Recognizing Reproducible Research

Keith E. Muller
Department of Health Outcomes and Biomedical Informatics
College of Medicine, University of Florida

Deborah H. Glueck
Department of Pediatrics
University of Colorado School of Medicine

UF | UNIVERSITY of FLORIDA

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgements of Current Support

Glueck, Muller, and Dabelea MPI of 9R01GM121081, *Methods and Software for Lifecourse Epidemiology Data and Sample Size Analysis* NIGMS, 08/16-06/20.

Muller and Glueck co-PI of 1R25GM111901, *A Master Course on Power for Multilevel and Longitudinal Health Behavior Studies*, OBSSR and NIGMS, 08/14-06/18.

Dr. Glueck and Muller are co-investigators on other NIH-funded projects.

Outline of Presentation

Motivating Problem: the Replication Crisis

Guideline 1. Explicitly Control Type I Errors (False Positives) and Type II Errors (False Negatives).

Guideline 2. Align the Goals, Design, Data Analysis, and Sample Size Analysis.

Guideline 3. Account for Uncertainty in Sample Size Computations.

Motivating Problem

The public, scientists, statisticians, and government officials share a growing concern about the reproducibility of health science.

What can we do about it?

Thesis: If we learn to recognize reproducible research, then we will conduct reproducible research and down weight irreproducible research.

What Is Science?

An *epistemology*:
a set of rules to decide what is true.

A way of knowing.

Science (Latin: *scio*, *sciere*, to know) is
defined by two features: *public* and *replicable*.

Scientific research requires specifying a model, a hypothesis, followed by systematic collection of information capable of falsifying the hypothesis.

All aspects must be *public* and *replicable*.

Helpful to Distinguish Reproducible From Replicable

The public dimension of science requires that any result be reproducible by others.

We describe a feature of a deterministic process, such as a confidence interval for a difference in means computed for a given set of data, as *reproducible* if another analyst can produce exactly the same value as originally reported.

We describe a feature of a probabilistic process based on observation, such as the time for cannon balls to fall to the ground, or the mass of the Higgs boson, as *replicable* if a similar experiment produces a similar value, allowing for reasonable statistical uncertainty.

We focus today primarily on replicability.

Commentators Often Mix Reproducible and Replicable

Guidelines for specifying a specific data analysis process or a data collection protocol address reproducibility.

Guidelines for controlling probabilities of decision errors address replicability.

We focus today primarily on replicability.

A Google search on "Reproducible Science New Yorker" finds articles in 2010, 2013, 2015, 2016, 2017, and 2018.

A Google search on "Reproducible Science New York Times" finds articles in 2011, 2014, 2015, 2016, 2017, and 2018.

Examples of Use of Term in Statistics

A Johns Hopkins Coursera "...course focuses on the concepts and tools behind reporting modern data analyses in a reproducible manner."

Biometric Society sessions, at least as early as 2011, had discussions by
David Banks (Duke, policy science),
Keith Baggerly (M.D. Anderson, forensic bioinformatics),
Frank Harrell (Vanderbilt, clinical trials),
of *reproducible analysis* and *replicable science* under the title of "reproducible research."

A. Lehman posted a short bibliography at
https://ccts.osu.edu/sites/default/files/inline-files/Articles on Irreproducible Research and P values.pdf

Scientists

A chronic concern of a subset of scientists in a subset of disciplines.

Behavioral scientists and their critics have been most vocal.

Historically many bench scientists claimed they did not need statistics.
A true statement when studying horseshoes, hand grenades and nuclear weapons; close is good enough.

High throughput assays (fMRI, genomics, metabolomics, etc.) have changed the tune.

Recently an eminent and gray-haired epigenetics friend was chagrined to admit he is reading a statistics book. Horrors!

Government

The Director of the National Institutes of Health (NIH) and his deputy (Collins and Tabak, 2014) outlined plans for NIH to rigorously address reproducibility in all projects.

The NIH changed reviewing and training requirements for applicants (*https://www.nih.gov/research-training/rigor-reproducibility*).

Methodological concerns included

poor study designs,
incorrect statistical analyses,
inappropriate sample size selection, and
misleading reporting.

Government (continued)

NIH concerns include reproducibility and replicability.

Recent elaborations of the NIH definition of a "clinical trial" now includes the majority of randomized studies in the behavioral sciences.

Regulatory standards of FDA and USEPA require reproducible protocols and analysis and enhance the chances of replicable research.

Features include scrupulous record keeping and archiving, the use of preregistration of studies at Clintrials.Gov, Manuals of Procedures, steering committees, Data and Safety Monitoring Boards.

They are not sufficiently widely used; some version needs to be universal.

Today we focus primarily on replicability.

SOLUTION TO PROBLEM:
In Planning and Evaluating:
Ensure Replicability by Following Statistical Guidelines

1. Explicitly control both Type I errors (false positives) and Type II errors (false negatives).
2. Align the goals, design, data analysis, and sample size analysis.
3. Account for uncertainty in inputs to sample size calculation.

Relevant Personal Bibliography

Muller, Christiansen, and Smith (1981) Guidelines for managing datasets, programs, and printouts in scientific research. *Computer Programs in Biomedicine*, 13, 281-288. **(Reproducibility)**

Muller, Barton, and Benignus (1984) Recommendations for appropriate statistical practice... *Neurotoxicology*, 5, 113-126.

Muller (1986) Design and analytical methods. Chapter 18, *Neurobehavioral Toxicology*, Z. Annau, ed., 404-423. Hopkins Press.

Chapter 11, *Selecting the Best Model*, in Muller and Fetterman (2002) *Regression and ANOVA: An Integrated Approach Using SAS® Software*.

Cheng, Edwards, Maldonado-Molina, Komro, and Muller (2010) Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in Medicine*, 29, 504-520. PMCID: PMC2811235

Relevant Personal Bibliography continued

Kairalla, Coffey, Thomann, and Muller (2012) Adaptive trial designs: a review.. *Trials*, 13(145). PMCID: PMC3519822

Guo, Logan, Glueck, and Muller (2013) Selecting a sample size for studies with repeated measures. *BMC Medical Research Methodology*, 13(100). PMCID: PMC3734029

Kreidler, Muller, Grunwald, Ringham, Coker-Dukowitz, Sakhadeo, Barón, and Glueck (2013) GLIMMPSE: online power computation for linear models ... *Journal of Statistical Software*, PMCID: PMC3882200

SampleSizeShop.org has some supporting articles, tutorials, talks and software.

UF UNIVERSITY of FLORIDA

Bibliography: Recent Sources

The rate of publication means any list would need to be updated monthly.

In the last year two special issues of journals have addressed the topic!

Proceedings of the National Academy of Sciences (2018) special issue on "Reproducibility of Research: Issues and Proposed Remedies." *http://www.pnas.org/content/115/11/2561*

The American Statistician, (2019) special issue on "Statistical Inference in the 21st Century: A World Beyond p < 0.05" *https://www.tandfonline.com/toc/utas20/73/sup1*

Bibliography: Continued

The journal issues cited and references in cited works provide entryways to the large literature.

Pubmed, Web of Science, GoogleScholar, etc., will keep you up to date.

Amy Lehman (Ohio State Biostatistics) posted a small bibliography at https://ccts.osu.edu/sites/default/files/inline-files/ Articles on Irreproducible Research and P values.pdf

Allison et al. (2016) Reproducibility: a tragedy of errors, *Nature*, 530 (7588), 27-29.

Iqbal, ... Ioannidis, (2016) Reproducible Research Practices and Transparency across the Biomedical Literature, *PLoS Biology* 14(1): e1002333

Guideline 1. Explicitly Control Type I Errors (False Positives) and Type II Errors (False Negatives).

Why do we care?

Reproducible science requires accounting for the myriad of ways the multiple testing problem arises in contemporary research.

To understand multiple tests requires understanding a single test.

The discussion is cast in terms of a pathology reading,
benign or malignant.

Traditional Description for Testing

		State of Nature	
		H_0	H_A
Decision	H_0	**correct negative**	**false negative**
	H_A	**false positive**	**correct positive**

probability of a type I error is α

Probability of a type II error is $\beta = 1$ - power

UF | UNIVERSITY of FLORIDA

Hidden Problem: Multiple Testing Can Inflate Type I Error Rate

Multiple testing happens in many ways:

- many outcomes, such as multiple cognitive performance tests,
- analyzing data in subgroups, and
- fitting many different models, such as with and without gender, age,...

Traditional simple test methods give biased estimates of Type I error.

Reproducible Science Requires Accounting for the Myriad Ways Multiple Testing Arises In Contemporary Research

Statisticians often joust about methods for estimation and inference.

Decision theory provides the overarching framework.

Differences and disagreements arise from differences in choice of probability models and *especially* in choice of cost functions.

The presenter strives to be closely aligned with Sir David Cox (*Principles of Statistical Inference*, Cambridge Press, 2006; also invited talks).

Rodgers and Shrout (2018) Psychology's replication crisis as scientific opportunity: a précis for policymakers, *Policy Insights from the Behavioral and Brain Sciences* 5 (1), 134-141.

Many Decry the Poor Rate of Replicability

Johnson, Payne, Wang, Asher and Mandal (2016) On the reproducibility of psychological science, *Journal of the American Statistical Association*.

"The results of this re-analysis provide a compelling argument for both increasing the threshold required for declaring scientific discoveries and for adopting statistical summaries of evidence that account for the high proportion of tested hypotheses that are false."

Johnson also has an article in the recent special issue of the *American Statistician*.

How Do We Recognize and Avoid the Problem?

Planning, planning, planning.

Follow Consort guidelines for clinical trials.
http://www.consort-statement.org

Follow Strobe guidelines for observational data.
https://www.strobe-statement.org

Use true split-sample methods to separate exploratory and confirmatory analyses (Chapter 11, *Selecting the Best Model*, Muller and Fetterman, 2002).

Accurate Reporting

All of what you have done must be fully revealed in your report.

Exploratory analysis is a critical tool of science (model selection).

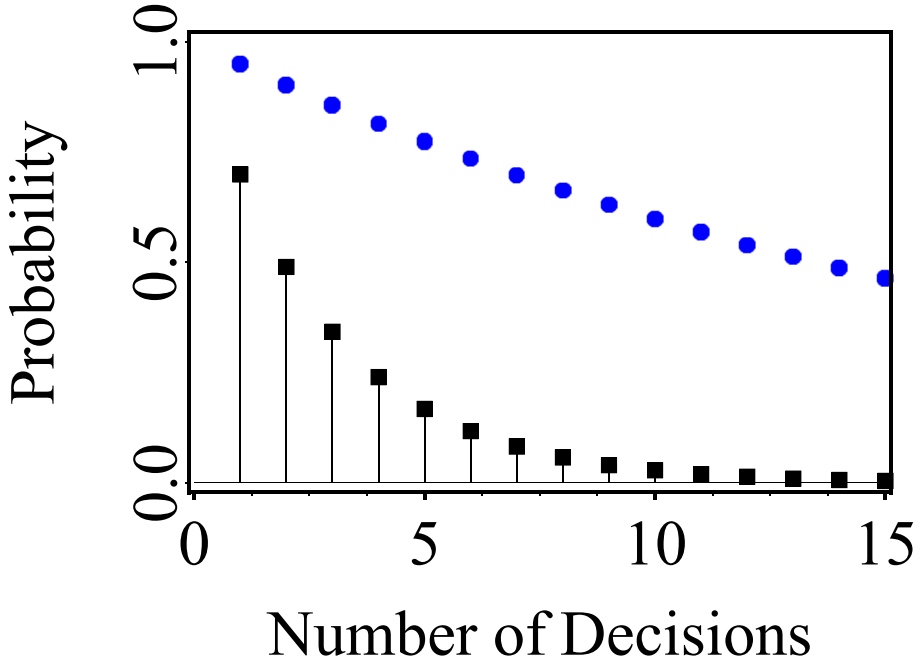
Misrepresenting selected exploratory analyses as confirmatory is one of the primary contributors to non-replicable research.

The error is not the use of exploratory approach.

The error is the failure to report the selection process which inflates decision error rates.

Inaccurate Reporting Under the Null

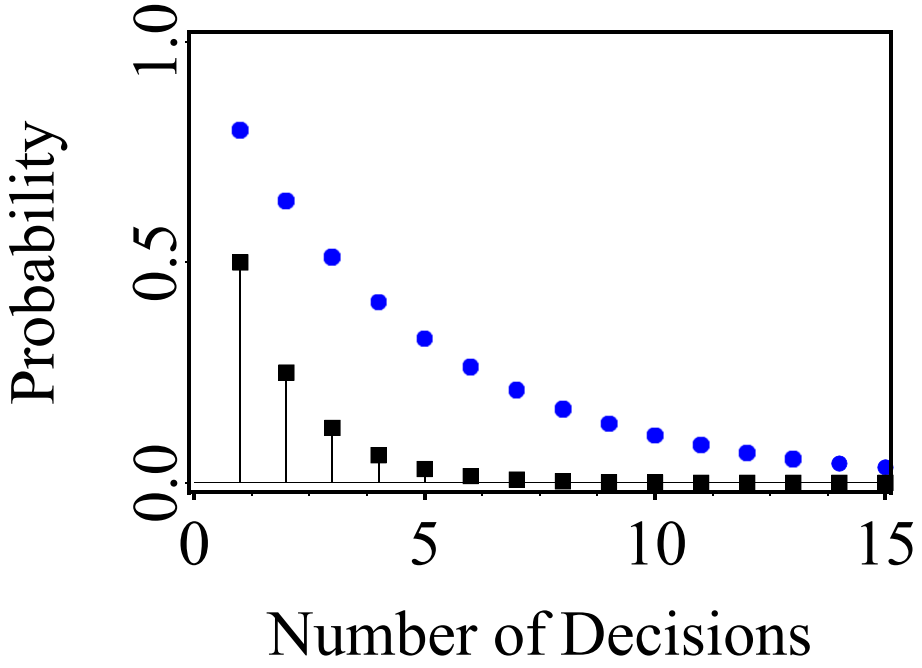
If the null hypothesis is true, the probability of a correct replication goes down as the type I error rate goes up and as the number of decisions goes up.



UF UNIVERSITY of FLORIDA

Inaccurate Reporting Under the Alternative

If the alternative hypothesis is true, the chance of a correct replication goes down as type II error rate goes up and as the number decisions goes up.



Achieving Accurate Reporting

The problem arises from a handful of statistical issues embedded in a culture allowing study details not being public, reproducible and replicable.

Rodgers and Shrout (2018) Psychology's replication crisis as scientific opportunity: a précis for policymakers, *Policy Insights from the Behavioral and Brain Sciences* 5 (1), 134-141.

Maxwell supported the same position with respect to statistical power in many papers: CV at *https://psychology.nd.edu/faculty/scott-e-maxwell/*.

Nothing is broken. We need culture change to get back to basics.

Mark Twain's comment on quitting smoking comes to mind.

Achieving Accurate Reporting

An important creator of non-replicable research is publication bias, which pre-registration (e.g. Clintrials.Gov) greatly helps to alleviate.

Iyengar and Greenhouse (1988) Selection models and the file drawer problem. *Statistical Science*, 3(1), 109-117.

Consider early work by Rosenthal in *Psychological Bulletin*, and the more recent work by Val. Johnson (recently in the *American Statistician*).

Guideline 2. Align the Goals, Design, Data Analysis, and Sample Size Analysis.

Conscientious scientists
  align scientific **goals** with the study **design** and
  align the study **design** with the **data analysis**.

Success requires enough time and support.

In contrast, aligning the sample size analysis
  has historically been difficult and often not done.

Lack of planning time, support, information, and methods are barriers.

How do we overcome the barriers?

Underfitting the Correlation Pattern Can Inflate the Type I Error Rate; Misaligned Design and Data Analysis

The most common example involves assuming simple models with equal correlations among variables when the pattern is complex.

Gurka, Edwards and Muller (2011) discussed linear mixed model examples in which the target type I error rate was 0.05 and the actual error rate was roughly 0.20.

Would patients be happy with a 0.20 false positive rate?

Gurka, Edwards, and Muller (2011) Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, 30, 2696–2707.

Align Sample Size Analysis with Data Analysis

Method must match method;
hypothesis must match hypothesis.

Muller, LaVange, Ramey and Ramey (1992, §5.2) gave two examples of >-test power being wrong for a repeated measures test.

The >-test power was *lower* than the actual power for a longitudinal study of child development.

The >-test power was *higher* than the actual power for a longitudinal study of kidney disease.

A Warning. Using a simplified hypothesis for a power analysis can lead to a sample size that is either too high or too low.

Align Sample Size Analysis with Data Analysis

Time-to-event ("survival") analysis based on proportional hazard models are widely used in large and small clinical trials.

The primary outcome is often the difference between treatments in median survival time, which is censored.

It is common to select a sample size for such studies based on a binomial outcome.

The risk of misalignment appears high due to computing power for the wrong hypothesis.

Align Sample Size Criterion *with Data Analysis*

Power and confidence interval width are different criteria.

Is the goal a confidence interval, a rejection of a null (power), or both?

Some authors incorrectly suggest using confidence interval width as a criterion to select sample size.

Jiroutek et al.(2003) give an example in which sample size could be about 4 times too large (106 vs. 24), or about 3 times too small, if one uses a width rather than a power criteria.

Barriers to Aligning Sample Size Analysis

There is a lack of statistical methods for many useful analyses.

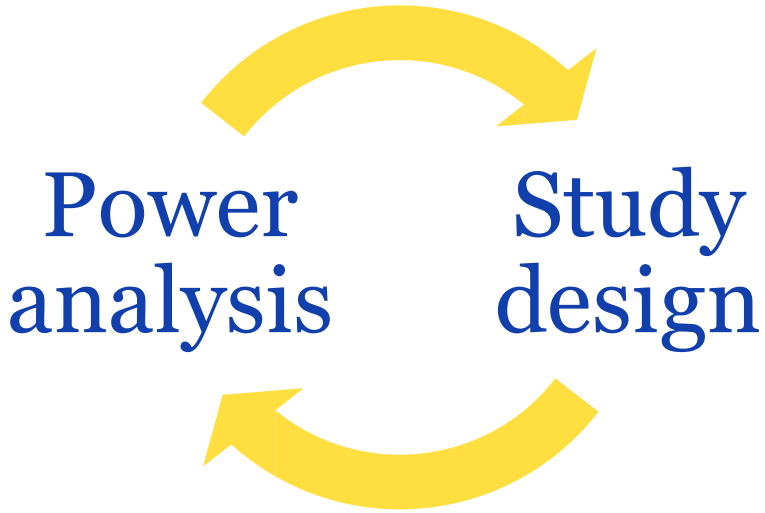
When there are sample size methods, there is often a lack of software for complex designs and hypotheses.

Our free GLIMMPSE software helps fill the gap for broad classes of linear models (**SampleSizeShop.org**).

Sample size calculations must be done after the study is designed. Many scientists do not realize how much time and effort is needed to create a credible sample size analysis.

Power Analysis Is an Iterative Process

Design changes are made in response to initial power analysis, which requires new power analyses.



A Shameless Plug for our Stuff

Recommend GLIMMPSE at **SampleSizeShop.org**

It provides power and sample size for continuous outcomes with any (factorial) combination of longitudinal and clustered (multilevel) design features.

It is free on the web.

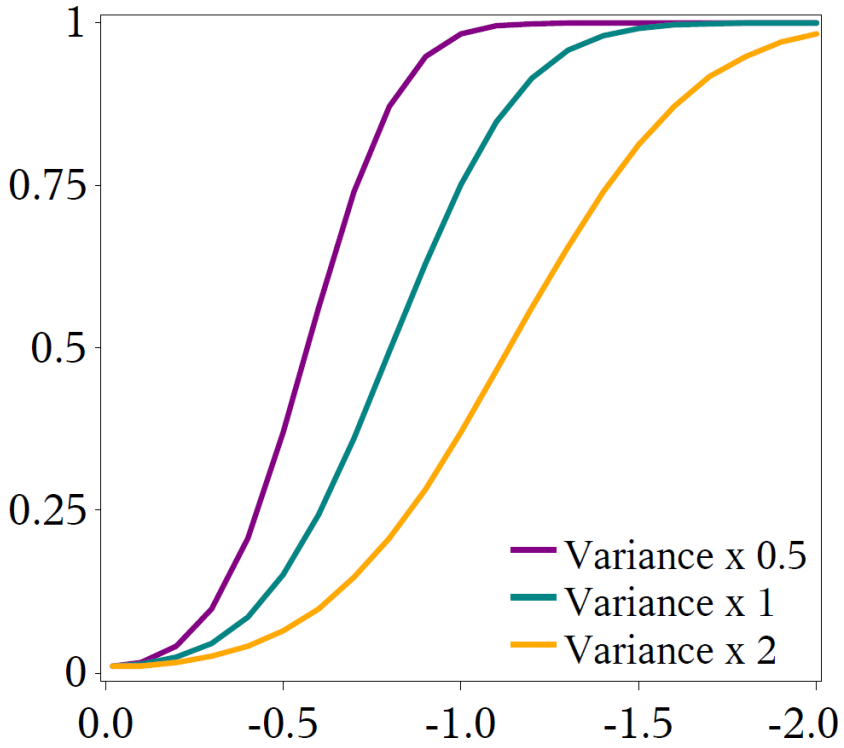
In contrast to much free software it has been validated by professional software engineers to professional standards.

It uses a point and click interface:
you only need a web browser on a computer, tablet, or smartphone.

It supports appropriately complex correlation patterns across time.

Guideline 3. Account for Uncertainty Sample Size Computations

Conduct Sensitivity Analyses for Speculated Values



UF | UNIVERSITY of FLORIDA

Account for Statistical Uncertainty

Account for uncertainty in inputs to sample size computations.

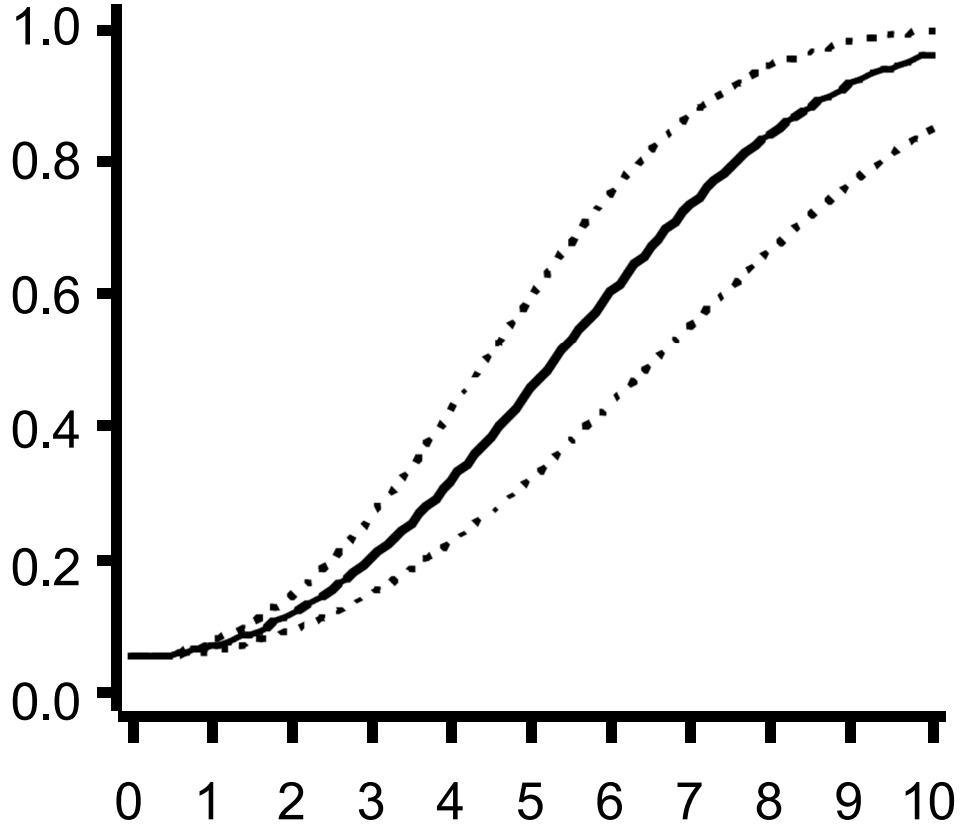
Values of means, variances and correlations from pilot studies are random; they are statistical estimates.

Power and sample size values are also statistical estimates when based on (random) estimates from pilot studies.

Taylor and Muller (1995) Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician*, 49, 43-47.

The method is implemented in GLIMMPSE at *SamplesizeShop.org* and illustrated on next page (plots are a user task).

Power as a Function of Mean Difference for Observed Variance (Black Line) and 95% Confidence Limits (Dotted Lines) (Taylor and Muller, 1995)



Investigators Often Plan Sequences of Studies

If the first study is *non-significant*, scientists may plan a larger study in an attempt to achieve significance.

A pessimistic power calculation give a sample size that is too large.

If the first study is *significant*, scientists may plan a new study to replicate the result.

An optimistically small sample size is the result.

Some Statistical Tools Can Help

Taylor and Muller (1996) Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods*, 25, 1595-1610.

Recent free software implements the methods in Taylor and Muller.

Anderson and Maxwell (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, 52, 305-324.

Conclusion: Following the Guidelines Will Solve the Problem

Extending the Guidelines to a Sequence of Studies

To protect replicability across a sequence of studies requires careful allocation of exploratory and confirmatory analyses within each.

Muller, Barton and Benignus (1984) described a *leapfrog design*. Our recommendation was to allow discussion only of *supportive* exploratory (same direction as the confirmatory) results in a paper.

Some forms of adaptive designs are very appealing; others are not.

Kairalla, Coffey, Thomann, and Muller (2012) Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13(145).

Recognizing Reproducible Research

Reproducible research is just science: public and replicable.

There are different ways to describe the same epistemology in an ethical world.

Published research can be judged to be reproducible and replicable, or not. Research can be planned to be reproducible and replicable, or not.

Although vexing, error variance is job security for statisticians and psychometricians.

Software and further readings are at **SampleSizeShop.org**.

Outline of Presentation

Motivating Problem: the Replication Crisis

Guideline 1. Explicitly Control Type I Errors (False Positives) and Type II Errors (False Negatives).

Guideline 2. Align the Goals, Design, Data Analysis, and Sample Size Analysis.

Guideline 3. Account for Uncertainty in Sample Size Computations.

Conclusion: Following the Guidelines Will Solve the Problem.

Questions?