From One Environment to Many: The Problem of Reproducibility of Experimental Results

by

Jinguang Lin

M.S., Guangzhou University of T.C.M., China, 2011

A REPORT

submitted in partial fulfillment of the
requirements for the degree

Master of Science

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Michael J. Higgins

Copyright

Abstract

When the same experiment is carried out in a different environment, the error term not only includes the random error within a given experiment, but it also includes the additional sources of variability that are introduced by conducting the same experiment in different environments. These differences include both natural factors such as location, time or weather and other factors such as personnel or equipment necessary to carry out this experiment. By considering the effect of changing experimental environments on the reproducibility of experiments, we try to figure out in what situations the initial experimental results will likely carry over to other environments. We examine how p-value, effect size, sample size, and the ratio of the standard deviation of environment by treatment interaction and the standard deviation of experimental error interact with one another, and as a whole, affect the experiment's reproducibility. We suggest that not only p-values but also the effect sizes and the *environmental effect ratio*—the ratio of the standard deviation of environment by treatment interaction and the standard deviation of experimental error—should be considered when researchers are making statistical inferences. Large effect sizes and/or small ratios of the environmental effect ratio favor high probability of reproducibility. If the environmental effect ratio is too large, the reproducibility probability may be reduced to just a coin toss, and if effect sizes are small, researchers should be very cautious about making inferences about reproducibility even if the observed p-value is small and sample size is large.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisors, Dr. James Higgins and Dr. Michael Higgins. They both are always available, and always patient to answer my questions. Thanks for their excellent guidance, support and help. I would like to thank Dr. Wei-Wen Hsu for his help, comments, time and serving as my supervisory committee member, and i would also like to thank other professors and staff in the Statistics department. Thanks for the teaching and help. I would like to thank K-State University. More than two years of living and studying here are full of love and fun. This will always be my precious memory. My deepest thanks also go to my family for all the support and love.

Chapter 1

Introduction

Researchers are more and more concerned with the reproducibility of scientific research since they have found a high probability of failure to reproduce the same results as other researchers. A survey of *Nature* readers found that about 70% of scientists have failed to reproduce other researchers' experiments, and more than 50% have failed to reproduce their own studies (Baker, 2016). Scientists have referred to these concerns as the problems of research reproducibility which are quite pervasive across all scientific domains (Begley and Ioannidis, 2015).

1.1 Reproducibility in Agriculture Research

Agriculture shoulders the responsibility of supplying safe and secure food for an exponentially growing population. From the time of R.A. Fisher at Rothamsted Experiment Station to today, agricultural researchers have recognized the difficulty in drawing valid statistical inferences from an experiment done under non-homogeneous conditions. They were early adopters of the techniques of blocking and randomization to mitigate the effects of confounding environmental factors, and now such techniques are recognized as the gold standard for research. However, even with this well-established systematic approach over several decades, many times, scientists found it difficult to reproduce other researcher's experiment in agri-

cultural research which is due to misconceptions and misuse of statistics (Bello and Renter, 2018). Incorrect or improper conduction in experimental design, methodology, data, or computer code as a whole put forward a big challenge for the reproducibility of research (Richard et al., 2016). To fix this problem, open science and standardized implementation of statistical principles and methods should be encouraged (Bello and Renter, 2018).

1.2 Reproducibility in Social Science Study

In social science, researchers want the treatment to be homogeneous across individuals, but at the same time they also worry about that the experimental sample do not represent the population as a whole. The infrequency of some data sources (Census, yields, and seasonal prices) may also degrade the power of studies (Richard et al., 2016). In 2015, an open science collaboration study was published on *Science* saying that researchers conducted replications of 100 experimental and correlational studies published in three psychology journals using original materials and found that less than 50% of the original findings were reproducible (Aarts et al., 2015). A study from Federal Reserve Board of USA, which tried to replicate 67 papers published in 13 well-regarded economics journals, concluded that economics research is usually not replicable; the study found that more than half of the papers are irreproducible despite using author-provided data and code. (Chang and Li, 2015). Even researchers that have big data from big projects of big companies, such like the Google's project of Google Flu Trends, have found that bad experimental design and bias of inference could be the blame for the failure of reproducibility (David et al., 2014; Bello and Renter, 2018).

In psychological studies, researchers need to predict and determine the internal or external conditions, such like cognition, affect, behavior or situation, for the stability and the variability of psychological phenomena because psychological phenomena may vary across time, situation and persons (Iso-Ahola, 2017). Bavel et al. (2016) analysized 100 replication attempts in psychology and found that replication success was associated with the extent of contextually sensitive variables (varying in time, culture, population, or location) which

might be a significant predictor of replication success even adjusting for effect size, or statistical power. Johnson et al. (2017) used a method based on z-transformed correlation coefficients to reanalyze 100 psychology experiments from *the Open Science Collaboration* and found that more than 90% of experimental hypotheses were incorrect.

With the problem of reproducibility in social science study a serious problem like this, the confidence in the research results should be a concern, and how to fix this problem needs in-depth discussion and intensive study. The proposed solutions should always be involved with open science and collaboration (Richard et al., 2016).

1.3 Reproducibility in Biomedical Science

The lack of reproducibility in the health science researches has been lamented by scientists (Boos and Stefanski, 2011). Every year, governments spend heavily on biomedical study hoping to find some ways to treat diseases or improve human health. At the same time, most of studies turn out to be irreproducible after researchers spend huge amounts of effort and time on it. The cause of irreproducibility in biomedical research is complicated. It could be due to the experiment design, the characteristic of the experimental samples, the natural variability in the biological experimental systems, the expertise of the personnel, the equipment, and/or the change of experimental environment. "*Huge expense of human research, ethical issues, the complex interaction of diet, genetics, environment and metabolic factors, the drastically change of time, geography and culture, strictly controlled experimental situation may cause the problems of lacking adequate controls, and make reproduction of results in real-life situations difficult and may limit generalizability of results to society in general*" (Richard et al., 2016). Prinz et al. (2011) used in-house target validation method to reproduce the results of 67 medical projects and found that 65 percent of results were not consistent with the original studies. These failures of reproducibility of biomedical research may risk loss of credibility with the public and in some way will hinder the development of biomedical research (Mullane et al., 2018).

This is a challenge that needs to be addressed globally through cooperation and collab-

oration between researchers and institutions (Partnership, 2016). Kaplan and Irvin (2015) suggests identifying the treatment effects should depend on transparent and impartial reporting of clinical trial results. Prinz et al. (2011) conjectured that heterogeneous experimental conditions caused by negligence over the experimental conditions could be an explanation for the failure of reproducibility despite self-correction and peer review, and recommended fulfillment of confirmatory validation studies should be prior to larger investment. Zarin et al. (2011) suggested that the study's purpose, recruitment status, design eligibility criteria, locations should be registered beforehand. In order to improve reproducibility, a number of measures such as greater openness and transparency of methods and data, collaboration, strict reporting and registering guidelines, peer review, automation, personnel training are encouraged (Partnership, 2016).

1.4 Statistical Significance and Reproducibility

When many scientists are putting a large amount of effort trying to get a significant result, do significant results really mean significant effects? Ioannidis (2005) noted "*A research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser pre-selection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical models; when there is greater financial and other interest and prejudice...*". For example, since the launch of *the clinicaltrials.gov registry* in 2000, which forced researchers to preregister their methodology, the percentage of significant positive results from large heart-disease clinical trials dropped from 57% to just 8% (Bavel et al., 2016). One possible explanation is that reduced error variance and more precise estimates of treatment effects were provided by newer clinical trial management methodologies (Kaplan and Irvin, 2015).

Journals tend to favor publication of statistically significant findings (Kaplan and Irvin, 2015). When researchers perform statistical inference, they likely rely on p-values to determine the "statistically significant" results. However, p-values may be misleading. Sample-

to-sample variability and ignoring variability in p-values is potentially misleading and could cause a lack of replication in studies; therefore in some situations the p-value should be adjusted (Boos and Stefanski, 2011). Based on the p-value method, Shenhav et al. (2015) mentioned a max (p-value) method (with $P1$, $P2$ as the p-values yielded from the original experiment and the reproduced experiment respectively) would be more precise. Boos and Stefanski (2011) argued that reporting the magnitudes of the p-values would make more sense. However, statistically significant p-values about treatment effects without further context do not have any practical significance (Bello and Renter, 2018). A small p-value, or statistical significance, does not imply a large size of an effect.

1.5 A Different Experimental Environment

In order to reproduce other researcher's experimental result, people need to carry out the experiment in the same or similar environment. However, only repeated experiments with similar findings will give "reproducibility" more sense (Goodman et al., 2016). Basic terms such as reproducibility, replicability, reliability, robustness, and generalizability are nearly identical and "reproducibility" consists of "methods reproducibility", "results reproducibility", and "inferential reproducibility" (Goodman et al., 2016). We will consider "inferential reproducibility".

When follow-up research fails to yield the same significant result as the initial research, the research concerns may focus on what went wrong with the experimental design, the methodology, the data, or the computer code. What perhaps is not as well-known is the fact that the follow-up research must be done in a different experimental environment. This can lead to failure to confirm the initial results even if two experiments are done flawlessly. The environment of an experiment includes natural factors such as weather, location, time, and other factors like the characteristic of the experimental units, the experimental protocol, the expertise of the personnel, and the equipment used to carry out the experiment. These uncontrollable and controllable factors can interact with each other and, as a whole, affect the outcome of an experiment in unique ways. From the time of R.A. Fisher when he

used the techniques of blocking and randomization to reduce the effects of confounding factors, to today, researchers have realized the difficulty in making valid statistical inferences from an experiment done under non-homogeneous environments. A biology multi-laboratory study figured out that genotype×laboratory interactions were unavoidable even under strict and careful standardization which implied that results from a single experiment might be idiosyncratic (Kafkafi et al., 2005).

When researchers finish an experiment and have significant results, they want their results to apply broadly. This is the solid foundation for the progress of science. If researchers can replicate the experiment under several randomly selected environments, then they can draw valid inferences about average treatment effects through the techniques of mixed models where the average is taken across environments. With limited coordinated community effort, and unavoidable existence of interaction across experimental environments, a mixed model where the size of interaction can be estimated from multi-laboratory results should be used if possible (Kafkafi et al., 2005). Even if researchers realize that there is interaction across experimental environments, it may not be possible to replicate the experiment because of costs, time, or other limitations. Thus the question is: under what conditions would it be reasonable to infer that results from an experiment done in one environment could be reproduced in another environment?

1.6 Introduction to Our Study

In this study, we define reproducibility as obtaining the same inference in a follow-up experiment as in the initial experiment. We examine how p-value, effect size, sample size, and variances between different environments interact with one another, and as a whole, affect the experiment's reproducibility. We propose three quantities that measure the likelihood that results of the initial experiment can be reproduced in the follow-up experiment: the probability of reproducibility, the adjusted p-value, and the adjusted confidence level. These quantities depend on four quantities: the *environmental effect ratio*—the ratio of the standard deviation of environment by treatment interaction and the standard deviation of

experimental error, the effect size, the level of significance of the test or the level of confidence, and the sample size. We focus on environment by treatment interaction and effect size.

Chapter 2

Model and Test Statistics

2.1 Model

We design an experiment to compare the effect of two treatments. Observations from the initial experiment are assumed to follow the model (we call it Model 2.1 here)

$$Y_{ij} = u_i + \epsilon_{ij}, i = 1, 2, j = 1, ..., n \tag{2.1}$$

where u_i is the mean of the ith treatment and the ϵ_{ij}'s are iid random variables distributed as $N(0, \sigma_e^2)$. We would like make inferences on $u_1 - u_2$.

In the follow-up experiment, when the same experiment is repeated, changing experimental conditions are assumed to affect the responses either in some systematic way or in ways unique to each treatment. We express this with a mixed model (we call it Model 2.2 here)

$$Y_{ij} = u_i + \theta + \delta_i + \epsilon_{ij}, i = 1, 2, j = 1, ..., n \tag{2.2}$$

The u_i's and the ϵ_{ij}'s follow the assumptions in Model 2.1. The term θ represents random sources of variability common to all observations such as weather, time or location. The δ's represent random sources of variability unique to each treatment such as varying expertise of the personnel in handling each treatment, variability caused by the equipment or procedures

that are unique to each treatment, and the interactions of the treatments with the climate or location under which the treatments are administered. The distribution of θ is $N(0, \sigma_\theta^2)$. Its distribution does not figure into the distribution of differences of the means. The δ_i are idependent and identically distributed $N(0, \sigma_B^2)$ (B denoting "between" between environments within treatment (or interaction). The random terms in Model 2.2 are assumed to be mutually independent. Since we want to confirm the result of the initial experiment, we assume the initial experiment is done perfectly in a flawless experimental environment, so the Model 2.1 excludes the terms of θ, and δ.

2.2 Test Statistics and Distribution

Note: A random variable NCT has a non-central t-distribution with degrees of freedom (df) if it can be expressed in the form

$$NCT = Y/\sqrt{V/df} \tag{2.3}$$

where Y has a $N(u, 1)$ distribution, V has a chi-square distribution with degrees of freedom df, and Y and V are independent. The parameter u is called the non-centrality parameter.

Now we will use the non-central t-distribution to compute the probability of reproducibility under Model 2.2 with the same assumptions on the terms we had before. Let us consider testing $H_0 : u_1 - u_2 = 0$ against $H_a : u_1 - u_2 \neq 0$. We will consider the case in which σ_e is unknown.

The difference of sample means can be expressed as:

$$(\overline{Y}_1 - \overline{Y}_2) = (u_1 - u_2) + (\overline{\epsilon}_1 - \overline{\epsilon}_2) + (\delta_1 - \delta_2) \tag{2.4}$$

where the bar notation indicates the sample mean of the respective random variables. This has a normal distribution with mean $u_1 - u_2$ and variance $2\sigma_e^2/n + 2\sigma_B^2$. Thus, the variable

Y defined by

$$Y = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} \tag{2.5}$$

has a normal distribution with mean

$$u = \frac{u_1 - u_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} \tag{2.6}$$

and standard deviation 1. The random variable V defined by $V = 2(n-1)(S_p^2/\sigma_e^2)$ has a chi-square distribution with $df = 2(n-1)$ degrees of freedom, where S_p is the "pooled" sample standard deviation from the observations taken on the two treatments. Thus, $S_p/\sigma_e = \sqrt{V/df}$. The test statistic is

$$t - stat = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{2/n}} \tag{2.7}$$

Multiply and divide the right hand side of 2.7 by $\sqrt{2\sigma_e^2/n + 2\sigma_B^2}$, and after some simplification, the $t - stat$ can be expressed as

$$t - stat = \frac{Y}{\sqrt{V/df}}\sqrt{1 + n\sigma_B^2/\sigma_e^2} = NCT\sqrt{1 + n\sigma_B^2/\sigma_e^2} \tag{2.8}$$

where under H_a, NCT has a non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter

$$u = \frac{u_1 - u_2}{\sqrt{2\sigma_e^2/n + 2\sigma_B^2}} = \frac{\sqrt{n}\Delta}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}} \tag{2.9}$$

where Δ is the effect size defined by

$$\Delta = (u_1 - u_2)/(\sigma_e\sqrt{2}). \tag{2.10}$$

The effect size is the expected value of the difference between two observations, one from treatment one and the other from treatment two, divided by the standard deviation of this difference. It occurs naturally in certain mathematical expressions in our discussion. Another definition of effect size is called Cohen's d (Cohen, 1988), which excludes the factor $\sqrt{2}$.

2.3 Probability of Reproducibility

2.3.1 Computation of Probability of Reproducibility

We define the *probability of reproducibility* as the probability that the follow-up experiment (Model 2.2) yields a significant result, assuming the initial experiment (Model 2.1) yields a significant result. Without loss of generality, we assume that $u_1 > u_2$ in testing $H_0 : u_1 - u_2 = 0$ against $H_a : u_1 - u_2 \neq 0$ using the $t - statistic$ and that the test is done at level of significant α which is set at the traditional .05 level or smaller. Let $t_{\alpha/2,df}$ denote the upper $\alpha/2$ percentage point of the Student's t-distribution. We consider the probability of reproducibility only in the case in which the initial test is statistically significant. If $t - stat \geq t_{\alpha/2,df}$ in the initial experiment, then we have a reproducible result if $t - stat \geq t_{\alpha/2,df}$ in the follow-up experiement. The probability of reproducibility in this case is $P(t - stat \geq t_{\alpha/2,df} | Model\ 2.2)$. In the event that the initial experiment is significant in the wrong direction, that is, $t - stat \leq -t_{\alpha/2,df}$, then reproducibility occurs if $t - stat \leq -t_{\alpha/2,df}$ in the follow-up experiment which would be a confirmation of an incorrect result. Since $P(t - stat \leq -t_{\alpha/2,df} | Model\ 2.2)$ is small, less than $\alpha/2$ when $u_1 > u_2$, we will only consider $t - stat \geq t_{\alpha/2,df}$ in computing the probability of reproducibility. Thus, for a two-sided test at level of significance α, the probability of reproducibility is

$$P(t - stat > t_{\alpha/2,df}) = P\left(NCT > t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \tag{2.11}$$

When the experiment is done the first time, we assume that there are no random components other than the ϵ's, so the distribution of the $t - statistic$ in this case can be attained by setting $\sigma_B = 0$ in the computation above. Thus, the probability of reaching the correct decision in the first experiment is

$$P(t - stat > t_{\alpha/2,df}) = P(NCT^* > t_{\alpha/2,df}) \tag{2.12}$$

where NCT^* has a non-central t-distribution with degree freedom $2(n-1)$ and non-centrality

parameter $\sqrt{n}\Delta$.

We denote $G_{df,u}(t)$ as the cumulative distribution function (*cdf*) of this non-central t-distribution, where *df* is the degrees of freedom, and u is the non-centrality parameter. Then the probability of reproducibility is

$$1 - G_{df,u}(t_{\alpha/2,df}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}) \tag{2.13}$$

The power of the initial test is approximately $P(t - stat \geq t_{\alpha/2,df}|Model\ 2.1)$, but that is not the case in the follow-up experiment. Under Model 2.2, the probability of significance in the wrong direction, i.e. $P(t - stat \leq -t_{\alpha/2,df}|Model\ 2.2)$, can be substantial and may be as large as .5 depending on the size of σ_B/σ_e—we call this term the *environmental effect ratio*. Thus, the probability of reproducibility is the probability of reaching correct conclusion in the follow-up experiment both in terms of rejecting the two-sided null hypothesis H_0 and doing so in the right direction. It is affected by four factors: α, n, Δ, and σ_B/σ_e.

Z-Test

When we consider the case in which σ_e is known, we still use the Model 2.1 and the Model 2.2 for the initial experiment and the follow-up experiment respectively and the assumptions based on both models still hold. The $z - statistic$ for Model 2.2 is

$$z - stat = (u_1 - u_2)/(\sigma_e\sqrt{2/n}) + (\bar{\epsilon}_1 - \bar{\epsilon}_2)/(\sigma_e\sqrt{2/n}) + (\delta_1 - \delta_2)/(\sigma_e\sqrt{2/n}) \tag{2.14}$$

from which we can see the distribution of $z - stat$

$$z - stat \sim N(\Delta\sqrt{n}, 1 + n\sigma_B^2/\sigma_e^2) \tag{2.15}$$

The probability of reproducibility is then

$$1 - \Phi\left(\frac{z_{\alpha/2} - \Delta\sqrt{n}}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}\right) \tag{2.16}$$

where Φ is the *cdf* of the standard normal distribution, and $z_{\alpha/2}$ is the $\alpha/2$ upper percentage point of the standard normal distribution.

Normal Approximation

The central and non-central t-distribution have degrees of freedom determined by the sample size. We shall use the normal distribution, as mentioned above, as an approximation for the central and non-central t-distribution to get an approximate formula when the sample size is sufficiently large such that a normal approximation to the central and non-central t-distribution is adequate. In this situation, under null hypothesis, the test statistic $t - stat \dot{\sim} N(0,1)$, and under alternative hypothesis, $t - stat \dot{\sim} N(u,1)$, where u is the non-centrality parameter. We shall use $t_{\alpha/2,df} \approx z_{\alpha/2}$, $t_{\beta,df} \approx z_{\beta}$, and

$$G_{df,u}(t_{\alpha/2,df}) \approx \Phi(z_{\alpha/2} - u). \tag{2.17}$$

where G and Φ is the *cdf* of non-central t-distribution and normal distribution respectively and $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of Φ.

2.3.2 The Effect of σ_B/σ_e on Probability of Reproducibility

We will assume that the initial experiment has sufficiently high power to detect an effect size of practical importance. We assume the initial experiment is well designed and well carried out and α is set at the traditional .05 level or smaller. We investigate conditions under which the follow-up experiment also produces a significant result. The follow-up experiment introduces the variance component σ_B^2 into the model, and it affects the probability distribution of the test statistic through the deviation ratio σ_B/σ_e. The size of σ_B/σ_e will depend on how well treatments can be administered consistently across environments and the interaction of the environment with the treatments. Less consistency and more interaction will result in larger values of σ_B/σ_e.

Example 1: Small n, Traditional α, $\Delta = 1.02$

Snedecor and Cochran (1989) illustrate the independent sample t-test with data from a study to compare the comb weights of male chicks given one of two hormone treatments. The sample size is $n = 11$, the sample means are 97g and 56g, and the pooled standard deviation is 12.14g. A test for differences of means gives $p = .003$ for the two-sided test. The observed effect size is 1.02.

To illustrate how σ_B/σ_e affects the probability of reproducibility, we consider an initial experiment like this one with a sample of size $n = 11$, $\alpha = .05$, and the true effect size $\Delta = 1.0$. Under Model 2.1, the $t - statistic$ has power .88 for this effect size, which we assume would meet the requirements of the researchers. The probability of reproducibility is $P(t - stat \geq t_{0.025,20}|Model\ 2.2)$. Figure 2.1 shows a plot of this probability and plots the probabilities of significance in the wrong direction, i.e. $P(t - stat \leq -t_{0.025,20}|Model\ 2.2)$, and non-significance as a function of σ_B/σ_e.

If $\sigma_B/\sigma_e = 0$, the probability of reproducibility is the power of initial test, .88, and small values of σ_B/σ_e will ensure that there is a high probability of reproducibility in the follow-up experiment. However, as σ_B/σ_e increases not only does the probability of reproducibility decrease, but the probability of significance in the wrong direction increases. It can be shown that for any given n, α,and Δ, these probabilities of reproducibility approach .5 as $\sigma_B/\sigma_e \to \infty$. Thus, the probability of obtaining a correct result for this effect size, while is .88 for the initial experiment, can be substantially degraded in the follow-up experiment. In the case of very large σ_B/σ_e, reproducibility is essentially determined by the toss of a coin.

Example 2: Small n, Traditional α, smaller effect size: $\Delta = 0.68$

Small et al. (2018) studied curcumin's anti-inflammatory properties, and curcumin's effect on memory in non-demented adults. The study used a randomized, double-blind, two-group parallel design and mixed general linear models controlling for age and education. The sample size for the treated group was 15, and the treatment group's effect size for improving experimental subjects' attention has a $d = 0.98$, $p - value < 0.0001$.(Small's study used the

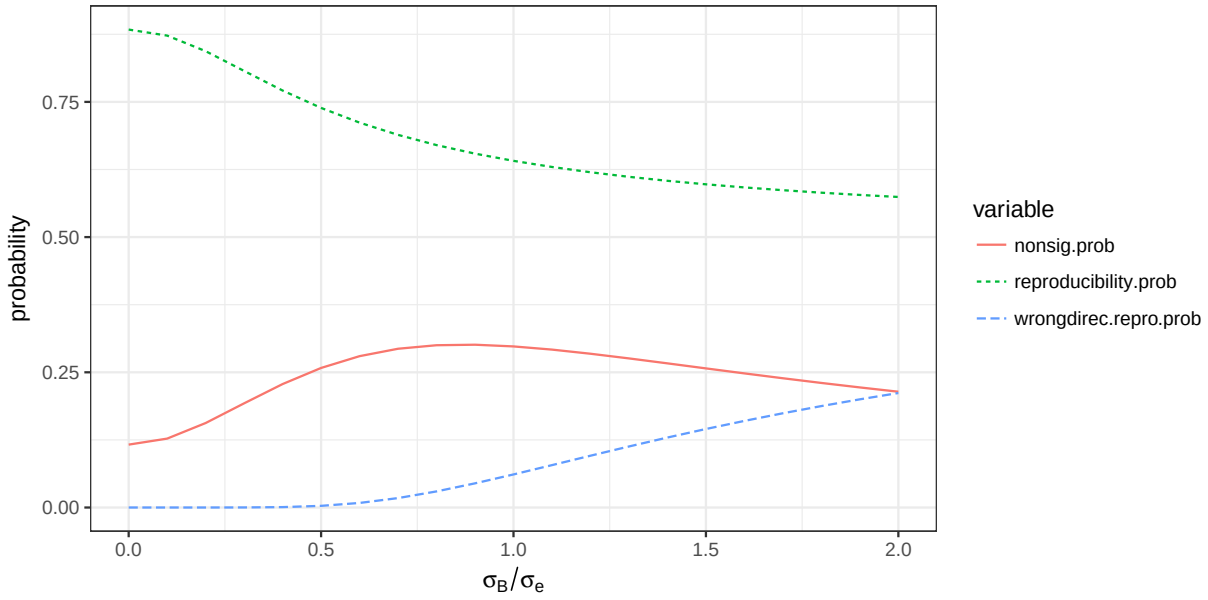

Figure 2.1: *Probability of reproducibility, probability of significance in the wrong direction, non-significance, vs.* σ_B/σ_e*,* $n = 11$*,* Δ*=1.0,* $\alpha = .05$*, initial power = .88*

Cohen's d as the effect size. We transform Cohen's d into the scale of the effect size we use in this study and have $\Delta = 0.68$). For this example we still use non-central t-distribution and the same test models . Figure 2.2 shows a plot of the probabilities of reproducibility, significance in the wrong direction and non-significance as a function of σ_B/σ_e. This plot has a similar general trend as Figure 2.1. But since the effect size is smaller, this experiment has power= .72 for this effect size under Model 2.1 and has a probability of non-significance as big as .28 respectively, which implies that for a fixed value of σ_B/σ_e, the power of finding significance, or the probability of reproducibility may decrease as effect size decreases.

Example 3: Large n and Small α

One might expect that when an initial experiment with a large sample size and small significance level yields a significant result, such result would be reproducible. However, this can be wrong when there is a small effect size. Table 2.1 shows a case in which n=500, α=.01, and Δ=.2 and 0.8.

The power of the initial test is .97 when Δ=.2. However, in the follow-up experiment with the same effect size in which σ_B/σ_e=.5, the probability of reproducibility is just .57.

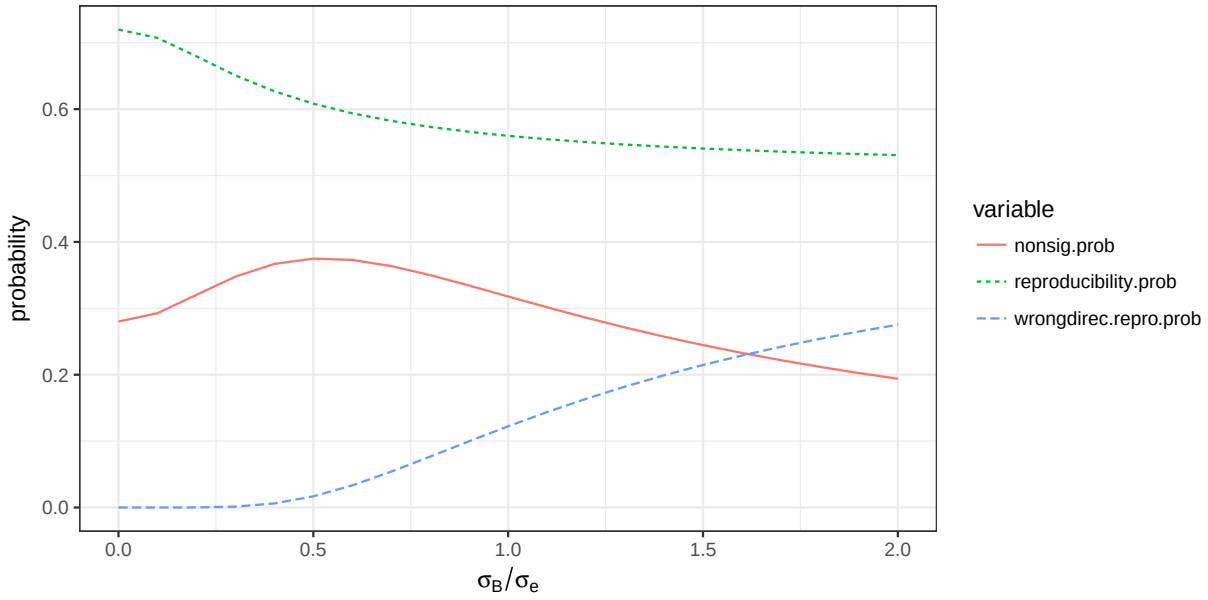

Figure 2.2: *Probability of reproducibility, probability of significance in the wrong direction, non-significance, vs σ_B/σ_e, n = 15, Δ=0.68, α = .05, initial power = .72*

n=500, α=.01,σ_B/σ_e = .5		
	Δ	
	0.2	0.8
Power of initial test: model (1)	0.97	1.00
Power of the follow-up test: model (2) two-tail rejection	0.83	0.95
Prob reproducibility: model (2) upper-tail rejection	0.57	0.91
Prob follow-up test significant in wrong direction	0.26	0.03
Prob follow-up non-significant	0.17	0.05

Table 2.1: *Example 3: a case in which n=500, α=.01, and Δ=.2 and 0.8.*

Also, the probability of finding a non-significant result is .17 and the probability of finding a significant result in the wrong direction is .26. If Δ is increased to .8, the probability of reproducibility is .91 and the probability of finding non-significance or significance in the wrong direction become much smaller.

From this example, we can conclude that to determine whether an experimental result is likely to be reproducible in a changing environment, we not only need to consider the level of significance and the power of the initial experiment but also need to take the effect size into consideration. Figure 2.3 shows plots of the probability of reproducibility and the probability of significance in the wrong direction versus σ_B/σ_e for n=500, α=.01, and Δ=.2

or .8 which illustrates that for a fixed σ_B/σ_e, a larger effect size may yield a bigger probability of reproducibility. In other words, large effect size may overcome a noisy environment.

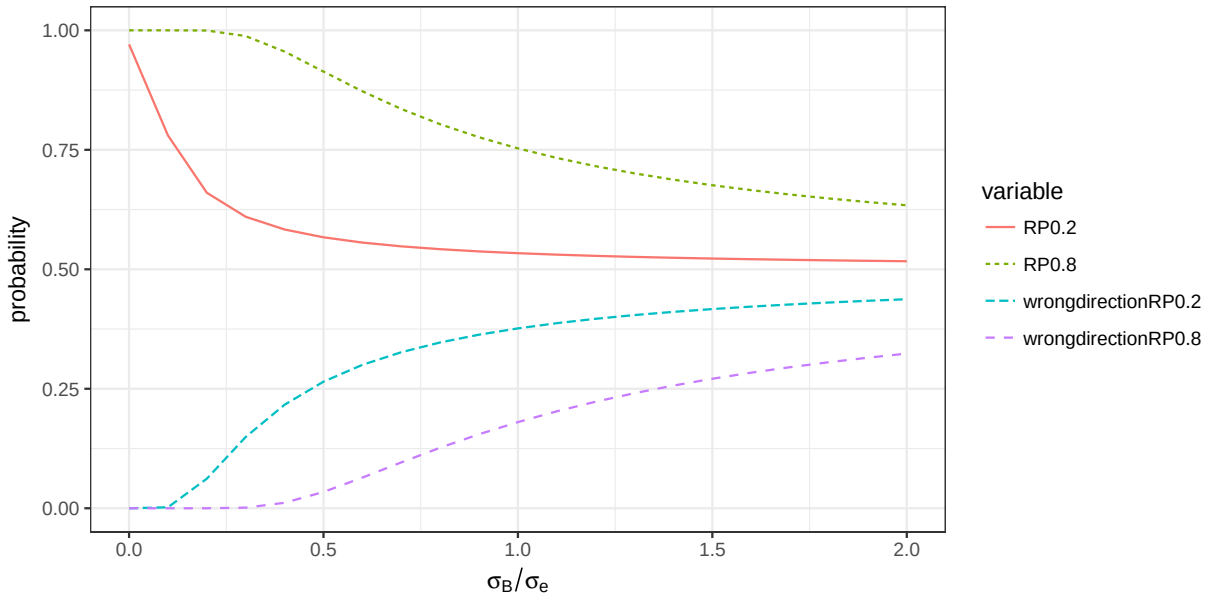

Figure 2.3: *Probability of reproducibility & probability of rejecting in the wrong direction vs* σ_B/σ_e, $n = 500$, Δ*=.2 and .8,* $\alpha = .01$, *initial power* $\geq .97$

Now, we can look at the limiting case as $n \to \infty$ to gain additional insight. By the Normal Approximation, we have

$$\lim_{n\to\infty} P(t\text{-}stat > t_{\alpha/2,df} | Model\ 2.2) = \lim_{n\to\infty} P(z\text{-}stat > z_{\alpha/2} | Model\ 2.2) \approx 1-\Phi(-\Delta/(\sigma_B/\sigma_e)) \tag{2.18}$$

Under Model 2.1, the probability of reproducibility approaches 1 as $n \to \infty$. This is not the case under Model 2.2. As Equation 2.18 shows, under Model 2.2, this limit turns out to depend on the Δ and σ_B/σ_e, but it is independent of the level of significance. Therefore, increasing the sample size or reducing the significance level will not assure getting a high reproducibility probability when random factors are in the follow-up experiment. In the other words, smaller values of σ_B/σ_e and/or larger effect sizes are the keys to obtaining high probabilities of reproducibility.

2.3.3 Relative Efficiency of the Initial and the Follow-up Experiments

In a statistical test in which the sample size is selected to achieve a desired power for a given effect size, the size of the sample is a measure of the efficiency of the test. The smaller the sample size, the more efficient the test is in achieving its objectives. If two tests are designed to test the same hypotheses and achieve the same power at the same level of significance, then the ratio of their respective sample sizes is a measure of the relative efficiency of the two tests. We will use this idea to get the relative efficiency of the initial and the follow-up experiment using the t-test.

Suppose for a given Δ and α, we would like the probabilities of reaching a correct conclusion for the initial and follow-up experiments to be the same. Because the follow-up experiment has a random component in it that is not in the initial experiment, the sample size will be larger for the follow-up experiment. Let n_I and n_F be the sample sizes necessary for the initial and follow-up experiments to have the same probability of a correct result respectively. Then, the relative efficiency of the follow-up experiment to the initial experiment is n_I/n_F. The sample size can be found by setting the probability of reproducibility equation 2.13 to $1-\beta$ and solving for n.

$$1 - G_{df,u}\left(t_{\alpha/2,df} \Big/ \sqrt{1+n\sigma_B^2/\sigma_e^2}\right) = 1-\beta \Longrightarrow G_{df,u=0}\left(t_{\alpha/2,df} \Big/ \sqrt{1+n\sigma_B^2/\sigma_e^2} - u\right) = \beta \tag{2.19}$$

Then we have

$$\begin{aligned} G^{-1}_{df,u=0}(\beta) &= t_{\alpha/2,df}/\sqrt{1+n\sigma_B^2/\sigma_e^2} - u \\ &= t_{\alpha/2,df}/\sqrt{1+n\sigma_B^2/\sigma_e^2} - \Delta\sqrt{n}/\sqrt{1+n\sigma_B^2/\sigma_e^2} \\ &= (t_{\alpha/2,df} - \Delta\sqrt{n})/\sqrt{1+n\sigma_B^2/\sigma_e^2} \end{aligned} \tag{2.20}$$

Thus,

$$\beta = G_{df,u=0}\left((t_{\alpha/2,df} - \Delta\sqrt{n})/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \tag{2.21}$$

Apply the Normal Approximation, then we have, for $\beta < .5$,

$$-z_\beta = \Phi^{-1}(\beta) \approx (z_{\alpha/2} - \Delta\sqrt{n})/\sqrt{1 + n\sigma_B^2/\sigma_e^2} \tag{2.22}$$

After some simplification, we have

$$z_{\alpha/2} + z_\beta\sqrt{1 + n\sigma_B^2/\sigma_e^2} \approx \Delta\sqrt{n} \tag{2.23}$$

For the initial experiment, $n_I = (z_{\alpha/2} + z_\beta)^2/\Delta^2$. For the follow-up experiment sample size, we need to solve for n_F in equation 2.23. Because the probability of reproducibility can be no greater than the expression on the right side of equation 2.18, it is possible that there is no solution to equation 2.23 in which case the relative efficiency is 0.

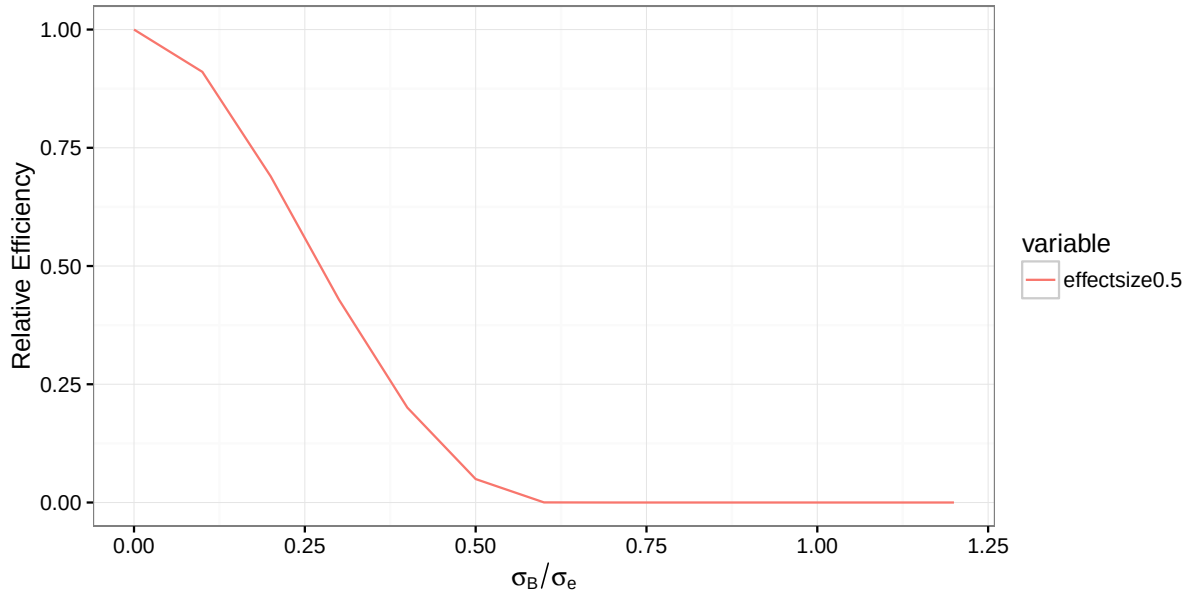

Figure 2.4: *Relative Efficiency vs* σ_B/σ_e, $1 - \beta = .8$, Δ*=0.5*, $\alpha = .05$

Figure 2.4 shows plots of the relative efficiencies for effect sizes of .5 versus σ_B/σ_e where the level of significance is .05 and the desired probability of reaching a correct conclusion

is .8. For instance, if the effect size is .5 and σ_B/σ_e is .2, the sample size are $n_I = 32$, $n_F = 46$, giving a relative efficiency $n_I/n_F = .70$. (Computed sample sizes are rounded up to the nearest integer). Thus, it takes a 44% larger sample size in the follow-up experiment than that is required in the initial experiment to achieve a probability .8 of getting a correct conclusion. If $\sigma_B/\sigma_e > .6$, $n_I/n_F \approx 0$, which means it is not possible to find a sample size for the follow-up experiment to have a power of .8.

This case of relative efficiency implies that there is a disadvantage for the researchers who carry out the follow-up experiment. That is, if the follow-up experiment is done at the same sample size as the initial experiment, it has a lower probability to find out the significant result. Researchers need to raise their sample size of the follow-up experiment to compensate the loss of efficiency. As with all sample size determinations, prior knowledge of the size of the variance component will be required to determine the sample size for the follow-up experiment.

2.4 Adjusted P-values and Adjusted Confidence Levels

The two common tools for making inferences about $u_1 - u_2$ are the p-values and confidence intervals, but what does a p-value or a level of confidence in the initial experiment tell us about these values in a follow-up experiment? We will show that small p-values or high levels of confidence are not enough to determine that the results of the initial experiment would carry over to a follow-up experiment. Instead, these must be interpreted in terms of the size of σ_B/σ_e and the estimated value of Δ.

2.4.1 Adjusted P-values

We define the observed effect size to be $\Delta^* = (\overline{y}_1 - \overline{y}_2)/(\sqrt{2}s_p)$, the lower-case letters denoting observed values of the sample means from two treatments. We assume $\Delta^* > 0$ consistent with our assumption that $u_1 > u_2$. The observed $t-stat = \dfrac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{2/n}} = \Delta^*\sqrt{n}$. The two-sided

p-value for the observations is

$$\begin{aligned} & P\left(|t-stat|>\Delta^*\sqrt{n}|Model\ 2.2, u_1-u_2=0\right) \\ =& P\left(\left|NCT\sqrt{1+n\sigma_B^2/\sigma_e^2}\right|>\Delta^*\sqrt{n}|Model\ 2.2, u_1-u_2=0\right) \end{aligned} \tag{2.24}$$

where NCT has a non-central t-distribution with $df = 2(n-1)$ and non-centrality parameter u. Under the null hypothesis, $u = 0$, NCT has a central t-distribution, with $df = 2(n-1)$. Then we have

$$\begin{aligned} & P\left(\left|NCT\sqrt{1+n\sigma_B^2/\sigma_e^2}\right|>\Delta^*\sqrt{n}|Model\ 2.2, u_1-u_2=0\right) \\ =& 2\left(1-G_{df,u=0}\left(\Delta^*\sqrt{n}/\sqrt{1+n\sigma_B^2/\sigma_e^2}\right)\right) \end{aligned} \tag{2.25}$$

We call this *the adjusted p-value*.

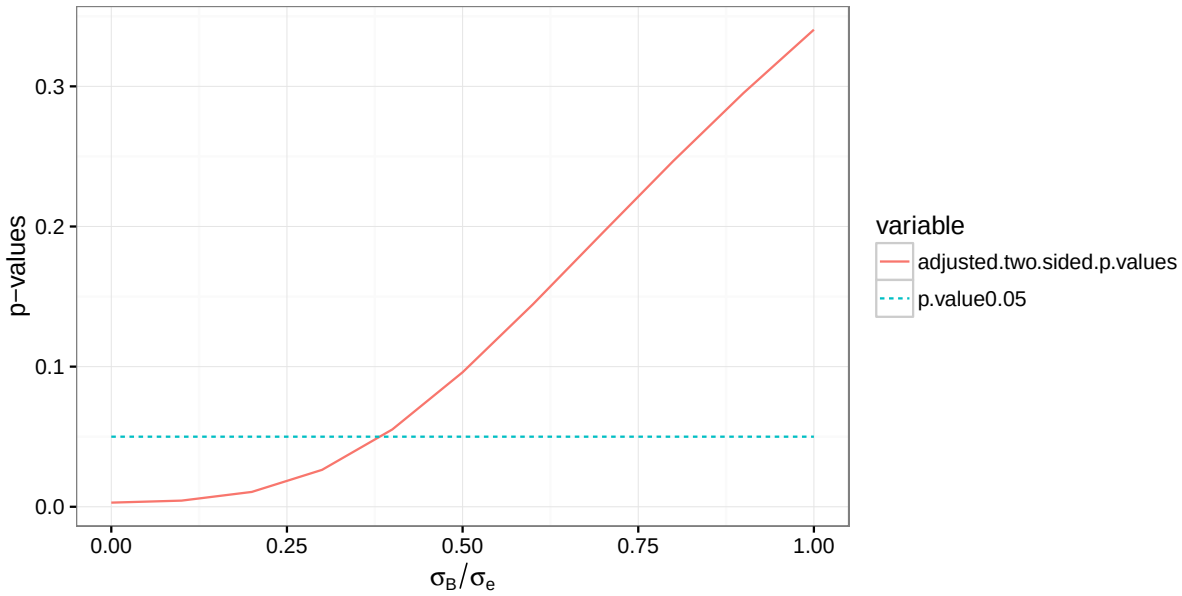

Figure 2.5: *Adjusted two-sided P-value vs* σ_B/σ_e, *n=11,* $\Delta^* = 1.02$

Figure 2.5 and Figure 2.6 show how the p-value of the initial experiment has to be adjusted in Examples 1 and 2, if Δ^* were to be observed in the follow-up experiment. Figure 2.5 shows that if $\sigma_B/\sigma_e < .38$, the inference can be made that the treatment effect would be significant

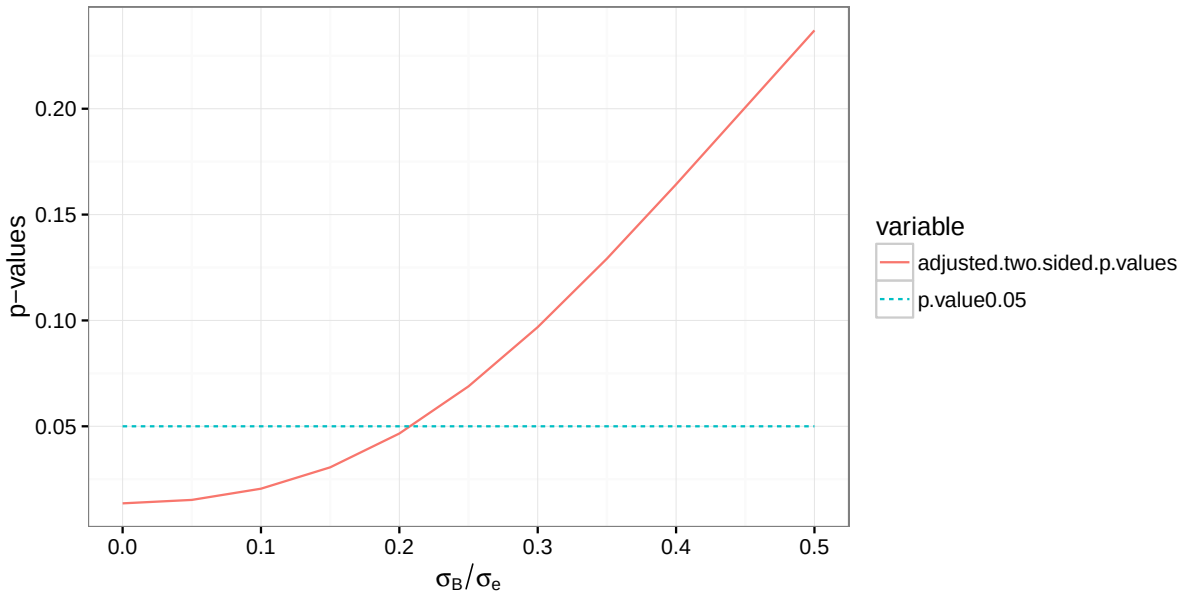

Figure 2.6: *Adjusted two-sided P-value vs* σ_B/σ_e, *n=15,* $\Delta^* = 0.68$

at the 5% level in the follow-up experiment for Example 1, where the $p - value = 0.003$ in the initial experiment indicates strong evidence for a treatment effect in the environment in which the experiment was conducted. However if $\sigma_B/\sigma_e > .38$, the adjusted p-value would show non-significance at the 5% level. Figure 2.6 shows that if $\sigma_B/\sigma_e > .21$, there is not enough evidence to infer that the treatment effect would be significant at the 5% level in the follow-up experiment for Example 2, while the the p-value of 0.014 in the initial experiment also indicates strong evidence for a treatment effect in the environment in which the experiment was conducted. This plot is similar to the p-value profile suggested by Perrett and Higgins (2006) in considering significance in non-replicated experiment.

When the sample size is large, we apply the normal approximation, then we have

$$P(|t - stat| > \Delta^*\sqrt{n}|Model\ 2.2, u_1 - u_2 = 0) \approx 2\left(1 - \Phi\left(\frac{\Delta^*\sqrt{n}}{\sqrt{1 + n\sigma_B^2/\sigma_e^2}}\right)\right) \tag{2.26}$$

If we take the limit of equation 2.26 as $n \to \infty$, the limit, we obtain the *asymptotic adjusted p-value*:

$$\lim_{n\to\infty} P(|t - stat| > \Delta^*\sqrt{n}) = 2\left(1 - \Phi\left(\Delta^*/(\sigma_B/\sigma_e)\right)\right) \tag{2.27}$$

For example, if we fix $\sigma_B/\sigma_e = .3$, and if $\Delta^* = .5$, then the limit of the adjusted p-value in a follow-up experiment is .096, and for any environment in which $\sigma_B/\sigma_e > .255$, it is impossible for the follow-up experiment to attain significance at the 5% level. If we increase $\Delta^* = .6$ and keep $\sigma_B/\sigma_e = .3$, the limit is .046. We can see the role that the observed effect size plays in conjunction with the size σ_B/σ_e in Figure 2.7. For small observed effect sizes and large values of σ_B/σ_e, it may be impossible for the asymptotic adjusted p-value to be less than .05. Hence, large effect size is a better indicator of reproducibility than small p-value.

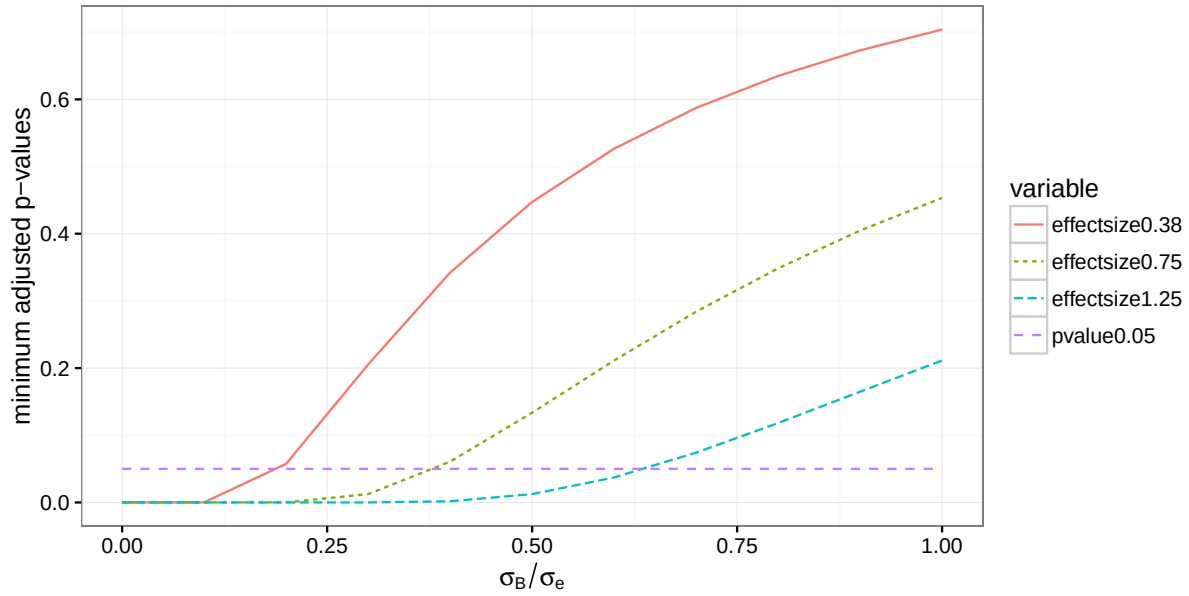

Figure 2.7: *Minimum Adjusted P-value vs* σ_B/σ_e

2.4.2 Adjusted Confidence Levels

Confidence intervals sometimes are more informative in hypothesis tests. Factors like sample sizes, significance levels of test, and the variability of data will affect the width of confidence intervals. For our study here, the random factors in a follow-up experiment will have an effect on the confidence intervals. In the other words, a confidence interval for the initial experiment and one for its follow-up will not have same level of confidence for $u_1 - u_2$ because of the existence of σ_B. Since the conditional mean of $\overline{Y}_1 - \overline{Y}_2$ in the follow-up experiment given θ, δ_1, and δ_2 is $(u_1 + \delta_1) - (u_2 + \delta_2)$, the confidence interval, $(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2,df} S_p \sqrt{2/n}$,

is biased by $\delta_1 - \delta_2$ as an interval for $u_1 - u_2$. If the level of confidence for an interval in the initial experiment is $1-\alpha$, based on the Equation 2.11, the true level of confidence $1-\alpha^*$ for the difference $u_1 - u_2$ in the follow-up experiment can be derived as follows:

$$
\begin{aligned}
& P\left(\left|\frac{(\overline{Y}_1 - \overline{Y}_2) - (u_1 - u_2)}{s_p\sqrt{2/n}}\right| \geq t_{\alpha/2}\right) \\
=& P\left(\left|NCT\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right| \geq t_{\alpha/2}\right) \\
=& P\left(|NCT| \geq t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \\
=& P\left(NCT \geq t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) + P\left(NCT \leq -t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \\
=& 1 - G_{df,u=0}\left(t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) + G_{df,u=0}\left(-t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \\
=& 1 - G_{df,u=0}\left(t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) + 1 - G_{df,u=0}\left(t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) \\
=& \alpha^*
\end{aligned}
\tag{2.28}
$$

Note, $u_1 - u_2 = 0 \rightarrow u = 0$. So NCT has a central t-distribution.

Based on this, we have

$$
1 - \alpha^* = 2G_{df,u=0}\left(t_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right) - 1 \tag{2.29}
$$

which is called *the adjusted confidence level*. When the sample size is large, apply the Normal Approximation, then we have

$$
1 - \alpha^* \approx 2\left(\Phi\left(z_{\alpha/2}/\sqrt{1 + n\sigma_B^2/\sigma_e^2}\right)\right) - 1 \tag{2.30}
$$

Note: For a central t-distribution, we denote its *cdf* as $G_{df,u=0}$. Then

$$
G_{df,u=0}(-t) = P\left(\frac{Z}{\sqrt{V/df}} \leq -t\right) = P\left(\frac{-Z}{\sqrt{V/df}} \geq t\right) = P\left(\frac{Z}{\sqrt{V/df}} \geq t\right) = 1 - G_{df,u=0}(t)
$$

and vice versa, which is based on the fact that Z and $-Z$ have the same distribution.

In Example 2 with $n = 15$, and $\alpha = .05$, the adjusted level of confidence for the initial 95% confidence interval is smaller than .95 by an amount determined by σ_B/σ_e. For instance, it is .88 for $\sigma_B/\sigma_e = .2$ and it is .64 for $\sigma_B/\sigma_e = .5$. As Figure 2.8 shows, the adjusted confidence level approaches 0 as $\sigma_B/\sigma_e \to \infty$.

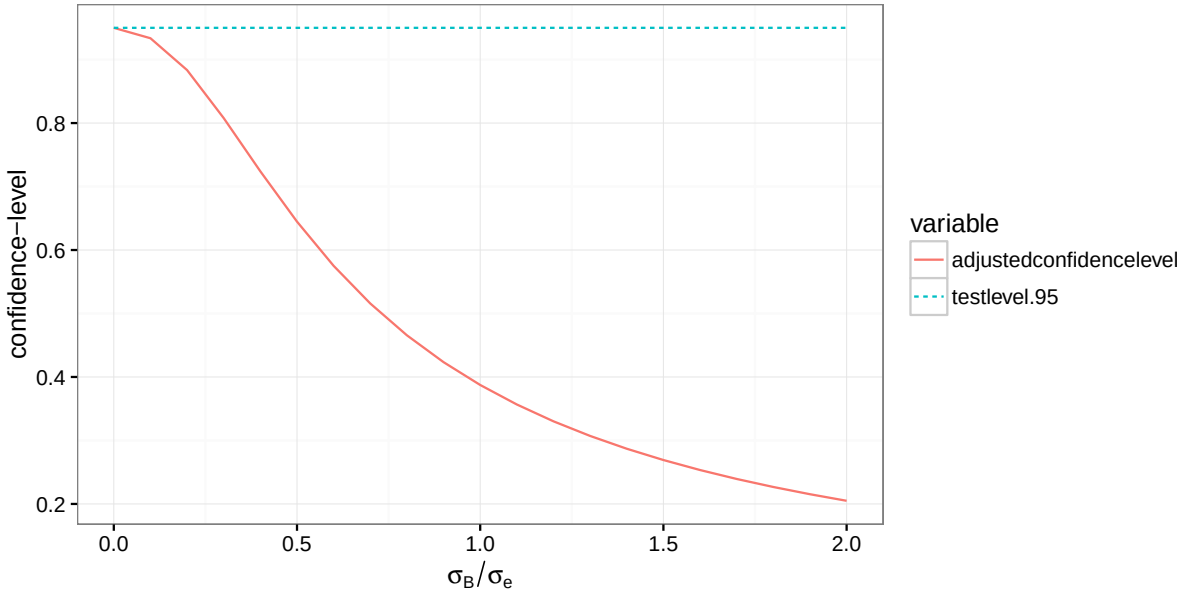

Figure 2.8: *Adjusted confidence level vs* σ_B/σ_e*, n=15,* $\Delta^* = 0.68$

2.4.3 Estimating and Interpreting σ_B/σ_e and Δ

The standard deviation ratio σ_B/σ_e and the effect size Δ play key roles in the reproducibility of research results. Having estimates of Δ and σ_B/σ_e will help researchers identify those results that are likely to carry over to other environments. Both σ_e and Δ can be estimated from the initial experiment, but that is not the case with σ_B. However,one might expect that similar experiments would have similar values of σ_B/σ_e. The following examples will give an indication of values of σ_B/σ_e that could occur in practice.

Example 4: An estimate of σ_B/σ_e

Even for an experiment at one facility, there is often variability in the environments that is accounted for in the design of the experiment by blocking. If the treatments appear more than once in each block, then the usual RCB analysis with fixed effect "treatment", random effects "block" and "block*treatment" would give us estimates of the desired components of variance σ_B^2 and σ_e^2. We used SAS PROC MIXED to analyze data in a randomized block design from Snedecor and Cochran (1989) in which there are 3 treatments measured each of 4 times in each of 5 blocks. The data are the number of wireworms in soil samples treated with either one of two fumigants or a control. We attained the estimates of the components of variance as shown in Table 2.2 and computed the estimate of σ_B/σ_e to be .65.

Covariance Parameter Estimates				
Cov Parm	Estimate	Se	Z Value	Pr>Z
blk	1.1052	2.4502	0.45	0.326
blk*trt($=\sigma_B^2$)	3.8559	3.1035	1.24	0.107
Residual($=\sigma_e^2$)	9.1056	1.9196	4.74	<.0001

Table 2.2: *Estimates of the components of variance for Example 4*

Relationship of σ_B/σ_e to Interclass Correlation

If we randomly select an environment then take observations according to Model 2.2, the observations within treatment are correlated because the random term $\theta + \delta_i$ is common to all the observations within the treatment. The correlation, called the interclass correlation, is given by

$$\rho = (\sigma_\theta^2 + \sigma_B^2)/(\sigma_\theta^2 + \sigma_B^2 + \sigma_e^2). \tag{2.31}$$

The smallest which this can be is when $\sigma_\theta^2 = 0$, so $\rho \geq \sigma_B^2/(\sigma_B^2 + \sigma_e^2)$. It follows that

$$\sigma_B/\sigma_e \leq \sqrt{\rho/(1-\rho)}. \tag{2.32}$$

Example 5: A bound on σ_B/σ_e

Perrett and Higgins (2006) estimated ρ for eight cultivars inoculated with spider mites in four greenhouses which are the environments for this example. The values ranged from 0 to .30 with a median of .12. If $\rho = .30$, then $\sigma_B/\sigma_e \leq .65$, and if $\rho = .12$, then $\sigma_B/\sigma_e \leq .37$. Even with a lack of statistical estimate of σ_B/σ_e, a bound on its value may be enough to indicate whether the results are likely to be reproducible. We recommend plots of the probability of reproducibility vs σ_B/σ_e as the Figure 2.1 or plots of the adjusted p-values vs σ_B/σ_e as in Figure 2.5 and Figure 2.6 to gauge the sensitivity of results to changing environments. With a reasonable bound on σ_B/σ_e, the researcher can place bounds on the probability of reproducibility and adjusted p-values.

Example 6: A multi-lab experiment

Kafkafi et al. (2005) proposed a mixed model where the size of interaction can be estimated from multi-laboratory data and an endpoint design is used to calculate the proportions of variance in which the total variance between the mice in a multilaboratory experiment was decomposed into between-genotype variability, between-laboratory variability, genotype×laboratory interaction variability, and within-group variability. The ratio of genotype×laboratory interaction variability and within-group variability is the ratio σ_B/σ_e in our study. We calculate ratios σ_B/σ_e for different endpoints respectively based on the graph of proportion of total variance. As Table 2.3 shows that the ratio σ_B/σ_e ranges from about 0 to .64. This is a multi-lab experiment, and by some extent, we may recognize this as the prior knowledge of the standard deviation ratio σ_B/σ_e under different experimental environments for this kind of experiment. In the future, it may help other researchers make more informed judgments for the reproducibility of results from a single experiment if these values are shown to be reasonable by more multi-lab experiments.

Ratios σ_B/σ_e for different endpoints (Kafkafi et al., 2005)			
Endpoint(Response)	σ_B/σ_e	Endpoint(Response)	σ_B/σ_e
lingering time	0.640	distance traveled	0.577
segment max speed	0.389	excursions	0.354
time for tum	0.146	radius of tum	0.316
segment length	0.513	center time	0.399
progression segments	0.149	segment acceleration	0.369
homebase occupancy	0	lingering mean speed	0.337
diversity	0	stops per excursion	0.155
lingering spatial spread	0.177	relative activity decrease	0
latency to half max speed	0		

Table 2.3: *A multi-lab experiment*

Interpreting Effect Sizes

The effect size is a measure of the separation between the distributions of the two treatments in the following sense. Let $m = (u_1 + u_2)/2$ be the midpoint between the means of the distributions of the two treatments. Let $p/2$ denote the upper-tail probability above m for the distribution with the smaller mean and equivalently the lower-tail probability below m for the distribution with the larger mean. The sum of these two probabilities p is what we define to be the overlap probability between the two distributions. The effect size can be expressed in terms of p as $\Delta = \sqrt{2}z_{p/2}$.

Table 2.4 shows overlap probabilities for effect sizes from .20 to 2.5. It also shows the asymptotic adjusted p-value $2(1 - \Phi(\Delta^*/(\sigma_B/\sigma_e)))$ for these effect sizes and for σ_B/σ_e set at either .25 or .75. For instance, for $\sigma_B/\sigma_e = .25$, an observed effect size $\Delta^* = .5$ has an asymptotic adjusted p-value of .05, but it would take $\Delta^* = 1.5$ for this to happen with $\sigma_B/\sigma_e = .75$.

Table 4: Overlap Probabilities, Min Adjusted P-values vs. Δ^*					
Observed effect size Δ^*	0.20	0.50	1.00	1.50	2.50
Overlap probability	0.89	0.72	0.48	0.29	0.08
Min adi p-value,$\sigma_B/\sigma_e = .25$	0.42	0.05	$<.01$	$<.01$	$<.01$
Min adi p-value,$\sigma_B/\sigma_e = .75$	0.79	0.50	0.18	0.05	$<.01$

Table 2.4: *Overlap probabilities for effect sizes from .20 to 2.5*

We define the effect size as $\Delta = (u_1 - u_2)/(\sigma_e\sqrt{2})$. Cohen's $d = \Delta\sqrt{2}$, is a measure of

effect size for comparison of two means. Cohen (1988) described effect sizes of .2, .5, and .8 as "small", "medium", and "large" respectively which are correspond to $\Delta = .14$, .35, and .57. Cohen (1988) also knowledged the danger of using terms like these. For different research domains, the benchmarks of d will enable a direct comparative experimental results. For domains like social science study, psychological, educational, or behavioral study, it is not always possible to get a significant result along with a large effect size, which contributes to the lack of reproducibility. Lipsey and Wilson (1994) listed the effect sizes (calculated in Cohen's d) of 302 studies about psychological, educational, and behavioral interventions from literature review, in which more than 62% of those studies have a $d \leq .54$, and the mean of effect sizes of all those studies is well below .54, both of which correspond to $\Delta \leq .38$. Coe (2002) thought the range of circumstances in which the effects of these 302 studies had been found might be limited. From Table 2.4 with $\sigma_B/\sigma_e = .75$, we see that none of the asymptotic adjusted p-value are less than .05. From Figure 2.7, we see that when $\sigma_B/\sigma_e = .18$, we need a Δ as big as at least .38 in order to get the asymptotic adjusted p-value less than .05, and if $\sigma_B/\sigma_e > .18$, a $\Delta = .38$ will never yield a significant asymptotic adjusted p-value. Since the effect sizes from research fields like social science, psychological, educational, or behavioral research might most of the time are not large enough to overcome the type of interaction that can occur in experiment, perhaps that is one of the reasons for lack of reproducibility in these fields.

Chapter 3

Summary and Conclusions

Since S.N.Goodman (1992) proposed the idea of using the replication probability to define the probability of repeating a statistically significant result, terms like reproducibility, replicability, and reliability have been used almost interchangeably (Goodman et al., 2016). Broadly, reproducibility means the results of scientific research should be able to withstand independent validation (Bello and Renter, 2018).

For our research purpose, we define the probability of reproducibility as the probability that the follow-up experiment yields a significant result, assuming the initial experiment yields a significant result. By considering the effect of changing experimental environments on this probability, we analyze how the reproducibility depends not only on the sample size and significance level, but also on the effect size and random environmental factors. Researchers would never want their results only to apply to their certain research facility at a certain time. They always want their results to apply broadly in different environment at different time. However, in a strict sense, statistical significance only applies to the environment in which the experiment is carried out. Environmental factors, either natural or those caused by the way in which the research is conducted, will unavoidably affect the outcome of the experiment and influence the reproducibility of the experiment. In order to extend conclusions to follow-up experimental environments which are different from the initial experiment, researchers must account for these environmental factors. We have proposed three measures which will help

the researchers to determine whether the results of an experiment in one environment may be reproduced by others: the probability of reproducibility, the adjusted p-value, and the adjusted confidence level. The environmental effect ratio—the size of the standard deviation ratio of random environmental factors and experimental error—and the effect size are key factors in determining whether a result from one environment is reproducible in another, and we have indicated situations in which reproducibility of experimental results can be expected and those that cannot.

A p-value or statistical significance does not measure the size of an effect or the importance of a result. Smaller p-values do not necessarily imply the presence of the larger or more important effects, and larger p- values do not imply a lack of importance or even lack of effect. If the sample size or measurement precision is high enough, small effect can yield a small p-value, but with small sample sizes or imprecise measurement, even large effects may produce large p-values (Wasserstein and Lazar, 2016). Large sample sizes and small significance levels from initial experiments are not necessarily strong evidence for the wide application of the results. If effect sizes are small, the probability of reproducibility may be unsatisfied, and the adjusted p-values may fall above the traditional .05 level even if the p-value of the initial experiment is highly significant.

When the same experiment is carried out in a different environment, the error term not only includes the random error within a given experiment, but it also includes the additional sources of variability that are introduced by conducting the same experiment in different environments. It is often impractical to repeat an experiment in different environments to determine significance across environments despite that being the right thing to do.

If one experiment is done in just one environment, a researcher may make use of prior knowledge or expert judgment to place bounds on σ_B/σ_e, which is a measure of the effect of the environment on reproducibility. Smaller values of σ_B/σ_e favor reproducibility of results, so researchers should control the size of σ_B/σ_e by reducing the size of σ_B in every possible way. This requires good experimental design, rigorous implementation of statistical principles and standard practice of methods, and good expertise of personnel (Bello and Renter, 2018). If the experiment will be conducted with a long timetable, strict consistency of the experimental

protocol needs to be implemented as unexpected errors may propogate over time. However, there are often uncontrollable elements that will affect the reproducibility even if researchers have done their best to control the controllable factors. Effect sizes can help the researcher decide whether results from the initial experiment will likely carry over to other environments. As the Table 2.4 shows, we need an effect size of at least 1.0 in order to feel confident about the reproducibility for values of σ_B/σ_e that occur in practice. Small effect sizes, even if statistically significant, should be viewed with caution. Plots of the probability of reproducibility vs σ_B/σ_e or plots of the adjusted p-value vs σ_B/σ_e can be used to indicate the sensitivity of results to changing environments and to assist in making judgments about the likelihood of reproducibility of results.

Ideally, researchers need to reproduce an experiment several times in randomly selected environments before they can say the results from the initial experiment are reproducible. Results from one experiment should be regarded as preliminary especially if σ_B/σ_e is expected to be large or effect size is small. Where possible, the components of variances should be reported along with means, standard errors, and effect sizes in research results. With the collaboration of research labs, we will have a further knowledge of the standard deviation ratio σ_B/σ_e or the environment by treatment interaction under different experimental environments, and then we can make more informed judgments for the reproducibility of results from a single experiment.

Bibliography

Alexander A. Aarts, Joanna E. Anderson, Christopher J. Anderson, and Peter R. Attridge. Estimating the reproducibility of psychological science. *Science*, 349:Issue 6251, 2015. doi: 10.1126/science.aac4716.

Monya Baker. Is there a reproducibility crisis? *Nature*, 533:452454, 2016.

Jay J. Van Bavel, Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. Contextual sensitivity in scientific reproducibility. *Proc Natl Acad Sci U S A.*, 112(23):6454–6459, 2016. doi: 10.1073/pnas.1521897113.

C. Glenn Begley and John P.A. Ioannidis. Reproducibility in science. *Circulation Research*, 116(1):116–126, 2015. ISSN 0009-7330. doi: 10.1161/CIRCRESAHA.114.303819. URL http://circres.ahajournals.org/content/116/1/116.

Nora M. Bello and David G. Renter. Invited review: Reproducible research from noisy data: Revisiting key statistical principles for the animal sciences. *Journal of Dairy Science*, 101 (7):5679–5701, 2018. doi: 10.3168/jds.2017-13978.

Dennis D. Boos and Leonard A. Stefanski. P-value precision and reproducibility. *The American Statistician*, 65(4):213–221, 2011. doi: 10.1198/tas.2011.10129.

Andrew C. Chang and Phillip Li. Is economics research replicable? sixty published papers from thirteen journals say "usually not". (2015-83), 2015. URL https://EconPapers.repec.org/RePEc:fip:fedgfe:2015-83.

Robert Coe. It's the effect size, stupid: what effect size is and why it is important. *Paper presented at the 2002 Annual Conference of the British Educational Research Association*, 2002. URL http://www.leeds.ac.uk/educol/documents/00002182.htm. AccessedMarch23,2012.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*, volume 2nd Ed. New Jersey: Hillsdale, 1988.

Lazer David, Kennedy Ryan, King Gary, and Vespignani Alessandro. Big data. the parable of google flu: traps in big data analysis. *Science*, 345(6193):148–149, 2014. doi: 10.1126/science.1248506.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016. ISSN 1946-6234. doi: 10.1126/scitranslmed.aaf5027. URL http://stm.sciencemag.org/content/8/341/341ps12.

JP. Ioannidis. Why most published research findings are false. *PLoS Med*, 2:124, 2005.

Seppo E. Iso-Ahola. Reproducibility in psychological science: When do psychological phenomena exist? *Frontiers in Psychology*, 8:879, 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00879. URL https://www.frontiersin.org/article/10.3389/fpsyg.2017.00879.

Valen E. Johnson, Richard D. Payne, Tianying Wang, and Alex Asher Soutrik Mandal. On the reproducibility of psychological science. *Journal of the American Statistical Association*, 112, 2017.

Neri Kafkafi, Yoav Benjamini, Anat Sakov, Greg I. Elmer, and Ilan Golani. Genotype–environment interactions in mouse behavior: A way out of the problem. *Proceedings of the National Academy of Sciences*, 102(12):4619–4624, 2005. ISSN 0027-8424. doi: 10.1073/pnas.0409554102. URL http://www.pnas.org/content/102/12/4619.

Robert M. Kaplan and Veronica L. Irvin. Likelihood of null effects of large nhlbi clinical trials has increased over time. *PLOS ONE*, 10:1–12, 08 2015. URL https://doi.org/10.1371/journal.pone.0132382.

Mark W. Lipsey and D.B. Wilson. The efficacy of psychological, educational, and behavioral

treatment. *American Psychologist*, 48(12):1181–1209, 1994. doi: 10.1037//0003-066X.48.12.1181.

Kevin Mullane, Michael J.Curtis, and MichaelWilliams. *Research in the Biomedical Sciences*. Academic Press, 2018. ISBN 978-0-12-804725-5. doi: https://doi.org/10.1016/B978-0-12-804725-5.00001-X. URL http://www.sciencedirect.com/science/article/pii/B978012804725500001X.

The InterAcademy Partnership. A call for action to improve the reproducibility of biomedical research. 2016. URL http://www.interacademies.org/File.aspx?id=30940.

Jamis J. Perrett and James J. Higgins. A method for analyzing unreplicated agricultural experiments. *Crop science*, 46(6):2482–2485, 2006.

Florian Prinz, Thomas Schlange, and Khusru Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10:712, 2011. doi: 10.1038/nrd3439-c1. URL http://dx.doi.org/10.1038/nrd3439-c1.

Mark Richard, Dawn Thilmany, McLellan Patsy, and BrannonAdriana Campa. Reproducibility and rigor in rees portfolio of research. 09 2016. doi: 10.13140/RG.2.2.24369.17761.

L. Shenhav, R. Heller, and Y. Benjamini. Quantifying replica-bility in systematic reviews: the r- value. *Technical Report,Tel-Aviv University*, 2015.

Gary W. Small, Prabha Siddarth, Zhaoping Li, Karen J. Miller, Linda Ercoli, Natacha D. Emerson, Jacqueline Martinez, Koon-Pong Wong, Jie Liu, David A. Merrill, Stephen T. Chen, Susanne M. Henning, Nagichettiar Satyamurthy, Sung-Cheng Huang, David Heber, and Jorge R. Barrio. Memory and brain amyloid and tau effects of a bioavailable form of curcumin in non-demented adults: A double-blind, placebo-controlled 18-month trial. *The American Journal of Geriatric Psychiatry*, 26(3):266 – 277, 2018. ISSN 1064-7481. doi: https://doi.org/10.1016/j.jagp.2017.10.010. URL http://www.sciencedirect.com/science/article/pii/S1064748117305110.

George W. Snedecor and William G. Cochran. *Statistical methods.* Iowa State University Press, 8th edition, 1989.

S.N.Goodman. A comment on replication, p-values and evidence. *Statistics in Medicine*, 11: 875–879, 1992.

Ronald L. Wasserstein and Nicole A. Lazar. The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. doi: 10.1080/00031305.2016.1154108.

Deborah A. Zarin, Tony Tse, Pharm.D. Rebecca J. Williams, Robert M. Califf, and Nicholas C. Ide. The clinicaltrials.gov results database update and key issues. *The New England Journal of Medicine*, 364:852–860, 2011. doi: 10.1056/NEJMsa1012065.

Appendix A

Code

```
[language=R]
#Figure 2.1:non-central t-distribution to compute probability of reproducibility
n=11, effectsize=1.0,alpha=0.05,vs. ratio of variances(0-2.0)
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 1.0,11,  alpha = .05)
tupperrej2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  lower.tail = FALSE, log.p = FALSE)
}
#experiment2 lower tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 1.0,11,  alpha = .05)
tlowerrej2 = function(param){
  ratio = param[1]
```

```
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(-reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  log.p = FALSE)
}

#experiment2 not rejection
utrej2 = apply(params2t,1, tupperrej2)
ltrej2 = apply(params2t,1, tlowerrej2)
notrej2 = 1-utrej2-ltrej2
round(cbind(utrej2,ltrej2,notrej2),digits = 4)
x<-c(seq(0,2.0,0.1))
reproducibility.prob<-utrej2
wrongdirec.repro.prob<-ltrej2
nonsig.prob<-notrej2
df <- data.frame(x,reproducibility.prob, wrongdirec.repro.prob, nonsig.prob)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot1.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot1<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("probability")
plot1 + theme_bw()
```

```
dev.off()

#Figure 2.2:non-central t-distribution to compute probability of reproducibility
#n=15,effectsize=0.68,alpha=0.05,vs. ratio of variances(0-2.0)
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.68, 15,  alpha = .05)
tupperrej2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)

  pt(reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  lower.tail = FALSE, log.p = FALSE)
}
#experiment2 lower tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.68, 15,  alpha = .05)
tlowerrej2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(-reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  log.p = FALSE)
}
utrej2 = apply(params2t,1, tupperrej2)
ltrej2 = apply(params2t,1, tlowerrej2)
```

```
notrej2 = 1-utrej2-ltrej2
round(cbind(utrej2,ltrej2,notrej2),digits = 4)
x<-c(seq(0,2.0,0.1))
reproducibility.prob<-utrej2
wrongdirec.repro.prob<-ltrej2
nonsig.prob<-notrej2
df <- data.frame(x,reproducibility.prob, wrongdirec.repro.prob, nonsig.prob)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot2.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot2<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("probability")
plot2 + theme_bw()
dev.off()

#Table 2.1: n=500,effectsize=0.2/0.8,alpha=0.05,vs. ratio of variances=0.5
#experiment2 upper tail rejection
params2t=expand.grid(c(0,0.5), c(0.2,0.8),500,  alpha = .01)
tupperrej2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
```

```
  pt(reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  lower.tail = FALSE, log.p = FALSE)
}
#experiment2 lower tail rejection
params2t=expand.grid(c(0,0.5), c(0.2,0.8),500,  alpha = .01)
tlowerrej2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(-reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  log.p = FALSE)
}
utrej2 = apply(params2t,1, tupperrej2)
ltrej2 = apply(params2t,1, tlowerrej2)
notrej2 = 1-utrej2-ltrej2
power=utrej2 +ltrej2
round(cbind(params2t,power,utrej2,ltrej2,notrej2),digits = 4)

#Figure 2.3: n=500,effectsize=0.2/0.8,alpha=0.01,vs. ratio of variances=0-2.0
#effectsize=0.2
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.2,500,  alpha = .01)
tupperrej0.2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
```

```
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)

  pt(reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  lower.tail = FALSE, log.p = FALSE)
}
#experiment2 lower tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.2,500,  alpha = .01)
tlowerrej0.2 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(-reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  log.p = FALSE)
}
RP0.2 = apply(params2t,1, tupperrej0.2)
wrongdirectionRP0.2 = apply(params2t,1, tlowerrej0.2)
#effectsize=1.0
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.8,500,  alpha = .01)
tupperrej0.8 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
```

```
  lower.tail = FALSE, log.p = FALSE)
}
#experiment2 lower tail rejection
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.8,500,  alpha = .01)
tlowerrej0.8 = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  pt(-reject, df=2*(nsize-1),sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),
  log.p = FALSE)
}
RP0.8 = apply(params2t,1, tupperrej0.8)
wrongdirectionRP0.8 = apply(params2t,1, tlowerrej0.8)
x<-c(seq(0,2.0,0.1))
df <- data.frame(x,RP0.8,wrongdirectionRP0.8,wrongdirectionRP0.2,RP0.2)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot3.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot3<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("probability")
plot3 + theme_bw()
dev.off()
```

```
#Figure 2.4: relative efficiency vs ratios
n0.5<-(qnorm(0.025)+qnorm(0.2))*(qnorm(0.025)+qnorm(0.2))/0.25
ratio0.5<-c(seq(0,0.6,0.1))
effectsize0.5<-0.5
n0.5<-((qnorm(0.2)+qnorm(0.025))^2)/((effectsize0.5^2))
a<-effectsize0.5^2-(qnorm(0.2)^2)*(ratio0.5^2)
b<-2*effectsize0.5*qnorm(0.025)
c<-qnorm(0.025)^2-(qnorm(0.2)^2)
N0.5<-((-b+sqrt(b^2-4*a*c))/(2*a))^2
N0.5
ratio0.5<-n0.5/N0.5
ratio0.5<-c(ratio0.5,rep(0,6))
x<-c(seq(0,1.2,0.1))
effectsize0.5<-ratio0.5
df <- data.frame(x,effectsize0.5)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot4.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot4<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("Relative Efficiency")
plot4 + theme_bw()
dev.off()
```

```
#Figure 2.5: adjusted p-value vs. ratio of variances: n=11, effectsize=1.02.
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,1.0,0.1)), 1.02,11,  alpha = .05)
tupperrej2pvalue = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  2*pt(sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),df=2*(nsize-1),
  lower.tail = FALSE, log.p = FALSE)
}
utrej2pvalues = apply(params2t,1, tupperrej2pvalue)
utrej2pvalues
x<-c(seq(0,1.0,0.1))
adjusted.two.sided.p.values<-utrej2pvalues
p.value0.05<-c(rep(0.05,11))
df <- data.frame(x,adjusted.two.sided.p.values,p.value0.05)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot5.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot5<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("p-values")
plot5 + theme_bw()
```

```
dev.off()

#Figure 2.6: adjusted p-value vs. ratio of variances: n=15, effectsize=0.68.
#experiment2 upper tail rejection
params2t=expand.grid(c(seq(0,0.5,0.05)), 0.68,15,  alpha = .05)
tupperrej2pvalue = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  2*pt(sqrt(nsize)*effectsize/sqrt(1+nsize*ratio*ratio),df=2*(nsize-1),
  lower.tail = FALSE, log.p = FALSE)
}
utrej2pvalues = apply(params2t,1, tupperrej2pvalue)
utrej2pvalues
x<-c(seq(0,0.5,0.05))
adjusted.two.sided.p.values<-utrej2pvalues
p.value0.05<-c(rep(0.05,11))
df <- data.frame(x,adjusted.two.sided.p.values,p.value0.05)
head(df)

library(ggplot2)
library(tidyr)
pdf("rplot6.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot6<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
```

```
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("p-values")
plot6 + theme_bw()
dev.off()
#Figure 2.7: minimum of adjusted p-values
params2=expand.grid(c(seq(0,1.0,0.1)), 0.38)
ureject2= function(param){
  ratio = param[1]
  effectsize = param[2]
  reject = param[3]
  2*(1-pnorm(effectsize/ratio))
}
effectsize0.38 = apply(params2,1, ureject2)

params2=expand.grid(c(seq(0,1.0,0.1)), 0.75)
ureject2= function(param){
  ratio = param[1]
  effectsize = param[2]
  reject = param[3]
  2*(1-pnorm(effectsize/ratio))
}
effectsize0.75 = apply(params2,1, ureject2)
params2=expand.grid(c(seq(0,1.0,0.1)), 1.25)
ureject2= function(param){
  ratio = param[1]
  effectsize = param[2]
  reject = param[3]
  2*(1-pnorm(effectsize/ratio))
}
```

```
effectsize1.25 = apply(params2,1, ureject2)
x<-c(seq(0,1.0,0.1))
pvalue0.05<-c(rep(0.05,11))
df <- data.frame(x,effectsize0.38,effectsize0.75,effectsize1.25,pvalue0.05)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot10.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot10<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("minimum adjusted p-values")
plot10 + theme_bw()
dev.off()

#confidence intervals
2*(pnorm(qnorm(0.975)))-1
2*(pnorm(qnorm(0.975)/(sqrt(1+15*0.04))))-1
2*(pnorm(qnorm(0.975)/(sqrt(1+15*0.25))))-1

#confidence intervals use t-distribution
params2t=expand.grid(c(0,0.2,0.5), 0.68,15,  alpha = .05)
tupperrej2pvalue = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
```

```
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  2*pt(reject,df=2*(nsize-1),lower.tail = TRUE, log.p = FALSE)-1
}
utrej2pvalues = apply(params2t,1, tupperrej2pvalue)
utrej2pvalues

#Figure 2.7: adjusted confidence level
params2t=expand.grid(c(seq(0,2.0,0.1)), 0.68,15,  alpha = .05)
tupperrej2pvalue = function(param){
  ratio = param[1]
  effectsize = param[2]
  nsize = param[3]
  newalpha = 1-param[4]/2
  reject = qt(newalpha,df = 2*(nsize-1))/sqrt(1+nsize*ratio*ratio)
  2*pt(reject,df=2*(nsize-1),lower.tail = TRUE, log.p = FALSE)-1
}
adjustedconfidencelevel = apply(params2t,1, tupperrej2pvalue)
adjustedconfidencelevel
x<-c(seq(0,2.0,0.1))
confidentlevel.95<-c(rep(0.95,21))
df <- data.frame(x,adjustedconfidencelevel,confidentlevel.95)
head(df)
library(ggplot2)
library(tidyr)
pdf("rplot7.pdf", height=4, width=8)
xy <- gather(df, key = variable, value = value, -x)
plot7<-ggplot(xy, aes(x = x, y = value, color = variable)) +
  geom_line(aes(linetype=variable))+
  scale_size_area() +
```

```
  xlab(expression(paste(sigma[B]/sigma[e]))) +
  ylab("confidence-level")
plot7 + theme_bw()
dev.off()

# Overlap probabilities for effect size of Table 2.4
(1-pnorm(c(0.2,0.5,1.0,1.5,2.5)/sqrt(2)))*2

#SAS code for table 2.2
DATA FUMIGANT;
INPUT trt $ blk $ rep1 rep2 rep3 rep4;
rep=1; WORMS=rep1; OUTPUT;
rep=2; WORMS=rep2; OUTPUT;
rep=3; WORMS=rep3; OUTPUT;
rep=4; WORMS=rep4; OUTPUT;
keep trt blk WORMS rep;
DATALINES;
C I 5 4 5 2
C II 0 9 3 3
C III 4 4 3 9
C IV 7 3 5 12
C V 4 9 8 6
S I 5 5 1 2
S II 6 4 5 4
S III 2 9 3 7
S IV 6 4 8 4
S V 2 9 7 3
O I 12 20 8 8
O II 7 4 4 5
```

```
O III 9 6 7 11
O IV 12 22 17 13
O V 7 8 5 9
;
RUN;
proc print;
run;
PROC MIXED data= FUMIGANT covtest;
CLASS trt blk rep;
model WORMS =trt;
random blk blk*trt;
run;
```