Bayesian sample size determination: a simulation study comparing sequential and a priori designs for logistic regression

Coen Willemsen (1476092)
Supervisor: Mirjam Moerbeek

23-6-2025

Introduction

In any statistical study — but particularly in social science research — determining the sample size is an important step, yet it is very often a difficult step as well. The size of the sample used for a statistical test is directly and positively related to its power, which is the probability that it will yield statistically significant results (Cohen, 1988).

The difficulty lies in the fact that the sample size should be large enough to reach sufficient power to detect the effect of interest, but avoid becoming so large that the study becomes inefficient or ethically dubious (Lenth, 2001). It is clear that too little power is undesirable, since a study that cannot detect the effect size that it is looking for (while in fact that effect does exist) can impede scientific progress and waste resources. However, too much power might not necessarily be seen as a disadvantage: why would it be bad to have a high probability of finding an effect?

But an overpowered study also wastes resources, since fewer subjects would have been necessary to draw the same conclusions. It could be the case that thereby resources are drawn away from research that needs it more. As alluded to in the opening sentence, the additional difficulty in social science research is that the sample usually consists of humans, and subjecting more people than necessary to an experiment may not be ethically justifiable. Even if the study does not involve a clinical trial with potential harmful side effects, it can still be argued that one ought to minimize the number of people used to conduct a study (Lenth, 2001). Furthermore, the cost (in time or money) associated with sampling too many subjects can be problematic in and of itself.

Sample size determination methods exist to address these issues, providing appropriate sample sizes based on statistical methods. For example, power analysis through a program such as G*Power provides a sample size based the desired power, effect size, and significance criterion (the α level) (Faul et al., 2007). This is done before the actual sampling, so this is known as an a priori method. There are also sequential designs, where the necessary sample size can be recomputed during the course of a study, based on interim results (Adcock, 1997).

Traditionally, such methods have been implemented within the frequentist framework. However, such implementations can suffer from practical problems, such as the need to set the maximum number of sample sizes that can be tried during a sequential design beforehand. The frequentist approach has also been criticized more fundamentally in the literature many times. For example, the null hypothesis is almost never actually in line with the research question and thus often misinterpreted (Cohen, 1994). Integral to null hypothesis significance testing are p-values, which are challenged for depending on data that is never observed and not actually quantifying statistical evidence (Wagenmakers, 2007). Though this subject will be revisited briefly in the methods section, a full discussion and comparison to Bayesian methods is outside the scope of this study.

Bayesian methods do, however, provide a compelling alternative. Bayesian sample size determination does not rely on the null hypothesis or p-values. Instead, informative hypotheses are formulated and support for those hypotheses can be quantified by Bayes factors (Kass and Raftery, 1995; Klugkist et al., 2005). As will be expanded upon later, Bayes factors only rely on the support found in the observed data and do not suffer from many of the same practical problems as frequentist methods.

Bayesian sample size determination methods have been steadily receiving more and more attention in the literature (e.g. Adcock, 1997; Brutti et al., 2008; Wang and Gelfand, 2002; Wilson et al., 2022; Zhang et al., 2011). There have only been a handful of simulation studies evaluating the impact of different input parameters and scenarios on such methods, though (Fu et al., 2020, 2022; Moerbeek, 2021). These simulation studies are necessary to know how Bayesian sample size determination methods behave under certain conditions, and when they might not function as expected. Most simulation studies have thus far focused mainly on simpler statistical tests, such as the t-test. Fu et al. (2022) have expanded their research to ANOVAs, too.

In this study, Bayesian sample size determination for a logistic regression is investigated through a simulation study, thereby extending the current literature by analyzing a more complex statistical test. Furthermore, past simulation research has focused mainly on either sequential designs (Moerbeek, 2021) or a priori methods (Fu et al., 2020). This study directly compares the two, which is another useful addition to the knowledge on Bayesian sample size determination methods. The following research questions lie at the heart of this study:

1. How often do the sequential and a priori sample size determination methods correctly show support for the hypothesis that was used to generate the data?

2. To what extent do both methods return similar sample sizes under similar conditions?

In answering these questions, it can be assessed whether researchers should have a preference for either method given their particular circumstance, or if they could for example just choose based on convenience. In the next section, the topics that have been introduced briefly will receive more attention, so that the sample size determination methods can be understood. The simulation setup will be described and the results discussed. Finally, there is space for a conclusion and discussion.

Simulation methods

Logistic regression

The statistical test chosen for this study was logistic regression. Previous simulation studies on Bayesian sample size determination have focused mainly on t-tests (Fu et al., 2020; Moerbeek, 2021). Fu et al. (2022) have expanded their research to ANOVA models. Recently, even analytical solutions for sample size determination for the t-test have been found (Pawel and Held, 2025).

The t-test offers computational and analytical convenience and is widely used to compare group means. However, in practice, more complicated models are often necessary.

Logistic regression is one of these more complicated models, modeling binary outcomes based on one or more predictors. It is, for example, commonly used in clinical trials and the social sciences (e.g. Del Prette et al., 2012; Ockene et al., 1991; Zardo and Collie, 2014). By examining Bayesian sample size determination for logistic regression, the existing methodological work is matched to the more commonly used statistical models.

The goal of logistic regression is to estimate a set of coefficients for the predictor variables that best explain the relationship with the binary outcome variable. This is done through maximum likelihood estimation, which finds the coefficients that maximize the probability of observing the data, given the predictor variables. This leads to a regression equation similar to that of linear regression. The difference is that a link function is necessary to map the values of the linear predictors (which can range from $-\infty$ to ∞) to the $[0,1]$ interval. This is because the outcome variable the is probability of some event occurring, for which only variables on that constrained interval are appropriate. Although different link functions exist, often the logit function is used, which is the natural logarithm of the odds of the outcome variable. The model for binary logistic regression as a whole then looks like this:

$$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}$$

Where the x's are the predictor variables, each with its own slope coefficient β. The number of predictors k can vary, as long as there is at least one. For simplicity, the sets of predictors and weights will be denoted as just $\beta_i x_i$, with the understanding that this represents the sum of all predictors and coefficients. In this model, π is the probability of the outcome of interest y happening given the predictor values x, which can be rewritten from Equation 1 in the following way:

$$\pi = P(y|x) = \frac{e^{\alpha+\beta_i x_i}}{1+e^{\alpha+\beta_i x_i}} \tag{2}$$

Here α and β are the model parameters to be estimated. The intercept α can be seen as the baseline logit rate of the event occurring (i.e. when x = 0). An α of 0 corresponds to a baseline probability of 0.5. Negative values for α give lower baseline rates, while positive values give higher rates. β is the effect of the predictor variable x. If it is positive, the predictor leads to a higher probability of the event occurring, while a negative value for β corresponds to a lower probability.

Informative hypotheses

Informative hypotheses are theory-driven statements that explicitly specify expectations about the model parameters of interest (Klugkist et al., 2005). These expectations can be based on previous findings, literature, or the researcher's beliefs (Gu et al., 2014; Van Lissa et al., 2020).

For example, one might study political participation and expect that education has a stronger effect than income, both being positive. Income, in turn, has a stronger effect than gender. The following informative hypothesis H_1 for the regression coefficients β would be formulated as such:

$$H_1 : \beta_{\text{education}} > \beta_{\text{income}} > \beta_{\text{gender}}$$

The above informative hypothesis is known as an inequality constrained hypothesis, since it defines its expectations using only inequalities (i.e., using symbols such as $<$ and $>$). It is also possible to formulate a hypothesis that constrains parameters to be equal ($=$) or approximately equal ($\approx$). The approximate equality holds if the value falls in some small predefined region δ around the value of interest. It is also possible that there is no expectation hypothesized between variables, that is, they are unconstrained. This means that it always holds. As an example, if an expectation for the difference between income and education is not given, but both are expected to be bigger than gender, which is approximately zero, then the informative hypothesis H_2 would specify:

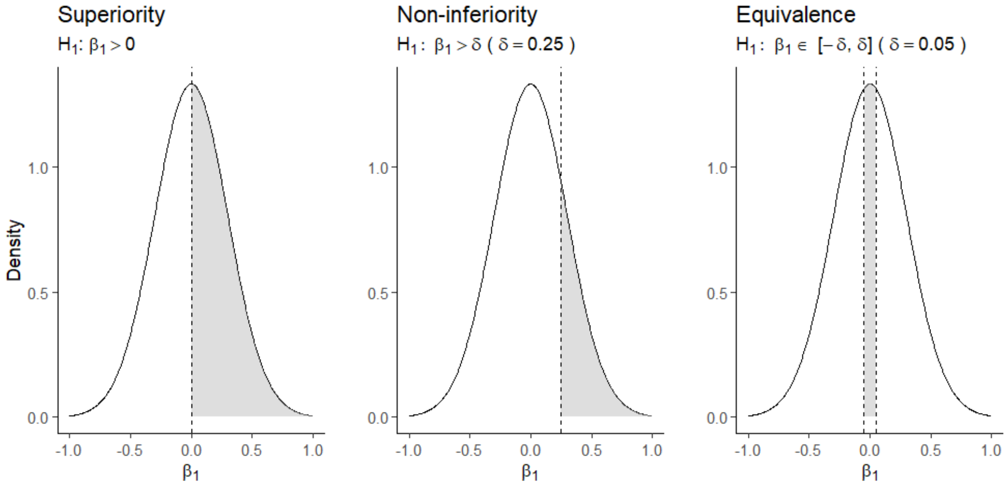
$$H_2 : (\beta_{\text{education}}, \beta_{\text{income}}) > \beta_{\text{gender}} \approx 0$$

where δ could be defined as 0.03, meaning that if the value for β_{gender} falls within $[-0.03, 0.03]$, it satisfies the equality constraint. This is useful, because the probability of a coefficient being an exact number is zero, under a continuous distribution. The comma between the education and income variables denotes an unconstrained hypothesis.

Different types of hypotheses lead to different research designs. Within clinical trials, these are commonly referred to superiority, non-inferiority, and equivalence designs (e.g. Cortegiani et al., 2020; Fukai et al., 2024; St-Jules et al., 2016). A superiority trial is designed to have as its hypothesis that the effect size is greater than 0, while a non-inferiority trial hypothesizes that the effect size is greater than some value δ, and the equivalence trial tests that the effect size falls within some region $[-\delta, \delta]$ (van Ravenzwaaij et al., 2019).

Informative hypotheses can easily be translated to such designs. In Figure 1, the different trial designs are displayed. The shaded region is the values that β_1 can take on to satisfy the hypotheses. In this case, a single regression coefficient is used for visualization, but any parameter could be tested. For example, the difference in means in a t-test could also be examined. Any value of δ could also be used. However, it is important to note that a superiority trial always tests for the parameter value being greater than 0. In this study, the structure of these trial categories will be used to assess sample size determination between hypothesis types.

Figure 1: Visualization of trial designs. The shaded region refers to the values that β_1 can take on to satisfy the hypothesis.



Informative hypotheses can be specified for a variety of models, not only regression. It is not necessary that a null hypothesis is included, which is an advantage of informative hypotheses in comparison to null hypothesis significance testing (NHST). The null-hypothesis in NHST specifies that all parameters equal 0. This is often not a relevant hypothesis in practice, and it exists only to be rejected in those cases (Van Lissa et al., 2020). If it is rejected, still there is no direct support found in favor of the alternative hypothesis (which is of actual interest to the researchers), and there is only direct evidence against the null hypothesis. Since informative hypotheses allow for direct testing, the results can also be interpreted in line with the substantive questions that are at their origin.

Bayes factors

Using Bayesian hypothesis testing, the relative support in the data for one hypothesis against another can be quantified. Instead of p-values as used in NHST, Bayes factors (BF) are used to evaluate hypotheses. More specifically, the way Bayes factors evaluate the support in the data D for two hypotheses H is with a ratio of their marginal likelihoods (Kass and Raftery, 1995).

$$BF_{12} = \frac{p(D|H_1)}{p(D|H_2)} \tag{3}$$

Kass and Raftery (1995) show how to arrive at this equation from Bayes' theorem, where the posterior distribution for the probability of the observed data occurring under H_k is:

$$p(H_k|D) = \frac{p(D|H_k)p(H_k)}{p(D|H_1)p(H_1) + p(D|H_2)p(H_2)}, \quad k = (1,2) \tag{4}$$

Dividing the posterior distributions by each other and simplifying returns Equation 5, a ratio that reflects the relative support for the competing hypotheses:

$$\frac{p(H_1|D)}{p(H_2|D)} = \frac{p(D|H_1)}{p(D|H_2)} \frac{p(H_1)}{p(H_2)} \tag{5}$$

The term $p(H_1)/p(H_2)$ here is the ratio of the prior distributions, which are formulated before observing the data. Therefore, the ratio $p(D|H_1)/p(D|H_2)$ represents directly the relative support provided by the data. As given in Equation 3, this term is the definition of the Bayes factor, quantifying the evidence for a hypothesis against another based purely on the information in the data.

A Bayes factor greater than 1 ($BF_{12} > 1$) indicates more support in the data for hypothesis H_1 than for H_2, whereas a value less than 1 implies the opposite. The amount of support is directly interpretable: $BF_{12} = 5$ implies that there is 5 times more support in the data for H_1 than for H_2. Bayes factors are reciprocal, in the sense that $BF_{21} = 1/BF_{12}$. So, in the case of $BF_{12} = 5$, BF_{21} would be 0.2, where the subscript denotes the 'direction' of the support.

Kass and Raftery (1995) also define guidelines for interpreting Bayes factors as follows: weak for a BF between 1 and 3, positive for a BF between 3 and 20, strong for a BF between 20 and 150, and very strong beyond that. It should be noted that these are guidelines and can differ based on context. For example, the stakes of an analysis may influence the choice of cutoff points. If the Bayes factor is used to decide whether or not to approve a new medical treatment, stronger evidence may be needed before a positive effect is concluded than if the decision was between two types of advertisement. Furthermore, there is an ongoing discussion in the literature regarding the place of such numerical thresholds in Bayesian statistics, with some scholars arguing against categorizing a continuous measure (Palfi and Dienes, 2020; Tendeiro et al., 2024; van der Linden and Chryst, 2017).

As described above in Equation 3, Bayes factors are computed as the ratio of the marginal likelihoods of the competing models. The marginal likelihood can be seen as the likelihood of observing the data, given the current hypothesis (Béland et al., 2012). However, estimation of the marginal likelihoods can be computationally intensive or even impossible. Klugkist et al. (2005) introduced a method to compute Bayes factors without the need to compute marginal likelihoods. This method makes use of the fit and complexity of the hypotheses at hand. The fit f_i is defined as the proportion of the posterior distribution that is in agreement with H_i, while the complexity c_i is the proportion of the prior distribution that is in agreement with H_i. Both quantities can take on any value in the range [0, 1].

Originally, this approach was developed to compare an informative hypothesis H_1 against the unconstrained hypothesis H_u. In that case, the Bayes factor is computed like this:

$$BF_{1u} = \frac{f_1}{c_1} \tag{6}$$

As detailed in Gu et al. (2017), if two hypotheses share the same unconstrained hypothesis, then their Bayes factors can be directly compared:

$$BF_{12} = \frac{BF_{1u}}{BF_{2u}} = \frac{f_1/c_1}{f_2/c_2} \tag{7}$$

And if and only if H_2 is the complement of H_1, its fit and complexity can be rewritten in terms of H_1. Two hypotheses are each others complement if they are mutually exclusive and exhaustive. Since this simulation study will only compare hypotheses against their complement, this is the version of the Bayes factor that will be used in this paper:

$$BF_{12} = \frac{f_1/c_1}{(1-f_1)/(1-c_1)} \tag{8}$$

Adjusted approximate fractional Bayes factors (AAFBF)

Having established how the Bayes factor is computed, it will now be shown how its two components, fit and complexity, are determined. The approximated adjusted fractional Bayes factor (AAFBF) approach will be used. For technical details, see Gu et al. (2017). The AAFBF is *fractional* because a fraction of the data is used to compute the prior distribution. It is *adjusted* in the sense that the prior distribution is adjusted to fall around the focal point, which is in the middle of the range space. This notion will be expanded upon later in this section. The *approximation* comes from the fact that the prior and posterior distributions are approximated to normal distributions, which again eases computation and allows for more general use.

The fit f_i describes to what extent the data support H_i. This is computed by integrating over that part of the posterior distribution that contains values in line with the hypothesis:

$$f_i = \int_{\beta_1 \in H_i} g_i(\beta_i|D)\; d\beta_i \tag{9}$$

where $\beta_i \in H_i$ denotes that the logistic regression coefficient β_i falls within the constraints of H_1 (for example, $\beta_1 > 0$), and g_i denotes the posterior distribution under H_i.

The complexity c_i represents how much of the prior distribution is consistent with H_i. Similar to the fit, this is defined as:

$$c_i = \int_{\beta_i \in H_i} h_i(\beta_i)\; d\beta_i \tag{10}$$

where h_i denotes the prior distribution under H_i. The complexity can be interpreted as the a priori likelihood of the model. Intuitively, more constrained hypotheses will generally have a lesser complexity. For example, the hypothesis $H_1 : \beta_1 > 0$ will have a lesser complexity than the hypothesis $H_1 : \beta_1 > 0.25$, if the prior distribution for both is the same. Of course, it could be the case that a more constrained hypothesis actually covers more area of the prior distribution than the less constrained hypothesis. This depends on the prior and hypotheses at hand. The greater the complexity, the smaller the Bayes factor (if the fit stays the same), so more parsimonious hypotheses will be preferred.

The posterior and prior distributions are approximated to normal distributions. The posterior distribution g_i is defined to have the maximum likelihood estimates of the regression coefficient as its mean, and the standard deviation is based on the estimate of the variance:

$$g_i(\beta_i|D) = \mathscr{N}(\hat{\beta}, \hat{\sigma^2}) \tag{11}$$

The prior distribution h_i is similarly normal, except that it is centered around the focal point β^* and has a σ^2 that is estimated by using a fraction b of the data. The focal point is defined to be the midpoint of the range space, which is the boundary between the regions where the hypothesis holds and does not hold. For inequality constrained hypotheses, this corresponds to the value where the constraint flips. For example, a superiority trial has $H_1 : \beta_1 > 0$. The alternative hypothesis then is $\beta_1 < 0$. The point $\beta_1 = 0$ is the midpoint and thus the focal point. Since superiority trials always test for $\beta > 0$, the focal point is always 0. For non-inferiority trials ($\beta_1 > \delta$), the midpoint is δ, through similar reasoning. For equivalence trials, the hypothesis is that the parameters falls in some range $[-\delta, \delta]$, where the midpoint is always 0 and thus the focal point is always 0 as well. Gu and colleagues (2017) explain this in more technical detail. The focal point and fractional estimation of the variance lead to the following definition of the prior distribution:

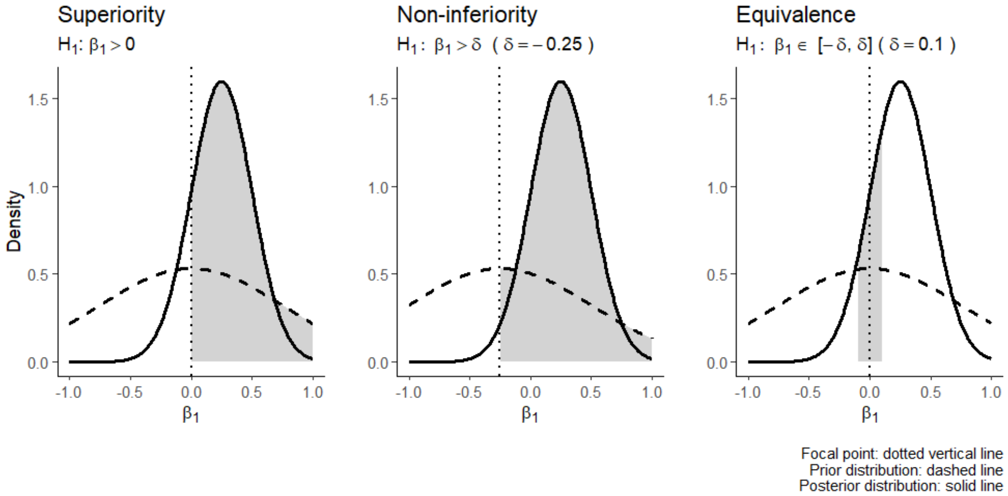
$$h_i(\beta_1|D) = \mathscr{N}(\beta^*,\ \frac{1}{b}\hat{\sigma^2}) \tag{12}$$

As can be seen in Figure 2, both the prior and posterior distribution are normalized. The shaded area corresponds to the area where the hypothesis is satisfied. Notice how the prior distribution is centered around the focal point (dotted vertical line), and that this is not 0 for the non-inferiority trial. Although there has been discussion in the literature about choices for b (see Gu et al. (2017) for an overview and analysis), the value used here is $b = \frac{1}{n}$. This means that the information of one subject is taken to compute the standard deviation, as is also done in e.g. Fu et al. (2020) and Moerbeek (2021). That practice is based on the minimal training sample idea, where the minimum amount of data needed to construct a proper prior distribution is taken, and the rest is left for hypothesis testing (Berger and Pericchi, 2004). A proper prior is one that integrates to 1 (such as a normal distribution), as opposed to an improper prior (such as an unbounded uniform distribution), which does not. Furthermore, when two inequality constrained hypotheses are tested, the Bayes factor is invariant to the choice of b. Choosing a default value is therefore sufficient for the purposes of this study.

Sample size determination

As discussed in the introduction, determining an appropriate sample size is an important aspect of social science research. This section presents the methods used for Bayesian sample size determination, namely a sequential and an a priori design. First, the sequential design will be established, along with the specific

Figure 2: Visualization of fit and complexity. The shaded region refers to the values that β_1 can take on to satisfy the hypothesis.



algorithm used to simulate data in this study. The a priori method will be described in a similar manner. Finally, the parameters of interest for both methods will be discussed, which leads to the simulation results.

Sequential design

The basic aim of a Bayesian sequential design is to find an appropriate sample size by conducting the statistical test at a given starting sample size, and continue adding subjects and testing until the Bayes factor is equal to or exceeds a predetermined threshold. When this threshold is reached, it is determined that enough support in the data to make informed statements about the hypotheses at hand is found. The number of subjects added between each sample size can be chosen arbitrarily, as fits the situation at hand. Furthermore, the number of tests before stopping is unbounded, since there is no Type I error to account for. These are advantages of Bayesian sequential designs over frequentist sequential designs. As detailed in e.g. Wassmer and Brannath (2016), such frequentist methods are more constrained. For example, the accepted probability of a Type I error (α) must be determined beforehand, which means that each intermittent test must have a set number of subjects added, as well as a maximum number of tests, so that α is not exceeded. If the final test statistic does not turn out to be significant, there cannot be any more tests conducted. Thus, Bayesian sequential designs are generally more flexible than frequentist methods.

In this study, the sequential design is based on the adjusted approximate fractional Bayes factor (AAFBF), as discussed above. This statistic is the basis for the decision rule, which determines whether more data is necessary or not. Since this is a simulation study, the data will be generated and therefore have some artificial structure, such as equal sized groups in the dichotomous independent variable. However, this is not a requirement if this method were to be applied to real-world data. Step for step, the simulation algorithm has the following form:

Algorithm 1 Simulating a sequential sample size determination method

```
Input: $N_{\text{min}}, N_{\text{max}}$: minimum and maximum sample sizes
Input: $N_{\text{step}}$: step size for sample increase
Input: $\text{AAFBF}_{\text{threshold}}$: evidence threshold

Output: Final sample size $N$ and evidence

$N \leftarrow N_{\text{min}}$
while $N < N_{max}$ do
    Generate balanced $x \in \{0,1\}$ of size $N$
    Compute $y = \text{logit}(\alpha + \beta x)$
    if $x$ or $y$ has no variance then
        $N \leftarrow N + N_{\text{step}}$ restart loop
    Compute logistic regression of $y$ on $x$
    Compute AAFBF
    if $(AAFBF \geq AAFBF_{threshold}) \vee (AAFBF \leq 1/AAFBF_{threshold})$ then
        break
    $N \leftarrow N + N_{\text{step}}$
return $N$, $AAFBF$
```

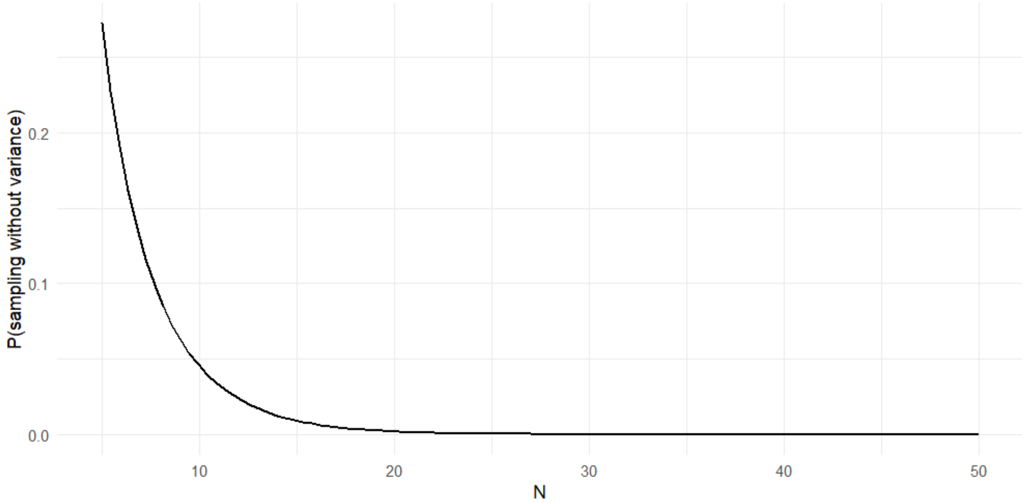
This is repeated a set number of times for each set of parameters, in this case 5000 times. Since there are 48 unique combinations of parameters under investigation, there were in the end 240.000 simulations of this algorithm conducted.

Some terms that have not been introduced earlier are important parameters for this

function. For example, N denotes the current sample size, while N_{step} represents the number of subjects that get added each iteration. $\text{AAFBF}_{\text{threshold}}$ refers to the minimum Bayes factor that is considered enough evidence to prefer one hypothesis over the other.
At the end of this section, there will be an overview of all parameters for both the sequential and a priori methods.

The check at step 5 is necessary to produce valid logistic regression models. If the x or y variable is generated without variance, there is no comparison group in the predictor or no pattern in the outcome variable to model. Though the probability of sampling without variance is small, it is non-negligible. The parameter setup in this study where sampling without variance is most likely, is when the sample size 10, the coefficient is 0, and the intercept is -1 (this is fixed). This gives an approximate probability of $P(\text{no variance in sample}) = 2 \times 0.5^{10} + 0.731^{10} + 0.269^{10} \approx 0.045.$[1] It should be noted that this probability decreases quickly as sample size increases, as can be seen in Figure 3.

Figure 3: Visualization of the probability of sampling either x or y without variance.



If there is a sample without variance, the sample size is increased by the step

[1] The probability of sampling y without variance depends on the sampling distribution in x, but here they can be treated as independent because $\beta_1 = 0$.

size, and the algorithm continues from step 1. Though other solutions would be possible (for example, resampling the data until there is variance), this choice was made to reflect how a researcher would deal with such problems in a real-world sequential design scenario. If it were the case that all subjects had the same value for the predictor or outcome variable, they would most likely increase their sample size as described here, and resampling would generally not be feasible.

Furthermore, the decision rule in step 9 states that the iteration will be ended if the AAFBF exceeds the threshold, but also if it reaches the inverse of the threshold ($\text{AAFBF} \leq 1/\text{AAFBF}_{\text{threshold}}$). As discussed earlier, Bayes factors are reciprocal, meaning that the inverse of a Bayes factor represents the support in the opposite direction. In this study, that corresponds to the support in the data for the complement of the hypothesis at hand.

If the threshold represents the amount of evidence that is needed to prefer one hypothesis over the other, then the inverse of the threshold should hold the same weight. This decreases the amount of iterations needed to simulate, since iterations that provide enough evidence for the complement of the hypothesis can conclude, instead of running until the sample hits N_{max}. It also allows for a more informative conclusion: instead of only being able to state that there was not enough evidence for H_1, it can also be stated that there was enough evidence to prefer $H_{\text{complement}}$.

A priori design

Though the goal of a priori sample size determination is the same, the key difference is that this method returns the final sample size not during the study, but beforehand. The appropriate sample size can only be computed given an effect size, since they are dependent (Cohen, 1992). Therefore, the subsequent other key difference that the a priori method necessitates specifying an expected effect size before any data is gathered. The sequential method implicitly uses the effect sizes estimated during the tests starting at N_{start}.

The need to specify effect sizes beforehand can be seen as a weakness, since the effect size is precisely what is unknown and estimating it is usually the goal of the study. There are ways to logically set an estimated effect size, for example based on expert knowledge, or previous studies. The minimum effect size that would be of interest could also be taken, so that a sample size with enough statistical power to find that minimum size is chosen. Such logical choices can provide reasonable justification for specifying effect size beforehand.

The a priori sample size determination method implemented in this paper requires setting an effect size as well, but also the probability of finding that effect. This is also known as the power or $1 - P(\text{Type II error})$ (Cohen, 1992). It is mainly based on the simulation study done on Bayesian sample size determination by Fu et al. (2020), where the parameter to set that probability is called η. Their research is extended here by determining sample size for logistic regression, as well as by the comparison to sequential designs.

The method works by doing a binary search for evidence across the space of sample sizes provided by the minimum and maximum sample sizes. At each point it searches, T simulations of the test are done. If it finds enough evidence, the sample size can be reduced and the simulations start again. If there is too little evidence, sample size is increased. At some point, the sample size will converge, where any step downward will lead to insufficient evidence. At this point, the final simulations are done and the power is reported. It is described in Algorithm 2. Please note that the subscript in H_c refers to the complementary hypothesis of H_1.

Algorithm 2 Simulating an a priori sample size determination method

Input: $N_{\min}, N_{\max}$: minimum and maximum sample sizes
Input: T: number of simulations per step
Input: $\text{AAFBF}_{\text{threshold}}$: evidence threshold
Input: η: posterior probability threshold

Output: Final sample size N and estimated power

```
while $N_{\max} - N_{\min} > 1$ do
    $N \leftarrow \lfloor (N_{\min} + N_{\max})/2 \rfloor$ // if first iteration:  $N \leftarrow N_{\mathtt{min}}$
    skipped $\leftarrow 0$
    for $T$ do
        Generate balanced $x \in \{0,1\}$
        Compute $y = \text{logit}(\alpha + \beta x)$
        if $x$ or $y$ has no variance then
            skipped $\leftarrow$ skipped $+1$  continue
        Compute logistic regression of $y$ on $x$
        Compute AAFBF
    if skipped $\geq 0.05T$ then
        $N \leftarrow N + 1$  restart loop
    $p_1 \leftarrow P(\text{AAFBF} \geq \text{AAFBF}_{\text{threshold}} \mid H_1)$
    $p_c \leftarrow P(\text{AAFBF} \leq 1/\text{AAFBF}_{\text{threshold}} \mid H_c)$
    if $(p_1 \wedge p_c) > \eta$ then
        $N_{\max} \leftarrow N$
    else
        $N_{\min} \leftarrow N$
return $N,\ p_1,\ p_c$
```

The binary search was an improvement suggested by Fu et al. (2020). The earlier version would increase the sample size with 1 if there was not enough evidence found, but this lead to long computing times since the T simulations would be run with every increase. Since the value for T is recommended to be set at 10.000, it is computationally heavy and thus time-intensive. The binary search decreases the number of iterations significantly, to maximally 12 (Fu et al., 2020).

As in Algorithm 1, there is a check for a sample without variance here as well, and for the same reason. However, since there are many samples taken (T), the probability of at least one having no variance is very high, and resampling every time or immediately increasing N is not desirable. Therefore, there is the somewhat arbitrary choice to only resample with $N+1$ if more than 5% of the samples contain no variance, rendering them unusable. There is 1 added to the new sample size to reflect what a researcher would most likely do in practice, as was also done in the sequential design. The choice is arbitrary in that there is no established literature on how much missing data is acceptable in similar studies.

However, the data is randomly generated and fixed to be balanced for the x variable. The missingness of x can thus reasonably be assumed to be Missing Completely At Random (MCAR), which means that listwise deletion will still produce unbiased results (Rubin, 1976).

The missingness for the y variable is dependent on x, especially when $\alpha \neq 0.5$. This means y may not strictly be assumed MCAR, and deleting listwise could introduce some bias. At the same time, this is mitigated by the fact that the predictor is equally distributed. The bias is expected to be minimal in practice, especially with the constraint of a maximum of 5% of samples missing.

As discussed earlier, to accept the sample size there must be a probability η that the AAFBF crosses the threshold value. This value of η was recommended to be set at 0.8 by Fu et al. (2020). Researchers working on higher-stakes outcomes might prefer to set it at 0.9, so that there is less risk of getting an inappropriate sample size.

Similar to the sequential design, there is a check that the AAFBF crosses the threshold for the hypothesis at hand, but also for its complement. In this case, the AAFBF must exceed the threshold under the condition that H_1 is correct (i.e. sampling under β), but it must also exceed the inverse of the threshold under the condition that H_c is correct (i.e. sampling under $-\beta$). Again, this makes use of the reciprocity of Bayes factors. This means that there are $2T$ samples generated and $2T$ logistic regressions computed per iteration, since both hypotheses need to be tested. If both thresholds are reached, and the sample size has converged, the probabilities of reaching the thresholds (p_1 and p_c) as well as the final N are outputted and the algorithm concludes.

Parameters

In Table 18, all input parameters for each model are displayed, along with the values that are simulated in this study.

Table 1: Simulation parameters. Shared parameters are used by both methods.

Parameter	**Description**	**Values**
A priori		
η	Posterior probability threshold	0.8, 0.9
T	Simulations per step	10000
Sequential		
N_{step}	Sample size increase per step	1, 4
$\texttt{nr}_{\texttt{it}}$	Number of iterations	5000
Shared		
N_{min}	Initial sample size	10, 50
N_{max}	Maximum sample size	100, 1000
$\text{AAFBF}_{\text{threshold}}$	Evidence threshold	5, 10
β_1	Effect size	0, 1
α	Intercept	-1
δ	Parameter range constraint	0.3
Hypothesis	Type of test	superiority, non-inferiority, equivalence

Most variables have been discussed earlier in the methods section, but $\texttt{nr}_{\texttt{it}}$ has not been named before. It determines how many times a parameter set (the collection of parameters used in an iteration) is simulated. Following an example set by Moerbeek (2021), it is set at 5000. It roughly corresponds to the parameter T for the a priori method.

The thresholds chosen for the AAFBF were set at 5 and 10 to emulate different cases in social science research. The threshold may be set at 5 in the situation where the stakes are quite low, but the researcher would like some indication as to which hypothesis is more likely given the data. This could be the case for a pilot study, for example.

An AAFBF of 10 reflects the situation in which a hypotheses receives ten times more support than its complement. This threshold was chosen to reflect the situation in which a researcher would like quite strong evidence that one hypothesis should be preferred.

The effect size of the dichotomous predictor β_1 is constant at 1. This reflects an effect size that is realistic, and for the purposes of this study well balanced between being moderate and findable. An example of an effect size like this for a dichotomous predictor can be found in Duplaga (2017), where it is the effect of making use of a mobile phone on the probability of using e-mail services on the Internet (though there were also other predictors involved there). There was an odds ratio of 2.78 reported, which corresponds to a coefficient of $ln(2.78) = 1.02$.

The intercept α was chosen to be -1, which corresponds to a baseline predicted probability of 0.27. Again, this value was chosen to reflect a real-world analysis. In another logistic regression in Duplaga (2017), the intercept has an odds ratio of 0.374, which would output a β of -0.98. The values for the logistic regression intercept and coefficient thus represent the situation in which the outcome variable is more likely to be 0, but the predictor variable has a positive effect.

The range constraint δ is relevant for the non-inferiority and equivalence test designs. In the non-inferiority case, it specifies the hypothesis that β should be greater than δ. The value for δ here is 0.3. Though a non-inferiority trial is often used to test if a parameter is greater than a certain negative value, it can just as well be used to test if it is above a positive value. Given that the value for β here is 1, it makes sense to test for non-inferiority of +0.3. Testing for -0.3 would be a less interesting case, since that expands the valid parameter range instead of restricting it when compared to the superiority trial (where $\delta = 0$ implicitly).
For the equivalence trial, the value of 0.3 means that the hypothesis is that $\beta \in [-0.3, 0.3]$. With a true β of 1, this means that this range is in fact misspecified, as the greatest mass of the posterior distribution will most likely fall around 1. That allows for a scenario in which the hypothesis might not be supported by the data to the extent that the superiority and non-inferiority trials are expected to be supported.

Furthermore, it should be noted that the N_{min} and N_{max} sample sizes are pairs: if the minimum sample size is 10, the maximum is 100. The same goes for 50 and 1000. This was done to reduce computation time, but still allow for larger sample sizes to be simulated. Apart from that, the simulation was set up such that every combination of variables formed a parameter set. The unvarying parameters (such as α) act as constants. This leads to 48 parameter sets for both algorithms.

The input parameters are known beforehand, but both methods output the variables of interest. This includes mainly the final N and Bayes factor, but others such as fail counts due to lack of variance will be considered as well. These will be analyzed in the next section, mainly through descriptive statistics and visual means.

Results

This section presents the results of the simulation study conducted using binary logistic regression. The goal was to evaluate and compare two Bayesian sample size determination methods: a sequential and an a priori design. To briefly reiterate the research questions:

1. How often do the sequential and a priori sample size determination methods correctly show support for the hypothesis that was used to generate the data?

2. To what extent do both methods return similar sample sizes under similar conditions?

The primary output variables are sample size and the Bayes factor. Counts of failed iterations due to samples without variance will also be reported. The simulated data as well as analysis scripts are publicly available on https://github.com/coenwil/bachthesis.

This section is structured as follows: first, the stability of the sequential design will be analyzed. Second, the performance of both methods in correctly supporting the data-generating hypothesis will be assessed. Finally, the sample sizes generated by each method will be compared across similar input parameters.

Sequential design

Stability of the results

Since the sequential method involves iterating 5000 times over the same parameter set, it is important to assess to what extent those iterations are similar. Relatively stable results imply a reliable and replicable method, which is necessary if the method is to be used in practice. Stability also strengthens the results of this simulation study, by reducing the likelihood that the patterns are due to random chance.

The stability of the output variables will be measured in coefficients of variation

(CV), which is the standard deviation divided by the mean. A lower CV implies a more stable variable, since the standard deviation is proportionally smaller. This measure was chosen because it scale independent, unlike the standard deviation. This means that variables with different units can be meaningfully compared, such as the Bayes factor and sample size.

The iterations were grouped by parameter set, and the coefficient of variation was calculated for each group. This results in 48 CV values per output variable, which is too many to present clearly in a table. As a summary, Table 2 reports the mean and standard deviation of the CVs for the sample size and Bayes factor. The CV values for each parameter set are included in a table in Appendix TODO though.

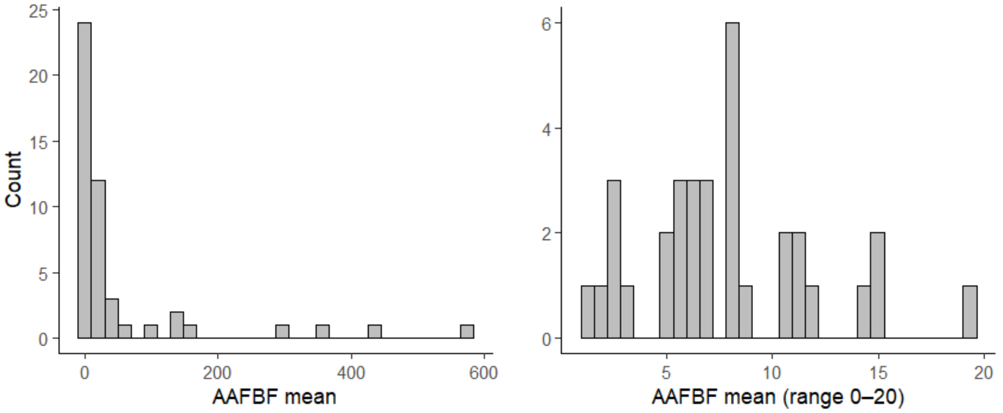
Table 2: Coefficient of variation (CV) summary statistics.

Variable	**CV mean**	**CV std. dev.**
N	0.18	0.13
AAFBF	2.96	2.59

The CV mean is highest for the Bayes factor, at 2.96. This implies that the standard deviation for the Bayes factor was on average 3 times larger than its mean. However, the mean for the CV of the sample size is still relatively low, which indicates that despite the high variability in the AAFBF, the actual sample sizes were quite stable.

A potential factor explaining the difference is that the Bayes factor is unbounded, while the sample size is constrained to be within N_{min} and N_{max}. An extreme amount of evidence can lead to an extreme value for the Bayes factor, but only result in a constrained sample size. Within iterations of the same parameter set, the AAFBF can therefore vary quite a bit while still resulting in a similar sample size. The distribution of AAFBF means (grouped by parameter set) is plotted in Figure 4 to allow for a visual investigation.

Figure 4: Distribution of grouped AAFBF means (full range and constrained to 0-20).



It can be seen that there are clear outliers, with the maximum value reaching a Bayes factor of 576. This occurred in a parameter set testing for superiority with a β_1 of 1 and a minimum sample size of 50. It makes sense that the combination of an easily satisfied hypothesis with a large initial sample size would lead to a large Bayes factor. Exemplary to the hypothesized explanation, this iteration had a CV of 5.24 for the Bayes factor yet a CV of just 0.01 for the sample size. The Pearson correlation between the mean of the Bayes factor and its standard deviation is $r = 0.95$, indicating that mostly the iterations with higher Bayes factors account for the higher variation. Given that the sample size remains stable with a low CV, the increased variation for the higher sample sizes appears inconsequential.

Across all 240 000 total simulations, there were 6928 counts of sampling without variance, which corresponds to about 2.9%. This roughly corresponds to the probability as estimated in the methods section, which was 4.5%. The probability is lower here, most likely due to the fact that half of the simulations were done with a starting sample size of 50, where the probability of sampling without variance is virtually 0.

Overall, it seems that there is no problematic instability in the simulations, and the values are most likely reliable as well as replicable. In the sections that follow, each parameter set (that has been iterated over 5000 times) will be regarded as its own data point, computed by taking the mean of its outcome variables. The sample size that it predicts will thus be the average across 5000 iterations, as well as for example the Bayes factor.

Correct support

Of the 48 parameter sets, 24 were 'correct'. That is, they found enough support in the data ($AAFBF \geq AAFBF_{\text{threshold}}$) for the hypothesis that aligned with the actual β_1. If the regression coefficient was 1, then the superiority and non-inferiority trials would be correct. If it was 0, then the equivalence trials would be correct.
Only half of the simulations being correct might seem like a small proportion. However, there were only 24 parameter sets that had a regression coefficient that aligned with the trial to begin with, and in those 24 cases the support was correctly found. So, the method actually performed very well when the trial was aligned with the generated data.

Concerning the 24 parameter sets that did not have an aligned generating coefficient and trial design, further inspection reveals that 12 were inconclusive results where β_1 did not align with the hypothesis. The inconclusive results were most often produced by the non-inferiority trial and least often for the superiority trial (6 and 2 instances, respectively). This is the desired behavior of this method.
However, that leaves 12 cases where the method had a conclusive Bayes factor while the coefficient was not in agreement with the hypothesis. This happened the most for the superiority trial, namely 6 times. This means that the Bayes factor was above the threshold while there was actually no effect in the data. It happened the least for the inferiority trial at 2 instances. It makes sense that it happened the most with the superiority trial, since it tests up to the border of how the data was actually generated; $H_1 : \beta_1 > 0$ while $\beta_1 = 0$. The non-inferiority trial has a 'buffer zone' of δ, which helped reduce the number of conclusive yet wrong iterations to 2. Across the trial designs, it happened about equally often for smaller or larger starting sample sizes.
The N_{step} input variable did not create any meaningful difference in the outcomes.

Overall, the sequential design was always able to find enough support in the data for the hypothesis at hand if the data were accordingly generated. However, if the hypothesis did not align with the generated data, the sequential design was inconclusive 12 out of 24 times, and the other 12 times it found undue support for the hypothesis at hand. This means that in 36 out of 48 cases it produced the desirable result.

A priori design

Since the a priori method iterates 10000 times at each sample size step, it is not necessary to also iterate over each parameter set as with the sequential design. Therefore, the stability of the results can be skipped here, as it is implicitly checked by the algorithm.

The sampling without variance in the apriori method was recorded different to the sequential design. As described in Algorithm REF ALGO2, there is only action taken if $\geq 5\%$ of the 10 000 iterations at a given sample size were missing due to being sampled without any variance. This happened in 19 out of 48 parameter sets, though it should be noted that this happened maximally once per set. Therefore, the overall prevalence should still be quite low, and should not have influenced the results in any meaningful way.

Conclusive results

For the a priori method, a conclusive result means getting both p_{H_1} and p_{H_c} above η, because only then can the sample size be anything less than the maximal sample size. A conclusive result implies a correct result in this case, since p_{H_1} and p_{H_c} test for the hypotheses at hand. However, it could be the case that the method is inconclusive while the data was generated according to the coefficient that is tested for.

In total, 15 out of 48 trials got a conclusive result. As with the sequential design, there were only 24 cases where the data was aligned with the hypothesis, so the proportion of conclusively correct trials was actually 15 out of 24. In those 9 inconclusive cases, there were 4 equivalence trials, 3 non-inferiority trials, and 2 superiority trials. It would therefore seem that more specific hypotheses have a higher likelihood of not getting enough support in the data.

Further splitting up the data reveals that when the sample size was large ($N_{\text{min}} = 50$ and $N_{\text{max}} = 1000$) and the generated data was aligned with the hypothesis, 11 out of 12 times the algorithm produced conclusive results. The one time it failed, was with an equivalence test with $\eta = 0.9$. More specifically, it had an estimated probability of getting enough support for the alternative hypothesis (p_{H_1}) of 0.997,

but the support against the complement lagged behind at 0.457.
Obviously, this leaves 4 out of 12 conclusive results for ($N_{\text{min}} = 10$ and $N_{\text{max}} = 100$). Half of these conclusive results were from superiority trials. When the equivalence and non-inferiority trials got conclusive results, they also had the $\eta = 0.8$, the lower value. At the same time, it is never the case that a parameter set that did not satisfy $\eta = 0.9$ did in fact satisfy $\eta = 0.8$, so the impact of that variable on the success rate was not substantial in this simulation study.

Similar results under similar conditions

While the input parameters for both methods differ slightly, they are similar enough to provide a basis for meaningful difference. The N_{step} variable in the sequential design is not shared with the a priori method, but it was largely inconsequential, so it can safely be ignored. Conversely, the η variable for the a priori algorithm does not exist in the sequential parameter sets, but it did not have a meaningful impact on the results for larger sample sizes or smaller sample sizes. Therefore, the focus can be on the shared input parameters.

To begin to understand the sample sizes given by both methods, only the conclusively correct results are taken here. The mean sample size provided by the sequential design was 41.2, while the a priori method provided a mean sample size of 182.9. It should be noted that this combines the larger and smaller sample sizes. Splitting those up for the sequential design returns a mean sample size of 27.4 for the smaller range, while the larger sample size range had a mean sample size of 54.9. The a priori method gave a mean sample size of 93.5 for the smaller range, but a mean value of 215.4 for the larger sample sizes.
These differ quite significantly, and the range of sample sizes seems to be especially important for the a priori method. These results, along with some others, are displayed in Table 3.

Table 3: Mean predicted sample sizes by method, sample size range, Bayes factor threshold, and hypothesis type

Method	**Total**	$\boldsymbol{N}_{\text{max}}$=100	$\boldsymbol{N}_{\text{max}}$=1000	$\boldsymbol{BF}_{\text{thr}}$=5	$\boldsymbol{BF}_{\text{thr}}$=10	**Sup.**	**Non-inf.**	**Equiv.**
Sequential	41.2	27.4	54.9	37.1	45.2	36.0	37.4	50.1
A priori	182.9	93.5	215.4	114.7	285.2	120.2	166.6	297.3

It should be noted that in this table, the values as displayed in the columns are fixed while the rest is free to vary. For example, the mean sample size given by the sequential method for the superiority test is 36, without further conditioning on e.g. sample size range. In any case, the between method variance is quite a bit greater than the within-method variance. The parameters that seem to have the most influence on the results are the sample size range, the Bayes factor threshold for the a priori design, and whether or not it concerned an equivalence trial.

Recalling the results of each method, the sequential design successfully supported the hypothesis in every case where the data aligned with it. However, it also wrongly found support for the hypothesis 12 times, when the data did not warrant that. The a priori method was able to find correct support 15 out of 24 times, which was improved to 11 out of 12 when only larger sample size ranges were considered. Notably, it did not provide undue support for a hypothesis that was not reflected in the data. It can therefore be posited that the sequential method achieved a perfect true positive rate (1.0), but with a false positive rate of 0.5. In contrast, the a priori method had a true positive rate of 0.63 overall — or 0.92 when restricted to more suitable parameter sets — and no false positives.

While the a priori method appears to require, very roughly, about three times the sample size of the sequential design, the influence of the parameter values seems broadly consistent across both methods. Finally, it seems to be the case that the a priori method does require a larger sample size range, but usually does not output sample sizes that are close to the maximum sample size. This is an important point for practical applications of this method, and will be discussed more in depth in the conclusion.

Conclusion

Having explored the performance of the two Bayesian sample size determination methods, this section now assesses the broader conclusions that can be drawn from the results.

Both algorithms necessarily have a different approach to picking a certain sample size, and what it means for that sample size to provide enough support for the hypothesis at hand. In the sequential design, the input data are the actual, observed data points. Meanwhile, in the a priori design, the input data are simulated, with the desired effect size as the coefficient.

With this in mind, it becomes clear why the sequential method might accept a certain sample size just providing a sufficient Bayes factor once as enough support, but the a priori method requires simulating a given sample size 10 000 times and checking if the proportion of iterations having a sufficient Bayes factor is at least η.

The outcome of this difference is that the sequential design generally outputs much lower sample sizes, and struggles less often with detecting an effect. At the same time, the a priori method seems to be much more resistant to finding support for a hypothesis that is not warranted by the data.

Should there be a preference for either one? It depends on the circumstances. First of all, one of the methods might not even be accessible for a given research design. For example, a longitudinal study that runs for, say, 10 years. It is probably not feasible to start with a low sample size and increase the sample if after 10 years it turns out that the sample size was too low.

Even when assuming that both methods are an option, which one should be picked still depends on circumstance. A researcher working on a high-stakes study (for example, a clinical trial for a new type of medicine with potential side effects) might look at the 0.5 false positive rate of the sequential design and reject using it as an option. However, they could use a non-inferiority trial with a δ that is far away from 0, so that the likelihood of the method falsely detecting an effect is reduced. They could also increase the Bayes factor threshold, for example.

At the same time, a researcher who is not in a position to use the sequential design might be opposed to the relatively high false negative rate of the a priori design, especially when sample sizes are low. They could then feel that neither method would be appropriate for them. Similarly, they could adjust η to be even lower than 0.80, or set the Bayes factor threshold lower, or even only check for sufficient support for H_1 and not check for support against the complement.

One might argue that the sequential design has a conceptual advantage, since it is based on observed rather than simulated values. That epistemological line of reasoning does seem to be valid. For example, Wagenmakers (2007) contrasts the Bayesian and frequentist approaches, and one of the advantages is argued to be the fact that the Bayesian method only relies on observed data. If the use of observed data is indeed valuable for a Bayesian, then this could reveal a principled reason to prefer the a priori method. Of course, there are counterarguments: simulated data is different from hypothesized data under H_0 and not necessarily antithetical to the Bayesian approach.

Regarding practical recommendations based on the results from this study, one of the key findings is that the a priori method more or less requires a sufficiently high maximum sample size. It might not actually output that size, but it is needed to properly search the sample size space. With a logistic regression coefficient of 1, the maximum sample size needed to be much higher than 100. The value 1000 here seems to work much better, but most likely there are more optimal choices.

Furthermore, both methods generally get the same influence from input parameter changes such as a higher required Bayes factor. This implies a potential for both methods to be used in tandem, with the same shared input parameters. In fact, that is what Fu et al. (2020) suggests in their simulation study of the t-test and Welch's test. That suggestion is taken over here. The most optimal algorithm is most likely one that takes some minimum effect size necessary for the study to be worthwhile, computes a sample size that has a probability of η of reaching the desired Bayes factor, and then transitions into a sequential design when data is actually observed.

Lastly, if the informative hypothesis is inequality constrained, it may be advisable to do a non-inferiority trial where δ is defined to lie just slightly beyond the parameter value of interest. This can help prevent support for the complement hypothesis when such support is not warranted, especially when the sequential design is used.

Discussion

Though this simulation study provided new insights in Bayesian sample size determination, there remain some questions. Expanding the parameter set that is simulated over would be valuable. For example, it would be interesting to analyze more moderate logistic regression coefficients, such as 0.5. Other intercepts could also have been analyzed. A Bayes factor threshold of 10 may also not be entirely accurate for high-stakes research, and one might actually want to have a threshold of 20, for example.

A key parameter for the a priori method is T. It determines how many times each sample size is iterated over, and as such is the main driver behind the long computation times. Further research could perhaps try some lower values of T, like 5000 or 1000, and see how stable the simulations are. If they are indeed stable, that could be an argument to recommend a lower T for the sake of computation.

Another suggestion for the a priori algorithm would be to test its performance without the second check for evidence against the complement (the second term in $(p_1 \wedge p_c) > \eta$). It could be the case that this improves the performance of the a priori design, though not necessarily so. A simulation study would be an excellent method to test that.

Lastly, this is also a call to extend Bayesian sample design methodology to more complex statistical models, such as multivariate logistic regression, but e.g. linear regression and multilevel models as well, all of which are extensively used in social science research.

References

Adcock, C. J. (1997). Sample size determination: A review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(2), 261–283. https://doi.org/10.1111/1467-9884.00082

Béland, S., Klugkist, I., Raîche, G., & Magis, D. (2012). A short introduction into bayesian evaluation of informative hypotheses as an alternative to exploratory comparisons of multiple group means. *Tutorials in Quantitative Methods for Psychology*, *8*(2), 122–126. https://doi.org/10.20982/tqmp.08.2.p122

Berger, J. O., & Pericchi, L. R. (2004). Training samples in objective bayesian model selection. *The Annals of Statistics*, *32*(3). https://doi.org/10.1214/009053604000000229

Brutti, P., De Santis, F., & Gubbiotti, S. (2008). Robust bayesian sample size determination in clinical trials. *Statistics in Medicine*, *27*(13), 2290–2306. https://doi.org/10.1002/sim.3175

Cohen, J. (1988, May). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. http://www.jstor.org/stable/20182143

Cohen, J. (1994). The earth is round (p ¡ .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

Cortegiani, A., Longhini, F., Madotto, F., Groff, P., Scala, R., Crimi, C., Carlucci, A., Bruni, A., Garofalo, E., Raineri, S. M., Tonelli, R., Comellini, V., Lupia, E., Vetrugno, L., Clini, E., Giarratano, A., Nava, S., Navalesi, P., & Gregoretti, C. (2020). High flow nasal therapy versus noninvasive ventilation as initial ventilatory strategy in copd exacerbation: A multicenter non-inferiority randomized trial. *Critical Care*, *24*(1). https://doi.org/10.1186/s13054-020-03409-0

Del Prette, Z. A. P., Prette, A. D., De Oliveira, L. A., Gresham, F. M., & Vance, M. J. (2012). Role of social performance in predicting learning problems: Prediction of risk using logistic regression analysis. *School Psychology International*, *33*(6), 615–630. https://doi.org/10.1177/0020715211430373

Duplaga, M. (2017). Digital divide among people with disabilities: Analysis of data from a nationwide study for determinants of internet use and activities performed online (Y.-K. Jan, Ed.). *PLOS ONE*, *12*(6), e0179825. https://doi.org/10.1371/journal.pone.0179825

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fu, Q., Hoijtink, H., & Moerbeek, M. (2020). Sample-size determination for the bayesian t test and welch's test using the approximate adjusted fractional bayes factor. *Behavior Research Methods*, *53*(1), 139–152. https://doi.org/10.3758/s13428-020-01408-1

Fu, Q., Moerbeek, M., & Hoijtink, H. (2022). Sample size determination for bayesian anovas with informative hypotheses. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.947768

Fukai, S., Mizusawa, Y., Noda, H., Tsujinaka, S., Maeda, Y., Hasebe, R., Eguchi, Y., Kanemitsu, R., Matsuzawa, N., Abe, I., Endo, Y., Fukui, T., Takayama, Y., Ichida, K., Inoue, K., Muto, Y., Watanabe, F., Futsuhara, K., Miyakura, Y., & Rikiyama, T. (2024). Superiority trial for the development of an ideal method for the closure of midline abdominal wall incisions to reduce the incidence of wound complications after elective gastroenterological surgery: Study protocol for a randomized controlled trial. *Trials*, *25*(1). https://doi.org/10.1186/s13063-024-08167-w

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511–527. https://doi.org/10.1037/met0000017

Gu, X., Mulder, J., & Hoijtink, H. (2017). Approximated adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261. https://doi.org/10.1111/bmsp.12110

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989x.10.4.477

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193. https://doi.org/10.1198/000313001317098149

Moerbeek, M. (2021). Bayesian updating: Increasing sample size during the course of a study. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01334-6

Ockene, J. K., Kristeller, J., Goldberg, R., Amick, T. L., Pekow, P. S., Hosmer, D., Quirk, M., & Kalan, K. (1991). Increasing the efficacy of physician-delivered smoking interventions: A randomized clinical trial. *Journal of General Internal Medicine*, *6*(1), 1–8. https://doi.org/10.1007/bf02599381

Palfi, B., & Dienes, Z. (2020). Why bayesian “evidence for h1” in one condition and bayesian “evidence for h0” in another condition does not mean good-enough bayesian evidence for a difference between the conditions. *Advances in Methods and Practices in Psychological Science*, *3*(3), 300–308. https://doi.org/10.1177/2515245920913019

Pawel, S., & Held, L. (2025). Closed-form power and sample size calculations for bayes factors. *The American Statistician*, 1–15. https://doi.org/10.1080/00031305.2025.2467919

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

St-Jules, D. E., Goldfarb, D. S., & Sevick, M. A. (2016). Nutrient non-equivalence: Does restricting high-potassium plant foods help to prevent hyperkalemia in hemodialysis patients? *Journal of Renal Nutrition*, *26*(5), 282–287. https://doi.org/10.1053/j.jrn.2016.02.005

Tendeiro, J. N., Kiers, H. A. L., Hoekstra, R., Wong, T. K., & Morey, R. D. (2024). Diagnosing the misuse of the bayes factor in applied research. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231213371

van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. A. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0699-7

van der Linden, S., & Chryst, B. (2017). No need for bayes factors: A fully bayesian evidence synthesis. *Frontiers in Applied Mathematics and Statistics*, *3*. https://doi.org/10.3389/fams.2017.00012

Van Lissa, C. J., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., & Hoijtink, H. (2020). Teacher's corner: Evaluating informative hypotheses using the bayes factor in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 292–301. https://doi.org/10.1080/10705511.2020.1745644

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin; Review*, *14*(5), 779–804. https://doi.org/10.3758/bf03194105

Wang, F., & Gelfand, A. E. (2002). A Simulation-Based Approach to Bayesian Sample Size Determination for Performance under a Given Model and for Separating Models. *Statistical Science*, *17*(2), 193–208. Retrieved June 23, 2025, from https://www.jstor.org/stable/3182824

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials*. Springer International Publishing. https://doi.org/10.1007/978-3-319-32562-0

Wilson, K. J., Williamson, S. F., Allen, A. J., Williams, C. J., Hellyer, T. P., & Lendrem, B. C. (2022). Bayesian sample size determination for diagnostic accuracy studies. *Statistics in Medicine*, *41*(15), 2908–2922. https://doi.org/10.1002/sim.9393

Zardo, P., & Collie, A. (2014). Predicting research use in a public health policy environment: Results of a logistic regression analysis. *Implementation Science*, *9*(1). https://doi.org/10.1186/s13012-014-0142-8

Zhang, X., Cutter, G., & Belin, T. (2011). Bayesian sample size determination under hypothesis tests. *Contemporary Clinical Trials*, *32*(3), 393–398. https://doi.org/10.1016/j.cct.2010.12.012