Determination of a pilot sample size to determine the sample size of a substantive trial

Scholastica Chinenye Obodo

A thesis submitted in Partial fulfilment of the requirement of the University of the West of England, Bristol, for the degree of Doctor of Philosophy

School of Computing & Creative Technologies,
University of the West of England, Bristol

June,2024

ABSTRACT

This research intends to determine the appropriate sample sizes for two-arm pilot studies to help correctly determine the required sample size for the corresponding substantive or definitive trial. Previous research in the area has been reviewed. R programming is used to undertake simulation studies. The first study is an investigation of the procedure proposed by Browne (1995). Results from this study confirm the work of Browne and show that alpha (nominal Type I error rate), beta (nominal Type II error rate), and effect size do not affect the error associated with estimation of sample size using the method. Desired coverage has a moderate effect, and the pilot sample size has a big effect on the error associated with the predicted sample size when using Browne's formula. In general, the approach of Browne (1995) is valid but gives very large incorrect estimates. For this reason, the Goldilocks ("just about right") approach was developed from quantifying the degree of underestimation and overestimation and this new approach may be used to guide researchers in controlling the excess margin. Study 2 compared the performance of four methods of sample size estimation when the Minimum Clinical Importance Difference (MCID) is unknown. The results showed that Browne's method does not work effectively when MCID is unknown. In these situations, the use of Hedge's correction was the best of four formulae but still led to overestimation. Hence the need for Study 3 where new methods are proposed using upper confidence limit of the coefficient of variation which was presented using examples and simulations. Using this new approach, the parameter had similar impact in error margin as in Browne's method and performed well for large effect sizes only. The multiplier critical table was developed in study 4 however it results in big sample sizes and therefore would not be used in practice.

To my beloved husband and my amazing Mum

To my deceased dear Dad Joseph

Acknowledgements

I extend my sincere gratitude to my supervisory team, Paul White and Deirdre Toher, for their invaluable guidance, unwavering support, and consistent assistance throughout my PhD journey. Their impact towards achieving this milestone remains profound.

I am also thankful to my loving and ever-supportive husband Chidi Obodo, and my adorable children, for their unwavering support, love and understanding.

Furthermore, I express my ineffable gratitude to my amazing mother, Juliana Ojii. The strength and courage you instilled in me while growing up have been the driving force behind my accomplishments. To my late father, Joseph Ojii, who celebrated every one of my academic achievements and motivated me towards excellence, I express my heartfelt gratitude. I am certain that I have made you proud. I equally owe a debt of gratitude to my ever-lovely siblings for their kind words and support.

To all my former educators, family members and friends who supported me in various ways towards achieving this success, I extend my sincere appreciation.

My heartfelt thanks also go to the former President of Nigeria, Goodluck Ebele Jonathan, for believing in us Nigerian youths and for planting this golden seed of empowerment, which has borne this great success. Additionally, I extend my gratitude to Petroleum Technology Development Fund (PTDF) for accepting the responsibility of sponsoring and nurturing this planted seed to germinate, grow and become fruitful.

Finally, my profound thanks go to Almighty God for His grace and strength throughout this journey.

Achievements

Published papers, Posters and Conference.

Obodo, S., Toher, D., and White, P. (2021). “Estimation of the two-group pilot sample size with a cautionary note on Browne’s formula”. *Journal of Applied Quantitative methods*. 16(3).

Obodo, S.C., Toher, D., and White, P. (2023). “The Just-about-right Pilot sample size to control error margin”. *International Journal of Statistics and Probability*.12(3), pp1-7.

September 2021: Poster presentation – “Investigating Browne’s method in sample size determination.” Fry Conference Series: Statistics at University of Bristol Conference. Future results and You 2021. **Won Bronze Medal.**

November 2021: Poster presentation – “Uncertainties in Browne’s method of sample size determination.” Southwest Doctoral Training Partnership (SWDTP) Conference. https://www.swdtp.ac.uk/thats-a-wrap-2021-student-conference-review/.

March 2022- Presented thesis- “Determination of a pilot sample size to determine the sample size of a substantive trial.” - Data Science & Mathematics Cluster Meeting at UWE Bristol.

March 2023- Presented thesis – “Determination of sample size for clinical trials” – UWE Three minutes thesis.

CONTENTS

List of Tables

List of figures

Commonly used parameters

Table 1: Commonly used notation, simulation parameters.

SYMBOL	MEANING
α	Alpha (The probability of a Type I error)
β	Beta (The probability of a Type II error)
m_1	Pilot sample size for randomised arm 1
m_2	Pilot sample size for randomised arm 2
n	Sample size
μ_1	Mean of distribution 1
μ_2	Mean of distribution 2
σ_1^2	Variance of distribution 1
σ_2^2	Variance of distribution 2
$\bar{x}_1$	Mean of sample 1
$\bar{x}_2$	Mean of sample 2
k	Multiplier in Browne's formula
ν	Degrees of freedom
s_1^2	Variance of sample 1
s_2^2	Variance of sample 2
s_p^2	Pooled Variance for sample 1 and sample 2
γ	Gamma (coverage)
r	Random allocation ratio
χ^2	Chi-Square
N_{true}	Smallest sample size which satisfies power requirements
$\widehat{N}$	Estimated total smallest sample which satisfies power requirements
δ	Cohen's delta

Chapter 1

Introduction

In 2022, there were more than 38,000 newly registered clinical trials, bringing the total to over 430,000 since the year 2000 on ClinicalTrials.gov, with a global estimated spending of more than US$68 billion by 2025 (Gehr et al., 2023). In addition, there are many other non-clinical trials (e.g., in psychology, or in the biological sciences) which do not require registration. A large percentage of the amount spent is on trials with an investigative medicinal product. A critical role in advancing medical knowledge and improving care of patients is done by clinical trials (Holford et al., 2010).

The effectiveness and safety of new treatments, medical services, drugs or identifying potential side effects or risk associated with interventions are the aims of these trials (Locock and Smith, 2011). There are phases in conducting the trials (Korn et al, 2012; Browne, 1995). In these drug investigations, trials are typically described as being Phase I, Phase II, Phase III, or Phase IV trials. In brief, a Phase I trial is a relatively small study (circa 10 to 50 participants) designed to find out what a new drug might do to the body and to identify safe dosage levels. Phase II trials, with typically 20 to 120 participants are designed to identify efficacy, find out more about side effects, and can be used to help plan for a more substantive or definitive Phase III trial. The Phase III trial is usually large; sample sizes are in hundreds or thousands and are designed to give substantive conclusions and to add to the knowledge base. Phase IV investigations are concerned with further understanding the long-term effects of drugs when used as a treatment. In non-drug studies, the terminology "pilot" and "definitive study" are used to reflect the equivalent of a Phase II study and a Phase III study respectively. However, a Phase III study might not give substantive conclusions if the sample size is incorrectly set too low or it may be too costly if the sample size is incorrectly pitched too high. Whitehead et al. (2016) explained that despite randomised controlled trials (RCTs) being considered the gold standard to determine the effectiveness of a novel intervention, they can be underpowered for the expected primary outcome measure if they fail to recruit sufficient participants.

The Goldilocks principle of getting the sample size "just about right" would therefore have great economic and social benefit. There is evidence that "just about right" is not happening in practice (Nayak, 2010). A systematic review of published RCTs with continuous outcomes found the population variation was underestimated in 80% of reported endpoints. According to Charles et al. (2009) a review of trials for both binary and continuous outcomes found that 25% of studies were vastly underpowered. The proposed work in this thesis will consider rectifying this problem with the most used design; the randomised controlled trial with two parallel arms.

For every clinical trial there is a need for a justification of the sample size to be used for the design (Julious, 2005), and the minimum number of participants needed for a clinical trial is determined from a sample size calculation (Campbell et al., 2010). In some research, there is no prior information upon which the sample size for a definitive or substantive trial can be justified leading to a need for more research to be carried out (Julious and Swank, 2005). For this reason, it is not uncommon for a pilot study to be carried out prior to a potentially costly large-scale investigation. One purpose of a pilot study may be to gather information on either the likely effect size, or the standard deviation and to then use that information to help estimate the sample size needed for a definitive or substantive trial (i.e., one which will either give essentially definitive conclusions with 90% power or higher or conclusions of some substance with at least 80% power). Although there are various rules of thumb for recommended pilot study sample sizes, there is no agreed way of determining the size of a pilot study to subsequently help determine the parameters needed to estimate the sample size for a larger definitive trial.

1.1 Motivation

According to Suresh and Chandrashekara (2012), if the appropriate sample size of a purportedly substantive trial is not determined it could lead to a study without statistical significance, and in particular, the chance of an inconclusive result is high when too few participants are used for a main trial (Halpern et al., 2002).

A study with a small sample size (underpowered) may produce incorrect or inconclusive results or unconvincing results and make the study process a failure. If a study is ostensibly designed to produce substantive (e.g., 80% power) or a definitive set of conclusions (e.g., 90% power or higher), but has a sample size incorrectly set too low, then this is deemed unethical as it means participants are being put through a data collection process which cannot meet its objectives. In contrast, an overly large sample size could lead to a waste of resources, and it is arguably unethical to put an increased number of participants through a research study when the same conclusions could have been determined with a smaller sample size. Relatedly, with an overly large sample size, too many people may be denied beneficial treatment if randomised to placebo or the control arm, if in fact the treatment confers benefit. If the treatment offers no benefit, and an overly large sample size is used, then this too may be seen as unethical. For these reasons, institutional ethics committees, review boards and funding councils use statistical expertise to help evaluate quantitative research including the proposed sample size. Unfortunately, textbooks do not give sufficient guidance on how to determine the sample size of a definitive trial (Prescott and Soeken, 1989). In addition, Browne (1995), Julious (2004), Julious (2005), and Sim and Lewis (2012), all acknowledge that there is a disagreement over what sample size should be used for pilot trials to inform the design of definitive randomised controlled trials; recommendations have been developed but there is no consensus. Whitehead et al. (2016) state that a pilot study can help predict more precisely parameters required for the sample size calculation such as the variance of the outcome and the dropout rate and present issues early on the trial development. Kelly et al. (2005) confirmed there is insufficient guidance for conducting pilot trials. Most of the recommendations focus on continuous outcome data (see Browne (1995), Kieser and Wassmer (1996), and Sim and Lewis (2012)).

1.2 Previous research in pilot sample size determination

The value of the standard deviation is not known ahead of the research; however, when this standard deviation is obtained from a pilot study, the values could be imprecise (Friede and Kieser, 2001). Pilot trials often estimate the standard deviation for a main trial imprecisely and more participants than necessary will be used for a trial if the anticipated standard deviation is estimated to be too high, and if the value for the standard deviation used is too low the trial will not have enough participants to achieve the required effect which could lead to misleading results or inconclusive findings (Whitehead et al., 2016). Using a standard deviation from a small pilot sample, the probability of getting the planned power can be as low as 40%; the rule of thumb of using a pilot sample size m of 30 or greater to estimate a parameter will not eliminate the problem, except when the effect size is quite large (Browne, 1995). The research results suggest the use of at least 80 per cent upper one-sided confidence limit on σ as the estimate of the standard deviation to guarantee an 80 per cent chance of achieving the planned power in the clinical trial (Browne, 1995). Sample size could be wrongly estimated if the imprecise value of standard deviation from a pilot study is used in the standard method for sample size determination (Friede and Kieser, 2001).

According to Whitehead et al. (2016), incorrectly estimating the sample size for a clinical trial could cause both ethical and financial challenges for a trial. Results for an external pilot trial sample size are offered which aim to lessen the overall trial sample size. It was found that the optimal pilot trial sample size increases with the size of the main trial. For a two arm study, Whitehead et al. (2016) proposed stepped rules of thumb for 90% powered main trials and that the sample size for a two-armed pilot trial to minimize the sample size for a two-armed pilot trial should be 150, 50, 30, and 20 per arm for standardized effect sizes (δ) of $\delta < 0.1, 0.1 \leq \delta < 0.3, 0.3 \leq \delta \leq 0.7 , \delta \geq 0.7$ respectively.

This overview of the literature shows that there have been previous adjustments made to sample size estimation formulae to improve sample sizes estimation however this adjustment led to inflated sample sizes (Whitehead et al 2016). The upper confidence limit approach (UCL) referred to as Browne's method (Browne, 1995) will be further discussed in section 1.2.1, the non-central *t*-distribution approach section will be covered in 1.2.2, Whitehead et al. (2016) research will be reviewed in section 1.2.3, and UCL review by Kieser and Wassmer 1996 will be covered in section 1.2.4.

1.2.1 The Upper Confidence Limit Approach (UCL)

The most influential work in this area is by Richard Browne (Browne 1995). The objective of the method is to better estimate sizes sample for trials while attaining the planned power. Using the standard deviation from a pilot sample size will likely lead to a study not achieving the planned power. To investigate the occurrence of actual power values that equal or surpass the planned power values when estimating standard deviation, Browne investigated the occurrence of actual power values that equal or surpass the planned power values when estimating the standard deviation from (a) $100(1-\gamma)$ percent upper one-sided confidence limit (UCL) on σ for several values of γ, and (b) the pilot sample standard deviation (unadjusted values). In this work, the one and two sample *t*-test, with an alternative hypothesis being a point specific value was used for the one arm and two arm design. It is assumed that this alternative hypothesis is precisely true in the simulations. In Browne's work, stochastic computer simulations are done for a nominal significance level (α)= 0.05 only and for m= 5, 10, 30, 50, 100. For a target $(1-\beta)$ = 0.8, Browne showed that the power for a definitive trial would be lower than wanted if the standard deviation from pilot sample is used in calculations of substantive sample sizes.

The alternative approach suggested by this method concluded that using an 80% upper one-sided confidence limit on the variance will help improve the chance of achieving the planned power in clinical trial 80% of the time, and more generally concluded that using a *P*% upper one-sided confidence limit on the variance will help improve the chance of achieving planned power in clinical trial up to *P*% of the time. This work of Browne is the leading approach for sample size estimation based on pilot estimates, **but** it does not identify the minimum sample size for a pilot study, nor does it consider whether the projected sample sizes are "too small" or "too large". A consideration of the limitations of Browne's work will help establish a "Just about right" (JAR) or Goldilocks approach to determining sample size.

For a definitive two group trial, the formula to determine the sample size for testing at the α significance level with power equal to $(1-\beta)$, assuming normally distributed data, is given by

$$N_{true} = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\beta)}\right)^2}{(\mu_1 - \mu_2)^2} \ .\sigma^2 \qquad (1.1)$$

where N_{true} denotes the minimum sample size needed; $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are standardized normal deviate for two-sided significance testing with nominal significance level α and required power $(1-\beta)$, σ^2 is the population variance for the outcome measure and is assumed to be equal for both arms, μ_1 is the mean for distribution 1, and μ_2 is the mean for distribution 2, and r is the allocation ratio(see Van Belle and Martin 1993, Whitehead et al, 2016). The N_{true} is for each arm and is rounded up to obtain an integer value.

In the above formula (1.1) it is assumed that all the parameters are known. In practice the parameters will not be known and will be estimated from small-scale pilot data. For assumed equal variances, a naive estimated sample size would then be

$$\hat{N} = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\beta)}\right)^2}{(\bar{x}_1 - \bar{x}_2)^2} \ .s^2 \tag{1.2}$$

where s^2 is sample variance of the pilot data and $\bar{x}_1$ and $\bar{x}_2$ are respectively the sample mean for group 1 and 2 respectively.

when the Minimum Important Clinical Difference (MCID) is known

$$\hat{N} = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\beta)}\right)^2}{(\mu_1 - \mu_2)^2} \ .s^2 \tag{1.3}$$

However, the estimated sample size, $\hat{N}_B$, using Browne's approach (Browne, 1995) is given by

$$\hat{N}_B = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\beta)}\right)^2}{(\mu_1 - \mu_2)^2} \ .ks^2 \tag{1.4}$$

where $k = (m_1 + \ m_2 - 2)/(\chi^2_{v,1-\gamma})$ is the multiplier in Browne’s method, m_1 is the pilot sample size for randomised arm 1, m_2 is the pilot sample size for randomised arm 2, $\chi^2_{v,1-\ \gamma}$ is the upper $100\,\gamma$ percentage point of the chi-square variate on v degrees of freedom, and s^2 is the pooled variance and r the allocation ration is equal to 1.

These formulae will form the basis of the initial investigations and be used in deriving conditions for the Goldilocks solution.

1.2.2 The Non-central t-distribution (NCT) Approach

Considering the use of s^2 (from sample) instead of σ^2 (from population) in the sample size calculation, Julious and Owen (2006) proposed a method for the sample size calculation known as the non-central t-distribution approach which accounts for this. However, in this approach the sample size is inflated based on the number of degrees of freedom df which the variance estimate is based on, the sample size per treatment arm in this approach for main trial is presented by

$$n \geq 2s^2 \frac{\left[tinv\left(1-\beta, df, t_{1-\alpha/2n-2}\right)\right]^2}{d^2} \tag{1.5}$$

where the inverse function of the cumulative distribution function with non-centrality parameter b on degrees of freedom (df) is given by $tinv\ (.,k,b\ )$, df is the degrees of freedom about the estimates of sample variance s^2, and d is the unknown difference in means. The estimated sample size will be increased if the estimate of the variance is based on only a few degrees of freedom (df). An increase in the degrees of freedom of the variance will lead to the estimated sample sizes being smaller.

Essentially, as the pilot sample size (and hence degrees of freedom) increase, the better the accuracy resulting in narrower confidence intervals and more accurate estimates.

According to Julious and Owen (2006) the method must be solved iteratively as n appears on both sides. With iterations starting at

$$n = \frac{2s^2\left[tinv\left(1-\beta, df\ , z_{1-\alpha/2}\right)\right]^2}{d^2} \tag{1.6}$$

The sample variance tends to the population variance as the degrees of freedom increases. According to Julious (2004), as the degrees of freedom becomes higher, the less sensitive calculation is to assumptions about the variance.
It is challenging specifying MCID, and a variance estimate hence specifying the standardized difference is preferred, which will lead to replacing d and s in the equation above with in equation 1.6

$$n = \frac{2\left[tinv\left(1-\beta, df\ , z_{1-\alpha/2}\right)\right]}{\delta^2} \tag{1.7}$$

To allow comparison of effect sizes across scales, Cohen (1992) proposed the use of $0.2, 0.5$ and 0.8 as small, medium, and large standardized effect sizes. This method however seen to lead to inflation of sample sizes by

$$IF = \frac{[tinv\left(1-\beta, df, z_{1-\alpha/2}\right)]^2}{(z_{1-\alpha/2} + z_{1-\beta})^2} \tag{1.8}$$

If the inflation factor (IF) is multiplied by the standard method in Eqtn 1.1, it will give the sample size as it would be when done with NCT method. The inflation factors depend on the type I error rate, the type II error rate and on pilot trial samples sizes. The inflation factor, using NCT approach for two-sided type I error rate of 5%, varying pilot trial sample sizes and power requirements of 80% and 90% are presented in Table 1.1. The table shows that as the pilot sample sizes increase the Inflation factor reduces, indicating that pilot sample sizes have an impact on the corresponding IF.It shows how much larger the sample size is compared to the standard calculation. Using pilot trial sample size of 50, as suggested in the tables implies the sample size will be inflated by 1.055 at 90% power and by 1.036 at 80% power.

Table 1.1: Inflation factor for sample size calculation for the NCT approach at 5% Type 1 error.

	Power	
Sample size for Pilot trial	90%	80%
20	1.156	1.099
24	1.125	1.080
30	1.097	1.062
40	1.071	1.045
50	1.056	1.036
70	1.039	1.025
100	1.027	1.017
200	1.013	1.008

This method therefore leads to inflation of sample size however it did not consider to what extent the sample sizes were overestimated as will be done in this research.

1.2.3 Upper Confidence Limit (UCL) and Non-central *t*-distribution (NCT) approaches comparism by Whitehead (2016)

UCL and NCT methods were investigated by Whitehead (2016) to find out how well these methods estimate sample sizes. UCL approach by Browne, as in equation 1.8,

$$\widehat{N}_B = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\beta)}\right)^2}{(\mu_1 - \mu_2)^2} \ . ks^2 \qquad (1.8)$$

was divided by the standard method of sample size calculation formula.

$$n = \frac{(r+1)\left(z_{1-\beta} \ + \ z_{1-\alpha/2}\right)^2 s^2}{rd^2} \qquad (1.9)$$

where s^2 is used as an estimate of σ^2. The research found UCL approach of sample size estimation to be larger by inflation factor.

$$B_{UCL}^2 = IF_{UCL} = \left[\frac{k}{\chi_{1-x,k}^2}\right]$$

and this depends on the sample size and the value of χ. The standardizes effect size, Type 1 error rate as well as the probability of achieving the required power (X) is set. The upper confidence limit taken in UCL approach which gives a $100X\%$ chance of achieving the required power for the trial. The eventually achieved probability levels using this approach are in Table 1.2.

Furthermore, using the NCT approach the Inflation Factor IF was obtained from Julious and Owen (2006) to investigate by how much the method overestimate compared to standard method.

$$IF = \frac{\left(2\left[tin\,v\left(1-\beta, k, z_{1-\alpha/2}\right)\right]^2 S^2\right)/d^2}{\left(2\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2 S^2\right)/d^2} \qquad (1.10)$$

Pilot trial sample size, type I error rate (α) and type II error rate (β) are all factors to be considered in inflation by NCT.

Table 1.2 presents inflation factors (IF) and proportion of confidence level (X) for the UCL approach that gives the same sample size as the NCT approach as summarized in the research of whitehead (2016). It shows that for 80 percent powered main trial, a researcher using a pilot sample size of 20 is 56% confident that the estimated sample size for the main trial will be inflated by a factor of 1.099 using either UCL or NCT. For a 90 percent powered main trial, a researcher using a pilot sample size of 20 is 62% confident that the estimated sample size for the main trial will be inflated by a factor of 1.156 using either UCL or NCT.

Table 1.2: Levels of X and IF that for the UCL approach that gives the same sample size as the NCT approach presented.

Sample size for pilot trial	Power 80%		Power 90%	
	X	$IF_{UCL/NCT}$	X	$IF_{UCL/NCT}$
20	0.566	1.099	0.622	1.156
24	0.560	1.080	0.611	1.125
30	0.553	1.062	0.599	1.097
40	0.546	1.045	0.586	1.071
50	0.541	1.036	0.577	1.056
70	0.534	1.025	0.565	1.039
100	0.529	1.017	0.554	1.027
200	0.520	1.008	0.538	1.013

The Inflation value shows these methods overestimate sample sizes, without achieving desired power supporting the need for further research in the area.

1.2.4 Keiser and Wassmer 1996 review of Upper Confidence limit approach

The research reviewed the UCL approach suggested by Browne 1995 using analytical considerations. It concludes that $100(1-\gamma)$ percent upper confidence interval for the population variance produces results with a probability of at least $1-\gamma$ of achieving the planned power $1-\beta$ thereby concurring with Browne's research. The research suggested adequate pilot sample size for $(1-\gamma)$=0.8 should be ranging 20 to 40 for an intended sample size of 80 to 250 hence adding to existing rules of thumb. The research however did not consider how much these processes overestimate the minimum required sample size.

1.3 Aim and Objectives

The overarching aim of this research is to estimate the sample size for a pilot study to help estimate the sample size of a definitive (90% power) or substantive (80% power) trial. This will help avoid overestimation of sample sizes that would waste resources, be unethical, and avoid underestimation of sample sizes that would lead to inaccurate, inconclusive, or unconvincing results and be unethical.

The prominent work in this area is due to Browne (1995). The proposed work will evaluate and extend the work of Browne. In addition, a new novel mathematical device to provide a Goldilocks solution is proposed; this new novel approach will be used to determine its effectiveness under a position when a researcher could specify a minimum clinically important difference or, through subject knowledge, estimate an effect size. Furthermore, the method of estimation of sample size using Upper confidence Limit of co-efficient of variation will be proposed and compared to Browne's approach. The critical values of k^2 multiplier was developed for when the true effect size is not known.

The objectives are:

1. Review of Browne's approach of sample size estimation (Study 1a) To determine how accurate Browne's approach is in sample size determination over a range of significance levels α =(0.01,0.05), power levels 1 – β =(0.8,0.9), pilot sample sizes m =(5,10,30,50,100), known effect sizes δ =(0.10,0.40,0.75), coverage levels $1-\gamma$ =(0.8,0.9). This review is given in Chapter 4.

2. Goldilocks method for sample size determination (Study 1b)
Propose and develop a method termed the "Just about Right" (JAR) method or the "Goldilocks approach" for sample size determination. Check for substantial merits of the proposed method compared with rules of thumb. Given in Chapter 5.

3. Comparison of Sample Size Estimation Methods when the Minimum Clinically Important Different (MCID) is unknown (Study 2).
This study compares: a naive estimate for sample sizes assuming σ^2, μ_1, μ_2 are unknown denoted as $(N_{N,C})$,an estimate for sample size using Cohen's d in Browne's approach denoted as $(N_{B,C})$**,**a naive estimate of sample size using Hedge's h denoted

as $(N_{N,H})$ and an estimate of sample size using Hedges' h in Browne's approach denoted as $(N_{B,H})$**.**The sample size estimates provided in $(N_{N,C})$, $(N_{B,C})$**,** $(N_{N,H})$ and $(N_{B,H})$ will be compared through simulation to determine which formula gives the most accurate sample size when MCID is unknown. The simulation parameters will be pilot sample size m = (8, 16, 32, 64, 128); α =(0.01, 0.05); β= (0.1, 0.2); $1-\gamma$= (0.8, 0.9) and δ = (0.10, 0.40, 0.75) chosen in close similarity to Browne's parameter combination. See Chapter 6.

4. Sample size estimation method developed from upper confidence limit of coefficient of variations (Study 3).

The study will develop three formulae for sample size estimation using coefficient of variation. They will be developed using coefficient of variation (*c*) considering the upper γ% confidence interval for *c* using the: Standard coefficient of variation formula $(N_{C,S})$**,** the McKay $(N_{C,M})$ and the Vangel formula $(N_{C,V})$ **.**Examples and simulations will be done to check the performance of the formulae and generate recommendations. See Chapter 7.

5. k^2 Modifier approach (Study 4)

This study will develop modified approach that uses k^2 modifier based on coverages for when MCID is unknown. See Chapter 8.

1.4 Research Questions

The study intends to answer the following questions:

1. How accurate is Browne's approach? (Study 1)
2. Does the newly proposed "Goldilocks approach" have substantial merit compared with rules-of-thumb ? (Study 1b)
3. Does $N_{N,C}, N_{B,C}, N_{N,H}, N_{B,H}$ methods of data estimation estimate with a tolerable level of error? (Study 2)
4. Does using the upper confidence limit of coefficient of variations approaches namely: Standard coefficient of variation $(N_{C,S})$, McKay $(N_{C,M})$ and the Vangel formula $(N_{C,V})$ estimate with less error than Brown's formula.

   (Study 3)

4. Does a modified approach that uses k^2 modifier based on coverages presents the proposed solution for when MCID is unknown? (Study 4)

1.5 Impact statement

If successful, this research will

(a) Add to the knowledge base for methodologists.

(b) Provide researchers with a means of conducting better science and empirical discovery in the two-arm randomised design.

(c) Help ethics committees when considering pilot research (Phase II) and substantive research (Phase III) in commonly encountered research.

(d) Help funding panels assess grant applications for a commonly used design.

(e) Help prevent underpowered research and avoid failure to demonstrate the true situation.

(f) Help prevent overly large sample sizes with attendant economic benefits.

(g) Help prevent too many participants being exposed to an intervention when the same conclusions could have been obtained from a smaller sample size (ethical benefit) or too many being assessed (e.g., depriving too many of a beneficial effect, or too many undergoing an ineffective invasive procedure or receiving an ineffective active treatment)

(h) Provide a springboard for extending the methodology to other designs and methods.

1.6 Outline of Thesis

This chapter began by introducing the concept of clinical trials and pilot studies. Some previous research in pilot sample size determination was reviewed, the motivation section explained the importance of the research, the aim, objectives, specific research question and the impact of the study was also discussed.

Chapter 2 gives a literature review of the research, and includes an outline of randomised trials, sample size for pilot studies, factors affecting powers of a test, pilot and feasibility studies, sample size for pilot studies and confidence intervals for means, standard deviations, cohen's d and coefficient of variation.

Chapter 3 contains methodological work involved in carrying out the research, discussed simulations and random number, Monte Carlo methods, number of iterations and research evaluation metrics.

Chapter 4 examines the sample size formula using Browne’s approach of sample size estimation, which is used for the two-group problem assuming normality, for known MCID. The results from Analysis of objective 1 are presented in tables and graphs, and their interpretation are discussed too (Study 1).

Chapter 5 explains the Goldilocks Principle, which is the “Just about right” (JAR) method of estimating sample size to ensure that an estimated sample size does not exceed certain margin for a lower or upper percentage of the required true, but unknown, sample size, which is objective 2. The results from this study led to two peer reviewed publications: Obodo et al. (2021) and Obodo et al. (2023) both given in Appendix A.

Chapter 6 presents Objective 3,which compares four methods of sample size estimation when the MCID is unknown to determine their performance in estimating sample sizes. The results from this are presented using tables and graphics (Study 2).

Chapter 7 elaborates on Objective 4, which involves the comparison of methods of sample size estimation using the upper confidence limit for the coefficient of variation (Study 3).The simulation design, results and summary are presented, followed by a comparism of the best of the three method to Browne’s method. The findings are presented in tables and graphs.

Chapter 8 presents objective 5, the proposed solution for when MCID is unknown, which involves developing critical values of the k^2 multiplier. However, the solution will not be feasible in practice (Study 4).

Chapter 9 This chapter gives a summary of the thesis, the findings and their implications, recommendation, and suggestion for further research.

Chapter 2

Literature Review and Preliminary investigation

The literature on randomised trials, sample size for pilot studies, factors affecting power of a test, pilot and feasibility studies, confidence intervals for means, standard deviation , Cohen's d and coefficient of variation are discussed. This chapter concludes with examples using coefficient of variation for confidence intervals, which will be utilized as a measure of variability in proposing new methods in study 3.

2.1 Randomised Trials

Ranjith (2005) highlighted that the term "randomised trial (RT)" and "randomised controlled trials (RCT)" are sometimes used synonymously. Note that RCT is "controlled" and not "controls" and can therefore describe studies that compare multiple treatment groups with each other in the absence of a control group. It is commonly recognized that the Randomised Controlled Trial is where allocation is controlled by randomisation irrespective of whether there is a control group or not. Bailey (2008), comments that there may be more than one treatment group, more than one control group, or both. As such, there are different types of randomised trials namely Randomised Controlled Trials (RCT), Randomised Clinical Trials (RCT) and Randomised Controlled Clinical Trials (RCCT).

Chalmers et al. (1981) states that a randomised controlled trial (RCT) is a type of scientific (often medical) experiment that aims to reduce certain sources of bias when testing the effectiveness or efficacy of new treatments by randomly allocating subjects to two or more groups while treating them differently and comparing with a measured response.

In a clinical trial, the intervention under review is typically contrasted with a control procedure or treatment. The control treatment accounts for the fact that feelings of excitement, acceptance, and anticipation will influence patient outcomes. A monitoring procedure should be used to ensure that the effect assessed is due to the intervention and not to the presence of the patient in the clinical trial itself, called the Hawthorne effect (Parsons, 1974).

Dettori (2010) explained that a randomised controlled trial (RCT), when done with a large enough sample, is an effective measure of reducing bias. Bias is an unintended distortion in the choice of patients, data collection, endpoint determination and final analysis (Chalmers et al.,1981). When there is a systematic difference between the true values and the result in the trial, bias is said to have occurred (Malone, 2014).

The bias of interest in clinical trial is the Hawthorne effect, however there are other types of bias namely assessment bias, confounding and attrition bias (Malone, 2014). However, randomization helps to reduce the different types of bias of a trial, and this is achieved by making the treatment groups comparable for a prognostic factor so that any contrast between the groups can be attributed to the intervention under investigation (Torgerson, 2008).

There are different designs for trials, for instance, crossover trials are where patients receive both treatments sequentially and randomisation is used to determine the order each of the participants receives interventions. A simpler design is the parallel design used in the comparison of treatment groups, where treatments are run concurrently each receiving one of the treatments to which patients are randomised. In sequential trials, results are monitored throughout the trial and the trial is stopped when one treatment is shown to be better or if it is not likely a difference will emerge. A factorial design allows us to investigate the effect of two treatments individually compared to control, compared to each other and when used in combination. Adaptive design allows the investigator to use accumulating data to modify the trial without deterring the validity and integrity of the trial (Chuang-Stein and Beltangady, 2010). The form of design thereby offers a high amount of flexibility to the investigators and could be argued they are more ethical, as they allow a trial to be stopped early if it shows that a treatment is inferior or superior so that participants then needed are not recruited into the trial.

The RCT is said to be dependent on the random allocation process. The process involves generating an unpredictable programming sequence and implementing the sequence in a way that conceals the treatment until the patients have been formally assigned to their group.

Hinkelmann and Kempthorne (1994), explained that in the design of experiments, treatments are applied to experimental units in a treatment group. In comparative experiments, members of a control group receive a standard treatment (treatment as usual, TAU), a placebo, usual care (UC), or no treatment at all.

According to Vickers (2019), RCTs are the least biased research designs for evaluating new technologies and results from such trials are used by decision makers, such as the National Institute for Health and Care Excellence (NICE), to guide policy and practice. These trials are to be well designed with good power to provide solutions to important clinical questions. Underpowered or overpowered trials pose both statistical, practical, and ethical problems.

Randomisation is a method of experimental control that has been used in human biological experiments and clinical trials. It insures against accidental bias and prevents selection bias. It produces comparable groups and eliminates the source of bias in treatment assignments (see Suresh, 2011). Schulz and Grimes (2002) outlined the benefits of randomisation in RCTs to include masking the identity of treatments from participants, assessors, and invigilators; it permits the use of probability theory to express the likelihood that difference in outcome between treatment groups merely indicates chance and eliminates bias in treatment assignment. Frane (1998), Altman and Bland (1999), all state that randomisation provides a basis for the statistical methods used in analyzing the data, ensure that each patient has an equal or pre-specified chance of receiving any of the treatments under study and generate comparable intervention groups, which are alike in all the important aspects except for the intervention each group receives. The summary of benefits of randomisation includes elimination of the selection bias, balances the groups with respect to many known and unknown confounding or prognostic variables, and forms the basis for statistical tests. In general, a randomised experiment is an essential tool for testing the efficacy or effectiveness of a treatment.

Frane (1998), Altman and Bland (1999), further explained that, in practice, studies involve the generation of a randomisation schedule which should be reproducible and usually includes obtaining random numbers and assigning random numbers to each subject or treatment conditions. For simple experiments with a small number of subjects, randomisation can be performed easily by assigning the random numbers from random number tables to the treatment conditions while for the large sample size situation, or if restricted randomisation or stratified randomisation is to be performed, then randomisation is done using a computer environment such as R or

SAS or similar. In general, there are various types of randomisation schemes, which include simple, block, stratified, and covariate adaptive randomisation.

Altman and Bland (1999) explained that simple randomisation proceeds as a single sequence of random assignments. This technique maintains complete randomness of the assignment of a subject to a particular group. For an equal allocation ratio, a simple method of simple randomisation is flipping a coin. For example, with two treatment groups (control versus treatment), the side of the coin (i.e., heads - control, tails - treatment) determines the assignment of each subject. Other physical methods include using a shuffled deck of cards (e.g., black - control, red - treatment) or throwing a six-sided dice (e.g., below and equal to 3 - control, over 3 - treatment). A random number table found in a statistics book, or computer-generated random numbers can also be used for simple randomisation of subjects. It is easy to implement in clinical research. For large sample sizes, simple randomisation can be trusted to generate similar numbers of subjects among groups. However, randomisation results could be problematic in relatively small sample size clinical research, resulting in an unequal number of participants among groups. Simple random allocation is the most basic and easiest approach that provides unpredictability of treatment assignment, treatment is made by chance without regard to prior allocation.

Frane (1998), and Altman and Bland (1999), state that the method designed to randomize subjects into groups that result in equal sample sizes and ensure balance in sample size across groups is *block* randomisation. Balance in sample size may be achieved using this method.

The need to control and balance the influence of covariates is addressed by stratified randomisation, this method is achieved by generating separate blocks for each combination of covariates. This method controls for possible influence of covariates that will affect the conclusion of clinical trials.

Fleiss, Levin, Paik (2013) and Zalene (1990) explained covariate adaptive randomisation as valid alternative randomisation method for clinical research. In this method a new participant is sequentially assigned to a particular treatment group by considering the specific covariates and previous assignments of participants.

The CONSORT (CONsolidated Standards of Reporting Trials) guideline is intended to improve the reporting of parallel group randomised controlled trials (RCTs), enabling readers to understand a trial's design, conduct analysis and interpretation, and to assess the validity of its results. This can only be achieved through complete adherence and transparency by authors. Shamseer et al. (2016) recommends that all submissions to journals should follow the CONSORT guidelines. The guideline in point 7a of CONSORT, states that authors will explain how sample size was calculated.

Relatedly, the SPIRIT (Standard Protocol Items Recommendation for Interventional) guide aims to improve quality of protocols and to enable accurate interpretation of trial results. The SPIRIT 2013 Statement provides evidence-based recommendations for the minimum content of a clinical trial report (Chan et al., 2013). SPIRIT is widely endorsed as an international standard for trial protocols. The SPIRIT Statement also details the systematic development methods of the SPIRIT guidance and the scope. The SPIRIT checklist is endorsed by high-ranking peer-reviewed journals, research institutions, the Cochrane Consumer Network, and some pharmaceutical companies.

The **EQUATOR (E**nhancing the **QUA**lity and **T**ransparency **O**f health **R**esearch) collaboration is an international collaboration that intends to promote accurate reporting of health research studies to enhance the values and reliability of medical research literature. Its network was established to raise awareness of the importance of good reporting of research monitoring the status of the quality of reporting of research studies in the health sciences literature and conducting research relating to issues that impact the quality of reporting of health research studies including sample size (Simera and Altman (2009), Simera et al. (2010)).

The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (**ICH**), give Statistical Principles for clinical trials (E9) and states that a clinical trial can provide reliable answer to questions addressed by the trial, only if the sample size is large enough (p.1923) (Lewis, 1999). The primary objective of the clinical trial is to be used in determining the sample size but if the trial size is determined based on a secondary objective, or safety questions it should be stated clearly and justified. A trial size based on primary efficacy questions will need a smaller number of subjects than one based on secondary objectives, basic safety questions and requirements.

2.2 Sample size for Pilot studies

The sample size calculation being done correctly is important for ensuring the validity, reliability, and integrity of results from a clinical trial. Chow et al., (2017); Cohen (2013) and Winer et al. (1971) explained that to achieve the sample size of a study the standard deviation of the population needs to be known. This is usually not possible ahead of the research, hence the importance of a pilot study to determine the value of the standard deviation that can be used in calculating the sample size for a larger trial (Campbell et al., 2010).

According to Teare et al. (2014) the main weakness when estimating key parameters from small sample sizes is the large sampling variation. Pilot sample sizes, however, often imprecisely estimate the standard deviation for the main trial (Whitehead et al., 2016). Some methods proposed for correcting the inaccurate estimation of sample size by adjusting the prediction of the variance from pilot trial to be used in main trial sample size calculation is further presented.

Recommendations for pilot sample size vary in the literature. Neiswiadomy (2002) recommends approximately 10 participants per group for a pilot study. Julious (2005) suggests 12 per group. Hertzog (2008) outlined that a sample of 20-25 would be needed if the aim of a pilot study is to demonstrate intervention efficacy in a single group. Birkett and Day (1994) suggested 20 per arm for internal pilot studies. Browne (1995) mentions that the use of 30 is commonplace. Kieser and Wassmer (1996), state that a pilot sample size of 20-40 is sufficient for a sample size between 80 and 250 per arm. Connelly (2008) gave a circular suggestion that the sample size for a pilot study should be 10% of the unknown sample size needed for a definitive study. According to Hertzog (2008) there is insufficient literature concerning guidance on the size of a pilot study and choosing 10% might not be adequate as there are numerous factors that could influence a study. Teare et al. (2014) recommends ≥70 for both arms. It is well known all the rules used in picking pilot sample sizes have limitations (Whitehead et al., 2016).

In the context of a single arm study, and two arm studies, Browne (1995) commented that to compute the sample size needed to compare population means with a planned power, the researcher requires the population standard deviation. However, the value of the population standard deviation is rarely known and must be estimated. It suggested that to run a pilot study to estimate sigma (σ). The distribution of the sample standard deviation is positively skewed and as such over 50 per cent of the time, the sample standard deviation will be less than sigma (σ) hence more likely to underestimate sigma. Using the sample standard deviation will therefore, on average, underestimate sample size.

Lee et al. (2014) emphasized the importance of estimating the sample size for pilot trials to minimize the overall sample size required for both pilot and main trial. Also, they suggest considering other factors like plausible estimates of clinical effect through confidence intervals.

Browne (1995) investigated the frequency with which the actual power equals or exceeds planned power values when the standard deviation is estimated using (a) the unadjusted value of the pilot sample standard deviation, or (b) 100(1 - γ) per cent upper one-sided confidence limits.

Browne considers the one sample design using the one sample *t*-test, with an alternative hypothesis being a point specific value. It is assumed that this alternative hypothesis is precisely true in the simulations. Simulations are only done for alpha = 0.05 and for the single sample design with pilot sample sizes equal to 5, 10, 30, 50, 100. For target power of 0.8, for unadjusted analyses, power is lower than wanted but it increases with increasing sample size and increasing effect size. For target power of 0.8, for a 50% Upper Confidence Limit (UCL), power exceeds 50% and rises with increasing sample size and increasing effect size e.g., 75% for a medium to large effect using m = 30. Similar conclusions are drawn for other UCLs. The same conclusions hold for a target power of 0.9. The above conclusions hold for the one-sample paired *t*-test assuming the point alternative hypothesis is true.

Browne (1995) showed that simply using the standard deviation from a pilot study reduces the chances of achieving planned power to as low as 40 percent and the rule of thumb that proposes the use of m equal to 30 to estimate a parameter for a sample group will not solve the problem except for where the effect size is large. The research concluded that at least 80% upper one sided confidence limit on the variance will help improve the chance of achieving planned power in clinical trial to 80%.

2.3 Factors affecting power of a test

The concept of null hypothesis testing has been long established. The Type I error rate, denoted by alpha (α), is the probability of incorrectly rejecting H_0 and claiming an effect to be real when in fact the null hypothesis, H_0, is true. It is the likelihood that the study will reject the null hypothesis, assuming the null hypothesis to be true (Dalgaard, 2008). Statistical tests are developed so that under idealistic settings, the error rate can be set at a pre-study selected level. Conventionally, the level is typically set at α = 0.05, but α = 0.1, α = 0.025, α = 0.01, and α = 0.001 are alternatives depending upon the situation, and in particular, significance levels in tests of assumptions may be radically different from significance levels used in either decision making or in testing research hypotheses. The Type I error c]an be thought of as a false positive.

The Type II error, beta (β), is the pre-study probability of failing to reject the null hypothesis when in fact the null hypothesis is false. This error is known as false negative. The pre-study probability of correctly rejecting a false null hypothesis is $1 - \beta$ and represents statistical power. Note that both α and β are pre -study concepts which are specified in the design phase prior to data collection.

For a given data set and a given test statistic, the *p*-value is defined as the largest significance level for which there is failure to reject the null hypothesis. At the 5% significance level, the null hypothesis is to be rejected if the *p*-value is less than 0.05. The *p*-value relates to the null hypothesis; *p*-values do not relate to the alternative hypothesis.

Cronbach et al, (1972), Marcoulides (1993), state that larger samples more accurately represent the characteristics of the population from which they are derived assuming an unbiased sampling mechanism. For example, a mean obtained from random sample of *n* = 10 would allow a better estimate for the population mean than a random sample of *n* = 3. Intuitively, unbiased big samples will better estimate a parameter compared with unbiased small samples. Accordingly, unbiased large samples will have a greater chance of rejecting a false null hypothesis than unbiased small samples. It therefore follows that increasing sample size is associated with increase in power subject to all other factors remaining constant. Accordingly, the power of a test increases with increases in sample size.

Type I errors are more serious than Type II errors, and for this reason, it is conventional to set α less than or equal to β. The pre-study significance level also affects power. A study with α equal 0.01 requires more evidence to reject, H_0 than the same study with α set to 0.05. The higher evidential threshold for α = 0.01 compared with α = 0.05 necessarily means that α = 0.01 has lower power. Hence, increasing alpha (α) is associated with increasing power assuming all other factors remain constant.

The standardized effect size δ denotes the strength of a relationship or the magnitude of a difference relative to the variation. If effects are large and clear, then they will be easy to detect and establish. If effects are small and unclear, then it will be difficult to detect and establish. The level of ease or difficulty is synonymous with power. Increase in effect size is associated with increase in power assuming all other factors remain constant.

Another factor affecting power is the chosen test statistics. Different test statistics are chosen to be optimal in certain considerations. For instance, under an assumption of normality, a parametric test will have greater power than its non-parametric counterpart. Likewise, in the absence of normality, an appropriate non-parametric test may be more powerful. Accordingly, a critical reason over the choice of statistics is to choose a test statistic that has got greater power, but which would not inflate the Type I error rate if the null hypothesis is true. Incorrectly choosing a test statistic would more than likely result in a decrease in power or a failure to control the Type I error rate.

The design of a study affects power. For example, if a researcher decides to maximize effect size by maximizing the difference between or among independent variable levels in a study examining the effect of caffeine on performance, the likely effect in performance will be more apparent when the researcher compares individuals who ingest widely different amounts of caffeine e.g. (450mg vs 0mg) than for comparison of (25mg vs 0mg). Cohen, (1992) states that error variance due to factors other than the independent variables, decreases the likelihood of detecting differences or relationships that exist. Cohen (1992) and Conover (1980) state that for designing a study were there is a choice of generating paired data or unpaired data then we would opt for paired design as it gives a direct focus on the phenomenon of interest and at the analysis phase we can remove or account for some variation providing the paired data does not affect the validity of the study (e.g. learning effects, fatigue effects and so on, which in a paired design, could compromise validity).

Outliers affect the power of a test. Typically, very large outliers may obscure important effects resulting in a decrease in power, but medium sized outliers may lead to an increase in power unless robust statistics are used (Derrick et al, 2017).

2.4 Pilot and Feasibility Studies

The NETSCC (NIHR Evaluation, Trials, and Studies Coordinating Centre) which oversees managing National Institute for Health and Care Research (NIHR) evaluation research in the UK, describe a pilot study as 'a miniature version of the main test to tell if the main study components will all work together.' A pilot study should focus on how the trial will progress. It notes that for a trial to be classified as a pilot there needs to be a proposal for future study (among other criteria) (NETSCC, 2012)

Feasibility studies are intended to determine whether it is feasible and appropriate to perform a larger study (Thabane et al., 2010). These studies focus on data collection, methods, recruitment, and retention. Potential problems that might occur during the main study can be identified by researchers at this stage and solutions suggested. It therefore helps ensure that the primary investigation is carried out successfully, efficiently and with minimal risk to research participants (Thabane et al., 2010)

Pilot and feasibility studies are critical components of clinical research because they provide valuable information that can be used to refine study designs, identify potential challenges, and optimize study processes (Thabane et al., 2010). The studies are conducted as a preliminary step before undertaking a larger definitive study to ensure the main study will be conducted efficiently and effectively as possible. They are used to test the study design, the intended outcomes and any intervention that are under investigation to improve inclusion/exclusion criteria, estimate sample size and adjust research topics and study designs (Thabane et al., 2010)

According to Prescott and Soeken (1989), the aim of a pilot study includes feasibility and to help plan a larger study. Likewise, according to Thabane et al. (2010), a pilot study, is a small-scale preliminary study conducted to evaluate feasibility, duration, cost, adverse events, and improve upon the study design prior to performance of a full-scale research project. However, other authors make a distinction between a pilot study and a feasibility study. Specifically, Justis and Kreigsmann (1979) and Georgakellos and Macris (2009) define a feasibility study as an assessment of the practicality of a proposed project or system. A feasibility study aims to objectively and rationally uncover the strengths and weaknesses of trial opportunities and threats present in the natural environment, the resources required to carry through, and ultimately the prospects for success.

A pilot study is like the main study in many ways, including assessment of primary outcome, it is sometimes the first phase of substantive study. Data from the pilot phase may contribute to the main trial and this case is referred to as internal pilot (Thabane et al., 2010). If at the end of a pilot study data is not used in the main trial, then it is said to be an external pilot (Lee et al., 2014).

Feasibility and pilot studies have certain limitations. Pilot studies are frequently of a limited scale and could not be an accurate representation of the research population. Considering this limitation, results of a pilot study should be interpreted with caution and not the best for definitive trials (Thabane et al., 2010).

Prescott and Soeken (1989), suggest that despite limitations of small-scale studies, researcher can increase the overall quality and rigour of their research by performing pilot and feasibility studies, ensuring that the study is conducted effectively, efficiently and with minimal risk to study participants.

In practice, it could be the case that a pilot study quite legitimately has an element of feasibility, and a feasibility study could include an element of piloting, and as such a study might not fall neatly into one or the other.

2.5 Confidence Intervals for Means, Standard Deviation, Cohen's d and Coefficient of Variation

A confidence interval alongside a statistical test improves the usefulness of a report, and it shows how precise an estimate is (Funder et al., 2014). This will be very helpful and the results from Confidence Intervals for Means, Standard Deviation, Cohen's *d* and Coefficient of Variation will be outlined to gain a better understanding of this research.

A confidence interval (CI) is an estimate computed from the statistics of observed data and can be used to propose a range of plausible values for an unknown parameter (e.g., the mean). The interval usually has a corresponding confidence level that the true parameter is in the range of that interval, and the confidence interval for an unknown parameter is obtained from sampling the distribution of a corresponding estimator, Dekking (2005). Using a confidence interval of 95% would mean that 95 percent of the confidence intervals would contain the unknown population parameter value under repeated sampling (Swinscow and Campbell, 2002).

The effect size and its confidence interval for a sample comparison contains more information than a p-value, which itself is also an estimate. An effect size that is estimated using data from a large sample size is likely to be more accurate than one estimated from a data of small sample size (assuming unbiased sampling). Hence, the concepts of confidence intervals may be used to quantify the error imposed on an effect size. The interpretation of the confidence interval for an effect size or a standardised effect size is the same as that in the case of the CI of the mean. For all hypothetically sampled data from the same population and using the same sampling method, a population effect size would fall in 95% of the calculated effect size. Providing the effect size (point estimate) and CI (the precision of effects) are essential to understand the magnitude of intended treatment effects (Hedge and Olkin,2014).

Confidence intervals for one sample mean

A confidence interval (CI) for the mean is a range of values in which the population mean is expected to lie a pre-specified proportion of the time. Under an assumption of normality, a confidence interval for a one sample mean can be calculated using

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s the sample standard deviation, *t* is the critical value of the *t* distribution and is based on $n-1$ degrees of freedom.

Example 1

For a sample of 3539 randomly selected participants in a clinical trial, the mean weight is 127.3kg with a standard deviation of 19.1kg. The 95% confidence interval is

$$= 127.3 \pm 1.96\frac{19.1}{\sqrt{3539}}$$

$$= 127.3 \pm 0.63$$

$$= (126.7,127.9)$$

An estimate for the mean weight of the population is 127.3kg, and we are 95% confident that the true mean is between 126.7kg and 127.9kg. The margin of error is very small 0.63 because the sample size is large.

Confidence Interval for two means

There are situations where it is of interest to compare two groups with respect to their mean scores on a scale outcome. These situations involve comparisons between two independent groups. For normally distributed data, the 95% CI for the difference in population means is calculated by

$$CI_{(1-\alpha)} = (\bar{x}_1 - \bar{x}_2\ )\ \pm \left(t_{a/2}\right) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where t represents the $t_{a/2}$ tabled critical value in the t distribution for $n_1 +\ n_2 - 2$ degree of freedom (df) and where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled variance and is an unbiased estimator of σ^2 . Similarly,

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

is the pooled standard deviation, where n_1 is sample size for population 1, s_1^2 the sample variance of sample 1, n_2 is the sample size of sample 2 and s_2^2 is the sample variance of sample 2.

Example 2

The following data relate to systolic blood pressure for randomly selected participants receiving a treatment for a medication and the placebo group. The 95% confidence interval for the difference in sample means of the systolic bp in mmHg is obtained as follows.

Treatment group (x_1)	102.9, 100.2, 97.4, 97.0, 100.5, 101.9, 97.4, 99.9, 104.4, 101.3
Placebo group (x_2)	100.2, 96.9, 100.9, 101.0, 100.1, 95.3, 97.5, 98.0, 97.0, 96.0

$\bar{x}_1 = 100.2900$ $\bar{x}_2 = 98.2900$

$s_1^2 = 6.1166$ $s_2^2 =\ 4.3966$

$s_p^2 = 5.2565$

$s_p = \sqrt{5.2565} = 2.2927$

For these data, the 95% confidence interval for the difference in means is

$$2 \pm 2.101(2.2927)(0.4472)$$
$$2 \pm 2.101(1.0253)$$
$$(-0.1541, 4.1541)$$

Our best estimate of the difference, the point estimate, is 2.0 mmHg. The standard error of the difference is 1.0253 mmHg, and the margin of error is 2.1531 mmHg. We are 95% confident that the difference in mean systolic blood pressures between treatment and placebo group is between -0.1541 and 4.1541 mmHg. In this sample, the treatment group have higher mean systolic blood pressures than control by 2.0 mmHg. Based on this interval, we also conclude that there is no statistically significant difference in mean systolic blood pressures between treatment and placebo group, because the 95% confidence interval includes the null value, zero.

Confidence interval for variance and standard deviation

For independent identically distributed normal random variables $X_1, X_2, \dots X_n$ it is well known that

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi_v^2$$

where

$$s^2 = \sum\nolimits_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$$

And

$v = n - 1$ (Zar, 1984).

Accordingly, two-sided confidence intervals for σ^2 may be obtained via

$$\Pr\left(\chi_{v,\alpha/2}^2 < \frac{(n-1)s^2}{\sigma^2} < \chi_{v,\,1-\alpha/2}^2\right) = 1 - \alpha$$

so that

$$\Pr\left(\frac{(n-1)s^2}{\chi_{v,\,1-\alpha/2}^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_{v,\,\alpha/2}^2}\right) = 1 - \alpha$$

Accordingly, the upper limit of the two-sided confidence interval for σ^2 is given by

$$u_1 = \frac{(n-1)s^2}{\chi_{v,\,\alpha/2}^2}$$

For the same situation, a one-sided $(1-\alpha)\,100\%$ confidence interval, may be obtained from

$$\Pr(0 < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{v,1-\alpha}) = 1 - \alpha$$

i.e.

$$1 - \Pr\left(\frac{(n-1)s^2}{\sigma^2} < \chi^2_{v,1-\alpha}\right) = 1 - \alpha$$

$$\Pr\left(\frac{(n-1)s^2}{\sigma^2} < \chi^2_{v,1-\alpha}\right) = \alpha$$

Hence, the upper limit of the one-sided confidence interval is given by

$$u_2 = \frac{(n-1)s^2}{\chi^2_{v,1-\alpha}}$$

Clearly for any give α

$$\chi^2_{v,1-\alpha} < \chi^2_{v,\,1-\alpha/2}$$

and it follows

$$u_2 < u_1$$

It is u_2 which is used in Browne's formula. It is worth noting that since the chi-square distribution is not symmetric, we will be obtaining confidence intervals that are not symmetric about the point estimate.

Example 3

A random sample of 17 participants using a new medication is selected. Their ages was taken as follows 56, 30, 34, 77, 55, 67, 45, 65, 44, 47, 49, 60, 63, 64, 55, 67, and 88. Assuming the data is normally distributed with unknown mean μ and unknown variance σ^2 then the 95% two-sided confidence interval for the variance σ^2 and standard deviation σ is derived as follows:

$$n = 17$$

$$s^2 = 216.4044$$

$$s = 14.7107$$

$$\frac{(n-1)s^2}{\chi^2_{v,1-\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{v,1-\alpha/2}}$$

$$\frac{(17-1)14.7107^2}{28.845} < \sigma^2 < \frac{(17-1)14.7107^2}{6.908}$$

$$120.0372 < \sigma^2 < 501.2268$$

For the 95% confidence interval for standard deviation take the square root of both sides

$$\sqrt{120.0372} < \sigma < \sqrt{501.2268}$$

$$10.9561 < \sigma < 22.3881$$

Hence, we can be 95% confident the standard deviation for the ages of the participants is between 10.96 years and 22.39 years.

Confidence interval for Cohen's d

Cohens' d is a standardized effect size for measuring the difference between two group means and it is used when comparing control group to treatment group and it is an effect size suitable for the independent samples t test. It is suggested that values 0.2,0.5 and 0.8 represent small, medium, and large effects respectively (Baguley, 2009).

Cohen originally defined effect size for comparing two independent groups based on equal sample sizes as

$$d = \frac{\bar{x}_1 - \bar{x}_2}{(s_1 + s_2)/2}$$

To accommodate different sample sizes Hedge gave a more general form as

$$d_h = (\bar{x}_1 - \bar{x}_2)/s_p$$

where s_p is given by

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

In the literature both d (a special case of d_h) and d_h are invariably referred to as Cohen's *d*, and both d and d_h are *biased* estimators of

$$\delta = (\mu_1 - \mu_2)/\sigma$$

Hedge derived an unbiased estimator for δ , given by

$$h^* = Q(n_1 + n_2 - 2)d_h$$

where a simple but accurate approximation of h is given by

$$h = \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) d_h$$

This latter form, h, is known as Hedges unbiased estimator for the standardized two sample effect size (Hedges and Olkin, 2014).

The confidence interval for Cohen's d at 95 % confidence interval is

$$d - 1.96\,\sigma(d), \qquad d + 1.96\,\sigma(d)$$

where

$$\sigma(d) = \sqrt{\frac{n_1 + n_2}{n_1 \times n_2} + \frac{d^2}{2(n_1 + n_2)}}$$

where n_1 is the sample size of group 1 and n_2 is the sample size of group 2.

Example 4

For a clinical trial, the body weight of 30 randomly selected participants were taken. The participants were classified into the treatment group and control group, 15 per group. The 95% confidence interval for Cohen's d for the data is given as

Treatment group (x_1)	75, 63, 54, 76, 83, 92, 54, 67, 58, 45, 56, 89, 57, 75, 76
Control group (x_2)	65, 87, 94, 65, 87, 65, 65, 98, 76, 56, 76, 45, 70, 62, 71

$\bar{x}_1 = 68.0000$ $\bar{x}_2 = 72.1333$

$s_1 = 14.0915$ $s_2 = 14.4611$

$\sigma(d) = 0.3585$ $s_p = 70.0971$

$d_h = (\bar{x}_1 - \bar{x}_2)/s_p$ $h = (-0.0589\ )$

The approximate 95% confidence interval for Cohen's d is

$d_h - 1.96 \times\ \sigma(d), d_h + 1.96\ \times\ \sigma(d)$

$(-0.0589) - 1.96 \times (0.3585), (-0.0589) + 1.96\ \times (0.3585)$

$(-0.7237, 0.6815)$

The researcher is 95% confident that the Cohen's d is between -0.72 and 0.68.

Coefficient of variation

For a population or distribution, the coefficient of variation is defined as

$$CV = \sigma/\mu$$

and is used with ratio data. A sample estimate of the coefficient of variation is given by

$$c = s/\bar{x}$$

In the context of Cohen's d and sample size estimation, it is worth noting that

$$\frac{1}{d} = \frac{s_p}{\bar{y}}$$

where s_p is the pooled standard deviation and

$$\bar{y} = \bar{x}_1 - \bar{x}_2$$

For normally distributed data, a naive confidence interval for the coefficient of variation would have

$$lower\ limit = c\sqrt{\frac{(n-1)}{\chi^2(1-\alpha/2, n-1)}}$$

$$upper\ limit = c\sqrt{\frac{(n-1)}{\chi^2(\alpha/2, n-1)}}$$

However, these limits have less than ideal coverage.

McKay (1932) provides a seemingly better estimate of confidence intervals whereby

$$lower\ limit = \frac{c}{\sqrt{\left(\frac{u_1}{n} - 1\right)c^2 + \frac{u_1}{(n-1)}}}$$

$$upper\ limit = \frac{c}{\sqrt{\left(\frac{u_2}{n} - 1\right)c^2 + \frac{u_2}{(n-1)}}}$$

where $u_1 = \chi^2_{(1-\alpha/2, n-1)}$ and $u_2 = \chi^2_{(\alpha/2, n-1)}$

To further improve on accuracy, Vangel (1996) gave alternative limits

$$lower\ limit = \frac{c}{\sqrt{\left(\frac{u_1+2}{n} - 1\right)c^2 + \frac{u_1}{(n-1)}}}$$

$$upper\ limit = \frac{c}{\sqrt{\left(\frac{u_2+2}{n} - 1\right)c^2 + \frac{u_2}{(n-1)}}}$$

Example 5

A clinical trial team randomly selected 15 participants for a clinical trial. The body weight of the participants was taken. The mean weight of the participants is 81kg and standard deviation 19.52kg. The 95% confidence interval for the coefficient of variation using Naive, McKay, and Vangel method is given below.

$$n = 15$$

$$\bar{x} = 81.00kg$$

$$s = 19.5192k$$

$$c = 0.2410$$

$$u_1 = 26.1190$$

$$u_2 = 5.629$$

The approximate 95% confidence interval

Using the Naive method

For normally distributed data, a naive confidence interval for the coefficient of variation would have

$$lower\ limit = \frac{s}{\bar{x}}\sqrt{\frac{(n-1)}{\chi^2_{(1-\alpha/2,n-1)}}}$$

$$= \frac{19.5192}{81}\sqrt{\frac{(15-1)}{\chi^2_{(1-0.025,15-1)}}} \quad = 0.1764$$

$$upper\ limit = \frac{s}{\bar{x}}\sqrt{\frac{(n-1)}{\chi^2_{(\alpha/2,n-1)}}}$$

$$= \frac{19.5192}{81}\sqrt{\frac{(15-1)}{\chi^2_{(0.025,15-1)}}} \quad = 0.3801$$

The 95% confidence interval using Naive method is (0.1764,0.3801)

McKay (1932) provides a seemingly better estimate of confidence intervals whereby

$$lower\ limit = \frac{c}{\sqrt{\left(\frac{u_1}{n}-1\right)c^2 + \frac{u_1}{(n-1)}}}$$

$$lower\ limit = \frac{0.2410}{\sqrt{\left(\frac{26.1190}{15}-1\right)0.2410^2 + \frac{26.1190}{(15-1)}}}$$

$$= \ 0.1744$$

$$upper\ limit = \frac{c}{\sqrt{\left(\frac{u_2}{n}-1\right)c^2 + \frac{u_2}{(n-1)}}}$$

$$upper\ limit = \frac{0.241}{\sqrt{\left(\frac{5.629}{15}-1\right)0.241^2 + \frac{26.110}{(15-1)}}}$$

$$= \ 0.3985$$

The 95% confidence interval using McKay's method is (0.1744,0.3985)

To further improve on accuracy, **Vangel (1996)** gave alternative limits:

$$lower\ limit = \frac{c}{\sqrt{\left(\frac{u_1 + 2}{n} - 1\right)c^2 + \frac{u_1}{(n-1)}}}$$

$$= \frac{0.2410}{\sqrt{\left(\frac{26.1190 + 2}{15} - 1\right)0.2410^2 + \frac{26.1190}{(15-1)}}}$$

$$= 0.1740$$

$$upper\ limit = \frac{c}{\sqrt{\left(\frac{u_2 + 2}{n} - 1\right)c^2 + \frac{u_2}{(n-1)}}}$$

$$= \frac{0.2410}{\sqrt{\left(\frac{5.629 + 2}{15} - 1\right)0.2410^2 + \frac{5.629}{(15-1)}}}$$

$$= 0.3944$$

The 95% confidence interval using Vangel method is (0.1740,3944).

The results from the three methods do not seem to differ much based on this example.

2.6 Summary

There was an extensive review of the literature on sample size, sample size for pilot studies, randomised trials, factors affecting power of a test, pilot and feasibility studies, confidence intervals, and coefficient of variations. The confidence interval and co-efficient of variations were better illustrated with practical examples. The reviewed literature elaborated on some challenges in standard deviation estimation. These findings in addition to literature in the first chapter will lead to investigation of new methods. The next chapter will present methodology for achieving this research.

Chapter 3

Methodology

The chapter presents the procedure for assessing the formula given by Browne (1995) and other proposed methods. It includes discussion of simulation and random numbers, Monte-Carlo simulations, the number of iterations, and research evaluation metrics.

3.1 Simulations and Random numbers

According to Banks et al, (2001), a simulation, in general, is an approximate imitation of the operation of a process or system. Sokolowski and Banks (2009) states that a simulation can be used to show the eventual real effects of alternative conditions and courses of action and is also used when the real system cannot be engaged, because it may not be accessible, or it may be dangerous or unacceptable to engage, or it is being designed but not yet built, or it may simply not exist. *In-silico* simulation studies use computer intensive procedures to assess the performance of a variety of statistical methods in relation to a known truth (Burton et al., 2006). Such evaluation cannot be achieved with studies of real data alone. The research details that when doing a simulation study, it is important to consider specific objectives of the study, determine the procedures for generating the data sets and the number of simulations to perform. These techniques provide empirical estimation of the sampling distribution of the parameters of interest that could not be achieved from a single study and enable the estimation of accuracy measures, such as the bias in the estimates of interest, as the truth is known. Whenever a new statistical method is developed, there are assumptions that need to be tested and confirmed. Statisticians use simulated data to test such assumptions or investigate the effect of their violation. Simulated data is cheap because it uses random numbers generated rather than primary collected data, it saves time as it is faster to achieve, and results of simulation statistics can approximate real result and may be replicated.

Physical random numbers are problematic because of the difficulty of storage, the need for frequent testing for randomness and the fact they are not generated in the computer but by an external source (see Ahrens et al., 1970). The physical random numbers that sound convincing from an axiomatic standpoint suffer from the practical deficiency that no concrete sequence used in a calculation can verifiably satisfy the definition. Therefore, one has taken recourse to sequences that make no pretense of being "random" in any meaningful sense of the term, but which can be readily generated in the computer by simple arithmetic algorithms, while still passing an assortment of statistical tests for randomness. The terms of such sequences are collectively (and loosely) called pseudo-random numbers (PRNs). It should be emphasized that no such sequence can perform well under all imaginable tests for randomness. Rather, the user of PRNs must be aware of the specific statistical properties that are desirable in a simulation calculation and choose PRN that are known to pass these tests. Overall, PRNs have a record of meeting any reasonably limited set of statistical requirements if adequately chosen for the particular purpose.

Barker et al. (2012), explained a pseudo-random number generator (PRNG) to be an algorithm for generating a sequence of numbers whose properties approximate the properties of sequences of random numbers. The PRNG is completely determined by an initial value hence the generated sequence is not truly random. The initial value is called the PRNG's seed (which may include truly random values) and will be represented as set.seed in the R programming codes in this research. The set.seed makes it possible to replicate the studies using the set value. PRNGs are central in applications such as simulations e.g. for the Monte Carlo simulation, (Maigne et al, 2004).

The generation of pseudo-random numbers is important for statistical computing (Scott, 2011). For well-tested pseudo-random generators for the uniform distribution, the probability integral transform may be employed to provide an exact algorithm for transformation to any desired probability distribution.

The generation of independent normally distributed random variables (Gaussian random variables) is paramount as an assumption of normality is frequently made in the development of parametric statistics. The Box–Muller transformation, (Box and Muller, 1958) is a pseudo-random number sampling method for generating pairs of independent, standard, normally distributed (zero expectation, unit variance) random numbers, given a source of uniformly distributed random numbers.

3.2 Monte-Carlo Methods

Experimental mathematics concerned with experiments on random numbers is known as Monte-Carlo methods and is a broad class of computation algorithms that rely on repeated random sampling to obtain numerical results. It is applicable in a field with variety of problems with limited resources of theoretical mathematics It uses randomness to solve problems that might be deterministic in principle (Kroese, 2014) and the terminology was made popular by von Neumann and Ulam (1951). It is very useful for simulating phenomena with significant uncertainty in inputs and systems with many coupled degrees of freedom. Random sampling of numbers is achieved by simulations and the simulation technique is an application of Monte Carlo methods .

The standard of Monte Carlo experiment in statistics were set by Cassey (2014). Monte Carlo, when used in applied statistics, gives the possibility of reducing error and infinitesimally small treatment effects as real data often do not fit abstract distribution (Maggio and Sawilowsky, 2014). Serlin (2000) explained that when a test statistic is proposed, the robustness for validity (Type 1 error rate) and efficiency (power) can be explored using simulation. Simulations in this research are performed in R, using various versions of it (R Core Team, 2020; R Core Team, 2022),with the R studio interface (R Studio Team, 2020; R Studio team 2022).The corresponding code for each study will be presented in Appendix B.

3.3 Number of iterations

A simulation will comprise a number of iterations. The number of iterations required is research dependent. The appropriate number of iterations is a critical aspect of conducting simulations in R. Inaccurate estimates may be obtained if an insufficient number of replications are used, (Kim, 2005). According to Kocak (2019) in Monte Carlo simulation studies, increasing the number of iterations helps in producing data with less error estimation and the number of iterations can be altered depending on the desired level of accuracy. Browne (1995) used 2000; Kieser and Wassmer (1996) used 2000, Whitehead et al., (2016) used 10,000, and in examining small sample behavior of various statistics Browne and Forsythe (1974) used 10,000, Wynants et al., (2015) used 100,000 in their research on a simulation study of sample size.

Considering the interest for best results, improved accuracy, and the range of values of the previous research in this area 100,000 iteration will be used in replication of Browne’s method (objective 1) in contrast to 2,000 previously used in their research and for the study of the other objectives in this research. The parameter combinations used in this research will be based on Browne’s parameter combination for effective comparison and will be described in detail for e]ach study objective.

3.4 Research Evaluation Metrics

Browne’s formula of sample size estimation was reviewed using simulation and the percentage of over and under estimation was considered. To evaluate the predictive accuracy of the formula in estimating the sample sizes the median percentage error will be used in this research.

Median Percentage error

The Median percentage error is used’ to assess the accuracy of the model which is crucial for informed decision using the model. It can be used to evaluate how accurate a model predicts data while considering precision. The MPE is used to evaluate the accuracy of a predictive model and it can be used as a robust measure of central tendency for error making it a very useful tool. It is determined by finding the median of the percentage errors (PE) of each observation where

$$PE = \frac{predicted\ value - actual\ value}{actual\ value} x\ 100$$

MPE provides direction of magnitude of error with positive values implying overestimation and negative indicating underestimation . The value of MPE being close to zero implies minimum bias and smaller values of MPE shows better model while large values show greater deviation (Hyndman and Koehler, 2006).

In the context of this research the percentage error for any iteration is given by

$$Percentage\ error = \frac{\hat{n} - n_{true}}{n_{true}}\ x\ 100$$

where $\hat{n}$ is the estimated value and n_{true} the true value. For any set of parameter combinations there will be Maximum Iteration (Maxiter) percentage errors. The median of this is the median percentage error.

Mean median percentage error (MMPE)

This evaluation technique combines both median and mean approaches advantages, it is equally used for the evaluation of accuracy of a model. It involves calculating the median of the percentage error across all observations and then taking the mean of each median percentage error. It uses the robustness of the median to mitigate the influence of the outliers, providing a balanced measure of forecast accuracy . This metric can help determine the most effective model by covering enhanced robustness against outliers while retaining interpretability.

$$\text{MMPE } = \frac{1}{n}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{PE_i}|$$

where n is the total number of observations and PE_i is the percentage of error for the i^{th} observation.

When the value of MMPE is close to zero it implies minimal bias, positive values imply overestimation and negative values imply underestimation. However, in the context of this research the MMPE is in relation to the regression model that will be developed for MPE and sample size. The regression model is represented by

$$y = \alpha + \beta x$$

The regression model of MPE for study one is given below

$$1/_{MPE} = \alpha + \beta\sqrt{m}$$

With the structural part of the regression model being for the mean MPE.where $1/_{MPE}$ is the dependent variable, α is the intercept, β is the slope and m is the sample size of the regression model. Hence, the regression is for the mean median percentage error and will be achieved using simulations in r programming.

Chapter 4

Review of Browne’s approach of sample size estimation

Browne’s approach of sample size estimation is reviewed using simulations to evaluate its effectiveness. The parameter combinations used is like Browne’s approach, however the number of iterations was increased to enhance accuracy of the results and considering other literatures. Additionally, the algorithm, flowchart is shown detailing the codes implementation that was done using r programming. The findings of the study are presented in tables and graphics for better analysis of study results.

4.1 Review of Browne’s approach of sample size estimation (Study 1a)

4.1.1 Sample size formulae

According to van Belle and Martin (1993), for the two-group problem (assuming normality), the formula to determine the sample size (n) for testing at the α significance level with power equal to 1 – β is given by

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2} \qquad (4.1)$$

where n denotes the sample size per arm needed; $Z_{1-\alpha/2}$ is the normal deviate for the alpha significance level, $Z_{1-\beta}$ is the normal deviate for power , σ_1^2 is the variance of population 1, σ_2^2 is the variance of population 2, μ_1 is the mean for population 1, and μ_2 is the mean for population 2. For proof of this formula see (Arnold 1990; Moore and McCabe 1989).

For equal variances the formula becomes

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2(2\sigma^2)}{(\mu_1 - \mu_2)^2} \qquad (4.2)$$

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2} \qquad (4.3)$$

where $\delta = \ (\mu_1 - \ \mu_2)/\sigma$ is the population standardized effect size.

In the above it is assumed that all the parameters are known. In practice the parameters will not be known and will be estimated from small scale pilot data. For assumed equal variances, a naive estimated sample size would then be

$$\hat{n} \ = \ \frac{(Z_{1-\ \alpha/2} + Z_{1-\beta})^2\ (2\hat{\sigma}^2)}{(\bar{x}_1 - \ \bar{x}_2)^2} \qquad (4.4)$$

$$= \ \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2} \qquad (4.5)$$

where d is the sample estimate of Cohen's d defined as the difference between two means divided by a standard deviation for the data, i.e., $d = (\bar{x}_1 - \bar{x}_2)/s$.

More generally, under an assumption of normality, for any given parameters, the required minimum sample size, N_{TRUE}, is given by Whitehead et.al. (2016) as

$$N_{true} = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\ \beta)}\right)^2}{(\mu_1 - \ \mu_2)^2} \ .\sigma^2 \qquad (4.6)$$

where r denotes the allocation ratio of participants between the placebo and the treatment group.

The estimated sample size, $\widehat{N}_B$, using Browne's approach would be given by

$$\widehat{N}_B = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\ \beta)}\right)^2}{(\mu_1 - \ \mu_2)^2} \ .ks^2 \qquad (4.7)$$

where the multiplier $k = \ (m_1 + \ m_2 - 2)/(\chi^2_{v,1-\gamma})$.

However, in most practical situations, the difference between distributional means might not be known, and hence a naive estimation of the sample size will be

$$\widehat{N} = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + \ Z_{(1-\ \beta)}\right)^2}{(\bar{x}_1 - \ \bar{x}_2)^2} \ .ks^2 \qquad (4.8)$$

The work of Browne (1995) provides a mechanism of estimating a sample size which will be equal to or exceed the desired true sample size with a preselected probability, $(1 - \gamma)$. For instance, $(1 - \gamma)$ = 0.8 would correspond to an estimated sample size which will be equal to or exceed, N_{true}, 80% of the time. However, Browne did not consider the magnitude of the difference between $\widehat{N} - N_{true}$. The first stage simulation would be to reproduce the work of Browne and to further consider the magnitude of $\widehat{N} - N_{true}$ which is how many times the estimate exceeds the true value. This was not considered by Browne.

In Browne's original work the null hypothesis is given by H_0: $\mu_1 = \mu_2$ and under an assumption of normality. Under the alternative hypothesis, $\mu_1 - \mu_2$ taken to be a prespecified minimally clinical important difference.

Table 4.1: Parameter combinations.

Factor	Number of Levels	Levels
Alpha(α)	2	0.01, 0.05
Beta (β)	2	0.1, 0.2
Coverage ($1-\gamma$)	2	0.8, 0.9
Effect size (δ)	3	0.1, 0.4, 0.75
Pilot sample size (m)	5	5,10, 30, 50,100

Hence the design corresponds to a 2 by 2 by 2 by 3 by 5 fully crossed design. 100,000 replicates will be conducted at each cell combination (contrast with Browne who undertook 2000 replicates per cell).

4.1.1.1: Algorithms for study one (a)

1. Start at set. seed 462
2. Set iteration =100,000
3. Define ranges of values of α=(0.01,0.05), β=(0.10,0.20), m= (5,10,30,50,100), $1-\gamma$ = (0.8,0.9), and δ =(0.1,0.40,0.75)
4. Work out $Z_{1-\alpha/2}$ and $Z_{1-\beta}$, $(2m-2)/\chi^2_{1-\gamma,2m-2}$
5. Loop over α, β,m, $1-\gamma$ and δ
6. Calculate N_{true}
7. Generate Sample 1 of size m using iid Normal (0,1)
8. Generate Sample 2 of size m using iid Normal (effect size, 1)
9. Calculate N-estimated
10. Store $(N_{est} - N_{true})$
11. Store Nest > (N_{true}+ P* N_{true}), (P= -0.2, -0.1, 0.0, 0.1, 0.2, 0.3, 0.5, 0.75, 1.0,

1.5, 2.0)

12. Loop for each combination for all iterations
13. Save results
14. End

The R code for the study is given in Appendix B and the flowchart is given in Figure 4.1.

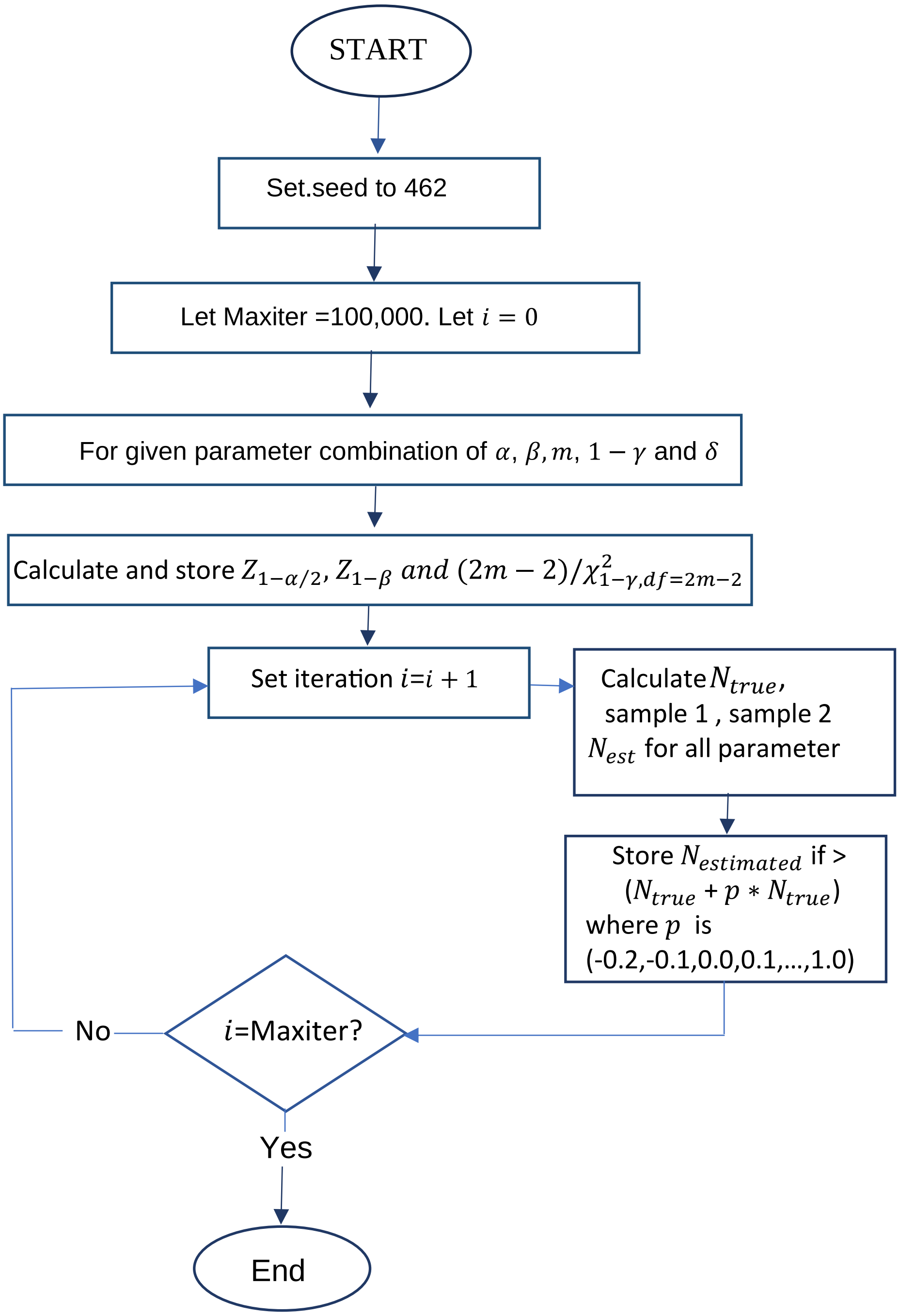

START
Set.seed to 462
Let Maxiter =100,000. Let $i = 0$
For given parameter combination of α, β, m, $1-\gamma$ and δ
Calculate and store $Z_{1-\alpha/2}, Z_{1-\beta}$ and $(2m-2)/\chi^2_{1-\gamma,df=2m-2}$
Set iteration i=$i+1$
Calculate N_{true}, sample 1 , sample 2 N_{est} for all parameter
Store $N_{estimated}$ if > $(N_{true} + p * N_{true})$ where p is (-0.2,-0.1,0.0,0.1,…,1.0)
i=Maxiter?
No
Yes
End

Figure 4.1: Flowchart for Study 1(a).

4.1.2 Results from Browne's method presented in tables

The simulation approach is used to generate results using Browne's method under the normality assumption. The simulation is done at varying parameter combinations and presented in the tables below. The results of the simulation are summarized in Table 4.2 through to Table 4.5. The remaining result not presented here is in Appendix C.

Table 4.2 considers the parameter settings $\alpha = 0.05$, $\beta = 0.2$, $(1-\gamma) = 0.8$,and, as required, 80% of the time the estimated sample size is equal to, or larger than, N_{true}. As the pilot sample size increases the degree of excess decreases and this is true for every effect size. With a very small pilot sample sizes of $m = 5$ per group, there is in excess of 50% chance of the estimated sample size being overestimated by more than 50%, there is in excess of 30% chance of the estimated sample size being in excess of 100%, and there is in excess of a 15% chance of the estimated sample size being in excess of 150%, and this is true for all effect sizes.

With a small pilot sample size of $m = 10$ per group there is in excess of a 30% chance of the sample being in excess of 50%, there is in excess of a 10% chance of the estimated sample size being over-estimated by 100%, there is in excess of 35% chance of the estimated sample size being over-estimated by 50%, and there is in excess of a 50% chance of the estimated sample size being in excess of 30%, and this is true for all effect sizes.

With a moderate pilot sample size of $m = 30$ per group, there is more than a 40% chance of the sample size being overestimating by more than 20%. Even with moderately large sample sizes of $m = 50$ per group, there is more than a 30% chance that the sample size will be overestimated by 20% and there is a non-trivial 10% chance that the estimated sample size will be more than 30%, and this is true for all effect sizes.

Even with a large sample size of $m = 100$ per group there is more than a 10% chance that the sample size will be overestimated by 20%.

Table 4.2: Proportion of time estimated sample size exceeds N_{true} +/- error at α = 0.05, β = 0.2, $(1-\gamma)$ = 0.8.

Sample size	Effect size	>-20%	$>N_{true}$	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.884	.799	.701	.650	.547	.325	.175
10	.10	.922	.800	.634	.544	.377	.107	.021
30	.10	.973	.799	.452	.289	.086	.001	.000
50	.10	.989	.797	.334	.159	.019	.000	.000
100	.10	.998	.801	.164	.035	.000	.000	.000
5	.40	.883	.799	.698	.646	.544	.325	.175
10	.40	.920	.800	.629	.539	.367	.105	.021
30	.40	.972	.801	.450	.284	.079	.001	.000
50	.40	.988	.801	.332	.154	.018	.000	.000
100	.40	.998	.801	.159	.032	.000	.000	.000
5	.75	.876	.801	.694	.638	.547	.325	.174
10	.75	.910	.801	.617	.520	.370	.106	.020
30	.75	.963	.798	.423	.255	.085	.001	.000
50	.75	.982	.799	.327	.126	.018	.000	.000
100	.75	.996	.799	.130	.023	.000	.000	.000

Table 4.3 uses α = 0.05, β = 0.1 and $1-\gamma$ = 0.8 and this is almost identical to Table 4.2 the only parameter that has changed is β. The results of both tables are almost identical, and hence β does not have an impact on the degree of excess.

Table 4.3: Proportion of time estimated sample size exceeds N_{true} +/- error at α= 0.05, β=0.1, $(1-\gamma)$=0.8.

Sample size	Effect size	>-20%	$>N_{true}$	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.884	.800	.701	.650	.549	.326	.175
10	.10	.922	.799	.631	.541	.372	.105	.021
30	.10	.974	.801	.454	.290	.085	.001	.000
50	.10	.989	.800	.338	.159	.019	.000	.000
100	.10	.998	.800	.166	.035	.000	.000	.000
5	.40	.883	.800	.701	.651	.545	.327	.175
10	.40	.918	.798	.627	.539	.367	.106	.021
30	.40	.971	.800	.452	.291	.083	.001	.000
50	.40	.987	.801	.329	.155	.017	.000	.000
100	.40	.998	.801	.158	.034	.000	.000	.000
5	.75	.886	.803	.702	.646	.538	.326	.174
10	.75	.921	.800	.628	.531	.354	.105	.020
30	.75	.972	.799	.445	.273	.072	.001	.000
50	.75	.988	.804	.331	.143	.014	.000	.000
100	.75	.998	.801	.155	.029	.000	.000	.000

Table 4.4 uses α= 0.01, β = 0.1 and $1-\gamma$ = 0.8 this is almost identical to Table 4.3 the only parameter that has changed is α . The proportion of excess in both tables are almost identical, this shows the degree of excess is not caused by changing α .

Table 4.4: Proportion of time estimated sample size exceeds N_{true} +/- error at α= 0.01, β=0.1, $(1-\gamma)$=0.8.

Sample size	Effect size	>-20%	$>N_{true}$	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.885	.800	.702	.651	.549	.328	.177
10	.10	.922	.801	.632	.543	.375	.107	.022
30	.10	.973	.799	.452	.291	.086	.001	.000
50	.10	.988	.800	.338	.160	.019	.000	.000
100	.10	.998	.803	.165	.035	.000	.000	.000
5	.40	.885	.800	.698	.648	.548	.325	.175
10	.40	.921	.800	.628	.541	.376	.107	.020
30	.40	.973	.799	.443	.286	.086	.001	.000
50	.40	.989	.801	.329	.157	.019	.000	.000
100	.40	.998	.801	.156	.034	.000	.000	.000
5	.75	.882	.801	.698	.650	.544	.325	.173
10	.75	.915	.802	.625	.539	.366	.107	.021
30	.75	.968	.799	.436	.283	.078	.001	.000
50	.75	.986	.800	.316	.152	.017	.000	.000
100	.75	.997	.800	.145	.032	.000	.000	.000

Table 4.5 uses α = 0.01, β = 0.1 and $1-\gamma$= 0.9 and these parameters are like the parameters in Table 4.4 except the parameter that has changed is coverage (from 80% to 90% coverage). Comparing the two tables there is a very large change in the proportion of error values. This shows that coverage does affect the error value in the outcome of the result.

Table 4.5: Percentage of time estimated sample size exceeds N_{true} +/- error α=0.01, β=0.1,$(1-\gamma)$ =0.9.

Sample size	Effect size	>-20%	$>N_{true}$	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.948	.901	.842	.808	.734	.541	.368
10	.10	.967	.900	.789	.721	.574	.246	.078
30	.10	.991	.900	.638	.471	.196	.005	.000
50	.10	.997	.901	.520	.303	.058	.000	.000
100	.10	.999	.900	.307	.092	.002	.000	.000
5	.40	.947	.902	.840	.807	.733	.539	.367
10	.40	.966	.901	.788	.722	.573	.247	.076
30	.40	.990	.898	.631	.467	.193	.005	.000
50	.40	.996	.899	.508	.299	.057	.000	.000
100	.40	1.000	.900	.293	.090	.002	.000	.000
5	.75	.944	.898	.835	.803	.727	.538	.362
10	.75	.963	.899	.781	.716	.561	.240	.072
30	.75	.988	.900	.620	.462	.181	.005	.000
50	.75	.995	.900	.499	.294	.052	.000	.000
100	.75	.999	.900	.278	.086	.002	.000	.000

Table 4.6 uses α = 0.05, β = 0.2 and $1-\gamma$ = 0.8 and shows the proportion of error is similar even when the effect size is varied. For under estimation by 20% the proportion are .884, .883 and .886 for effect size 0.1, 0.4 and 0.75 respectively, the results are very similar. For over estimation by 20% the proportions are 0.701, 0.698 and 0.694 respectively which are all approximately 0.7. This shows effect size does not have an impact on the proportion of error generated in sample size estimation using Browne’s method.

Table 4.6: Proportion of time estimated sample size exceeds N_{true} +/- error at α=0.01, β=0.2, $(1-\gamma)$ =0.9 considering impact of effect size for different percentages of error.

m	% under or overestimation	δ 0.1	 0.4	 0.75
5	-20	.884	.883	.886
10		.922	.918	.921
30		.974	.971	.972
50		.989	.987	.988
100		.998	.998	.998
5	0	.799	.799	.800
10		.800	.800	.801
30		.799	.801	.798
50		.797	.801	.799
100		.801	.801	.799
5	20	.701	.698	.694
10		.634	.629	.694
30		.452	.450	.423
50		.334	.332	.327
100		.164	.159	.130
5	30	.650	.646	.634
10		.544	.539	.520
30		.289	.284	.255
50		.159	.154	.126
100		.035	.032	.023
5	100	.325	.325	.325
10		.107	.105	.106
30		.000	.001	.000
50		.000	.000	.000
100		.000	.000	.000
5	150	.175	.175	.174
10		.021	.021	.020
30		.000	.000	.000
50		.000	.000	.000
100		.000	.000	.000

4.1.3 Graphical representation of results from study 1

By way of illustation, Figure 4.2 shows the proportion of overestimation of sample size by more than 30%. It can be seen that there is no difference due to $\boldsymbol{\delta}$, no difference attributable to $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, and hence these three parameters do not affect proportion of error. There is change due to coverage with 0.9 coverage level giving higher error in sample size estimation. Figure 4.2 shows also that as the sample size increases the error values reduces and unlikely to overestimate by 30% with pilot sample sizes as large as $\boldsymbol{m} = \mathbf{100}$ the value tends to zero at that point.

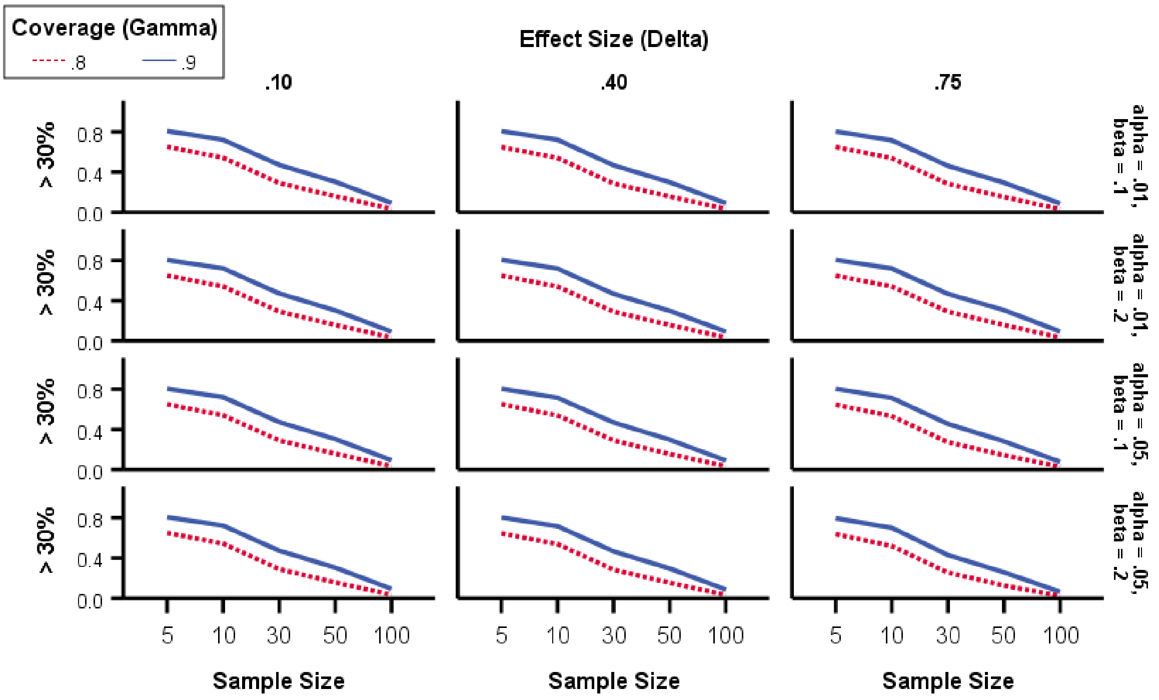


Figure 4.2: Graphical representation for $N_{true} > 30\%$.

Figure 4.3 shows the proportion of overestimation of sample size by more than 100%. There is no change due to δ, the coverage showed a change (higher error for higher values of coverage), but the error tends to zero with increasing pilot sample size. For $m = 30$ the curves flatten i.e., $, m \geq 30$ for the parameter combinations ensures that more than 100% overestimation is unlikely.

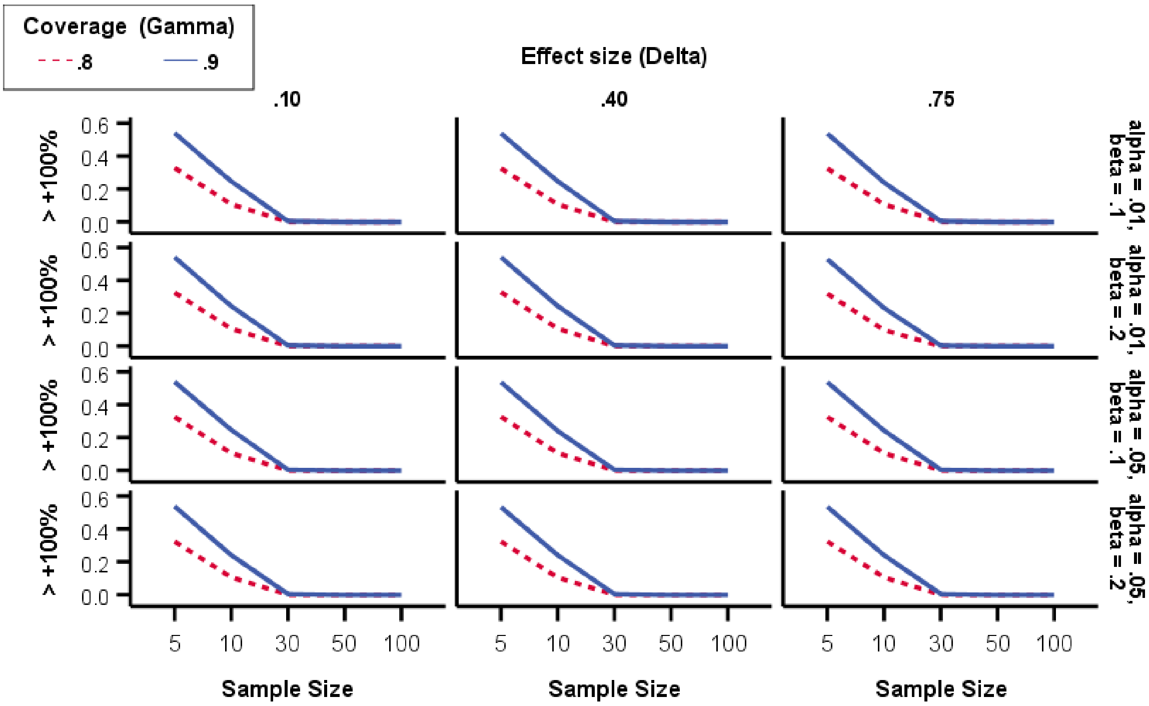


Figure 4.3: Proportion of times that the sample size is overestimated by more than 100%.

Figure 4.4 is the proportion of times that the sample size is overestimated by more than 150%. For pilot sample size ≥ 30 the error flattens. It is unlikely that the error will exceed 150% of the true required sample size. It further shows the effect size does not cause a change, there is change due to coverage and there is change as the pilot sample sizes increased. When the sample size becomes 30 both error margins flatten.

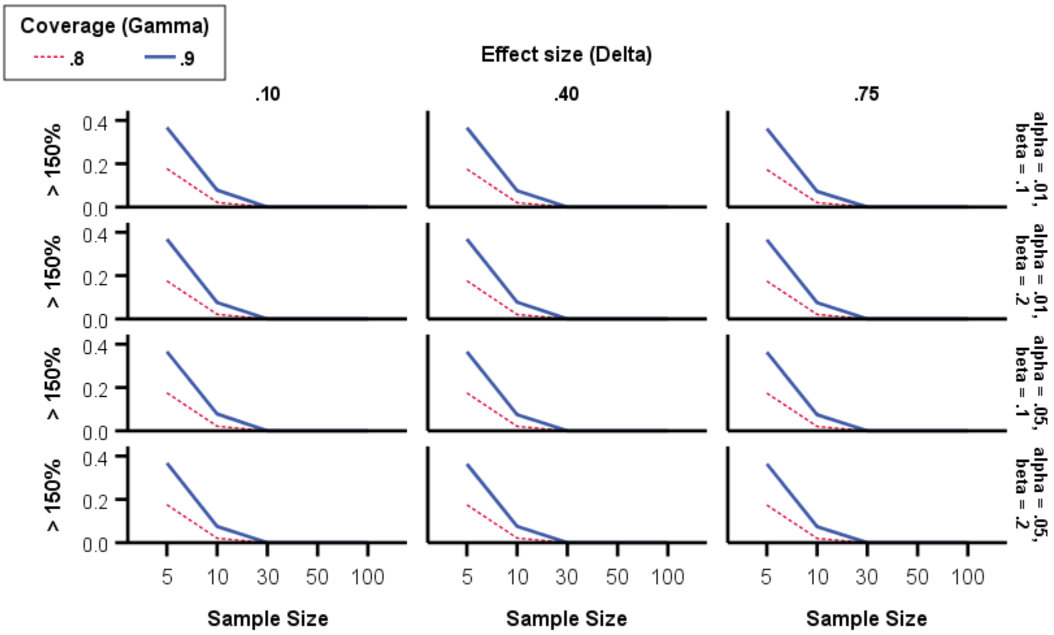


Figure 4.4: Proportion of times that the sample size is overestimated by more than 150%.

4.2 Median Percentage Error of study 1

MPE is used to evaluate the performance of a formula in sample size prediction. This indicates the percentage by which the sample size deviates from the true or optimal sample size. The median percentage error will show the impact of different parameters. The median percentage error will be used to measure the accuracy of Browne's approach for sample size estimation.

4.2.1 Median Percentage error (MPE) for study 1a in tables

Table 4.7 and 4.8 gives median percentage error summary for the various parameter combinations.

At β=.10, m=5, δ=.10 and α=$.01$ the MPE is approximately 60%, other parameters remaining the same and α changes to .05 the MPE remains the same. For α=$.01$, m=10, δ=.40 and β=.10, the MPE is approximately 35%, other parameters remaining the same and when β changes to .20 the MPE remains the same. This result shows that median percentage error is not dependent on α or β as the median percentage error value approximately same despite the changes in both α and β values. There is negligible impact by δ.

For α=.01, β=.10, δ=.10, m=5 the MPE is approximately 60%, when pilot sample size m=10 is approximately 35%, at 30 is approximately 17% , at 50 is approximately 13% and at 100 is approximately 9% . The MPE is decreasing with increase in pilot sample size hence it is said to have an impact. Table 4.8 shows MPE for coverage of 0.9 and the error increases with increase in coverage hence coverage has an impact on the MPE. Both Table 4.7 and 4.8 show that Browne's method can lead to overestimation of sample sizes

Table 4.7: Median percentage error for study one at $(1-\gamma)$ = 0.80

						m		
$(1-\gamma)$	δ	α	β	5	10	30	50	100
.80	.10	.01	.10	60.49	34.96	17.39	13.04	8.96
			.20	59.60	34.97	17.32	13.03	8.92
		.05	.10	60.25	34.85	17.40	12.98	8.92
			.20	59.32	34.42	17.53	13.10	8.95
	.40	.01	.10	59.88	34.77	17.39	13.09	8.97
			.20	60.59	34.55	17.45	13.19	8.99
		.05	.10	59.24	34.34	16.68	12.46	8.49
			.20	58.80	33.71	16.55	12.03	7.89
	.75	.01	.10	60.49	34.21	17.25	12.97	8.75
			.20	58.05	33.41	16.08	11.81	7.77
		.05	.10	57.33	32.68	15.46	11.08	7.16
			.20	59.43	34.70	17.03	12.68	8.61

Table 4.8: Median percentage error for study one at $(1-\gamma)$ = 0.9.

						m		
$(1-\gamma)$	δ	α	β	5	10	30	50	100
.90	.10	.01	.10	110.21	59.93	28.32	20.82	14.04
			.20	110.79	59.81	28.27	20.92	14.08
		.05	.10	109.87	59.13	28.40	20.81	14.01
			.20	110.29	59.82	28.23	20.81	14.08
	.40	.01	.10	110.76	59.34	28.40	20.82	14.07
			.20	109.98	59.82	28.35	20.90	14.08
		.05	.10	108.99	58.47	27.57	20.22	13.49
			.20	108.77	58.05	27.13	19.82	13.05
	.75	.01	.10	109.62	59.19	27.92	20.53	13.91
			.20	107.98	57.75	26.65	19.50	12.84
		.05	.10	106.64	56.81	26.08	18.80	12.16
			.20	109.98	59.29	27.98	20.52	13.65

4.3 SUMMARY

Browne’s method of sample size estimation was explained. Using simulation, the results were replicated at various parameter combinations, the proportion of under/over estimation results showed that sample size could be estimated by this method however it could lead to over estimation of sample sizes by over 100%. α and β were seen to have little to no impact on the error margin, δ had negligible impact, $1-\gamma$ and sample size had great impact on the proportion of error. As m increased the margin of error reduced. The MPE was calculated and showed the same impact by the parameters as in the over/under estimation tables.

CHAPTER 5

The Goldilocks ("Just about right") approach

It is of ethical importance in sample size estimation to have research done with the least possible sample size but sufficiently large to draw firm conclusions. Results from study one showed sample sizes could be easily overestimated by up to 100% using Browne's method. This chapter is interested in controlling the margin of error. Using regression of Median percentage error (MPE) of study one, pilot sample sizes are proposed for different MPEs. To control error margin a new Goldilocks "Just about right" interval for the different error margins is developed. Results from this study led to two peer reviewed publications, Obodo et al. (2021) and Obodo et al. (2023) both given in Appendix A.

5.1 Regression of MPE of study one to control margin of error

Using the data from Study 1a, a regression model of MPE will be used to explain the range of values for pilot sample at different MPEs. The margin of error will be the width of the interval within which the true sample size value lies while considering certain upper and lower values. A plot of the median percentage error at each cell combination against the square root of pilot sample size for coverage $(1-\gamma)$ = 0.8 is used to generate the result as given in Figure 5.1.

The accuracy of mean and other parameters is related to the square root of the sample size (n). Example the 95% Confidence interval for a mean is

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

where $\bar{x}$ is the sample mean, s the sample standard deviation, *t* the critical value of the *t* distribution and is based on $n-1$ degrees of freedom. Hence, a relationship against $1/\sqrt{n}$ would be expected to hold with the error; or on arrangement for $1/error$ to be related to $\sqrt{n}$. Furthermore, the regression model utilizes the inverse of the MPE due to the inverse relationship between MPE and pilot sample size. This suggests that an increase in the pilot sample size will result in a decrease in the MPE. Therefore, using an inverse relationship will give the best model, considering how change of pilot sample sizes affects MPE of interest in this research. Additionally, A square root function is applied to the pilot sample size in the model. This is due to the diminishing relationship between MPE and pilot sample size, where an increase in the pilot sample size leads to diminishing effect on MPE. Hence, using the square root function will improve the linearity of the model, given the linearity nature of regression model been used (Kutner, 2005).

The regression of the inverse of the MPE against the square root of pilot sample size per arm m for the simulation data is given by

$$^{1}/_{MPE} = -0.01208 + 0.01298\sqrt{m} \qquad (5.1)$$

(R^2 = 0.985) and which on re-arrangement gives

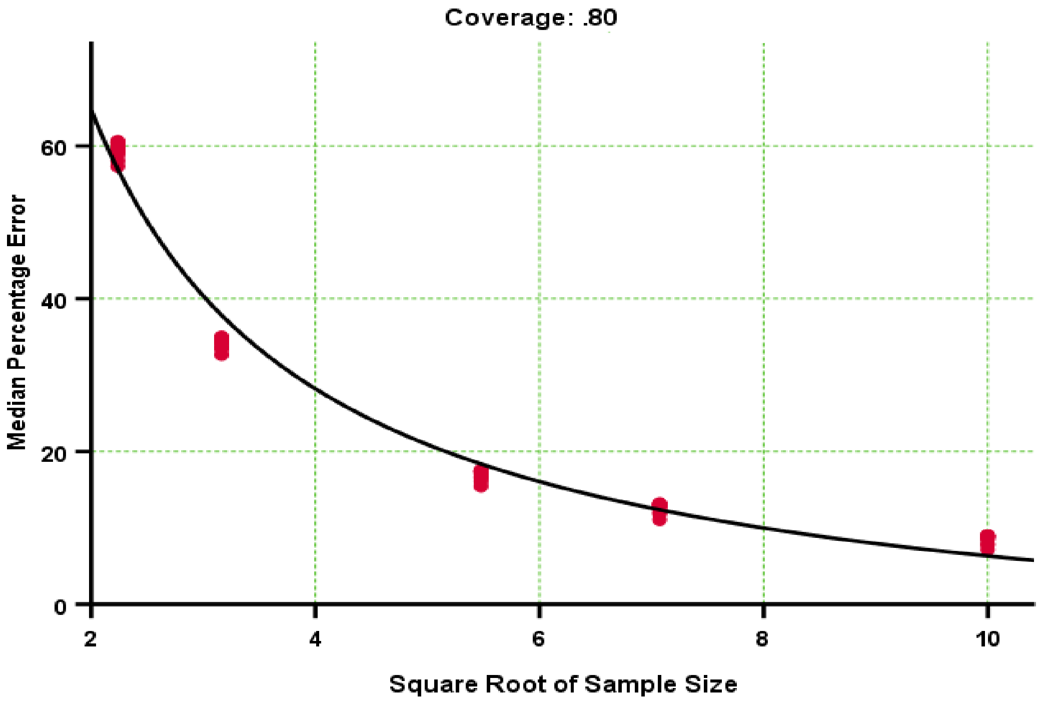
$$m = \left(0.930663 + \frac{1}{0.01298 \times MPE}\right)^2 \qquad (5.2)$$

Repeating the process for coverage $1 - \gamma$ = 0.9 gives the regression equation

$$^{1}/_{MPE} = -0.009339 + 0.00828\sqrt{m} \qquad (5.3)$$

(R^2 = 0.994) which on re-arrangement gives

$$m = \left(1.1280 + \frac{1}{0.00828 \times MPE}\right)^2 \qquad (5.4)$$



.

Figure 5.1: Graphical representation of MPEs against the square root of the pilot sample size $(\sqrt{m})$

Hence, via regression the relationship between median MPE and pilot sample size may be quantified. Table 5.1 quantifies the pilot sample size to maintain the median percentage error at a required level for either $(1-\gamma)$ = 0.8 and for $(1-\gamma)$ = 0.9. Thus, for instance, if a researcher opts for 80% coverage and wishes to control the MPE to be no more 10% of true sample size, then a pilot sample size of m = 75 per arm would be needed, but for 90% coverage the minimum sample size to have a MPE of 10% would be 174 per arm. Furthermore, if the researcher opts for 80% coverage and the MPE to be no more than 20% of the true sample size, then the pilot sample size of m= 23 per arm would be needed, but for 90% coverage the minimum sample size to have a median error of 20% would be 52.

Table 5.1: Median percentage error at various pilot sample size per arm.

Median Percentage Error	Pilot sample size per arm at 80% coverage	Pilot sample size per arm at 90% coverage
4	408	980
5	267	639
6	190	452
7	143	337
8	112	262
9	90	211
10	75	174
12	54	125
14	42	95
16	33	75
18	28	62
20	23	52
22	20	44
24	18	38

Table 5.2 shows the mean median percentage error (MMPE) for a given sample size for 80% and 90% coverage. From this table, as the pilot sample size is increased the MMPEs value is seen to decrease implying increased accuracy with increasing pilot sample size. For instance, considering 80% coverage at m = 25 per arm the

corresponding MMPEs is 19 while an increase of m = 50 the median percentage error decreases to 13 showing a better accuracy as the pilot sample size is increased.

Table 5.2: Mean Median percentage error.

m	80% Coverage	90% Coverage
5	59	109
10	35	60
15	26	44
20	22	36
25	19	31
30	17	28
35	16	25
40	14	23
45	14	22
50	13	20

Table 5.1 and Table 5.2 and their corresponding regressions equations therefore have value that will be helpful to a research team and funders to decide on how big a pilot sample size should be. Table 5.3 gives the Mean Median percentage error values that will be associated with different rules of thumb at 80% coverage. In comparism to Julious (2005) it suggests that if the pilot sample size of 12 is used then the sample sizes will have a corresponding 24 MMPE at 80% coverage.

Table 5.3: Comparison of some rules of thumb to MMPE at 80 % coverage.

Proposed by	Pilot sample size per group	MMPE
Teare et al. (2014)	35	16
Sim and Lewis (2012)	≥ 28	14
Browne (1995)	15	26
Kieser and Wassmer (1996)	10-20	35-26
Julious (2005)	12	24

The results above suggest specific MMPE error for different at different pilot samples sizes and compares to rules of thumb but does not give an upper or lower control range for research that intend to be within certain upper or lower limit range of error hence a Goldilocks Approach will be developed.

5.2 The Goldilocks ("Just about right") approach (Study 1b)

In any given situation it may be desirable to

(a) Ensure that the estimated sample size, $\widehat{N}$ does not exceeds a lower percentage of the true sample size $N_L = N + q_L N$ $(-1 < q_L < 0)$ with a desired probability λ_1.

(b) Ensure that the estimated sample size, $\widehat{N}$ does not exceed an upper percentage of the true sample size $N_U = N + q_U N$ $(q_U > 0)$ with a desired probability λ_2.

where N_L and N_U are the lower and upper sample sizes,N is the true sample size,q_L and q_U are the lower and upper proportions.

For instance, suppose that the true sample size being estimated is 1000. Requirements may be to ensure there is a 95% chance that the estimate for the sample size is above the required sample size minus 10% (i.e., 95% chance of being above 900) and a 20% chance of not going beyond the true value plus 30% (i.e., not going beyond 1300). In this example $N_L = 900$, $q_L =$ -0.1, λ_1 = 0.95, $N_U = 1300$, $q_U =$ 0.3, λ_2 = 0.20.

It is postulated, that both requirement (a) and requirement (b) can be satisfied by a judicious choice of pilot sample size. The estimate of the sample size under requirement (a) would equal the estimate of the sample size under requirement (b) if $(1 + q_L)N_L = (1 + q_U)N_U$ i.e.

when $(1 + q_L)/(1 + q_U)$ = N_U/k_L.

The left-hand side of this equation is based on researcher specification; the right-hand side of the equation depends on sample size. We will solve this equation and hence estimate sample size for given requirements. This will be verified by simulation.

An investigator chosen "Just About Right" JAR interval is operationalized to be $[n - \lambda_1 n, n + \lambda_2 n]$ where $\lambda_1, \lambda_2 \in [0, 1]$, are investigator chosen parameters to prevent the degree of underpowering (λ_1) and degree of overpowering (λ_2). The aim is for trialists to be able to justify a pilot sample size and to make a statement to the effect of "*The proposed two group pilot study will have a sample size of m per arm. This sample size is chosen so that the resultant power calculations for a larger study will* *have $100(1-\gamma)\%$ chance of exceeding the minimum required sample size and which in a two-sided test with significance level α will have $100(1-\beta)\%$ power for detecting a difference between arms assuming a MCID of $(\mu_1-\mu_2)$. This proposed pilot sample size of m per arm will ensure that the estimated sample size will lie in the interval $1-\lambda_1 n$ to $1-\lambda_2 n$, with probability π providing a safeguard over under- and over-powering.*" For this statement we consider $\alpha=(0.01, 0.05)$, power $(1-\beta)=(0.8, 0.9)$, coverage $(1-\gamma)=(0.8, 0.9)$, as it is known that these factors do not affect the estimated sample size (as given in Chapter 4). Furthermore, the lower bounds $\lambda_1=(0.1, 0.2)$ and upper bounds $\lambda_2=(0.1, 0.2, 0.3)$ for any chosen level of π and any MCID will be used.

Inspection of Table 5.4, 5.5, 5.6, 5.7 and Figure 5.2, shows the percentage within any given interval monotonically increases with increasing pilot sample size for each of $(1-\gamma)$ = 0.8 and 0.9. It is also clear that the percentage in any given interval is greater for $(1-\gamma)$ = 0.8 compared with $(1-\gamma)$ = 0.9 and this is only to be expected since, for any estimated sample size, the sample size for when coverage is 0.9 must be greater than the sample size when a tolerance for coverage is set to be equal to 0.8. The percentage of instances within an interval is particularly sensitive to the upper bound λ_2 which naturally follows from the positively skewed chi-square distribution used in the estimation process.

Table 5.4: Percentage of simulation instances $100\hat{\pi}$ is in the interval $[n-\lambda_1 n,\ n+\lambda_2 n]$ for λ_1=0.1,0.2,λ_2=0.1,0.2,0.3, m= 5(5)35, $(1-\gamma)$=0.8.

	Coverage = 0.8					
	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
m	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
5	11.1	16.5	22.0	15.8	20.9	24.9
10	14.1	22.3	31.0	19.4	27.6	36.0
15	16.9	27.7	39.2	22.4	33.7	45.7
20	19.5	32.8	46.6	25.3	39.2	54.2
25	22.1	37.7	53.3	28.0	44.3	61.5
30	24.6	42.4	59.2	30.6	49.1	67.7
35	27.1	46.8	64.4	33.1	53.5	72.9

Table 5.5: Percentage of simulation instances $100\hat{\pi}$ is in the interval $[n - \lambda_1 n,\ n + \lambda_2 n\,]$ for λ_1=0.1,0.2,λ_2=0.1,0.2,0.3, m= 40(10)100, $(1-\gamma)$=0.8.

Coverage = 0.8						
	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
m	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
40	29.6	50.9	69.0	35.5	57.5	77.3
50	34.4	58.4	76.5	40.1	64.6	83.9
60	39.1	64.9	82.1	44.4	70.5	88.4
70	43.5	70.4	86.2	48.4	75.4	91.6
80	47.8	75.0	89.4	52.2	79.4	93.8
90	51.8	78.9	91.7	55.7	82.7	95.4
100	55.6	82.2	93.5	59.0	85.5	96.5

Table 5.6: Percentage of simulation instances $100\hat{\pi}$ is in the interval $[n - \lambda_1 n,\ n + \lambda_2 n\,]$ for λ_1=0.1,0.2,λ_2=0.1,0.2,0.3, m= 5(5)35, $(1-\gamma)$=0.9 in the interval.

Coverage = 0.9						
m	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
5	6.6	10.1	13.6	9.0	12.5	15.7
10	8.4	14.1	20.4	11.1	16.8	23.2
15	10.1	17.9	27.2	13.0	20.9	30.4
20	11.8	21.7	33.9	14.8	24.9	37.5
25	13.5	25.5	40.4	17.0	28.8	44.1
30	15.2	29.3	46.6	18.2	32.6	50.4
35	16.9	33.1	52.4	20.0	36.4	56.1

Table 5.7: Percentage of simulation instances $100\hat{\pi}$ is in the interval $[n - \lambda_1 n,\ n + \lambda_2 n\,]$ for λ_1=0.1,0.2,λ_2=0.1,0.2,0.3, m= 40(10)100 $(1-\gamma)$ =0.9 in the interval.

	Coverage = 0.9					
m	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
40	18.7	36.8	57.7	22.0	40.0	61.3
50	22.2	44.0	66.9	25.0	47.0	70.0
60	25.7	50.8	74.3	29.0	53.3	76.9
70	29.3	57.0	80.0	31.4	59.1	82.2
80	32.9	62.6	84.5	34.6	64.2	86.1
90	36.5	67.5	87.9	37.8	68.8	89.2
100	40.0	71.9	90.5	40.8	72.8	91.6

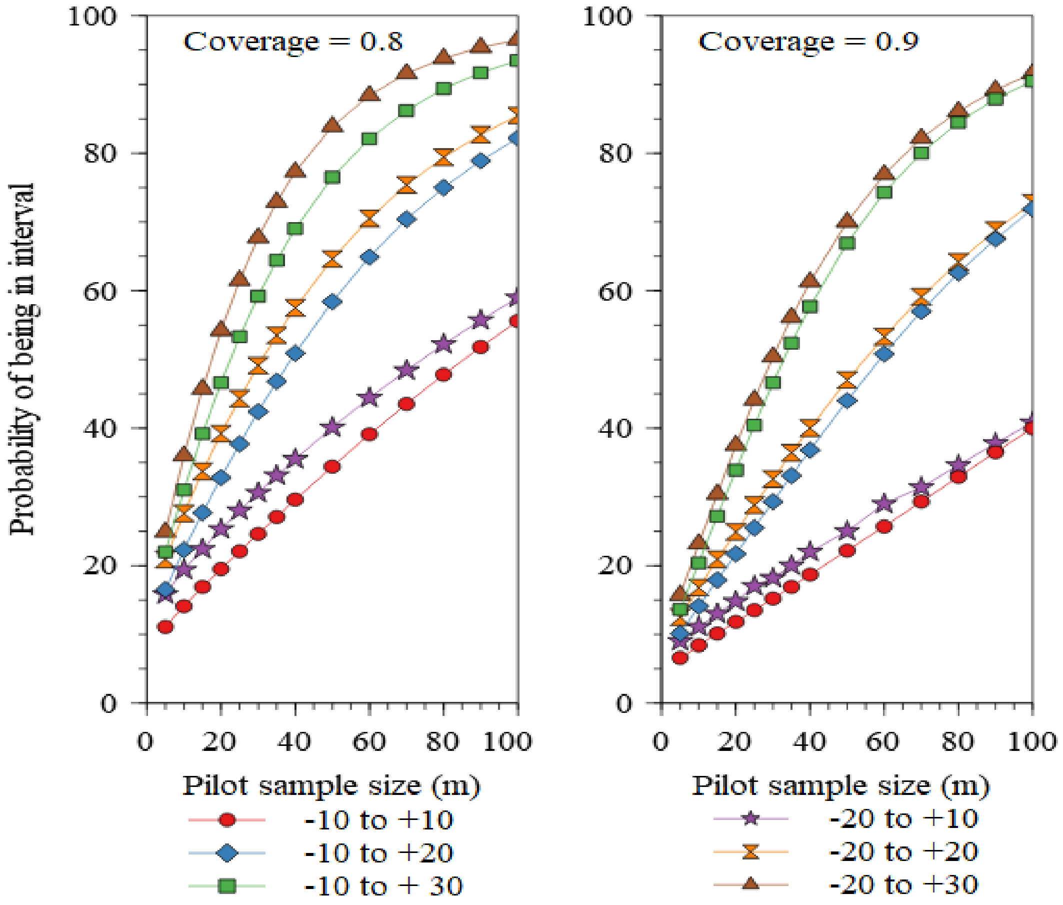


Figure 5.2: Scatter plot for probability of sample size being in various intervals of pilot sample sizes.

Monotonic trends between $\hat{\pi}$ and m per arm.

Table 5.10 is the monotonic trends between $\hat{\pi}$ and pilot per arm sample size m, for each interval $[n - \lambda_1 n,\ n + \lambda_2 n\,]$ and each level of coverage modelled using linear regression with the functional form

$$\ln(\hat{\pi}) = b_0 + b_1\sqrt{m}$$

Thus, for instance, when coverage = 0.8 and the interval $n \pm 0.1n$ is considered then it is readily verified that $\ln(\hat{\pi}) = -2.745 + 0.297\sqrt{m}$ and that the overall goodness-of-fit, $100R^2$, is 96.3% obtained from Table 5.8. Table 5.8 and Table 5.9 presents the estimated intercepts, gradients, and goodness of fit for λ_1= 0.1, 0.2; $\lambda_2 = 0.1, 0.2, 0.3$ for $(1-\gamma)$= 0.8 and $(1-\gamma)$ = 0.9 respectively.

Table 5.8: Regression equations of the form $ln(\pi) = b_0 + b_1\sqrt{m}$ given estimated intercept (b_0), gradient (b_1) for $(1-\gamma)$ = 0.8.

$(1-\gamma)$ = 0.8				
Lower Percentage $(100\,\lambda_1)$	Upper Percentage $(100\,\lambda_2)$	Intercept	Gradient	R- Squared
10	10	-2.745	.297	.963
10	20	-2.531	.406	.988
10	30	-2.399	.506	.993
10	40	-2.094	.543	.981
10	50	-1.697	.527	.952
20	10	-2.256	.262	.954
20	20	-2.228	.400	.989
20	30	-2.375	.569	.997
20	40	-2.613	.759	.997
20	50	-2.557	.853	.998

Table 5.9: Regression equations of the form $ln(\pi)$ = $b_0 + b_1\sqrt{m}$ given estimated intercept (b_0), gradient (b_1) for $(1-\gamma)$ = 0.9.

$(1-\gamma)$ = 0.9				
Lower Percentage $(100\,\lambda_1)$	Upper Percentage $(100\,\lambda_2)$	Intercept	Gradient	R- Squared
10	10	-3.306	.290	.957
10	20	-3.082	.402	.984
10	30	-3.029	.528	.995
10	40	-3.028	.656	.998
10	50	-2.827	.712	.991
20	10	-2.872	.250	.955
20	20	-2.795	.378	.986
20	30	-2.856	.524	.995
20	40	-3.108	.716	.996
20	50	-3.450	.919	.993

For any level of coverage and any interval, any regression equation in Table 5.6 may be re-written in terms of pilot sample size i.e., $m = [\ln(\hat{\pi}) - b_0]/ b_1)$^2. Solution of this will give an estimated pilot sample size per arm, m, for any required percentage for the given interval.

Table 5.10 shows the pilot sample size per arm (m) needed to have a required probability (π) of being in each interval $[n - \lambda_1 n,\ n + \lambda_2 n\ ]$ for coverage of 0.8 or coverage 0.9. Thus, for instance, if an investigator requires an 80% chance of not being underpowered for a definitive trial (coverage = 0.8) and requires a 70% chance (π = 0.7) of being within ± 10% of the true required sample size (λ_1 = 0.1, λ_2 = 0.1) then a sample size per arm (m) of 65 is needed for any given MCID.

Table 5.10: Pilot sample size (m) required for a required proportion (π) to be in the interval $[n - \lambda_1 n,\ n + \lambda_2 n\ ]$.

	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
π	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
$(1-\gamma)$= 0.8						
0.50	48	20	11	36	15	9
0.55	52	23	13	40	17	10
0.60	56	25	14	44	18	11
0.65	61	27	15	49	20	12
0.70	65	29	16	53	21	13
0.75	68	31	17	56	23	13
0.80	72	32	18	60	25	14
0.90	79	36	h21	67	28	16
$(1-\gamma)$= 0.9						
0.50	81	35	20	76	31	17
0.55	87	38	21	83	34	18
0.60	93	41	23	89	37	20
0.65	98	43	24	95	39	22
0.70	103	46	26	101	42	22
0.75	108	48	27	107	43	24
0.80	113	50	28	112	46	25
0.90	121	54	31	122	51	27

If an investigator requires a 90% chance of not being underpowered for a definitive trial $(1-\gamma)$ = 0.9 and requires a 60% chance $(\pi = 0.6)$ of being within the interval $\pm 10\%$ that is $(\lambda_1 = 0.1,\ \lambda_2 = 0.1)$ of the true required sample size then a pilot sample size per arm m of 93 is needed. The other intervals can be interpreted in the same order.

Table 5.11 shows comparism of some rules of thumb to proposed goldilocks interval at $(1-\gamma) = 0.8$. It shows that using Julious (2005) suggested 12 pilot sample per arm at 80% coverage, the researcher will achieve only a 55% chance $(\pi = .55)$ of being within the interval of $(\lambda_1 = 0.2,\ \lambda_2 = 0.3)$ of the true sample size.

Table 5.11: Comparison of some rules of thumb for pilot sample size and Goldilocks proportion range of Table 5.10 at $(1-\gamma) = 0.8$.

Proposed by	Pilot sample size per arm	π	λ_1, λ_2
Teare et al. (2014)	35	.50	0.2,0.1
Sim and Lewis (2012)	≥ 28	.90	0.2,0.2
Browne (1995)	15	.50	0.2,0.2
Kieser and Wassmer (1996)	10-20	.55	0.2,0.3, 0.1,0.2
Julious (2005)	12	.65	0.2,0.3

5.3 Summary

Considering the level of excess recorded in the Study 1 results, a model was developed using regression of the median percentage error to show the corresponding median percentage error associated with different pilot sample sizes. The mean median percentage error for different pilot sample sizes at 80% and 90% coverage was also developed. The Goldilocks 'Just about right' approach was further developed to allows researcher to select a pilot sample size to control the error (MPE and interval error), this ensures sample sizes do not go below a certain lower percentage or exceed a certain upper percentage of the true sample size for a desired probability. The findings were compared to the rules of thumb that suggested pilot sample sizes without considering the corresponding error values. The findings will be very helpful to researchers in making informed decision when choosing pilot sample sizes. The next chapter will consider applying Browne's method and naive methods using Cohen's d and Hedge's h when the minimum clinical importance difference (MCID) is unknown.

CHAPTER 6

Comparison of sample size Estimation Methods when Minimum Clinically Important Different (MCID) is unknown (Study 2).

This study develops and investigates how Naive and modified Browne's formulae operate when the MCID cannot be specified in advance. The formulae are developed using Browne's method, Cohen's d , and Hedge's h . They give an estimation for sample size and will be compared based on how accurately they estimate sample size. Their percentage greater than smallest sample size that satisfies power requirements N_{true} will be reviewed. The Median percentage error associated with each formula will be used to evaluate their performance. The formulae are respectively:

(a) the naive estimate for the sample size using Cohen's d when σ^2, μ_1, μ_2 are unknown known as Naive-Cohen.

$$N_{N,C} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{d^2}$$

(b) using Cohen's d in Browne's formula known as Browne-Cohen

$$N_{B,C} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 k}{d^2}$$

where k is the Browne multiplier and depends on the level of coverage, $(1-\gamma)$. d is a biased estimator of $\delta = (\mu_1 - \mu_2)/\sigma$ whereas the correction proposed by Hedge is unbiased.

Hence a naive estimate of sample size, using Hedge's h instead of Cohen's d would be

(c) Known as Naive-Hedge

$$N_{N,H} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{h^2}$$

and

(d) using Hedges' h in Browne's approach known as Browne-Hedge

$$N_{B,H} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 k}{h^2}$$

All $N_{,C}$, $N_{B,C}$, $N_{N,H}$ and $N_{B,H}$ were generated for the two-sample case, assuming equal variances, and under 1:1 allocation.

6.1 Pre-Study Hypothesis for Study 2

By inspection Cohen's d is larger in magnitude than Hedge's h. Accordingly for any given pilot samples $N_{N,C}$ (the naive estimate for the sample size using Cohen's d) will be smaller than $N_{N,H}$ (the naive estimate of sample size using Hedge's h) i.e., $N_{N,C} < N_{N,H}$. By the same reasoning, $N_{B,C}$ (the estimated sample size using Cohen's d in Browne's formula) will be smaller than the estimate $N_{B,H}$ (the estimated sample size using Hedge's h in Browne's method) i.e $N_{B,C} < N_{B,H}$.

Since, $kN_{N,C} = N_{B,C}$ and $kN_{N,H} = N_{B,H}$

with $k = (2m-2)/\chi^2_{1-\gamma,m_1+m_2-2} > 1$ then it follows

$$N_{N,C} < N_{N,H} < N_{B,C} < N_{B,H}$$

and by virtue of finite machine precision during simulation there may be instance were

$$N_{N,C} \leq N_{N,H} < N_{B,C} \leq N_{B,H}$$

In study 1, $\mu_1 - \mu_2$, is assumed known and

$$N_{est} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 ks^2}{(\mu_1 - \mu_2)^2} \tag{6.1}$$

The comparable estimate in using study 2 is

$$N_{B,C} = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 ks^2}{(\bar{x}_1 - \bar{x}_2)^2} \tag{6.2}$$

The ratio of these two estimates is

$$\frac{N_{est}}{N_{B,C}} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{(\mu_1 - \mu_2)^2} = \frac{(\bar{x}_1 - \bar{x}_2)^2}{\delta^2} \tag{6.3}$$

when $\sigma^2 = 1$ (as per simulation parameters).

6.1.1 Non-central chi-square distribution for two-sample case for sample size estimation

In accordance with Casella and Berger (2002), if X_1 and X_2 are independent normal random variables with means $\omega_1 - \omega_2$ and common variance, σ^2 then $(X_1 - X_2)^2$ has a scaled non-central chi square distribution with one degree of freedom and non-centrality parameter

$$\lambda = \left(\frac{\omega_1 - \omega_2}{\sigma}\right)^2 \tag{6.4}$$

The mean and variance of the non- central chi- square distribution are given by

Mean: $(1+\lambda)$

Variance: $(1+2\lambda)$

and with positive skew

When considering the scaling parameter σ^2 the mean and variance is

Mean: $\sigma^2(1+\lambda)$

Variance: $\sigma^4(1+2\lambda)$

Accordingly

$$\frac{(\bar{x}_1 - \bar{x}_2)^2}{\delta^2} \tag{6.5}$$

has a scaled non- central chi square distribution with scaling parameter.

$$\frac{2}{m\delta^2} \tag{6.6}$$

and hence with mean

$$= \frac{2}{m\delta^2} + 1 \tag{6.7}$$

and with variance

$$\frac{4}{\delta^4 m^2} + \frac{8}{\delta^2 m^3} \tag{6.8}$$

Hence, on average, the estimated sample sizes for study 2 will be smaller than those estimated when $\mu_1 - \mu_2$ is known and then this degree of excess is a function of m (degree of excess diminishes with increasing sample size). The simulation will investigate the degree of difference.

The formulae would be compared by simulation to view their impact on sample size estimation. Simulation parameters are presented in Table 6.1 with the design corresponding to a 2 by 2 by 2 by 2 by 3 by 5 fully crossed design. 100,000 replicates will be conducted at each cell combination.

Table 6.1: Parameter combinations for study 2.

Factor	Number of Levels	Levels
Alpha(α)	2	0.01, 0.05
Beta (β)	2	0.1, 0.2
Coverage $(1-\gamma)$	2	0.8, 0.9,
Effect size (δ)	3	0.1, 0.4, 0.75
Pilot sample size (m)	5	8, 16, 32, 64, 128

6.1.2 Algorithm for study two

1. Start at set.seed (100)
2. Set number of iterations 100,000
3. Create the vector for given parameter combination of α= (0.01,0.05), β =(0.1,0.2), $(1-\gamma)$ =(0.8,0.9), δ =(0.1,0.4,0.75), and m =(8,16,32,64, 128).
4. Loop over the parameters
5. Define qnorm $(1-(\alpha/2)$ and qnorm $(1- \beta\ )$
6. Define $(2*\mathrm{m}-2)/\mathrm{qchisq}(1-\mathrm{coverage}, \mathrm{df}=(2*\mathrm{m}-2))$
7. Calculate N_{true}
8. Calculate $N_{B,C}, N_{B,H}$, $N_{N,C}$, and $N_{N,H}$
9. Calculate percentage greater than N_{true} and the Median percentage error Get results summary for data. frame (N_{true}, $N_{B,C}, N_{B,H}$, $N_{N,C}$, and $N_{N,H}$, percentage greater than N_{true} and the Median percentage error)
10. Loop for each combination for all iterations
11. Create summary results for each combination for all iteration
12. Save results
13. End

6.2 Percentage greater than N_{true} for estimated sample size and Median Percentage Error (MPE) of Study 2

6.2.1 Percentage greater than N_{true} for estimated sample size using $N_{B,C}$, $N_{B,H}$, $N_{N,C}$, and $N_{N,H}$

Percentage greater than N_{true} calculates the percentage by which a predicted or estimated value exceeds the true or actual value. It is computed as follows:

Percentage Greater than N_{true} value = (Number of values greater than N_{true} /Total number of values) *100

This will be done by simulation, and the result will be used to determine the magnitude of overestimation. This outcome will present the percentage by which the results from $N_{B,C}$, $N_{B,H}$, $N_{N,C}$, and $N_{N,H}$ are greater than N_{true} for different α,β,$(1-\gamma)$,δ and m scenarios. Result will help determine the formulae that appears to perform better.

Results

Table 6.2 shows the percentage greater than N_{true} at 80% coverage for the four formulae, from the result the approaches generally lead to overestimation of sample sizes. It further shows that this overestimation is not due to α or β. δ and m has an impact in the level of overestimation. Browne-Hedge $N_{B,H}$ seems to perform better based on the result. This will be elaborated in further detail using graphs.

Table 6.2: Percentage Greater than N_{true} at α = 0.01, and $(1-\gamma)$ = 0.8.

		Browne-Cohen		Browne-Hedge		Naive-Cohen		Naive-Hedge	
		β		β		β		β	
m	δ	.10	.20	.10	.20	.10	.20	.10	.20
8	.10	18.5	18.8	19.6	19.5	15.3	15.2	16.1	16.1
	.40	51.8	51.8	54.5	54.4	43.6	43.5	45.9	45.8
	.75	60.5	60.5	64.2	64.1	48.6	48.5	51.8	51.8
16	.10	24.0	23.8	24.2	24.4	21.2	21.0	21.7	21.6
	.40	54.6	54.6	24.6	55.9	48.3	48.4	49.6	49.6
	.75	59.5	59.5	55.9	61.7	49.0	48.9	51.1	51.0
32	.10	31.1	31.0	61.8	31.4	28.7	28.6	51.1	29.0
	.40	55.1	55.1	31.5	55.9	49.6	49.6	29.1	50.3
	.75	59.4	58.6	55.9	60.1	49.5	48.8	50.4	50.2
64	.10	38.9	39.0	39.2	39.2	37.0	37.0	37.2	37.2
	.40	55.0	55.0	55.6	55.6	49.6	49.8	50.4	50.3
	.75	58.9	58.1	59.9	59.1	49.5	48.7	50.5	49.7
128	.10	46.1	46.2	46.2	46.3	44.4	44.6	44.6	44.7
	.40	55.0	54.8	55.4	55.1	49.9	49.8	50.3	50.1
	.75	58.7	57.8	58.7	57.8	49.6	48.6	50.2	49.3

The percentage achieved overall is not up to desired 80% coverage, the simulation finding agrees with the theory which is always underneath the coverage level. A further elaboration of the impact of various parameters based on table 6.2 is presented

Beta impact

Considering impact of β, the percentage greater than N_{true} for **Naive-Cohen** approach, with $\alpha = .01$, $m = 8$, $\delta = .10$ at β =.10, is approximately 15%. This percentage remains the same when β is changed to .20. For **Browne-Cohen** approach, with $\alpha = .01$, $m = 16$, $\delta = .40$ at $\beta = .10$, the percentage greater than N_{true} is approximately 54%. This percentage remains the same when β is changed to .20. **Naive-Hedge** approach, with $\alpha = .05$, $m = 32$, δ=.75 at $\beta = .10$, the percentage greater than N_{true} is approximately 50%. Again, this percentage remains constant when β is changed to .20. Finally for **Browne-Hedge** approach with $\alpha = .05$

, $m = 128$, $\delta = .10$ at $\beta = .10$ the percentage greater than N_{true} is approximately 46% and remains the same when β changes to .20. It indicates that β does not have an effect in the result this is graphically presented in Figure 6.1.

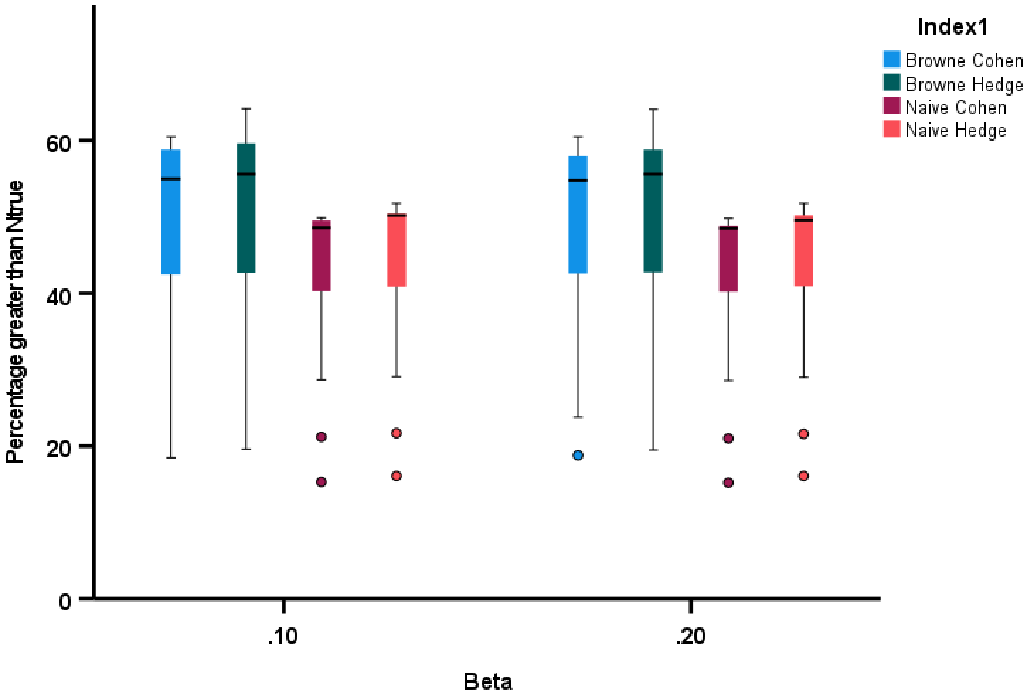


Figure 6.1: Impact of β on percentage greater than N_{true} , at α = 0.01, $(1-\gamma)$ = 0.8,0.9,m = (8,16,32, 64,128), and δ = (0.10,0.40,0.75).

Effect sizes impact

Considering the impact of δ, the **Naive-Cohen** approach, with $\alpha = .01$, $\beta = .10$, $m = 8$, and $\delta = .10$, results in a percentage greater than N_{true} of approximately 15%. When is $\delta = .4,$ the percentage rises to 44%, and for $\delta = .75$ it reaches 49%. For **Browne-Cohen** approach, with $\alpha = .05$ $\beta = .10$, $m = 16$, and $\delta = .10$ percentage greater than N_{true} is approximately 24%. With $\delta = .4$,it becomes 54% and for $\delta = .75$ it rises to 59%. In the case of **Naive-Hedge** approach with $\alpha = .05$ $\beta = .10$, $m = 32$, and $\delta = .10$ percentage greater than N_{true} is approximately 29%. With $\delta = .4$ the percentage becomes 50% and at $\delta = .75$ it stays the same. Using **Browne-Hedge** approach, with $\alpha = .05$ $\beta = .20$, $m = 32$, and $\delta = .10$ the percentage greater than N_{true} is approximately 31%. With $\delta = .4,$ the percentage increases to 56% and for $\delta = .75$, it rises to 60%. Hence δ has an impact in the result. The variance in performance is less for the larger δ but still a tendency to underestimate with 50:50 to be underpowered using these methods providing δ is ≥ 0.4.

Figure 6.2 further concurs there is an impact due to δ, with the range of value is smaller for δ of 0.4 and 0.74.

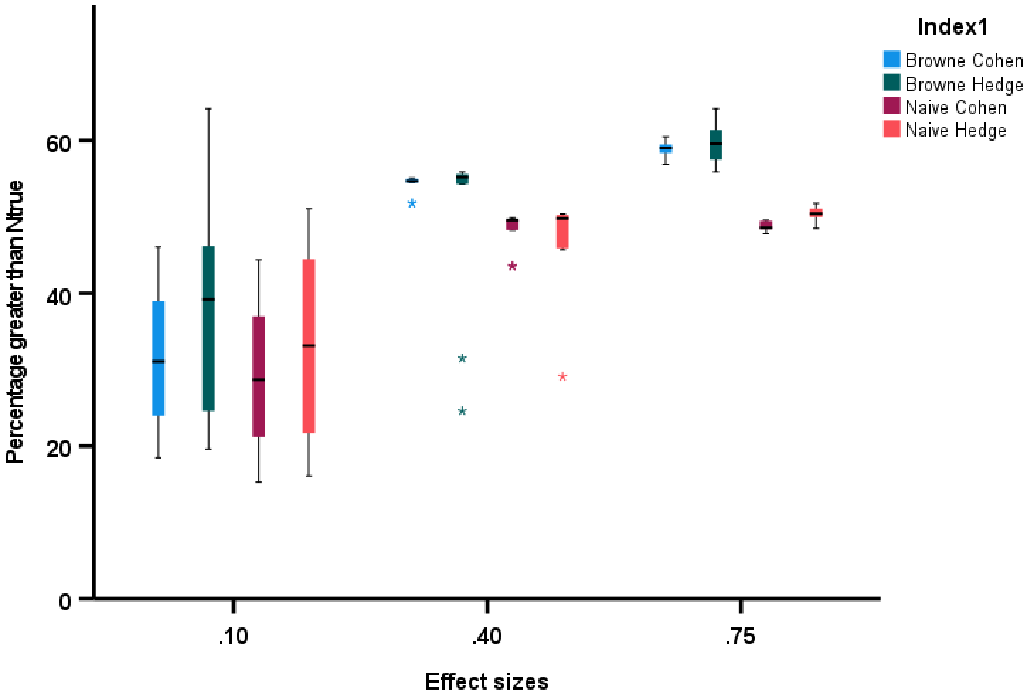


Figure 6.2: Impact of δ on percentage greater than N_{true} at α = (0.01,0.05), β = (0.1, 0.2), $1 - \gamma$ = (0.8.0.9), and m =(8,16,32, 64,128).

Pilot sample size

Considering the impact of m, for **Naive-Cohen** approach, with $\alpha = .01 \quad \beta = .10, \delta = .10$ and $m = 8$ the percentage greater than N_{true} is approximately 15%. Other parameter remaining the same at $m = 16$ the percentage greater than N_{true} becomes 21%, $m = 32$ it becomes 28%, $m = 64$ it is 37%, finally at $m = 128$ it reaches 48%. There is similar pattern for **Browne-Cohen, Naive-Hedge** and **Browne Hedge approaches** hence there is impact due to m.

Figure 6.3 Shows the impact of m in the different formulae. The box plot shows there is impact in the result due to changes in m. The variance decreases with an increase in pilot sample sizes for all the formulae.

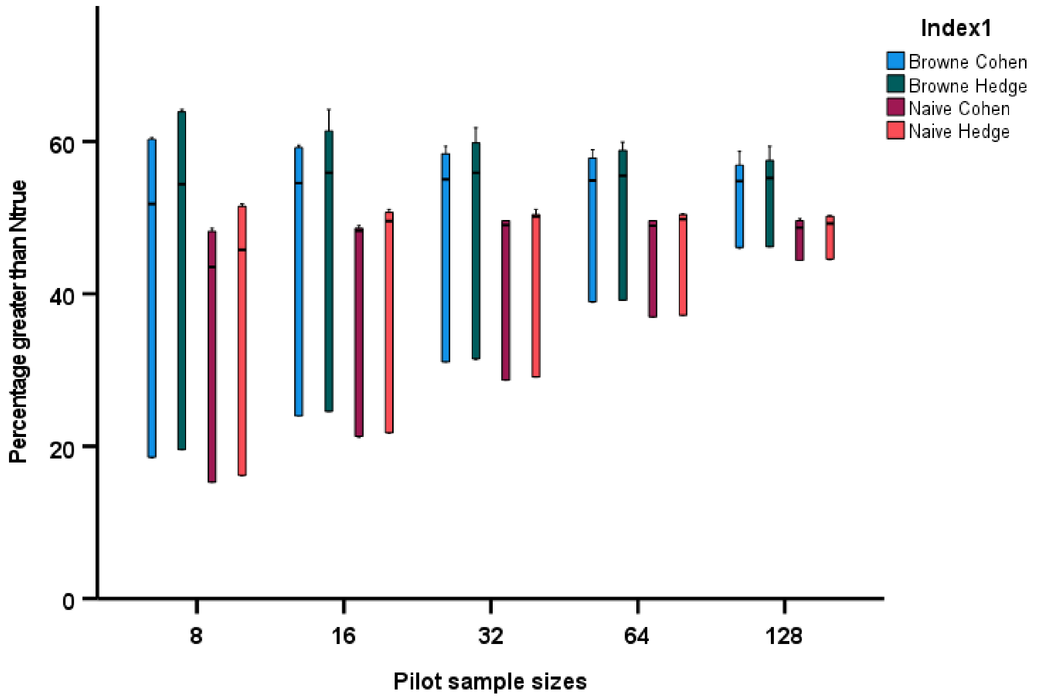


Figure 6.3: Impact of pilot of sample sizes on percentage greater than N_{true} at α = (0.01,0.05), β = (0.1, 0.2), $(1-\gamma)$ = (0.8,0.9) and δ = (0.10, 0.40, 0.75).

Alpha impact in percentage greater than N_{true}

The parameter combination for Table 6.3 is like those of Table 6.2, with the only difference being the change in α value from 0.01 to 0.05. The results are however similar and the similarity in result suggest that there is no significant impact due to α value.

In the **Naive-Cohen** approach with $\beta = .10$, $m = 8$, $\delta = .10$ at $\alpha = .01$, the percentage greater than N_{true} is approximately 15%. The percentage remains the same when α is changed to .05. For the **Browne-Cohen** approach with $\beta = .10$, m=16, $\delta = .40$ and $\alpha = .01$ the percentage greater than N_{True} is approximately 48%, and it remains the same when α is changed to .05. Similarly, the **Naive-Hedge** approach shows that with $\beta = .20, m = 32$ $\delta = .75$ at $\alpha = .01$, the percentage greater than N_{True} is approximately 49%. Again, the percentage remain the same when α is changed to .05. Finally, for **Browne-Hedge** approach with $\beta = .20$=,m=64, $\delta = .40$ at α =.01, the percentage greater than N_{true} is approximately 55% and remains the same when α changes to .05. These observations indicate that α has no impact in the in sample size estimation using these formulae.

Table 6.3: Percentage Greater than N_{true} at α = 0.05, and $(1-\gamma)$ = 0.8.

		Browne-Cohen		Browne-Hedge		Naive-Cohen		Naive-Hedge	
		β		β		β		β	
m	δ	.10	.20	.10	.20	.10	.20	.10	.20
8	.10	18.5	18.8	19.6	19.5	15.3	15.2	16.1	16.1
	.40	51.8	51.8	54.5	54.4	43.6	43.5	45.9	45.8
	.75	60.5	60.5	64.2	64.1	48.6	48.5	51.8	51.8
16	.10	24.0	23.8	24.2	24.4	21.2	21.0	21.7	21.6
	.40	54.6	54.6	24.6	55.9	48.3	48.4	49.6	49.6
	.75	59.5	59.5	55.9	61.7	49.0	48.9	51.1	51.0
32	.10	31.1	31.0	61.8	31.4	28.7	28.6	51.1	29.0
	.40	55.1	55.1	31.5	55.9	49.6	49.6	29.1	50.3
	.75	59.4	58.6	55.9	60.1	49.5	48.8	50.4	50.2
64	.10	38.9	39.0	39.2	39.2	37.0	37.0	37.2	37.2
	.40	55.0	55.0	55.6	55.6	49.6	49.8	50.4	50.3
	.75	58.9	58.1	59.9	59.1	49.5	48.7	50.5	49.7
128	.10	46.1	46.2	46.2	46.3	44.4	44.6	44.6	44.7
	.40	55.0	54.8	55.4	55.1	49.9	49.8	50.3	50.1
	.75	58.7	57.8	59.4	58.5	49.6	48.6	50.2	49.3

Figure 6.4 further illustrates there is no impact by α in result by the four formulae as the box-plot are similar for the four formulae for both α of .01 and .05

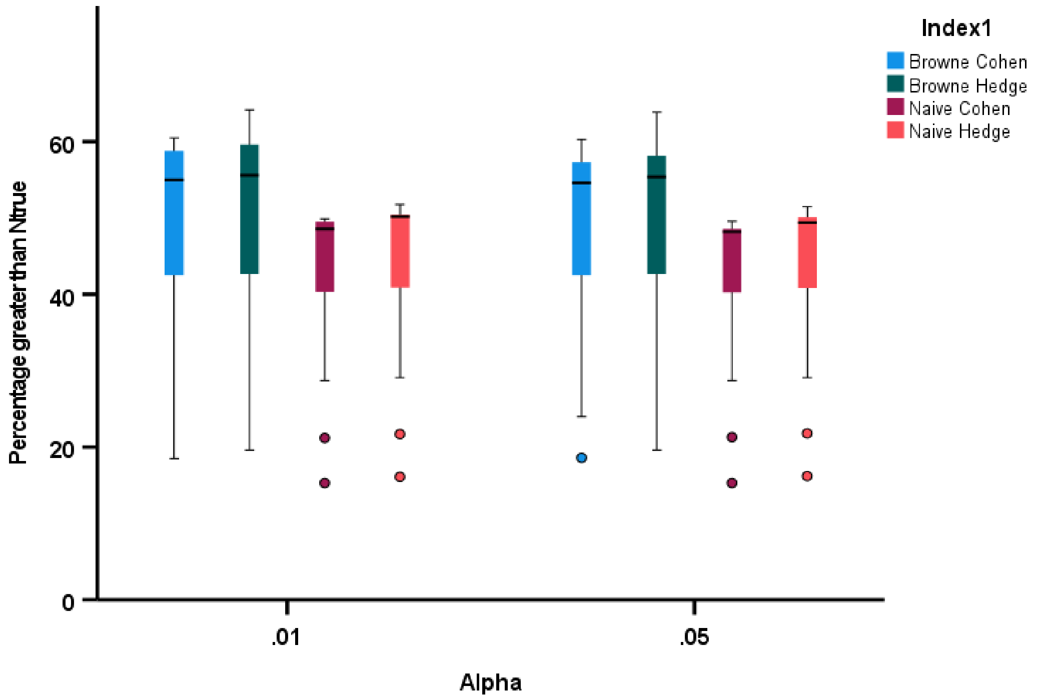


Figure 6.4: Impact of α on percentage greater than N_{true} at β = (0.1, 0.2), $(1-\gamma)$ = 0.8,0.9,m = (8,16, 32, 64,128), and δ = (0.10, 0.40,0.75).

Table 6.4 and 6.5 shows results for percentage Greater than N_{true} for $N_{B,C}$, $N_{B,H}$, $N_{N,C}$, and $N_{N,H}$ at coverage of 0.9 as against 0.8 in table 6.2 and 6.3. This result shows coverage has an impact on the percentage greater than N_{true} as the values are slightly different despite all other parameters remaining the same. The percentage greater than N_{true} changes slightly.

Coverage impact in overestimation of sample sizes using table 6.4 and 6.2.

The Naïve-Cohen and Naive-Hedge do not have any adjustment for coverage. So, they perform the same for both 0.8 and 0.9 coverage. They simply tend to underestimate. The Browne-Cohen and Browne-Hedge shows a slight impact in the results due to change in coverage.

Table 6.4: Percentage Greater than N_{true} at α = 0.01 and $(1-\gamma)$ = 0.9.

		Browne-Cohen		Browne-Hedge		Naive-Cohen		Nave-Hedge	
		β		β		β		β	
m	δ	.10	.20	.10	.20	.10	.20	.10	.20
8	.10	20.4	20.4	21.6	21.5	15.3	15.2	16.2	16.2
	.40	56.4	56.6	59.0	59.2	43.6	43.7	45.9	45.9
	.75	67.1	67.1	70.8	70.6	48.7	48.6	52.0	52.0
16	.10	25.4	25.4	26.1	26.0	21.2	21.0	21.8	21.6
	.40	57.8	58.0	59.2	59.4	48.3	48.4	49.4	49.6
	.75	65.4	64.8	67.6	67.1	49.0	48.7	51.4	50.8
32	.10	32.4	32.3	32.8	32.7	28.8	28.7	29.1	29.0
	.40	58.1	58.0	58.9	58.3	49.6	49.6	50.5	50.4
	.75	64.4	63.8	65.9	65.3	49.5	48.8	50.9	50.5
64	.10	40.1	40.1	40.3	40.3	37.0	37.0	37.2	37.3
	.40	57.8	57.7	58.3	58.1	49.6	49.8	50.4	50.3
	.75	63.8	62.9	64.8	63.9	49.5	48.7	50.6	49.7
128	.10	47.1	47.0	47.2	47.1	44.7	44.6	44.8	44.6
	.40	57.5	57.7	57.9	58.1	49.9	49.8	50.3	50.1
	.75	63.2	62.4	63.9	63.1	49.4	48.7	50.1	49.3

The parameter combination for Table 6.3 is like Table 6.5, with the only difference being the change in coverage value from 0.8 to 0.9. The results show slight difference for Browne approaches suggesting an impact due to coverage in them.

Table 6.5: Percentage Greater than N_{true} at α = 0.05 and $(1-\gamma)$ = 0.9.

		Browne-Cohen		Browne-Hedge		Naive-Cohen		Naive-Hedge	
		β		β		β		β	
m	δ	.10	.20	.10	.20	.10	.20	.10	.20
8	.10	20.4	20.5	21.6	21.6	15.4	15.5	16.2	16.3
	.40	56.1	56.1	58.8	58.8	43.4	43.4	45.6	45.7
	.75	66.5	67.0	70.1	70.1	48.1	48.6	51.4	51.9
16	.10	25.5	21.8	26.1	26.2	21.3	21.3	21.8	21.8
	.40	57.9	49.7	59.3	59.3	48.4	48.2	49.5	49.4
	.75	65.3	51.0	67.0	67.0	48.9	49.2	50.7	51.2
32	.10	32.3	32.3	32.7	32.7	28.7	28.7	29.1	29.0
	.40	57.9	57.9	58.8	58.8	49.5	49.5	50.3	50.3
	.75	63.3	64.3	64.8	64.8	48.5	49.3	49.9	50.8
64	.10	40.1	37.3	40.3	40.3	37.0	37.0	37.2	37.2
	.40	57.5	50.2	58.1	58.1	49.7	49.5	50.1	50.1
	.75	62.5	49.3	63.5	63.5	48.3	49.3	50.3	50.3
128	.10	46.5	47.1	47.1	47.1	44.5	44.6	44.6	44.7
	.40	57.4	56.9	57.8	57.8	49.8	49.6	50.1	47.6
	.75	61.5	63.2	62.1	62.1	47.7	49.4	48.5	62.1

6.2.2 Graphical presentation of result showing comparism of the four formulae using Percentage greater than N_{true} for all formulae

Figure 6.5 shows the percentage of overestimation using the four different formulae. The plot showed skewed distribution for all. The median value for the Naive approaches is below median value of about 50% of the times and barely above the median hence they will have sample size estimates that are not evenly distributed around the median. The Browne Cohen and Browne Hedges plot show a more balanced distribution with values below and above the median value of 50% which make them better estimators with Browne Hedges as the best estimator based on these results and graphic.

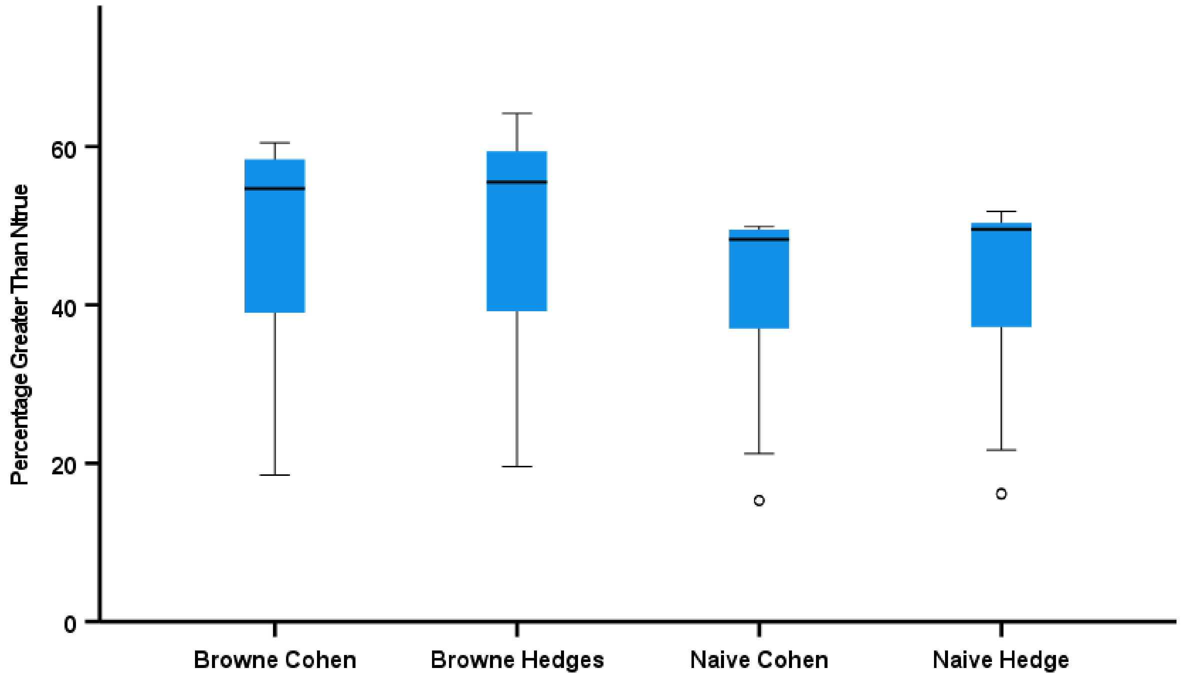


Figure 6.5: Box plot for Percentage Greater than N_{true} at α = (0.01,0.05) , β = (0.1,0.2), $(1-\gamma)$= (0.8,0.9) m = (8, 16, 32, 64,128), and δ = (0.10, 0.40, 0.75).

6.2.3 MPE of Study 2 presented in Tables and Graphs

The overall accuracy of predictions or estimates of the four formulae can be determined using the median percentage error. The MPE is used to quantify the average deviation of the four formulae from N_{true} and this is usually expressed as a percentage and can help to check the under and overestimation of sample sizes by the four formulae. The median percentage error for $N_{B,C}, N_{B,H}$, $N_{N,C}$, and $N_{N,H}$ at different parameter combination will be presented in tables and graphs to further study them.

Results of MPE

For Table 6.6 and 6.7 the MPE for **Naive Cohen's** shows negative values for the different combinations of the various parameters hence indicating on average they are less likely than not to achieve or exceed the true value. Browne **Cohen's** shows for effect sizes of 0.10 the MPE values are negative and this is same for all pilot sample sizes. However, the MPE are positive for 0.4 and 0.75 effect sizes showing that on average they are more likely than not to achieve or exceed the true value.

The **Naive Hedges** shows negative values for effect size of .10 and .40 sample sizes for pilot sample sizes of 8 and 16. For effect size of 0.75,the values are positive for all pilot sample sizes. The results of **Browne Hedge's**, shows negative values for 0.10 effect size and positive for 0.40 and 0.75 effect sizes and this is applicable for both 0.80 and 0.90 coverages. This finding will be a useful guide to researchers in decision making of parameters to pick in research.

Table 6.6: Median percentage error for Study 2 at $\beta = 0.1$ and $(1-\gamma)$ = 0.8.

			Naive-Cohen		Browne-Cohen		Naive-Hedge		Browne-Hedge	
			α		α		α		α	
β	m	δ	.01	.05	.01	.05	.01	.05	.01	.05
.10	8	.10	-92.0	-92.0	-88.1	-88.1	-91.0	-91.0	-86.7	-86.7
		.40	-26.3	-26.5	8.9	8.6	-17.6	-17.7	21.7	21.3
		.75	-3.8	-5.3	41.5	41.2	7.5	7.0	58.5	57.0
	16	.10	-84.2	-84.1	-79.7	-79.6	-83.3	-83.3	-78.6	-78.5
		.40	-6.5	-6.6	20.3	19.9	-1.4	-1.8	26.5	26.0
		.75	-1.9	-2.6	26.4	26.3	3.8	2.6	32.7	31.6
	32	.10	-70.3	-70.5	-65.0	-65.1	-69.6	-69.7	-64.1	-64.2
		.40	-.9	-1.0	17.0	16.9	1.6	1.3	19.9	19.4
		.75	.0	-.9	18.2	15.8	1.9	.9	20.8	18.4
	64	.10	-22.9	-23.2	-16.7	-16.9	-22.5	-22.7	-16.2	-16.4
		.40	.2	-.3	8.4	7.8	.9	.3	9.0	8.6
		.75	.0	.0	8.8	7.9	1.9	.0	9.4	7.9
	128	.10	-22.9	-23.2	-16.7	-16.9	-22.5	-22.7	-16.2	-16.4
		.40	.2	-.3	8.4	7.8	.9	.3	9.0	8.6
		.75	.0	.0	8.8	7.9	1.9	.0	9.4	7.9

The parameter combination for Table 6.6 is like those in Table 6.7, with the only difference being the change in β value from 0.10 to 0.20. The results are however similar and the similarity in results suggest that there is no significant impact due to β.

Table 6.7: Median percentage error for study 2 at β = 0.2, and $(1-\gamma)$ = 0.8.

			Nave-Cohen		Browne-Cohen		Naive-Hedge		Browne-Hedge	
			α		α		α		α	
β	m	δ	.01	.05	.01	.05	.01	.05	.01	.05
.20	8	.10	-92.0	-92.0	-88.1	-88.2	-91.0	-91.0	-86.7	-16.4
		.40	-26.3	-26.6	8.7	8.43	-17.8	-17.8	21.5	21.2
		.75	-4.8	-3.6	41.3	42.9	7.1	7.1	57.9	59.5
	16	.10	-84.2	-84.1	-79.8	-79.5	-83.4	-83.2	-78.7	-78.5
		.40	-6.2	-6.7	20.5	19.9	-1.4	-1.7	26.7	25.9
		.75	-1.6	.0	26.2	28.6	3.2	3.6	33.3	33.3
	32	.10	-70.4	-70.3	-65.0	-64.9	-69.6	-69.5	-64.2	-64.0
		.40	-.9	-1.3	16.9	16.5	1.4	1.0	19.6	19.5
		.75	.0	.0	16.7	17.9	2.4	3.6	19.0	21.4
	64	.10	-49.0	-48.8	-42.8	-42.6	-48.4	-48.2	-42.2	-41.9
		.40	.0	-1.0	11.6	10.8	.9	.0	13.0	12.1
		.75	.0	.0	11.9	14.3	.0	3.6	11.9	14.3
	128	.10	-22.6	-22.8	-16.3	-16.5	-22.1	-22.3	-15.8	-16.0
		.40	.2	-.3	8.0	7.7	.5	.0	8.4	8.1
		.75	.0	.0	7.1	10.7	.0	3.6	9.5	10.7

The parameter combination for Table 6.8 is like Table 6.6, with the only difference being the change in coverage value from 0.80 to 0.90. The results are however different, and this suggests that there is significant impact on the MPE due to coverage for Browne based approaches.

Table 6.8: Median percentage error for study 2 at β = 0.1 and $(1-\gamma)$ = 0.9.

			Naive-Cohen		Brown-Cohen		Naive-Hedge		Browne-Hedge	
β	m	δ	α		α		α		α	
			0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
.10	8	.10	-91.0	-92.0	-88.1	-85.6	-91.0	-91.0	-83.8	-83.9
		.40	-25.8	-26.8	8.9	30.9	-17.2	-18.4	48.4	46.5
		.75	-3.1	-5.3	41.5	70.2	8.2	6.1	93.7	89.5
	16	.10	-84.1	-84.1	-76.8	-76.9	-83.3	-83.3	-75.6	-75.7
		.40	-6.1	-6.1	35.5	37.1	-2.2	-.8	42.7	43.9
		.75	-1.8	-1.8	44.0	43.0	4.4	3.5	51.6	50.0
	32	.10	-61.7	-61.7	-61.9	-61.9	-69.5	-69.6	-60.8	-61.0
		.40	27.4	27.4	26.3	26.8	1.6	1.3	30.6	30.3
		.75	28.3	28.3	26.3	26.3	1.9	.9	32.1	28.9
	64	.10	-48.5	-48.5	-39.0	-38.9	-48.0	-47.9	-38.3	-38.2
		.40	-.3	-.3	18.6	17.9	1.1	.8	19.9	19.4
		.75	.0	.0	18.9	18.4	1.9	.0	20.8	18.4
	128	.10	-22.9	-22.9	-12.4	-13.2	-21.5	-22.5	-11.9	-12.7
		.40	.0	.0	12.7	12.4	.7	.5	13.4	13.1
		.75	.0	.0	13.2	10.5	1.9	.0	13.2	13.2

The parameter combination for Table 6.7 are like those of Table 6.8, with the only difference being the change in $(1-\gamma)$ value from 0.80 to 0.90. The results are however different, this suggests that there is significant impact on the MPE due to coverage. Figure 6.6 shows the impact of coverage on MPE. The range of the median percentage error is seen to increase as the coverage changes from 0.8 to 0.9. This shows there is impact on MPE caused by coverage. Increase in coverage leads to increase in range of MPEs.

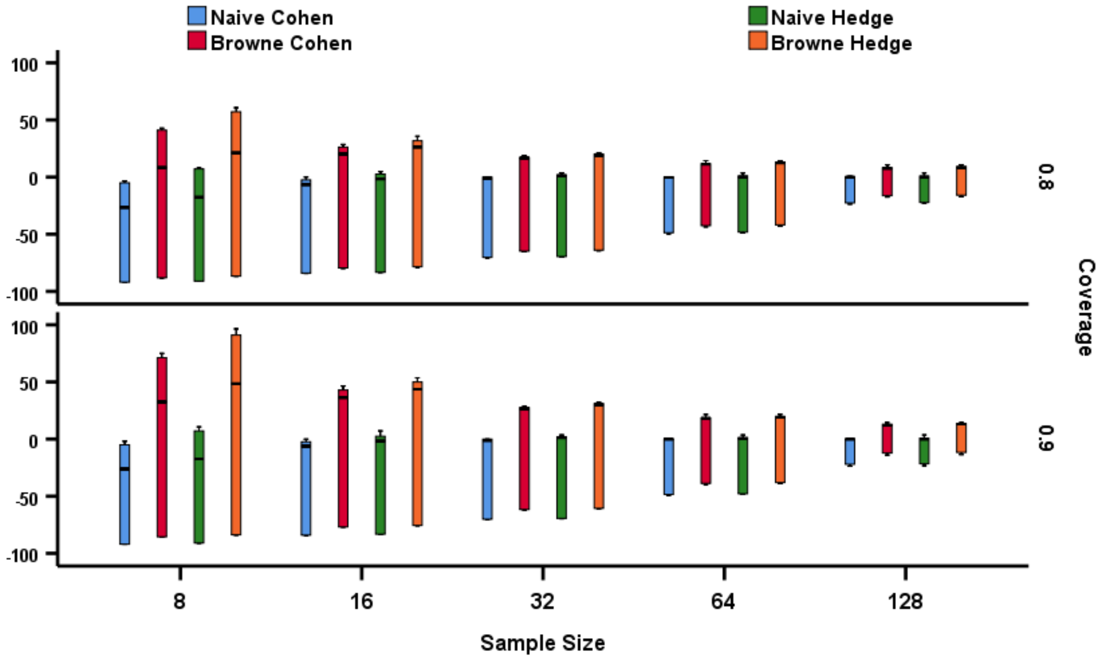


Figure 6.6**:** Median Percentage Error for study 2 showing the impact of coverage.

6.2.4   Graphical comparison of MPE results for study 2

The result of the study median percentage errors for each formula are graphically represented for better understanding. Figure 6.7 shows that Browne Cohen leads to overestimation or underestimation of sample sizes with the median value being above zero. The Browne Hedges show a similar range. The Naive Cohen approach has the median value and upper range below 0 which implies the method will lead to underestimation in all cases. The Naive Hedges show median value at 0 point with the range all below that and implies the method will lead to underestimation most of the time. Median percentage errors being less than zero for both Naive Cohen and Naive Hedges implies that the naive methods underestimate sample sizes. This plot clearly shows the Browne Hedges method as the most suitable of the four for sample size estimation as it's better skewness will give chance for better estimation of sample sizes.

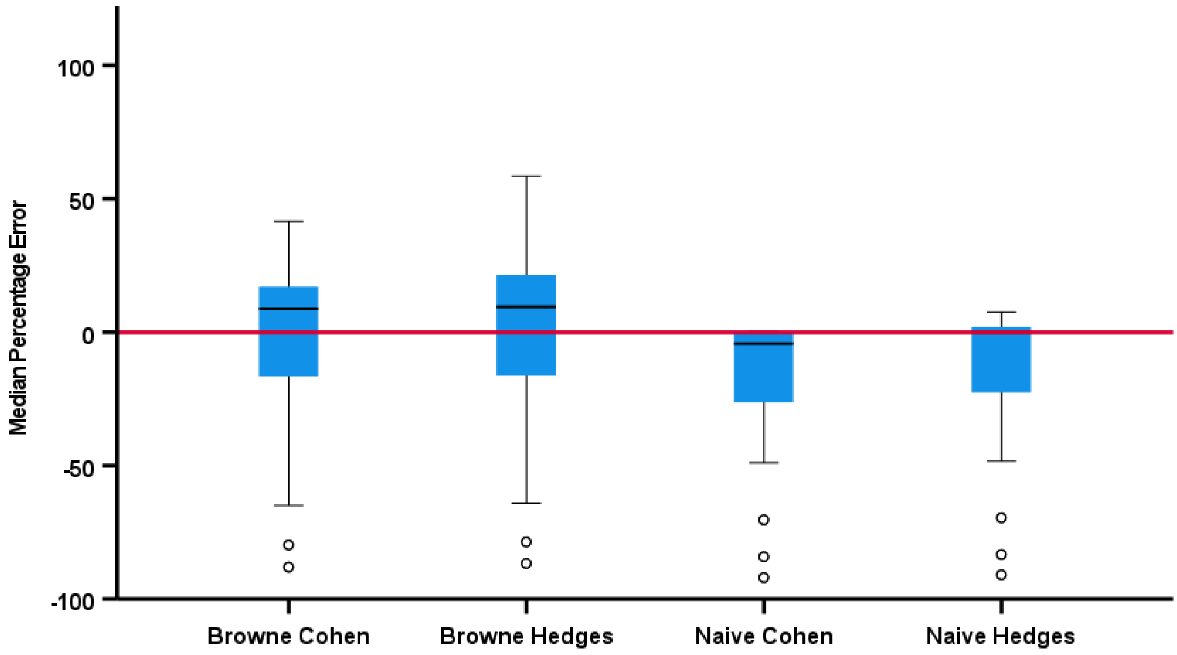


Figure 6.7: Median percentage error of study 2 at α =(0.01,0.05) *,* β = 0.1*,* $(1-\gamma)$ = *0.8,*m = (8, 16, 32, 64,128)*, and* δ = (0.10, 0.40, 0.75).

Figure 6.8 shows the impact of alpha, beta, and pilot sample sizes in the MPE. It shows there is no impact due to alpha as the plots remain the same when alpha values change from 0.01 to 0.05, no impact due to beta as the plots remain the same when Beta changes from 0.1 to 0.2. The range of the plot is seen to reduce as pilot sample sizes changes from 8 to 128, this implies there is impact due to pilot sample sizes. These results are in concurrence with the percentage greater than N_{true} findings.

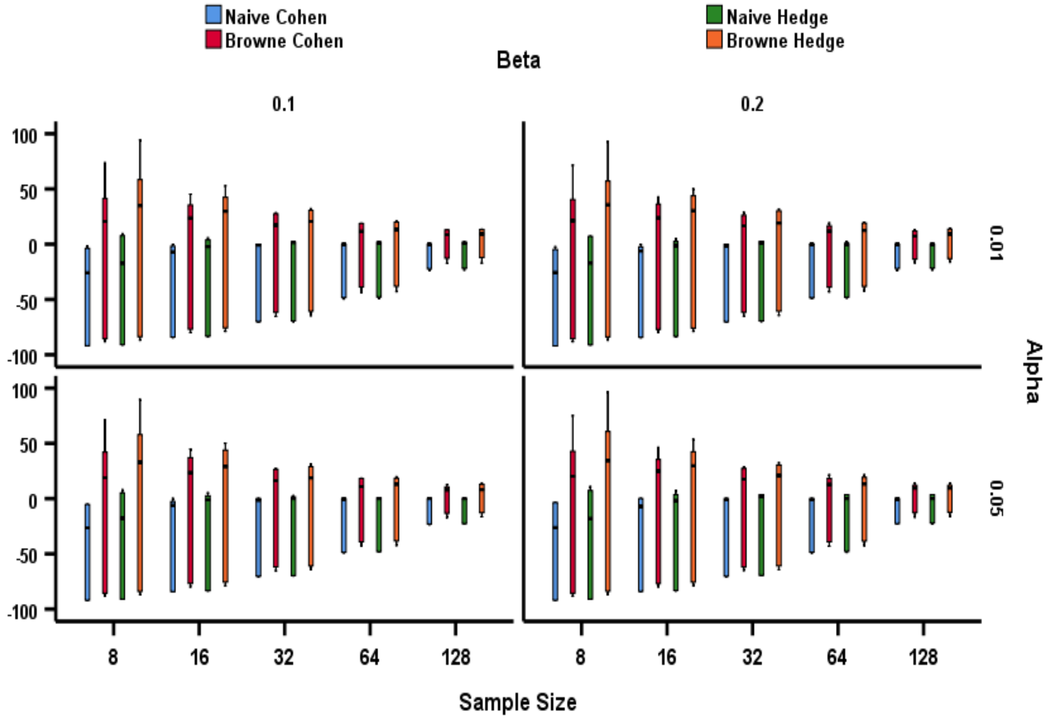


Figure 6.8: Median Percentage error $N_{N,C}, N_{B,H}, N_{B,C}$, and $N_{N,H}$ at different parameter combinations.

Figure 6.9 and 6.10 considered MPE for alpha and beta concurrently results showed that error across all settings is ≤ 0 for $N_{N,C}$ and $N_{N,H}$ but median percentage error > 0 for $N_{B,C}$ and $N_{B,H}$. This shows that naive approaches underestimate whereas previous work has shown a Brown correction can produce large estimates.

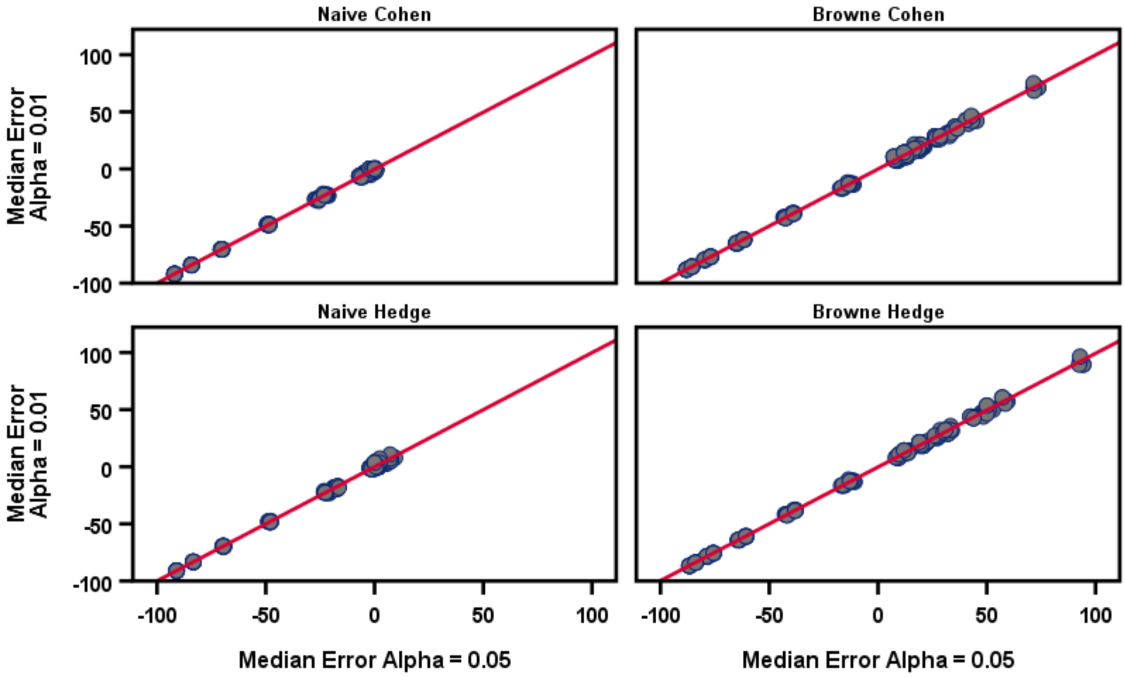


Figure 6.9: Median percentage error of the methods at alpha (0.01, 0.05).

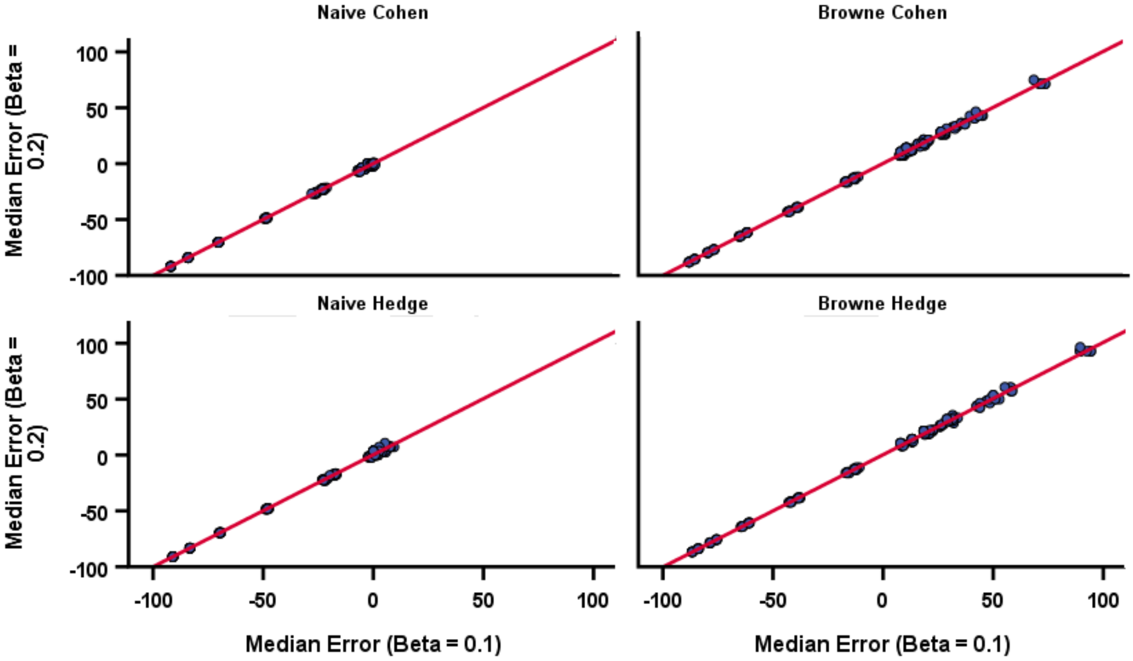


Figure 6.10: Median percentage errors at Beta (0.1,0.2).

6.3 Summary

This chapter presented four formulae for sample size estimation when MCID is unknown $N_{B,H}$, $N_{B,C}$, $N_{N,H}$and $N_{N,C}$. Simulation results for study 2, represented in both tables and graphics, shows that Browne's method, when μ_1 and μ_2 are unknown, does not work effectively. Additionally, using Hedge's correction does not completely solve the problem. The results are not dependent on α or β; that increase in $(1-\gamma)$ gives worse sample size estimates, $(1-\gamma)$ and m has an impact in the MPE by the formulae. The naive methods will lead to underestimation of sample sizes most of the time and Browne Hedges' is seen to be the best of the four formulae with the best range of skewness. results show that in 99.9% of instances the sample sizes followed the prior reasoned logical order of $N_{B,H}$, $N_{B,C}, N_{N,H}, and\ N_{N,C}$ 100% of the time we have $N_{B,H}$ >= $N_{B,C}$> $N_{N,H}$>= $N_{N,C}$, this follows from the formula. The next chapter will develop a method of sample size estimation using upper confidence limit for coefficient of variations.

CHAPTER 7

Sample size estimation method developed from upper confidence limit of coefficient of variations (Study 3).

The study will develop three formulae for sample size estimation using the upper $\gamma\%$ confidence interval for improved versions of coefficient of variation c^2,a measure of variability discussed in Chapter two (2.6) of this research. The new formulae will be improved:

(a) Standard coefficient of variation approach

(b) McKay approach

(c) Vangel approach

The most efficient in the new formulae will be compared to Browne's method N_B. They will be presented by examples and simulation to consider their performance in sample size estimation and recommendations based on findings will be generated.

7.1 Theorem of study 3

The sample size formula is given by

$$n = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2 . \frac{\sigma^2}{(\mu_1 - \mu_2)^2} \tag{7.1}$$

Where σ^2, μ_1 and μ_2 are unknown but may be estimated, the estimated sample size given is given by

$$n_{est} = 2(Z_{1-\alpha/2} + Z_{1-\beta})^2 . \varphi^2 \tag{7.2}$$

where φ^2 is the $\gamma\%$ upper confidence limit for $\sigma^2/(\mu_1 - \mu_2)^2$ .Under an assumption of normality.

Consider $\bar{Y} = \bar{x}_1 - \bar{x}_2$

$$\bar{Y} \sim N(\mu_1 - \mu_2, \sigma_{\bar{Y}}^2)$$

where

$$\sigma_{\bar{Y}}^2 = \sigma^2/m + \sigma^2/m = 2\sigma^2/m \tag{7.3}$$

The coefficient of variation for $\bar{Y}$ is

$$C_{\bar{Y}} = \frac{\sigma_{\bar{Y}}}{\mu_{\bar{Y}}} \tag{7.4}$$

$$= \frac{\sqrt{2\sigma^2/m}}{(\mu_1 - \mu_2)} \tag{7.5}$$

Hence

$$C_{\bar{Y}}^2 = \frac{2\sigma^2/m}{(\mu_1 - \mu_2)^2} \tag{7.6}$$

$$\frac{mC_{\bar{Y}}^2}{2} = \frac{\sigma^2}{(\mu_1 - \mu_2)^2} \tag{7.7}$$

The upper $\gamma\%$ confidence interval for $\sigma^2/(\mu_1 - \mu_2)^2$ is therefore given by the upper $\gamma\%$ confidence interval $mC_{\bar{Y}}^2/2$.

The $\gamma\%$ confidence interval for $C_{\bar{Y}}$ may be given by a few formulae which includes the standard formula, McKay, and Vangel. Using the standard $\gamma\%$ standard co-efficient formula

$$C_{\bar{Y}} \pm t_\gamma \frac{(1 + 1/4n)}{\sqrt{2n}} \tag{7.8}$$

(See Anderson et al., 2020).

In terms of m where $n = 2m$ this confidence interval is given by

$$C_{\bar{Y}} \pm t_\gamma \frac{(1 + 1/8m)}{2\sqrt{m}} \tag{7.9}$$

If $\bar{y} = abs(\bar{x}_1 - \bar{x}_2)$

then

$$C_u = C_{\bar{Y}} + t_\gamma \frac{(1 + 1/8m)}{2\sqrt{m}} \tag{7.10}$$

and

$$\varphi^2 = \frac{m}{2} C_u^2 \tag{7.11}$$

Therefore, by substituting into equation 7.2, the Standard formula is

$$N_{C,S} = (Z_{1-\alpha/2} + Z_{1-\beta})^2 . mC_u^2 \tag{7.12}$$

Subsequently for McKay

$$N_{C,M} = (Z_{1-\alpha/2} + Z_{1-\beta})^2 . mC_{um}^2 \tag{7.13}$$

where C_{um}^2 is upper co-efficient of variation of McKay

$$N_{C,M} = \left((Z_{1-\alpha/2} + Z_{1-\beta})^2 . m . C_{\bar{Y}}^2 . \left(\left(\frac{u_2}{2m} - 1 \right) C_{\bar{Y}}^2 + \frac{u_2}{(2m-2)} \right) \right) \tag{7.14}$$

and $u_2 = \chi^2_{(1-\gamma,2m-2)}$

For the Vangel approach,

$$N_{C,M} = (Z_{1-\alpha/2} + Z_{1-\beta})^2.mC_{uv}^2$$

where C_{uv}^2 is upper co-efficient of variation of Vangel

$$N_{C,V} = \left((Z_{1-\alpha/2} + Z_{1-\beta})^2.m.C_{\bar{Y}}^2.\left(\left(\frac{u_2+2}{2m} - 1\right)C_{\bar{Y}}^2 + \frac{u_2}{(2m-2)}\right)\right) \qquad (7.15)$$

Study 3 will use the same parameters for sample size, effect size, alpha, beta and gamma as Study 2.

Example

For a randomised controlled clinical pilot trial with two groups, the data of the weight of the participants for the two groups were taken and presented below

Grp1(kg)	100,105,109,98,90,87,103,109,113,106,95,106,115,97,99,103,106,118,117,90
Grp2(kg)	106,110,108,103,92,93,100,1]09,115,107,94,103,120,102,103,103,107,119,120,93

Using $\alpha = 0.05$ and $\beta = 0.1$

Estimate the required sample size for the substantive trial using the upper 80% confidence interval for c using the

a. Standard coefficient of variation formula $(N_{C,S})$

b. McKay method $(N_{C,M})$

c. Vangel method $(N_{C,V})$

d. Browne's method

$$\alpha = 0.05$$

$$\beta = 0.2$$

$$m_1 + m_2 = m = 40$$

$$(\bar{x}_1 - \bar{x}_2) = 2$$

Solution

a. Sample size calculation using Standard coefficient of variation formula $(N_{C,S})$

$$N_{C,S} = (Z_{1-\alpha/2} + Z_{1-\beta})^2.mC_u^2$$

$$C_{\bar{Y}}^2 = \frac{2s_p^2/m}{(\bar{x}_1 - \bar{x}_2)^2}$$

$$C_{\bar{Y}}^2 = \frac{2(77.38)/40}{(2)^2}$$

$$C_{\bar{Y}}^2 = 0.9672$$

$$C_{\bar{Y}} = 0.9835$$

$$C_u = C_{\bar{Y}} + t_{\gamma} \frac{(1 + 1/8m)}{2\sqrt{m}}$$

$$C_u = 0.9835 + 0.8512 \frac{(1 + 1/8(20)}{2\sqrt{20}}$$

$$C_u = 1.0793$$

$$C_u^2 = 0.9835 + 0.8512.0.1125$$

$$C_u^2 = 1.0793$$

$$N_{C,S} = (1.9599 + 1.2816)^2 . 20 . (1.0793)$$

$$N_{C,S} = 226$$

b. McKay method ($N_{C,M}$)

$$N_{M,V} = (Z_{1-\alpha/2} + Z_{1-\beta})^2 . m . C_{\bar{Y}}^2 . \left(\frac{u_2}{2m} - 1\right) C_{\bar{Y}}^2 + \frac{u_2}{(2m-2)}$$

$$u_2 = 30.5373$$

$$= (1.959 + 1.282)^2 . 20(0.9672)(0.5670)$$

=115

c. Vangel method ($N_{C,V}$)

$$N_{C,V} = (Z_{1-\alpha/2} + Z_{1-\beta})^2 . m C_{\bar{Y}}^2 . \left(\frac{u_2 + 2}{2m} - 1\right) C_{\bar{Y}}^2 + \frac{u_2}{(2m-2)}$$

$$= (1.959 + 1.282)^2 . 20 . (0.9672)(0.6231)$$

$$= 127$$

e. Browne formula (N_B)

$$N_B = \frac{1 + r}{r} . \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta)}\right)^2}{(\bar{x}_1 - \bar{x}_2)^2} . k s^2$$

where $k = (m_1 + m_2 - 2)/(\chi^2_{1-\gamma,\nu})$, and $\chi^2_{1-\gamma,\nu}$ is the upper 80% point of the chi-square variate on υ degrees of freedom.

$$\chi^2_{1-\gamma,\nu} = 45.076$$

$$k = \frac{38}{30.5373} = 1.2444$$

$$N_B = 2 . \frac{(1.9599 + 1.2816)^2}{(2)^2} . (1.2444)(77.38)$$

$$N_B = 505$$

From the result the sample size that should be used for the substantive trial using Standard coefficient of variation formula $(N_{C,S})$, McKay method $(N_{C,M})$**,** Vangel method $(N_{C,V})$, and Brown formula (N_B) are respectively 226,115,127 and 505. However, there are other factors that need to be considered including attaining the desired power for the research before concluding on the sample size hence the need to further view the details of this methods using simulations. The parameter combination for the simulation is given as follows.

Table 7.1: Parameter combinations for study 3.

FACTOR	Number of Levels	LEVELS
Alpha(α)	2	0.01,0.05
Beta (β)	2	0.1,0.2
Coverage $(1-\gamma)$	2	0.8, 0.9,
Effect size (δ)	3	0.1, 0.4,0.75
Pilot sample size (m)	5	8, 16, 32, 64, 128

7.1.1 Algorithm for study 3

1. Set the seed at 101
2. Set the number of iterations 100,000
14. Create the vector for α= (0.01,0.05), β =(0.1,0.2), $(1-\gamma)$ =(0.8,0.9), δ =(0.1,0.4,0.75), and m =(8,16,32,64, 128).
3. Loop over the parameters
4. For each alpha value calculate the k1 = $Z_{(1-\alpha/2)}$ using the qnorm function
5. For each beta value calculate the value calculate k2 = $Z_{(1-\beta)}$ using the qnorm function
6. For each coverage probability, loop over the sample size
7. For each sample size calculate the critical t value using the qt function and calculate the chi square using qchisq.
8. Calculate the true sample size required using $Ntrue$ formula.
9. Initialize empty vectors for all thresholds

10. Loop over the iteration
11. Using rnorm function with mean 0 and the effect size generate two random samples.
12. Calculate the pooled var, difference in means.
13. Calculate the estimated sample size using the Standard approach (Repeat process for the McKay and the Vangel formula)
14. For GreaterThanN_{true} vector if the estimated sample size is greater than or equal the true samplesize append 1 otherwise 0
15. If the estimated sample size is greater than or equal to a threshold append 1 for the various thresholds, otherwise append 0.
16. Repeat 11 to 14 for all iterations.
17. Give an output of the results showing the number of times estimated sample size exceeds the true sample size and the various threshold for each combination of parameters.
18. End

7.2 Study 3 results

The findings from Standard coefficient of variation formula $N_{C,S}$ in tables

Table 7.2 illustrates that for an δ of 0.1 the percentage greater than $Ntrue$ is 16% at m=8, indicating a severe underestimation. At m = 16, the percentage increases to 22%, still indicating underestimation but not as severe as at m = 8. For m = 128 the percentage > $Ntrue$ is 45%, implying about half the time it will underestimate and half overestimate.

For δ of 0.4 the percentage greater than $Ntrue$ at m=8 is 50%, indicating half the time it will underestimate and half overestimate. At m = 16, the percentage becomes 55%, underestimation about 45% of the time. For m = 128 the percentage > $Ntrue$ is 67% suggesting underestimation around 33% of the time.

For large δ of 0.75 of the time the estimated sample size is larger than $Ntrue$ about 80% for $m = 32, 64 \text{ and } 128$. The result suggests the standard approach demonstrates there is underestimation when effect sizes are small, and this is partially corrected by increasing sample size. However, this approach generally shows overestimation which is not corrected by increasing pilot sample size.

To illustrate this percentage greater than 50% is chosen and graphically presented in Figure 7.1 and 7.2 to show the general impact of pilot sample size and effect size.

Table 7.2: Proportion of time estimated sample size exceeds N_{true} +/- error at $\alpha = 0.05$, $\beta = 0.2$, and $(1 - \gamma) = 0.8$.

m	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.180	.161	.145	.139	.129	.110	.110
16	.10	.247	.220	.201	.193	.181	.159	.159
32	.10	.330	.294	.269	.258	.242	.213	.213
64	.10	.417	.378	.348	.334	.310	.269	.269
128	.10	.499	.448	.411	.397	.370	.318	.318
8	.40	.546	.491	.446	.428	.394	.342	.301
16	.40	.616	.548	.494	.473	.437	.373	.327
32	.40	.673	.583	.513	.482	.434	.354	.301
64	.40	.726	.610	.508	.470	.405	.296	.227
128	.40	.812	.662	.526	.467	.372	.229	.150
8	.75	.756	.667	.584	.550	.501	.406	.345
16	.75	.838	.733	.625	.576	.509	.377	.295
32	.75	.918	.809	.674	.610	.517	.332	.230
64	.75	.928	.898	.743	.660	.522	.264	.138
128	.75	.997	.960	.811	.704	.520	.185	.062

Figure 7.1 shows there is large change caused by m as the proportion of excess value for 8,16,32,64 and 128 varies greatly. The proportion for $m = 8$ between .1 to .5, $m = 16$ is between .08 to .55, $m = 32$ between .25 to .45, $m = 64$ between .30 to .50, and $m = 128$ between .30 to .50. The range of the proportion of excess is seen to reduce as the pilot sample sizes increases. Implying an impact in variance of under /over estimation by the formulae as the pilot sample size increases.

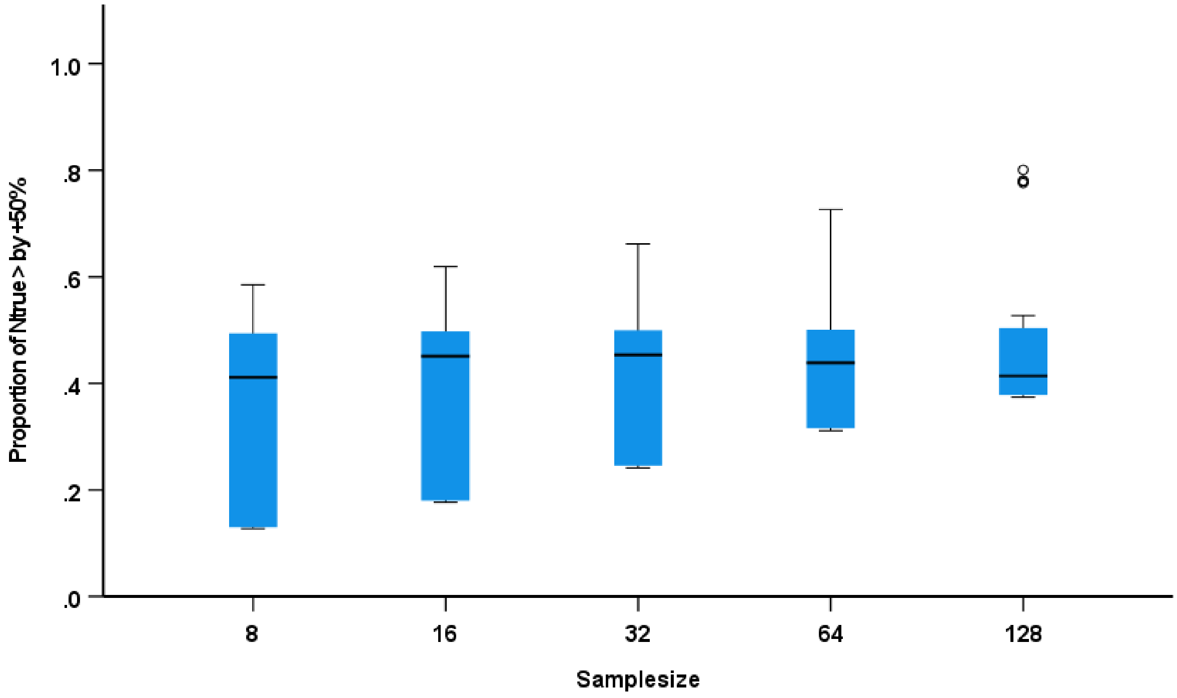


Figure 7.1: Proportion > N_{true} +50% for samplesize.

Figure 7.2 shows there is change caused by δ as the proportion value for both 0.1,0.4 and 0.75 varies. For effect size of 0.10 the proportion ranges from 0.1 to 0.3 with the median value of about 0.22, for 0.40 the proportion ranges from 0.38 to 0.41 with the median value of about 0.4 and for 0.75 the proportion ranges from 0.38 to 0.41 with the median value of about 0.44 to .58 with median value of 0.45.

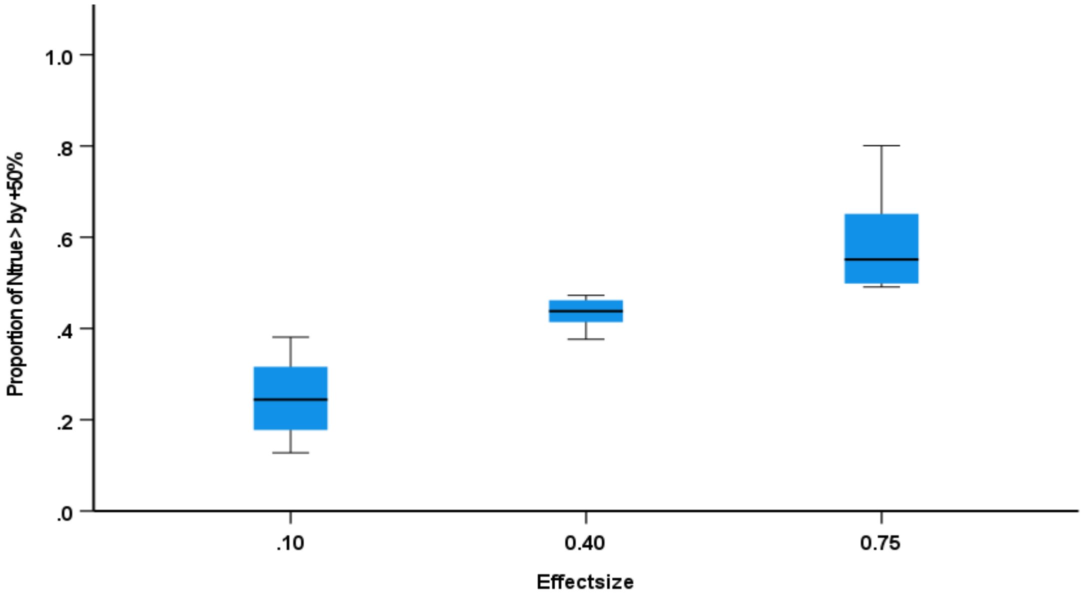


Figure 7.2: Proportion > N_{true} +50% for effect size.

For Table 7.3 there is change in beta values to 0.1 as against 0.2 in Table 7.2 however the results presented in the table are similar. This shows that β has little to no effect in the over-under estimation of sample sizes by this standard formula.

Table 7.3: Proportion of time estimated sample size exceeds $Ntrue$ +/- error at $\alpha = 0.05$, $\beta = 0.1, (1 - \gamma) = 0.8$.

m	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.174	.156	.141	.136	.126	.109	.098
16	.10	.247	.221	.201	.193	.180	.156	.140
32	.10	.330	.295	.270	.260	.242	.210	.189
64	.10	.424	.382	.351	.337	.315	.274	.245
128	.10	.508	.459	.420	.404	.377	.328	.294
8	.40	.550	.493	.448	.430	.399	.345	.306
16	.40	.619	.550	.496	.473	.438	.372	.328
32	.40	.676	.584	.512	.484	.436	.354	.301
64	.40	.740	.619	.518	.478	.413	.302	.237
128	.40	.819	.667	.527	.471	.378	.231	.154
8	.75	.746	.656	.574	.540	.490	.396	.336
16	.75	.830	.721	.609	.562	.492	.362	.283
32	.75	.914	.799	.659	.595	.498	.316	.214
64	.75	.973	.882	.721	.636	.497	.250	.134
128	.75	.997	.953	.794l	.686	.497	.172	.058

Figure 7.3 shows there is no change in over/under estimation caused by β as the proportion value for both .10 and 0.20 are similar.

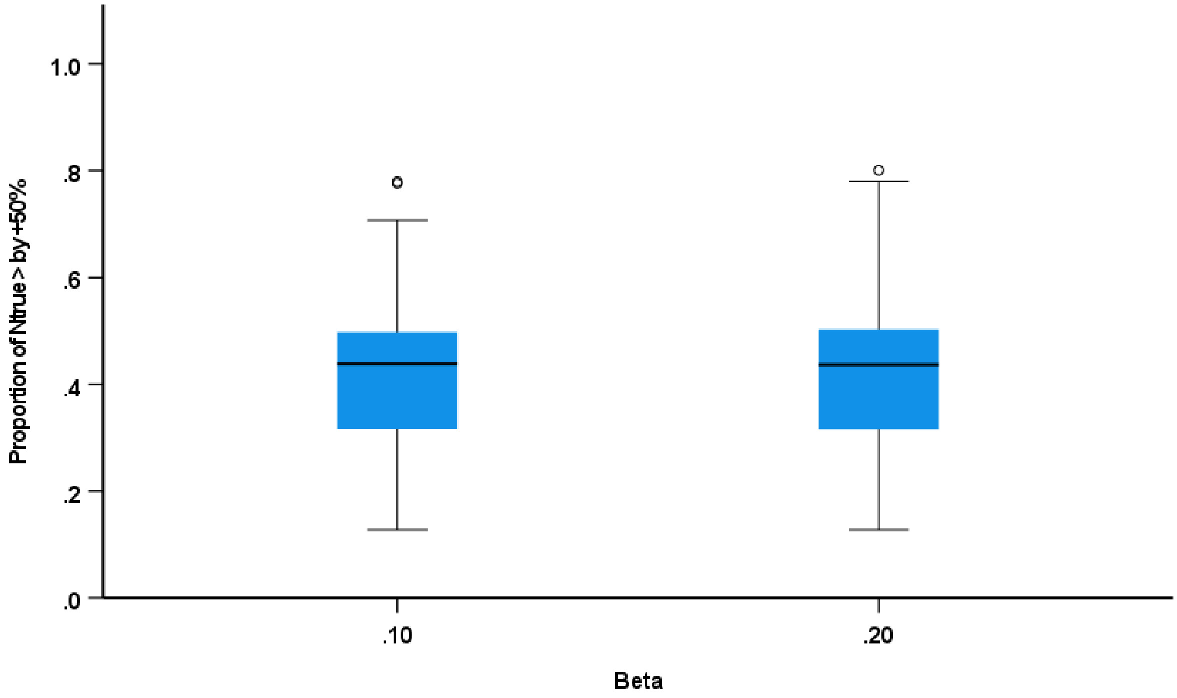


Figure 7.3: Proportion $> N_{true}$ +50% for Beta.

Table 7.4 used $\alpha = 0.01$ as against results from table 7.3 where $\alpha = 0.05$. Results displayed in both tables are however similar and this shows there is no effect caused by α in proportion of over or underestimation of sample sizes this will be further graphically presented.

Table 7.4*:* Proportion of time estimated sample size exceeds N_{true} +/- error at $\alpha = 0.01$, $\beta = 0.1$,$(1 - \gamma) = 0.8$.

m	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.176	.157	.143	.137	.128	.110	.098
16	.10	.243	.217	.198	.191	.178	.153	.137
32	.10	.328	.294	.270	.259	.242	.209	.187
64	.10	.426	.383	.351	.338	.317	.275	.246
128	.10	.506	.456	.418	.402	.376	.326	.293
8	.40	.554	.496	.450	.432	.402	.347	.309
16	.40	.618	.550	.496	.475	.439	.374	.331
32	.40	.676	.585	.514	.487	.440	.357	.304
64	.40	.740	.619	.519	.480	.414	.305	.239
128	.40	.819	.667	.530	.475	.383	.236	.157
8	.75	.751	.659	.579	.549	.493	.402	.339
16	.75	.835	.724	.617	.574	.496	.366	.286
32	.75	.917	.803	.666	.610	.500	.322	.218
64	.75	.975	.888	.730	.655	.503	.257	.136
128	.75	.997	.957	.810	.716	.506	.182	.062

Figure 7.4 shows there is little to no change caused by Alpha as the proportion value for both 0.01 and 0.05 are quit similar, ranging from 0.3 to 0.5 with the median value of about 0.45.

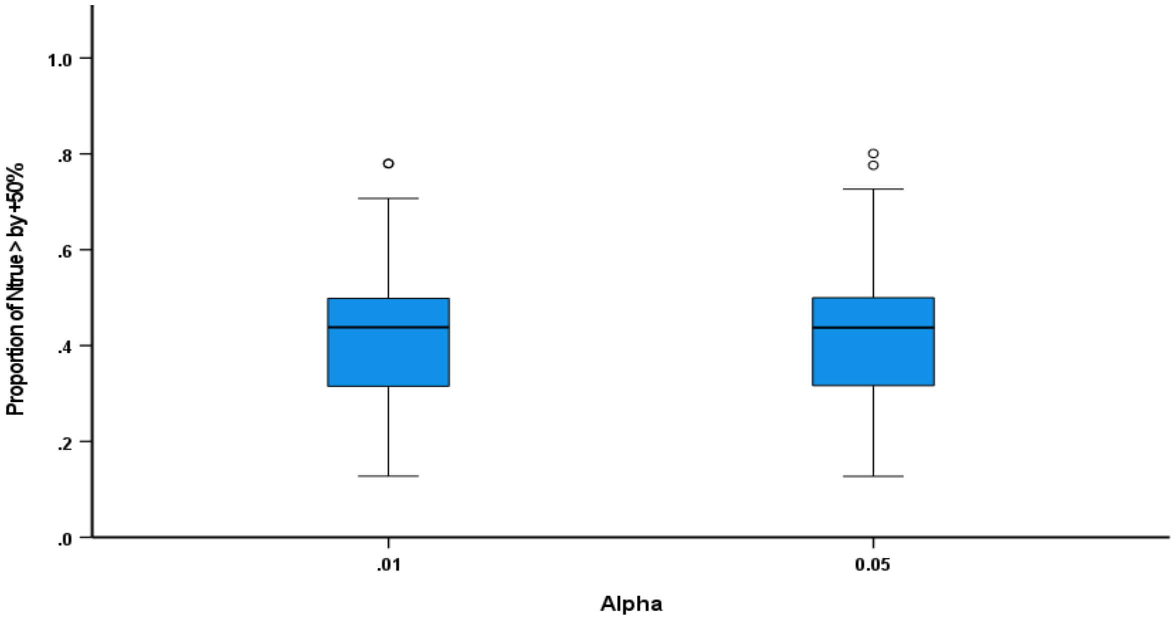


Figure 7.4: Proportion > N_{true} +50% for Alpha.

Table 7.5 used similar parameters as table 7.3 with the only for coverage changing to 0.9. There is a slight change increase in the result values implying there is impact caused by change in coverage values. The findings are however same as the results shows underestimations with only large effect size attaining the desired coverage at $m = 32, 64$ and 128.

Table 7.5: Proportion of time estimated sample size exceeds N_{true} +/- error at $\alpha = 0.05$, $\beta = 0.1, (1 - \gamma) = 0.9$.

m	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.178	.159	.144	.138	.128	.111	.099
16	.10	.251	.224	.204	.196	.182	.158	.141
32	.10	.336	.300	.274	.263	.245	.212	.191
64	.10	.431	.388	.355	.342	.319	.277	.248
128	.10	.516	.466	.426	.409	.382	.332	.297
8	.40	.594	.529	.479	.459	.424	.363	.321
16	.40	.669	.592	.532	.507	.465	.392	.344
32	.40	.738	.638	.558	.525	.471	.379	.319
64	.40	.815	.691	.581	.536	.461	.334	.259
128	.40	.898	.760	.615	.553	.446	.271	.178
8	.75	.867	.774	.679	.638	.576	.455	.380
16	.75	.940	.853	.743	.688	.605	.439	.337
32	.75	.985	.931	.820	.757	.648	.418	.280
64	.75	.999	.982	.901	.836	.704	.384	.207
128	.75	1.000	.998	.966	.916	.774	.341	.125

Figure 7.5 shows there is change caused by $(1-\gamma)$ as the proportion value for both 0.8, and 0.9 varies. For coverage of .80 the proportion ranges from .3 to .5 with the median value of about 0.42, and for .90 the proportion ranges from 0.30 to 0.61 with the median value of about 0.42.

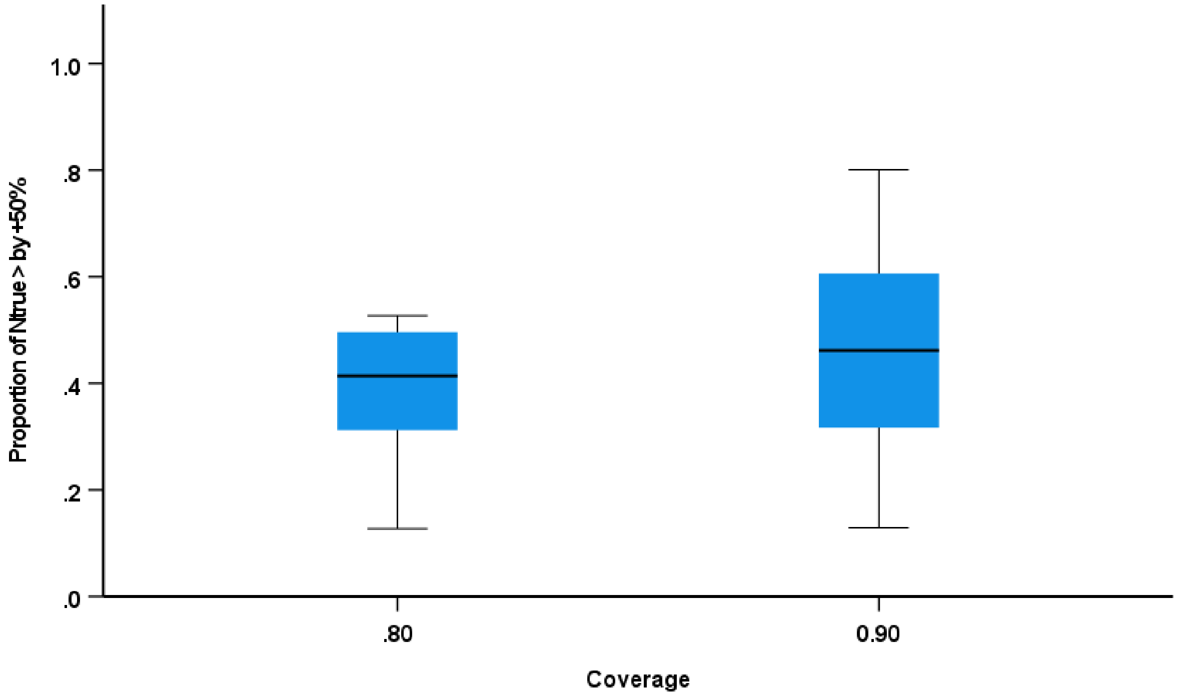


Figure 7.5: Proportion > N_{true} +50% for Coverage.

The findings from McKay formula $N_{C,M}$ in tables

Table 7.6 illustrates that for an δ of 0.1 the percentage greater than N_{true} is 25% at m=8, indicating a severe underestimation. At m = 16, the percentage increases to 30%, still indicating underestimation but not as severe as at m = 8. For m = 128 the percentage > N_{true} is 48%, implying about half the time it will underestimate and half overestimate.

For δ of 0.4 the percentage greater than N_{true} at m=8 is 51%, indicating half the time it will underestimate and half overestimate. At m = 16, the percentage becomes 55%, indicating underestimation about 45% of the time. For m = 128 the percentage > N_{true} is 55% suggesting underestimation around 45% of the time.

For large δ of 0.75 percentage greater than N_{true} at m=8 is 58%, indicating about 42% of the time it will underestimate. At m = 16, the percentage becomes 58%, indicating underestimation about 42% of the time. For m = 128 the percentage > N_{true} is 58% suggesting underestimation around 42% of the time. The result suggests the Mckay approach demonstrates there is underestimation with at all effect sizes and the power is not achieved at any parameter combination.

Table 7.6: Proportion of time estimated sample size exceeds N_{true} +/- error at $\alpha = 0.01$, $\beta = 0.1$,$(1-\gamma) = 0.8$ using McKay formula.

m	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.263	.245	.231	.226	.216	.198	.185
16	.10	.322	.299	.280	.273	.260	.237	.221
32	.10	.387	.356	.333	.324	.309	.279	.259
64	.10	.457	.419	.390	.378	.357	.319	.297
128	.10	.522	.476	.441	.414	.402	.356	.324
8	.40	.559	.514	.478	.463	.437	.389	.356
16	.40	.607	.549	.506	.488	.456	.399	.361
32	.40	.631	.555	.495	.473	.433	.361	.314
64	.40	.657	.550	.468	.437	.381	.287	.231
128	.40	.695	.551	.434	.436	.317	.202	.140
8	.75	.646	.579	.522	.501	.459	.391	.343
16	.75	.679	.588	.509	.479	.424	.331	.270
32	.75	.717	.591	.479	.437	.361	.243	.174
64	.75	.764	.597	.436	.378	.279	.145	.081
128	.75	.825	.581	.382	.306	.185	.059	.021

The findings from Vangel formula $N_{C,V}$ in tables

Table 7.7 illustrates that for an δ of 0.1 the percentage greater than N_{true} is 27% at m=8, indicating a severe underestimation. At m = 32, the percentage increases to 32%, still indicating underestimation but not as severe as at m = 8. For m = 128 the percentage > N_{true} is 48%, implying about half the time it will underestimate and half overestimate.

For δ of 0.4 the percentage greater than N_{true} at m=8 is 53%, indicating 47% of the time it will underestimate. At m = 16, the percentage becomes 56%, indicating underestimation about 14% of the time. For m = 128 the percentage > $Ntrue$ is 55% suggesting underestimation around 45% of the time.

For large δ of 0.75 percentage greater than N_{true} at m=8 is 55%, indicating about 45% of the time it will underestimate. At m = 16, the percentage becomes 59%, indicating underestimation about 41% of the time. For m = 128 the percentage > N_{true} is 60% suggesting underestimation around 40% of the time. The result suggests the Vangel approach demonstrates there is underestimation at all effect size, and the power is not achieved at any parameter combination.

Table 7.7: Proportion of time estimated sample size exceeds N_{true}+/- error at $\alpha = 0.01$, $\beta = 0.1$,$(1 - \gamma) = 0.8$ using Vangel formula.

Sample Size	Effect Size	>-20%	> N_{true}	>+20%	>+30%	>+50%	>+100%	>+150%
8	.10	.293	.274	.259	.253	.243	.224	.211
16	.10	.339	.315	.298	.291	.277	.252	.236
32	.10	.396	.365	.343	.334	.318	.289	.268
64	.10	.461	.423	.395	.383	.362	.324	.299
128	.10	.523	.477	.443	.429	.404	.358	.327
8	.40	.575	.531	.496	.493	.482	.457	.411
16	.40	.613	.556	.513	.513	.496	.464	.409
32	.40	.633	.558	.498	.498	.476	.436	.365
64	.40	.658	.552	.469	.469	.438	.382	.289
128	.40	.695	.551	.435	.435	.392	.318	.202
8	.75	.653	.588	.532	.512	.471	.402	.356
16	.75	.683	.592	.514	.484	.428	.336	.276
32	.75	.718	.597	.481	.439	.363	.244	.176
64	.75	.824	.600	.437	.379	.280	.145	.082
128	.75	.657	.590	.383	.307	.186	.060	.022

Results from Standard coefficient of variation formula, McKay formula and Vangel formula shows the formula developed from co-efficient of variation also lead to over/under estimation of sample sizes. The standard method is seen to be the only that can attain the required power only for large effect sizes and hence it will be compared to Browne's method.

7.3 Comparison of Browne's approach of sample size estimation (N_B) and sample size estimation approach developed from upper confidence limit of standard coefficient of variation $(N_{C,S})$

Table 7.8 shows both approaches achieved the desired 80% coverage at large effect size of 0.75.

Table 7.8: Percentage of time estimated sample size exceeds N_{true} +/- error at $\alpha = 0.05$, $\beta = 0.2, (1-\gamma) = 0.8$.

		$N_{C,S}$	N_B	$N_{C,S}$	N_B	$N_{C,S}$	N_B
Sample Size	Effect Size	$> N_{true}$	$> N_{true}$	>+50%	>+50%	>+100%	>+100%
8	.10	.161	.801	.129	.436	.110	.166
16	.10	.220	.798	.181	.240	.159	.026
32	.10	.294	.798	.242	.073	.213	.001
64	.10	.378	.799	.310	.006	.269	.000
128	.10	.448	.801	.370	.000	.318	.000
8	.40	.491	.795	.394	.421	.342	.162
16	.40	.548	.790	.437	.229	.373	.024
32	.40	.583	.786	.434	.065	.354	.000
64	.40	.610	.782	.405	.005	.296	.000
128	.40	.662	.771	.372	.000	.229	.000
8	.75	.667	.795	.501	.430	.406	.166
16	.75	.733	.798	.509	.237	.377	.025
32	.75	.809	.794	.517	.071	.332	.000
64	.75	.898	.795	.522	.006	.264	.000
128	.75	.960	.789	.520	.000	.185	.000

Table 7.9 is same parameter as table 7.8 with only the $(1-\gamma)$ value changing to 0.9 and the results vary slightly showing there is impact due to coverage. Both approaches achieved 90% coverage at large effect size of 0.75

Table 7.9: Percentage of time estimated sample size exceeds $Ntrue$ +/- error at $\alpha = 0.01$, $\beta = 0.1$,$(1 - \gamma) = 0.9$.

		$\mathrm{N_{C,S}}$	$\mathrm{N_B}$	$\mathrm{N_{C,S}}$	$\mathrm{N_B}$	$\mathrm{N_{C,S}}$	$\mathrm{N_B}$
Sample Size	Effect Size	$> N_{true}$	$> N_{true}$	>+50%	>+50%	>+100%	>+100%
8	.10	.160	.900	.130	.631	.112	.339
16	.10	.221	.901	.180	.431	.155	.085
32	.10	.299	.900	.244	.420	.211	.003
64	.10	.389	.900	.320	.171	.278	.000
128	.10	.463	.901	.380	.024	.330	.000
8	.40	.532	.901	.427	.634	.365	.340
16	.40	.591	.899	.466	.419	.395	.084
32	.40	.637	.900	.475	.175	.381	.003
64	.40	.690	.901	.462	.026	.336	.000
128	.40	.759	.900	.450	.000	.276	.000
8	.75	.778	.899	.577	.624	.461	.340
16	.75	.858	.898	.608	.408	.445	.084
32	.75	.934	.900	.650	.163	.427	.003
64	.75	.983	.898	.708	.020	.397	.000
128	.75	.999	.896	.782	.000	.357	.000

7.4 Summary

New formulae for sample size estimation using upper confidence limit of co-efficient of variations was developed .They are namely (a) the upper γ% confidence interval for c^2 using the Standard coefficient of variation formula ($N_{C,S}$), (b) the upper γ% confidence interval for c using the McKay ($N_{C,M}$), and (c) the upper γ% confidence interval for c using the Vangel formula ($N_{C,V}$).They were studied using examples and simulations for various combinations of α , β, $(1 - \gamma), \delta$ and m.The findings was presented in table and graphically. The result show that the Standard method can be used for sample size estimation however the desired coverage is attained for only large effect sizes of 0.75. The $N_{C,S}$ formula presented better outcome compared to the other two formulae and hence was compared to Browne's method. Similar, to Browne's method this method can lead to over or under estimation of sample sizes.

CHAPTER 8

Sample size determination using k^2 modifier based on coverage

Modification for sample size determination for two-armed trial is further discussed. The study 4 of this research which develops modified approach using k^2 modifier based on coverages when the Minimum Clinical Importance Difference is unknown is presented. Standard method for sample size estimation is compared to a modified method based on observed data using k^2 modifier values to achieve desired coverage, where k is Browne's multiplier discussed in chapter one (1.2.1).

8.1 Pre study hypothesis for Study 4

For the two-armed RCT with normally distributed data, then

$$N_{true} = 2\left(Z_{1-\alpha/2} + Z_{\beta}\right)^2 \frac{\sigma^2}{(\mu_1 - \mu_2)^2} = \frac{2\left(Z_{1-\alpha/2} + Z_{\beta}\right)^2}{\delta^2} \qquad (8.1)$$

Suppose we consider

$$N_{est} = 2\left(Z_{1-\alpha/2} + Z_{\beta}\right)^2 \frac{s^2}{(\bar{x}_1 - \bar{x}_2)^2} = \frac{2\left(Z_{1-\alpha/2} + Z_{\beta}\right)^2}{d^2} \qquad (8.2)$$

In this case $N_{est} > N_{true}$ when

$$\frac{1}{d^2} > \frac{1}{\delta^2} \qquad (8.3)$$

or equivalently when

$$\delta^2 > d^2 \qquad (8.4)$$

In other words when Cohen's d is smaller than the true effect size (δ) then the process will overestimate the sample size; and when Cohen's d is larger than the effect size (δ) then the process will underestimate the sample size. Consider $\delta = 1$. The distribution of d^2 for $m = 5$ is shown in Figure 8.1 (based on 100,000 simulation instances). From the histogram, the 80-th percentile for m = 5 and δ = 1 is estimated to be 2.85. Hence, dividing d^2 by is 2.85 would ensure 80% of the values would be smaller than 1.

Let $k^2(1 - \lambda, m, \delta) = k^2(0.8, 5, 1)$ = 2.85, (8.5)

then

$$N_{modified} = 2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \frac{s^2}{(\bar{x}_1 - \bar{x}_2)^2}\, k^2_{0.8,5,1} \tag{8.6}$$

$$= \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{d^2}\, k^2_{0.8,5,1} \tag{8.7}$$

$$= \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{d^2}\, 2.85 \tag{8.8}$$

Equation 8.8 would provide an estimate of N_{true} which would have 80% coverage provided $\delta = 1$.

$$N_{modified} = 2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \frac{s^2}{(\bar{x}_1 - \bar{x}_2)^2}\, k^2_{0.7,5,1} \tag{8.9}$$

Equation 8.9 would provide estimate of N_{true} which would have 70% coverage.

$$N_{modified} = 2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \frac{s^2}{(\bar{x}_1 - \bar{x}_2)^2}\, k^2_{0.6,5,1} \tag{8.10}$$

Equation 8.10 would provide an estimate of N_{true} which would have 60% coverage provided $\delta = 1$.

8.2 Graphical representation of study 4.

The finding for 80% coverage is graphically shown in Figure 8.1

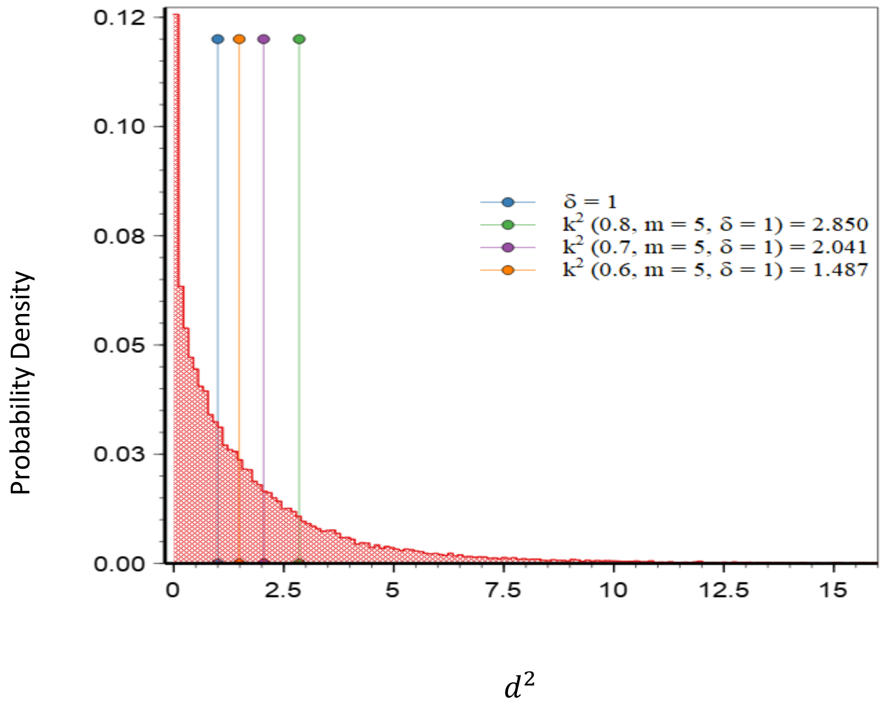


Figure 8.1: Graphical representation of probability density for varying d^2.

Now consider when δ = 0.05. Figure 8.2 shows that when $\delta = 0.5$ the distribution of d^2 has an estimated 80-th percentile of 0.5216 and the distribution of d^2/δ^2 when $\delta = 0.5$ has a percentile of 2.1046. Hence $k^2(1 - \gamma = 0.8, m = 30,\ \delta = 0.5)$ = 2.105

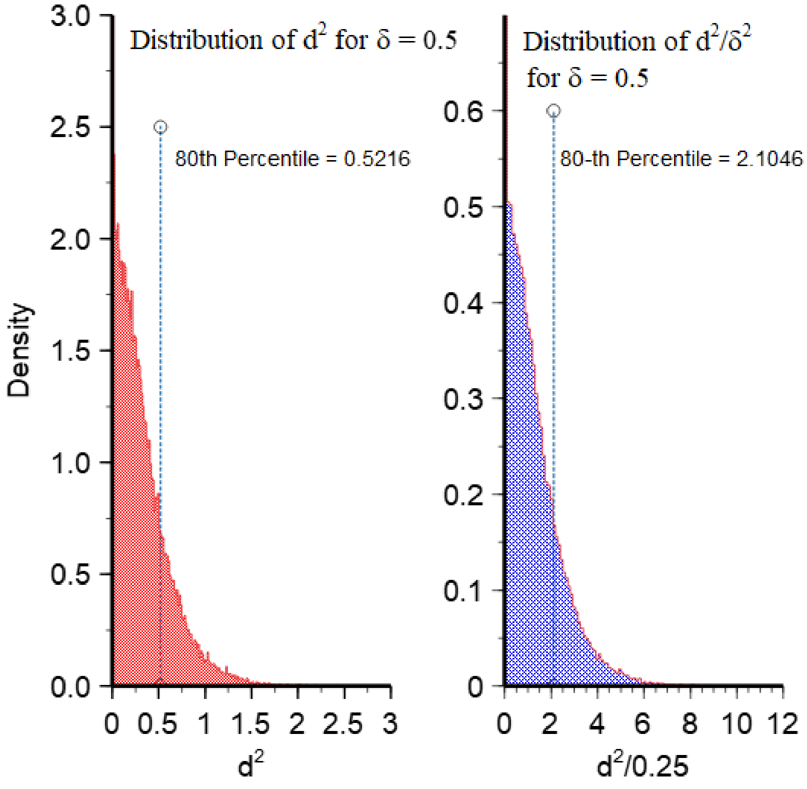


Figure 8.2: Representation for δ = 0.05 when d^2 has an estimated 80-th percentile.

As a third example consider δ = 0.2, m = 40. The 60th percentile of $d^2$is estimated to be 0.07235, the 70th percentile of d^2 is estimated to be 0.1058, the 80th percentile of d^2 is estimated to be 0.15544. Hence, dividing the values of the above percentiles by $\delta^2 = 0.2^2 = 0.04$ gives

$k^2(1 - \gamma = 0.6, m = 40, \delta = 0.2)$ is estimated to be 1.809

$k^2(1 - \gamma = 0.7, m = 40, \delta = 0.2)$ is estimated to be 2.645

$k^2(1 - \gamma = 0.8, m = 40, \delta = 0.2)$ is estimated to be 3.886

The 80% percentile transformation is shown in the graphic below.

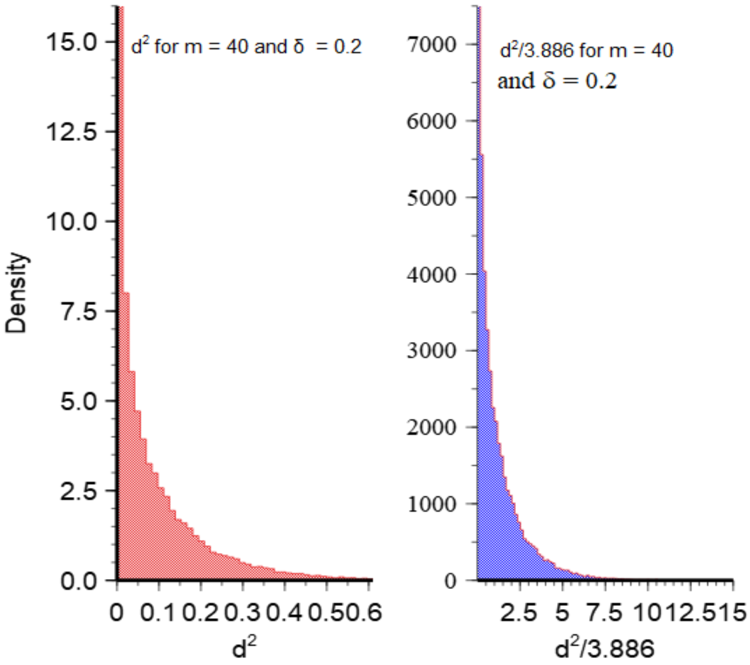


Figure 8.3: Representation of density at 80% percentile transformation.

Tables 8.1 to 8.3 provides an estimate of N_{true} which would have 80% coverage providing $\delta = 1$. In the same way, $k^2(0.6, 5, 1)$ is estimated to be 1.487, and $k^2(0.7, m, 1)$ is estimated to be 2.041 as shown in the tables. Table 8.1 provides Modifier k^2 values for 60% coverage, which when used in modified method will ensure 60% coverage is achieved for example $k^2(0.6, 15, 1)$ is 1.2282 and multiplying same using modified method will ensure a sample size with 60% coverage.

While Table 8.2 provides Modifier k^2 values for 70% coverage, which when used in modified method will ensure 70 % coverage is achieved for example $k^2(0.7, 20, 1)$ is 1.4220 and multiplying same using modified method will ensure a sample size with 70% coverage.

Table 8.1: Modifier k^2 $(0.6, m, \delta)$ for 60% coverage.

m	δ					
	0.1	0.2	0.3	0.4	0.5	1.0
5	33.0976	8.9543	4.5122	2.9532	2.3111	1.5318
10	15.0447	4.4519	2.5117	1.8675	1.6202	1.3135
15	10.2908	3.2261	1.9662	1.5937	1.4359	1.2282
20	7.9663	2.7076	1.7841	1.5006	1.3883	1.2049
25	6.4924	2.2632	1.5790	1.3966	1.3162	1.1671
30	5.5809	2.0397	1.4925	1.3475	1.2774	1.1473
35	5.1016	1.9841	1.5329	1.3848	1.3091	1.1614
40	4.4161	1.8923	1.4831	1.3492	1.2777	1.1445
50	3.6747	1.6641	1.3819	1.2819	1.2286	1.1260
60	3.1299	1.5498	1.3294	1.2417	1.1947	1.1007
70	2.9146	1.5332	1.3302	1.2465	1.2013	1.1056
80	2.6187	1.4918	1.3109	1.2340	11839	1.0969
90	2.4397	1.4152	1.2702	1.1987	1.1595	1.0826
100	2.2029	1.4046	1.2612	1.1928	1.1542	1.0816

Table 8.2: Modifier $k^2(0.7, m, \delta)$ for 70% coverage.

m	δ					
	0.1	0.2	0.3	0.4	0.5	1.0
5	49.8961	13.598	6.8363	4.5429	3.5106	2.0888
10	23.6521	6.8063	3.7682	2.7813	2.3169	1.6515
15	15.8028	4.7842	2.9100	2.2677	1.9697	1.4839
20	12.2835	4.0452	2.5642	2.0718	1.8427	1.4220
25	9.9356	3.4264	2.2912	1.9060	1.7040	1.3609
30	8.3515	3.0522	2.1600	1.8281	1.6460	1.3244
35	7.6212	2.9543	2.1197	1.7979	1.6228	1.3179
40	6.6411	2.6902	1.9915	1.7130	1.5571	1.2851
50	5.6492	2.4505	1.8781	1.6353	1.5011	1.2532
60	4.7377	2.1978	1.7396	1.5340	1.4227	1.2141
70	4.4272	2.1529	1.7109	1.5174	1.4105	1.2113
80	3.9570	2.0676	1.6664	1.4829	1.3820	1.1949
90	3.6989	1.9639	1.6149	1.4480	1.3530	1.1777
100	3.3431	1.8780	1.5599	1.4109	1.3281	1.1676

Table 8.3 provides Modifier k^2 values for 80% coverage, which when used in modified method will ensure 80 % coverage is achieved for example $k^2(0.8, 50, 1)$ is 1.4163 and multiplying same using modified method will ensure a sample size with 80% coverage.

Table 8.3: Modifier k^2 $(0.8, m, \delta)$ for 80% coverage.

m	δ					
	0.1	0.2	0.3	0.4	0.5	1.0
5	79.3654	21.4529	10.7778	7.1006	5.3040	2.9311
10	36.8651	10.6793	5.8641	4.1698	3.3592	2.0980
15	24.3492	7.4068	4.3801	3.2894	2.7473	1.8413
20	18.9876	6.0954	3.7380	2.8765	2.4287	1.6943
25	15.3601	5.1867	3.3192	2.6161	2.2542	1.6093
30	12.9273	4.5719	3.0217	2.4105	2.0901	1.5337
35	11.5061	4.3322	2.9219	2.3475	2.0404	1.5074
40	10.3247	3.9459	2.7353	2.2369	1.9584	1.4655
50	8.5697	3.5587	2.5242	2.0854	1.8452	1.4163
60	7.3339	3.1618	2.3097	1.9401	1.7338	1.3584
70	6.6934	3.0544	2.2476	1.8942	1.7013	1.3445
80	5.9455	2.8248	2.1244	1.8110	1.6360	1.3171
90	5.5567	2.7166	2.0547	1.7575	1.5931	1.2931
100	5.1031	2.5805	1.9823	1.7100	1.5551	1.2797

8.3 Summary

This chapter presented a modified method for sample size estimation to ensure the expected coverage is achieved. Findings were presented with graphics and tables. Table 8.1,8.2 and 8.3 shows that the modifier needed is very much dependent on coverage and the unknown true effect size. However, the correction factor seems large and therefore although the problem can be corrected the extent of the correction means that either extraordinarily large pilot sample sizes would be needed or the projected sample size for the full study would be extraordinarily large.

CHAPTER 9

Conclusions and Recommendations

Estimation of pilot sample sizes and sample sizes for substantive studies was researched. Interest with issues arising from over/under estimation of sample sizes in clinical research have driven this research as both could lead to either waste of limited resources or unethical situation leading to participant undergoing a process that is completely avoidable. Numerous formulae exist for samples sizes estimation; however, this study focuses on the Upper Confidence limit developed by Browne in 1995. The focus of the research is to propose a solution to control error margin in sample size estimation. This will serve as a guide to researchers and yield better research results. This final chapter presents a summary of the research findings, recommendations, and potential areas for further research.

9.1 Findings and their implications

There are numerous suggested formulae for estimating sample sizes in clinical trials; however, they have inherent limitations since some necessary parameter needed for the calculation are not known before the research begins. Researchers, therefore, depend on pilot studies to obtain the parameter value, such as standard deviation, required for the calculation. However, the standard deviation from pilot studies could be inaccurate. Currently, different rules of thumb are used by researchers to determine pilot sample sizes, but they are all known to have their limitations.

Browne developed a formula that suggests using at least 80% one-side confidence limits on standard deviation as the estimate of standard deviation to ensure at least 80% chance of achieving the planned power in the clinical trial. The degree of over/underestimation of sample sizes by the formula was not considered.

This research in study 1a investigates Browne 1995's formula for sample size estimation in study one, using the following parameter combinations significance levels α = (0.01,0.05), power level 1 – β= (0.8,0.9), pilot sample sizes m = (5,10,30,50,100), known standardized effect sizes δ = (0.10,0.40,0.75), coverage levels $(1-\gamma)$ = (0.8,0.9). To improve accuracy of findings,100,000 iteration was used, in contrast to Browne's original research, which employed 2000 iteration. The results show that using Browne's 1995 formula would achieve the planned power but could lead to massive overestimation of sample sizes by over 100percent and underestimation of sample sizes more than minus 20percent. Consequently, the research further investigated the corresponding median percentage error associated to the formula and confirmed its over/underestimation tendencies. This implies that in the case of overestimation, the research team using Browne's formula could spend double the cost needed to achieve a research result. Conversely, in the case of underestimation, the research team could end up with inconclusive results. Both situations are unethical and should be avoided.

Based on the finding of study one(a) and considering its implications, this research developed a Goldilocks ("just about right") model for pilot sample size estimation in study one(b), aiming to control both the lower and upper bound of the error margin. This model will serve as guide for researchers in managing the range of error when choosing sample sizes, enable them to choose pilot sample sizes that correspond to a desired coverage level and error margin for their research, and this will enhance decision making for researchers. This presents an advantage for Browne's formula and the rules of thumb that do not consider the level of over/under estimation of sample sizes.

Study two developed and studied four formulae for sample size estimation when the Minimum Clinical Importance Difference (MCID) is unknown. These formulae are as follows: Naive-Cohen: the naive estimate for the sample size using Cohen's d when σ^2, μ_1, μ_2 are unknown($N_{N,C}$), Browne-Cohen: using Cohen's d in Browne's approach ($N_{B,C}$), Naive Hedge-a naive estimate using Hedge's h ($N_{N,H}$), and Browne Hedge- using Hedges' h in Browne's approach ($N_{B,H}$). The percentage greater than N_{true} and Median percentage error result for each formula were tabulated and the findings showed that the Naive estimates lead to massive underestimation of sample sizes. The box plots of Browne-Hedges showed a better range of variance compared to the other formulae suggesting it is the best of them all for sample size estimation for a substantive trial when MCID is unknown. The formulae can be used with caution, considering their tendency for both overestimation and underestimation, thus highlighting the need for a further study.

Using the upper confidence limit for co-efficient of variation study three develop and investigates three formulae for sample size estimation namely Standard coefficient of variation formula ($N_{C,S}$),McKay formula ($N_{C,M}$),and Vangel formula ($N_{C,V}$). Result shows that the formulae work best for only large effect sizes (0.75) and that is the point where the desired coverage can be achieved . Based on the results the relatively best of the three formula is the standard formula and it was compared to results from Browne's formula . The study showed a level of over/under estimation using it and does not perform better than Browne as the coverage level is not attained at small effect sizes.

The study proceeded to conduct study 4 the final study, where a modifier correction critical table was created to lead the expected coverage when MCID is unknown. This table ensures the desired coverage is attained at each point of sample size estimation, thus addressing the problem of coverage not been attained by some parameter combination. However, it was observed that this approach leads to a massive overestimation of sample sizes which is not practical in real life research.

9.2 Recommendation

The following recommendations are made after careful consideration of the findings from this research. Firstly, researchers should be informed of the limitation of the different methods of sample size estimation and exiting rules of thumb for pilot sample sizes, to ensure they have control over under/over estimation of sample sizes. Secondly when using Browne's 1995 formula for sample size estimation leads to over/under estimation of sample sizes, the proposed Goldilocks ("just about right") model and tables should be employed to control the over/under estimation of sample sizes and improve research outputs.

Also, when MCID is unknown the Browne-Hedge formula is the best of the proposed formulae but equally lead to over/under estimation of sample sizes hence the result in this research can serve as guide to control the range of error.

For sample size estimation formulae developed using upper confidence limit of co-efficient of variation a close comparism of the formulae showed the Standard coefficient of variation formula ($N_{C,S}$) is the best. The result compared to Browne method showed it is efficient as Browne's Formula for large effect sizes only and can be used inter changeably.

Furthermore, the k^2 multiplier table developed in study four can be utilized when there is particular interest is in attaining a certain desired coverage level.

However, it should be noted that this formula will lead to massive overestimation of sample sizes, which is not cost effective but can guarantee accurate result.

Finally, the findings of this research will be valuable in providing proper guidance to researchers in controlling over/under estimation of sample sizes. This will result in resources saving, and ethical standard compliant outcomes.

9.3 Further research

This research was conducted based on assumption of normality and equal variance. Normal distribution and continuous outcome. Further research in this area could explore for non-normal distribution including binary outcomes and variance heterogeneity, do simulations for other two arm pre- post- RCT design with repeated measures ANCOVA as the analysis strategy. A study of co-efficient of variation using other adjustments methods like Generalized Confidence interval by Liu 2012 can be explored. The margin of error models for the method when MCID is unknown and when using upper confidence limit of coefficient of variation can be studied and developed.

REFERENCES

Ahrens, J.H., Dieter, U. and Grube, A. (1970). “Pseudo-random numbers”. *Computing*, 6(1-2), pp.121-138.

Altman, D.G. and Bland, J.M. (1999). “Treatment allocation in controlled trials: why randomise?.” *British Medical Journal*,318(7192), pp.1209-1209.

Arnold, S.F.,1990 *Mathematical statistics*. Prentice- Hall Englewood Cliffs, NJ.

Anderson, D.R., Sweeney, D.J. and Williams, T.A., (2020). Statistics for business and economics.14th edition.

Baguley, T. (2009). “Standardized or simple effect size: What should be reported?”. *British journal of psychology,* 100(3), pp.603-617.

Bailey, R.A. (2008). *Design of comparative experiments*. Cambridge University Press.

Banks, J., Carson, J., Nelson, B., and Nicol, D.(2001). *Discrete-Event System Simulation.* Prentice Hall. p. 3. ISBN 978-0-13-088702-3.

Barker, E., Barker, W., Burr, W., Polk, W., and Smid, M. (2012). "Recommendation for Key Management" (PDF). *NIST Special Publication* 800-57.

Birkett, M.A., and Day, S.J., (1994). “Internal pilot studies for estimating sample size”. *Statistics in medicine*, 13(23 ‐ 24), pp.2455-2463.

Box, G.E.P. and Muller M.E., (1958). “A note on the generation of random normal deviates”. *Ann Math* Stat, 29: 610–611.

Browne, R.H., (1995). “On the use of a pilot sample for sample size determination”. *Statistics in Medicine*, 14(17), pp.1933-1940.

Brown, M.B. and Forsythe, A.B., (1974). “Robust tests for the equality of variances”. *Journal of the American statistical association*, 69(346), pp.364-367.

Burton, A., Altman, D.G., Royston, P. and Holder, R.L., (2006). “The design of simulation studies in medical statistics”. *Statistics in Medicine*, 25(24), pp.4279-4292.

Campbell, M.J., Machin, D. and Walters, S.J., (2010). *Medical statistics: a textbook for the health sciences*. John Wiley & Sons.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* 2nd edition. Belmont , CA : Duxbury Press.

Cassey, A.J. and Smith, B.O., (2014). “Simulating confidence for the Ellison–Glaeser index”. *Journal of Urban Economics,* 81, pp.85-103.

Chalmers, T.C., Smith Jr, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D. and Ambroz, A.,(1981). “A method for assessing the quality of a randomised control trial”. *Controlled clinical trials*, 2(1), pp.31-49.

Chan, A.W., Tetzlaff, J.M., Altman, D.G., Laupacis, A., Gøtzsche, P.C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J.A. and Doré, C.J., (2013). “SPIRIT 2013 statement: defining standard protocol items for clinical trials”. *Annals of internal medicine*, 158(3), pp.200-207.

Charles, P., Giraudeau, B., Dechartres, A., Baron, G. and Ravaud, P., (2009). “Reporting of sample size calculation in randomised controlled trials”. *British Medical Journal*, 338, p.b1732.

Chow, S.C., Shao, J., Wang, H. and Lokhnygina, Y., (2017). *Sample size calculations in clinical research*. CRC press.

Chuang-Stein, C. and Beltangady, M., (2010). “FDA draft guidance on adaptive design clinical trials: Pfizer's perspective”. *Journal of biopharmaceutical statistics*, 20(6), pp.1143-1149.

Cohen, J., (1992). “Quantitative methods in psychology: A power primer”*. Psychol. Bull*., 112, pp.1155- 1159.

Cohen, J., (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cronbach, L. J. , Gleser, G. C., Nanda, H., & Rajaratnam, N., (1972). *The dependability of behavioral measurements: Theory generalizability for scores and profiles*. New York: Wiley.

Conover, W.J. and Conover, W.J., (1980). Practical nonparametric statistics.

Connelly, L.M (2008). “Pilot studies”. *Med. Surg. Nursing*, 17(6), 411-2.

Dalgaard, P., (2008). *Power and the computation of sample size. In Introductory Statistics with R* (pp. 155-162). Springer, New York, NY.

Derrick, B., Broad, A., Toher, D. and White, P., (2017). “The impact of an extreme observation in a paired samples design”. *Metodološki zvezki-Advances in Methodology and Statistics*, 14(2).

Dettori, J., (2010). “The random allocation process: two things you need to know. Evidence-based spine”. *care journal*, 1(03), pp.7-9.

Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P. and Meester, L.E., (2005). A Modern Introduction to Probability and Statistics: Understanding why and how. *Springer Science & Business Media.*

Fleiss, J.L., Levin, B. and Paik, M.C., (2013). Statistical methods for rates and proportions. *John Wiley & Sons; New York*.

Frane, J.W., (1998). “A method of biased coin Randomisation, its implementation, and its validation”. *Drug information journal: DIJ/Drug Information Association*, 32(2), pp.423-432.

Friede, T. and Kieser, M., (2001). “A comparison of methods for adaptive sample size adjustment”. *Statistics in medicine*, 20(24), pp.3861-3873.

Funder, D.C., Levine, J.M., Mackie, D.M., Morf, C.C., Sansone, C., Vazire, S. and West, S.G., (2014). “ Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice”. *Personality and Social Psychology Review*, 18(1), pp.3-12.

Gehr, S., Balasubramaniam, N.K. and Russmann, C., (2023). “Use of mobile diagnostics and digital clinical trials in cardiology”. *Nature Medicine*, 29(4), pp.781-784.

Georgakellos, D.A. and Macris, A.M., (2009). “Application of the semantic learning approach in the feasibility studies preparation training process”. *Information Systems Management,* 26(3), pp.231-240.

Halpern, S.D., Karlawish, J.H. and Berlin, J.A., (2002). “The continuing unethical conduct of underpowered clinical trials”. *Jama*, 288(3), pp.358-362.

Hedge L.V. and Olkin I.(2014). Statistical methods for meta-analysis. Orlando: *Academic Press Inc*; p. 53 -95.

Hertzog, M.A., (2008). “Considerations in determining sample size for pilot studies”. *Research in nursing & health*, 31(2), pp.180-191.

Hinkelmann, K. and Kempthorne, O., (1994). Design and analysis of experiments (Vol. 1). *New York: Wiley*.

Holford, N., Ma, S.C. and Ploeger, B.A., (2010). “Clinical trial simulation: a review”. Clinical Pharmacology & Therapeutics, 88(2), pp.166-182.

Hyndman, R.J. and Koehler, A.B., (2006). “Another look at measures of forecast accuracy.” *International journal of forecasting*, *22*(4), pp.679-688.

Julious, S.A., (2004). “Sample sizes for clinical trials with normal data”. *Statistics in medicine*, 23(12), pp.1921-1986.

Julious, S.A., (2005). “Sample size of 12 per group rule of thumb for a pilot study. Pharmaceutical Statistics”: *The Journal of Applied Statistics in the Pharmaceutical Industry*, 4(4), pp.287-291.

Julious, S.A. and Swank, D.J., (2005). “Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan”. *Pharmaceutical Statistics*, 4(1), pp.37-46.

Julious, S.A. and Owen, R.J., (2006). “Sample size calculations for clinical studies allowing for uncertainty about the variance”. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 5(1), pp.29-37.

Justis, R.Y. and Kreigsmann, B., (1979). “The feasibility study as a tool for venture analysis”. *Journal of Small Business Management* (pre-1986), 17(000001), p.35.

Kelly, P.J., Stallard, N. and Todd, S., (2005). “An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several”. *Journal of biopharmaceutical statistics*, 15(4), pp.641-658.

Kieser, M. and Wassmer, G., (1996). “On the use of the upper confidence limit for the variance from a pilot sample for sample size determination”. *Biometrical journal*, 38(8), pp.941-949.

Kroese, D.P., Brereton, T., Taimre, T. and Botev, Z.I., (2014). “Why the Monte Carlo method is so important today. Wiley Interdisciplinary Reviews”. *Computational Statistics*, 6(6), pp.386-392.

Kim, M.J., (2005). Number of replications required in control chart Monte Carlo simulation studies. *University of Northern Colorado*.

Korn, E.L., Freidlin, B., Abrams, J.S. and Halabi, S., (2012). “Design issues in randomised phase II/III trials”. *Journal of Clinical Oncology*, 30(6), p.667.

Kocak, D., (2019). A method to increase the power of Monte Carlo method: Increasing the number of iteration. *Pedagogical Research*, *5*(1), p.em0049.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W., (2005). *Applied linear statistical models*. McGraw-hill.

Lee, E.C., Whitehead, A.L., Jacques, R.M. and Julious, S.A., (2014). “The statistical interpretation of pilot trials: should significance thresholds be reconsidered?”. *BMC medical research methodology,* 14(1), p.41.

Locock, L. and Smith, L., (2011). “Personal benefit, or benefiting others? Deciding whether to take part in clinical trials”. *Clinical trials*, 8(1), pp.85-93.

Lewis, J.A., (1999). “Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline”*. Statistics in medicine*, 18(15), pp.1903-1942.

Marcoulides, G.A., (1993). “Maximizing power in generalizability studies under budget constraints.” *Journal of Educational Statistics*, 18(2), pp.197-2.

Maigne, L., Hill, D., Calvat, P., Breton, V., Reuillon, R., Lazaro, D., Legre, Y. and Donnarieix, D., (2004). Parallelization of Monte Carlo simulations and submission to a grid environment. *Parallel processing letters*, 14(02), pp.177-196.

Maggio, S. and Sawilowsky, S., (2014). “JMASM 33: A Two Dependent Samples Maximum Test Calculator: Excel”. *Journal of Modern Applied Statistical Methods,* 13(1), p.32.

Malone, H., Nicholl, H. and Tracey, C., (2014). “Awareness and minimisation of systematic bias in research”. *British Journal of Nursing*, *23*(5), pp.279- 282.

McKay A.T. (1932). “Distribution of the coefficient of variation and the extended t distribution”. *Journal of the Royal Statistical Society*, Vol 95, 695-698.

Moore, D.S., and McCabe, G.P., (1989). *Introduction to the practice of statistics*. WH Freeman/Times Books/Henry Holt & Co.

Nayak, B.K., (2010). “Understanding the relevance of sample size calculation”. Indian *journal of ophthalmology*, 58(6), p.469.

Neiswiadomy, R.M. (2002). Foundations of nursing research. Upper saddle River, *NJ:Pearson Education.*

NETSCC. (2012). Glossary: Feasibility and Pilot Studies [Online]. Available: http://www.netscc.ac.uk/glossary/ [Accessed 8th October 2021].

Obodo, S., Toher, D., and White, P. (2021). “Estimation of the two-group pilot sample size with a cautionary note on Browne’s formula”. *Journal of Applied Quantitative methods*. 16(3).

Obodo, S.C., Toher, D., and White, P. (2023). “The Just-about-right Pilot sample size to control error margin”. *International Journal of Statistics and Probability*.12(3), pp1-7.

Parsons, H.M., (1974). “What happened at Hawthorne?: New evidence suggests the Hawthorne effect resulted from operant reinforcement contingencies”. *Science*, 183(4128), pp.922-932.

Prescott, P.A. and Soeken, K.L., (1989). “The potential uses of pilot work”. *Nursing Research*, 38(1), p.60-62.

Ranjith, G. (2005). “Interferon-Î±-Induced Depression: When a Randomised Trial Is Not a Randomised Controlled Trial”. *Psychotherapy and psychosomatics*, 74(6), p.387.

R Core team (2020). *R:A language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL:http://www.R-project.org/.

R Core team (2022). *R:A language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL:http://www.R-project.org/.

R Studio Team (2020). *RStudio: Integrated Development Environment for R.* RStudio, Inc. Boston, MA. URL:http://www.rstudio.com/.

R Studio Team (2022). *RStudio: Integrated Development Environment for R.* RStudio, Inc. Boston, MA. URL:http://www.rstudio.com/.

Sim, J. and Lewis, M.(2012). “The size of a pilot study for a clinical trial should calculated in relation to considerations of precision and efficiency”. *Journal of clinical epidemiolog*,65(3), pp.301-308.

Serlin, R.C., (2000). “Testing for robustness in Monte Carlo studies”. *Psychological methods*, 5(2), p.230.

Scott, D.W., (2011). Box–muller transformation. Wiley Interdisciplinary Reviews: *Computational Statistics*, 3(2), pp.177-179.

Schulz, K.F. and Grimes, D.A., (2002). “Generation of allocation sequences in randomised trials: chance, not choice”. *The Lancet*, 359(9305), pp.515-519.

Sokolowski, J.A. and Banks, C.M. (2009). Principles of Modeling and Simulation. *John Wiley & Son. p.* 6. ISBN 978-0-470-28943-3.

Sim, J. and Lewis, M.(2012). “The size of a pilot study for a clinical trial should calculated in relation to considerations of precision and efficiency”. Journal of clinical epidemiology 65(3), pp.301-308.

Simera, I. and Altman, D.G., (2009). Writing a research article that is “fit for purpose”: EQUATOR Network and reporting guidelines. *BMJ Evidence-Based Medicine*, *14*(5), pp.132-134.

Simera, I., Moher, D., Hoey, J., Schulz, K.F. and Altman, D.G. (2010). “A catalogue of reporting guidelines for health research”. *European journal of clinical investigation*, 40(1), pp.35-53.

Shamseer, L., Hopewell, S., Altman, D.G., Moher, D. and Schulz, K.F., (2016). “Update on the endorsement of CONSORT by high impact factor journals: a survey of journal “Instructions to Authors”.*Trials*,17, pp.1-8.

Suresh, K.P., (2011). “An overview of randomisation techniques: an unbiased assessment of outcome in clinical research”. *Journal of Human Reproductive Sciences*, 4(1), p.8 - 11.

Suresh, K.P. and Chandrashekara, S. (2012). “Sample size estimation and power analysis for clinical research studies”. *Journal of human reproductive sciences*, 5(1), p.7 - 13.

Swinscow, T.D.V. and Campbell, M.J. (2002). “Statistics at square one” . London: *British medical journal.* pp. 111-25

Teare, M.D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A. and Walters, S.J. (2014). “Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study”. *Trials*, 15(1), p.264.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., Thabane, M., Giangregorio, L. and Goldsmith, C.H. (2010). “A tutorial on pilot studies: the what, why and how”. *BMC medical research methodology*, 10(1), p.1.

Torgerson, D., (2008). “Designing randomised trials in health, education and the social sciences: an introduction”. *Springer*.

Van Belle, G. and Martin, D.C., (1993). “Sample size as a function of coefficient of variation and ratio of means”. *The American Statistician*, 47(3), pp.165-167.

Vangel, M.G., (1996). “Confidence intervals for a normal coefficient of variation”. *The American Statistician*, 50(1), pp.21-26.

Vickers, A., 2019. “An evaluation of survival curve extrapolation techniques using long-term observational cancer data”. *Medical Decision Making*, 39(8), pp.926-938.

Von Neumann, J. and Ualam, S (1951). “Monte Carlo method”. *National Bureau of Standards Applied Series* 12,p.35.

Whitehead, A., 2016. “Sample Size Justifications for Pilot Trials of Publicly Funded Randomised Controlled Trials”. *(Doctoral dissertation, University of Sheffield).*

Whitehead, A.L., Julious, S.A., Cooper, C.L. and Campbell, M.J. (2016). “Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable". *Statistical methods in medical research*, 25(3), pp.1057-1073.

Winer, B.J., Brown, D. and Michels, K.1971. Statistical principles in experimental design. *New York McGraw-Hill*, pp.169-172.

Wynants, L., Bouwmeester, W., Moons, K.G.M., Moerbeek, M., Timmerman, D., Van Huffel, S., Van Calster, B. and Vergouwe, Y., (2015). "A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data". *Journal of clinical epidemiology*, *68*(12), pp.1406-1414.

Zalene, M., (1990). "Randomised consent designs for clinical trials: An update". *Stat Med*, 9(6), pp.645-656.

Zar, J.H. (1984). Biostatistical Analysis. Second Edition. Prentice-Hall. Englewood Cliffs, *New Jersay.*

Appendix A:Output

A1

Obodo, S., Toher, D. and White, P 2021., “Estimation of the two-group pilot sample size with a cautionary note on Browne’s formula”.
Journal of Applied Quantitative methods.16(3).
Submitted version

Estimation of the two-group pilot sample size with a cautionary note on Browne's formula

Scholastica OBODO
PhD Candidate
University of the West of England, Bristol
E-mail: Scholastica.Obodo@uwe.ac.uk

Deirdre TOHER
PhD, Senior Lecturer
University of the West of England, Bristol
E-mail: Deirdre.Toher@uwe.ac.uk

Paul WHITE
PhD, Associate Professor
University of the West of England, Bristol
E-mail: Paul.White@uwe.ac.uk

ABSTRACT

Using data obtained from a pilot study, Browne (1995) proposed a procedure for estimating the sample size needed for a definitive two-arm randomised controlled trial when the minimal important difference is specified. Simulations confirm these findings. The results attributable to Browne are extended to consider the degree of error attached to sample size estimation using this procedure. A consideration of the error provides a simple mechanism to estimate the sample size needed for a pilot study so as to control the degree of error in the follow-on substantive trial.

Key words: sample size, estimation, pilot study

INTRODUCTION

A pilot study is often the first step in a research protocol and is typically a smaller – scale study that aids in the planning and modification of the main study (In 2017). The pilot study is often done to aid in the development of a future substantive or definitive trial i.e. one with at least 80% power or at least 90% power (Thabane et.al. 2010). A vexing question is how big the pilot sample size needs to be to help accurately determine the sample size for a follow-on substantive trial.

Sample information for a pilot study, such as the sample variance, may be used to help inform a power calculation for a follow-on trial. However, pilot studies invariably use small sample sizes and the main weakness when estimating key parameters from small sample sizes is the large sampling variation (Teare et. al., 2014). Using data from a pilot study to calculate a future sample size may result in any planned definitive trial being severely underpowered (sample size too small) or overpowered (sample size too large). If the sample size for a planned definitive trial is too small then there is great chance of inconclusive results (Machin et al., 2011) and it is unethical to put too few people through a trial in these circumstances (Halpern et. al. 2002). Too large a sample size would lead to resources been wasted, more patients than necessary could be given a treatment which will later be proven to be inferior; or an effective treatment may mean too many in the control arm being denied the effective treatment, or an effective treatment may be delayed from being released on to the market (Julious, 2009). Too large a sample size or too small a sample size may therefore be considered unethical.

Hertzog (2008) laments that there is little published guidance for how large a pilot sample size should be. Rules of thumb for helping in pilot size estimation are available. For instance, in two arm studies, Nieswiadomy (2002) recommends obtaining approximately 10 participants per arm, Birkett and Day (1994) suggest 20, Browne indicates that 30 is commonplace. Kieser and Wassmer (1996) suggest 20 to 40 to be used when main trial requires between 80 to 250, Teare et al, (2014) suggest ≥70, Sim and Lewis (2011) suggest ≥ 55 use for small to medium effect sizes. Others suggest the pilot sample size should be 10% of the unknown final study size (Connelly, 2008). Whitehead et al (2016) considered stepped rules of thumb with pilot sample sizes being a function of anticipated effect size. However, it is recognised that in practice, the final decision on a pilot sample size will be guided by cost and time constraints as well as by size and variability of the population.

A common design is the two-group parallel design with 1:1 randomisation and with an assumed normally distributed outcome variable. In this case, if the two groups are to be compared using the two-sample t-distribution then the sample size per arm (*n*) needed to have (1 – β) *100% power at the alpha (α) significance level is given by

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

where $Z_{1-\alpha/2}$ is the normal deviate for the α significance level, $Z_{1-\beta}$ is the normal deviate for power (1 – β)μ_i and σ_i^2 are the means and variances of distribution *i*, (*i* = 1, 2) respectively. For equal variances the formula becomes

$$n = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2\sigma^2}{(\mu_1 - \mu_2)^2}$$

(see Van Belle and Martin, 1993).

In practice the population means and variances will not be known, although the effect size $\mu_1 - \mu_2$ may be hypothesised, often based on the minimally important difference (MID), and the pooled variance may be estimated from sample pilot data. Browne (1995) considered this formula when (i) the pooled sample variance is used to estimate σ^2 (unadjusted analyses) and (ii) when the 100$(1 - \gamma)$ per cent upper one-sided confidence limit is used as an estimate of σ^2 (adjusted analyses). Using simulation, the parameters alpha = 0.05, beta = 0.1, 0.2, sample sizes per arm *m* = 5, 10, 30, 50,100 and upper one-sided confidence limits with
γ = 0.1, 0.2, 0.4, 0.5 and standardized effect size = 0.2, 0.3, 0.75 were considered. For a target power of 0.8, Browne found that the resulting predicted sample sizes in the unadjusted formula frequently resulted in power lower than wanted but with increases in the percentage of times target power was achieved with increasing pilot sample size and increasing standardised effect size. In adjusted analyses, for a given coverage γ, the required sample size (*n*) to produce at least 80% power was achieved 100* $(1-\gamma)$ % of the time. However, the magnitude and degree of excess between the predicted sample size $\hat{n}$ and the true sample size n was not considered by Browne, nor was a mechanism on how to determine an appropriate sample size for the pilot study. This article aims to determine how accurate Browne's approach is in sample size determination over a range of significance levels (α = 0.01, 0.05), power levels (1 – β= 0.8, 0.9), pilot sample sizes (*m* = 5, 10, 30, 50, 100), standardised effect sizes (δ = 0.10, 0.40, 0.75), and coverage levels ($1 - \gamma$ = 0.8, 0.9) and to determine pilot sample size (*m* per arm) to control median percentage error between predicted sample size and the true minimum unknown sample size need to satisfy power requirements.

SIMULATION DESIGN

We consider the two-sample parallel RCT design with 1:1 randomisation with a normally distributed outcome variable. In common with Browne (1995) we consider when the specified mean difference $\mu_1 - \mu_2$ is equal to 0.10, 0.40, and 0.75 with a common unit standard deviation so that standardised effect size, Cohen's δ is 0.10, 0.40, and 0.75, corresponding to what is often considered “small”, “medium” and “large” effect sizes. Likewise, in common with Browne, we consider when the planned power for the follow-on study is 1 – β = 0.8,0.9, and when pilot sample size is *m* = 5, 10, 30, 50, 100 per arm. Browne considered when the nominal significance level to be used in the follow-on trial is α = 0.05; we additionally extend the simulation to include α = 0.01. Hence the design corresponds to a 2 by 2 by 2 by 3 by 5 fully crossed design (see Table 1). 100,000 replicates will be conducted at each cell combination (contrast with Browne who undertook 2000 replicates per cell).

Table 1: Parameter combinations

Factor	Number of Levels	Levels
Significance level (α)	2	0.01, 0.05
Power (1 – β)	2	0.8, 0.9
Coverage level ($1 - \gamma$)	2	0.8, 0.9
Standardised Effect size (δ)	3	0.10, 0.40, 0.75
Pilot sample size per arm (*m*)	5	5, 10, 30, 50, 100

The predicted sample size for the follow-on study, under Browne's adjust approach is given by

$$\hat{n} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 k s^2}{(\mu_1 - \mu_2)^2}$$

where s^2 is the pooled sample variance based on m_1+ $m_2 - 2$ = $2m - 2$ degrees of freedom. and *k* = $(2m-2)/\chi^2_{,}(2m-2, 1-\gamma)$ (i.e., $k\ s^2$ is the pilot sample derived one-sided upper

$100 \times (1-\gamma)$ % confidence limit for σ^2 assuming normality). The output from the simulation will record the percentage of time the predicted sample size $\hat{n}$ will exceed n and how frequently $\hat{n}$ will exceed $n + pn$ where *p* = -0.2, 0.0, 0.2, 0.3, 0.5, 1.0, 1.5. The median percentage error at each cell combination of the design will be recorded.

RESULTS

Table 2 considers the parameter settings $\alpha = 0.05$, $\beta = 0.2$, $1-\gamma = 0.8$, and records the percentage of times the predicted per arm sample size for the follow-on study ($\hat{n}$) exceeds the required sample size (n) for each level of pilot sample size and standardised effect size. As required, and as given previously by Browne (1995), 80% of the time the estimated sample size $\hat{n}$ is equal to, or larger than n. As the pilot sample size (*m*) increases the degree of excess decreases and this is true for every effect size. With a very small sample sizes of $m = 5$ per group, there is in excess of a 50% chance of the estimated sample size being in excess of 50%, there is in excess of a 30% chance of the estimated sample size being in excess of 100%, and there is in excess of a 15% chance of the estimated sample size being in excess of 150%, and this is true for all effect sizes (see Table 2)

Table 2: Percentage of time estimated sample size exceeds $n \pm pn$ (*p* = -0.2, +0.2, +0.3, +0.5, +1.0, +1.5) for α= 0.05, β=0.2, $1-\gamma$ = 0.8

Pilot Sample size (m)	Effect size	>-20%	>n	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.884	.799	.701	.650	.547	.325	.175
10	.10	.922	.800	.634	.544	.377	.107	.021
30	.10	.973	.799	.452	.289	.086	.001	.000
50	.10	.989	.797	.334	.159	.019	.000	.000
100	.10	.998	.801	.164	.035	.000	.000	.000
5	.40	.883	.799	.698	.646	.544	.325	.175
10	.40	.920	.800	.629	.539	.367	.105	.021
30	.40	.972	.801	.450	.284	.079	.001	.000
50	.40	.988	.801	.332	.154	.018	.000	.000

100	.40	.998	.801	.159	.032	.000	.000	.000
5	.75	.876	.801	.694	.638	.547	.325	.174
10	.75	.910	.801	.617	.520	.370	.106	.020
30	.75	.963	.798	.423	.255	.085	.001	.000
50	.75	.982	.799	.327	.126	.018	.000	.000
100	.75	.996	.799	.130	.023	.000	.000	.000

Equally, from Table 2, with a small sample size of $m = 10$ per group there is more than a 30% chance of the sample being in excess of 50%, there is more than a 10% chance of the estimated sample size being in excess of 100%, there is more than a 35% chance of the estimated sample size being in excess of 50%, and there is more than a 50% chance of the estimated sample size being in excess of 30%, and this is true for all effect sizes.

With a moderate sample size of $m = 30$ per group, there is more than a 25% chance of the estimated sample size being in excess of 50%.

With a moderately large sample size i.e., $m = 50$ per group, there is more than a 30% chance of the estimated sample size being in excess of 20%, and more than a 10% chance of the estimated sample size being in excess of 30%, and this is true for all effect sizes.

Even with large sample pilot samples sizes of $m = 100$ per group there is in excess of 10% chance of the sample being overestimated by 20%.

Note that with sample sizes of m = 30 or smaller, there is a non-trivial chance of underestimating by more than 20% of n.

Table 2 summarises the data when α = 0.05, $1 - \gamma$ = 0.8 and β = 0.20 and Table 3 uses α = 0.05, $1 - \gamma$ = 0.8 and β = 0.10. Note that the results in both tables are almost identical, hence there is very little effect caused by changing β. The corresponding tables with α = 0.01 are equally near identical as to when α = 0.05 indicating significance level does not affect the degree of excess of the algorithm and for parsimony of exposition these additional tables have been suppressed in this note.

Table 3: Percentage of time estimated sample size exceeds $n + pn$ (p = *-0.2, +0.2, +0.3, +0.5, +1.0, +1.5),* α*= 0.05,* β*=0.1, 1 – γ = 0.8*

Pilot Sample size (m)	Effect size	>-20%	> n	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.884	.800	.701	.650	.549	.326	.175
10	.10	.922	.799	.631	.541	.372	.105	.021
30	.10	.974	.801	.454	.290	.085	.001	.000
50	.10	.989	.800	.338	.159	.019	.000	.000
100	.10	.998	.800	.166	.035	.000	.000	.000
5	.40	.883	.800	.701	.651	.545	.327	.175
10	.40	.918	.798	.627	.539	.367	.106	.021
30	.40	.971	.800	.452	.291	.083	.001	.000
50	.40	.987	.801	.329	.155	.017	.000	.000
100	.40	.998	.801	.158	.034	.000	.000	.000
5	.75	.886	.803	.702	.646	.538	.326	.174
10	.75	.921	.800	.628	.531	.354	.105	.020
30	.75	.972	.799	.445	.273	.072	.001	.000
50	.75	.988	.804	.331	.143	.014	.000	.000
100	.75	.998	.801	.155	.029	.000	.000	.000

If attention is restricted to any degree of excess then the same qualitative patterns are observed. For instance, Figure 1 is a plot of the percentage of times the predicted per arm sample size $\hat{n}$ is double the required sample size n. The panels of Figure 1, defined by the factorial combination of α and β are near identical at each level of effect size indicating the small effect both α and β have on the degree of excess, and show that the degree of excess is (a) greater when 1 - γ increases from 0.8 to 0.9 and (b) diminishes with increasing pilot sample size (*m*).

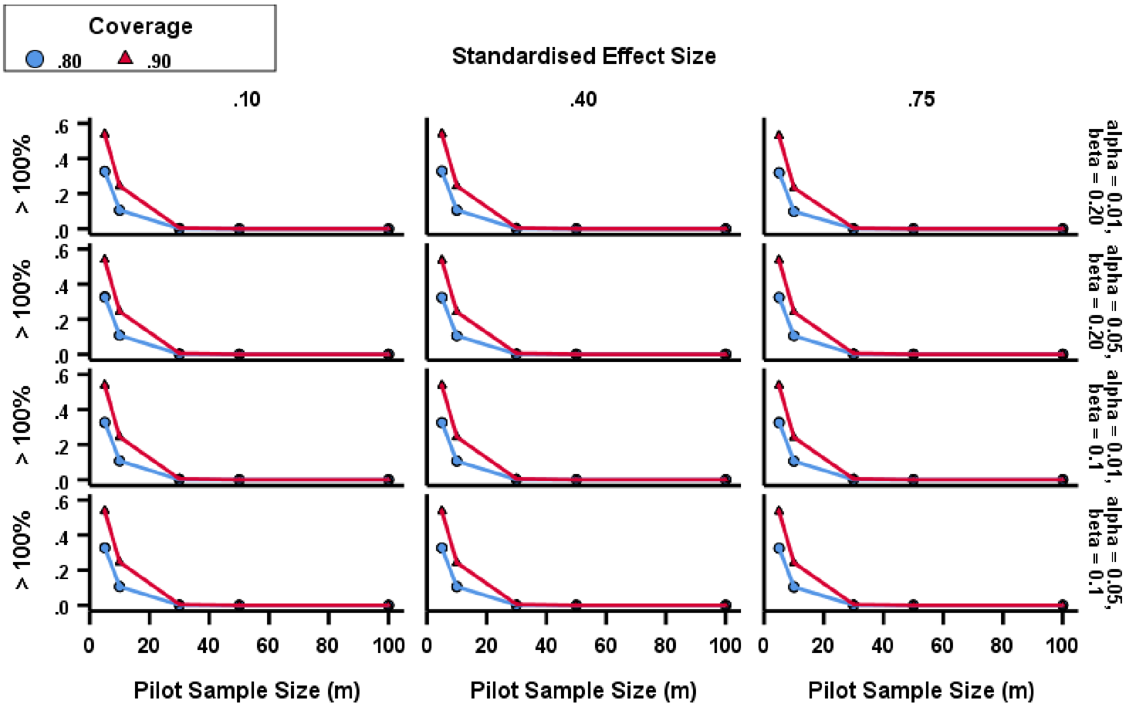


Figure 1 Graphical representation for $\hat{n} > 2n$ for standardised effect size δ = 0.10, 0.40, 0.75; coverage $1 - \gamma$ = 0.8, 0.9; significance level α = 0.01, 0.05; β = 0.1, 0.2 and per arm pilot sample size *m* = 5, 10, 30, 50, 100

Equally, Figure 2, (using $1 - \gamma$ = 0.8) graphically depicts that the degree of excess is not dependent on effect size. In summary the degree of excess is dependent on (a) pilot sample size m and level of coverage $1 - \gamma$.

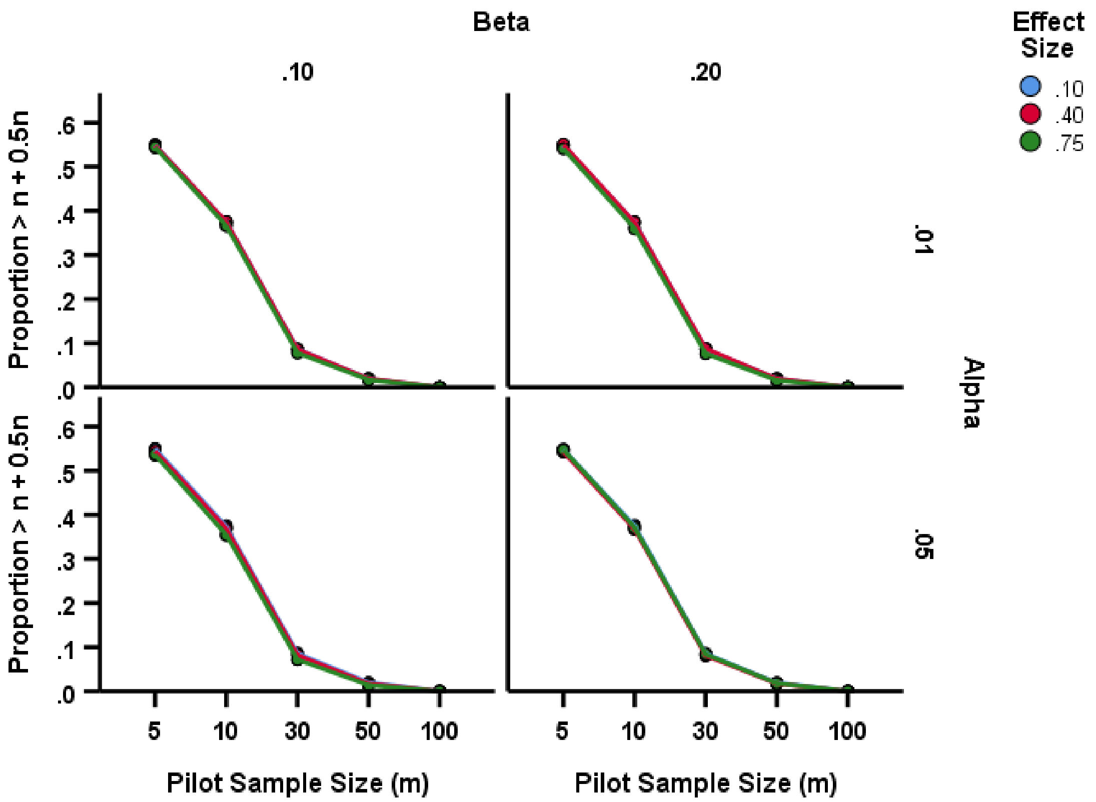


Figure 2 Graphical representation for $\hat{n} > 1.5n$ for coverage $1 - \gamma$ = 0.8 *and for* standardised effect size δ = 0.10,0.40,0.75*;* α = 0.01, 0.05*;* β = 0.1, 0.2 and *per arm* pilot sample size *m* = 5, 10, 30, 50, 100

Table 4 summarizes median percentage error at each cell combination. Specifically, median percentage error is not dependent on alpha, nor dependent on beta, and the effect of standardised effect size appears negligible. Equally, the median percentage error decreases with increasing sample size and increases with increasing coverage.

Table 4 *Median percentage error at different alpha, beta, coverage, effect size, and sample sizes.*

				Pilot Sample size (m)				
Coverage	Effect size	Alpha	Beta	5	10	30	50	100
.80	.10	.01	.10	60.49	34.96	17.39	13.04	8.96
			.20	59.60	34.97	17.32	13.03	8.92
		.05	.10	60.25	34.85	17.40	12.98	8.92
			.20	59.32	34.42	17.53	13.10	8.95
	.40	.01	.10	59.88	34.77	17.39	13.09	8.97
			.20	60.59	34.55	17.45	13.19	8.99
		.05	.10	59.24	34.34	16.68	12.46	8.49
			.20	58.80	33.71	16.55	12.03	7.89
	.75	.01	.10	60.49	34.21	17.25	12.97	8.75
			.20	58.05	33.41	16.08	11.81	7.77
		.05	.10	57.33	32.68	15.46	11.08	7.16
			.20	59.43	34.70	17.03	12.68	8.61
.90	.10	.01	.10	110.21	59.93	28.32	20.82	14.04
			.20	110.79	59.81	28.27	20.92	14.08
		.05	.10	109.87	59.13	28.40	20.81	14.01
			.20	110.29	59.82	28.23	20.81	14.08
	.40	.01	.10	110.76	59.34	28.40	20.82	14.07
			.20	109.98	59.82	28.35	20.90	14.08
		.05	.10	108.99	58.47	27.57	20.22	13.49
			.20	108.77	58.05	27.13	19.82	13.05
	.75	.01	.10	109.62	59.19	27.92	20.53	13.91
			.20	107.98	57.75	26.65	19.50	12.84
		.05	.10	106.64	56.81	26.08	18.80	12.16
			.20	109.98	59.29	27.98	20.52	13.65

A plot of the median percentage error at each cell combination against the square root of pilot sample size for coverage $1 - \gamma$ = 0.8 is given in Figure 3. The regression of the inverse of the median percentage error (MPE) against the square root of pilot sample size per arm (m) for the simulation data is given by

$$ {}^{1}/_{MPE} = -0.01208 + 0.01298\sqrt{m} $$

(R^2 = 0.985) and which on re-arrangement gives

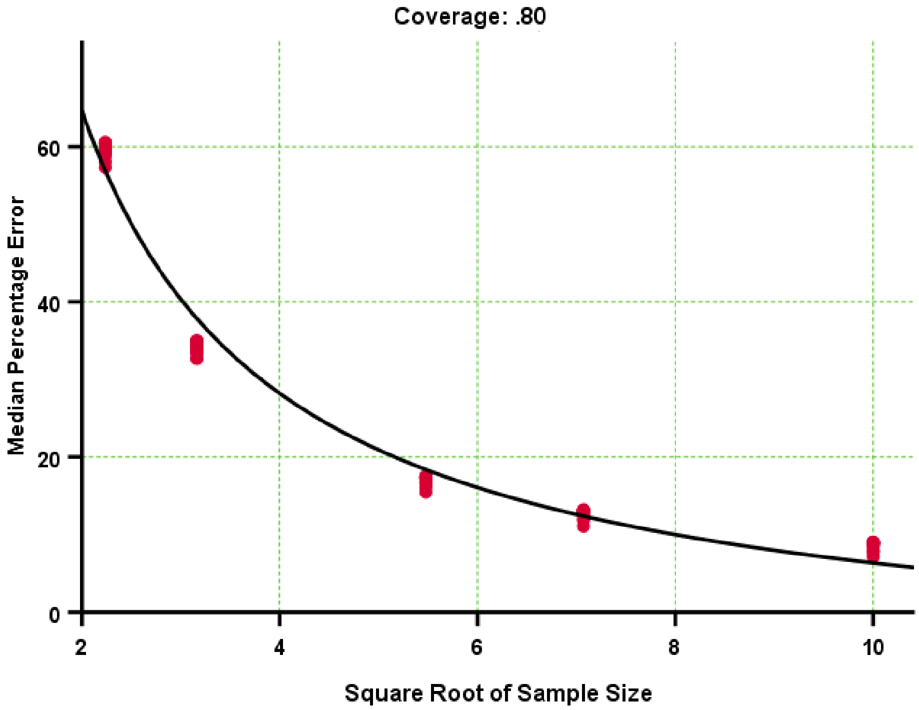
$$ m = \left(0.930663 + \frac{1}{0.01298 \times MPE}\right)^2 $$

Repeating the process coverage γ = 0.9 gives the regression equation

$$ {}^{1}/_{MPE} = -0.009339 + 0.00828\sqrt{m} $$

(R^2 = 0.994) which on re-arrangement gives

$$ m = \left(1.1280 + \frac{1}{0.00828 \times MPE}\right)^2 $$



.

Figure 3: Graphical representation of median percentage errors against the square root of the pilot sample size ($\sqrt{m}$)

Hence, via regression the relationship between median percentage error and pilot sample size may be quantified. Table 5 quantifies the pilot sample size to maintain the median percentage error at a required level for either $1 - \gamma$ = 0.8 and for $1 - \gamma$ = 0.9

Table 5: Median percentage error and pilot sample size per arm at 80% and 90% coverage

Median Percentage Error	Pilot sample size per arm at 80% coverage	Pilot sample size per arm at 90% coverage
4	408	980
5	267	639
6	190	452
7	143	337
8	112	262
9	90	211
10	75	174
12	54	125
14	42	95
16	33	75
18	28	62
20	23	52
22	20	44
24	18	38

Thus, for instance, if a researcher opts for 80% coverage and wishes to control median percentage error to be no more 10% of true sample size, then a pilot sample size of *m* = 75 per arm would be needed, but for 90% coverage the minimum sample size to have a median percentage error of 10% would be 174 per arm.

Table 6 shows the mean median percentage error (MPE) for a given sample size for 80% and 90% coverage; for instance, if *m* = 25 per arm is used, then with 80% coverage the mean MPE would be 19% but would be 31% if 90% coverage is used.

Table 6 Mean Median percentage error for a given pilot sample size (m) per arm

M	80% Coverage	90% Coverage
5	59	109
10	35	60
15	26	44
20	22	36
25	19	31
30	17	28
35	16	25
40	14	23
45	14	22
50	13	20

Conclusions

The simulations support the earlier work of Browne i.e. for situations when the effect size $\mu_1 - \mu_2$ is known (hypothesised or MID) then the 100*$(1 - \gamma)$ % upper confidence interval for the variance may be used to determine sample size for a larger study and have 100*$(1 - \gamma)$ % coverage (0.8 or 0.9) However, the simulations also show that the degree of excess from this approach may result in intolerably large predicted sample sizes leading to a study being vastly overpowered. This degree of excess is very much pronounced when pilot sample sizes are small (e.g., *m* = 10 per arm) and the degree of excess is seen to decrease with increasing sample size.

The degree of excess may be characterised by the median percentage error (MPE). For a given level of coverage, the MPE is inversely proportional to the square root of pilot sample size (*m*) with a high degree of accuracy. The regression model permits a pilot sample size (*m*) to be determined for any degree of median percentage error. For instance, suppose a research team is planning a pilot study to help determine the sample size for a follow-on substantive study. They may elect to have 80% coverage of the sample size; however if they want to ensure that median percentage error is no more than, say 16%, then a pilot sample size of m = 33 or large would be needed for each arm of the pilot study. Table 5 and Table 6 and their corresponding regressions equations therefore have value in helping a research team and funders to decide on how big a pilot sample size should be.

References

Browne, R.H., (1995). On the use of a pilot sample for sample size determination. *Statistics in medicine*, *14*(17),.1933-1940.

Birkett, M.A. and Day, S.J., 1994. Internal pilot studies for estimating sample size. *Statistics in medicine*, *13*(23-24), pp.2455-2463.

Connelly, L.M., 2008. Pilot studies. *Medsurg Nursing*, *17*(6), 411 - 412

Halpern, S.D., Karlawish, J.H. and Berlin, J.A., (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3),.358-362.

Hertzog, M.A. (2008). Considerations in determining sample size for pilot studies. *Research in nursing & health*, *31*(2),.180-191.

In J (2017) Introduction of a pilot study, *Korean journal of anesthesiology*, 70(6), 601 – 605.

Julious, S.A. (2009) *Sample sizes for clinical trials*. CRC Press.

Kieser, M. and Wassmer, G., (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical journal*, *38*(8), 941-949.

Machin, D., Campbell, M.J., Tan, S.B. and Tan, S.H. (2011). *Sample size tables for clinical studies*. John Wiley & Sons.

Nieswiadomy, R.M. (2002). Foundations of nursing research (4th ed.). Upper Saddle River, NJ: Pearson Education.

Sim, J. and Lewis, M., (2012). The size of a pilot study for a clinical trial should be calculated in relation to considerations of precision and efficiency. *Journal of clinical epidemiology*, *65*(3), 301-308.

Teare, M.D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A. and Walters, S.J., (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, *15*(1), 1-13.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., Thabane, M., Giangregorio, L. and Goldsmith, C.H., (2010). A tutorial on pilot studies: the what, why and how. *BMC medical research methodology*, *10*(1),.1.

Van Belle, G. and Martin, D.C., (1993). Sample size as a function of coefficient of variation and ratio of means. *The American Statistician*, *47*(3), 165-167.

Whitehead, A.L., Julious, S.A., Cooper, C.L. and Campbell, M.J., (2016). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. *Statistical methods in medical research*, *25*(3), 1057-1073.

A2

Obodo, S.C., Toher, D. and White, P 2023., “The Just-about-right Pilot sample size to control error margin”. International Journal of Statistics and Probability.12(3). pp. 1-7.

Published version

International Journal of Statistics and Probability; Vol. 12, No. 3; May 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Just-About-Right Pilot Sample Size to Control the Error Margin

Scholastica C. Obodo[1], Deirdre Toher[1], Paul White[1]

[1] Department of Data Science and Mathematics, University of the West of England, Bristol, United Kingdom

Correspondence: Paul White, Department of Data Science and Mathematics, University of the West of England, Bristol, United Kingdom

Received: February 6, 2023 Accepted: April 20, 2023 Online Published: May 23, 2023
doi:10.5539/ijsp.v12n3p1 URL: https://doi.org/10.5539/ijsp.v12n3p1

Abstract

In practice, the required sample size for a two-arm randomised controlled trial cannot always be determined pre-study with great accuracy. This lack of accuracy has economic, ethical and scientific implications. The sample size for a pilot study is an important consideration in helping the decision making for the sample size of a follow-on trial. Consideration of under- and over-estimation of the sample size results in the idea of a Just-About-Right (JAR) sample size. For studies involving a minimally clinical important difference (MCID) we present the pilot sample sizes to meet investigator desired JAR considerations.

Keywords: just-about-right, overestimation, pilot study, power, sample size, underestimation

1. Introduction

1.1 Incorrect Estimation of Sample Sizes

A pilot study is a small-scale investigation designed to test the feasibility of methods and procedures for later use on a larger scale (Thabane et al, 2010). In clinical studies, a pilot randomised controlled trial (RCT) could be used to help in the planning of a proposed substantive RCT (power = 0.8) or definitive RCT (power >= 0.9). The pilot RCT provides a means to collect preliminary data on safety, is used to assess the recruitment rate and the degree of participant retention, provides data on willingness to be randomised, and crucially, to provide estimates of variation in outcomes measures to assist the decision-making process for the sample size of the follow-on trial (Lancaster et al, 2004, Ln 2005, Arnold et al, 2009). This latter consideration begs the question, on how to determine the optimal RCT pilot sample size for any given context, with the aim of being able to estimate accurately the sample size requirements of the proposed follow-on RCT.

One of the most common errors in any type of empirical scientific research is an insufficient sample size (Makin, 2019). Small sample sizes can lead to Type Two errors (false negatives) and in practice this is especially true when combined with moderately low or low effect sizes. Small sample sizes can leave a research community in some doubt as to whether effects are real. There is also the position that it is unethical to ask participants to commit to taking part in a study which is insufficiently powered to meet objectives (Altman 1980, Halpern et. al. 2002). In addition, any such study would be an uneconomic use of resources. On the contrary, having too large a sample size could also be problematic. A sample size might be considered too large if the same quality of conclusions could have been obtained with a much smaller sample size. If the sample size is too large, then this too may be considered an uneconomic use of resources and it may be deemed unethical to be randomly allocating any excess sample size to control or intervention irrespective of whether intervention confers a benefit or not. In summary, for any substantive or definitive trial, the sample size should be sufficient to achieve worthwhile results, but not so large as to involve unnecessary recruitment of participants. Guidance is needed to allow research teams, ethics committees, funding panels, data monitoring committees, and protocol reviewers to evaluate whether a study intends to recruit too many participants (overpowered) or too few participants (underpowered) and it is important to get a just-about-right (JAR) sample size which is not too small, not too large but just-about-right.

A well conducted pilot study could be instrumental in helping to determine a JAR sample size for the follow-on study. Extant literature provides some rules-of-thumb for pilot sample size estimation. Julious (2005) noted that the marginal additional sample information content decreases with each unit increase in sample size and recommended a sample size of at least n = 12 per group. Similarly, Birkett and Day (1994) suggest 20 per arm, Kieser and Wassmer (1996) suggest between 20 to 40 per arm be used when the main trial requires between 80 to 250, Teare et al, (2014) suggest ≥70, and Browne (1995) indicates that n = 30 per arm is commonplace practice. However, the prevailing sentiment is that a simple one-size-fits-all solution or one rule-of-thumb would be inadequate when context specific considerations apply.

In terms of context specific considerations, Browne (1995), considered determining sample size for a two-arm parallel RCT study when (a) a pilot study is used to collect preliminary data on outcome variation and (b) the minimum clinically important difference (MCID) is pre-specified and (c) the follow-on study is to be adequately powered to detect an effect and (d) an assumption of normally distributed outcome data can be made. For these situations, the required per arm sample size, n, is given by

$$n = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$$

(

1) where $\mu_1 - \mu_2$ is the true mean difference or MCID, $Z_{1-\alpha/2}$, and $Z_{1-\beta}$ are standardised normal deviates for two-sided significance testing with nominal significance level α and required power $1-\beta$, and σ^2 is the population variance for the outcome measure assumed to be equal between arms. Although the MCID might be specified by hypothesis, the true population variance σ^2 would be unknown. The pilot study would provide a sample estimate for σ^2, but this sample estimate s^2, would most likely underestimate the population variance σ^2, since $(m_1 + m_2 - 2)\, s^2/\sigma^2 \backsim \chi^2_{m_1+m_2-2}$ where m_1 and m_2 denotes the sample sizes in the two arms of the pilot study. It is well known that chi-square distributions are positively skewed, hence using s^2 in place of σ^2 in the above formula would typically produce an estimated sample size lower than truly required. For this reason, Browne (1995) cautiously suggested estimating and replacing σ^2 in the sample size formula with the estimated $100(1-\gamma)$ per cent one-sided upper confidence limit (UCL) for σ^2. Specifically, the sample size per arm, for 1:1 randomisation under Browne's suggested approach is given by

$$\hat{n} = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2\, ks^2}{(\mu_1 - \mu_2)_2} \qquad (2)$$

where s^2 is the sample pooled variance and ks^2 is the $100(1-\gamma)$ percent one-sided upper confidence limit (UCL) for σ^2. The quantity $100(1-\gamma)$ is the "coverage" i.e., the percentage of times that the predicted sample size per arm, $\hat{n}$, would exceed the true required sample size per arm n. From a practical perspective, Browne advocated a coverage of 80% (0.8) or a coverage of 90% (0.9).

1.2 Browne's Method

Simulation work conducted by Browne (1995) and Obodo et al, (2021) confirms that the approach considered by Browne has merit, achieving the required coverage of 0.8 or 0.9 as appropriate, for $\alpha = 0.01$, $\alpha = 0.05$, $\beta = 0.2$, $\beta = 0.1$, and for a range of effect sizes (small, medium, large) and for a range of pilot sample sizes between 5 and 100. However, Obodo et al, (2021) show that the procedure can produce underpowered studies, or frequently produce an intolerably large degree of excess, and that the extent of the problem depends on pilot sample size per arm (m), level of coverage $(1-\gamma)$ but not on significance level $\alpha = 0.01, 0.05$, nor on power $1-\beta = 0.8, 0.9$, nor on MCID. Both coverage and pilot sample size are at the control of an investigator at the trial planning stage.

We therefore sought to quantify the relationship between pilot sample size and JAR requirements for coverage of 0.8 and 0.9 separately.

We operationalise an investigator chosen JAR interval to be $[n - \lambda_1 n, n + \lambda_2 n\ ]$ where λ_1, $\lambda_2 \in [0, 1]$, are investigator chosen parameters to prevent the degree of underpowering (λ_1) and degree of overpowering (λ_2). We aim for trialists to be able to justify pilot sample size and to make a statement to the effect of "The proposed two group pilot study will have a sample size of m per arm. This sample size is chosen so that the resultant power calculations for a larger study will have $100(1 - \gamma)\%$ chance of exceeding the minimum required sample size and which in a two-sided test with significance level α will have $100(1 - \beta)\%$ power for detecting a difference between arms assuming a MCID of $(\mu_1 - \mu_2)$. This proposed pilot sample size of m per arm will ensure that the estimated sample size will lie in the interval $(1 - \lambda_1)n$ to $(1 + \lambda_2)n$, with probability π providing a safeguard for under- and over- powering." In this statement we consider $\alpha = 0.01, 0.05$, power $(1 - \beta) = 0.8, 0.9$, coverage $(1 - \gamma) = 0.8, 0.9$, any value for MCID, lower bounds $\lambda_1 = 0.1, 0.2$ and upper bounds $\lambda_2 = 0.1, 0.2, 0.3, 0.4$ for any chosen level of π.

Monte Carlo Simulation Design

The Monte Carlo simulations are informed by Browne (1995) and mimic the design given by Obodo et al, (2021). In brief, we consider the two-arm parallel RCT with 1:1 randomisation which is to be analysed using the independent samples t-test (equal variances assumed, two-sided, alpha = 0.05, 0.01). The true sample size for the RCT is calculated for desired power (0.8 or 0.9), for a specified MCID corresponding to a small, medium or large effect (0.1, 0.4, 0.75) assuming equal variances $(\sigma^2 = 1)$ under an assumption of normality.

For pilot samples sizes (m = 5, 10, 30, 50, 100) the 80% and 90% upper one-sided confidence limit for the pooled sample variance is used in Browne's formula. The percentage of times that the estimated sample size, $\hat{n}$ is in the interval $[n - \lambda_1 n, n + \lambda_2 n\ ]$ is recorded for λ_1 = 0.1, 0.2, and λ_2 = 0.1, 0.2, 0.3, 0.4, 0.5. Table 1 summarises the factor levels for the 2 by 2 by 2 by 3 by 5 fully crossed design.

Table 1. Parameter combinations

FACTOR	Number of Levels	LEVELS
Power	2	0.8, 0.9
Significance level	2	0.01, 0.05
Coverage level	2	0.8, 0.9
Effect size	3	0.10, 0.40, 0.75
Pilot sample size	5	5, 10, 30, 50, 100

Simulation was done using the R programming language with 100,000 replicates (as against Browne 1995 who used 2,000 replicates) for each cell of the design to obtain more precise simulation values.

Results

Table 2 summarises the percentage of times the estimated sample size, $\hat{n}$, for the follow-on study would be in the interval $[n - \lambda_1 n, n + \lambda_2 n\ ]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3 for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9. Simulation percentages are aggregated over significance level α = 0.01, 0.05, over prior reasoned statistical power

$(1 - \beta)$ = 0.8, 0.9 and assumed effect size $\mu_1 - \mu_2$ = 0.1, 0.4, 0.75 as it is known that these factors do not affect the estimated sample size (Obodo et al, 2021).

Inspection of Table 2 and Figure 1, clearly shows the percentage within any given interval monotonically increases with increasing pilot sample size for each of coverage = 0.8 and for coverage = 0.9. It is also clear that the percentage in any given interval is greater for coverage = 0.8 compared with coverage = 0.9 and this is only to be expected since, for any estimated sample size, the sample size for when coverage is 0.9 must be greater than the sample size when a tolerance for coverage is set to be equal to 0.8. Table 2 and Figure 1 show that the percentage of instances within an interval is particularly sensitive to the upper bound λ_2 which naturally follows from the positively skewed chi-square distribution used in the estimation process.

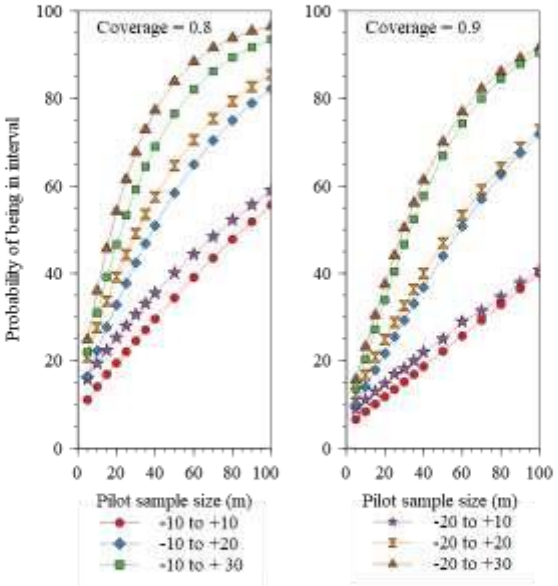


Figure 1. Percentage of simulation instances $100\hat{\pi}$ in the interval $[n - \lambda_1 n, n + \lambda_2 n\,]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3, for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9

Table 2. Percentage of simulation instances $100\hat{\pi}$ in the interval $[n - \lambda_1 n, n + \lambda_2 n\,]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3, for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9

	Coverage = 0.8					
	λ_1 0.1	λ_1 0.1	λ_1 0.1		λ_1 0.2	λ_1 0.2
m	λ_2 0.1	λ_2 0.2	λ_2 0.3		λ_2 0.2	λ_2 0.3
5	11.1	16.5	22.0		20.9	24.9
10	14.1	22.3	31.0		27.6	36.0
15	16.9	27.7	39.2		33.7	45.7
20	19.5	32.8	46.6		39.2	54.2
25	22.1	37.7	53.3		44.3	61.5
30	24.6	42.4	59.2		49.1	67.7
35	27.1	46.8	64.4		53.5	72.9
40	29.6	50.9	69.0		57.5	77.3
50	34.4	58.4	76.5		64.6	83.9
60	39.1	64.9	82.1		70.5	88.4
70	43.5	70.4	86.2		75.4	91.6
80	47.8	75.0	89.4		79.4	93.8
90	51.8	78.9	91.7	55.7	82.7	95.4
100	55.6	82.2	93.5	59.0	85.5	96.5

	Coverage = 0.9					
m	λ_1 0.1	λ_1 0.1	λ_1 0.1		λ_1 0.2	λ_1 0.2
	λ_2 0.1	λ_2 0.2	λ_2 0.3		λ_2 0.2	λ_2 0.3
5	6.6	10.1	13.6		12.5	15.7
10	8.4	14.1	20.4		16.8	23.2
15	10.1	17.9	27.2		20.9	30.4
20	11.8	21.7	33.9		24.9	37.5
26	13.5	25.5	40.4		28.8	44.1
30	15.2	29.3	46.6		32.6	50.4
35	16.9	33.1	52.4		36.4	56.1
40	18.7	36.8	57.7		40.0	61.3
50	22.2	44.0	66.9		47.0	70.0
60	25.7	50.8	74.3		53.3	76.9
70	29.3	57.0	80.0		59.1	82.2
80	32.9	62.6	84.5		64.2	86.1
90	36.5	67.5	87.9		68.8	89.2
100	40.0	71.9	90.5		72.8	91.6

The monotonic trends between $\hat{\pi}$ and pilot per arm sample size and each level of coverage has been modelled using linear regression with the functional form . Thus, for instance, when coverage = 0.8 and the interval $n \pm 0.1n$ is considered then it is readily verified that $\ln(\hat{\pi}) = -2.745 + 0.297\sqrt{m}$ and that the overall goodness-of-fit, $100R^2$, is 96.3%. Table 3 provides the estimated intercepts, gradients and goodness of fit for λ_1= 0.1, 0.2; $\lambda_2 = 0.1, 0.2, 0.3, 0.4$ 0.5 for coverage 0.8 and coverage 0.9.

Table 3. Regression equations of the form $\ln(\pi) = b_0 + b_1\sqrt{m}$ giving estimated intercept (b_0), gradient (b_1), coefficient of determination (R-squared) for λ_1 = 0.1, 0.2, and λ_2 = 0.1, 0.2, 0.3, 0.4, 0.5

Coverage = 0.8				
Lower Percentage ($100\lambda_1$)	Upper Percentage ($100\ \lambda_2$)	Intercept	Gradient	R- Squared
10	10	-2.745	.297	.963
10	20	-2.531	.406	.988
10	30	-2.399	.506	.993
10	40	-2.094	.543	.981
10	50	-1.697	.527	.952
20	10	-2.256	.262	.954
20	20	-2.228	.400	.989
20	30	-2.375	.569	.997
20	40	-2.613	.759	.997
20	50	-2.557	.853	.998

Coverage = 0.9				
Lower Percentage ($100\ \lambda_1$)	Upper Percentage ($100\ \lambda_2$)	Intercept	Gradient	R- Squared
10	10	-3.306	.290	.957
10	20	-3.082	.402	.984
10	30	-3.029	.528	.995
10	40	-3.028	.656	.998
10	50	-2.827	.712	.991
20	10	-2.872	.250	.955
20	20	-2.795	.378	.986
20	30	-2.856	.524	.995
20	40	-3.108	.716	.996
20	50	-3.450	.919	.993

For any level of coverage and any interval, any regression equation in Table 3 may be re-written in terms of pilot sample size i.e., $m = ([\ln(\hat{\pi}) - b_0]/\ b_1)$^2. Solution of this will give an estimated pilot sample size per arm, m, for any required percentage for the given interval.

Table 4 shows the pilot sample size per arm (m) needed to have a required probability (π) of being in a given interval $[n - \lambda_1 n, n + \lambda_2 n\ ]$ for coverage of 0.8 or coverage 0.9. Thus, for instance, if an investigator requires an 80% chance of not being underpowered for a definitive trial (coverage = 0.8) and requires a 70% chance $(\pi = 0.7)$ of being within $\pm$ 10% of the true required sample size $(\lambda_1 = 0.1, \lambda_2 = 0.1)$ then a sample size per arm (m) of 65 is needed for any given MCID.

Table 4. Pilot sample size (m) required for a required proportion (π) to be in the interval $[n - \lambda_1 n, n + \lambda_2 n]$ for a given coverage.

	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
π	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
			Coverage = 0.8			
0.50	48	20	11	36	15	9
0.55	52	23	13	40	17	10
0.60	56	25	14	44	18	11
0.65	61	27	15	49	20	12
0.70	65	29	16	53	21	13
0.75	68	31	17	56	23	13
0.80	72	32	18	60	25	14
0.90	79	36	21	67	28	16
			Coverage = 0.9			
0.50	81	35	20	76	31	17
0.55	87	38	21	83	34	18
0.60	93	41	23	89	37	20
0.65	98	43	24	95	39	22
0.70	103	46	26	101	42	22
0.75	108	48	27	107	43	24
0.80	113	50	28	112	46	25
0.90	121	54	31	122	51	27

Discussion and Conclusion

Pilot studies are conducted for a variety of reasons. One such reason is to help determine variation in outcome measures to help plan the required sample size for a large-scale substantive or definitive follow-on study. The preceding sections consider the situation where the MCID can be pre-specified for a scale outcome variable and an assumption of normality is reasonable.

Sample size may be calculated if parameters are either known or can be reasonably estimated. For instance, in a two-arm study, if for example MCID = 0.2, variance = 1, alpha = 0.05, beta = 0.10, then the required sample size may be verified to n = 526 per arm (complete data set after any missing data). In practice the variation of the outcome measure may not be known but may be estimated by collecting pilot data. In these regards, Browne's method, may be used to estimate a required sample size with either 80% or 90% coverage i.e. the estimated sample size has an 80% or 90% chance of exceeding the required minimum sample size. A problem with this approach is the chance of underestimating the required sample size, or in having an estimated sample size which far exceeds the required sample size (see Obodo et al, 2021). We considered a strategy to curb these excesses so that estimated sample sizes would be "not too small" and "not too large" in comparison to the true required but unknown sample size, by considering a just-about-right (JAR) sample size. The chosen coverage (say 80% or 90%) is not dependent on pilot sample size. However, with a given level of coverage a researcher may wish to ensure that the probability of the margin of error attached to any estimate is pre-specified to be within an interval around the true required sample size e.g. 70% chance of being within 10% of the required sample size. By inspection of Table 4, if 80% coverage is required with a 70% chance of being within +/- 10% of true sample size, then the pilot study would require at least m = 65 per arm. The protocol may then contain a summary "The proposed two group pilot study will have a sample size of 65 per arm. This sample size is chosen so that the resultant power calculations for sample size in a larger study will have an 80% chance of exceeding the minimum required sample size and which in a two-sided test with significance level α will have $100(1-\beta)\%$ power for detecting an effect assuming an MCID of $(\mu_1 - \mu_2)$.

This proposed pilot sample size of 65 per arm will ensure that the estimated sample size will have a 70% chance of being in an interval of +/- 10% of the true required sample size providing a safeguard over under- and over- powering."

The pilot sample sizes given in this article (Table 4) is predicated on an MCID. If the true effect size exceeds the MCID then the follow-on study is likely to be overpowered to detect a difference (the lesser of the two possible errors). If the true effect is smaller than the MCID then any effect smaller than the MCID is not of clinical interest and may go undetected.

The pilot sample sizes given in this article are based on assumptions of normality and equal variance. In these regards, the practical utility of the pilot sample size recommendations needs further investigation for variance heterogeneity and non-normal distributions including binary outcomes. In a similar way, other simulations may consider the two arm pre- post- RCT design with repeated measures ANCOVA as the analysis strategy. As such the given pilot sample sizes are restricted to the stated assumptions with a direct parametric comparison between the two groups.

References

Arnold, D. M., Burns, K. E., Adhikari, N. K., Kho, M. E., Meade, M. O., & Cook, D. J. (2009). The design and interpretation of pilot trials in clinical research in critical care. *Critical care medicine*, *37*(1), S69-S74. https://doi.org/10.1097/CCM.0b013e3181920e33

Altman, D. G. (1980). Statistics and ethics in medical research: III How large a sample? *British medical journal*, *281*(6251), 1336. https://doi.org/10.1136/bmj.281.6251.1336

Birkett, M. A., & Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in medicine*, *13*(23-24), 2455-2463. https://doi.org/10.1002/sim.4780132309

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in medicine*, *14*(17), 1933-1940.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3), 358-362. https://doi.org/10.1001/jama.288.3.358

In, J. (2017). Introduction of a pilot study. *Korean journal of anesthesiology*, *70*(6), 601-605.
https://doi.org/10.4097/kjae.2017.70.6.601

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, *4*(4), 287-291. https://doi.org/10.1002/pst.185

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical journal*, *38*(8), 941-949. https://doi.org/10.1002/bimj.4710380806

Lancaster, G. A., Dodd, S., & Williamson, P. R. (2004). Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice*, *10*(2), 307-312. https://doi.org/10.1111/j..2002.384.doc.x

Makin, T.R. and Orban de Xivry, J.J., 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *Elife*, *8*, p.e48175. https://doi.org/10.7554/eLife.48175

Obodo, S., Toher, D., & White, P. (2021). Estimation of the two-group pilot sample size with a cautionary note on Browne's formula. *Journal of Applied Quantitative Methods*, *16*(3).

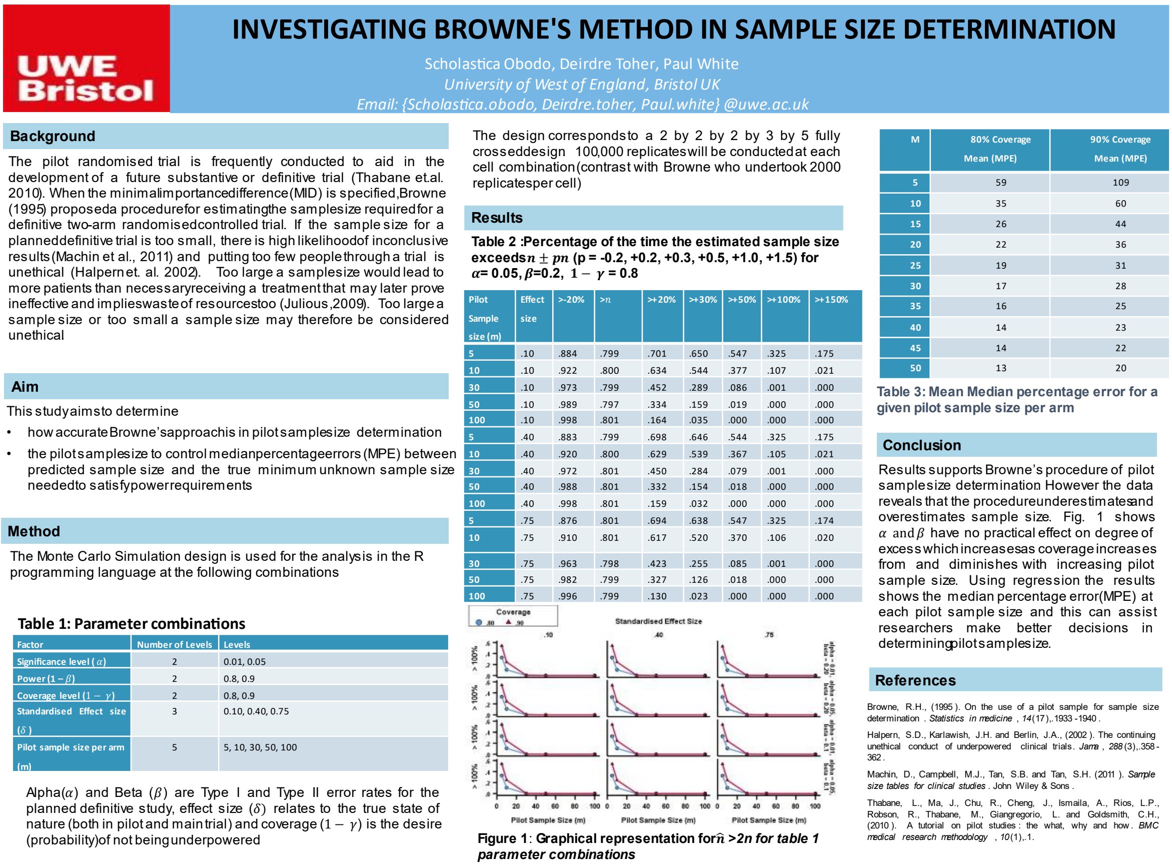
Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, *15*, 1-13. https://doi.org/10.1186/1745-6215-15-264

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., ... & Goldsmith, C. H. (2010). A tutorial on pilot studies:
the what, why and how. *BMC medical research methodology*, *10*, 1-10. https://doi.org/10.1186/1471-2288-10-1

Copyrights

A3

Poster 1: Investigating Browne's method in sample size determination presented at University of Bristol. Statistics at Bristol: Future Results and You 2021 16th & 17th September- **AWARD WINNING POSTER.**



INVESTIGATING BROWNE'S METHOD IN SAMPLE SIZE DETERMINATION

UWE Bristol
Scholastica Obodo, Deirdre Toher, Paul White
University of West of England, Bristol UK
Email: {Scholastica.obodo, Deirdre.toher, Paul.white} @uwe.ac.uk

Background

The pilot randomised trial is frequently conducted to aid in the development of a future substantive or definitive trial (Thabane et.al. 2010). When the minimalimportancedifference(MID) is specified,Browne (1995) proposeda procedurefor estimatingthe samplesize requiredfor a definitive two-arm randomisedcontrolled trial. If the sample size for a planneddefinitive trial is too small, there is high likelihoodof inconclusive results(Machin et al., 2011) and putting too few peoplethrough a trial is unethical (Halpernet. al. 2002). Too large a samplesize would lead to more patients than necessaryreceiving a treatmentthat may later prove ineffective and implieswasteof resourcestoo (Julious,2009). Too large a sample size or too small a sample size may therefore be considered unethical

Aim

This studyaimsto determine

- howaccurateBrowne'sapproachis in pilotsamplesize determination
- the pilotsamplesize to controlmedianpercentageerrors(MPE) between predicted sample size and the true minimum unknown sample size neededto satisfypowerrequirements

Method

The Monte Carlo Simulation design is used for the analysis in the R programming language at the following combinations

Table 1: Parameter combinations

Factor	Number of Levels	Levels
Significance level (α)	2	0.01, 0.05
Power $(1-\beta)$	2	0.8, 0.9
Coverage level $(1-\gamma)$	2	0.8, 0.9
Standardised Effect size (δ)	3	0.10, 0.40, 0.75
Pilot sample size per arm (m)	5	5, 10, 30, 50, 100

Alpha(α) and Beta (β) are Type I and Type II error rates for the planned definitive study, effect size (δ) relates to the true state of nature (both in pilot and main trial) and coverage $(1-\gamma)$ is the desire (probability)of not beingunderpowered

The design correspondsto a 2 by 2 by 2 by 3 by 5 fully crosseddesign 100,000 replicateswill be conductedat each cell combination(contrast with Browne who undertook 2000 replicatesper cell)

Results

Table 2 :Percentage of the time the estimated sample size exceeds $n \pm pn$ (p = -0.2, +0.2, +0.3, +0.5, +1.0, +1.5) for α= 0.05, β=0.2, $1-\gamma$ = 0.8

Pilot Sample size (m)	Effect size	>-20%	>n	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.884	.799	.701	.650	.547	.325	.175
10	.10	.922	.800	.634	.544	.377	.107	.021
30	.10	.973	.799	.452	.289	.086	.001	.000
50	.10	.989	.797	.334	.159	.019	.000	.000
100	.10	.998	.801	.164	.035	.000	.000	.000
5	.40	.883	.799	.698	.646	.544	.325	.175
10	.40	.920	.800	.629	.539	.367	.105	.021
30	.40	.972	.801	.450	.284	.079	.001	.000
50	.40	.988	.801	.332	.154	.018	.000	.000
100	.40	.998	.801	.159	.032	.000	.000	.000
5	.75	.876	.801	.694	.638	.547	.325	.174
10	.75	.910	.801	.617	.520	.370	.106	.020
30	.75	.963	.798	.423	.255	.085	.001	.000
50	.75	.982	.799	.327	.126	.018	.000	.000
100	.75	.996	.799	.130	.023	.000	.000	.000

Figure 1: **Graphical representation for $\hat{n}$ *>2n for table 1 parameter combinations***

M	80% Coverage Mean (MPE)	90% Coverage Mean (MPE)
5	59	109
10	35	60
15	26	44
20	22	36
25	19	31
30	17	28
35	16	25
40	14	23
45	14	22
50	13	20

Table 3: Mean Median percentage error for a given pilot sample size per arm

Conclusion

Results supports Browne's procedure of pilot samplesize determination However the data reveals that the procedureunderestimatesand overestimates sample size. Fig. 1 shows α and β have no practical effect on degree of excesswhich increasesas coverage increases from and diminishes with increasing pilot sample size. Using regression the results shows the median percentage error(MPE) at each pilot sample size and this can assist researchers make better decisions in determiningpilotsamplesize.

References

Browne, R.H., (1995). On the use of a pilot sample for sample size determination . *Statistics in medicine* , *14*(17),.1933 -1940 .

Halpern, S.D., Karlawish, J.H. and Berlin, J.A., (2002). The continuing unethical conduct of underpowered clinical trials. *Jama* , *288*(3),.358 -362 .

Machin, D., Campbell, M.J., Tan, S.B. and Tan, S.H. (2011). *Sample size tables for clinical studies* . John Wiley & Sons .

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L.P., Robson, R., Thabane, M., Giangregorio, L. and Goldsmith, C.H., (2010). A tutorial on pilot studies : the what, why and how . *BMC medical research methodology* , *10*(1),.1.

Award winning Poster feedback

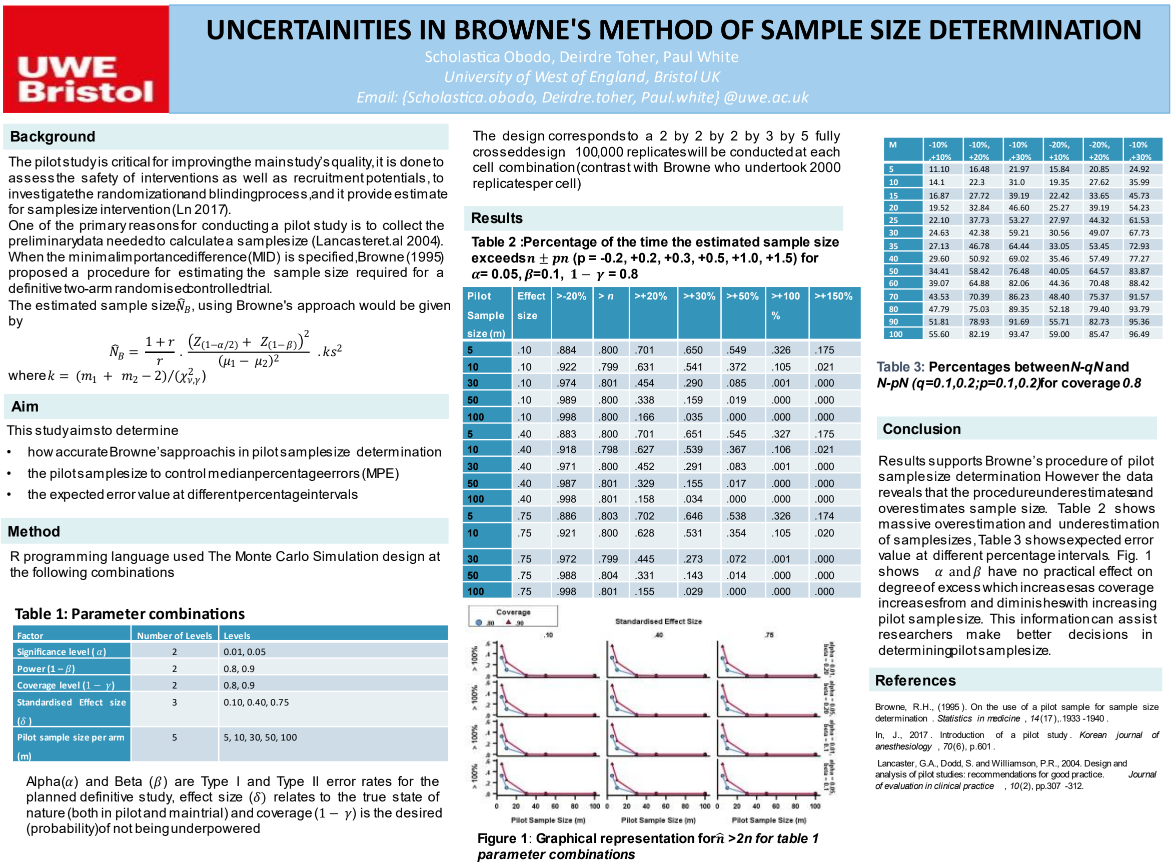
Dear Scholastica Obodo,
On behalf of the organising committee I would like to thank you for presenting your poster at the recent FRY conference in Bristol. We very much enjoyed talking to you about it, and would like to make the award of a Bronze Medal:
Do feel free to put this on your CV and thank you again for participating.

All the best,

Professor Oliver Johnson
Oliver Johnson, Professor of Information Theory
Director of Institute for Statistical Science and MSc in Maths of Cybersecurity
<https://www.bristolmathsresearch.org/statistical-science/>
School of Maths, Fry Building, Univ. of Bristol, Woodland Road, Bristol, BS8 1UG

A4

Poster 2: South West Doctoral Training Partnership (SWDTP) Conference. presentation – “Uncertainties in Browne’s method of sample size determination.” https://www.swdtp.ac.uk/thats-a-wrap-2021-student-conference-review



UWE Bristol

UNCERTAINITIES IN BROWNE'S METHOD OF SAMPLE SIZE DETERMINATION

Scholastica Obodo, Deirdre Toher, Paul White
University of West of England, Bristol UK
Email: {Scholastica.obodo, Deirdre.toher, Paul.white} @uwe.ac.uk

Background

The pilot studyis critical for improvingthe mainstudy’squality,it is doneto assessthe safety of interventions as well as recruitment potentials, to investigatethe randomizationand blindingprocess,and it provide estimate for samplesize intervention(Ln 2017).

One of the primary reasons for conducting a pilot study is to collect the preliminarydata neededto calculatea samplesize (Lancasteret.al 2004). When the minimalimportancedifference(MID) is specified,Browne (1995) proposed a procedure for estimating the sample size required for a definitive two-arm randomisedcontrolledtrial.

The estimated sample size$\widehat{N}_B$, using Browne's approach would be given by

$$\widehat{N}_B = \frac{1+r}{r} \cdot \frac{\left(Z_{(1-\alpha/2)} + Z_{(1-\beta)}\right)^2}{(\mu_1 - \mu_2)^2} \cdot ks^2$$

where$k = (m_1 + m_2 - 2)/(\chi^2_{v,\gamma})$

Aim

This studyaimsto determine

- how accurate Browne’sapproachis in pilot samplesize determination
- the pilotsamplesize to control medianpercentageerrors (MPE)
- the expected error value at differentpercentageintervals

Method

R programming language used The Monte Carlo Simulation design at the following combinations

Table 1: Parameter combinations

Factor	Number of Levels	Levels
Significance level (α)	2	0.01, 0.05
Power ($1-\beta$)	2	0.8, 0.9
Coverage level ($1-\gamma$)	2	0.8, 0.9
Standardised Effect size (δ)	3	0.10, 0.40, 0.75
Pilot sample size per arm (m)	5	5, 10, 30, 50, 100

Alpha(α) and Beta (β) are Type I and Type II error rates for the planned definitive study, effect size (δ) relates to the true state of nature (both in pilotand maintrial) and coverage ($1-\gamma$) is the desired (probability)of not beingunderpowered

The design correspondsto a 2 by 2 by 2 by 3 by 5 fully crosseddesign 100,000 replicateswill be conductedat each cell combination(contrast with Browne who undertook 2000 replicatesper cell)

Results

Table 2 :Percentage of the time the estimated sample size exceeds$n \pm pn$ (p = -0.2, +0.2, +0.3, +0.5, +1.0, +1.5) for α= 0.05, β=0.1, $1-\gamma$ = 0.8

Pilot Sample size (m)	Effect size	>-20%	> *n*	>+20%	>+30%	>+50%	>+100 %	>+150%
5	.10	.884	.800	.701	.650	.549	.326	.175
10	.10	.922	.799	.631	.541	.372	.105	.021
30	.10	.974	.801	.454	.290	.085	.001	.000
50	.10	.989	.800	.338	.159	.019	.000	.000
100	.10	.998	.800	.166	.035	.000	.000	.000
5	.40	.883	.800	.701	.651	.545	.327	.175
10	.40	.918	.798	.627	.539	.367	.106	.021
30	.40	.971	.800	.452	.291	.083	.001	.000
50	.40	.987	.801	.329	.155	.017	.000	.000
100	.40	.998	.801	.158	.034	.000	.000	.000
5	.75	.886	.803	.702	.646	.538	.326	.174
10	.75	.921	.800	.628	.531	.354	.105	.020
30	.75	.972	.799	.445	.273	.072	.001	.000
50	.75	.988	.804	.331	.143	.014	.000	.000
100	.75	.998	.801	.155	.029	.000	.000	.000

Figure 1: Graphical representation for$\hat{n}$ >2*n* *for table 1 parameter combinations*

M	-10% ,+10%	-10%, +20%	-10% ,+30%	-20%, +10%	-20%, +20%	-10% ,+30%
5	11.10	16.48	21.97	15.84	20.85	24.92
10	14.1	22.3	31.0	19.35	27.62	35.99
15	16.87	27.72	39.19	22.42	33.65	45.73
20	19.52	32.84	46.60	25.27	39.19	54.23
25	22.10	37.73	53.27	27.97	44.32	61.53
30	24.63	42.38	59.21	30.56	49.07	67.73
35	27.13	46.78	64.44	33.05	53.45	72.93
40	29.60	50.92	69.02	35.46	57.49	77.27
50	34.41	58.42	76.48	40.05	64.57	83.87
60	39.07	64.88	82.06	44.36	70.48	88.42
70	43.53	70.39	86.23	48.40	75.37	91.57
80	47.79	75.03	89.35	52.18	79.40	93.79
90	51.81	78.93	91.69	55.71	82.73	95.36
100	55.60	82.19	93.47	59.00	85.47	96.49

Table 3: Percentages between*N-qN* and *N-pN (q=0.1,0.2;p=0.1,0.2)*for coverage *0.8*

Conclusion

Results supports Browne’s procedure of pilot samplesize determination However the data reveals that the procedureunderestimatesand overestimates sample size. Table 2 shows massive overestimation and underestimation of samplesizes, Table 3 showsexpected error value at different percentage intervals. Fig. 1 shows α and β have no practical effect on degreeof excess which increasesas coverage increasesfrom and diminisheswith increasing pilot samplesize. This information can assist researchers make better decisions in determiningpilotsamplesize.

References

Browne, R.H., (1995). On the use of a pilot sample for sample size determination . *Statistics in medicine* , *14*(17),.1933 -1940 .

In, J., 2017 . Introduction of a pilot study . *Korean journal of anesthesiology* , *70*(6), p.601 .

Lancaster, G.A., Dodd, S. and Williamson, P.R., 2004. Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice* , *10*(2), pp.307 -312.

A5

Three minutes thesis feedback

UWE Bristol
Three Minute Thesis 2023
Feedback to participants

The panel was impressed by your evident enthusiasm for your research topic, and your presentation style was crystal clear and highly engaging.

We were also impressed by the clarity with which you articulated your research question. You showed in-depth, constructive engagement with a research topic which very much needs to be done.

We really liked how you anchored your research in a real-life example – this is a great way of making your research relatable, especially for non-specialist audiences. When talking about your research to non-specialists (especially non-scientists!) do also remember to explain any technical terms.

Overall, a very clear and well delivered presentation in which passion and energy were to the fore.

Thank you very much for giving us the opportunity to hear your presentation, and we hope this feedback helps you feel encouraged and equipped to talk about your research to a wide range of different audiences.

With best wishes,

The Judging Panel
29 March 2023

Appendix B: R codes for study 1 to 3

Study 1

```
set. seed(462)                         #set value for replication
k21<- 1   r=1                          #where r is ratio of allocation
Niter<-100000                           #for 100000 iterations
Effect sizes<-c (.1,.4,.75)
Coverage<-c (.8,.9)
ALPHA<-c (.05,.01)
BETA<-c (.1,.2)
SampleSizes<-c (5,10,30,50,100)
For (a in 1: length (ALPHA)) {
alpha<-ALPHA[a]
k22<- qnorm(1-(alpha/2))
for (b in 1: length (BETA)) {
beta<-BETA[b]
k23<-qnorm(1-beta)
for (l1 in 1: length (Coverage)) {
 coverage<-Coverage[l1]
for (l in 1: length (SampleSizes)) {
 m<-SampleSizes[l]
  k25<- (2*m-2)/qchisq (1-coverage, df=(2*m-2))
  for (j in 1: length (Effect sizes)) {
  effect size<-Effect sizes[j]
  Ntrue<- 2*(k22+k23) *(k22+k23)/(effect size*effect size)
                                                               # rep (NA, Niter)
                                                              creates a vector of Niter
                                                              (e.g. 100000) NAs
                                                              ready to take values
                                                              later
SamplePooledVariance<-rep (NA, Niter)
Nestimated<-rep (NA, Niter)
GreaterThanNtrue<-rep (NA, Niter)
GreaterThanM20percent<- rep (NA, Niter)
```

```
GreaterThanM10percent<- rep (NA, Niter)
GreaterThanP10percent<- rep (NA, Niter)
GreaterThanP20percent<- rep (NA, Niter)
GreaterThanP30percent<-rep (NA, Niter)
GreaterThanP40percent<-rep (NA, Niter)
GreaterThanP50percent<-rep (NA, Niter)
GreaterThanP75percent<-rep (NA, Niter)
GreaterThanP100percent<-rep (NA, Niter)
GreaterThanP125percent<-rep (NA, Niter)
GreaterThanP150percent<-rep (NA, Niter)
GreaterThanP175percent<-rep (NA, Niter)
GreaterThanP200percent<-rep (NA, Niter)

For (i in 1:Niter) {
        Sample1<-rnorm (m, mean = 0, sd=1) # sample m observations from a normal distribution with mean=0, sd=1
        Sample2<-rnorm (m, mean=effect size, sd=1) # sample m observations from a Normal distribution with mean=k24 (effect size), sd=1
        sd1<-sd (Sample1) # calculate the standard deviation of Sample 1
        sd2<-sd (Sample2)
        pooledvar<-(sd1*sd1+sd2*sd2)/2 # TO DO comment
        SamplePooledVariance[i]<-pooledvar
        Nest<- ((1+k20) *(1+k21)/k21) *(((k22+k23)/effect size)^2)*k25*pooledvar
        Nestimated[i]<-Nest #estimated N
          GreaterThanNTrue[i]<-ifelse (Nest<NTrue,0,1) # if estimated N is less than Ntrue, record 0, else record 1
          k31<-ceiling (0.8*Ntrue)
          GreaterThanM20percent[i]<-ifelse (Nest<k31,0,1)
          k32<-ceiling (0.9*NTrue)
          GreaterThanM10percent[i]<-ifelse (Nest<k32,0,1)
          k33<-ceiling (1.1*NTrue)
          GreaterThanP10percent[i]<-ifelse (Nest<k33,0,1)
          k34<-ceiling (1.2*NTrue)
          GreaterThanP20percent[i]<-ifelse (Nest<k34,0,1)
          k35<-ceiling (1.3*NTrue)
          GreaterThanP30percent[i]<-ifelse (Nest<k35,0,1)
          k36<-ceiling (1.4*NTrue)
          GreaterThanP40percent[i]<-ifelse (Nest<k36,0,1)
          k37<-ceiling (1.5*NTrue)
          GreaterThanP50percent[i]<-ifelse (Nest<k37,0,1)
          k38<-ceiling (1.75*NTrue)
          GreaterThanP75percent[i]<-ifelse (Nest<k38,0,1)
          k39<-ceiling(2*NTrue)
          GreaterThanP100percent[i]<-ifelse (Nest<k39,0,1)
          k40<-ceiling (2.25*NTrue)
          GreaterThanP125percent[i]<-ifelse (Nest<k40,0,1)
          k41<-ceiling (2.5*NTrue)
          GreaterThanP150percent[i]<-ifelse (Nest<k41,0,1)
          k42<-ceiling (2.75*NTrue)
          GreaterThanP175percent[i]<-ifelse (Nest<k42,0,1)
          k43<-ceiling(3*NTrue)
          GreaterThanP200percent[i]<-ifelse (Nest<k43,0,1)
         }
 Ntrue
 results<-data.frame(GreaterThanM20percent, GreaterThanM10percent,
          GreaterThanNTrue, GreaterThanP10percent,
          GreaterThanP20percent,GreaterThanP30percent,
          GreaterThanP40percent,GreaterThanP50percent,
          GreaterThanP75percent,GreaterThanP100percent,
          GreaterThanP125percent, GreaterThanP150percent,
          GreaterThanP175percent, GreaterThanP200percent)
                                                      #View(results)
 colMeans(results)

 ForFile<-t(c(alpha,beta,m,effect
size,coverage,colMeans(results),min(SamplePooledVariance),max(SamplePooledVa
riance)))
 Colnames(ForFile)[1:5]<-c("alpha","beta","m","effect size","coverage")
 colnames(ForFile)[20:21]<-c("minPooledVar", "minPooledVar")
 write.table(ForFile,
            "H:/Test_all.csv",sep = ",",
            append=TRUE,col.names = FALSE,row.names = FALSE)
} # effect size
} # samplesize
} # coverage
} # beta
} # alpha
```

Study 2

```
set.seed(101)
k20<- 0 # q=0
k21<- 1 # r=1
Niter<-100,000
# effect size
SampleSizes<-c(8,16,32,64,128)
effect sizes<-c(0.1,.4,.75)
G<-c(.8,.9) # coverage, used with Chisq distribution.
alpha<-c(.01,.05)
B<-c(.1,.2)
For (a in 1: length (ALPHA)) {
  alpha<-ALPHA[a]
  k22<- qnorm(1-(alpha))
  for (b in 1: length(B)) {
    beta<-B[b]
    k23<-qnorm(1-beta)

       for (l1 in 1:length(G)) {
      coverage<-G[l1]
      for (l in 1: length(SampleSizes)){
```

```
m<-SampleSizes[l]
k25<- (2*m-2)/qchisq(1-coverage,df=(2*m-2))
for (j in 1: length (Effect sizes)) {
  effect size<-Effect sizes[j]
  Ntrue<- ceiling(2*(k22+k23) *(k22+k23)/(effect size*effect size) # calculate Ntrue
  h<- (1- (3/((8*m)-9)))
  NNC<-NBC<-NNH<-NBH<-rep (NA, Niter)
  for (i in 1: Niter) {
    Sample1<-rnorm (m, mean = 0, sd=1) # sample m observations from a normal distribution with mean=0, sd=1
    Sample2<-rnorm(m, mean=effect size, sd=1) # sample m observations from a normal distribution with mean=k24 (effect size), sd=1
    sd1<-sd (Sample1) # calculate the standard deviation of Sample 1
    sd2<-sd (Sample2)
    pooledvar<-(sd1*sd1+sd2*sd2)/2 # To do comment
    SamplePooledVariance[i]<-pooledvar
    diffsamplemeans<-mean (Sample2)-mean (Sample1) # xbar_2 - xbar_1
    d <- (diffsamplemeans)/sqrt(pooledvar)
    dh<-d*h
    #NNC is N where d is cohen's d Naive, Cohen
    NNC[i]<-ceiling(2*((k22+k23) ^2)/(d^2))
    #NBC is N where Browne's approach is used, Cohen's formula, Browns scaling.
    NBC[i]<-ceiling(2*k25*((k22+k23) ^2)/(d^2)
    #NNH is where Hedges h is used instead of Cohen’s d, Hedges naively
    NNH[i]<-ceiling(2*((k22+k23) ^2)/(dh^2))
    #NBH is where Hedges h is used instead of cohen's d; Hedges, Browns scaling.
    NBH[i]<-ceiling(2*k25*((k22+k23) ^2)/(dh^2))
  } # closes i
  results<-data. frame(Ntrue,NNC,NBC,NNH,NBH) #View(results)
  results$alpha<-alpha
  results$beta<-beta
```

```
    results$m<-m
    results$coverage<-coverage
    results$effect size<-effect size  #   summaryResults<-t(apply(results,2,median)
    t1<-100*apply((results[,2:5]-results[,1])/results[,1],2,mean) # mean percentage error
    t2<-100*apply((results[,2:5]-results[,1])/results[,1],2,median)#Median percentage Error # results [,1] is NTrue
    t3<-100*colMeans(results[,2:5]>results[,1])
    t4<-100*colMeans(results[,2:5]<.9*results[,1])
    t5<-100*colMeans(results [,2:5]>1.1*results [,1])
    t6<-100*colMeans(results[,2:5]>1.2*results [,1])
    names(t1) <- paste0(names(t1),"_MeanPE")
    names(t2) <- paste0(names(t2),"_MedianPE")
    names(t3) <- paste0(names(t3),"_PercentGreaterNTrue")
    names(t4) <- paste0(names(t4),"_PercentLess.9NTrue")
    names(t5) <- paste0(names(t5),"_PercentGreater1.1NTrue")
    names(t6) <- paste0(names(t6),"_PercentGreater1.2NTrue")


    t1<-t(t1)
    t2<-t(t2)
    t3<-t(t3)
    t4<-t(t4)
    t5<-t(t5)
    t6<-t(t6)
    summaryResults<-data. frame (Ntrue, results [1,6:10], t1, t2, t3, t4, t5, t6)
    if(j==1&l==1&l1==1&b==1&a==1) {
      write.table(results,"std2Naive_and_Brown_fullresults.csv”, sep = ",",
               col. names = TRUE, row. names = FALSE)
      write. table (summaryResults,"Naive_and_Brown_summaryresults.csv”, sep = ",",
               col. names = TRUE, row.names = FALSE)
    } else {
```

```
write.table(summaryResults,"std2Naive_and_Brown_summaryresults.csv",sep = ",",
                        col. names = FALSE, row. names = FALSE, append = TRUE)
          }
        } # closes j
      } # closes l
    } # closes l1
  } # closes b
} # closes a
```

Study 3

Standard Coefficient of variation

```
set.seed(100)
Niter<-100000
SampleSizes<-c (8,16,32,64,128)      # loop over l
Effect sizes<-c (.1,.4,.75)           # loop over e
Coverage<-c (0.8)                     # coverage loop over l1
Alpha<-c (.01,.05)                    # loop over a
Beta<-c (.1,.2)                       # loop over b
for (e in 1: length (Effect sizes)) {
  effect size<-Effect sizes[e]
for (a in 1: length (Alpha)) {
    alpha<-Alpha[a]
    k1<- qnorm(1-(alpha/2),0,1)            # z_(1-alpha/2)
for (b in 1: length (Beta)) {
      beta<-Beta[b]
      k2<-qnorm(1-beta,0,1)                   # z_(1-beta)
for (l1 in 1: length (Coverage)) {
        coverage<-Coverage[l1]
for (l in 1: length (SampleSizes)) {
          m<-SampleSizes[l]
          tcritical<-qt(coverage,2*m-2)
          NTrue<- ceiling(2*(k1+k2) *(k1+k2)/(effect size*effect size)) #Ntrue calculation
```

```
 #kount1<-0
 Nestimated<-rep (NA, Niter)
GreaterThanNTrue<-rep (NA, Niter)
GreaterThanM20percent<-rep (NA, Niter)
GreaterThanM10percent<-rep (NA, Niter)
GreaterThanP10percent<-rep (NA, Niter)
GreaterThanP20percent<-rep (NA, Niter)
GreaterThanP30percent<-rep (NA, Niter)
GreaterThanP40percent<-rep (NA, Niter)
GreaterThanP50percent<-rep (NA, Niter)
GreaterThanP75percent<-rep (NA, Niter)
GreaterThanP100percent<-rep (NA, Niter)
GreaterThanP125percent<-rep (NA, Niter)
GreaterThanP150percent<-rep (NA, Niter)
GreaterThanP175percent<-rep (NA, Niter)
GreaterThanP200percent<-rep (NA, Niter)
 For (i in 1: Niter) {
   Sample1<-rnorm (m, mean = 0, sd=1)
   Sample2<-rnorm (m, mean=effect size, sd=1)
   var1<-var (Sample1)
   var2<-var (Sample2)
   pooledvar<-((m-1) *var1+(m-1) *var2)/(m+m-2)
   diffsamplemeans<-abs (mean (Sample2)-mean (Sample1)) # xbar_2 - xbar_1
   cy<-sqrt(2*pooledvar/m)/diffsamplemeans
   cu<-cy+tcritical*(1+1/(8*m))/(2*sqrt(m))
   cusquared<-cu^2
   Nest<-ceiling((k1+k2) ^2*m*cusquared)                    #Nest calculation
  # To Do: comment
   Nestimated[i]<-Nest
   GreaterThanNTrue[i]<-ifelse (Nest<NTrue,0,1)             #if (nest>NT) kount1
                                                                 <-kount1+1
   k31<-ceiling (0.8*NTrue)
   GreaterThanM20percent[i]<-ifelse (Nest<k31,0,1)
   k32<-ceiling (0.9*NTrue)
   GreaterThanM10percent[i]<-ifelse (Nest<k32,0,1)
   k33<-ceiling (1.1*NTrue)
   GreaterThanP10percent[i]<-ifelse (Nest<k33,0,1)
   k34<-ceiling (1.2*NTrue)
   GreaterThanP20percent[i]<-ifelse (Nest<k34,0,1)
   k35<-ceiling (1.3*NTrue)
   GreaterThanP30percent[i]<-ifelse (Nest<k35,0,1)
   k36<-ceiling (1.4*NTrue)
   GreaterThanP40percent[i]<-ifelse (Nest<k36,0,1)
   k37<-ceiling (1.5*NTrue)
   GreaterThanP50percent[i]<-ifelse (Nest<k37,0,1)
   k38<-ceiling (1.75*NTrue)
   GreaterThanP75percent[i]<-ifelse (Nest<k38,0,1)
   k39<-ceiling(2*NTrue)
   GreaterThanP100percent[i]<-ifelse (Nest<k39,0,1)
   k40<-ceiling (2.25*NTrue)
   GreaterThanP125percent[i]<-ifelse (Nest<k40,0,1)
   k41<-ceiling (2.5*NTrue)
   GreaterThanP150percent[i]<-ifelse (Nest<k41,0,1)
   k42<-ceiling (2.75*NTrue)
   GreaterThanP175percent[i]<-ifelse (Nest<k42,0,1)
   k43<-ceiling(3*NTrue)
   GreaterThanP200percent[i]<-ifelse (Nest<k43,0,1)
   #print(kount1)
 } # closes i
 NTrue
 results<-data.frame(GreaterThanM20percent, GreaterThanM10percent,
                     GreaterThanNTrue, GreaterThanP10percent,
                     GreaterThanP20percent, GreaterThanP30percent,
                     GreaterThanP40percent, GreaterThanP50percent,
                     GreaterThanP75percent, GreaterThanP100percent,
                     GreaterThanP125percent, GreaterThanP150percent,
                     GreaterThanP175percent, GreaterThanP200percent)
 colMeans(results)
```

```
          ForFile<-
            t (c (alpha, beta, m,effect size,coverage,colMeans(results)))
          colnames (ForFile) [1:5] <-c("alpha","beta","m","effect size","coverage")
                  write. table (ForFile,"C:/Users/sc-obodo/Documents/Test81.csv”, sep =
           ",”, append=TRUE, col. names = FALSE, row. names = FALSE)
        } # closes l
      } # closes l1
    } # closes b
  } # closes a
} # closes e
```

McKay formula

```
set. seed (101)
Niter<-100000
SampleSizes<-c (8,16,32,64,128)            # loop over l
Effect sizes<-c (0.1,0.4,0.75)            # loop over e
Coverage<-c (0.8)                         # coverage, used with Chisq distribution.
Alpha<-c (0.01,0.05)                      # loop over a
Beta<-c (0.1,0.2)                         # loop over b

for (e in 1: length (Effect sizes)) {
  effect size<-Effect sizes[e]
  for (a in 1: length (Alpha)) {
    alpha<-Alpha[a]
    k1<- qnorm(1-(alpha/2),0,1) # z_(1-alpha/2)
    for (b in 1: length (Beta)) {
      beta<-Beta[b]
      k2<-qnorm(1-beta,0,1) # z_(1-beta)
      for (l1 in 1: length (Coverage)) {
        coverage<-Coverage[l1]
        for (l in 1: length (SampleSizes)) {
          m<-SampleSizes[l]
          k3<-(k1+k2) ^2
```

```
u2<-qchisq (1-coverage, df=(2*m-2))
NTrue<- ceiling(2*(k1+k2) *(k1+k2)/(effect size*effect size))
Nestimated<-rep (NA, Niter)
GreaterThanNTrue<-rep (NA, Niter)
GreaterThanM20percent<-rep (NA, Niter)
GreaterThanM10percent<-rep (NA, Niter)
GreaterThanP10percent<-rep (NA, Niter)
GreaterThanP20percent<-rep (NA, Niter)
GreaterThanP30percent<-rep (NA, Niter)
GreaterThanP40percent<-rep (NA, Niter)
GreaterThanP50percent<-rep (NA, Niter)
GreaterThanP75percent<-rep (NA, Niter)
GreaterThanP100percent<-rep (NA, Niter)
GreaterThanP125percent<-rep (NA, Niter)
GreaterThanP150percent<-rep (NA, Niter)
GreaterThanP175percent<-rep (NA, Niter)
GreaterThanP200percent<-rep (NA, Niter)

for (i in 1: Niter) {
  Sample1<-rnorm (m, mean = 0, sd=1)
  Sample2<-rnorm (m, mean=effect size, sd=1)
  var1<-var (Sample1)
  var2<-var (Sample2)
  pooledvar<-((m-1) *var1+(m-1) *var2)/(m+m-2)
  diffsamplemeans<-abs (mean (Sample2)-mean (Sample1)) # xbar_2 - xbar_1
  csquared<-(2*pooledvar)/ (m* diffsamplemeans^2)
  k4<- csquared
  k8<-((u2)/(2*m)-1)
  k9<-k8*(k4)
  k10<-(u2/(2*m-2))
  k11<-(k4/(k9+k10))
  Nest<-ceiling(k3*m*k11)            #Mckay Nest
  Nestimated[i]<-Nest
  GreaterThanNTrue[i]<-ifelse (Nest<NTrue,0,1) #if (nest>NT) kount1
```

```
                                                      <-kount1+1
            k31<-ceiling (0.8*NTrue)
            GreaterThanM20percent[i]<-ifelse (Nest<k31,0,1)
            k32<-ceiling (0.9*NTrue)
            GreaterThanM10percent[i]<-ifelse (Nest<k32,0,1)
            k33<-ceiling (1.1*NTrue)
            GreaterThanP10percent[i]<-ifelse (Nest<k33,0,1)
            k34<-ceiling (1.2*NTrue)
            GreaterThanP20percent[i]<-ifelse (Nest<k34,0,1)
            k35<-ceiling (1.3*NTrue)
            GreaterThanP30percent[i]<-ifelse (Nest<k35,0,1)
            k36<-ceiling (1.4*NTrue)
            GreaterThanP40percent[i]<-ifelse (Nest<k36,0,1)
            k37<-ceiling (1.5*NTrue)
            GreaterThanP50percent[i]<-ifelse (Nest<k37,0,1)
            k38<-ceiling (1.75*NTrue)
            GreaterThanP75percent[i]<-ifelse (Nest<k38,0,1)
            k39<-ceiling(2*NTrue)
            GreaterThanP100percent[i]<-ifelse (Nest<k39,0,1)
            k40<-ceiling (2.25*NTrue)
            GreaterThanP125percent[i]<-ifelse (Nest<k40,0,1)
            k41<-ceiling (2.5*NTrue)
            GreaterThanP150percent[i]<-ifelse (Nest<k41,0,1)
            k42<-ceiling (2.75*NTrue)
            GreaterThanP175percent[i]<-ifelse (Nest<k42,0,1)
            k43<-ceiling(3*NTrue)
  GreaterThanP200percent[i]<-ifelse (Nest<k43,0,1)
} # closes i
NTrue
results<-data. frame (GreaterThanM20percent, GreaterThanM10percent,
                GreaterThanNTrue, GreaterThanP10percent,
                GreaterThanP20percent, GreaterThanP30percent,
                GreaterThanP40percent, GreaterThanP50percent,
                GreaterThanP75percent, GreaterThanP100percent,
                        GreaterThanP125percent, GreaterThanP150percent,
                        GreaterThanP175percent, GreaterThanP200percent)
          colMeans(results)
          ForFile<-
            t (c (alpha, beta, m, effect size, coverage, colMeans(results)))
          colnames (ForFile) [1:5] <-c("alpha","beta","m","effect size","coverage")
          write. table (ForFile,"C:/Users/sc-obodo/Documents/Test64.csv”, sep = ",",
                      append=TRUE, col. names = FALSE, row.names = FALSE)
        } # closes l
      } # closes l1
    } # closes b
  } # closes a
} # closes e
```

Appendix C: Other result from study one

Table C1: Percentage of time estimated sample size exceeds Ntrue +/- error at α= 0.01, β=0.2, γ=0.9

Sample size	Effect size	>-20%	>N True	>+20%	>+30%	>+50%	>+100%	>+150%
5	.10	.947	.900	.839	.805	.732	.541	.368
10	.10	.965	.898	.790	.721	.572	.244	.076
30	.10	.990	.901	.640	.472	.195	.005	.000
50	.10	.997	.901	.520	.303	.058	.000	.000
‘100	.10	1.000	.899	.310	.092	.002	.000	.000
5	.40	.947	.901	.838	.806	.733	.541	.369
10	.40	.966	.899	.786	.720	.571	.245	.077
30	.40	.990	.900	.629	.467	.193	.005	.000
50	.40	.996	.901	.506	.299	.059	.000	.000
100	.40	.999	.901	.292	.093	.002	.000	.000
5	.75	.943	.901	.839	.806	.726	.529	.366
10	.75	.962	.899	.785	.720	.559	.233	.075
30	.75	.987	.899	.630	.468	.178	.004	.000
50	.75	.994	.900	.512	.306	.049	.000	.000
100	.75	.999	.900	.298	.093	.002	.000	.000

Appendix D

Other achievements and Awards during the PhD

March 2020- Presented -Telling your research story-UWE Bristol event.

October 2021-Attended-7th Trial Methodology Symposium Ireland and UK (online)

October 2022-Attended- Black heroes of mathematics conference.

October 2022- Attended- Royal Statistical Society conference.

Awards

Black Hall of Fame recipient in 2020-2021 for instrumental contribution to UWE Bristol community and the wider society.

Nominated for committee member of the year in 2021 for outstanding support to fellow students as a committee President.

Organizing committee member and chair of session for 2021 South West Doctoral Training Partnership (SWDTP) Conference 2021- Getting Through It: Uncertainty in research. https://www.swdtp.ac.uk/our-students/student-conferences/swdtp-conference-2021-getting-through-it-uncertainty-in-research/

2020 – 2021- President of SU society. UWE Bristol

2020 – 2021 -Student council Member. UWE Bristol

2021 - 2023-Interview shared Tips for RD1 and RD2 success featured as resource material in Research Skills development programme on UWE Blackboard.

2021-202- Associate lecturer – Supporting tutorial sessions at UWE Bristol.