = rstatix::get_test_label(MurrarB_ANCOVA, detailed = TRUE),
       caption = rstatix::get_pwc_label(pwc_B))
```



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 moderately 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 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 Checking the 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

- homogeneity of variance

These are depicted in the flowchart, below.

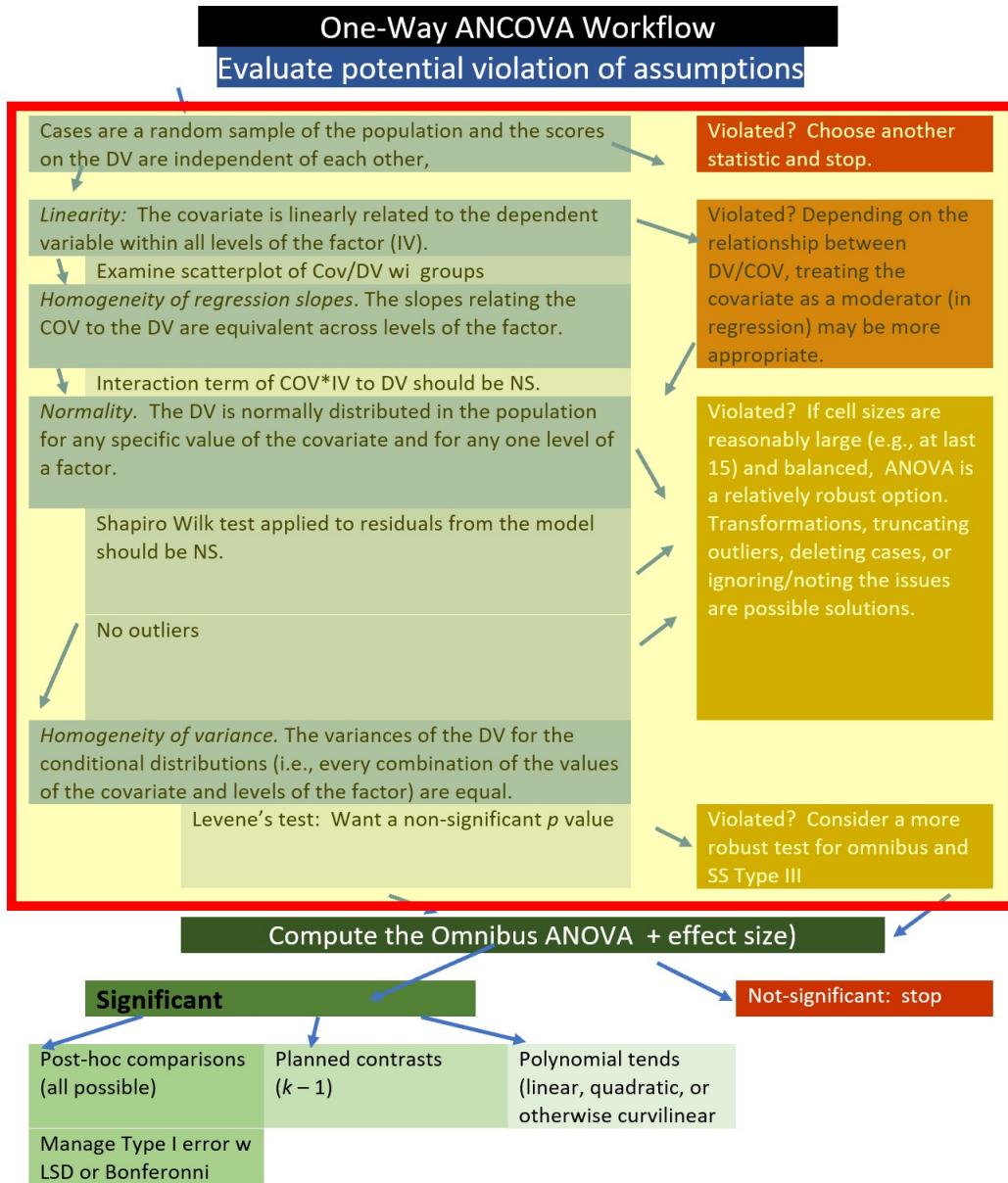


Figure 11.5: Image of the ANCOVA workflow, showing our current place in the process

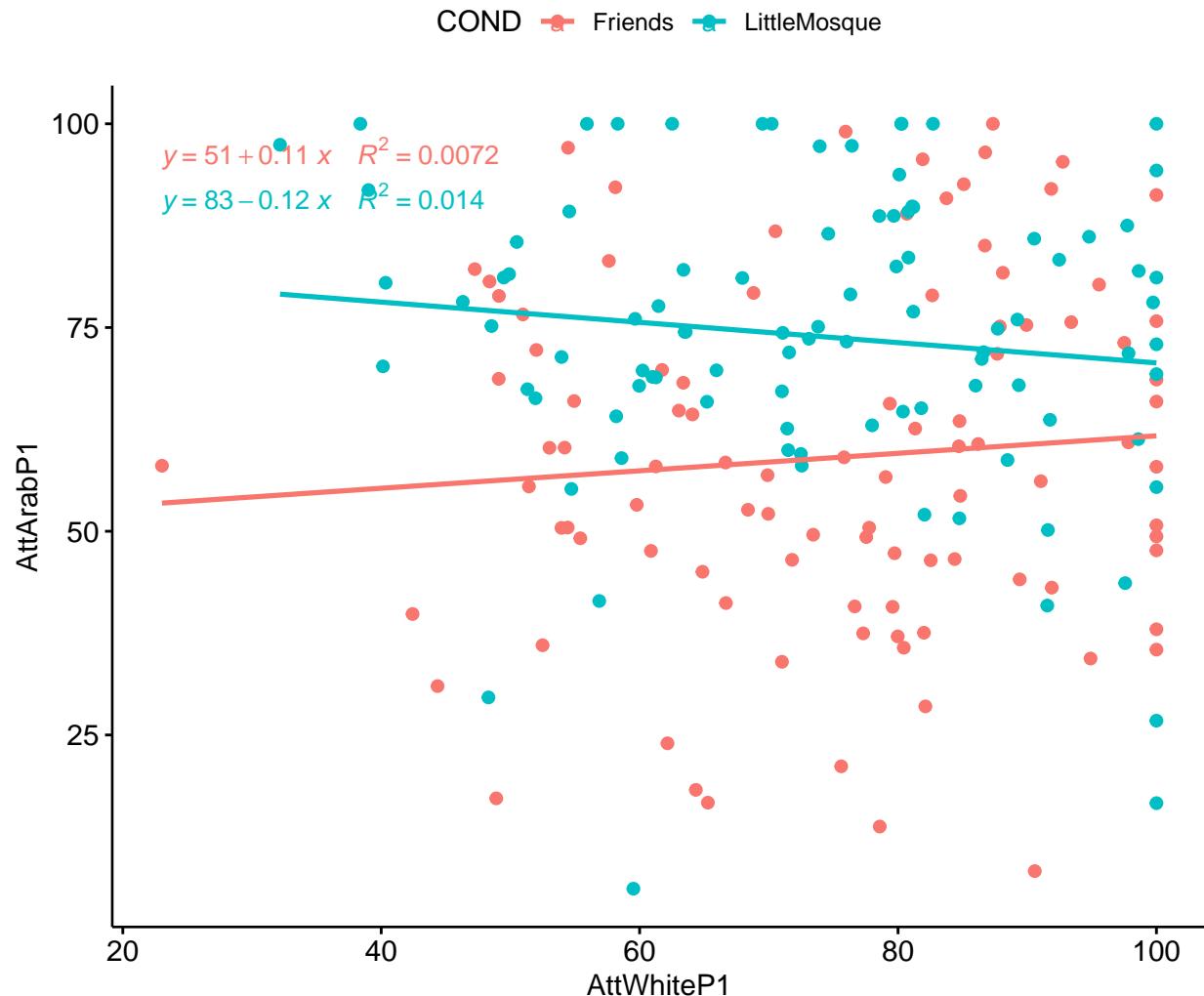
11.5.2.1 Linearity assumption

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))
```

`geom_smooth()` using formula 'y ~ x'



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.

```
Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ COND * AttWhiteP1)
```

Coefficient covariances computed by hccm()

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 rom 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 .cooks.d .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()

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 [2016] has a great section on some of these options.

11.5.2.6 Write-up 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].

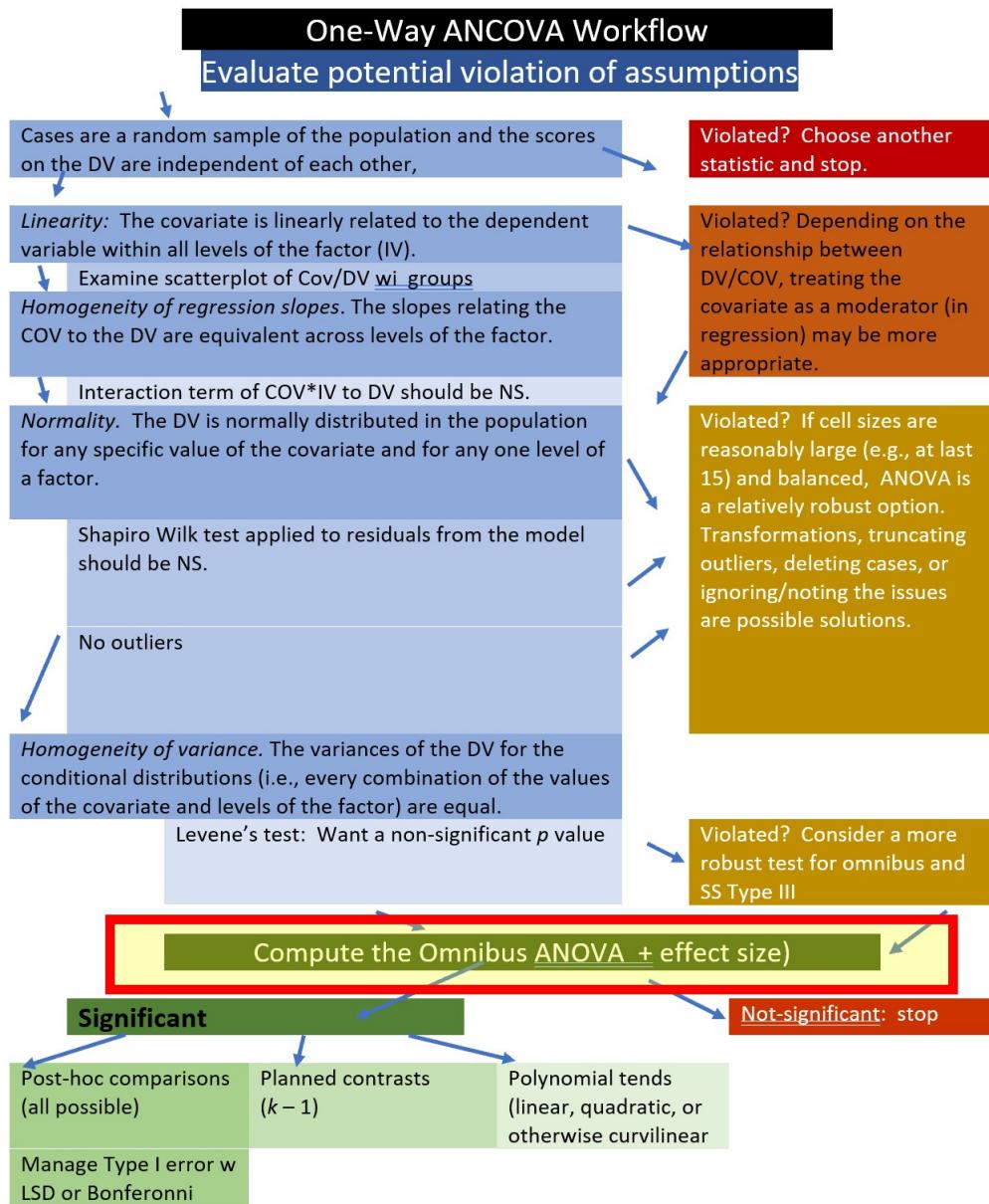


Figure 11.6: Image of the ANCOVA workflow, showing our current place in the process.

```
WhCov_ANCOVA <- Murrar_wide %>%
  rstatix::anova_test(AttArabP1 ~ AttWhiteP1 + COND)
```

Coefficient covariances computed by hccm()

```
rstatix::get_anova_table(WhCov_ANCOVA)
```

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.

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	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

11.5.5 Toward an APA style results section

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")
```

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

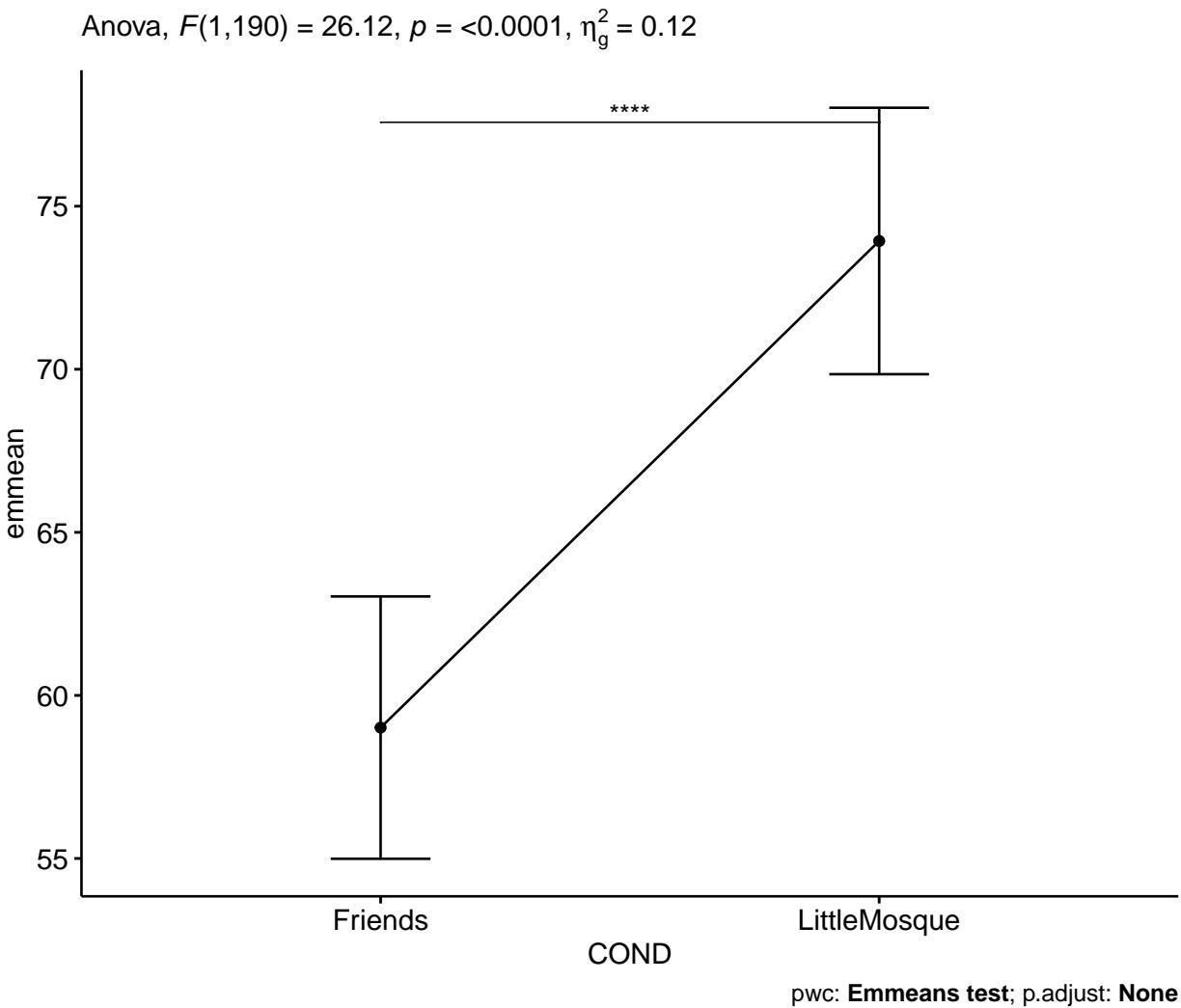
Unadjusted and Covariate-Adjusted Descriptive Statistics

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") +
  geom_errorbar(aes(ymin = conf.low, ymax = conf.high), width = 0.2) +
  ggpubr::stat_pvalue_manual(pwc_B, hide.ns = TRUE, tip.length = FALSE) +
  labs(subtitle = rstatix::get_test_label(WhCov_ANCOVA, detailed = TRUE),
       caption = rstatix::get_pwc_label(pwc_cond))
```



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. 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 “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). 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. Check and, if needed, format data	5	_____
2. Evaluate statistical assumptions	5	_____
3. Conduct omnibus ANCOVA (w effect size)	5	_____
4. If the IV has three or more levels, conduct follow-up tests	5	_____
5. Present means and covariate-adjusted means; interpret them	5	_____
6. APA style results with table(s) and figure	5	_____
7. Explanation to grader	5	_____
Totals	35	_____

References

Bibliography

How can I do post-hoc pairwise comparisons in R? | R FAQ. URL <https://stats.idre.ucla.edu/r/faq/how-can-i-do-post-hoc-pairwise-comparisons-in-r/>.

Anna Lisa Amodeo, Simona Picariello, Paolo Valerio, and Cristiano Scandurra. Empowering transgender 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, a. URL <https://www.datanovia.com/en/lessons/ancova-in-r/>.

- Datanovia. Repeated Measures ANOVA in R: The Ultimate Guide, b. URL <https://www.datanovia.com/en/lessons/repeated-measures-anova-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, seventh edition. edition, 2014a. ISBN 978-0-205-95860-3.
- Samuel B. Green and Neil J. Salkind. *Using SPSS for Windows and Macintosh: analyzing and understanding data*. Pearson, Boston, seventh edition. edition, 2014b. ISBN 978-0-205-95860-3.
- 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/>.
- Robert I. Kabacoff. Power Analysis, 2017. URL <https://www.statmethods.net/stats/power.html>.
- Rex B. Kline. *Principles and practice of structural equation modeling*. Guilford Publications, New York, UNITED STATES, 4th edition, 2016. 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 Djamaludin 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>.
- 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/>.