Unveiling pitfalls and exploring alternatives in the use of pilot studies for sample size estimation

by
Chris Ji

A Research Paper
presented to the University of Waterloo
in partial fulfillment of the requirements for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada
April 5, 2024

Contents

Abstract

Pilot studies are used to estimate effect sizes, which in turn are used in power calculations to determine the sample size needed for the main study to achieve a prespecified power and significance level. In this paper we explore the pitfalls of using small pilot studies to perform these estimates. Additionally, we examine three alternatives to determine a sufficient sample size needed for the main study, the corridor of stability, which utilizes bootstrapping to determine a sample size at which the estimate of the effect size will become stable, as well as two Bayesian metrics, the average coverage criterion, and the average length criterion, which involve controlling statistics based on the posterior distribution of the effect size. All three of these metrics are more robust than current methods to determine sample sizes and effect sizes from small pilot studies. Both Bayesian metrics are unaffected by sample size, and hence may be able to bypass the need for pilot studies altogether.

keywords: Effect Size, Pilot Study, Power Analysis, Sample Size, Simulation

1 Introduction

1.1 Background

Pilot studies serve as invaluable precursors to clinical trials, with multifaceted purposes. Firstly, pilot studies help researches assess various dimensions of a study's feasibility, such as the process, recruitment, randomization, and overall scientific validity. It can also help assess various logistical concerns, such as available resources, management strategies, implementation, and retention of the study (Leon et al., 2010). Secondly, and of more interest to statisticians, pilot studies play a pivotal role in estimating a treatment's effect size. This estimate can be used to determine whether a larger scale study should even be performed, and if so, the estimate is required to determine the sample size needed to achieve a certain power and significance level in such a study. In the interest of simplicity, we shall be assuming a randomized clinical trial with 2 groups of equal size, but it should be noted that the results shown here generalize to more complicated trials.

Effect size is commonly measured through Cohen's d (Cohen, 1977), which is defined as

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s},$$

where $\bar{x}_i$ is the mean of group i and $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ is the pooled variance, with $s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{i,j} - \bar{x}_i)^2$ representing the variance within each group. Researchers classify effect sizes as small, medium, or large based on d taking values of $0.2, 0.5$, and 0.8, respectively (Cohen, 1977). Determining

whether or not a treatment is effective corresponds to testing the null hypothesis $H_0 : d = 0$. Accompanying this hypothesis test are two fundamental statistical concepts: significance level and power. Controlling the significance level controls the type I error, the probability of rejecting H_0 when it is true, and controlling the power of a test controls the type II error, the probability of not rejecting H_0 when it is false.

The power of our test is given approximately by (Chow et al., 2007)

$$1 - \beta = \Phi\left(\frac{d\sqrt{n}}{s} - Z_{\alpha/2}\right), \tag{1}$$

where Φ is the cumulative standard normal distribution function, s is the pooled sample standard deviation given above, and $Z_{\alpha/2}$ is the upper $\frac{\alpha}{2}$th quantile of the standard normal distribution. Pilot studies are the most common way to estimate d and σ, and hence calculate n, the sample size required for a test with significance level α and power $1 - \beta$ (other methods include using effect sizes from related studies).

1.2 Literature Review

Many authors, such as Lakens and Evers (2014), have cautioned against using pilot studies to estimate effect sizes. An a-priori power analysis, which is used to calculate sample size needed, assumes there is a true effect of a specified size. This specified size is ideally estimated from a pilot study. However, the low sample sizes of pilot studies lead to estimated effect sizes with considerably large standard errors. Hence, it is quite possible that the true effect size is either severely underestimated, leading to the study

being aborted, or the true effect size is severely overestimated, leading to an underpowered main study.

Lakens and Evers (2014) recommended novel alternatives such as utilizing the corridor of stability, sequential analysis, v statistic, and p curve analysis. Others, such as Biesanz and Schrager (2010), suggest using a Bayesian framework to better incorporate the uncertainty in an effect size's estimate into the sample size estimate. Kelter (2020) also explores Bayesian alternatives to traditional frequentist null hypothesis significance testing (NHST).

The corridor of stability (Schönbrodt and Perugini, 2013) utilizes bootstrap to calculate a sample size such that the magnitude of a correlation can be expected to be stable. Sequential analysis involves shifting the analysis to be done intermittently, as data is collected. The v statistic offers an alternative metric to judge results. The v statistic gives the probability that the model based on the data is more accurate than a random one (Lakens and Evers, 2014). p curve is a meta-analytic technique, which entails analyzing the distribution of p values for a set of studies, and comparing it to a uniform distribution of p values that would correspond to a null effect (Simonsohn et al., 2014).

In this paper we perform some simulations to show the severe variance in sample size estimates that stem from pilot studies. We then adapt the corridor of stability approach for treatment effect sizes, and a couple of Bayesian methodologies to estimate sample size. Furthermore, since normal distributions are quite rare in practice (Micceri, 1989), we also explore the effect

of skewness on these sample size estimates. Through simulation, we show that all of these methods are independent of both effect size and pilot study sample size.

2 Limitations of Traditional Power Calculations

In this section, we outline how pilot studies are currently used to estimate the required main study sample sizes. We also perform some simulation studies that illustrate the wide range of results.

2.1 Normal Distributions

Solving (1) for n, we get that the sample size needed for a study with true effect size d to obtain significance level α and power β is

$$n = \frac{2 \cdot (Z_{\alpha/2} + Z_{1-\beta})^2 \cdot \sigma^2}{d^2}. \tag{2}$$

Pilot studies are used to obtain $\hat{\sigma}$ and $\hat{d}$, estimates of the true σ and d, which are then used to estimate the approximate sample size needed for their main study. However, the problem with this traditional method arises and is clear from this formula: if $\hat{d}$ is too large, it can result in an estimated sample size that is too small for the main study, ultimately leading to an underpowered study. Conversely, if $\hat{d}$ is too small, it can result in a larger-than-necessary sample size, and usually results in the study being aborted due to lack of resources.

To illustrate the magnitude of this problem, a simulation study was performed. 10,000 pilot studies were simulated, with $n = 10, 25$, and 50 for each group. For the first group, data was generated from a $N(0, 1)$ distribution, and for the second group, data was generated from a $N(0.2, 1)$ distribution, yielding in a small effect size ($d = 0.2$). $\hat{\sigma}$ and $\hat{d}$ were then calculated using the equations from 1.1 for each pilot study, and (1) was used to calculate the resulting sample size needed for the main study. The distributions of estimated effect sizes from the pilot studies are shown in Figure 1, with some summary statistics displayed in Table 1. The distribution of n is shown in Figure 2, with a black dot representing the true sample size needed to achieve a power of 0.8 with significance level 0.95 and a small effect size ($n = 393$). Medium and large effect sizes yield similar results.

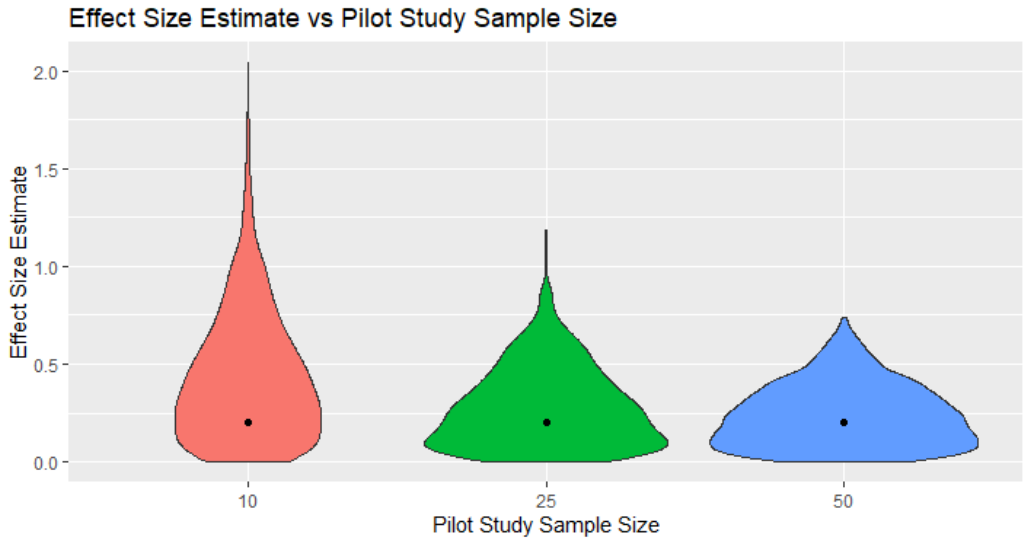

Figure 1: Violin plots depicting estimated effect size for a true small effect size ($d = 0.2$). Effect size was estimated through three different pilot study sample sizes, $n = 10, 25$, and 50. The dot represents the true effect size.

Table 1: Some summary statistics of the estimated effect size.

Pilot Size	Min	Q1	Median	Mean	Q3	max	SD[1]
10	0.25×10^{-4}	0.17	0.36	0.43	0.61	2.43	0.34
25	0.10×10^{-4}	0.12	0.25	0.29	0.42	1.65	0.22
50	0.42×10^{-5}	0.10	0.21	0.24	0.34	0.99	0.16

[1]SD represents the standard deviation of the effect size estimates.

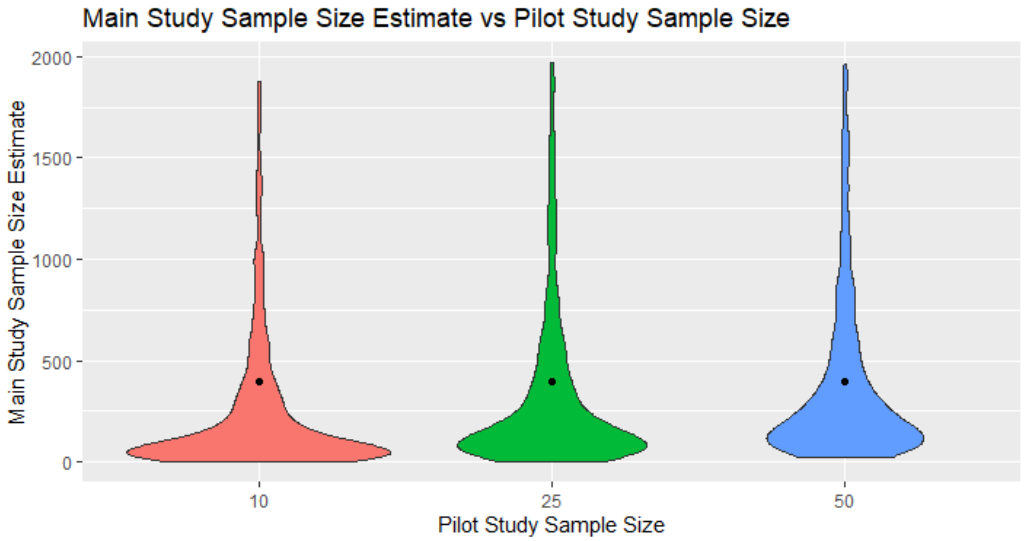

Figure 2: Violin plots depicting estimated sample size for a true small effect size($d = 0.2$). Effect size was estimated through three different pilot study sample sizes, $n = 10, 25$, and 50. The dot represents the true sample size needed, calculated from the true effect size. The vertical axis is capped at $n = 2000$.

Table 2: Some summary statistics of the estimated main study sample size. All observations above 2000 are truncated to 2000.

Pilot Size	Min	Q1	Median	Mean	Q3	m[1]	SD[2]
10	4	43	123	447.6	523	0.13	687.5
25	0	91	248	649.8	1071.5	0.20	745.2
50	0	137	351	747.9	1485.2	0.21	758.9

[1] m represents the proportion of observations greater than 2000.

[2] SD represents the standard deviation of the main study sample size estimates.

Predictably, as the pilot study size increases, the variance of the estimated effect size decreases. This results in a decrease in variance of the estimated sample size needed for the main study. However, for all three

pilot study sample sizes, the sample size needed for the main study is substantially underestimated. As the pilot sample size increases, the average estimated sample size needed in the main study approaches the theoretical sample size needed, but the majority of pilot studies lead to estimates that are too low, and a few lead to estimates that extremely high. Figure 3 and Table 3 show the same simulation, with a large effect size. While the mode sample size estimate is accurate with a larger effect size, the variance is still exceptionally high.

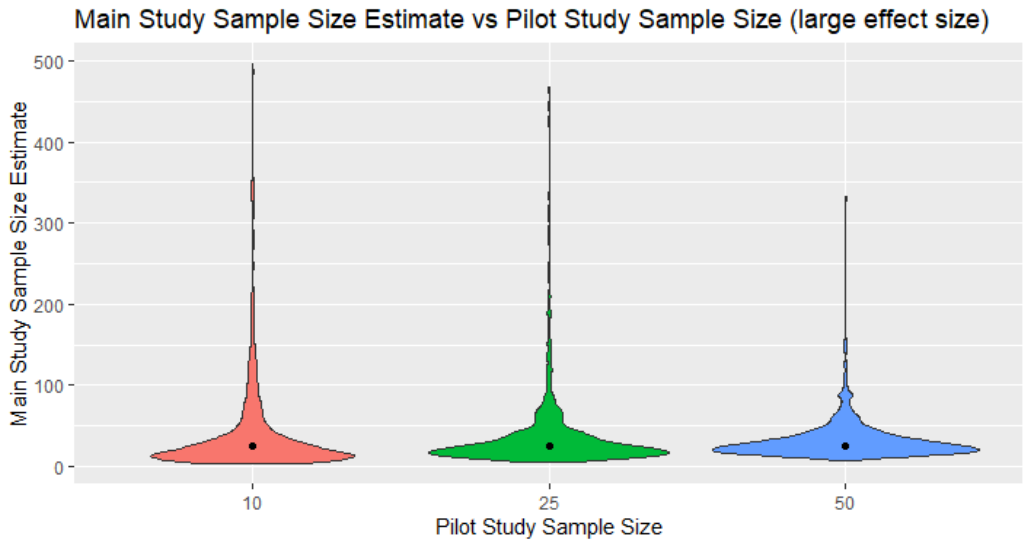

Figure 3: Violin plots depicting estimated sample size for a large effect size ($d = 0.8$). Effect size was estimated through three different pilot study sample sizes, $n = 10, 25$, and 50. The dot represents the true sample size needed, calculated from the true effect size. The vertical axis is capped at $n = 500$.

Table 3: Some summary statistics of the estimated main study sample size. All observations above 2000 are truncated to 2000.

Pilot Size	Min	Q1	Median	Mean	Q3	m[1]	SD[2]
10	3	12	23	75.2	56	626	130.2
25	5	16	24	44.14	42	137	70.0
50	7	19	25	32.88	37	14	31.7

[1] m represents the proportion of observations greater than 2000.

[2] SD represents the standard deviation of the main study sample size estimates.

2.2 Skewed Distributions

We also explore the affect of skewness on the sample size estimates. The lognormal distribution was chosen as an exemplary skewed distribution due to its simplicity. Since the majority of real life datasets have skewness within $(-1, 1)$ (Cain et al., 2016), we simulate datasets with skewness of 0.5 and 1. For a skewness of 0.5, we simulate a lognormal distribution with $\mu = 1.787, \sigma = 0.164$, and for a skewness of 1, we simulate a lognormal distribution with $\mu = 1.083, \sigma = 0.314$. These parameters are chosen so they both have unit variance. Then, $d = 0.2, 0.5, 0.8$ are added to half the observations to represent the treatment group. We then estimate the effect size as above. The distribution of estimated effect sizes for $d = 0.2$, with a skewness of 0.5 and 1 are shown below.

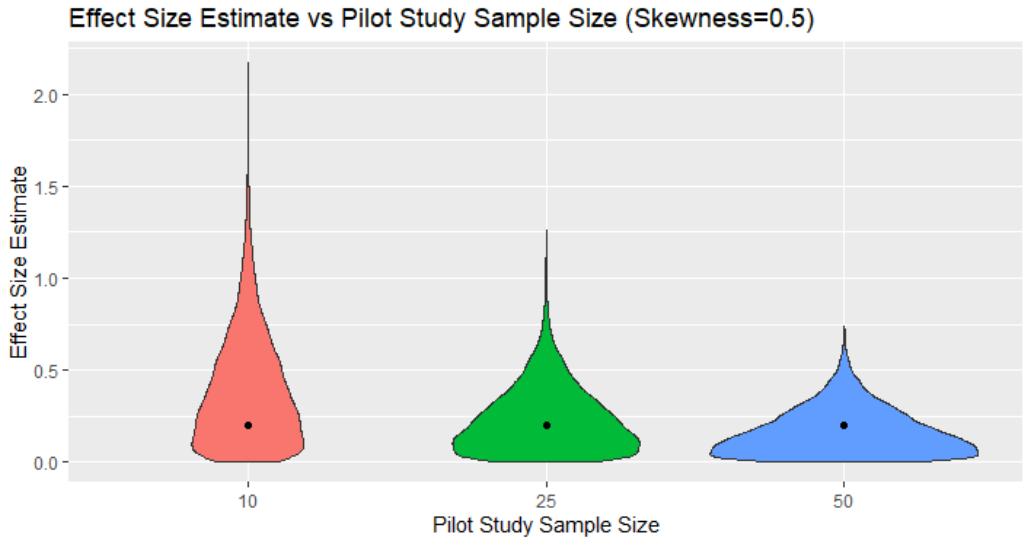

Figure 4: Violin plots depicting estimated effect size for a small effect size ($d = 0.2$), with a log normal distribution with skewness of 1. Effect size was estimated through three different pilot study sample sizes, $n = 10, 25$, and 50. The dot represents the true effect size.

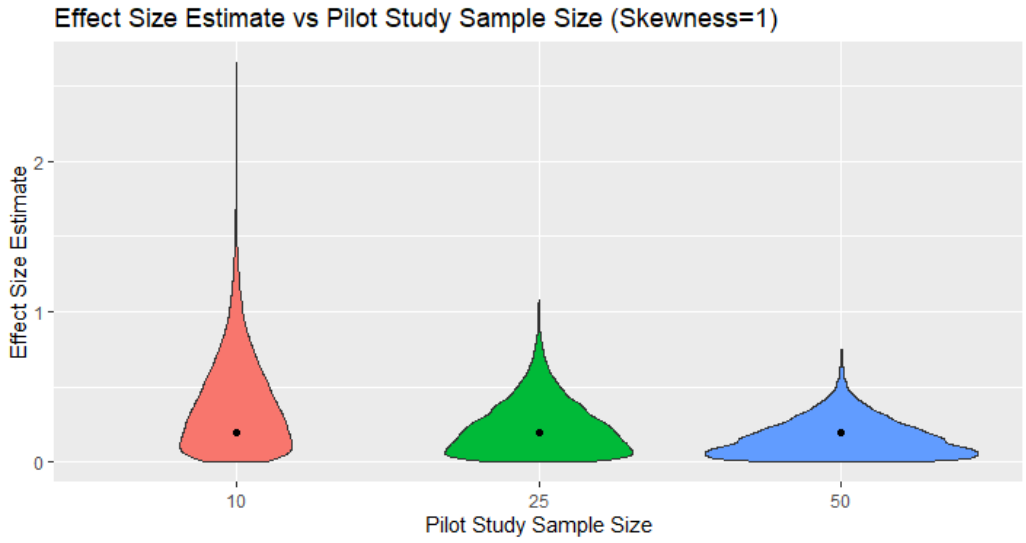

Figure 5: Violin plots depicting estimated effect size for a small effect size($d = 0.2$) , with a log normal distribution with skewness of 0.5. Effect size was estimated through three different pilot study sample sizes, $n = 10, 25$, and 50. The dot represents the true effect size.

Due to way the simulation was set up, the results are identical to the symmetric distributions, resulting in the effect size being a difference between group means. Hence the estimated sample sizes are identical as well.

3 Alternative Approaches to Determining Sample Size

Above, we showed that calculating sample sizes following a purely frequentist framework fails due to the low sample sizes of the typical pilot study. In this section, we explore a novel approach involving bootstrapping (Schönbrodt and Perugini, 2013) in 3.1, and explore some Bayesian methods to estimate

sample size in 3.2 (Chow et al., 2007).

3.1 Corridor of Stability

In the pursuit of more robust methods of determining sample sizes, one interesting approach is the "Corridor of Stability" (COS). The COS introduces a novel perspective on estimating sample sizes by establishing a concept of stability around the true effect size. At its core, the COS is a width w around the true effect size within which variations are considered acceptable.

To implement the COS method, note that from a pilot study with sample size n, we can obtain n estimates of d, which correspond to estimating d using only $1, 2, \ldots, n$ units of the sample. These estimates form a trajectory of the effect size, indicating how it fluctuates as the sample size increases. A critical concept within the COS framework is the "Point of Stability" (POS), which represents the sample size at which the trajectory of effect size estimates no longer exits the COS. In other words, the fluctuations around the true effect size become practically insignificant, and hence our estimate of the effect size is stable.

By bootstrapping from the original pilot study data and repeating this process, we can get a distribution of POS. This distribution captures the uncertainty associated with the determination of the POS. Percentiles of this distribution represent a value for which, for example, 80% of all trajectories are stabilized by at least that point. Hence percentiles are homologous to confidence in the POS. An illustration of this concept is shown in Figure

3, with a small effect size ($d = 0.2$), and a COS defined as ± 0.1 ($w = 0.1$) around the true effect size. Figures 4 and 5 show similar results, with medium ($d = 0.5$) and large ($d = 0.8$) effect sizes, with the same width of $w = 0.1$.

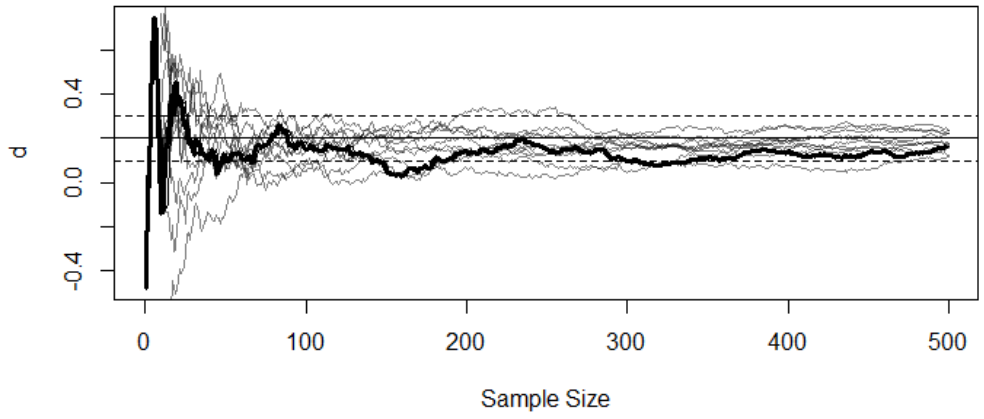

Figure 6: Original (thick black line) and bootstrapped trajectories of estimated effect size. Dashed lines represent the COS of $(0.1, 0.3)$ around the true effect size of $d = 0.2$.

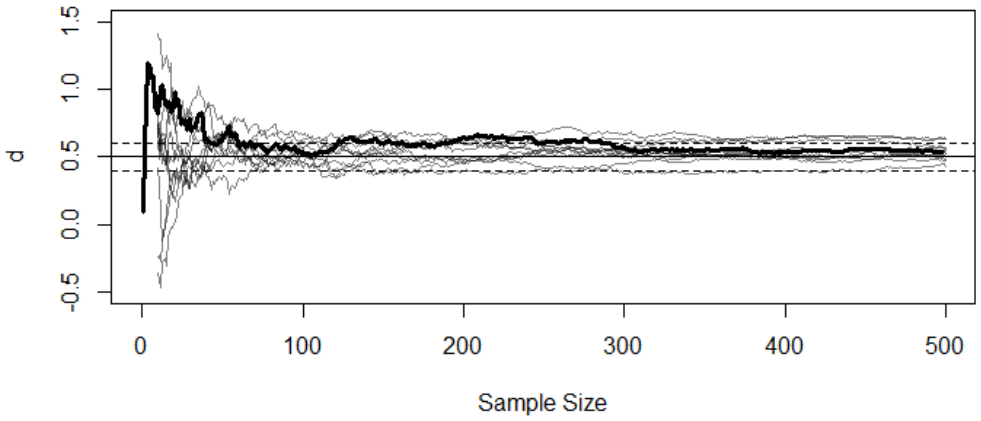

Figure 7: Original (thick black line) and bootstrapped trajectories of estimated effect size. Dashed lines represent the COS of $(0.4, 0.6)$ around the true effect size of $d = 0.5$.

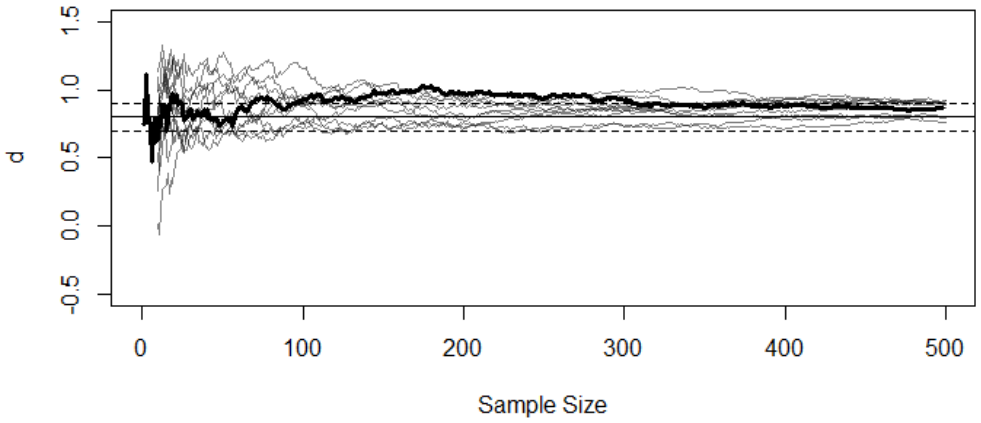

Figure 8: Original (thick black line) and bootstrapped trajectories of estimated effect size. Dashed lines represent the COS of $(0.7, 0.9)$ around the true effect size of $d = 0.8$.

To compute the distribution of POS values, first two Gaussian distributions with 500,000 observations each were generated, the first with $\mu = 0$ and $\sigma^2 = 1$, and the second with $\mu = 0.2, 0.5, 0.8$ and $\sigma^2 = 1$, for the different effect sizes. Then, $B = 1000$ bootstrap samples with $n = 1000$ sample size were drawn. For each bootstrap sample, the effect size was estimated for every sample size n, from 10 to 1000.

Table 2 shows 0.8, 0.9, and 0.95 confidence points of stability for small, medium, and large effect sizes with $w = 0.1$ and $w = 0.15$. As one would expect, required sample size increases with smaller width and greater confidence. However, surprisingly, the true effect size does not effect the sample size at which the estimate of the effect size stabilizes. This encourages the use of the corridor of stability, as, after selecting a desired level of confidence and width of the corridor before calculating the effect size, the result of the effect size will not affect the sample size needed.

Table 4: The points of stability for different widths, levels of confidence, and effect sizes. For each level of confidence, $w = 0.1$ in the left column, and $w = 0.15$ in the right.

	Level of confidence					
d	0.8		0.9		0.95	
0.2	524	237	727	338	908	445
0.5	529	236	742	337	920	447
0.8	563	254	778	363	950	481

3.1.1 Skewed Distributions

The above simulation was then repeated with the log-normal distributions introduced in Section 2.1, to show how the results differ for a skewed data set. Specifically, two log-normal distributions with 1,000,000 observations each were generated, the first with $\mu = 1.787$ and $\sigma = 0.164$, and the second with $\mu = 1.083$ and $\sigma = 0.314$. The first set represents a data set with unit variance and a skewness of 0.5, and the second with unit variance and a skewness of 1. Then, $d = 0.2, 0.5, 0.8$ was added to half of these observations, to represent the second treatment group, and then bootstrap samples were taken and effect sizes were estimated as above. Tables 3 and 4 show the results for the distributions with skewness 0.5 and 1, respectively.

Table 5: The points of stability for different widths, levels of confidence, and effect sizes, for log-normal distribution with skewness of 0.5. For each level of confidence, $w = 0.1$ in the left column, and $w = 0.15$ in the right.

	Level of confidence					
d	0.8		0.9		0.95	
0.2	517	235	731	334	914	442
0.5	541	243	751	347	931	462
0.8	567	254	788	364	956	482

While the estimated sample sizes are larger for the skewed distributions, the patterns seen for the symmetric normal distribution still hold. Additionally, the results are quite comparable to the results above. Hence, the COS is robust to skewed data.

Table 6: The points of stability for different widths, levels of confidence, and effect sizes, for log-normal distribution with skewness of 1. For each level of confidence, $w = 0.1$ in the left column, and $w = 0.15$ in the right.

	Level of confidence					
d	0.8		0.9		0.95	
0.2	539	234	755	341	925	448
0.5	550	250	759	359	944	468
0.8	571	261	793	372	965	489

3.2 Bayesian Methods

Any sample size calculation using any frequentist approach ultimately assumes that the values of the true parameter can be known. With limited data as in pilot studies, the estimates are not guaranteed to be precise, and will likely come with much uncertainty. It can be inappropriate, therefore, to use type I and type II error rates to calculate sample sizes.

In this section we consider two Bayesian metrics that can be used to determine sample size: the average coverage criterion (ACC) and the average length criterion (ALC). The ACC controls the coverage rate $1 - \alpha$ of length l highest posterior density (HPD) intervals, and the ALC controls length l HPD intervals to have coverage rate $1 - \alpha$. Often, these two metrics are considered along with the worst outcome criterion (WOC), but it has been shown that sample sizes determined by the WOC are significantly greater than sample sizes determined by the ACC and ALC for the same l and α (Cao et al., 2009). Hence, for our purposes of keeping sample sizes feasibly low, we do not consider the WOC.

For both methods, it will be assumed that the two treatment groups have means μ_j, $j = 1, 2$, and a common variance λ. We will assume a conjugate prior for the precision, i.e. $\lambda^{-1/2} \sim \Gamma(v, \beta)$, for some v and β, which implies that $\mu_j|\lambda \sim N(\mu_{0j}, n_{0j}\lambda)$, where μ_{0j} and n_{0j} are hyperparameters. Practically, the n_{0j}'s control the weight given to the prior means μ_{0j} for the estimates, and can be intepreted as having n_{0j} subjects with mean μ_{0j} prior to the study. Hence, for our purposes, we set n_{0j} to be equal to the pilot study size. Additionally, since we are assuming equal sample size allocation, we set $n_{01} = n_{02} = n_0$.

Since the Gamma distribution has mean $\frac{v}{\beta}$ and variance $\frac{v}{\beta^2}$, we obtain estimates of $\hat{v}$ and $\hat{\beta}$ by taking the variance of our pilot study, as well as bootstrapping from our pilot study to obtain an estimate of the variance of the variance.

3.2.1 Average Coverage Criterion

The ACC (Adcock, 1988) determines the smallest sample size such that for a fixed length l, the expected coverage level is at least $1 - \alpha$. Specifically, it is the smallest integer n that satisfies

$$\int_{\chi} \left\{ \int_{a}^{a+l} f(\theta|x, n) d\theta \right\} f(x) dx \geq 1 - \alpha,$$

where a is some statistic that can be chosen based on the data, and χ is the data space.

Joseph and Bélisle (1997) showed that the sample size needed, under equal sample size allocation, is given by

$$n \geq \frac{-B + \sqrt{B^2 - 4AC}}{2A}, \tag{3}$$

where

$$\begin{aligned} A &= \frac{vl^2}{4} \\ B &= \frac{n_0 vl^2}{2} - 2\beta t^2_{2v,1-\alpha/2} \\ C &= \frac{n_0^2 vl^2}{4} - 2\beta n_0 t^2_{2v,1-\alpha/2}, \end{aligned}$$

where $t_{n,\alpha}$ represents the α^{th} quantile of a t-distribution with n degrees of freedom.

The second simulation in 2.1 is repeated. Specifically, 10,000 pilot studies were simulated, with $n = 10, 25$, and 50 for each group. The first group follows an $N(0, 1)$ distribution, and the second group follows an $N(d, 1)$ distribution, for $d = 0.2, 0.5$, and 0.8. For each pilot study, the pooled variance is calculated, and $B = 5000$ bootstrap samples are drawn. For each bootstrap sample, the pooled variance is estimated, then the variance of the bootstrap variances is calculated. These values are used to obtain v and β. Equation (3) is then used to obtain the estimated main study sample size needed for $l = 0.3, \alpha = 0.05$. The results of the simulation were identical for all d. Violin plots of the distribution of main sample size estimates, as well as some summary statistics, are shown in Figure 9 and Table 7.

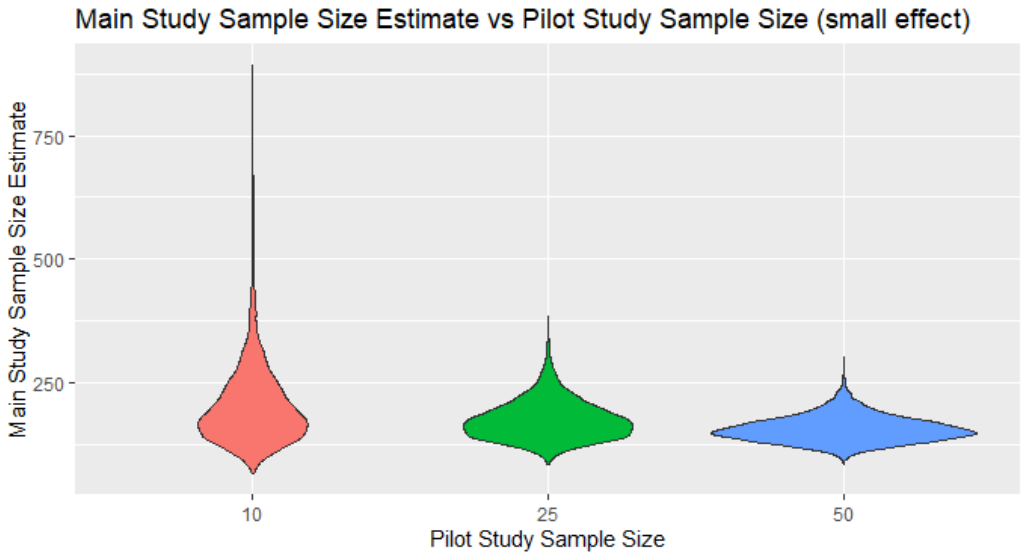

Figure 9: Violin plots for estimated sample size using ACC ($l = 0.3$) for a small ($d = 0.2$) effect size, with pilot studies of size $n = 10, 25$, and 50.

Table 7: Some summary statistics of the estimated main study sample size using the ACC.

Pilot Size	Min	Q1	Median	Mean	Q3	max	SD[1]
10	64	147	184	201.9	238	894	79.1
25	85	145	168	173.4	195	384	39.7
50	81	137	152	155.3	170	302	26.4

[1]SD represents the standard deviation of the main study sample size estimates.

Expectantly, as the effect size grows, the average estimated sample size needed for the main study shrinks. However, unlike the results in Section 2, the mode of the main study sample size estimate is similar across the pilot study sample sizes. The variance of the main study sample size estimate is also substantially lesser for all pilot study sizes.

The fact that the sample size estimated attained by the ACC is not

affected by the true effect size may eliminate the need for pilot studies altogether. Recall that the main goal of pilot studies is to get a rudimentary estimate of the effect size, which can be used in a power analysis to determine the sample size needed to accurately estimate the effect size. However, if researchers re-frame their hypothesis in terms of the ACC instead of NHST, then researchers can determine a sample size needed just from setting α and l.

3.2.2 Average Length Criterion

The ALC (Joseph and Bélisle, 1997) determines the smallest sample size such that for a fixed nominal coverage level $1-\alpha$, the expected length of the corresponding HPD interval is at most l. Specifically, the ALC is the smallest integer n that satisfies $\int_\chi l'(x,n)f(x)dx \leq l$, where $l'(x,n)$ is the length of the $1-\alpha$ credible interval, which is determined by solving $\int_a^{a+l'(x,n)} f(\theta|x,n)d\theta = 1-\alpha$. Once again, a is some statistic that can be chosen based on the data, and χ is the data space.

Joseph and Bélisle (1997) showed that, for $v > 0.5$, the sample size needed, under equal sample size allocation, can be found by solving the inequality

$$2t_{2n+2v,1-\alpha/2}\sqrt{\frac{2\beta(2n+2n_0)}{(2n+2v)(n+n_0)^2}} \times \frac{\Gamma\left(\frac{2n+2v}{2}\right)\Gamma\left(\frac{2v-1}{2}\right)}{\Gamma\left(\frac{2n+2v-1}{2}\right)\Gamma(v)} \leq 1 \qquad (4)$$

To solve this, bisectional search was used.

The simulation as above was repeated, with n being found using Equation (4) instead of Equation (3). Similar to the ACC, the ALC was not affected

by the true effect size. The results for $l = 0.3, \alpha = 0.05$, $d = 0.2$, and $n_0 = 10, 25$, and 50 are shown in Figure 10 and Table 8.

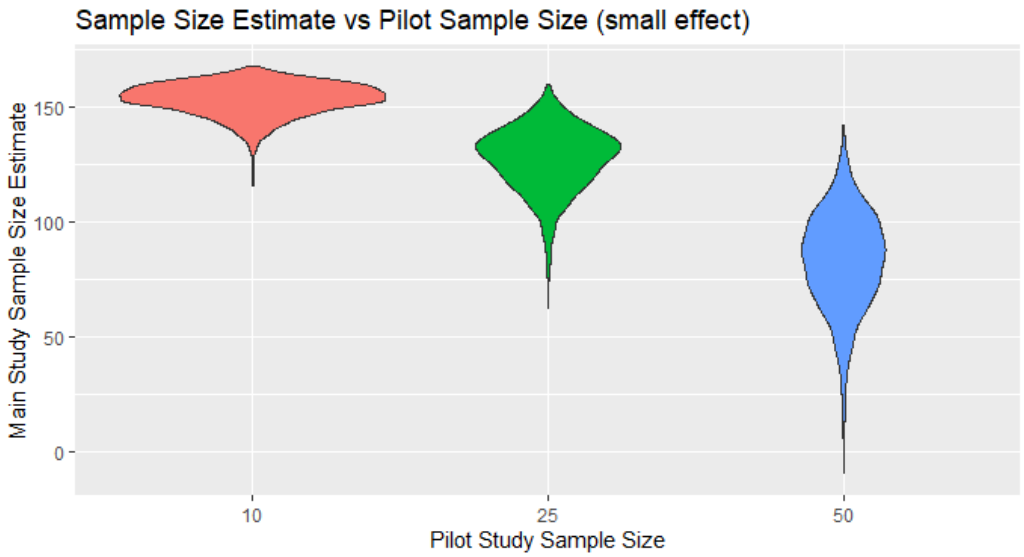

Figure 10: Violin plots for estimated sample size using ALC ($l = 0.3$) for a medium ($d = 0.5$) effect size, with pilot studies of size $n = 10, 25$, and 50.

Table 8: Some summary statistics of the estimated main study sample size using ALC.

Pilot Size	Min	Q1	Median	Mean	Q3	max	SD[1]
10	116	150	154	153.5	159	168	7.1
25	62	119	129	127.4	137	160	14.0
50	-10	70	85	83.09	99	142	21.6

[1]SD represents the standard deviation of the main study sample size estimates.

Similar to the ACC, the ALC does not seem to be affected by the effect size. However, in contrast to the ACC, the variance in the main study sample size estimate increases as the pilot study sample size increases. Additionally, the centre of the distribution of main study sample sizes decreases as the

pilot study sample size increases, although this can be explained by the ALC utilizing the prior information of n_0 more effectively than the ACC. Note that for a pilot study sample size of 50, some results were negative. This corresponds to some pilot sample sizes of 50 being unnecessarily large to achieve the average coverage of HPD intervals of length 0.3 to be at least 0.95.

3.2.3 Skewed distributions

The simulations in Section 3.2.1 and Section 3.2.2 are repeated, using the lognormal distributions described in Section 2.1. Since Section 3.2.1 and Section 3.2.2 showed that the sample size estimates are not affected by the effect size, only the results for large effect size are shown below.

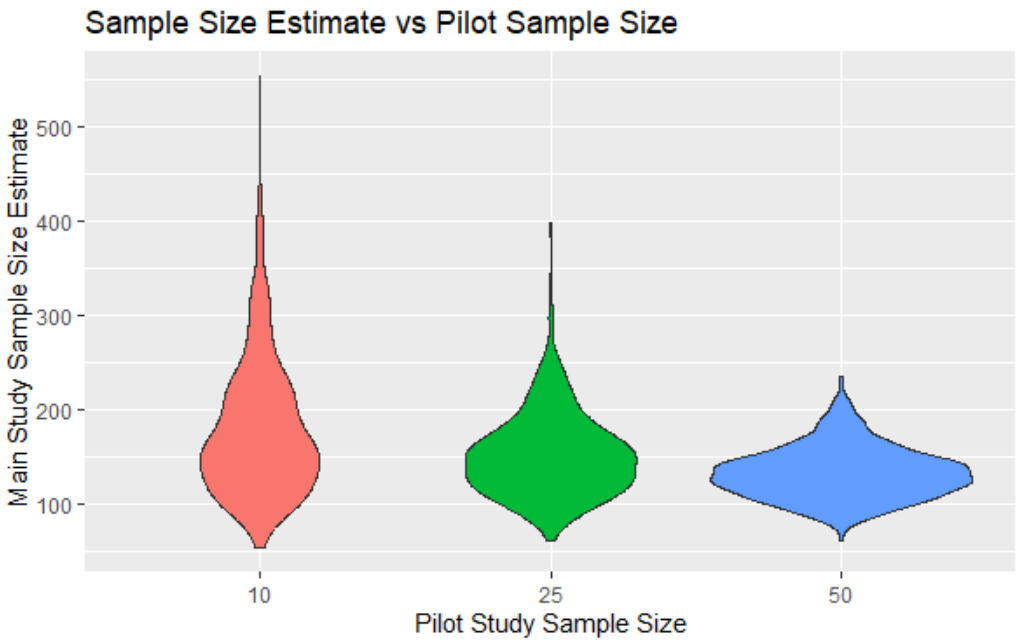

Figure 11: Violin plots for estimated sample size using ACC ($l = 0.3$) for a large ($d = 0.8$) effect size, with pilot studies of size $n = 10, 25$, and 50. Data is generated from a log normal with skewness equal to 1.

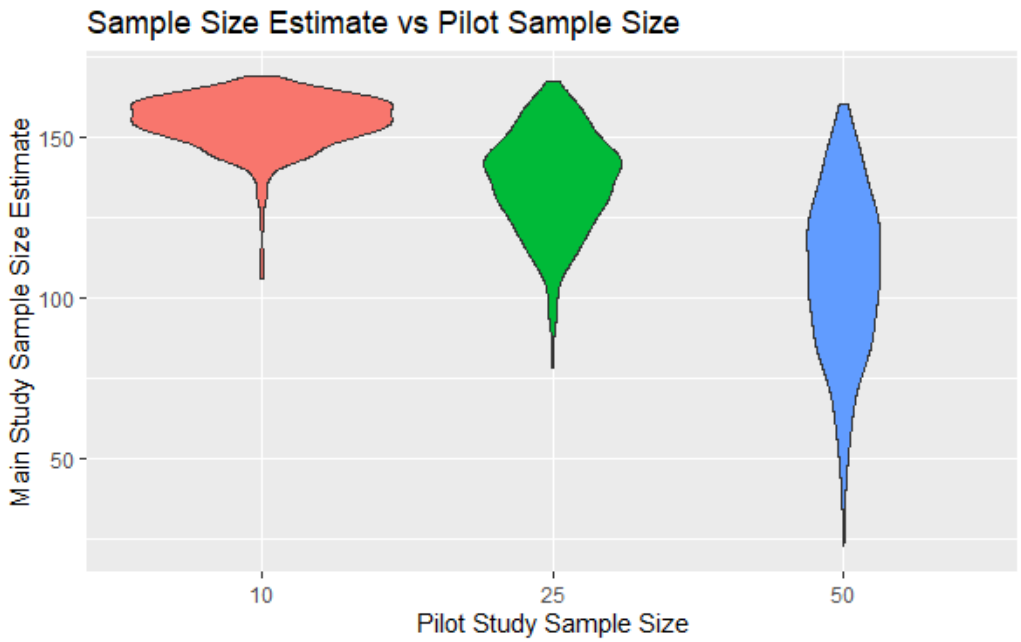

Figure 12: Violin plots for estimated sample size using ALC ($l = 0.3$) for a large ($d = 0.8$) effect size, with pilot studies of size $n = 10, 25$, and 50. Data is generated from a log normal with skewness equal to 1.

The distributions for both metrics are noticeably heavier tailed, but otherwise very similar to the distributions using symmetric data. Most importantly, the ranges for estimated sample size under the skewed data are the same as the ranges under symmetric data, and the modes are also the same. Therefore, the ACC and ALC are reasonably robust to skewed data.

4 Concluding Remarks

In Section 2 we investigated the shortcomings of current practices to estimate main study sample size using pilot studies. Due to small pilot study sizes,

estimates of effect size can vary drastically, which lead to an extremely skewed distribution of plausible effect sizes. Additionally, the estimated main study sample size deviate substantially based on the effect size, which is the purpose of the pilot study in the first place. In the following sections, we explored some alternatives that are more robust at smaller pilot study sample sizes, and do not depend on the effect size.

In Section 3.1, we examined the corridor of stability to obtain sample size estimates. The corridor of stability uses bootstrapping along with a predefined width around the true effect size, to estimate a sample size at which effect size estimates become stable. These estimates have a wide range depending on the level of confidence and width of corridor required. These estimates are comparatively not affected by effect size, and also are not affected by skewed data.

In Section 3.2, we explored two Bayesian alternatives to estimating sample size: the ACC and the ALC. The ACC produced sample size estimates similar to traditional power calculations, and the distribution of estimated sample sizes had greatly reduced variance. The ALC similarly had a far smaller range in its distribution of estimated sample sizes. The sample size estimates from both Bayesian metrics were not significantly different under skewed data.

Neither Bayesian metric were affected by the effect size itself. This means that researchers can arrive at a sample size estimate without having to do a pilot study at all, simply by setting the length l and coverage rate $1 - \alpha$ desired of the ACC or ALC. However, this would require researchers to move

on from traditional NHST to framing their hypotheses in terms of the ACC or ALC.

Acknowledgments

I would like to thank Yeying Zhu for her endless patience, understanding, and willingness to provide guidance and support. I would also like to thank my friends and family for humouring my rants about mathematics, without which I would not have the passion and curiosity I have today.

References

Adcock, C.J. "A Bayesian Approach to Calculating Sample Sizes." *The Statistician* 37 (1988): 433 – 439.

Biesanz, Jeremy and Sheree Schrager. "Sample Size Planning with Effect Size Estimates." (2010). Unpublished manuscript, University of British Columbia.

Cain, Meghan, Zhiyong Zhang, and Ke-Hai Yuan. "Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation." *Behavior Research Methods* 17 (2016).

Cao, Jing, J. Jack Lee, and Susan Alber. "Comparison of Bayesian sample size criteria: ACC, ALC, and WOC." *Journal of Statistical Planning and Inference* 139 (2009): 4111–4122.

Chow, Shein-Chung, Jun Shao, and Hansheng Wang. *Sample size calculations in clinical research (2nd ed.).* 2007.

Cohen, Jacob. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press, 1977.

Joseph, Lawrence and Patrick Bélisle. "Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means." *Journal of the Royal Statistical Society. Series D (The Statistician)* 46 (1997): 209–226.

Kelter, Riko. "Analysis of Bayesian posterior significance and effect size indices for the two-sample t-test to support reproducible medical research." *BMC Medical Research Methodology* 20 (2020).

Lakens, Daniël and Ellen R. K. Evers. "Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies." PMID: 26173264. *Perspectives on Psychological Science* 9 (2014): 278–292.

Leon, Andrew, Lori Davis, and Helena Kraemer. "The Role and Interpretation of Pilot Studies in Clinical Research." *Journal of psychiatric research* 45 (2010): 626–9.

Micceri, Theodore. "The unicorn, the normal curve, and other improbable creatures.." *Psychological Bulletin* 105 (1989): 156–166.

Schönbrodt, Felix D. and Marco Perugini. "At what sample size do correlations stabilize?." *Journal of Research in Personality* 47 (2013): 609–612.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. "p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results." PMID: 26186117. *Perspectives on Psychological Science* 9 (2014): 666–681.