

Summer 2022

Null Hypothesis Statistical Testing: A Survey of the History, Critiques, and Alternative Methodologies

Bradley David Rogers

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>



Part of the [Educational Psychology Commons](#)

Recommended Citation

Rogers, B. D.(2022). *Null Hypothesis Statistical Testing: A Survey of the History, Critiques, and Alternative Methodologies*. (Doctoral dissertation). Retrieved from <https://scholarcommons.sc.edu/etd/7004>

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

NULL HYPOTHESIS STATISTICAL TESTING: A SURVEY OF THE HISTORY,
CRITIQUES, AND ALTERNATIVE METHODOLOGIES

by

Bradley David Rogers

Bachelor of Arts
Brewton-Parker College, 2002

Master of Education
Lincoln Memorial University, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2021

Accepted by:

Michael Seaman, Major Professor

Christine DiStefano, Committee Member

Melissa Duffy, Committee Member

Katie Wolfe, Committee Member

Tracey L. Weldon, Vice Provost and Dean of the Graduate School

© Copyright by Bradley David Rogers, 2022

All Rights Reserved.

ACKNOWLEDGEMENTS

Above all, I thank my great God and Savior Jesus Christ who with providential condescension used even the process of dissertation writing and program of study to effect the repentance and salvation of a wretch. I am grateful beyond measure. Thine own of thine own we offer unto Thee.

I would like to thank my wife for her unqualified and undeserved love and patience. I would have been unable to complete this work without you. You make me better. I am also grateful to my children, Annora and Emilia, for the vision of beauty and purpose they have brought into my life. You provide joy even in the darkest moments of frustration and doubt. To my Mother, I am eternally grateful for your support and love.

I am also deeply grateful to my supervisor, Dr. Michael Seaman, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and improved the quality of my work. A special thanks to the members of my committee. Your feedback and insight was invaluable. Lastly, I would like to extent my gratitude to my colleagues and friends for their support during this process.

ABSTRACT

Considered normative from the second half of the 20th century (Danziger, 1990), null hypothesis statistical testing (NHST) has received consistent, largely unheeded criticism. Critiques have received more attention in recent years with the recognition of the replication crisis in the social sciences and the American Statistical Society's statement and special issue on p values (Wasserstein & Lazar, 2016; Wasserstein et al., 2019). This paper seeks to provide a framework for understanding the issue by looking at the history of NHST and investigating some alternatives to the methodology. A brief review of the history, spread, and critiques of NHST is provided to help place the issues in context. Supplemental and alternative methods are described to demonstrate the range of available options to NHST. A selection of viable alternative methodologies and supplements to NHST are demonstrated using simulated data sets. Those methods that provide decision rules are compared with NHST using Monte Carlo method simulations.

TABLE OF CONTENTS

Acknowledgements	iii
Abstract	iv
List of Tables	vi
List of Figures	vii
List of Symbols	xi
List of Abbreviations	xiii
Chapter 1 Introduction	1
Chapter 2 Null Hypothesis Statistical Testing: A Survey of the History, Critiques, and Alternatives.....	6
Chapter 3 An Introduction and Demonstration of Select Proposed Alternatives to Null Hypothesis Statistical Testing	65
Chapter 4 A Monte Carlo Methods Comparison of Conventional NHST with Three Alternative Methodologies.....	129
Chapter 5 Conclusion.....	188
References.....	200
Appendix A: Simulation Functions R Code	219
Appendix B: Monte Carlo Simulation R Code	241
Appendix C: Figures R Code	253

LIST OF TABLES

Table 2.1 Social Sciences Adoption of NHST from 1945 – 2007	15
Table 2.2 Management Sciences Adoption of NHST from 1945 – 2007	16
Table 3.1 Critiques of NHST Addressed by Each Proposed Method.....	78
Table 3.2 Two-group Design Parameters	94
Table 3.3 Three-group Design Parameters	95
Table 3.4 2x2 Between-subjects Design Parameters	96
Table 4.1 Linear Model Effect Parameters	145
Table 4.2 Decision Criteria by Method.....	147
Table 4.3 Type I Error Under Normality Condition	150
Table 4.4 Type I Error Under Violations of Homogeneity	153
Table 4.5 Type I Error Under Distributional Violations.....	157
Table 4.6 Summary of Results for Type I Error	160
Table 4.7 Power Under Normality Condition	162
Table 4.8 Power for Two-group linear model under violations of Homogeneity	169
Table 4.9 Real Power under violations of Homogeneity ratio: K = 3 group	170
Table 4.10 Power for Two by two between groups linear model under violations of Homogeneity.....	171
Table 4.11 Power Under Skewed Distribution Condition	179
Table 4.12 Power Under Bimodal Distribution Condition	181

Table 4.13 Summary of Results for Power	182
Table 5.1 Example of Power by Pop Effect and Sample Size	192
Table 5.2 Approximate Sample Size for 80%er by Population Effect	193

LIST OF FIGURES

Figure 2.1 Image Recall Experiment Pattern Sample	58
Figure 2.2 Image Recall Experiment Ordinal Pattern Sample.....	59
Figure 2.3 Deep Structure Matrices for Sex and a Four Point Likert Style Scale	60
Figure 3.1 Expected Ordinal Pattern.....	90
Figure 3.2 Two-by-Two Between Subjects Multi-unit Frequency Histogram.....	122
Figure 4.1 Illustration of point null hypothesis and interval null with interval estimate	140
Figure 4.2 Type I Error and CI by Linear Model Under the Normality Condition.....	151
Figure 4.3 Type I Error Ratios by Linear Model Under the Normality Condition.....	152
Figure 4.4 Type I Error and CI by Linear Model Under Violations of Homogeneity	154
Figure 4.5 Type I Error Ratios by Linear Model Under Violations of Homogeneity: 2:1	155
Figure 4.6 Type I Error Ratios by Linear Model Under Violations of Homogeneity: 4:1	156
Figure 4.7 Type I Error for Skewed and Bimodal Distributions by Linear Model.....	158
Figure 4.8 Skewed Distribution Type I Error Ratios by Linear Model	159
Figure 4.9 Bimodal Type I Error Ratios by Linear Model	159
Figure 4.10 Power Ratios for Two-group Linear Model by Population Effect Size Under Normality Condition	163

Figure 4.11 Power Ratios for K = 3 group linear model by Population Effect Size Under Normal Condition	164
Figure 4.12 Power for 2x2 Between Group Linear Model by Population Effect Size Under Normality Condition	165
Figure 4.13 Two-group Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)	167
Figure 4.14 Power and CI for Two-group Linear Model by Population Effect Size and Var. Ratio	169
Figure 4.15 Power and CI for K=3 Group linear model by Population Effect Size and Var. Ratio	170
Figure 4.16 Power and CI for 2x2 Between Group Linear model by Population Effect Size and Var. Ratio	171
Figure 4.17 Power ratios for Two-group Liner Model by Population Effect Size and Variance Ratio	173
Figure 4.18 Power Ratios for K = 3 linear model by Population Effect Size and Variance Ratio	174
Figure 4.19 Power Ratios for 2x2 Between Group linear model by Population Effect Size and Variance Ratio	174
Figure 4.20 Two Group linear model with 2:1 Var. Ratio Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)	175
Figure 4.21 Two Group linear model with 4:1 Var. Ratio Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)	176
Figure 4.22 Power Ratios by Linear Model, Population Effect Size, and Variance Ratio	178
Figure 4.23 Power Rates with CI by Linear Model, Distribution, and Population Effect Size	180

LIST OF SYMBOLS

H_0	Statistical null hypothesis
H_A	Statistical alternative hypothesis
H_M	Neyman-Pearson main hypothesis
α	Type I error rate
β	Type II error rate
p	Probability value of the observed data under the null hypothesis
s	Shannon information value
r	Pearson correlation coefficient
n	Sample size
c	Chance value
k	Number of experimental groups
μ	Population mean
σ	Population standard deviation
d	Cohen's d effect size
η^2	Eta-squared effect size
t	Sample t statistic
M	Sample mean
SD	Sample standard deviation
F	F statistic
p_δ	Second generation p value statistic

\approx Approximately

δ For SGPV, the distance between intervals in δ units, where $\delta = \frac{1}{2}|H_0|$.

LIST OF ABBREVIATIONS

AJPH.....	American Journal of Public Health
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
APA.....	American Psychological Association
ASA.....	American Statistical Association
BASP.....	Journal of Basic and Applied Psychology
BF.....	Bayes Factor
BFB	Bayes Factor Bound
BPR	Binary Procrustes Rotation
CI.....	Confidence Interval
D.....	Data
I	Interval Estimate
MANCOVA.....	Multivariate Analysis of Covariance
MES	Minimum Effect Size
MESP	Minimum Effect Size Plus <i>p</i> value
MPSD.....	Minimum Practically Significant Difference
NHST	Null Hypothesis Statistical Test
OOM.....	Observation Orientated Modelling
PCC.....	Percent Correct Classifications

Pr	Probability
PM.....	Pub Med
PMC	Pub Med Central
SCD.....	Single-case designs
SGPV	Second Generation p value
TFSI	Task Force on Statistical Inference

CHAPTER 1

INTRODUCTION

The standard inferential methodology in the social sciences and related fields is null hypothesis statistical testing (NHST) used in conjunction with experimental/control group experimentation. In its basic form, a sample is selected, ideally at random, from a population of interest. Each of the n participants in the sample is randomly assigned to one of two or more experimental treatment conditions, consisting of treatment and control groups. These conditions comprise the independent variable. Random selection and random assignment will mitigate any selection effect and attending nuisance variables, though either or both methods are difficult to realize in much social science research where the researcher does not have complete control of the experimental conditions. Each participant's performance in their respective experimental condition to which they were assigned is recorded, usually using a psychological instrument such as a test, a Likert scale, or a checklist; this is the dependent variable. The average dependent variable performance for all n subjects in their assigned IV condition is then calculated. Descriptive statistics are evaluated to determine where there is a detectable difference between the sample groups.

Researchers, however, do not only wish to only determine whether there are differences between sample groups but rather whether those differences are present among other similar groups. In the NHST paradigm, this is understood as inferring from the sample statistics to parameters of a hypothesized population of which the sample is

representative. In NHST, two mutually exclusive hypotheses are posited: the null hypothesis (H_0) and the alternative hypothesis (H_A). The alternative hypothesis is reflective of the research hypothesis, whereas the null hypothesis is a negation of the alternative hypothesis. Typically, researchers will, without much consideration, define H_0 as a nil-null hypothesis specifying the absence of an effect. With the null hypothesis so defined, the H_A takes the form of its opposite, indicating the presence of an effect within a population, understood as a difference between groups or the presence of an association. There is typically no attempt to make a precise prediction about the H_A such as the size of an effect or association. For instance, if a researcher predicts that there will be a difference between the means of two groups, the alternative hypothesis states that the two means are different, and the null hypothesis indicates that the means are precisely equal. Researchers will select a Type I error rate or alpha that is deemed acceptable, customarily a rate of $\alpha = .05$ though other values may be used. Type I errors occur if a null hypothesis is rejected when the null hypothesis is true. A test statistic (e.g., t , F , χ^2) is calculated along with a corresponding p value. The p value is the probability of obtaining a test result at least as extreme as the obtained result if the null hypothesis is true for the population from which the sample data were drawn. If the obtained p value is less than or equal to the selected alpha level, then the null hypothesis is rejected, and the results are declared to be “statistically significant” since the observed data is deemed to be sufficiently unlikely if the null hypothesis is true. Since the null and alternative hypotheses are mutually exclusive, the determination of the probable falsity of the null hypothesis allows for the acceptance of the alternative hypothesis and for inferential claims to be made regarding the research hypothesis. If the p value is greater than the

alpha level, then the null hypothesis is not rejected, and the results are considered inconclusive.

Statement of the Problem

Considered normative from the second half of the 20th century (Danziger, 1990; Gigerenzer and Murray, 1987), the NHST research paradigm is not without its critics. As noted by Kline (2013), the assumptions of NHST are far more restrictive than is commonly thought by many researchers and are often not met, which results in potentially biased p values. NHST is not just a test of the H_0 but of a statistical model comprised of several assumptions, the violation of which is reflected in the p value. A p value is the result not only of the probability of the data given the H_0 , but of random variation and violations of model assumptions, none of which can be differentiated by the p value (Amrhein, Tafimow, & Greenland). Null hypothesis statistical tests also have methodological limitations related to sample size sensitivity and the type of null hypothesis (nil-null), which is typically used. There is also broad interpretational misunderstanding by researchers of what claims are valid when a null hypothesis is rejected, which manifests as misinterpretations or exaggerations of inferences. Lastly, because of professional pressure to publish and a publication bias in favor of novel statistically significant findings, there is evidence that some researchers have taken to using questionable practices such as data mining, the selective reworking of data, or performing multiple statistical analyses to identify patterns in data that produce statistically significant results.

Criticism of NHST is long-standing and has existed for as long as the methodology itself (Kline, 2013); however, over the last decade, two concurrent

phenomena have contributed to an increased awareness of the problems with NHST as it is commonly practiced as well as added urgency to calls for reform of methodological practices. No longer an esoteric topic only considered by statisticians and a subset of researchers, the problems with NHST methodology moved from little-read academic journals into popular publications and online blogs ScienceNews (Siegfried 2010, 2014), Nautilus (Siegfried 2013), and Nature (Nuzzo, 2014). Simultaneously, the replication crisis afflicting many scientific fields, especially the social sciences, was definitively demonstrated and widely reported (Wasserstein & Lazar, 2016). In response, some academic publications have begun to reconsider methodological requirements for publication, discouraging, or in some cases banning, NHST or key features thereof while at the same time encouraging supplemental or alternative methods (International Committee of Medical Journal Editors, 2010; Trafimow, 2015). The American Statistical Association, the primary professional organization for statisticians, took unprecedented action in 2016 and 2019 and issued statements on the use and misuse of p values and accompanied the latter with articles containing suggestions for ways to augment NHST as well as alternative or supplemental methodologies.

This dissertation aims to add to the understanding of alternatives and supplements to NHST by exploring an exemplary sample of methodologies drawing from methods mentioned in the various ASA articles, including notable options not discussed therein. An attempt will be made to understand the various methods by investigating each and identifying which problems and criticisms of NHST a methodology attempts to address. It will also explore each methodology's comparative strengths and weaknesses by simulating various conditions and parameterizations, drawing a sample, and then

applying each method to the sample. The following research questions will guide the investigation:

1. How well do a set of exemplar alternative methods to NHST address the known problems and limitations of NHST, such as assumption violations, sample size and nil-null limitations, dichotomization of decisions, misinterpretation, and aggregation?
2. How does a set of exemplar alternative methods to NHST differ in identifying an effect and correcting NHST issues when applied to simulated data sets with varying effect sizes, sample sizes, and distributions?
3. What are the relative strengths and weaknesses of each method relative to NHST and each other?

CHAPTER 2

NULL HYPOTHESIS STATISTICAL TESTING: A SURVEY OF THE HISTORY, CRITIQUES, AND ALTERNATIVES¹

Null hypothesis statistical testing (NHST) has been the normative research methodology in social science since the second half of the 20th century (Danziger, 1990). Criticism of NHST is long-standing and has existed for as long as the methodology itself (Kline, 2013). However, over the last decade, two concurrent phenomena have contributed to an increased awareness of the problems with NHST as it is commonly practiced as well as added urgency to calls for reform of methodological practices. No longer an esoteric topic only considered by statisticians and a subset of researchers, the problems with NHST methodology moved from little-read academic journals into popular publications and online blogs (Cumming, 2013; Nuzzo, 2014; Siegfried 2010, 2014). Simultaneously, the replication crisis afflicting many scientific fields, especially the social sciences, was definitively demonstrated and widely reported (Wasserstein & Lazar, 2016). In response, some academic publications and professional organizations have renewed reconsideration of the role of NHST and p values in research, recommending restrictive methodological requirements for publication, discouraging, or in some cases banning, NHST or key features thereof while at the same time encouraging

¹ B. D. Rogers. Not yet submitted for publication.

supplemental or alternative methods (International Committee of Medical Journal Editors, 2010; Trafimow, 2015). Given these developments, the subject of this investigation is the origin, history, and criticism of, as well as a sampling of alternatives to NHST methodology.

History and Spread

The rudiments of NHST have a long pedigree, with predecessors to modern NHST methods appearing in scientific papers as early as the 18th century. Daniel Bernoulli, a polymath and pioneer of probability theory, conducted significance tests in a 1735 report on the randomness of planetary motion, which provided the first known example of the calculation of a p value corresponding to a non-extreme value of a test statistic (Hald, 1998). Likewise, Pierre-Simon Laplace performed hypothesis tests on the motion and origin of comets (Hald, 1998). Yet it was in the 20th century that the core elements of NHST were formally developed. Karl Pearson (1900) systematized a procedure with his χ^2 test of significance and the corresponding concept of the p value. Ronald Fisher built on the work of Pearson and others such as William Gosset, who developed the z test and, by extension, the t -test, to suggest a significance test with a null hypothesis, a p value, and an arbitrary significance value. Fisher viewed inductive inference as the “only process by which new knowledge comes into the world” (Fisher, 1966) and considered p values and significance tests as an objective inferential methodology that was key to advancing scientific knowledge (Hubbard & Bayarri, 2003). Fisher considered statistical analyses such as significance tests to be essential to the development of the social sciences as these techniques, in his opinion, could serve to raise them to the rank of actual sciences, such as the natural sciences (Fisher, 1970).

Fisher was responsible for popularizing the use of significance tests and p values with the publication of multiple editions of his books *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935).

Fisher's Test of Significance

Fisher's significance testing model is not NHST as it has come to be practiced. There are many similarities but also important differences, as I will discuss below. Fisher's model posits a null hypothesis (H_0) which can be a nil hypothesis of no relationship or effect or any other hypothesis a researcher wishes to investigate. A correlation of .3, a reduction of five cigarettes smoked per day, or an increase of 50 points on the SAT mathematics assessment are each potential null hypotheses (Fisher, 1955, 1956; Gigerenzer, 2004). The sample of the H_0 comes from a hypothetical infinite population with a known sampling distribution. Fisher's model is not a test of competing hypotheses and does not consider an explicit alternative hypothesis. A p value is calculated under the null hypothesis for the obtained data, $\Pr(D | H_0)$. A level of significance is then used to identify "statistically significant results." Under Fisher's model, a level of significance does not need to be set a priori, nor must it be consistent for all testing occasions. Fisher suggested using a significance value of $p \leq 0.05$ as a convenient standard, especially with novel research (Carver, 1978; Fisher, 1925), but his model allows for significance values to be flexible from case to case depending upon the reasoned judgment of researchers (Fisher, 1956). For instance, a p value of .06 is considered to convey about the same degree of evidence as a value of .04, and either may be used as a significance cut-off value. If a researcher determines that a much lower value is needed, a significance value of 0.01 or .001 may be used (Fisher, 1925). Fisher's model

also considers p values to be gradable, meaning that the smaller the p value, the greater the evidence against H_0 . Despite suggesting a significance level of .05, Fisher also advised reporting the obtained p value and for researchers to consider the level of sensitivity related to sample size in relation to the obtained value.

A significant result is interpreted as indicating that either a rare event has occurred or the H_0 does not satisfactorily explain the research results and thus, is false. If it is concluded that the H_0 is false, it follows that the experiment demonstrated a positive result. Fisher was insistent that a low p value from a single experiment alone does not prove that whatever has been manipulated in the experiment explains the obtained results. Instead, it provides inferential evidence against the null hypothesis. A robust research design, which includes randomization, was considered of paramount importance to Fisher in order to make plausible inferences. He also insisted that further research was needed, even in the event of a significant test result, to establish that the effects were due to the treatment. Fisher considered significant results to be single data points and suggested combining the results of significance tests across studies, which ideally would include significant and non-significant results (Fisher, 1932). For Fisher, an effect is deemed to be established when well-designed repeated experiments rarely fail to yield statistically significant results.

Neyman-Pearson Test of Hypotheses

Jerzy Neyman and Egon Sharpe Pearson attempted to improve upon Fisher's procedure (Fisher, 1955; Pearson, 1955) but ultimately developed a markedly different approach to data testing. The primary innovation of Neyman-Pearson is the introduction of explicit alternative hypotheses. The hypothesis considered to be the most likely or the

most important to the research question is designated the main hypothesis (H_M) under the Neyman-Pearson model. The alternative hypothesis is conceptualized as a second population distribution that sits alongside the main hypothesis population distribution on the same continuum of values. The test is not one of significance of the data under the H_0 , like Fisher's, but one of competing hypotheses. As a result, controlling decision errors over the long run is very important and resulted in introducing the now familiar and important concepts of Type I error, the maximum allowable Type I error rate (α), effect size, Type II error, the maximum allowable Type II error rate (β), and power. Extending from the testing of more than one hypothesis, Neyman-Pearson introduced the concept of Type I error, wrongly rejecting the main hypothesis, and an a priori α defined as the probability of committing a Type I error. Alpha is not a level of significance but an error probability. The presence of the H_A also allows for the introduction of an effect size, which is considered the difference between the H_A and H_M . Researchers are to select an expected effect size value a priori.

Though the Neyman-Pearson model considers two competing hypotheses, it only tests the data under the main hypothesis. Neyman-Pearson's H_M is very similar to Fisher's H_0 and was at times referred to as a null hypothesis by Neyman-Pearson. There are, however, critical differences between the concepts. Neyman-Pearson's H_M must be considered in the experimental design stage. Further, the selected minimum effect size (MES) is incorporated into the analysis by using a priori power analysis and the potential use of a null band (e.g., $H_M: M1-M2 = 0 \pm \text{MES}$). The power analysis involves using α , β and the MES to calculate the necessary sample size for a selected level of statistical power. Power is defined as $1 - \beta$. Neyman-Pearson also recommended considering the

power of a given test and selecting the optimal test as the one that would have the highest power. Because the test is a decision between two competing hypotheses, the decision is binary, and the decision rule is fixed and inflexible. Further, evidence of the probability of the sample data is of little relevance in the Neyman-Pearson model; therefore, a critical or rejection region for the test statistic is preferred to a p value. In considering the interpretation of the findings of a Neyman-Pearson hypothesis test, it must be noted that it is better understood as an example of decision theory, not a theory of inference. Neyman-Pearson did not interpret the outcomes of hypothesis tests as providing any information about the truth or falsity of the tested hypothesis; rather, the model is conceptualized as a means of controlling and minimizing decision errors in the repeated application of the same test.

Null Hypothesis Statistical Tests

NHST procedure is a hybrid of Fisher's test of significance and elements of Neyman-Pearson hypothesis testing. In NHST, there are two mutually exclusive hypotheses posited, the null hypothesis (H_0) and the alternative hypothesis (H_A). Typically, researchers will, without much consideration, define H_0 as a nil-null hypothesis specifying the absence of an effect. With the null hypothesis so defined, the H_A takes the form of its opposite, indicating the presence of an effect within a population, understood as a difference between groups or the presence of an association. There is typically no attempt to make a precise prediction about the H_A such as an exact mean or the size of an effect or association. For instance, if a researcher predicts that there will be a difference between the means of two groups, the alternative hypothesis states that the two means are different, and the null hypothesis indicates that the means are precisely

equal. Researchers will select a Type I error rate or alpha, customarily a rate of $\alpha = .05$. A p value is calculated, and if the obtained p value is less than or equal to the selected alpha level, which is treated as Fisher's level of significance, then the null hypothesis is rejected, and the results are declared to be "statistically significant." Since the null and alternative hypotheses are mutually exclusive, the determination of the probable falsity of the null hypothesis allows for the acceptance of the alternative hypothesis and for inferential claims to be made regarding the research hypothesis. Such inferential claims are inconsistent with both Fisher's test of significance, which considers the test as providing inferential evidence about the H_0 only, and the Neyman-Pearson model, which eschews any inference altogether. If the p value is greater than the alpha level, then the null hypothesis is not rejected, and the results are considered inconclusive. Contrary to Fisher's significance testing, NHST posits an alternative hypothesis and then, contrary to both methods, uses a p -value to decide between the null hypothesis and an alternative hypothesis. The probability of a Type I error (α) is not based on judgment and careful consideration of loss functions as in the Neyman-Pearson model but is mechanically set at .05. Further, the probability of a Type II error (β) is usually not considered since a priori power analysis is used infrequently, with researchers instead relying on "rules of thumb" or common in-field practice to determine sample size.

Spread of NHST

The importance of Fisher's books in the spread of significance tests is difficult to overstate, with Savage (1954) calling them the two most influential works of the 20th century in the development of statistics. Indeed, a drastic change occurred in social science methodology in the years immediately following the first editions of Fisher's

books. Gigerenzer and Murray (1987) coined the term *inference revolution* to describe the dramatic shift in the research practices of psychology between 1940 and 1955. In the years preceding 1940, the dominant research tradition followed the methods of Wilhelm Wundt, generally considered the founder of experimental psychology, which involved experimentation and detailed observation of individual participants. When replicated, these experiments were thought to reveal general psychological principles or laws. The inference revolution saw the replacement of the Wundtian tradition with group experiments and the analysis of aggregate statistics using inferential methods, especially significance tests. Danziger (1990), who wrote a history of this shift in psychological research methodology, interpreted the change as resulting from social pressures on psychologists, both internal and external, to achieve the respectability associated with being considered a valid scientific discipline and to show practical utility. Treatment group experimentation and significance tests were thought to answer both these needs, and the shift to these methods was dramatic. Surveys of published articles in psychological journals provide evidence to support Gigerenzer's inference revolution hypothesis. Using the American Journal of Psychology as an exemplar, Gigerenzer calculated that the percentage of published empirical studies rose from 25% to 80% between 1915 and 1950 and that during the same period, the percentage of published single-case articles decreased from 80% to 17% (Gigerenzer, 2000). Null hypothesis statistical tests also saw a marked increase; from 1934 to 1950, there were only 17 articles published using NHST, yet by 1955, 80% of the articles surveyed used some form of the procedure (Rucci & Tweney, 1980). Data gathered from the Journal of Applied Psychology showed that between 1940 and 1945, only 25% of empirical studies

used significance tests with p values, whereas the amount leapt to 59.7% from 1950 to 1954 (Hubbard et al., 1997). Hubbard and Ryan (2000) provide further evidence of an inference revolution in their analysis of twelve American Psychological Association journals where only 4% of empirical studies reported p values from 1935 to 1939; this increased to 22.7% from 1940 to 1945 and jumped to 71.7% for the period from 1950 to 1954.

The use of both empirical data and NHST continued to increase in psychology in the years following the so-called inference revolution, assuming a near-monopolistic status, yet the phenomenal spread of the methodology was not limited to psychology. A survey by the decade of top journals in the social and management sciences from 1945 to 2007 provides evidence that the increase of empirical data and NHST in the social and management sciences followed a similar trajectory to that of psychology (Hubbard, 2016). Table 2.1 shows the results from the social sciences, and Table 2.2 displays the results from the management sciences. Papers were evaluated for the type of data, either empirical or non-empirical, and of the empirical studies, whether the data were analyzed using NHST. All told, some 10,874 papers were evaluated from 16 leading social science journals and 4,594 articles from 18 management science journals. For the period from 1945 to 1949, 70% of published psychological research articles used empirical data, and of those, 62.5% used NHST. From 2000 to 2007, 91% of published articles were empirical studies, with 95.4% using NHST.

Table 2.1 Social Sciences Adoption of NHST from 1945 - 2007

<u>Years</u>	<u>Psychology</u>		<u>Sociology</u>		<u>Geography</u>		<u>Political Science</u>	
	<u>Empirical</u>	<u>NHT</u>	<u>Empirical</u>	<u>NHST</u>	<u>Empirical</u>	<u>NHST</u>	<u>Empirical</u>	<u>NHST</u>
	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>
1945 -								
1949	70.5	62.5	34.1	26.8	34.5	0.0	2.4	0.0
1950 -								
1959	85.0	81.3	52.1	50.6	29.8	4.2	16.5	14.3
1960 -								
1969	89.0	86.9	70.0	51.0	49.5	16.2	36.7	30.0
1970 -								
1979	85.9	90.4	75.9	68.8	55.0	39.4	52.0	57.9
1980 -								
1989	78.3	91.9	77.3	86.8	67.2	58.3	63.0	72.6
1990 -								
1999	80.3	91.9	82.8	91.6	58.8	54.8	59.9	91.6
2000 -								
2007	91.0	95.4	90.8	90.3	69.6	42.9	64.9	93.6

The social science data show that the adoption of empirical studies and NHST in psychology and sociology follow a similar trend. In both fields, there was an expansion of empirical studies and a correlating increase of NHST. Though sociology researchers were several decades behind psychologists, the percentages are nearly the same by the decades of 1980 to 1989. From 2000 to 2007, over 90% of all published research in both psychology and sociology were empirical studies, and over 90% of these were analyzed using NHST. Political science research from 2000 to 2007 was not yet as dominated by empirical studies as psychology and sociology, with 64.9% empirical studies, but of those, 93.4% used NHST. Geography research seems to be an outlier with a similar number of empirical studies as political science from 1970 to 2007, but a relatively modest percent of empirical studies in the field used NHST; in fact, there was a decrease from the decade of 1990 to 1999, with 54.8%, to 2000 to 2007 with 42.9%. Hubbard speculates that the limited use of NHST by geography researchers can be accounted for

by the discipline's use of spatial statistics that offer fewer opportunities for hypothesis testing (Hubbard, 2016).

The trend observed in the social sciences is not unique and can be observed in other fields of study, from management science to biomedical research. Evidence of a similar rise and expansion of empirical studies with a corresponding increase in the use of NHST is seen in management science data in Table 2.2. Notable is the swiftness of the adoption of NHST in accounting, economics, and finance from the decade of 1950-1959 to 1970-1979. During these years, NHST went from a little-used statistical tool with 0.0%, 13.8%, and 10.0% usage, respectively, to assume the position as the dominant research technique in empirical studies with 86.2%, 74.6%, and 85.2% use. Management researchers preceded the other fields by several decades in adopting both empirical studies and NHST.

Table 2.2 Management Sciences Adoption of NHST from 1945 - 2007

<u>Years</u>	<u>Accounting</u>		<u>Economics</u>		<u>Finance</u>		<u>Management</u>	
	<u>Empirical</u>	<u>NHST</u>	<u>Empirical</u>	<u>NHST</u>	<u>Empirical</u>	<u>NHST</u>	<u>Empirical</u>	<u>NHST</u>
	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>	<u>%</u>
1945 - 1949	8.3	0.0	21.3	0.0	29.2	0.0	34.2	53.8
1950 - 1959	11.6	0.0	30.1	13.8	50.0	10.0	52.4	67.7
1960 - 1969	16.9	42.9	39.0	52.3	51.2	54.7	50.7	69.8
1970 - 1979	38.9	86.2	45.8	74.6	47.1	85.2	75.5	81.8
1980 - 1989	70.1	96.8	55.4	83.1	62.9	89.4	79.1	89.3
1990 - 1999	75.5	98.3	66.3	85.7	75.9	95.2	80.3	91.7
2000 - 2007	77.2	98.6	71.0	89.1	80.8	95.0	83.1	85.6

The evidence of an expansion of NHST in biomedical research is also quite strong, though it is not as dominant. A survey of all biomedical literature from 1990 to 2015 was conducted using PubMed (PM) and PubMed Central (PMC) databases (Chavalarias et al., 2016). Over 16 million items (articles and other resources) were surveyed in PM and 844,00 full-text articles from PMC. Text mining was used to assess the presence of p values in PM article abstracts. The analysis revealed that the percentage of reported p values increased from 8% in 1990 to 17% in 2015. Also noteworthy is that certain types of research in biomedical studies use NHST at a much higher rate. Articles using randomized controlled trials, for instance, reported p values 60% of the time in their abstracts. When complete texts in PMC were analyzed, the total number of articles containing p values was 51.1%, and even this is an underestimate since the text mining technique could not detect p values contained in tables or figures. Regardless, the expansion and wide use of NHST in biomedical research are evident, and, as noted by Ioannidis (2019), it is by far the dominant method of statistical inference used in the discipline. Of 100 papers selected from the PMC dataset for manual evaluation, Ioannidis found that 55 reported p values, but only 4 used other methods of statistical inference (Ioannidis, 2019).

Criticisms

Violations of Assumptions

Null hypothesis statistical tests require that certain assumptions be met to ensure that p value calculations and inferences are accurate. The p value is the prior conditional probability of observing the obtained data, or data even more inconsistent with the null hypothesis, if the null hypothesis is true and other assumptions are met. For many,

perhaps most, NHST, these assumptions include

1. Random sampling
2. Normal distribution of errors
3. Equality of population error variances (homoscedasticity)
4. Independence of observations
5. Sampling and measurement error are the only sources of error (Kline, 2013; Loftus, 1996).

Such assumptions are more restrictive than is commonly thought by many researchers and are often not met (Kline, 2013), resulting in potentially biased p values. When p values are lower than they should be because of violations of assumptions, there is a positive bias meaning that the Type I error rate is higher than the stated α level.

Conversely, if p values are too high, there is a negative bias resulting in inflated Type II errors. (Kline, 2013).

As one example, the requirement of random sampling from known populations, which is crucial to make valid inferences, is rarely met in psychology and other social sciences. The samples are frequently not random but convenience samples, often drawn from homogenous groups, such as university students, or they consist of non-random groupings such as students in school classrooms. Further, researchers frequently do not attempt to define the population from which the sample is allegedly selected (Kline, 2013). All of this results in a too-conservative estimation of standard errors that causes an under-estimate of p values and overstatement of statistical significance (Reichardt & Gollob, 1999).

The assumption of normality is also crucial for many of the most common

methods using NHST, the violation of which affects both p values and power that occurs whether group sizes are equal or not (Erceg- Hurn & Mirosevich, 2008; Kline, 2013).

Micceri (1989) examined 440 large data sets from psychological and educational literature, which were composed of a variety of ability and aptitude measures (e.g., math and reading tests) as well as psychometric measures (e.g., scales measuring personality, anxiety, anger, satisfaction, locus of control). None of the examined data were found to be normally distributed. The distributions were most frequently multimodal, skewed, and leptokurtic. Micceri's study found that data typically used in psychological experiments more closely resembles an exponential distribution than a normal one (Micceri, 1989).

Other surveys have confirmed that commonly used psychological variables have asymmetrical and skewed distributions (Bradley, 1977; Erceg-Hurn & Mirosevich, 2008; Taylor, 1985).

Homoscedasticity, the equality of population variances, is also an assumption of many procedures and is frequently not valid. The degree of heteroscedasticity, inequality of population variances, can be thought of as a variance ratio. Equal variances would result in a variance ratio of 1:1, meaning that the populations are homoscedastic.

Keselman (1998) conducted a survey of ANOVA analyses from 17 educational and developmental psychology journals to determine if there was evidence of homoscedasticity. A sample variance ratio was calculated for each study by dividing the largest variance by the smallest. Keselman found that in one-way design studies, the median variance ratio was 2.25:1 with a mean ratio of 4:1. In factorial designs, the variance ratio median was 2.89:1, and the mean was 7.84:1. Erceg-Hurn & Mirosevich (2008) did a more informal review but found similar violations of homoscedasticity. Two

issues of the *Journal of Experimental Psychology: General* and the *Journal of Experimental Psychology: Human Perception and Performance* were reviewed, and 28 studies were selected. Most of the sample variance ratios were between 2:1 and 4:1; however, several were extreme, with variance ratios greater than 10:1, and four studies had variance ratios exceeding 39:1. Though these ratios are based on sample rather than population variances, and the assumption of homoscedasticity is about population variances, they suggest that such an assumption is routinely violated.

The problem of assumption violations is exasperated because there is evidence that researchers seldom verify many of the assumptions of NHST and thus neither adjust their analyses to suit the data better nor report the violations in research articles (Greenland et al., 2016). Keselman (1998) reviewed over 400 articles published in 17 prominent journals of psychology from 1994 to 1995, focusing on those using ANOVA, ANCOVA, and MANCOVA, to determine whether researchers reported investigating if their data met the assumptions of the analyses performed. Keselman found that of the 61 instances in which univariate ANOVA was used, 11.4% of articles referenced validating distributional assumptions, and only 8.1% reported validating equality of variance. Only 4.9% of the articles assessed both distributional and homogeneity assumptions. Of the 79 articles in which MANCOVA was used, only 6.3% validated distributional assumptions, and none mentioned variance homogeneity. Forty-eight articles mentioned using ANCOVA, but of these, only 4.1% reported validating distributional assumptions, and 8.3% validated equality of variance. Similarly, in the 226 articles containing repeated measures designs, 15.5% of researchers mentioned validating distributional assumptions, and 0.4% reported on variance assumptions. Osborne (2013) suggested that an

overestimation of the robustness to violations of assumptions may be in part responsible for this tendency.

Sensitivity to Sample Size

Sensitivity to sample size is a widely recognized limitation of NHST (Boster, 2002; Cohen, 1994). Small sample sizes often lack sufficient power to detect even strong effects and fail to have p values smaller than the critical value (i.e., a Type II error is made). Alternatively, when a sample size is large, trivial effects can produce impressively small p values. Levine et al. (2008) provide some accessible examples of how this looks practically. For the relationship $r = .40$ when $n = 20$, a two-tailed significance test at $p = .05$ is not statistically significant; yet, $r = .07$, a dramatically weaker effect, is statistically significant if $n = 1000$. An even more stark example is if a two-tailed NHST is conducted using $\alpha = .05$ with an observed effect of exactly $r = .25$, the results are statistically significant when $n = 63$ but not when $n = 61$. P values do not indicate the magnitude of an effect but rather are a function of effect size and sample size and of other contributors of statistical power (directionality of the alternative hypothesis, experimental design, the test statistic, and measurement reliability; Kline, 2013), all of which have little to do with the credibility of one's predictions or theory (Meehl, 1986). As Ableson (1985) noted, the theoretical and practical value of a result relies more on the magnitude of the effect than on the likelihood of the data under the null hypothesis.

This sensitivity to sample size has several undesirable consequences. One is the paradox that with increased precision in the form of statistical power due to large samples, there is a greater possibility that effects that would be considered to have little practical or scientific value are statistically significant. Another consequence is the

possibility that researchers can generate significance by increasing sample sizes, a form of questionable practice called *p* hacking (Simmons et al., 2011). A similar problem is that the internet and related technologies allow for the collection of immense volumes of data. These massive data sets present a new challenge in that studies can be “overpowered.” Massive data sets also increase the number of analyses that can be conducted, and with relatively lenient thresholds, such as $p = 0.05$, make statistically significant results much easier to obtain and the risk of false positives much greater (Ioannidis, 2019).

The point or Nil-null

Another important criticism of NHST is that the null hypothesis is commonly defined as a nil hypothesis, meaning that the hypothesis is that some parameter is precisely zero or that some set of parameters is precisely equal. (Cohen, 1994; Meehl, 1978). In practice, however, the differences between means or the observed correlation of any variables, no matter how seemingly unrelated, will never be precisely zero. Randomization does not balance the effects of all extraneous variables, nor should it be expected that in correlation studies that there will be no uncontrolled variables (Meehl, 1978). The nil-null hypothesis is nearly always false, so rejecting it is neither impressive nor informative. It will always be rejected, given enough statistical power (Cohen, 1994). This is more concerning when considering the effect sample size has on power and the increasing availability of massive data sets. Given that the nil-null is always false and granted a large enough sample, NHST will always render a “statistically significant” result regardless of the value of the alternative hypothesis (Meehl 1986). Substantive Type I error rates are considerably increased with an increase in precision or power, and,

as a result, the rejection of the nil-null has little practical importance.

Power and Error Rates

Another criticism of NHST is that low statistical power has led to unacceptably high Type II error rates in research, especially in the social sciences (Boster, 2002; Hunter, 1997; Schmidt, 1996). Reviews of literature in the social and medical sciences have estimated that the statistical power of published research is between .4 and .6 (Cohen, 1962, 1988; Schmidt, 1997; Sedlmeier & Gigerenzer, 1992). However, some fields have much lower estimated ranges. The statistical power of published neuroscience literature has an estimated range of .08 to .31 (Button, 2013). Suppose studies have an average statistical power of .5. This implies that there is a 50% chance of correctly rejecting the null hypothesis. Even when an effect is present with a statistical power of .5, it will only be detected 50% of the time. As noted by Schmidt and Hunter (1997), an unbiased coin flip would be as effective and much cheaper. Low power has historically been a problem because of small sample sizes, and researchers have sought to increase power by increasing sample sizes (Schmidt, 1997). This is not a compelling solution. First, many types of research simply could not be done because of cost or other sampling factors if sample sizes adequate to attain sufficient statistical power were required. Second, large samples give more credence to the nil-null are always false criticism, resulting in just as serious a problem as low power and Type II errors.

Misunderstanding and Abuse

Given the ubiquity of NHST in the social sciences, one would hope that practitioners would possess a thorough and accurate understanding of the methodology. There is evidence, however, that this is not the case. There are many false beliefs

regarding NHST, most of which are misunderstandings or exaggerations of what can be inferred from the decision to reject or fail to reject the null hypothesis. Consideration will be limited to a discussion of what Kline (2013) identifies as the “Big Five” false beliefs about NHST. These are an extension of the work of Haller and Krauss (2002), who found that between 80% and 90% of psychology students and professors endorse at least one of the “Big Five” false beliefs.

The odds-against-chance fallacy is the false belief that the p value is the probability that the results were the consequence of sampling error. Put simply, if $p < .05$, then it is thought that there is less than a 5% chance that the results were obtained due to chance alone. An extension of this error is to sort results into two categories: those that are $p > .05$ and misconstrued as due to “chance” and those that are $p > .05$ and considered due to the presence of real effects. This is wrong because p values are calculated under the assumption that the null hypothesis is true and that sampling error is the cause of any deviation from the null hypothesis. Thus, the sampling error is always the cause of the p value rather than the p value being the probability of sampling error.

Local Type I error fallacy is the belief that when $p < .05$ given that $\alpha = .05$, then the probability of committing a Type I error is less than 5%. Over two-thirds of psychology professors and students have been found to hold to this fallacy (Haller & Krauss, 2002). According to Pollard (1993), this false belief stems in part from the confusion of the conditional probability of a Type I error with the conditional posterior probability of a Type I error given that the null hypothesis has been rejected. Further confusion is related to the nature of the p value, which is the conditional probability of the data given a true null hypothesis and does not apply to a specific binary decision to

reject or retain a null hypothesis.

Inverse probability fallacy is the belief that the p value measures the likelihood that the null hypothesis is true given the data, expressed as $\Pr(H_0|D)$. A researcher who interprets $p < .05$ as indicating that the null is true with a probability that is less than .05 has committed this error. This belief is an inversion of what the p value actually does. P -values give the conditional probability of the data given the null hypothesis, $\Pr(D|H_0)$ not the conditional probability of the hypothesis given the data. Haller and Krauss (2002) report that approximately one-third of psychology professors and students hold to this fallacy.

The validity fallacy is like the inverse probability fallacy but is instead related to the alternative hypothesis rather than the null hypothesis. This fallacy states that the probability that H_1 is true is $1-p$. Thus, if a p value of less than .05 is calculated, then it is believed that the probability that H_1 is true is greater than 0.95. This grossly overstates the information provided by the p value. The probability $1-p$ is the likelihood of getting a result less extreme than one attained if the null hypothesis is true. More than half of psychology professors and students have been found to endorse this fallacy (Haller & Krauss, 2002)

The replicability fallacy is the false belief that $1-p$ is the probability of finding the same statistically significant result in a replication of the study. If a p less than 0.05 were calculated, a researcher who holds to this fallacy would infer that the probability of replicating the result is greater than 0.95. Approximately half of the psychology professors and students endorsed this fallacy. This fallacy can have dire effects as it can lead researchers who subscribe to it to place unwarranted confidence in the conclusions

of single research studies. As a result, “significant” findings are often imbued with a sense of finality as if they have established a settled fact (De Long & Lang, 1992; Hubbard, 2016;). Hubbard (2016) speculates that this fallacy may contribute to the tendency of researchers to over-generalize the scope of their findings.

Publication Bias

Researchers in the social sciences have long suggested that there exists an editorial reviewer bias against the publication of so-called negative or null results (Bakan, 1966; Bakker et al., 2012; Hubbard & Amrstrong, 1997; Ioannidis & Trikalinos, 2007). Broadly, results that are inconsistent with predictions are not considered to be publishable, but under the NHST research paradigm, this means results that are not statistically significant because $p > .05$ are not publishable. Such a bias is alarming because it can seriously distort the content of empirical literature. This is known as the “file drawer problem” and can call into question the findings of meta-analyses that rely upon published literature. Evidence of editorial reviewer bias against null results is decades-long and consistent. Kerr, Tolliver, and Petree (1977) surveyed 429 advisory board members of 19 social science journals to gather common reasons for manuscript rejection or acceptance. They found that even when a manuscript was judged by editors and reviewers to be competent both in style and substance and relevant to the current interests of the field, findings related to $p > .05$ nonetheless markedly lowered the likelihood of publication. Atkinson, Furlong, and Wampold (1982) built on these findings by asking 101 consulting editors of two leading psychology journals to review three versions of an article that differed only in the level of statistical significance reported: not statistically significant, almost statistically significant, statistically significant. They

found that editors were three times as likely to reject the articles reporting no, or almost, statistically significant results as the one with statistically significant results. Epstein (2004) replicated this study in the field of social work, corroborating Atkins et al.'s findings. A more recent study by Fanelli (2010) involving all major social and medical science disciplines provides further support for the existence of publication bias. Fanelli randomly sampled 1,316 articles from publications across scientific fields then evaluated whether negative findings were reported. Only 17.6% of articles reported negative results. Also, Fanelli found that the researchers with a higher publication output reported fewer null results than less successful researchers.

There is evidence that bias against negative or non-statistically significant results influences how researchers respond to non-significant findings. There are at least two ways. One is that researchers can choose not to submit negative results for publication. Though the percentage of research that is not submitted for publication is difficult to estimate, there is evidence of this tendency among researchers. Greenwald (1975), using data from 36 articles submitted to a prominent psychology journal, estimated that the probability of researchers submitting papers for publication with null results was .06 compared with .59 for those with "significant" results. Coursol and Wagner (1986) used 609 responses to a survey of counseling psychologists asking questions about their publication habits. They found that 82% of articles with positive results were submitted for publication, while 43% of those with negative or neutral findings were submitted. In another study, a survey of authors who published articles in psychology journals found that 61% agreed that "research that is not statistically significant has little likelihood of being published" (Kupfersmid & Fiala, 1991). This tendency of researchers to avoid

submitting negative results for publication, though possibly a result of editorial reviewer bias, nonetheless works in tandem with such bias to misrepresent the volume of research in any given field (Ferguson & Heene, 2012).

The second way that editorial reviewer bias can affect researchers' behavior is to incentivize the search for "significant" results. Researchers may decide to persevere in perusing a study after attaining a negative result in the hope they might attain statistical significance. Greenwald (1975) asked a sample of psychology researchers whether they were likely to conduct an exact or modified replication of a study under certain conditions. Sixty-two percent of researchers said they would attempt a replication of the results that were not statistically significant, while 36% stated that they would replicate after a statistically significant finding. One reason this is problematic is that even under ideal circumstances, NHST with an α level of .05 will produce an average of a Type I error in 1 out of every 20 replications. Another is that when the goal is to attain statistically significant results to publish those results, the prior studies resulting in non-rejection of the null hypothesis would not be published, thus leaving holes in the research literature. The incentive to find "significant" results can also lead to other questionable behaviors such as data mining or *p* hacking. Data mining/*p* hacking is the selective reworking of data or the performance of multiple statistical analyses to either create or identify patterns in data that will yield statistically significant outcomes (Brodeur et al., 2012; Cumming, 2014; Gadbury, 2012; Simmons, 2011). Though the extent of *p* hacking is difficult to ascertain, there is some evidence that it is widespread. Head et al. (2015) used text mining on accessible articles in the PubMed database to create distributions of published *p* values called *p* curves (Simonsohn et al., 2011) and then investigated

whether values clustered around the “significance” threshold of .05. They reasoned that if this occurred, there was evidence of p hacking. They concluded that the practice is widespread across disciplines published in PubMed. Practices that are considered data mining/ p hacking include: performing multiple analyses while searching for one that yields statistically significant results; sub-setting data to search for statistically significant values (Simmons, 2014); conducting analyses during data collection to decide whether to continue collecting data (Gadbury & Allison, 2012; John, 2012); ceasing to collect data during an experiment once statistical significance is achieved (Bastardi, 2011); collecting data on multiple dependent variables then post-analysis choosing to report only those which yield statistical significance (John, 2012); deciding to exclude outliers post-analysis (John, 2012); and deciding to exclude covariates post-analysis (Simmons, 2011). The use of these spurious techniques can have dramatic effects. Simmons (2011) used simulations to investigate how using four different p hacking tactics (choosing among dependent variables, manipulating sample size, manipulating covariates, reporting subsets of data) affects Type I error rates. The results showed that when using any single technique, Type I errors when the nominal level is set at 5% are increased to between 7.7% and 12.6%. When combinations of these techniques are used, the Type I error rate can be as high as 60%.

Aggregated Statistics

Since the triumph of the “inference revolution,” data analysis of psychological processes has been predominantly conducted using methods that focus on the variation between subjects (Danziger, 1990; Molenaar, 2009). Inter-individual variation is used to derive aggregate statistics such as means or correlations that are calculated by pooling

across subjects. Standard statistical methods, whether they be cross-sectional or longitudinal, single or multi-level designs, all focus on the analysis of inter-individual variation. The results obtained from samples are (it is hoped) inferable, that is, generalizable, to a clearly defined population of which the sample is representative. It is important to note that it is the aggregate statistic found in the sample that is inferred to the aggregate population parameter.

It will be helpful to discuss and define generalization, or rather the types of generalization, more clearly, as this is the crux of the critique of aggregate statistics. Firestone (1993) differentiated several types of generalization that can be considered statistical and analytic. I will also discuss a third type, nomian. I will not discuss Firestone's case-to-case generalization because it is only applied to case studies using single-subject designs.

Statistical generalization is the inference of a sample statistic to a population parameter. It relies on sampling and probability theory. It requires random, representative sampling, among other assumptions that I previously discussed. Further, it deals in aggregates derived from inter-individual variation.

Analytic generalization occurs when researchers attempt to generalize from particular findings to a broader theory. When generalizing to a theory, researchers must make predictions based upon a theory and then attempt to confirm those predictions. This can differ radically from statistical generalization in that theories are not confined to specific populations but can be intended to apply across a range of populations and settings. This type of generalization does not necessarily exclude statistical generalization, but it does not rely upon it. Further, the two can be employed in concert.

For instance, if researchers discover no relationship between salary and job satisfaction for university professors who had no recent job offers but a strong relationship for those with recent job offers, they might first statistically generalize this correlation to the population of university professors. Concurrently, they could analytically generalize the findings and consider the results to be evidentiary support for cognitive dissonance theory (Firestone, 1993; Pfeffer & Lawler, 1980). This support, of course, would not be considered definitive proof but only evidence supporting the theory, as single studies generally provide weak support for a theory (Kiess & Bloomquist, 1985). Replications both under exact and different conditions would be needed to strengthen the claim.

There is a generalization that seeks to use research findings to identify laws defined as that which is common to all with “each individual a particular in which the general is manifest” (Bakan, 1966; Lamiell, 2019). I refer to this as *nomian generalization*, *nomoi* being Greek for “laws.” The laws of nomian generalization are common to all the investigated individuals and, at least provisionally, are generalized to all non-investigated individuals. This type of generalization is explicitly associated with traditional experimental psychology as used by its founders, such as Fechner, Ebbinghaus, and Wundt (Bakan, 1966). It is implicit in much of present psychology, as will be demonstrated by returning to the above example of research on university professors. The researchers located a strong relationship between recent job offers and job satisfaction. They statistically generalized this finding to the population of university professors and analytically generalized it as support for the theory of cognitive dissonance. The theory of cognitive dissonance is a theory of cognitive processes within individual minds; it is not a theory of group tendencies or of demography. Therefore, in

this example, the analytic generalization is also a form of nomian generalization, and for this to be appropriate, the strong relationship identified must be present in each individual. If it is not, then the use of analytic generalization from this particular study is improper. The study could still provide evidence of general tendencies of a population of university professors, even though it may not apply to individuals.

The use of results from statistical methods that rely on aggregated statistics to support theories about individual psychology is not uncommon (Grice, 2011; Lamiell, 2019). As Molenaar (2009) points out, it might seem self-evident that inferences about affairs at the population level represent findings that apply to each individual in the population. However, applying results obtained from pooling across individuals to each individual in the population requires a level shift from inter-individual variation to intra-individual variation. For this shift to be valid, specific conditions must be met. These conditions were determined as extensions of the classical ergodic theorems, which were derived in the early 1930s in ergodic theory research. This branch of mathematics was initially developed to address problems in statistical physics of when inter-individual variation can be applied to individuals.

The first condition of ergodicity is the homogeneity of the population, such that the same statistical model is valid for every individual within a population. The features of a given statistical model describing the data must be invariant across all individuals (Molenaar, 2009). As an example, consider, for instance, a one-factor model for an achievement assessment of several items. The single factor should explain the correlations between the item scores. Factor loadings are the regression weights that indicate the strength of the relationships of the observed variables. To meet the

homogeneity condition of ergodicity, the number of factors, as well as the factor loadings, must be invariant across all individuals. If this condition is not met, then even though the one-factor model can explain the correlations between item scores at the aggregated inter-individual level, it does not necessarily apply at the intra-individual level. Thus, for a given participant, a model of any number of factors might be a better fit than the one-factor model.

An empirical example of a non-ergodic factor model can be found in the work of Borkenau and Ostendorf (1998), who investigated the ergodicity of the Big Five personality factors (neuroticism, extroversion, agreeableness, conscientiousness, and intellect). The Big Five personality traits are well-established and widely used by researchers as well as by companies in the United States of America and Western Europe (Chen, 2018). Borkenau and Ostendorf conducted a replicated time series design experiment with 22 participants who were measured over 90 consecutive days. Participants were administered equivalent forms of a 30-item personality test on each day. They found that, as expected, the five factors did indeed explain the cross-sectional inter-individual correlations between the item scores. However, the five factors did not explain the correlations between the item scores of each individual participant. None of the intra-individual factor models had five factors, nor were the patterns of the factor loadings consistent across individuals (Borkenau & Ostendorf, 1998). Participants were found to have 2, 3, 6, and even 8 factor structures. The five-factor model simply did not hold for any of the individuals. The study has been replicated, yielding similar findings (Cervone, 2005; Cervone & Shoda, 1999; Epstein, 2010).

The second condition for ergodicity is stationarity and is achieved when a process, psychological or otherwise, has constant statistical characteristics over time that practically would manifest when statistical parameters of data are invariant across all time points. Any developmental process necessarily violates this condition. Many psychological processes tend to be developmental; cognitive information processing, emotion, and motor behavior are all pertinent examples that are non-ergodic (Molenaar, 2009). These processes are person-specific and thus differ from variables occurring in a population such as sex, socioeconomic status, or experimental condition.

The violation of either of the ergodic conditions is sufficient to render processes non-ergodic. Because of the incongruence of population-based inter-individual variation models with intra-individual variation in non-ergodic processes, it is crucial to avoid making claims about individual processes based on population-based findings (Lamiell, 2019; Molenaar, 2009). As Lamiell (2019) demonstrates, even a relatively high correlation of $r = 0.6$ reveals little about individuals and can have vast inconsistency at the individual level. Because of this, researchers must use analytic methods that allow for the investigation of individual variations if they wish to analytically generalize to theories that address non-ergodic processes or if they wish to extend their findings using nomian generalization.

Attempts at Reform

The criticism of NHST as a research methodology has existed for as long as the methodology has been around (Kline, 2013); however, this criticism did not stop its wide adoption and spread to become the dominant methodology in many social and medical sciences (Hubbard, 2019). The recognition of the shortcomings of NHST has, at times,

provoked attempts to limit or ban its use along with suggestions of alternative methods to use in its stead. One of the first such efforts was taken up by Kenneth J. Rothman, who, from 1984 to 1987, was the assistant editor of the *American Journal of Public Health*. Rothman, in revise and resubmit letters, advised authors to remove from their manuscripts all references to p values (Fidler et al. 2004). Articles published in AJPH that used NHST as the reported inferential methodology dropped from 63% in 1982 to 5% in 1986-1989 (Fidler et al., 2004). There was, however, evidence that in some articles that the interpretation was based on unreported use of NHST (Fidler et al., 2004). The use of confidence intervals increased from 10% to 54% for the same period (Fidler et al., 2004). Rothman later founded the journal *Epidemiology* where in 1998, he issued an outright prohibition on the use of significance tests (Rothman, 1998) that was generally well-received by researchers submitting to the journal (Fidler et al., 2004). The editor of *Memory and Cognition*, Geoffrey R. Loftus (1993), attempted to institute similar guidelines in 1993 that strongly discouraged the use of NHST and required the reporting of confidence intervals; there was no outright prohibition on NHST, but it was stated that such a ban would be considered if it was deemed appropriate. He received significant resistance, with researchers, at times, refusing to interpret or even calculate their own confidence intervals (Fidler et al., 2004; Kline, 2013). During this period, several prominent social science journals, including the *Journal of Experimental Education* (Thompson, 1993), *Psychological Science* (Shrout, 1997), and *Research in the Schools* (Kaufman, 1998), published special sections or issues discussing a possible ban on NHST.

The first major organized response to the criticism of NHST was the convening of

the Task Force on Statistical Inference (TFSI) in 1996 by the Board of Scientific Affairs of the American Psychological Association. The TFSI was tasked with considering the issues related to NHST and with offering suggestions for the then-upcoming fifth edition of the *APA Publication Manual*. The TFSI stopped short of a ban on NHST but did offer important suggestions to mitigate misuse (Wilkinson & The APA Task Force on Statistical Inference, 1999). Most of the recommendations were included in the fifth edition of the *APA Publication Manual*. These included the recommendations always to report adequate descriptive statistics, “almost always” report effect sizes, and a strong recommendation to report confidence intervals, but no requirement to do so. The sixth edition of the *Publication Manual*, published in 2010, uses similar language, preserving but not extending the recommendations. Similar recommendations were included in the *Uniform Requirements for Manuscripts of Medical Journals* (International Committee of Medical Journal Editors, 1997), which included recommendations against the use of NHST as the primary method of inference, the use of descriptive statistics, and the use of indicators of measurement error and uncertainty such as confidence intervals. A later version also called for journal editors to take seriously the publication of “non-significant results” (International Committee of Medical Journal Editors, 2010).

A series of recent developments represent a swell in the criticism of NHST and the reform-minded response to it. First, discussions of the problems with NHST methodology moved from little-read academic journals into popular publications and online blogs. *ScienceNews* (Siegfried 2010, 2014), *Nautilus* (Siegfried 2013), and *Nature* (Nuzzo, 2014), among others, all published articles highly critical of hypothesis testing and questioning the validity of findings based on the methodology. The article in *Nature*

is, to date, the most viewed online article published on the website (“*Scientific method: Statistical errors - Overview of attention for news story in Nature*”, 2021). Occurring concurrently was the increased awareness of the replication crisis afflicting many scientific fields, especially the social sciences (Wasserstein & Lazar, 2016). In 2011, the Center for Open Science at the University of Virginia began the Reproducibility Project: Psychology, a collaboration of 270 contributing researchers whose goal was to reproduce 100 experimental and correlational studies published in four major psychological journals. The results, published in 2015, were striking. Only 36.1% of replications yielded significant effects, whereas 97% of the original studies had significant effects. Further, the mean effect size of the replicated studies was approximately half that initially reported.

Reacting in part to the persistence and intensification of these concerns, David Trafimow, who assumed the editorship of *Basic and Applied Social Psychology* (BASP) in 2014, established new policies for article submissions which included no requirement of inferential statistical procedures, openness to publishing null effects, and openness to publishing research which contradicts previously published work (Trafimow, 2014). He states that this is because NHST “has been shown to be logically invalid and to provide little information about the actual likelihood of either the null or experimental hypothesis” (Trafimow, 2014). There was no absolute ban on NHST or other inferential methods, and though NHST procedure was clearly discouraged, authors were free to decide which, if any, inferential procedures to use. Authors always were required to provide complete descriptive statistics, including effect sizes (Trafimow, 2014). A subsequent editorial published one year later revealed that these changes were merely the

first step and were, in fact, a grace period (Trafimow, 2015). BASP editors banned the use of NHST, as well as confidence intervals, explaining that confidence intervals are subject to the same logical flaws, misunderstandings, and misuses as NHST (Trafimow, 2015). Bayesian procedures were also critiqued though not banned but were limited to some appropriate applications which would be considered on a case-by-case basis. No inferential statistical procedures were required for publication in BASP, but submissions must have included robust descriptive statistics and effect sizes. Large sample sizes were encouraged to reduce sampling error, but no sample size requirement was stated.

The editors of the *American Statistician*, a journal of the American Statistical Association (ASA), concerned with confusion and even doubt about the validity of the scientific endeavor arising from the twin issues of the problems surrounding NHST and the replication crisis, took the unprecedented step of issuing a statement on p values (Wasserstein & Lazar, 2016). The ASA had never previously issued a statement taking positions on specific matters of statistical practice (Wasserstein & Lazar, 2016). The ASA statement sought to clearly state several widely agreed upon principles related to the proper interpretation and use of p values. While it did not call for the cessation of NHST or p values, the statement did, however, contain several strong cautions against common misuses and misunderstandings of p values. These included:

- Avoid concluding whether an effect or association was found based only upon “bright line” dichotomous rules such as using a $p \leq .05$ threshold to determine “statistical significance.”
- Avoid concluding that an association or effect exists or is absent based only upon “statistical significance.”

- Do not believe that a p value provides the probability that a hypothesis is true or that a result was the product of chance alone.
- Do not conclude anything related to scientific or practical importance based only upon “statistical significance.”

These cautions were strong enough that the ASA editors feared that the statement would be viewed by researchers as asking them to “tear out the beams and struts holding up the edifice of modern scientific research” (Wasserstein & Lazar, 2019). Considering this, the ASA held a symposium in 2017 and issued a call for papers to suggest a path forward. A second statement was openly published along with 43 articles, each addressing some aspect of the crisis and offering suggestions of what can be done by researchers, editors, and policymakers. The statement advises researchers to become used to uncertainty, to be thoughtful in research design and selection of statistical methods, especially alternatives to p values, to be open and transparent, and to be modest, welcoming criticism and attempts to reproduce findings (Wasserstein & Lazar, 2019). While there is much agreement within the ASA statement and the attending papers on these suggestions, as well as the extent of the problem and the need to alter the existing research paradigm, there is no consensus about how this is to be done or what shape it will take. There is then a wide-open door to the development and use of alternative statistical methodologies that avoid the pitfalls of NHST and align with the principles outlined in the 2019 ASA statement.

Alternative Methodologies

ASA Recommendations

The 2019 ASA statement advises researchers to welcome uncertainty, be thoughtful, open and transparent, and modest regarding criticism and attempts to reproduce findings (Wasserstein & Lazar, 2019). Part of practicing thoughtful research, according to the authors of the statement, is considering multiple approaches for solving research problems or conducting analyses. The first step in accomplishing this goal, according to the ASA statement, is to abandon some practices researchers have come to rely upon regarding the use and interpretation of p values. Key among these is to leave aside, in nearly all situations, the dichotomization, or indeed any categorization, of p values around an arbitrary value and to forsake entirely the use of the language of “significance” and “non-significance” and the associated concepts of research worthy or unworthy of reporting based on “significance.” Without a decision rule centered upon some arbitrary critical value, p values are to be treated as continuous values and reported and interpreted as such. Further, the ASA statement cautions against treating confidence intervals in a dichotomized manner or as a substitute for a dichotomization of p values.

The special issue also includes specific methodological suggestions for supplementing or replacing p values drawn from several research papers included in the issue. The authors of these papers agree that the conventional methods of using p values in the NHST paradigm need reform, but they differ as to the degree and kind of reform that is needed. Indeed, several authors do not entirely agree with the suggestions of the 2019 ASA statement. Some contributors (Benjamin & Berger, 2019; Matthews, 2019) are convinced that because of the ubiquity and intransigence of NHST that the best approach

is an incremental approach. They expect researchers to continue doing what they have been doing heretofore: conducting hypothesis tests, using a dichotomized decision rule, using the language of significance., etc. The goal of this incremental approach is to offer “easy but impactful changes” that might realistically be adopted within the NHST paradigm instead of offering new or complicated alternatives. Others (e.g., Amrhein et al., 2019; Hurlbert et al., 2019; McShane et al., 2019) prefer a sterner approach (e.g., prohibiting the use of dichotomization, significance categories, and even statistical inference to parameters), which they deem appropriate given the severity of the problem. Even here, the authors do not suggest banning p values but instead allowing them to be used descriptively as continuous values, which are thoroughly explained and interpreted. Each suggestion offers some improvement over the prevailing practice with a focus primarily on increasing the stringency of the criteria for how inferences are made and lessening the tendency to misinterpret p values. This can take the form of composite null hypotheses, lower critical values, the use of Bayesian priors, or other methods of estimating practical importance. What all of these have in common is that they do not address other important foundational problems with the NHST paradigm, such as violation of assumptions, the treatment of non-continuous data as if it were continuous, and the problem of aggregation. The ASA statement does, however, say that researchers should use statistical techniques that are appropriate to their data and that going forward, different disciplines may take different routes in a post p value research environment. What follows is a series of brief descriptions of the methodological suggestions from the ASA special issue.

Interpretational Aids

Methods categorized as interpretational aids assume that researchers will not quickly or easily be persuaded away from NHST or the use of p values as a summarizing statistic. There is no expectation that researchers will radically alter much of what they do when using NHST. The hope, however, is that the long-standing confusion about the interpretation of p values can be ameliorated by transforming them into more easily understood statistics. Further, there is a recognition of the difficulties researchers who are accustomed to a dichotomous decision rule will face while attempting to interpret a p value. These methods are proposed with the intent to help researchers better understand and interpret the strength of evidence provided by p values.

Bayes Factor Methods

Three different papers in the 2019 *American Statistician* special issue suggest supplementing NHST with Bayesian statistics. Benjamin and Berger suggest using a Bayes factor bound (BFB) that gives the maximum odds in favor of the alternative hypothesis relative to the null hypothesis assuming each hypothesis is equally likely. Colquhoun (2019) recommends calculating the False Positive Risk statistic, which is interpreted as the probability that the obtained results occurred by chance. Matthews (2019) suggests the Analysis of Credibility statistic, which uses the Bayes theorem to derive a posterior distribution for the claimed effect. Each of these has at heart a similar calculation using Bayes theorem with the p value and addressing in slightly nuanced ways the strength of evidence. Because of this similarity, I will discuss one of these in more detail to exemplify the use of Bayes-based approaches. All three methods are easy

to interpret, but the Bayes factor bound is far simpler to calculate than the other two options and, because of its simplicity, may be more likely to be adopted by researchers.

The Bayes factor bound is a function of the p -value, using the natural logarithm and its base, the constant e .

$$BF \leq BFB = \frac{1}{-e p \log p}$$

The BFB is interpreted as the maximum odds of observing the data under the alternative hypothesis to observing it under the null hypothesis. Calculating the BFB does not require specifying a particular alternative hypothesis because it is an upper bound across a large class of reasonable alternative hypotheses. The value of the BFB is that it transforms the p value into a statistic that clarifies the amount of evidence against the H_0 provided by the p value.

For example, the data-based odds in favor of the H_A for $p = .05$ is 2.45 to 1, meaning that the data is 2.45 times more likely under that H_A than under the H_0 . The BFB can also be transformed into a posterior probability using equation 2, assuming the prior probabilities of H_0 to H_A are equal.

$$\Pr (H_A | p) = BFB/(1 + BFB)$$

For $p = .05$, the posterior probability of the H_A is 0.71, meaning that there is a 71% chance that the alternative hypothesis is true.

Shannon Information Values

Greenland (2019) recommends transforming p values into Shannon information values known as s values, a statistic drawn from information theory. The transformation is accomplished by taking the negative log of the p value $-\log_2(p)$. By using s values, Greenland hopes to present a more intuitive statistic than p values. Greenland points to

two areas where s values are an improvement. P values are directionally and conceptually mismatched; that is, as the value of p decreases, the evidence against the null hypothesis increases. Also, p values are not scaled so that the difference between values is consistent. For example, the difference between 0.01 and 0.10 is not the same as the difference between 0.90 and 0.99. The s value, by contrast, is a continuous measure of bits of information against the test hypothesis. An example is that a p value of 0.05 yields an s value of 4.3, indicating that there are 4.3 bits of binary digits of information against the test hypothesis. Each bit as a binary digit has only two outcomes; therefore, it is easy to think of them in familiar terms, such as a coinflip. According to Greenland (2019), if we round to the nearest integer, then four bits of information should be interpreted as no more surprising than getting all heads in four fair coin tosses.

Decision Criteria Methods

Alternative decision criteria methodologies posit different and generally more demanding criteria for making an inference within the NHST paradigm. These are put forward as a response to the problems of replication and the divergence between statistical and practical significance. The methods rest upon the assumption that since researchers are likely to continue to use hypothesis tests with decision rules, different and more difficult decision rules will mitigate some of the problems with $p \leq .05$ dichotomization.

Decrease Decision Threshold to 0.005

A methodological suggestion to reduce the decision threshold for inference from 0.05 to 0.005 for new discoveries has been suggested several times in the decade between 2010 and 2020 (Greenwald et al. 1996; Johnson 2013). Most forcefully was in 2018, in a

comment article published in the journal *Nature Human Behavior* that listed seventy-two authors and co-authors (Benjamin et al., 2018). Benjamin and Berger, who are the first and second authors, repeated the suggestion in the 2019 ASA supplemental issue on p values (Benjamin & Berger, 2019). The authors make the case that the conventional NHST research tradition will not change easily or quickly. As a result, they put forward several modest suggestions for improving statistical practice when using the NHST framework. It is noteworthy that the authors numbered their recommendations in a fractional form (0.1, 0.2, and 0.3) to indicate that each is only a small step away from current practice (Benjamin & Berger, 2019). The change of the significance threshold is numbered 0.1.

Benjamin and Berger's suggestion is to require a significance threshold of 0.005 for novel research. P values less than or equal to 0.005 may be called "significant," and those between 0.05 and 0.005 may be referred to as "suggestive." Less stringent significance thresholds can and should be used for confirmation or replication research, or as is the case in genetics or physics, much more stringent thresholds may be appropriate. The primary concern is to begin to address the strength of evidence question. The authors consider the replication crises to be the result of a high number of Type I errors resulting from inflated claims of findings based on the 0.05 significance threshold. Despite the fact that the authors agree that the use of dichotomous decision rules is inappropriate, Benjamin and Berger nonetheless consider a change in the threshold for significance from 0.05 to 0.005 to be a marked improvement. To support the selection of 0.005, Benjamin and Berger use Bayesian statistical methods to calculate the probability of a true H_0 at a given p value. Assuming that the null hypothesis and the alternative

hypothesis are equally likely (prior odds), for a $p = 0.05$, there is at least a 29% chance that the null hypothesis is true. This reduces to a 7% chance when using a threshold of 0.005. They also present evidence of potentially dramatic decreases in estimated false positive rates. For example, the false positive rate is greater than 33% with prior odds of 1:5 and a p value threshold of 0.05, regardless of the level of statistical power. Reducing the threshold to 0.005 would reduce the minimum false positive rate to 5% (Benjamin et al., 2018).

A concern related to transitioning to a 0.005 significance threshold is the relationship between statistical power and sample size. For most statistical tests to maintain statistical power of 80% with an $\alpha = 0.005$, sample sizes would have to increase by approximately 70% (Benjamin et al., 2018). Such an increase would mean that many researchers would likely find it impossible to conduct their studies using current experimental designs and funding. The proponents of the 0.005 threshold consider this to be an acceptable cost of improving the quality of research. Further, they suggest that such seemingly negative effects could lead to improvements in research as the resources which would have been spent on studies based on false premises and yielding questionable results could now be allocated to research featuring better designs and analyses.

Roaming Alpha

Another proposal somewhat related to the reduction of the significance level is to abandon using a fixed alpha and allow it to vary depending on a predetermined set of criteria. Gannon et al. (2019) propose a new hypothesis testing procedure that combines frequentist and Bayesian tools to provide a significance level that is a function of sample size. This is obtained from a generalized form of the Neyman–Pearson Lemma that

minimizes a linear combination of α , the probability of rejecting a true null hypothesis, and β , the probability of failing to reject a false null, instead of fixing α and minimizing β (Gannon et al., 2019). The calculations of this procedure are complex, which may limit accessibility to researchers lacking extensive statistical training. A similar but more accessible and exclusively frequentist technique was proposed by Mudge et al. (2012). Here an α is selected to minimize the combination of Type I and Type II error given an *a priori* critical effect size. The calculations involved are similar to power analysis, familiar to most researchers.

Second Generation p value

Blume et al. (2019) propose replacing *p* values with a second-generation *p* value (SGPV) that incorporates measures of practical significance into its computation. The SGPV is calculated from an interval null hypothesis that represents a range of differences that would be practically or scientifically inconsequential. Researchers would construct the interval null hypothesis by specifying in advance a range of effects that they consider to be without practical or scientific importance. Blume reasons that this would help to eliminate the conflict between scientific and statistical significance. This is an application of the “good enough principle” proposed by Serlin and Lapsley (1985). Blume’s innovation is in devising a statistic that summarizes the test results in light of the “good enough belt” proposed by Serlin and Lapsley. Like *p* values, SGPV values range from 0 to 1. An SGPV of 0 indicates that the data are incompatible with any of the null hypotheses or, stated alternately, that the data only support meaningful effects. An SGPV of 1 indicates that the data only support the null hypotheses or trivially null effects. An SGPV between 0 and 1 is considered inconclusive to varying degrees, an SGPV of 0.5

being the most inconclusive. In addition to the SGPV, Blume suggests reporting detailed descriptions of the findings in the form of an interval estimate of effect size and noting its proximity to the composite null hypothesis. Blume argues that the SGPV provides a high-level descriptive summary of the data that is straightforward and, since it mimics p values, is familiar to researchers.

Minimum Effect Plus p value

Goodman et al. (2019) propose using a hybrid criterion of minimum effect size plus p value (MESP). In this scheme, researchers are tasked with selecting a minimum practically significant difference (MPSD) between the parameter and the null hypothesis prior to conducting research. It is assumed that researchers would have sufficient knowledge of past findings and competence in their field of expertise to be able to select a scientifically or practically meaningful MPSD and that it would be selected in good faith. The MPSD is then paired with a conventional p value methodology to construct a hybrid decision rule. The MESP method indicates that to reject a null hypothesis, each of two conditions must be satisfied: (1) $p \leq \alpha$ and (2) the observed effect size (i.e., the absolute value of the sample mean – null mean) \geq MPSD. A benefit mentioned by Goodman is that the MESP method is practicable without complicated calculations using existing statistical software. In a simulation study, Goodman found that the MESP method maintained power roughly equivalent to NHST with $p < 0.05$, except with high nominal power where MESP power is lower. Goodman defines power uniquely by first constructing an interval null or “thick null” and calculating the proportion of times a method correctly identifies the presence or absence of the true effect within the interval

null. An additional benefit mentioned by Goodman is that the MESP method is practicable without complicated calculations using existing statistical software.

Deemphasize or Abandon Significance Testing

Methodologists who suggest moving away from statistical testing consider such tests to be philosophically flawed or practically misused to such a degree that much more will be gained in losing or limiting them than by persisting in their use. Common to each of the recommendations presented below is a suggestion to make no claims about inferences to hypothetical parameters. Such inference is to be avoided altogether. Instead, these suggestions focus on reporting the obtained results, their estimates, and looking to replication as the prime means of inference.

Prediction of Future Observables

Billheimer (2019) suggests abandoning inferences about parameters and instead focusing on the prediction of future observables and their associated uncertainty. He argues that the importance of any treatment or explanatory group effect is realized only through the distribution of future observables. The primary purpose of statistical inference in this understanding is to predict realizable values that have not yet been observed based on values that were observed. Billheimer argues that this change in inferential focus will improve scientific accuracy and reproducibility by shifting the focus from “finding differences” among hypothetical parameters to predicting observable events based on current scientific understanding. The predictive distribution is estimated using Bayesian methods relying on the Finetti representation theorem. The calculations are relatively complex, but the interpretation is intuitive. For example, with a two-group mean difference experiment, the predictive distribution can be calculated from the

obtained sample, which provides the probability of a mean difference, either positive or negative, given a certain sample size. The predictive probability of effect sizes is also possible using a Monte Carlo method of sampling the predictive distributions at a given sample size.

Interval Estimates

Amrhein et al. (2019) advocate abandoning the language of significance and stopping dichotomization of both p values and interval estimates. They also advocate treating all inferential statistics as very unstable local descriptions of relationships between models and obtained data rather than as providing generalizable inferences, keeping in mind that statistics such as p values are tests not just of a single hypothesis but of all the assumptions. Given all this and the authors' shared opinion that researchers' long-standing habits of thought regarding NHST and p values are unlikely to change quickly, Amrhein makes a case for the benefits of at least a temporary ban on statistical tests. The authors further advocate for interpreting confidence intervals not as a range of values that contain the true parameter at a selected level of confidence but as "compatibility intervals showing the effect sizes most compatible with the data according to their p values under the model used to compute the interval" (Amrhein et al., 2019, p. 265). Extending from the conceptual to the more concrete, Amrhein has several specific methodological recommendations. Estimates, not tests, should be emphasized and interpreted, explicitly discussing both the lower and upper bounds of interval estimates. The precise values of statistics should be reported, not inequalities. For example, researchers should report the exact p value such as " $p = 0.02$ " not " $p < 0.05$ ". Language of "significance" or "confidence" should not be used to describe results. Lastly, there

should be an acknowledgment that statistical results describe relations among assumptions and data in a given study and that scientific generalization is unwarranted.

Building and Testing Models

Amrhein's (2019) conceptualization of the H_0 and the various assumptions associated with it as one model of many possible models of the obtained data has important implications beyond the explicit suggestions related to interval estimation. In fact, such a conceptualization is an example of what Rogers (2010) considers a quiet revolution that has been taking place concurrent with the NHST controversy from the 1990s. This change in thinking has involved a transition from the Fisher/Neyman-Person NHST paradigm to one based on building and comparing statistical/mathematical models. The key distinction is that the philosophical basis of NHST is the H_0 where evidence of the research hypothesis, embedded within the H_A , depends upon its rejection. However, from a model-building perspective, the researcher's hypothesis is explicitly modeled and compared with other competing models, which may or may not include a null model as a reasonable competitor. The epistemological focal point is shifted in the model-building perspective from the H_0 , which in NHST is assumed until it can be rejected in favor of a broad and relatively ill-defined alternative to the current research model. Tong (2019) suggests that most research is exploratory in nature and should be used for the purpose of identifying patterns in data that can be used in model construction and refinement. Inference then is not applicable to most research and is reserved for study protocols and statistical models that are fully prespecified (Gelman, 2016; Tong, 2019).

Mathematical models generally have two characteristics: a model matches the reality it describes in some important ways and is simpler than that reality. Models are

judged based on their fit to reality, operationalized as the obtained data, with simpler models preferred when models fit the data approximately well. Fit statistics such as chi-square, the root mean square error of approximation, and the Akaike information criterion (AIC), among others, are used to compare model fit.

If this perspective is already present, one may wonder why then NHST is still a problem. The answer is that there is a great deal of inertia associated with NHST, as discussed previously. The concepts associated with the modeling approach are still spreading to applied researchers. Further adoption of a modeling methodology is hindered by the modeling perspective receiving little attention in most statistics or methods instruction.

Other Options not Included in the ASA Issue

The suggestions published in the *American Statistician* supplemental issue were not intended to be exhaustive of the options available to researchers. Rather, the editors were addressing real concerns about the 2016 statement (Wasserstein & Lazar, 2016, 2019) related to criticism that more than a scolding about the limits and misuse of p-values was needed. In the 2019 issue, ideas of a way forward were offered, an attempt to be yea saying as well as nay saying. However, the breadth of the discussion, from methodological and interpretational to suggestions for publishing and institutional reform, necessarily required limits to the discussion. Important methodologies and alternatives to NHST were lightly mentioned or left out entirely. What follows is a brief treatment of some of the alternatives not mentioned in the *American Statistician* articles. The methodologies discussed were selected to reflect the variety of options available to

researchers, with each method moving further away philosophically from the NHST framework.

Nonparametric Methods

Parametric and nonparametric methods are two broad classifications of statistical methodologies. Parametric procedures assume that sample data are drawn from a population having a probability distribution with a defined shape and a fixed set of parameters. Most well-known and regularly used statistical methods are part of the parametric family. NHST commonly uses methods within the parametric family of methods, though this need not be the case. It is the use of NHST in conjunction with parametric methods that necessitate the assumptions discussed previously, especially the normality assumption and sample size requirements.

Nonparametric methods do not assume a fixed set of parameters and use relaxed assumptions about the shape of the probability distribution of the population. Many nonparametric methods do not make any assumptions about distribution shape. Because of this, nonparametric procedures are sometimes referred to as “distribution-free.” This does not imply that nonparametric methods are assumption-free; they do indeed have assumptions, though there are often fewer assumptions than required by parametric procedures. Typical assumptions include random sampling from a population, independence or dependence of samples, and the symmetry of probability distributions. These are not required for all nonparametric procedures, and some procedures have fewer or more assumptions. In short, the assumptions are generally fewer in number and less challenging to meet than those of parametric procedures. Further, nearly all the most frequently used parametric procedures have nonparametric analogues: Wilcoxon rank-

sum test instead of a two-sample t-test, Wilcoxon signed-rank test instead of a paired t-test, Kruskal-Wallis instead of a one-way ANOVA, Friedman test instead of repeated measures ANOVA.

Nonparametric procedures that are analogous to standard parametric procedures have not been widely used by social science researchers, even though they have fewer assumptions that are often easier to justify. The reasons for this likely lie, at least in part, in two perceived drawbacks of nonparametric methods. First, these methods are generally less statistically powerful than their analogous parametric procedures when parametric assumptions are met. It is important to note, however, that nonparametric procedures can be more powerful than parametric procedures when those assumptions are not met, which, as was discussed previously, happens frequently. Secondly, results of nonparametric tests are often thought to be more difficult for researchers to interpret because nonparametric tests often use transformations of values, such as ranks of scores, in the data instead of the actual data. For example, interpreting that the difference between the mean ranks of two groups is three is less intuitive than interpreting that the mean difference in test scores on a one hundred item assessment is 10.

The classes of nonparametric procedures extend beyond methods that are analogous to parametric procedures. For instance, there are methods for constructing sampling distributions from the observed data. Randomization or permutation tests build, rather than assume, the sampling distribution by “permuting” the observed data. More precisely, the observed data are shuffled by assigning different outcome values to each observation from among the set of observed outcomes. In randomization tests, this is done without replacement. A similar procedure called bootstrapping includes the

replacement of assigned values. The recombination process allows for the creation of new data sets that can represent all of the possible alternative treatment assignments, and the test statistic in the observed data can be assessed based on where it falls relative to all of the possible outcomes. Thus, the likelihood of obtaining the observed test statistic from the sample can be ascertained. While a complete randomization test requires the calculation of all possible combinations of the data (which can become quite large), an approximate test can easily be conducted by simply drawing a very large number of resamples.

The inferences drawn from randomization tests are different from those of parametric methods. Randomization tests do not lead to inferences from a sample to a population. In fact, randomization tests do not have any concern with populations at all; therefore, there is no need to define a hypothetical population or to estimate its characteristics. It follows then that there is no assumption of random sampling from a population. Instead, randomization tests allow one to know how likely an outcome is from all possible combinations of the actual observed data. Randomization tests are also free from sample size requirements, so free that they can be employed for single-subject designs. The only required assumption for group or single-subject designs is the randomization to conditions. Randomization tests work on the logical principle that if cases were randomly assigned to treatments, and if treatments have absolutely no effect on scores, then a particular score is just as likely to have appeared under one condition than under any other. Randomization to conditions then is extraordinarily important in randomization tests; not that it is not crucial in other tests of group difference, but in randomization tests, the emphasis is increased.

Observation Oriented Modeling

Observation Orientated Modeling (OOM) is a methodological approach developed by James Grice consisting of a collection of analytic methods explicitly based on a realist philosophical perspective (Grice, 2011). OOM analyses are accessible through a freely available software package. The approach is consistent with the ASA framework in as much as it eschews the language of significance and dichotomized decision rules. Indeed, the OOM approach has much in common with the suggestions of Amrhein et al., who suggest it may be necessary to abandon hypothesis tests and inferences to population parameters entirely. OOM also attempts to address other criticisms of NHST, which the ASA statement and attending methods do not, such as that of aggregation. Grice asserts that the methodology and metaphysics of the prevailing research tradition have retarded the accumulation of genuinely scientific knowledge. OOM is advanced as a collection of alternative analytical methods based upon different philosophical assumptions that place the analytical focus on identifying patterns of individual observations and developing and testing hypotheses about the psychological processes that give rise to those patterns. This is in contrast with the variable and parameter orientated approach of psychology's prevailing research paradigm. This contrast is an extension of the fundamental differences in the philosophies undergirding the two approaches, a detailed treatment of which is beyond the scope of this proposal; however, Arocha (2020) provides a useful presentation of the differences between scientific realism and the empiricism and operationalism of the present paradigm. From this basis, Grice posits several principles of research whereby researchers should focus upon phenomena that are real, accurate, repeated, and observable and that researchers

should use integrated causal models that require deep consideration of causation which, following moderate realism, adhere within individuals not variables, incorporate statistical outliers, and avoid aggregation (Grice, 2011). As Grice (2011) explains:

observation oriented modeling shifts the focus of analysis away from computed aggregates such as means and variances onto the observations themselves. In other words, the focus is shifted to the people, specific behaviors, animals, things, and events under investigation. The psychologist instead worries less about fulfilling untenable assumptions... and thinks more about the patterns of ordered observations relative to a competing perspective of chance (p.40).

Observation Oriented Modeling analytical procedures all function to build and test theoretical models that predict patterns of behaviors of individuals. Most OOM analyses require researchers to predict patterns a priori based on theoretical models that are then analyzed using the data to determine the accuracy of the prediction and relative rarity of the outcomes. Predictions can be as precise as theoretical models allow, but all must predict individuals' behavior, not aggregations. For example, consider an experiment on recalling images using two experimental conditions based on the types of sound (white noise, silence) to which participants were exposed during recall. Using OOM software, a researcher could choose a precise pattern based on theory that predicts exactly the number of images recalled by individuals in each group. The example in Figure 2.1 shows an expected pattern for the white noise and silence conditions. The prediction indicates that individuals in the white noise condition will have counts between 15 and 18, while those in the silence condition will be between 21 and 25. The pattern need not be a range, could have overlap between the two conditions, or have gaps

in a single condition. The count along the left-hand side is the actual range of counts in the observed data.

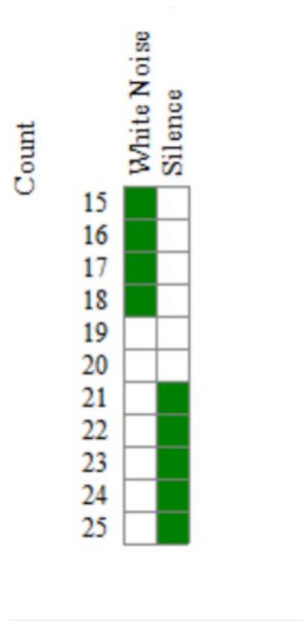


Figure 2.1 Image Recall Experiment Pattern Sample

An OOM procedure called a pattern analysis would then be conducted by calculating the percentage of observations that match the predicted pattern, called the Percent Correct Classifications (PCC). The PCC is not an aggregated statistic such as a mean or variance; instead, it is a percentage of individual observations in a sample with responses that match the expected pattern. In the OOM framework, the PCC serves the same function as an effect size in conventional statistical analyses. A PCC can range from 0 to 1, where 1 would indicate that all observations are accounted for by the predicted pattern. If a theory is not refined enough to predict precise patterns, it usually will be able to indicate the expected ordinal relation of pairs of individuals from each group. For instance, in the example above, a researcher could expect that individuals in

the white noise condition will recall fewer images than those in the silence condition.

Figure 2.2 shows how this pattern is indicated in the OOM software.



Figure 2.2 Recall Experiment Ordinal Pattern Sample

Ordinal pattern analysis is conducted to test the predicted pattern, whereby each individual in the white noise condition is paired with every individual in the silence condition. The ordinal relationship of pairs is evaluated, and a PCC is calculated. Further evaluation of patterns is possible using a post hoc analysis, which uses Binary Procrustes Rotation (BPR), a modified form of Procrustes rotation first proposed by Green (1952). At the ground of BPR are what Grice calls the deep structures of data. Deep structures are represented as matrices of zeros and ones that can be manipulated to identify patterns in the data. These structures are obtained by translating data into binary code similar to dummy or effect coding. For example, the deep structure of biological sex is represented using two columns, one corresponding to male and the other to female, with each row corresponding to an observation. Each column has a binary coding of “1” if positive or “0” if negative. The coding for the biological sex of an individual who is male, if the column order is male and female, would be “1 0”, a female observation would be “0 1”. Note that sex is not represented in a single column with “1” for one sex and “0” for the other. A 4-point Likert style scale with categories of strongly disagree,

disagree, agree, and strongly agree would be similarly coded: a strongly disagree response would be coded as “1 0 0 0” whereas a disagree response would be coded “0 1 0 0”. Example matrices for sex and Likert-style responses are displayed below.

$$\text{Sex} = {}_6\mathbf{Y}_2 \begin{vmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{vmatrix} \quad \text{Likert Scale} = {}_6\mathbf{X}_4 \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{vmatrix}$$

Figure 2.3 Deep Structure Matrices for Sex and a Four Point Likert Style Scale

The deep structure matrices must be appropriately arranged to perform BPR: observed categories, whether people, animals, or behaviors, are assigned to the rows of matrices, and the units of observations, such as sex or level of agreement, are assigned to the columns. Further, a target matrix must be selected. For instance, in the above sample, the sex matrix would be the target matrix indicating that male and female responses should be different. In an experimental design, the matrix containing the control and experimental groups would be the target matrix. Other more complex target matrices are possible depending upon the nature of the data and the expectations of outcomes. Pairs of deep structure matrices are rotated to agreement using BPR transformations. Binary Procrustes Rotation seeks to align the columns so that the co-occurrences of 1's in the two deep structure matrices are maximized (Grice, 2011). A normalization process is used to ensure that the rotated matrix values do not exceed 1 and to minimize the effects of large discrepancies in the numbers of observations for different units of observations.

Finally, a PCC is calculated, which indicates the percent of individual observations that matched the target matrix.

To determine what value of PCC might be considered a success, the OOM methodology uses a resampling procedure that randomly reshuffles the data and recalculates the PCC value. The process is repeated a selected number of times (Grice recommends 500 to 1000), and a probability statistic is calculated called the chance value or c-value. The chance value is the frequency with which the resampling yields results at least as accurate as the initial data. This provides a measure of the rarity of the outcome achieved, given the data. If the randomized data sets fit the selected pattern as well or better than the actual data, then a high c-value (closer to 1) will result, indicating that the observed outcomes are common. Conversely, a low c-value (closer to 0) results when the randomized data set does not fit the selected pattern as well as the actual data, indicating that the pattern of observations is unlikely. It should be noted that the sample size does have an effect on the usefulness of the c-value statistic. Small data sets allow for fewer random permutations from which comparisons may be obtained. As a result, the c-value is less informative as to the probability of the obtained outcome. Inferential caution is needed and should be exercised regardless of the size of data sets. Further, cut-off values should not be used, or the result would be a similar error as that of NHST (Grice, 2017).

Single-case Designs

Single-case designs (SCD), also known as single-subject, small n , and $N=1$ designs, are a type of research methodology characterized by the repeated assessment of a particular phenomenon exhibited by a single “case.” A case may refer to a particular person, a social unit (e.g., family, support group), or an organizational body. Further, the

designation “single case” should not be construed to mean that only a single subject participates in a given study. Several subjects may participate in a single case study; however, data will not be aggregated or combined. Instead, each subject’s data are collected and analyzed individually and may be viewed as a replication of the study. Single-case designs differ from group designs in that all data collected and analyses conducted are done at the level of the individual unit, not at the group level. Foundationally, the basic assumption of single-case design research is that behavior takes place at the level of the individual case and that the unit of analysis is properly the individual case. Single-case research thus avoids the problems associated with aggregation and the assumptions needed for inference to population parameters previously discussed.

There are limitations to the single-case design model. Researchers have traditionally used visual analysis as the primary method of analyzing single-case design data. Visually analyzing graphic representations can be subject to disagreement or error. Ottenbacher (1993) conducted a meta-analysis that revealed a mean of .58 (range: .39-.84) interrater agreement. Ninci (2015) conducted a similar analysis and found a higher inter-rater agreement of .76. Further, there are inferential concerns about the ability to generalize from single-case research to other contexts. Replication can obviously help, but there has also been work done to develop standardized effect sizes that could allow for the results of different studies on similar questions to be combined or compared in a meta-analysis.

Standardized effect sizes were developed to address these types of concerns with experimental group methods. Standardized effect sizes calculate the magnitude of a

treatment effect and place it on a standard scale, which allows for the results of different studies on similar questions to be combined or compared. This will enable researchers to interpret the results from SCDs using the same conventions as those used with other experimental designs: calculation of effect sizes, confidence intervals, and significance tests. It also allows for the combination and comparison of results from different SCDs, such as those investigating the same intervention on the same behavior but using different measurement systems (e.g., percent or count) and other experimental designs (Hedges et al., 2012, 2013; Pustejovsky et al., 2014).

Three standardized effect sizes have recently been developed for use with SCDs. These allow for comparisons between cases, that is, between different designs or studies. Hedges et al. (2012, 2013) created an effect size with two slightly different calculations: one for multiple baseline designs and one for reversal designs. Hedges effect size accounts for serial dependence of the data and allows for power analysis. The effect size assumes at least three cases, no trend, that the effect is constant across cases, and that the phase mean's residuals are normally distributed. Pustejovsky et al. (2014) further developed the work of Hedges et al. This effect size has all of the assumptions of Hedges effect size, including normality of residuals, but it allows for trend, treatment effects, and the interaction of the two to vary across cases. However, it does not allow for power analysis, and to calculate the additional effects, a sample size of at least 12 is required. Lastly, Swaminathan et al. (2014) created a Bayesian-based effect size estimator that also allows for trend and interaction effects but has the benefit of more stable parameter estimation. The Swaminathan effect size estimator assumes normally distributed residuals and requires at least three cases.

Conclusion

The NHST research paradigm in the social sciences developed and was popularized in part as a means to gain the creditability and respect associated with being a “true” scientific discipline such as the physical sciences. Group experimental designs, quantitative measures, and complex statistical procedures were thought to provide the necessary rigor and objectivity to achieve this goal. However, the limitations of the paradigm and misuse and misunderstanding on the part of some researchers resulted in a replication crisis and an undermining of the credibility of social science research. These developments might seem to be a cause for despair, but in truth, they offer a great opportunity for social science researchers. Following the ASA statement's advice, researchers should increasingly feel free to experiment with alternative methods, many of which are detailed in this article. This experimentation can and should draw from Bayesian, nonparametric, non-inferential, and other procedures appropriate to the research design and questions the researcher hopes to address. This freedom also comes with the burden of thinking more deeply and differently about design and analysis than researchers steeped in the NHST paradigm might be accustomed.

CHAPTER 3

AN INTRODUCTION TO AND A DEMONSTRATION OF SELECT PROPOSED ALTERNATIVES TO NULL HYPOTHESIS STATISTICAL TESTING²

² B.D. Rogers. Not yet submitted for publication.

Abstract

Considered normative from the second half of the 20th century (Danziger, 1990), null hypothesis statistical testing (NHST) has received consistent, largely unheeded criticism. Critiques have received more attention in recent years with the recognition of the replication crisis in the social sciences and the American Statistical Society's statement and special issue on p values (Wasserstein & Lazar, 2016; Wasserstein et al., 2019). This paper explores seven viable alternative methodologies and supplements to NHST by thoroughly explaining the methods and demonstrating them with the analysis of simulated data and looking for strengths and weaknesses, and how each method aims to improve upon NHST. The findings suggest that no single method is a panacea, but rather that multiple methods provide specific benefits over NHST. The use of different methods in different experimental contexts and the combination of methods are suggested.

Null hypothesis statistical testing has been the dominant research methodology in social science since the second half of the 20th century. Criticism of NHST is long-standing, existing nearly as long as the methodology (Kline, 2013). However, over the last decade, two concurrent phenomena have contributed to a broader awareness of the problems with NHST methods and added urgency to calls for reform of methodological practices. No longer an esoteric topic considered only by statisticians and a subset of researchers, the problems with NHST methodology have moved from little-read academic journals into popular publications and online blogs (Cumming, 2013; Nuzzo, 2014; Siegfried, 2010, 2014). Simultaneously, the replication crisis afflicting many scientific fields, especially the social sciences, was definitively demonstrated and widely

reported (Open Science Collaboration, 2012; Pashler & Wagenmakers, 2012; Wasserstein & Lazar, 2016). In response, some academic publications and professional organizations have renewed reconsideration of the role of NHST and p values in research, recommending restrictive methodological requirements for publication, discouraging, or in some cases banning, NHST or key features thereof while at the same time encouraging supplemental or alternative methods (International Committee of Medical Journal Editors, 2010; Trafimow, 2015; Wasserstein & Lazar, 2016; Wasserstein et al., 2019).

Before considering a selection of supplements or alternatives to NHST methods, I will review some of the major reasons these methods are thought necessary, along with the problems each method is attempting to address. There are many criticisms of the NHST research paradigm, but these can be categorized as involving the violation of assumptions, sample size limitations, problems with the nil-null, the dichotomization of decisions, problems with aggregation, and misinterpretation.

Violations of Assumptions

Null hypothesis statistical tests require that certain assumptions be met to ensure that p value calculations and inferences are accurate. The p value is the prior conditional probability of observing the obtained data, or data even more inconsistent with the null hypothesis if the null hypothesis is true and other assumptions are met. For many hypotheses tests, these assumptions include (a) random sampling, (b) normal distribution of errors, (c) equality of population error variances (homoscedasticity), (d) independence of observations, and (e) sampling and measurement error as the only sources of error (Kline, 2013; Loftus, 1996). Such assumptions are more restrictive than is frequently

thought by many researchers and are often not met (Kline, 2013), resulting in potentially biased p values. When p values are lower than they should be because of violations of assumptions, there is a positive bias meaning that the Type I error rate is higher than the stated α level. Conversely, if p values are too high, a negative bias results in inflated Type II errors. (Kline, 2013).

The problem of assumption violations is exasperated because there is evidence that researchers seldom explore whether they have met the assumptions of NHST and thus neither adjust their analyses to suit the data nor report the violations in research articles (Greenland et al., 2016). Keselman (1998) reviewed over 400 articles published in 17 prominent journals of psychology from 1994 to 1995, focusing on those with ANOVA-type designs to determine whether researchers reported investigating if their data met the assumptions of the analyses performed. Keselman found that of the 61 instances in which univariate ANOVA was used, 11.4% of articles referenced validating distributional assumptions, and only 8.1% reported validating equality of variance. Only 4.9% of the articles assessed both distributional and homogeneity assumptions. Of the 79 articles in which MANCOVA was used, only 6.3% validated distributional assumptions, and none mentioned variance homogeneity. Forty-eight articles mentioned using ANCOVA, but of these, only 4.1% reported validating distributional assumptions, and 8.3% validated equality of variance. Similarly, in the 226 articles containing repeated measures designs, 15.5% of researchers mentioned validating distributional assumptions, and 0.4% reported on variance assumptions. Osborne (2013) suggested that an overestimation of the robustness to violations of assumptions may partly be responsible for this tendency.

Sensitivity to Sample Size

Small sample sizes often lack sufficient power to detect even strong effects and fail to have p values smaller than the critical value (i.e., a Type II error is made).

Alternatively, when a sample size is large, trivial effects can produce impressively small p values. Levine et al. (2008) provide some accessible examples of how this looks practically. For the relationship $r = .40$ with $n = 20$, a two-tailed significance test at $p = .05$ is not statistically significant, whereas $r = .07$ with $n = 1000$, a dramatically weaker effect, is statistically significant. If a two-tailed NHST is conducted using $\alpha = .05$ with an observed effect of exactly $r = .25$, the results are statistically significant when $n = 63$ but not when $n = 61$. P values do not indicate the magnitude of an effect but rather are a function of effect size and sample size and of other contributors of statistical power such as directionality of the alternative hypothesis, experimental design, the test statistic, measurement reliability, and whether assumptions are met (Kline, 2013).

This sensitivity to sample size has several undesirable consequences. One is the paradox that with increased precision in the form of statistical power due to large samples, there is a greater possibility that practically unimportant or scientifically meaningless effects will be statistically significant. One result is that researchers can “chase significance” by increasing sample sizes, a form of questionable practice called p hacking (Simmons et al., 2011). Similarly, the advent of “big data” has led to the increased availability of massive data sets that result in studies being overpowered. These massive data sets increase the number of analyses that can be conducted and, with relatively lenient thresholds, such as $p = 0.05$, make statistically significant results much easier to obtain and the risk of false positives much greater (Ioannidis, 2019).

The Point or Nil–Null Hypothesis

The nil-null hypothesis is when the null hypothesis (H_0) is that some parameter is precisely zero or that a set of parameters are precisely equal. (Cohen, 1994; Meehl, 1978). In practice, the differences between means or the observed correlation of any variables, no matter how seemingly unrelated, will never be 0.00 out to the n th decimal place. Randomization does not perfectly balance the effects of all extraneous variables, nor should it be expected that in correlation studies there will be no uncontrolled third variables (Meehl, 1978). The nil-null hypothesis is nearly always false, so rejecting it is neither impressive nor informative. It will always be rejected, given enough statistical power (Cohen, 1994). This is more concerning when considering the effect sample size has on power and the increasing availability of massive data sets. Given that the nil-null hypothesis is always false and granted a large enough sample, NHST will always render a “significant” result regardless of the value of the alternative hypothesis, creating a divergence of statistical and practical significance (Meehl 1986).

Dichotomization of Decisions

A dichotomized decision rule is the practice of reducing the evidence necessary for scientific inference to mechanical “bright line” rules such as $p \leq 0.05$. In NHST, those findings on one side are deemed “statistically significant” with the implication that they are also meaningful or true, while those on the other are “not significant” and therefore not meaningful or true. Though convenient and straightforward, such a practice can lead to poor decision-making and erroneous beliefs. A p -value cannot definitively demonstrate the plausibility, presence, truth, or importance of an association or effect (Wasserstein et al., 2019). Also, when using a dichotomized decision rule, dramatic

differences in interpretation are made from inconsequential differences in p value estimates (are 0.051 and 0.049 really meaningfully different?). Furthermore, there is evidence that an editorial reviewer bias exists against the publication of non-significant results (Bakan, 1966; Bakker et al., 2012; Hubbard & Amrstrong, 1997; Ioannidis & Trikalinos, 2007). Such a bias distorts the published literature and incentivizes researchers to generate statistically significant results. This can encourage questionable research behaviors, which can distort research findings or help researchers achieve statistically significant results. Such questionable practices can involve manipulating the data collection, analysis, and reporting processes. Though not definitive, these in part consist of (a) collecting more data after an analysis reveals non-significance, (b) stopping data collection after achieving the desired result, (c) excluding data after looking at the effect they have on a specific research finding, (d) data mining/ p hacking, (e) failing to report all dependent measures that are relevant for a finding, and (f) falsifying data. John et al. (2012) and Fiedler and Schwarz (2016) provide evidence of researchers' wide use of such practices.

Misunderstanding

There are many false beliefs regarding NHST, most of which are misunderstandings or exaggerations of what can be inferred from the decision to reject or fail to reject the null hypothesis. Consideration here is limited to what has been identified as the “Big Five” false beliefs about NHST (Kline, 2013; Haller and Krauss, 2002).

These are

- The odds-against-chance fallacy is the false belief that the p value is the probability that the results were the consequence of sampling error.

- Local Type I error fallacy is the belief that when $p < .05$ given that $\alpha = .05$, then the probability of committing a Type I error is less than 5%.
- Inverse probability fallacy is the belief that the p value measures the likelihood that the null hypothesis is true given the data or $\Pr(H_0|D)$.
- The validity fallacy is the belief that the p value gives the likelihood of the alternative hypothesis rather than the null hypothesis. This fallacy states that the probability that H_1 is true is $1 - p$.
- The replicability fallacy is the false belief that $1 - p$ is the probability of finding the same statistically significant result in a replication of the study.

Haller and Kraus (2002) found that between 80% and 90% of psychology students and professors endorse at least one of the “Big Five” false beliefs. Cassidy et al. (2019) found that 89% of introductory psychological textbooks incorrectly define statistical significance, usually committing the odds against chance fallacy, providing evidence that the tools for training psychologists may contribute to the confusion.

Aggregated Statistics

Since the “inference revolution” described by Danziger (1990) and Gigerenzer and Murray (1987), data analysis of psychological processes has been predominantly conducted using NHST methods that focus on the variation between subjects. Inter-individual variation is used to derive aggregate statistics such as means or correlations that are calculated by pooling across subjects. Standard statistical methods, whether they be cross-sectional or longitudinal, single or multi-level designs, all focus on the analysis of inter-individual variation. The results of such analyses are to be generalized to a clearly defined population of which the sample is representative. It is important to note

that in this analysis framework, the aggregate statistic found in the sample is properly inferred to the aggregate population parameter.

Social scientific research questions and theories are often concerned with the psychology or actions of individuals, and the results from statistical methods that rely on aggregated statistics are frequently used to support theories about individual psychology (Grice, 2011; Lamiell, 2019). As Molenaar (2009) points out, it might seem self-evident that inferences about affairs at the population level represent findings that apply to each individual in the population. However, applying results obtained from pooling across individuals to each individual in the population requires a level shift from inter-individual variation to intra-individual variation. For this shift to be valid, specific conditions must be met. These conditions were determined as extensions of the classical ergodic theorems, which were derived in the early 1930s in ergodic theory research.

The first condition of ergodicity is the homogeneity of the population, such that the same statistical model is valid for every individual within a population. The features of a given statistical model describing the data must be invariant across all individuals (Molenaar, 2009). The second condition for ergodicity is stationarity and is achieved when a process, psychological or otherwise, has constant statistical characteristics over time that practically would manifest when statistical parameters of data are invariant across all time points. Any developmental process necessarily violates this condition.

Addressing Problems with NHST

Methodologies put forward to address the issues with NHST can be characterized as ranging from those that work within the NHST paradigm seeking to improve it to those that consider NHST so flawed as to abandon it altogether. Methods supplemental to

NHST are thought to be more palatable to applied researchers accustomed to NHST and thus more readily adopted by them. Such methods either aid in the interpretation of p values or posit alternative criteria for making decisions.

Interpretational Aids

Interpretational aids seek to mitigate the long-standing confusion associated with the interpretation of p values. Many such suggestions are drawn from Bayesian statistics, such as the Bayes factor bound (Benjamin & Berger, 2019), the false positive risk (Colquoun, 2019), and the analysis of credibility (Mathews, 2019). There are also non-Bayesian suggestions, such as the s value (Greenland, 2019), which derive from information theory. Regardless, shared by each methodology is the mathematical transformation of the p value into a more informative and easily understood statistic. Such methods are proposed to help researchers better understand and interpret the strength of evidence provided by p values. They do not address any of the critiques of the NHST process, such as violation of assumptions, sample size sensitivity, or problems with a nil-null hypothesis.

Decision Criteria

Alternative decision criteria methodologies posit different and generally more demanding criteria for making inferences within the NHST paradigm. These are principally a response to the problems of replication, which is often blamed on Type I error and the divergence between statistical and practical significance. These methods begin with the assumption that since researchers are likely to continue to use hypothesis tests with decision rules, different and more difficult decision rules will mitigate some of the problems with $p \leq 0.05$ dichotomized decisions.

Changes can be relatively minor such as increasing the threshold for significance. P values of 0.01 or 0.001 were suggested by Fisher as early as 1925 when required by research design and professional judgment. Benjamin and Berger (2019) more recently suggested replacing the 0.05 threshold with 0.005 for new discoveries and have had notable backing, gathering seventy co-authors (Benjamin et al., 2018) for a supportive comment article in the journal *Nature Human Behavior*. Gannon et al. (2019) propose a procedure that combines frequentist and Bayesian tools to provide a roaming significance level that is a function of sample size. Effects sizes have also been suggested as an alternative decision criterion. Huberty (2002) noted that researchers often employ effect sizes with p values somewhat informally as a decision criterion. Goodman et al. (2019) propose a formalization of this combined decision rule where researchers use $p \leq 0.05$ with a selected effect size cutoff value. Other suggestions involve altering the null hypothesis to form a null hypothesis belt or band that contains hypotheses that are practically without import (Serlin & Lapsley, 1985). Blume et al. (2019) suggest a refinement of the null belt as a second-generation p value (SGPV) that incorporates an interval null into its computation.

Deemphasize or Abandon Significance Testing

Methodologists who suggest moving away from statistical testing consider such tests philosophically flawed or practically misused to such a degree that much more will be gained in losing or limiting them than by persisting in their use. Common to each is a suggestion to make no claims about inferences to hypothetical parameters; such inference should be avoided altogether. These methodologies can vary widely as what they chiefly have in common is that they are not NHST.

Some suggest using intervals as an alternative, being careful to abandon the language of significance and the use of dichotomization (Amrhein et al., 2019). Others, such as Billheimer (2019), suggest abandoning inferences about parameters and instead focusing on predicting future observables and their associated uncertainty. Rogers (2010) contends that methods of building and testing models have already gained such wide use as to constitute a quiet revolution that has been taking place concurrent with the NHST controversy from the 1990s. Alternative paradigms such as single-case designs or qualitative methodologies are also widely used alternatives that address some of the issues with NHST, such as aggregation, that are not addressed by other methodologies. Observation Oriented Modeling is a notable alternative that seeks to integrate model building, a focus on the individual, and non-parametric methods into an accessible alternative to NHST.

Alternative methodologies

A loosely representative collection of alternative methodologies was selected to include in this study drawn from the three categories discussed in the above section, with two or three methods from each category. The selection of methodologies was not systematic, though that is not to say that selection criteria were not used. An attempt was made to include methods that belong to obvious sub-categories. When there were multiple examples of a sub-category, those requiring a lower level of expertise to calculate and exhibiting greater ease of interpretation were favored.

The Bayes factor bound and s value methods are the interpretational aides selected. Most interpretational aids rely upon the Bayes theorem to affect a transformation of the p value. The Bayes factor bound uses a more straightforward

transformation calculation than other Bayesian interpretational aides yet functions similarly. This is a characteristic that may help in adopting this method and is the reason for its inclusion. The *s* value is unique as it is the only interpretational aid method that is not Bayesian in origin. The alternative decision criteria methods were selected to represent the three sub-categories of decision criteria: alteration of the significance level or alpha ($p \leq 0.005$), use of an effect size measure (MESP), and use of a null band (SGPV). The vast diversity of methods that deemphasize or abandon significance testing presented more of a problem for selection than the other two categories. A discussion and comparison of radically different research paradigms such as single-case designs or qualitative methods are beyond the scope of this paper and were thus excluded. For the remainder, the two criteria already mentioned, as well as the preference of the author, were the determining factors in selecting interval estimation and Observation Oriented Modeling.

Each of the methods selected address some of the critiques of the NHST paradigm. Table 3.1 provides a summary of which of the seven critiques outlined in the literature review are attempted to be addressed by each of the methods included in this study. Details of the functioning of each method and the specifics of how each goes about addressing the various critiques will be discussed in the method descriptions that follow or in the demonstration and discussion sections.

Table 3.1 Critiques of NHST addressed by each proposed method

	Strength of evidence	Misinterpretation	Dichotomization	Sample size	Nil-null	Aggregate	Violation of assumptions
$p \leq 0.005$	✓	X	✓/X	X	X	X	X
Bayes Factor Bound	✓	✓	✓	X	X	X	X
S-values	✓	✓	✓	X	X	X	X
MESP	✓	✓	X	✓	✓	X	X
SGPV	✓	✓	✓	✓	✓	X	X
Interval estimation	✓	✓	✓	✓	✓	X	X
OOM	✓	✓	✓	✓	✓	✓	✓

Bayes Factor Bound

Bayesian statistical methods are often suggested as alternatives to NHST. For instance, Newman and Krull (2015) detail the potential benefits of Bayes factors over NHST when methods such as regression analysis or confidence intervals are inappropriate given the data or research questions. Bayesian methods, however, have a different philosophical underpinning than the frequentist methods commonly used by social science researchers who may find it challenging to restructure their thinking about statistics, specifically the use of prior distributions.

With this in mind, Benjamin and Berger (2019) suggest a calculation to transform the p value into a Bayes factor bound (BFB). The Bayes factor bound is calculated simply using the p -value, the natural logarithm, and its base, the constant e .

$$BF \leq BFB = \frac{1}{-e p \log p}$$

The BFB is interpreted as the maximum odds of observing the data under the alternative hypothesis to observing it under the null hypothesis, assuming that the hypotheses are equally likely. For example, a BFB of 2.45 to 1 indicates that the data is at most 2.45 times more likely under the H_A than under the H_0 . Calculating the BFB does not require specifying a particular alternative hypothesis because it is an upper bound across a large class of reasonable alternative hypotheses. The BFB can also be transformed into a posterior probability assuming the prior probabilities of H_0 to H_A are equal.

$$\Pr(H_A | p) = BFB / (1 + BFB)$$

For a BFB of 2.45, the posterior probability of the H_A is calculated to be 0.71, meaning there is a 71% chance that the alternative hypothesis is true given the data.

The potential value of the BFB is that it uses the p value to derive a statistic that clarifies the amount of evidence against the H_0 provided by the p value and thus can be an interpretive corrective to common misinterpretations of p values.

S value

The Shannon information value (Greenland, 2019; McKay, 2003), also known as an s value, is a statistic drawn from information theory pioneered by Claude Shannon (Shannon, 1948). The s value is a continuous measure of information in binary bits where each bit, as a binary digit, has only two outcomes. Greenland (2019) recommends transforming p values into s values to aid in interpretation. The transformation is accomplished by taking the negative log of the p value $-\log_2(p)$. By using s values, Greenland hopes to present a more intuitive statistic than p values. Greenland points to two areas where s values are an improvement. P values are directionally and conceptually mismatched; that is, as the value of p decreases, the evidence against the null hypothesis increases. Also, P values are not scaled so that the difference between values is consistent. For example, the difference between 0.01 and 0.10 is not the same as the difference between 0.90 and 0.99. The s value, by contrast, is a continuous measure of bits of information against the test hypothesis, which increases in a direct relationship. Further, since each bit is a binary digit and has only two outcomes, there is no problem with unequal scaling. Further, when rounded to the nearest integer, binary bits are intuitive to interpret; it is easy to think of them in familiar terms, such as a coin flip.

The s value works similarly in any analysis that provides the p statistic, which is common with any interpretational aid that is a transformation of the p value. Reporting of results can follow typical NHST procedure, except the exact p value associated with the

test statistic should be reported and transformed into an s value that is then reported in binary outcomes. The probability of consecutive binary outcomes should also be given. Confidence intervals can also be interpreted using s values.

Decrease Decision Threshold to 0.005

The methodological suggestion to reduce the decision threshold for inference from 0.05 to 0.005 for discoveries is not new (Greenwald et al. 1996; Johnson 2013) but was most forcefully made in 2018 in a comment article published in the journal *Nature Human Behavior* which listed seventy-two authors and co-authors (Benjamin et al., 2018). Benjamin and Berger, who are the first and second authors, repeated the suggestion in the 2019 ASA supplemental issue on p values (Benjamin & Berger, 2019). The authors make the case that the conventional NHST research tradition will not change easily or quickly, and therefore, incremental steps are necessary to move researchers away from the NHST paradigm.

A threshold for significance of 0.005 is suggested only for novel research. Further, Benjamin and Berger (2019) suggest a trichotomization of outcomes as a replacement for the dichotomization of NHST. According to this scheme, p values less than or equal to 0.005 are considered “significant,” and those between 0.05 and 0.005 are considered “suggestive.” Higher values are considered to be “non-significant.” Less stringent significance thresholds can be used for confirmation or replication research, or much more stringent ones may be appropriate, as is the case in genetics or physics. The primary concern is to begin to address the strength of evidence question. The authors consider the replication crises to be the result of a high number of Type I errors resulting from inflated claims of findings based on the 0.05 significance threshold. Despite wide

agreement that the use of dichotomous decision rules is inappropriate (Benjamin et al., 2018; Benjamin & Berger, 2019), increasing the threshold for significance from 0.05 to 0.005 is considered a marked improvement. To support the selection of 0.005, Benjamin and Berger use Bayesian statistical methods to calculate the probability of a true H_0 at a given p value. Assuming that the null hypothesis and the alternative hypothesis are equally likely (prior odds), for a $p = 0.05$, there is at least a 29% that the null hypothesis is true. This reduces to 7% when a threshold of 0.005 is used. They also present evidence of potentially dramatic decreases in estimated false positive rates. For example, the false positive rate is greater than 33% with prior odds of 1:5 and a p value threshold of 0.05, regardless of the level of statistical power. Reducing the threshold to 0.005 would reduce the minimum false positive rate to 5% (Benjamin et al., 2018).

Minimum Effect Plus p value

The use of effect size measures has long been suggested to correct the potential divergence of practical and statistical significance in NHST (Cohen, 1988; Thompson, 2002; Wilkinson, L., & the Task Force on Statistical Inference, 1999). Suggestions typically have taken the form of reporting an effect size alongside the p value. Huberty (2002) noted that researchers had employed two approaches to using effect sizes with p values: (a) using the p value ($p \leq 0.05$) to determine statistical significance and then the effect size to indicate the magnitude of an effect and (b) considering the p value and effect size jointly for determining if an effect is real. Recently, Goodman et al. (2019) have proposed a formalization of Huberty's second observation and recommend considering the p value and effect size together for statistical inference. Goodman calls this hybrid criterion minimum effect size plus p value (MESP). In this scheme,

researchers are tasked with selecting a minimum practically significant difference (MPSD) between the null hypothesis of no effect and a limit of practical effect prior to conducting research. It is assumed that researchers would have sufficient knowledge of past findings and competence in their field of expertise to be able to select a scientifically or practically meaningful MPSD and that it would be selected in good faith. The MPSD is then paired with a conventional p value methodology to construct a hybrid decision rule. The MESP method indicates that to reject a null hypothesis, each of two conditions must be satisfied: (1) $p \text{ value} \leq \alpha = 0.05$ and (2) the observed effect size \geq MPSD. A benefit of the MESP method is that it is practicable using existing statistical software without additional calculations since all software packages already provide a p value and an estimate of effect size.

Though MESP uses the conventional α of 0.05, there is no standard suggestion of a minimum practical effect. It is, therefore, necessary for researchers to think deeply when selecting a minimum practically significant difference. Unlike the mechanistic application of $p \leq 0.05$, research context and researcher knowledge are key to selecting an MPSD, and different researchers may not come to the same conclusions about what value best expresses practical significance. It is then incumbent upon researchers to make their best case for the value selected.

Second-generation p value

Serlin and Lapsley (1985) suggested using a “good enough” principle in selecting a null hypothesis as an alternative to the nil null or point hypotheses. Researchers would establish an interval null hypothesis of all those values that were not deemed to be practically meaningful. Consistent with the good enough principle, Blume et al. (2019)

propose replacing p values with a second-generation p value (SGPV) that incorporates an interval null into its computation. Blume's innovation is in devising a statistic that summarizes the test results in light of the “good enough belt” similar to that proposed by Selin and Lapsley. The SGPV is a measure of the data-supported hypotheses that are also scientifically or practically null hypotheses. Researchers would construct the interval null hypothesis by specifying in advance a range of effects that they consider to be without practical or scientific importance. Data-supported hypotheses are identified using interval estimates such as confidence intervals though other interval estimates can be used, such as support intervals or credible intervals.

Suppose there is a parameter of interest θ . Let $I = [\theta_l, \theta_u]$ be the interval estimate of θ and $|I| = \theta_l - \theta_u$ be the length of the estimate. Let the interval null hypotheses be H_0 and its length $|H_0|$. The SGPV is p_δ . The SGPV calculation is

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max \left(\frac{|I|}{2|H_0|}, 1 \right)$$

$$= \begin{cases} \frac{|I \cap H_0|}{|I|} & \text{when } |I| \leq 2|H_0| \\ \frac{1}{2} \frac{|I \cap H_0|}{|H_0|} & \text{when } |I| > 2|H_0| \end{cases}$$

where $I \cap H_0$ is the overlap of the data-based and null intervals. When the width of $|I| \leq 2|H_0|$ then the SGPV is a fraction of I that are H_0 . In instances when $|I| > 2|H_0|$, the interval estimate is very long and will often extend to either side of the null interval. The SGPV tends to be small when this occurs and does not properly reflect the inconclusive nature of the data. The correction term, $\frac{1}{2} \frac{|I \cap H_0|}{|H_0|}$, is used in these instances. Blume (2018) indicates that while many factors can result in an analysis that requires the correction term, typically, it will only be needed in severely underpowered studies. There is a

package available called “sgpv” (Welty et al., 2020) for R statistical software (R Core Team, 2020) that calculates the SGPV and automates the application of the correction term.

Like p values, the SGPV values range from 0 to 1. An SGPV of 1 indicates that the data only support the null hypotheses or trivially null effects. An SGPV of 0 indicates that the data are incompatible with any of the null hypotheses or, stated alternately, that the data only support meaningful effects. An SGPV between 0 and 1 is considered inconclusive to varying degrees. An SGPV of 0.5 is the most inconclusive, with the amount of inconclusiveness decreasing as SGPV nears either extreme. For example, when $p_\delta \approx 0.2$ the data could be interpreted as trending toward supporting a certain alternative hypothesis, or when $p_\delta \approx 0.1$, the data could be said to be suggestive of a meaningful effect but not definitively. Blume (2019), however, states that while descriptors of SGPV magnitude might be helpful as communicators of results, they are not essential since the ending states or the SGPV are well defined.

It is also important to consider that two studies with equal SGPVs do not necessarily represent equal amounts of statistical evidence. For instance, two studies, each with $p_\delta = 0$, can have very different distances between their respective interval estimates and null intervals. It can, then, be helpful to have a way of ranking studies with $p_\delta = 0$ by their strength of evidence. Blume (2018) proposes using the delta-gap, which is the distance between intervals in δ units, where $\delta = \frac{1}{2}|H_0|$. The delta gap should always be reported along with the $p_\delta = 0$. In addition, because the SGPV is a summary statistic designed to communicate information cogently, Blume also suggests that for the purposes of scientific discussion and policy decisions, detailed descriptions of the

findings should be provided in the form of an interval estimate of effect size and noting its proximity to the composite null hypothesis.

The null interval is the most important element of the SGPV and must be chosen with care and intention. Like with the selection of the MPSD, researchers must clearly indicate the values of the null interval and provide a rationale for their choice. Such information can be positioned as part of the method section of the study report or can be stated prior to the delivery of the results. In addition to past findings and subject matter expertise, Blume recommends that researchers might also consider measurement error, the gravity of the findings, and research community standards when selecting the null interval.

Interval Estimates

Interval estimates such as confidence intervals have long been suggested as a supplementary corrective to some of the problems associated with NHST (International Committee of Medical Journal Editors, 1997; Wilkinson & The APA Task Force on Statistical Inference, 1999). Because the intervals contain a range of plausible effects, they can shift the focus away from the dichotomization of significance and toward the observed effects. However, in practice, confidence intervals can be interpreted dichotomously where one declares statistical significance if an interval does not contain zero and non-significance when zero is within the interval (Kline, 2013). While this may be a natural interpretation for those steeped in the dichotomous thinking of NHST, it merely repeats the error and ignores the wealth of information contained within the interval. In 2019, Amrhein, Greenland, and McShane published a comment in the journal *Nature* with over 800 signatories calling for the cessation of the use of the concept of

statistical significance and advocating for interval estimates. These methodologists and practitioners see interval estimates as a viable alternative to NHST. However, both in the comment article and in a similar article in the *American Statistician* (Amrhein, Trafimow, & Greenland, 2019), strong suggestions are provided for how best to do so while avoiding using intervals as a mask for NHST. The suggested framework requires a shift in how researchers think about not only significance and p values but inferential statistics in general.

Amrhein et al. (2019) first suggest abandoning the language of significance and stopping the dichotomization of both p values and interval estimates. They do not think that p values should be abandoned but rather advocate for treating all inferential statistics as very unstable local descriptions of relationships between models and obtained data rather than providing generalizable inferences, keeping in mind that statistics such as p values are tests, not just of a single hypothesis but of all the assumptions of the model. The importance of this cannot be overstated. The p value here is understood as a descriptive statistic indicating the fit of a particular set of data to a model, which assumes not only a null hypothesis but a range of other assumptions, any one of which could affect the model fit. A small p value is the result of a combination of random variation and violations of model assumptions. However, it does not indicate which (if any) assumption is violated. It could indeed result because the null hypothesis is false, but it can also mean that the model was not correctly specified, sampling was not random, the analyses used led to a small p value (downward p hacking), or the measurement instrument did not measure what it was thought to measure (Amrhein, 2018).

Further, to ward against researchers using confidence intervals as de facto tests of significance, Amrhein et al. (2019) advocate for interpreting confidence intervals not as a range of values that contain the true parameter at a selected level of confidence but as “compatibility intervals showing the effect sizes most compatible with the data according to their p values, under the model used to compute the interval” (Amrhein et al., 2019, p. 265). In addition, Amrhein has several specific methodological recommendations. Estimates, not tests, should be emphasized and interpreted, explicitly discussing both the lower and upper bounds of interval estimates. The precise values of statistics should be reported, not inequalities; for example, the exact p value such as “ $p = 0.02$ ” not “ $p < .05$ ”. Amrhein et al. also recommend using s values, likelihood ratios, and Bayesian methods to assist in further explaining and interpreting the data. However, here the focus will be kept on interval estimates since some of these methods have been discussed separately in this paper and because Amrhein’s primary concern is with interval estimates.

It is important to remember that the analysis focuses on the interpretation of intervals and not on a statistical test. In fact, the statistical test does not need to be reported at all. The nil null hypothesis is considered as part of the interval together with a range of other hypotheses and can be explicitly discussed, but one must be careful not to make claims about findings based only on the presence or absence of the nil null in an interval. The language used to discuss the interval estimate is also changed. Following Amrhein et al. (2019), “confidence intervals” become “compatibility intervals,” and language suggesting the presence of a true effect within the interval or of support of any

single value within an interval is avoided. Instead, intervals are discussed in terms of compatibility with the data or the model used to compute the interval.

Observation Oriented Modeling

Observation Orientated Modelling (OOM) is a methodological approach developed by James Grice consisting of a collection of analytic methods posited as alternatives to NHST (Grice, 2011). OOM analyses are accessible through a freely available software package. The OOM approach has much in common with the suggestions of Amrhein et al. in that it abandons hypothesis tests and inferences to population parameters. Further, OOM analyses are assumption-free and can be used with data that are inappropriate under NHST. OOM also attempts to address the aggregation criticism of NSHT by focusing the analysis on individuals instead of aggregate statistics. Grice asserts that the methodology and metaphysics of the prevailing research tradition have retarded the accumulation of genuinely scientific knowledge. OOM is advanced as a collection of alternative analytical methods based upon different philosophical assumptions that place the analytical focus on identifying patterns of individual observations and developing and testing hypotheses about the psychological processes that give rise to those patterns. There is no inference to population parameters, and generalization is achieved only through replication. This is in contrast with the variable and parameter orientated approach of NHST. This contrast is an extension of the fundamental differences in the philosophies undergirding the two approaches; Arocha (2020) provides a useful presentation of the differences between moderate scientific realism, which undergirds OOM, and the empiricism and operationalism of the NHST paradigm. From this basis, Grice posits several principles of research whereby

researchers should focus upon phenomena that are real, accurate, repeated, and observable and that researchers should use integrated causal models that require deep consideration of causation which, following moderate realism, adhere within individuals not variables, incorporate statistical outliers, and avoid aggregation (Grice, 2011).

Observation Oriented Modeling's analytical procedures all function to build and test theoretical models that predict patterns of behaviors of individuals. Most OOM analyses require researchers to predict patterns a priori based on theoretical models, which are then analyzed using the data to determine the accuracy of the prediction and relative rarity of the outcomes. Predictions can be as precise as theoretical models allow, but all must predict individuals' behavior, not aggregations. A detailed examination of the construction of models is beyond the scope of the present paper, which will focus only on the analytical methods.

For example, consider the two-group data; using OOM software, a researcher could choose a precise pattern based on theory that predicts exactly the ACT composite scores of individuals in each group. If a theoretical model is not refined enough to predict precise patterns, usually, one will at least be able to indicate the expected ordinal relation of pairs of individuals from each group. For instance, in the two-group example, a researcher could expect that individuals in the after-school tutoring condition will have higher ACT composite scores than those in the control condition. Figure 3.1 shows how this pattern is indicated in the OOM software.

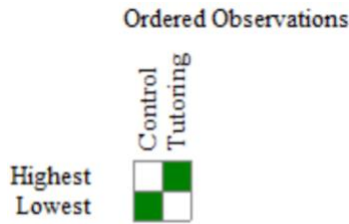


Figure 3.1 Expected Ordinal Pattern

Ordinal pattern analysis is conducted to test the predicted pattern, whereby each individual in the after-school tutoring condition is paired with every individual in the control condition. Our anticipated pattern is that a student in the afterschool tutoring condition will have a higher ACT composite score than any student in the control condition. The ordinal relationship of pairs is evaluated, and the percentage of observations that match the predicted pattern is calculated, called the Percent Correct Classification (PCC). The PCC serves the same function as an effect size in conventional statistical analyses. A PCC can range from 0 to 1, where a value of 1 indicates that all observations are accounted for by the predicted pattern. The OOM methodology uses a resampling procedure to determine what value of PCC might be considered a success. The procedure, known as a randomization test, randomly reshuffles the data and recalculates the PCC value. The process is repeated a selected number of times (Grice recommends 500 to 1000), and a probability statistic is calculated called the chance value or c-value. The chance value is the frequency with which the resampling yields results at least as accurate to the researcher determined pattern as the initial data. This provides a measure of the rarity of the outcome achieved, given the data. If the randomized data sets fit the selected pattern as well or better than the actual data, then a high c-value (closer to

1) will result, indicating that the observed outcomes are common. Conversely, a low c-value (closer to 0) results when the randomized data set does not fit the selected pattern as well as the actual data, indicating that the pattern of observations is unlikely. It should be noted that the sample size does have an effect on the usefulness of the c-value statistic. Small data sets allow for fewer random permutations from which comparisons may be obtained. As a result, the c-value is less informative as to the probability of the obtained outcome. Inferential caution is needed and should be exercised regardless of the size of data sets. Further, cut-off values should not be used, or the result would be a similar error as that of NHST (Grice, 2017).

Method

The seven representative alternative methods will be demonstrated using simulated data sets. The simulations feature three common experimental designs that typically use NHST analysis: two independent groups, three independent groups, and two-by-two between subjects. Examples will be drawn from each of the experimental designs that will help demonstrate the application of the various methods. The three simulated design examples will be analyzed using each of the seven methodologies; however, not every method will be reported for each design at the same level of detail since some methods, particularly interpretational aids, function the same regardless of experimental design or hypothesis test used. Hypothetical research scenarios will be provided for the simulated sets modeled on actual studies to aid in demonstrating the workings of the various methods, some of which incorporate elements of subjectivity and particularity that can only be demonstrated in an actual research context.

Except for OOM, all demonstrated methods will use a t-test with the two-group design and an ANOVA with the three-group and two-by-two between-subject designs. The results will be written in a format consistent with that which would appear in a formal research report following each method's conventions and recommendations. The purpose will be to demonstrate how each method functions in analyzing and interpreting data. Also, comments will be provided discussing the use of methodology within each particular context.

Data Simulation

All data sets for each experimental design were simulated using R statistical software version 4.0.2 (R Core Team, 2020) and supplementary packages “broom” (Robinson, Hayes, and Couch, 2021), “effectsize” (Ben-Shachar & Makowski, 2020), “sgpv” (Welty et al., 2020), and “tidyverse” (Wickham et al., 2019). All simulated data are inspired by actual research, which is relied upon to determine the condition parameters. Further, each experimental design is given a research scenario inspired by real experiments. However, the simulations are not intended to, and indeed do not exactly match or emulate any given study. Instead, the simulations provide an occasion to demonstrate the various analytical methods under plausible research scenarios.

The two-group experimental design data was simulated using the following linear model.

$$y_{ik} = \mu + \alpha_k + \varepsilon_{ik}$$

where μ_k is the mean, α_k is the effect parameter, and ε_{ik} is the error term. The scores of the control group (y_{i1}) were simulated with $\mu = 22$, $\alpha_1 = -1$, and ε_{i1} with a sample of 100, a mean of 0, and a standard deviation of 5.9. The treatment group scores (y_{i2}) were

simulated with $\mu = 22$, $\alpha_2 = 1$, and ε_{i1} with a sample size of 100, a mean of 0, and a standard deviation of 5.9. The parameters for the two-group experimental design conditions are shown in Table 3.2. All parameters were selected to reflect actual research on test preparation and the ACT test drawing especially from the work of Moore, Sanchez, and San Pedro (2018) and the meta-analytical study of Montgomery and Lilly, J. (2012). Each condition was randomly sampled from an independent normal distribution to simulate sampling variability realistically.

Table 3.2 Two-group Design Parameters

	μ_k	α_k	σ_k	n_k
Control Group	21	-1	5.9	100
Treatment Group	23	+1	5.9	100

The research scenario is a two-group post-test design having a treatment group that received six weeks of after-school ACT tutoring instruction and a control group that received no instruction. The sample is drawn from high school juniors who intend to take the ACT for the first time. The students were administered the ACT at the end of the 6-week tutoring program, the composite scores of which are the outcome measure. The goal of the research is to determine whether this particular tutoring program is more effective at preparing students to take the ACT than no preparatory instruction.

The three-group experimental design data was also simulated using the linear model

$$y_{ik} = \mu + \alpha_k + \varepsilon_{ik}$$

The scores of the control group (y_{i1}), treatment group 1 (y_{i2}), and treatment group 2 (y_{i3}), were independently simulated with $\mu = 22$, and effect sizes $\alpha_1 = -1$, $\alpha_2 = -1$,

and $\alpha_3 = 2$. The residuals ε_{ik} for all groups were simulated with a sample size of 53, a mean of 0, and a standard deviation of 5.9. The parameters for the three-group experimental design conditions are shown in Table 3.2. These parameters were also selected drawing from the work of Moore, Sanchez, and San Pedro (2018) and the meta-analytical study of Montgomery and Lilly, J. (2012). Each of the three conditions was randomly sampled from an independent normal distribution to simulate sampling variability realistically.

Table 3.3 Three-group Design Parameters

	μ_k	α_k	σ_k	n_k
Control Group	21	-1	5.9	53
Treatment Group 1	21	-1	5.9	53
Treatment Group 2	24	+2	5.9	53

The research scenario is a three-group posttest design having a treatment group one which received six weeks of after-school ACT tutoring instruction, treatment group two, which was given access to self-directed online ACT preparation, and a control group that received no instruction of any kind. The sample is drawn from high school juniors who intend to take the ACT for the first time. The students were administered the ACT at the end of the 6-week tutoring program. The composite ACT scores, the scale of which ranges from 0 to 36, are the outcome measure. The goal of the research is to determine if there are differences in the effectiveness of in-person and online tutoring in preparing students to take the ACT.

The two-by-two between-subjects data linear model is

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

where μ is the grand mean, α_j is a main effect parameter for the first factor, β_k is the effect parameter for the second factor, γ_{jk} is the interaction effect parameter and ε_{ijk} is the error term. Each factor has two conditions resulting in a total of four simulation conditions. The simulation parameters of each condition are shown in table 3.4. The error term for all conditions was simulated with a sample size of 32, a mean of 0, and a standard deviation of 3. The parameters and scenario are borrowed in slightly modified form from a study involving the influence of study method and familiarity with the words upon vocabulary test performance (Garber, 2021).

Table 3.4 2x2 Between-subjects Design Parameters

	μ_{jk}	α_i	β_k	γ_{ik}	σ_{jk}	n_{jk}
Passages/Unfamiliar Words	12	-.8	-1	+.5	3	32
Passages/Familiar Words	12	-.8	+1	-.5	3	32
Definitions/Unfamiliar Words	12	+.8	-1	-.5	3	32
Definitions/Familiar Words	12	+.8	+1	+.5	3	32

In this scenario, 128 middle school participants were randomly assigned to two type of study conditions in order to prepare for a 20-item vocabulary assessment. One group was given passages to read which contained the words upon which they would be assessed, while the other group was provided the definitions of the words to read. Students were also randomly assigned to a word type condition wherein one group would be assessed upon unfamiliar words and the other group upon familiar words. This results in four experimental conditions to which the students were assigned: passages with unfamiliar words, passages with familiar words, definitions with unfamiliar words, and definitions with familiar words.

Limitations

The present study is limited in that it is a demonstration of methodologies intended to provide examples of the application of alternative methods to NHST within the context of common designs. As such, the study does not allow for the comparison of outcome trends of the various methods. Such a comparison is better accomplished through Monte Carlo methods simulation studies. Further, the analytical outcomes from a single example should not be taken to be typical or representational but only examples of the application of the method. Also, because the focus is on the demonstration of the methods when a given method is applied in the same manner regardless of the experimental design, one example will be considered sufficient. Interpretational methods are applied and function identically regardless of the experimental design because they merely transform a p value into a more easily interpreted form. Therefore, in the interest of parsimony, interpretational methods will not be demonstrated across all three experimental designs.

Demonstration of Selected Alternatives

Demonstration of methods is accomplished by analyzing the data of each experimental design using each respective method. A description of the analysis of the data for each method is provided. A research article style report of the findings based on the analysis for each method is also provided to demonstrate how the results could be presented in a formal context.

Two-group Design

The two-group design is among the simplest experimental designs. Its benefit is to chiefly be found in that it allows for the demonstration of the proposed methods in an

uncomplicated experimental context, that is when there is only one effect size of interest. Further, the effect is easy to conceptualize since it is the difference between the means of the two groups. This understanding of effect, however, does not hold for OOM, which eschews using aggregate statistics. This difference is further elucidated in the OOM demonstration. The benefits of the simplicity of the two-group design, however, do still help in demonstrating the OOM methodology.

Bayes Factor Bound

The BFB is an interpretation aid and does not change most of the fundamentals of analysis, which are firmly within the NHST paradigm. For a two-group design, a t -test is conducted, and a p value, test statistic, and descriptive statistics such as means and standard deviations are reported. However, the use of BFB requires reporting the p value as a continuous statistic instead of merely indicating whether it is less than a given cut-off value. This shift away from the dichotomy of significance found in NHST is amplified by transforming the p value into the BFB, which is also reported and aids in interpretation by allowing for a consideration of the evidence in favor of the alternative hypothesis as odds and probabilities. This change puts the onus upon the researcher and the reader of the findings to deeply consider how determinate one should consider the results of a single study that reports a certain percentage chance that the H_0 is true. Whereas, under traditional NHST, a researcher might declare the results to be “significant” with a $p \leq 0.05$ and thus reject the H_0 , here the researcher must be content that the transformed p value indicated odds that the H_0 is true. Definitive statements of the inferential truth or falsehood of a finding are more difficult when considering the strength of evidence using the BFB.

For example, the simulated data for the two-group design resulted in a p value of 0.01, which would be considered significant using the typical NHST significance value of $p \leq 0.05$. However, the BFB is 6.67, indicating that there is at most a 6.67 to 1 chance that the H_A is true. This BFB corresponds to a probability of 87%, which is high, but it bears considering that there is still a least a 13% chance that the H_0 is true. The results of the analysis are below.

An independent samples t-test was performed comparing the mean ACT composite scores of high school juniors who received after-school ACT tutoring and those who received no preparation services. Those in the tutoring condition ($M = 22.8$, $SD = 5.9$, $n = 100$) had a higher sample mean score than those who did not receive tutoring ($M = 20.8$, $SD = 5.5$, $n = 100$). The t-test provides evidence against the null hypothesis of no effect ($t(197) = 2.52$, $p = 0.01$). The attained p value corresponds to an upper Bayes factor bound of 6.67. There are then odds of at most 6.67 to 1 that the alternative hypothesis that there is a mean difference between the two conditions is true, which assuming that the null and alternative hypotheses were originally equally likely corresponds to a probability of 87% that the alternative hypothesis is true.

S value

Much like the BFB, the s value is not meant to replace the p value but to aid in interpretation while gently leading researchers away from dichotomous decisions. It, too, accomplishes this by requiring the reporting and consideration of the attained p value then transforming it into a more intuitive measure. In this example, the p value of 0.01 has an s value of 6.3. Fractional bits of information are confusing, so following the

recommendation of Greenland (2019), the s value is rounded to the nearest integer for reporting. Because the s value uses binary digits, the probability of consecutive outcomes can be calculated and reported. It is left up to the researcher's judgment whether the probability is sufficiently extreme to make any inferential claim. However, as a supplement, the s value method does not provide a framework for making such decisions. The s value can also be used to help interpret the parameter estimates within a confidence interval. The point estimate is the most compatible with the data (meaning it has the least refutational information against it), while those values near the limits have more information against them with a maximum of 4.3 when a 95% CI is used. This is because the limits of a 95% CI have a $p < .05$, which equals an s value of 4.3.

Again, consider the example of the two-group design simulation.

An independent samples t -test was performed comparing the mean ACT composite scores of high school juniors who received after-school ACT tutoring and those who received no preparation services. Those in the tutoring condition ($M = 22.8$, $SD = 5.9$, $n = 100$) had a higher sample mean score than those who did not receive tutoring ($M = 20.8$, $SD = 5.5$, $n = 100$). The t -test provides evidence against the null hypothesis of no effect ($t(197) = 2.52$, $p = 0.01$). The attained p value corresponds to an s value of 6.3, indicating that there are 6 bits of binary information against the null hypothesis, making it as likely as getting heads on six consecutive fair coin tosses. The probability of such a result is 1.6% and provides strong evidence against the null hypothesis of no effect. The raw effect is 2.04 with a 95% CI[0.44, 3.63]. The most extreme values within the 95% CI, 0.44 and 3.63, have at most 4.3 bits of information against them. No

parameter value inside the 95% confidence interval has more than 4.3 bits of information against it. All other values within the CI are increasingly likely given the data, with the point estimate of 2.04 being the most likely.

Decrease Decision Threshold to 0.005

The alternate decision threshold method suggested by Benjamin and Berger is similar to NHST in most respects. It differs in that it uses a reduced significance level and a trichotomized decision rule. The attained p value with the two-group experimental design does not meet the significance cut-off of 0.005, but it is less than 0.05 and, within the trichotomized framework of Benjamin and Berger, is considered to be “suggestive.” The strength of evidence is thus considered weaker than “significant” but still meaningful. The confidence interval is included in the report and considered to provide additional evidence. It is good practice to discuss the CI bounds but is avoided in this example since it is not mentioned in the literature suggesting the 0.005 cut-off. The mention of the need for additional evidence or research is necessary to emphasize the relative weakness of the evidence compared to that of a $p \leq 0.005$ and to make it clear that a $p \leq 0.05$ is insufficient for strong inferential claims. Let us now consider the two-group example:

An independent samples t -test was performed comparing the mean ACT composite scores of high school juniors who received after-school ACT tutoring and those who received no preparation services. Those in the tutoring condition ($M = 22.8$, $SD = 5.9$, $n = 100$) had a higher sample mean score than those who did not receive tutoring ($M = 20.8$, $SD = 5.5$, $n = 100$). The t -test provides suggestive evidence against the null hypothesis of no effect ($t(197) = 2.52$,

$p = 0.01$) as the p value is less than 0.05 but does not meet the $p = 0.005$ significance threshold. The raw effect is 2.04 with a 95% CI[0.44, 3.63] which provides evidence of an effect; however, additional evidence is required to make a more determinate claim regarding the size of the population effect.

Minimum Effect Plus p value

MESP is a dual decision criteria methodology using both the p value and a minimum practically significant effect size. To conduct an analysis, an MPSD must be determined. For this analysis, there is a wealth of research on the effects of retesting and test preparation on scores of the ACT and other similar achievement tests. This information, together with the researcher's expert judgment, needs to be brought together to make a determination of an MPSD, which then must be detailed in the research reporting. For the analysis of the ACT data, evidence from meta-analyses investigating the effects of preparation and retesting on cognitive assessments similar to the ACT were consulted. The results showed an average score gain of .25 standard deviations from either test preparation or retesting, with test preparation showing no benefit over retesting. This corresponds to approximately 1.2 – 1.5 ACT composite score points (Montgomery & Lily, 2012; Powers, 1993). Research looking specifically at the ACT is scarce but yielded similar results of between a 1 and 2 composite score point increase (Andrews & Ziomek, 1998; Moore, Sanchez, & San Pedro, 2018). Taking all this together, the standardized increase of .25 was used together with the highest ACT standard deviation (5.9) reported by the ACT organization from the 2000 - 01 to the 2020 - 21 testing years to calculate a raw effect of 1.5 rounded to the tenth decimal place. The highest reported standard deviation was used because the standard deviation has steadily

increased on the ACT from year to year for the period. The raw effect is used instead of the standardized effect measure because the raw composite score increase is a more meaningful measure in the context of a standardized test. A standardized measure of effect will be preferable in other contexts.

The analysis follows the standard procedure of reporting the test statistic and the p value, with the p value being judged as either “significant” or “not significant” based upon a cut-off value of $p \leq 0.05$. This determination is not enough to declare findings practically significant as the effect size must also be greater than the MPSD. For this analysis, the MESP determined that the findings were practically significant with a p value of 0.01 and an effect of $d = 0.36$ or 2.04 composite score points. Confidence intervals are not reported or discussed in this methodology.

The t test example analysis is shown below.

An independent samples t -test was performed with $\alpha = 0.05$ and a minimum practically significant difference (MPSD) of 1.5 composite score points or a .25 standardized effect. Meta-analyses investigating the effects of preparation and retesting on cognitive assessments similar to the ACT consistently reveal a score gain of .25 standard deviations or between 1.2 and 1.5 ACT composite score points (Montgomery & Lily, 2012; Powers, 1993). Research looking specifically at the ACT is scarce but has yielded similar results of between 1 and 2 composite score point increase (Andrews & Ziomek, 1998; Moore, Sanchez, & San Pedro, 2018). Based on previous findings, a standardized MPSD of .25 standardized deviation increase was used for this analysis.

The t-test shows a statistically and practically significant effect ($t(197) = 2.52$, $p = 0.01$) as the p value is less than 0.05 and the standardized effect is $d = 0.36$, larger than the MPSD of 0.25. The raw effect is 2.04 scale score points, larger than the maximum increase of 1.5 observed in previous research involving preparation and retesting. The observed effect of the tutoring program was greater than the expected effect of other preparation programs or simple retesting. The results provide strong evidence of a practically significant effect of the preparation program.

Second-generation p value

The second-generation p value (p_δ) is a summary statistic providing the proportion of data-supported hypotheses that are also scientifically or practically null hypotheses. An interval null hypothesis is central to the functioning of the SGPV and must be determined by researchers, preferably during the design phase. For this analysis, the method of determining the null interval references the same data as that used to select the MPSD in the MESP method. Here a change of score greater than .25 standard deviations will be used to indicate practical importance. The null interval then will have bounds of -.25 and .25. Considering a large decrease is important, as opposed to only considering a positive score increase since this hypothetical preparation method is new with no existing research on its effects on ACT scores.

Below is an example related to the two-group post-test simulation. In the example, note that the null interval is discussed and justified prior to reporting results. Such vital information must be explicitly relayed since the interval estimate and the SGPV alone do not communicate it. The unstandardized and standardized effects

confidence intervals are both reported for clarity, though reporting both is not strictly necessary. Finally, the overlap of the data-supported interval with the null interval is briefly discussed to flesh out the meaning of the SGPV. Had there been evidence of an effect, indicated by a p_δ of 0, then the distance between intervals should be mentioned along with the delta gap to indicate the strength of evidence and allow for comparisons to similar studies using SGPV. Researchers will likely need to flesh out the meaning of both the p_δ and the delta gap as part of the discussion of their findings.

Meta-analyses investigating the effects of retesting and preparation on cognitive assessments similar to the ACT reveal that for both retesting and preparation, an average score gain of .25 standard deviations or between 1.2 and 1.5 ACT composite score points (Montgomery & Lily, 2012; Powers, 1993). Research looking specifically at the ACT has yielded similar results of between 1 and 2 composite score point increase (Andrews & Ziomek, 1998; Moore, Sanchez, & San Pedro, 2018). The standard deviation of the ACT has maintained or increased every year from 2000-19, ranging from 4.5 to 5.9 SD. A standardized effect was thus deemed most appropriate for the null interval.

An independent samples t -test was performed using a null interval of $[-.25, .25]$ in Cohen's d standardized effect. No evidence of a scientifically meaningful effect of the after-school ACT tutoring program was found, Cohen's d 95% CI $[0.076, 0.63]$, $p_\delta = 0.31$). The point estimate of effect was $d = .36$ or 2.04 composite score points. However, when considering the intervals, a p_δ of 0.31 reveals that 31% of the data-supported hypotheses are also null hypotheses providing weak evidence of an effect.

Interval Estimates

Interval estimates, particularly confidence intervals, should be familiar to most researchers. However, following the suggestions of Amrhein et al., the interpretation and specific language used may be quite different. As seen in the example below, there is an avoidance of discussion of a true parameter. Instead, the intervals are discussed as providing the effect sizes most compatible with the data under the model used to compute the interval (Amrhein et al., 2019, p. 265). Also, the consideration is not of the presence of a nil value within the interval but rather of all the compatible estimates. The range of estimates is explicitly discussed, emphasizing both the lower and upper bounds. That is not to say that the point estimate is of no value; on the contrary, it is the value most compatible with the data under the model and is noted as such.

For this data, the standardized point effect estimate is 0.36, which provides evidence that there is a meaningful effect from the tutoring program. However, there is a relatively wide range of other values that are also highly compatible with the data, from trivial ($d = 0.08$) to very high ($d = 3.63$). Because of this, despite the evidence of the point estimate, the results must be considered inconclusive. Of course, with this data, we are again using the evidence of prior research to inform our determination instead of an arbitrary cut-off or the presence of a nil value within the interval. In other contexts, researchers will either need to determine standards of meaningfulness and explain them or simply report the range of compatible effects without further comment on meaningfulness. The example write-up is shown below.

Analyses using an independent samples t test were conducted to investigate the effects of an ACT after-school tutoring program on ACT composite scores. The

unstandardized effect estimate of 2.04 is slightly higher than the upper bound of the expected range of 1-2 ACT composite score points. The standardized effect is $d = 0.36$, higher than the average effect of .25 observed in similar experiments. However, other effects highly compatible with the sample data, given the model, reveal that the results are inconclusive with regard to the magnitude of the effect. Within a compatibility interval of 95%, unstandardized effects were observed, ranging from 0.44 (a trivial difference) to 3.63 (more than 1.5 times the maximum expected amount). Compatible standardized effects in the 95% compatibility interval range from 0.08 to 0.64.

Observation Oriented Modeling

The two-group data is analyzed using the OOM procedure ordinal pattern analysis. Ordinal pattern analysis was selected because our model cannot predict the precise ACT scores that each individual should achieve in each group. We are only predicting the relationship of scores of individuals in the control group to those in the treatment group: namely, individuals in the control group will have lower scores than those in the treatment group. If a higher level of precision were possible, a different procedure called a pattern analysis would be used that allows for the selection of precise patterns of scores by grouping. Note that in the report, descriptive statistics of central tendency contain meaningful information and should be reported even though OOM analysis focuses on individuals. Also, the percent correct classification (PCC) is a descriptive statistic relating the percent of times in a pairwise comparison that an individual in the experiential condition had a higher score than an individual in the control condition. The PCC should not be construed as inferential to a parameter. The

PCC can be considered similarly to an effect size, indicating the strength of an effect. The rarity of the PCC is indicated by the c value that results from a randomization test. There is, however, no cut-off mentioned, only whether the c value seems to be relatively rare or not based on the researchers' judgment. Further, it should be noted that simply because a PCC is rare, it does not follow that it is also particularly meaningful. A trivially small PCC can be rare, given the data. The write-up of the results is below.

Analyses were performed to investigate the effect of an after-school tutoring program on individual students' ACT achievement. Descriptive statistics for each group suggest a modest effect: Control condition ($n = 100$, $M = 20.75$, $SD = 5.51$), After-school condition ($n = 100$, $M = 22.87$, $SD = 5.92$). Ordinal pattern analysis was performed using an expected pattern of higher ACT composite scores among individuals in the after-school tutoring condition compared to the control condition. The pattern derives from our theory that supplementary ACT specific tutoring will better prepare a student to take the ACT than if the student had only participated in regular educational courses that contained no ACT specific preparation. Ten thousand pairs of observations were compared and had a percent correct classification of 56.73, a result somewhat higher than the 50% that would be expected if individuals in each group performed equally well. A randomization test with 1000 trials was conducted to investigate the rarity of the obtained PCC, given the data. A PCC range of 33.24 – 60.55 was calculated. Of these, 19 were greater than or equal to the observed PCC of 56.73, resulting in a c value of 0.02.

The analysis results indicate that our model can only predict 56.73% of the outcomes of the pairwise comparisons of individuals from each group, indicating a weak effect of the tutoring program on individual performance. However, the c value is relatively rare, granting evidence that the observed PCC is unusual and may not be the result of chance variation. Further research is needed to account for the overperformance of individuals in the control group and the underperformance of some individuals in the experimental group, specifically, the identification of characteristics or exposures that will aid in the construction of more detailed and accurate models.

Three-group Design

The three-group design has a more complicated structure and analysis than the two-group design. The introduction of a third group makes possible the demonstration of the methodologies with omnibus hypotheses as well as multiple comparisons.

Decrease Decision Threshold to 0.005

The three-group simulated data can help to demonstrate how this method is used with an ANOVA analysis with multiple comparisons. The application is straightforward since we are simply using the p value with a more stringent standard of significance. The analysis and reporting are otherwise identical to NHST for both the ANOVA and the post hoc contrast analysis. It is, however, important to notice that in the example below, because the attained p value is less than the significance level of 0.005, the results are reported just as they would be in NHST. This is in contrast with the finding of the two-group example, which fell into a third category between non-significant and significant and was considered “suggestive.”

A one-way ANOVA was conducted to investigate the effects of ACT preparation programs on mean ACT composite scores. The analysis of variance showed a statistically significant effect of ACT preparation on ACT scores ($F(2,156) = 5.78, p = 0.004$) using a 0.005 significance level. Post hoc contrasts using the Tukey HSD test indicated that the after-school tutoring condition ($M = 22.9, SD = 6.1, n = 53$) was significantly different ($p = 0.003$) than the control condition ($M = 19.2, SD = 5.2, n = 53$). However, the online tutoring condition was not statistically different ($p = 0.15$) from the control condition. Further, there was no statistically significant difference ($p = .46$) between the after-school tutoring and online tutoring conditions.

Minimum Effect Plus p value

When conducting more complex analyses with MESP, researchers will have to think deeply and creatively about how to determine the MPSD. The ANOVA analysis used with the three-group design presents a unique challenge since there are no references in the literature for meaningful effects with an omnibus test. One option is to use the average standardized effect of $d = 0.25$ as a reference for determining the MPSD for the ANOVA. This can be accomplished by converting the Cohen's d to an effect size coefficient appropriate for an ANOVA analysis. Other options could be to use the lower ($d = .17$) or upper ($d = .34$) bounds of observed effects or to use recommendations such as those offered by Cohen for meaningful effects. Further, the type of ANOVA appropriate effect size measure must also be selected. The η^2 is used in this analysis merely because it is one of the more widely used effect size calculations. There are several other effect sizes, such as the ω^2 or ε^2 , that would be appropriate as

well. All of this is, of course, an imperfect solution. However, an effect size measure must be selected for the MPSD and whatever is chosen must be explained and justified.

To demonstrate how these choices might affect the results of the test, we will briefly consider what would happen if researchers chose the average effect ($\eta_2 = 0.015$), the lower effect bound ($\eta_2 = 0.007$), or the upper effect bound ($\eta_2 = 0.028$) as the MPSD. The ANOVA results provide a p value of 0.004 and an effect of $\eta_2 = 0.069$. The p value is less than the MESP cut-off value of 0.05 and is considered statistically significant. The observed effect of 0.069 is greater than any of the MPSDs under consideration and is practically significant regardless of choice. In this example, the selection of the effect size cut-off value, even a high one, makes no difference at all.

The write-up of the results is below. For this example, the average effect based on a d of 0.25 was used for the MPSD for both the ANOVA test and the post hoc contrasts using the Tukey HSD test.

A one-way ANOVA was conducted to investigate the effects of ACT preparation programs on mean ACT composite scores. The analysis of variance showed a statistically and practically significant effect of ACT preparation on ACT scores ($F(2,156) = 5.78, p = 0.004$) with an effect of $\eta_2 = 0.069$ using a 0.05 significance level and an MPSD of 0.015.

A post hoc contrasts analysis using the Tukey HSD test and a MPSD of $d = 0.25$ indicated that the after school tutoring condition ($M = 22.9, SD = 6.1, n = 53$) was significantly and practically different ($p = 0.003, d = 0.66$) than the control condition ($M = 19.2, SD = 5.2, n = 53$). However, the online tutoring condition was not statistically different ($p = 0.15$) from the control condition though it did have a practically significant effect $d = 0.37$. Similarly, there was no statistically significant difference ($p = .46$)

between the after-school tutoring and online tutoring conditions though there was an effect marginally higher than the MPSD ($d = 0.29$).

Second-generation p value

Much like the MPSD in the MESP methodology, researchers will have to think deeply and creatively about how to set the null interval when using the SGPV method. An important distinction, however, is that whereas the MESP uses a cutoff value, the SGPV instead has a null interval which necessarily consists of a lower and upper bound. The value of both bounds must be thoughtfully considered. It is easy and natural to select symmetrical bounds, as was done with the two-group example. However, symmetry is not necessary for the SGPV, and asymmetrical bounds should be used when appropriate. Further, besides having a strong rationale for the values of the null interval bounds, the scale of the standardized effect size measure must also be considered since some, such as Cohen's d , allow for negative effects while others, such as η_2 , do not.

Assuming the same rationales discussed in selecting an MPSD for the omnibus test, researchers might use either the average effect ($\eta_2 = 0.015$), the lower effect bound ($\eta_2 = 0.007$), or the upper effect bound ($\eta_2 = 0.028$) for the upper bound of the null interval. The lower bound will be set to 0 because the η_2 scale is 0 to 1, and there is no reason to consider small effects to be practically significant with this data. Also, since contrasts are being conducted, a separate null interval must be selected for these as well. A null range of - 0.25 to 0.25 based on an average effect of $d = 0.25$ is used, assuming the reasoning expressed in the two-group example. Further, it should be noted that the SGPV method does not require any adjustments for multiple comparisons.

As with the MESP, each null range option will be demonstrated to show how researcher choices might affect the reported outcomes using the SGPV. A null interval constructed using the average effect ($\eta_2 = 0.015$) yields a p_δ of 0.22, which though inconclusive, can be interpreted as trending toward supporting the H_A . Based on this, one should expect that a more narrow null interval using the lower effect bound ($\eta_2 = 0.007$) will provide somewhat stronger support for the H_A . Indeed, this is what is observed, but the support is much stronger with a p_δ of 0 and a delta gap of 0.37. The delta gap statistic indicates the distance between intervals in delta units and is useful for comparing this result with other results that also have a p_δ of 0. It should not be considered as an indicator of strength of effect. Lastly, a null interval using the upper effect bound ($\eta_2 = 0.028$) is inclusive, with a p_δ of 0.35. Unlike with the three-group MESP example, researcher decisions of how to set the null interval can drastically affect the reported outcomes.

A sample write-up is provided below using an average effect null interval of $\eta_2 = [0, 0.015]$ for the ANOVA and $d = [-.25, .25]$ for the Tukey HSD test.

A one-way ANOVA was conducted to investigate the effects of ACT preparation programs on mean ACT composite scores. Literature suggests that the effects of retesting or preparation on ACT composite scores range from 1 – 2 composite score points with an average standardized effect of $d = 0.25$. Using the average effect and transforming it to an ANOVA effect size coefficient results in an η_2 of 0.015. The null interval of $\eta^2 = [0, 0.015]$ was used for the analysis.

The results of the analysis of variance evidence are inconclusive (η_2 95% CI [0.008, 0.15], $p_\delta = 0.22$). However, the results can be considered to provide some evidence

trending toward supporting meaningful effects of the treatment. This can be said because the overlap of data-supported hypotheses with null hypotheses though not insignificant (22%), is still not so large as to rule out meaningful effects, especially in the context of an omnibus test. Post hoc contrasts were conducted using the Tukey HSD and a null bound of Cohen's d $[-.25, .25]$. The test indicates that there is evidence tending toward meaningful effects of the after-school tutoring condition relative to the control condition (Cohen's d 95% CI $[0.2, 1.1]$, $SGPV = 0.07$). The results were inconclusive for the online tutoring condition and the control condition (Cohen's d 95% CI $[-0.1, 0.86]$, $SGPV = 0.39$). The results for the after-school to online tutoring condition were also inconclusive (Cohen's d 95% CI $[-0.16, 0.73]$, $SGPV = 0.47$).

Interval Estimates

The execution of the interval estimates method using the three-group data does not differ significantly from that of the two-group data. Compatibility intervals are discussed, paying close attention to both the upper and the lower bounds as well as the point estimate. The only meaningful difference is that method is applied to both the omnibus test and a post hoc test of contrasts. Regardless of the context, Intervals are interpreted similarly, as can be seen in the example below.

A one-way ANOVA was conducted to investigate the effects of ACT preparation programs on mean ACT composite scores. The analysis of variance provides evidence ($F(2,156) = 5.78$) that a nil effect of ACT preparation on ACT scores is not very compatible with the sample data given the model and model assumptions. The standardized effect estimate most compatible with the sample data is $\eta^2 = 0.07$. However, other highly compatible effects range from 0.008, a negligible

effect, to 0.15, indicating the proportion of variance in score means attributable to group assignment range from .8% to 15%.

Post hoc comparisons of the after-school school tutoring condition ($M = 22.9$, $SD = 6.1$, $n = 53$) and the control condition ($M = 19.2$, $SD = 5.2$, $n = 53$) indicate that an effect estimate most compatible with the data of 3.71 composite score points, $d = .66$ standardized. Other highly compatible effects within a 95% compatibility interval range from 1.12 – 6.30 composite score points, Cohen's d 95% CI [0.2, 1.1]. This provides evidence that the effect parameter is consistent with, and potentially much higher than, the expected range of 1-2 composite score points.

A comparison of the online tutoring condition and the control condition renders an effect estimate most compatible with the data of 2.06 and a CI 95% [- 0.53 – 4.65]. The standardized effect estimate is $d = .37$ with a compatibility interval of Cohen's d 95% CI [-0.1, 0.86]. The compatible effects are inconclusive evidence of the type and magnitude of the effect. The lower bound indicates a possible decrease in the score, while the higher bound is more than double the highest expected gain based upon the literature.

Comparison of the after-school tutoring and online tutoring conditions indicated an effect most compatible with the data of 1.66 ($d = .29$). Other highly compatible effects range from -0.93 to 4.24, Cohen's d 95% CI [-0.16, 0.73]. The CI provides inconclusive evidence with compatible effects of nearly one composite score point increase of the online condition over the after-school condition to a more than a four-point increase in favor of the after-school condition.

Observation Oriented Modeling

The ordinal pattern analysis is used when there are two or more categorical groupings and an outcome measure that is at least ordinal and can be used with data that under NHST would be analyzed by a t test or an ANOVA. The three-group data analysis is similar to the two-group analysis, except the ordinal relation of three groups will need to be specified instead of two. Here it can reasonably be expected that the students in the control group will have the lowest scores while those in the online tutoring to be lower than those in the after-school tutoring condition but higher than the control group. The ordinal analysis will render results of the overall accuracy of our prediction with a PCC and c value for all pairs of observations, which functions somewhat like an omnibus test in an ANOVA. The ordinal analysis then gives the results for each pair of conditions, much like the contrasts in an ANOVA analysis. A brief description of an analysis report is below.

For the three-group data, our model predicts that the ordinal relation of individual scores is that those in the online tutoring condition will be higher than the control condition while those in the after-school tutoring condition will be higher than the control and online tutoring condition. The PCC of the ordinal pattern analysis is 60.51, indicating that our prediction is correct for 60.51% of the pairwise comparisons. The c value is less than 0.001, a rare result given the data. The comparison of individuals in the control and online conditions had a PCC of 57.67 and a c value of 0.06. The control and after-school condition results were a PCC of 66.18 and c value = 0. The pairwise comparison results of the online to the after-school condition had a PCC of 57.67 and a c value of 0.04.

The results provide evidence that our model, though simple, is able to predict the ordinal relation of student scores with over 60% accuracy. This is especially true of students who are in the after-school condition when compared to those in the control condition, which was accurate for over 66% of comparisons. The model did less well when comparing the online condition to the control or after-school conditions.

Two by Two Between Subjects Design

The two-by-two between subjects design offers an opportunity to demonstrate the methods in a more complicated context. The design allows for the presence of an interaction effect as well as the potential for two main effects if no interaction is present.

Minimum Effect Plus p value

In the t test example, researchers could have in mind a precise raw effect in terms of composite score points that could be considered meaningful. However, this type of precision using the raw effect is not always possible, especially with more complex analyses. As an example, we will consider the two-by-two factorial ANOVA analysis. This analysis forces the researcher to consider the main effects for two independent variables and an interaction effect of those variables, neither of which can easily be thought of in terms of simple raw effects like a difference in composite scores. A standardized effect size can be used instead; however, this can present some difficulties with determining the MPSD. Researchers will have to rely on past research, or if that is lacking, general guidelines of effect and professional judgment to determine the MPSD. For this analysis, a small to moderate effect of $\eta^2 = .035$ is used, relying on Cohen's guidelines. This effect size was deemed appropriate given the lack of information

regarding similar studies in the existing literature. Further setting the MPSD too low would effectively change the MESP into an NHST since the p value would be the sole determining factor of “significance.” Whatever amount is decided upon should be supported and discussed. If contrast analyses are conducted, the MPSD can be set much like it is in the t test example. Turning to the factorial ANOVA example:

A between groups factorial ANOVA using an $\alpha = 0.05$ and a minimum practically significant difference of $\eta_2 = .035$ was conducted to examine how type of study and word type relate to performance on a vocabulary test. Since there is little existing research on word type and study type to draw from to indicate a meaningful omnibus effect size and given the sample size $n = 128$ and strength of power (.8), Cohen’s guidelines (small = .01, moderate = .06, large = .14) for η_2 were used to select a small/moderate effect size for the MPSD. The interaction of word type and type of study ($F(1,124) = 4.2$, $p = 0.04$, $MSe = 8.64$, $\eta_2 = .032$) was not statistically and meaningfully significant as the effect was less than MPSD of $\eta_2 = .035$. There was a main effect of type of study ($F(1,124) = 17.8$, $p = 0.00005$, $\eta_2 = .13$), with better overall performance following study using definitions ($M = 12.6$, $SD = 3.2$) than using passages ($M = 10.5$, $SD = 3.1$). There was also a main effect of word type, $F(1,124) = 13.9$, $p = 0.0003$, $\eta_2 = .1$) with better overall performance with familiar ($M = 12.5$, $SD = 3.2$) than with unfamiliar ($M = 10.6$, $SD = 3.1$) words.

Second-generation p value

For the two-by-two between-subject design, we will use the same effect size ($\eta_2 = 0.035$) to construct the null interval as was used for the MPSD in the MESP

method. This will help further demonstrate how even though the same value might be used, it can cause quite different results depending on the methodology. Recall that the MESP found no meaningfully significant interaction effect but did find both main effects to be statistically and meaningfully significant. The SGPV, by contrast, using a null interval of η_2 [0, 0.035] only found the effect of type of study to have strong evidence of scientific meaningfulness ($F(1,124) = 17.8$, η_2 95% CI [0.036, 0.24], SGPV = 0.00). The interaction effect is the most inconclusive ($F(1,124) = 4.2$, η_2 95% CI [0, 0.12], SGPV = 0.5) and word type ($F(1,124) = 17.8$, η_2 95% CI [0.02, 0.21], SGPV = 0.17) also lacking strong evidence of meaningful effects. The difference in outcomes of the two methods results from comparing the null interval to the interval estimate in SGPV instead of using a cut-off value as in MESP. The overlap of the intervals is considered in the SGPV method to indicate uncertainty, but intervals are ignored in the MESP method just as they are in NHST and any method that uses cut-off values. Results derived from null intervals will behave quite differently from those of a cut-off value, even when using a similar metric. This important distinction should be considered when selecting either the MPSP or constructing a null interval.

A write-up of the results is below.

A between-group factorial ANOVA was conducted to examine how the type of study and word type relate to performance on a vocabulary test. Since there is little existing research on word type and study type to draw from to determine a meaningful null interval for interaction or main effects, Cohen's guidelines (small = .01, moderate = .06, large = .14) for η_2 were used to select a small/moderate effect of 0.035 for the upper bound with a lower bound of zero. A

null interval, $\eta^2 = [0, 0.035]$ with a small/moderate upper bound seems likely to include most practically negligible effects. A more refined null interval should be selected in future studies as the body of literature around this area becomes more robust.

The interaction of word type and type of study ($F(1,124) = 4.2$, η^2 95% CI $[0, 0.12]$, $SPGV = 0.5$) was inconclusive with the $SPGV$ of 0.5 indicating that 50% of the data-supported hypotheses are also null hypotheses. Strong evidence was found to support the scientific meaningfulness of the main effect of type of study ($F(1,124) = 17.8$, η^2 95% CI $[0.036, 0.23]$, $SGPV = 0.00$). The main effect of word type lacks strong evidence of an effect ($F(1,124) = 13.9$, η^2 95% CI $[0.02, 0.2]$, $SGPV = 0.17$), but with and $SGPV$ of 0.17 can be considered to be trending toward the alternative hypothesis. Such evidence should not be dismissed especially given the exploratory nature of the present study.

Observation Oriented Modeling

The ordinal pattern analysis can also be used with the two-by-two between-subjects data allowing for an analysis of interaction and main effects. In our simulated data set, one group of students was given words with which they were familiar, while a second group was given unfamiliar words. Further, students were then given either the definitions of the words or passages containing the words as study material. In the ordinal analysis, the ordinal relationship of the outcome measure, words correctly defined, must be determined for the familiarity condition, the study condition, and the study condition under each level of familiarity. For example, researchers could reasonably expect students assigned familiar words to perform better at the spelling test than those assigned

unfamiliar words. They also might expect students assigned the definitions as study aids to perform better compared to those given passages to read. Both of these correspond to the main effects in an ANOVA. Lastly, researchers could predict that those students provided with familiar words would perform no better regardless of the study condition they were assigned, but those given unfamiliar words would have higher scores if they were given definitions as a study aid. This is similar to the interaction effect in ANOVA. See the results report below.

The results of the ordinal analysis for the word type condition resulted in a PCC of 59.96 and a c value of 0.001 with a PCC range of 29.13 – 60.94. A very rare result even though the accuracy of the model at predicting the ordinal relation of student scores was only 59.96%. The analysis of the study condition resulted in a PCC of 57.08, less than the accuracy for the word type condition. The randomization test resulted in a c value of 0.01 and a PCC range of 30.08 – 58.84. An analysis of the interaction of word type and study condition was also conducted. The prediction that of students given unfamiliar words, those given definitions would score higher than those given paragraphs was not supported by the analysis. The PCC was 48.24 with a c value of 0.24. The range of PCC scores from the randomization test was 19.53 – 65.82. Our prediction that students given familiar words would see no difference in scores regardless of study type was also not supported by the analysis having a PCC of 12.11 and c value of 0.80.

Further evaluation of patterns is possible with a post hoc analysis, which uses Binary Procrustes Rotation (BPR), a modified form of Procrustes rotation first proposed by Green (1952). It attempts to conform the deep structure units of one set of

observations to the deep structure units of a second set of observations. Deep structures are matrices of zeros and ones that can be manipulated to identify patterns in the data. These structures are obtained by translating data into binary code similar to dummy or effect coding. For the two-by-two between subjects design, researchers may want to investigate whether there are any patterns that the analysis can detect related to the interaction of word type with study type. The analysis will attempt to conform (or classify) the outcome measure to the four groupings (unfamiliar/passage, familiar/passage, unfamiliar/definition, familiar/definition) created from the word type and study type interactions. Looking at the results, the BPR was able to classify 41% of the observations, a PCC of 41.41, with a c value of 0.19. The results then are not that uncommon given the data.

The BPR, however, might still offer some help in understanding the data. A graphic visual called a multi-Unit Frequency Histogram shows the distribution of observations in the four groups and is color-coded to indicate those that the BPR was able to classify. One can observe in the example found in Figure 3.2 that the distributions of observations in the four categories are largely the same and overlapping, which explains why it was difficult for the BPR to identify any patterns in the data. It does, however, appear that the distribution of both “familiar” groups is skewed in favor of the higher scores, with the familiar/definition having higher scoring students than even the familiar/passage group. Familiarity seems to be a stronger predictor of higher scores than the type of study, but some students may have a small advantage when provided definitions as a study tool. This could be a source of further investigation into the characteristics of those students with the hope of constructing a more advanced model.

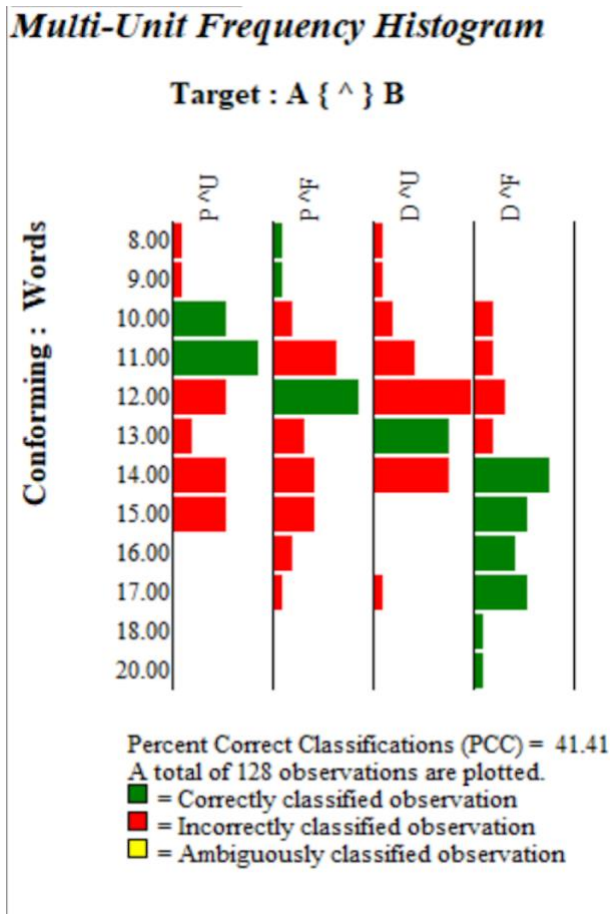


Figure 3.2 Two-by-two Between Subjects Multi-unit Frequency Histogram

Conclusion

The methodologies explored each attempt to address particular flaws in the NHST paradigm. Some methodologies are limited quite intentionally to only addressing a particular aspect of NHST. Others are intended to address multiple flaws in the paradigm. Researchers may wish to use one or more of these methods in conjunction and will find a clear and concise explanation of the methodologies to be useful. That is what this paper has attempted. Now it remains to consider the methodologies comparatively using the critiques discussed previously as a guide to understanding what each method offers.

All of the methods attempt to address the strength of evidence question. The BFB and s value, because they are interpretive aids, transform and restate the p value in different terms, with the intention of better communicating how much evidence is provided by the p statistic. They do not set any criteria for inference or decision making nor alter the functioning of the p value, which can help in conceptualizing how much evidence is provided by a p value but leaves aside the degree of evidence considered sufficient. The $p \leq 0.005$, the MESP, and the SGPV, by contrast, set alternative decision criteria which should render decisions possessing greater strength of evidence. The more demanding significance level $p \leq 0.005$ is a harder hurdle to jump than the typical level set in NHST; however, there is a bit of uncertainty regarding MESP and SGPV since deciding what constitutes trivial effects is researcher determined. One can easily imagine researchers selecting an MPSD so small that the findings differ little from those from conventional NHST. Comparing a null interval with an interval estimate in SPGV makes such a situation less likely, but it is still possible to set a too permissive null bound given the particulars of the research question or past findings. Researchers will need to think deeply about such issues when using either MESP or SGPV. Similar to SGPV, Interval estimation allows for the consideration of the strength of evidence by presenting the range of compatible hypotheses, which can help researchers consider the relative uncertainty of their findings. Unlike SGPV, there is no null bound set for determining which values are meaningful. OOM operates under a different paradigm but nonetheless offers the PCC and the c value statistics for determining the relative strength of a finding.

Except for NHST with a $p \leq 0.005$ cut-off, all of the methods provide some element that either aids in the correct interpretation of p values or replaces the p value

with a more easily interpreted methodology. For interpretational aid methods, the sole purpose is to assist with interpreting the strength of evidence provided by the p value, and both examples do it well. Whether it is the odds of observing the data under the alternative hypothesis to observing it under the null hypothesis (BFB) or the evidence against the null in binary bits (s value), each method provides a more intuitive means of understanding the evidence provided by the p value statistic. This is especially important if the p value is reported as a continuous statistic. Further, such interpretational aids can be used together with other methods such as the $p \leq 0.005$ cut-off or the MESP. The MESP provides perhaps the least improvement since it does not explain the p value nor offer an alternative statistic. Its only contribution is to force the consideration of meaningful effect sizes along with the p statistic. Supplementing MESP with an interpretational aid to assist in understanding the p value reported as a continuous statistic can be easily justified. The SGPV, Interval estimates, and OOM each eschew the use of p -values and offer methods and statistics that are far more intuitive than the p statistic.

A dichotomized decision rule is abandoned by all the methods except MESP and possibly NHST with a $p \leq 0.005$ cut-off. The latter is considered “possible” because Benjamin (2018, 2019) suggests a trichotomized decision framework that could be adhered to but could just as easily become dichotomized by researchers who come to view $p \leq 0.005$ as the threshold for meaningful findings. The other methods avoid using a hard threshold for decisions, instead relying on researchers’ expertise for evaluating the meaning of a result and allowing for uncertainty.

The issues arising from the sample size paradox and the nil-null are addressed by MESP, SGPV, interval estimates, and OOM. The MESP and SGPV both restrict findings

to meaningful effects. In MESP, a dual null is proposed such that even with an extremely small p value, a finding would not be deemed meaningful if it failed to meet the MPSD criteria. The SPGV is even more restrictive since it has an interval null and requires that the null and interval estimate have no overlap to find unequivocal evidence for the null hypothesis only. Interval estimates have no null hypothesis at all, and though the width of the interval will narrow with sample size, it is up to the researcher to determine whether the results are meaningful. OOM, like interval estimates, does not test a null hypothesis. Further, sample size also has nearly no bearing on OOM analysis; even single subject designs can be analyzed with certain procedures. The only caveat is that the c -value statistic is less meaningful with small data since they allow for fewer random permutations.

Only the OOM procedure attempts to address aggregation and the violation of assumptions. OOM avoids the problem of aggregation by focusing on the patterns of individual observations, not on the values of means or variances. All the other methods retain the focus on aggregate statistics in their analyses. Also, since OOM is generally assumption-free and uses descriptive statistics, it avoids many of the fundamental issues associated with NHST. That being said, OOM is not a panacea. As discussed previously, it is an attempt to address the NHST problem by offering a simplified methodology more appropriate to applied research, more appropriate for most data, and more understandable both to applied researchers and readers.

The field of statistics is increasingly moving away from NHST and p values. The social sciences are slowly adapting. The methods discussed in this paper provide a sampling of methodological alternatives which could be suitable, but the question

remains open as to which methodologies should be adopted and how these might affect the research landscape. An answer to the first question is expressed by Wasserstein et al. (2019), who, in laying out the principles of a post p value world, advise researchers to be “statistically thoughtful,” by which they mean having clearly expressed objectives as to whether studies are exploratory or rigidly preplanned and to consider selecting analytical methods, including the use of multiple methods, that best match those objectives. For instance, methods with hard decision rules or, as with MESP or SGPV, that require the existence of prior research for the thoughtful selection of a cut-off or construction of a null interval could be considered a poor match for exploratory research. This type of research might be better served with a discussion of intervals combined with a report of the point estimate with a continuous p value whose strength of evidence is interpreted using an aide such as an s value. A researcher might also use OOM in these situations if the focus of research is more on patterns of individuals than on means and aggregates. One can imagine any number of scenarios. This type of thoughtfulness and openness to a range of methods alone constitute a shift in prevailing research habits.

A further consequence of adopting some of these alternatives is the acceptance of a level of uncertainty and, with it, a limit upon the claims that can be made regarding outcomes. Interval estimates and SGPV, for instance, both consider the range of plausible hypotheses supported by the data. Claims of certainty such as are made with NHST are difficult when discussing multiple plausible data-supported hypotheses. This is also the case with OOM, which makes no claims to inference to a parameter but instead seeks to identify patterns of data at the level of the individual. Further, uncertainty and thoughtfulness are inherent in selecting decision criteria such as the MPSD or the null

interval. The discussion of how these decisions are made can be of great potential benefit in inspiring a wider consideration of the meaningfulness of results.

Conceptual shifts in thoughtfulness and acceptance of uncertainty are less likely to occur with methods that closely adhere to the NHST paradigm. Increasing the difficulty of the decision criteria from 0.05 to 0.005, for instance, would do little on these fronts since it would maintain the NHST paradigm. That is not to say that there would be no effects from such a change as indeed there would be. First, it would be much more difficult to meet the new criteria for “significance.” If the three levels of significance scheme proposed by Benjamin & Berger (2019) were used, much research would be considered “suggestive.” To achieve “significance,” statistical power would need to be increased, which would likely result in studies with much larger sample sizes.

Interpretational aides, by contrast, while supplementing NHST, could be a bit of a trojan horse and work to slowly undermine the NHST paradigm by making clear the paucity of evidence provided by $p = 0.05$. However, this is likely the intent of such methods as Benjamin & Berger (2019) clearly state that the BFB is considered an intermediate method intended to move researchers away from NHST slowly.

It is hoped that this study and others like it will provide researchers with enough familiarity with alternative methods to feel comfortable integrating them into their research programs. The methods discussed here are not exhaustive and are not intended to be. Other very good alternatives and supplements to NHST exist, but whichever methods are used will be a welcome shift away from NHST

CHAPTER 4³

A MONTE CARLO METHODS COMPARISON OF CONVENTIONAL NHST WITH THREE ALTERNATIVE METHODOLOGIES

³ B. D. Rogers. Not yet submitted for publication.

Abstract

A Monte Carlo methods simulation was used to compare NHST with a $p \leq 0.05$ decision rule with three alternative analytical methodologies: NHST with $p \leq 0.005$, second-generation p value (SGPV), and minimum effect size plus p value (MESP). Results indicate that NHST with $p \leq 0.005$ maintained a consistently low Type I error near 0.005 except with a variance ratio of 4:1. There was also, as expected, an accompanying decrease in power with NSHT $p \leq 0.005$ achieving between 12% and 59% of the power observed with NHST $p \leq 0.05$. SGPV showed improvement upon NHST in that it filters out trivial effects while maintaining a low Type I error rate. MESP showed little improvement on NHST as the results precisely mirrored those of NHST with $p \leq .05$, except in certain conditions with a large sample size. Ultimately, the results show that no one method is a panacea, and researchers should be encouraged to use a range of analytical tools.

Null hypothesis statistical testing has been the dominant research and inferential methodology in the social sciences and related fields since the second half of the 20th century (Danziger, 1990; Gigerenzer and Murray, 1987). NHST, despite its ubiquity, has been surrounded by persistent controversy and vocal, though largely unheeded, criticism. As noted by Kline (2013), the assumptions of NHST are far more restrictive than is commonly thought by many researchers and are often not met, which results in potentially biased p values. NHST is not simply a test of the null hypothesis (H_0), though, in applied analysis, it is used as such. Instead, it is a test of a statistical model based on a range of assumptions, the violations of which are reflected in the p value. A p value is the result not only of the probability of the data given the H_0 , but of random

variation and violations of model assumptions, none of which can be differentiated by the p statistic (Amrhein, Trafimow, & Greenland). Null hypothesis statistical tests also have methodological limitations related to sample size sensitivity and the null hypothesis (nil-null) typically used. Researchers also have a broad interpretational misunderstanding of what can be claimed when a null hypothesis is rejected, which manifests as misinterpretations or exaggerations of inferences. Lastly, because of professional pressure to publish and a publication bias in favor of novel statistically significant findings, there is evidence that some researchers have taken to using questionable practices such as data mining, the selective reworking of data, or performing multiple statistical analyses to identify patterns in data that produce statistically significant results.

Over the last decade, two concurrent phenomena have contributed to an increased awareness of the problems with NHST and added urgency to calls for reform of methodological practices. First, there is a broader awareness of the problems with NHST as articles detailing the methodology issues have been published in popular publications and online blogs (Cumming, 2013; Nuzzo, 2014; Siegfried, 2010 & 2014). At the same time, the replication crisis afflicting many scientific fields, especially the social sciences, has been definitively demonstrated and widely reported (Wasserstein & Lazar, 2016). In response, some academic publications have begun to reconsider methodological requirements for publication, discouraging, or in some cases banning, NHST or key features thereof while at the same time encouraging supplemental or alternative methods (International Committee of Medical Journal Editors, 2010; Trafimow, 2015). The suggested methodologies are characterizable as ranging from those that work within the NHST paradigm seeking to improve it to those that abandon NHST altogether. Methods

supplemental to NHST are supposed to be more palatable to applied researchers accustomed to NHST and thus more readily adopted by them. These methods can be characterized as either aiding in the interpretation of p values or offering alternative decision criteria.

Interpretational aids seek to help mitigate the long-standing confusion about the interpretation of p values. Some aids are drawn from Bayesian statistics, such as the Bayes factor bound (Benjamin & Berger, 2019), the false positive risk (Colquoun, 2019), and the analysis of credibility (Mathews, 2019). There are also non-Bayesian suggestions, such as the s value (Greenland, 2019), which derives from information theory. Regardless, shared by each methodology is the mathematical transformation of the p value into a more informative and easily understood statistic. Such methods are proposed to help researchers better understand and interpret the strength of evidence provided by p values. They do not address any of the critiques of the NHST process, such as violation of assumptions, sample size sensitivity, or problems with a nil-null hypothesis.

By contrast, alternative decision criteria methodologies posit different and generally more demanding criteria for making inferences within the NHST paradigm. These are principally a response to the problems of replication, which is often blamed on Type I error and the divergence between statistical and practical significance. The methods rest upon the assumption that since researchers are likely to continue to use hypothesis tests with decision rules, different and more difficult decision rules will mitigate some of the problems with $p \leq .05$ dichotomized decisions. Further, because the decision criterion is altered in various ways, the manner that the tests respond to violations of assumptions and other inferential issues related to power and the null

hypothesis could be changed or improved as well. This will be the focus of this article. However, before looking at methods in depth, I will review of some of the specific problems related to NHST.

Problems with NHST

Violations of Assumptions

Null hypothesis statistical tests require that certain assumptions be met to ensure that p value calculations and inferences are accurate. The p value is the prior conditional probability of observing the obtained data, or data even more inconsistent with the null hypothesis if the null hypothesis is true and other assumptions are met. For many, perhaps most, NHST these assumptions include (a) random sampling, (b) normal distribution of errors, (c) equality of population error variances (homoscedasticity), (d) independence of observations, and (e) sampling and measurement error are the only sources of error (Kline, 2013; Loftus, 1996). Such assumptions are more restrictive than is frequently thought by many researchers and are often not met (Kline, 2013), resulting in potentially biased p values. When p values are lower than they should be because of violations of assumptions, there is a positive bias meaning that the Type I error rate is higher than the stated α level. Conversely, if p values are too high, there is a negative bias resulting in inflated Type II errors. (Kline, 2013).

The problem of assumption violations is exasperated because there is evidence that researchers rarely investigate the plausibility of many of the assumptions of NHST and thus neither adjust their analyses to suit the data nor report the violations in research articles (Greenland et al., 2016). Keselman (1998) reviewed over 400 articles published in 17 prominent journals of psychology from 1994 to 1995, focusing on those with

ANOVA designs to determine whether researchers reported investigating whether their data met the assumptions of the analyses performed. Keselman found that of the 61 instances in which univariate ANOVA was used, 11.4% of articles referenced investigating distributional assumptions, and only 8.1% reported checking equality of variance. Only 4.9% of the papers assessed both distributional and homogeneity assumptions. Of the 79 articles in which MANCOVA was used, only 6.3% assessed the plausibility of distributional assumptions, and none mentioned evaluating variance homogeneity. Forty-eight articles mentioned using ANCOVA, but of these only 4.1% reported assessing distributional assumptions, while 8.3% reported looking into equality of variance. Similarly, in the 226 articles containing repeated measures designs, 15.5% of researchers mentioned investigating distributional assumptions and 0.4% reported on variance assumptions. Osborne (2013) suggested that an overestimation of the robustness to violations of assumptions may partly be responsible for this tendency.

Sensitivity to Sample Size

Small sample sizes often lack sufficient power to detect even strong effects and fail to have p values smaller than the critical value even when there is an effect (i.e., a Type II error is made). Alternatively, when a sample size is large, trivial effects can produce impressively small p values. Levine et al. (2008) provide some accessible examples of how this looks practically. For the relationship $r = .40$ with $n = 20$, a two-tailed significance test at $p = .05$ is not statistically significant; whereas, $r = .07$ with $n = 1000$, a dramatically weaker effect, is statistically significant. An even more stark example is if a two-tailed NHST is conducted using $\alpha = .05$ with an observed effect of exactly $r = .25$, the results are statistically significant when $n = 63$ but not when $n = 61$. P

values do not indicate the magnitude of an effect but rather are a function of effect size and sample size and of other contributors of statistical power such as directionality of the alternative hypothesis, experimental design, the test statistic, measurement reliability, and whether assumptions are met (Kline, 2013).

This sensitivity to sample size has several undesirable consequences. One is the paradox that with increased precision in the form of statistical power due to large samples, there is a greater possibility of a statistically significant finding even when effects are not practically meaningful. Another consequence is that researchers can chase statistical significance by increasing sample sizes, an example of questionable practices called *p* hacking (Simmons et al., 2011). Similarly, the advent of “big data” has led to the increased availability of massive data sets, which presents a new challenge in that studies can be “overpowered.” These massive data sets increase the number of analyses that can be conducted, and with relatively lenient thresholds, such as $p = 0.05$, make statistically significant results much easier to obtain and the risk of false positives much greater (Ioannidis, 2019).

The point or Nil-null

The nil-null is when the null hypothesis (H_0) is that some parameter is precisely zero or that the parameters in some set of parameters are exactly equal. (Cohen, 1994; Meehl, 1978). However, in practice, the differences between means or the observed correlation of any variables, no matter how seemingly unrelated, will never be 0.00 out to the n th decimal place. Randomization does not perfectly balance the effects of all extraneous variables, nor should it be expected that in correlation studies that there will be no uncontrolled third variables (Meehl, 1978). The nil-null hypothesis is nearly always

false, so rejecting it is neither impressive nor informative. It will always be rejected, given enough statistical power (Cohen, 1994). This is more concerning when considered together with the effect sample size has on power and the increasing availability of massive data sets. Given that the nil-null is always false and granted a large enough sample, NHST will always render a “significant” result regardless of the value of the alternative hypothesis creating a divergence of statistical and practical significance (Meehl 1986).

Alternative Methodologies

A wide range of alternative methodologies to NHST exist. Some of these methodologies discourage decisions about the meaningfulness of a finding. This is not a flaw in the methods but a beneficial limitation intended to prevent the thoughtless application of a universal decision rule, such as with $p \leq 0.05$. However, other methods do contain decision rules, so it is important to investigate whether they improve upon NHST. It is this latter class of methods that will be the focus of the present study.

Decrease Decision Threshold to 0.005

The methodological suggestion to reduce the decision threshold for inference from 0.05 to 0.005 is not novel (Greenwald et al. 1996; Johnson 2013) but was most forcefully made in 2018 in a comment article published in the journal *Nature Human Behavior* which listed seventy-two authors and co-authors (Benjamin et al., 2018). Benjamin and Berger, the first and second authors, repeated the suggestion in the 2019 ASA supplemental issue on p values (Benjamin & Berger, 2019). The authors make the case that incremental steps are necessary to move researchers away from the NHST paradigm.

The authors consider the replication crises to be the result of a high number of Type I errors resulting from inflated claims of findings based on the 0.05 significance threshold. Despite broad agreement that the use of dichotomous decision rules is inappropriate (Benjamin et al., 2018; Benjamin & Berger, 2019), increasing the stringency of the threshold for significance from 0.05 to 0.005 is considered to be a marked improvement. To support the selection of 0.005, Benjamin and Berger use Bayesian statistical methods to calculate the probability of a true H_0 at a given p value. Assuming that the null hypothesis and the alternative hypothesis are equally likely (prior odds), for a $p = 0.05$ there is at least a 29% chance that the null hypothesis is true. Using a threshold of 0.005 reduces the probability to 7%. They also present evidence of potentially dramatic decreases in estimated false positive rates. For example, the false positive rate is greater than 33% with prior odds of 1:5 and a p value threshold of 0.05, regardless of the level of statistical power. Reducing the threshold to 0.005 would reduce the minimum false positive rate to 5% (Benjamin et al., 2018).

Minimum Effect Plus p value

The use of effect size measures has long been suggested to correct the potential divergence of practical and statistical significance in NHST (Cohen, 1988; Thompson, 2002; Wilkinson, L., & the Task Force on Statistical Inference, 1999). Suggestions typically have taken the form of reporting an effect size alongside the p value. Huberty (2002) noted that researchers had employed two approaches to using effect sizes with p values: (a) using the p value ($p \leq 0.05$) to determine statistical significance and then the effect size to indicate the magnitude of an effect and (b) considering the p value and effect size jointly for determining if an effect is real. Recently, Goodman et al. (2019)

have proposed a formalization of Huberty's second category and recommend considering the p value and effect size together for statistical inference. Goodman calls this hybrid criterion "minimum effect size plus p value" (MESP). In this scheme, researchers are tasked with selecting a minimum practically significant difference (MPSD) between the null hypothesis of no effect and a limit of practical effect before conducting research. It is assumed that researchers would have sufficient knowledge of past findings and competence in their field of expertise to select a scientifically or practically meaningful MPSD and that it would be selected in good faith. The MPSD is then paired with a conventional p value methodology to construct a hybrid decision rule. The MESP method indicates that to reject a null hypothesis, each of two conditions must be satisfied: (1) p value $\leq \alpha = 0.05$ and (2) the observed effect size \geq MPSD. In a simulation study, Goodman found that the MESP method maintained more consistent true power under the tested conditions than did the conventional p value. MESP true power did weaken in low nominal power cases, though Goodman stated that knowing this, researchers could "respond accordingly." An additional benefit mentioned by Goodman is that the MESP method is practicable without complicated calculations using existing statistical software.

Unlike the mechanical application of $p \leq 0.05$, research context and researcher knowledge are key to selecting an MPSD. Different researchers may not come to the same conclusions regarding selecting the MPSD, even in the same research project with the same data. It is then incumbent upon researchers to make their best case for the value selected.

Second-generation p value

Serlin and Lapsley (1985) suggested using a “good enough” principle in selecting a null hypothesis as an alternative to the nil null or point hypotheses. Researchers would establish a null hypothesis that is an interval of all those values that are not deemed to be practically meaningful. Following this suggestion, Blume et al. (2019) propose replacing p values with a second-generation p value (SGPV) that incorporates an interval null into its computation. Blume’s innovation is devising a statistic that summarizes the test results in light of the “good enough belt” proposed by Selin and Lapsley. The SGPV is a measure of the data-supported hypotheses that are also scientifically or practically null hypotheses. Researchers would construct the interval null hypothesis by specifying in advance a range of effects that they consider to be without practical or scientific importance. Data-supported hypotheses are identified using interval estimates such as confidence intervals though other interval estimates can be used, such as support intervals (Blume et al., 2019; Wagenmakers et al., 2020) or credible intervals (Hespanhol et al., 2019).

Suppose there is a parameter of interest θ . Let $I = [\theta_l, \theta_u]$ be the interval estimate of θ and $|I| = \theta_l - \theta_u$ be the length of the estimate. Let the interval null hypothesis be H_0 and its length be $|H_0|$. The SGPV is p_δ . The SGPV calculation is

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max \left(\frac{|I|}{2|H_0|}, 1 \right)$$

where $I \cap H_0$ is the overlap of the data-based and null intervals.

Figure 4.1 can help conceptualize what the SGPV calculation is doing (reproduced from Blume et al. 2018). The top image visualizes the familiar confidence interval around a point estimate, \hat{H} , and its relationship to a point null hypothesis, H_0 .

The bottom image directly relates to how SGPV functions. Here there is a confidence interval but without a point estimate, with bounds $[CI^-, CI^+]$. Further, a null hypothesis interval with bounds $[H_0^-, H_0^+]$ replaces the point null hypothesis. The overlap of the interval estimate and the interval null is the essence of the SGPV.

When the interval estimate is fully contained within the null interval, the data support only null hypotheses. When the interval estimate and null set do not overlap, the data are considered to be incompatible with the null hypothesis. When the interval null and interval estimate partially overlap, the data are considered to be inconclusive.

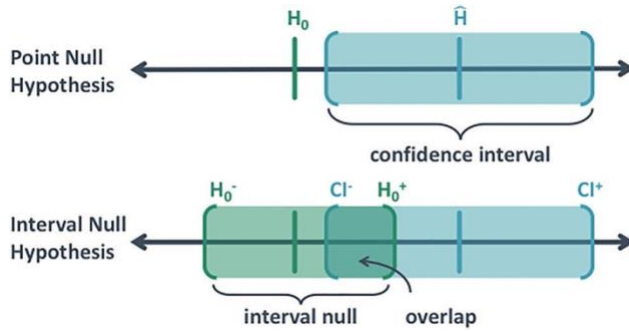


Figure 4.1 Illustration of point null hypothesis and interval null with interval estimate

Each of the above scenarios assumes that the width of $|I| \leq 2 |H_0|$. In instances when $|I| > 2 |H_0|$, the interval estimate is very long and will often extend to either side of the null interval. The SGPV tends to be small when this occurs and does not correctly reflect the inconclusive nature of the data. In these cases, an alternate

calculation, $\frac{1}{2} \frac{|I \cap H_0|}{|H_0|}$, is used. Blume (2018) indicates that while many factors can result in an analysis that requires the correction term, it will typically only be needed in severely

underpowered studies. There is a helpful R package (`sgpv`) designed to aid researchers in conducting the SGPV analysis that automates this process (Welty et al., 2020).

Like p values, the SGPV values range from 0 to 1. An SGPV of 1 indicates that the data only support the null hypotheses or trivially null effects. An SGPV of 0 indicates that the data are incompatible with any of the null hypotheses or, stated alternately, that the data only support meaningful effects. An SGPV between 0 and 1 is considered inconclusive to varying degrees. An SGPV of 0.5 is the most inconclusive, with the amount of inconclusiveness decreasing as SGPV nears either extreme. For example, when $p_\delta \approx 0.2$ the data could be interpreted as trending toward supporting the alternative hypothesis, or when $p_\delta \approx 0.1$, the data could be said to be suggestive of a meaningful effect but not definitively. Blume, however, states that while descriptors of SGPV magnitude might be helpful as communicators of results, they are not essential since the ending states or the SGPV are well defined.

It is also important to consider that two studies with equal SGPVs do not necessarily represent equal amounts of statistical evidence. For instance, two studies, each with $p_\delta = 0$, can have very different distances between their respective interval estimates and null intervals. It can be helpful to have a way of ranking studies with $p_\delta = 0$ by their strength of evidence. Blume (2018) proposes using the delta-gap, which is the distance between intervals in δ units, where $\delta = \frac{1}{2}|H_0|$. The delta gap should always be reported along with the $p_\delta = 0$. In addition, because the SGPV is a summary statistic designed to communicate information cogently, Blume also suggests that for scientific discussion and policy decisions, detailed descriptions of the findings should be provided

in the form of an interval estimate of effect size and noting its proximity to the composite null hypothesis.

The null interval is the most important element of the SGPV and must be chosen with care and intention. Like with the selection of the MPSD, researchers must clearly indicate the values of the null interval and provide a rationale for their choice. Such information can be positioned as part of the method section of the study report or stated before the delivery of the results. In addition to past findings and subject matter expertise, Blume recommends that researchers also consider measurement error, the gravity of the findings, and research community standards when selecting the null interval.

Method

A Monte Carlo methods simulation was used to compare NHST with a $p \leq .05$ decision rule with the alternative methods profiled in the literature review: NHST with a $p \leq .005$, SGPV, and MESP. Monte Carlo methods are experiments that use simulated random numbers to estimate some functions of a probability distribution. Among the principal benefits of Monte Carlo methods is the ability to perform multiple iterations limited only by computing power, which helps mitigate sampling error. Further, and perhaps more important, this method allows one to control the number and values of random/independent variables, thus allowing for the control of elements that would otherwise be random. The independent variables used in this simulation were linear model, sample size, effect size, type of population distribution, and variance ratio. Other potential variables such as random selection and independence of observations were considered since they are mentioned in the literature critical of NHST. However, I made

the decision to limit the number of variables in order to control the size of the simulation, chiefly due to computing power limitations.

The experimental design, when assuming the normal distribution of residuals, is a completely crossed 4x3x3x3 resulting in 108 combinations. The violation of homogeneity reflected in the variance ratio variable could not be paired with distributional violations; thus, a 4x3x3x2 design was used with the skewed and bimodal distributions resulting in 72 combinations. A total of 180 combinations were generated for each of the four methods, which were iterated 10,000 times.

Independent Variables

Linear Model

Three common experimental designs, henceforth called “linear models” to distinguish them from the experimental design of this simulation experiment, typically used in NHST analysis were selected: two-group, three-group, and two by two between-subjects models. These linear models are widely used in the social sciences, and each has an analytical element lacking in each of the others that makes them somewhat representational of other designs. The two-group model allows for the direct comparison of means; the three-group model has an omnibus test for equality of means; the two by two between-group model allows for an interaction effect.

The data were simulated separately for each linear model condition. The two-group used the linear equation

$$y_{ik} = \mu + \alpha_k + \varepsilon_{ik}$$

where μ is the population mean, α_k is the group effect, and ε_{ik} is the error term. Some elements were held constant and are detailed here. Variable elements are discussed in

their own subsections. Population means were held constant at 100 with no group effect in the control group and variable effects in the treatment group. The residual distribution was simulated, having a mean of 0. A standard deviation of 15 was used for the control group and was varied for the treatment group.

The three-group model used the same linear equation. Two of the groups were simulated with a population mean of 100 with no group effect and a residual distribution with a mean of 0 and a standard deviation of 15. The third group, henceforth called the comparison group, was manipulated to manifest the experimental conditions related to effect size and variance ratio.

The two by two between-group model was simulated with the equation

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

where μ is the population mean, α_j is the first main effect, β_k is the second main effect, γ_{jk} is the interaction effect, and ε_{ijk} is error. A population mean of 100 was again used with no main effects at α_j and β_k . Because the interaction effect is the only effect of interest, α_j and β_k were held constant at 0. The interaction effect coefficient was manipulated to correspond to the selected η^2 . The residual distribution was simulated with a mean of 0 and a standard deviation of 15 except when varied as part of the variance ratio condition.

Effect Size

Effect size was varied to investigate both Type I error, the detection of a false effect, and power under the various conditions. For Type I error, the simulations were conducted under the null hypothesis by setting the effect to nil. To investigate power, Cohen's rules of thumb for small, medium, and large effects were used as a guide: $d = .2$,

.5, .8 and $\eta^2 = .001, .06, .14$. Cohen's recommendations were not used because of any authoritative value but simply for convenience and familiarity. The effects for each of the three linear models are displayed in Table 4.1. The effects, α_k , corresponding to Cohen's small, moderate, and large effect sizes are shown for the two-group and three-group linear models. The interaction effect, γ_{jk} , is also displayed for the small, moderate, and large effect sizes. The effect γ_{jk} is the only effect of interest for the 2x2 between-group linear model and is thus the only effect shown. Other effects for this model were constrained to zero.

Table 4.1 Linear Model Effect Parameters

		α_k/γ_{jk} (small Effect)	α_k/γ_{jk} (moderate Effect)	α_k/γ_{jk} (large Effect)
Two-group model	Control Group	-1.5	-3.75	-6
	Treatment Group	+1.5	+3.75	+6
Three- group model	Control Group	-1.5	-3.75	-6
	Treatment Group 1	-1.5	-3.75	-6
	Treatment Group 2	+1.5	+3.75	+6
2x2 Between- group model	Factor A ₁ / Factor B ₁	+1.5	+3.8	+6.05
	Factor A ₁ / Factor B ₂	-1.5	-3.8	-6.05
	Factor A ₂ / Factor B ₁	-1.5	+3.8	-6.05
	Factor A ₂ / Factor B ₂	+1.5	-3.8	+6.05

Sample Size

Three sample sizes were used that correspond to small, medium, and large samples (20, 50, 500). Sizes were selected, on the one hand, to reflect reasonable sample sizes in the social sciences and, on the other hand, criticisms related to statistical power. A sample of 20 is often viewed as a minimum sample needed for statistical tests. An average sample is approximately 50, and 500 is generally considered a very large sample. These amounts were chosen to investigate the sample size effect on each of the methods.

Population Distribution

The population distributions used were normal, skewed, and bimodal. These distributions are drawn from the literature, which indicates that distributions in psychological and educational research are typically non-normal, frequently multimodal, or skewed (Micceri, 1989). The skewed distribution used had a population skew of 1 and a standard deviation of 15. The bimodal distribution was sampled from a population with modes of 87 and 113 with a standard deviation of 15.

Variance Ratios

Variance ratios of 1:1, 2:1, and 4:1 were selected. The variance ratios are drawn from the surveys of Keselman (1998) and Erceg-Hurn & Mirosevich (2008), which both found that published one-way ANOVA research in psychological and educational journals have average variance ratios between 2:1 and 4:1.

Outcome Variables

The data were analyzed using a three-step process. First, the appropriate statistical test for each of the three linear models was conducted. Because the methods under investigation are layered upon an NHST framework, tests typical of the NHST paradigm

were used: a t -test with the two-group model, a one-way ANOVA with the three-group model, and a two-way ANOVA with the two by two between-group model. The results of each test were then interpreted using each of the four methods, with each method rendering a decision about the presence of an effect. The decision criteria of the various methods are displayed in Table 4.2.

Table 4.2 Decision Criteria by Method

Method	Researcher Determined	Decision Criteria
Conventional NHST	-	$p \leq 0.05$ (reject H_0)
More stringent NHST	-	$p \leq 0.005$ (reject H_0)
SGPV	Null bound: $d = (.1, .35)$; $\eta^2 = (.005, .035)$	$p_\delta = 0$ $0 < p_\delta < 1$ (inconclusive)
MESP	Minimum effect (MPSD): $d = (.1, .35)$; $\eta^2 = (.005, .035)$	$p \leq .05$ & effect \geq MPSD

The conventional NHST framework typically uses the $p \leq 0.05$ decision rule to reject the H_0 . The more stringent NHST uses the decision rule of $p \leq 0.005$. Each of these is relatively uncomplicated and standardized. This is not the case for both the SGPV and the MESP, which require the selection of meaningful effect sizes. In an experimental research context, researchers will draw upon past findings and their knowledge and expertise to determine which effects are trivial and which are not. Neither of those conditions applies to simulated data. For demonstration purposes, two effect size amounts will be used to show how researchers' decisions can affect outcomes. Values were chosen drawing from Cohen's guidelines for effects sizes d and η^2 . Specifically, the effect size values selected are the midpoints between no effect and Cohen's small effect size and the midpoint between Cohen's small and medium effect size. Though arbitrary,

these values are reasonable since researchers are probably more likely to consider effects in the very small to moderately small range to be trivial.

The outcome measures for comparing the methods are Type I error and statistical power. Type I error is the false identification of an effect when one is not present at the population level. Type I error was calculated by summing the times each method identified an effect when the population effect was set to zero and dividing that by the number of iterations. Statistical power is an indicator of the ability of a test to detect an effect when one is present. Real statistical power was similarly calculated by summing the number of times each method detected an effect when an effect was present at the population level and dividing by the number of iterations.

Software

The generation of simulated data and all calculations were accomplished using R statistical software version 4.0.2 (R Core Team, 2020) with RStudio version 1.3.1093 and R packages “broom” (Robinson, Hayes, and Couch, 2021), “car” (Fox & Weisberg, 2019), “effectsize” (Ben-Shachar & Makowski, 2020), “fGarch” (Wuertz et al., 2020), “MonteCarlo” (Leschinski, 2019), “sgpv” (Welty et al., 2020), “tidyverse” (Wickham et al., 2019), and “psych” (Revelle, 2021).

Data Generation

Residuals were generated using three different random number generators corresponding to the distributional requirements of the population distribution condition. The base R (R Core Team, 2020) function “rnorm” was used for normal distributions. The function “rsnorm” found in the “fGarch” (Wuertz et al., 2020) package was used for skewed distributions. A proprietary function was written to generate the bimodal

distribution that relies in part on the base R “rlnorm” function to generate random numbers. This function is provided in Appendix A

Sample Size

Sample size in the context of a Monte Carlo simulation is the number of iterations or repetitions of the simulation. The required sample size was estimated using the following formula:

$$n = \frac{(1.96 \times \sigma)^2}{E^2}$$

Where n is the required sample size, σ is the population standard deviation, and E is the preferred margin of error around a given mean.

The population standard deviation was estimated by running the simulation 200 times and using the derived sample standard deviation which was calculated at approximately 0.03. The desired E selected was less than .01. The minimum sample size was calculated using an E of 0.01, 0.0075, and 0.005 with the intention of using the largest sample that could be run given the limitations of computing power. The resulting sample sizes rounded to the nearest thousand were 6,000; 10,000; and 24,000. A sample size of 10,000 was the largest that could reasonably be run and thus was selected.

Analysis

Outcomes were analyzed using three different methods. Confidence intervals were calculated for the Type I error and power. Estimates and confidence intervals are displayed in tables organized by experimental condition. Further, since the purpose of this study is to compare alternative methodologies to standard NHST, ratios of the Type I error and power rates achieved by the alternative methods relative to those calculated for NHST were computed. These ratios are displayed in figures sorted by experimental

condition. Lastly, the power outcome variable indicates the rate at which an effect was detected when there actually is a population effect. It does not indicate the accuracy of the effect estimate. To aid with understanding the accuracy with which the methods detected effects, figures of the distributions of effect estimates were generated for each method together with an overlapping distribution of detected effect estimates for conventional NHST.

Results

Type I Error

Table 4.3 shows the false detection rates or Type I error rates for each method under normality conditions. The maximum error rates for both conventional NHST with $p \leq 0.05$ and NHST with $p \leq 0.005$ held consistently near their respective α levels, regardless of linear model. Error rates for both SGPV methods show a pattern of decreasing values with an increase of sample size. MESP methods match exactly the values of conventional NHST ($p \leq 0.05$) except for MESP 2 and a large sample size, $n = 500$.

Table 4.3 Type I Error Under Normality Condition

Linear model	<i>N</i>	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	20	0.0487	0.0050	0.0309	0.0098	0.0487	0.0487
	50	0.0472	0.0048	0.0219	0.0020	0.0472	0.0472
	500	0.0481	0.0048	0.0017	0.00	0.0481	0.0001
Three-group	20	0.0496	0.0051	0.0224	0.0118	0.0496	0.0496
	50	0.0441	0.0028	0.0152	0.0018	0.0441	0.0441
	500	0.0529	0.0039	0.0002	0.0000	0.0529	0.0000
2x2 between-group	20	0.0489	0.0039	0.0195	0.0060	0.0489	0.0489
	50	0.0514	0.0053	0.0131	0.0013	0.0514	0.0514
	500	0.0497	0.0053	0.0005	0.0000	0.0497	0.0000

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

Figure 4.2 displays the Type I error rates by linear model and method as a bar plot. Confidence interval bounds are represented by the error bands at the tip of each

respective bar. The range of the confidence bounds do not exceed 0.009. Clustering by methods helps display the influence of sample size on the Type I error of each method. As was seen in the raw numbers, the two NHST based methods held closely to their prescribed α levels regardless of sample size with little variation. The SGPV methods, however, had a decrease in Type I error with an increase in sample size. Also, Type I errors were lessened when a wider null bound was used, as can be seen when comparing SGPV1 to the more stringent SGPV2. Notably, the Type I error rates for even the more permissive SGPV1 are less than those of NHST ($p \leq 0.05$) regardless of sample size.

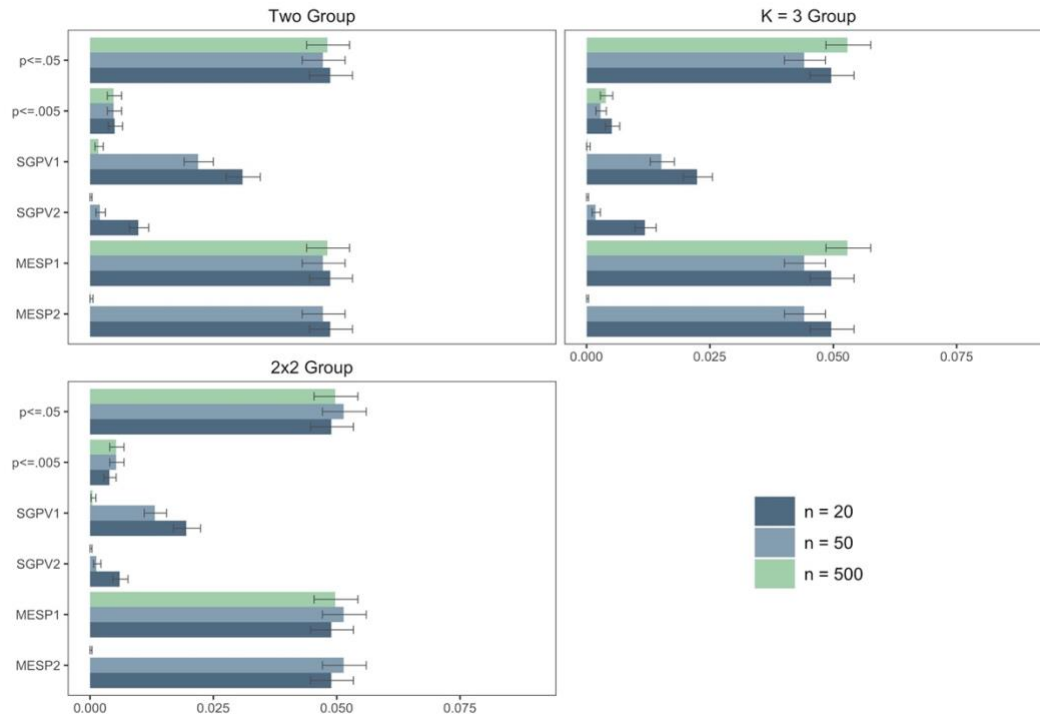


Figure 4.2 Type I Error and CI by Linear Model Under Normality Condition

While the exact calculated rates are informative, just as important is how each method compares with NHST ($p \leq 0.05$). Figure 4.3 displays the ratio of Type I error rate of each method relative to that of NHST ($p \leq 0.05$). The MESP 1 ($d = .1/\eta^2 = .005$) tracks exactly with the $p \leq 0.05$, indicating that the inclusion of a small effect size cut-off

has no influence on Type I error at any of the tested sample sizes or models. Interestingly, the MESP 2 ($d = .35/\eta^2 = .035$) sample sizes of 20 and 50 also mirror the $p \leq 0.05$ results for all models, but the Type I error ratio is exactly or nearly zero at a sample size of 500 for all linear models. The MESP 2 result indicates that even with a small/moderate effect cut-off, the MESP does not affect Type I error rates for $n = 20$ or 50 sample sizes. It is only with a large sample, $n = 500$, that MESP 2 shows any difference in false detection over NHST $p \leq 0.05$. The improvement is because the detected false effects were less than the MESP trivial effect cut-off of $d = .35/\eta^2 = .035$.

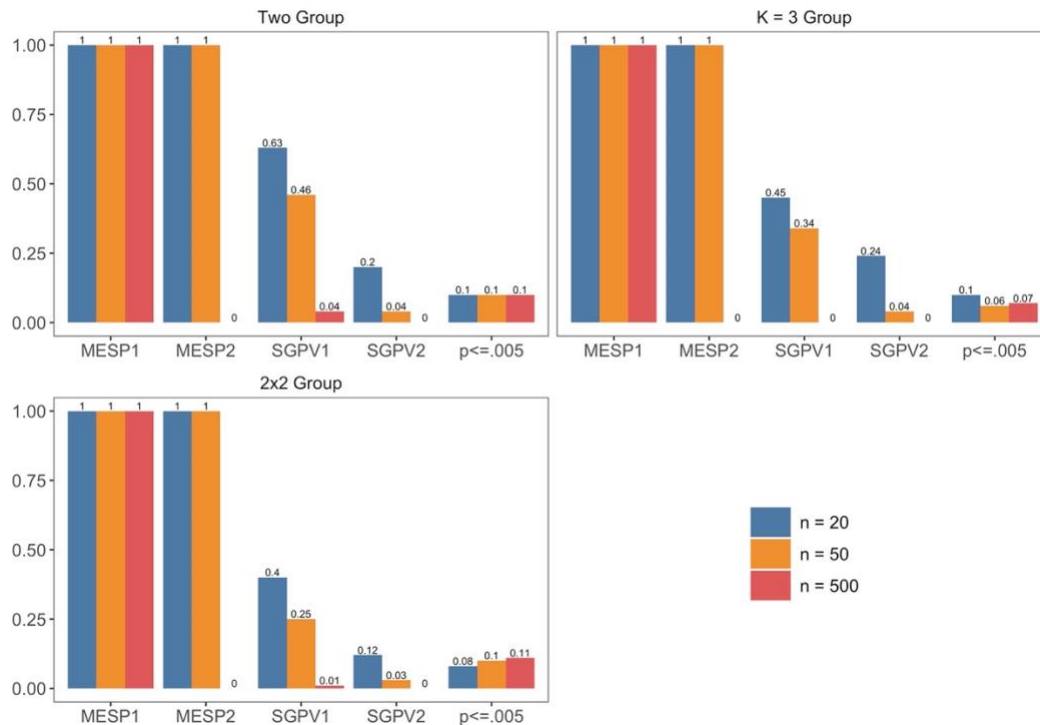


Figure 4.3 Type I Error Ratios by Linear Model Under Normality Condition

The SGPV 1 and SGPV 2 each had lower Type I error rates than the NHST ($p \leq 0.05$), with the two-group condition having higher rates than the more similar three-group and 2x2 group conditions. The SGPV 1 predictably had higher Type I error than

the more restrictive SGPV 2. Notably, for both SGPV methods, there is a proportionate decrease in false detection as the sample size increases. A similar trend was observed in MESP2 but only with a sample size of $n = 500$. This trend is consistent regardless of the size of the null band selected. NHST with $p \leq 0.005$ held a reasonably constant rate with ratios ranging from 0.06 to 0.1 relative to conventional NHST. This is expected as both NHST methods maintained Type I error rates consistent with their respective α levels.

Table 4.4 Type I Error Under Violations of Homogeneity

Linear Model	Var. ratio	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	2:1	20	0.0484	0.0048	0.0322	0.0114	0.0484	0.0484
		50	0.0486	0.0055	0.0231	0.0024	0.0486	0.0486
		500	0.0471	0.0054	0.0023	0.0000	0.0471	0.0000
	4:1	20	0.0542	0.0047	0.0339	0.0119	0.0542	0.0542
		50	0.0471	0.0045	0.0209	0.0031	0.0471	0.0471
		500	0.0497	0.0047	0.0021	0.0000	0.0497	0.0000
Three-group	2:1	20	0.0598	0.0100	0.0305	0.0173	0.0598	0.0598
		50	0.0581	0.0087	0.0225	0.0053	0.0581	0.0581
		500	0.0565	0.0102	0.0028	0.0000	0.0565	0.0006
	4:1	20	0.0808	0.0213	0.0491	0.0322	0.0808	0.0808
		50	0.0771	0.0191	0.0382	0.0146	0.0771	0.0771
		500	0.0734	0.0167	0.0065	0.0000	0.0734	0.0026
2x2 between-group	2:1	20	0.0576	0.0071	0.0247	0.0101	0.0576	0.0576
		50	0.0526	0.0058	0.0166	0.0019	0.0526	0.0526
		500	0.0495	0.0054	0.0005	0.0000	0.0495	0.0001
	4:1	20	0.0730	0.0142	0.0395	0.0200	0.0730	0.0730
		50	0.0601	0.0083	0.0189	0.0034	0.0601	0.0601
		500	0.0536	0.0043	0.0004	0.0000	0.0536	0.0001

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

Table 4.4 shows the Type I error rates under conditions of violations of the assumptions of homogeneity. Figure 4.4 displays the same data as a bar plot with confidence interval bands. The confidence interval bounds do not exceed a range of 0.01. Rates for conventional NHST ($p \leq 0.05$) held closely to 0.05 for all linear models under the 2:1 variance ratio condition. The 4:1 variance ratio condition, however, did influence the Type I error rates of the $k = 3$ and 2x2 between-group linear models, with both

showing an increase above the prescribed α level. The $k = 3$ model showed a notable increase to as much as 0.08 with the small sample size, which decreased as sample size increased. The 2x2 between-group model behaved similarly but with slightly less inflation. The NHST ($p \leq 0.005$) model had a similar pattern, with the most Type I error inflation occurring with the three-group model under the 4:1 variance ratio condition. SGPV methods followed the same pattern already observed of decreasing errors with increasing sample size. The MESP methods again matched the NHST ($p \leq 0.005$) except with MESP2 at a sample of $n = 500$.

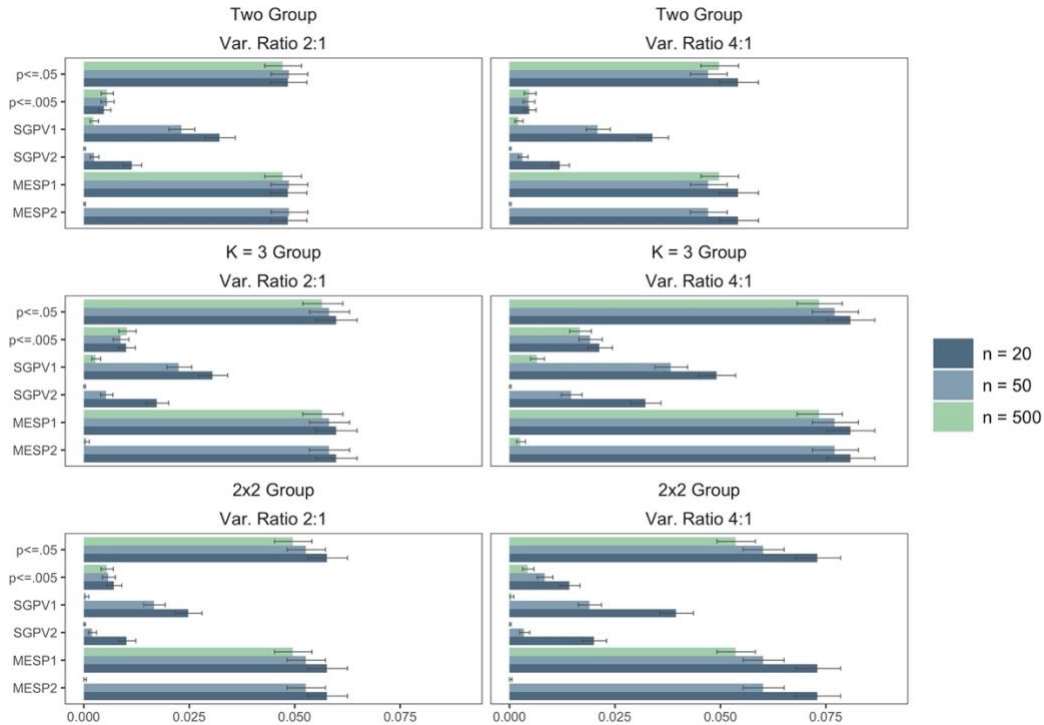


Figure 4.4 Type I Error and CI by Linear Model Under Violations of Homogeneity

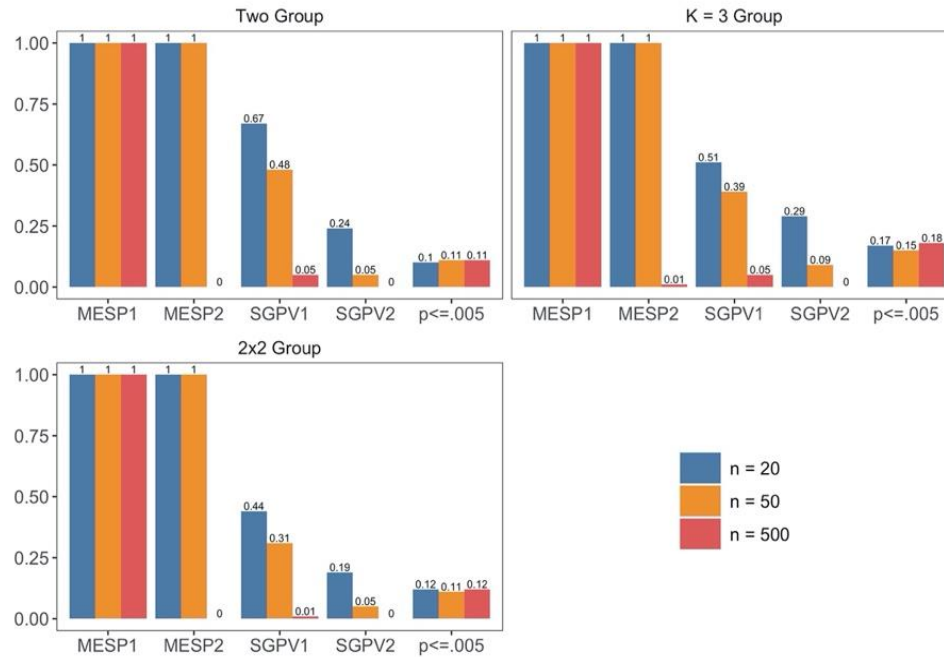


Figure 4.5 Type I Error Ratios by Linear Model Under Violations of Homogeneity: 2:1

The ratios of Type I error rates for the various methods relative to NHST ($p \leq 0.05$) under violations of homogeneity with a variance ratio of 2:1 are displayed in Figure 4.5. MESP 1 ($d = .1/\eta^2 = .005$) again consistently mirrored conventional NHST ($p \leq 0.05$). MESP 2 ($d = .35/\eta^2 = .035$) also tracked with NHST ($p \leq 0.05$) except at a sample of $n = 500$, where the rate dropped to zero or nearly zero, consistent with what was observed under normal conditions. SGPV displayed the same general pattern; however, there is a marginal increase in error, up to .06, relative to NHST over what was observed under normal conditions. Type I rates were consistently lower than conventional NHST for both SGPV 1 and SGPV 2, with the ratios again decreasing with an increase in sample size. False detection with the more stringent NHST ($p \leq 0.005$) held close to a ratio of 0.1 for two group and 2x2 group conditions. There was, however, a slight increase of Type I errors relative to conventional NHST for the three-group

condition. This increase is accounted for by the error inflation observed for NHST ($p \leq 0.005$), somewhat greater than that observed for conventional NHST ($p \leq 0.05$).

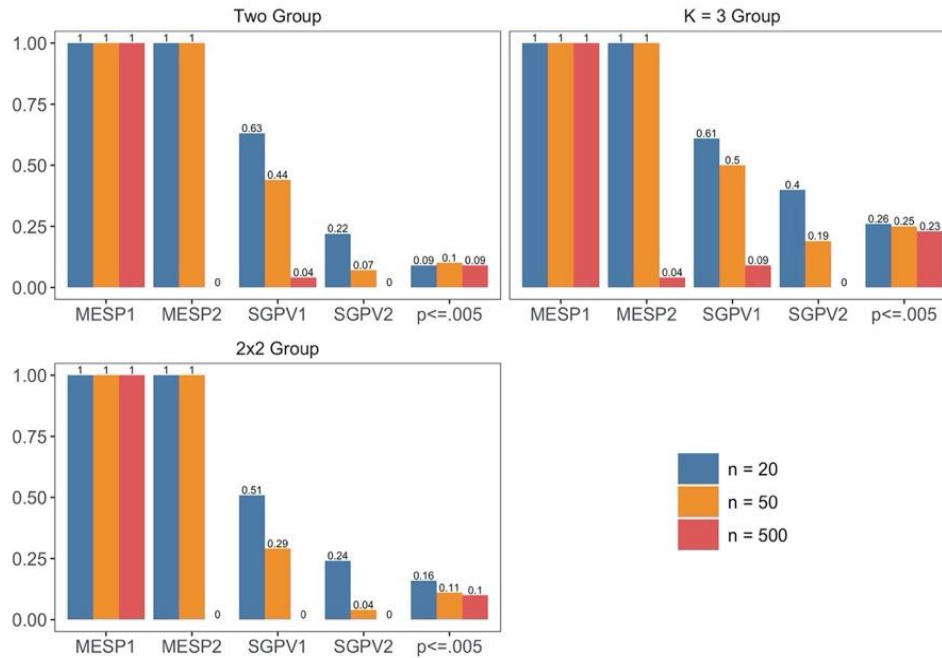


Figure 4.6 Type I Error Ratios by Linear Model Under Violations of Homogeneity: 4:1

Figure 4.6 Type I Error Ratios by Linear Model Under Violations of Homogeneity: 4:1 The ratios of Type I errors under the heterogeneous condition with a variance ratio of 4:1 are shown in Figure 4.6. The same general pattern is observed, as was seen in Figure 4.5. There is, however, a notable increase in the ratios for SGPV1, SGPV2, and NHST ($p \leq 0.005$) in the three-group condition over the amounts observed under the heterogeneous condition with a variance ratio of 2:1. Ratios increased for all sample size conditions for each of the three methods. SGPV1 and SGPV2 had increases of between .04 and .11 points with increases of at least .1 except at a sample size of $n = 500$, an indication that there is a sample size influence. NHST ($p \leq 0.005$) had increases from between .05 and .1 points relative to conventional NHST ($p \leq 0.05$).

Table 4.5 and Figure 4.7 show the Type I error rates under conditions of violations of distributional assumptions. Rate amounts and patterns matched very closely those observed under normal conditions. The error rates for both conventional NHST ($p \leq 0.05$) and NHST ($p \leq 0.005$) held consistently near the respective α levels for both skewed and bimodal distributions. Observed patterns for both SGPV methods show decreasing values with an increase in sample size. MESP methods again match exactly the values of conventional NHST ($p \leq 0.05$) except for MESP 2 and a large sample size of $n = 500$.

Table 4.5 Type I Error Under Distributional Violations

Linear Model	Distr.	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	skewed	20	0.0448	0.0037	0.0275	0.0076	0.0448	0.0448
		50	0.0484	0.0050	0.0211	0.0025	0.0484	0.0484
		500	0.0493	0.0049	0.0020	0.0000	0.0493	0.0000
	bimodal	20	0.0495	0.0050	0.0321	0.0107	0.0495	0.0495
		50	0.0523	0.0045	0.0221	0.0019	0.0523	0.0523
		500	0.0478	0.0047	0.0019	0.0000	0.0478	0.0001
Three-group	skewed	20	0.0458	0.0045	0.0195	0.0096	0.0458	0.0458
		50	0.0485	0.0056	0.0180	0.0035	0.0485	0.0485
		500	0.0489	0.0052	0.0005	0.0000	0.0489	0.0000
	bimodal	20	0.0510	0.0057	0.0243	0.0112	0.0510	0.0510
		50	0.0499	0.0050	0.0180	0.0028	0.0499	0.0499
		500	0.0528	0.0062	0.0009	0.0000	0.0528	0.0002
2x2 between-group	skewed	20	0.0497	0.0043	0.0190	0.0075	0.0497	0.0497
		50	0.0556	0.0049	0.0139	0.0020	0.0556	0.0556
		500	0.0524	0.0043	0.0004	0.0000	0.0524	0.0000
	bimodal	20	0.0510	0.0063	0.0215	0.0097	0.0510	0.0510
		50	0.0488	0.0046	0.0124	0.0011	0.0488	0.0488
		500	0.0475	0.0042	0.0006	0.0000	0.0475	0.0000

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

Figure 4.8 and Figure 4.9 display the ratios of the detection rates with skewed and bimodal population distributions. The ratios varied little in amounts or patterns from those observed with a normal distribution. Here also, the MESP 1 ($d = .1/\eta^2 = .005$) tracks exactly with the $p \leq 0.05$. MESP 2 ($d = .35/\eta^2 = .035$) also matches the $p \leq$

0.05 results except for a sample size of 500, as was observed in other conditions. The SGPV 1 and SGPV 2 had lower Type I error rates than the NHST with $p \leq 0.05$. Both SGPV methods again displayed an inverse relationship to sample size. NHST with $p \leq 0.005$ held a reasonably constant rate with ratios ranging from 0.08 to 0.13 relative to conventional NHST.

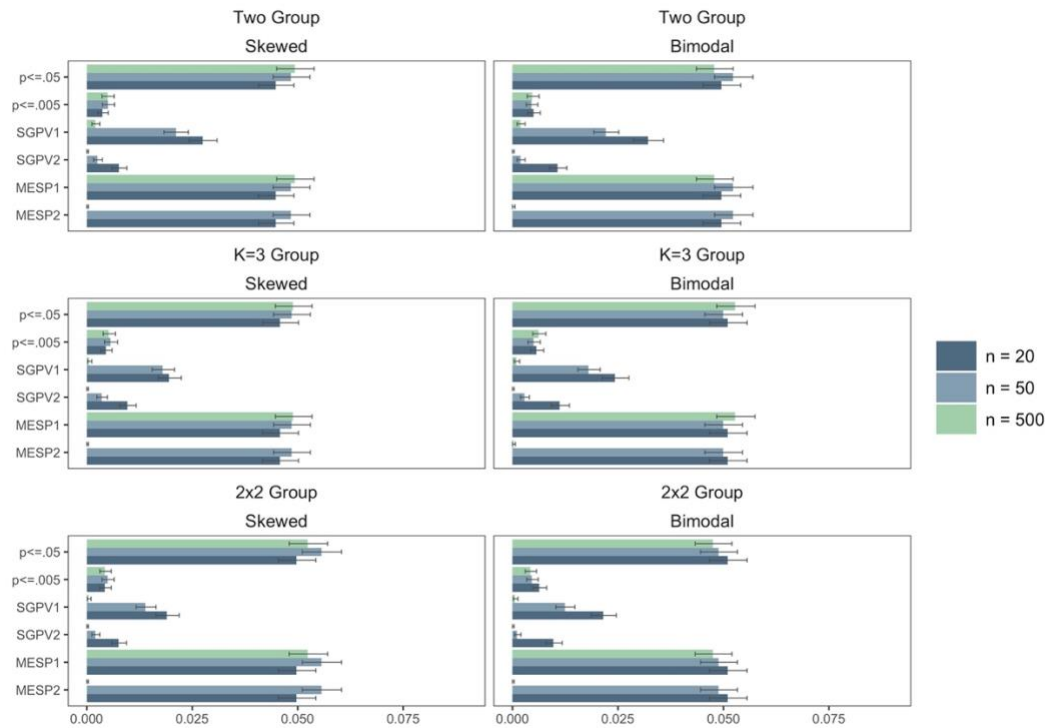


Figure 4.7 Type I Error for Skewed and Bimodal Distributions by Linear Model

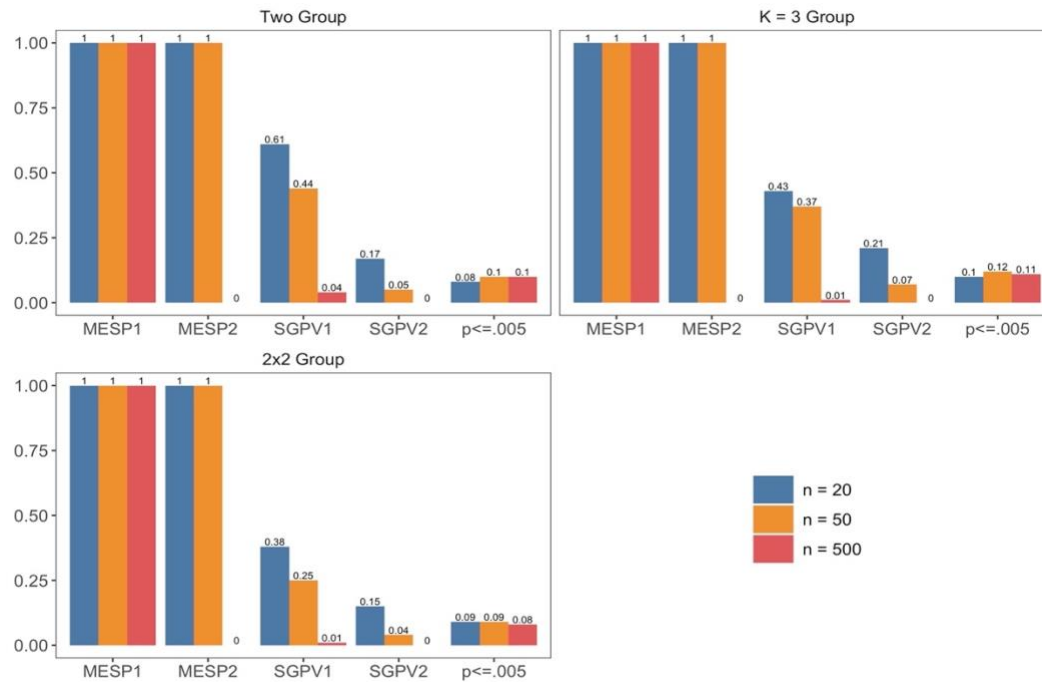


Figure 4.8 Skewed Distribution Type I Error Ratios by Linear Model

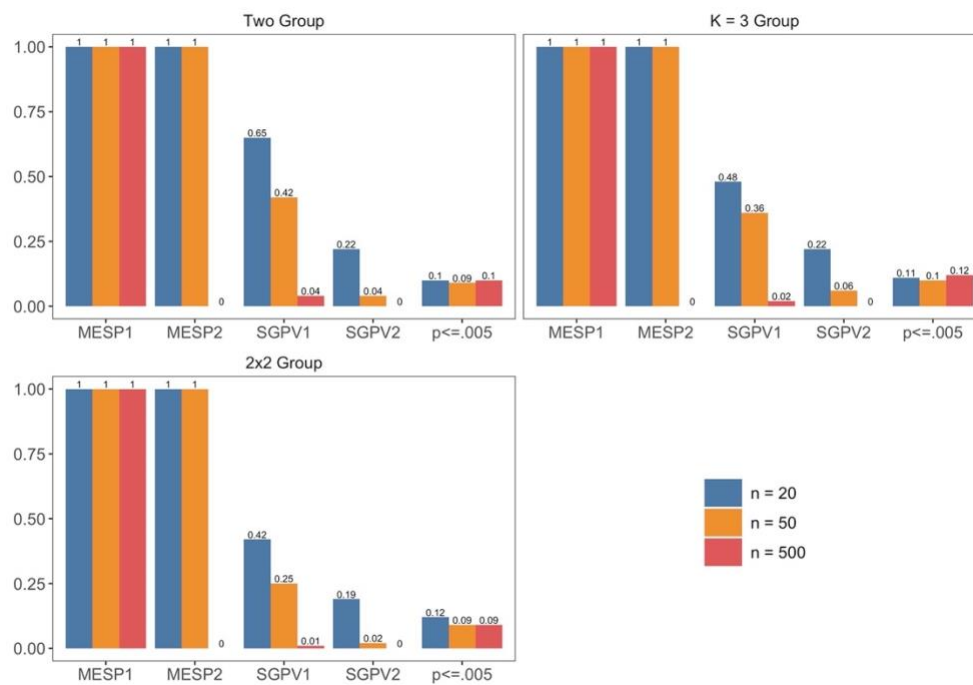


Figure 4.9 Bimodal Type I Error Ratios by Linear Model

Table 4.6 Summary of Results for Type I Error

Independent Variable	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP2
Linear Model	Type I error held consistently near or 0.005	Type I error lower than NHST $p \leq 0.05$ regardless of model. Range of 0.0005 - 0.03 across all models.	Type I error lower than NHST $p \leq 0.05$ regardless of model. Range of 0.00 - 0.012 across all models.	Type I error held near 0.05 and tracked precisely with NHST $p \leq 0.05$.	There was no effect of linear model on Type I error relative to NHST $p \leq 0.05$.
Sample Size	Type I error held consistently near or 0.005	Type I error held lower than NHST $p \leq 0.05$ regardless of sample size. Error rates decreased with increase in sample size.	Type I error held lower than NHST $p \leq 0.05$ regardless of sample size. Error rates decreased with increase in sample size.	Type I error tracked precisely with NHST $p \leq 0.05$.	Type I error was identical with NHST $p \leq 0.05$ at the low and moderate sample size. Error rates at the large sample size were consistently near 0.
Variance Ratio	Slight error inflation was observed with 2:1 ratio with ANOVA based analysis. Much greater error inflation was observed with 4:1 ratio and ANOVA based analysis; an increase proportionally greater than NHST $p \leq 0.05$.	Type I error held lower than NHST $p \leq 0.05$. Slight error inflation was observed with 2:1 ratio with ANOVA based analysis. Greater error inflation was observed with 4:1 ratio and ANOVA based analysis; an increase proportionally greater than NHST $p \leq 0.05$.	Type I error held lower than NHST $p \leq 0.05$. Slight error inflation was observed with 2:1 ratio with ANOVA based analysis. Greater error inflation was observed with 4:1 ratio and ANOVA based analysis; an increase proportionally greater than NHST $p \leq 0.05$.	Type I error tracked precisely with NHST $p \leq 0.05$.	There was no effect of variance ratio on Type I error relative to NHST $p \leq 0.05$.
Population Distribution	Type I error held consistently near or 0.005 for both skewed and bimodal distributions.	Type I error lower than NHST $p \leq 0.05$ regardless of model. There was no observed effect of population distribution.	Type I error lower than NHST $p \leq 0.05$ regardless of model. There was no observed effect of population distribution.	Type I error tracked precisely with NHST $p \leq 0.05$.	There was no effect of distribution on Type I error relative to NHST $p \leq 0.05$.

Statistical Power

Table 4.7 shows the power of each method under normal conditions.

Conventional NHST ($p \leq 0.05$) behaves as expected, with power increasing as a function of effect size and sample size regardless of model. NHST ($p \leq 0.005$) shows a large reduction in power at lower sample and effect sizes when compared with NHST ($p \leq 0.005$). As can be seen, power levels are relatively low using NHST ($p \leq 0.005$) even with moderate effects or sample sizes, only exceeding .47 with a sample of 500 and at least a moderate effect. This, however, is not a flaw in the stricter method, but rather, it behaves exactly as it is intended. The purpose of using NHST ($p \leq 0.005$) is to reduce Type I error which necessarily requires a reduction in power except at high sample sizes or with very strong effects. All three models for SGPV 1 showed an increase in power with an increase in sample size, a pattern that held with all effect sizes. This pattern held for SGPV 2 for moderate and high effects sizes. SGPV 2 with a low effect had decreasing power with an increase in sample size. This occurs because, with SGPV, an increase of precision afforded by an increase in the sample leads to a lower rate of detection when the population effect is near or less than the interval bound. The MESP methods match those of NHST ($p \leq 0.005$) except with a high sample size and small or moderate effect size exhibiting only a small difference with a moderate population effect.

Table 4.7 Power Under Normality Condition

Linear Model	Pop effect d/η^2	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	.2	20	0.0700	0.0086	0.0456	0.0168	0.0700	0.0700
		50	0.1066	0.0160	0.0563	0.0076	0.1066	0.1066
		500	0.6015	0.2833	0.2066	0.0000	0.6015	0.0512
	.5	20	0.1729	0.0291	0.1255	0.0525	0.1729	0.1729
		50	0.4045	0.1339	0.2802	0.0825	0.4045	0.4045
		500	0.9999	0.9974	0.9947	0.3947	0.9999	0.9549
	.8	20	0.3841	0.1054	0.3065	0.1613	0.3841	0.3841
		50	0.7868	0.4673	0.6749	0.3561	0.7868	0.7868
		500	1.0000	1.0000	1.0000	0.9993	1.0000	0.9999
Three-group	.01	20	0.0647	0.0085	0.0302	0.0174	0.0647	0.0647
		50	0.0836	0.0123	0.0331	0.0073	0.0836	0.0836
		500	0.4622	0.1773	0.0800	0.0000	0.4622	0.0281
	.06	20	0.1326	0.0192	0.0715	0.0419	0.1326	0.1326
		50	0.3221	0.0896	0.1836	0.0617	0.3221	0.3221
		500	0.9992	0.9931	0.9768	0.2734	0.9992	0.9367
	.14	20	0.2941	0.0687	0.1833	0.1203	0.2941	0.2941
		50	0.7036	0.3570	0.5351	0.2929	0.7036	0.7036
		500	1.0000	1.0000	1.0000	0.9972	1.0000	1.0000
2x2 between-group	.01	20	0.0722	0.0088	0.0345	0.0136	0.0722	0.0722
		50	0.1114	0.0193	0.0416	0.0057	0.1114	0.1114
		500	0.6136	0.2935	0.1018	0.0000	0.6136	0.0242
	.06	20	0.1861	0.0336	0.0987	0.0490	0.1861	0.1861
		50	0.4333	0.1466	0.2457	0.0686	0.4333	0.4333
		500	0.9999	0.9982	0.9834	0.2921	0.9999	0.9179
	.14	20	0.3906	0.1105	0.2514	0.1456	0.3906	0.3906
		50	0.8140	0.4961	0.6398	0.3252	0.8140	0.8140
		500	1.0000	1.0000	1.0000	0.9965	1.0000	1.0000

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

Figure 4.10, Figure 4.11, and Figure 4.12 display the ratios of observed power for each method relative to NHST ($p \leq 0.05$) by experimental condition and population effect size. The MESP1 method's power results track with the NHST ($p \leq 0.05$) results regardless of linear model or effect size. This holds consistently at every level for MESP 1, revealing that setting a low effect size cutoff is no different from conventional NHST under normal conditions. MESP 2 also mirrors NHST ($p \leq 0.05$) power except under conditions of large sample size ($n = 500$). Even with a large sample, there is a notable difference only when the effect size is small, with ratios of 0.09, 0.06, and 0.04 for the

two-group, three-group, and 2x2 between-group conditions. With a moderate effect, power is still high, with a ratio of at least .92 across models.

SGPV 1 generally has low power under conditions of low population effect across models. With a low effect and $n = 20$, power ratios range from 0.42 - 0.65, with less power than NHST ($p \leq 0.05$) by more than half for three-group and two by two between-group models. In fact, SGPV 1 with the two-group model had higher power ratios than with three-group model or 2x2 between-group models at all effect size amounts. The same pattern was not observed with SGPV 2.

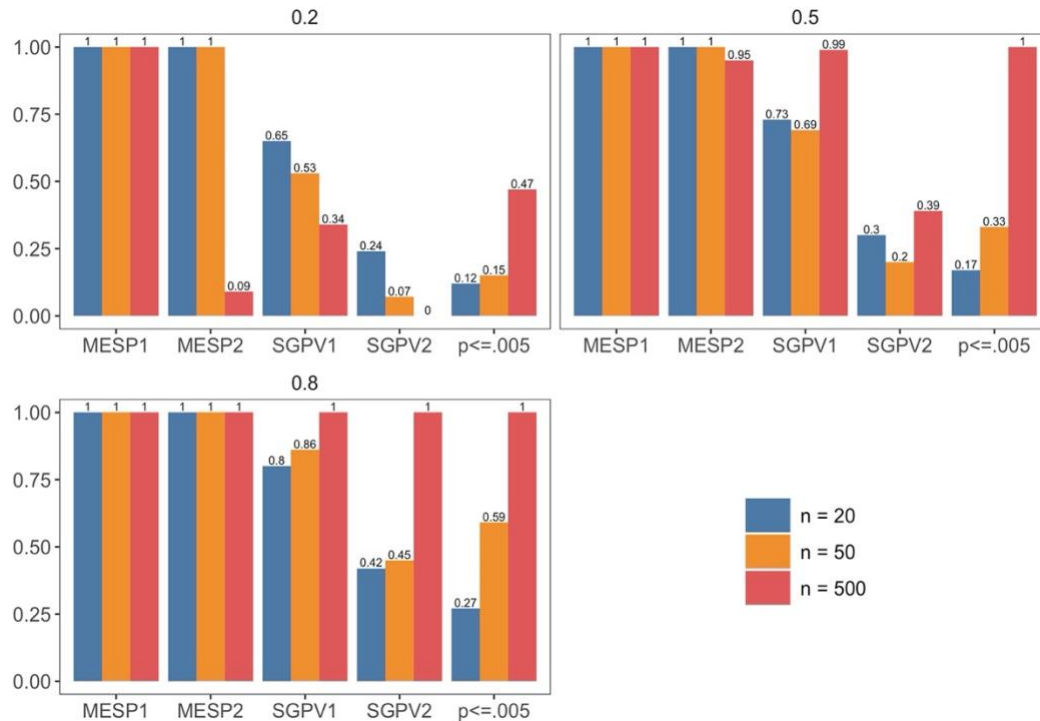


Figure 4.10 Power Ratios for Two-group Linear Model by Population Effect Size Under Normality Condition

For all three linear models, both SGPV methods with the small effect size showed decreasing power ratios with an increase in sample size. The opposite pattern is observed with the large effect sizes for SGPV 1. The moderate effect size power ratios might have

a positive relationship to sample size as well, but it is difficult to discern as the low and moderate sample size ratios are nearly flat. It is likely that the small interval between $n = 20$ and $n = 50$ is masking the relationship. SGPV 2 ($d = .35/\eta^2 = .035$) had low power ratios at the low population effect size which is expected with the wider interval null of SGPV2. With a moderate population effect, the power ratios of SGPV 2 displayed a “U” shaped pattern associated with sample size. This pattern occurred despite real power calculations increasing with sample size, as seen in Table 4.7. SGPV 2 with a large effect had the same pattern as SGPV 1 with a moderate effect, similar ratios at $n = 20$ and 50 , and a large ratio, approaching 1, with $n = 500$.

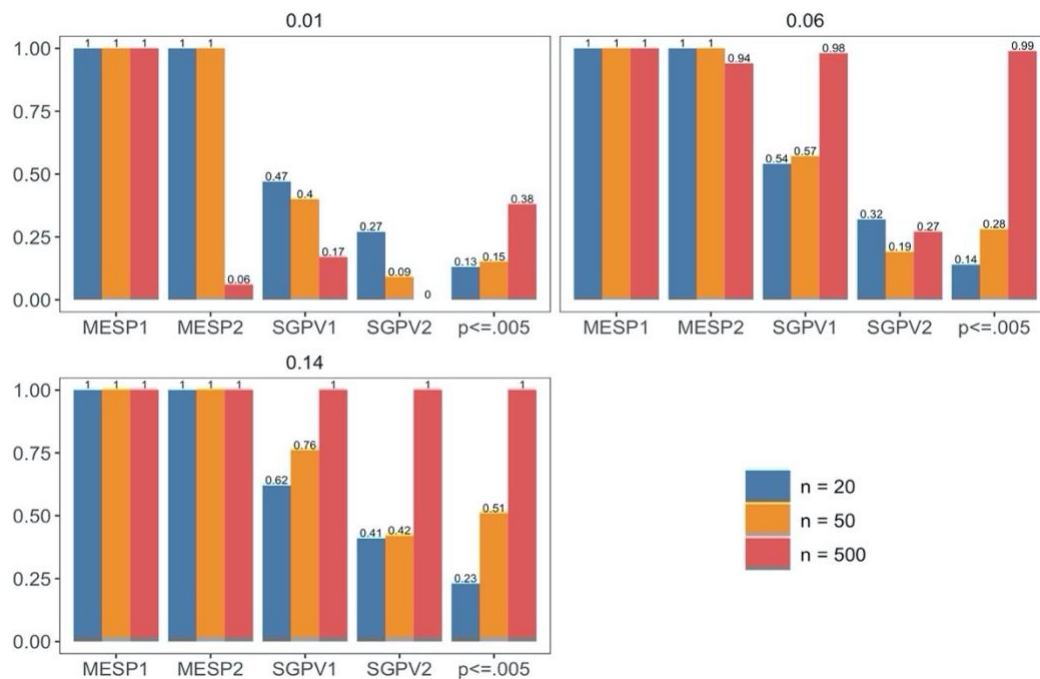


Figure 4.11 Power Ratios for Three-group linear model by Pop. Effect Size Under Normal Condition

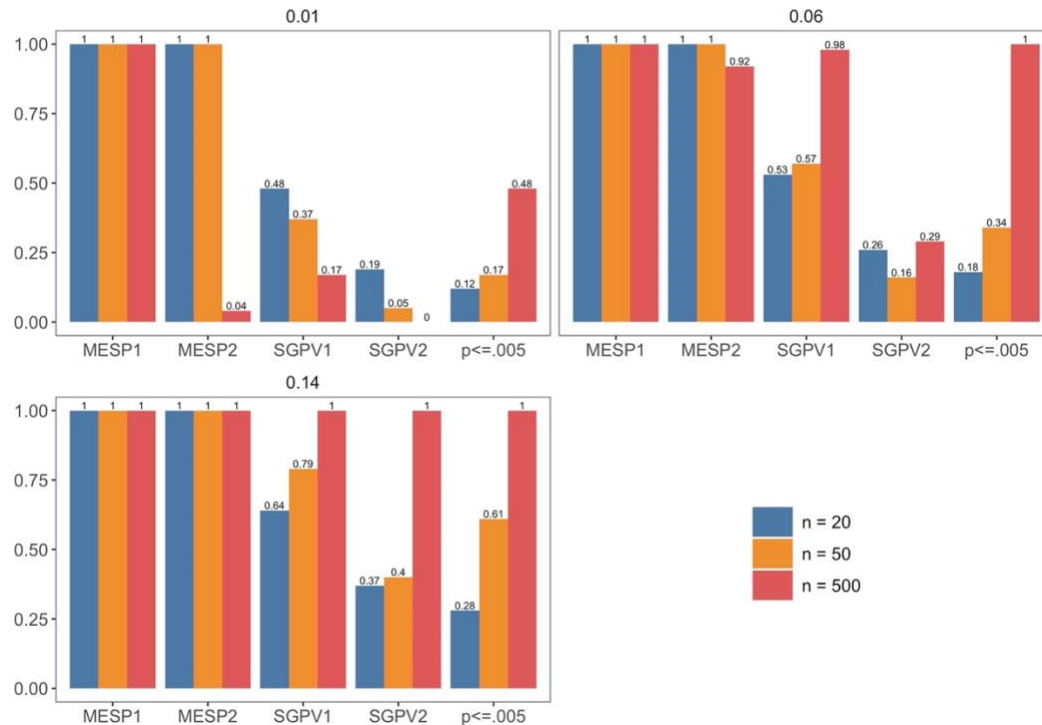


Figure 4.12 Power for 2x2 between-group Linear Model by Population Effect Size Under Normality Condition

Power, as defined here, is the detection of an effect without regard to the estimate's accuracy. Effect estimates, when compared with the population effect, can give an indication of the general accuracy of a statistical method. Figure 4.13 displays histograms of the unstandardized estimated effects when a method detected an effect for each of the three linear models. Only the two-group model is shown as the general trends are consistent for all linear models. The only notable difference between the two-group model and other linear models is that the standardized effect scale used in the $k = 3$ and 2x2 between-group models does not allow for negative effects. The plot of NHST ($p \leq 0.05$) estimates is displayed in light green with that of the compared method in blue. The mean estimate for NHST ($p \leq 0.05$) is indicated with a black vertical line, while the

mean of all other methods is in red. Plots of MESP 1 and MESP 2 with $n = 20$ and $n = 50$ are omitted since the estimates are identical to NHST ($p \leq 0.05$).

In Figure 4.13, the distribution of unstandardized estimated effects for NHST ($p \leq 0.005$) contains several discernable patterns. Fewer effects were detected relative to conventional NHST for all conditions except those with a sample size of 500 and moderate or large population effects in which the distribution of effect estimates for both methods match precisely. Also, the mean estimates are inflated relative to the population effects for both conventional NHST and NHST ($p \leq 0.005$) with the small and moderate samples. Further, with the exception of the $n = 500$ sample with moderate and large population effects, the NHST ($p \leq 0.005$) distributions are narrower and tend to be right-skewed. Consequently, the mean effect estimates for NHST ($p \leq 0.005$) are higher than those of NHST ($p \leq 0.05$). The more stringent criteria of NHST ($p \leq 0.005$) require higher estimated effects for detection except with a sample of $n = 500$ with moderate or large effects. This results in fewer detections of effect, but when effects are detected, effect estimates are overestimated compared to the population effect and also higher than the estimates of conventional NHST.

The distributions of effect estimates for SGPV 1 clearly display the influence of using a narrow null interval $(-1.5, 1.5)$. It should be noted that looking at the point estimate cannot display the full range of what the interval method is doing; nonetheless, it is helpful to focus on the effect estimates for comparative purposes. The method successfully eliminates smaller effects, as can be seen by the lack of overlap of the distributions on the left of all plots except for $n = 500$ with moderate or large population effects. Further, SGPV 1 eliminates more small effects and, therefore, differs more from

NHST ($p \leq 0.05$) when the population effect is close to the null interval bound and has a high sample size. The distribution of estimates for SGPV 1 differs less from conventional NHST when the sample size is small or moderate with a population effect near the null interval. The means of estimated effects are close to those of NHST ($p \leq 0.05$) but generally higher with the exception of those for $n = 500$ with moderate or large effects, which are identical.

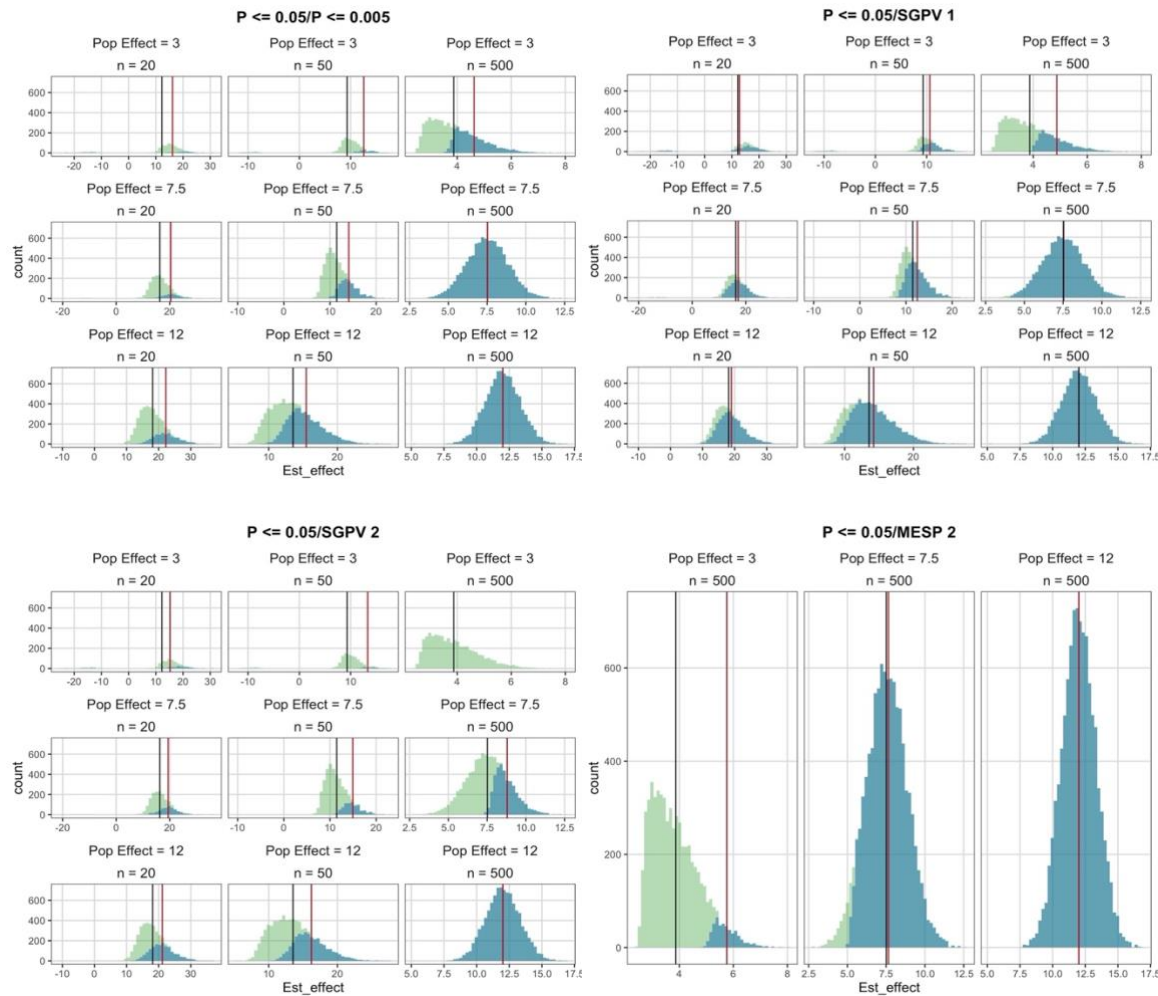


Figure 4.13 Two-group Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)

SGPV 2 has a much wider null interval (-5.25, 5.25) than SGPV 1. As a consequence, far more effects are eliminated as being trivial, with much fewer detections generally relative to NHST ($p \leq 0.05$). This is clearly demonstrated with the small population effect, where very few effects were detected at all sample sizes. A sample size influence is, however, observable as there are some effects detected at low and moderate sample sizes but none at $n = 500$. These also are notably inflated relative to the true population effect. The moderate and large population effect sizes show the same pattern as SGPV 1 with the elimination of smaller effect estimates and a consistently higher mean estimate than NHST ($p \leq 0.05$). The two method's distributions match precisely with a large population effect and an $n = 500$ sample size.

MESP methods had distributions of effects identical to NHST ($p \leq 0.05$) for all conditions except with MESP 2 with a sample size of $n = 500$. MESP 2 had an unstandardized MPSD of 5.25 points. The functioning of the effect cut-off is seen with both the small and moderate population effect sizes, where the distinction between MESP and NHST ($p \leq 0.05$) occurs at precisely 5.25 points. The distributions otherwise are identical.

Table 4.8, Table 4.9, and Table 4.10 show the power results under violations of homogeneity. There is a general decrease in power for all methods when compared with the normal conditions. Further, the decrease is greater as the variance ratio becomes more imbalanced. Figures 4.12, 4.13, and 4.14 contain bar plots of the power rates for both homogeneous and heterogeneous conditions providing a clear visual demonstration of the decrease of power as the variance ratio becomes more pronounced. Also, the plots clearly demonstrate that this pattern holds for all analytical methods.

Table 4.8 Power for Two-group linear model under violations of Homogeneity

Var. ratio	Pop effect d	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
2:1	.2	20	0.0612	0.0062	0.0414	0.0137	0.0612	0.0612
		50	0.0719	0.0085	0.0351	0.0057	0.0719	0.0719
		500	0.2944	0.0869	0.0511	0.0000	0.2944	0.0057
	.5	20	0.1004	0.0135	0.0709	0.0269	0.1004	0.1004
		50	0.1907	0.0416	0.1182	0.0231	0.1907	0.1907
		500	0.9394	0.7558	0.6651	0.0099	0.9394	0.3499
	.8	20	0.1810	0.0361	0.1364	0.0642	0.1810	0.1810
		50	0.4092	0.1369	0.2928	0.0886	0.4092	0.4092
		500	1.0000	0.9978	0.9948	0.4228	1.0000	0.9563
4:1	.2	20	0.0494	0.0053	0.0344	0.0111	0.0494	0.0494
		50	0.0549	0.0061	0.0248	0.0033	0.0549	0.0549
		500	0.1186	0.0223	0.0121	0.0000	0.1186	0.0015
	.5	20	0.0685	0.0080	0.0468	0.0207	0.0685	0.0685
		50	0.0878	0.0117	0.0452	0.0062	0.0878	0.0878
		500	0.4855	0.1834	0.1242	0.0001	0.4855	0.0239
	.8	20	0.0843	0.0113	0.0589	0.0254	0.0843	0.0843
		50	0.1572	0.0296	0.0906	0.0184	0.1572	0.1572
		500	0.8588	0.5982	0.5008	0.0024	0.8588	0.2084

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

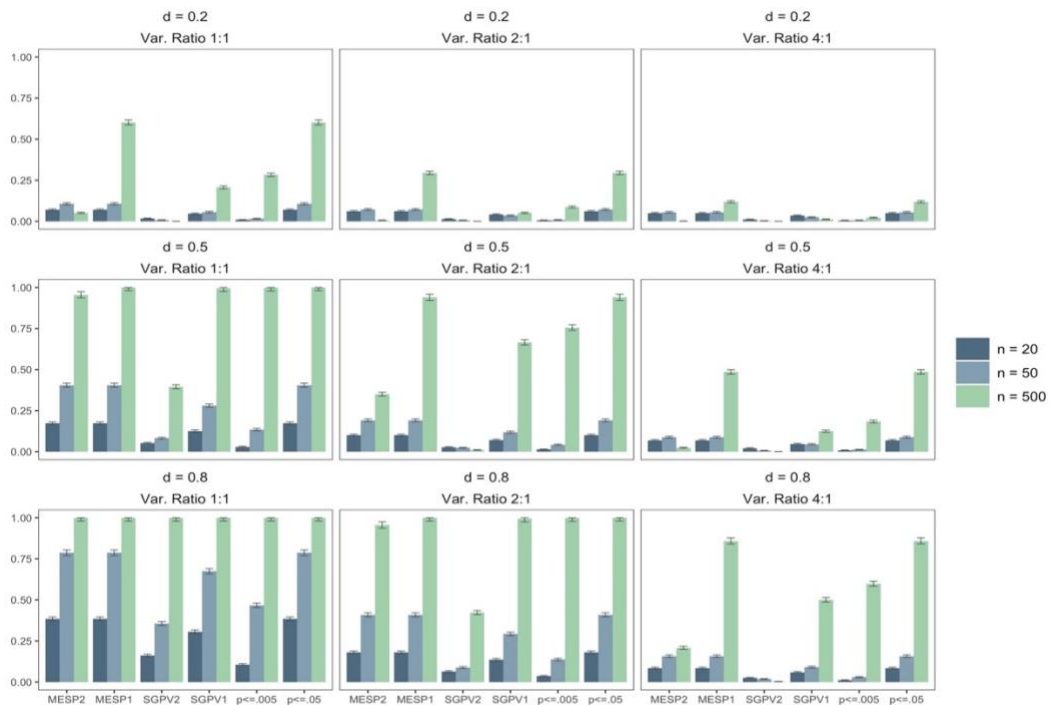


Figure 4.14 Power and CI for Two-group Linear Model by Pop. Effect Size and Var. Ratio

Table 4.9 Power for Three-group linear model under violations of Homogeneity

Var. ratio	Pop effect η^2	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
2:1	.01	20	0.0682	0.0116	0.0359	0.0210	0.0682	0.0682
		50	0.0791	0.0148	0.0364	0.0092	0.0791	0.0791
		500	0.2541	0.0868	0.0398	0.0000	0.2541	0.0151
	.06	20	0.1031	0.0189	0.0578	0.0353	0.1031	0.1031
		50	0.1886	0.0543	0.1048	0.0382	0.1886	0.1886
		500	0.9024	0.7396	0.5934	0.0326	0.9024	0.4364
	.14	20	0.1810	0.0480	0.1144	0.0791	0.1810	0.1810
		50	0.3832	0.1562	0.2543	0.1232	0.3832	0.3832
		500	0.9995	0.9958	0.9856	0.5321	0.9995	0.9632
4:1	.01	20	0.0774	0.0193	0.0437	0.0302	0.0774	0.0774
		50	0.0842	0.0202	0.0417	0.0139	0.0842	0.0842
		500	0.1334	0.0424	0.0186	0.0000	0.1334	0.0081
	.06	20	0.0955	0.0248	0.0618	0.0401	0.0955	0.0955
		50	0.1192	0.0321	0.0641	0.0235	0.1192	0.1192
		500	0.4752	0.2432	0.1465	0.0015	0.4752	0.0790
	.14	20	0.1163	0.0349	0.0752	0.0532	0.1163	0.1163
		50	0.1816	0.0639	0.1116	0.0505	0.1816	0.1816
		500	0.8262	0.6342	0.4903	0.0267	0.8262	0.3548

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

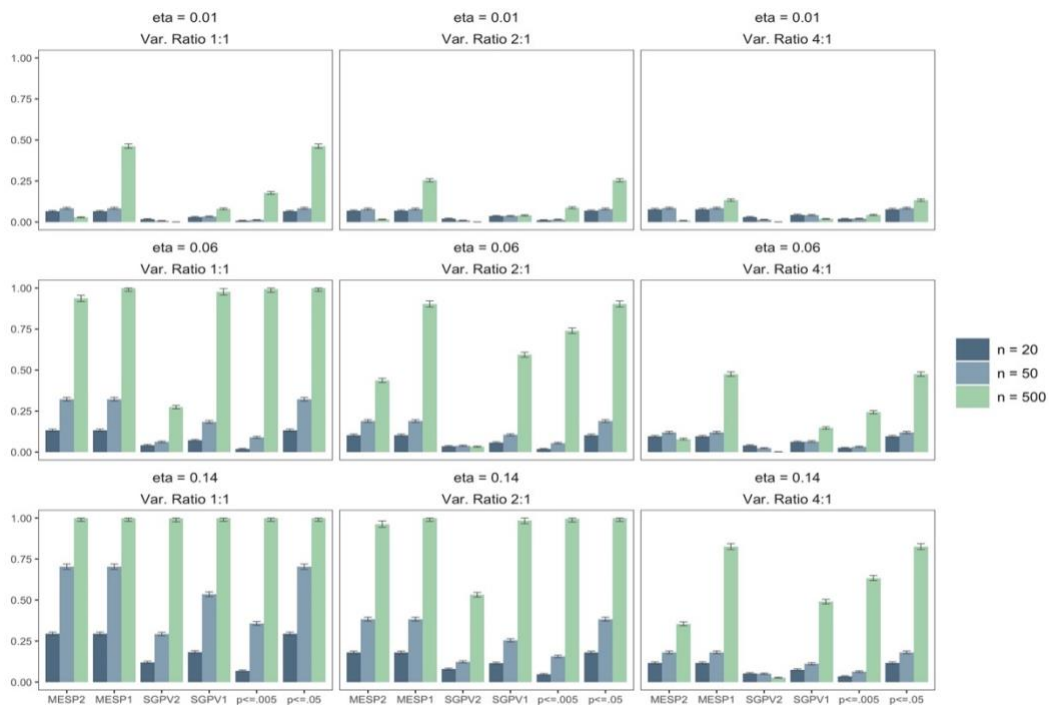


Figure 4.15 Power and CI for Three-group linear model by Pop. Effect Size and Var. Ratio

Table 4.10 Power for Two by two between-group linear model under violations of Homogeneity

Var. ratio	Pop effect η^2	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
2:1	.01	20	0.0666	0.0100	0.0327	0.0144	0.0656	0.0656
		50	0.0848	0.0145	0.0300	0.0056	0.0848	0.0848
		500	0.4016	0.1344	0.0352	0.0000	0.4016	0.0067
	.06	20	0.1358	0.0266	0.0730	0.0364	0.1358	0.1358
		50	0.2748	0.0740	0.1315	0.0321	0.2748	0.2748
		500	0.9884	0.9239	0.7650	0.0286	0.9884	0.5217
	.14	20	0.2608	0.0651	0.1569	0.0879	0.2608	0.2608
		50	0.5779	0.2490	0.3711	0.1394	0.5779	0.5779
		500	0.1000	0.1000	0.9995	0.7095	0.1000	0.9945
4:1	.01	20	0.0792	0.0162	0.0411	0.0210	0.0792	0.0792
		50	0.0740	0.0109	0.0254	0.0047	0.0740	0.0740
		500	0.1809	0.0413	0.0074	0.0000	0.1809	0.0007
	.06	20	0.1114	0.0262	0.0617	0.0340	0.1114	0.1114
		50	0.1385	0.0336	0.0638	0.0129	0.1385	0.1385
		500	0.7466	0.4183	0.01780	0.0004	0.7466	0.0566
	.14	20	0.1549	0.0411	0.0906	0.0533	0.1549	0.1549
		50	0.2686	0.0813	0.1347	0.0387	0.2686	0.2686
		500	0.9842	0.9018	0.7170	0.0220	0.9842	0.4637

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

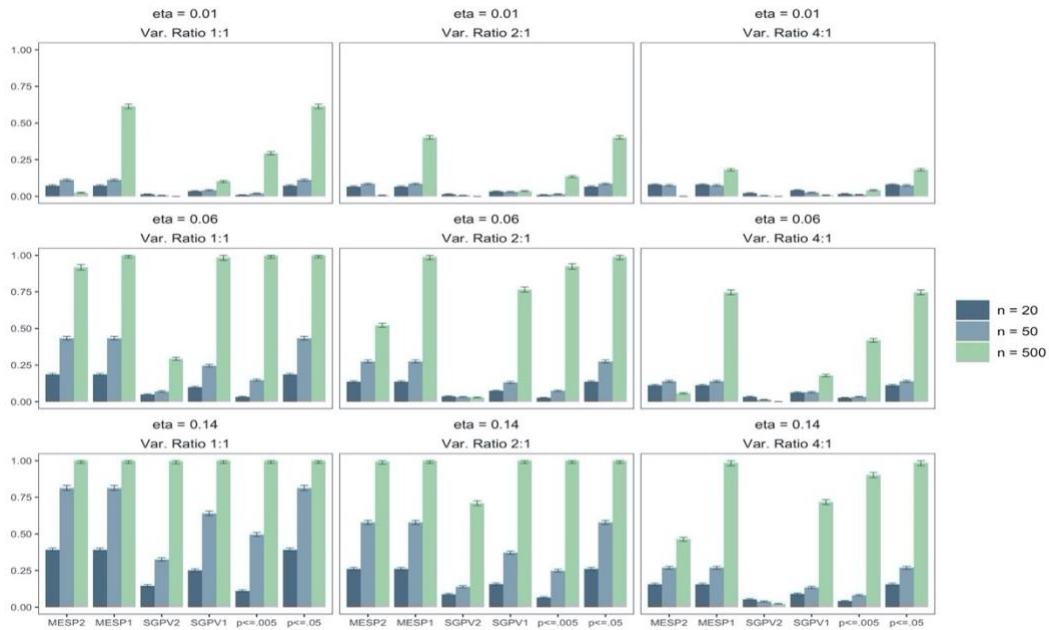


Figure 4.16 Power and CI for 2x2 between-group Linear model by Pop. Effect Size and Var. Ratio

Figure 4.17 displays the ratios of observed power for each method relative to NHST ($p \leq 0.05$) by population effect size and variance ratio for the two-group linear model. The MESP1 method's power results are identical with the NHST ($p \leq 0.05$) results regardless of linear model or variance ratio. With MESP 2, the pattern holds in that it mirrors NHST ($p \leq 0.05$) except with the large sample size ($n = 500$). However, the power ratios are dramatically lower with the moderate and large sample size than that which was observed under the normal condition, where a ratio of 0.95 was achieved with a moderate population effect. A power ratio of 0.37 with a variance ratio of 2:1 and 0.05 with a variance ratio of 4:1 indicates that when the variances are imbalanced many more small effects are detected using a significance level of 0.05 and are then eliminated by the MPSE. This further provides evidence of less accurate estimates with an increased variance ratio.

The power ratios for the SGPV methods with the low population effect are similar for all variance ratios except with the sample size of $n = 500$. Here there is a pattern of reducing ratios as the variance ratio becomes more imbalanced. With a moderate and large effect, reductions in the power ratios can be observed at both the $n = 50$ and $n = 500$ sample size with greater loss as the variance ratios and sample size increase.

NHST ($p \leq 0.005$) method generally had lower power ratios compared to the normal homogeneous condition. The differences were greatest with a large sample size and increased with increased heterogeneity. This trend held except for the large effect size and a variance ratio of 2:1.

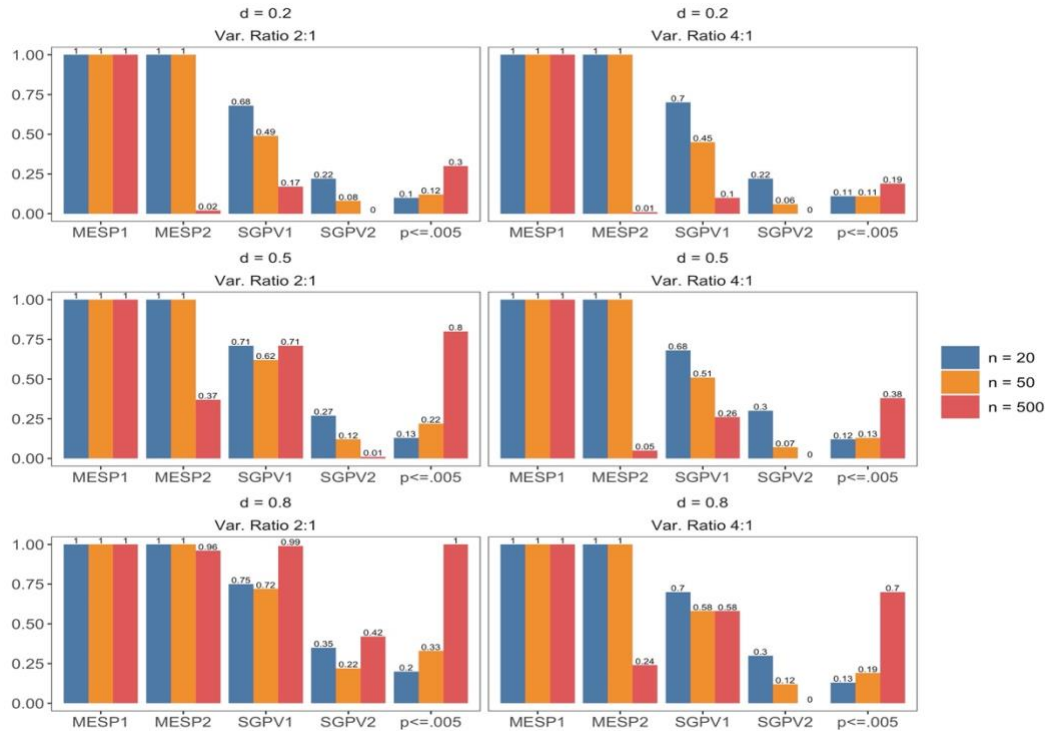


Figure 4.17 Power ratios for Two-group Liner Model by Population Effect Size and Variance Ratio

NHST ($p \leq 0.005$) method generally had lower power ratios compared to the normal homogeneous condition. The differences were greatest with a large sample size and increased with increased heterogeneity. This trend held except for the large effect size and a variance ratio of 2:1.

Figures 4.18 and 4.19 show the power ratio results for the three-group and 2x2 between-group models. The trends observed in the two-group model held for each of these models as well. Of note, however, is that the power ratios are consistently lower for the 2x2 group model than for the two-group or three-group models.

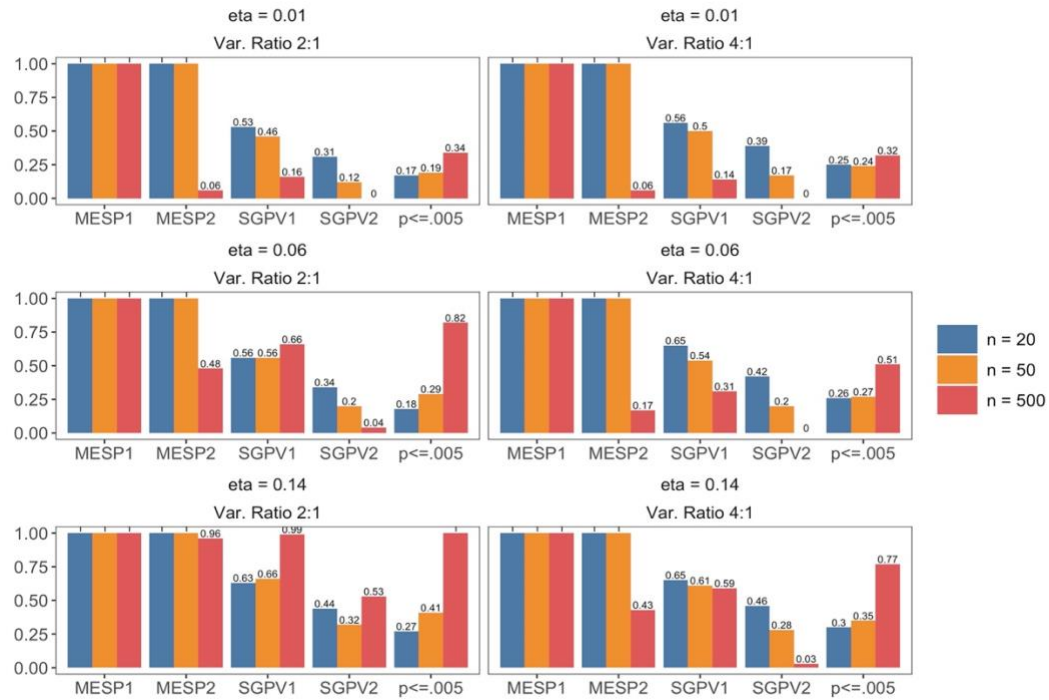


Figure 4.18 Power Ratios for Three-group linear model by Population Effect Size and Variance Ratio

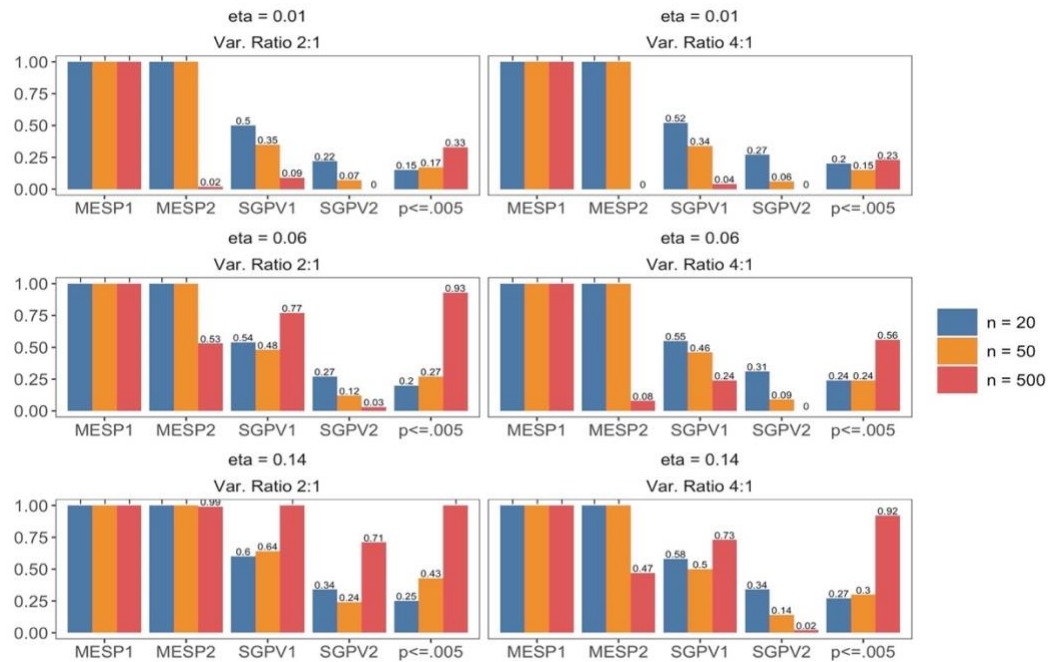


Figure 4.19 Power Ratios for 2x2 between-group linear model by Population Effect Size and Variance Ratio

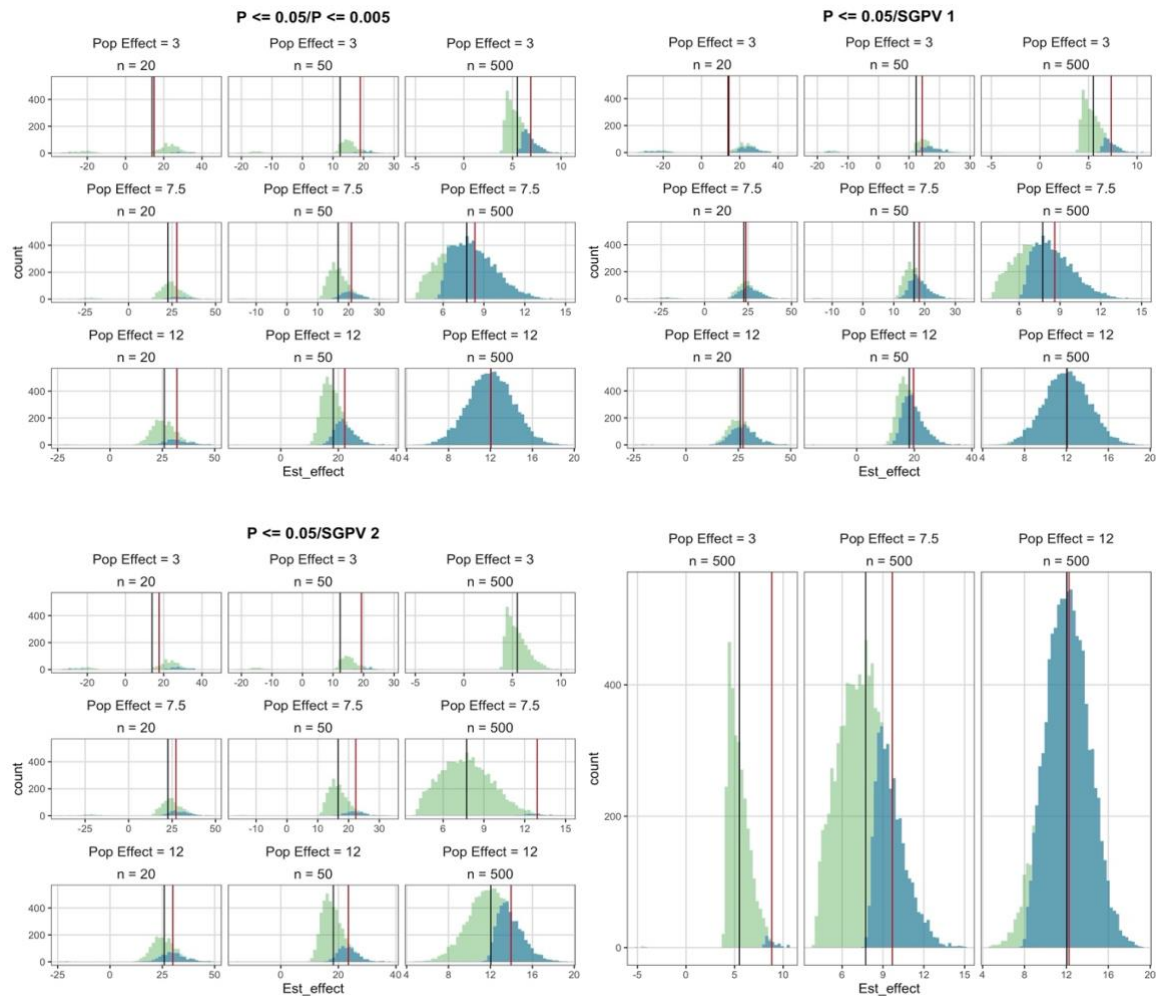


Figure 4.20 Two Group linear model with 2:1 Var. Ratio Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)

Figure 4.20 displays the distribution of unstandardized estimated effects for the two-group model and 2:1 variance ratio. All methods, including NHST ($p \leq 0.05$), detected fewer effects relative to those detected under homogeneous conditions. Further, the detected effects are inflated at the low and moderate sample sizes. At the large sample size, estimates are closer to the population effect but are less accurate than with under homogeneous conditions. Also, all of the alternative methods diverge from NHST ($p \leq 0.05$) at all conditions except with $n = 500$ and a large population effect with NHST ($p \leq$

0.005) and SGPV 1. Lastly, and most notably, MESP 2 eliminated effects that are greater than the unstandardized MPSD at the small population effect size.

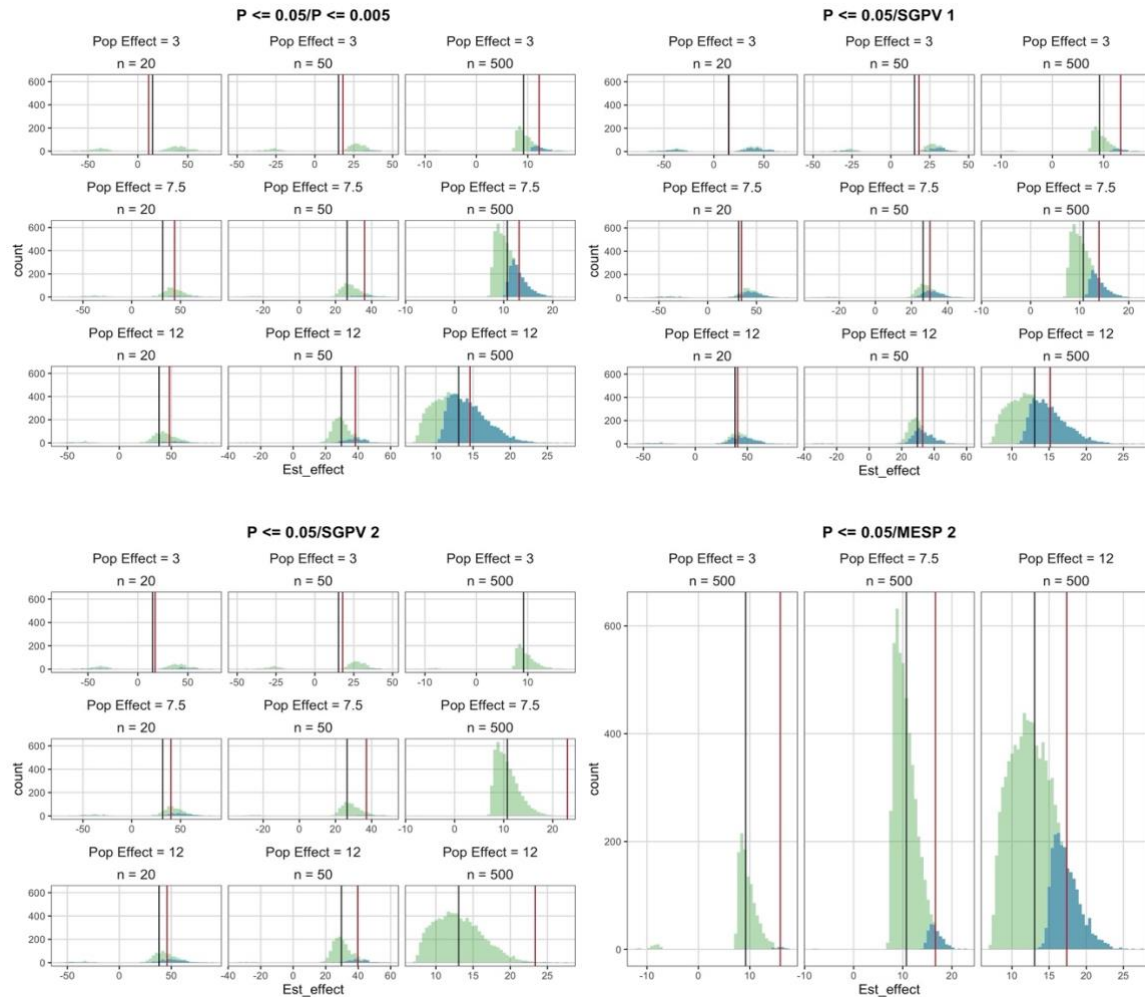


Figure 4.21 Two Group linear model with 4:1 Var. Ratio Effect Estimate Distributions by Method Compared with NHST ($p \leq 0.05$)

Figure 4.21 displays the distribution of unstandardized estimated effects for the 4:1 variance ratio. The trends seen in Figure 4.20 are present in the results found in Figure 4.21, but they are intensified. Far fewer effects were detected relative to those detected under homogeneous conditions. Further, the inflation at the low and moderate

sample sizes is greater than observed in Figure 4.20. The inflation effect is also present at the $n = 500$ sample size but is moderated with an increase of population effect. The effect observed in Figure 4.20 where MESP 2 eliminated effects that are greater than the unstandardized MPSD is seen at all population effect sizes but is more extreme with small and moderate population effect sizes.

Figure 4.22 displays histograms of the power ratios for all linear models with skewed and bimodal distributions by method. The patterns are no different from those observed with the normal distribution. There does not seem to be any difference in how each method responds with skewed or bimodal distribution compared with the normal distribution results.

Table 4.11 and Table 4.12 display the realized power for all models using a skewed and bimodal distribution. The power rates under both conditions do not differ notably from each other nor from the rates under the normality condition. They are included here in the interest of thoroughness. Figure 4.23 displays power rates for these conditions as bar plots.

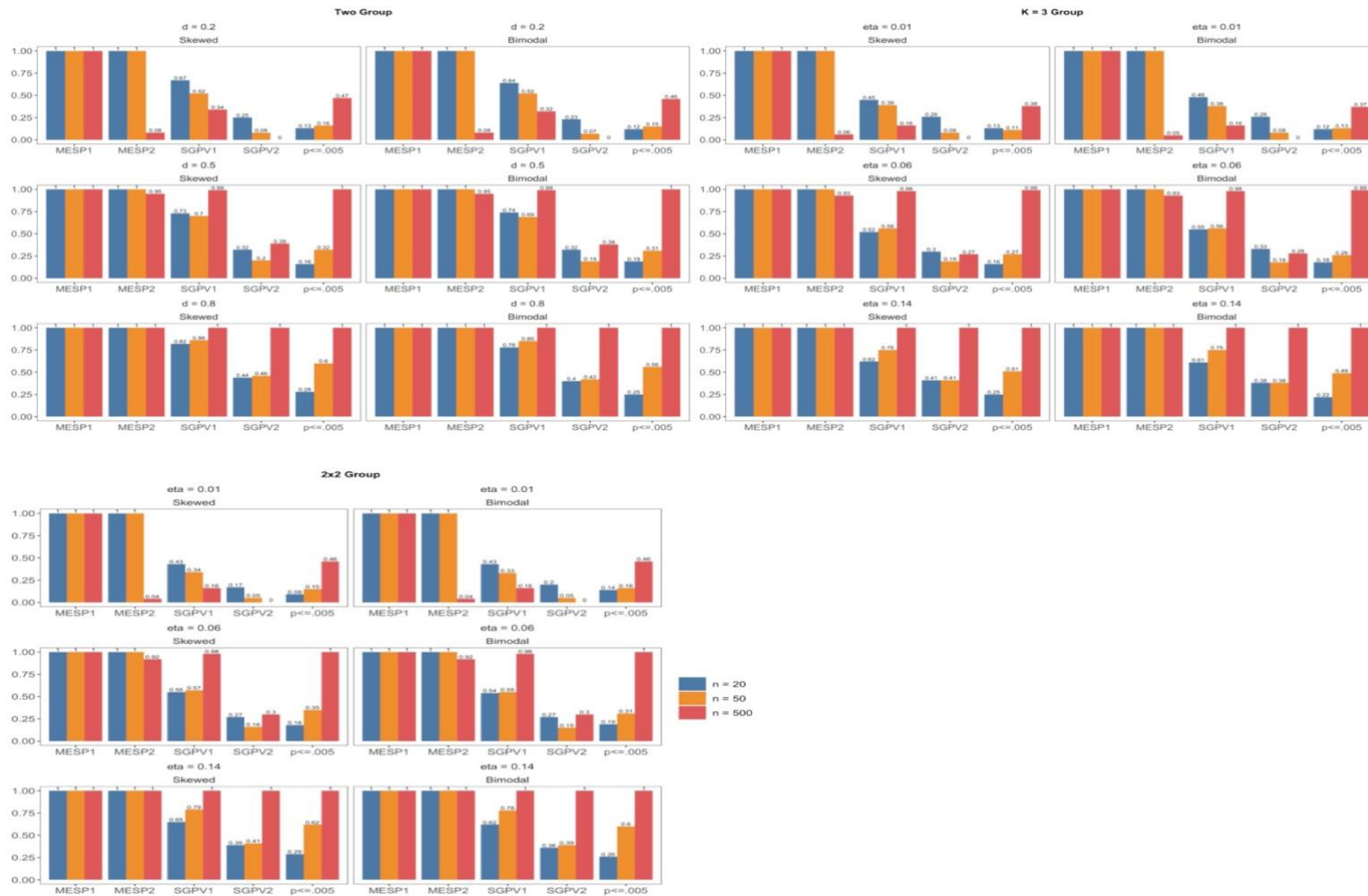


Figure 4.22 Power Ratios by Linear Model, Population Effect Size, and Variance Ratio

Table 4.11 Power Under Skewed Distribution Condition

Linear Model	Pop effect d/η^2	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	.2	20	0.0673	0.0086	0.0450	0.0169	0.0673	0.0673
		50	0.1051	0.0171	0.0548	0.0080	0.1051	0.1051
		500	0.6039	0.2815	0.2031	0.0001	0.6039	0.0471
	.5	20	0.1918	0.0309	0.1393	0.0611	0.1918	0.1918
		50	0.4061	0.1315	0.2843	0.0808	0.4061	0.4061
		500	0.9998	0.9973	0.9931	0.3861	0.9998	0.9533
	.8	20	0.3964	0.1122	0.3239	0.1760	0.3964	0.3964
		50	0.7937	0.4724	0.6829	0.3648	0.7937	0.7937
		500	1.0000	1.0000	1.0000	0.9989	1.0000	1.0000
Three-group	.01	20	0.0614	0.0081	0.0275	0.0161	0.0614	0.0614
		50	0.0784	0.0090	0.0302	0.0060	0.0784	0.0784
		500	0.4527	0.1703	0.0729	0.0000	0.4527	0.0261
	.06	20	0.1488	0.0231	0.0775	0.0446	0.1488	0.1488
		50	0.3237	0.0874	0.1827	0.0626	0.3237	0.3237
		500	0.9998	0.9943	0.9778	0.2729	0.9998	0.9348
	.14	20	0.3138	0.0778	0.1955	0.1285	0.3138	0.3138
		50	0.7062	0.3568	0.5301	0.2899	0.7062	0.7062
		500	1.0000	1.0000	1.0000	0.9973	1.0000	1.0000
2x2 between-group	.01	20	0.0746	0.0070	0.0322	0.0129	0.0746	0.0746
		50	0.1100	0.0160	0.0370	0.0055	0.1100	0.1100
		500	0.6155	0.2839	0.0995	0.0000	0.6155	0.0235
	.06	20	0.1931	0.0351	0.1053	0.0520	0.1931	0.1931
		50	0.4322	0.1519	0.2481	0.0706	0.4322	0.4322
		500	0.9997	0.9977	0.9825	0.2972	0.9997	0.9206
	.14	20	0.4111	0.1196	0.2679	0.1600	0.4111	0.4111
		50	0.8200	0.5071	0.6513	0.3334	0.8200	0.8200
		500	1.0000	1.0000	1.0000	0.9957	1.0000	1.0000

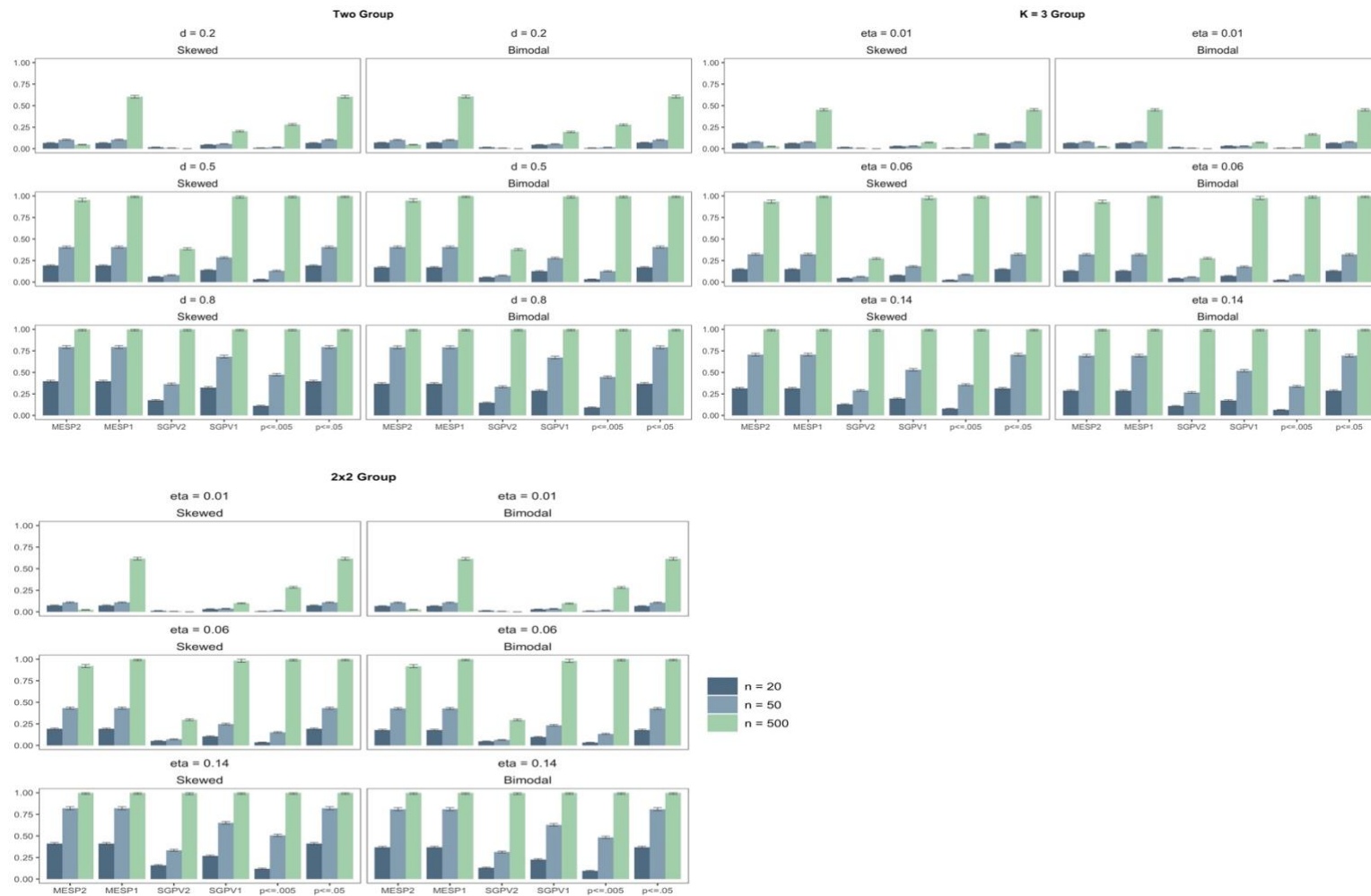


Figure 4.23 Power Rates with CI by Linear Model, Distribution, and Population Effect Size

Table 4.12 Power Under Bimodal Distribution Condition

Linear Model	Pop effect d/η^2	n	$p \leq 0.05$	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP 2
Two-group	.2	20	0.0703	0.0083	0.0453	0.0161	0.0703	0.0703
		50	0.1029	0.0155	0.0530	0.0068	0.1029	0.1029
		500	0.6059	0.2782	0.1943	0.0000	0.6059	0.0466
	.5	20	0.1711	0.0322	0.1264	0.0553	0.1711	0.1711
		50	0.4070	0.1258	0.2790	0.0767	0.4070	0.4070
		500	0.9999	0.9966	0.9934	0.3795	0.9999	0.9478
	.8	20	0.3693	0.0930	0.2885	0.1481	0.3693	0.3693
		50	0.7908	0.4460	0.6725	0.3322	0.7908	0.7908
		500	1.0000	1.0000	1.0000	0.9992	1.0000	1.0000
Three-group	.01	20	0.0644	0.0076	0.0307	0.0167	0.0644	0.0644
		50	0.0795	0.0100	0.0302	0.0066	0.0795	0.0795
		500	0.4515	0.1671	0.0730	0.0000	0.4515	0.0242
	.06	20	0.1308	0.0234	0.0715	0.0432	0.1308	0.1308
		50	0.3201	0.0838	0.1789	0.0580	0.3201	0.3201
		500	0.9996	0.9931	0.9762	0.2765	0.9996	0.9327
	.14	20	0.2876	0.0629	0.1745	0.1095	0.2876	0.2876
		50	0.6946	0.3375	0.5193	0.2664	0.6946	0.6946
		500	1.0000	1.0000	1.0000	0.9967	1.0000	1.0000
2x2 between-group	.01	20	0.0665	0.0093	0.0287	0.0136	0.0665	0.0665
		50	0.1081	0.0176	0.0354	0.0058	0.1081	0.1081
		500	0.6138	0.2817	0.0968	0.0000	0.6138	0.0253
	.06	20	0.1793	0.0335	0.0973	0.0476	0.1793	0.1793
		50	0.4272	0.1334	0.2339	0.0631	0.4272	0.4272
		500	1.0000	0.9980	0.9805	0.2957	1.0000	0.9192
	.14	20	0.3682	0.0952	0.2265	0.1314	0.3682	0.3682
		50	0.8097	0.4834	0.6279	0.3131	0.8097	0.8097
		500	1.0000	1.0000	1.0000	0.9971	1.0000	1.0000

Note: SGPV1 and MESP1 use $d = .1$ or $\eta^2 = .005$; SGPV2 and MESP2 use $d = .35$ or $\eta^2 = .035$

Table 4.13 Summary of Results for Power

Independent Variable	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP2
Linear Model	Linear model had little effect on power rates relative to NHST.	The two-group model had higher power ratios than the three-group model or 2x2 between-group models at all effect sizes and sample sizes.	Linear model had little effect on power rates relative to NHST.	No effect of linear model on power rates relative to NHST.	No effect of linear model on power rates relative to NHST.
Effect Size	Very low power relative to NHST $p \leq 0.05$ (0.12 - 0.59 ratio) except with moderate or large effects with $n = 500$. Power increased with increase in effect size.	Inverse relation of power ratios and sample size for pop. effect sizes near or within the null interval. Direct relationship of power ratios and sample size with pop. effect size outside the null interval.	Inverse relation of power ratios and sample size for low pop. effect. Power ratios displayed a “U” shaped pattern with moderate population effect, despite real power calculations increasing with sample size. Possible direct relationship of power ratios and sample size with large pop. effect size.	Power tracked precisely with NHST $p \leq 0.05$.	Power identical with NHST $p \leq 0.05$ at the low and moderate sample size. Very low power ratios (0.04 – 0.09) with low effect and $n = 500$. High power ratios (0.92 – 0.95) with moderate pop. effect and $n = 500$. Power ratios of 1 with large effect and $n = 500$.

(Continued)

Table 4.13 Summary of Results for Power

Independent Variable	$p \leq 0.005$	SGPV 1	SGPV 2	MESP 1	MESP2
Sample Size	Very low power relative to NHST $p \leq 0.05$ (0.12 - 0.59 ratio) except with $n = 500$ and moderate or large.. Power increased with increase in sample size.	Power ratios showed an inverse relation of detection rates and sample size for pop. effect sizes near or within the null interval. Direct relationship of detection rates and sample size with pop. effect size outside the null interval.	Inverse relation of power ratios and sample size for low pop. effect. Power ratios had “U” shaped pattern with moderate population effect, despite real power calculations increasing with sample size. Possible direct relationship of power ratios and sample size with large pop. effect size.	Power tracked precisely with NHST $p \leq 0.05$.	Power identical with NHST $p \leq 0.05$ at the low and moderate sample size. Only $n = 500$ had power ratios lower than 1, and only for small and moderate effect sizes.
Variance Ratio	Generally lower power ratios compared to the normal homogeneous condition. Differences were greatest with a large sample size and increased with increased heterogeneity.	Marginally lower power ratios at low and moderate effects and sample sizes. Much lower power ratios are $n = 500$ which increased with larger effect sizes and increase heterogeneity.	Generally lower power ratios compared to the normal homogeneous condition. Differences were greatest with a large sample size and increased with increased heterogeneity.	Power tracked precisely with NHST $p \leq 0.05$.	Power identical with NHST $p \leq 0.05$ at the low and moderate sample size.
Population Distribution	Power did not differ notably from normality condition.	Power did not differ notably from normality condition.	Power did not differ notably from normality condition.	Power did not differ notably from normality condition.	Power did not differ notably from normality condition.

Conclusion

This study provides helpful information for evaluating the functioning of select methods intended to improve upon the NHST methodological framework. Broadly there are three elements related to each method for which the study sought to provide some clarity. Among the most fundamental is whether a given method does indeed function as intended. The second is how the Type I Error and power results of a method compare with NHST. Lastly, the study looked at whether the methodologies functioned differently or offered any advantage over NHST under conditions that violated assumptions.

The more stringent NHST ($p \leq 0.005$) method performed as expected. The purpose of the method is to respond to the reproducibility crisis by reducing the rate of false inferences relative to NHST ($p \leq 0.005$). Success in this can be suggested by looking at the Type I error rates and real power results. Type I error rates held consistently at 0.005 or less under normal conditions with no pattern of difference between linear models. There was some slight inflation of Type I error when homogeneity was violated and even less with violations of distribution assumptions that closely matched those of the normal condition. The inflation proportionally tracked with the same type of inflation using NSHT ($p \leq 0.05$), indicating no real advantage or disadvantage over conventional NHST when assumptions of homogeneity are violated. Real power was predictably lower, which should not necessarily be considered a detriment. The practical result of using this method for researchers is that large sample sizes will not only be encouraged but will be necessary to detect an effect.

In nearly every condition, with one notable exception, MESP mirrored exactly the results of NHST ($p \leq 0.05$). MESP establishes a practically meaningful effect size then

uses a two-criteria decision rule using the MPSD together with $p \leq 0.05$ to select meaningful findings. However, in the simulation, when a finding had a p value less than 0.05, it also nearly always had an effect greater than the practically meaningful effect size. Only with a very large sample size, $n = 500$, a moderate minimum practically significant difference, $d = .35/\eta^2 = .035$, and a low or moderate population effect did MESP produce results different from NHST ($p \leq 0.05$). Type I error and power were lower in these instances since MESP functioned to eliminate findings with effects below the MPSD. The evidence presents serious questions regarding the practical usefulness of MESP for applied researchers in the social sciences. Moderate to large effects are unlikely to be considered trivial or practically non-significant in the field of social science. Therefore, if MESP functions nearly identically to NHST ($p \leq 0.05$) except under conditions of extremely large sample sizes or the selection of large MPSD, it is unlikely to improve upon NHST in most research settings.

The second-generation p -value effectively eliminates practically meaningless effects while also improving on NHST in other respects. For instance, Type I error rates were lower than those of NHST regardless of condition with both sizes of null interval tested. Slight inflation of Type I Error was observed when homogeneity was violated, and as with all other methods, the inflation proportionally tracked with that observed with NSHT ($p \leq 0.05$). Importantly, the Type I error rate decreased inversely proportionate to the increase of sample size. This effect is caused by the increase in precision gained from larger samples which manifests as narrower interval estimates. Because SGPV is a measure of the overlap of the null interval and interval estimate, it directly reflects this narrowing of the interval in lower Type I error rates. Further, SGPV had Type I error

rates lower than NHST when using either of the two selected effect sizes to construct the null interval. This contrasts with MESP, which tracked exactly with NHST when a small effect was used as the MPSD, a difference accounted for by the use of a null interval in SGPV instead of a cutoff value. A similar effect was observed in the SGPV's ability to detect an effect when one was present; that is, realized power increased with the sample size. However, when the population effect was close to but outside of the null interval bound, the ability to detect the effect was low even with a large sample size, $n = 500$. Further, the detection of these effects decreased with sample size relative to NHST($p \leq 0.05$). This again is a consequence of the interval overlap method of SGPV. If a population effect is close to the null interval bound, it is likely that the interval estimate will both contain the population effect and have some overlap with the null bound. This overlap will diminish with the sample size. This difficulty in determining the presence of marginally non-trivial effects is mitigated somewhat by the ability of the SGPV statistic to allow for consideration of the percent of overlap of the null interval and the interval estimate, an ability not investigated in this simulation study. When the reverse was true, that the population effect was marginally within the null interval, the SGPV did a very good job of identifying the effect as trivial. It should be noted that in the simulation, the distance of the population effects for the two tested null intervals is not the same. For example, the small effect $d = .2$ was .1 standard effect outside the null when the bound was $d = [-.1; .1]$ but was .15 within the null when the bound was $d = [-.35; .35]$. SGPV had a large decrease in the ability to detect an effect when homogeneity was violated. This is the result of the decrease in precision and accuracy that are caused by the assumption violation. The reduction in real power is merely a reflection of the higher

standard of SGPV, which does not evaluate the point estimate but the overlap of null and interval estimates.

For researchers ensconced in the NHST paradigm, the more stringent NHST ($p \leq 0.005$) and the SGPV offer some methodological improvements. Both methods are, by design, somewhat limited in what they attempt to do. Neither method addressed every problem with NHST, and indeed, none of the methods here discussed attempt to. Of the two methods, SGPV is the better. SGPV's null interval approach effectively answers the problems of a nil-null and simple dichotomization based on a cut-off value. The null interval also helps in mitigating the problem of large sample sizes finding significant but trivial effects. It filters out trivial effects, as defined by the null interval while maintaining a low Type I error rate that, though not constant and decreases with increasing sample sizes, is consistently lower than NHST ($p \leq 0.05$). Even more, it requires researchers to think deeply about what constitutes trivial effects, something that is a limitation in a context-free simulation study such as this one. Deep consideration of the practical meaning of effects is not needed when using a dichotomous decision rule regardless of the cutoff selected. Ultimately, no one method is a panacea. It is hoped that this small study can contribute to an understanding of the working of some of the tools available.

CHAPTER 5

CONCLUSION

The NHST research paradigm was adopted in the social sciences primarily to gain the credibility and respect associated with being a “true” scientific discipline such as the physical sciences. Group experimental designs, quantitative measures, and complex statistical procedures were thought to provide the necessary rigor and objectivity to achieve this goal. However, the limitations of the paradigm and misuse by some researchers have resulted in a replication crisis and an undermining of the credibility of social science research. These developments might seem to be a cause for despair, but they offer an excellent opportunity for social science researchers. Following the ASA statement’s advice, they should increasingly feel free to experiment with alternative methods. This experimentation can and should draw from Bayesian, nonparametric, non-inferential, and other methodologies appropriate to a given research design and the questions a researcher hopes to address. Of course, this freedom also comes with the burden of thinking much more deeply about design and analysis than researchers steeped in the NHST paradigm might be accustomed.

This dissertation attempts to look closely at select alternative methods to understand their purpose and how they compare to NHST. This is accomplished by demonstrating seven methodologies and further investigating methods with decision rules using a Monte-Carlo simulation. The methods fall into three broad categories

interpretational aids, alternate decision criteria, and methods that eschew significance testing and statistical inference.

Interpretational Aids

Interpretational aid methods can help interpret the evidence provided by the p statistic by transforming the p value into a more easily understood statistic. The Bayes Factor Bound (BFB) and s value transform and restate the p value in different terms to better communicate how much evidence the p statistic provides. The BFB provides the maximum odds in favor of the alternative hypothesis relative to the null hypothesis assuming each hypothesis is equally likely. The s value, using information theory, is a continuous measure of bits of information against the test hypothesis. Each can also be helpful in gently reorienting researchers away from dichotomized thinking based on “significance” by encouraging researchers to convert the obtained p values into alternate expressions.

Further, such interpretational aids can be used together with other methods or in any instance when a p value is reported. One can easily imagine an s value or BFB together with the more stringent decision criteria $p \leq 0.005$ or MESP. A p value can be reported together with the SGPV or interval estimation, thus retaining the familiar p statistic and providing an interpretive aide to enhance understanding. Further, since these methods affect a transformation of the p value, they work similarly in any analysis that provides the p statistic. For example, in an ANOVA, the exact p value associated with the F statistic would be reported and then transformed into a BFB with posterior probability or an s value. The same procedure would occur for main effects and interaction effects. The process of evaluating contrasts is precisely the same.

Comparing the two methods, the BFB has some shortcomings, which the s value does not. First, it requires a determination of the prior odds of the H_0 relative to the H_A . Since this is generally unknown, Benjamin and Berger recommend assuming equal odds (Benjamin & Berger, 2019). Also, the BFB can be a confusing metric when attempting to interpret large p values. With small p values, there is an inverse relationship with the BFB. As a p value decreases, the BFB increases, indicating higher odds in favor of the H_A . Likewise, when the p value increases, the BFB provides corresponding lower odds in favor of the H_A ; but only up to a point. At approximately $p = .368$, the BFB is precisely 1, indicating no evidence for or against either hypothesis. It then increases as the p value increases, now providing the odds in favor of the H_0 . The shift can be confusing when one is unaccustomed or unaware.

The s value, by contrast, does not require any assumptions about prior odds and is a direct transformation of the p value. This can be viewed as a detriment only if one has evidence of the prior odds of H_0 relative to the H_A . Further, unlike the BFB, it is a continuous measure that increases as the evidence against the null hypothesis increases. All of this taken together, of the two interpretational aids, the s value is in most cases to be preferred over the BFB, though only marginally. Proper training on the BFB could help researchers to be aware of the shift in the BFB scale. More importantly, if one has evidence of prior odds, then the BFB should be used instead of the s value.

Both interpretational aid methods, however, do have limitations. Most notably, they do not address some of the more fundamental problems with NHST. Poor research practices involving the violations of assumptions can persist and then use an interpretational aid to interpret potentially spurious results. Since they are derived from

the p value, they too are affected by the effects of sample sizes. Interpretational aids also do not address the problems associated with the nil-null hypothesis. Lastly, inferential claims or any claims regarding the strength of evidence are left solely to the judgment of the researcher, and though this is a limitation, it is not considered to be a negative but rather a positive feature. For example, the p value of 0.01 has an s value of 6.3. The probability of six consecutive outcomes is 1.6%. Whether this probability is considered rare and meaningful is left to the expert judgment of the researcher, who may simply report the value without comment or alternately make claims based on theory, past research, or experience.

Alternate Decision Criteria

Alternative decision criteria methods use different means than NHST ($p \leq 0.005$) to render decisions that possess greater strength of evidence. NHST ($p \leq 0.005$) reduces the decision threshold for inference from 0.05 to 0.005. The minimum effect size plus p value method uses a dual significance threshold of $p \leq 0.05$ and a minimum effect size value called the minimum practically significant difference (MPSD). The second-generation p value (SGPV) uses a summary statistic and a researcher determined null band to report the percentage of data-supported hypotheses that are also scientifically or practically null hypotheses. Based on the results of the Monte-Carlo simulation, NHST ($p \leq 0.005$) and SGPV both accomplish the goal of effectively providing more stringent decision criteria. The MESP, by contrast, performed no differently than conventional NHST in most tested conditions. Only with a large sample size, $n = 500$, a moderate minimum practically significant difference (MPSD), $d = .35 / \eta^2 = .035$ and a low or moderate population effect was there any divergence from the results of conventional

NHST. MESP then functions well at countering the increased sensitivity of very large samples, and researchers working with such samples should find MESP to be useful. However, researchers who are unlikely to consider small to moderate effects trivial or do not have access to very large samples will not find MESP a viable alternative to NHST.

The NHST ($p \leq 0.005$) method is a response to the reproducibility crisis, which has the goal of reducing the rate of false inferences. If implemented, it should have this effect. Indeed, the results from the Monte-Carlo simulation show that for the linear models employed, the Type I error rates hold reasonably well at or near 0.005 even under violations of assumptions of normality and homogeneity of variance. Real power, however, was low, never exceeding .47, under any condition except with moderate or large effects and a large sample size. The practical result of reduced power for researchers is that large sample sizes will be necessary to achieve the needed power to find an effect in most cases. Table 5.1 provides an example of the difference in power for the tested population effect sizes and sample sizes for the two-group linear model.

Table 5.1 Example of Power by Pop Effect and Sample Size

Pop effect d	n	$p \leq 0.05$	$p \leq 0.005$
.2	20	0.0700	0.0086
	50	0.1066	0.0160
	500	0.6015	0.2833
.5	20	0.1729	0.0291
	50	0.4045	0.1339
	500	0.9999	0.9974
.8	20	0.3841	0.1054
	50	0.7868	0.4673
	500	1.0000	1.0000

Notably, even with an n of 500, a very large sample for social science research, a small effect of $d = .2$ is accurately detected only 28% of the time compared with 60% with $p \leq 0.05$. Also of note are the large gaps between an n of 50 and 500 with

$p \leq 0.005$, a result and limitation of the chosen experimental design. A follow-up simulation was conducted to determine the needed sample size to achieve 0.8 power at each of the three effect sizes, the results are in Table 5.2. The required sample to achieve .8 power with a $p \leq 0.005$ cut-off value is at least 63% greater than that needed with $p \leq 0.05$. As discussed in the review of literature, even though such a requirement may be disruptive to research programs that find it challenging to attain large samples, the benefits of reducing Type I error outweigh the loss of research programs that render spurious results. Given all that, using this method still perpetuates the use of NHST and is, at best, a stop-gap solution.

Table 5.2 Approximate Sample Size for 80% Power by Population Effect Size

Pop effect d	$p \leq 0.05$	$p \leq 0.005$
.2	800	1350
.5	130	230
.8	55	90

The second-generation p -value is among the more promising alternate decision methods evaluated. The SGPV's null interval approach effectively answers the problems of a nil-null and simple dichotomization based on a cut-off value. The null interval also helps mitigate the problem of large sample sizes finding significant but trivial effects. This effect was seen in the Monte-Carlo simulation results. Real power increased with the sample size; however, when the population effect was close to but outside of the null interval bounds, the ability to detect an effect was low even with a large sample size, $n=500$, an indication that the SGPV is conservative in effect detection when the effect is near the null bound. The SGPV also offers a clearer metric than the p value which does an adequate job of summarizing the relationship of the null interval to the data-based interval estimate. The delta gap statistic, by contrast, is not as intuitive but does offer a

means of comparing the strength of evidence provided with an $SGPV = 0$. One drawback is that the effectiveness of $SGPV$ hinges in large part on the selection of the null interval and the prior identification of what constitutes trivial effects. The $SGPV$ also does not address all the major critiques of $NHST$; assumption violations and the question of aggregation are left aside.

Deemphasize or Abandon Significance Testing

The interval estimate approach can avoid the problems of a nil-null and simple dichotomization based on a cut-off value when interpreted carefully. However, any interval can be interpreted dichotomously around the nil-null, and researchers accustomed to $NHST$ may intuitively interpret them in such a manner. The use of intervals can further aid in avoiding misinterpretation or unwarranted claims based upon the p value of a single test hypothesis. Considering the range of compatible hypotheses can also help researchers consider the relative uncertainty of their findings. This uncertainty is amplified by re-conceptualizing p values as indicators of model fit, taking into account all of the assumptions, known and unknown, that affect the model. The use of an interval estimate approach has much in common with $SGPV$. Though it does not require a determination of a null interval, one can imagine selecting a null interval or even reporting an $SGPV$ along with an interval estimate analysis. Like the other methods discussed thus far, the interval interpretation method does not address assumption violations and the question of aggregation.

OOM attempts to address the most fundamental issues associated with $NHST$. It avoids the assumptions of $NHST$, and since it uses descriptive statistics is generally assumption-free. OOM abandons inference to parameters in favor of identifying and

predicting patterns in the data. There is no null hypothesis, nil or otherwise, that is tested. Sample size also has nearly no bearing on analysis; in fact, single-subject designs can be analyzed with certain OOM procedures. However, it should be noted that the sample size does affect the usefulness of the c-value statistic. Small data sets allow fewer random permutations from which comparisons may be obtained. As a result, the c-value is less informative as to the probability of the obtained outcome. Finally, the focus of OOM is on the patterns of individual observations, not on the value of means or variances; the problem of aggregation is then avoided.

Some researchers accustomed to NHST or other more advanced analyses may find OOM lacking sophistication. This, in part, is by design as OOM is specifically developed to be understandable and valuable to applied researchers. Further, the developers of the method consider the statistical simplicity of OOM to be a strength. This opinion results from the critique that statistical analyses too often rely upon advanced procedures when simpler methods could address the research questions just as well.

Limitations

Both studies contained in this dissertation are limited by the inclusion of only certain of the available alternative methods to NHST. Though a guiding criterion was used, it does not preclude the likelihood that another researcher would have, with good reason and much justification, chosen other alternative methods. The selection of methods must then be viewed as limited and somewhat idiosyncratic. A further limitation is that only those methods having a decision rule were used in the Monte Carlo simulation study. A more clever researcher may have devised an outcome variable that

would have allowed for the inclusion of the interpretational aid and interval estimation methods, but this researcher could not.

Future Research

Devising an outcome variable other than statistical power or Type I error, concepts intimately connected with the NHST paradigm, could allow not only a comparison of methods without a decision rule but also offer deeper insights for comparison with NHST. Another opportunity is related to the ability of SGPV to detect effects that are near the null band. The Monte Carlo methods study showed that SGPV is conservative in detecting an effect that is near but just outside of a narrow null band and that the ability to detect the effect increases with sample size. A question worth exploring is whether this trend generally holds, for instance, with large effects just outside of a wide null band. Lastly, as has been seen with the persistence of NSHT, there is a strong human element in research that should not be ignored. Another opportunity is to investigate whether researchers are using alternative methods and how they are using them. Such a study might be possible using published research articles and text mining techniques.

Discussion

Researchers are increasingly becoming aware of the problems with NHST and p values. Necessarily, many questions exist about what the future holds for statistical analysis and how to proceed going forward. Several viable alternatives and supplements are discussed in this dissertation; each offers something that improves upon NSHT. There is, however, the possibility of misuse of these methods, whether voluntary or in ignorance, just as there has been a misuse of NHST. Interpretational methods, for instance, are misused unless they are reported with the attained continuous p value.

Merely reporting whether a p value is less than a given cut-off is insufficient even if an interpretational method is used. That being said, if a researcher is going to the trouble of using an interpretational aid, he will almost certainly understand the need to report the attained p value. More concerning is the possibility of misuse of the alternate decision criteria methodologies. Because each method establishes something like a decision rule, the temptation exists to make overwrought claims of certainty much like those made with NHST. It is important, therefore, to avoid the language of certainty and instead discuss findings in terms of evidence provided by the given sample. This is likely more naturally done with SGPV because of its null band, but great diligence is needed when using with NHST ($p \leq 0.005$) or MESP. Those trained in the NHST paradigm must also be disciplined when using the interval estimation method not to interpret intervals dichotomously. For instance, simply because an interval contains a value of 0, it does not mean that there was no effect. It only indicates that the value is one of several hypotheses compatible with the sampled data and should be discussed as such.

Also important to consider is that nearly all of the methods chosen for this dissertation rest upon an NHST framework. Thus, those methods are beholden to the same set of assumptions as NHST, the violation of which is reflected in the estimates and summary statistics of those methods. These statistics, like the p value, reflect in an undifferentiated manner not only random variation but also violations of model assumptions. Only the interval estimation method and OOM attempt to deal with the problem of violation of assumptions, though both do this in markedly different ways. As presented by Amrhein et al. (2019), interval estimation encourages the interpretation of the p value as an unstable local description of relationships between the model and the

obtained data. Simply, it is interpreted as a model fit statistic. Observation Oriented Modeling is the sole method that does not rest upon the NHST framework and is designed partly to address the problem of violation of assumptions. This is accomplished via OOM's reliance upon descriptive statistics and randomization tests. It should be noted that sample size does affect the usefulness of the c -value statistic derived from the randomization tests. Small data sets allow for fewer random permutations from which comparisons may be obtained. As a result, the c -value is less informative as to the probability of the obtained outcome.

Ultimately, whichever method researchers use largely depends on their ability to move away from the NSHT paradigm. Methods such as OOM, interval estimation, as presented by Amrhein et al. (2019), or statistical model construction and testing, which was discussed in the review of literature, all require a major philosophical shift. The other methods reviewed generally work within the NHST framework. Yet, those with subjective elements such as SGPV and MESP force deeper thinking about effects than the ritualized NHST process. Ultimately researchers need to feel free to explore alternatives and not feel beholden to any one methodology.

Following the recommendations of the ASA, being humble and embracing uncertainty can allow for the use of multiple methodologies in the analysis of data, even if they do not render a definitive or consistent assessment of the findings. A simple example of multiple methods, which could be easily implemented and one from which researchers who are reluctant to move too far from NHST would benefit, is using an interpretation aid, such as the s value, together with a continuous p value. Even better, a researcher could use an interval estimate to discuss the range of data-supported hypotheses and

could still report a point estimate with its p value together with an s value. These two examples are likely best suited to new research or research lacking strong predictive theory. However, if a researcher has adequate existing evidence or a predictive theoretical model and is using NHST based tests, there is no reason not to use the SGPV, a method superior to NHST ($p \leq 0.05$). Because SGPV is a summary statistic of the overlap of a null interval with the data-supported interval, it would pair naturally with the interval estimation method's discussion of data-supported hypotheses. Further, suppose a researcher wished to explore both the aggregate effect of a treatment and how that effect was manifested at the individual level. In that case, OOM could be used together with any number of combinations of the other methods which analyze the outcomes of aggregate-based statistical tests. Such freedom would undoubtedly be an improvement over the reliance on a single statistic and overwrought claims of a finding. It is hoped that this dissertation, in some small way, contributes to the discussion of alternative methodologies.

REFERENCES

- Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97(1), 129–133. <https://doi.org/10.1037/0033-2909.97.1.129>
- ACT. (2021, June 24). *National Norms for ACT Test Scores Reported During the 2020-21 Reporting Year*. <https://www.act.org/content/dam/act/unsecured/documents/MultipleChoiceStemComposite.pdf>
- Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(Sup1), 262–270. <https://doi.org/10.1080/00031305.2018.1543137>
- Arocha, J. F. (2020). Scientific realism and the issue of variability in behavior. *Theory and Psychology*, 1–14. <https://doi.org/10.1177/0959354320935972>
- Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29(2), 189–194.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423–437. <https://doi.org/10.1037/h0020412>
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543–554.

- <https://doi.org/10.1177/1745691612459060>
- Bastardi, A., Uhlmann, E., & Ross, L. (2011). Wishful thinking: Belief, desire, and the motivated evaluation of scientific evidence. *Psychological Science*, 22, 731–732. <https://doi.org/10.1177/0956797611406447>
- Ben-Shachar M., Lüdtke D., & Makowski D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, 5(56), 2815. doi: 10.21105/joss.02815
- Benjamin, D. & Berger, J. (2019). Three recommendations for improving the use of *p*-values. *The American Statistician*, 73(Sup1), 186–191. <https://doi.org/10.1080/00031305.2018.1543135>
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., ... Zinman, J. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10. <https://doi.org/10.1038/s41562-017-0189-z>
- Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73(sup1), 291-295. <https://doi.org/10.1080/00031305.2018.1518270>
- Blume, J., Greevy, R., Welty, V., Smith, J., & DuPont, W. (2019). An Introduction to second generation *p*-value. *The American Statistician*, 73(Sup1), 157–167. <https://doi.org/10.1080/00031305.2018.1537893>
- Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time? *Journal of Research in Personality*, 32(2), 202–221. <https://doi.org/10.1006/jrpe.1997.2206>

- Boster, F. J. (2002). On making progress in communication science. *Human Communication Research*, 28(4), 473–490. <https://doi.org/10.1093/hcr/28.4.473>
- Bradley, J. W. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, 31, 147–150. <https://doi.org/10.2307/2683535>
- Brodeur A., Le, M., Sangnier, M., & Zylberberg, Y. (2016). Star Wars: The empirics strike back. *American Economic Journal: Applied Economic*, 8(1), 1–32. <https://doi.org/10.1257/app.20150044>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14, 365–376. <https://doi.org/10.1038/nrn3475>
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48(3): 378–399. <https://doi.org/10.17763/haer.48.3.t490261645281841>
- Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology*, 56(1), 423–452. <https://doi.org/10.1146/annurev.psych.56.091103.070133>
- Cervone, D., & Shoda, Y. (1999). Beyond traits in the study of personality coherence. *Current Directions in Psychological Science*, 8(1), 27–32. <https://doi.org/10.1111/1467-8721.00007>
- Chavalarias, D., Wallach, J.D., Li, A.H., & Ioannidis, J.P. (2016). Evolution of reporting *p* values in the biomedical literature, 1990–2015. *JAMA*, 315(11), 1141–1148. <https://doi.org/10.1001/jama.2016.1952>
- Chen, A. (2018, October 10). *How Accurate Are Personality Tests?* Scientific American.

- <https://www.scientificamerican.com/article/how-accurate-are-personality-tests/>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145–153.
<https://doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 50(12), 1103–1103. <https://doi.org/10.1037/0003-066x.50.12.1103>
- Colquhoun, D. (2019). The false positive risk: A proposal concerning what to do about p values. *The American Statistician*, 73(Sup1), 192–201. <https://doi.org/10.1080/00031305.2018.1529622>
- Coursol, A., Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice*, 17(2), 136–137. <https://doi.org/10.1037/0735-7028.17.2.136>
- Cumming, G. (2013). *The Problem With p Values: How Significant are They, Really?* The Conversation. <https://theconversation.com/the-problem-with-p-values-how-significant-are-they-really-20029>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Danziger, K. (1990). *Constructing the Subject*. Cambridge University Press.
- De Long, J.B., & Lang, K. (1992). Are all economic hypotheses false? *Journal of Political Economy*, 100(6), 1257–1272. <https://doi.org/10.1086/261860>

- Epstein, S. (2010). The Big Five Model: Grandiose ideas about surface traits as the foundation of a general theory of personality. *Psychological Inquiry*, 21(1), 34–39. <https://doi.org/10.1080/10478401003648682>
- Epstein, W. M. (2004). Conformational response bias and the quality of the editorial processes among American social work journals. *Research on Social Work Practice*, 14(6), 450–458. <https://doi.org/10.1177/1049731504265838>
- Erceg-Hurn, D., & Mirosevich, V. (2008). Modern, robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591–601. <https://doi.org/10.1037/0003-066x.63.7.591>
- Fanelli, D. (2010). Do pressures to publish increase scientists bias? An empirical support from U.S. states data. *PLoS ONE*, 5(4), 1–7. <https://doi.org/10.1371/journal.pone.0010271>
- Ferguson, C. J., & Heene, M. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17(1), 120–128. <https://doi.org/10.1037/a0024445>
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals but can't make them think. *Psychological Science*, 15(2), 119–126. <https://doi.org/10.1111/j.0963-7214.2004.01502008.x>
- Firestone, W. A. (1993). Alternative arguments for generalizing from data as applied to qualitative research. *Educational Researcher*, 22(4), 16–23. <https://doi.org/10.3102/0013189x022004016>
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

- Fisher, R. A. (1966). *The design of experiments* (8th ed.). Hafner Press.
- Fisher, R. A. (1955). *Statistical methods and scientific induction*. Oliver & Boyd
- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd
- Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.
- Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Oliver & Boyd.
- Fisher, R. A. (1970). *Statistical methods for research workers* (14th ed.). Hafner Press.
- Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (3rd ed.). Sage.
<https://socialsciences.mcmaster.ca/jfox/Books/Companion/>
- Gadbury, G. L., & Allison, D. B. (2012). Inappropriate fiddling with statistical analyses to obtain a desirable p value: Tests to detect its presence in published literature. *PLoS ONE*, 7(10), 1–9. <https://doi.org/10.1371/journal.pone.0046363>
- Gannon, A. G., de Branganca Pereira, C. A., Polpo, A. (2019). Blending Bayesian and classical tools to define optimal sample-size-dependent significance levels. *The American Statistician*, 73(Sup1), 213–222.
<https://doi.org/10.1080/00031305.2018.1518268>
- Garber, C. (2021). *The ANOVA for 2x2 Independent Groups Factorial Design*.
<https://psych.unl.edu/psycrs/handcomp/hcbgfact.PDF>
- Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, 71, 369–382.
<https://doi.org/10.1111/j.1751-5823.2003.tb00203.x>
- Goodman, W., Spruill, S., & Komaroff, E. (2019). A proposed hybrid effect size plus p value criterion: Empirical evidence supporting its use. *The American Statistician*, 73(Sup1), 26–30. <https://doi.org/10.1080/00031305.2018.1564697>

- Green, B. F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17, 429–440. <https://doi.org/10.1007/BF02288918>
- Greenland, S. (2019). Valid p -values behave exactly as they should: Some misleading criticisms of p -values and their resolution with s -values. *The American Statistician*, 73(Sup1), 106–114. <https://doi.org/10.1080/00031305.2018.1529625>
- Greenwald, A. G. (1975). Consequences of prejudice against null hypothesis. *Psychological Bulletin*, 82(1), 1–20. <https://doi.org/10.1037/10109-033>
- Greenwald, A. G., Gonzalez, R., Harris, R. J., Guthrie, D. (1996). Effect sizes and p values: What should be reported and what should be replicated? *Psychophysiology*, 33(2), 175–183. <https://doi.org/10.1111/j.1469-8986.1996.tb02121.x>
- Gigerenzer, G., & Murray, D. (1987). *Cognition as intuitive statistics*. Psychology Press. <https://doi.org/10.2307/2348863>
- Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195153729.001.0001>
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. <https://doi.org/10.1016/j.socec.2004.09.033>
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 33–350. <https://doi.org/10.1007/s10654-016-0149-3>
- Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. Academic Press.

- Grice, J. W., Yepez, M., Wilson, N. L., & Shoda, Y. (2017). Observation-Oriented Modeling: Going beyond “Is it all a matter of chance”? *Educational and Psychological Measurement*, 77(5), 855–867.
<https://doi.org/10.1177/0013164416667985>
- Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. John Wiley & Sons. <https://doi.org/10.2307/1271022>
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7(1), 1–20.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, 13(3).
<https://doi.org/10.1371/journal.pbio.1002106>
- Hedges, L.V., Pustejovsky, J., & Shadish, W.R. (2012). A Standardized Mean Difference Effect Size for Single-Case Designs. *Research Synthesis Methods*, 3(3), 224–239.
<https://doi.org/10.1002/jrsm.1052>
- Hedges, L.V., Pustejovsky, J., & Shadish, W.R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, 4(4), 324–341. <https://doi.org/10.1002/jrsm.1086>
- Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian journal of physical therapy*, 23(4), 290–301.
<https://doi.org/10.1016/j.bjpt.2018.12.006>

- Hubbard, R. T. (2016). *Corrupt research: The case for reconceptualizing empirical management and social science*. Sage Publications.
<https://doi.org/10.1080/1051712X.2017.1313672>
- Hubbard, R. T. (2019). Will the ASA's efforts to improve statistical practice be successful? Some evidence to the contrary. *The American Statistician*, 73(Sup1), 31–35. <https://doi.org/10.1080/00031305.2018.1497540>
- Hubbard, R., & Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports*, 80(1), 337–338. <https://doi.org/10.2466/pr0.1997.80.1.337>
- Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (ps) versus errors (α s) in classical statistical testing. *The American Statistician*, 57(3), 171–178. <https://doi.org/10.1198/0003130031856>
- Hubbard, R., Parsa, R. A., & Luthy, M. R. (1997). The spread of statistical significance testing in psychology: The case of the Journal of Applied Psychology, 1917–1994. *Theory & Psychology*, 7(4), 545–554.
<https://doi.org/10.1177/0959354397074006>
- Hubbard, R. & Ryan, P.A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects (with discussion). *Educational and Psychological Measurement*, 60(5), 661–696.
<https://doi.org/10.1177/00131640021970808>
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8(1), 3–7. <https://doi.org/10.1111/j.1467-9280.1997.tb00534.x>

Hurlbert, S., Levine, R., Utts, J. (2019). Coup de grâce for a tough old bull: ‘Statistically significant’ expires. *The American Statistician*, 73(Sup1), 353–357.

<https://doi.org/10.1080/00031305.2018.1543616>

International Committee of Medical Journal Editors. (1997). Uniform requirements for manuscripts submitted to biomedical journals. *Annals of Internal Medicine*, 126(1), 36–47. [https://doi-org.pallas2.tcl.sc.edu/10.7326/0003-4819-126-1-](https://doi-org.pallas2.tcl.sc.edu/10.7326/0003-4819-126-1-199701010-00006)

[199701010-00006](https://doi-org.pallas2.tcl.sc.edu/10.7326/0003-4819-126-1-199701010-00006)

International Committee of Medical Journal Editors. (2010). Uniform requirements for manuscripts submitted to biomedical journals. *Journal of Pharmacology and Pharmacotherapeutics*, 1(1), 42–58.

<https://login.pallas2.tcl.sc.edu/login?url=https://www.proquest.com/scholarly-journals/uniform-requirements-manuscripts-submitted/docview/851251579/se-2?accountid=13965>

Ioannidis, J. P. (2019). What have we (not) learnt from millions of scientific papers with *p* values? *The American Statistician*, 73(Sup1), 20–25.

<https://doi.org/10.1080/00031305.2018.1447512>

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4(3), 245–253.

<https://doi.org/10.1177/1740774507079441>

John L.K., Loewenstein G., Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532. <https://doi.org/10.1177/0956797611430953>

- Johnson, V. E. (2013). Revised standards for statistical evidence. *The Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.
<https://doi.org/10.1073/pnas.1313476110>
- Kaufman, A. (1998). Introduction to the special issue on statistical significance testing. *Research in the Schools*, 5(2), 1–2.
- Kerr, S., Tolliver, J., Petree, D. (1977). Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1), 132–141. <https://doi.org/10.5465/255467>
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petosky, M. D., Keselma, J. C., Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386. <https://doi.org/10.3102/00346543068003350>
- Kiess, H. O., & Bloomquist, D. W. (1985). *Psychological research methods: a conceptual approach*. Allyn and Bacon.
- Kline, R. B. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*. American Psychological Association. <https://doi.org/10.1037/14136-000>
- Kupfersmid, J., & Fiala, M. (1991). A survey of attitudes and behaviors of authors who publish in psychology and education journals. *American Psychologist*, 46(3), 249–250. <https://doi.org/10.1037/0003-066X.46.3.249>
- Lamiell, J. T. (2019). *Psychology's misuse of statistics and persistent dismissal of its critics*. Palgrave Macmillan. <https://doi.org/10.1007/978-3-030-12131-0>

- Leschinski, C. (2019). MonteCarlo: Automatic Parallelized Monte Carlo Simulations. R package version 1.0.6. <https://CRAN.R-project.org/package=MonteCarlo>
- Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, 34(2), 171–187.
<https://doi.org/10.1111/j.1468-2958.2008.00317.x>
- Loftus, G. R. (1993). Editorial comment. *Memory and Cognition*, 21(1), 1–3.
<https://doi.org/10.3758/BF03211158>
- Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5(6), 161–171.
<https://doi.org/10.1111/1467-8721.ep11512376>
- Matthews, R. (2019). Moving toward the post $p < 0.05$ era via the analysis of credibility. *The American Statistician*, 73(Sup1), 202–212.
<https://doi.org/10.1080/00031305.2018.1543136>
- MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*, Cambridge. Cambridge University Press.
<http://www.inference.org.uk/mackay/itila/book.html>
- McShane, B., Gal, D., Gelman, A., Robert, C., & Tackett, J. (2019). Abandon statistical significance. *The American Statistician*, 73(Sup1), 235–245. <https://doi.org/10.1080/00031305.2018.1527253>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>

- Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50(3), 370–375.
https://doi.org/10.1207/s15327752jpa5003_6
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108–141.
https://doi.org/10.1207/s15327965pli0102_1
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166. <https://doi.org/10.1037/0033-2909.105.1.156>
- Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, 18(2), 112–117.
<https://doi.org/10.1111/j.1467-8721.2009.01619.x>
- Montgomery, P., & Lilly, J. (2012). Systematic reviews of the effects of preparatory courses on university entrance examinations in high school-age students. *International Journal of Social Welfare*, 21(1), 3–12.
- Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE*, 7(2), 1–7.
<https://doi.org/10.1371/journal.pone.0032734>
- National Center for Education Statistics. (2021, June 24). *Number and percentage of graduates taking the ACT test; average scores and standard deviations, by sex and race/ethnicity; and percentage of test takers with selected composite scores and planned fields of postsecondary study: Selected years, 1995 through 2018*.
https://nces.ed.gov/programs/digest/d19/tables/dt19_226.50.asp

- Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case Data. *Behavior Modification*, 39(4), 510–541.
<https://doi.org/10.1080/10705511.2011.607714>
- Nuzzo, R. (2014), Scientific Method: Statistical Errors. *Nature*, 506(7487), 150–152.
<https://doi.org/10.1038/506150a>
- Osborne, J. W. (2013) Is data cleaning and the testing of assumptions relevant in the 21st century? *Frontiers in Psychology*, 4(370), 1–18.
<https://doi.org/10.3389/fpsyg.2013.00370>
- Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation*, 98(1), 135–142.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175.
<https://doi.org/10.1080/14786440009463897>
- Pfeffer, J., & Lawler, J. (1980). Effects of job alternatives, extrinsic rewards, and behavioral commitment on attitude toward the organization: A field test of the insufficient justification paradigm. *Administrative Science Quarterly*, 25(1), 38.
<https://doi.org/10.2307/2392225>
- Pollard, P. (1993). How significant is “significance”? In G. Karen & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 449–460). Psychology Press. <https://doi.org/10.4324/9781315799582>

- Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*, 12(2), 24-30.
- Pustejovsky, J.E., Hedges, L.V., & Shadish, W.R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, 39(5), 368–393.
<https://doi.org/10.3102/1076998614547577>
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Reichardt, C., & Gollob, H. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychological Methods*, 4(1), 117–128. <https://doi.org/10.1037/1082-989X.4.1.117>
- Revelle, W. (2021) psych: Procedures for Personality and Psychological Research. R package version 2.1.6. <https://CRAN.R-project.org/package=psych>.
- Robinson, D., Hayes, A., & Couch, S. (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.7. <https://CRAN.R-project.org/package=broom>
- Rothman, K. (1998). Writing for epidemiology. *Epidemiology*, 9(3), 333–337.
<https://www.jstor.org/stable/3703065>
- Rucci, A., & Tweney, R. (1980). Analysis of variance and the ‘second discipline’ of scientific psychology: A historical account. *Psychological Bulletin*, 87(1), 166–184. <https://doi.org/10.1037/0033-2909.87.1.166>

- Savage, L. J. (1954). *The foundations of statistics*. John Wiley & Sons.
<https://doi.org/10.1002/nav.3800010316>
- Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309–316.
<https://doi.org/10.1037/0033-2909.105.2.309>
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist*, 40 (1), 73-83.
<https://doi.org/10.1037/0003-066X.40.1.73>
- Scientific method: Statistical errors - Overview of attention for news story in Nature*. (2021, February 3). Altmetric. Retrieved February, 3, 2021.
<https://www.altmetric.com/details/2115792#score>
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379–423, 623–656. DOI: 10.1002/j.1538-7305.1948.tb00917.x.
- Siegfried, T. (2010). Odds are, it's wrong: Science fails to face the shortcoming of statistics. *Science News*, 177(7), 26–29.
<https://doi.org/10.1002/scin.5591770721C>
- Siegfried, T. (2013, August 22). *Science's significant stats problem: Researchers' rituals for assessing probability may mislead as much as enlighten*. Nautilus.
<https://nautil.us/issue/4/the-unlikely/sciences-significant-stats-problem>
- Siegfried, T. (2014, February 7). *To make science better, watch out for statistical flaws*. Science News Context Blog. <https://www.sciencenews.org/blog/context/make-science-better-watch-out-statistical-flaws>

- Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(1), 1359–1366.
<https://doi.org/10.1037/e519702015-014>
- Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547.
<https://doi.org/10.1037/a0033242>
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2), 115–129. <https://doi.org/10.1037/1082-989X.1.2.115>
- Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik & J. Steiger (Eds). *What if There Were no Significance Tests?* (pp. 38–64). Hillsdale, NJ England: Lawrence Erlbaum Associates.
https://www.researchgate.net/profile/Frank_Schmidt10/publication/285020701_What_if_there_were_no_significance_tests/links/570bd5e208ae8883a1ffdc0c.pdf
- Shrout, P. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, 8(1), 1–2.
<https://doi.org/10.1111/j.1467-9280.1997.tb00533.x>
- Swaminathan, H., Rogers, H.J., & Horner, R.H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, 52(2), 213–230. <https://doi.org/10.1016/j.jsp.2013.12.002>

- Taylor, J. M. G. (1985). Measures of location of skew distributions obtained through Box-Cox transformations. *Journal of the American Statistical Association*, 80(390), 427–432. <https://doi.org/10.2307/2287909>
- Thompson, B. (1993). Forward. *The Journal of Experimental Education*, 61(4), 285–286. <https://www.jstor.org/stable/20152381>
- Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, 73(sup1), 246 –261. <https://doi.org/10.1080/00031305.2018.1518264>
- Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology*, 36(1), 1–2. <https://doi.org/10.1080/01973533.2014.865505>
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1–2. <https://doi.org/10.1080/01973533.2015.1012991>
- Valerie Welty, Rebecca Irlmeier, Thomas Stewart, Robert Greevy, Jr., Lucy D'Agostino McGowan, & Jeffrey Blume (2020). sgpv: Calculate Second-Generation p-Values and Associated Measures. R package version 1.1.0. <https://CRAN.R-project.org/package=sgpv>
- Wagenmakers, E., Gronau, Q., Dablander, F., & Erkenntnis, A. (2020). The Support Interval. *Erkenn*. <https://doi.org/10.1007/s10670-019-00209-z>
- Wasserstein, R., & Lazar, N. (2016). The ASA statement on *p* values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>

Wasserstein, R. L., Schirm A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician*, 73(Sup1), 1–19.

<https://doi.org/10.1080/00031305.2019.1583913>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R.,
Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller,
E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ...
Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*,
4(43), 1686. <https://doi.org/10.21105/joss.01686>

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical
methods in psychology journals: Guidelines and explanations. *American
Psychologist*, 54, 594-604. <https://doi.org/10.1037/0003-066X.54.8.594>

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2020).
fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R
package version 3042.83.2. <https://CRAN.R-project.org/package=fGarch>

APPENDIX A

SIMULATION FUNCTIONS R CODE

```
1. #bimodal distribution creation function
2.
3. bimodalDistFunc<-function (n,cpct, mu1, mu2, sig1, sig2) {
4.   y0 <- rlnorm(n,mean=mu1, sd = sig1)
5.   y1 <- rlnorm(n,mean=mu2, sd = sig2)
6.
7.   flag <- rbinom(n,size=1,prob=cpct)
8.   y <- y0*(1 - flag) + y1*flag
9. }
```

```
1. # Two Group Simulation Function
2.
3. # S1 = Standard deviation 1
4. # S2 = Standard deviation 2
5. # D = desired pop effect size as Cohen's d
6. # n = sample size
7. #distr = population distribution
8.
9. sim_t3<-function(S1,S2,D,n,distr=c(rnorm,rsnorm,bimodal)){
10.
11.   library(tidyverse)
12.   library(effectsize)
13.   library(sgpv)
14.   library(fGarch)
15.   library(psych)
16.
17.   options(scipen=999)
18.
19.   M1 <- 100          # Group 1
20.   M2 <- 100 + (D*S1) # Group 2
21.
22.   m <- c(M1,M2)      # means for each group
23.   # sample size for each group
24.
25.   ## generate Normal deviates for each group
26.   if(distr=="rnorm"){ #normal distribution
27.     grp1 <- rnorm(n/2,mean=m[1],sd=S1)
28.     grp2 <- rnorm(n/2,mean=m[2],sd=S2)
29.
30.     # run t-test
31.     t.test<-t.test(grp2,grp1)
32.
33.     p<-t.test$p.value
34.
35.     #Sample means
36.     sm1<-mean(grp1)
37.     sd1<-sd(grp1)
38. }
```

```

39.   sm2<-mean(grp2)
40.   sd2<-sd(grp2)
41.
42.   #Pooled sd
43.   psd<-sqrt((sd1^2 + sd2^2)/2)
44.
45.   #std Effect size estimate
46.   d<-(mean(grp2)-mean(grp1))/psd
47.
48.   #D Effect size CI
49.   D.CI_L<-t.test$conf.int[1]/psd
50.   D.CI_H<-t.test$conf.int[2]/psd
51.
52.   #Raw effect estimate
53.   raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
54.
55.   #Effect size CI
56.   CI_L<-t.test$conf.int[1]
57.   CI_H<-t.test$conf.int[2]
58.
59.   #Second generation p-value
60.   sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
61.
62.   sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
63.
64.   pop_diff<-(D*S1)
65.
66.   #return items
67.   return(list("cohens_D" =d,
68.             "D.CI_LOW"=D.CI_L,
69.             "D.CI_HI" = D.CI_H,
70.             "Pop_effect" =pop_diff,
71.             "Est_effect" = raw_effect,
72.             "CI_LOW"=CI_L,
73.             "CI_HI" = CI_H,
74.             "pvalue" = p,
75.             "sgpvalue.1"=sgp.1$p.delta,
76.             "sgp.1_d.gap"=sgp.1$delta.gap,
77.             "sgpvalue.35"=sgp.35$p.delta,
78.             "sgpv.35_d.gap"=sgp.35$delta.gap))
79. }else if(distr=="rsnorm"){           #Skewed distribution
80.
81.   grp1 <- rsnorm(n/2,mean=m[1],sd=15,xi=60)
82.   grp2 <- rsnorm(n/2,mean=m[2],sd=15,xi=60)
83.
84.   sk1<-skew(grp1)
85.   sk2<-skew(grp2)
86.
87.   # run t-test
88.   t.test<-t.test(grp2,grp1)
89.
90.   #p value
91.   p<-t.test$p.value
92.
93.   #Sample means
94.   sm1<-mean(grp1)
95.   sd1<-sd(grp1)
96.
97.   sm2<-mean(grp2)
98.   sd2<-sd(grp2)
99.
100.   #Pooled sd
101.   psd<-sqrt((sd1^2 + sd2^2)/2)
102.
103.   #std Effect size estimate

```

```

104.     d<-(mean(grp2)-mean(grp1))/psd
105.
106.     #D Effect size CI
107.     D.CI_L<-t.test$conf.int[1]/psd
108.     D.CI_H<-t.test$conf.int[2]/psd
109.
110.     #Raw effect estimate
111.     raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
112.
113.     #Effect size CI
114.     CI_L<-t.test$conf.int[1]
115.     CI_H<-t.test$conf.int[2]
116.
117.     #Second generation p-value
118.     sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
119.
120.     sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
121.
122.     pop_diff<-(D*S1)
123.
124.     #return items
125.     return(list("G1.skew"=sk1,
126.                "G2.skew"=sk2,
127.                "cohens_D" =d,
128.                "D.CI_LOW"=D.CI_L,
129.                "D.CI_HI" = D.CI_H,
130.                "Pop_effect" =pop_diff,
131.                "Est_effect" = raw_effect,
132.                "CI_LOW"=CI_L,
133.                "CI_HI" = CI_H,
134.                "pvalue" = p,
135.                "sgpvalue.1"=sgp.1$p.delta,
136.                "sgp.1_d.gap"=sgp.1$delta.gap,
137.                "sgpvalue.35"=sgp.35$p.delta,
138.                "sgpv.35_d.gap"=sgp.35$delta.gap))
139.
140.   }else if(distr=="bimodal" & D==0){ #bimodal distribution with no effect
141.
142.       #.0 (0)= 113.044 & 87
143.
144.       grp1 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=87,mu2=113.044,
145. sig1=7.5,sig2=7.5))
146.       grp2 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=87,mu2=113.044,
147. sig1=7.5,sig2=7.5))
148.
149.       # run t-test
150.       t.test<-t.test(grp2,grp1)
151.
152.       p<-t.test$p.value
153.
154.       #Sample means
155.       sm1<-mean(grp1)
156.       sd1<-sd(grp1)
157.
158.       sm2<-mean(grp2)
159.       sd2<-sd(grp2)
160.
161.       #Pooled sd
162.       psd<-sqrt((sd1^2 + sd2^2)/2)
163.
164.       #std Effect size estimate
165.       d<-(mean(grp2)-mean(grp1))/psd
166.

```

```

167.   #D Effect size CI
168.   D.CI_L<-t.test$conf.int[1]/psd
169.   D.CI_H<-t.test$conf.int[2]/psd
170.
171.   #Raw effect estimate
172.   raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
173.
174.   #Effect size CI
175.   CI_L<-t.test$conf.int[1]
176.   CI_H<-t.test$conf.int[2]
177.
178.   #Second generation p-value
179.   sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
180.
181.   sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
182.
183.   pop_diff<-(D*S1)
184.
185.   #return items
186.   return(list("SD1" = sd1,
187.             "SD2" = sd2,
188.             "cohens_D" =d,
189.             "D.CI_LOW"=D.CI_L,
190.             "D.CI_HI" = D.CI_H,
191.             "Pop_effect" =pop_diff,
192.             "Est_effect" = raw_effect,
193.             "CI_LOW"=CI_L,
194.             "CI_HI" = CI_H,
195.             "pvalue" = p,
196.             "sgpvalue.1"=sgp.1$p.delta,
197.             "sgp.1_d.gap"=sgp.1$delta.gap,
198.             "sgpvalue.35"=sgp.35$p.delta,
199.             "sgpv.35_d.gap"=sgp.35$delta.gap))
200.
201.   }else if(distr=="bimodal" & D==.2){#bimodal distribution with d = .2
202.
203.     #.0 (0)= 113.044 & 87
204.     #.2 (3) = 116.008 & 90
205.
206.
207.     grp1 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=87,mu2=113.044,
208. sig1=7.5,sig2=7.5))
209.     grp2 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=90,mu2=116.008,
210. sig1=7.5,sig2=7.5))
211.
212.     # run t-test
213.     t.test<-t.test(grp2,grp1)
214.
215.     p<-t.test$p.value
216.
217.     #Sample means
218.     sm1<-mean(grp1)
219.     sd1<-sd(grp1)
220.
221.     sm2<-mean(grp2)
222.     sd2<-sd(grp2)
223.
224.     #Pooled sd
225.     psd<-sqrt((sd1^2 + sd2^2)/2)
226.
227.     #std Effect size estimate
228.     d<-(mean(grp2)-mean(grp1))/psd
229.
230.     #D Effect size CI

```

```

230. D.CI_L<-t.test$conf.int[1]/psd
231. D.CI_H<-t.test$conf.int[2]/psd
232.
233. #Raw effect estimate
234. raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
235.
236. #Effect size CI
237. CI_L<-t.test$conf.int[1]
238. CI_H<-t.test$conf.int[2]
239.
240. #Second generation p-value
241. sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
242.
243. sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
244.
245. pop_diff<-(D*S1)
246.
247. #return items
248. return(list("SD1" = sd1,
249.             "SD2" = sd2,
250.             "cohens_D" =d,
251.             "D.CI_LOW"=D.CI_L,
252.             "D.CI_HI" = D.CI_H,
253.             "Pop_effect" =pop_diff,
254.             "Est_effect" = raw_effect,
255.             "CI_LOW"=CI_L,
256.             "CI_HI" = CI_H,
257.             "pvalue" = p,
258.             "sgpvalue.1"=sgp.1$p.delta,
259.             "sgp.1_d.gap"=sgp.1$delta.gap,
260.             "sgpvalue.35"=sgp.35$p.delta,
261.             "sgpv.35_d.gap"=sgp.35$delta.gap))
262.
263.
264. }else if(distr=="bimodal" & D==.5){ #bimodal distribution with d = .5
265.
266.     #.0 (0)= 113.044 & 87
267.     #.5 (7.5) = 120.5 & 94.5
268.
269.
270.     grp1 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=87,mu2=113.044,
271. sig1=7.5,sig2=7.5))
272.
273.     grp2 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=94.5,mu2=120.5,
274. sig1=7.5,sig2=7.5))
275.
276.
277. # run t-test
278. t.test<-t.test(grp2,grp1)
279.
280. p<-t.test$p.value
281.
282. #Sample means
283. sm1<-mean(grp1)
284. sd1<-sd(grp1)
285.
286. sm2<-mean(grp2)
287. sd2<-sd(grp2)
288.
289. #Pooled sd
290. psd<-sqrt((sd1^2 + sd2^2)/2)
291.
292. #std Effect size estimate

```

```

293.     d<-(mean(grp2)-mean(grp1))/psd
294.
295.     #D Effect size CI
296.     D.CI_L<-t.test$conf.int[1]/psd
297.     D.CI_H<-t.test$conf.int[2]/psd
298.
299.     #Raw effect estimate
300.     raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
301.
302.     #Effect size CI
303.     CI_L<-t.test$conf.int[1]
304.     CI_H<-t.test$conf.int[2]
305.
306.     #Second generation p-value
307.     sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
308.
309.     sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
310.
311.     pop_diff<-(D*S1)
312.
313.     #return items
314.     return(list("SD1" = sd1,
315.               "SD2" = sd2,
316.               "cohens_D" =d,
317.               "D.CI_LOW"=D.CI_L,
318.               "D.CI_HI" = D.CI_H,
319.               "Pop_effect" =pop_diff,
320.               "Est_effect" = raw_effect,
321.               "CI_LOW"=CI_L,
322.               "CI_HI" = CI_H,
323.               "pvalue" = p,
324.               "sgpvalue.1"=sgp.1$p.delta,
325.               "sgp.1_d.gap"=sgp.1$delta.gap,
326.               "sgpvalue.35"=sgp.35$p.delta,
327.               "sgpv.35_d.gap"=sgp.35$delta.gap))
328.
329.   }else if(distr=="bimodal" & D==.8) #bimodal distribution with d = .8
330.
331.       #.0 (0)= 113.044 & 87
332.
333.       #.8 (12) = 125 & 99
334.
335.       grp1 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=87,mu2=113.044,
336. sig1=7.5,sig2=7.5))
337.
338.       grp2 <- log(bimodalDistFunc(n=n/2,cpct =.5,mu1=99,mu2=125, sig1=7.5,sig2=7.5))
339.
340.
341.     # run t-test
342.     t.test<-t.test(grp2,grp1)
343.
344.     p<-t.test$p.value
345.
346.     #Sample means
347.     sm1<-mean(grp1)
348.     sd1<-sd(grp1)
349.
350.     sm2<-mean(grp2)
351.     sd2<-sd(grp2)
352.
353.     #Pooled sd
354.     psd<-sqrt((sd1^2 + sd2^2)/2)
355.
356.     #std Effect size estimate

```

```

357.   d<-(mean(grp2)-mean(grp1))/psd
358.
359.   #D Effect size CI
360.   D.CI_L<-t.test$conf.int[1]/psd
361.   D.CI_H<-t.test$conf.int[2]/psd
362.
363.   #Raw effect estimate
364.   raw_effect<-(t.test$estimate[1]-t.test$estimate[2])%>%unnname
365.
366.   #Effect size CI
367.   CI_L<-t.test$conf.int[1]
368.   CI_H<-t.test$conf.int[2]
369.
370.   #Second generation p-value
371.   sgp.1<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.1, null.hi = .1)
372.
373.   sgp.35<-sgpvalue(D.CI_L,D.CI_H,null.lo = -.35, null.hi = .35)
374.
375.   pop_diff<-(D*S1)
376.
377.   #return items
378.   return(list("SD1" = sd1,
379.             "SD2" = sd2,
380.             "cohens_D" =d,
381.             "D.CI_LOW"=D.CI_L,
382.             "D.CI_HI" = D.CI_H,
383.             "Pop_effect" =pop_diff,
384.             "Est_effect" = raw_effect,
385.             "CI_LOW"=CI_L,
386.             "CI_HI" = CI_H,
387.             "pvalue" = p,
388.             "sgpvalue.1"=sgp.1$p.delta,
389.             "sgp.1_d.gap"=sgp.1$delta.gap,
390.             "sgpvalue.35"=sgp.35$p.delta,
391.             "sgpv.35_d.gap"=sgp.35$delta.gap))
392.
393. }
394.

```

```

1.  #3 Group Simulation Function
2.
3.  #S1 = standard deviation 1
4.  #S2 = standard deviation 2
5.  #S3 = standard deviation 3
6.  #D = Desired pop effect size equivalent to Cohen's D
7.  #n = sample size
8.  #distr = Pop distribution
9.
10.
11. sim_anova<-function(S1,S2,S3,D,n,distr=c(rnorm,rsnorm,bimodal)){
12.
13.   library(tidyverse)
14.   library(sgpv)
15.   library(broom)
16.   library(stats)
17.   library(fGarch)
18.   library(effectsize)
19.   library(psych)
20.
21.   options(scipen=999)
22.
23.   M1 <- 100           # Group 1
24.   M2 <- 100           # Group 2
25.   M3 <- 100 +(D*S1)  # Group 2

```



```

26.
27. m <- c(M1,M2,M3)      # means for each group
28.
29. if(distr=="rnorm"){
30. ## generate Normal deviates for each group
31. grp1 <- rnorm(round(n/3)-1,mean=m[1],sd=S1)%>%as_tibble
32. grp2 <- rnorm(round(n/3) ,mean=m[2],sd=S2)%>%as_tibble
33. grp3 <- rnorm(round(n/3) ,mean=m[3],sd=S3)%>%as_tibble
34.
35.
36. #bind data
37. data<-bind_rows(grp1,grp2,grp3,.id="group")
38.
39. #set as factor
40. data$group <- as.factor(data$group)
41.
42. #descriptive stats: count, mean, sd
43. disc<-group_by(data, group) %>%
44.   summarise(
45.     count = n(),
46.     mean = mean(value, na.rm = TRUE),
47.     sd = sd(value, na.rm = TRUE)
48.   )
49.
50. # Analysis of variance
51. model<-aov(value ~ group, data = data)
52.
53. #Anova table
54. sum<-tidy(model)
55.
56. #eta2
57. eta_ci<-eta_squared(model,partial = T,ci=.95)
58.
59. #pvalue
60. p<-sum$p.value[1]
61.
62. #Second generation p-value limit .005 eta2
63. sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
64.
65. #Second generation p-value limit .035 eta2
66. sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
67.
68. #Second generation p-value limit .06 eta2
69. sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)
70.
71. HSD<-tidy(TukeyHSD(model))
72.
73.
74. #return items
75. return(list("Mean1" = disc$mean[1],
76.             "Mean2" = disc$mean[2],
77.             "Mean3" = disc$mean[3],
78.             "Sd1"   = disc$sd[1],
79.             "Sd2"   = disc$sd[2],
80.             "Sd3"   = disc$sd[3],
81.             "eta2"  = eta_ci$Eta_Sq_partial,
82.             "eta2.CI_LOW" = eta_ci$CI_low,
83.             "eta2.CI_HI" = eta_ci$CI_high,
84.             "pvalue" = p,
85.             "sgpvalue.005"=sgp.1$p.delta,
86.             "sgpvalue.035"=sgp.35$p.delta,
87.             "sgpvalue.06"=sgp.06$p.delta,
88.             "C_estimate_2_1" = HSD$estimate[1],
89.             "C_CONF_LOW_2_1" = HSD$conf.low[1],
90.             "C_CONF_HI_2_1"  = HSD$conf.high[1],

```

```

91.         "C_pvalue_2_1" = HSD$adj.p.value[1],
92.         "C_estimate_3_1" = HSD$estimate[2],
93.         "C_CONF_LOW_3_1" = HSD$conf.low[2],
94.         "C_CONF_HI_3_1" = HSD$conf.high[2],
95.         "C_pvalue_3_1" = HSD$adj.p.value[2],
96.         "C_estimate_3_2" = HSD$estimate[3],
97.         "C_CONF_LOW_3_2" = HSD$conf.low[3],
98.         "C_CONF_HI_3_2" = HSD$conf.high[3],
99.         "C_pvalue_3_2" = HSD$adj.p.value[3]))
100.
101. }else if(distr=="rsnorm"){
102.   grp1 <- rsnorm(round(n/3)-1,mean=m[1],sd=15,xi=60)%>%as_tibble
103.   grp2 <- rsnorm(round(n/3) ,mean=m[2],sd=15,xi=60)%>%as_tibble
104.   grp3 <- rsnorm(round(n/3) ,mean=m[3],sd=15,xi=60)%>%as_tibble
105.
106.
107.   #bind data
108.   data<-bind_rows(grp1,grp2,grp3,.id="group")
109.
110.   #set as factor
111.   data$group <- as.factor(data$group)
112.
113.   #descriptive stats: count, mean, sd
114.   disc<-group_by(data, group) %>%
115.     summarise(
116.       count = n(),
117.       mean = mean(value, na.rm = TRUE),
118.       sd = sd(value, na.rm = TRUE)
119.     )
120.
121.   # Analysis of variance
122.   model<-aov(value ~ group, data = data)
123.
124.   #Anova table
125.   sum<-tidy(model)
126.
127.   #eta2
128.   eta_ci<-eta_squared(model,partial = T,ci=.95)
129.
130.   #pvalue
131.   p<-sum$p.value[1]
132.
133.   #Second generation p-value limit .005 eta2
134.   sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
135.
136.   #Second generation p-value limit .035 eta2
137.   sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
138.
139.   #Second generation p-value limit .06 eta2
140.   sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)
141.
142.   HSD<-tidy(TukeyHSD(model))
143.
144.
145.
146.
147.   #return items
148.   return(list("Mean1" = disc$mean[1],
149.             "Mean2" = disc$mean[2],
150.             "Mean3" = disc$mean[3],
151.             "Sd1" = disc$sd[1],
152.             "Sd2" = disc$sd[2],
153.             "Sd3" = disc$sd[3],
154.             "eta2" = eta_ci$Eta_Sq_partial,
155.             "eta2.CI_LOW" = eta_ci$CI_low,

```

```

156.         "eta2.CI_HI" = eta_ci$CI_high,
157.         "pvalue" = p,
158.         "sgpvalue.005" = sgp.1$p.delta,
159.         "sgpvalue.035" = sgp.35$p.delta,
160.         "sgpvalue.06" = sgp.06$p.delta,
161.         "C_estimate_2_1" = HSD$estimate[1],
162.         "C_CONF_LOW_2_1" = HSD$conf.low[1],
163.         "C_CONF_HI_2_1" = HSD$conf.high[1],
164.         "C_pvalue_2_1" = HSD$adj.p.value[1],
165.         "C_estimate_3_1" = HSD$estimate[2],
166.         "C_CONF_LOW_3_1" = HSD$conf.low[2],
167.         "C_CONF_HI_3_1" = HSD$conf.high[2],
168.         "C_pvalue_3_1" = HSD$adj.p.value[2],
169.         "C_estimate_3_2" = HSD$estimate[3],
170.         "C_CONF_LOW_3_2" = HSD$conf.low[3],
171.         "C_CONF_HI_3_2" = HSD$conf.high[3],
172.         "C_pvalue_3_2" = HSD$adj.p.value[3]))
173.
174.   }else if(distr=="bimodal" & D==0){
175.
176.     #.0 (0)= 113.044 & 87
177.
178.     grp1 <- log(bimodalDistFunc(n=round(n/3)-1, cpct = .5, mu1=87, mu2=113.044,
179.       sig1=7.5, sig2=7.5))%>%as_tibble
180.     grp2 <- log(bimodalDistFunc(n=round(n/3) , cpct = .5, mu1=87, mu2=113.044,
181.       sig1=7.5, sig2=7.5))%>%as_tibble
182.     grp3 <- log(bimodalDistFunc(n=round(n/3) , cpct = .5, mu1=87, mu2=113.044,
183.       sig1=7.5, sig2=7.5))%>%as_tibble
184.
185.     #bind data
186.     data<-bind_rows(grp1,grp2,grp3,.id="group")
187.
188.     #set as factor
189.     data$group <- as.factor(data$group)
190.
191.     #descriptive stats: count, mean, sd
192.     disc<-group_by(data, group) %>%
193.       summarise(
194.         count = n(),
195.         mean = mean(value, na.rm = TRUE),
196.         sd = sd(value, na.rm = TRUE)
197.       )
198.
199.     # Analysis of variance
200.     model<-aov(value ~ group, data = data)
201.
202.     #Anova table
203.     sum<-tidy(model)
204.
205.     #eta2
206.     eta_ci<-eta_squared(model,partial = T,ci=.95)
207.
208.     #pvalue
209.     p<-sum$p.value[1]
210.
211.     #Second generation p-value limit .005 eta2
212.     sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
213.
214.     #Second generation p-value limit .035 eta2
215.     sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
216.
217.     #Second generation p-value limit .06 eta2
218.     sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)

```

```

218.   HSD<-tidy(TukeyHSD(model))
219.
220.
221.
222.
223.   #return items
224.   return(list("Mean1" = disc$mean[1],
225.             "Mean2" = disc$mean[2],
226.             "Mean3" = disc$mean[3],
227.             "Sd1"   = disc$sd[1],
228.             "Sd2"   = disc$sd[2],
229.             "Sd3"   = disc$sd[3],
230.             "eta2"  = eta_ci$Eta_Sq_partial,
231.             "eta2.CI_LOW" = eta_ci$CI_low,
232.             "eta2.CI_HI" = eta_ci$CI_high,
233.             "pvalue" = p,
234.             "sgpvalue.005"=sgp.1$p.delta,
235.             "sgpvalue.035"=sgp.35$p.delta,
236.             "sgpvalue.06"=sgp.06$p.delta,
237.             "C_estimate_2_1" = HSD$estimate[1],
238.             "C_CONF_LOW_2_1" = HSD$conf.low[1],
239.             "C_CONF_HI_2_1"  = HSD$conf.high[1],
240.             "C_pvalue_2_1"   = HSD$adj.p.value[1],
241.             "C_estimate_3_1" = HSD$estimate[2],
242.             "C_CONF_LOW_3_1" = HSD$conf.low[2],
243.             "C_CONF_HI_3_1"  = HSD$conf.high[2],
244.             "C_pvalue_3_1"   = HSD$adj.p.value[2],
245.             "C_estimate_3_2" = HSD$estimate[3],
246.             "C_CONF_LOW_3_2" = HSD$conf.low[3],
247.             "C_CONF_HI_3_2"  = HSD$conf.high[3],
248.             "C_pvalue_3_2"   = HSD$adj.p.value[3]))
249.
250.
251.   }else if(distr=="bimodal" & D==.201){
252.
253.     #.0 (0)= 113.044 & 87
254.     #.2 (3.015) = 116.016 & 90.016
255.
256.     grp1 <- log(bimodalDistFunc(n=round(n/3)-1,cpct =.5,mu1=87,mu2=113.044,
257.       sig1=7.5,sig2=7.5))%>%as_tibble
258.     grp2 <- log(bimodalDistFunc(n=round(n/3) ,cpct =.5,mu1=87,mu2=113.044,
259.       sig1=7.5,sig2=7.5))%>%as_tibble
260.     grp3 <- log(bimodalDistFunc(n=round(n/3) ,cpct =.5,mu1=90.016,mu2=116.016,
261.       sig1=7.5,sig2=7.5))%>%as_tibble
262.
263.     #bind data
264.     data<-bind_rows(grp1,grp2,grp3,.id="group")
265.
266.     #set as factor
267.     data$group <- as.factor(data$group)
268.
269.     #descriptive stats: count, mean, sd
270.     disc<-group_by(data, group) %>%
271.       summarise(
272.         count = n(),
273.         mean = mean(value, na.rm = TRUE),
274.         sd = sd(value, na.rm = TRUE)
275.       )
276.
277.     # Analysis of variance
278.     model<-aov(value ~ group, data = data)
279.
280.     #Anova table
281.     sum<-tidy(model)

```

```

280.   #eta2
281.   eta_ci<-eta_squared(model,partial = T,ci=.95)
282.
283.   #pvalue
284.   p<-sum$p.value[1]
285.
286.   #Second generation p-value limit .005 eta2
287.   sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
288.
289.   #Second generation p-value limit .035 eta2
290.   sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
291.
292.   #Second generation p-value limit .06 eta2
293.   sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)
294.
295.   HSD<-tidy(TukeyHSD(model))
296.
297.
298.
299.
300.   #return items
301.   return(list("Mean1" = disc$mean[1],
302.              "Mean2" = disc$mean[2],
303.              "Mean3" = disc$mean[3],
304.              "Sd1"   = disc$sd[1],
305.              "Sd2"   = disc$sd[2],
306.              "Sd3"   = disc$sd[3],
307.              "eta2"  = eta_ci$Eta_Sq_partial,
308.              "eta2.CI_LOW" = eta_ci$CI_low,
309.              "eta2.CI_HI" = eta_ci$CI_high,
310.              "pvalue" = p,
311.              "sgpvalue.005"=sgp.1$p.delta,
312.              "sgpvalue.035"=sgp.35$p.delta,
313.              "sgpvalue.06"=sgp.06$p.delta,
314.              "C_estimate_2_1" = HSD$estimate[1],
315.              "C_CONF_LOW_2_1" = HSD$conf.low[1],
316.              "C_CONF_HI_2_1"  = HSD$conf.high[1],
317.              "C_pvalue_2_1"   = HSD$adj.p.value[1],
318.              "C_estimate_3_1" = HSD$estimate[2],
319.              "C_CONF_LOW_3_1" = HSD$conf.low[2],
320.              "C_CONF_HI_3_1"  = HSD$conf.high[2],
321.              "C_pvalue_3_1"   = HSD$adj.p.value[2],
322.              "C_estimate_3_2" = HSD$estimate[3],
323.              "C_CONF_LOW_3_2" = HSD$conf.low[3],
324.              "C_CONF_HI_3_2"  = HSD$conf.high[3],
325.              "C_pvalue_3_2"   = HSD$adj.p.value[3]))
326.
327. }else if(distr=="bimodal" & D==.535825){
328.
329.   #.0 (0)= 113.044 & 87
330.   #.5 (8.037) = 121.039 & 95.039
331.
332.   grp1 <- log(bimodalDistFunc(n=round(n/3)-1,cpct =.5,mu1=87, mu2=113.044,
333.     sig1=7.5,sig2=7.5))%>%as_tibble
334.   grp2 <- log(bimodalDistFunc(n=round(n/3) ,cpct =.5,mu1=87, mu2=113.044,
335.     sig1=7.5,sig2=7.5))%>%as_tibble
336.   grp3 <- log(bimodalDistFunc(n=round(n/3) ,cpct
337.     =.5,mu1=95.039,mu2=121.039,sig1=7.5,sig2=7.5))%>%as_tibble
338.
339.   #bind data
340.   data<-bind_rows(grp1,grp2,grp3,.id="group")
341.
342.   #set as factor
343.   data$group <- as.factor(data$group)

```

```

342.
343. #descriptive stats: count, mean, sd
344. disc<-group_by(data, group) %>%
345.   summarise(
346.     count = n(),
347.     mean = mean(value, na.rm = TRUE),
348.     sd = sd(value, na.rm = TRUE)
349.   )
350.
351. # Analysis of variance
352. model<-aov(value ~ group, data = data)
353.
354. #Anova table
355. sum<-tidy(model)
356.
357. #eta2
358. eta_ci<-eta_squared(model,partial = T,ci=.95)
359.
360. #pvalue
361. p<-sum$p.value[1]
362.
363. #Second generation p-value limit .005 eta2
364. sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
365.
366. #Second generation p-value limit .035 eta2
367. sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
368.
369. #Second generation p-value limit .06 eta2
370. sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)
371.
372. HSD<-tidy(TukeyHSD(model))
373.
374.
375.
376.
377. #return items
378. return(list("Mean1" = disc$mean[1],
379.            "Mean2" = disc$mean[2],
380.            "Mean3" = disc$mean[3],
381.            "Sd1"   = disc$sd[1],
382.            "Sd2"   = disc$sd[2],
383.            "Sd3"   = disc$sd[3],
384.            "eta2"  = eta_ci$Eta_Sq_partial,
385.            "eta2.CI_LOW" = eta_ci$CI_low,
386.            "eta2.CI_HI" = eta_ci$CI_high,
387.            "pvalue" = p,
388.            "sgpvalue.005"=sgp.1$p.delta,
389.            "sgpvalue.035"=sgp.35$p.delta,
390.            "sgpvalue.06"=sgp.06$p.delta,
391.            "C_estimate_2_1" = HSD$estimate[1],
392.            "C_CONF_LOW_2_1" = HSD$conf.low[1],
393.            "C_CONF_HI_2_1"  = HSD$conf.high[1],
394.            "C_pvalue_2_1"   = HSD$adj.p.value[1],
395.            "C_estimate_3_1" = HSD$estimate[2],
396.            "C_CONF_LOW_3_1" = HSD$conf.low[2],
397.            "C_CONF_HI_3_1"  = HSD$conf.high[2],
398.            "C_pvalue_3_1"   = HSD$adj.p.value[2],
399.            "C_estimate_3_2" = HSD$estimate[3],
400.            "C_CONF_LOW_3_2" = HSD$conf.low[3],
401.            "C_CONF_HI_3_2"  = HSD$conf.high[3],
402.            "C_pvalue_3_2"   = HSD$adj.p.value[3]))
403.
404. }else if(distr=="bimodal" & D==.857)
405.
406.   #.0 (0)= 113.044 & 87

```

```

407.   #.8 (12.855) = 125.855 & 99.855
408.
409.   grp1 <- log(bimodalDistFunc(n=round(n/3)-1,cpct =.5,mu1=87,mu2=113.044,
    sig1=7.5,sig2=7.5))%>%as_tibble
410.   grp2 <- log(bimodalDistFunc(n=round(n/3) ,cpct =.5,mu1=87,mu2=113.044,
    sig1=7.5,sig2=7.5))%>%as_tibble
411.   grp3 <- log(bimodalDistFunc(n=round(n/3) ,cpct =.5,mu1=99.855,mu2=125.855,
    sig1=7.5,sig2=7.5))%>%as_tibble
412.
413.
414.   #bind data
415.   data<-bind_rows(grp1,grp2,grp3,.id="group")
416.
417.   #set as factor
418.   data$group <- as.factor(data$group)
419.
420.   #descriptive stats: count, mean, sd
421.   disc<-group_by(data, group) %>%
422.     summarise(
423.       count = n(),
424.       mean = mean(value, na.rm = TRUE),
425.       sd = sd(value, na.rm = TRUE)
426.     )
427.
428.
429.   # Analysis of variance
430.   model<-aov(value ~ group, data = data)
431.
432.   #Anova table
433.   sum<-tidy(model)
434.
435.   #eta2
436.   eta_ci<-eta_squared(model,partial = T,ci=.95)
437.
438.   #pvalue
439.   p<-sum$p.value[1]
440.
441.   #Second generation p-value limit .005 eta2
442.   sgp.1<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .005)
443.
444.   #Second generation p-value limit .035 eta2
445.   sgp.35<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .035)
446.
447.   #Second generation p-value limit .06 eta2
448.   sgp.06<-sgpvalue(eta_ci$CI_low,eta_ci$CI_high,null.lo = 0, null.hi = .06)
449.
450.   HSD<-tidy(TukeyHSD(model))
451.
452.
453.   #return items
454.   return(list("Mean1" = disc$mean[1],
455.             "Mean2" = disc$mean[2],
456.             "Mean3" = disc$mean[3],
457.             "Sd1" = disc$sd[1],
458.             "Sd2" = disc$sd[2],
459.             "Sd3" = disc$sd[3],
460.             "eta2" = eta_ci$Eta_Sq_partial,
461.             "eta2.CI_LOW" = eta_ci$CI_low,
462.             "eta2.CI_HI" = eta_ci$CI_high,
463.             "pvalue" = p,
464.             "sgpvalue.005"=sgp.1$p.delta,
465.             "sgpvalue.035"=sgp.35$p.delta,
466.             "sgpvalue.06"=sgp.06$p.delta,
467.             "C_estimate_2_1" = HSD$estimate[1],
468.             "C_CONF_LOW_2_1" = HSD$conf.low[1],

```

```

469.         "C_CONF_HI_2_1" = HSD$conf.high[1],
470.         "C_pvalue_2_1" = HSD$adj.p.value[1],
471.         "C_estimate_3_1" = HSD$estimate[2],
472.         "C_CONF_LOW_3_1" = HSD$conf.low[2],
473.         "C_CONF_HI_3_1" = HSD$conf.high[2],
474.         "C_pvalue_3_1" = HSD$adj.p.value[2],
475.         "C_estimate_3_2" = HSD$estimate[3],
476.         "C_CONF_LOW_3_2" = HSD$conf.low[3],
477.         "C_CONF_HI_3_2" = HSD$conf.high[3],
478.         "C_pvalue_3_2" = HSD$adj.p.value[3]))
479.     }

```

```

1. #2x2 Between Group Simulation Function
2.
3. #S1 = standard deviation 1
4. #S2 = standard deviation 2
5. #n = sample size
6. #gamma = interaction effect
7. #####
8. #(eta = .01; gamma = 1.511),
9. #(eta = .06; gamma = 3.8),
10. #(eta = .14; gamma = 6.05 )
11. #####
12. #n = sample size
13. #distr = Pop distribution
14.
15. n<-100
16. gamma<-6
17. S1<-15
18. S2<-15
19. sim_factor2(S1=15,S2=15,gamma=6,n=100,distr="bimodal")
20.
21.
22. sim_factor2<-function(S1,S2,gamma,n,distr=c("rnorm","rsnorm","bimodal")){
23.
24.   library(tidyverse)
25.   library(sgpv)
26.   library(effectsize)
27.   library(psych)
28.   library(broom)
29.   library(fGarch)
30.
31.   options(scipen=999)
32.
33.   mu = 100
34.   alpha = 0
35.   beta = 0
36.
37.   if(distr=="rnorm"){
38.
39.     # Generate your data using the regression equation
40.     y11 = mu + alpha + beta + gamma + (rnorm(n/4, 0, S1))
41.     y12 = mu + alpha + -(beta) + -gamma + (rnorm(n/4, 0, S1))
42.     y21 = mu + -(alpha) + beta + -gamma + (rnorm(n/4, 0, S1))
43.     y22 = mu + -(alpha) + -(beta) + gamma + (rnorm(n/4, 0, S2))
44.
45.     #Factor labels
46.     A = c(rep(c(0), n/2), rep(c(1), n/2)) # '0' 20 times, '1' 20 times
47.     B = c(rep(c(0), n/4), rep(c(1), n/4), rep(c(0), n/4), rep(c(1),n/4)) # '0'x10,
         '1'x10, '0'x10, '1'x10
48.

```



```

49. #bind score data
50. data<-rbind(y11%>as.tibble,y12%>as.tibble,y21%>as.tibble,y22%>as.tibble)
51.
52. # Join the variables in a data frame
53. data <- data.frame(cbind(data,A,B))
54. data <- data%>%select(`y`=value, A,B)
55.
56. #Descriptives
57. meanA<-data%>%group_by(A) %>%
58.   summarise(
59.     count = n(),
60.     mean = mean(y, na.rm = TRUE),
61.     sd = sd(y, na.rm = TRUE)
62.   )
63.
64.
65. meanB<-data%>%group_by(B) %>%
66.   summarise(
67.     count = n(),
68.     mean = mean(y, na.rm = TRUE),
69.     sd = sd(y, na.rm = TRUE)
70.   )
71.
72.
73. meanAB<-data%>%group_by(A,B) %>%
74.   summarise(
75.     count = n(),
76.     mean = mean(y, na.rm = TRUE),
77.     sd = sd(y, na.rm = TRUE)
78.   )
79.
80. #ANOVA model
81. model = aov(y ~ A*B, data=data)
82.
83. #Model summary
84. sum<-tidy(model)
85.
86. #effect size
87. eta_sq<-eta_squared(model,partial = T,ci=.95)
88.
89. #####
90. #####
91. #main effect A#
92. #####
93. #####
94. #Second generation p-value limit .005 eta2
95. sgp.005_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
96. .005)
97.
98. #Second generation p-value limit .035 eta2
99. sgp.035_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
100. .035)
101.
102. #####
103. #####
104. #main effect B#
105. #####
106. #####
107. #Second generation p-value limit .005 eta2
108. sgp.005_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
109. .005)

```

```

103.
104.   #Second generation p-value limit .035 eta2
105.   sgp.035_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
106.   .035)
107.
108.   #####
109.   #Interaction effect AB#
110.   #####
111.   #Second generation p-value limit .005 eta2
112.   sgp.005_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
113.   .005)
114.   #Second generation p-value limit .035 eta2
115.   sgp.035_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
116.   .035)
117.   #####
118.   return(list("mean_A_0" = meanA$mean[1],
119.               "mean_A_1" = meanA$mean[2],
120.               "mean_B_0" = meanB$mean[1],
121.               "mean_B_1" = meanB$mean[2],
122.               "mean_A0_B0" = meanAB$mean[1],
123.               "mean_A0_B1" = meanAB$mean[2],
124.               "mean_B1_B0" = meanAB$mean[3],
125.               "mean_A1_B1" = meanAB$mean[4],
126.               "SD_A_0" = meanA$sd[1],
127.               "SD_A_1" = meanA$sd[2],
128.               "SD_B_0" = meanB$sd[1],
129.               "SD_B_1" = meanB$sd[2],
130.               "SD_A0_B0" = meanAB$sd[1],
131.               "SD_A0_B1" = meanAB$sd[2],
132.               "SD_B1_B0" = meanAB$sd[3],
133.               "SD_A1_B1" = meanAB$sd[4],
134.               "eta2_A" = eta_sq$Eta2_partial[1],
135.               "eta2_A.CI_LOW" = eta_sq$CI_low[1],
136.               "eta2_A.CI_HI" = eta_sq$CI_high[1],
137.               "eta2_B" = eta_sq$Eta2_partial[2],
138.               "eta2_B.CI_LOW" = eta_sq$CI_low[2],
139.               "eta2_B.CI_HI" = eta_sq$CI_high[2],
140.               "eta2_AB" = eta_sq$Eta2_partial[3],
141.               "eta2_AB.CI_LOW" = eta_sq$CI_low[3],
142.               "eta2_AB.CI_HI" = eta_sq$CI_high[3],
143.               "pvalue_A" = sum$p.value[1],
144.               "pvalue_B" = sum$p.value[2],
145.               "pvalue_AB" = sum$p.value[3],
146.               "sgpvalue.005_A"=sgp.005_A$p.delta,
147.               "sgpvalue.035_A"=sgp.035_A$p.delta,
148.               "sgpvalue.005_B"=sgp.005_B$p.delta,
149.               "sgpvalue.035_B"=sgp.035_B$p.delta,
150.               "sgpvalue.005_AB"=sgp.005_AB$p.delta,
151.               "sgpvalue.035_AB"=sgp.035_AB$p.delta))
152.
153. }else if(distr=="rsnorm"){
154.
155.   # Generate your data using the regression equation
156.   y11 = mu + alpha + beta + gamma + (rsnorm(n/4, 0, S1,xi = 60))
157.   y12 = mu + alpha + -(beta) + -gamma + (rsnorm(n/4, 0, S1,xi = 60))
158.   y21 = mu + -(alpha) + beta + -gamma + (rsnorm(n/4, 0, S1,xi = 60))

```

```

159.   y22 = mu + -(alpha) + -(beta) +   gamma + (rsnorm(n/4, 0, S1,xi = 60))
160.   #Factor labels
161.   A = c(rep(c(0), n/2), rep(c(1), n/2)) # '0' 20 times, '1' 20 times
162.   B = c(rep(c(0), n/4), rep(c(1), n/4), rep(c(0), n/4), rep(c(1),n/4)) # '0'x10,
    '1'x10, '0'x10, '1'x10
163.
164.   #bind score data
165.   data<-rbind(y11%>%as.tibble,y12%>%as.tibble,y21%>%as.tibble,y22%>%as.tibble)
166.
167.   # Join the variables in a data frame
168.   data <- data.frame(cbind(data,A,B))
169.   data <- data%>%select(`y`=value, A,B)
170.
171.   #Descriptives
172.   meanA<-data%>%group_by(A) %>%
173.     summarise(
174.       count = n(),
175.       mean = mean(y, na.rm = TRUE),
176.       sd = sd(y, na.rm = TRUE)
177.     )
178.
179.
180.   meanB<-data%>%group_by(B) %>%
181.     summarise(
182.       count = n(),
183.       mean = mean(y, na.rm = TRUE),
184.       sd = sd(y, na.rm = TRUE)
185.     )
186.
187.
188.   meanAB<-data%>%group_by(A,B) %>%
189.     summarise(
190.       count = n(),
191.       mean = mean(y, na.rm = TRUE),
192.       sd = sd(y, na.rm = TRUE)
193.     )
194.
195.   #ANOVA model
196.   model = aov(y ~ A*B, data=data)
197.
198.   #Model summary
199.   sum<-tidy(model)
200.
201.   #effect size
202.   eta_sq<-eta_squared(model,partial = T,ci=.95)
203.
204.   #####
205.   #main effect A#
206.   #####
207.   #Second generation p-value limit .005 eta2
208.   sgp.005_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
    .005)
209.
210.   #Second generation p-value limit .035 eta2
211.   sgp.035_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
    .035)
212.
213.   #####
214.   #main effect B#

```

```

215. #####
#####
216. #Second generation p-value limit .005 eta2
217. sgp.005_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
.005)
218.
219. #Second generation p-value limit .035 eta2
220. sgp.035_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
.035)
221.
222. #####
#####
223. #Interaction effect AB#
224. #####
#####
225. #Second generation p-value limit .005 eta2
226. sgp.005_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
.005)
227.
228. #Second generation p-value limit .035 eta2
229. sgp.035_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
.035)
230.
231. #####
#####
232.
233. return(list("mean_A_0" = meanA$mean[1],
234.            "mean_A_1" = meanA$mean[2],
235.            "mean_B_0" = meanB$mean[1],
236.            "mean_B_1" = meanB$mean[2],
237.            "mean_A0_B0" = meanAB$mean[1],
238.            "mean_A0_B1" = meanAB$mean[2],
239.            "mean_B1_B0" = meanAB$mean[3],
240.            "mean_A1_B1" = meanAB$mean[4],
241.            "SD_A_0" = meanA$sd[1],
242.            "SD_A_1" = meanA$sd[2],
243.            "SD_B_0" = meanB$sd[1],
244.            "SD_B_1" = meanB$sd[2],
245.            "SD_A0_B0" = meanAB$sd[1],
246.            "SD_A0_B1" = meanAB$sd[2],
247.            "SD_B1_B0" = meanAB$sd[3],
248.            "SD_A1_B1" = meanAB$sd[4],
249.            "eta2_A" = eta_sq$Eta2_partial[1],
250.            "eta2_A.CI_LOW" = eta_sq$CI_low[1],
251.            "eta2_A.CI_HI" = eta_sq$CI_high[1],
252.            "eta2_B" = eta_sq$Eta2_partial[2],
253.            "eta2_B.CI_LOW" = eta_sq$CI_low[2],
254.            "eta2_B.CI_HI" = eta_sq$CI_high[2],
255.            "eta2_AB" = eta_sq$Eta2_partial[3],
256.            "eta2_AB.CI_LOW" = eta_sq$CI_low[3],
257.            "eta2_AB.CI_HI" = eta_sq$CI_high[3],
258.            "pvalue_A" = sum$p.value[1],
259.            "pvalue_B" = sum$p.value[2],
260.            "pvalue_AB" = sum$p.value[3],
261.            "sgpvalue.005_A"=sgp.005_A$p.delta,
262.            "sgpvalue.035_A"=sgp.035_A$p.delta,
263.            "sgpvalue.005_B"=sgp.005_B$p.delta,
264.            "sgpvalue.035_B"=sgp.035_B$p.delta,
265.            "sgpvalue.005_AB"=sgp.005_AB$p.delta,
266.            "sgpvalue.035_AB"=sgp.035_AB$p.delta))
267.

```

```

268. }else if(distr=="bimodal")
269.
270. # Generate your data using the regression equation
271. y11 = mu + alpha + beta + gamma + log(bimodalDistFunc(n=n/4,cpct =.5,mu1=-
13,mu2=13, sig1=7.5,sig2=7.5))
272. y12 = mu + alpha + -(beta) + -gamma + log(bimodalDistFunc(n=n/4,cpct =.5,mu1=-
13,mu2=13, sig1=7.5,sig2=7.5))
273. y21 = mu + -(alpha) + beta + -gamma + log(bimodalDistFunc(n=n/4,cpct =.5,mu1=-
13,mu2=13, sig1=7.5,sig2=7.5))
274. y22 = mu + -(alpha) + -(beta) + gamma + log(bimodalDistFunc(n=n/4,cpct =.5,mu1=-
13,mu2=13, sig1=7.5,sig2=7.5))
275.
276. #Factor labels
277. A = c(rep(c(0), n/2), rep(c(1), n/2)) # '0' 20 times, '1' 20 times
278. B = c(rep(c(0), n/4), rep(c(1), n/4), rep(c(0), n/4), rep(c(1),n/4)) # '0'x10,
'1'x10, '0'x10, '1'x10
279.
280. #bind score data
281. data<-rbind(y11%>%as.tibble,y12%>%as.tibble,y21%>%as.tibble,y22%>%as.tibble)
282.
283. # Join the variables in a data frame
284. data <- data.frame(cbind(data,A,B))
285. data <- data%>%select(`y`=value, A,B)
286.
287. #Descriptives
288. meanA<-data%>%group_by(A) %>%
289. summarise(
290. count = n(),
291. mean = mean(y, na.rm = TRUE),
292. sd = sd(y, na.rm = TRUE)
293. )
294.
295.
296. meanB<-data%>%group_by(B) %>%
297. summarise(
298. count = n(),
299. mean = mean(y, na.rm = TRUE),
300. sd = sd(y, na.rm = TRUE)
301. )
302.
303.
304. meanAB<-data%>%group_by(A,B) %>%
305. summarise(
306. count = n(),
307. mean = mean(y, na.rm = TRUE),
308. sd = sd(y, na.rm = TRUE)
309. )
310.
311. #ANOVA model
312. model = aov(y ~ A*B, data=data)
313.
314. #Model summary
315. sum<-tidy(model)
316.
317. #effect size
318. eta_sq<-eta_squared(model,partial = T,ci=.95)
319.
320. #####
#####
321. #main effect A#
322. #####
#####
323. #Second generation p-value limit .005 eta2

```

```

324.   sgp.005_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
    .005)
325.
326.   #Second generation p-value limit .035 eta2
327.   sgp.035_A<-sgpvalue(eta_sq$CI_low[1],eta_sq$CI_high[1],null.lo = 0, null.hi =
    .035)
328.
329.   #####
    #####
330.   #main effect B#
331.   #####
    #####
332.   #Second generation p-value limit .005 eta2
333.   sgp.005_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
    .005)
334.
335.   #Second generation p-value limit .035 eta2
336.   sgp.035_B<-sgpvalue(eta_sq$CI_low[2],eta_sq$CI_high[2],null.lo = 0, null.hi =
    .035)
337.
338.   #####
    #####
339.   #Interaction effect AB#
340.   #####
    #####
341.   #Second generation p-value limit .005 eta2
342.   sgp.005_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
    .005)
343.
344.   #Second generation p-value limit .035 eta2
345.   sgp.035_AB<-sgpvalue(eta_sq$CI_low[3],eta_sq$CI_high[3],null.lo = 0, null.hi =
    .035)
346.
347.   #####
    #####
348.
349.   return(list("mean_A_0" = meanA$mean[1],
350.             "mean_A_1" = meanA$mean[2],
351.             "mean_B_0" = meanB$mean[1],
352.             "mean_B_1" = meanB$mean[2],
353.             "mean_A0_B0" = meanAB$mean[1],
354.             "mean_A0_B1" = meanAB$mean[2],
355.             "mean_B1_B0" = meanAB$mean[3],
356.             "mean_A1_B1" = meanAB$mean[4],
357.             "SD_A_0" = meanA$sd[1],
358.             "SD_A_1" = meanA$sd[2],
359.             "SD_B_0" = meanB$sd[1],
360.             "SD_B_1" = meanB$sd[2],
361.             "SD_A0_B0" = meanAB$sd[1],
362.             "SD_A0_B1" = meanAB$sd[2],
363.             "SD_B1_B0" = meanAB$sd[3],
364.             "SD_A1_B1" = meanAB$sd[4],
365.             "eta2_A" = eta_sq$Eta2_partial[1],
366.             "eta2_A.CI_LOW" = eta_sq$CI_low[1],
367.             "eta2_A.CI_HI" = eta_sq$CI_high[1],
368.             "eta2_B" = eta_sq$Eta2_partial[2],
369.             "eta2_B.CI_LOW" = eta_sq$CI_low[2],
370.             "eta2_B.CI_HI" = eta_sq$CI_high[2],
371.             "eta2_AB" = eta_sq$Eta2_partial[3],
372.             "eta2_AB.CI_LOW" = eta_sq$CI_low[3],

```

```

373.         "eta2_AB.CI_HI" = eta_sq$CI_high[3],
374.         "pvalue_A"      = sum$p.value[1],
375.         "pvalue_B"      = sum$p.value[2],
376.         "pvalue_AB"     = sum$p.value[3],
377.         "sgpvalue.005_A"=sgp.005_A$p.delta,
378.         "sgpvalue.035_A"=sgp.035_A$p.delta,
379.         "sgpvalue.005_B"=sgp.005_B$p.delta,
380.         "sgpvalue.035_B"=sgp.035_B$p.delta,
381.         "sgpvalue.005_AB"=sgp.005_AB$p.delta,
382.         "sgpvalue.035_AB"=sgp.035_AB$p.delta))
383.
384.     }
385.

```

APPENDIX B

MONTE CARLO SIMULATION R CODE

```

1. #TWO GROUP MONTE CARLO SIMULATION
2.
3. #load MonteCarlo package
4. library(MonteCarlo)
5.
6. #set parameters
7. #RUN DISTS SEPARATELY BECAUSE SKEW AND BIMODAL DO NOT VARY SD
8. #####
9. #NORMAL DIST          SKEW DIST          BIMODAL DIST
10. #S1_grid<-15          #S1_grid<-15          #S1_grid<-15
11. #S2_grid<-c(15,30,60) #S2_grid<-c(15)          #S2_grid<-c(15)
12. #D_grid<-c(.2,.5,.8)  #D_grid<-c(.2,.5,.8)    #D_grid<-c(.2,.5,.8)
13. #n_grid<-c(20,50,500) #n_grid<-c(20,50,500)    #n_grid<-c(20,50,500)
14. #dist_grid<-"rnorm"    #dist_grid<-"rsnorm"    #dist_grid<-"bimodal"
15. #####
16.
17. S1_grid<-15
18. S2_grid<-c(15,30,60)
19. D_grid<-c(.2,.5,.8)
20. n_grid<-c(20,50,500)
21. dist_grid<-c("rnorm")
22.
23. nrep<-10000 #10,000 rep
24.
25. # collect parameter grids in list:
26. param_list=list("S1"=S1_grid, "S2"=S2_grid, "D"=D_grid, "n" = n_grid,
  "distr"=dist_grid)
27.
28. #set seed
29. set.seed(101)
30.
31. #run MonteCarlo and save as object
32. MC_result_T<-MonteCarlo(func=sim_t3, nrep=nrep, param_list=param_list)
33.
34. #look at parameter summary
35. summary(MC_result_T)
36.
37.
38. #name sim_data_T_rnorm for normal dist
39. #name sim_data_T_rsnorm for skewed dist
40. #name sim_data_T_bimodal for bimodal dist
41.
42. sim_data_T_rnorm<-MakeFrame(MC_result_T)
43.
44. #MonteCarlo function interferes with tidyverse: re-install and load tidyverse
45.
46. install.packages("tidyverse")

```



```

47. library(tidyverse)
48.
49. #place data object processed for results
50.
51. #1    sim_data_T_rnorm
52. #2    sim_data_T_rsnorm
53. #3    sim_data_T_bimodal
54.
55. sim_data<-
56.
57. #####
58. #####
59. #CONSTRUCT RESULTS TABLE OF TYPE I ERROR AND POWER
60. #####
61. #####
62.
63. #pvalue<=.05 COUNT
64. Results_Table<-
  sim_data%>%filter(pvalue<=.05)%>%group_by(distr,D,n,S1,S2)%>%count(name="p<=.05")%>%
65.   mutate(`p<=.05_TI/Power`=`p<=.05`/nrep)%>%left_join(
66.     #pvalue<=.005 COUNT
67.
  sim_data%>%filter(pvalue<=.005)%>%group_by(D,n,S1,S2)%>%count(name="p<=.005")%>%
68.   mutate(`p<=.005_TI/Power`=`p<=.005`/nrep))%>%left_join(
69.
  #SGPV.1==0 COUNT
70.   sim_data%>%filter(sgpvalue.1==
71. 0)%>%group_by(D,n,S1,S2)%>%count(name="SGPV.1=0")%>%
72.   mutate(`SGPV.1=0_TI/Power`=`SGPV.1=0`/nrep))%>%left_join(
73.
  #SGPV.1<=.1 COUNT
74.   sim_data%>%filter(sgpvalue.1<=
75. 0.1)%>%group_by(D,n,S1,S2)%>%count(name="SGPV.1<=.1")%>%
76.   mutate(`SGPV.1<=.1_TI/Power`=`SGPV.1<=.1`/nrep))%>%left_join(
77.
  #SGPV.3<=0 COUNT
78.   sim_data%>%filter(sgpvalue.35==
79. 0)%>%group_by(D,n,S1,S2)%>%count(name="SGPV.35=0")%>%
80.   mutate(`SGPV.35=0_TI/Power`=`SGPV.35=0`/nrep))%>%left_join(
81.
  #SGPV.3<=.1 COUNT
82.   sim_data%>%filter(sgpvalue.35<=
83. 0.1)%>%group_by(D,n,S1,S2)%>%count(name="SGPV.35<=.1")%>%
84.   mutate(`SGPV.35<=.1_TI/Power`=`SGPV.35<=.1`/nrep))%>%left_join(
85.
  #MESP.1 Count
86.   sim_data%>%filter(pvalue<=.05 & (cohens_D>= .1|cohens_D<=-
87. 0.1))%>%group_by(D,n,S1,S2)%>%count(name="MESP.1")%>%
88.   mutate(`MESP.1_TI/Power`=`MESP.1`/nrep))%>%left_join(
89.
  #MESP.35
90.   sim_data%>%filter(pvalue<=.05 & (cohens_D>= .35|cohens_D<=-
91. 0.35))%>%group_by(D,n,S1,S2)%>%count(name="MESP.35")%>%
92.   mutate(`MESP.35_TI/Power`=`MESP.35`/nrep))
93.
94. #view
95. Results_Table%>%view
96.
97.
98. #Rename and set na to 0
99. T_rnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
100.
101. T_rsnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
102.
103. T_bimodal_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)

```

```

104.
105.
106.
107. #DATA
108. #normal dist data
109. saveRDS(sim_data_T_rnorm,"T_sim_rnorm.rds")
110.
111. #normal dist results table
112. saveRDS(T_rnorm_TI.PWR.Results,"T_rnorm_TI.PWR.Results.rds")
113.
114. #normal dist data
115. saveRDS(sim_anova_data,"T_sim_rsnorm.rds")
116.
117. #normal skew dist results table
118. saveRDS(T_rsnorm_TI.PWR.Results,"T_rsnorm_TI.PWR.Results.rds")
119.
120.
121. #bimodal dist data
122. saveRDS(sim_anova_data,"T_sim_bimodal.rds")
123.
124. #bimodal dist results table
125. saveRDS(T_bimodal_TI.PWR.Results,"T_bimodal_TI.PWR.Results.rds")
126.

```

```

1. #3 GROUP MONTE CARLO SIMULATION
2.
3. #load MonteCarlo package
4. library(MonteCarlo)
5.
6. #D = .201(eta = .01),.535825(eta = .06),.857 (eta = .14)
7.
8. #set parameters
9. #RUN DISTS SEPARATELY AS SKEW AND BIMODAL DO NOT VARY SD
10. #####
    #####
11. #NORMAL DIST                SKEW DIST                BIMODAL DIST
12. #S1_grid<-15                #S1_grid<-15                #S1_grid<-15
13. #S2_grid<-15                #S2_grid<-15                #S2_grid<-15
14. #S3_grid<-c(15,30,60)      #S3_grid<-15                #S3_grid<-15
15. #D_grid<-c(0,.201,.535825,.857) #D_grid<-c(0,.201,.535825,.857) #D_grid<-
    c(0,.201,.535825,.857)
16. #n_grid<-c(20,50,500)      #n_grid<-c(20,50,500)      #n_grid<-
    c(20,50,500)
17. #dist_grid<-"rnorm"        #dist_grid<-"rsnorm"        #dist_grid<-
    "bimodal"
18. #####
    #####
19.
20. S1_grid<-c(15)
21. S2_grid<-c(15)
22. S3_grid<-c(15,30,60)
23. D_grid<-c(0,.201,.535825,.857)
24. n_grid<-c(20,50,500)
25. dist_grid<-c("rnorm")
26.
27. nrep<-10000 #10,000 rep
28.
29. # collect parameter grids in list:
30. param_list=list("S1"=S1_grid, "S2"=S2_grid,"S3"=S3_grid, "D"=D_grid, "n" =
    n_grid,"distr"=dist_grid)
31.
32. #set seed
33. set.seed(101)
34.

```

```

35. #run MonteCarlo and save as object
36. MC_result<-MonteCarlo(func=sim_anova, nrep=nrep, param_list=param_list)
37.
38. #look at parameter summary
39. summary(MC_result)
40.
41.
42. #name sim_data_anova_rnorm for normal dist
43. #name sim_data_anova_rsnorm for skewed dist
44. #name sim_data_anova_bimodal for bimodal dist
45.
46. sim_data_anova_rnorm<-MakeFrame(MC_result)
47.
48. #Recode D as POP ETA2
49. sim_data_anova_rnorm<-sim_data_anova_rnorm%>%mutate(Pop_eta2 = case_when(D == 0
~ 0,
50.                                     D == 0.201
~ .01,
51.                                     D ==
0.535825 ~ .06,
52.                                     D == 0.857
~ .14),.after = "Sd3")
53.
54.
55. #MonteCarlo function interferes with tidyverse: re-install and load tidyverse
56. install.packages("tidyverse")
57. library(tidyverse)
58.
59. #place data object processed for results
60.
61. #1    sim_data_anova_rnorm
62. #2    sim_data_anova_rsnorm
63. #3    sim_data_anova_bimodal
64.
65. sim_anova_data<-
66.
67.
68. #####
69. #####
70. #CONSTRUCT RESULTS TABLE OF TYPE I ERROR AND POWER
71. #####
72. #####
73.
74. #pvalue<=.05 COUNT
75. Results_Table<-sim_anova_data%>%filter(pvalue<=.05)%>%group_by(distr,
Pop_eta2,n,S1,S2,S3)%>%count(name="p<=.05")%>%
76.   mutate(`p<=.05_TI/Power` = `p<=.05`/nrep)%>%left_join(
77.
78.   #pvalue<=.005 COUNT
79.
sim_anova_data%>%filter(pvalue<=.005)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(
name="p<=.005")%>%
80.   mutate(`p<=.005_TI/Power` = `p<=.005`/nrep))%>%left_join(
81.
82.
83.   #SGPV.005==0 COUNT
84.   sim_anova_data%>%filter(sgpvalue.005==
0)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.005=0")%>%
85.   mutate(`SGPV.005=0_TI/Power` = `SGPV.005=0`/nrep))%>%left_join(
86.
87.
88.   #SGPV.005<=.1 COUNT
89.   sim_anova_data%>%filter(sgpvalue.005<=
.1)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.005<=.1")%>%
90.   mutate(`SGPV.005<=.1_TI/Power` = `SGPV.005<=.1`/nrep))%>%left_join(

```

```

91.
92.
93.           #SGPV.035<=0 COUNT
94.           sim_anova_data%>%filter(sgpvalue.035==
0)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.035=0")%>%
95.           mutate(`SGPV.035=0_TI/Power`= `SGPV.035=0`/nrep))%>%left_join(
96.
97.
98.           #SGPV.035<=.1 COUNT
99.           sim_anova_data%>%filter(sgpvalue.035<=
.1)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.035<=.1")%>%
100.          mutate(`SGPV.035<=.1_TI/Power`=
`SGPV.035<=.1`/nrep))%>%left_join(
101.
102.
103.          #MESP.005 Count
104.          sim_anova_data%>%filter(pvalue<=.05 & eta2>=
.005)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="MESP.005")%>%
105.          mutate(`MESP.005_TI/Power`=
`MESP.005`/nrep))%>%left_join(
106.
107.
108.          #MESP.035
109.          sim_anova_data%>%filter(pvalue<=.05 & eta2 >=
.035)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="MESP.035")%>%
110.          mutate(`MESP.035_TI/Power`= `MESP.035`/nrep))
111.
112.
113. #view
114. Results_Table%>%view
115.
116. #Rename and set na to 0
117. ANOVA_rnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
118.
119. ANOVA_rsnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
120.
121. ANOVA_bimodal_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
122.
123. #DATA
124. #normal dist data
125. saveRDS(sim_anova_data,"ANOVA_sim_rnorm.rds")
126.
127. #normal dist results table
128. saveRDS(ANOVA_rnorm_TI.PWR.Results,"ANOVA_rnorm_TI.PWR.Results.rds")
129.
130. #normal dist data
131. saveRDS(sim_anova_data,"ANOVA_sim_rsnorm.rds")
132.
133. #normal skew dist results table
134. saveRDS(ANOVA_rsnorm_TI.PWR.Results,"ANOVA_rsnorm_TI.PWR.Results.rds")
135.
136.
137. #bimodal dist data
138. saveRDS(sim_anova_data,"ANOVA_sim_bimodal.rds")
139.
140. #bimodal dist results table
141. saveRDS(ANOVA_bimodal_TI.PWR.Results,"ANOVA_bimodal_TI.PWR.Results.rds")
142.

```

```

1. #K = 3 GROUP MONTE CARLO SIMULATION
2.
3. #load MonteCarlo package
4. library(MonteCarlo)
5.

```

```

6. #D = .201(eta = .01),.535825(eta = .06),.857 (eta = .14)
7.
8. #set parameters
9. #RUN DISTS SEPARATELY AS SKEW AND BIMODAL DO NOT VARY SD
10. #####
#####
11. #NORMAL DIST          SKEW DIST          BIMODAL DIST
12. #S1_grid<-15          #S1_grid<-15          #S1_grid<-15
13. #S2_grid<-15          #S2_grid<-15          #S2_grid<-15
14. #S3_grid<-c(15,30,60) #S3_grid<-15          #S3_grid<-15
15. #D_grid<-c(0,.201,.535825,.857) #D_grid<-c(0,.201,.535825,.857) #D_grid<-
c(0,.201,.535825,.857)
16. #n_grid<-c(20,50,500) #n_grid<-c(20,50,500) #n_grid<-
c(20,50,500)
17. #dist_grid<-"rnorm"    #dist_grid<-"rsnorm"    #dist_grid<-
"bimodal"
18. #####
#####
19.
20. S1_grid<-c(15)
21. S2_grid<-c(15)
22. S3_grid<-c(15,30,60)
23. D_grid<-c(0,.201,.535825,.857)
24. n_grid<-c(20,50,500)
25. dist_grid<-c("rnorm")
26.
27. nrep<-10000 #10,000 rep
28.
29. # collect parameter grids in list:
30. param_list=list("S1"=S1_grid, "S2"=S2_grid,"S3"=S3_grid, "D"=D_grid, "n" =
n_grid,"distr"=dist_grid)
31.
32. #set seed
33. set.seed(101)
34.
35. #run MonteCarlo and save as object
36. MC_result<-MonteCarlo(func=sim_anova, nrep=nrep, param_list=param_list)
37.
38. #look at parameter summary
39. summary(MC_result)
40.
41.
42. #name sim_data_anova_rnorm for normal dist
43. #name sim_data_anova_rsnorm for skewed dist
44. #name sim_data_anova_bimodal for bimodal dist
45.
46. sim_data_anova_rnorm<-MakeFrame(MC_result)
47.
48. #Recode D as POP ETA2
49. sim_data_anova_rnorm<-sim_data_anova_rnorm%>%mutate(Pop_eta2 = case_when(D == 0
~ 0,
50.                                     D == 0.201
~ .01,
51.                                     D ==
0.535825 ~ .06,
52.                                     D == 0.857
~ .14),.after = "Sd3")
53.
54.
55. #MonteCarlo function interferes with tidyverse: re-install and load tidyverse
56. install.packages("tidyverse")
57. library(tidyverse)
58.
59. #place data object processed for results
60.

```

```

61. #1      sim_data_anova_rnorm
62. #2      sim_data_anova_rsnorm
63. #3      sim_data_anova_bimodal
64.
65. sim_anova_data<-
66.
67.
68. #####
69. #####
70. #CONSTRUCT RESULTS TABLE OF TYPE I ERROR AND POWER
71. #####
72. #####
73.
74. #pvalue<=.05 COUNT
75. Results_Table<-sim_anova_data%>%filter(pvalue<=.05)%>%group_by(distr,
  Pop_eta2,n,S1,S2,S3)%>%count(name="p<=.05")%>%
76.   mutate(`p<=.05_TI/Power` = `p<=.05`/nrep)%>%left_join(
77.
78.   #pvalue<=.005 COUNT
79.
  sim_anova_data%>%filter(pvalue<=.005)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(
    name="p<=.005")%>%
80.   mutate(`p<=.005_TI/Power` = `p<=.005`/nrep))%>%left_join(
81.
82.
83.   #SGPV.005==0 COUNT
84.   sim_anova_data%>%filter(sgpvalue.005==
85.   0)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.005=0")%>%
    mutate(`SGPV.005=0_TI/Power` = `SGPV.005=0`/nrep))%>%left_join(
86.
87.
88.   #SGPV.005<=.1 COUNT
89.   sim_anova_data%>%filter(sgpvalue.005<=
90.   .1)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.005<=.1")%>%
    mutate(`SGPV.005<=.1_TI/Power` = `SGPV.005<=.1`/nrep))%>%left_join(
91.
92.
93.   #SGPV.035<=0 COUNT
94.   sim_anova_data%>%filter(sgpvalue.035==
95.   0)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.035=0")%>%
    mutate(`SGPV.035=0_TI/Power` = `SGPV.035=0`/nrep))%>%left_join(
96.
97.
98.   #SGPV.035<=.1 COUNT
99.   sim_anova_data%>%filter(sgpvalue.035<=
100.  .1)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="SGPV.035<=.1")%>%
    mutate(`SGPV.035<=.1_TI/Power` =
101.  `SGPV.035<=.1`/nrep))%>%left_join(
102.
103.
104.   #MESP.005 Count
105.   sim_anova_data%>%filter(pvalue<=.05 & eta2>=
106.   .005)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="MESP.005")%>%
    mutate(`MESP.005_TI/Power` =
107.  `MESP.005`/nrep))%>%left_join(
108.
109.
110.   #MESP.035
111.   sim_anova_data%>%filter(pvalue<=.05 & eta2 >=
112.   .035)%>%group_by(distr,Pop_eta2,n,S1,S2,S3)%>%count(name="MESP.035")%>%
    mutate(`MESP.035_TI/Power` = `MESP.035`/nrep))
113.
114. #view
115. Results_Table%>%view

```

```

115.
116. #Rename and set na to 0
117. ANOVA_rnorm_TI.PWR.Results<-Results_Table<-Results_Table%% replace(is.na(.), 0)
118.
119. ANOVA_rsnorm_TI.PWR.Results<-Results_Table<-Results_Table%% replace(is.na(.), 0)
120.
121. ANOVA_bimodal_TI.PWR.Results<-Results_Table<-Results_Table%% replace(is.na(.), 0)
122.
123. #DATA
124. #normal dist data
125. saveRDS(sim_anova_data,"ANOVA_sim_rnorm.rds")
126.
127. #normal dist results table
128. saveRDS(ANOVA_rnorm_TI.PWR.Results,"ANOVA_rnorm_TI.PWR.Results.rds")
129.
130. #normal dist data
131. saveRDS(sim_anova_data,"ANOVA_sim_rsnorm.rds")
132.
133. #normal skew dist results table
134. saveRDS(ANOVA_rsnorm_TI.PWR.Results,"ANOVA_rsnorm_TI.PWR.Results.rds")
135.
136.
137. #bimodal dist data
138. saveRDS(sim_anova_data,"ANOVA_sim_bimodal.rds")
139.
140. #bimodal dist results table
141. saveRDS(ANOVA_bimodal_TI.PWR.Results,"ANOVA_bimodal_TI.PWR.Results.rds")
142.

```

```

1. #2x2 GROUP MONTECARLO SIMULATION
2.
3. #load MonteCarlo package
4. library(MonteCarlo)
5.
6.
7. #(eta = .01; gamma = 1.511),
8. #(eta = .06; gamma = 3.8),
9. #(eta = .14; gamma = 6.05 )
10. #alpha=100,beta1=0,beta2=0,beta3,n,S1,S2,distr
11.
12. #RUN DISTS SEPARATELY AS SKEW AND BIMODAL DO NOT VARY SD
13. #####
14. #NORMAL DIST                                SKEW DIST                                BIMODAL
15. #S1_grid<-15                                #S1_grid<-na
16. #S2_grid<-c(15,30,60)                        #S2_grid<-na
17. #gamma_grid<-c(0,1.511,3.8,6.05)              #gamma_grid<-c(0,1.511,3.8,6.05)
18. #n_grid<-c(20,50,500)                        #n_grid<-c(20,50,500)
19. #dist_grid<-"rnorm"                          #dist_grid<-"rsnorm"
20. #####
21.
22.
23. #set parameters normal
24. nrm.S1_grid<-c(15)
25. nrm.S2_grid<-c(15,30,60)
26. nrm.gamma_grid<-c(0,1.511,3.8,6.05)
27. nrm.n_grid<-c(20,52,500)

```

```

28. nrm.dist_grid<-c("rnorm")
29.
30. #set parameters Skew
31. skw.S1_grid<-c(15)
32. skw.S2_grid<-c(15)
33. skw.gamma_grid<-c(0,1.511,3.8,6.05)
34. skw.n_grid<-c(20,52,500)
35. skw.dist_grid<-c("rsnorm")
36.
37. #set parameters bimodal
38. bmd.S1_grid<-c(15)
39. bmd.S2_grid<-c(15)
40. bmd.gamma_grid<-c(0,1.511,3.8,6.05)
41. bmd.n_grid<-c(20,52,500)
42. bmd.dist_grid<-c("bimodal")
43.
44. #set number of reps
45. nrep<-10000 #10,000 rep
46.
47.
48. # collect parameter grids in list:
49. param_list.nrm=list("S1"=nrm.S1_grid, "S2"=nrm.S2_grid,"gamma" =nrm.gamma_grid, "n"
= nrm.n_grid, "distr"=nrm.dist_grid)
50.
51. param_list.skw=list("S1"=skw.S1_grid, "S2"=skw.S2_grid,"gamma" =skw.gamma_grid, "n"
= skw.n_grid, "distr"=skw.dist_grid)
52.
53. param_list.bmd=list("S1"=bmd.S1_grid, "S2"=bmd.S2_grid,"gamma" =bmd.gamma_grid, "n"
= bmd.n_grid, "distr"=bmd.dist_grid)
54.
55. #set seed
56. set.seed(101)
57.
58. #run MonteCarlo and save as object
59. MC_result<-MonteCarlo(func=sim_factor2, nrep=nrep, param_list=param_list.nrm)
60.
61. #look at parameter summary
62. summary(MC_result)
63.
64.
65. #name sim_data_factor_rnorm for normal dist
66. sim_data_factor_rnorm<-MakeFrame(MC_result)
67.
68. #name sim_data_factor_rsnorm for skewed dist
69. sim_data_factor_rsnorm<-MakeFrame(MC_result)
70.
71. #name sim_data_factor_bimodal for bimodal dist
72. sim_data_factor_bimodal<-MakeFrame(MC_result)
73.
74. sim_data_factor_bimodal%>%view
75.
76. #Recode D as POP ETA2
77. #rnorm
78. sim_data_factor_rnorm<-sim_data_factor_rnorm%>%mutate(Pop_eta2 = case_when(gamma ==
0 ~ 0,
79.                                     gamma ==
3.8 ~ .06,
80.                                     gamma ==
1.511 ~ .01,
81.                                     gamma ==
6.05 ~ .14),.after = "S2")
82. #rsnorm
83. sim_data_factor_rsnorm<-sim_data_factor_rsnorm%>%mutate(Pop_eta2 = case_when(gamma
== 0 ~ 0,

```



```

84.                                     gamma
== 1.511 ~ .01,
85.                                     gamma
== 3.8 ~ .06,
86.                                     gamma
== 6.05 ~ .14),.after = "S2")
87.
88. #bimodal
89. sim_data_factor_bimodal<-sim_data_factor_bimodal%>%mutate(Pop_eta2 = case_when(gamma
== 0 ~ 0,
90.                                     gamma
== 1.511 ~ .01,
91.                                     gamma
== 3.8 ~ .06,
92.                                     gamma
== 6.05 ~ .14),.after = "S2")
93.
94.
95. #MonteCarlo function interferes with tidyverse: re-install and load tidyverse
96. install.packages("tidyverse")
97. library(tidyverse)
98.
99.
100. #place data object processed for results
101.
102. #1 sim_data_factor_rnorm
103. sim_factor_data<-sim_data_factor_rnorm
104.
105. #2 sim_data_factor_rsnorm
106. sim_factor_data<-sim_data_factor_rsnorm
107.
108. #3 sim_data_factor_bimodal
109. sim_factor_data<-sim_data_factor_bimodal
110.
111.
112. #####
113. #####
114. #CONSTRUCT RESULTS TABLE OF TYPE I ERROR AND POWER
115. #####
116. #####
117.
118. #pvalue_AB<=.05 COUNT
119. Results_Table<-
sim_factor_data%>%filter(pvalue_AB<=.05)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(
name="p<=.05")%>%
120. mutate(`p<=.05_TI/Power` = `p<=.05`/nrep)%>%left_join(
121.
122. #pvalue_AB<=.005 COUNT
123.
sim_factor_data%>%filter(pvalue_AB<=.005)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count
(name="p<=.005")%>%
124. mutate(`p<=.005_TI/Power` = `p<=.005`/nrep))%>%left_join(
125.
126.
127. #SGPV.005==0 COUNT
128. sim_factor_data%>%filter(sgpvalue.005_AB==
0)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="SGPV.005=0")%>%
129. mutate(`SGPV.005=0_TI/Power` = `SGPV.005=0`/nrep))%>%left_join(
130.
131.
132. #SGPV.005<=.1 COUNT
133. sim_factor_data%>%filter(sgpvalue.005_AB<=
.1)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="SGPV.005<=.1")%>%
134. mutate(`SGPV.005<=.1_TI/Power` = `SGPV.005<=.1`/nrep))%>%left_join(
135.

```

```

136.
137.           #SGPV.035<=0 COUNT
138.           sim_factor_data%>%filter(sgpvalue.035_AB==
139.           0)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="SGPV.035=0")%>%
140.           mutate(`SGPV.035=0_TI/Power`= `SGPV.035=0`/nrep))%>%left_join(
141.
142.           #SGPV.035<=.1 COUNT
143.           sim_factor_data%>%filter(sgpvalue.035_AB<=
144.           .1)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="SGPV.035<=.1")%>%
145.           mutate(`SGPV.035<=.1_TI/Power`=
146.           `SGPV.035<=.1`/nrep))%>%left_join(
147.
148.           #MESP.005 Count
149.           sim_factor_data%>%filter(pvalue_AB<=.05 & eta2_AB>=
150.           .005)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="MESP.005")%>%
151.           mutate(`MESP.005_TI/Power`=
152.           `MESP.005`/nrep))%>%left_join(
153.
154.           #MESP.035
155.           sim_factor_data%>%filter(pvalue_AB<=.05 & eta2_AB >=
156.           .035)%>%group_by(distr,Pop_eta2,n,S1,S2)%>%count(name="MESP.035")%>%
157.           mutate(`MESP.035_TI/Power`= `MESP.035`/nrep))
158.
159. #view
160. Results_Table%>%view
161.
162. #Rename and set na to 0
163. Factor_rnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
164.
165. Factor_rsnorm_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.), 0)
166.
167. Factor_bimodal_TI.PWR.Results<-Results_Table<-Results_Table%>% replace(is.na(.),
168. 0)
169.
170. #DATA
171.
172. #normal dist save
173. saveRDS(sim_data_factor_rnorm,"Data/factor_sim_rnorm.rds")
174.
175. #normal dist results table
176. saveRDS(Factor_rnorm_TI.PWR.Results,"Data/Results/Factor_rnorm_TI.PWR.Results.rds"
177. )
178.
179. #skew dist data
180. saveRDS(sim_data_factor_rsnorm,"Data/factor_sim_rsnorm.rds")
181.
182. #skew dist results table
183. saveRDS(Factor_rsnorm_TI.PWR.Results,"Data/Results/Factor_rsnorm_TI.PWR.Results.rds")
184.
185.
186. #bimodal dist data
187. saveRDS(sim_data_factor_bimodal,"Data/factor_sim_bimodal.rds")
188.
189. #bimodal dist results table

```

```
190. saveRDS(Factor_bimodal_TI.PWR.Results,"Data/Results/Factor_bimodal_TI.PWR.Results.  
rds")
```

APPENDIX C

FIGURES R CODE

```
1. #FIGURES
2.
3. #LOAD PACKAGES
4. library(tidyverse)
5. library("ggthemes")
6. library(jtools)
7. library(lemon)
8. library("cowplot")
9.
10. #####
11. #####PREPARE DATA#####
12. #####
13.
14.
15. #####
16. #####Two Group#####
17. #####
18.
19.
20. #SET NA to 0
21. T_rnorm_TI.PWR.Results<-T_rnorm_TI.PWR.Results%% replace(is.na(.), 0)
22.
23. #ratio of TI and Power Method/p<=.05
24. Ratio_Normal_Tlerror<-
25. T_rnorm_TI.PWR.Results%%select(D,n,`p<=.05_TI/Power`, `p<=.005_TI/Power`,
26. `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
27. `MESP.35_TI/Power`)%%
28. mutate(
29. across(c(`p<=.005_TI/Power`,
30. `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
31. `MESP.35_TI/Power`),~(./`p<=.05_TI/Power`)%%round(digits = 2)))
32.
33. #S2 = 15
34. # reorder data
35. S2.15_Tlerror<-Ratio_Normal_Tlerror%%filter(D==0 &
36. S2==15)%%ungroup%%select(!`p<=.05_TI/Power`)%%
37. pivot_longer(!c("distr", "D", "S1", "S2", "D", "n"),names_to = "method", values_to =
38. "ratio")
39.
40. #set factors n and method
41. S2.15_Tlerror<-S2.15_Tlerror%%mutate(n=factor(n,levels = c("20", "50", "500"),
42. labels=c("n = 20", "n = 50", "n = 500")),
43. method=factor(method,levels =
44. c('MESP.1_TI/Power', 'MESP.35_TI/Power', 'SGPV.1=0_TI/Power', 'SGPV.35=0_TI/Power', 'p<=
45. .005_TI/Power'),
46. labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005')))
```

```

44. #S2 = 30
45. # reorder data
46. S2.30_T1error<-Ratio_Normal_T1error%>%filter(D==0 &
S2==30)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
47. pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
"ratio")
48.
49. #set factor or method
50. S2.30_T1error<-S2.30_T1error%>%mutate(n=factor(n,levels = c("20","50","500"),
labels=c("n = 20","n = 50","n = 500")),
51. method=factor(method,levels =
c('MESP.1_TI/Power','MESP.35_TI/Power','SGPV.1=0_TI/Power','SGPV.35=0_TI/Power','p<=
.005_TI/Power'),
52. labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
53.
54.
55. #S2 = 60
56. # reorder data
57. S2.60_T1error<-Ratio_Normal_T1error%>%filter(D==0 &
S2==60)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
58. pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
"ratio")
59.
60. #set factor or method
61. S2.60_T1error<-S2.60_T1error%>%mutate(n=factor(n,levels = c("20","50","500"),
labels=c("n = 20","n = 50","n = 500")),
62. method=factor(method,levels =
c('MESP.1_TI/Power','MESP.35_TI/Power','SGPV.1=0_TI/Power','SGPV.35=0_TI/Power','p<=
.005_TI/Power'),
63. labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
64.
65.
66. #####
67. #####K = 3 Group#####
68. #####
69.
70. #DATA
71.
72. #view names
73. ANOVA_rnorm_TI.PWR.Results%>%names
74.
75. #Set NA to 0
76. ANOVA_rnorm_TI.PWR.Results<-ANOVA_rnorm_TI.PWR.Results%>% replace(is.na(.), 0)
77.
78. #ratio of TI and Power Method/p<=.05
79. K3.Ratio_Normal_T1error<-
ANOVA_rnorm_TI.PWR.Results%>%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power`,
80. `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
`MESP.035_TI/Power`)%>%
81. mutate(across(c(`p<=.005_TI/Power`,
82. `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
83. `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%>%round(digits = 2)))
84.
85.
86. # S3 = 15
87. k3.S3.15_T1error<-K3.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
S3==15)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
88. pivot_longer(!c("distr","Pop_eta2","S1","S2","S3","n"),names_to = "method",
values_to = "ratio")
89.
90. #set factors n and method
91. k3.S3.15_T1error<-k3.S3.15_T1error%>%mutate(n=factor(n,levels = c("20","50","500"),
labels=c("n = 20","n = 50","n = 500")),

```

```

92.                                     method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
    ', 'p<=.005_TI/Power'),
93.                                     labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
94.
95.
96.
97. # S3 = 30
98. k3.S3.30_T1error<-K3.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
    S3==30)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
99.   pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "S3", "n"), names_to = "method",
    values_to = "ratio")
100.
101. #set factor or method
102. k3.S3.30_T1error<-k3.S3.30_T1error%>%mutate(n=factor(n, levels =
    c("20", "50", "500"), labels=c("n = 20", "n = 50", "n = 500"))),
103.                                     method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
    ', 'p<=.005_TI/Power'),
104.                                     labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
105.
106.
107.
108.
109. # S3 = 60
110. k3.S3.60_T1error<-K3.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
    S3==60)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
111.   pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "S3", "n"), names_to = "method",
    values_to = "ratio")
112.
113. #set factor or method
114. k3.S3.60_T1error<-k3.S3.60_T1error%>%mutate(n=factor(n, levels =
    c("20", "50", "500"), labels=c("n = 20", "n = 50", "n = 500"))),
115.                                     method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
    ', 'p<=.005_TI/Power'),
116.                                     labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
117.
118.
119.
120.
121. #####
122. #####2X2 factorial Group#####
123. #####
124.
125. #DATA
126.
127. #view names
128. Factor_rnorm_TI.PWR.Results%>%names
129.
130. #Set NA to 0
131. Factor_rnorm_TI.PWR.Results<-Factor_rnorm_TI.PWR.Results%>% replace(is.na(.), 0)
132.
133. #ratio of TI and Power Method/p<=.05
134. Fct.Ratio_Normal_T1error<-
    Factor_rnorm_TI.PWR.Results%>%select(Pop_eta2, n, `p<=.05_TI/Power`, `p<=.005_TI/Power`
135. ,
    `SGPV.005=0_TI/Power`, `SGPV.035=0_TI/Power`, `MESP.005_TI/Power`,
136. `MESP.035_TI/Power`)%>%
137.   mutate(across(c(`p<=.005_TI/Power`,
138. `SGPV.005=0_TI/Power`, `SGPV.035=0_TI/Power`, `MESP.005_TI/Power`,

```

```

139.           `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%>%round(digits =
140. 2)))
141. # S3 = 15
142. Fct.S2.15_T1error<-Fct.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
143. S2==15)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
144. pivot_longer(!c("distr","Pop_eta2","S1","S2","n"),names_to = "method", values_to
145. = "ratio")
146. #set factors n and method
147. Fct.S2.15_T1error<-Fct.S2.15_T1error%>%mutate(n=factor(n,levels =
148. c("20","52","500"), labels=c("n = 20","n = 50","n = 500")),
149. method=factor(method,levels =
150. c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
151. ','p<=.005_TI/Power'),
152. labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
153. # S3 = 30
154. Fct.S2.30_T1error<-Fct.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
155. S2==30)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
156. pivot_longer(!c("distr","Pop_eta2","S1","S2","n"),names_to = "method", values_to
157. = "ratio")
158. #set factor or method
159. Fct.S2.30_T1error<-Fct.S2.30_T1error%>%mutate(n=factor(n,levels =
160. c("20","52","500"), labels=c("n = 20","n = 50","n = 500")),
161. method=factor(method,levels =
162. c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
163. ','p<=.005_TI/Power'),
164. labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
165. # S3 = 60
166. Fct.S2.60_T1error<-Fct.Ratio_Normal_T1error%>%filter(Pop_eta2==0 &
167. S2==60)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
168. pivot_longer(!c("distr","Pop_eta2","S1","S2","n"),names_to = "method", values_to
169. = "ratio")
170. #set factor or method
171. Fct.S2.60_T1error<-Fct.S2.60_T1error%>%mutate(n=factor(n,levels =
172. c("20","52","500"), labels=c("n = 20","n = 50","n = 500")),
173. method=factor(method,levels =
174. c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
175. ','p<=.005_TI/Power'),
176. labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
177. #####
178. #####
179. #Combine data
180. normal<-bind_rows(S2.15_T1error%>%mutate(Type="Two Group"),
181. k3.S3.15_T1error%>%mutate(Type="K = 3 Group"),
182. Fct.S2.15_T1error%>%mutate(Type="2x2 Group"))%>%
183. mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
184. Htg.30<-bind_rows(S2.30_T1error%>%mutate(Type="Two Group"),

```

```

185.         k3.S3.30_T1error%>%mutate(Type="K = 3 Group"),
186.         Fct.S2.30_T1error%>%mutate(Type="2x2 Group"))%>%
187.     mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
188.
189.
190. Htg.60<-bind_rows(S2.60_T1error%>%mutate(Type="Two Group"),
191.                   k3.S3.60_T1error%>%mutate(Type="K = 3 Group"),
192.                   Fct.S2.60_T1error%>%mutate(Type="2x2 Group"))%>%
193.     mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
194.
195. #####
196. #####
197. #####
198.
199. #GRAPHS
200. normal<-ggplot(normal,aes(fill= `n`, y=method, x=ratio))+
201.   geom_bar(position="dodge", stat="identity")+
202.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "free_x")+
203.   scale_fill_tableau()+
204.   xlab(NULL)+ylab(NULL)+
205.   theme_apo()+
206.   theme(axis.text = element_text(size =12))+
207.   coord_flip()+
208.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
  position_dodge(.9))+
209.   theme(plot.title = element_text(hjust=.5,size=12))
210. normal<-reposition_legend(normalg,'center', panel='panel-2-2')
211.
212. Htg.30g<-ggplot(Htg.30,aes(fill= `n`, y=method, x=ratio))+
213.   geom_bar(position="dodge", stat="identity")+
214.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "free_x")+
215.   scale_fill_tableau()+
216.   xlab(NULL)+ylab(NULL)+
217.   theme_apo()+
218.   theme(axis.text = element_text(size =12))+
219.   coord_flip()+
220.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
  position_dodge(.9))+
221.   theme(plot.title = element_text(hjust=.5,size=12))
222. Htg.30g<-reposition_legend(Htg.30g,'center', panel='panel-2-2')
223.
224.
225.
226.
227. Htg.60g<-ggplot(Htg.60,aes(fill= `n`, y=method, x=ratio))+
228.   geom_bar(position="dodge", stat="identity")+
229.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "free_x")+
230.   scale_fill_tableau()+
231.   xlab(NULL)+ylab(NULL)+
232.   theme_apo()+
233.   theme(axis.text = element_text(size =12))+
234.   coord_flip()+
235.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
  position_dodge(.9))+
236.   theme(plot.title = element_text(hjust=.5,size=12))
237. Htg.60g<-reposition_legend(Htg.60g,'center', panel='panel-2-2')
238.
239.
240.
241.
242.
243. #Save plots
244. ggsave(plot=normalg,width=10,height = 7,"/Users/MyMac/OneDrive - University of
  South Carolina/R code/Dissertation simulation/Plots/Type1_Error/Norm_S15.jpeg")
245.

```



```

246. ggsave(plot=Htg.30g,width=10,height = 7,"/Users/MyMac/OneDrive - University of
    South Carolina/R code/Dissertation simulation/Plots/Type1_Error/Norm_S30.jpeg")
247.
248. ggsave(plot=Htg.60g,width=10,height = 7,"/Users/MyMac/OneDrive - University of
    South Carolina/R code/Dissertation simulation/Plots/Type1_Error/Norm_S60.jpeg")
249.
250.
251.
252.
253. #####
254. ###SKEW and Bi-modal Dist###
255. #####
256.
257. #####
258. #####TWO GROUP#####
259. #####
260.
261. #SKEWED DATA
262.
263. #SET NA to 0
264. T_rsnorm_TI.PWR.Results<-T_rsnorm_TI.PWR.Results%>% replace(is.na(.), 0)
265.
266. #ratio of TI and Power Method/p<=.05
267. T_Ratio_Skew_Tlerror<-
    T_rsnorm_TI.PWR.Results%>%select(D,n,`p<=.05_TI/Power`, `p<=.005_TI/Power`,
268. `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
    `MESP.35_TI/Power`)%>%
269. mutate(across(c(`p<=.005_TI/Power`,
    `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
270. `MESP.35_TI/Power`),~(. / `p<=.05_TI/Power`) %>% round(digits = 2)))
271.
272. #Skew
273. # reorder data
274. T_Ratio_Skew_Tlerror<-
    T_Ratio_Skew_Tlerror%>%filter(D==0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
275. pivot_longer(!c("distr", "D", "S1", "S2", "D", "n"),names_to = "method", values_to =
    "ratio")
276.
277. #set factors n and method
278. T_Ratio_Skew_Tlerror<-T_Ratio_Skew_Tlerror%>%mutate(n=factor(n,levels =
    c("20", "50", "500"), labels=c("n = 20", "n = 50", "n = 500")),
279. method=factor(method,levels =
    c('MESP.1_TI/Power', 'MESP.35_TI/Power', 'SGPV.1=0_TI/Power', 'SGPV.35=0_TI/Power', 'p<=
    .005_TI/Power'),
280. labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
281.
282.
283.
284.
285.
286.
287. #Bimodal Data
288.
289. #SET NA to 0
290. T_bimodal_TI.PWR.Results<-T_bimodal_TI.PWR.Results%>% replace(is.na(.), 0)
291.
292. #ratio of TI and Power Method/p<=.05
293. T_Ratio_Bimodal_Tlerror<-
    T_bimodal_TI.PWR.Results%>%select(D,n,`p<=.05_TI/Power`, `p<=.005_TI/Power`,
294. `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
    `MESP.35_TI/Power`)%>%
295. mutate(across(c(`p<=.005_TI/Power`,
    `SGPV.1=0_TI/Power`, `SGPV.35=0_TI/Power`, `MESP.1_TI/Power`,
296. `MESP.35_TI/Power`),~(. / `p<=.05_TI/Power`) %>% round(digits = 2)))
297.
298.

```

```

299.
300.
301.
302. # reorder data
303. T_Ratio_Bimodal_Tlerror<-
  T_Ratio_Bimodal_Tlerror%>%filter(D==0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
304.   pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
    "ratio")
305.
306. #set factors n and method
307. T_Ratio_Bimodal_Tlerror<-T_Ratio_Bimodal_Tlerror%>%mutate(n=factor(n,levels =
  c("20","50","500"), labels=c("n = 20","n = 50","n = 500")),
308.   method=factor(method,levels =
    c('MESP.1_TI/Power','MESP.35_TI/Power','SGPV.1=0_TI/Power','SGPV.35=0_TI/Power','p<=
    .005_TI/Power'),
309.   labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
310.
311.
312.
313. #####
314. #####K = 3 GROUP#####
315. #####
316.
317.
318. #SKEWED DATA
319.
320. #SET NA to 0
321. ANOVA_rsnorm_TI.PWR.Results<-ANOVA_rsnorm_TI.PWR.Results%>% replace(is.na(.), 0)
322.
323. #ratio of TI and Power Method/p<=.05
324. K3_Ratio_Skew_Tlerror<-
  ANOVA_rsnorm_TI.PWR.Results%>%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power`
325. ,
  `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
326.   `MESP.035_TI/Power`)%>%
327.   mutate(across(c(`p<=.005_TI/Power`,
328.     `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
329.     `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`) %>%round(digits =
    2)))
330.
331.
332. # reorder data
333. K3_Ratio_Skew_Tlerror<-
  K3_Ratio_Skew_Tlerror%>%filter(Pop_eta2==0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
334.   pivot_longer(!c("distr","Pop_eta2","S1","S2","S3","n"),names_to = "method",
    values_to = "ratio")
335.
336. #set factors n and method
337. K3_Ratio_Skew_Tlerror<-K3_Ratio_Skew_Tlerror%>%mutate(n=factor(n,levels =
  c("20","50","500"), labels=c("n = 20","n = 50","n = 500")),
338.   method=factor(method,levels =
    c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
    ','p<=.005_TI/Power'),
339.   labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
340.
341.
342.
343.
344.
345.
346. #Bimodal Data

```

```

347.
348. #SET NA to 0
349. ANOVA_bimodal_TI.PWR.Results<-ANOVA_bimodal_TI.PWR.Results%% replace(is.na(.), 0)
350.
351. #ratio of TI and Power Method/p<=.05
352. K3_Ratio_Bimodal_T1error<-
      ANOVA_bimodal_TI.PWR.Results%%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power`
      ,
353.
354. `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
      `MESP.035_TI/Power`)%>%
355.       mutate(across(c(`p<=.005_TI/Power`,
356.       `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
357.       `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%%round(digits =
      2)))
358.
359.
360.
361. # reorder data
362. K3_Ratio_Bimodal_T1error<-
      K3_Ratio_Bimodal_T1error%%filter(Pop_eta2==0)%>%ungroup%%select(!`p<=.05_TI/Power`
      )%%
363.       pivot_longer(!c("distr","Pop_eta2","S1","S2","S3","Pop_eta2","n"),names_to =
      "method", values_to = "ratio")
364.
365. #set factors n and method
366. K3_Ratio_Bimodal_T1error<-K3_Ratio_Bimodal_T1error%%mutate(n=factor(n,levels =
      c("20","50","500"), labels=c("n = 20","n = 50","n = 500")),
367.       method=factor(method,levels =
      c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power`
      ',`p<=.005_TI/Power`)),
368.       labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))
369.
370.
371.
372. #####
373. #####2x2 GROUP#####
374. #####
375.
376. #SKEWED DATA
377.
378. #SET NA to 0
379. Factor_rsnorm_TI.PWR.Results<-Factor_rsnorm_TI.PWR.Results%% replace(is.na(.), 0)
380.
381. #ratio of TI and Power Method/p<=.05
382. Factor_Ratio_Skew_T1error<-
      Factor_rsnorm_TI.PWR.Results%%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power`
      ,
383.
384. `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
      `MESP.035_TI/Power`)%>%
385.       mutate(across(c(`p<=.005_TI/Power`,
386.       `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,
387.       `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%%round(digits =
      2)))
388.
389.
390. # reorder data
391. Factor_Ratio_Skew_T1error<-
      Factor_Ratio_Skew_T1error%%filter(Pop_eta2==0)%>%ungroup%%select(!`p<=.05_TI/Power`
      )%%
392.       pivot_longer(!c("distr","Pop_eta2","S1","S2","n"),names_to = "method", values_to
      = "ratio")
393.

```

```

394. #set factors n and method
395. Factor_Ratio_Skew_T1error<-Factor_Ratio_Skew_T1error%>%mutate(n=factor(n,levels =
  c("20","52","500"), labels=c("n = 20","n = 50","n = 500")),
396.
  method=factor(method,levels =
    c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
    ','p<=.005_TI/Power'),
397.
  labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
398.
399.
400. #Bimodal Data
401.
402. #SET NA to 0
403. Factor_bimodal_TI.PWR.Results<-Factor_bimodal_TI.PWR.Results%>% replace(is.na(.),
  0)
404.
405. #ratio of TI and Power Method/p<=.05
406. Factor_Ratio_Bimodal_T1error<-
  Factor_bimodal_TI.PWR.Results%>%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power`,
  `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,`MESP.035_TI/Power`)%>%
  mutate(across(c(`p<=.005_TI/Power`,`SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,`MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%>%round(digits =
  2))))
412.
413.
414.
415. # reorder data
416. Factor_Bimodal_T1error<-
  Factor_Ratio_Bimodal_T1error%>%filter(Pop_eta2==0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
  pivot_longer(!c("distr","Pop_eta2","S1","S2","n"),names_to = "method", values_to = "ratio")
418.
419. #set factors n and method
420. Factor_Bimodal_T1error<-Factor_Bimodal_T1error%>%mutate(n=factor(n,levels =
  c("20","52","500"), labels=c("n = 20","n = 50","n = 500")),
  method=factor(method,levels =
  c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
  ','p<=.005_TI/Power'),
422.
  labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))))
423.
424.
425.
426.
427.
428.
429. #Combine data
430. skew<-bind_rows(T_Ratio_Skew_T1error%>%mutate(Type="Two Group"),
  K3_Ratio_Skew_T1error%>%mutate(Type="K = 3 Group"),
  Factor_Ratio_Skew_T1error%>%mutate(Type="2x2 Group"))%>%
433. mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
434.
435. bimod<-bind_rows(T_Ratio_Bimodal_T1error%>%mutate(Type="Two Group"),
  K3_Ratio_Bimodal_T1error%>%mutate(Type="K = 3 Group"),
  Factor_Bimodal_T1error%>%mutate(Type="2x2 Group"))%>%
438. mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
439.
440.
441. #GRAPHS

```

```

442. skewg<-ggplot(skew,aes(fill= `n`, y=method, x=ratio))+
443.   geom_bar(position="dodge", stat="identity")+
444.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "free_x")+
445.   scale_fill_tableau()+
446.   xlab(NULL)+ylab(NULL)+
447.   theme_apo()+
448.   theme(axis.text = element_text(size =12))+
449.   coord_flip()+
450.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
  position_dodge(.9))+
451.   theme(plot.title = element_text(hjust=.5,size=12))
452. skewg<-reposition_legend(skewg,'center', panel='panel-2-2')
453.
454. bimodg<-ggplot(bimod,aes(fill= `n`, y=method, x=ratio))+
455.   geom_bar(position="dodge", stat="identity")+
456.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "free_x")+
457.   scale_fill_tableau()+
458.   xlab(NULL)+ylab(NULL)+
459.   theme_apo()+
460.   theme(axis.text = element_text(size =12))+
461.   coord_flip()+
462.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
  position_dodge(.9))+
463.   theme(plot.title = element_text(hjust=.5,size=12))
464. bimodg<-reposition_legend(bimodg,'center', panel='panel-2-2')
465.
466.
467.
468. #Save plots
469. ggsave(plot=skewg,
470.   width=10,height=7,
471.   "/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
  simulation/Plots/Type1_Error/Skewed_S15.jpeg")
472.
473. ggsave(plot=bimodg,
474.   width=10,height=7,
475.   "/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
  simulation/Plots/Type1_Error/Bimodal_S15.jpeg")
476.
477.
478.
479. #####
480. #####POWER#####
481. #####
482. #SET NA to 0
483. T_rnorm_TI.PWR.Results<-T_rnorm_TI.PWR.Results%>% replace(is.na(.), 0)
484.
485.
486. #S2 = 15
487. # reorder data
488. S2.15_Power<-Ratio_Normal_T1error%>%filter(D!=0 &
  S2==15)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
489.   pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
  "ratio")
490.
491. #set factors n and method
492. S2.15_Power<-S2.15_Power%>%mutate(n=factor(n,levels = c("20","50","500"),
  labels=c("n = 20","n = 50","n = 500")),
  method=factor(method,levels =
  c('MESP.1_TI/Power','MESP.35_TI/Power','SGPV.1=0_TI/Power','SGPV.35=0_TI/Power','p<=
  .005_TI/Power'),
494.   labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005'))
495.   )
496.

```

```

497.
498.
499. K3.S2.15_Power<-K3.Ratio_Normal_T1error%>%filter(Pop_eta2!=0 &
    S3==15)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
500.   pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "S3", "n"), names_to = "method",
    values_to = "ratio")
501.
502. #set factors n and method
503. K3.S2.15_Power<-K3.S2.15_Power%>%mutate(n=factor(n, levels = c("20", "50", "500"),
    labels=c("n = 20", "n = 50", "n = 500")),
504.   method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power',
    'p<=.005_TI/Power'),
505.   labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
506.
507.
508. Fct.Ratio_Normal_T1error
509. Fct.S2.15_Power<-Fct.Ratio_Normal_T1error%>%filter(Pop_eta2!=0 &
    S2==15)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
510.   pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "n"), names_to = "method", values_to
    = "ratio")
511.
512. #set factors n and method
513. Fct.S2.15_Power<-Fct.S2.15_Power%>%mutate(n=factor(n, levels = c("20", "52", "500"),
    labels=c("n = 20", "n = 50", "n = 500")),
514.   method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power',
    'p<=.005_TI/Power'),
515.   labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
516.
517.
518.
519. #Power Normal S2 15 grouped bar plot
520.
521. G2.normal.S15_Power<-ggplot(S2.15_Power, aes(fill= `n`, y=method, x=ratio))+
522.   geom_bar(position="dodge", stat="identity")+
523.   facet_wrap(~ D, nrow=2, ncol=2, scales = "free_x")+
524.   scale_fill_tableau()+
525.   xlab(NULL)+ylab(NULL)+
526.   theme_apo()+
527.   theme(axis.text = element_text(size = 12))+
528.   coord_flip()+
529.   geom_text(aes(label=ratio), vjust=-.2, hjust=.5, size=2.5, color="black", position =
    position_dodge(.9))+
530.   theme(plot.title = element_text(hjust=.5, size=12))
531. G2.normal.S15_Power<-reposition_legend(G2.normal.S15_Power, 'center', panel='panel-
    2-2')
532.
533. library(lemon)
534. ggsave(plot=G2.normal.S15_Power, "/Users/MyMac/OneDrive - University of South
    Carolina/R code/Dissertation simulation/Plots/G2.normal.S15_Power.jpeg")
535.
536. K3.normal.S15_Power<-ggplot(K3.S2.15_Power, aes(fill= `n`, y=method, x=ratio))+
537.   geom_bar(position="dodge", stat="identity")+
538.   facet_wrap(~ Pop_eta2, nrow=2, ncol=2, scales = "free_x")+
539.   scale_fill_tableau()+
540.   xlab(NULL)+ylab(NULL)+
541.   theme_apo()+
542.   theme(axis.text = element_text(size = 12))+
543.   coord_flip()+
544.   geom_text(aes(label=ratio), vjust=-.2, hjust=.5, size=2.5, color="black", position =
    position_dodge(.9))+
545.   theme(plot.title = element_text(hjust=.5, size=12))

```

```

546. K3.normal.S15_Power<-reposition_legend(K3.normal.S15_Power,'center', panel='panel-
2-2')
547.
548.
549. ggsave(plot=K3.normal.S15_Power,"/Users/MyMac/OneDrive - University of South
Carolina/R code/Dissertation simulation/Plots/K3.normal.S15_Power.jpeg")
550.
551. Fct.normal.S15_Power<-ggplot(Fct.S2.15_Power,aes(fill= `n`, y=method, x=ratio))+
552.   geom_bar(position="dodge", stat="identity")+
553.   facet_wrap(~ Pop_eta2,nrow=2,ncol=2,scales = "free_x")+
554.   scale_fill_tableau()+
555.   xlab(NULL)+ylab(NULL)+
556.   theme_apo()+
557.   theme(axis.text = element_text(size =12))+
558.   coord_flip()+
559.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
position_dodge(.9))+
560.   theme(plot.title = element_text(hjust=.5,size=12))
561. Fct.normal.S15_Power<-reposition_legend(Fct.normal.S15_Power,'center',
panel='panel-2-2')
562.
563.
564. ggsave(plot=Fct.normal.S15_Power,width=10,height = 7,
565.   "/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/Fct.normal.S15_Power.jpeg")
566.
567.
568. #DATA
569.
570. #Two Group Data
571. Ratio_Normal_Power<-
Ratio_Normal_TIerror%>%filter(D!=0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
572.   pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
"ratio")%>%
573.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500"))),
574.   method=factor(method,levels =
c('MESP.1_TI/Power','MESP.35_TI/Power','SGPV.1=0_TI/Power','SGPV.35=0_TI/Power','p<=
.005_TI/Power'),
575.     labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005')),
576.   D=factor(D,levels = c("0.2","0.5","0.8"),
577.     labels = c("d = 0.2","d = 0.5","d = 0.8")),
578.   S2 = factor(S2,levels = c("15","30","60"),
579.     labels = c("Var. Ratio 1:1","Var. Ratio 2:1","Var. Ratio
4:1")))
580.
581. #K =3 Data
582. K3.Ratio_Normal_Power<-
K3.Ratio_Normal_TIerror%>%filter(Pop_eta2!=0)%>%ungroup%>%select(-
`p<=.05_TI/Power`)%>%
583.   pivot_longer(!c("distr","Pop_eta2","S1","S2","S3","n"),names_to = "method",
values_to = "ratio")%>%
584.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500"))),
585.   method=factor(method,levels =
c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
','p<=.005_TI/Power'),
586.     labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005')),
587.   Pop_eta2=factor(Pop_eta2,levels = c("0.01","0.06","0.14"),
588.     labels = c("eta = 0.01","eta = 0.06","eta = 0.14")),
589.   S3 = factor(S3,levels = c("15","30","60"),
590.     labels = c("Var. Ratio 1:1","Var. Ratio 2:1","Var. Ratio
4:1")))
591.
592. #2x2 Data

```

```

593. Fct.Ratio_Normal_Power<-
    Fct.Ratio_Normal_T1error%>%filter(Pop_eta2!=0)%>%ungroup%>%select(-
    `p<=.05_TI/Power`)%>%
594.   pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "n"), names_to = "method", values_to
    = "ratio")%>%
595.   mutate(n=factor(n, levels = c("20", "52", "500"), labels=c("n = 20", "n = 50", "n =
    500")),
596.   method=factor(method, levels =
    c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
    ', 'p<=.005_TI/Power')),
597.   labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005')),
598.   Pop_eta2=factor(Pop_eta2, levels = c("0.01", "0.06", "0.14"),
599.   labels = c("eta = 0.01", "eta = 0.06", "eta = 0.14")),
600.   S2 = factor(S2, levels = c("15", "30", "60"),
601.   labels = c("Var. Ratio 1:1", "Var. Ratio 2:1", "Var. Ratio
    4:1"))))
602.
603.
604.
605. #TWO GROUP HETEROGENEOUS PLOT
606. G2.normal.S30.60_Power<-ggplot(Ratio_Normal_Power%>%filter(S2!="Var. Ratio
    1:1"), aes(fill= `n`, y=method, x=ratio))+
607.   geom_bar(position="dodge", stat="identity")+
608.   facet_wrap(~ D+S2, nrow=4, ncol=2, scales = "free_x")+
609.   scale_fill_tableau()+
610.   xlab(NULL)+ylab(NULL)+
611.   theme_apa()+
612.   theme(axis.text = element_text(size =12))+
613.   coord_flip()+
614.   geom_text(aes(label=ratio), vjust=-.2, hjust=.5 ,size=2.5,color="black", position =
    position_dodge(.9))+
615.   theme(plot.title = element_text(hjust=.5, size=12))
616.
617.
618. ggsave(plot=G2.normal.S30.60_Power,
619.   width=10.5,height=9, units="in"
620.   ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
    simulation/Plots/G2.normal.S30.60_Power.jpeg")
621.
622.
623.
624. #K = 3 HETEROGENEOUS PLOT
625.
626. K3.normal.S30.60_Power<-ggplot(K3.Ratio_Normal_Power%>%filter(S3!="Var. Ratio
    1:1"), aes(fill= `n`, y=method, x=ratio))+
627.   geom_bar(position="dodge", stat="identity", show.legend = TRUE)+
628.   facet_wrap(~ Pop_eta2+S3, nrow=4, ncol=2, scales = "free_x")+
629.   scale_fill_tableau()+
630.   xlab(NULL)+ylab(NULL)+
631.   theme_apa()+
632.   theme(axis.text = element_text(size =12))+
633.   coord_flip()+
634.   geom_text(aes(label=ratio), vjust=-.2, hjust=.5 ,size=2.5,color="black", position =
    position_dodge(.9))+
635.   theme(plot.title = element_text(hjust=.5, size=12))
636.
637. ggsave(plot=K3.normal.S30.60_Power,
638.   width=10,height=7, units="in"
639.   ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
    simulation/Plots/Power/K3.normal.S30.60_Power.jpeg")
640.
641.
642. #2x2 HETEROGENEOUS PLOT
643.

```



```

644. Fct.normal.S30.60_Power<-ggplot(Fct.Ratio_Normal_Power%%filter(S2!="Var. Ratio
1:1"),aes(fill= `n`, y=method, x=ratio))+
645.   geom_bar(position="dodge", stat="identity")+
646.   facet_wrap(~ Pop_eta2+S2,nrow=4,ncol=2,scales = "free_x")+
647.   scale_fill_tableau()+
648.   xlab(NULL)+ylab(NULL)+
649.   theme_apa()+
650.   theme(axis.text = element_text(size =12))+
651.   coord_flip()+
652.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
position_dodge(.9))+
653.   theme(plot.title = element_text(hjust=.5,size=12))
654.
655. ggsave(plot=Fct.normal.S30.60_Power,
656.   width=10,height=7, units="in"
657.   ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/Fct.normal.S30.60_Power.jpeg")
658.
659.
660.
661.
662. T_Ratio_Skew.Bimodal_T1error<-bind_rows(T_Ratio_Skew_T1error,
663.   T_Ratio_Bimodal_T1error2)
664. T_Ratio_Skew.Bimodal_T1error%%view
665. Ratio_Skew.Bimodal_Power<-
T_Ratio_Skew.Bimodal_T1error%%filter(D!=0)%%ungroup%%select(!`p<=.05_TI/Power`)%>
%
666.   pivot_longer(!c("distr","D","S1","S2","D","n"),names_to = "method", values_to =
"ratio")%%>%
667.   mutate(D=factor(D,levels = c("0.2","0.5","0.8"),
668.     labels = c("d = 0.2","d = 0.5","d = 0.8")),
669.     distr = factor(distr,levels = c("distr=rsnorm","distr=bimodal"),
670.     labels = c("Skewed","Bimodal")))
671.
672.
673.
674. K3_Ratio_Skew.Bimodal_T1error<-bind_rows(K3_Ratio_Skew_T1error,
675.   K3_Ratio_Bimodal_T1error)
676.
677. K3.Ratio_Skew.Bimodal_Power<-
K3_Ratio_Skew.Bimodal_T1error%%filter(Pop_eta2!=0)%%ungroup%%select(-
`p<=.05_TI/Power`)%>%
678.   pivot_longer(!c("distr","Pop_eta2","S1","S2","S3","n"),names_to = "method",
values_to = "ratio")%%>%
679.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500")),
680.     method=factor(method,levels =
c('MESP.005_TI/Power','MESP.035_TI/Power','SGPV.005=0_TI/Power','SGPV.035=0_TI/Power
','p<=.005_TI/Power'),
681.     labels=c('MESP1','MESP2','SGPV1','SGPV2','p<=.005')),
682.     Pop_eta2=factor(Pop_eta2,levels = c("0.01","0.06","0.14"),
683.     labels = c("eta = 0.01","eta = 0.06","eta = 0.14")),
684.     distr = factor(distr,levels = c("distr=rsnorm","distr=bimodal"),
685.     labels = c("Skewed","Bimodal")))
686.
687. #2x2 Data
688. Factor_Ratio_Skew_T1error2<-
Factor_rsnorm_TI.PWR.Results%%select(Pop_eta2,n,`p<=.05_TI/Power`,`p<=.005_TI/Power
`,`
689. `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,`
690. `MESP.035_TI/Power`)%>%
691.   mutate(across(c(`p<=.005_TI/Power`,`
692.     `SGPV.005=0_TI/Power`,`SGPV.035=0_TI/Power`,`MESP.005_TI/Power`,`

```

```

693.           `MESP.035_TI/Power`),~(. / `p<=.05_TI/Power`)%>%round(digits =
694. 2)))
695.
696. # reorder data
697. Factor_Ratio_Skew_T1error2<-
  Factor_Ratio_Skew_T1error2%>%filter(Pop_eta2!=0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
698. pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "n"), names_to = "method", values_to
  = "ratio")
699.
700. #set factors n and method
701. Factor_Ratio_Skew_T1error2<-Factor_Ratio_Skew_T1error2%>%mutate(n=factor(n, levels
  = c("20", "52", "500"), labels=c("n = 20", "n = 50", "n = 500")),
702.
  method=factor(method, levels =
  c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
  ', 'p<=.005_TI/Power'),
703.
  labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
704.
705.
706. Fct_Bimodal_T1error2<-
  Factor_Ratio_Bimodal_T1error%>%filter(Pop_eta2!=0)%>%ungroup%>%select(!`p<=.05_TI/Power`)%>%
707. pivot_longer(!c("distr", "Pop_eta2", "S1", "S2", "n"), names_to = "method", values_to
  = "ratio")
708.
709. Fct_Ratio_Bimodal_T1error2%>%view
710.
711. #set factors n and method
712. Fct_Ratio_Bimodal_T1error2<-Fct_Bimodal_T1error2%>%mutate(n=factor(n, levels =
  c("20", "52", "500"), labels=c("n = 20", "n = 50", "n = 500")),
713.
  method=factor(method, levels =
  c('MESP.005_TI/Power', 'MESP.035_TI/Power', 'SGPV.005=0_TI/Power', 'SGPV.035=0_TI/Power
  ', 'p<=.005_TI/Power'),
714.
  labels=c('MESP1', 'MESP2', 'SGPV1', 'SGPV2', 'p<=.005'))))
715.
716. Factor_Ratio_Skew_T1error%>%view
717. Fct_Ratio_Skew.Bimodal_T1error<-bind_rows(Factor_Ratio_Skew_T1error2,
718.                                           Fct_Ratio_Bimodal_T1error2)
719.
720.
721. Fct.Ratio_Skew.Bimodal_Power<-
  Fct_Ratio_Skew.Bimodal_T1error%>%filter(Pop_eta2!=0)%>%ungroup%>%
722. mutate(distr = factor(distr, levels =
  c("distr=rsnorm", "distr=bimodal"),
723.
  labels = c("Skewed", "Bimodal")),
724.
  Pop_eta2=factor(Pop_eta2, levels = c("0.01", "0.06", "0.14")),
725.
  labels = c("eta = 0.01", "eta = 0.06", "eta = 0.14"))
726.
727.
728.
729. #TWO GROUP DIST PLOT
730. G2.Skew.Bimodal_Power<-ggplot(Ratio_Skew.Bimodal_Power, aes(fill= `n`, y=method,
  x=ratio))+
731.   geom_bar(position="dodge", stat="identity", show.legend = F)+
732.   facet_wrap(~ D+distr, nrow=4, ncol=2, scales = "free_x")+
733.   scale_fill_tableau()+
734.   xlab(NULL)+ylab(NULL)+
735.   theme_apa()+
736.   theme(axis.text = element_text(size = 12))+
737.   coord_flip()+

```

```

738.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
position_dodge(.9))+
739.   theme(plot.title = element_text(hjust=.5,size=12))+
740.   ggtitle("Two Group")
741.
742.
743.   ggsave(plot=G2.Skew.Bimodal_Power,
744.         width=10.5,height=9, units="in"
745.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/G2.Skew.Bimodal_Power.jpeg")
746.
747.
748.
749.   #K = 3 DIST PLOT
750.
751.   K3.Skew.Bimodal_Power<-ggplot(K3.Ratio_Skew.Bimodal_Power,aes(fill= `n`, y=method,
x=ratio))+
752.   geom_bar(position="dodge", stat="identity",show.legend = F)+
753.   facet_wrap(~ Pop_eta2+distr,nrow=4,ncol=2,scales = "free_x")+
754.   scale_fill_tableau()+
755.   xlab(NULL)+ylab(NULL)+
756.   theme_apo()+
757.   theme(axis.text = element_text(size =12))+
758.   coord_flip()+
759.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
position_dodge(.9))+
760.   theme(plot.title = element_text(hjust=.5,size=12))+
761.   ggtitle("K = 3 Group")
762.
763.   ggsave(plot=K3.Skew.Bimodal_Power,
764.         width=10.5,height=9, units="in"
765.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/K3.Skew.Bimodal_Power.jpeg")
766.
767.   Fct.Ratio_Skew.Bimodal_Power%>%view
768.   #2x2 DIST PLOT Fct.Ratio_Skew.Bimodal_Power
769.
770.   Fct.Skew.Bimodal_Power<-ggplot(Fct.Ratio_Skew.Bimodal_Power,aes(fill= `n`,
y=method, x=ratio))+
771.   geom_bar(position="dodge", stat="identity")+
772.   facet_wrap(~ Pop_eta2+distr,nrow=4,ncol=2,scales = "free_x")+
773.   scale_fill_tableau()+
774.   xlab(NULL)+ylab(NULL)+
775.   theme_apo()+
776.   theme(axis.text = element_text(size =12))+
777.   coord_flip()+
778.   geom_text(aes(label=ratio),vjust=-.2,hjust=.5 ,size=2.5,color="black",position =
position_dodge(.9))+
779.   theme(plot.title = element_text(hjust=.5,size=12))+
780.   ggtitle("2x2 Group")
781.
782.   ggsave(plot=Fct.Skew.Bimodal_Power,
783.         width=10.5,height=9, units="in"
784.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/Fct.Skew.Bimodal_Power.jpeg")
785.

```

```

1. #NORMAL DISTRIBUTION FREQUENCY HISTOGRAM PLOTS
2.
3. #LOAD PACKAGES
4. library(tidyverse)
5. library("ggthemes")
6. library(jtools) #theme_apo
7. library(cowplot)#background_grid

```

```

8.
9.
10. #####
11. #####NORMAL DATA 2 Group#####
12. #####
13.
14. #LOAD DATA FILE IF NOT IN R ENVIRONMENT
15. T_sim_rnorm<-readRDS("~/OneDrive - University of South Carolina/R code/Dissertation
simulation/DATA/T_sim_rnorm.rds")
16.
17. #EFFECT SIZE MEANS and SD for V-line on plot
18. data<-T_sim_rnorm
19. means<-left_join(data%>%filter(pvalue<=.05 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%summarise(p.mean=mean(Est_effect),
20.
p.sd=sd(Est_effect)),
21.
22.
23. data%>%filter(pvalue<=.005&
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%summarise(p005.mean=mean(Est_effect),
24.
p005.sd=sd(Est_effect))%>%left_join(
25.
26.
27. data%>%filter(sgpvalue.1==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%summarise(sgp1.mean=mean(Est_effect),
28.
sgp1.sd=sd(Est_effect))%>%left_join(
29.
30.
31. data%>%filter(sgpvalue.35==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%summarise(sgp35.mean=mean(Est_effect),
32.
sgp35.sd=sd(Est_effect))%>%left_join(
33.
34.
35.
36. data%>%filter(pvalue<=.05 & D!=0 & (cohens_D>= .1|cohens_D<=-
.1))%>%group_by(S2,D,n,Pop_effect)%>%summarise(MESP1.mean=mean(Est_effect),
37.
MESP1.sd=sd(Est_effect))%>%left_join(
38.
39. data%>%filter(pvalue<=.05 & D!=0 & (cohens_D>= .35|cohens_D<=-
.35))%>%group_by(S2,D,n,Pop_effect)%>%summarise(MESP35.mean=mean(Est_effect),
40.
MESP35.sd=sd(Est_effect))%>%view
41. #####
42. #####NORMAL SET Factors#####
43. #####
44.
45. #Data
46. #Set Factors
47. T_sim_rnorm<-T_sim_rnorm%>%
48.   mutate(Pop_effect.ftr=factor(Pop_effect,levels = c("0","3","7.5","12"),
49.     labels=c("Pop Effect = 0","Pop Effect = 3","Pop
Effect = 7.5","Pop Effect = 12")),
50.     n.ftr=factor(n,levels=c("20","50","500"),
51.       labels=c("n = 20","n = 50","n = 500")))
52.
53. means<-means%>%mutate(Pop_effect.ftr=factor(Pop_effect,levels = c("3","7.5","12"),
54.   labels=c("Pop Effect = 3","Pop Effect =
7.5","Pop Effect = 12")),
55.     n.ftr=factor(n,levels=c("20","50","500"),
56.       labels=c("n = 20","n = 50","n = 500")))
57.

```

```

58.
59. #2G Normal P<.005/P<.05
60. ggplot(T_sim_rnorm%%filter(D!=0 & S2==15), aes(x=Est_effect)) +
61.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
62.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
63.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.005&D!=0 & S2==15),fill =
64.     "#1f77b4",bins=50,alpha = 6/10)+
65.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
66.   theme_apa()+background_grid()+
67.   geom_vline(data=means%%filter(S2==15),
68.     aes(xintercept = p.mean),alpha=9/10)+
69.   geom_vline(data=means%%filter(S2==15),
70.     aes(xintercept = p005.mean),color="#B10318")+
71.   ggtitle("P <= 0.05/P <= 0.005")+
72.   theme(plot.title = element_text(hjust = 0.5))
73.
74. #2G Normal SGPV1/P<.05
75. ggplot(T_sim_rnorm%%filter(D!=0 & S2==15), aes(x=Est_effect)) +
76.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
77.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
78.   geom_histogram(data=T_sim_rnorm%%filter(sgpvalue.1==0&D!=0 & S2==15),fill =
79.     "#1f77b4",bins=50,alpha = 6/10)+
80.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
81.   theme_apa()+background_grid()+
82.   geom_vline(data=means%%filter(S2==15),
83.     aes(xintercept = sgp1.mean),color="#B10318")+
84.   geom_vline(data=means%%filter(S2==15),
85.     aes(xintercept = p.mean),alpha=9/10)+
86.   ggtitle("P <= 0.05/SGPV 1")+
87.   theme(plot.title = element_text(hjust = 0.5))
88.
89.
90.
91. #2G Normal SGPV2/P<.05
92. ggplot(T_sim_rnorm%%filter(D!=0 & S2==15), aes(x=Est_effect)) +
93.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
94.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
95.   geom_histogram(data=T_sim_rnorm%%filter(sgpvalue.35==0&D!=0 & S2==15),fill =
96.     "#1f77b4",bins=50,alpha = 6/10)+
97.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
98.   theme_apa()+background_grid()+
99.   geom_vline(data=means%%filter(S2==15),
100.     aes(xintercept = p.mean),alpha=9/10)+
101.   geom_vline(data=means%%filter(S2==15),
102.     aes(xintercept = sgp35.mean),color="#B10318")+
103.   ggtitle("P <= 0.05/SGPV 2")+
104.   theme(plot.title = element_text(hjust = 0.5))
105.
106.
107. #2G Normal MESP2/P<.05
108. #Only for n=500
109. ggplot(T_sim_rnorm%%filter(D!=0 & S2==15&n==500), aes(x=Est_effect)) +
110.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
111.     S2==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
112.   geom_histogram(data=T_sim_rnorm%%filter(pvalue<=0.05 & n==500 & (cohens_D>=
113.     .35|cohens_D<=-.35)&D!=0 & S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+

```

```

112.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
113.   theme_apo()+background_grid()+
114.   geom_vline(data=means%>%filter(S2==15&n==500),
115.             aes(xintercept = p.mean),alpha=9/10)+
116.   geom_vline(data=means%>%filter(S2==15&n==500),
117.             aes(xintercept = MESP35.mean),color="#B10318")+
118.   ggtitle("P <= 0.05/MESP 2")+
119.   theme(plot.title = element_text(hjust = 0.5))
120.
121.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.normal_MESP2.jpeg")
122.
123.
124. #####
125. #####2G 2:1 Var Ratio DATA#####
126. #####
127.
128.
129.
130. #Normal P<.005/P<.05
131. ggplot(T_sim_rnorm%>%filter(D!=0 & S2==30), aes(x=Est_effect)) +
132.   geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+
133.   geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.005&D!=0 & S2==30),fill =
"#1f77b4",bins=50,alpha = 6/10)+
134.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
135.   theme_apo()+background_grid()+
136.   geom_vline(data=means%>%filter(S2==30),
137.             aes(xintercept = p.mean),alpha=9/10)+
138.   geom_vline(data=means%>%filter(S2==30),
139.             aes(xintercept = p005.mean),color="#B10318")+
140.   ggtitle("P <= 0.05/P <= 0.005")+
141.   theme(plot.title = element_text(hjust = 0.5))
142.
143.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S30_p005.jpeg")
144.
145. #Normal SGPV1/P<.05
146. ggplot(T_sim_rnorm%>%filter(D!=0 & S2==30), aes(x=Est_effect)) +
147.   geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+
148.   geom_histogram(data=T_sim_rnorm%>%filter(sgpvalue.1==0&D!=0 & S2==30),fill =
"#1f77b4",bins=50,alpha = 6/10)+
149.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
150.   theme_apo()+background_grid()+
151.   geom_vline(data=means%>%filter(S2==30),
152.             aes(xintercept = sgp1.mean),color="#B10318")+
153.   geom_vline(data=means%>%filter(S2==30),
154.             aes(xintercept = p.mean),alpha=9/10)+
155.   ggtitle("P <= 0.05/SGPV 1")+
156.   theme(plot.title = element_text(hjust = 0.5))
157.
158.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S30_SGPV1.jpeg")
159.
160.
161.
162.
163. #Normal SGPV2/P<.05
164. ggplot(T_sim_rnorm%>%filter(D!=0 & S2==30), aes(x=Est_effect)) +
165.   geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+
166.   geom_histogram(data=T_sim_rnorm%>%filter(sgpvalue.35==0&D!=0 & S2==30),fill =
"#1f77b4",bins=50,alpha = 6/10)+
167.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+

```

```

168.   theme_apa()+background_grid()+
169.   geom_vline(data=means%>%filter(S2==30),
170.             aes(xintercept = p.mean),alpha=9/10)+
171.   geom_vline(data=means%>%filter(S2==30),
172.             aes(xintercept = sgp35.mean),color="#B10318")+
173.   ggtitle("P <= 0.05/SGPV 2")+
174.   theme(plot.title = element_text(hjust = 0.5))
175.
176.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S30_SGPV2.jpeg")
177.
178.
179.
180.   #Normal MESP2/P<.05
181.   #Only for n=500
182.   ggplot(T_sim_rnorm%>%filter(D!=0 & S2==30&n==500), aes(x=Est_effect)) +
183.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==30&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
184.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05 & n==500 & (cohens_D>=
.35|cohens_D<=-.35)&D!=0 & S2==30),fill = "#1f77b4",bins=50,alpha = 6/10)+
185.     facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
186.     theme_apa()+background_grid()+
187.     geom_vline(data=means%>%filter(S2==30&n==500),
188.               aes(xintercept = p.mean),alpha=9/10)+
189.     geom_vline(data=means%>%filter(S2==30&n==500),
190.               aes(xintercept = MESP35.mean),color="#B10318")
191.   ggtitle("P <= 0.05/MESP 2")+
192.   theme(plot.title = element_text(hjust = 0.5))
193.
194.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S30_MESP2.jpeg")
195.
196.
197.   #####
198.   #####2G 4:1 Var Ratio DATA#####
199.   #####
200.
201.
202.
203.   #Normal P<.005/P<.05
204.   ggplot(T_sim_rnorm%>%filter(D!=0 & S2==60), aes(x=Est_effect)) +
205.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==60),bins=50,fill="#2CA02C",alpha = 4/10)+
206.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.005&D!=0 & S2==60),fill =
"#1f77b4",bins=50,alpha = 6/10)+
207.     facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
208.     theme_apa()+background_grid()+
209.     geom_vline(data=means%>%filter(S2==60),
210.               aes(xintercept = p.mean),alpha=9/10)+
211.     geom_vline(data=means%>%filter(S2==60),
212.               aes(xintercept = p005.mean),color="#B10318")+
213.     ggtitle("P <= 0.05/P <= 0.005")+
214.     theme(plot.title = element_text(hjust = 0.5))
215.
216.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S60_p005.jpeg")
217.
218.
219.   #Normal SGPV1/P<.05
220.   ggplot(T_sim_rnorm%>%filter(D!=0 & S2==60), aes(x=Est_effect)) +
221.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==60),bins=50,fill="#2CA02C",alpha = 4/10)+
222.     geom_histogram(data=T_sim_rnorm%>%filter(sgpvalue.1==0&D!=0 & S2==60),fill =
"#1f77b4",bins=50,alpha = 6/10)+
223.     facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+

```



```

224.   theme_apo()+background_grid()+
225.   geom_vline(data=means%>%filter(S2==60),
226.             aes(xintercept = sgp1.mean),color="#B10318")+
227.   geom_vline(data=means%>%filter(S2==60),
228.             aes(xintercept = p.mean),alpha=9/10)+
229.   ggtitle("P <= 0.05/SGPV 1")+
230.   theme(plot.title = element_text(hjust = 0.5))
231.
232.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S60_SGPV1.jpeg")
233.
234.
235.
236.   #Normal SGPV2/P<.05
237.   ggplot(T_sim_rnorm%>%filter(D!=0 & S2==60), aes(x=Est_effect)) +
238.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==60),bins=50,fill="#2CA02C",alpha = 4/10)+
239.     geom_histogram(data=T_sim_rnorm%>%filter(sgpvalue.35==0&D!=0 & S2==60),fill =
"#1f77b4",bins=50,alpha = 6/10)+
240.     facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
241.     theme_apo()+background_grid()+
242.     geom_vline(data=means%>%filter(S2==60),
243.               aes(xintercept = p.mean),alpha=9/10)+
244.     geom_vline(data=means%>%filter(S2==60),
245.               aes(xintercept = sgp35.mean),color="#B10318")+
246.     ggtitle("P <= 0.05/SGPV 2")+
247.     theme(plot.title = element_text(hjust = 0.5))
248.
249.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S60_SGPV2.jpeg")
250.
251.
252.   #Normal MESP2/P<.05
253.   #Only for n=500
254.   ggplot(T_sim_rnorm%>%filter(D!=0 & S2==60&n==500), aes(x=Est_effect)) +
255.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S2==60&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
256.     geom_histogram(data=T_sim_rnorm%>%filter(pvalue<=0.05 & n==500 & (cohens_D>=
.35|cohens_D<=-.35)&D!=0 & S2==60),fill = "#1f77b4",bins=50,alpha = 6/10)+
257.     facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
258.     theme_apo()+background_grid()+
259.     geom_vline(data=means%>%filter(S2==60&n==500),
260.               aes(xintercept = p.mean),alpha=9/10)+
261.     geom_vline(data=means%>%filter(S2==60&n==500),
262.               aes(xintercept = MESP35.mean),color="#B10318")+
263.     ggtitle("P <= 0.05/MESP 2")+
264.     theme(plot.title = element_text(hjust = 0.5))
265.
266.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/G2.S60_MESP2.jpeg")
267.
268.
269.   #####
270.   #####K=3 GROUP NORMAL DATA#####
271.   #####
272.
273.   #EFFECT SIZE MEANS and SD
274.
275.   k3.means<-left_join(ANOVA_sim_rnorm%>%filter(pvalue<=.05 &
D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(p.mean=mean(eta2),
276.   p.sd=sd(eta2)),
277.
278.

```



```

279. ANOVA_sim_rnorm%>%filter(pvalue<=.005&
  D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(p005.mean=mean(eta2),
280.   p005.sd=sd(eta2))%>%left_join(
281.
282.
283. ANOVA_sim_rnorm%>%filter(sgpvalue.005==0 &
  D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp1.mean=mean(eta2),
284.   sgp005.sd=sd(eta2))%>%left_join(
285.
286.
287. ANOVA_sim_rnorm%>%filter(sgpvalue.035==0 &
  D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp35.mean=mean(eta2),
288.   sgp035.sd=sd(eta2))%>%left_join(
289.
290.
291.
292. ANOVA_sim_rnorm%>%filter(pvalue<=.05 & D!=0 & (eta2>= .005|eta2<=-
  .005))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP1.mean=mean(eta2),
293.   MESP1.sd=sd(eta2))%>%left_join(
294.
295. ANOVA_sim_rnorm%>%filter(pvalue<=.05 & D!=0 & (eta2>= .035|eta2<=-
  .035))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP35.mean=mean(eta2),
296.   MESP35.sd=sd(eta2))%>%view
297.
298. #####
299. #####K=3 GROUP NORMAL DATA SET Factors#####
300. #####
301. ANOVA_sim_rnorm%>%head%>%view
302. ANOVA_sim_rnorm%>%count(Pop_eta2)
303. #Data
304. #Set Factors
305. ANOVA_sim_rnorm<-ANOVA_sim_rnorm%>%
306.   mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
307.     labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
  eta^2 = 0.06","Pop eta^2 = 0.14")),
308.   n.ftr=factor(n,levels=c("20","50","500"),
309.     labels=c("n = 20","n = 50","n = 500")))
310.
311. k3.means<-k3.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2,levels =
  c("0.01","0.06","0.14"),
312.   labels=c("Pop eta^2 = 0.01","Pop eta^2
  = 0.06","Pop eta^2 = 0.14")),
313.   n.ftr=factor(n,levels=c("20","50","500"),
314.     labels=c("n = 20","n = 50","n = 500")))
315.
316. ?scales
317. #Normal P<.005/P<.05
318. ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
319.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
  S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
320.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.005&D!=0 & S3==15),fill =
  "#1f77b4",bins=50,alpha = 6/10)+
321.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales="free_x")+
322.   theme_apo()+background_grid()+
323.   geom_vline(data=k3.means%>%filter(S3==15),
324.     aes(xintercept = p.mean),alpha=9/10)+
325.   geom_vline(data=k3.means%>%filter(S3==15),
326.     aes(xintercept = p005.mean),color="#B10318")+
327.   ggtitle("P <= 0.05/P <= 0.005")+
328.   theme(plot.title = element_text(hjust = 0.5))
329.
330. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
  simulation/Plots/Histograms/NORMAL/K3.normal_p005.jpeg")
331.

```

```

332. #Normal SGPV005/P<.05
333. ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
334.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
   S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
335.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(sgpvalue.005==0&D!=0 & S3==15),fill
   = "#1f77b4",bins=50,alpha = 6/10)+
336.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
337.   theme_apo()+background_grid()+
338.   geom_vline(data=k3.means%>%filter(S3==15),
339.     aes(xintercept = sgp1.mean),color="#B10318")+
340.   geom_vline(data=k3.means%>%filter(S3==15),
341.     aes(xintercept = p.mean),alpha=9/10)+
342.   ggtitle("P <= 0.05/SGPV 1")+
343.   theme(plot.title = element_text(hjust = 0.5))
344.
345. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.normal_SGPV1.jpeg")
346.
347.
348.
349. #Normal SGPV2/P<.05
350. ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
351.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
   S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
352.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(sgpvalue.035==0&D!=0 & S3==15),fill
   = "#1f77b4",bins=50,alpha = 6/10)+
353.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
354.   theme_apo()+background_grid()+
355.   geom_vline(data=k3.means%>%filter(S3==15),
356.     aes(xintercept = p.mean),alpha=9/10)+
357.   geom_vline(data=k3.means%>%filter(S3==15),
358.     aes(xintercept = sgp35.mean),color="#B10318")+
359.   ggtitle("P <= 0.05/SGPV 2")+
360.   theme(plot.title = element_text(hjust = 0.5))
361.
362. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.normal_SGPV2.jpeg")
363.
364.
365. #Normal MESP2/P<.05
366. #Only for n=500
367. ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==15&n==500), aes(x=eta2)) +
368.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
   S3==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
369.   geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05 & n==500 & (eta2>=
   .035|eta2<=-.035)&D!=0 & S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
370.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
371.   theme_apo()+background_grid()+
372.   geom_vline(data=k3.means%>%filter(S3==15&n==500),
373.     aes(xintercept = p.mean),alpha=9/10)+
374.   geom_vline(data=k3.means%>%filter(S3==15&n==500),
375.     aes(xintercept = MESP35.mean),color="#B10318")+
376.   ggtitle("P <= 0.05/MESP 2")+
377.   theme(plot.title = element_text(hjust = 0.5))
378.
379. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.normal_MESP2.jpeg")
380.
381.
382.
383. #####
384. #####K=3 GROUP 2:1 Var Ratio DATA#####
385. #####
386.
387.

```

```

388.
389. #Normal P<.005/P<.05
390. ggplot(ANOVA_sim_rnorm%%filter(D!=0 & S3==30), aes(x=eta2)) +
391.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
   S3==30),bins=50,fill="#2CA02C",alpha = 4/10)+
392.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.005&D!=0 & S3==30),fill =
   "#1f77b4",bins=50,alpha = 6/10)+
393.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
394.   theme_apa()+background_grid()+
395.   geom_vline(data=k3.means%%filter(S3==30),
396.             aes(xintercept = p.mean),alpha=9/10)+
397.   geom_vline(data=k3.means%%filter(S3==30),
398.             aes(xintercept = p005.mean),color="#B10318")+
399.   ggtitle("P <= 0.05/P <= 0.005")+
400.   theme(plot.title = element_text(hjust = 0.5))
401.
402. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S30_p005.jpeg")
403.
404. #Normal SGPV1/P<.05
405. ggplot(ANOVA_sim_rnorm%%filter(D!=0 & S3==30), aes(x=eta2)) +
406.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
   S3==30),bins=50,fill="#2CA02C",alpha = 4/10)+
407.   geom_histogram(data=ANOVA_sim_rnorm%%filter(sgpvalue.005==0&D!=0 & S3==30),fill
   = "#1f77b4",bins=50,alpha = 6/10)+
408.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
409.   theme_apa()+background_grid()+
410.   geom_vline(data=k3.means%%filter(S3==30),
411.             aes(xintercept = sgp1.mean),color="#B10318")+
412.   geom_vline(data=k3.means%%filter(S3==30),
413.             aes(xintercept = p.mean),alpha=9/10)+
414.   ggtitle("P <= 0.05/SGPV 1")+
415.   theme(plot.title = element_text(hjust = 0.5))
416.
417. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S30_SGPV1.jpeg")
418.
419.
420.
421. #Normal SGPV2/P<.05
422. ggplot(ANOVA_sim_rnorm%%filter(D!=0 & S3==30), aes(x=eta2)) +
423.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
   S3==30),bins=50,fill="#2CA02C",alpha = 4/10)+
424.   geom_histogram(data=ANOVA_sim_rnorm%%filter(sgpvalue.035==0&D!=0 & S3==30),fill
   = "#1f77b4",bins=50,alpha = 6/10)+
425.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
426.   theme_apa()+background_grid()+
427.   geom_vline(data=k3.means%%filter(S3==30),
428.             aes(xintercept = p.mean),alpha=9/10)+
429.   geom_vline(data=k3.means%%filter(S3==30),
430.             aes(xintercept = sgp35.mean),color="#B10318")+
431.   ggtitle("P <= 0.05/SGPV 2")+
432.   theme(plot.title = element_text(hjust = 0.5))
433.
434. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S30_SGPV2.jpeg")
435.
436.
437. #Normal MESP2/P<.05
438. #Only for n=500
439. ggplot(ANOVA_sim_rnorm%%filter(D!=0 & S3==30&n==500), aes(x=eta2)) +
440.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.05&D!=0 &
   S3==30&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
441.   geom_histogram(data=ANOVA_sim_rnorm%%filter(pvalue<=0.05 & n==500 & (eta2>=
   .035|eta2<=-.035)&D!=0 & S3==30),fill = "#1f77b4",bins=50,alpha = 6/10)+

```

```

442.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
443.   theme_apo()+background_grid()+
444.   geom_vline(data=k3.means%>%filter(S3==30&n==500),
445.             aes(xintercept = p.mean),alpha=9/10)+
446.   geom_vline(data=k3.means%>%filter(S3==30&n==500),
447.             aes(xintercept = MESP35.mean),color="#B10318")+
448.   ggtitle("P <= 0.05/MESP 2")+
449.   theme(plot.title = element_text(hjust = 0.5))
450.
451.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S30_MESP2.jpeg")
452.
453.   #####
454.   #####K=3 GROUP 4:1 Var Ratio DATA#####
455.   #####
456.
457.
458.
459.   #Normal P<.005/P<.05
460.   ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==60), aes(x=eta2)) +
461.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S3==60),bins=50,fill="#2CA02C",alpha = 4/10)+
462.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.005&D!=0 & S3==60),fill =
"#1f77b4",bins=50,alpha = 6/10)+
463.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
464.     theme_apo()+background_grid()+
465.     geom_vline(data=k3.means%>%filter(S3==60),
466.               aes(xintercept = p.mean),alpha=9/10)+
467.     geom_vline(data=k3.means%>%filter(S3==60),
468.               aes(xintercept = p005.mean),color="#B10318")+
469.     ggtitle("P <= 0.05/P <= 0.005")+
470.     theme(plot.title = element_text(hjust = 0.5))
471.
472.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S60_p005.jpeg")
473.
474.   #Normal SGPV1/P<.05
475.   ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==60), aes(x=eta2)) +
476.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S3==60),bins=50,fill="#2CA02C",alpha = 4/10)+
477.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(sgpvalue.005==0&D!=0 & S3==60),fill
= "#1f77b4",bins=50,alpha = 6/10)+
478.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
479.     theme_apo()+background_grid()+
480.     geom_vline(data=k3.means%>%filter(S3==60),
481.               aes(xintercept = sgp1.mean),color="#B10318")+
482.     geom_vline(data=k3.means%>%filter(S3==60),
483.               aes(xintercept = p.mean),alpha=9/10)+
484.     ggtitle("P <= 0.05/SGPV 1")+
485.     theme(plot.title = element_text(hjust = 0.5))
486.
487.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S60_SGPV1.jpeg")
488.
489.
490.
491.   #Normal SGPV2/P<.05
492.   ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==60), aes(x=eta2)) +
493.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S3==60),bins=50,fill="#2CA02C",alpha = 4/10)+
494.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(sgpvalue.035==0&D!=0 & S3==60),fill
= "#1f77b4",bins=50,alpha = 6/10)+
495.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
496.     theme_apo()+background_grid()+
497.     geom_vline(data=k3.means%>%filter(S3==60),

```

```

498.         aes(xintercept = p.mean), alpha=9/10)+
499.     geom_vline(data=k3.means%>%filter(S3==60),
500.         aes(xintercept = sgp35.mean), color="#B10318")+
501.     ggtitle("P <= 0.05/SGPV 2")+
502.     theme(plot.title = element_text(hjust = 0.5))
503.
504.     ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S60_SGPV2.jpeg")
505.
506.
507.     #Normal MESP2/P<.05
508.     #Only for n=500
509.     ggplot(ANOVA_sim_rnorm%>%filter(D!=0 & S3==60&n==500), aes(x=eta2)) +
510.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05&D!=0 &
S3==60&n==500), bins=50, fill="#2CA02C", alpha = 4/10)+
511.     geom_histogram(data=ANOVA_sim_rnorm%>%filter(pvalue<=0.05 & n==500 & (eta2>=
.035|eta2<=-.035)&D!=0 & S3==60), fill = "#1f77b4", bins=50, alpha = 6/10)+
512.     facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=1, ncol=3, scales = "free_x")+
513.     theme_apl()+background_grid()+
514.     geom_vline(data=k3.means%>%filter(S3==60&n==500),
515.         aes(xintercept = p.mean), alpha=9/10)+
516.     geom_vline(data=k3.means%>%filter(S3==60&n==500),
517.         aes(xintercept = MESP35.mean), color="#B10318")+
518.     ggtitle("P <= 0.05/MESP 2")+
519.     theme(plot.title = element_text(hjust = 0.5))
520.
521.     ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/K3.S60_MESP2.jpeg")
522.
523.
524.
525.
526.
527.
528.
529.
530.
531. #####
532. #####2x2 GROUP NORMAL DATA#####
533. #####
534.
535. #EFFECT SIZE MEANS and SD
536.
537.
538. fct.means<-left_join(factor_sim_rnorm%>%filter(pvalue_AB<=.05 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
539.     summarise(p.mean=mean(eta2_AB),p.sd=sd(eta2_AB)),
540.
541.     factor_sim_rnorm%>%filter(pvalue_AB<=.005&
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
542.     summarise(p005.mean=mean(eta2_AB),p005.sd=sd(eta2_AB)))%>%left_join(
543.
544.     factor_sim_rnorm%>%filter(sgpvalue.005_AB==0 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
545.     summarise(sgp1.mean=mean(eta2_AB),sgp005.sd=sd(eta2_AB)))%>%left_join(
546.
547.     factor_sim_rnorm%>%filter(sgpvalue.035_AB==0 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
548.     summarise(sgp35.mean=mean(eta2_AB),sgp035.sd=sd(eta2_AB)))%>%left_join(
549.
550.     factor_sim_rnorm%>%filter(pvalue_AB<=.05 & Pop_eta2!=0 & (eta2_AB>=
.005|eta2_AB<=-.005))%>%group_by(S2,n,Pop_eta2)%>%
551.     summarise(MESP1.mean=mean(eta2_AB),MESP1.sd=sd(eta2_AB)))%>%left_join(

```

```

552. #####
553. factor_sim_rnorm%>%filter(pvalue_AB<=.05 & Pop_eta2!=0 & (eta2_AB>=
.035|eta2_AB<=-.035))%>%group_by(S2,n,Pop_eta2)%>%
554. summarise(MESP35.mean=mean(eta2_AB),MESP35.sd=sd(eta2_AB))%>%view
555.
556. #####
557. #####
558. #####
559.
560. #####
561. #####2x2 GROUP NORMAL DATA SET Factors#####
562. #####
563. #####
564. factor_sim_rnorm%>%head%>%view
565.
566. #Data
567. #Set Factors
568. factor_sim_rnorm<-factor_sim_rnorm%>%
569. mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
570. labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
eta^2 = 0.06","Pop eta^2 = 0.14")),
571. n.ftr=factor(n,levels=c("20","50","500"),
572. labels=c("n = 20","n = 50","n = 500")))
573.
574. fct.means<-fct.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2,levels =
c("0.01","0.06","0.14"),
575. labels=c("Pop eta^2 = 0.01","Pop
eta^2 = 0.06","Pop eta^2 = 0.14")),
576. n.ftr=factor(n,levels=c("20","50","500"),
577. labels=c("n = 20","n = 50","n = 500")))
578.
579.
580. #Normal P<.005/P<.05
581. ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
582. geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=.05&Pop_eta2!=0 &
S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
583. geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=.005&Pop_eta2!=0 &
S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
584. facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales="free_x")+
585. theme_apa()+background_grid()+
586. geom_vline(data=fct.means%>%filter(S2==15),
587. aes(xintercept = p.mean),alpha=9/10)+
588. geom_vline(data=fct.means%>%filter(S2==15),
589. aes(xintercept = p005.mean),color="#B10318")+
590. ggtitle("P <= 0.05/P <= 0.005")+
591. theme(plot.title = element_text(hjust = 0.5))
592.
593. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.normal_p005.jpeg")
594.
595. #Normal SGPV005/P<.05
596. ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
597. geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=.05&Pop_eta2!=0 &
S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
598. geom_histogram(data=factor_sim_rnorm%>%filter(sgpvalue.005_AB==0&Pop_eta2!=0 &
S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
599. facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
600. theme_apa()+background_grid()+
601. geom_vline(data=fct.means%>%filter(S2==15),

```



```

602.         aes(xintercept = sgp1.mean),color="#B10318")+
603.   geom_vline(data=fct.means%>%filter(S2==15),
604.             aes(xintercept = p.mean),alpha=9/10)+
605.   ggtitle("P <= 0.05/SGPV 1")+
606.   theme(plot.title = element_text(hjust = 0.5))
607.
608.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.normal_SGPV1.jpeg")
609.
610.
611.
612.
613.   #Normal SGPV2/P<.05
614.   ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
615.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
616.     geom_histogram(data=factor_sim_rnorm%>%filter(sgpvalue.035_AB==0&Pop_eta2!=0 &
S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
617.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
618.     theme_apo()+background_grid()+
619.     geom_vline(data=fct.means%>%filter(S2==15),
620.               aes(xintercept = p.mean),alpha=9/10)+
621.     geom_vline(data=fct.means%>%filter(S2==15),
622.               aes(xintercept = sgp35.mean),color="#B10318")+
623.     ggtitle("P <= 0.05/SGPV 2")+
624.     theme(plot.title = element_text(hjust = 0.5))
625.
626.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.normal_SGPV2.jpeg")
627.
628.
629.
630.   #Normal MESP2/P<.05
631.   #Only for n=500
632.   ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==15&n==500), aes(x=eta2_AB)) +
633.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
634.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05 & n==500 &
(eta2_AB>= .035|eta2_AB<=-.035)&Pop_eta2!=0 & S2==15),fill = "#1f77b4",bins=50,alpha
= 6/10)+
635.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
636.     theme_apo()+background_grid()+
637.     geom_vline(data=fct.means%>%filter(S2==15&n==500),
638.               aes(xintercept = p.mean),alpha=9/10)+
639.     geom_vline(data=fct.means%>%filter(S2==15&n==500),
640.               aes(xintercept = MESP35.mean),color="#B10318")+
641.     ggtitle("P <= 0.05/MESP 2")+
642.     theme(plot.title = element_text(hjust = 0.5))
643.
644.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.normal_MESP2.jpeg")
645.
646.
647.
648.
649.   #####
650.   #####2x2 GROUP 2:1 Var Ratio DATA#####
651.   #####
652.
653.
654.
655.   #Normal P<.005/P<.05
656.   ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==30), aes(x=eta2_AB)) +
657.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+

```

```

658.   geom_histogram(data=factor_sim_rnorm%%filter(pvalue_AB<=0.005&Pop_eta2!=0 &
S2==30),fill = "#1f77b4",bins=50,alpha = 6/10)+
659.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
660.   theme_apache()+background_grid()+
661.   geom_vline(data=fct.means%%filter(S2==30),
662.             aes(xintercept = p.mean),alpha=9/10)+
663.   geom_vline(data=fct.means%%filter(S2==30),
664.             aes(xintercept = p005.mean),color="#B10318")+
665.   ggtitle("P <= 0.05/P <= 0.005")+
666.   theme(plot.title = element_text(hjust = 0.5))
667.
668.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S30_p005.jpeg")
669.
670.
671.   #Normal SGPV1/P<.05
672.   ggplot(factor_sim_rnorm%%filter(Pop_eta2!=0 & S2==30), aes(x=eta2_AB)) +
673.     geom_histogram(data=factor_sim_rnorm%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+
674.     geom_histogram(data=factor_sim_rnorm%%filter(sgpvalue.005_AB==0&Pop_eta2!=0 &
S2==30),fill = "#1f77b4",bins=50,alpha = 6/10)+
675.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
676.     theme_apache()+background_grid()+
677.     geom_vline(data=fct.means%%filter(S2==30),
678.               aes(xintercept = sgp1.mean),color="#B10318")+
679.     geom_vline(data=fct.means%%filter(S2==30),
680.               aes(xintercept = p.mean),alpha=9/10)+
681.     ggtitle("P <= 0.05/SGPV 1")+
682.     theme(plot.title = element_text(hjust = 0.5))
683.
684.     ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S30_SGPV1.jpeg")
685.
686.
687.
688.   #Normal SGPV2/P<.05
689.   ggplot(factor_sim_rnorm%%filter(Pop_eta2!=0 & S2==30), aes(x=eta2_AB)) +
690.     geom_histogram(data=factor_sim_rnorm%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==30),bins=50,fill="#2CA02C",alpha = 4/10)+
691.     geom_histogram(data=factor_sim_rnorm%%filter(sgpvalue.035_AB==0&Pop_eta2!=0 &
S2==30),fill = "#1f77b4",bins=50,alpha = 6/10)+
692.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
693.     theme_apache()+background_grid()+
694.     geom_vline(data=fct.means%%filter(S2==30),
695.               aes(xintercept = p.mean),alpha=9/10)+
696.     geom_vline(data=fct.means%%filter(S2==30),
697.               aes(xintercept = sgp35.mean),color="#B10318")+
698.     ggtitle("P <= 0.05/SGPV 2")+
699.     theme(plot.title = element_text(hjust = 0.5))
700.
701.     ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S30_SGPV2.jpeg")
702.
703.
704.   #Normal MESP2/P<.05
705.   #Only for n=500
706.   ggplot(factor_sim_rnorm%%filter(Pop_eta2!=0 & S2==30&n==500), aes(x=eta2_AB)) +
707.     geom_histogram(data=factor_sim_rnorm%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==30&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
708.     geom_histogram(data=factor_sim_rnorm%%filter(pvalue_AB<=0.05 & n==500 &
(eta2_AB>= .035|eta2_AB<=-.035)&Pop_eta2!=0 & S2==30),fill = "#1f77b4",bins=50,alpha
= 6/10)+
709.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
710.     theme_apache()+background_grid()+
711.     geom_vline(data=fct.means%%filter(S2==30&n==500),

```



```

712.         aes(xintercept = p.mean), alpha=9/10)+
713.   geom_vline(data=fct.means%>%filter(S2==30&n==500),
714.             aes(xintercept = MESP35.mean), color="#B10318")+
715.   ggtitle("P <= 0.05/MESP 2")+
716.   theme(plot.title = element_text(hjust = 0.5))
717.
718.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S30_MESP2.jpeg")
719.
720. #####
721. #####2x2 GROUP 4:1 Var Ratio DATA#####
722. #####
723.
724.
725.
726. #Normal P<.005/P<.05
727. ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==60), aes(x=eta2_AB)) +
728.   geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==60), bins=50, fill="#2CA02C", alpha = 4/10)+
729.   geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.005&Pop_eta2!=0 &
S2==60), fill = "#1f77b4", bins=50, alpha = 6/10)+
730.   facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales = "free_x")+
731.   theme_apa()+background_grid()+
732.   geom_vline(data=fct.means%>%filter(S2==60),
733.             aes(xintercept = p.mean), alpha=9/10)+
734.   geom_vline(data=fct.means%>%filter(S2==60),
735.             aes(xintercept = p005.mean), color="#B10318")+
736.   ggtitle("P <= 0.05/P <= 0.005")+
737.   theme(plot.title = element_text(hjust = 0.5))
738.
739.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S60_p005.jpeg")
740.
741.
742. #Normal SGPV1/P<.05
743. ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==60), aes(x=eta2_AB)) +
744.   geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==60), bins=50, fill="#2CA02C", alpha = 4/10)+
745.   geom_histogram(data=factor_sim_rnorm%>%filter(sgpvalue.005_AB==0&Pop_eta2!=0 &
S2==60), fill = "#1f77b4", bins=50, alpha = 6/10)+
746.   facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales = "free_x")+
747.   theme_apa()+background_grid()+
748.   geom_vline(data=fct.means%>%filter(S2==60),
749.             aes(xintercept = sgp1.mean), color="#B10318")+
750.   geom_vline(data=fct.means%>%filter(S2==60),
751.             aes(xintercept = p.mean), alpha=9/10)+
752.   ggtitle("P <= 0.05/SGPV 1")+
753.   theme(plot.title = element_text(hjust = 0.5))
754.
755.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S60_SGPV1.jpeg")
756.
757.
758.
759.
760. #Normal SGPV2/P<.05
761. ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==60), aes(x=eta2_AB)) +
762.   geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==60), bins=50, fill="#2CA02C", alpha = 4/10)+
763.   geom_histogram(data=factor_sim_rnorm%>%filter(sgpvalue.035_AB==0&Pop_eta2!=0 &
S2==60), fill = "#1f77b4", bins=50, alpha = 6/10)+
764.   facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales = "free_x")+
765.   theme_apa()+background_grid()+
766.   geom_vline(data=fct.means%>%filter(S2==60),
767.             aes(xintercept = p.mean), alpha=9/10)+

```

```

768.   geom_vline(data=fct.means%>%filter(S2==60),
769.             aes(xintercept = sgp35.mean),color="#B10318")+
770.   ggtitle("P <= 0.05/SGPV 2")+
771.   theme(plot.title = element_text(hjust = 0.5))
772.
773.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S60_SGPV2.jpeg")
774.
775.
776.
777.   #Normal MESP2/P<.05
778.   #Only for n=500
779.   ggplot(factor_sim_rnorm%>%filter(Pop_eta2!=0 & S2==60&n==500), aes(x=eta2_AB)) +
780.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==60&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
781.     geom_histogram(data=factor_sim_rnorm%>%filter(pvalue_AB<=0.05 & n==500 &
(eta2_AB>= .035|eta2_AB<=-.035)&Pop_eta2!=0 & S2==60),fill = "#1f77b4",bins=50,alpha
= 6/10)+
782.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
783.     theme_apo()+background_grid()+
784.     geom_vline(data=fct.means%>%filter(S2==60&n==500),
785.               aes(xintercept = p.mean),alpha=9/10)+
786.     geom_vline(data=fct.means%>%filter(S2==60&n==500),
787.               aes(xintercept = MESP35.mean),color="#B10318")+
788.     ggtitle("P <= 0.05/MESP 2")+
789.     theme(plot.title = element_text(hjust = 0.5))
790.
791.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/NORMAL/Fct.S60_MESP2.jpeg")
792.

```

```

1.  #SKEWED DISTRIBUTION FREQUENCY HISTOGRAM PLOTS
2.
3.  #LOAD PACKAGES
4.  library(tidyverse)
5.  library("ggthemes")
6.  library(jtools) #theme_apo
7.  library(cowplot)#background_grid
8.
9.  #####
10. #####2GROUP SKEWED DATA#####
11. #####
12.
13. #LOAD DATA FILE IF NOT IN R ENVIRONMENT
14. T_sim_rsnorm<-readRDS("~/OneDrive - University of South Carolina/R code/Dissertation
simulation/DATA/T_sim_rsnorm.rds")
15.
16. #EFFECT SIZE MEANS and SD for V-line on plot
17.
18. data1<-T_sim_rsnorm
19. means.skew<-left_join(data1%>%filter(pvalue<=.05 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
20.                   summarise(p.mean=mean(Est_effect),p.sd=sd(Est_effect)),
21.
22.
23.                   data1%>%filter(pvalue<=.005&
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
24.                   summarise(p005.mean=mean(Est_effect),p005.sd=sd(Est_effect)))%>%left_join(
25.
26.
27.                   data1%>%filter(sgpvalue.1==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%

```

```

28. summarise(sgp1.mean=mean(Est_effect),sgp1.sd=sd(Est_effect)))%>%left_join(
29.
30.
31.           data1%>%filter(sgpvalue.35==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
32. summarise(sgp35.mean=mean(Est_effect),sgp35.sd=sd(Est_effect)))%>%left_join(
33.
34.
35.           data1%>%filter(pvalue<=.05 & D!=0 & (cohens_D>=
.1|cohens_D<=-.1))%>%group_by(S2,D,n,Pop_effect)%>%
36. summarise(MESP1.mean=mean(Est_effect),MESP1.sd=sd(Est_effect)))%>%left_join(
37.
38.           data1%>%filter(pvalue<=.05 & D!=0 &
(cohens_D>= .35|cohens_D<=-.35))%>%group_by(S2,D,n,Pop_effect)%>%
39. summarise(MESP35.mean=mean(Est_effect),MESP35.sd=sd(Est_effect)))%>%view
40.
41.
42. #####
43. #####2GROUP SKEWED SET Factors#####
44. #####
45.
46. #Data
47. #Set Factors
48. T_sim_rsnorm<-T_sim_rsnorm%>%
49.   mutate(Pop_effect.ftr=factor(Pop_effect,levels = c("0","3","7.5","12"),
50.     labels=c("Pop Effect = 0","Pop Effect = 3","Pop
Effect = 7.5","Pop Effect = 12")),
51.     n.ftr=factor(n,levels=c("20","50","500"),
52.       labels=c("n = 20","n = 50","n = 500")))
53.
54. means.skew<-means.skew%>%mutate(Pop_effect.ftr=factor(Pop_effect,levels =
c("3","7.5","12"),
55.     labels=c("Pop Effect = 3","Pop
Effect = 7.5","Pop Effect = 12")),
56.     n.ftr=factor(n,levels=c("20","50","500"),
57.       labels=c("n = 20","n = 50","n = 500")))
58.
59.
60.
61.
62.
63.
64. #2G SKEW P<.005/P<.05
65. ggplot(T_sim_rsnorm%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
66.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
67.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.005&D!=0 & S2==15),fill =
"#1f77b4",bins=50,alpha = 6/10)+
68.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
69.   theme_apa()+background_grid()+
70.   geom_vline(data=means.skew%>%filter(S2==15),
71.     aes(xintercept = p.mean),alpha=9/10)+
72.   geom_vline(data=means.skew%>%filter(S2==15),
73.     aes(xintercept = p005.mean),color="#B10318")+
74.   ggtitle("P <= 0.05/P <= 0.005")+
75.   theme(plot.title = element_text(hjust = 0.5))
76.
77.
78. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Skew_p005.jpeg")
79.

```

```

80.
81. #2G SKEW SGPV1/P<.05
82. ggplot(T_sim_rsnorm%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
83.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
      S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
84.   geom_histogram(data=T_sim_rsnorm%>%filter(sgpvalue.1==0&D!=0 & S2==15),fill =
      "#1f77b4",bins=50,alpha = 6/10)+
85.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
86.   theme_apa()+background_grid()+
87.   geom_vline(data=means.skew%>%filter(S2==15),
88.     aes(xintercept = sgp1.mean),color="#B10318")+
89.   geom_vline(data=means.skew%>%filter(S2==15),
90.     aes(xintercept = p.mean),alpha=9/10)+
91.   ggtitle("P <= 0.05/SGPV 1")+
92.   theme(plot.title = element_text(hjust = 0.5))
93.
94. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Skew_SGPV1.jpeg")
95.
96.
97. #2G SKEW SGPV2/P<.05
98. ggplot(T_sim_rsnorm%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
99.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
      S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
100.   geom_histogram(data=T_sim_rsnorm%>%filter(sgpvalue.35==0&D!=0 & S2==15),fill =
      "#1f77b4",bins=50,alpha = 6/10)+
101.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
102.   theme_apa()+background_grid()+
103.   geom_vline(data=means.skew%>%filter(S2==15),
104.     aes(xintercept = p.mean),alpha=9/10)+
105.   geom_vline(data=means.skew%>%filter(S2==15),
106.     aes(xintercept = sgp35.mean),color="#B10318")+
107.   ggtitle("P <= 0.05/SGPV 2")+
108.   theme(plot.title = element_text(hjust = 0.5))
109.
110. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Skew_SGPV2.jpeg")
111.
112.
113. #2G Normal MESP2/P<.05
114. #Only for n=500
115. ggplot(T_sim_rsnorm%>%filter(D!=0 & S2==15&n==500), aes(x=Est_effect)) +
116.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
      S2==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
117.   geom_histogram(data=T_sim_rsnorm%>%filter(pvalue<=0.05 & n==500 & (cohens_D>=
      .35|cohens_D<=-.35)&D!=0 & S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
118.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
119.   theme_apa()+background_grid()+
120.   geom_vline(data=means.skew%>%filter(S2==15&n==500),
121.     aes(xintercept = p.mean),alpha=9/10)+
122.   geom_vline(data=means.skew%>%filter(S2==15&n==500),
123.     aes(xintercept = MESP35.mean),color="#B10318")+
124.   ggtitle("P <= 0.05/MESP 2")+
125.   theme(plot.title = element_text(hjust = 0.5))
126.
127. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Skew_MESP2.jpeg")
128.
129.
130.
131.
132.
133.
134.
135.

```

```

136.
137.
138. #####
139. #####K=3 GROUP SKEW DATA#####
140. #####
141. ANOVA_sim_rsnorm%>%head%>%view
142. ANOVA_sim_rsnorm<-ANOVA_sim_rsnorm%>%mutate(Pop_eta2=case_when(D==0.0 ~ 0,
143.                                     D==0.2 ~0.01,
144.                                     D==0.5 ~0.06,
145.                                     D==0.8 ~0.14))
146. #EFFECT SIZE MEANS and SD
147.
148. k3.skew.means<-left_join(ANOVA_sim_rsnorm%>%filter(pvalue<=.05 &
149. D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%
150. summarise(p.mean=mean(eta2),p.sd=sd(eta2)),
151. ANOVA_sim_rsnorm%>%filter(pvalue<=.005& D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%
152. summarise(p005.mean=mean(eta2),p005.sd=sd(eta2)))%>%left_join(
153.
154.
155. ANOVA_sim_rsnorm%>%filter(sgpvalue.005==0 &
156. D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp1.mean=mean(eta2),
157. sgp005.sd=sd(eta2)))%>%left_join(
158.
159. ANOVA_sim_rsnorm%>%filter(sgpvalue.035==0 &
160. D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp35.mean=mean(eta2),
161. sgp035.sd=sd(eta2)))%>%left_join(
162. ANOVA_sim_rsnorm%>%filter(pvalue<=.05 & D!=0 & (eta2>= .005|eta2<=-
163. .005))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP1.mean=mean(eta2),
164. MESP1.sd=sd(eta2)))%>%left_join(
165. ANOVA_sim_rsnorm%>%filter(pvalue<=.05 & D!=0 & (eta2>= .035|eta2<=-
166. .035))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP35.mean=mean(eta2),
167. MESP35.sd=sd(eta2)))%>%view
168. #####
169. #####K=3 GROUP SKEW DATA SET Factors#####
170. #####
171.
172. #Data
173. #Set Factors
174. ANOVA_sim_rsnorm<-ANOVA_sim_rsnorm%>%
175. mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
176. labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
177. eta^2 = 0.06","Pop eta^2 = 0.14")),
178. n.ftr=factor(n,levels=c("20","50","500"),
179. labels=c("n = 20","n = 50","n = 500")))
180. k3.skew.means<-k3.skew.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2,levels =
181. c("0.01","0.06","0.14"),
182. labels=c("Pop eta^2 = 0.01","Pop
183. eta^2 = 0.06","Pop eta^2 = 0.14")),
184. n.ftr=factor(n,levels=c("20","50","500"),
185. labels=c("n = 20","n = 50","n = 500")))
186. #Normal P<.005/P<.05
187. ggplot(ANOVA_sim_rsnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
188. geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
189. S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
190. geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.005&D!=0 & S3==15),fill
191. = "#1f77b4",bins=50,alpha = 6/10)+
192. facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales="free_x")+

```

```

191.   theme_apo()+background_grid()+
192.   geom_vline(data=k3.skew.means%>%filter(S3==15),
193.             aes(xintercept = p.mean),alpha=9/10)+
194.   geom_vline(data=k3.skew.means%>%filter(S3==15),
195.             aes(xintercept = p005.mean),color="#B10318")+
196.   ggtitle("P <= 0.05/P <= 0.005")+
197.   theme(plot.title = element_text(hjust = 0.5))
198.
199.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/K3.Skew_p005.jpeg")
200.
201.
202.   #Normal SGPV005/P<.05
203.   ggplot(ANOVA_sim_rsnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
204.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
205.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(sgpvalue.005==0&D!=0 &
S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
206.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
207.     theme_apo()+background_grid()+
208.     geom_vline(data=k3.skew.means%>%filter(S3==15),
209.               aes(xintercept = sgp1.mean),color="#B10318")+
210.     geom_vline(data=k3.skew.means%>%filter(S3==15),
211.               aes(xintercept = p.mean),alpha=9/10)+
212.     ggtitle("P <= 0.05/SGPV 2")+
213.     theme(plot.title = element_text(hjust = 0.5))
214.
215.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/K3.Skew_SGPV1.jpeg")
216.
217.
218.   #Normal SGPV2/P<.05
219.   ggplot(ANOVA_sim_rsnorm%>%filter(D!=0 & S3==15), aes(x=eta2)) +
220.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
221.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(sgpvalue.035==0&D!=0 &
S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
222.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
223.     theme_apo()+background_grid()+
224.     geom_vline(data=k3.skew.means%>%filter(S3==15),
225.               aes(xintercept = p.mean),alpha=9/10)+
226.     geom_vline(data=k3.skew.means%>%filter(S3==15),
227.               aes(xintercept = sgp35.mean),color="#B10318")+
228.     ggtitle("P <= 0.05/SGPV 2")+
229.     theme(plot.title = element_text(hjust = 0.5))
230.
231.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/K3.Skew_SGPV2.jpeg")
232.
233.
234.   #Normal MESP2/P<.05
235.   #Only for n=500
236.   ggplot(ANOVA_sim_rsnorm%>%filter(D!=0 & S3==15&n==500), aes(x=eta2)) +
237.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.05&D!=0 &
S3==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
238.     geom_histogram(data=ANOVA_sim_rsnorm%>%filter(pvalue<=0.05 & n==500 & (eta2>=
.035|eta2<=-.035)&D!=0 & S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
239.     facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
240.     theme_apo()+background_grid()+
241.     geom_vline(data=k3.skew.means%>%filter(S3==15&n==500),
242.               aes(xintercept = p.mean),alpha=9/10)+
243.     geom_vline(data=k3.skew.means%>%filter(S3==15&n==500),
244.               aes(xintercept = MESP35.mean),color="#B10318")+
245.     ggtitle("P <= 0.05/MESP 2")+
246.     theme(plot.title = element_text(hjust = 0.5))

```

```

247.
248. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/K3.Skew_MESP2.jpeg")
249.
250.
251.
252. #####
253. #####2x2 GROUP SKEW DATA#####
254. #####
255.
256. #EFFECT SIZE MEANS and SD
257.
258.
259. fct.skew.means<-left_join(factor_sim_rsnorm%>%filter(pvalue_AB<=.05 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
260. summarise(p.mean=mean(eta2_AB),p.sd=sd(eta2_AB)),
261.
262. factor_sim_rsnorm%>%filter(pvalue_AB<=.005&
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
263. summarise(p005.mean=mean(eta2_AB),p005.sd=sd(eta2_AB)))%>%left_join(
264.
265. factor_sim_rsnorm%>%filter(sgpvalue.005_AB==0 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
266. summarise(sgp1.mean=mean(eta2_AB),sgp005.sd=sd(eta2_AB)))%>%left_join(
267.
268. factor_sim_rsnorm%>%filter(sgpvalue.035_AB==0 &
Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
269. summarise(sgp35.mean=mean(eta2_AB),sgp035.sd=sd(eta2_AB)))%>%left_join(
270.
271. factor_sim_rsnorm%>%filter(pvalue_AB<=.05 &
Pop_eta2!=0 & (eta2_AB>= .005|eta2_AB<=-.005))%>%group_by(S2,n,Pop_eta2)%>%
272. summarise(MESP1.mean=mean(eta2_AB),MESP1.sd=sd(eta2_AB)))%>%left_join(
273.
274. #####
275. factor_sim_rsnorm%>%filter(pvalue_AB<=.05 & Pop_eta2!=0 & (eta2_AB>= .035|eta2_AB<=-
.035))%>%group_by(S2,n,Pop_eta2)%>%
276. summarise(MESP35.mean=mean(eta2_AB),MESP35.sd=sd(eta2_AB)))%>%view
277.
278. #####
#####
#####
279. #####
#####
#####
280.
281.
282. #####
283. #####2x2 GROUP SKEW DATA SET Factors#####
284. #####
285. factor_sim_rsnorm%>%head%>%view
286.
287. #Data
288. #Set Factors
289. factor_sim_rsnorm<-factor_sim_rsnorm%>%
290. mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
291. labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
eta^2 = 0.06","Pop eta^2 = 0.14")),
292. n.ftr=factor(n,levels=c("20","50","500"),

```



```

293.           labels=c("n = 20", "n = 50", "n = 500"))
294.
295.   fct.skew.means<-fct.skew.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2, levels =
c("0.01", "0.06", "0.14"),
296.                                           labels=c("Pop eta^2 =
0.01", "Pop eta^2 = 0.06", "Pop eta^2 = 0.14")),
297.                                           n.ftr=factor(n, levels=c("20", "50", "500"),
298.                                           labels=c("n = 20", "n = 50", "n
= 500")))
299.
300.
301.   #Normal P<.005/P<.05
302.   ggplot(factor_sim_rsnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
303.     geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15), bins=50, fill="#2CA02C", alpha = 4/10)+
304.     geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.005&Pop_eta2!=0 &
S2==15), fill = "#1f77b4", bins=50, alpha = 6/10)+
305.     facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales="free_x")+
306.     theme_apo()+background_grid()+
307.     geom_vline(data=fct.skew.means%>%filter(S2==15),
308.               aes(xintercept = p.mean), alpha=9/10)+
309.     geom_vline(data=fct.skew.means%>%filter(S2==15),
310.               aes(xintercept = p005.mean), color="#B10318")+
311.     ggtitle("P <= 0.05/P<.005")+
312.     theme(plot.title = element_text(hjust = 0.5))
313.
314.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/Fct.Skew_p005.jpeg")
315.
316.
317.   #Normal SGPV005/P<.05
318.   ggplot(factor_sim_rsnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
319.     geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15), bins=50, fill="#2CA02C", alpha = 4/10)+
320.     geom_histogram(data=factor_sim_rsnorm%>%filter(sgpvalue.005_AB==0&Pop_eta2!=0 &
S2==15), fill = "#1f77b4", bins=50, alpha = 6/10)+
321.     facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales = "free_x")+
322.     theme_apo()+background_grid()+
323.     geom_vline(data=fct.skew.means%>%filter(S2==15),
324.               aes(xintercept = sgp1.mean), color="#B10318")+
325.     geom_vline(data=fct.skew.means%>%filter(S2==15),
326.               aes(xintercept = p.mean), alpha=9/10)+
327.     ggtitle("P <= 0.05/SGPV 1")+
328.     theme(plot.title = element_text(hjust = 0.5))
329.
330.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/Fct.Skew_SGPV1.jpeg")
331.
332.
333.
334.
335.   #Normal SGPV2/P<.05
336.   ggplot(factor_sim_rsnorm%>%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
337.     geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15), bins=50, fill="#2CA02C", alpha = 4/10)+
338.     geom_histogram(data=factor_sim_rsnorm%>%filter(sgpvalue.035_AB==0&Pop_eta2!=0 &
S2==15), fill = "#1f77b4", bins=50, alpha = 6/10)+
339.     facet_wrap(~ Pop_eta2.ftr+n.ftr, nrow=4, ncol=3, scales = "free_x")+
340.     theme_apo()+background_grid()+
341.     geom_vline(data=fct.skew.means%>%filter(S2==15),
342.               aes(xintercept = p.mean), alpha=9/10)+
343.     geom_vline(data=fct.skew.means%>%filter(S2==15),
344.               aes(xintercept = sgp35.mean), color="#B10318")+
345.     ggtitle("P <= 0.05/SGPV 2")+
346.     theme(plot.title = element_text(hjust = 0.5))

```



```

347.
348. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/Fct.Skew_SGPV2.jpeg")
349.
350.
351.
352. #Normal MESP2/P<.05
353. #Only for n=500
354. ggplot(factor_sim_rsnorm%>%filter(Pop_eta2!=0 & S2==15&n==500), aes(x=eta2_AB)) +
355.   geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15&n==500), bins=50, fill="#2CA02C", alpha = 4/10)+
356.   geom_histogram(data=factor_sim_rsnorm%>%filter(pvalue_AB<=0.05 & n==500 &
(eta2_AB>= .035|eta2_AB<=-.035)&Pop_eta2!=0 & S2==15), fill = "#1f77b4", bins=50, alpha
= 6/10)+
357.   facet_wrap(~ Pop_eta2, ftr+n.ftr, nrow=1, ncol=3, scales = "free_x")+
358.   theme_apa()+background_grid()+
359.   geom_vline(data=fct.skew.means%>%filter(S2==15&n==500),
360.             aes(xintercept = p.mean), alpha=9/10)+
361.   geom_vline(data=fct.skew.means%>%filter(S2==15&n==500),
362.             aes(xintercept = MESP35.mean), color="#B10318")+
363.   ggtitle("P <= 0.05/MESP 2")+
364.   theme(plot.title = element_text(hjust = 0.5))
365.
366. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/Fct.Skew_MESP2.jpeg")
367.

```

```

1. #BIMODAL DISTRIBUTION FREQUENCY HISTOGRAM PLOTS
2.
3. #LOAD PACKAGES
4. library(tidyverse)
5. library("ggthemes")
6. library(jtools) #theme_apa
7. library(cowplot)#background_grid
8.
9. #####
10. #####2GROUP BIMODAL DATA#####
11. #####
12.
13. #LOAD DATA FILE IF NOT IN R ENVIRONMENT
14. T_sim_rsnorm<-readRDS("~/OneDrive - University of South Carolina/R code/Dissertation
simulation/DATA/T_sim_bimodal.rds")
15.
16. #EFFECT SIZE MEANS and SD for V-line on plot
17. data2<-T_sim_bimodal
18. means.bimodal<-left_join(data2%>%filter(pvalue<=.05 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
19.   summarise(p.mean=mean(Est_effect),p.sd=sd(Est_effect)),
20.
21.   data2%>%filter(pvalue<=.005&
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
22.   summarise(p005.mean=mean(Est_effect),p005.sd=sd(Est_effect)))%>%left_join(
23.
24.   data2%>%filter(sgpvalue.1==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
25.   summarise(sgp1.mean=mean(Est_effect),sgp1.sd=sd(Est_effect)))%>%left_join(
26.
27.   data2%>%filter(sgpvalue.35==0 &
D!=0)%>%group_by(S2,D,n,Pop_effect)%>%
28.   summarise(sgp35.mean=mean(Est_effect),sgp35.sd=sd(Est_effect)))%>%left_join(
29.

```

```

30. data2%>%filter(pvalue<=.05 & D!=0 & (cohens_D>=
31. .1|cohens_D<=-.1))%>%group_by(S2,D,n,Pop_effect)%>%
32. summarise(MESP1.mean=mean(Est_effect),MESP1.sd=sd(Est_effect))%>%left_join(
33. data2%>%filter(pvalue<=.05 & D!=0 &
34. (cohens_D>= .35|cohens_D<=-.35))%>%group_by(S2,D,n,Pop_effect)%>%
35. summarise(MESP35.mean=mean(Est_effect),MESP35.sd=sd(Est_effect))%>%view
36.
37.
38. #####
39. #####2GROUP BIMODAL SET Factors#####
40. #####
41.
42. #Data
43. #Set Factors
44. T_sim_bimodal<-T_sim_bimodal%>%
45. mutate(Pop_effect.ftr=factor(Pop_effect,levels = c("0","3","7.5","12"),
46. labels=c("Pop Effect = 0","Pop Effect = 3","Pop
47. Effect = 7.5","Pop Effect = 12")),
48. n.ftr=factor(n,levels=c("20","50","500"),
49. labels=c("n = 20","n = 50","n = 500")))
50. means.bimodal<-means.bimodal%>%mutate(Pop_effect.ftr=factor(Pop_effect,levels =
51. c("3","7.5","12"), labels=c("Pop Effect =
52. 3","Pop Effect = 7.5","Pop Effect = 12")),
53. n.ftr=factor(n,levels=c("20","50","500"),
54. labels=c("n = 20","n = 50","n =
55. 500")))
56.
57.
58. #2G Normal P<.005/P<.05
59. ggplot(T_sim_bimodal%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
60. geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.05&D!=0 &
61. S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
62. geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.005&D!=0 & S2==15),fill =
63. "#1f77b4",bins=50,alpha = 6/10)+
64. facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
65. theme_apa()+background_grid()+
66. geom_vline(data=means.bimodal%>%filter(S2==15),
67. aes(xintercept = p.mean),alpha=9/10)+
68. geom_vline(data=means.bimodal%>%filter(S2==15),
69. aes(xintercept = p005.mean),color="#B10318")+
70. ggtitle("P <= 0.05/P <= 0.005")+
71. theme(plot.title = element_text(hjust = 0.5))
72.
73.
74. #2G Normal SGPV1/P<.05
75. ggplot(T_sim_bimodal%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
76. geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.05&D!=0 &
77. S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
78. geom_histogram(data=T_sim_bimodal%>%filter(sgpvalue.1==0&D!=0 & S2==15),fill =
79. "#1f77b4",bins=50,alpha = 6/10)+
80. facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
81. theme_apa()+background_grid()+
82. geom_vline(data=means.bimodal%>%filter(S2==15),
83. aes(xintercept = sgp1.mean),color="#B10318")+

```

```

82.   geom_vline(data=means.bimodal%>%filter(S2==15),
83.             aes(xintercept = p.mean),alpha=9/10)+
84. ggtitle("P <= 0.05/SGPV 1")+
85.   theme(plot.title = element_text(hjust = 0.5))
86.
87. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Bimodal_SGPV1.jpeg")
88.
89.
90.
91.
92. #2G Normal SGPV2/P<.05
93. ggplot(T_sim_bimodal%>%filter(D!=0 & S2==15), aes(x=Est_effect)) +
94.   geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.05&D!=0 &
S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
95.   geom_histogram(data=T_sim_bimodal%>%filter(sgpvalue.35==0&D!=0 & S2==15),fill =
"#1f77b4",bins=50,alpha = 6/10)+
96.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
97.   theme_apa()+background_grid()+
98.   geom_vline(data=means.bimodal%>%filter(S2==15),
99.             aes(xintercept = p.mean),alpha=9/10)+
100.  geom_vline(data=means.bimodal%>%filter(S2==15),
101.            aes(xintercept = sgp35.mean),color="#B10318")+
102.  ggtitle("P <= 0.05/SGPV 2")+
103.  theme(plot.title = element_text(hjust = 0.5))
104.
105. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Bimodal_SGPV2.jpeg")
106.
107.
108.
109. #2G Normal MESP2/P<.05
110. #Only for n=500
111. ggplot(T_sim_bimodal%>%filter(D!=0 & S2==15&n==500), aes(x=Est_effect)) +
112.   geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.05&D!=0 &
S2==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
113.   geom_histogram(data=T_sim_bimodal%>%filter(pvalue<=0.05 & n==500 & (cohens_D>=
.35|cohens_D<=-.35)&D!=0 & S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
114.   facet_wrap(~ Pop_effect.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
115.   theme_apa()+background_grid()+
116.   geom_vline(data=means.bimodal%>%filter(S2==15&n==500),
117.             aes(xintercept = p.mean),alpha=9/10)+
118.   geom_vline(data=means.bimodal%>%filter(S2==15&n==500),
119.             aes(xintercept = MESP35.mean),color="#B10318")+
120.   ggtitle("P <= 0.05/SGPV 2")+
121.   theme(plot.title = element_text(hjust = 0.5))
122.
123. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/G2.Bimodal_SGPV2.jpeg")
124.
125.
126. #####
127. #####K=3 GROUP SKEW DATA#####
128. #####
129. ANOVA_sim_bimodal%>%head%>%view
130.
131. #EFFECT SIZE MEANS and SD
132.
133. k3.bimodal.means<-left_join(ANOVA_sim_bimodal%>%filter(pvalue<=.05 &
D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%
134.   summarise(p.mean=mean(eta2),p.sd=sd(eta2)),
135.
136.   ANOVA_sim_bimodal%>%filter(pvalue<=.005&
D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%

```

```

137. summarise(p005.mean=mean(eta2),p005.sd=sd(eta2)))%>%left_join(
138.
139.
140. ANOVA_sim_bimodal%>%filter(sgpvalue.005==0 &
D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp1.mean=mean(eta2),
141. sgp005.sd=sd(eta2)))%>%left_join(
142.
143.
144. ANOVA_sim_bimodal%>%filter(sgpvalue.035==0 &
D!=0)%>%group_by(S3,D,n,Pop_eta2)%>%summarise(sgp35.mean=mean(eta2),
145. sgp035.sd=sd(eta2)))%>%left_join(
146.
147. ANOVA_sim_bimodal%>%filter(pvalue<=.05 & D!=0 & (eta2>= .005|eta2<=
.005))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP1.mean=mean(eta2),
148. MESP1.sd=sd(eta2)))%>%left_join(
149.
150. ANOVA_sim_bimodal%>%filter(pvalue<=.05 & D!=0 & (eta2>= .035|eta2<=
.035))%>%group_by(S3,D,n,Pop_eta2)%>%summarise(MESP35.mean=mean(eta2),
151. MESP35.sd=sd(eta2)))%>%view
152.
153. #####
154. #####K=3 GROUP SKEW DATA SET Factors#####
155. #####
156.
157. #Data
158. #Set Factors
159. ANOVA_sim_bimodal<-ANOVA_sim_bimodal%>%
160. mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
161. labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
eta^2 = 0.06","Pop eta^2 = 0.14")),
162. n.ftr=factor(n,levels=c("20","50","500"),
163. labels=c("n = 20","n = 50","n = 500")))
164.
165. k3.bimodal.means<-k3.bimodal.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2,levels =
c("0.01","0.06","0.14"),
166. labels=c("Pop eta^2 =
0.01","Pop eta^2 = 0.06","Pop eta^2 = 0.14")),
167. n.ftr=factor(n,levels=c("20","50","500"),
168. labels=c("n = 20","n = 50","n =
500")))
169.
170.
171. #Normal P<.005/P<.05
172. ggplot(ANOVA_sim_bimodal%>%filter(D!=0 & S3==15), aes(x=eta2)) +
173. geom_histogram(data=ANOVA_sim_bimodal%>%filter(pvalue<=0.05&D!=0 &
S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
174. geom_histogram(data=ANOVA_sim_bimodal%>%filter(pvalue<=0.005&D!=0 & S3==15),fill
="#1f77b4",bins=50,alpha = 6/10)+
175. facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales="free_x")+
176. theme_apo()+background_grid()+
177. geom_vline(data=k3.bimodal.means%>%filter(S3==15),
178. aes(xintercept = p.mean),alpha=9/10)+
179. geom_vline(data=k3.bimodal.means%>%filter(S3==15),
180. aes(xintercept = p005.mean),color="#B10318")+
181. ggtitle("P <= 0.05/P <= 0.005")+
182. theme(plot.title = element_text(hjust = 0.5))
183.

```

```

184. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/K3.Bimodal_p005.jpeg")
185.
186.
187. #Normal SGPV005/P<.05
188. ggplot(ANOVA_sim_bimodal%%filter(D!=0 & S3==15), aes(x=eta2)) +
189.   geom_histogram(data=ANOVA_sim_bimodal%%filter(pvalue<=0.05&D!=0 &
S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
190.   geom_histogram(data=ANOVA_sim_bimodal%%filter(sgpvalue.005==0&D!=0 &
S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
191.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
192.   theme_apo()+background_grid()+
193.   geom_vline(data=k3.bimodal.means%%filter(S3==15),
194.             aes(xintercept = sgp1.mean),color="#B10318")+
195.   geom_vline(data=k3.bimodal.means%%filter(S3==15),
196.             aes(xintercept = p.mean),alpha=9/10)+
197.   ggtitle("P <= 0.05/SGPV 1")+
198.   theme(plot.title = element_text(hjust = 0.5))
199.
200. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/BIMODAL/K3.Bimodal_SGPV1.jpeg")
201.
202.
203.
204. #Normal SGPV2/P<.05
205. ggplot(ANOVA_sim_bimodal%%filter(D!=0 & S3==15), aes(x=eta2)) +
206.   geom_histogram(data=ANOVA_sim_bimodal%%filter(pvalue<=0.05&D!=0 &
S3==15),bins=50,fill="#2CA02C",alpha = 4/10)+
207.   geom_histogram(data=ANOVA_sim_bimodal%%filter(sgpvalue.035==0&D!=0 &
S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
208.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
209.   theme_apo()+background_grid()+
210.   geom_vline(data=k3.bimodal.means%%filter(S3==15),
211.             aes(xintercept = p.mean),alpha=9/10)+
212.   geom_vline(data=k3.bimodal.means%%filter(S3==15),
213.             aes(xintercept = sgp35.mean),color="#B10318")+
214.   ggtitle("P <= 0.05/SGPV 2")+
215.   theme(plot.title = element_text(hjust = 0.5))
216.
217. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/BIMODAL/K3.Bimodal_SGPV2.jpeg")
218.
219.
220. #Normal MESP2/P<.05
221. #Only for n=500
222. ggplot(ANOVA_sim_bimodal%%filter(D!=0 & S3==15&n==500), aes(x=eta2)) +
223.   geom_histogram(data=ANOVA_sim_bimodal%%filter(pvalue<=0.05&D!=0 &
S3==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
224.   geom_histogram(data=ANOVA_sim_bimodal%%filter(pvalue<=0.05 & n==500 & (eta2>=
.035|eta2<=-.035)&D!=0 & S3==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
225.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
226.   theme_apo()+background_grid()+
227.   geom_vline(data=k3.bimodal.means%%filter(S3==15&n==500),
228.             aes(xintercept = p.mean),alpha=9/10)+
229.   geom_vline(data=k3.bimodal.means%%filter(S3==15&n==500),
230.             aes(xintercept = MESP35.mean),color="#B10318")+
231.   ggtitle("P <= 0.05/MESP 2")+
232.   theme(plot.title = element_text(hjust = 0.5))
233.
234. ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/BIMODAL/K3.Bimodal_MESP2.jpeg")
235.
236. #####
237. #####2x2 GROUP SKEW DATA#####
238. #####

```

```

239.
240. #EFFECT SIZE MEANS and SD
241.
242.
243. fct.bimodal.means<-left_join(factor_sim_bimodal%>%filter(pvalue_AB<=.05 &
  Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
244.     summarise(p.mean=mean(eta2_AB),p.sd=sd(eta2_AB)),
245.
246.     factor_sim_bimodal%>%filter(pvalue_AB<=.005&
  Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
247.     summarise(p005.mean=mean(eta2_AB),p005.sd=sd(eta2_AB)))%>%left_join(
248.
249.     factor_sim_bimodal%>%filter(sgpvalue.005_AB==0 &
  Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
250.     summarise(sgp1.mean=mean(eta2_AB),sgp005.sd=sd(eta2_AB)))%>%left_join(
251.
252.     factor_sim_bimodal%>%filter(sgpvalue.035_AB==0 &
  Pop_eta2!=0)%>%group_by(S2,n,Pop_eta2)%>%
253.     summarise(sgp35.mean=mean(eta2_AB),sgp035.sd=sd(eta2_AB)))%>%left_join(
254.
255.     factor_sim_bimodal%>%filter(pvalue_AB<=.05 &
  Pop_eta2!=0 & (eta2_AB>= .005|eta2_AB<=-.005))%>%group_by(S2,n,Pop_eta2)%>%
256.     summarise(MESP1.mean=mean(eta2_AB),MESP1.sd=sd(eta2_AB)))%>%left_join(
257.
258.     #####
  factor_sim_bimodal%>%filter(pvalue_AB<=.05 & Pop_eta2!=0 & (eta2_AB>=
  .035|eta2_AB<=-.035))%>%group_by(S2,n,Pop_eta2)%>%
259.     summarise(MESP35.mean=mean(eta2_AB),MESP35.sd=sd(eta2_AB)))%>%view
260.
261.
262. #####
  #####
  #####
263. #####
  #####
  #####
264.
265.
266. #####
267. #####2x2 GROUP BIMODAL DATA SET Factors#####
268. #####
269. factor_sim_bimodal%>%head%>%view
270.
271. #Data
272. #Set Factors
273. factor_sim_bimodal<-factor_sim_bimodal%>%
274.   mutate(Pop_eta2.ftr=factor(Pop_eta2,levels = c("0","0.01","0.06","0.14"),
275.     labels=c("Pop eta^2 = 0","Pop eta^2 = 0.01","Pop
  eta^2 = 0.06","Pop eta^2 = 0.14")),
276.     n.ftr=factor(n,levels=c("20","50","500"),
277.     labels=c("n = 20","n = 50","n = 500")))
278.
279. fct.bimodal.means<-fct.bimodal.means%>%mutate(Pop_eta2.ftr=factor(Pop_eta2,levels
  = c("0.01","0.06","0.14"),
280.     labels=c("Pop eta^2 =
  0.01","Pop eta^2 = 0.06","Pop eta^2 = 0.14")),
281.     n.ftr=factor(n,levels=c("20","50","500"),
282.     labels=c("n = 20","n = 50","n
  = 500")))

```

```

283.
284.
285. #Normal P<.005/P<.05
286. ggplot(factor_sim_bimodal%%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
287.   geom_histogram(data=factor_sim_bimodal%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
288.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
289.   geom_histogram(data=factor_sim_bimodal%%filter(pvalue_AB<=0.005&Pop_eta2!=0 &
290.     S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
291.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales="free_x")+
292.   theme_apo()+background_grid()+
293.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
294.     aes(xintercept = p.mean),alpha=9/10)+
295.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
296.     aes(xintercept = p005.mean),color="#B10318")+
297.   ggtitle("P <= 0.05/P <= 0.005")+
298.   theme(plot.title = element_text(hjust = 0.5))
299.
300.
301. #Normal SGPV005/P<.05
302. ggplot(factor_sim_bimodal%%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
303.   geom_histogram(data=factor_sim_bimodal%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
304.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
305.   geom_histogram(data=factor_sim_bimodal%%filter(sgpvalue.005_AB==0&Pop_eta2!=0 &
306.     S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
307.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
308.   theme_apo()+background_grid()+
309.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
310.     aes(xintercept = sgp1.mean),color="#B10318")+
311.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
312.     aes(xintercept = p.mean),alpha=9/10)+
313.   ggtitle("P <= 0.05/SGPV 1")+
314.   theme(plot.title = element_text(hjust = 0.5))
315.
316.
317.
318.
319. #Normal SGPV2/P<.05
320. ggplot(factor_sim_bimodal%%filter(Pop_eta2!=0 & S2==15), aes(x=eta2_AB)) +
321.   geom_histogram(data=factor_sim_bimodal%%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
322.     S2==15),bins=50,fill="#2CA02C",alpha = 4/10)+
323.   geom_histogram(data=factor_sim_bimodal%%filter(sgpvalue.035_AB==0&Pop_eta2!=0 &
324.     S2==15),fill = "#1f77b4",bins=50,alpha = 6/10)+
325.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=4,ncol=3,scales = "free_x")+
326.   theme_apo()+background_grid()+
327.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
328.     aes(xintercept = p.mean),alpha=9/10)+
329.   geom_vline(data=fct.bimodal.means%%filter(S2==15),
330.     aes(xintercept = sgp35.mean),color="#B10318")+
331.   ggtitle("P <= 0.05/SGPV 2")+
332.   theme(plot.title = element_text(hjust = 0.5))
333.
334.
335.
336. #Normal MESP2/P<.05
337. #Only for n=500
338. ggplot(factor_sim_bimodal%%filter(Pop_eta2!=0 & S2==15&n==500), aes(x=eta2_AB)) +

```



```

339.   geom_histogram(data=factor_sim_bimodal%>%filter(pvalue_AB<=0.05&Pop_eta2!=0 &
S2==15&n==500),bins=50,fill="#2CA02C",alpha = 4/10)+
340.   geom_histogram(data=factor_sim_bimodal%>%filter(pvalue_AB<=0.05 & n==500 &
(eta2_AB>= .035|eta2_AB<=-.035)&Pop_eta2!=0 & S2==15),fill = "#1f77b4",bins=50,alpha
= 6/10)+
341.   facet_wrap(~ Pop_eta2.ftr+n.ftr,nrow=1,ncol=3,scales = "free_x")+
342.   theme_apo()+background_grid()+
343.   geom_vline(data=fct.bimodal.means%>%filter(S2==15&n==500),
344.             aes(xintercept = p.mean),alpha=9/10)+
345.   geom_vline(data=fct.bimodal.means%>%filter(S2==15&n==500),
346.             aes(xintercept = MESP35.mean),color="#B10318")+
347.   ggtitle("P <= 0.05/MESP 2")+
348.   theme(plot.title = element_text(hjust = 0.5))
349.
350.   ggsave("/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Histograms/BIMODAL/Fct.Bimodal_MESP2.jpeg")
351.

```

```

1. #####
2. #Confidence Interval plots
3. #####
4.
5. #T_Normal
6. #####
7. #T_names
8. T_nms<-c("p<=.05_TI/Power",
9.          "p<=.005_TI/Power",
10.         "SGPV.1=0_TI/Power",
11.         "SGPV.35=0_TI/Power",
12.         "MESP.1_TI/Power",
13.         "MESP.35_TI/Power")
14.
15. #Prep data and calculate CI
16. X<-T_rnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
17.   mutate(`p<=.05.ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`,.after
= `p<=.05_TI/Power`)%>%
18.   mutate(`p<=.005.ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`,.after =
`p<=.005_TI/Power`)%>%
19.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.1=0_TI/Power`,.after =
`SGPV.1=0_TI/Power`)%>%
20.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.35=0_TI/Power`,.after =
`SGPV.35=0_TI/Power`)%>%
21.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.1_TI/Power`,.after
= `MESP.1_TI/Power`)%>%
22.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.35_TI/Power`,.after
= `MESP.35_TI/Power`)%>%
23.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`,.after
= `p<=.05_TI/Power`)%>%
24.   mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`,.after =
`p<=.005_TI/Power`)%>%
25.   mutate(`SGPV.1=0.ic_L` = `SGPV.1=0_TI/Power` - `SGPV.1=0_CI_L`,.after =
`SGPV.1=0_TI/Power`)%>%
26.   mutate(`SGPV.35=0.ic_L` = `SGPV.35=0_TI/Power` - `SGPV.35=0_CI_L`,.after =
`SGPV.35=0_TI/Power`)%>%
27.   mutate(`MESP.1.ic_L` = `MESP.1_TI/Power` - `MESP.1_CI_L`,.after
= `MESP.1_TI/Power`)%>%
28.   mutate(`MESP.35.ic_L` = `MESP.35_TI/Power` - `MESP.35_CI_L`,.after
= `MESP.35_TI/Power`)%>%
29.   select(distr,D,n,S1,S2,
30.          "p<=.05_TI/Power", `p<=.05.ic_H`, `p<=.05.ic_L`,
31.          "p<=.005_TI/Power", `p<=.005.ic_H`, `p<=.005.ic_L`,
32.          "SGPV.1=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
33.          "SGPV.35=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
34.          "MESP.1_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,

```



```

35.       "MESP.35_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
36.
37. #Wrangle data
38. test.ci<-X%>%select("distr","D","n","S1","S2",ends_with("Power"))%>%
39.   ungroup%>%
40.   pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%>%
41.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500"))),
42.         method=factor(method,levels =
c('MESP.35_TI/Power','MESP.1_TI/Power','SGPV.35=0_TI/Power','SGPV.1=0_TI/Power','p<=
.005_TI/Power','p<=.05_TI/Power'),
43.         labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05'))))
44.
45. #Combine with ic data
46. T_rnorm_ci.fig<-
  cbind(test.ci,X%>%select("distr","D","n","S1","S2",ends_with("ic_H"))%>%
47.    ungroup%>%
48.    pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%>%
49.    rename(ic_h=value)%>%select(-c(distr,D,n,S1,S2,method))%>%
50.    cbind(X%>%select("distr","D","n","S1","S2",ends_with("ic_L"))%>%
51.      ungroup%>%
52.      pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%>%
53.      rename(ic_l=value)%>%select(-c(distr,D,n,S1,S2,method)))
54.
55.
56. #####
57. #K3 Normal
58. #####
59. ANOVA_rnorm_TI.PWR.Results_CI%>%names
60.
61. #T_names
62. AN_nms<-c("p<=.05_TI/Power",
63.           "p<=.005_TI/Power",
64.           "SGPV.005=0_TI/Power",
65.           "SGPV.035=0_TI/Power",
66.           "MESP.005_TI/Power",
67.           "MESP.035_TI/Power")
68.
69. #Prep data and calculate IC
70. X<-ANOVA_rnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
71.   mutate(`p<=.05.ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`, .after =
`p<=.05_TI/Power`)%>%
72.   mutate(`p<=.005.ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`, .after =
`p<=.005_TI/Power`)%>%
73.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.005=0_TI/Power`, .after =
`SGPV.005=0_TI/Power`)%>%
74.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.035=0_TI/Power`, .after =
`SGPV.035=0_TI/Power`)%>%
75.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.005_TI/Power`, .after =
`MESP.005_TI/Power`)%>%
76.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.035_TI/Power`, .after =
`MESP.035_TI/Power`)%>%
77.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`, .after =
`p<=.05_TI/Power`)%>%
78.   mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`, .after =
`p<=.005_TI/Power`)%>%
79.   mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
`SGPV.005=0_TI/Power`)%>%
80.   mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
`SGPV.035=0_TI/Power`)%>%
81.   mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
`MESP.005_TI/Power`)%>%
82.   mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
`MESP.035_TI/Power`)%>%
83.   select(distr,Pop_eta2,n,S1,S2,S3,

```

```

84.         "p<=.05_TI/Power",      `p<=.05.ic_H`      , `p<=.05.ic_L`,
85.         "p<=.005_TI/Power",     `p<=.005.ic_H`   , `p<=.005.ic_L`,
86.         "SGPV.005=0_TI/Power",   `SGPV.1=0.ic_H`  , `SGPV.1=0.ic_L`,
87.         "SGPV.035=0_TI/Power",   `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
88.         "MESP.005_TI/Power",     `MESP.1.ic_H`    , `MESP.1.ic_L`,
89.         "MESP.035_TI/Power",     `MESP.35.ic_H`   , `MESP.35.ic_L`)
90.
91. #Wrangle data
92. test.ci<-X%>%select("distr","Pop_eta2","n","S1","S2","S3",ends_with("Power"))%>%
93.   ungroup%>%
94.   pivot_longer(!c("distr","Pop_eta2","n","S1","S2","S3"),names_to = "method")%>%
95.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
96.     method=factor(method,levels =
97.       c('MESP.035_TI/Power','MESP.005_TI/Power','SGPV.035=0_TI/Power','SGPV.005=0_TI/Power
98.       ','p<=.005_TI/Power','p<=.05_TI/Power')),
99.       labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05'))))
100. #Combine with ic data
101. ANOVA_rnorm_ci.fig<-
102.   cbind(test.ci,X%>%select("distr","Pop_eta2","n","S1","S2","S3",ends_with("ic_H"))%>%
103.     ungroup%>%
104.     pivot_longer(!c("distr","Pop_eta2","n","S1","S2","S3"),names_to = "method")%>%
105.     rename(ic_h=value)%>%select(-
106.       c(distr,Pop_eta2,n,S1,S2,S3,method)))%>%
107.     cbind(X%>%select("distr","Pop_eta2","n","S1","S2","S3",ends_with("ic_L"))%>%
108.       ungroup%>%
109.       pivot_longer(!c("distr","Pop_eta2","n","S1","S2","S3"),names_to =
110.         "method")%>%
111.       rename(ic_l=value)%>%select(-c(distr,Pop_eta2,n,S1,S2,S3,method)))
112. #####
113. #Factor Normal
114. #####
115. Factor_rnorm_TI.PWR.Results_CI%>%names
116. #T_names
117. AN_nms<-c("p<=.05_TI/Power",
118.           "p<=.005_TI/Power",
119.           "SGPV.005=0_TI/Power",
120.           "SGPV.035=0_TI/Power",
121.           "MESP.005_TI/Power",
122.           "MESP.035_TI/Power")
123. #Prep data and calculate IC
124. X<-Factor_rnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
125.   mutate(`p<=.05.ic_H`      = `p<=.05_CI_H`      - `p<=.05_TI/Power`,      .after =
126.     `p<=.05_TI/Power`)%>%
127.   mutate(`p<=.005.ic_H`     = `p<=.005_CI_H`     - `p<=.005_TI/Power`,     .after =
128.     `p<=.005_TI/Power`)%>%
129.   mutate(`SGPV.1=0.ic_H`    = `SGPV.1=0_CI_H`    - `SGPV.005=0_TI/Power`, .after =
130.     `SGPV.005=0_TI/Power`)%>%
131.   mutate(`SGPV.35=0.ic_H`   = `SGPV.35=0_CI_H`   - `SGPV.035=0_TI/Power`, .after =
132.     `SGPV.035=0_TI/Power`)%>%
133.   mutate(`MESP.1.ic_H`     = `MESP.1_CI_H`     - `MESP.005_TI/Power`, .after =
134.     `MESP.005_TI/Power`)%>%
135.   mutate(`MESP.35.ic_H`    = `MESP.35_CI_H`    - `MESP.035_TI/Power`, .after =
136.     `MESP.035_TI/Power`)%>%
137.   mutate(`p<=.05.ic_L`     = `p<=.05_TI/Power`   - `p<=.05_CI_L`,      .after =
138.     `p<=.05_TI/Power`)%>%
139.   mutate(`p<=.005.ic_L`    = `p<=.005_TI/Power`  - `p<=.005_CI_L`,      .after =
140.     `p<=.005_TI/Power`)%>%

```

```

133. mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
`SGPV.005=0_TI/Power`)%>%
134. mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
`SGPV.035=0_TI/Power`)%>%
135. mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
`MESP.005_TI/Power`)%>%
136. mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
`MESP.035_TI/Power`)%>%
137. select(distr, Pop_eta2, n, S1, S2,
138.         "p<=.05.ic_H", `p<=.05.ic_H`, `p<=.05.ic_L`,
139.         "p<=.005.ic_H", `p<=.005.ic_H`, `p<=.005.ic_L`,
140.         "SGPV.005=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
141.         "SGPV.035=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
142.         "MESP.005_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,
143.         "MESP.035_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
144.
145. #Wrangle data
146. test.ci<-X%>%select("distr", "Pop_eta2", "n", "S1", "S2", ends_with("Power"))%>%
147. ungroup%>%
148. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2"), names_to = "method")%>%
149. mutate(n=factor(n, levels = c("20", "52", "500"), labels=c("n = 20", "n = 50", "n =
500"))),
150.         method=factor(method, levels =
c('MESP.035_TI/Power', 'MESP.005_TI/Power', 'SGPV.035=0_TI/Power', 'SGPV.005=0_TI/Power',
'p<=.005_TI/Power', 'p<=.05_TI/Power')),
151. labels=c('MESP2', 'MESP1', 'SGPV2', 'SGPV1', 'p<=.005', 'p<=.05'))
152.
153. #Combine with ic data
154. Factor_rnorm_ci.fig<-
cbind(test.ci, X%>%select("distr", "Pop_eta2", "n", "S1", "S2", ends_with("ic_H"))%>%
155. ungroup%>%
156. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2"), names_to = "method")%>%
157.         rename(ic_h=value)%>%select(-
c(distr, Pop_eta2, n, S1, S2, method)))%>%
158. cbind(X%>%select("distr", "Pop_eta2", "n", "S1", "S2", ends_with("ic_L"))%>%
159. ungroup%>%
160. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2"), names_to =
"method")%>%
161.         rename(ic_l=value)%>%select(-c(distr, Pop_eta2, n, S1, S2, method)))
162.
163.
164. test<-bind_rows(T_rnorm_ci.fig%>%mutate(Type="Two Group"),
165.                 ANOVA_rnorm_ci.fig%>%mutate(Type="K = 3 Group"),
166.                 Factor_rnorm_ci.fig%>%mutate(Type="2x2 Group"))%>%
167. mutate(Type=factor(Type, levels = c("Two Group", "K = 3 Group", "2x2 Group")))
168.
169. test<-test%>%mutate(ic=ic_h+ic_l)
170. test%>%summarise(range=range(ic), mean=mean(ic))
171. test%>%filter(ic >0.02)%>%view
172. T1e_norm.S15_ci%>%mutate(ic=ic_h+ic_l)%>%summarise(range=range(ic), mean=mean(ic))
173.
174. #####
175. #S15 Normal T1E #
176. #####
177.
178. #Combine all rnorm data
179.
180. T1e_norm.S15_ci<-bind_rows(T_rnorm_ci.fig%>%filter(S2==15 &
D==0)%>%mutate(Type="Two Group"),
181.                             ANOVA_rnorm_ci.fig%>%filter(S3==15 & Pop_eta2
==0)%>%mutate(Type="K = 3 Group"),
182.                             Factor_rnorm_ci.fig%>%filter(S2==15 & Pop_eta2
==0)%>%mutate(Type="2x2 Group"))%>%

```

```

183.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
184.
185.
186.
187. #Create plot
188. norm.S15.CI<-ggplot(T1e_norm.S15_ci,aes(fill= `n`, y=method, x=value))+
189.   geom_bar(position="dodge", stat="identity")+
190.   xlim(0,.09)+
191.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
192.                 colour="grey30",
193.                 size=.3,width=.5,
194.                 position=position_dodge(.9))+
195.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "fixed")+
196.   scale_fill_tableau(palette="Miller Stone")+
197.   xlab(NULL)+ylab(NULL)+
198.   theme_apa()+
199.   theme(axis.text = element_text(size =9))+
200.   theme(plot.title = element_text(hjust=.5,size=12))
201. #reposition legend
202. norm.S15.CI<-reposition_legend(norm.S15.CI, 'center', panel='panel-2-2')
203. T1e_norm.S60_ci%>%view
204. #SAVE PLOT
205. ggsave(plot=norm.S15.CI,width=10,height = 7,"/Users/MyMac/OneDrive - University of
  South Carolina/R code/Dissertation simulation/Plots/Type1_Error/CI
  plots/norm.S15.CI.jpeg")
206.
207.
208. #####
209. #S30 and S60 Normal T1E #
210. #####
211.
212. T1e_norm.S30_ci<-bind_rows(T_rnorm_ci.fig%>%filter(S2==30 &
  D==0)%>%mutate(Type="Two Group"),
213.                            ANOVA_rnorm_ci.fig%>%filter(S3==30 & Pop_eta2
  ==0)%>%mutate(Type="K = 3 Group"),
214.                            Factor_rnorm_ci.fig%>%filter(S2==30 & Pop_eta2
  ==0)%>%mutate(Type="2x2 Group"))%>%
215.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
216.
217.
218. T1e_norm.S60_ci<-bind_rows(T_rnorm_ci.fig%>%filter(S2==60 &
  D==0)%>%mutate(Type="Two Group"),
219.                            ANOVA_rnorm_ci.fig%>%filter(S3==60 & Pop_eta2
  ==0)%>%mutate(Type="K = 3 Group"),
220.                            Factor_rnorm_ci.fig%>%filter(S2==60 & Pop_eta2
  ==0)%>%mutate(Type="2x2 Group"))%>%
221.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
222.
223.
224. T1e_norm.S30_ci%>%mutate(var.ratio="2:1")
225.
226. T1e_norm.S60_ci%>%mutate(var.ratio="4:1")
227.
228. T1e_norm.S30.S60<-
  bind_rows(T1e_norm.S30_ci%>%mutate(var.ratio="2:1"),T1e_norm.S60_ci%>%mutate(var.ratio="4:1"))%>%
229.   mutate(var.ratio=factor(var.ratio,levels = c("2:1","4:1"),labels = c("Var. Ratio
  2:1","Var. Ratio 4:1")))
230.
231.
232. norm.S30.S60.CI<-ggplot(T1e_norm.S30.S60,aes(fill= `n`, y=method, x=value))+
233.   geom_bar(position="dodge", stat="identity")+
234.   xlim(0,.09)+
235.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
236.                 colour="grey30",

```

```

237.             size=.3,width=.5,
238.             position=position_dodge(.9))+
239.   facet_wrap(~ Type+var.ratio,nrow=3,ncol=2,scales = "fixed")+
240.   scale_fill_tableau(palette="Miller Stone")+
241.   xlab(NULL)+ylab(NULL)+
242.   theme_apa()+
243.   theme(axis.text = element_text(size = 9))+
244.   theme(plot.title = element_text(hjust=.5,size=12))
245.
246.   ggsave(plot=testplot,
247.         width=10,height=7,
248.         "/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Type1_Error/CI plots/norm.S30.60.CI.jpeg")
249.
250.
251. #####
252. #####POWER#####
253. #####
254.
255.   pwr_norm_ci<-bind_rows(T_rnorm_ci.fig%>%filter(D!=0)%>%mutate(Type="Two Group"),
256.                         ANOVA_rnorm_ci.fig%>%filter(Pop_eta2 !=0)%>%mutate(Type="K
= 3 Group"),
257.                         Factor_rnorm_ci.fig%>%filter(Pop_eta2
!=0)%>%mutate(Type="2x2 Group"))%>%
258.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")),
259.   D=factor(D,levels = c("0.2","0.5","0.8"),
260.   labels = c("d = 0.2","d = 0.5","d = 0.8")),
261.   S2 = factor(S2,levels = c("15","30","60"),
262.   labels = c("Var. Ratio 1:1","Var. Ratio 2:1","Var. Ratio 4:1")),
263.   Pop_eta2=factor(Pop_eta2,levels = c("0.01","0.06","0.14"),
264.   labels = c("eta = 0.01","eta = 0.06","eta = 0.14")),
265.   S3 = factor(S3,levels = c("15","30","60"),
266.   labels = c("Var. Ratio 1:1","Var. Ratio 2:1","Var. Ratio 4:1")))
267.
268.
269.   #TWO Group
270.   G2.normal_Power_CI<-ggplot(pwr_norm_ci%>%filter(Type=="Two Group"),aes(fill= `n`,
y=method, x=value))+
271.   geom_bar(position="dodge", stat="identity")+
272.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
273.   colour="grey30",
274.   size=.3,width=.5,
275.   position=position_dodge(.9))+
276.   facet_wrap(~ D+S2,nrow=4,ncol=3,scales = "fixed")+
277.   scale_fill_tableau(palette="Miller Stone")+
278.   xlab(NULL)+ylab(NULL)+
279.   theme_apa()+
280.   coord_flip()+
281.   theme(axis.text = element_text(size = 9))+
282.   theme(plot.title = element_text(hjust=.5,size=12))
283.
284.   ggsave(plot=G2.normal_Power_CI,
285.         width=12,height=9, units="in"
286.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/G2.normal_Power_CI.jpeg")
287.
288.
289.
290.
291.   #K=3 ANOVA
292.   K3.normal_Power_CI<-ggplot(pwr_norm_ci%>%filter(Type=="K = 3 Group"),aes(fill=
`n`, y=method, x=value))+
293.   geom_bar(position="dodge", stat="identity")+
294.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
295.   colour="grey30",

```

```

296.             size=.3,width=.5,
297.             position=position_dodge(.9))+
298.   facet_wrap(~ Pop_eta2+S3,nrow=4,ncol=3,scales = "fixed")+
299.   scale_fill_tableau(palette="Miller Stone")+
300.   xlab(NULL)+ylab(NULL)+
301.   theme_apa()+
302.   coord_flip()+
303.   theme(axis.text = element_text(size = 9))+
304.   theme(plot.title = element_text(hjust=.5,size=12))
305.
306.   ggsave(plot=K3.normal_Power_CI,
307.         width=12,height=9, units="in"
308.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/K3.normal_Power_CI.jpeg")
309.
310.
311.
312.   #Factorial normal ANOVA
313.   Fct.normal_Power_CI<-ggplot(pwr_norm_ci%>%filter(Type=="2x2 Group"),aes(fill= `n`,
y=method, x=value))+
314.   geom_bar(position="dodge", stat="identity")+
315.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
316.               colour="grey30",
317.               size=.3,width=.5,
318.               position=position_dodge(.9))+
319.   facet_wrap(~ Pop_eta2+S2,nrow=4,ncol=3,scales = "fixed")+
320.   scale_fill_tableau(palette="Miller Stone")+
321.   xlab(NULL)+ylab(NULL)+
322.   theme_apa()+
323.   coord_flip()+
324.   theme(axis.text = element_text(size = 9))+
325.   theme(plot.title = element_text(hjust=.5,size=12))
326.
327.   ggsave(plot=Fct.normal_Power_CI,
328.         width=12,height=9, units="in"
329.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/Fct.normal_Power_CI.jpeg")
330.
331.   #####
332.   #T_Skew
333.   #####
334.   #T_names
335.   T_nms<-c("p<=.05_TI/Power",
336.            "p<=.005_TI/Power",
337.            "SGPV.1=0_TI/Power",
338.            "SGPV.35=0_TI/Power",
339.            "MESP.1_TI/Power",
340.            "MESP.35_TI/Power")
341.
342.   #Prep data and calculate IC
343.   X<-T_rsnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
344.   mutate(`p<=.05.ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`,.after
= `p<=.05_TI/Power`)%>%
345.   mutate(`p<=.005.ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`,.after =
`p<=.005_TI/Power`)%>%
346.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.1=0_TI/Power`,.after =
`SGPV.1=0_TI/Power`)%>%
347.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.35=0_TI/Power`,.after =
`SGPV.35=0_TI/Power`)%>%
348.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.1_TI/Power`,.after
= `MESP.1_TI/Power`)%>%
349.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.35_TI/Power`,.after
= `MESP.35_TI/Power`)%>%
350.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`,.after
= `p<=.05_TI/Power`)%>%

```

```

351.   mutate(`p<=.005_ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`,.after =
`p<=.005_TI/Power`)%>%
352.   mutate(`SGPV.1=0_ic_L` = `SGPV.1=0_TI/Power` - `SGPV.1=0_CI_L`,.after =
`SGPV.1=0_TI/Power`)%>%
353.   mutate(`SGPV.35=0_ic_L` = `SGPV.35=0_TI/Power` - `SGPV.35=0_CI_L`,.after =
`SGPV.35=0_TI/Power`)%>%
354.   mutate(`MESP.1_ic_L` = `MESP.1_TI/Power` - `MESP.1_CI_L`,.after
= `MESP.1_TI/Power`)%>%
355.   mutate(`MESP.35_ic_L` = `MESP.35_TI/Power` - `MESP.35_CI_L`,.after
= `MESP.35_TI/Power`)%>%
356.   select(distr,D,n,S1,S2,
357.          "p<=.05_TI/Power",    `p<=.05_ic_H`,    `p<=.05_ic_L`,
358.          "p<=.005_TI/Power",   `p<=.005_ic_H`,   `p<=.005_ic_L`,
359.          "SGPV.1=0_TI/Power",   `SGPV.1=0_ic_H`,   `SGPV.1=0_ic_L`,
360.          "SGPV.35=0_TI/Power",  `SGPV.35=0_ic_H`, `SGPV.35=0_ic_L`,
361.          "MESP.1_TI/Power",     `MESP.1_ic_H`,    `MESP.1_ic_L`,
362.          "MESP.35_TI/Power",    `MESP.35_ic_H`,   `MESP.35_ic_L`)
363.
364. #Wrangle data
365. test.ci<-X%>%select("distr","D","n","S1","S2",ends_with("Power"))%>%
366.   ungroup%>%
367.   pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%>%
368.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500"))),
369.   method=factor(method,levels =
c('MESP.35_TI/Power','MESP.1_TI/Power','SGPV.35=0_TI/Power','SGPV.1=0_TI/Power','p<=
.005_TI/Power','p<=.05_TI/Power'),
370.   labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05'))
371.
372. #Combine with ic data
373. T_rsnorm_ci.fig<-
cbind(test.ci,X%>%select("distr","D","n","S1","S2",ends_with("ic_H"))%>%
374.   ungroup%>%
375.   pivot_longer(!c("distr","D","n","S1","S2"),names_to =
"method")%>%
376.   rename(ic_h=value)%>%select(-
c(distr,D,n,S1,S2,method))%>%
377.   cbind(X%>%select("distr","D","n","S1","S2",ends_with("ic_L"))%>%
378.   ungroup%>%
379.   pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%>%
380.   rename(ic_l=value)%>%select(-c(distr,D,n,S1,S2,method))
381.
382.
383. #####
384. #K3 Skew
385. #####
386. ANOVA_rsnorm_TI.PWR.Results_CI%>%names
387.
388. #T_names
389. AN_nms<-c("p<=.05_TI/Power",
390.           "p<=.005_TI/Power",
391.           "SGPV.005=0_TI/Power",
392.           "SGPV.035=0_TI/Power",
393.           "MESP.005_TI/Power",
394.           "MESP.035_TI/Power")
395.
396. #Prep data and calculate IC
397. X<-ANOVA_rsnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
398.   mutate(`p<=.05_ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`, .after =
`p<=.05_TI/Power`)%>%
399.   mutate(`p<=.005_ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`, .after =
`p<=.005_TI/Power`)%>%
400.   mutate(`SGPV.1=0_ic_H` = `SGPV.1=0_CI_H` - `SGPV.005=0_TI/Power`, .after =
`SGPV.005=0_TI/Power`)%>%

```



```

401. mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.035=0_TI/Power`, .after =
`SGPV.035=0_TI/Power`)>%
402. mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.005_TI/Power`, .after =
`MESP.005_TI/Power`)>%
403. mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.035_TI/Power`, .after =
`MESP.035_TI/Power`)>%
404. mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`, .after =
`p<=.05_TI/Power`)>%
405. mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`, .after =
`p<=.005_TI/Power`)>%
406. mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
`SGPV.005=0_TI/Power`)>%
407. mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
`SGPV.035=0_TI/Power`)>%
408. mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
`MESP.005_TI/Power`)>%
409. mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
`MESP.035_TI/Power`)>%
410. select(distr, Pop_eta2, n, S1, S2, S3,
411. "p<=.05_TI/Power", `p<=.05.ic_H`, `p<=.05.ic_L`,
412. "p<=.005_TI/Power", `p<=.005.ic_H`, `p<=.005.ic_L`,
413. "SGPV.005=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
414. "SGPV.035=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
415. "MESP.005_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,
416. "MESP.035_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
417.
418. #Wrangle data
419. test.ci<-X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("Power"))>%
420. ungroup%>%
421. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to = "method")>%
422. mutate(n=factor(n, levels = c("20", "50", "500"), labels=c("n = 20", "n = 50", "n =
500"))),
423. method=factor(method, levels =
c('MESP.035_TI/Power', 'MESP.005_TI/Power', 'SGPV.035=0_TI/Power', 'SGPV.005=0_TI/Power',
'p<=.005_TI/Power', 'p<=.05_TI/Power')),
424. labels=c('MESP2', 'MESP1', 'SGPV2', 'SGPV1', 'p<=.005', 'p<=.05'))
425.
426. #Combine with ic data
427. ANOVA_rsnorm_ci.fig<-
cbind(test.ci, X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("ic_H"))>%
428. ungroup%>%
429. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to = "method")>%
430. rename(ic_h=value)%>%select(-
c(distr, Pop_eta2, n, S1, S2, S3, method))>%
431. cbind(X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("ic_L"))>%
432. ungroup%>%
433. pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to =
"method")>%
434. rename(ic_l=value)%>%select(-c(distr, Pop_eta2, n, S1, S2, S3, method))
435.
436. #####
437. #Factor Skew
438. #####
439.
440. Factor_rsnorm_TI.PWR.Results_CI%>%names
441.
442. #T_names
443. AN_nms<-c("p<=.05_TI/Power",
444. "p<=.005_TI/Power",
445. "SGPV.005=0_TI/Power",
446. "SGPV.035=0_TI/Power",
447. "MESP.005_TI/Power",
448. "MESP.035_TI/Power")

```



```

449.
450. #Prep data and calculate IC
451. X<-Factor_rsnorm_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
452.   mutate(`p<=.05.ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`, .after =
`p<=.05_TI/Power`)%>%
453.   mutate(`p<=.005.ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`, .after =
`p<=.005_TI/Power`)%>%
454.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.005=0_TI/Power`, .after =
`SGPV.005=0_TI/Power`)%>%
455.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.035=0_TI/Power`, .after =
`SGPV.035=0_TI/Power`)%>%
456.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.005_TI/Power`, .after =
`MESP.005_TI/Power`)%>%
457.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.035_TI/Power`, .after =
`MESP.035_TI/Power`)%>%
458.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`, .after =
`p<=.05_TI/Power`)%>%
459.   mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`, .after =
`p<=.005_TI/Power`)%>%
460.   mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
`SGPV.005=0_TI/Power`)%>%
461.   mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
`SGPV.035=0_TI/Power`)%>%
462.   mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
`MESP.005_TI/Power`)%>%
463.   mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
`MESP.035_TI/Power`)%>%
464.   select(distr,Pop_eta2,n,S1,S2,
465.     "p<=.05.ic_H", `p<=.05.ic_H`, `p<=.05.ic_L`,
466.     "p<=.005.ic_H", `p<=.005.ic_H`, `p<=.005.ic_L`,
467.     "SGPV.005=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
468.     "SGPV.035=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
469.     "MESP.005_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,
470.     "MESP.035_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
471.
472. #Wrangle data
473. test.ci<-X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("Power"))%>%
474.   ungroup%>%
475.   pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to = "method")%>%
476.   mutate(n=factor(n,levels = c("20","52","500"), labels=c("n = 20","n = 50","n =
500"))),
477.   method=factor(method,levels =
c('MESP.035_TI/Power','MESP.005_TI/Power','SGPV.035=0_TI/Power','SGPV.005=0_TI/Power
','p<=.005_TI/Power','p<=.05_TI/Power'),
478.   labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05')))
479.
480. #Combine with ic data
481. Factor_rsnorm_ci.fig<-
cbind(test.ci,X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("ic_H"))%>%
482.   ungroup%>%
483.   pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to = "method")%>%
484.   rename(ic_h=value)%>%select(-
c(distr,Pop_eta2,n,S1,S2,method))%>%
485.   cbind(X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("ic_L"))%>%
486.   ungroup%>%
487.   pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to =
"method")%>%
488.   rename(ic_l=value)%>%select(-c(distr,Pop_eta2,n,S1,S2,method)))
489.
490.
491.
492.
493. #####

```

```

494. #S15 Skew T1E #
495. #####
496.
497. #Combine all rsnorm data
498.
499. T1e_skew.S15_ci<-bind_rows(T_rsnorm_ci.fig%>%filter(S2==15 &
D==0)%>%mutate(Type="Two Group"),
500.                        ANOVA_rsnorm_ci.fig%>%filter(S3==15 & Pop_eta2
==0)%>%mutate(Type="K = 3 Group"),
501.                        Factor_rsnorm_ci.fig%>%filter(S2==15 & Pop_eta2
==0)%>%mutate(Type="2x2 Group"))%>%
502.      mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
503.
504.
505.
506. #Create plot
507. skew.S15.CI<-ggplot(T1e_skew.S15_ci,aes(fill= `n`, y=method, x=value))+
508.   geom_bar(position="dodge", stat="identity")+
509.   xlim(0,.09)+
510.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
511.                 colour="grey30",
512.                 size=.3,width=.5,
513.                 position=position_dodge(.9))+
514.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "fixed")+
515.   scale_fill_tableau(palette="Miller Stone")+
516.   xlab(NULL)+ylab(NULL)+
517.   theme_apa()+
518.   theme(axis.text = element_text(size =9))+
519.   theme(plot.title = element_text(hjust=.5,size=12))
520. #reposition legend
521. skew.S15.CI<-reposition_legend(skew.S15.CI,'center', panel='panel-2-2')
522.
523. #SAVE PLOT
524. ggsave(plot=skew.S15.CI,"/Users/MyMac/OneDrive - University of South Carolina/R
code/Dissertation simulation/Plots/Type1_Error/CI plots/skew.S15.CI.jpeg")
525.
526.
527.
528. #####
529. #####
530. #####
531. #T_Bimodal
532. #####
533. #T_names
534. T_nms<-c("p<=.05_TI/Power",
535.           "p<=.005_TI/Power",
536.           "SGPV.1=0_TI/Power",
537.           "SGPV.35=0_TI/Power",
538.           "MESP.1_TI/Power",
539.           "MESP.35_TI/Power")
540.
541. #Prep data and calculate IC
542. X<-T_bimodal_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
543.   mutate(`p<=.05_ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`,.after
= `p<=.05_TI/Power`)%>%
544.   mutate(`p<=.005_ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`,.after =
`p<=.005_TI/Power`)%>%
545.   mutate(`SGPV.1=0_ic_H` = `SGPV.1=0_CI_H` - `SGPV.1=0_TI/Power`,.after =
`SGPV.1=0_TI/Power`)%>%
546.   mutate(`SGPV.35=0_ic_H` = `SGPV.35=0_CI_H` - `SGPV.35=0_TI/Power`,.after =
`SGPV.35=0_TI/Power`)%>%
547.   mutate(`MESP.1_ic_H` = `MESP.1_CI_H` - `MESP.1_TI/Power`,.after
= `MESP.1_TI/Power`)%>%
548.   mutate(`MESP.35_ic_H` = `MESP.35_CI_H` - `MESP.35_TI/Power`,.after
= `MESP.35_TI/Power`)%>%

```

```

549.   mutate(`p<=.05.ic_L`      = `p<=.05_TI/Power` - `p<=.05_CI_L`,.after
= `p<=.05_TI/Power`)%%>%
550.   mutate(`p<=.005.ic_L`     = `p<=.005_TI/Power` - `p<=.005_CI_L`,.after =
`p<=.005_TI/Power`)%%>%
551.   mutate(`SGPV.1=0.ic_L`    = `SGPV.1=0_TI/Power` - `SGPV.1=0_CI_L`,.after =
`SGPV.1=0_TI/Power`)%%>%
552.   mutate(`SGPV.35=0.ic_L`   = `SGPV.35=0_TI/Power` - `SGPV.35=0_CI_L`,.after =
`SGPV.35=0_TI/Power`)%%>%
553.   mutate(`MESP.1.ic_L`     = `MESP.1_TI/Power` - `MESP.1_CI_L`,.after
= `MESP.1_TI/Power`)%%>%
554.   mutate(`MESP.35.ic_L`    = `MESP.35_TI/Power` - `MESP.35_CI_L`,.after
= `MESP.35_TI/Power`)%%>%
555.   select(distr,D,n,S1,S2,
556.          "p<=.05_TI/Power",    `p<=.05.ic_H`,    `p<=.05.ic_L`,
557.          "p<=.005_TI/Power",   `p<=.005.ic_H`, `p<=.005.ic_L`,
558.          "SGPV.1=0_TI/Power",  `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
559.          "SGPV.35=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
560.          "MESP.1_TI/Power",    `MESP.1.ic_H`,  `MESP.1.ic_L`,
561.          "MESP.35_TI/Power",   `MESP.35.ic_H`, `MESP.35.ic_L`)
562.
563. #Wrangle data
564. test.ci<-X%%>%select("distr","D","n","S1","S2",ends_with("Power"))%%>%
565.   ungroup%%>%
566.   pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%%>%
567.   mutate(n=factor(n,levels = c("20","50","500"), labels=c("n = 20","n = 50","n =
500")),
568.          method=factor(method,levels =
c('MESP.35_TI/Power','MESP.1_TI/Power','SGPV.35=0_TI/Power','SGPV.1=0_TI/Power','p<=
.005_TI/Power','p<=.05_TI/Power'),
569.          labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05')))
570.
571. #Combine with ic data
572. T_bimodal_ci.fig<-
cbind(test.ci,X%%>%select("distr","D","n","S1","S2",ends_with("ic_H"))%%>%
573.   ungroup%%>%
574.   pivot_longer(!c("distr","D","n","S1","S2"),names_to =
"method")%%>%
575.   rename(ic_h=value)%%>%select(-
c(distr,D,n,S1,S2,method))%%>%
576.   cbind(X%%>%select("distr","D","n","S1","S2",ends_with("ic_L"))%%>%
577.   ungroup%%>%
578.   pivot_longer(!c("distr","D","n","S1","S2"),names_to = "method")%%>%
579.   rename(ic_l=value)%%>%select(-c(distr,D,n,S1,S2,method)))
580.
581.
582. #####
583. #K3 Bimodal
584. #####
585. ANOVA_bimodal_TI.PWR.Results_CI%%>%names
586.
587. #T_names
588. AN_nms<-c("p<=.05_TI/Power",
589.           "p<=.005_TI/Power",
590.           "SGPV.005=0_TI/Power",
591.           "SGPV.035=0_TI/Power",
592.           "MESP.005_TI/Power",
593.           "MESP.035_TI/Power")
594.
595. #Prep data and calculate IC
596. X<-ANOVA_bimodal_TI.PWR.Results_CI%%>%mutate(across(.cols=everything(),round,4))%%>%
597.   mutate(`p<=.05.ic_H`      = `p<=.05_CI_H`      - `p<=.05_TI/Power`,      .after =
`p<=.05_TI/Power`)%%>%
598.   mutate(`p<=.005.ic_H`    = `p<=.005_CI_H`    - `p<=.005_TI/Power`,      .after =
`p<=.005_TI/Power`)%%>%

```

```

599.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.005=0_TI/Power`, .after =
`SGPV.005=0_TI/Power`)%>%
600.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.035=0_TI/Power`, .after =
`SGPV.035=0_TI/Power`)%>%
601.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.005_TI/Power`, .after =
`MESP.005_TI/Power`)%>%
602.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.035_TI/Power`, .after =
`MESP.035_TI/Power`)%>%
603.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`, .after =
`p<=.05_TI/Power`)%>%
604.   mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`, .after =
`p<=.005_TI/Power`)%>%
605.   mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
`SGPV.005=0_TI/Power`)%>%
606.   mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
`SGPV.035=0_TI/Power`)%>%
607.   mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
`MESP.005_TI/Power`)%>%
608.   mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
`MESP.035_TI/Power`)%>%
609.   select(distr, Pop_eta2, n, S1, S2, S3,
610.          "p<=.05_TI/Power", `p<=.05.ic_H`, `p<=.05.ic_L`,
611.          "p<=.005_TI/Power", `p<=.005.ic_H`, `p<=.005.ic_L`,
612.          "SGPV.005=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
613.          "SGPV.035=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
614.          "MESP.005_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,
615.          "MESP.035_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
616.
617. #Wrangle data
618. test.ci<-X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("Power"))%>%
619.   ungroup%>%
620.   pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to = "method")%>%
621.   mutate(n=factor(n, levels = c("20", "50", "500"), labels=c("n = 20", "n = 50", "n =
500"))),
622.   method=factor(method, levels =
c('MESP.035_TI/Power', 'MESP.005_TI/Power', 'SGPV.035=0_TI/Power', 'SGPV.005=0_TI/Power',
'p<=.005_TI/Power', 'p<=.05_TI/Power')),
623.   labels=c('MESP2', 'MESP1', 'SGPV2', 'SGPV1', 'p<=.005', 'p<=.05'))
624.
625. #Combine with ic data
626. ANOVA_bimodal_ci.fig<-
cbind(test.ci, X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("ic_H"))%>%
627.   ungroup%>%
628.   pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to = "method")%>%
629.   rename(ic_h=value)%>%select(-
c(distr, Pop_eta2, n, S1, S2, S3, method))%>%
630.   cbind(X%>%select("distr", "Pop_eta2", "n", "S1", "S2", "S3", ends_with("ic_L"))%>%
631.     ungroup%>%
632.     pivot_longer(!c("distr", "Pop_eta2", "n", "S1", "S2", "S3"), names_to =
"method")%>%
633.     rename(ic_l=value)%>%select(-c(distr, Pop_eta2, n, S1, S2, S3, method)))
634.
635. #####
636. #Factor Bimodal
637. #####
638.
639. Factor_bimodal_TI.PWR.Results_CI%>%names
640.
641. #T_names
642. AN_nms<-c("p<=.05_TI/Power",
643.           "p<=.005_TI/Power",
644.           "SGPV.005=0_TI/Power",
645.           "SGPV.035=0_TI/Power",

```

```

646.         "MESP.005_TI/Power",
647.         "MESP.035_TI/Power")
648.
649. #Prep data and calculate IC
650. X<-
  Factor_bimodal_TI.PWR.Results_CI%>%mutate(across(.cols=everything(),round,4))%>%
651.   mutate(`p<=.05.ic_H` = `p<=.05_CI_H` - `p<=.05_TI/Power`, .after =
     `p<=.05_TI/Power`)%>%
652.   mutate(`p<=.005.ic_H` = `p<=.005_CI_H` - `p<=.005_TI/Power`, .after =
     `p<=.005_TI/Power`)%>%
653.   mutate(`SGPV.1=0.ic_H` = `SGPV.1=0_CI_H` - `SGPV.005=0_TI/Power`, .after =
     `SGPV.005=0_TI/Power`)%>%
654.   mutate(`SGPV.35=0.ic_H` = `SGPV.35=0_CI_H` - `SGPV.035=0_TI/Power`, .after =
     `SGPV.035=0_TI/Power`)%>%
655.   mutate(`MESP.1.ic_H` = `MESP.1_CI_H` - `MESP.005_TI/Power`, .after =
     `MESP.005_TI/Power`)%>%
656.   mutate(`MESP.35.ic_H` = `MESP.35_CI_H` - `MESP.035_TI/Power`, .after =
     `MESP.035_TI/Power`)%>%
657.   mutate(`p<=.05.ic_L` = `p<=.05_TI/Power` - `p<=.05_CI_L`, .after =
     `p<=.05_TI/Power`)%>%
658.   mutate(`p<=.005.ic_L` = `p<=.005_TI/Power` - `p<=.005_CI_L`, .after =
     `p<=.005_TI/Power`)%>%
659.   mutate(`SGPV.1=0.ic_L` = `SGPV.005=0_TI/Power` - `SGPV.1=0_CI_L`, .after =
     `SGPV.005=0_TI/Power`)%>%
660.   mutate(`SGPV.35=0.ic_L` = `SGPV.035=0_TI/Power` - `SGPV.35=0_CI_L`, .after =
     `SGPV.035=0_TI/Power`)%>%
661.   mutate(`MESP.1.ic_L` = `MESP.005_TI/Power` - `MESP.1_CI_L`, .after =
     `MESP.005_TI/Power`)%>%
662.   mutate(`MESP.35.ic_L` = `MESP.035_TI/Power` - `MESP.35_CI_L`, .after =
     `MESP.035_TI/Power`)%>%
663.   select(distr,Pop_eta2,n,S1,S2,
664.          "p<=.05_TI/Power", `p<=.05.ic_H`, `p<=.05.ic_L`,
665.          "p<=.005_TI/Power", `p<=.005.ic_H`, `p<=.005.ic_L`,
666.          "SGPV.005=0_TI/Power", `SGPV.1=0.ic_H`, `SGPV.1=0.ic_L`,
667.          "SGPV.035=0_TI/Power", `SGPV.35=0.ic_H`, `SGPV.35=0.ic_L`,
668.          "MESP.005_TI/Power", `MESP.1.ic_H`, `MESP.1.ic_L`,
669.          "MESP.035_TI/Power", `MESP.35.ic_H`, `MESP.35.ic_L`)
670.
671. test.ci<-X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("Power"))%>%
672.   ungroup%>%
673.   pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to = "method")%>%
674.   mutate(n=factor(n,levels = c("20","52","500"), labels=c("n = 20","n = 50","n =
675.     500"))),
     method=factor(method,levels =
676.       c('MESP.035_TI/Power','MESP.005_TI/Power','SGPV.035=0_TI/Power','SGPV.005=0_TI/Power',
677.         'p<=.005_TI/Power','p<=.05_TI/Power'),
     labels=c('MESP2','MESP1','SGPV2','SGPV1','p<=.005','p<=.05'))
678. #Combine with ic data
679. Factor_bimodal_ci.fig<-
  cbind(test.ci,X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("ic_H"))%>%
680.     ungroup%>%
681.     pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to = "method")%>%
682.     rename(ic_h=value)%>%select(-
683.       c(distr,Pop_eta2,n,S1,S2,method))%>%
684.     cbind(X%>%select("distr","Pop_eta2","n","S1","S2",ends_with("ic_L"))%>%
685.       ungroup%>%
686.       pivot_longer(!c("distr","Pop_eta2","n","S1","S2"),names_to =
687.         "method")%>%
688.       rename(ic_l=value)%>%select(-c(distr,Pop_eta2,n,S1,S2,method)))
689.

```

```

690.
691. T_rsnorm_ci.fig
692. T_bimodal_ci.fig
693. ANOVA_rsnorm_ci.fig
694. ANOVA_bimodal_ci.fig
695. Factor_rsnorm_ci.fig
696. Factor_bimodal_ci.fig
697.
698. rsnorm.bimodal_ci.fig<-bind_rows(T_rsnorm_ci.fig%%mutate(Type="Two
Group"),T_bimodal_ci.fig%%mutate(Type="Two Group"),
699.
ANOVA_rsnorm_ci.fig%%mutate(Type="K=3"),ANOVA_bimodal_ci.fig%%mutate(Type="K=3"),
700.
Factor_rsnorm_ci.fig%%mutate(Type="2x2"),Factor_bimodal_ci.fig%%mutate(Type="2x2")
)%%%
701. mutate(distr=factor(distr,levels=c("distr=rsnorm","distr=bimodal"),
702. labels = c("Skewed","Bimodal")))%%%
703. mutate(Type=factor(Type,levels=c("Two Group","K=3","2x2"),
704. labels = c("Two Group","K=3 Group","2x2 Group")))
705.
706. rsnorm.bimodal_ci.fig%%names
707. rsnorm.bimodal<-ggplot(rsnorm.bimodal_ci.fig%%filter(D==0 |
Pop_eta2==0),aes(fill= `n`, y=method, x=value))+
708. geom_bar(position="dodge", stat="identity")+
709. xlim(0,.09)+
710. geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
711. colour="grey30",
712. size=.3,width=.5,
713. position=position_dodge(.9))+
714. facet_wrap(~ Type+distr,nrow=3,ncol=2,scales = "fixed")+
715. scale_fill_tableau(palette="Miller Stone")+
716. xlab(NULL)+ylab(NULL)+
717. theme_apa()+
718. theme(axis.text = element_text(size =9))+
719. theme(plot.title = element_text(hjust=.5,size=12))
720.
721. ggsave(plot=rsnorm.bimodal,width=10,height=7,"/Users/MyMac/OneDrive - University
of South Carolina/R code/Dissertation simulation/Plots/Type1_Error/CI
plots/rsnorm.bimodal.CI.jpeg")
722.
723.
724. #####
725. #S15 Bimodal T1E #
726. #####
727.
728. #Combine all rsnorm data
729.
730. T1e_bimodal.S15_ci<-bind_rows(T_bimodal_ci.fig%%filter(S2==15 &
D==0)%%%mutate(Type="Two Group"),
731. ANOVA_bimodal_ci.fig%%filter(S3==15 & Pop_eta2
==0)%%%mutate(Type="K = 3 Group"),
732. Factor_bimodal_ci.fig%%filter(S2==15 & Pop_eta2
==0)%%%mutate(Type="2x2 Group")))%%%
733. mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")),
734. D=factor(D,levels = c("0.2","0.5","0.8"),
735. labels = c("d = 0.2","d = 0.5","d = 0.8")),
736. Pop_eta2=factor(Pop_eta2,levels = c("0.01","0.06","0.14"),
737. labels = c("eta = 0.01","eta = 0.06","eta = 0.14")),
738. distr = factor(distr,levels = c("distr=rsnorm","distr=bimodal"),
739. labels = c("Skewed","Bimodal")))
740.
741.
742. #Create plot
743. bimodal.S15.CI<-ggplot(T1e_bimodal.S15_ci,aes(fill= `n`, y=method, x=value))+
744. geom_bar(position="dodge", stat="identity")+

```

```

745.   xlim(0,.09)+
746.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
747.                 colour="grey30",
748.                 size=.3,width=.5,
749.                 position=position_dodge(.9))+
750.   facet_wrap(~ Type,nrow=2,ncol=2,scales = "fixed")+
751.   scale_fill_tableau(palette="Miller Stone")+
752.   xlab(NULL)+ylab(NULL)+
753.   theme_apache()+
754.   theme(axis.text = element_text(size =9))+
755.   theme(plot.title = element_text(hjust=.5,size=12))
756.
757. #reposition legend
758. bimodal.S15.CI<-reposition_legend(bimodal.S15.CI,'center', panel='panel-2-2')
759.
760. #SAVE PLOT
761. ggsave(plot=bimodal.S15.CI,"/Users/MyMac/OneDrive - University of South Carolina/R
code/Dissertation simulation/Plots/Type1_Error/CI plots/bimodal.S15.CI.jpeg")
762.
763.
764.
765. #####
766. #####POWER#####
767. #####
768.
769. rsnorm_ci<-bind_rows(T_rsnorm_ci.fig%>%mutate(Type="Two Group"),
770.                     ANOVA_rsnorm_ci.fig%>%mutate(Type="K = 3 Group"),
771.                     Factor_rsnorm_ci.fig%>%mutate(Type="2x2 Group"))%>%
772.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
773.
774. fct_ci<-bind_rows(T_bimodal_ci.fig%>%mutate(Type="Two Group"),
775.                  ANOVA_bimodal_ci.fig%>%mutate(Type="K = 3 Group"),
776.                  Factor_bimodal_ci.fig%>%mutate(Type="2x2 Group"))%>%
777.   mutate(Type=factor(Type,levels = c("Two Group","K = 3 Group","2x2 Group")))
778.
779. skew.fct_ci<-bind_rows(rsnorm_ci,fct_ci) %>%
780.   mutate(D=factor(D,levels = c("0.2","0.5","0.8"),
781.                     labels = c("d = 0.2","d = 0.5","d = 0.8")),
782.          Pop_eta2=factor(Pop_eta2,levels = c("0.01","0.06","0.14"),
783.                           labels = c("eta = 0.01","eta = 0.06","eta = 0.14")),
784.          distr = factor(distr,levels = c("distr=rsnorm","distr=bimodal"),
785.                          labels = c("Skewed","Bimodal")))
786.
787.
788. #TWO GROUP
789.
790. skew.fct_ci%>%filter(Type=="Two Group")%>%view
791.
792. g2.skew.bimodal<-ggplot(skew.fct_ci%>%filter(Type=="Two Group" & D!=0),aes(fill=
`n`, y=method, x=value))+
793.   geom_bar(position="dodge", stat="identity",show.legend = FALSE)+
794.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
795.                 colour="grey30",
796.                 size=.3,width=.5,
797.                 position=position_dodge(.9))+
798.   facet_wrap(~ D+distr,nrow=3,ncol=2,scales = "fixed")+
799.   scale_fill_tableau(palette="Miller Stone")+
800.   xlab(NULL)+ylab(NULL)+
801.   theme_apache()+
802.   coord_flip()+
803.   theme(axis.text = element_text(size =9))+
804.   theme(plot.title = element_text(hjust=.5,size=12))+
805.   ggtitle("Two Group")
806.
807.

```



```

808. ggsave(plot=g2.skew.bimodal,
809.         width=10,height=7, units="in"
810.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/G2.skew.bimodal_Power_CI.jpeg")
811.
812.
813.
814.
815. #K=3 Group
816.
817. k3.skew.bimodal<-ggplot(skew.fct_ci%>%filter(Type=="K = 3 Group" & Pop_eta2
!=0),aes(fill= `n`, y=method, x=value))+
818.   geom_bar(position="dodge", stat="identity",show.legend = FALSE)+
819.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
820.                colour="grey30",
821.                size=.3,width=.5,
822.                position=position_dodge(.9))+
823.   facet_wrap(~ Pop_eta2+distr,nrow=3,ncol=2,scales = "fixed")+
824.   scale_fill_tableau(palette="Miller Stone")+
825.   xlab(NULL)+ylab(NULL)+
826.   theme_apa()+
827.   coord_flip()+
828.   theme(axis.text = element_text(size =9))+
829.   theme(plot.title = element_text(hjust=.5,size=12))+
830.   ggtitle("K = 3 Group")
831.
832.
833. ggsave(plot=k3.skew.bimodal,
834.         width=10,height=7, units="in"
835.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/K3.skew.bimodal_Power_CI.jpeg")
836.
837.
838. #2x2 Group
839.
840. fct.skew.bimodal<-ggplot(skew.fct_ci%>%filter(Type=="2x2 Group" & Pop_eta2
!=0),aes(fill= `n`, y=method, x=value))+
841.   geom_bar(position="dodge", stat="identity")+
842.   geom_errorbar(aes(xmin=value-ic_l, xmax=value+ic_h),
843.                colour="grey30",
844.                size=.3,width=.5,
845.                position=position_dodge(.9))+
846.   facet_wrap(~ Pop_eta2+distr,nrow=3,ncol=2,scales = "fixed")+
847.   scale_fill_tableau(palette="Miller Stone")+
848.   xlab(NULL)+ylab(NULL)+
849.   theme_apa()+
850.   coord_flip()+
851.   theme(axis.text = element_text(size =9))+
852.   theme(plot.title = element_text(hjust=.5,size=12))+
853.   ggtitle("2x2 Group")
854.
855.
856. ggsave(plot=fct.skew.bimodal,
857.         width=10,height=7, units="in"
858.         ,"/Users/MyMac/OneDrive - University of South Carolina/R code/Dissertation
simulation/Plots/Power/CI plots/Fct.skew.bimodal_Power_CI.jpeg")
859.
860.
861.
862. Factor_rnorm_TI.PWR.Results%>%names
863. Factor_rnorm_TI.PWR.Results%>%filter(Pop_eta2==0 & S2==15)%>%
864.   select(pwr.names)%>%
865.   view
866.
867.

```



```

868. Factor_rnorm_TI.PWR.Results%>%names
869.
870. pwr.names<-c("distr", "Pop_eta2", "n", "S1", "S2",
871.             "p<=.05_TI/Power", "p<=.005_TI/Power",
872.             "SGPV.005=0_TI/Power", "SGPV.035=0_TI/Power",
873.             "MESP.005_TI/Power", "MESP.035_TI/Power")
874.
875.
876.
877. #T1 ERROR
878. Factor_rnorm_TI.PWR.Results%>%names
879. Factor_rnorm_TI.PWR.Results%>%filter(Pop_eta2==0 & S2==15)%>%
880.   select(pwr.names)%>%
881.   view
882.
883.
884. Factor_rnorm_TI.PWR.Results%>%names
885. Factor_rnorm_TI.PWR.Results%>%filter(Pop_eta2==0 & S2!=15)%>%
886.   select(pwr.names)%>%
887.   view
888.
889.
890. Factor_rsnorm_TI.PWR.Results%>%filter(Pop_eta2==0 & S2==15)%>%
891.   select(pwr.names)%>%
892.   view
893.
894. Factor_bimodal_TI.PWR.Results%>%filter(Pop_eta2==0 & S2==15)%>%
895.   select(pwr.names)%>%
896.   view
897.
898.
899. #POWER
900.
901. Factor_rnorm_TI.PWR.Results%>%names
902. Factor_rnorm_TI.PWR.Results%>%filter(Pop_eta2!=0 & S2==15)%>%
903.   select(pwr.names)%>%
904.   view
905.
906.
907. Factor_rnorm_TI.PWR.Results%>%names
908. Factor_rnorm_TI.PWR.Results%>%filter(Pop_eta2==.14 & S2==60)%>%
909.   select(pwr.names)%>%
910.   view
911.
912.
913. Factor_rsnorm_TI.PWR.Results%>%filter(Pop_eta2==0.14 & S2==15)%>%
914.   select(pwr.names)%>%
915.   view
916.
917. Factor_bimodal_TI.PWR.Results%>%filter(Pop_eta2==0.14 & S2==15)%>%
918.   select(pwr.names)%>%
919.   view
920. T_rnorm_ci.fig%>%names
921. #####
922. T_rnorm_ci.range<-T_rnorm_ci.fig%>%mutate(ci.range=ic_h+ic_l)%>%
923.   mutate(Effect=factor(D,levels = c(0,0.2,0.5,.8),
924.                               labels =c("none","small","medium","large")))%>%
925.   mutate(model="two_group")%>%
926.   select(distr,model,Effect,n,S2,method,ci.range)
927.
928. ANOVA_rnorm.ci.range<-ANOVA_rnorm_ci.fig%>%mutate(ci.range=ic_h+ic_l)%>%
929.   mutate(Effect=factor(Pop_eta2,levels = c(0,0.01,0.06,0.14),
930.                               labels = c("none","small","medium","large")))%>%

```

```

931.   mutate(S2=S3)%>%mutate(model="k=3")%>%select(distr,model,Effect,n,S2,method,ci.range
)
932.
933.   Factor_rnorm.ci.range<-Factor_rnorm_ci.fig%>%mutate(ci.range=ic_h+ic_l)%>%
934.     mutate(Effect=factor(Pop_eta2,levels = c(0,0.01,0.06,0.14),
935.       labels = c("none","small","medium","large")))%>%
936.     mutate(model="2x2")%>%select(distr,model,Effect,n,S2,method,ci.range)
937.
938.   rnorm.ci.range<-
     bind_rows(T_rnorm.ci.range,ANOVA_rnorm.ci.range,Factor_rnorm.ci.range)%>%
939.     mutate(model=factor(model,levels = c("two_group","k=3","2x2"),
940.       labels = c("two_group","k=3","2x2")))
941.   rnorm.ci.range%>%view
942.   rnorm.ci.range%>%filter(Effect=="none")%>%group_by(model)%>%summarize(range=range(
     ci.range))
943.

```