BAYESIAN SAMPLE SIZE CALCULATIONS FOR COHORT AND CASE-CONTROL STUDIES

by

Cyr Emile M'lan

Department of Mathematics and Statistics
McGill University, Montreal

July 2002

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-88519-4
Our file Notre référence
ISBN: 0-612-88519-4

Abstract

Sample size determination is one of the most important statistical issues in the early stages of any investigation that anticipates statistical analyses.

In this thesis, we examine Bayesian sample size determination methodology for interval estimation. Four major epidemiological study designs, cohort, case-control, cross-sectional and matched pair are the focus. We study three Bayesian sample size criteria: the **average length criterion** (**ALC**), the **average coverage criterion** (**ACC**) and the **worst outcome criterion** (**WOC**) as well as various extensions of these criteria. In addition, a simple cost function is included as part of our sample size calculations for cohort and case-controls studies. We also examine the important design issue of the choice of the optimal ratio of controls per case in case-control settings or non-exposed to exposed in cohort settings.

The main difficulties with Bayesian sample size calculation problems are often at the computational level. Thus, this thesis is concerned, to a considerable extent, with presenting sample size methods that are computationally efficient.

Résumé

La détermination de la taille de l'échantillon est l'une des questions les plus importantes dans les phases premières de toute étude qui prévoit des analyses statistiques.

Dans cette thèse, nous examinons des méthodologies bayésiennes de calcul de taille d'échantillon pour estimation d'intervalles. Quatre études épidémiologiques majeures, à savoir les études de cohortes, les études de cas-témoins, les études de données transversales, et les études de données pairées forment le point d'intérêt principal. Nous étudions trois critères bayésiens de taille d'échantillon: le critère de longueur moyenne (**ALC**), le critère de recouvrement moyen (**ACC**), et le critère du plus mauvais scenario (**WOC**) aussi bien que divers prolongements de ces derniers. En outre, une fonction de coût simple est incluse en tant qu'élément intégrant dans nos calculs de taille d'échantillon pour les études de cohortes et les études de cas-témoins. Notre protocole d'étude addresse également l'importante question du choix du rapport optimal du nombre de sujets témoins pour chaque sujet cas dans les études de cas-témoins ou du nombre de sujets non-exposés pour chaque sujet exposé dans les études de cohortes.

Les difficultés principales avec les méthodes bayésiennes de détermination

de taille d'échantillon réside essentiellement au niveau calculatoire. Ainsi, une bonne partie de cette thèse porte sur la présentation de méthodes de calcul de taille d'échantillon numériquement efficace.

Acknowledgments

I would first like to offer my deepest gratitude to my co-supervisors Professor David Wolfson and Professor Lawrence Joseph, for their many suggestions, guidance, and advice. Their constant support, both moral and financial has helped me considerably to elevate my knowledge of statistical theory and computations. From the innermost recesses of my soul, I know I owe them a lot for what I have become today and they will always be my models.

I would like to thank Professor Keith Worsley and Professor Masoud Asgharian, for valuable discussions and encouragements that I benefited from. Many thanks also to Christina Wolfson for giving me a chance to work on an important project about reevaluating the effect of length-bias on the survival of patients after the onset of dementia.

I am also thankful to all my close-friends, including Dembélé Lassina and Traoré Issouf for their true friendship. Thanks to my office-mate and sport-partner Jonathan Taylor, for giving me a a unique opportunity to celebrate a Thanksgiving day with his family and also for showing me the many faces of Montreal. For making my stay at McGill University so delightful, I would like to thanks the personnel, the professors, and all the students in

the department.

At last, I wish my parents were here. I missed them all the time, even in joyful days. Time goes by but I always think of them as if I left them yesterday. Jess, your daddy loves you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Epidemiology has become an indispensable tool for both clinical research and the study of public health. Epidemiology contributes to the understanding of how disease transmission and pathogenesis relates to environmental or other agents. Epidemiological designs are also used to identify gene-drug interactions, unrecognized drug hazards in the fields of pharmacogenetics and pharmacovigilance (Weiss et al., 2001; Ashby et al., 1998), and many other areas. Case-control, cohort, cross-sectional and matched pairs studies are well-established epidemiological designs for investigating these and other diverse issues, hundreds of such epidemiological studies are carried out each year around the world. Rothman and Greenland, 1998, provide a modern overview of epidemiology.

In order for epidemiological studies to best serve their purposes, they need to be carefully designed. One of the key components for a successful design is the determination of the appropriate sample size, that is, the number of subjects that need to be observed in order to ensure sufficient precision in the estimation of important parameters. Sample size estimation is almost always a difficult problem, not only for technical or computational reasons, but also because the choice often depends on the context of the proposed study, and could take into account both practical and ethical consideration, in addition to those of statistical precision. Lenth, 2001, discusses practical guidelines for effective sample size determination. Further, the solution can take several forms, depending on the specific criterion used. For example, one can consider Bayesian or frequentist criteria, and within each paradigm one must also choose the main outcome parameters, and whether the criterion is based on interval width, power, or some other criterion. In all cases, sample size determination requires a careful gathering of prior information from experts.

In this thesis, we develop sample size methods for four of the most commonly used epidemiological designs, including case-control, cohort, matched pairs, and cross-sectional designs, in the context of 2 × 2 contingency tables. Traditionally, sample size problems have been presented in anticipation of a statistical analysis centred on hypothesis testing. In recent years, however, there has been a dramatic shift in philosophy towards interval estimation in the epidemiology literature (O'Neill, 1984, Gardner and Altman, 1986, Goodman and Berlin, 1994, Hoenig and Heisey, 2001, Cesana et al., 2001). Almost all statistical analyses now include interval estimates of the parameters of interest. Now, clearly, sample size calculations should be based on a criterion that, in spirit, matches the eventual analysis. Thus, for example, it is well known that sample sizes provided by power calculations are often too small to guarantee that the parameter of interest is precisely estimated. In this thesis, we therefore develop sample size methodology for interval estimation. The main idea underlying this methodology is that sample sizes are determined on the basis of expected confidence-interval width in the frequentist paradigm, and expected credible-interval width or coverage, in the Bayesian framework.

For reasons that we discuss fully in chapter 2, this thesis takes a Bayesian approach to sample size determination. From this viewpoint, a prior distribution is formed about the unknown parameters, which summarizes the information available from past studies or expert opinion. Bayesian approaches are natural in epidemiology because information collected from past data is often available. Explicit Bayesian sample size calculations for cohort, case-control, cross-sectional, and matched pairs studies have not been previously addressed in the literature, although various Bayesian analyses for the 2×2 tables are presented in Aitchison and Bacon-Shone, 1981, Latorre, 1982, Nurminen and Mutanen, 1987, Zelen and Parker, 1986, Marshall, 1988, Franck et al., 1988, Carlin, 1992, Ashby et al., 1993, Walters, 1997, and Hashemi et al., 1997. In this thesis, we find sample sizes for all of the above designs under five different criteria. Although, some of these *criteria* have been previously defined, others are new, and all are applied, for the first time in this thesis, to the estimation of risk ratios and odds ratios in above settings.

As in much recent work on Bayesian analysis, the statement of the problem, i.e., the statement of how each criterion should be applied to each study design, is relatively straightforward. On the other hand, the practical application, meaning the actual calculation of the sample sizes required for each combination of design and sample size criterion, is usually where the main challenges lie. While exact results are occasionally available and are derived in this thesis, we often resort to combinations of analytic and numerical approximations.

Our principal contributions to sample size calculation methodology are found in Chapters 3 and 4. The focus of this thesis, presented in Chapter 4, is sample size determination for 2×2 tables arising in the context of the two sample problem (i.e., finding sample sizes for both exposed and unexposed subjects in cohort studies, or for both case and control groups in case-control studies). Nevertheless, in Chapter 3, we first consider sample size methods for a simpler one sample problem, where we assume some parameters are a priori exactly known. We begin with the one sample problem because the methods developed for the two sample problem are, for the most part, adaptations of those we develop for the one sample problem.

In Chapter 2, we first review the definitions of cohort, case-control, cross-sectional, and matched pairs studies. We then review frequentist and Bayesian analyses for the parameters of interest for these four designs. We next present a literature review of both frequentist and Bayesian sample size methodologies, accompanied by a critical discussion of the various choices of statistical criterion functions. Chapter 2 ends with a collection of other results useful for Chapters 3 and 4.

In Chapter 3, we develop exact methods for Bayesian sample size determination for exposure- and case-only studies for five different Bayesian interval based sample size criteria. Exposure- and case-only studies are special cases of cohort and case-control studies, respectively, where one only needs to collect information on exposed and diseased subjects, respectively. Since exact methods are not always feasible, and can often be accurately and efficiently replaced by third order approximations, we derive third order approximations to the criterion functions for each setting. We also provide explicit plug-in sample size formulae for these designs. Aside from exact and third order approximations, we also use Monte Carlo simulations to estimate the criterion functions. We compare these three methods, and show that they yield slightly different sample sizes.

Chapter 4 is devoted to the two sample problem for cohort, case-control, cross-sectional and matched pairs designs. We derive sample size formulae for these four epidemiological designs. Here, since exact methods are not possible, we mainly rely on Monte Carlo simulation approaches to sample size calculation. For this reason, the steps within each Monte Carlo simulation algorithm are described carefully.

In summary, in this thesis we present, for the first time, methods for Bayesian sample size determination for the most common epidemiologic designs. We consider a wide variety of criterion functions, and place emphasis on ease of implementation. At the same time we do not hide practical difficulties, where they arise.

Chapter 2

Literature review of the design and analysis of case-control and cohort studies

Most studies in epidemiology can be classified into one of two types, experimental or observational.

- Almost all experimental studies include randomization or other manipulation of treatments or exposures by researchers. Clinical trials are examples where one randomly assigns treatments to subjects. Such trials are mainly carried out to quantify the effectiveness of new approaches to treatment and prevention.

- In observational studies exposure to putative risk factors or treatments occurs "naturally", without direct manipulation by researchers. Cohort and case-control studies are usually examples of observational studies, although some clinical trials can also be cohort studies.

Cohort and case-control studies can be further subdivided by their relation to time:

- A prospective study is one in which exposure and all or most covariates are measured before disease manifests, with ascertainment of disease status occurring during a future follow-up period.
- A retrospective investigation is one in which exposure is determined after the identification of disease status.

Increasingly, studies are relying on a combination of both prospective and retrospective elements to gather information. A third type of non-randomized design is the cross-sectional study, in which we view a snapshot of a well identified population at a certain point in time. Both diseases and exposure information (for example, drug use or chemical exposure) are collected simultaneously for each study subject, so that different exposure subpopulations may be compared with respect to their disease prevalence. It is, however difficult to ascertain a temporal relationship between the exposure and the onset of disease in cross-sectional studies. Randomized trials, while relatively free of bias compared to observational studies, are accompanied by problems of medical ethics, feasibility, cost, and logistical considerations. These limitations have turned epidemiologists toward observational cohort and case-control studies for the establishment of etiological relationships, despite their potential for producing biased results. Cohort and case-control studies are the two most well-established and documented observational epidemiologic designs for studying exposure-disease relationships (Breslow and Day, 1980; Breslow and Day, 1988; Schlesselman, 1982; Rothman, 1986; Gordis, 1996; Rothman and Greenland, 1998; Rosenberg et al., 2000). We define these designs more formally in section 2.2, as they represent the main focus of the sample size methods developed for this thesis.

In section 2.1, we set forth some quantitative measures of disease and exposure frequency, and association between disease and exposure. In section 2.2, we briefly review cohort and case-control studies. Section 2.3 is devoted to an overview of frequentist analysis of risk ratios and odds ratios while section 2.4 describes a Bayesian counterpart. We finally come to the problem of sample size methodology in section 2.5. Section 2.6 contains a summary while section 2.7 presents a variety of preliminary results needed previously appearing in the literature which are essential to the rest of the thesis.

2.1 Measures of disease occurrence and the association between disease and exposures

Measures of disease frequency are the building blocks of any epidemiological investigation. They tell us how a disease is related to exposure in a given population. These measures, especially the comparative measures of the relative risk and the odds ratio, are the main parameters of interest in this thesis. Sample size derivation for case-control and cohort studies, the main topic of this thesis, will be based on these measures of association.

2.1.1 Incidence and prevalence

Disease risk is principally expressed through the **incidence proportion** or simply the **risk** $p_{(t_0,t_1)}$ $(0 \leq p_{(t_0,t_1)} \leq 1)$, i.e., the fraction of new cases that occur during a specified period of time (t_0, t_1) in a population at risk for developing a disease. More precisely, let T be the random variable describing the time at which an individual develops a disease or a change in health status. Then

$$p_{(t_0,t_1)} \equiv \Pr(T \in (t_0, t_1) | \text{ individual does not die from any other cause in } (t_0, t_1)). \tag{2.1.1}$$

Assuming exchangeability of the subjects in the population (see Gelman et al., 1995), disease outcomes for individuals in the interval (t_0, t_1) could be thought of as i.i.d. random events conditional on $p_{(t_0,t_1)}$. When the time period is clear, we will denote the parameter $p_{(t_0,t_1)}$ by p.

The second measure of disease occurrence is the prevalence. The **prevalence** of a disease, which we denote by p', is the proportion of disease cases in a population at a given point in time or during a well defined time period, irrespective of the disease onset time. Prevalence is an important and useful measure of the burden of disease in a population, and is the most frequent measure obtained in cross-sectional studies.

2.1.2 Relative measures of disease occurrence

The question of whether or not a particular health condition is more likely under one set of exposures than another is very common in epidemiology. For example, Wilkins and Sinks, 1990 used a case-control study of parental occupation and intracranial neoplasms in childhood to investigate whether the case parents were more likely than the control parents to have a job before, during, or after pregnancy, that involved exposure to N-nitroso compounds. Such questions invariably involve two proportions, p_1 and p_0, the probabilities of the disease occurrence under exposure and non-exposure, respectively, and are best approached using a relative measure of disease occurrence.

We now define two measures that are central to this thesis.

2.1.2.1 Relative Risk

For simplicity, suppose we have two exposures, E_1 and E_0, that are thought to be associated with the occurrence of disease D. Let E and D be bivariate random variables describing the marginal exposure and the marginal disease status, i.e.,

$$E = \begin{cases} 1, & \text{if individual is exposed to } E_1, \\ 0, & \text{if individual is exposed to } E_0, \end{cases} \tag{2.1.2}$$

$$D = \begin{cases} 1, & \text{if individual has the disease } D, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1.3}$$

Let $p_1 = \Pr(D = 1|E = 1)$ denote the disease risk among individuals exposed to E_1, and let $p_0 = \Pr(D = 1|E = 0)$ denote the disease risk among the individuals exposed to E_0. We define the **relative risk** or **risk ratio**, R, of E_1 relative to E_0 as

$$R = \frac{p_1}{p_0} = \frac{\Pr(D = 1|E = 1)}{\Pr(D = 1|E = 0)}. \tag{2.1.4}$$

This ratio indicates how manyfold the risk of disease in the group E_1 is increased. If $R > 1$ we refer to a positive association while $R < 1$ indicates a negative association. When $R = 1$ we say there is no association.

Note also that there is no particular reason in equation (2.1.4) to choose p_1/p_0 over p_0/p_1, $(1-p_0)/(1-p_1)$, or $(1-p_1)/(1-p_0)$. Those four parameters all carry the same information, although their interpretations are slightly different. The symmetry in these definitions imply a similar symmetry in the statistical properties of estimators of these quantities.

2.1.2.2 Odds Ratio

Another measure of importance is the odds ratio. The ratio $o_0 = \dfrac{p_0}{1-p_0}$ and $o_1 = \dfrac{p_1}{1-p_1}$ are called the odds of disease in E_0 and E_1, respectively. The odds ratio of disease, ψ_d, is defined as the ratio of o_1 to o_0

$$\begin{aligned}\psi_d = \frac{o_1}{o_0} &= \frac{p_1(1-p_0)}{p_0(1-p_1)} \\ &= \frac{\Pr(D=1|E=1)[1-\Pr(D=1|E=0)]}{\Pr(D=1|E=0)[1-\Pr(D=1|E=1)]}.\end{aligned} \tag{2.1.5}$$

The interpretation of ψ_d is only qualitatively similar to R. Unfortunately, the odds ratio cannot be taken as a direct substitute for the risk ratio in general, although if p_0 and p_1 are both small, say, less than 0.2 then $\psi_d \approx R$; this is known as the rare disease assumption. In general, as shown by Davies et al., 1998, $\psi_d < R \iff R < 1$, and $\psi_d > R \iff R > 1$. The discrepancy between ψ_d and R increases as the departure of p_1 and R from unity is increased. For instance, $p_1 = 0.95$ and $p_0 = 0.8$ gives $R = 1.1875$ with $\psi_d = 4.75$. For $p_1 = .9$ and $p_0 = .1$, however, we get $R = 9$ and $\psi_d = 81$.

As with the relative risk, there is no particular reason to choose $p_1(1-p_0)/p_0(1-p_1)$ over its reciprocal $p_0(1-p_1)/p_1(1-p_0)$.

Corresponding to an odds ratio for disease, an odds ratio for exposure ψ_e can also be defined: Let $p_1' = \Pr(E=1|D=1)$ and $p_0' = \Pr(E=1|D=0)$. Define

$$\psi_e = \frac{p_1'(1-p_0')}{p_0'(1-p_1')} = \frac{\Pr(E=1|D=1)[1-\Pr(E=1|D=0)]}{\Pr(E=1|D=0)[1-\Pr(E=1|D=1)]}. \qquad (2.1.6)$$

Now a simple application of Bayes' theorem shows that $\psi_d = \psi_e$ (Cornfield, 1951). Since ψ_e is estimated directly from a case-control study, while ψ_d is the parameter of real interest, the equality of ψ_d and ψ_e means that a case-control study may be used as a reasonable substitute for a cohort study. A time-dependent version of $\psi_d = \psi_e$, relating the instantaneous odds ratios of disease and exposure is given by Prentice and Breslow, 1978.

An algebraic relationship binding the proportions p_1 and p_0 is easily seen to be described by the pair of equations,

$$p_1 = R \cdot p_0 \qquad \text{and} \qquad \frac{1}{p_1} = 1 - \frac{1}{\psi_d} + \frac{1}{\psi_d}\cdot\frac{1}{p_0}. \qquad (2.1.7)$$

Other measures of association between exposure and diseases include the risk difference or attributable risk $\delta = p_1 - p_0$, the preventive fraction $PF_e = \dfrac{p_0 - p_1}{p_0} = 1 - R$, and the attributable proportion or fraction $AP_e = \dfrac{p_1 - p_0}{p_1} = 1 - \dfrac{1}{R}$, where p_1 and p_0 are defined as in the definition of the risk ratio. These will not be extensively discussed in this thesis. See Miettinen, 1976; Walker, 1976; Rothman, 1986; and Blackwelder, 1993; for descriptions and other properties of these parameters as well as estimation procedures for them.

The parameters PF_e and AP_e are both linear combinations of R, so sample size computations based on them will be similar to those based on R. We will discuss this further in chapter 4.

2.2 Cohort and Case-Control Studies

In this section, we will formally define both a cohort study and a case-control study.

2.2.1 Cohort Studies

There are essentially two types of cohort studies, namely prospective and retrospective cohort studies. The former, which is perhaps more common, proceeds as follows. One identifies a disease-free representative sample of the population of interest, with various levels or combinations of exposures. The cohort is then followed dynamically forward for a period of time and disease status is recorded, usually along with other suspected cofactors of interest, for each individual in the study. This experimental design can be very costly, time consuming and logistically complex.

A retrospective cohort study is often less expensive to conduct. For example one may have already gathered a large amount of data on individuals living in certain community, for example, through government registries or through insurance company data. Suppose one is interested in learning if drivers using a cell phone while driving are more likely to be involved in traffic accidents than those who do not. From the database of the SAAQ

(Société d'Assurance Automobile du Quebec) combined with cell phone data, one could independently sample two groups of drivers, those who have a cellular and those who do not, recording whether each driver was involved in an accident or not. Other possible confounders such as driver's age and gender, age of the car, number of accidents during the last 24 months could be included in an analysis.

The cohort study facilitates the direct estimation of the effects of exposures on disease incidence or death rates. Table 2.1 serves as a summary of a simple cohort study with two exposure groups. Sampling theory for $\widehat{R}$ and $\widehat{\psi_d}$ will be discussed in section 2.3 and 2.4.

Now, since disease incidence is often low, very long follow-up times and large samples sizes are typically necessary in a prospective cohort study in order to observe a reasonable number of cases. Hence, time and funding constraints often dictate that another sampling approach is needed. Therefore in the next section we consider case-control studies.

2.2.2 Case-Control Design

Case-control designs are perhaps the most common designs used in epidemiology. In their simplest forms, retrospective or traditional case-control studies use a sample of cases of disease or of deaths that occur during an accrual time period, and a sample of control subjects alive and free of disease from the population. The investigator then "looks back" to record exposure status. In the simplest form of case-control studies, the data collected may be summarized in a 2×2 table as in Table 2.1. If the conditional probability of

Table 2.1: **Generic 2 × 2 table for exposure-disease outcomes**

	$E = 1$	$E = 0$	Total
$D = 1$	a	b	$\mathbf{n}_1 = a + b$
$D = 0$	c	d	$\mathbf{n}_0 = c + d$
Total	$\mathbf{m}_1 = a + c$	$\mathbf{m}_0 = b + d$	$\mathbf{N} = a + b + c + d.$

exposure in the cases is larger than among the controls, this indicates that exposure increases the probability of getting the disease. For example, the study described in the previous subsection can be adapted to fit the retrospective case-control scheme. Suppose that the SAAQ wishes to ascertain if drivers having a mobile phone are more likely to be involved in an accident than their counterparts who do not own a mobile phone. Using SAAQ records one could identify all the drivers involved in a collision and draw a sample from drivers who have never been involved in a collision. Retrospectively, the researcher would record whether each driver was exposed (had a mobile phone) or was unexposed (did not have a mobile phone).

In practice, case-control studies may be divided into three types depending on how cases are accumulated: cumulative incidence, prevalence, or incidence density. Retrospective case-control studies are of the first type when incident cases are accumulated during a well defined study period, from the population of interest. After deciding to keep all or a random sample of the cases, one draws the controls from the non-cases in the population. This is often the case for a localized outbreak of an infectious disease and or food poisoning in a restaurant or a social gathering.

The second type of a retrospective case-control design, the prevalence study, is rarely used, but has the benefit of yielding cases more quickly. A prevalence study is a study in which we ascertain all members of a population in a given time period and record their disease and exposure status. A typical example is the case-control study of infants with malformations.

Unlike the first two types, incidence-density case-control studies do not require any rare disease assumptions. These studies are based on the idea of matching each case on time, with one or more randomly selected controls (with or without replacement) alive at the onset time of cases. Such studies are ordinarily termed nested case-control studies (Mantel, 1973; Thomas, 1977; Liddell et al., 1977, Robins et al., 1989; Langholz and Clayton, 1994). Sufficient conditions under which the odds ratio is a good estimate of the relative risk (often called the incidence-densities ratio) are given by Greenland and Thomas, 1982. For instance, in addition to the constancy of the risk ratio in (t_0, t_1), it may be necessary that the exposure proportion also remain constant over that study period. A more general design (Prentice and Breslow, 1978; Lubin and Gail, 1984, Eq. 3.4) assumes a proportional hazard model for the exposed and unexposed groups.

Another matching strategy used in both case-control and cohort studies is the technique of matching on covariates other than time (Rothman and Greenland, 1998; Kupper et al., 1981). Matching refers to the pairing of one or more controls to the cases in a case-control design or that of one or more unexposed to the exposed group in a cohort with regard to possible confounding factors. Matched studies do not require the stability of the exposure proportion over time in order to consistently estimate the relative risk.

Other design strategies are described in Sato, 1992a, Sato, 1992b, Weinberg and Wacholder, 1990, Langholz and Thomas, 1990 and Langholz and Thomas, 1991, Rothman and Greenland, 1998. One of particular relevance to this thesis is the case-only design (Umbach and Weinberg, 1997, Greenland, 1999) used quite often in genetic epidemiology. It is based on the idea of sampling only cases and using prior information on exposure to replace the information provided by the controls en route to estimating parameters of interest. A benefit of case-only designs is that more effort can be devoted to the gathering of accurate information on the cases, but of course, the prior information on the controls needs to be sound. We shall address these designs in chapter 3.

Having given a brief overview of cohort and case-control studies we move on to their statistical analysis. We restrict ourselves to 2 × 2 tables, and, as usual, the analysis can be approached by frequentist or Bayesian methods. Frequentist methodology can be unconditional or conditional. We will present, briefly, the frequentist unconditional and Bayesian approaches. The conditional approaches are based on the idea of ancillary statistics and make the assumption that all the marginal totals n_0, n_1, m_0, m_1 (see Table 2.1) are fixed. They are described in Breslow and Day, 1980.

2.3 Frequentist analysis of 2×2 tables

In the sequel, we suppress the dependence of the analysis on the time interval over which cases are accrued, for notational convenience. We further restrict our attention to the simple but important case of dichotomous exposure (E) and disease (D) variables, which can be represented by the 2 × 2 contingency table shown in Table 2.1.

Following the notation of Marshall, 1988, the data from Table 2.1 will be notated as $\mathtt{T} = (\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d})$. $N = a + b + c + d$ is called the total sample size of the 2 × 2 table. The premiere goal of the thesis is to present a Bayesian approach to the selection of N, under different scenarios and a variety of sample size criteria. In case-control studies, we focus on $N = n_0 + n_1$ (respectively, the sample sizes for controls and cases), whereas in a cohort study, we consider $N = m_0 + m_1$ (sample sizes for unexposed and exposed).

2.3.1 Cohort sampling

Since subjects recruited in an unmatched cohort design are sampled according to their exposure status, Table 2.1 must be looked at columnwise. In their simplest form, the data from the exposed and unexposed samples will be assumed to be independently binomially distributed. That is,

$$\sum_{i=1}^{m_1} D_i \sim \mathbf{Bin}(m_1, p_1) \qquad \text{given } E = 1, \text{ and} \tag{2.3.1}$$

$$\sum_{j=1}^{m_0} D_j \sim \mathbf{Bin}(m_0, p_0) \qquad \text{given } E = 0, \tag{2.3.2}$$

where $p_0 = \Pr(D = 1|E = 0)$ and $p_1 = \Pr(D = 1|E = 1)$ are the success probabilities, while D_i and D_j are the individual disease outcomes for exposed and unexposed, $i = 1, \cdots, m_1$, $j = 1, \cdots, m_0$. The usual summary parameter is the risk ratio $R = \frac{p_1}{p_0}$. The unconditional maximum likelihood estimator (umle) of R is easily seen to be $\widehat{R} = \frac{am_0}{bm_1}$, and the asymptotic variance (Gart and Nam, 1988) $\mathbf{Var}(\ln \widehat{R}) \approx \frac{q_1}{m_1 p_1} + \frac{q_0}{m_0 p_0}$, where $q_i = 1 - p_i$, $i = 0, 1$. Similarly, the umle of $R_1 = \frac{p_0}{p_1}$, $R_2 = \frac{1-p_1}{1-p_0}$, and $R_3 = \frac{1-p_0}{1-p_1}$ can be derived along with corresponding asymptotic variances. As expected, $\mathbf{Var}(\ln \widehat{R}) \approx \mathbf{Var}(\ln \widehat{R}_1)$ while $\mathbf{Var}(\ln \widehat{R}_2) \approx \mathbf{Var}(\ln \widehat{R}_3)$. However $\mathbf{Var}(\ln \widehat{R}_1) \not\approx \mathbf{Var}(\ln \widehat{R}_3)$, which suggests that sample size based on these parameters might be dissimilar. The Bayesian sample size calculations corresponding to R_1, R_2 and R_3 will be described in chapter 4.

2.3.2 Case-control sampling

In the traditional case-control design, Table 2.1 must be looked at by row because subjects are sampled according to their disease status. The data from the cases and the controls are usually assumed to be independently binomially distributed. Formally:

$$\sum_{i=1}^{n_1} E_i \sim \mathbf{Bin}(n_1, p'_1) \qquad \text{for the cases, and} \tag{2.3.3}$$

$$\sum_{j=1}^{n_0} E_j \sim \mathbf{Bin}(n_0, p'_0) \qquad \text{for the controls,} \tag{2.3.4}$$

where $p'_0 = \Pr(E = 1|D = 0)$ and $p'_1 = \Pr(E = 1|D = 1)$ are success probabilities, and E_i and E_j the cases' and the controls' respective exposure outcomes $i = 1, \cdots, n_1$, $j = 1, \cdots, n_0$. The most often used summary parameter is the exposure odds ratio $\psi_e = \frac{p'_1(1-p'_0)}{p'_0(1-p'_1)}$. The umle is well known to be $\widehat{\psi_e} = \frac{ad}{bc}$, and an approximate variance is $\mathbf{Var}(\ln(\widehat{\psi_e})) \approx \frac{1}{n_0 p'_0 q'_0} + \frac{1}{n_1 p'_1 q'_1}$, where $q'_i = 1 - p'_i$, $i = 0, 1$. Fortunately here, the umle for $\psi_1 = \frac{1}{\psi_e} = \frac{p'_0(1-p'_1)}{p'_1(1-p'_0)}$, $\widehat{\psi_1} = \frac{bc}{ad}$, has the same limiting variance, so that the frequentist sample sizes for both parameters are expected to be the same. The corresponding Bayesian sample size calculation will be described in chapter 4.

2.3.3 Cross-sectional sampling

In a cross-sectional study, disease and exposure outcomes are measured simultaneously. The usual framework of a cross-sectional study is to regard the outcomes contained in Table 2.1 as governed by a multinomial distribution.

Let $p_{11} = \Pr(D = 1, E = 1)$, $p_{10} = \Pr(D = 1, E = 0)$, $p_{01} = \Pr(D = 0, E = 1)$, and $p_{00} = \Pr(D = 0, E = 0) = 1 - p_{11} - p_{10} - p_{01}$. The probability function of the table $\mathtt{T} = (\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d})$ is given by

$$f_{\mathrm{T}}(a, b, c, d) = \frac{N!}{a!\, b!\, c!\, d!}\, p_{11}^{a}\, p_{10}^{b}\, p_{01}^{c}\, p_{00}^{d}. \tag{2.3.5}$$

Under this sampling scheme, the umle of the odds ratio or cross ratio $\theta = \dfrac{p_{11}p_{00}}{p_{10}p_{01}}$ is given by $\widehat{\theta} = \dfrac{ad}{bc}$, and its variance satisfies $\lim\limits_{N \longrightarrow \infty} N \times \mathbf{Var}(\log(\hat{\theta})) = \dfrac{1}{p_{11}} + \dfrac{1}{p_{10}} + \dfrac{1}{p_{01}} + \dfrac{1}{p_{00}}$. Again, this asymptotic variance is invariant under the inverse relation.

Note that the identities

$$p_0 = \frac{p_{10}}{p_{10} + p_{00}} \qquad p_1 = \frac{p_{11}}{p_{11} + p_{01}} \tag{2.3.6}$$

$$p_0' = \frac{p_{01}}{p_{01} + p_{00}} \qquad p_1' = \frac{p_{11}}{p_{11} + p_{10}} \tag{2.3.7}$$

gives the relationships between the parameters of a cross-section study and the parameter of a cohort study and a case-control study.

2.3.4 Pair-matched sampling

The outcomes for pair-matched sampling can be represented as in Table 2.2, where a denotes the number of matched case and control pairs with the designation $(+,+)$, that is such that both members of the pair are exposed with similar definitions for b, c and d. Here N is the number of pairs and $2N$ is the total sample size for this design. Let p'_{11}, p'_{10}, p'_{01}, and p'_{00} be the probabilities of the pairs $(+,+)$, $(+,-)$, $(-,+)$ and $(-,-)$, respectively. In pair-matched terminology, a and d are the numbers of concordant pairs whereas b and c the numbers of discordant pairs and the sum $n = b + c$ is the effective sample size which is most often of interest. It has been shown by Ejigou and McHugh, 1977 that the odds ratio of exposure for a matched pair design is given by

$$\psi'_e = \frac{p'_{10}}{p'_{01}}, \tag{2.3.8}$$

under the assumption of homogeneity of the odds ratio across pairs i.e, there is no multiplicative interaction between exposure and the matching variable. A prospective study gives a similar table except for example that the outcome a is the number of (disease, disease) pairs, with similar definitions for cells b, c, and d. A similar relation $\psi'_d = \dfrac{p_{10}}{p_{01}}$ for prospective studies is described in Ejigou and McHugh, 1977.

The table of data represented by $\mathtt{T} = \left(\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\right)$ can again be modelled by a multinomial distribution. The umle of the odds ratio is $\widehat{\psi}'_e = \dfrac{b}{c}$ and its limiting variance is $\lim\limits_{N\longrightarrow\infty} N \times \mathbf{Var}(\log(\widehat{\psi}'_e)) = \dfrac{\psi'_e + 1}{p'_{01}\psi'_e} = \dfrac{1}{p'_{01}} + \dfrac{1}{p'_{10}}$. Likewise, $\widehat{\psi}'_d = \dfrac{b}{c}$ and $\lim\limits_{N\longrightarrow\infty} N \times \mathbf{Var}(\log(\widehat{\psi}'_d)) = \dfrac{\psi'_d + 1}{p_{01}\psi'_d} = \dfrac{1}{p_{01}} + \dfrac{1}{p_{10}}$. Again, this

Table 2.2: **Case-control matched-pair table where (+) represents exposure and (−) represents non-exposure.**

		Controls	
		+	-
Cases	+	a	b
	-	c	d

limiting variance is symmetrical with respect to ψ'_e and $\dfrac{1}{\psi'_e}$.

Note that these coefficients might not always be defined, since some cells might contain zero subjects. It has, therefore, been suggested to slightly modify the counts by adding some nonnegative constant to all cells entries, to get a table $\mathtt{T} = \left(\mathtt{a} + \varepsilon_{\mathtt{a}},\ \mathtt{b} + \varepsilon_{\mathtt{b}},\ \mathtt{c} + \varepsilon_{\mathtt{c}},\ \mathtt{d} + \varepsilon_{\mathtt{d}}\right)$. The three most popular adjustments for unmatched designs are $\mathtt{T} = \left(\mathtt{a},\ \mathtt{b} + 1,\ \mathtt{c} + 1,\ \mathtt{d}\right)$, $\mathtt{T} = \left(\mathtt{a} + \frac{1}{2},\ \mathtt{b} + \frac{1}{2},\ \mathtt{c} + \frac{1}{2},\ \mathtt{d} + \frac{1}{2}\right)$, and $\mathtt{T} = \left(\mathtt{a} + \frac{1}{2},\ \mathtt{b} + 1,\ \mathtt{c} + 1,\ \mathtt{d} + \frac{1}{2}\right)$, which correspond to using a total effective sample size of $N+2$ or $N+3$. In general, the estimators from these adjusted table don't grossly distort the information contained in the data when the cell entries are balanced, say $\log(\psi_e) < 4.0$. Comparative evaluations of the small bias properties of these estimators were given by Jewell, 1984, Jewell, 1986 and Walker and Cook, 1991.

Coverage probabilities and lengths of the confidence intervals of the log-odds parameter are investigated by Agresti, 1999, who also described two smoothing estimators for the p_{ij}'s. For example, he suggests $\widehat{p}_{11} = \dfrac{N}{N+4\varepsilon}\left(\dfrac{a}{N}\right) +$

$\frac{4\varepsilon}{N+4\varepsilon}\left(\frac{1}{4}\right)$ and $\widetilde{p}_{11} = \frac{N}{N+4\varepsilon}\left(\frac{a}{N}\right) + \frac{4\varepsilon}{N+4\varepsilon}\left(\frac{n_1 m_1}{N^2}\right)$, $\varepsilon > 0$ for a table $\mathrm{T} = \left(\mathrm{a}+\varepsilon,\ \mathrm{b}+\varepsilon,\ \mathrm{c}+\varepsilon,\ \mathrm{d}+\varepsilon\right)$. Adding the same constant ε to each cell is perhaps the best compromise for the risk ratio because this conservatively biases all ratios, R, R_1, R_2, and R_3, towards the null value of unity. This problem has led researchers to Bayesian methodology to analyze data of the type presented in Tables 2.1 and 2.2. Bayesian approaches require specification of a prior distribution across all unknown parameters. In return, Bayesian analyses allow the computation of credible intervals as opposed to confidence intervals, or the calculation of the probability that, say, R is above some clinical or etiological meaningful threshold, giving therefore a better way to summarize the effectiveness of any intervention (Franck et al., 1988). We now briefly review Bayesian inference for 2 × 2 tables.

2.4 Bayesian inferences for 2×2 tables

For a Bayesian methodologist, the aim is to obtain the posterior density (probability function)

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta), \qquad \text{(by Bayes' rule)} \qquad (2.4.1)$$

by updating the prior density (probability function) $p(\theta)$ with evidence from the data as reflected through the likelihood $p(y|\theta)$. Continuing in this spirit, Bayesian methods can be built in a complex hierarchical fashion which may require Markov chain Monte Carlo techniques (Brooks, 1998) to approximate joint and marginal posterior distributions. The book by Gelman et al., 1995 gives a readable account of practical Bayesian methods.

At the present time, many papers concerned with inferences about the odds ratio and the risk ratio have taken a fully Bayesian approach. Lindley, 1964, appears to be the pioneer of such methodology, with subsequent work by Leonard, 1972, Leonard, 1975, Leonard, 1977, Aitchison and Bacon-Shone, 1981, Latorre, 1982, Nurminen and Mutanen, 1987, Zelen and Parker, 1986, Marshall, 1988, Franck et al., 1988, Maritz, 1989, Carlin, 1992, Ashby et al., 1993, Walters, 1997, and Hashemi et al., 1997.

In general, the parameters arising from cohort and case-control are summarized using independent Beta prior distributions, while the cross-sectional and the matching designs usually use Dirichlet prior distributions, giving rise to the well known Beta-Binomial and Dirichlet-Multinomial models. We now briefly describe these models.

2.4.1 Cohort analysis

Bayesian methods for the analysis of binomial data rely on the specification of suitable priors for success probabilities. As previously mentioned, independent Beta (**Be**) priors are most commonly used for p_1 and p_0. That is, we assume that p_1 and p_0 are independent with

$$p_1 \sim \mathbf{Be}(a', c') \qquad \text{and} \qquad p_0 \sim \mathbf{Be}(b', d'), \qquad a', b', c', d' > 0. \qquad (2.4.2)$$

This conjugate model for the joint prior distribution of (p_0, p_1) can be regarded as equivalent to having observed an additional dataset $\mathtt{T}' = \left(\mathtt{a}', \mathtt{b}', \mathtt{c}', \mathtt{d}'\right)$. If the data are as in Table 2.1, then we have an augmented data set overall of $\mathtt{T}'' = \mathtt{T} + \mathtt{T}' = \left(\mathtt{a}'', \mathtt{b}'', \mathtt{c}'', \mathtt{d}''\right)$, where $a'' = a + a'$; and so on. Therefore, when independent Beta priors are used, elicitation of the priors for the binomial probabilities is equivalent to specifying a', b', c', d' for a fictitious sample size of $N' = a' + b' + c' + d'$ often called the **prior sample size**. The parameters of these Beta prior distributions can be elicited, for example, by a quick pilot study or by matching Beta quantiles with elicited means, modes, percentiles, or interquartile ranges, as described in Chaloner and Duncan, 1983, Chaloner and Duncan, 1987, Maritz, 1989, McCulloch, 1988, Franck et al., 1988, Spiegelhalter et al., 1994, and Bedrick et al., 1997. Non-informative (also called diffuse, vague, or reference) priors can take on various forms depending on how they are derived. The most common choices are: $\mathbf{Be}(0,0)$, $\mathbf{Be}(0,1)$, $\mathbf{Be}(1,0)$, $\mathbf{Be}(1,1)$ and $\mathbf{Be}(1/2,1/2)$. See Jeffreys, 1961; Bernardo, 1979; Kass and Wasserman, 1996; Bernardo and Ramón, 1998.

Historically, the parameter $0 < \gamma = \dfrac{R}{R+1} = \dfrac{p_1}{p_1 + p_0} < 1$ was the first studied. Aitchison and Bacon-Shone, 1981 derived the exact posterior density of γ:

$$p_\gamma(\gamma \,|\, \mathrm{T}'') = \begin{cases} \dfrac{1}{K} \dfrac{\gamma^{a''-1}}{(1-\gamma)^{a''+1}} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1} \left[1 - \frac{\gamma}{1-\gamma} \cdot Z\right]^{c''-1} dZ, \\ \hfill 0 < \gamma \leq \dfrac{1}{2} \\ \dfrac{1}{K} \dfrac{(1-\gamma)^{b''-1}}{\gamma^{b''+1}} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1} \left[1 - \frac{1-\gamma}{\gamma} \cdot Z\right]^{d''-1} dZ \\ \hfill \gamma > \dfrac{1}{2}. \end{cases} \tag{2.4.3}$$

where $K = \mathbf{B}(a'', c'') \times \mathbf{B}(b'', d'')$ and where $\mathbf{B}(.,.)$ is the Beta function. Note the symmetry property,

$$p_\gamma(\gamma \,|\, a'', b'', c'', d'') = p_\gamma(1-\gamma \,|\, c'', d'', a'', b''). \tag{2.4.4}$$

These authors suggest the use of $\gamma = \gamma(R)$ which is a monotonic function of R, in order to overcome the unbounded nature of R. The non-symmetric nature of R often makes credible intervals very wide and hard to interpret.

From equation (2.4.3), it can be shown that the posterior distribution for the risk ratio $R = \dfrac{p_1}{p_0} = \dfrac{\gamma}{1-\gamma}$ is,

$$p_R(R \,|\, \mathrm{T}'') = \begin{cases} \dfrac{R^{a''-1}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1}\left[1 - R\cdot Z\right]^{c''-1} dZ, \\ \hfill 0 < R \leq 1 \\[2ex] \dfrac{R^{-(b''+1)}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1}\left[1 - \dfrac{1}{R}\cdot Z\right]^{d''-1} dZ, \\ \hfill R > 1. \end{cases} \tag{2.4.5}$$

When discussing sample size calculations for the risk ratio, equation (2.4.5) will be central.

Sixteen years later, another form of the posterior density of R, equal to (2.4.5) up to a transformation, was independently derived by Hashemi et al., 1997, along with a normal approximation to the posterior distribution. They also derived approximate highest posterior density (HPD) regions for R and described the steps for the exact computation of HPD regions. These will be discussed fully in section 4.1.1.1.

In the context of randomized trials, Franck et al., 1988 use the prior distributions $p_0 \sim \mathbf{Be}(\alpha_0, \beta_0)$ and $R \sim \mathbf{Be}(\alpha_1, \beta_1)$, respectively on p_0 and R, to fit the constraint $0 < p_1 < p_0 < 1$ or equivalently $0 < R < 1$. With data summarized by a table $\mathrm{T} = \left(\mathrm{a}_1, \mathrm{b}_1, \mathrm{c}_1, \mathrm{d}_1\right)$, this model leads a joint posterior density

$$\begin{aligned} p_{R,p_0}(R, p_0 | \mathrm{T}) \;\; &\propto \;\; p_0^{\alpha_0+a_1+b_1-1}(1-p_0)^{\beta_0+d_1-1} R^{\alpha_1+a_1-1}(1-R)^{\beta_1-1} \times \\ &(1-Rp_0)^{c_1}, \qquad\qquad 0 < p_0,\, R < 1. \qquad (2.4.6) \end{aligned}$$

Note that there are two errors in equation (7) by Franck et al., 1988. The factor $(1-R)^{\beta-1}$ is missing and the power of $(1-Rp_0)$ should be $n-y$ in their notation. By integrating over p_0, one gets

$$\begin{aligned} p_R(R| \mathrm{T}) \propto R^{\alpha_1+a_1-1}&(1-R)^{\beta_1-1} \times \\ &\int_0^1 p^{\alpha_0+a_1+b_1-1}(1-p)^{\beta_0+d_1-1}(1-Rp)^{c_1}\, dp, \;\; 0 < R < 1. \quad (2.4.7) \end{aligned}$$

It is obvious that the family of posterior distributions underlying equation (2.4.7) contain those from equation (2.4.5) when $R < 1$ as a proper subset when c' is an integer. More precisely, for equation (2.4.7) to be equal to the first row of equation (2.4.5), the following conditions must hold:

$$\begin{aligned} &\beta_1 = 1, \qquad \alpha_1 + a_1 = a + a' \qquad c_1 = c + c' - 1, \\ &\alpha_0 + a_1 + b_1 = a + a' + b + b', \qquad \beta_0 + d_1 = d + d'. \qquad (2.4.8) \end{aligned}$$

One such match is obtained by choosing

$$\begin{aligned} &a_1 = a, \qquad b_1 = b, \qquad c_1 = c + c' - 1, \qquad d_1 = d \\ &\alpha_0 = a' + b', \qquad \beta_0 = d', \qquad \alpha_1 = a' \qquad \beta_1 = 1. \qquad (2.4.9) \end{aligned}$$

We discuss this distribution more formally in subsection 4.1.1.2 of chapter 4.

Now that we have given the distribution of R, let's examine the conditions under which the frequentist and Bayesian posterior point estimates are equal. Of course, even here interpretations of the inferences differ between the two paradigms. Following this, we will describe how to derive the distribution of the other relative risk parameters, R_1, R_2 and R_3, from section 2.3.1.

We have

$$\mathrm{E}(R|\,\mathrm{T}'') = \frac{(a+a')(m_0+b'+d'-1)}{(m_1+a'+c')(b+b'-1)} \tag{2.4.10}$$

which yields the frequentist estimate $\widehat{R} = \dfrac{am_0}{bm_1}$ of R by choosing $\mathrm{T}' = \big(0, 1, 0, 0\big)$. Similar estimates can be given for the other corrected estimates, which then also match the corresponding frequentist estimates.

Finally, we note that to get the posterior distribution of R_1, R_2 and R_3, one has only to replace the posterior augmented data $\mathrm{T}'' = \big(\mathrm{a}'', \mathrm{b}'', \mathrm{c}'', \mathrm{d}''\big)$ by $\mathrm{T}_1'' = \big(\mathrm{b}'', \mathrm{a}'', \mathrm{d}'', \mathrm{c}''\big)$, $\mathrm{T}_2'' = \big(\mathrm{c}'', \mathrm{d}'', \mathrm{a}'', \mathrm{b}''\big)$, and $\mathrm{T}_3'' = \big(\mathrm{d}'', \mathrm{c}'', \mathrm{b}'', \mathrm{a}''\big)$ respectively, in equation (2.4.5). Obviously $T'' = T_1''$ if and only if $a'' = b''$and $c'' = d''$, thus R and its reciprocal R_1 have the same posterior distribution. Similar relations can be derived between the parameters R_2 and R_3.

2.4.2 Case-control and cross-sectional analysis

In case-control studies the odds ratio, ψ_e, is the parameter that is directly estimable from the data. The posterior distribution function for the odds ratio ψ_e was obtained by Hora and Kelley, 1983, while the posterior density for the odds ratio was derived by Zelen and Parker, 1986, Marshall, 1988, and Hashemi et al., 1997. In what follows it is assumed that p_0' and p_1' are independent Beta (**Be**):

$$p_1' \sim \mathbf{Be}(a', b') \qquad \text{and} \qquad p_0' \sim \mathbf{Be}(c', d'), \qquad a', b', c', d' > 0. \tag{2.4.11}$$

These authors show that the posterior density function is

$$p_{\psi_e}(\psi \,|\, \mathrm{T}'') = \begin{cases} \dfrac{\psi^{a''-1}}{K} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+\psi y)^{a''+b''}}\, dy, & 0 < \psi < 1, \\ \\ \dfrac{\psi^{-(b''+1)}}{K} \displaystyle\int_0^1 \frac{y^{b''+d''-1}(1-y)^{a''+d''-1}}{\left(1-y+\dfrac{y}{\psi}\right)^{a''+b''}}\, dy, & \psi \geq 1, \end{cases} \tag{2.4.12}$$

where $K = \mathbf{B}(a'', b'')\, \mathbf{B}(c'', d'')$ and where again $\mathbf{B}(.,.)$ is the Beta function. Expression (2.4.12) can be rewritten using a hypergeometric function, as

$$p_{\psi_e}(\psi \,|\, \mathrm{T}'') = \begin{cases} \dfrac{\psi^{a''-1}}{C}\, {}_2F_1\left(a''+b'',\, a''+c'';\, N'';\, 1-\psi\right), & \psi \leq 1, \\ \\ \dfrac{\psi^{-b''-1}}{C}\, {}_2F_1\left(b''+d'',\, a''+b'';\, N'';\, 1-\dfrac{1}{\psi}\right), & \psi > 1, \end{cases} \tag{2.4.13}$$

where $C = \dfrac{\mathbf{B}(a'', b'')\, \mathbf{B}(c'', d'')}{\mathbf{B}(a''+c'', b''+d'')}$. Here $N'' = N + N'$ is called the **effective sample size** or the **extended sample size** (Adcock, 1992) and

$$\begin{aligned} {}_2F_1(a, b; c; z) &= \frac{1}{\mathbf{B}(b, c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-zt)^{-a} dt, & (2.4.14) \\ &= \sum_{i=0}^{\infty} \frac{\Gamma(a+i)}{\Gamma(a)} \frac{\Gamma(b+i)}{\Gamma(b)} \frac{\Gamma(c)}{\Gamma(c+i)} \frac{z^i}{i!}, \quad |z| < 1, \quad c > b > 0,\ a > 0. \\ &= {}_2F_1(b, a; c; z) \end{aligned}$$

From (2.4.12), it can be shown that the posterior expected value of ψ_e is

$$\mathbf{E}(\psi_e |\, \mathrm{T}'') = \frac{(a+a')(d+d')}{(c+c'-1)(b+b'-1)}, \tag{2.4.15}$$

matching again the umle $\hat{\psi}_e = \dfrac{ad}{bc}$ when $\mathrm{T}' = \big(0, 1, 1, 0\big)$ (Marshall, 1988). Corresponding results for other estimates similarly hold.

A useful property of the posterior density of ψ_e is that it is invariant: $p_{\psi_e}(\psi \,|\, \mathrm{T}'') = p_{\psi_e}(\psi \,|\, \mathrm{T}''_1)$, where $\mathrm{T}''_1 = \left(\mathrm{d}'', \mathrm{c}'', \mathrm{b}'', \mathrm{a}''\right)$. This follows from the invariance of k-th moment of the posterior density derived by Marshall, 1988,

$$\mathbb{E}(\psi_e^k \,|\, \mathrm{T}'') = \frac{\Gamma(a''+k)\Gamma(b''-k)\Gamma(c''-k)\Gamma(d''+k)}{\Gamma(a'')\Gamma(b'')\Gamma(c'')\Gamma(d'')}, \qquad 0 \le k \le \min(b'', c''). \tag{2.4.16}$$

Indeed, such a result is expected since the point estimate of the parameter ψ_e for such a table is equal to the estimate corresponding to the table T'', so a natural symmetry exists.

Again, equation (2.4.22) for the moments shows that inverting the parameter ψ_e is equivalent to using a table $\mathrm{T}''_2 = \left(\mathrm{c}'', \mathrm{d}'', \mathrm{a}'', \mathrm{b}''\right)$ or $\mathrm{T}''_3 = \left(\mathrm{b}'', \mathrm{a}'', \mathrm{d}'', \mathrm{c}''\right)$, i.e., using $p_{\psi_e}(\psi \,|\, T''_2)$ as the density function for the odds ratio.

A log normal-based approximation for the posterior distribution of ψ_e is described by Zelen and Parker, 1986, Marshall, 1988, and Hashemi et al., 1997.

The following relations are always true, whether or not the rare disease assumption holds:

$$R = \frac{p_1}{p_0} = \frac{p'_1}{1-p'_1}\frac{1-p_e}{p_e}, \tag{2.4.17}$$

$$p_e - p'_0 = (p'_1 - p'_0)p_d, \tag{2.4.18}$$

where $p_1 = \Pr(D = 1|E = 0)$, $p_0 = \Pr(D = 1|E = 0)$, $p'_1 = \Pr(E = 1|D = 1)$, $p'_0 = \Pr(E = 1|D = 0)$, $p_e = \Pr(E = 1)$, and $p_d = \Pr(D = 1)$. Therefore, in particular, under the rare disease assumption $p_e \approx p'_0$, so that the probability of exposure in the control series in case-control sampling can be approximated by p_e. This idea formed the basis of the analysis by Marshall who used equation (2.4.12) to carry out a posterior analysis for the risk ratio R, including the computation of posterior credible intervals.

It is possible to carry out a case-only design if sufficient prior information on the non-case subjects is available. It is difficult to imagine how a frequentist analysis would incorporate such prior information, except perhaps, in the trivial case where a perfect point estimate is assumed.

In a cross-sectional study, by using a Multinomial-Dirichlet model we again obtain equation (2.4.12) as the posterior distribution for the odds ratio (Latorre, 1982). Such a result is not surprising in light of the following proposition.

Proposition 2.4.1. *Let*

$$(p_{11}, p_{10}, p_{01}, p_{00}) \sim \textit{Dirichlet}\ (a_{11}, a_{10}, a_{01}, a_{00}). \tag{2.4.19}$$

a) *Define the random variables* $w = p_{00} + p_{01}$, $p_1' = \dfrac{p_{11}}{p_{11} + p_{10}}$ *and* $p_0' = \dfrac{p_{01}}{p_{01} + p_{00}}$. *Then* w, p_1' *and* p_0' *are independent random variables with*

$$\begin{aligned} w &\sim \boldsymbol{Be}(a_{11} + a_{10}, a_{01} + a_{00}) \\ p_1' &\sim \boldsymbol{Be}(a_{11}, a_{10}) \\ p_0' &\sim \boldsymbol{Be}(a_{01}, a_{00}). \end{aligned} \tag{2.4.20}$$

Conversely, given equation (2.4.20) *and* $p_{11} = wp_1'$, $p_{10} = w(1 - p_1')$, $p_{01} = (1 - w)p_0'$, $p_{00} = (1 - w)(1 - p_0')$, *equation* (2.4.19) *follows.*

b) Similarly, define the random variables $v = p_{11}+p_{01}$, $p_1 = \dfrac{p_{11}}{p_{11}+p_{01}}$ *and* $p_0 = \dfrac{p_{10}}{p_{10}+p_{00}}$. *Then* v, p_1 *and* p_0 *are independent random variables with*

$$\begin{aligned} v &\sim Be(a_{11}+a_{01}, a_{10}+a_{00}) \\ p_1 &\sim Be(a_{11}, a_{01}) \\ p_0 &\sim Be(a_{10}, a_{00}). \end{aligned} \tag{2.4.21}$$

Conversely, given equation (2.4.21) *and* $p_{11} = vp_1$, $p_{10} = (1-v)p_0$, $p_{01} = v(1-p_1)$, $p_{00} = (1-v)(1-p_0)$, *equation* (2.4.19) *follows.*

Proof. We will only prove **a)**, the proof for b is similar. The density of $(p_{11}, p_{10}, p_{01}, p_{00})$ is

$$f(p_{11}, p_{10}, p_{01}) \propto p_{11}^{a_{11}-1}\, p_{10}^{a_{10}-1}\, p_{01}^{a_{01}-1}\, p_{00}^{a_{00}-1}, \tag{2.4.22}$$

with p_{11}, p_{10}, $p_{01}, p_{00} > 0$ and $p_{11}+p_{10}+p_{01}+p_{00} = 1$. Using the definitions of the variables w, p_1' and p_0', one easily derives the relations: $p_{11} = wp_1'$, $p_{10} = w(1-p_1')$, $p_{01} = (1-w)p_0'$, and $p_{00} = (1-w)(1-p_0')$, yielding a Jacobian of $J = w(1-w)$. Thus,

$$f(w, p_1', p_0') \propto w^{a_{11}+a_{10}-1}(1-w)^{a_{01}+a_{00}-1}\, p_1'^{a_{11}-1}(1-p_1')^{a_{10}-1}\, p_0'^{a_{11}-1}(1-p_0')^{a_{00}-1}, \tag{2.4.23}$$

where w, p_1 and p_0 are independent random variables and equation (2.4.20) follows. The proof of the converse goes through the same steps in reverse order. □

Although its proof is straightforward, Proposition 2.4.1 does not seem to have been previously stated. In a cross-sectional study, we have

$$(p_{11}, p_{10}, p_{01}, p_{00}) \,|\, T'' \sim \text{Dirichlet } (a'', b'', c'', d'').$$

Hence $p_1' = \dfrac{p_{11}}{p_{11}+p_{10}}$ and $p_0' = \dfrac{p_{01}}{p_{01}+p_{00}}$ are independently distributed as

$$p_1' \sim \mathbf{Be}(a'', b''), \qquad p_0' \sim \mathbf{Be}(c'', d''), \tag{2.4.24}$$

as given by equation (2.4.12). Similarly, $p_1' = \dfrac{p_{11}}{p_{11}+p_{01}} \sim \mathbf{Be}(a'', c'')$ is independent of $p_0' = \dfrac{p_{10}}{p_{10}+p_{00}} \sim \mathbf{Be}(b'', d'')$ as given by equation (2.4.4). The result by Latorre, 1982 is therefore completed with the relation $\psi = \dfrac{p_{11}p_{00}}{p_{10}p_{01}} = \dfrac{p_1'(1-p_0')}{p_0'(1-p_1')} = \dfrac{p_1(1-p_0)}{p_0(1-p_1)}$. Proposition 2.4.1 provides an elegant and powerful property linking case-control, cohort and the cross-sectional studies, in that the invariance of the posterior distribution is shown across all study types. Proposition 2.4.1 provides an elegant and effortless way to simulate data from the Dirichlet distribution with four parameters. In the general case of Dirichlet distributions with $k > 3$ parameters, the algorithm by Devroye, 1986, based on the Gamma distribution is suitable. Another unexpected consequence of Proposition 2.4.1 is that the risk ratio from a cross-sectional study, defined by $R = \dfrac{p_{11}(p_{10}+p_{00})}{p_{10}(p_{11}+p_{01})}$, will also follow the distribution given by equation (2.4.5), as does $R = \dfrac{p_1}{p_0}$, computed from a cohort design. These results do not seem to have been previously discussed. In fact, no one has attempted to derive the posterior distribution of R in a cross-sectional setting, perhaps because a straightforward proof using transformations and Jacobians is very messy.

Despite the equivalence shown in Proposition 2.4.1 it should be remembered that the two situations arise from different sampling schemes.

2.4.3 Matched case-control and cohort analyses

In a matched pair analysis, under again the Dirichlet-Multinomial model, the posterior density of the odds ratio ψ'_e and ψ'_d are the same, that is

$$f_{\psi'_e}(\psi) = \frac{1}{\mathrm{B}(b'', c'')} \frac{\psi^{b''-1}}{(1+\psi)^{b''+c''}}. \tag{2.4.25}$$

This density is that of a type-II scalar beta random variable. With some algebra, it is easy to show that $\frac{c''}{b''}\psi \sim F_{2b'',2c''}$. Actually, this is also the density of the odds $\frac{p}{1-p}$ of a proportion p (Lindley, 1964) obeying the Beta-Binomial setup with posterior distribution $\mathbf{Be}(a'', b'')$. This distribution was also derived by Bernardo and Ramón, 1998 in a more general set up as the ratio $\phi = \theta_i/\theta_j$ of two parameters of a multinomial distribution with $k > 3$ parameters.

Now that we have described our four main designs in terms of their modelling and analyses, it is time to move to the subject of primary interest, which is sample size determination for estimating the risk ratio and the odds ratio. In a cohort design, sample size investigation is equivalent to finding m_1 and m_0 while in case-control studies we are interested in ascertaining n_1 and n_0, although $N = m_0 + m_1 = n_0 + n_1$ (refer to Table 2.1). In cross-sectional and matched studies, the quantity of importance is $N = a + b + c + d$ (refer to Table 2.2). Occasionally, we might be concerned with $n = b + c$, the number of discordant pairs in a pair-matched design necessary to detect a predetermined association.

2.5 Review of sample size methodology

One of the most important statistical questions in the early stages of virtually any study is the choice of sample size. Indeed, large sample sizes can be wasteful of resources whereas small sample sizes are often not large enough for accurate estimation of the parameters of interest.

Both frequentist and Bayesian sample size determination methods can be partitioned into at least 3 broad types: prediction, hypothesis testing, and estimation, each approach usually resulting in different sample sizes. For example, it is well acknowledged (O'Neill, 1984) that sample sizes arising from power considerations are often not sufficient for estimation purposes. In general, the formulation of a sample size criterion depends heavily on many quantities, including the specifics of the study design, the methodology used, the modelling of covariates, the main parameter under study, estimates of that parameter and its variance, the statistic employed, and the desired precision of estimation. It is worth noting that the majority of frequentist sample size results in the literature to date are based on asymptotics.

In some cases, it may be more important to ensure accurate prediction for future subjects, rather than ensuring accurate parameter estimates. Frequentist and Bayesian designs for prediction are usually combined with optimal designs for estimation and prediction (Silvey, 1980; Atkinson and Donev, 1992) and covered extensively in the review by Chaloner and Verdinelli, 1995. Optimal experimental designs for binary data are described in Chaloner and Larntz, 1989 and Abdelbasit and Plackett, 1983, while optimal designs for multinomial logistic regression are given by Zocchi and Atkinson, 1999. These papers discuss the A-, C-, and D- optimality criteria as applied to problems of the selection of covariates in either a linear or a non-linear problem. These approaches are not discussed further in this thesis. Optimal allocation for Bayesian inference about an odds ratio based on the normal distribution was treated by Brooks, 1987.

Frequentist power calculations (Schlesselman, 1982; and Breslow and Day, 1988; Wickramaratne, 1995), currently the most commonly used procedures in practice, are based on the specification of an alternative hypothesis, and type I and II error probability levels α and β. For example, sample size required in a cohort design assuming $m_1 = m_0$ for the hypothesis $H_0 : R = 1$ against $H_A : R > 1$ satisfies the equation

$$m_1 = \frac{\left[z_\alpha\sqrt{2\bar{p}\bar{q}} + z_\beta\sqrt{p_0[1 + R - p_0(1 + R^2)]}\right]^2}{[p_0(R-1)]^2}, \tag{2.5.1}$$

where $\bar{p} = \frac{1}{2}p_0(1+R)$, $\bar{q} = 1-\bar{p}$, and z_α and z_β are the usual quantiles of the normal distribution. Schork and Williams, 1980, Parker and Bregman, 1986, Connett et al., 1987, Connor, 1987, Fleiss, 1988, Dupont, 1988, Royston, 1993, Ejigou, 1996, Julious and Campbell, 1998 have also investigated power issues for matched studies. Power-based sample size requirements for case-only designs to detect gene-environment interactions are the raison d'etre of the paper by Yang et al., 1997. Bayesian sample sizes derived from hypothesis testing for binomial experiments based on Bayes factors (Kass and Raftery, 1995) is developed by Katsis and Toman, 1999. A different philosophy was taken by Walker, 1977. Rather than searching for the smallest sample size, Walker choose to specify the magnitude of the smallest detectable relative risk for given α, β and n.

Although power-based sample size methods are very popular, they have often been criticized because they carry with them all of the drawbacks associated with p-values. In addition, type I and II errors do not often reflect the true purpose of a study. For instance, most statistical analysts agree that confidence intervals are more informative than p-values, and the design of a study should match the eventual analysis. We are therefore left with sample size calculation for interval estimation.

2.5.1 Frequentist fixed width coverage-based sample size calculation

In this section, we present the main frequentist results for sample size calculation when estimating a parameter of interest to within a fixed distance l of the true parameter value, with fixed confidence coefficient $1-\alpha$. In the sequel we shall present alternative Bayesian methods.

The following is a presentation of the work by O'Neill, 1984 on case-control designs. Let $g = \dfrac{n_0}{n_1}$ be the predetermined ratio of controls to cases, $p_1' = \Pr(E = 1|D = 1)$ and $p_0' = \Pr(E = 1|D = 0)$ be the exposure probabilities among cases and controls, respectively, $q_i' = 1 - p_i'$, $i = 0, 1$, and $\psi_e = \dfrac{p_1'(1-p_0')}{p_0'(1-p_1')}$ be the exposure odds ratio. An approximate $100(1-\alpha)\%$ two-sided confidence interval (CI) for $\log(\psi_e)$ is

$$\log(\widehat{\psi}_e) \pm z_{1-\alpha/2}\sqrt{\mathbf{Var}(\log(\widehat{\psi}_e))}. \tag{2.5.2}$$

An alternative approach is to derive a CI for ψ_e directly by

$$\widehat{\psi}_e \pm z_{1-\alpha/2}\sqrt{\mathbf{Var}(\widehat{\psi}_e)}. \tag{2.5.3}$$

$\mathbf{Var}(\log(\widehat{\psi}_e)) \simeq \dfrac{1}{n_1 p_1' q_1'} + \dfrac{1}{n_0 p_0' q_0'}$ and $\mathbf{Var}(\widehat{\psi}_e) \simeq \psi_e^2 \mathbf{Var}(\log(\widehat{\psi}_e))$ may be obtained by means of the delta method. One reasonable way to determine the sample sizes n_0 and n_1 is to equate the length of the CIs derived from equations (2.5.2) and (2.5.3), respectively, with $2l$. Using (2.5.2) given ψ_e and p_0' gives the following solution for n_1:

$$n_1 = \frac{z^2_{1-\alpha/2}}{p_0' q_0'} \frac{1}{l^2} \left\{ \frac{(q_0' + p_0'\psi_e)^2}{\psi_e} + \frac{1}{g} \right\}. \tag{2.5.4}$$

Using equation (2.5.3) gives the following solution for n_1:

$$n_1 = \frac{z^2_{1-\alpha/2}}{p_0' q_0'} \left(\frac{\psi_e}{l} \right)^2 \left\{ \frac{(q_0' + p_0'\psi_e)^2}{\psi_e} + \frac{1}{g} \right\}. \tag{2.5.5}$$

Another approximate CI for ψ_e, termed the logit CI, has better asymptotic coverage properties for moderate sample sizes. This interval is given by $\widehat{\psi}_e \exp\left\{ \pm z_{1-\alpha/2}\sqrt{\mathbf{Var}(\log(\widehat{\psi}_e))} \right\}$, with a width $\widehat{W} = 2\psi_e \sinh(d)$ conditional on ψ_e, where $d = z_{1-\alpha/2}\sqrt{\mathbf{Var}(\log(\widehat{\psi}_e))}$. The corresponding sample size for a fixed length $2l$, conditional on ψ_e and p_0', is

$$n_1 = \frac{z^2_{1-\alpha/2}}{p_0' q_0'} \frac{\left\{ \dfrac{(q_0' + p_0'\psi_e)^2}{\psi_e} + \dfrac{1}{g} \right\}}{\left\{ \operatorname{arcsinh}\left(\dfrac{l}{\psi_e} \right) \right\}^2}. \tag{2.5.6}$$

The ratio of the sample sizes given by equation (2.5.5) and (2.5.6) respectively, is $\left\{ \dfrac{\operatorname{arcsinh}(\rho)}{\rho} \right\}^2$, where $\rho = \dfrac{l}{\psi_e}$. A simple analysis of the first derivative of the function $\rho \longmapsto \dfrac{\operatorname{arcsinh}(\rho)}{\rho}$ reveals that the curve associated with this function is roughly cusp shaped and bounded by 0 and 1 as described by

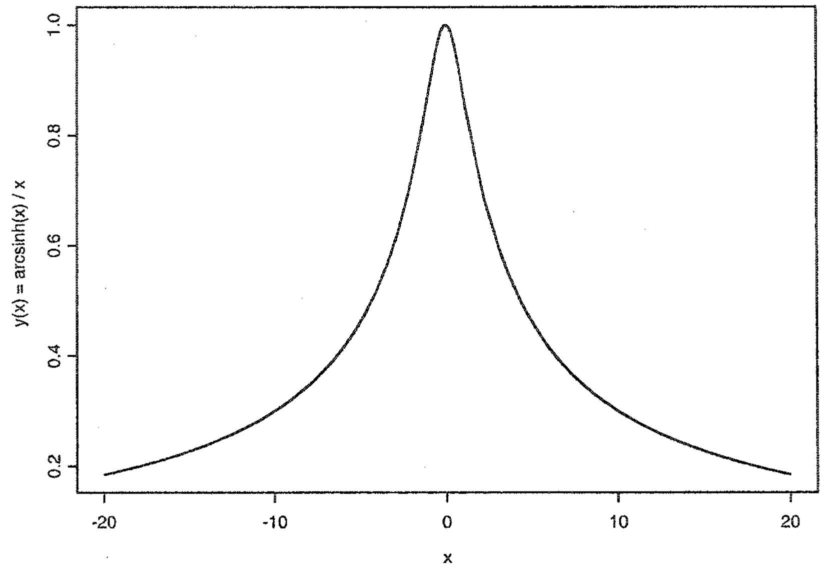


Figure 2.5.1: Graph of the function $f(x) = \dfrac{\operatorname{arcsinh}(x)}{x}$.

Figure 2.5.1. This shows that although the logit CI for ψ_e results in larger sample sizes, the CI used in equation (2.5.3) is known to have poor coverage properties for moderate sample sizes. Both equations (2.5.5) and (2.5.6) depend on l only through ρ, which suggests that we can specify ρ rather than l.

The derivation of these equations for cohort, cross-sectional and matched designs requires $\dfrac{1}{p_0'q_0'}\left\{\dfrac{(q_0'+p_0'\psi_e)^2}{\psi_e}+\dfrac{1}{g}\right\}$ be replaced by the asymptotic variances of either R, ψ, ψ_e' or ψ_d' as given in section 2.3, and ψ_e in $\left(\dfrac{\psi_e}{l}\right)^2$ by either R, ψ, ψ_e' or ψ_d', respectively.

Despite the usefulness of these and other similar sample size formulae, they do not fully take into account the stochastic nature of CI's. This can lead to severe underestimation of sample size required for a specified length or coverage (Satten and Kupper, 1990). Having exposed the problem, these authors proposed computing sample sizes based on tolerance probabilities. For example, if the confidence interval that one is intending to use after the collection of data is the logit CI in a case-control design, then such a criteria is equivalent to finding the smallest n_1 such that

$$\Pr\left[z_{1-\alpha/2}\sqrt{\frac{1}{A}+\frac{1}{n_1-A}+\frac{1}{C}+\frac{1}{kn_1-C}}\leq l\right]\geq 1-\varepsilon, \qquad (2.5.7)$$

where A and C represent the random variables whose outcomes are $0 \leq a \leq n_1$ and $0 \leq c \leq kn_1$, respectively (see Table 2.1). The left hand-size of equation (2.5.7) can be solved explicitly or estimated by means of Monte Carlo simulation. This will not be discussed further here, as we prefer Bayesian approaches that will solve the same problem, while also fully incorporating prior information.

In sample size calculation problems for 2 × 2 tables, one can also consider the choice of the optimal ratio of controls per case, $g = \frac{n_0}{n_1}$ in case-control studies or the ratio of non-exposed to exposed $g = \frac{m_0}{m_1}$, in cohort studies. Traditionally, g is set to a convenient constant g_0 chosen by the investigator before the collection of the data. The sample size of interest is then $N(g_0) = n_1 + n_0 = (g_0 + 1)n_0$. In practice, $g = 1$ is often the choice, corresponding to a design with equal sample sizes between cases and controls or exposed and non-exposed. This choice for g is often not optimal, especially when cost is incorporated into the calculation of sample sizes. Indeed, it is well known that when the costs of the exposed, and the non-exposed groups in cohort studies or of the case and the control groups in case-control studies, respectively c_1 and c_0, are very different, substantial savings can be made by using unequal sample sizes (Gail et al., 1976; Meydrech and Kupper, 1978). Moreover, even when $c_1 = c_0$, choosing the optimal ratio g leads to better designs in terms asymptotic variance (Walker, 1977). Since the choice of g is an element of the design it is natural to incorporate the search for g into the sample size calculation problem. This idea has been considered by numerous authors who use a crude direct search method to approximately solve the optimality problem. First, choose a grid of values for g and then for each g in that grid find the minimal $N(g)$. At the end, choose the ratio g that leads to the overall minimal $N(g)$. It is clear that the proposed solution is equivalent to finding, for example, the overall minimal sample size (n_1, n_0) that minimize the objective function $N = c_1 n_1 + c_0 n_0$ in case-control studies.

In the context of hypothesis testing, Gail et al., 1976 have demonstrated that the optimal ratio g follows what he called the square root rule, that is $g = \sqrt{\eta r}$, where $r = \frac{c_1}{c_0}$ and $\eta = \frac{p_1' q_1'}{p_0' q_0'} = \frac{\psi_e}{(q_0' + p_0' \psi_e)^2}$ for a case-control study and $\eta = \frac{p_1 q_0}{p_0 q_1} = \psi_d = \frac{R q_0}{1 - R p_0}$ for a cohort study. These results apply also to equations (2.5.4), (2.5.5), and (2.5.6) (proof available from the author). Unequal sample sizes are also investigated by Blackwelder, 1993, Nam and Fears, 1992a,Nam and Fears, 1992b, Fleiss, 1973, and Fleiss et al., 1980.

Most of the frequentist sample size formulae rely on the adequacy (Leemis and Trivedi, 1996) of the normal approximation to the binomial distribution. Often, when the rare disease assumption is valid, some refinements need to be made (Casagrande et al., 1978; Lemeshow et al., 1981). These problems can be minimized by computing the exact powers, CIs, and coverages as reported by Wickramaratne, 1995, or by Monte Carlo estimates of the power. Frequentist sample size estimation requires accurate point estimates of the parameters of interest, which are often unknown at the planning stage of the experiment. As discussed below, Bayesian sample size determination methods do not share many of these problems. They assume a prior distribution rather than a single point estimate. CI's are easily replaced by credible intervals from the marginal posterior distribution of interest. Importantly, they fully take into account the stochastic nature of the dataset.

2.5.2 Decision theoretic Bayesian sample size methods

Bayesian decision-oriented sample size calculation methods, the only "fully" Bayesian approach according to some authors (Lindley, 1956; Berger, 1985; Lindley, 1997; Bernardo, 1997), rely on maximizing an expected utility function over the set of all possible designs of size $n \geq 0$ and over all decisions $d \in \mathcal{D}$. The following notation is valid for this subsection. Let X be a random variable with density $p(x|\theta)$ and assume that a priori $\theta \sim p(\theta), \theta \in \Theta$. Let $\mathbf{x} = (x_1, \cdots, x_n) \in \mathcal{X}$ represent n realizations of X. After observing $\mathbf{x}$, we wish to make a decision $d \in \mathcal{D}$ about θ. Maximizing the expected utility approaches are based on finding the minimal n which maximizes

$$\int_{\mathcal{X}} \left\{ \max_{d \in \mathcal{D}} \int_{\Theta} u(n, \mathbf{x}, d, \theta) p(\theta|\mathbf{x}, n) \right\} p(\mathbf{x}|n) d\theta d\mathbf{x}, \qquad (2.5.8)$$

where $u(n, x, d, \theta)$ is the utility function reflecting the merit in choosing n, after having observed x and having taken the decision d about θ, when θ is the true value of the parameter. Here $p(\mathrm{x}|n) = \int_\Theta p(\mathrm{x}|\theta, n)p(\theta)d\theta$ is the pre-posterior predictive distribution, that is, the distribution of the not yet observed data x. A common form for u is

$$u(n, \mathrm{x}, d, \theta) = K\delta(n, \mathrm{x}, d, \theta) - Lw(n, \mathrm{x}, d) - cn, \qquad (2.5.9)$$

for an interval d, where $\delta(n, \mathrm{x}, d, \theta) = 1$ if $\theta \in \mathcal{D}$, $\delta(n, \mathrm{x}, d, \theta) = 0$ otherwise, and $w(n, \mathrm{x}, d)$ is the width of the interval d. The quantities $K, L > 0$ are two positive constants balancing high coverage against low width, and $c \geq 0$ the common cost associated with observing each subject. Basically, the cost is balanced with precision.

Shao, 1989, Müller, 1998 and Müller and Parmigiani, 1995 suggest that equation (2.5.8) can be approximately solved numerically using Monte Carlo simulations or MCMC methods, and by passing a smooth curve through the plot of the estimates. We will show how this idea can be used to help us solve the sample size problems posed in chapter 3 and 4, although full Bayesian decision theoretic sample size methods are not considered further in this thesis.

2.5.3 Bayesian sample size calculation for estimation problems

Various non-decision theoretic Bayesian criteria for sample size determination have appeared in the literature. The first three which we discuss, developed by Pham-Gia and Turkkan, 1992, are closely related to the idea of minimizing the Bayes risk. They all find the smallest n such that each of the following criterion functions,

$$\begin{aligned} \mathrm{PGT}-(i): \quad \psi_1(n) &= \max_{0\leq \mathbf{x}\leq n} \mathbf{Var}(\theta|\,\mathbf{x}, n) && (2.5.10) \\ \mathrm{PGT}-(ii): \quad \psi_2(n) &= \mathbf{E}\left[\mathbf{Var}(\theta|\,\mathbf{x}, n)\right] && (2.5.11) \\ \mathrm{PGT}-(iii): \quad \psi_3(n) &= -\frac{1}{c\mathbf{Var}\left[\mathbf{E}(\theta|\mathbf{x}, n)\right]} && (2.5.12) \end{aligned}$$

is smaller than $\epsilon > 0$, a predetermined level of precision.

Other criteria of relevance to this thesis are based on the idea of fixing the width or the coverage of credible intervals and searching for the minimum n for which the average coverage or the average length attains some specified probability level $1-\alpha$ or length l, respectively. The first of these criteria dates back to Adcock, 1987 and was further discussed in Adcock, 1988, Adcock, 1992, Adcock, 1993, Adcock, 1995, Adcock, 1997. He proposed seeking the minimum n such that

$$\int_{\mathcal{X}} \left\{ \int_{R(\mathbf{x})} p(\theta|\mathbf{x}, n) d\theta \right\} p(\mathbf{x}|n) d\mathbf{x} \geq 1-\alpha, \qquad (2.5.13)$$

where $R(\mathbf{x}) = [\mathbf{E}(\theta|\,\mathbf{x}) - l/2,\, \mathbf{E}(\theta|\,\mathbf{x}) + l/2]$ is a central symmetric credible interval of length l, called the tolerance region.

In the same spirit, Joseph et al., 1995 derived three criteria involving HPD credible regions. HPD regions of content $1-\alpha$ (see e.g. Box and Tiao, 1992) have the properties that $\Pr(\theta \in \mathtt{HPD}(\mathbf{x}, n)\,|\,\mathbf{x}, n) = 1-\alpha$ and for $\theta_1 \in \mathtt{HPD}(\mathbf{x}, n)$, $\theta_2 \notin \mathtt{HPD}(\mathbf{x}, n)$, $p(\theta_1|\,\mathbf{x}, n) \geq p(\theta_2|\,\mathbf{x}, n)$. HPD regions have the smallest volume for any fixed coverage level. Sample size criteria based on HPD regions are important to minimizing the sample size when the distribution is very skewed, and return similar samples sizes to symmetric regions for less skewed distributions, and so are optimal in this sense. Since the sample size determinations proposed in this thesis are based on these criteria we now review them.

2.5.3.1 Average Coverage Criterion

The average coverage criterion (**ACC**) finds the smallest integer n such that the average of all posterior coverage probabilities of $\mathtt{HPD}(\mathbf{x}, n)$ intervals of length l given $\mathbf{x}$ and n, over the predictive distribution $p(\mathbf{x}|n)$, is at least $1-\alpha$: That is, the **ACC** seeks the minimum n such that

$$\int_{\mathcal{X}} \left\{ \int_{\mathtt{HPD}(\mathbf{x},n)} p(\theta|\,\mathbf{x}, n)d\theta \right\} p(\mathbf{x}|n)d\mathbf{x} \geq 1-\alpha. \tag{2.5.14}$$

2.5.3.2 Average Length Criterion

The **average length criterion** (**ALC**), finds the smallest integer n such that the average of all lengths $l(\mathbf{x}, n)$ of $\mathtt{HPD}(\mathbf{x}, n)$ intervals of coverage $1-\alpha$ given x and n, over the predictive distribution $p(\mathbf{x}|n)$, is at most l: That is, the **ALC** seeks the minimum n such that

$$\int_{\mathcal{X}} l(\mathbf{x}, n)p(\mathbf{x}|n)d\mathbf{x} \leq l. \tag{2.5.15}$$

2.5.3.3 Worst Outcome Criterion

Sometimes, we might not want to average the coverage probabilities over all possible values of $\mathbf{x}$ as in equations (2.5.14) or (2.5.15), preferring a stricter criterion that finds the minimum n such that each coverage probability is larger than $1-\alpha$ for all $\mathbf{x}$ belonging to some pre-specified set $\mathcal{S} \subset \mathcal{X}$. This leads to the **worst outcome criterion** (**WOC**), which seeks the smallest n such that

$$\inf_{\mathbf{x}\in\mathcal{S}} \int_{\text{HPD}(\mathbf{x},n)} p(\theta|\,\mathbf{x},n)d\theta \geq 1-\alpha. \tag{2.5.16}$$

Closed form formulae in the case of a normal likelihood, given by Joseph and Bélisle, 1997, have demonstrated that this criterion approximately agrees numerically with frequentist results under certain weak prior conditions, and when the variance is assumed known. All seven criteria described respectively by equations (2.5.10), (2.5.11), (2.5.12), (2.5.13), (2.5.14), (2.5.15), and (2.5.16) may be extended to multivariate densities usually involving nuisance parameters. For example, if σ is a nuisance parameter, we can seek the n such that

$$\int_{\mathbf{x},\sigma} \left\{ \int_{\text{HPD}(\mathbf{x},\sigma,n)} p(\theta|\,\mathbf{x},\sigma,n)\,d\theta \right\} p(\mathbf{x},\sigma|\,n)\,d\mathbf{x}\,d\sigma \;\geq\; 1-\alpha, \tag{2.5.17}$$

or, equivalently, such that

$$\int_{\mathbf{x}} \left\{ \int_{\sigma} \int_{\text{HPD}(\mathbf{x},\sigma,n)} p(\theta|\,\mathbf{x},\sigma,n)\,p(\sigma|\,\mathbf{x})\,d\theta\,d\sigma \right\} p(\mathbf{x}|n)d\mathbf{x} \geq 1-\alpha, \tag{2.5.18}$$

where $\theta \sim p(\theta|\sigma)$, $\sigma \sim p(\sigma)$, so that $p(\theta,\sigma) = p(\theta|\sigma)p(\sigma)$. Essentially, we integrate over the nuisance parameter.

Although criticized as not being a fully Bayesian decision theoretic approach (Lindley, 1997) because either the length or the coverage is fixed before the search for the minimal sample size, it is easily seen that both the **ALC** and the **ACC** are limiting cases of maximizing the expected utility approaches (see section 2.5.2) as K or L tends to zero and $c = 0$. Indeed, we may place the **ACC** criterion within a decision problem framework as follows. One seeks the minimal n that satisfies

$$\int \left\{ \max_{C(\mathbf{x},l)\in\mathcal{I}(l)} \int_{C(\mathbf{x},l)} p(\theta|\,\mathbf{x}, n)d\theta \right\} p(\mathbf{x}|n)d\mathbf{x} \geq 1 - \alpha, \qquad (2.5.19)$$

where $\mathcal{I}(l)$ is the set of all posterior credible intervals of length l. If we set $L = 0$ in equation (2.5.9), then we conjecture that (2.5.10) would often converge to zero when the posterior second moments are defined for all n and $\mathbf{x}$ as $n \longrightarrow \infty$, although the proof does not seem straightforward. Maximizing (2.5.19) as in equation (2.5.13) creates an MEU criterion. Similarly for the **ALC** criterion.

2.6 Conclusion

We have just described practical Bayesian criteria for sample size determination. In the next chapter, we will show how some of these criteria, particularly **ACC**, **ALC**, **WOC** and formulae (2.5.17) or (2.5.18) can be implemented for sample size determination when estimating the risk and odds ratios. We will also consider the question of case-only designs. We will show how these criteria can be generalized to a much larger family of criteria, expose difficulties arising in their implementation, and suggest ideas for circumventing them. We will consider Monte Carlo curve fitting to improve the accuracy of sample size estimates as well. While developing our methodologies, we shall also consider transformations of the risk and the odds ratio. To close this chapter, we present some known results on HPD intervals and unimodality, important background to the results presented in chapter 3 and 4.

2.7 Useful definitions, results, and algorithms

2.7.1 General definitions and results

Given a scalar parameter $\theta \in \Theta$, let $\mathbf{x}_n = (x_1, \cdots, x_n)$ be n exchangeable realizations of a random variable X distributed as $p(x|\,\theta)$, $x \in \mathcal{X}$. We assume that θ is a continuous random variable distributed according to $p(\theta)$ (prior distribution). Let $(\theta_{\min}, \theta_{\max})$, with possibly, $\theta_{\min} = -\infty$, and/or $\theta_{\max} = \infty$, be the support on which $p(\theta) > 0$. The results presented here are mainly associated with the construction of credible intervals based on the posterior distribution $p(\theta|\,\mathbf{x}_n)$,

$$p(\theta|\,\mathbf{x}_n) = \frac{p(\mathbf{x}_n|\,\theta)\,p(\theta)}{\int_\Theta p(\mathbf{x}_n|\,\theta)p(\theta)d\theta}, \tag{2.7.1}$$

which is by assumption absolutely continuous with respect to Lebesgue measure on an interval $(\theta_{\min}, \theta_{\max})$. The majority of posterior distributions employed in this thesis are *unimodal* as will be proved in chapters 3 and 4. Consequently, the HPD regions studied here are simple intervals. For this reason, in the sequel we will refer to HPD intervals rather than the more general HPD regions.

Some proofs about unimodality require the notion of strongly unimodal distributions. According to Dharmadhikari and Joag-dev, 1988, we have the following.

Definition 2.7.1. The random variable θ or its distribution $p(\theta|\,\mathrm{x}_n)$ is strongly unimodal on $(\theta_{\min}, \theta_{\max})$ if and only if $p(\theta|\,\mathrm{x}_n)$ is continuous and log-concave $(\theta_{\min}, \theta_{\max})$, that is if and only if $\log\bigl(p(\theta|\,\mathrm{x}_n)\bigr)$ is concave. If $\log\bigl(p(\theta|\,\mathrm{x}_n)\bigr)$ has a second derivative on $(\hat{\theta}_{\mathrm{x}_n}, \theta_{\max})$, then this property is equivalent to

$$\frac{\partial^2 \log\bigl(p(\theta|\,\mathrm{x}_n)\bigr)}{d\theta^2} \leq 0.$$

Obviously, a strongly unimodal distribution is unimodal. The notion can be generalized to higher dimensions. The family of $\mathbf{Be}(a, b)$, $a, b \geq 1$ distributions is strongly unimodal. The family of distributions given by (2.4.25) with $b'', c'' \geq 1$, the type-I Beta is not strongly unimodal but is unimodal with a unique mode at $\dfrac{b'' - 1}{c'' + 1}$. The following proposition will help us later to show that the posterior distribution of the risk and the odds ratios are unimodal.

Proposition 2.7.1 (Dharmadhikari and Joag-dev, 1988).

- *All the moments of a strongly unimodal distribution are finite.*
- *All the marginal distributions of a strongly unimodal multivariate distribution are strongly unimodal.*
- *The set of all strongly unimodal distributions is closed under convolutions. If X_1 and X_2 are two independent strongly unimodal random variables, then so are aX_1+b and X_2-X_1, where a, b are real constants.*

Let q_α be the α-th percentile of $p(\theta|\,\mathrm{x}_n)$, i.e. q_α satisfies the equation $\Pr(\theta_{min} < \theta < q_\alpha|\,\mathrm{x}_n) = \alpha$. Before giving the basic properties of HPD intervals, we give a definition of an HPD interval, different from, but equivalent to, the definition given in subsection 2.5.3.

Definition 2.7.2. A highest posterior density (HPD) interval $[\theta_L, \theta_U]$ is an interval containing all points satisfying the equation $p(\theta|\,\mathbf{x}_n) \geq c$, where $c > 0$ is a constant chosen in a way that guarantees the interval $[\theta_L, \theta_U]$ to have a desired length l or coverage $1 - \alpha$. That is $\theta_U - \theta_L = l$ or $\Pr(\theta_L \leq \theta \leq \theta_U|\,\mathbf{x}_n) = 1 - \alpha$.

We will denote by $\mathtt{HPD}(\mathbf{x}_n, n, l)$ an HPD interval of length l and $\mathtt{HPD}(\mathbf{x}_n, n, 1-\alpha)$ an HPD interval of coverage $1-\alpha$, given x and n. If a sufficient statistic is available, it will also be designated by $\mathbf{x}_n$. When a sufficient statistic exists, the posterior density depends on the data only through that statistic (Berger, 1985).

Before completing our general definitions, the following is a simple, but powerful property of HPD intervals. It will be used for case-only designs in chapter 3. The proof is straightforward.

Theorem 2.7.2. *Let Y, be a scalar random variable with density $f_Y(y)$ and $Z = aY + b$, a linear transformation of Y. Let $Lgth_Y(\mathtt{HPD}(1-\alpha))$ (resp. $cvg_Y(\mathtt{HPD}(l))$ be the length (resp. the coverage) of the HPD interval or an equal-tailed of coverage $1-\alpha$ (resp. of length l) for Y. A similar definition holds for $Lgth_Z(\mathtt{HPD}(1-\alpha))$ (resp. $cvg_Z(\mathtt{HPD}(l))$ associated with Z. Then*

$$cvg_Z(\mathtt{HPD}(l)) = cvg_Y\left(\mathtt{HPD}\left(\frac{l}{|a|}\right)\right), \qquad (2.7.2)$$

$$Lgth_Z(\mathtt{HPD}(1-\alpha)) = |a| \times Lgth_Y(\mathtt{HPD}(1-\alpha)). \qquad (2.7.3)$$

Remark 2.7.1. If $Y \sim \mathrm{Be}(a,b)$, then $1-Y$ and Y have HPD intervals with the same length (coverage) given fixed values for coverage (length). Hence, the average length (coverage) for the Beta-Binomial model for a given coverage (length) are equal, irrespective of the choice of $\mathrm{Be}(a,b)$ or $\mathrm{Be}(b,a)$ as the prior distribution. This result will be proved in subsection 3.2.2.

We shall also consider the posterior equal-tailed interval, often called the central posterior interval. This is the most common type of interval encountered in practice, and is, in fact, preferred by Gelman et al., 1995. Equal-tailed $[\theta_U, \theta_L]$ intervals are defined by the equation

$$\Pr(\theta < \theta_L | \mathbf{x}_n) = \Pr(\theta > \theta_U | \mathbf{x}_n) = \frac{\alpha}{2}, \tag{2.7.4}$$

in other words, $\theta_L = q_{\alpha/2}$ and $\theta_U = q_{1-\alpha/2}$. These intervals do not have the same optimality property as HPD intervals, but are easier to compute. More importantly, for fixed coverage $1-\alpha$, $[g(q_{\alpha/2}), g(q_{1-\alpha/2})]$ is the corresponding posterior equal-tailed intervals for the random variable $Y = g(\theta)$, where g is monotonic increasing, an invariance property not shared by HPD intervals. Asymptotically, under certain regularity conditions, HPD and central posterior intervals are equivalent. For sample size for purposes, HPD and equal-tailed intervals often provide similar answers.

Now that we have provided definitions of HPD and equal-tailed intervals, we next give several omnibus algorithms useful for their computation. These derivative-free algorithms require that the posterior density be strictly increasing and then decreasing. Although the omnibus algorithms are slower than derivative-based techniques, they are sometimes the only alternative, for example, when derivative-based procedures fail to converge owing to unavoidably poor choices starting values. Derivative-based algorithms for three specific parameters are discussed in section 3.2.3 along with their drawbacks and challenges.

2.7.2 Omnibus algorithms for the computation of HPD and central posterior intervals

2.7.2.1 Case where the length, l, is fixed

The following bisectional search algorithm finds the two solutions of the equation $f_{\theta|\mathbf{x}_n}(u) = f_{\theta|\mathbf{x}_n}(u+l)$ for u, with l known and then compute the coverage of the interval determined by these two points. Let $\hat{\theta}$ be the mode of $f_{\theta|\mathbf{x}_n}$.

Algorithm 1

Initialisation step: Set $\theta_L = \max(\theta_{\min}, \hat{\theta} - l)$ and $\theta_U = \hat{\theta}$.

Main step:

1. Define $\text{med}_L = (\theta_L + \theta_U)/2$ and $\text{med}_U = \min(\text{med}_L + l, \theta_{\max})$.

2. Compute $f_L = f_{\theta|\mathbf{x}_n}(\text{med}_L)$ and $f_U = f_{\theta|\mathbf{x}_n}(\text{med}_U)$.

3. If $f_U \geq f_L$ then $\theta_L = \text{med}_L - \epsilon$. Otherwise set $\theta_U = \text{med}_L + \epsilon$, where ϵ is the machine precision.

4. Repeat steps 1, 2, and 3 until $|f_L - f_U|$ and $|\theta_U - \theta_L|$ reach the desired precision.

5. Finally compute $\text{cvg}(l) = \Pr(\text{med}_L < p < \text{med}_U)$ as the desired coverage.

For a posterior equal-tailed interval, the equation to solve is $\Pr(\theta < u) = \Pr(\theta > u + l)$. The initial values of θ_L and θ_U are set to $\theta_L = \theta_{\min}$ and $\theta_U = \theta_{\max}$ while $f_L = \Pr(\theta < \text{med}_L)$, and $f_U = \Pr(\theta > \theta_{\max})$. At the end, we compute the coverage as $\Pr(\text{med}_L < p < \text{med}_U) = 1 - (f_L + f_U)$.

The above HPD algorithm requires that the mode be known while the algorithm for equal-tailed intervals requires that $\theta_{\min}$ and $\theta_{\max}$ be finite. When this is not the case and when there exists a monotonic one-to-one transformation $g(\theta)$ with inverse g^{-1} of the random variable of θ that is bounded in both directions by $g_{\min} > -\infty$ and $g_{\max} < \infty$, the above algorithms can still be applied with the following minor adjustments. The initial values are set to be $g_L = g_{\min}$ and $g_U = g_{\max}$. Step 1 is replaced by $\text{med}_L = (g_L + g_U)/2$ and $\text{med}_U = g(g^{-1}(\text{med}_L) + l)$ and step 2 by $f_L = \Pr(\theta < \text{med}_L)$ and $f_U = \Pr(\theta > \text{med}_U)$.

2.7.2.2 Case where the coverage level is fixed at $1 - \alpha$

This subsection is concerned with the determination of the two solutions θ_L and θ_U such that $f_{\theta|\mathrm{x}_n}(\theta_L) = f_{\theta|\mathrm{x}_n}(\theta_U)$ and $\Pr(\theta_L < \theta < \theta_U) = 1 - \alpha$ for the HPD intervals. When $\theta_{\min}$ and $\theta_{\max}$ are finite, the following algorithm applies.

Algorithm 2

Initialisation step: Set $\text{len}_L = 0.0$ and $\text{len}_U = \theta_{\max} - \theta_{min}$.

Main step:

1. Let $l = (\text{len}_L + \text{len}_U)/2$.
2. Apply Algorithm 1 to obtain the coverage $\text{cvg}(l)$ in step 5.
3. If $\text{cvg}(l) < 1 - \alpha$ then $\text{len}_L = l - \epsilon$. Otherwise set $\text{len}_U = l + \epsilon$.
4. Repeat the steps 2, 3, and 4 until $|\text{len}_L - \text{len}_U|$ and $|\text{cvg}(1) - (1 - \alpha)|$ reach the desired precision.
5. Compute length $= \text{len}_U$.

As might be expected, this algorithm converges very slowly, but is the only viable alternative in some instances. It can be refined to fit the case where the random variable is unbounded in both directions and where numerical computations of percentiles are unavailable. One starts with any possible value of l, and multiplies it by 2 if the coverage is not attained until we have one such bound, len_U, and the main step can then proceed.

Another alternative is to turn to simulation-based Monte Carlo algorithms. These require being able to simulate variables from the distribution $f(\theta|\,\mathbf{x}_n)$ and rely on the posterior distribution's being strictly unimodal.

2.7.2.3 Monte Carlo algorithms

Simulation-based algorithms for HPD intervals, discussed in Tanner (1996), are rewritten here to give the following algorithms for fixed length and fixed coverage situations. Let $\theta_1 < \cdots < \theta_M$ be M random values from $f_{\theta|\mathbf{x}_n}$. Let $(f_1, \cdots, f_M)$ where $f_i = f_{\theta|\mathbf{x}_n}(\theta_i)$, $i = 1, \cdots, M$, $c_{\max} = \max(f_i)$, and $c_{\min} = \min(f_i)$. Algorithm 3 and 4 assume that $f_{\theta|\mathbf{x}_n}$ is unimodal.

Algorithm 3 (fixed length)

1. Set $c = (c_{\text{min}} + c_{\text{max}})/2$ and find the vectors I_θ of indexes $i = 1, \cdots, M$ for which $f_i \geq c$. Let $L = \min_i I_\theta$ and $U = \max_i I_\theta$. Set $\text{len}_\theta = \theta_U - \theta_L$.
2. If $\text{len}_\theta < l$ then set $c_{\text{min}} = c - \epsilon$. Otherwise set $c_{\text{max}} = c + \epsilon$,
3. Repeat 1 and 2 until $|\text{len}_\theta - l|$ and $|c_{\text{max}} - c_{\text{min}}|$ reach the desired precision. Often this step can not be done exactly so one has to set an upper bound on the number of iterations.
4. Compute $(U - L + 1)/M$ as an estimate of the coverage.

Algorithm 4 (fixed coverage)

1. Set $c = (c_{\text{min}} + c_{\text{max}})/2$ and find the vectors I_θ of indexes $i = 1, \cdots, M$ for which $f_i \geq c$. Let $L = \min_i I_\theta$ and $U = \max_i I_\theta$.
2. Set $\text{cvg}_\theta = (U - L + 1)/M$.
3. If $\text{cvg}_\theta < 1 - \alpha$ then set $c_{\text{min}} = c - \epsilon$. Otherwise set $c_{\text{max}} = c + \epsilon$,
4. Repeat 1, 2 and 3 until $|\text{cvg}_\theta - (1 - \alpha)|$ and $|c_{\text{max}} - c_{\text{min}}|$ reach the desired precision. The same comments as in step 3 of Algorithm 3 apply here.
5. Compute $\theta_U - \theta_L$ as an estimate of the length.

We now turn to approximation of HPD and equal-tailed left and right tails. The approximations are important because they can reduce the burden of computation of the exact credible intervals.

2.7.3 Approximation of HPD and equal-tailed left and right tails when the coverage is given

First order approximations for both HPD and equal-tailed intervals are based on the asymptotic normality of the posterior distribution. Conditions under which such results holds are described in most books on Bayesian statistics. Thus,

$$\text{Lgth}_{\theta|\mathbf{x}_n} \approx 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(\theta|\,\mathbf{x}_n)}, \tag{2.7.5}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$-th percentile of the normal distribution. Although these first order approximations are often sufficient, they have three fundamental limitations. First, they do not distinguish between the two credible intervals of interest, HPD and equal-tailed intervals. Since one goal of this thesis is to compare sample sizes from both types of credible interval, one needs two distinct approximations. Second, for unbounded random variables, this limiting result holds only for very large sample sizes, suggesting approximations with higher order terms may be useful. Third, and possibly most critical, we have the fact that posterior variances are not always defined, and for those distributions we are simply left with no approximations for these credible intervals. Fortunately, higher order of approximations have been derived by various authors (Welch and Peers, 1963; Peers, 1968; Mukerjee and Dey, 1993; Severini, 1991).

The following large sample approximations hold under certain regularity conditions (see Peers, 1968). Let $\hat{\theta}$ be the posterior mode and $z = z_{1-\alpha/2}$ for a given coverage level $1-\alpha$ and let

$$w_j = n^{-\frac{j}{2}} \left\{ \frac{\partial^j \log p(\mathbf{x}_n|\hat{\theta})}{\partial \hat{\theta}^j} \right\}, \qquad (j = 2, 3, 4). \qquad (2.7.6)$$

Let $[\theta_L^{(1)}, \theta_U^{(1)}]$ and $[\theta_L^{(3)}, \theta_U^{(3)}]$ be the equal-tailed and HPD intervals.

2.7.3.1 HPD intervals

We have

$$\begin{aligned}(-nw_2)^{\frac{1}{2}}(\theta_U^{(3)} - \hat{\theta}) \approx z &+ \frac{z^2}{6} w_3(-w_2)^{-\frac{3}{2}} + (-nw_2)^{-\frac{1}{2}} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\} \\ &+ \frac{z^3 + 3z}{24} w_4(-w_2)^{-2} + \frac{z}{-2nw_2} \left\{ \frac{d^2 \log p(\hat{\theta})}{d\hat{\theta}^2} \right\} \\ &+ \frac{5(z^3 + 3z)}{72} w_3^2(-w_2)^{-3} + \frac{z}{2\sqrt{n}} w_3(-w_2)^{-2} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\}.\end{aligned} \qquad (2.7.7)$$

2.7.3.2 Equal-tailed intervals

We have

$$\begin{aligned}(-nw_2)^{\frac{1}{2}}(\theta_U^{(1)} - \hat{\theta}) \approx z &+ \frac{z^2 + 2}{6} w_3(-w_2)^{-\frac{3}{2}} + (-nw_2)^{-\frac{1}{2}} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\} \\ &+ \frac{z^3 + 3z}{24} w_4(-w_2)^{-2} + \frac{z}{-2nw_2} \left\{ \frac{d^2 \log p(\hat{\theta})}{d\hat{\theta}^2} \right\} \\ &+ \frac{5z^3 + 19z}{72} w_3^2(-w_2)^{-3} + \frac{z}{2\sqrt{n}} w_3(-w_2)^{-2} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\}.\end{aligned} \qquad (2.7.8)$$

$\theta_L^{(1)}$ and $\theta_L^{(3)}$ are obtained by replacing z by $-z$ in equation (2.7.7) and (2.7.8).

2.7.3.3 Approximate lengths

As a consequence of the approximations (2.7.7) and (2.7.8), we have

$$(-nw_2)^{\frac{1}{2}}\left\{\theta_U^{(3)}-\theta_L^{(3)}\right\} \approx 2z+\frac{z^3+3z}{12}w_4(-w_2)^{-2}+\frac{z}{-nw_2}\frac{d^2\log p(\hat{\theta})}{d\hat{\theta}^2}$$
$$+\frac{5(z^3+3z)}{36}w_3^2(-w_2)^{-3}+\frac{z}{\sqrt{n}}w_3(-w_2)^{-2}\frac{d\log p(\hat{\theta})}{d\hat{\theta}}, \tag{2.7.9}$$

and

$$(-nw_2)^{\frac{1}{2}}\left(\theta_U^{(1)}-\theta_L^{(1)}\right) \approx 2z+\frac{z^3+3z}{12}w_4(-w_2)^{-2}+\frac{z}{-nw_2}\frac{d^2\log p(\hat{\theta})}{d\hat{\theta}^2}$$
$$+\frac{5z^3+19z}{36}w_3^2(-w_2)^{-3}+\frac{z}{\sqrt{n}}w_3(-w_2)^{-2}\frac{d\log p(\hat{\theta})}{d\hat{\theta}}. \tag{2.7.10}$$

Therefore, the positive difference between the lengths of posterior equal-tailed and HPD intervals is

$$\left\{\left(\theta_U^{(1)}-\theta_L^{(1)}\right)-\left(\theta_U^{(3)}-\theta_L^{(3)}\right)\right\} \approx \frac{z}{9}(-nw_2)^{-\frac{1}{2}}w_3^2(-w_2)^{-3}. \tag{2.7.11}$$

Chapter 3

Bayesian interval-based sample size determination for estimating risk ratios for exposure-only designs and odds ratios for case-only designs

In this chapter, we investigate Bayesian sample size determination for case-only and exposure-only designs. Let $p_0 = \Pr(D = 1|E = 0)$ and $p_1 = \Pr(D = 1|E = 1)$ be the "success" probabilities, conditional on exposure status, in a cohort study and $p_0' = \Pr(E = 1|D = 0)$ and $p_1' = \Pr(E = 1|D = 1)$ be the "success" probabilities, conditional on disease status, in a case-control study where D and E are the disease and exposure variables, as described in section 2.3. Let $q_i = 1 - p_i$, $i = 0, 1$ and $q'_i = 1 - p'_i$, $i = 0, 1$ represent the "failure" probabilities. For notational convenience, when the context is clear we drop the prime superscripts. We are particularly interested in three parameters, namely the risk ratio, $R = \frac{p_1}{p_0}$, in cohort settings, and the odds ratio, $\psi_e = \frac{p_1 q_0}{p_0 q_1}$, and the log-odds ratio, $\log(\psi_e)$, in case-control settings. For both cohort and case-control settings, we assume that p_0 is known. Therefore, we do not need to sample observations from the non-exposed subjects or the control subjects. For these reasons, these designs will be referred to in a broad sense as exposure-only (often known as exposure-series) and case-only designs (often known as case-series), or case-exposure designs according to Hogue et al., 1986. This definition of case-only designs is different from the current terminology used in genetics; there case-only designs arise in the two sample problem. We discuss this matter in chapter 4. We denote by n the sample size of interest in both designs ($n = m_1$ in the cohort design and $n = n_1$ in the case-control design).

Sample sizes for the parameters of interest are investigated in two particular contexts. In the first scenario, we do not impose any particular restrictions on p_1 whereas in the second scenario, we allow additional information of the type $p_1 < p_0$ or $p_1 > p_0$. These, in turn, imply that $R < 1$ and $\psi_e < 1$ or $R > 1$ and $\psi_e > 1$. The log-odds ratio will be studied only in the first scenario, since on the log-scale differences between the cases $R < 1$ and $R > 1$ essentially disappear. Bayesian sample size determination will be based on both HPD and equal-tailed credible intervals.

This chapter starts with a presentation of several sample size criteria in section 3.1. Section 3.2 discusses sample size calculations when p_1 is unrestricted, while section 3.3 deals with the case when p_1 is restricted. Section 3.2 is divided into 8 subsections. In subsection 3.2.1, it is pointed out that there is a bridge between the sample size provided by an exposure or a case-only designs and the sample size provided by a one sample problem. In subsection 3.2.2, we briefly introduce preliminary results about the pre-posterior predictive distribution. In subsection 3.2.3, we discuss some derivative-based algorithms for the exact computation of HPD intervals for R, ψ_e and $\log(\psi_e)$, when the coverage is given, for the determination of the exact sample sizes. In subsection 3.2.4, we present approximate methods to estimate the tails of the HPD and equal-tailed intervals for the calculation of approximate sample sizes. This is followed by two subsections on Monte Carlo procedures for finding sample sizes. Finally, we use all of these preliminary results to derive methods for calculating sample sizes. In section 3.3, we first present the prior-likelihood models used for the case when p_1 is restricted along with the resulting posterior distribution. We then present methods using Monte Carlo simulations.

3.1 Bayesian criteria for sample size

Every sample size calculation problem begins with the definition of a criterion. In the chapter, the following criteria will be used.

The first criterion is the k-th average coverage criterion **ACC**$_k$, ($1 < k < \infty$ integer). Here, the statistician fixes the desired length of the HPD or the equal-tailed interval at l. The posterior coverage probabilities of these credible intervals are then averaged with respect to the pre-posterior distribution $p_{X_n}(x_n)$ under the L_k-norm. This leads to searching for the minimum n such that

$$\left(\int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{\int_{\theta \in \mathtt{HPD}(\mathbf{x}_n, n, l)} p(\theta | \, \mathbf{x}_n) \, d\theta \right\}^k p_{X_n}(\mathbf{x}_n) \, d\mathbf{x}_n \right)^{1/k} \geq 1 - \alpha, \qquad (3.1.1)$$

where $\mathcal{X}_n = \{0, 1, \cdots, n\}$. The second measure, the k-th average length criterion ($\mathbf{ALC}_k$), fixes the coverage level of each HPD or equal-tailed interval at $1 - \alpha$, and averages their lengths over the distribution $p_{X_n}(x_n)$. This leads to seeking the minimum n such that

$$\left(\int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{\int_{\theta \in \mathtt{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\}^k p_{X_n}(\mathbf{x}_n) \, d\mathbf{x}_n \right)^{1/k} \leq l. \qquad (3.1.2)$$

These two measures are obvious generalizations of the **ALC** and **ACC** ($k = 1$) by Joseph et al., 1995 as discussed in chapter 2, and serve to unify various isolated criteria. For instance, PGT-(i) and PGT-(ii) (defined in subsection 2.5.3), are now related to the **WOC** and **ALC**. This link has not been previously recognized. At the risk of digressing briefly, we establish in the proposition below the link between $\mathbf{ALC}_2$ and PGT-(ii).

Proposition 3.1.1. *Under the regularity conditions for the asymptotic normality of the posterior distributions, the $\boldsymbol{ALC}_2$ for fixed length l is asymptotically equivalent to the average posterior variance criterion, PGT-(ii), when the target level of precision, ϵ, is taken to be $\frac{l^2}{4\, z_{1-\alpha/2}^2}$.*

Proof. Without loss of generality, we can assume that the $\mathbf{ALC}_2$ consists of solving the equation

$$\int_{\mathcal{X}_n} \left\{ \int_{\theta \in \mathtt{HPD}(\mathrm{x}_n, n, 1-\alpha)} d\theta \right\}^2 p_{X_n}(\mathrm{x}_n)\, d\mathrm{x}_n \;=\; l^2 \tag{3.1.3}$$

with n continuous rather than n discrete, as in (3.1.2). Under the same regularity conditions which assure asymptotic normality of the posterior distribution, we have

$$\begin{aligned}\int_{\mathcal{X}_n} \left\{ \int_{\theta \in \mathtt{HPD}(\mathrm{x}_n, n, 1-\alpha)} d\theta \right\}^2 p_{X_n}(\mathrm{x}_n)\, dx \;&\approx\; 4z^2_{1-\alpha/2} \int_{\mathcal{X}_n} \mathbf{Var}(\theta|\,\mathrm{x}_n) p_{X_n}(\mathrm{x}_n)\, d\mathrm{x}_n \\ &=\; 4z^2_{1-\alpha/2}\, \mathbf{E}_{X_n}[\mathbf{Var}(\theta|\,\mathrm{x}_n)].\end{aligned}$$

Therefore, equation (3.1.3) is asymptotically equivalent to

$$\mathbf{E}_{X_n}[\mathbf{Var}(\theta|\,\mathrm{x}_n)] \;=\; \frac{l^2}{4\,z^2_{1-\alpha/2}}.$$

□

Expected coverage and length, as given by the left sides of equations (3.1.1) and (3.1.2) when k=1, change at roughly the rate of $\frac{1}{\sqrt{n}}$, and hence slowly for large n. An often applied technique under such circumstances when increasing the differences between consecutive values of n, especially for expected length, is to raise all lengths to a power k before averaging them. This leads, not surprisingly, to a different solution, compared with that obtained when $k = 1$. Indeed, the content of Proposition 3.1.2 below is that the $\mathbf{ALC}_k$ leads to a larger sample size as k increases, whereas the opposite trend occurs when working with the $\mathbf{ACC}_k$.

Proposition 3.1.2. *Let $n(k, 1-\alpha, l)$ and $m(k, l, 1-\alpha)$ denote the optimal* *sample sizes under the* $\boldsymbol{ALC_k}$ *and the* $\boldsymbol{ACC_k}$*, respectively. Then*

$$\begin{aligned} n(k+1, 1-\alpha, l) &> n(k, 1-\alpha, l), \quad \textit{and,} \qquad (3.1.4)\\ m(k+1, l, 1-\alpha) &< m(k, l, 1-\alpha). \qquad (3.1.5) \end{aligned}$$

Proof. The proof of this proposition is entirely based on the natural ordering of the L_k-norm. Let $n = n(k+1, 1-\alpha, l)$. Then

$$\left(\int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \mathrm{HPD}(\mathbf{x}_n, n, l)} p(\theta | \, \mathbf{x}_n) \, d\theta \right\}^{k+1} p_{X_n}(\mathbf{x}_n) \, d\mathbf{x}_n \right)^{1/(k+1)} \leq l.$$

This implies also that n satisfies

$$\left(\int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \mathrm{HPD}(\mathbf{x}_n, n, l)} p(\theta | \, \mathbf{x}_n) \, d\theta \right\}^{k} p_{X_n}(\mathbf{x}_n) \, d\mathbf{x}_n \right)^{1/k} \leq l, \qquad (3.1.6)$$

because the L_k-norm increases monotonically as k increases. Therefore $n(k+1, l, 1-\alpha)$ is larger than the smallest bound of all n satisfying equation 3.1.6, that is, $n(k, l, 1-\alpha)$. Similarly for the $\mathbf{ACC}_k$, but with inequalities reversed. □

In general, sample size criteria should match the inferential techniques used in the analysis. Therefore, **ALC** and **ACC** are more natural than $\mathbf{ALC}_k$ and $\mathbf{ACC}_k$ for $k > 1$. Nevertheless, as k-moments criteria, the $\mathbf{ALC}_k$ and $\mathbf{ACC}_k$ remain of interest since, as we have seen, they can be related to other sample size criteria that have been proposed.

The third criterion is the worst outcome criterion, **WOC**, defined in subsection 2.5.3.3. We temporarily rename this criterion to be the worst coverage outcome criterion, **WCOC**, to emphasize its link with the **ACC**. In a similar spirit, define a worst length outcome criterion, **WLOC**, a close relative of the **ALC**, defined as follows: The **WLOC** requires that we find the minimum n that satisfies the inequality

$$\sup_{\mathbf{x}_n \in \mathcal{S}_n} \left\{ \int_{\theta \in \mathrm{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\} \leq l, \tag{3.1.7}$$

where $\mathcal{S}_n \subset \mathcal{X}_n$, is a pre-specified $100(1-\gamma)\%$ pre-posterior predictive credible region for $\mathbf{x}_n$.

Proposition 3.1.3 shows that the **WOC** is not only a coverage criterion but also a length criterion.

Proposition 3.1.3. *When $\mathcal{S}_n = \mathcal{X}_n$, **WCOC**= **WLOC**= **ALC**$_\infty$.*

Proof. It is clear that **WLOC** = **WCOC** since both criteria lead to choosing the minimum value of $\mathcal{S}$ where

$$\mathcal{S} = \left\{ n : \sup_{\mathbf{x}_n \in \mathcal{X}_n} \left(\int_{\mathrm{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right) \leq l, \inf_{\mathbf{x}_n \in \mathcal{X}_n} \left(\int_{\mathrm{HPD}(\mathbf{x}_n, n, l)} p(\theta | \, \mathbf{x}_n, n) d\theta \right) \geq 1 - \alpha \right\}.$$

Let $w_1, w_2, \cdots, w_m$ and $a_1, a_2, \cdots, a_m$ be a sequence of m non-negative real numbers with $\sum_{i=1}^m w_i = 1$ and $\sup a_i < \infty$. Then,

$$\lim_{k \longrightarrow \infty} \left(\sum_{i=1}^m w_i a_i^k \right)^{1/k} = \sup a_i, \text{ for} \tag{3.1.8}$$

let j be the index i such that $a_j = \sup a_i$. Then, we have

$$w_j^{1/k} a_j \leq \left(\sum_{i=1}^m w_i a_i^k \right)^{1/k} \leq a_j,$$

and equation (3.1.8) follows as we take the limit on $k \longrightarrow \infty$. A straightforward application of equation (3.1.8) to our problem where the pre-posterior predictive distribution is always a discrete mass function leads to

$$\sup_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \mathrm{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\} = \lim_{k \longrightarrow \infty} \left(\int_{\mathcal{X}_n} \left\{ \int_{\theta \in \mathrm{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\}^k p_{X_n}(\mathbf{x}_n) \, d\mathbf{x}_n \right)^{1/k},$$

for any given n, and the proof is complete. □

Proposition 3.1.4, which complements Proposition 3.1.1, draws a link between the **WOC** and PGT-(i).

Proposition 3.1.4. *Under the regularity conditions for the asymptotic normality of the posterior distributions, the* **WOC** *for fixed length l and coverage $1-\alpha$ is asymptotically equivalent to the maximum posterior variances criterion, PGT-(i), when the target level of precision, ϵ, is taken to be $\dfrac{l^2}{4\,z^2_{1-\alpha/2}}$.*

Proof. Under the regularity conditions which assure asymptotic normality of the posterior distribution, on a continuous scale, the **WOC** solves the equation

$$\begin{aligned} l^2 = \left(\sup_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \text{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\} \right)^2 &= \sup_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \text{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\}^2 \\ &\approx 4 z^2_{1-\alpha/2} \sup_{\mathbf{x}_n \in \mathcal{X}_n} \mathbf{Var}(\theta | \, \mathbf{x}_n). \end{aligned}$$

□

In the same vein, we may define the median coverage outcome criterion, **MCOC**, and the median length outcome criterion, **MLOC**, as respectively, the minimum n such that

$$\text{med}_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \text{HPD}(\mathbf{x}_n, n, l)} p(\theta | \, \mathbf{x}_n) \, d\theta \right\} \geq 1 - \alpha, \tag{3.1.9}$$

and

$$\text{med}_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \int_{\theta \in \text{HPD}(\mathbf{x}_n, n, 1-\alpha)} d\theta \right\} \leq l. \tag{3.1.10}$$

Although we define all sample size criteria in terms of HPD intervals, many practitioners prefer to use equal-tailed intervals in place of HPD intervals. Indeed, equal-tailed intervals are easier to compute especially from a Monte

Carlo simulation viewpoint. In our experience, sample sizes based on equal-tailed intervals are often much more computationally efficient than those using HPD intervals. Second, samples size based on equal-tailed intervals can often be used as good estimates of those based on HPD intervals. Thirdly, sample sizes based on equal-tailed intervals will always be equal to or overestimate sample sizes based on HPD intervals.

Now that we have clearly stated our sample size criteria, we are ready to proceed with sample size determination. Although the sample size calculation now reduces to a purely computational problem as is often the case with modern Bayesian inference, there are many practical obstacles to overcome. In section 3.2, we address the question of sample size determination from three perspectives: exact computation, approximate computation and Monte-Carlo estimates. These three approaches are described in subsections 3.2.3, 3.2.4, 3.2.5, 3.2.6 and 3.2.7. Before heading to these subsections, we first establish a straightforward relation between sample size determination for estimating R in exposure-only settings and ψ_e or $\phi_e = \log(\psi_e)$ in case-only settings and sample size calculation for estimating $p_1, \omega = \dfrac{p_1}{1-p_1}$, and $\phi = \log(\omega)$ in the one sample problem (in subsection 3.2.1), respectively. Given the bridge between case-only or exposure-only problems and one sample problems, we may transfer attention to that of finding sample sizes for estimating the parameters p_1, ω, and ϕ in the one sample problem. In order to compute sample sizes, we first need to discuss the common pre-posterior distribution corresponding to R, ψ_e or ϕ_e. We do this in subsection 3.2.2.

3.2 Sample size when p_1 is unrestricted

3.2.1 Bridge between exposure and case-only designs and the one sample problem

The following important theorem tells us that from a theoretical point of view, within the one sample problem framework, the sample size problems for inference about the proportion, p_1, the odds $\omega = \frac{p_1}{1-p_1}$, and the log-odds $\phi = \log(\omega) = \log\left(\frac{p_1}{1-p_1}\right)$ are simply related to those of R, ψ_e and $\log(\psi_e)$, respectively.

Theorem 3.2.1. *Let Z be a scalar random variable, $Z = cX + d$ be a linear transformation of X, and $1 < k < \infty$ an integer. Let $n_Z(k, l, 1-\alpha)$ and $n_X(k, l, 1-\alpha)$ be the sample sizes for estimating X and Z respectively, using the **ACC_k**. Similarly, define $m_Z(k, 1-\alpha, l)$ and $m_X(k, 1-\alpha, l)$ to be the sample sizes for estimating X and Z using the **ALC_k**. The credible intervals used are assumed either to be* HPD *or equal-tailed intervals. Then*

$$n_Z(k, l, 1-\alpha) = n_X\left(k, \frac{l}{|c|}, 1-\alpha\right) \tag{3.2.1}$$

$$m_Z(k, 1-\alpha, l) = m_X\left(k, 1-\alpha, \frac{l}{|c|}\right). \tag{3.2.2}$$

*Similar results hold for the criteria **WOC**, **MCOC** and **MLOC**.*

Proof. The proof of these results is a straightforward application of Theorem 2.7.2. We will only prove equation (3.2.2). According to equation (2.7.2), we have the following result for the length of the HPD interval:

$$\text{Lgth}_Z\left(\text{HPD}(1-\alpha)\right) = |c| \times \text{Lgth}_X\left(\text{HPD}(1-\alpha)\right).$$

Therefore

$$\mathbf{E}\left[\mathrm{Lgth}_Z^k\left(\mathtt{HPD}(1-\alpha)\right)\right] = |c|^k \times \mathbf{E}\left[\mathrm{Lgth}_X^k\left(\mathtt{HPD}(1-\alpha)\right)\right],$$

and the result follows. □

Corollary 3.2.2 below is the bridge between the exposure-only and case-only designs and the one sample problem.

Corollary 3.2.2. *For case-only and exposure-only designs, with* p_0 *fixed, the sample sizes for estimating* R, ψ_e, *and* $\log(\psi_e)$ *may be obtained from the single parameter problem by setting,* $d = 0$ *and* $c = \dfrac{1}{p_0}$, $c = \dfrac{q_0}{p_0}$, $c = 1$, *respectively in* (3.2.1) *and* (3.2.2).

3.2.2 Pre-posterior predictive distribution

Since all our sample size criteria require averaging over the pre-posterior predictive distribution we turn our attention to its computation.

From now until the end of section 3.2, we assume the following prior-likelihood model: p_1 is distributed as $\mathbf{Be}(a, b)$, $a, b > 0$, and $X_n | \, p_1$ is $\mathbf{Bin}(n, p_1)$.

It is easily seen that the prior-likelihood model discussed above yields the Beta-Binomial pre-posterior predictive distribution irrespective of the parameter p_1, $\omega = \dfrac{p_1}{1-p_1}$, or $\phi = \log(\psi)$ under consideration. We use the notation $X_n \sim \mathcal{BB}(n, a, b)$. This distribution has probability mass function

$$p_{X_n}(x_n|n,a,b) = \binom{n}{x_n}\frac{\mathbf{Be}(a+x_n, n+b-x_n)}{\mathbf{Be}(a,b)}, \quad x_n = 0, 1, \cdots, n. \quad (3.2.3)$$

Setting $a = b = 1$ results in the uniform distribution on $\{0, 1, \cdots, n\}$. The expected value and variance of X_n, given a and b, are

$$\begin{aligned}\mathbf{E}(X_n) &= n\mathbf{E}_{p_1}(p_1) = n\frac{a}{a+b}, &(3.2.4)\\ \mathbf{Var}(X_n) &= n^2\mathbf{Var}_{p_1}(p_1) + n\mathbf{E}_{p_1}[p_1(1-p_1)]\\ &= \frac{abn^2}{(a+b)^2(a+b+1)} + \frac{abn}{(a+b)(a+b+1)}. &(3.2.5)\end{aligned}$$

An important property of the Beta-Binomial family of distributions, which will be used later, is

$$p_{X_n}(x_n|n,a,b) = p_{X_n}(n-x_n|n,b,a), \qquad x_n = 0,1,\cdots,n. \qquad (3.2.6)$$

Equation (3.2.6), combined with the fact that $1-p_1 \sim \mathbf{Be}(b,a)$, gives rise to the following symmetrical property given by Theorem 3.2.3. It is useful because we are sometimes not able to compute the sample size for a given pair (a,b) but are able to compute the corresponding sample size for the pair (b,a).

Theorem 3.2.3. *Let $n_k(a,b,l,1-\alpha)$ and $m_k(a,b,1-\alpha,l)$ be the optimal sample sizes for estimating the proportion p_1 using the $\boldsymbol{ALC_k}$ and $\boldsymbol{ACC_k}$ $(1 \leq k \leq \infty)$ criteria. Then*

$$\begin{aligned}n_k(l,a,b,1-\alpha) &= n_k(l,b,a,1-\alpha), \textit{ and} &(3.2.7)\\ m_k(1-\alpha,a,b,l) &= m_k(1-\alpha,b,a,l). &(3.2.8)\end{aligned}$$

Equivalent results hold for the criteria ***WCOC***, ***WLOC***, ***MCOC*** *and* ***MLOC***.

Similar properties hold when estimation of the parameter $\phi = \log\left(\frac{p_1}{1-p_1}\right) = -\log\left(\frac{1-p_1}{p_1}\right)$, is of interest.

To recap, sample size determination is primarily based on computing exactly, approximately, or using Monte Carlo simulations, the following criterion functions associated with the various criteria of interest:

$$\begin{aligned}
\mathrm{acc}_k(n,a,b) &= \left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_{\theta\in\mathtt{HPD}(\mathbf{x}_n,n,l)} p(\theta|\,\mathbf{x}_n,a,b)\,d\theta\right\}^k p_{X_n}(\mathbf{x}_n|\,n,a,b)\,dx\right)^{1/k},\\
\mathrm{alc}_k(n,a,b) &= \left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_{\theta\in\mathtt{HPD}(\mathbf{x}_n,n,1-\alpha)} d\theta\right\}^k p_{X_n}(\mathbf{x}_n|\,n,a,b)\,dx\right)^{1/k},\\
\mathrm{woc}(n,a,b) &= \inf_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_{\theta\in\mathtt{HPD}(\mathbf{x}_n,n,l)} p(\theta|\,\mathbf{x}_n,a,b)\,d\theta\right\},\\
\mathrm{mcoc}(n,a,b) &= \mathrm{med}_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_{\theta\in\mathtt{HPD}(\mathbf{x}_n,n,l)} p(\theta|\,\mathbf{x}_n)\,d\theta\right\},\\
\mathrm{mloc}(n,a,b) &= \mathrm{med}_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_{\theta\in\mathtt{HPD}(\mathbf{x}_n,n,1-\alpha)} d\theta\right\},
\end{aligned}$$

where θ is either p_1, ω or ϕ and where $p(\theta|\,\mathbf{x}_n)$ is the posterior distribution of θ. We use the simpler notation $\mathrm{alc}(n,a,b)$ and $\mathrm{acc}(n,a,b)$ when $k=1$. Then, using a bisectional search strategy, we find n such that, for example, $\mathrm{alc}_k(n-1,a,b) > l$ and $\mathrm{alc}_k(n,a,b) \leq l$. In order to compute these criterion functions exactly, we need to compute HPD intervals exactly in order to determine their coverage or length for a given length or coverage, respectively.

3.2.3 Exact computation of HPD intervals

When the coverage is fixed, the calculation of HPD intervals is very intensive. It is therefore important to develop fast entirely derivative-based approaches to the exact computation of HPD intervals for p_1, ω and ϕ. As far as equal-tailed intervals are concerned, we take advantage of the availability of derivative-based algorithms in most statistical and mathematical software packages to find the required quantiles. On the contrary, when the length is fixed, the omnibus algorithm 1 for the calculation of HPD and equal-tailed intervals presented in subsection 2.7.2.1 is reasonably fast, so that there is no need to consider this case further. Below we show how fast derivative-based algorithms can be used to compute HPD intervals when the coverage is given.

The posterior density of the parameter of interest p_1 is well-known by conjugacy, to be

$$f_{p_1}(p_1) = K(\mathrm{x}_n, n, a, b)\, p^{a+\mathrm{x}_n-1}\,(1-p)^{n+b-\mathrm{x}_n-1},\ 0<p<1, \tag{3.2.9}$$

where $\dfrac{1}{K(\mathrm{x}_n,n,a,b)} = \mathbf{B}(a+\mathrm{x}_n, n+b-\mathrm{x}_n),\ \mathrm{x}_n \in \{0,1,\cdots,n\},\ a,b>0,$ and where again $\mathbf{B}(.,.)$ is the Beta function. The others follow by simple transformation of variables:

$$\begin{aligned} f_\omega(\omega) &= K(\mathrm{x}_n,n,a,b)\,\frac{\omega^{a+\mathrm{x}_n-1}}{(1+\omega)^{n+a+b}}, & 0<\omega<\infty, \qquad (3.2.10)\\ f_\phi(\phi) &= K(\mathrm{x}_n,n,a,b)\,\frac{e^{(a+\mathrm{x}_n)\,\phi}}{(1+e^{\phi})^{n+a+b}}, & -\infty<\phi<\infty. \quad (3.2.11)\end{aligned}$$

It is well known (Dharmadhikari and Joag-dev, 1988) that the variable p_1 is strongly unimodal when $a+\mathrm{x}_n, n+b-\mathrm{x}_n \geq 1$ while the variable ω is unimodal irrespective of x_n. For ϕ, we have $\dfrac{\partial^2 f_\phi(\phi)}{\partial\phi^2} = -\dfrac{(n+a+b)e^{\phi}}{(1+e^{\phi})^2}$. Therefore ϕ is strongly unimodal and hence unimodal irrespective of x_n. Let F_{p_1}, F_ω and F_ϕ, denote the respective distribution functions of p_1, ω, and ϕ. It has been suggested by Hashemi et al., 1997, to compute HPD intervals by means of the derivative-free Nelder-Mead algorithm by forcing, for instance, the function

$$G(t,z) = \left|F_{p_1}(z) - F_{p_1}(t) - (1-\alpha)\right| + \left|f_{p_1}(z) - f_{p_1}(t)\right| \tag{3.2.12}$$

to equal zero. The two such subroutines **DUVMGS** and **DUMPOL** (Nelder and Murray, 1965) are available in the **ISML library**. Obvious starting values for the numerical solution of all the equations given here, are $q_{\alpha/2}$ and $q_{1-\alpha/2}$, the percentiles of f_{p_1}. Although slow, these subroutines are suitable if one has to compute only a few HPD intervals. When convergence is not reached due to poor starting values, the researcher is left with the option of trying various initial values and eventually convergence will be reached. Such a suggestion however, is not efficient for large n, and an automatic derivative-based algorithm must, therefore, be found.

In order to implement such an algorithm and avoid the non-differentiability of the absolute value function, we propose the following adjustments to the proposition by Hashemi et al., 1997. Instead of the objective function (3.2.12), use

$$G(t,z) = \eta\big[F_{p_1}(z) - F_{p_1}(t) - (1-\alpha)\big]^2 + \big[f_{p_1}(z) - f_{p_1}(t)\big]^2, \qquad (3.2.13)$$

where $\eta > 0$. Note that f_{p_1} can be replaced by $\log f_{p_1}$ because the equalities $f_{p_1}(z) = f_{p_1}(t)$ and $\log f_{p_1}(z) = \log f_{p_1}(t)$ are equivalent. The modification which introduces the square rather than the absolute value, includes another modification through the introduction of the multiplicative constant, η. This constant is introduced to balance the large difference between $|F_{p_1}(z) - F_{p_1}(t) - (1-\alpha)|$ and $|f_{p_1}(z) - f_{p_1}(t)|$, which is amplified by the square. Various derivative-based subroutines are available from the **ISML library** but only three subroutines turned out to be satisfactory with the objective function on the right hand side of (3.2.13). The first two are the subroutines **DUMINF** and **DBCONF** (Gill and Murray, 1976; Denis and Schnabel, 1983) for which η was held fixed at 100. One should note that the choice of η is critical since we might reach convergence but towards the wrong values. It is, therefore, also important to consider slow omnibus algorithms to make sure that convergence is toward the proper values. The third subroutine is **DNEQNF** (More et al., 1980). Although **DUMINF** is easily the fastest subroutine, we prefer **DNEQNF** since it is more accurate. Indeed, accuracy is important, for the reduction of computation errors when averaging over the pre-posterior distribution. We describe what each subroutine does in appendix H.

To use (3.2.13) and to accommodate the computation of the exact HPD intervals for the parameters ω and ϕ, we take advantage of the following observation. Let $[\omega_1, \omega_2]$ (resp. $[\phi_1, \phi_2]$), denote an HPD interval for ω (resp. ϕ), and set $u_1 = \dfrac{\omega_1}{1+\omega_1}$, $u_2 = \dfrac{\omega_2}{1+\omega_2}$ and $v_1 = \dfrac{\exp(\phi_1)}{1+\exp(\phi_1)}$, $v_2 = \dfrac{\exp(\phi_2)}{1+\exp(\phi_2)}$. Then, u_1, u_2, v_1, v_2 satisfies the equations

$$\begin{cases} g_1(u_1) = (u_1)^{a+\mathrm{x}_n-1}(1-u_1)^{n+b-x_n+1} = (u_2)^{a+\mathrm{x}_n-1}(1-u_2)^{n+b+\mathrm{x}_n+1} = g_1(u_2), \\ g_2(v_2) = (v_1)^{a+\mathrm{x}_n}(1-v_1)^{n+b+\mathrm{x}_n} = (v_2)^{a+\mathrm{x}_n}(1-v_2)^{n+b+\mathrm{x}_n} = g_2(v_2), \end{cases} \tag{3.2.14}$$

and we are left with minimizing the following functions to zero:

$$\begin{cases} G_1(t,z) = & \eta\left[F_{p_1}(z) - F_{p_1}(t) - (1-\alpha)\right]^2 + \left[g_1(z) - g_1(t)\right]^2, \\ G_2(t,z) = & \eta\left[F_{p_1}(z) - F_{p_1}(t) - (1-\alpha)\right]^2 + \left[g_2(z) - g_2(t)\right]^2, \end{cases} \tag{3.2.15}$$

to get u_1, u_2, v_1, v_2. We implemented these minimization techniques and others in this thesis using Visual Fortran 6.1 (Compaq), which includes most **ISML libraries** (International Mathematical and Statistical Libraries, Inc, IMSL Library, Houston, TX, 1991).

Although the use of (3.2.15) offers a big improvement over the omnibus algorithm 2 in computation time, the starting values are sometimes insufficient to ensure rapid convergence, especially when $n > 10000$. One major obstacle with the ISML subroutines is that any non-progress in five iterative steps stops the program, leaving us therefore with no output. We discuss these practical matters further in subsections 3.2.6.

All of these inconveniences lead us to consider approximate methods. Approximate methods are techniques used frequently in traditional Bayesian sample size calculation problems when covariates are of interest (see Chaloner and Verdinelli, 1995). These techniques turn the discrete problem of sample size derivation into a continuous one, often allowing the use of calculus.

3.2.4 Approximate methods

There are, broadly speaking, two ways to approximate a solution to the sample size problem. The first approach is based on an asymptotic result for the limiting distribution of the pre-posterior predictive distribution, and the second approach is based on the asymptotic expansion of the posterior distribution. We give these results in subsections 3.2.4.1 and 3.2.4.2, before describing in 3.2.4.3 how they can be applied to our problem.

3.2.4.1 Asymptotic distribution of the pre-posterior predictive distribution

Theorem 3.2.4 below, which is a straightforward application of Khintchin's weak law of large numbers, does not seem to have been formulated before. This is surprising in view of its simplicity.

Theorem 3.2.4. *Let $X_1, \cdots, X_n$ be n exchangeable random variables such that $X_i | \theta \sim f_X(x|\theta)$, $i = 1, \ldots, n$, $\theta \sim f(\theta)$, $E(\|X\| \,|\, \theta) < \infty$, and let $Z = E(X|\theta)$ except on a set of measure zero with respect to $f(\theta)$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n}{n} \rightarrow^P Z. \tag{3.2.16}$$

Proof. We have

$$\lim_{n\to\infty} \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon\right] = \lim_{n\to\infty} \int_\Theta \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon \,\middle|\, \theta\right] f(\theta)\, d\theta \tag{3.2.17}$$

$$= \int_\Theta \lim_{n\to\infty} \Pr\left[\left\|\frac{S_n}{n} - Z\right\| > \varepsilon \,\middle|\, \theta\right] f(\theta)\, d\theta \tag{3.2.18}$$

$$= 0,$$

by first using Lebesgue's dominated convergence theorem to interchange the limit and the integral in expression (3.2.17) to get (3.2.18), and then Khintchin's theorem to evaluate the limit inside the integral. □

It is worth noting that Theorem 3.2.4 does not make any assumptions about the dimension of the random variables $X_1, \cdots, X_n$ and θ.

Srivastava and Wu, 1993 discuss a similar result for the Beta-Binomial model. They suggest that a moment generating function argument may be used to prove this special case.

Theorem 3.2.4 combined with the fact that the random variable $\frac{X_n}{n}$ is uniformly bounded by 1 yields

Corollary 3.2.5.

$$\begin{aligned}\lim_{n\to\infty} \boldsymbol{E}\left(\frac{X_n}{n}\right) &= \frac{a}{a+b}\\ \lim_{n\to\infty} \mathrm{Var}\left(\frac{X_n}{n}\right) &= \frac{ab}{(a+b)^2(a+b+1)}.\end{aligned}$$

We will investigate the accuracy of this approximation in section 3.2.4.3.

3.2.4.2 Approximate left and right tails of HPD and equal-tailed intervals for p_1, ω, and ϕ

Alternatively, we may approximate the lengths of the HPD intervals by using a third order asymptotic expansion of the posterior distribution. As we will see in subsection 3.2.7.2, the sample sizes based on these approximations are very close to those obtained using the exact method.

We first approximate the left and right tails of the HPD and equal-tailed intervals for the proportion, the odds, and the log-odds, and then compute their lengths. For any of these approximations to hold, it is crucial that the posterior distribution has a unique maximum in the interior of its set of definition. Therefore, when estimating the proportion p_1, we need $a > 1$ and $b > 1$ and when estimating the odds ω, we need $a > 1$.

Although one might contemplate using the expressions given in subsection 2.7.3 to obtain approximate lengths of the HPD and equal-tailed intervals for the parameters p_1, ω, and ϕ, unfortunately this cannot be achieved directly. We therefore propose a modification of the procedure that overcomes these difficulties, which may be illustrated by considering, for example, the proportion p_1. Consider the third order Taylor expansion

$$\begin{aligned}(-nw_2)^{\frac{1}{2}}(\theta_U^{(3)} - \hat{\theta}) \approx z &+ \frac{z^2}{6} w_3(-w_2)^{-\frac{3}{2}} + (-nw_2)^{-\frac{1}{2}} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\} \\ &+ \frac{z^3+3z}{24} w_4(-w_2)^{-2} + \frac{z}{-2nw_2} \left\{ \frac{d^2 \log p(\hat{\theta})}{d\hat{\theta}^2} \right\} \\ &+ \frac{5(z^3+3z)}{72} w_3^2(-w_2)^{-3} + \frac{z}{2\sqrt{n}} w_3(-w_2)^{-2} \left\{ \frac{d \log p(\hat{\theta})}{d\hat{\theta}} \right\}.\end{aligned}$$

One of the troublesome terms when $\theta = p_1$ is

$$\frac{1}{2}\{-Nw_2\}^{-1} \frac{d^2 \log p(\hat{p}_1)}{d\hat{p}_1^2} = -\frac{1}{2n}\left\{(a-1)\frac{n-\mathrm{x}_n}{\mathrm{x}_n} + (b-1)\frac{\mathrm{x}_n}{n-\mathrm{x}_n}\right\}.$$

which is undefined for $\mathrm{x}_n = 0$ and $\mathrm{x}_n = n$. Similar difficulties occur when the odds and log-odds are under considerations. This causes a problem since in computing $\mathrm{alc}_k(n, a, b)$ we must average these approximate lengths from $x = 0$ to $x = n$. We propose, below, a solution to this difficulty.

Proportion

Let Model 1 be $p_1 \sim \mathbf{Be}(a, b)$ and $\mathrm{x}_n | p_1 \sim \mathbf{Bin}(n, p_1)$, and let $L_1(\mathrm{x}_n, n, a, b)$ be the length of the HPD interval given n, x_n, a, b. Similarly, let Model 2 be $p_1 \sim \mathbf{Be}(0, 0)$ and $\mathrm{y}_N | p_1 \sim \mathbf{Bin}(N, p_1)$, and let $L_2(\mathrm{y}_N, N, 0, 0)$ be the length of the HPD interval given $N, \mathrm{y}_N, 0, 0$, where $N = n + a + b$. Suppose for the moment that a and b are integers.

Now notice that $L_1(\mathrm{x}_n, n, a, b) = L_2(a + \mathrm{x}_n, N, 0, 0)$, $\mathrm{x}_n = 0, 1, \cdots, n$ since the posterior distributions under Model 1 and Model 2 are the same. Hence

$$\begin{aligned}\mathrm{alc}_k^k(n,a,b) &= \sum_{\mathrm{x}_n=0}^{n} L_1^k(\mathrm{x}_n, n, a, b)\, p_{X_n}(\mathrm{x}_n|\, n, a, b),\\ &= \sum_{\mathrm{x}_n=0}^{n} L_2^k(a+\mathrm{x}_n, N, 0, 0)\, p_{X_n}(\mathrm{x}_n|\, n, a, b),\\ &\approx \sum_{\mathrm{x}_n=0}^{n} \left\{L_2^{approx}(\mathrm{x}_n + a, N, 0, 0)\right\}^k p_{X_n}(\mathrm{x}_n|\, n, a, b),\end{aligned}$$

where $L_2^{approx}(\mathrm{y}_N, N, 0, 0)$ is the third order approximate HPD length using the expressions given in subsection 2.7.3. Since the sum of the terms $L_2(\mathrm{y}_N, N, 0, 0)$ goes from $\mathrm{y}_N = a + x_n$ to $\mathrm{y}_N = n + a$ which are well-defined, the difficulties at $\mathrm{x}_n = 0$ and $\mathrm{x}_n = n$ are avoided. Although, the idea behind using $L_2^{approx}(\mathrm{y}_N, N, 0, 0)$ is based on the assumption that a and b are integers, we found that the third order approximate HPD and equal-tailed lengths obtained applies as well for any $a, b > 1$. We give below the third order approximations to the credible intervals lengths using equations (2.2.19) and (2.2.20). In Appendix A, we give all the components involved in the computation of the approximate left and right tails of the HPD and equal-tailed intervals (equations (2.7.7) and (2.7.8)). These components are also used for the computations in equations (2.7.9) and (2.7.10), which are essential to the derivation of approximate lengths.

Using equation (2.7.9) together with the adjustment discussed above and the results in appendix A, one gets an approximate HPD of length

$$\begin{aligned}L_{\mathrm{HPD}}(p_1, z, \mathrm{x}_n) &= \frac{2}{N\sqrt{v_1(\mathrm{x}_n)}}\left\{z - (z^3+3z)\frac{v_2(\mathrm{x}_n)-1}{4N} + z\frac{v_2(\mathrm{x}_n)}{2N} + \right.\\ &\qquad \left. 5(z^3+3z)\frac{v_2(\mathrm{x}_n)-2}{18N} - z\frac{v_2(\mathrm{x}_n)-2}{N}\right\},\end{aligned} \tag{3.2.19}$$

where $v_1(\mathrm{x}_n) = \frac{1}{\mathrm{x}_n + a} + \frac{1}{n+b-\mathrm{x}_n}$, $v_2(\mathrm{x}_n) = \frac{n+b-\mathrm{x}_n}{\mathrm{x}_n+a} + \frac{\mathrm{x}_n+a}{n+b-\mathrm{x}_n}$, $\mathrm{x}_n = 0, 1, \cdots, n$, and $z = z_{1-\alpha/2}$.

For an equal-tailed interval from expression (2.7.10), one obtains

$$L\text{eq}(p_1, z, \mathrm{x}_n) = L_{\mathrm{HPD}}(p_1, z, \mathrm{x}_n) + 4z\frac{v_2(\mathrm{x}_n)}{9N^2\sqrt{v_1(\mathrm{x}_n)}}. \tag{3.2.20}$$

For completeness, we present approximate credible interval lengths for the odds and log-odds.

Odds

Using equations (2.7.9) and (2.7.10) together with the results in Appendix A, one gets an approximate HPD of length

$$\begin{aligned} L_{\mathrm{HPD}}(\omega, z, \mathrm{x}_n) &= 2\sqrt{v_3(\mathrm{x}_n)}\left\{z - (z^3+3z)\frac{v_2(\mathrm{x}_n)+1}{4(n+b-\mathrm{x}_n)} + \frac{z}{2}v_1(\mathrm{x}_n) + \right. \\ &\left. 5(z^3+3z)\frac{v_2(\mathrm{x}_n)+2}{18(n+b-\mathrm{x}_n)} - z\left(v_1(\mathrm{x}_n) + \frac{1}{n+b-\mathrm{x}_n}\right)\right\}, \end{aligned} \tag{3.2.21}$$

and an approximate equal-tailed interval of length

$$L\text{eq}(\omega, z, \mathrm{x}_n) = L_{\mathrm{HPD}}(\omega, z, \mathrm{x}_n) + z\frac{v_2(\mathrm{x}_n)+2}{9(n+b-\mathrm{x}_n)}\sqrt{v_3(\mathrm{x}_n)}, \tag{3.2.22}$$

where $N = n+a+b$, $v_1(\mathrm{x}_n) = \frac{1}{\mathrm{x}_n+a} + \frac{1}{n+b-\mathrm{x}_n}$, $v_2(\mathrm{x}_n) = \frac{N}{\mathrm{x}_n+a} + \frac{\mathrm{x}_n+a}{N}$, and $v_3(\mathrm{x}_n) = \frac{N(\mathrm{x}_n+a)}{(n+b-\mathrm{x}_n)^3}$, $\mathrm{x}_n = 0, 1, \cdots, n$.

Log-odds

Using equations (2.7.9) and (2.7.10) together with the results in Appendix A, one gets the approximate lengths

$$\begin{aligned} L_{\text{HPD}}(\phi, z, \mathrm{x}_n) &= 2\sqrt{v_1(\mathrm{x}_n)}\left\{z - (z^3+3z)\frac{v_2(\mathrm{x}_n)-4}{24N} + 5(z^3+3z)\frac{v_2(\mathrm{x}_n)-2}{72N}\right\} \\ &= 2z\sqrt{v_1(\mathrm{x}_n)}\left\{1 + \frac{z^2+3}{36N}(v_2(\mathrm{x}_n)+1)\right\} \qquad (3.2.23) \end{aligned}$$

and

$$L\text{eq}(\phi, z, \mathrm{x}_n) = L_{\text{HPD}}(\phi, z, \mathrm{x}_n) + z\frac{v_2(\mathrm{x}_n)-2}{9N}\sqrt{v_1(\mathrm{x}_n)}, \qquad (3.2.24)$$

where $N = n+a+b$, $v_1(\mathrm{x}_n) = \dfrac{1}{\mathrm{x}_n + a} + \dfrac{1}{n+b-\mathrm{x}_n}$ and $v_2(\mathrm{x}_n) = \dfrac{n+b-\mathrm{x}_n}{\mathrm{x}_n+a} + \dfrac{\mathrm{x}_n+a}{n+b-\mathrm{x}_n}$, $\mathrm{x}_n = 0, 1, \cdots, n$.

In subsection 3.2.7.2, we will examine how sample sizes derived using these approximations perform compared with those based on exact methods. We shall, first, however, assess the effect of using the limiting distribution of the pre-posterior predictive distribution in place of its exact counterpart. By turning the discrete problem into a continuous sample size problem, we are able to exploit the manageable form of the limiting pre-posterior predictive distribution, to obtain approximate sample sizes for the **ALC** and **ALC**$_2$. For the sake of readability, we relegate the approximate sample size formulae for the **MLOC** to appendix D. In subsection 3.2.4.4 we discuss how a regression analysis might help to derive or improve sample size calculations based on Monte Carlo estimates.

3.2.4.3 Approximate sample size formula for the $\mathbf{ALC}_k$ for $k = 1, 2$

The proportion, p_1: We begin by pointing out the similarity for large sample sizes between the approach suggested by Pham-Gia and Turkkan, 1992 and the approach based on $\mathbf{ALC}_2$. For the **ALC** however, we must resort to a different and more sophisticated technique.

The posterior variance of the parameter p_1, obtained by using the density in equation 3.2.9, is $\mathbf{Var}(p_1 | \mathbf{x}_n, n, a, b) = \dfrac{(a+\mathbf{x}_n)(n+b-\mathbf{x}_n)}{(n+a+b)^2(n+a+b+1)}$, and is denoted by $\mathbf{Var}_{p_1}(\mathbf{x}_n, n, a, b)$ for $\mathbf{x}_n = 0, \cdots, n$. The pre-posterior average variance for p_1, derived by Pham-Gia and Turkkan, 1992, is

$$\mathbf{E}_{X_n}\left[\mathbf{Var}_{p_1}(X_n, n, a, b)\right] = \frac{c_{p_1}(a,b)}{N}, \tag{3.2.25}$$

where $c_{p_1}(a,b) = \dfrac{\mathbf{B}(a+1,b+1)}{\mathbf{B}(a,b)} = \dfrac{ab}{(a+b)(a+b+1)}$, and where $N = n + a + b$. This implies that

$$\sqrt{n+a+b}\Big\{\mathbf{E}_{X_n}\left[\mathbf{Var}_{p_1}(X_n, n, a, b)\right]\Big\}^{1/2} = \sqrt{c_{p_1}(a,b)}.$$

Let $l_{p_1}(\mathbf{x}_n, n, a, b) = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}_{p_1}(\mathbf{x}_n, n, a, b)}$ be the first order approximation of the length of the HPD or equal-tailed interval. Note that this first order approximation is only valid for all $\mathbf{x}_n = 0, 1, \cdots, n$ when $a > 1$ and $b > 1$. Such an approximation was suggested by Pham-Gia and Turkkan, 1992 as a substitute for the length of credible intervals. Application to the $\mathbf{ALC}_2$ requires that we solve the equation $\left[\mathbf{E}_{X_n} l^2_{p_1}(X_n, n, a, b)\right]^{1/2} = l$, for n, giving an approximate sample size of

$$n_{p_1} = 4\frac{z^2_{1-\alpha/2}}{l^2}\, c_{p_1}(a,b) - a - b, \qquad a, b > 1. \tag{3.2.26}$$

This approximate sample size corresponds to the exact sample size formula under PGT-(ii)(Pham-Gia and Turkkan, 1992) when the target level of accuracy ϵ is taken to be $\dfrac{l^2}{4\,z^2_{1-\alpha/2}}$ (see subsection 2.5.3).

Although it is straightforward to obtain approximate sample size formula for the $\mathbf{ALC}_2$ and indeed this has essentially been obtained by Pham-Gia and Turkkan, 1992 in a different guise, it is very difficult to derive approximate sample size formula for the **ALC** using the exact pre-posterior predictive distribution. Fortunately, Theorems 3.2.6, 3.2.7 and Corollary 3.2.8 suggest a solution to this problem.

Theorem 3.2.6. *If $X_n \longrightarrow^d X$ and the X_n are uniformly integrable, then X is integrable and*

$$\lim_n \boldsymbol{E}[X_n] = \boldsymbol{E}[X].$$

Proof. See Billingsley, 1995. □

Note that X_n uniformly bounded implies X_n uniformly integrable.

Theorem 3.2.7. *Suppose that $X_n \longrightarrow^d X$ and h_n and h are Borel functions. Let E be the set of x for which $h_n(x_n) \longrightarrow h(x)$ fails for some sequence $x_n \longrightarrow x$. Suppose that E is a Borel set and $P[X \in E] = 0$. Then $h_n(X_n) \longrightarrow^d h(X)$.*

Proof. See Billingsley, 1995, exercise 25.8, p.340. □

It is obvious that this theorem can be generalized to a sequence with more than one argument, i.e., of the form $h_n(X_n, Y_n, \cdots, Z_n)$. This generalization is used in chapter 4 to find approximate sample size formulae for the two sample problem.

We now use both Theorems 3.2.6 and 3.2.7 to establish an important corollary which will be used to derive approximate sample sizes for p_1 under the **ALC**.

Corollary 3.2.8.

$$\lim_n \sqrt{n+a+b}\boldsymbol{E}_{X_n}[l_{p_1}(X_n,n,a,b)] = 2z_{1-\alpha/2}\frac{\boldsymbol{B}(a+1/2,b+1/2)}{\boldsymbol{B}(a,b)}.$$

Proof. Set $Y_n = \dfrac{X_n}{n}$, and let $\mathcal{F}_n = \left\{0, \frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}, 1\right\}$ be the set of points where the mass function of Y_n is positive. According to Theorem 3.2.4, we have $Y_n \longrightarrow^d p_1$. Let $V_n = h_n(Y_n)$, $h_n(y) = \sqrt{\dfrac{(a+ny)(n+b-ny)}{N^2}}$ and $h(y) = \sqrt{y(1-y)}$, $y \in [0,1]$. Theorem 3.2.7 suggests that $V_n \longrightarrow^d h(p_1)$. We have

$$\begin{aligned}\sqrt{N}\mathbf{E}_{X_n}[l_{p_1}(X_n,n,a,b)] &= \frac{2z_{1-\alpha/2}N}{\sqrt{N(N+1)}}\sum_{y\in\mathcal{F}_n} h_n(y)\, p_{Y_n}(y|n,a,b), \\ &= \frac{2z_{1-\alpha/2}N}{\sqrt{N(N+1)}}\mathbf{E}[V_n].\end{aligned}$$

Since the sequence of random variables, V_n, are uniformly bounded, Theorem 3.2.6 implies that,

$$\begin{aligned}\lim_n \mathbf{E}[V_n] &= \int_0^1 h(y)\frac{y^{a-1}(1-y)^{b-1}}{\mathbf{B}(a,b)}\,dy \\ &= \frac{\mathbf{B}(a+1/2,b+1/2)}{\mathbf{B}(a,b)},\end{aligned}$$

which completes the proof. □

Corollary 3.2.8 suggests that an approximate sample size using the **ALC** may be obtained by solving $\mathbf{E}_{X_n}\left[l_{p_1}(X_n,n,a,b)\right] = l$ for n using the fact that

$\dfrac{\mathbf{E}_{X_n}\left[l_{p_1}(X_n,n,a,b)\right]}{2z_{1-\alpha/2}} \approx \dfrac{\mathrm{B}(a+1/2,b+1/2)}{\mathrm{B}(a,b)\sqrt{n+a+b}}$. One obtains

$$n_{p_1} = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(a+1/2,b+1/2)}{\mathrm{B}(a,b)}\right)^2 - a - b, \quad a,b>1. \qquad (3.2.27)$$

In general, for the $\mathbf{ALC}_k$, it is easy to establish

Corollary 3.2.9.

$$\lim_n \sqrt{n+a+b}\Big\{\boldsymbol{E}_{X_n}[l^k_{p_1}(X_n,n,a,b)]\Big\}^{1/k} = 2z_{1-\alpha/2}\left\{\frac{B(a+k/2,b+k/2)}{B(a,b)}\right\}^{1/k}.$$

Proof. The proof is similar to that of Corollary 3.2.8 where $h(y)=y^{k/2}(1-y)^{k/2}$ and $h_n(y)=\dfrac{(a+ny)^{k/2}(n+b-ny)^{k/2}}{N^k}$, $y\in[0,1]$. □

As before, using Corollary 3.2.9 we obtain,

$$n_{p_1} = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(a+k/2,b+k/2)}{\mathrm{B}(a,b)}\right)^{2/k} - a - b \qquad (3.2.28)$$

as an approximate sample size for the $\mathbf{ALC}_k$. These approximate sample sizes for the $\mathbf{ALC}_k$ increase as k increases since

$$\begin{aligned}\sqrt{n_{p_1}+a+b} &= 2\frac{z_{1-\alpha/2}}{l}\left(\frac{\mathrm{B}(a+k/2,b+k/2)}{\mathrm{B}(a,b)}\right)^{1/k}\\ &= 2\frac{z_{1-\alpha/2}}{l}\left(\int_0^1\{x^{1/2}(1-x)^{1/2}\}^k\frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)}\right)^{1/k},\end{aligned}$$

and the last expression is essentially an L_k-norm. If we take the limit as k goes to infinity, we obtain an approximate sample size for the **WOC** of

$$n_{p_1} = \frac{z^2_{1-\alpha/2}}{l^2} - a - b, \qquad a,b>1. \qquad (3.2.29)$$

Adcock, 1987 derived a similar sample size formula using the quantile of a chi-square distribution. In practice, this approximate formula gives an accurate estimate of the exact sample size with an overestimation or underestimation

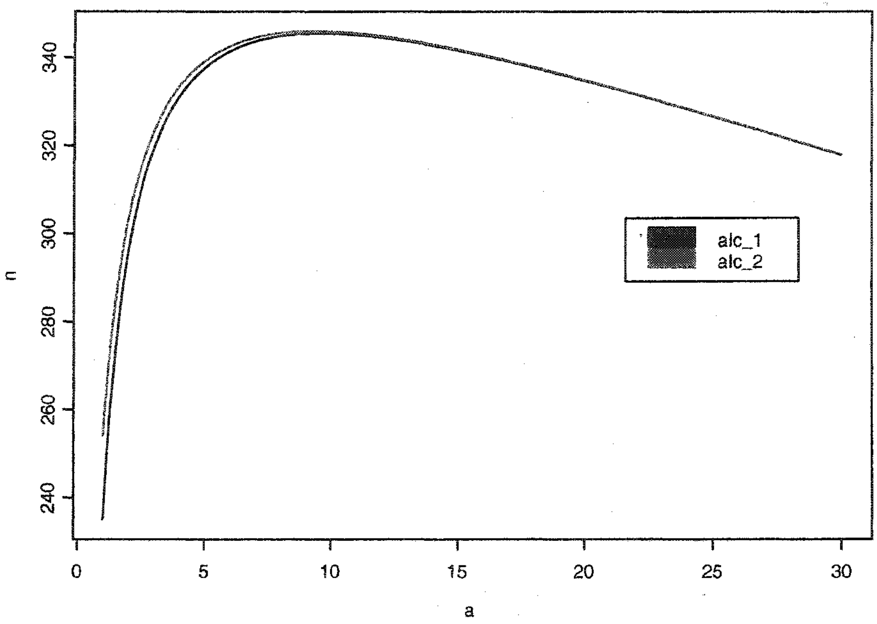


Figure 3.2.1: Graph of the approximate sample size for the **ALC** (in black) and **ALC**$_2$ (in blue) as a function of $1.0 \leq a \leq 30.0,\ b = a,\ l = .1,\ 1 - \alpha = .95$ when p_1 is of interest.

of 1 or 2 when $a, b \geq 1$. Note that unlike equation 3.2.27, the approximate sample sizes in equation (3.2.29) depend on a and b only through $a + b$. Consequently, the prior parameters a and b need not be known if $a + b$ is known. Equation (3.2.29) shows that the Bayesian sample size is smaller than the corresponding frequentist sample size. This reduction in sample size comes about because the incorporation of prior information through $a + b$ permits a reduction of the sample size by this amount.

Figure 3.2.1 displays the approximate sample sizes for the **ALC** and the

$\mathbf{ALC}_2$ as a function of $1.0 < a < 30.0$, on a continuous scale for $l = .1$, $1-\alpha = .95$, and $b = a$. We can see that the sample sizes for the **ALC** and $\mathbf{ALC}_2$ are similar. As prior parameters increase, although it is expected that the sample size should decrease because of increase in prior information, this is not the case for $a < 9.3$. It should be remembered, however, that there are counteracting influences on the sample size: first, the larger $a + b$ the more precise the prior information. On the other hand, the variance of these data increases with increasing $a+b$. In the beginning the influence of the variance prevails, while later the peakedness of the prior prevails.

We retrace the steps followed for p_1 in deriving sample size formulae for odds and log-odds.

The Odds, ω : When $b > 2$, the posterior variance of ω, obtained by using the posterior density in equation 3.2.10, can be approximated by

$$\begin{aligned}\mathbf{Var}(\omega|\,\mathrm{x}_n, n, a, b) &= \frac{(a+\mathrm{x}_n)(n+a+b-1)}{(n+b-\mathrm{x}_n-1)^2(n+b-\mathrm{x}_n-2)}, \qquad (3.2.30)\\ &\approx \frac{(a+\mathrm{x}_n)(n+a+b-2)}{(n+b-\mathrm{x}_n-1)(n+b-\mathrm{x}_n-2)(n+b-\mathrm{x}_n-3)},\\ &\approx \frac{(a+\mathrm{x}_n)(N-1)}{(n+b-\mathrm{x}_n-1)^3},\end{aligned}$$

denoted here by $\mathbf{Var}_\omega(\mathrm{x}_n, n, a, b)$ for $\mathrm{x}_n = 0, \cdots, n$. The derivative of $\log \mathbf{Var}(\omega|\,\mathrm{x}_n, n, a, b)$ as a continuous function of x_n in $[0, n]$ is $\dfrac{1}{n+b-\mathrm{x}_n-2} + \dfrac{2}{n+b-\mathrm{x}_n-1} + \dfrac{1}{a+\mathrm{x}_n} > 0$ for $b > 2$. Hence these posterior variances increase and satisfy

$$\frac{a(N-1)}{(n+b-2)(n+b-1)^2} \leq \mathbf{Var}_\omega(\mathrm{x}_n, n, a, b) \leq \frac{(n+a)(N-1)}{(b-2)(b-1)^2},$$

where $N = n + a + b$. Since the maximum value of the posterior variances increase with n, the criterion function $\mathrm{wloc}(n, a, b)$ diverges as n increases. Therefore, the **WLOC** is not useful in this context.

With this background, we are now ready to derive approximate sample size formulae for ω under the $\mathbf{ALC}_2$, **ALC**, and finally under the $\mathbf{ALC}_k$ for $k \geq 1$. We begin with special cases of $k = 1, 2$ for illustrative purposes only, recognizing that the general result covers these cases.

The pre-posterior average variance for $b > 3$ is

$$\begin{aligned}
\mathbf{E}_{X_n}\left[\mathbf{Var}_\omega(X_n, n, a, b)\right] &= \sum_{\mathbf{x}_n=0}^{n} \mathbf{Var}_\omega(\mathbf{x}_n, n, a, b)\, p_{X_n}(\mathbf{x}_n | n, a, b), \\
&\lessapprox \sum_{\mathbf{x}_n=0}^{n} \binom{n}{\mathbf{x}_n} \frac{\mathbf{Be}(a + 1 + \mathbf{x}_n, n + b - 3 - \mathbf{x}_n)}{(N-1)\mathbf{Be}(a, b)}, \\
&= \frac{c_\omega(a, b)}{N - 1}, \qquad (3.2.31)
\end{aligned}$$

where $c_\omega(a, b) = \dfrac{\mathbf{B}(a + 1, b - 3)}{\mathbf{B}(a, b)} = \dfrac{a(a + b - 1)(a + b - 2)}{(b - 3)(b - 2)(b - 1)}$, implying that $\lim_n \sqrt{n + a + b}\left\{\mathbf{E}_{X_n}\left[\mathbf{Var}_\omega(X_n, n, a, b)\right]\right\}^{1/2} = \sqrt{c_\omega(a, b)}$. Again this can be proved using Theorem 3.2.6.

Let

$$l_\omega(\mathbf{x}_n, n, a, b) = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}_\omega(X_n, n, a, b)}$$

be the first order of approximation of the length of the credible interval for ω which is defined for all $\mathbf{x}_n = 0, 1, \cdots, n$ only when $a > 1$. An approximate sample size for the $\mathbf{ALC}_2$ is obtained by solving the equation $\mathbf{E}_{X_n}\left[l^2_\omega(X_n, n, a, b)\right] = l^2$, in n. This yields an approximate sample size

$$n_\omega = 4\frac{z^2_{1-\alpha/2}}{l^2}\, c_\omega(a, b) - a - b, \qquad a > 1. \qquad (3.2.32)$$

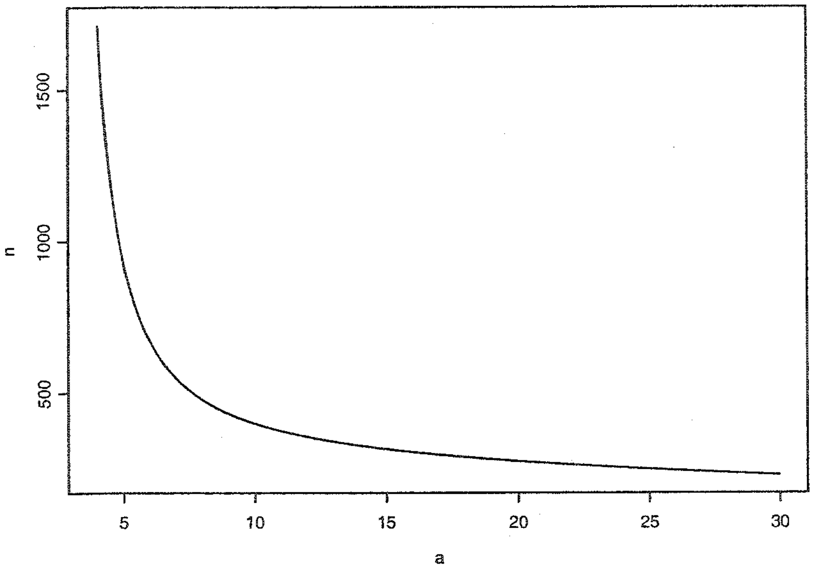


Figure 3.2.2: Graph of the approximate sample sizes for the $\mathbf{ALC}_2$ as a function of a for $4.0 \leq a \leq 30.0$, $b = a$, $l = .5$ and $1 - \alpha = .95$ when ω is of interest.

Figure 3.2.2 displays the approximate sample sizes for the $\mathbf{ALC}_2$ as a function of a for $4.0 < a < 30.0$ on a continuous scale for $l = .5$, $1 - \alpha = .95$, and $b = a$. These sample sizes decrease as $a = b$ increases.

In order to obtain an approximate sample size formula for the **ALC**, we need another corollary to Theorem 3.2.6.

Corollary 3.2.10. *For $b \geq 2$,*

$$\lim_n \sqrt{n + a + b} E_{X_n}[l_\omega(X_n, n, a, b)] = 2z_{1-\alpha/2} \frac{B(a + 1/2, b - 3/2)}{B(a, b)}.$$

Proof. Let $Y_n = \dfrac{X_n}{n}$, $\mathcal{F}_n = \{0, \frac{1}{n}, \cdots, \frac{n-1}{n}, 1\}$, and $h(y) = \sqrt{y(1-y)}$, $y \in$

$[0,1]$. Set $h_n(y) = \sqrt{\dfrac{(a+ny)(n+b-2-ny)}{N^2}}$. We have

$$\begin{aligned}
\frac{\mathbf{E}_{X_n}[l_\omega(X_n,n,a,b)]}{2z_{1-\alpha/2}} &= \sum_{y\in\mathcal{F}_n}\sqrt{\frac{(N-1)(a+ny)}{(n+b-ny-1)^2(n+b-ny-2)}}\,p_{Y_n}(y|n,a,b) \\
&= c_N\sum_{y\in\mathcal{F}_n} h_n(y)\,p_{Y_n}(y|n,a,b-2).
\end{aligned}$$

where $c_N = \dfrac{N}{(N-2)\sqrt{N-1}}$ and $\lim_n c_N\sqrt{N} = 1$. We have shown in the proof of Corollary, that 3.2.8

$$\begin{aligned}
\lim_n \sum_{y\in\mathcal{F}_n} f_n(y)\,p_{Y_n}(y|n,a,b-2) &= \int_0^1 f(y)\frac{y^{a-1}(1-y)^{b-3}}{\mathbf{B}(a,b)}\,dy \\
&= \frac{\mathbf{B}(a+1/2,b-3/2)}{\mathbf{B}(a,b)},
\end{aligned}$$

which completes the proof. □

In similar fashion to the result for p_1, we obtain

$$n_\omega = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathbf{B}(a+1/2,b-3/2)}{\mathbf{B}(a,b)}\right)^2 - a - b, \qquad a>1,\ b>2. \quad (3.2.33)$$

Figure 3.2.3 displays the approximate sample size for the **ALC** as a function of a for $3.0 < a < 30.0$ on a continuous scale for $l = .5$, $1-\alpha = .95$, and $b = a$. Comparing Figures 3.1.2 and 3.1.3 we see that the sample sizes from the **ALC** and **ALC**$_2$ are quite different.

In general, for $k \geq 1$, by applying Corollary B.0.7 in appendix B, we may show

$$n_\omega = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathbf{B}(a+k/2,b-3k/2)}{\mathbf{B}(a,b)}\right)^{2/k} - a - b, \qquad a>1,\ b>3k/2, \quad (3.2.34)$$

to be an approximate sample size for the **ALC**$_k$. The approximation in (3.2.34) improves as b increases.

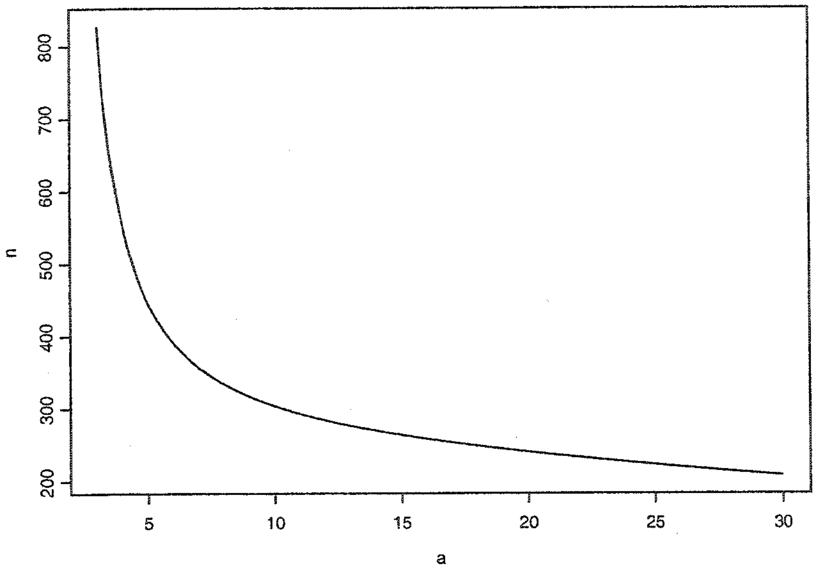


Figure 3.2.3: Graph of the approximate sample sizes for **ALC** as a function of a for $3.0 \leq a \leq 30.0$, $b = a$, $l = .5$ and $1 - \alpha = .95$ when ω is of interest.

The Log-odds, ϕ : For given n, a, b, it is well known (Johnson et al., 1994) that the posterior variance of ϕ is

$$\begin{aligned} \mathbf{Var}(\phi|\, \mathbf{x}_n, n, a, b) &= \Psi'(a + \mathbf{x}_n) + \Psi'(n + b - \mathbf{x}_n), \\ &\simeq \frac{1}{a + \mathbf{x}_n - 1} + \frac{1}{n + b - \mathbf{x}_n - 1}, \quad a, b > 1, \quad (3.2.35) \end{aligned}$$

where $\Psi'(x) = \dfrac{\Psi(x)}{dx}$ is the trigamma function, $\Psi(x) = \dfrac{d \log \Gamma(x)}{dx}$, and $\mathbf{x}_n = 0, \cdots, n$. Therefore a first order approximation to the credible intervals is given by

$$l_\phi(\mathbf{x}_n, n, a, b) = 2z_{1-\alpha/2}\sqrt{\frac{N-2}{(a + \mathbf{x}_n - 1)(n + b - \mathbf{x}_n - 1)}}, \qquad a, b > 1.$$

Sparing the reader the details which are essentially the same as those contained in the derivation of p_1 and ω, an approximation to the sample size formula for estimating ϕ using the $\mathbf{ALC}_k$ is given by

$$n_\phi = 4\frac{z_{1-\alpha/2}^2}{l^2}\left(\frac{\mathrm{B}(a-k/2,b-k/2)}{\mathrm{B}(a,b)}\right)^{2/k} - a - b, \qquad a,b > k/2, \quad (3.2.36)$$

since

$$\lim_n \sqrt{n+a+b}\left\{\mathrm{E}_{X_n}[l_\phi^k(X_n,n,a,b)]\right\}^{1/k} = 2z_{1-\alpha/2}\left(\frac{\mathrm{B}(a-k/2,b-k/2)}{\mathrm{B}(a,b)}\right)^{1/k}.$$

Although Corollaries 3.2.8, 3.2.9, 3.2.10, and B.0.7 are enough to derive approximate sample sizes, it is often prudent to consider higher order expansions of the criterion function $\mathrm{alc}_k(n,a,b)$. These improved approximations are given in appendix C, and this matter is discussed further in subsection 3.2.5.

So far we have mainly discussed exact and approximate HPD and equal-tailed intervals computations en route to finding $\mathbf{ALC}_k$ sample sizes. For the **ACC**, results analogous to those for the $\mathbf{ALC}_k$ are, unfortunately, not apparent. We turn briefly to a discussion of an exact and an approximate method for the $\mathbf{ACC}_k$.

3.2.4.4 Approximate sample size for the $\mathbf{ACC}_k$

We have potentially two methods to determine the coverage of an HPD interval when the length is fixed. These coverages must then, of course, be averaged, and equated to our specified average coverage. The first method is to use the omnibus procedure described in subsection 2.7.2 to determine the coverage of the HPD interval of length l in combination with the preposterior distribution. The second method, which we term the first order approximation, proposed in this thesis for the first time, is as follows. Consider, for the moment the parameter p_1. The idea is to use the approximate length of the credible interval, $l_{p_1}(\mathbf{x}_n|\, n, a, b) = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}_{p_1}(\mathbf{x}_n, n, a, b)}$ to recover an approximate value of the unknown coverage $1-\alpha$, that led to the quantile $z_{1-\alpha/2}$, given that we set $l_{p_1}(\mathbf{x}_n|\, n, a, b) \approx l$. First, for example, solve the equation $2z_{1-\alpha/2}\sqrt{\mathbf{Var}_{p_1}(\mathbf{x}_n, n, a, b)} = l$ in $z_{1-\alpha/2}$ and, thereby recover α. The approximate coverage $1-\alpha$ is taken to be $1-2\Phi(-t)$ where $t = \dfrac{l}{2\sqrt{\mathbf{Var}_{p_1}(\mathbf{x}_n, n, a, b)}}$ and where Φ is the cumulative distribution of a standard normal distribution.

Finally, approximate sample size formulae for the **MCOC** which are identical to those for the **MLOC** may be obtained and are given in appendix D.

3.2.5 Monte Carlo procedures

Since the Monte Carlo approaches to sample size calculation that we use are similar to those described in Joseph et al., 1995, they will not be the object of exhaustive discussion here. The main idea, described, for brevity, through the $\mathbf{ALC}_k$, is that one uses a bisectional search strategy to locate an integer n such that for given a coverage $1-\alpha$, $\widetilde{\text{alc}}_k(n, a, b) \leq l$ and $\widetilde{\text{alc}}_k(n-1, a, b) > l$, where $\widetilde{\text{alc}}_k(n, a, b)$ is a Monte Carlo estimate of $\text{alc}_k(n, a, b)$.

We sketch the algorithm when estimating p_1 is of interest. The algorithms for ω and ϕ are very similar. For each step in the bisectional search over n:

- Simulate p_{1i}, $i = 1, \cdots, m$ from a $\mathbf{Be}(a, b)$.

- For each p_{1i}, simulate an observation $x_i \sim \mathbf{Bin}(n, p_{1i})$.

- For each i, compute the length l_i of the HPD interval given $1 - \alpha$ using Algorithm 4 in subsection 2.7.2.3. Let M be the size of the Monte Carlo simulation used in Algorithm 4.

- compute $\widetilde{\mathrm{alc}}_k(n, a, b) = \frac{1}{m} \sum_1^m l_i$.

To reduce Monte Carlo errors, Joseph et al., 1997 used an average of 10 estimated sample sizes. A similar strategy can also be used for the $\mathbf{ACC}_k$, **WOC**, **MLOC**, and the **MCOC**.

A general criticism about this bisectional search algorithm is that it tends to underestimate the sample size when m and M are small, say $m, M < 2000$. There is an explanation for this phenomenon. For instance, if the range of the simulated values is less than the specified length, then that range is set to be the HPD interval length and is obviously an underestimate. This behavior is less prominent with proportions than it is with odds and log-odds because the former, unlike the others, are bounded. In the Monte Carlo simulation algorithm, it is not clear a priori what combinations of m and M provides adequate accuracy in the final sample size estimate as this depends on the particularities of the problem. Some preliminary runs are therefore suggested in order to obtain an idea of the variability.

Table 3.1 presents a summary of sample means, $(\bar{n})$, standard deviations (std_{n_j}), biases, mean square errors $(\sqrt{\mathrm{MSE}})$, and approximate 95% confidence intervals for n, based on 500 sample sizes n_j, $j = 1, \cdots, 500$, generated

Table 3.1: **Monte Carlo-based sample size calculation for various m and M for the ALC when the odds is under consideration. We set $a = b = 3$, $1 - \alpha = .95$, $l = .5$. The exact sample size is 816.**

$m = M$	$\bar{n}$	std_{n_j}	bias	$\text{std}_{\bar{n}}$	$\sqrt{\text{MSE}}$	95% confidence interval for n
500	783.20	66.63	-32.80	2.98	32.94	777.36—789.04
1000	795.24	217.19	-20.76	9.71	22.92	776.20—814.28
2000	817.16	125.30	1.16	5.60	5.72	806.17—828.14
5000	817.18	92.88	1.18	4.15	4.31	809.04—825.32
10000	818.18	72.96	2.18	3.26	3.92	811.78—824.57

using the Monte Carlo approach for each of $m = M = 500$, 1000, 2000, 5000, 10000. For instance, $\bar{n} = \frac{1}{500}\sum_{j=1}^{500} n_j$ and $\text{std}^2_{n_j} = \textbf{Var}(n_j)$. The parameter specification for the **ALC** for estimating ω are: $a = b = 3$, $1-\alpha = .95$, $l = .5$. Under these specifications, the exact sample size is 816, as determined by using the exact computation of the criterion function described in subsection 3.2.3. One marked characteristic of the estimated standard deviations for all independent and identically distributed random variable n_j in Table 3.1 is that they are all large, although estimated sample means are excellent when $m, M > 2000$. There is a large bias when $m = M = 500$ and $m = M = 1000$, which indicates $m, M \leq 1000$ is insufficient.

Another way to reduce potential biases induced by Monte Carlo procedures and also to account for the uncertainty in Monte Carlo procedures is to adapt a version of the regression approach developed by Müller and Parmigiani, 1995 and Müller, 1998 to the present situation.

3.2.6 Regression approach

Consider the proportion. Equation (C.1.5) in Appendix C suggests we fit a general regression equation of the type

$$\mathbf{E}_{X_n}\left[l_{p_1}(X_n)\right] = e_1\frac{1}{n^{1/2}} + e_2\frac{1}{n^{3/2}} + \cdots + e_h\frac{1}{n^{(2h-1)/2}}$$

to the alc(n, a, b) function, where e_i, $i = 1, \ldots, h$ are the regression coefficients and $l_{p_1} = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(p_1 |\, \mathbf{x}_n, n, a, b)}$, $a, b > 1$. Since, for large n, HPD and equal-tailed intervals are close, we could set

$$\text{alc}_k(n, a, b) = e_1\frac{1}{n^{1/2}} + e_2\frac{1}{n^{3/2}} + \cdots + e_h\frac{1}{n^{(2h-1)/2}}, \qquad k \geq 1. \qquad (3.2.37)$$

The idea now is to equate the specified l to the right hand side of (3.2.37) and to solve for n once the estimated regression coefficients $\hat{e}_1, \ldots, \hat{e}_h$ have been obtained, as follows.

Compute g pairs of $(n, \text{alc}(n, a, b))$ and then using least squares fit equation (3.2.37) to these data. Finally, the equation $\hat{e}_1\frac{1}{n^{1/2}} + \hat{e}_2\frac{1}{n^{3/2}} + \cdots + \hat{e}_h\frac{1}{n^{(2h-1)/2}} = l$ may be solved for n given h, the number of regression terms included, using a software package such as Maple. Table 3.2 presents the results for this approach when $(a, b) = (1, 1)$, $1 - \alpha = .95$, and $l = .1$ or $l = .02$, and $h = 1, \ldots, 7$.

For various h, we obtain approximate sample sizes for both HPD and equal-tailed intervals. For $l = .1$, $h = 3$ is sufficient to provide accurate solutions for both intervals. For $l = .02$, $h = 4$ is sufficient to provide an accurate solution for an equal-tailed interval while $h = 6$ for an HPD interval. Although we have not provided a formal stopping rule for h which depends on l, $1 - \alpha$, and the parameter of interest, a simple monitoring of the increase in sample sizes as

Table 3.2: **Approximate regression based sample sizes for estimating the parameter p_1 using HPD and equal-tailed intervals for various h and l when $(a, b) = (1, 1)$.**

HPD			equal-tailed		
h	$l = .1$	$l = .02$	h	$l = .1$	$l = .02$
1	182	4400	1	189	4608
2	221	5489	2	227	5649
3	231	5792	3	234	5878
4	233	5880	4	235	5917
5	234	5906	5	235	5922
6	234	5917	6	235	5922
7	234	5917	7	235	5922
exact	234	5921	exact	235	5922

h increases reveals when to stop. For instance, we might want the change in consecutive approximate sample size estimates not to be greater than 5. Values of h larger than 7 are not probably justified in view of the marginal gain in accuracy at the expense of increasing computational burden.

Remark 3.2.1. Although in the above example the equation (3.2.38)

$$\frac{\hat{e}_1}{n^{1/2}} + \frac{\hat{e}_2}{n^{3/2}} + \cdots + \frac{\hat{e}_h}{n^{(2h-1)/2}} = l \tag{3.2.38}$$

has exactly one real solution in n, in general we may face the difficulty of having to choose the optimal solution from a set of several that satisfy equation (3.2.38). When there is more than one real solution to equation

(3.2.38), it is hard to ascertain which corresponds to the optimal sample size unless we know an interval in which the true solution lies. The 10 estimated Monte Carlo sample sizes, for instance (see section 3.2.5), might suggest which of the roots is the correct choice.

In practice, we fit the regression equation (3.2.37) to g random pairs $(n, \widetilde{\text{alc}}_k(n, a, b))$ where $\widetilde{\text{alc}}_k(n, a, b)$ is a Monte Carlo estimate of $\text{alc}_k(n, a, b)$. We now illustrate this regression approach for the odds, with $a = b = 3$, $1 - \alpha = .95$, $l = .5$. The exact sample size is 816 and the approximate sample size given by equation (3.2.32) is 828. Table 3.1 suggests that one generates g random points n_i, $i = 1, 2, \cdots, g$, in the set $\{700, 701, \cdots, 1000\}$ and then compute $\widetilde{\text{alc}}(n_i, 3, 3)$ for each point n_i. With $g = 2000$ and $M = m = 1000$, we obtain an approximate sample size of 811 for $h = 1, 2, 3$. With $g = 5000$ and $m = M = 500$, we obtain 806, 806, 807, and 808 for $h = 1, 2, 3$, and 4, respectively. In both situations, we found that $g \geq 500$ is a reasonable choice if one wants an estimate accurate to within 15 units of 816. These two choices of g show that $h = 1$ is sufficient. This is not surprising as in subsection 3.2.7.1 we show empirically that there is often, but clearly not always, an approximate linear relationship between $\dfrac{1}{\text{alc}^2(n, a, b)}$ and n.

The above regression approach differs slightly from that given by Müller and Parmigiani, 1995 and Müller, 1998. These authors advocate a fit of a parametric curve to the pairs (n_i, l_i), where l_i's are Monte Carlo estimates of the HPD length (l_i are obtained third step of the algorithm sketched in subsection 3.2.5). This idea can be seen to correspond to the regression method proposed when $m = 1$. The motivation of these above authors for

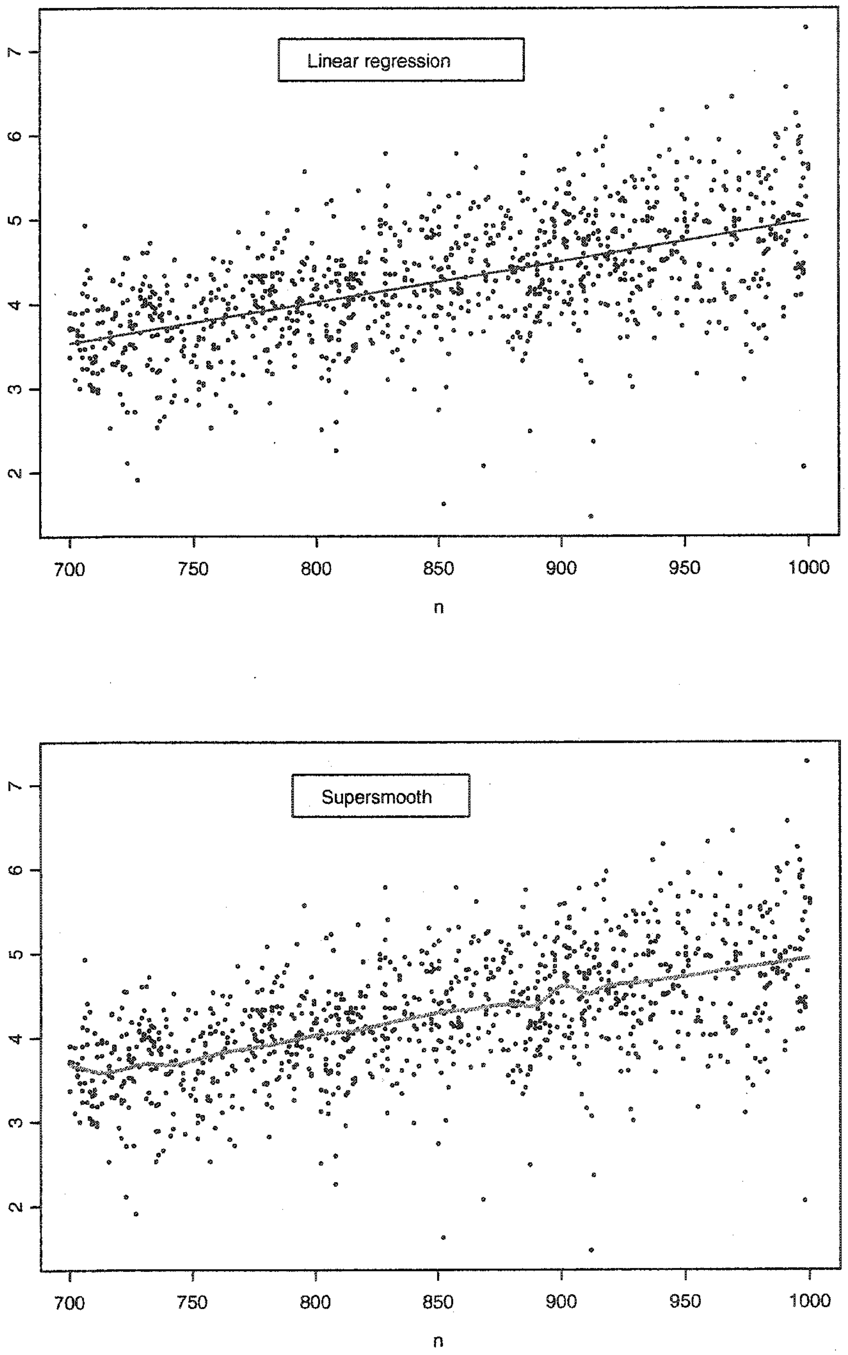


Figure 3.2.4: Graphs of the Monte Carlo pairs $\left(n, \dfrac{1}{\widetilde{\mathrm{alc}}^2(n,3,3)}\right)$.

using a regression approach is different from ours. They advocate a regression approach when the evaluation of the expected loss for various designs may be difficult and costly. We propose a regression approach to reduce the "noise" inherent in Monte Carlo methods. When a parametric form for the regression equation is unknown, these authors suggest the use of a smoothing curve. As indicated by Chaloner and Verdinelli, 1995, this alternative proposal is sometimes hard to implement effectively since different smoothers might yield different curves. Moreover, Müller and Parmigiani, 1995 recognize that their methods will carry more variability. Fortunately, however, the equation (3.2.37) suggests a parametric form for the regression function when the $\mathbf{ALC}_k$ is used. We shall exploit this observation to apply the idea by Müller and Parmigiani, 1995. Returning to our example with $(a, b) = (3, 3)$ and $M = 1000$, $g = 5000$, we got the following first order, $(h = 1)$, regression-based sample sizes of 832, 796, 781, 771, 808, 788, 787, 777, 785, 781 when using the first 500, 1000, 1500, $\cdots$, 5000 rows of the generated dataset composed of (n_i, l_i), $700 \leq n_i \leq 1000$. With $M = 10000$, $g = 20000$ and an increment of 2500 instead of 500, we obtained 765, 797, 777, 779, 781, 781, 798, 802 which shows that increasing the parameter M and g does not result necessarily in improvement in the final sample size. Using a second order expansion, $h = 2$, did not change the estimated sample sizes.

Our major contribution in this subsection is that we provide the form of the regression equations, so that one can in addition check if the regression equation holds by plotting $\dfrac{1}{\widetilde{\text{alc}}^{2}(n, a, b)}$ against n. Figure 2.2.5 provides two plots of $\left(n, \dfrac{1}{\widetilde{\text{alc}}^2(n,3,3)}\right)$ against n along with the simple regression linear regression line and the use of a supersmoother. In practice, any of the most popular smoothers in Splus, namely, linear fit, polynomial fit, natural splines, lowess, or supersmooth will do an adequate job.

Remark 3.2.2. In practice, the regression equation

$$\text{alc}_k(n,a,b) = e_1\frac{1}{n^{1/2}} + e_2\frac{1}{n^{3/2}} + \cdots + e_h\frac{1}{n^{(2h-1)/2}}, \qquad k \geq 1 \qquad (3.2.39)$$

can be applied for any $a, b > 0$ when considering the proportion, for $a > 0$ and $b > 3k/2$ when considering the odds, and $a, b > k/2$ when considering the log-odds. Empirical evidence suggest that equation (3.2.39) does not hold when $k < b \leq 3k/2$, for the odds and when $0 < a \leq k/2$ or $0 < b \leq k/2$ for the log-odds, because the leading term is no longer $\dfrac{1}{n^{1/2}}$. Rather, there exists $0 < \lambda < 1/2$ (often unknown) such that

$$\text{alc}_k(n,a,b) = e_1\frac{1}{n^{\lambda}} + e_2\frac{1}{n^{\lambda+1}} + \cdots + e_h\frac{1}{n^{\lambda+h-1}}, \qquad k \geq 1. \qquad (3.2.40)$$

Therefore, one should be careful near the boundaries $3k/2$ and $k/2$. In general, equation (3.2.39) works rather well when $b > 3(k+1)/2$ and $a, b > (k+1)/2$. In general, λ can be estimated using a bisection search strategy. For the case $h = 1$ and given a coverage of $1-\alpha$ and a length l , one can estimate λ using the regression equation $\log(\text{alc}_k(n,a,b)) = \mu - \lambda\log(n)$. In that case, the sample size is easily derived as $n = \exp\dfrac{\hat{\mu} - \log(l)}{\hat{\lambda}}$.

Equation (3.2.39) also applies to the two sample problems discussed in chapter 4 for all the cases where we have derived sample size formulae. Note that in contrast to the one sample problem, regression-based sample size calculations are the only reasonable alternative for the two sample problems. In all cases where equation (3.2.39) applies, often a simple plot of Monte Carlo estimates $\dfrac{1}{\widetilde{\text{alc}}_k^2(n,a,b)}$ against n reveals a linear relationship, suggesting rather a simpler regression model of the form $\dfrac{1}{\widetilde{\text{alc}}_k^2(n,a,b)} = e_1 + e_2 n$, and hence a straightforward approach to sample size calculations.

3.2.7 Applications and Results

This section is divided into two subsections. Subsection 3.2.7.1 is mainly concerned with graphical displays of the objective functions $\text{alc}_k(n,a,b)$ and $\text{acc}_k(n,a,b)$ with respect to n, using the HPD, the equal-tailed and the first order approximation to Bayesian credible intervals. In subsection 3.2.7.2, we compare sample sizes arising from the various sample size criteria, and the different approximations presented in subsection 3.2.4.

3.2.7.1 Displays of the various sample size criteria functions as a function of n

It is useful to plot the criterion functions used in the computation of sample sizes in order to observe their forms. The top halves of figures 3.2.5, 3.2.6 and 3.2.7 display $\text{alc}(n,a,b)$ as a function of n, while the bottom halves plot $\dfrac{1}{\text{alc}^2(n,a,b)}$ versus n. The bottom halves were suggested by the expansion equations in appendix C. The prior parameters for the proportion, the odds and the log-odds parameters are $a = b = 1$ and $a = b = 2$, and $a = b = 3$, respectively. An immediate conclusion is that the three different ways of computing credible intervals give very similar sample sizes, at least for these

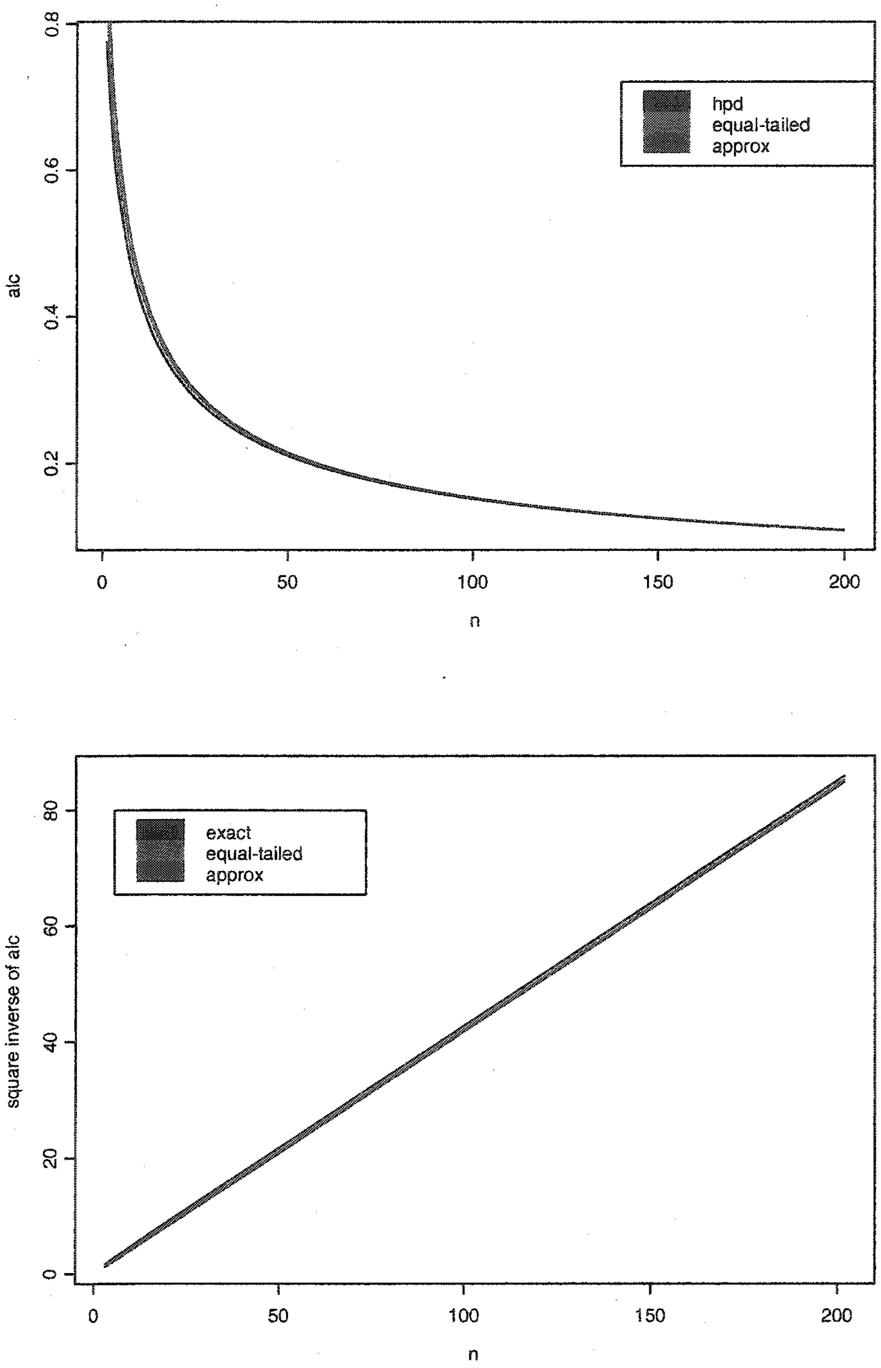


Figure 3.2.5: Graphs of $\text{alc}\,(n, 1, 1)$ (top) and $\dfrac{1}{\text{alc}^2(n, 1, 1)}$ (bottom) for the proportion p_1 against $n = 1, \cdots, 200$.

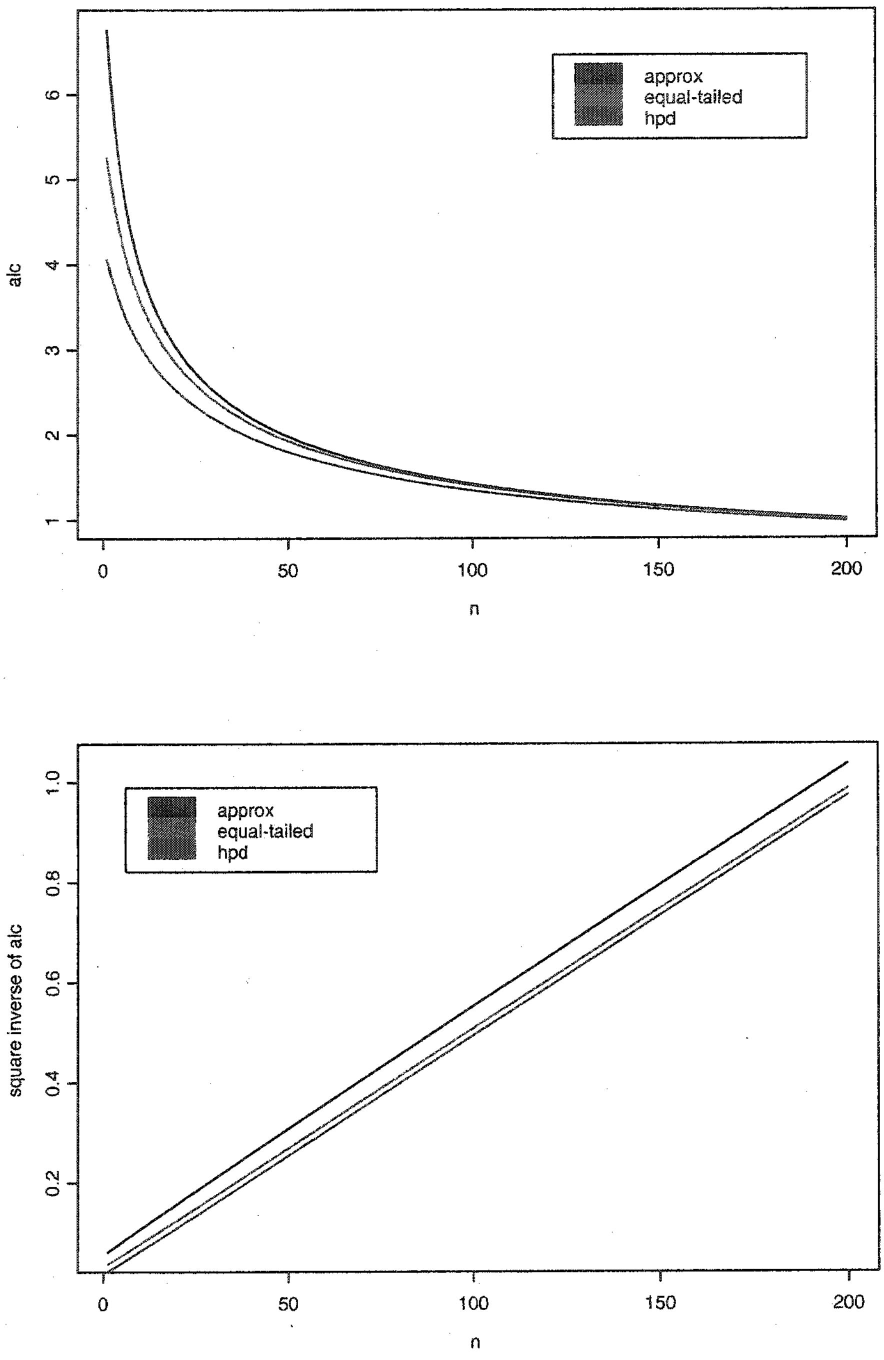


Figure 3.2.6: Graphs of alc$(n, 3, 3)$ (top) and $\dfrac{1}{\text{alc}^2(n,3,3)}$ (bottom) for the odds ω against $n = 1, \cdots, 500$.

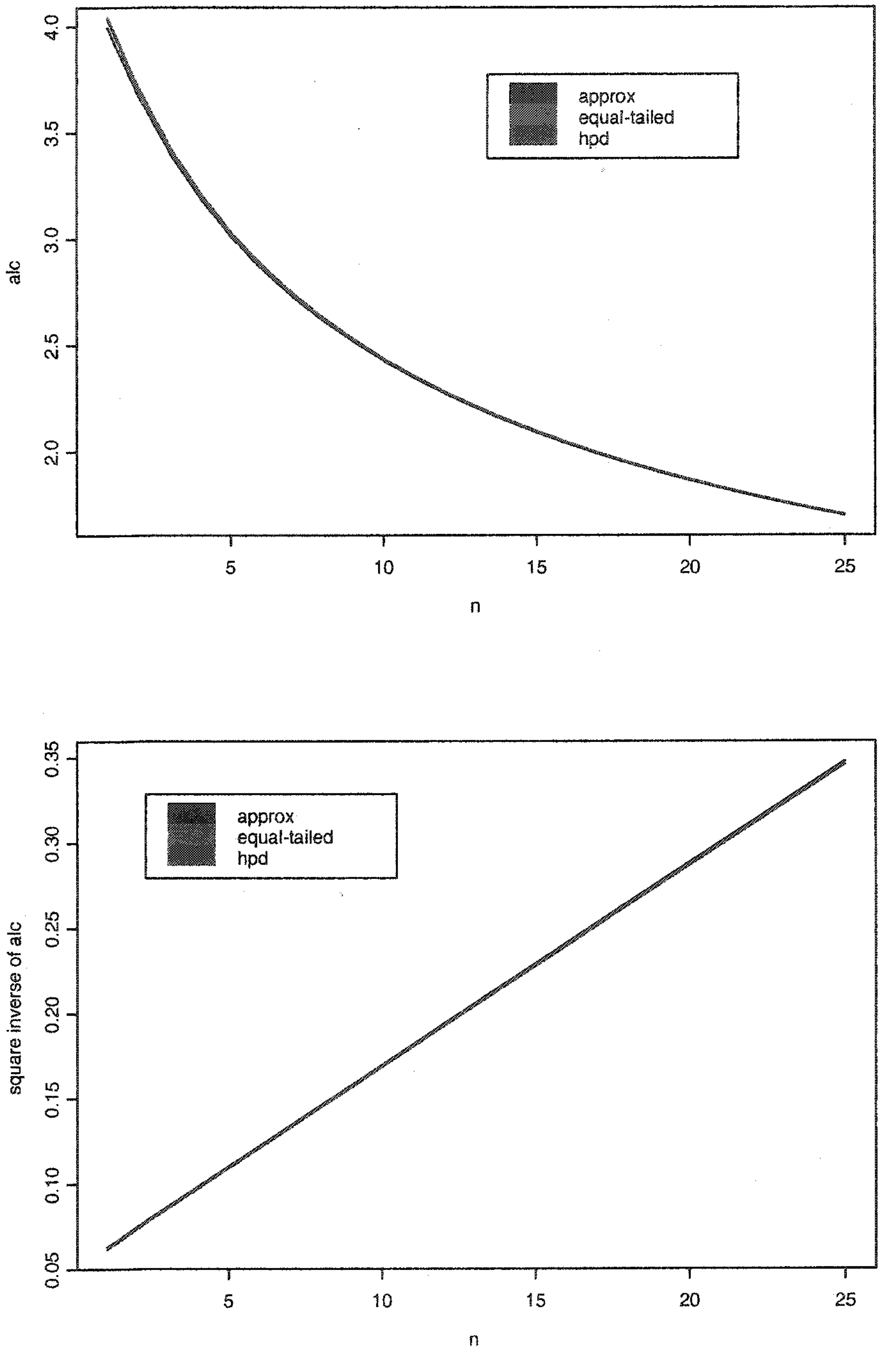


Figure 3.2.7: Graphs of alc$(n, 2, 2)$ (top) and $\dfrac{1}{\text{alc}^2(n,2,2)}$ (bottom) for the log-odds ϕ against $n = 1, \cdots, 25$.

values for a and b. Similar behavior was observed by Joseph et al., 1995 for a single proportion.

The linear relations observed across those plots is somewhat striking for the odds and log-odds parameters, and is extremely useful in practice. One use of these curves is that the sample size can be approximated by solving the equation $\dfrac{1}{\mathrm{alc}_k^2(n,a,b)} = \dfrac{1}{l^2}$ graphically, by recovering the coordinates of the intersection of the linear curves $\left(n, \dfrac{1}{\mathrm{alc}_k^2(n,a,b)}\right)$ and $\left(n, \dfrac{1}{l^2}\right)$, instead of graphically solving the less convenient original problem $\mathrm{alc}_k(n,a,b) = l$. These linear plots also suggest approximating sample sizes using the intercept and slope of these curves which might be estimated by least squares methodology. For instance, in the case of the proportion p_1, we have

$$\frac{1}{\mathrm{alc}^2(n,1,1)} = 1.5907273 + 0.422017\,n,$$

using the data composed of the pairs $(n, \mathrm{alc}(n,1,1))$, $n = 1, 2, \cdots, 3000, 3002, \cdots, 5000, 5005, \cdots, 10000$. This leads to a sample size formula

$$n = -3.7693386 + \frac{2.3695693}{l^2},$$

when solving the equation $\dfrac{1}{\mathrm{alc}^2(n,1,1)} = \dfrac{1}{l^2}$. When using equal-tailed intervals, the sample size formula becomes

$$n = -2.1352672 + \frac{2.3696086}{l^2}.$$

Both slopes are only negligibly different from the asymptotic slope 2.3696920 in equation (3.2.35).

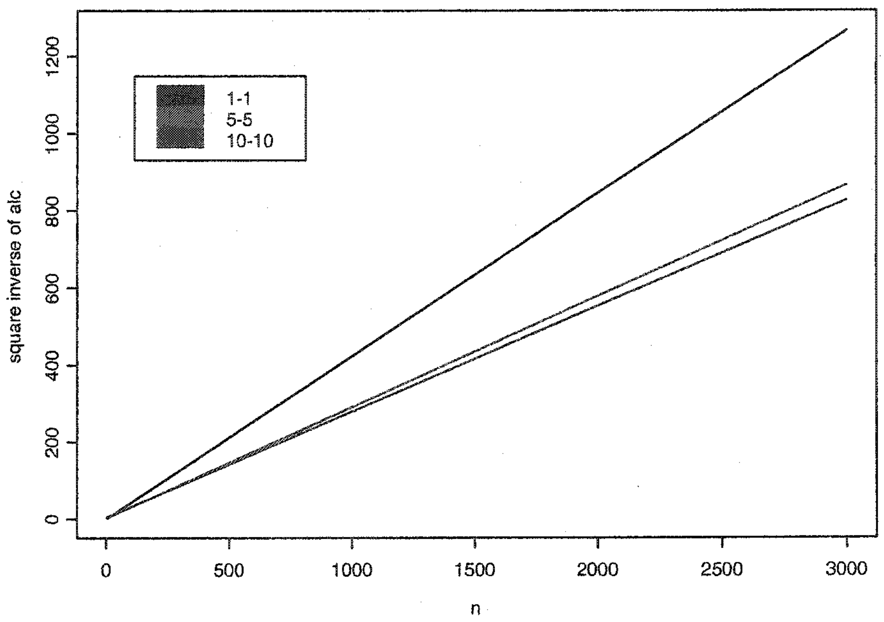
Consider the sample size for estimating the proportion p_1 with $(a,b) = (1,1)$, $1-\alpha = .95$, and $l = .05$, the odds ω with $(a,b) = (3,3)$, $1-\alpha = .95$,

Table 3.3: **Comparison of the approximate sample sizes for various m.**

Proportion $\mathbf{p_1}$: $(\mathbf{a},\mathbf{b}) = (\mathbf{1},\mathbf{1})$, $\mathbf{1}-\alpha = \mathbf{.95}$, $\mathbf{l} = \mathbf{.05}$							
m	200	500	1000	2000	10000	exact	eq. (3.2.27)
n	943	944	944	944	944	945	946
Odds ω : $(\mathbf{a},\mathbf{b}) = (\mathbf{3},\mathbf{3})$, $\mathbf{1}-\alpha = \mathbf{.95}$, $\mathbf{l} = \mathbf{.4}$							
m	200	500	1000	2000	10000	exact	eq. (3.2.33)
n	1267	1279	1283	1284	1284	1284	1296
Log-odds ϕ : $(\mathbf{a},\mathbf{b}) = (\mathbf{2},\mathbf{2})$, $\mathbf{1}-\alpha = \mathbf{.95}$, $\mathbf{l} = \mathbf{.4}$							
m	200	500	1000	2000	10000	exact	eq. (3.2.36)
n	528	528	528	528	528	529	530

and $l = .4$, and the log-odds ϕ with $(a, b) = (2, 2)$, $1 - \alpha = .95$, and $l = .4$, using the intercept and the slope of these linear curves based on the sequence $\left(n, \dfrac{1}{\text{alc}_k^2(n, a, b)}\right)_{1 \leq n \leq m}$ for various m. Results based on HPD intervals are presented in Table 3.3. Table 3.3 suggests that, in general, $m \geq 500$ provides sufficient accuracy when the true sample size is about one thousand, but even values as low as $m = 200$ are reasonably close.

Another important use of the linear plots discussed above is that they can serve for sensitivity analysis of the choice of prior distribution. The linear relation suggests that any qualitative statement on the comparison of two couples (a_1, b_1) and (a_2, b_2) for a given length and coverage will provide information for all lengths given that coverage. For instance, we can easily



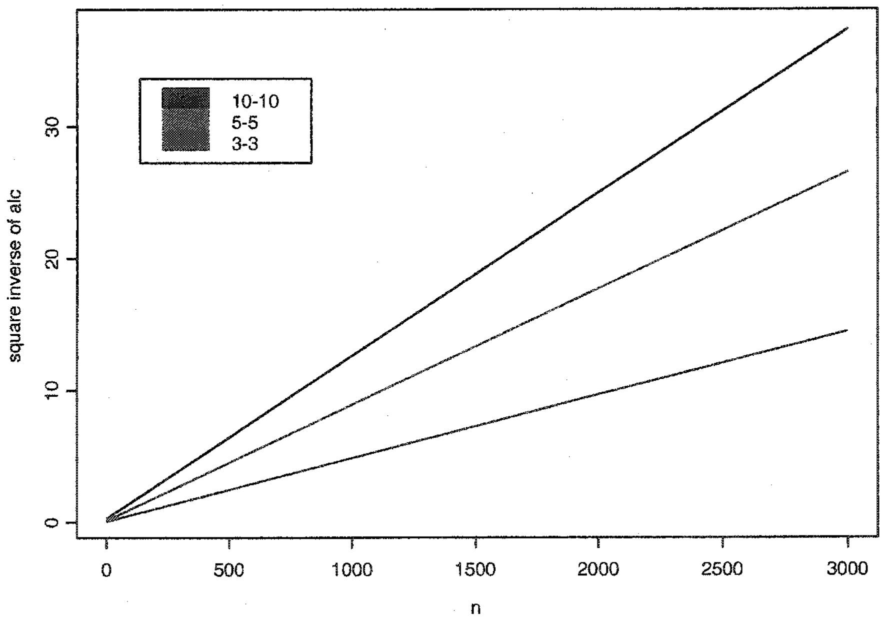


Figure 3.2.8: Graph of $\dfrac{1}{\text{alc}^2(n,a,b)}$ for the proportion p_1 (top) and the odds ω (bottom) as function of n for (a,b) as specified in the legend.

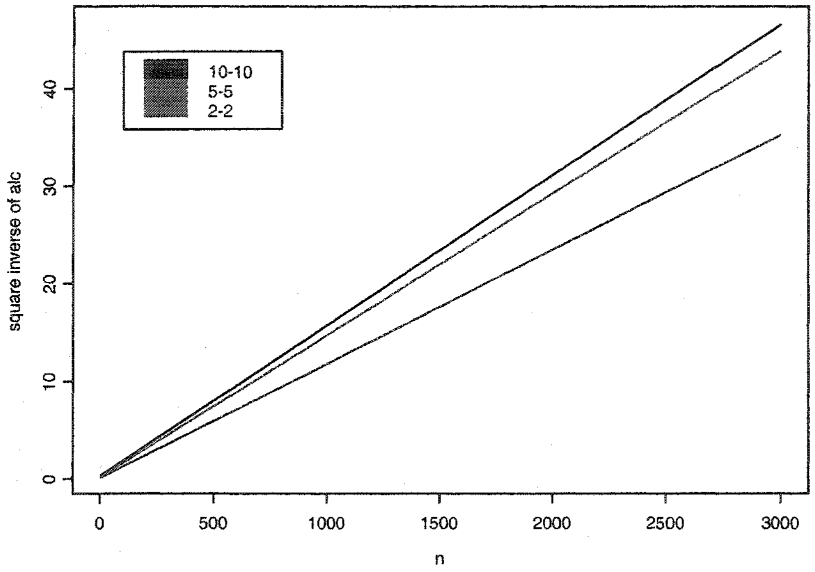


Figure 3.2.9: Graph of $\dfrac{1}{\text{alc}^2(n,a,b)}$ for the log-odds as function of n for (a,b) as specified in the legend.

read from these curves which prior parameters lead to smaller sample sizes for a given coverage. In Figures 3.2.8 and 3.2.9, we compare the linear plots for various values of $a = b$. The top half of Figure 3.2.8 suggests that the prior information contained in the Beta prior with parameter $(a, b) = (5, 5)$ is much closer to the "information", in a vague sense, contained in that of $(a, b) = (10, 10)$ than it is to the "information" conveyed by $(a, b) = (1, 1)$ when considering the proportion p_1.

The $\text{alc}(n, a, b)$ function for the odds ratio does not converge to zero as $n \to \infty$, but is rather is an increasing function of n for $(a, b) = (2, 1)$ and $(1, 1)$ as illustrated by Figure 3.2.10. We might be tempted to explain

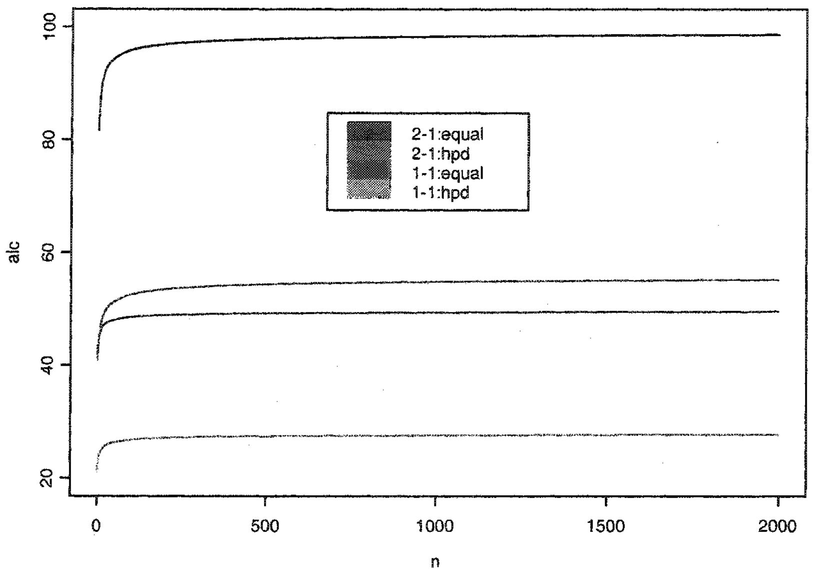


Figure 3.2.10: Graph of $\text{alc}(n, a, b)$ for the odds ω as a function of n for (a, b) as specified in the legend. $1 - \alpha = .95$ here.

such anomalies with the fact that one or more posterior variances are not defined, but this argument is insufficient since some posterior variances are also undefined for the case $(1, 2)$ where there is convergence. In general, for the **ALC**, we find empirically that $b > 1$ is necessary to have convergence to zero of the criterion function for ω. All these anomalies point to the fact that the $\text{alc}_k(n, a, b)$ criterion functions do not always converge to zero as $n \to \infty$, especially when the variance is not defined. Table 3.4 on page 115 presents several prior parameter values for which there is both convergence to zero and a linear relationship between $\dfrac{1}{\text{alc}_k^2(n, a, b)}$ and n. It seems that for the odds there is convergence to zero of the criterion function $\text{alc}_k(n, a, b)$ when

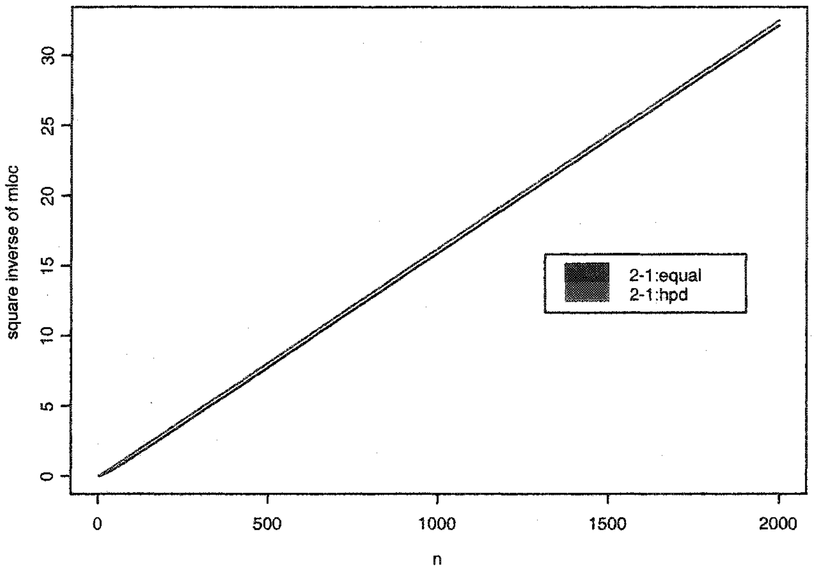


Figure 3.2.11: Graph of mloc(n, a, b) for the odds ω as function of n for (a, b) as specified in the legend.

$b > k$. We think that this phenomenon is related to the existence of the k-th moment of the parameter under consideration.

Although the criterion alc$_k$ does not converge for $(a, b) = (2, 1)$ and $(1, 1)$, similar behavior was not observed when considering the criterion **MLOC** (see Figure 3.2.11), which always converges. This means that the **MLOC** may be used to find the sample size regardless of the prior parameters used, and in particular, in the cases where the **ALC**$_k$ breaks down.

Figure 3.2.12 displays the criterion functions alc$(n, 1, 1)$, mloc$(n, 1, 1)$, and woc$(n, 1, 1)$ when p_1 is under consideration.

The following is a brief summary of the observed linearity relationships we

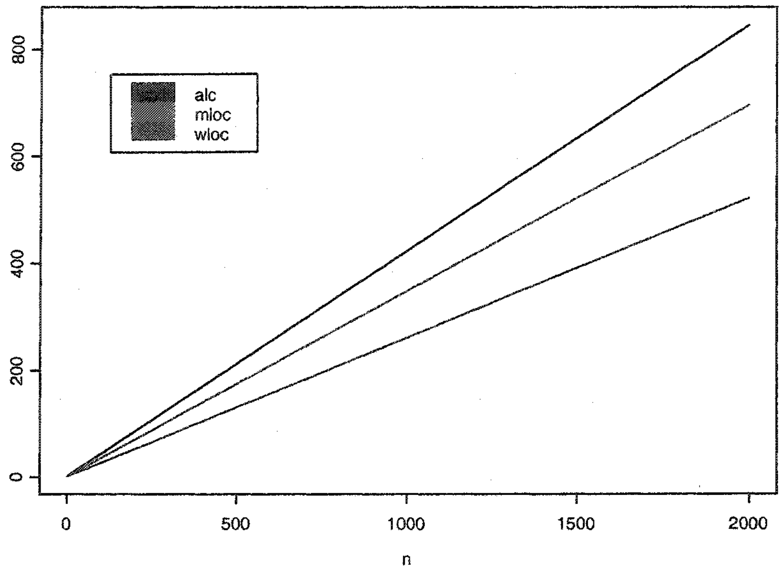


Figure 3.2.12: Graphs of alc$(n,1,1)$, mloc$(n,1,1)$, and mloc$(n,1,1)$ for the proportion p_1 as a functions of n.

have observed. Although we show empirically that there is a linear relation between the criterion function $\frac{1}{\text{alc}^2(n,a,b)}$ and n for limited values of (a,b), this linear relation seems to be very common to the $\mathbf{ALC}_k, k \geq 1$ and the **MLOC**. For proportions, we can include the **WOC**. In the case of **WOC** and **MLOC**, the approximate sample size formulae derived in subsections 3.2.4.3 and D.1 based on the posterior variances corroborates that conjecture. These formulae go even further by showing that there is unique slope irrespective of (a,b). For the odds, there exists a linear relation between $\frac{1}{\text{alc}_k^2(n,a,b)}$ and n whenever $b > 3k/2$ while for the log-odds, it happens when $a, b > k/2$. For other cases, empirical evidence not displayed here suggest that there is a linear relation between $\dfrac{1}{\mathrm{alc}_k^\lambda(n,a,b)}$ and n for some $\lambda \geq 2$ when the k-moment the parameter under consideration is defined.

A linear relation was not observed with the log-odds parameter for $(a,b) = (2,2)$ when using **WOC** although there is convergence. There is a more general pattern for the **WOC** when estimating the log-odds parameter since the existence of a linear relation implies that $\lim_n \mathrm{woc}(n,a,b) = 0$, which is not the case as $\max_{x_n} \mathbf{Var}(\phi|\mathbf{x}_n, n, a, b) > \max(\Psi'(a), \Psi'(b))$, where $\Psi'(x)$ is the trigamma function defined in section 3.2.4.3.

Figure 3.2.13 is typical of the displays that one obtains for $\mathrm{acc}(n,a,b)$ versus n. We have not been able to find the sort of linear relations observed before, for, say, $\dfrac{1}{\mathrm{acc}_k^2(n,a,b)}$ versus n.

Table 3.4: **Some prior parameters that imply convergence to zero and linearity of the criterion function $\mathrm{alc}_k(n,a,b)$ with respect to n.**

	proportion	odds	log-odds
convergence	$a,b>0$	$b>k$	$a,b>0$
linearity	$a,b>0$	$b>3k/2$	$a,b>k/2$

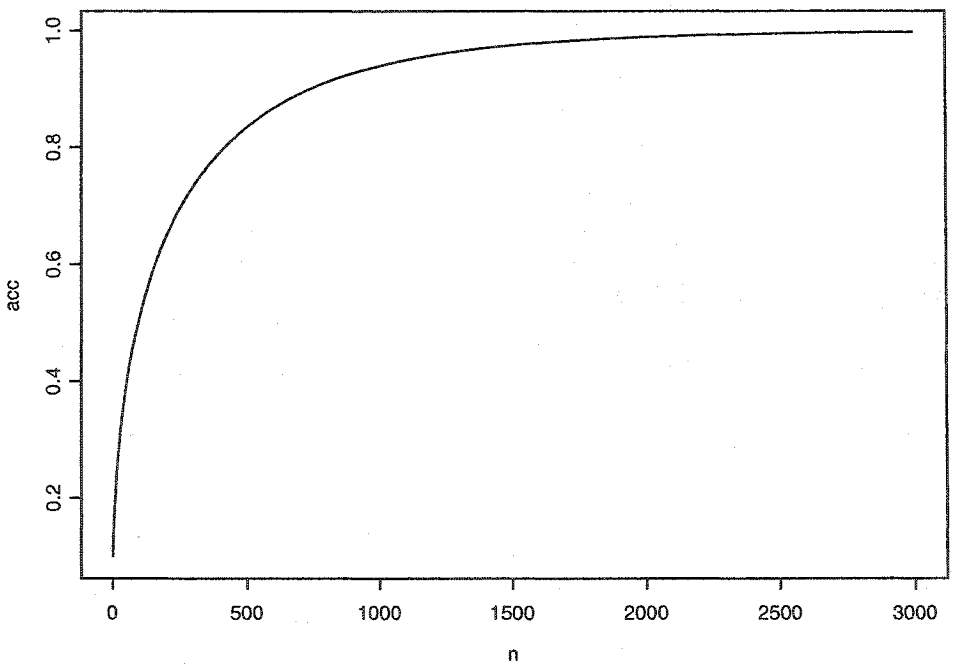


Figure 3.2.13: Graph of acc$(n, 1, 1)$ for the proportion p_1 as a function of n.

3.2.7.2 Comparison between the exact sample sizes and the various approximate sample sizes

In this subsection, we compare sample sizes resulting from the approximations developed in subsections 3.2.4 and 3.2.5 to the corresponding exact values for $1-\alpha = .90, .95$, and .99 and various lengths. The chosen prior parameters are $(1,1)$, $(5,5)$, and $(10,10)$ for the proportion, $(3,3)$, $(5,5)$, and $(10,10)$ for the odds, and $(2,2)$, $(5,5)$, and $(10,10)$ for the log-odds. The sample size formulae in use in this subsection are given by equations (3.2.28), (3.2.34), (3.2.36), (D.1.1), (D.2.1), and (D.3.1). The sample sizes are displayed in column 4 of Tables J.1-J.18 in appendix J. When estimating ϕ, the column labelled "limiting" under $\mathbf{ALC}_k$ use the criterion function

$$\widetilde{\mathrm{alc}}_k(n,a,b) = \left(\int_0^1 \left\{\int_{\phi\in\mathtt{HPD}(n,x,1-\alpha)} d\phi\right\}^k \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a,b)}\right)^{1/k} \quad (3.2.41)$$

where $\mathtt{HPD}(n,x,1-\alpha)$ represents the $\mathtt{HPD}$ interval of coverage $1-\alpha$ for estimating ϕ from the posterior distribution

$$f_\phi(\phi) = \frac{1}{\mathbf{Be}(nx+a, n(1-x)+b)} \frac{e^{(a+nx)\phi}}{(1+e^\phi)^{n+a+b}}. \quad (3.2.42)$$

For this criterion, one uses the limiting distribution of the pre-posterior predictive distribution in place of the pre-posterior predictive distribution. This criterion function can be interpreted as an approximation or an alternative criterion function to $\mathrm{alc}(n,a,b)$. Similarly for limiting sizes for the $\mathbf{ACC}_k$ and the parameters p_1 and ψ. One advantage of this criterion is that we can relax the requirement that n has to be an integer, thereby transforming the discrete problem of sample size computations into a continuous one. This allows the use of any minimization algorithms for smooth functions to be used.

Empty cells in Tables J.1-J.18 in appendix J are cells for which we were not able to derive the required sample sizes, often because the sample size is extremely large. We found empirically that **MLOC** = **MCOC** for the proportion, the odds and the log-odds, but have not yet been able to prove this formally.

For the parameter p_1 (see Tables J.1-J.6), the approximate sample size formulae in equations (3.2.26), (3.2.29), and (D.2.1) perform excellently compared to the exact solutions. The first and the more accurate third order approximation of HPD intervals turned out to be very good, although we did not find any large differences between HPD and equal-tailed intervals.

For odds (see Tables J.7-J.12), as expected, the third order approximation captures the difference between HPD and equal-tailed intervals well when using the $\mathbf{ALC}_k$. Often, for large l, the first order approximation tends to overestimate the optimal sample size. The third order approximation does poorly with the **MLOC**. The limiting distribution seems to give a slightly different estimate although the discrepancy is reduced as the prior parameters are increased.

For the log-odds (see Tables J.13-J.18), ϕ, the approximate values are uniformly close to the true values. This might be due to the fact that the distribution of ϕ approaches the normal distribution more rapidly than do the distributions of p_1 and ψ. The limiting distribution worked also performed well.

As an expected general behavior, we found that the larger the prior parameters the closer these approximations were to the true sample size.

Tables J.1-J.18 show that there are no large differences in the sample sizes based on HPD and equal-tailed intervals, when low information or symmetric priors are used, but this is not always the case, as illustrated by Table 3.5, where the first four rows support the necessity of a third order approximation. The discrepancy between the sample sizes provided by HPD and equal-tailed intervals can be substantial as is demonstrated by the last four rows for the odds. Unfortunately, in none of the remaining cases did the third order-based and the limiting distribution approach remain close to the exact values. Such behavior is somehow predictable in light of the requirement for the odds that $a > 1$ in order to make use of the third order approximation. Although, there is no such theoretical conditions on the log-odds parameter, in practice, we found that $a, b \geq k/2$ is necessary for the third order approximation to perform well.

As promised in the last paragraph of subsection 3.2.4.3, we now provide some general guidelines about which sample size approaches to use in practice. Since **MCOC** = **MLOC**, we only discuss the **MLOC**, but note that exact sample size computation for the **MLOC** should be replaced by exact computation for the **MCOC**, which is always faster.

3.2.8 General guidelines

- **For proportions:** When $a, b \geq 1$:

Table 3.5: **Comparison of the exact and third order approximation sample sizes when estimating ω and ϕ under ALC.**

Odds, ω							
$(a, b, 1-\alpha, l)$	**Exact**			**3rd order**		**limiting**	
	HPD	equal	formula	HPD	equal	HPD	equal
$(15.0, 3.0, .95, 5.0)$	458	494	497	453	487	394	426
$(15.0, 3.0, .95, 4.0)$	745	784	786	741	777	674	709
$(2.0, 2.0, .95, 1.0)$	652	716	764	633	683	559	602
$(1.0, 2.0, .95, 0.5)$	529	572	604	516	551	466	496
$(3.5, 1.6, .95, 6.0)$	408	638	2647	304	430	176	274
$(3.5, 1.5, .95, 6.0)$	1063	1787	N/A	748	1074	694	450
$(1.0, 1.4, .95, 3.0)$	292	646	N/A	154	271	73	154
$(1.0, 1.2, .95, 8.0)$	218	2153	N/A	1	85		29
$(3.0, 3.0, .95, 1.5)$	688	832	N/A	654	769	439	509
Log-odds, ϕ							
$(0.5, 0.5, .95, 1.0)$	662	725	N/A	649	702		561
$(0.3, 0.4, .95, 1.5)$	1851	2366	N/A	1791	2214		1376
$(0.3, 3.7, .95, 3.5)$	579	804	N/A	557	755		396

1. For the **ALC**$_k$, **WLOC** and the **MLOC**, use the approximate sample size formulae in equations (3.2.28), (3.2.29), and (D.2.1), respectively.

2. For the **ACC**$_k$, use the limiting distribution approach in subsection 3.2.4.4.

In other cases, for the **ACC**$_k$ and the **MLOC**, use the exact computation while for the **ALC**$_k$, use either the exact computation in subsection 3.2.3 or the Monte Carlo approach in subsections 3.2.5 and 3.2.6.

- **For the odds:**

  1. For the **ALC**$_k$, when $a, b \geq \dfrac{3(k+1)}{2}$, use the third order approximation to the length of the credible intervals in equations (3.2.21) and (3.2.22). These provide better approximations to the sample sizes than the approximate sample size formula given by equation (3.2.34). Equation (3.2.34) should only be used if $a, b \gg \dfrac{3(k+1)}{2}$. Otherwise, use the exact computation or the regression-based Monte Carlo approach.

  2. For the **ACC**$_k$ and the **MLOC**, use the exact computation.

- **For the log-odds:**

  1. For the **ALC**$_k$ and the **MLOC**: when $a, b \geq \dfrac{k+1}{2}$, use either the approximate sample size formulae in equations (3.2.34) and (D.3.1), the limiting distribution approach described by equation (3.2.42), or the third order approximation described by equations

(3.2.23) and (3.2.24). All three methods are equally efficient and accurate. Otherwise, use the exact computation or the regression-based Monte Carlo approach.

2. For the **ACC**$_k$: when $a, b \geq 1$, use any approach: exact, first order or limiting distribution. Elsewhere, use the exact computation.

Note that the log-odds is the only parameter where the limiting distribution approach often works well for all criteria.

In conclusion, the different approximations used to estimate the true sample sizes generally work well, but it is important to remember the conditions under which they apply.

3.3 Sample size when p_1 is restricted

In this subsection, we return to the problem of Bayesian sample size calculations, using the five criteria defined at the beginning of the chapter, namely **ALC**$_k$, **ACC**$_k$ $(1 < k < \infty)$, **WOC**, **MLOC** and **MCOC**, but consider the case where p_0 is exactly known and p_1 is restricted to the interval $p_1 < p_0$ or $p_1 > p_0$. It is widely acknowledged that cohort and case-control studies are rarely carried out in incomplete isolation. Therefore, as pointed by Marshall, 1988, "there may some sound biological or epidemiological factors which limits the size of the risk". Ignoring prior information leads to applying the same inference and sample size mechanism to both plausible and implausible exposures. We start in subsection 3.3.1 by suggesting types of prior distributions that could be used for the restricted unknown parameter

p_1, and explore the pre-posterior predictive distributions and posterior distributions that results from these choices. The computation of the sample sizes is Monte Carlo based, as exact or formulae based methods appear difficult. We discuss various examples in section 3.3.1.2.

3.3.1 Choice of prior distributions, and the predictive and posterior distributions

Contrary to section 3.2, we assume we know more than the usual $0 < p_1 < 1$ bound. We assume, for instance, that the risk ratio $R = \frac{p_1}{p_0} < 1$ or $R = \frac{p_1}{p_0} > 1$, although more generally, we might have bounds $0 < f < p_1 < h < 1$. Similarly for the odds ratio. Such assumptions are often made in sample size computations that are based on power considerations (Blackwelder, 1993), but are relevant in the Bayesian context, where this information must be incorporated into our Beta prior distribution for the parameter p_1. Two prior developed to accommodate these constraints have been proposed in the literature. The first proposed prior is the generalized beta distribution with four parameters, also called the Pearson type I distribution (Carnahan, 1989), while the second (Smith, 1975), perhaps more natural, is a truncated or incomplete Beta distribution on the interval (f, h) as the prior distribution. The latter maintains conjugacy with a binomial likelihood.

3.3.1.1 Generalized Beta Distribution

The generalized Beta distribution has density

$$f_{p_1}(p) = \frac{(p-f)^{a-1}(h-p)^{b-1}}{\mathbf{Be}(a,b)(h-f)^{a+b-1}}, \qquad 0 < f < p < h < 1, \qquad (3.3.1)$$

where h and f are the endpoints of the interval of support, and $a > 0$ and $b > 0$ are the shape parameters. This distribution will be denoted by $\mathbf{Be}(a, b, f, h)$. Note that $\dfrac{p_1 - f}{h - f}$ is a $\mathbf{Be}(a, b)$ random variable. It is easily seen that $1 - p \sim \mathbf{Be}(b, a, h, f)$. This prior distribution, combined with the likelihood function $\binom{n}{x} p_1^x (1-p_1)^{n-x}$ gives a posterior distribution

$$\begin{aligned} f_{p_1}(p|\mathbf{x}_n, n, a, b, f, h) &= \binom{n}{x_n} \frac{p^{\mathbf{x}_n}(1-p)^{n-\mathbf{x}_n}}{p_{X_n}(\mathbf{x}_n)} \frac{(p-f)^{a-1}(h-p)^{b-1}}{\mathbf{Be}(a,b)(h-f)^{a+b-1}}, \\ &\propto p^{\mathbf{x}_n}(1-p)^{n-\mathbf{x}_n}(p-f)^{a-1}(h-p)^{b-1}, \quad f < p < h, \end{aligned} \qquad (3.3.2)$$

where $f, h > 0$ and where p_{X_n} is the pre-posterior predictive distribution whose expression is given below. Although straightforward, this prior/likelihood model does not seem to have been previously discussed in detail.

Pre-posterior predictive distribution: The pre-posterior predictive distribution, p_{X_n}, has mass function

$$\begin{aligned}
p_{X_n}(\mathrm{x}_n|\, n, a, b, f, h) &= \binom{n}{\mathrm{x}_n} \int_f^h \frac{p^{\mathrm{x}_n}(1-p)^{n-\mathrm{x}_n}(p-f)^{a-1}(h-p)^{b-1}}{\mathbf{Be}(a,b)(h-f)^{a+b-1}}\, dp, \\
&= \binom{n}{\mathrm{x}_n} \int_{1-h}^{1-f} \frac{(1-p)^{\mathrm{x}_n} p^{n-\mathrm{x}_n} \{p-(1-h)\}^{a-1}\{(1-f)-p\}^{a-1}}{\mathbf{Be}(b,a)(h-f)^{a+b-1}}\, dp, \\
&= \frac{\binom{n}{\mathrm{x}_n}}{\mathbf{Be}(a,b)} \int_0^1 p^{a-1}(1-p)^{b-1} \times \\
&\qquad \{f+(h-f)p\}^{\mathrm{x}_n}\{1-f-(h-f)p\}^{n-\mathrm{x}_n}\, dp, \\
&= \frac{\binom{n}{\mathrm{x}_n}}{\mathbf{Be}(a,b)} \int_0^1 p^{b-1}(1-p)^{a-1} \times \\
&\qquad \{1-h+(h-f)p\}^{n-\mathrm{x}_n}\{h-(h-f)p\}^{\mathrm{x}_n}\, dp, \qquad (3.3.3)
\end{aligned}$$

for $\mathrm{x}_n = 0, 1, \ldots, n$. We also have the asymptotic result $\dfrac{X_n}{n} \rightarrow^d \text{Beta}(a, b, h, f)$ or equivalently $\dfrac{X_n - nf}{nh} \rightarrow^d \text{Beta}(a, b)$, as $n \longrightarrow \infty$.

We now consider the two special cases, $f = 0$ and $h = p_0$, and $f = p_0$ and $h = 1$, with respective corresponding posteriors

$$f_{p_1}(p|\mathrm{x}_n, n, a, b, 0, p_0) \propto (p_0 - p)^{b-1} p^{a+\mathrm{x}_n-1}(1-p)^{n-\mathrm{x}_n}, \qquad 0 < p < p_0, \tag{3.3.4}$$

and

$$f_{p_1}(p|\mathrm{x}_n, n, a, b, p_0, 1) \propto (p - p_0)^{a-1} p^{\mathrm{x}_n}(1-p)^{n+b-\mathrm{x}_n-1}, \qquad p_0 < p < 1. \tag{3.3.5}$$

These range restrictions correspond to the information that $R = \dfrac{p_1}{p_0} < 1$ and $R > 1$ in the cohort design, and $\psi_e = \dfrac{p_1 q_0}{q_1 p_0} < 1$ and $\psi_e > 1$, in the case control design, respectively.

Posterior distributions of R and ψ_e: The posterior distribution of R is proportional to

$$\begin{aligned} f_R(R|\mathbf{x}_n, n, a, b, 0, p_0) &\propto p_0^{a+b+\mathbf{x}_n-1} R^{a+\mathbf{x}_n-1}(1-R)^{b-1}(1-Rp_0)^{n-\mathbf{x}_n}, \\ & \qquad 0 < R < 1, \end{aligned} \tag{3.3.6}$$

when $f = 0$ and $h = p_0$, and

$$\begin{aligned} f_R(R|\mathbf{x}_n, n, a, b, p_0, 1) &\propto p_0^{a+\mathbf{x}_n} R^{\mathbf{x}_n}(R-1)^{a-1}(1-Rp_0)^{n+b-\mathbf{x}_n-1}, \\ & \qquad 1 < R < \frac{1}{p_0}, \end{aligned} \tag{3.3.7}$$

when $f = p_0$ and $h = 1$.

Similarly, the posterior distribution of ψ_e is

$$\begin{aligned} f_{\psi_e}(\psi|\mathbf{x}_n, n, a, b, 0, p_0) &\propto p_0^{a+b+\mathbf{x}_n-1} q_0^{n+b-\mathbf{x}_n} \frac{\psi^{a+\mathbf{x}_n-1}(1-\psi)^{b-1}}{(q_0+p_0\psi)^{n+a+b}}, \\ & \qquad 0 < \psi < 1, \end{aligned} \tag{3.3.8}$$

when $f = 0$ and $h = p_0$, and

$$\begin{aligned} f_{\psi_e}(\psi|\mathbf{x}_n, n, a, b, p_0, 1) &\propto p_0^{a+\mathbf{x}_n} q_0^{n+a+b-\mathbf{x}_n-1} \frac{\psi^{\mathbf{x}_n}(\psi-1)^{a-1}}{(q_0+p_0\psi)^{n+a+b}}, \\ & \qquad 1 < \psi < \infty, \end{aligned} \tag{3.3.9}$$

when $f = p_0$ and $h = 1$.

As we will show below, all four of the above posterior distributions are unimodal. Unimodality is a central property in this thesis since it is required by all the algorithms for the computation of HPD intervals.

Unimodality of R and ψ_e: The unimodality of R and ψ_e will follow as a corollary to Theorem 3.3.2 below. This corollary does not appear to have been stated elsewhere.

Lemma 3.3.1. *Let U be strongly unimodal random variable with an absolutely continuous density f_U. Then $V = \exp(U)$ is unimodal.*

Proof. The density of V is easily seen to be $f_V(v) = \dfrac{f_U(\log(v))}{v}$ and consequently $\log(f_V(v)) = \log(f_U(\log(v))) - \log(v)$. The differentiation of $\log(f_V(v))$ with respect to v gives

$$\frac{\partial \log(f_V(v))}{dv} = \frac{1}{v}\left\{\frac{f_U'(\log(v))}{f_U(\log(v))} - 1\right\}, \tag{3.3.10}$$

where $f_U'(u) = \dfrac{\partial f_U(u)}{du}$. Since U is strongly unimodal, $\dfrac{f_U'}{f_U}$ is decreasing. Therefore the right hand side of equation (3.3.10) can have at most one change in sign. If the sign does change, the change must be from positive to negative. This shows that $\exp(U)$ is unimodal. □

Theorem 3.3.2. *Let $p \sim \boldsymbol{Be}(a, b, f, h)$ and $\mathrm{x}_n|p \sim \boldsymbol{Bin}(n, p)$, with $0 \leq f < g \leq 1$, $a, b \geq 1$. Let $\omega = \dfrac{p}{1-p}$ and $\phi = \log(\psi)$. Then*

(i) The posterior distributions of p and ϕ given x_n are strongly unimodal (and therefore unimodal).

(ii) The posterior distribution of ω given x_n is unimodal.

Proof. The posterior density of p is given by equation (3.3.2). The strong unimodality of the posterior distribution of p given x_n is a direct consequence of

$$\frac{\partial^2 \log(f_p(p|\mathrm{x}_n, n, f, h, a, b))}{dp^2} = -\frac{\mathrm{x}_n}{p^2} - \frac{n-\mathrm{x}_n}{(1-p)^2} - \frac{a-1}{(p-f)^2} - \frac{b-1}{(h-p)^2} \leq 0.$$

The posterior density of ϕ given x_n is

$$f_\phi(\phi|\mathrm{x}_n, n, f, h, a, b) \;\propto\; \frac{\exp(\mathrm{x}_n\phi)}{(1+\exp(\phi))^{n+2}}\left\{\frac{\exp\phi}{1+\exp\phi} - f\right\}^{a-1} \times \left\{h - \frac{\exp\phi}{1+\exp\phi}\right\}^{b-1},$$

where $\log\left(\frac{f}{1-f}\right) < \omega < \log\left(\frac{h}{1-h}\right)$. The result follows since

$$\begin{aligned}\frac{(1+e^{\phi})^2}{e^{\phi}}\,\frac{\partial^2 \log\big(f_\phi(\phi|\mathrm{x}_n, n, f, h, a, b)\big)}{d\phi^2} &= -(n+2)-(a-1)\frac{(1-f)e^{2\phi}+f}{\left[f+(f-1)e^{\phi}\right]^2}\\ &\quad -(b-1)\frac{(1-h)e^{2\phi}+h}{\left[h+(h-1)e^{\phi}\right]^2} \leq 0.\end{aligned}$$

The unimodality of ω is then a straightforward application of Lemma 3.3.1.

□

Corollary 3.3.3. *Under the conditions of Theorem 3.3.2, R and ψ_e are unimodal.*

Proof. We have observed that R and ψ_e are linear transformations of the random variables defined in Theorem 3.3.2 under the conditions of interest. Since linear transformations of unimodal random variables are also unimodal, the unimodality of R and ψ_e follow. □

Simulating observations from the posterior distributions of R and ψ_e: There is no trivial way to simulate observations from equations (3.3.7), (3.3.8), (3.3.9) and (3.3.10). One possibility is to employ the inverse cumulative distribution technique on a grid of points lying in the interval (f, h), and another is the SIR-like algorithm proposed by Ross, 1996. A further idea, which is described below for the first time, is based on regarding the posterior distribution of R as a finite mixture of independent Beta distributions. Indeed, the binomial expansion

$$\begin{aligned}\left(1-Rp_0\right)^{n-\mathrm{x}_n} &= \left[q_0+p_0(1-R)\right]^{n-\mathrm{x}_n}\\ &= \frac{1}{n-\mathrm{x}_n+1}\sum_{j=0}^{n-\mathrm{x}_n}\frac{p_0^j\,q_0^{n-\mathrm{x}_n-j}}{\mathbf{B}(j+1, n-x_n-j+1)}\,(1-R)^j\end{aligned}$$

implies

$$\begin{aligned} f_R(R|\mathbf{x}_n, n, a, b, 0, p_0) &\propto p_0^{a+b+\mathbf{x}_n-1} R^{a+\mathbf{x}_n-1}(1-R)^{b-1}(1-Rp_0)^{n-\mathbf{x}_n}, \\ &\propto \sum_{j=0}^{n-\mathbf{x}_n} \frac{p_0^{a+b+\mathbf{x}_n+j-1}\, q_0^{n-\mathbf{x}_n-j}\, R^{a+\mathbf{x}_n-1}(1-R)^{b+j-1}}{(n-\mathbf{x}_n+1)\mathrm{B}(j+1, n-x_n-j+1)}, \\ &= \sum_{j=0}^{n-\mathbf{x}_n} w_j \frac{R^{a+\mathbf{x}_n-1}(1-R)^{b+j-1}}{\mathbf{Be}(a+\mathbf{x}_n, b+j)}, \qquad 0 < R < 1, \end{aligned}$$

where $w_j = \dfrac{v_j}{\sum_{k=0}^{n-\mathbf{x}_n} v_k}$ and $v_j = \dfrac{\mathbf{Be}(a+\mathbf{x}_n, b+j)}{\mathrm{B}(j+1, n-x_n-j+1)}\, p_0^{a+b+\mathbf{x}_n+j-1}\, q_0^{n-\mathbf{x}_n-j}$.

For the case of the posterior distribution with $1 < R < \dfrac{1}{p_0}$, that is

$$f_R(R|\mathbf{x}_n, n, a, b, p_0, 1) \propto p_0^{a+\mathbf{x}_n} R^{\mathbf{x}_n}(R-1)^{a-1}(1-Rp_0)^{n+b-\mathbf{x}_n-1},$$

first simulate observations from the random variable $0 < T = \dfrac{p_0}{q_0}(R-1) < 1$ and then compute $R = \dfrac{q_0}{p_0}T + 1$. Now we describe how one can generate an observation from the random variable T. We have

$$\begin{aligned} f_T(t|\mathbf{x}_n, n, a, b, p_0, 1) &\propto p_0^{\mathbf{x}_n}\, q_0^{n+a+b-\mathbf{x}_n-1}\, t^{a-1}\, (1-t)^{n+b-\mathbf{x}_n-1} \left(1+\frac{q_0}{p_0}t\right)^{\mathbf{x}_n}, \\ &\propto \frac{1}{\mathbf{x}_n+1} \sum_{j=0}^{\mathbf{x}_n} \frac{p_0^{\mathbf{x}_n-j}\, q_0^{n+a+b+j-\mathbf{x}_n-1}}{\mathrm{B}(j+1, \mathbf{x}_n-j+1)}\, t^{a+j-1}\, (1-t)^{n+b-\mathbf{x}_n-1} \\ &= \sum_{j=0}^{\mathbf{x}_n} w_j \frac{t^{a+j-1}\, (1-t)^{n+b-\mathbf{x}_n-1}}{\mathbf{Be}(a+j, n+b-\mathbf{x}_n)}, \end{aligned}$$

where $w_j = \dfrac{v_j}{\sum_{k=0}^{\mathbf{x}_n} v_k}$ and $v_j = \dfrac{\mathrm{B}(a+j, n+b-\mathbf{x}_n)}{\mathrm{B}(j+1, \mathbf{x}_n-j+1)}\, p_0^{\mathbf{x}_n-j}\, q_0^{n+a+b+j-\mathbf{x}_n-1}$.

To simulate random observations ψ_j from equations (3.3.9) and (3.3.10), we first simulate observations R_j from (3.3.7) and (3.3.8) as described above using the appropriate mixture of Beta distributions, and then set $\psi_j = \dfrac{q_0\, R}{1-p_0 R}$.

The generation of random samples of R and ψ_e can be used to estimate sample sizes via Monte Carlo algorithms as we will see in subsection 3.3.2.

3.3.1.2 Incomplete Beta Family

Smith (1975) proposed the probability density function

$$f_{p_1}(p) = \frac{p^{a-1}(1-p)^{b-1}}{k_{f,h}(a,b)}, \qquad 0 \leq f < p < h \leq 1, \tag{3.3.11}$$

where $a, b > 0$ are the shape parameters,

$$k_{f,h}(a,b) \;=\; k_{0,h}(a,b) - k_{0,f}(a,b), \tag{3.3.12}$$

and

$$k_{0,\pi_0}(a,b) = \int_0^{\pi_0} x^{a-1}(1-x)^{b-1}\, dp. \tag{3.3.13}$$

This distribution will be denoted by IBeta(a, b, f, h). It is easily seen that $1 - p_1 \sim$ IBeta$(b, a, 1-h, 1-f)$.

Posterior distribution of p_1: The incomplete Beta distribution combined with the Binomial(n, p) likelihood at x_n ($\mathrm{x}_n = 0, 1, \ldots, n$) gives a posterior distribution which is IBeta$(a + \mathrm{x}_n, b + n - \mathrm{x}_n, f, h)$. Its comparatively simple form will allow us to use the inverse cumulative distribution function technique to simulate observations from this posterior distribution.

Recall that the cases of interest to us are $0 < R < 1$ or $1 < R < \dfrac{1}{p_0}$ in the case-control setting, and $0 < \psi_e < 1$ or $1 < \psi_e < \infty$ in the case-control setting.

Under these conditions, the posterior distributions of $R = \dfrac{p_1}{p_0}$ in the cohort study is proportional to

$$f_R(R|\mathrm{x}_n, n, a, b, f, h) \propto \frac{p_0^{a+\mathrm{x}_n} R^{a+\mathrm{x}_n-1}(1-Rp_0)^{b+n-\mathrm{x}_n-1}}{k_{f,h}(a+\mathrm{x}_n, b+n-\mathrm{x}_n)} \tag{3.3.14}$$

for both intervals $0 < R < 1$ and $1 < R < \frac{1}{p_0}$. For the case-control study, we obtain that the posterior distribution of $\psi_e = \frac{p'_1 q'_0}{p'_0 q'_1}$ is proportional to

$$f_{\psi_e}(\psi|\mathrm{x}_n, n, a, b, f, h) \propto \frac{p_0'^{a+\mathrm{x}_n}\, q_0'^{b+n-\mathrm{x}_n}}{k_{f,h}(a+\mathrm{x}_n, n+b-\mathrm{x}_n)} \frac{\psi^{a+\mathrm{x}_n-1}}{(q'_0 + p'_0\psi)^{n+a+b}}. \qquad (3.3.15)$$

To simulate random variables from equations (3.3.14) and (3.3.15), we will use the cumulative distribution technique. When b is an integer, then $f_R(R|\mathrm{x}_n, n, a, b, f, h),\ 0 < R < 1$ can be regarded as a finite mixture of independent Beta distributions,

$$f_R(R|\mathrm{x}_n, n, a, b, f, h) = \sum_{j=0}^{b+n-\mathrm{x}_n-1} w_j \frac{R^{a+\mathrm{x}_n}(1-R)^{j+1}}{\mathrm{Be}(a+\mathrm{x}_n, j+1)}, \qquad 0 < R < (3.3.16)$$

with $w_j = \frac{v_j}{\sum_{k=0}^{b+n-\mathrm{x}_n-1} v_k}$ and $v_j = \frac{\mathrm{B}(a+\mathrm{x}_n, j+1)}{\mathrm{B}(b+n-\mathrm{x}_n-j, j+1)}\, p_0^{a+\mathrm{x}_n+j-1}\, q_0^{b+n-\mathrm{x}_n-j-1}$.
For $f_R(R|\mathrm{x}_n, n, a, b, f, h),\ 1 < R < \frac{1}{p_0}$, with a an integer we have

$$f_T(t|\mathrm{x}_n, n, a, b, f, h) = \sum_{j=0}^{a+\mathrm{x}_n-1} w_j \frac{t^{n+b-\mathrm{x}_n}(1-t)^{j+1}}{\mathrm{Be}(n+b-\mathrm{x}_n, j+1)}, \qquad (3.3.17)$$

where $w_j = \frac{v_j}{\sum_{k=0}^{a+\mathrm{x}_n-1} v_k}$ and $v_j = \frac{\mathrm{B}(b+n-\mathrm{x}_n, j+1)}{\mathrm{B}(a+\mathrm{x}_n-j, j+1)}\, p_0^{a+\mathrm{x}_n-j-1}\, q_0^{b+n-\mathrm{x}_n+j}$
and $T = \frac{p_0}{q_0}(R-1)$

If b is large, the above decomposition can be regarded as a close approximation to the posterior distribution of R.

Pre-posterior predictive distribution: The pre-posterior predictive (marginal) distribution associated with the incomplete beta prior density is

$$p_{X_n}(\mathrm{x}_n|\, n, a, b, f, h) = \binom{n}{\mathrm{x}_n} \frac{k_{f,h}(a+\mathrm{x}_n, n+b-\mathrm{x}_n)}{k_{f,h}(a, b)}. \qquad (3.3.18)$$

Again $\frac{X_n}{n} \to^d \mathrm{IBeta}(a, b, f, h)$.

Table 3.6: **Monte Carlo-based sample size calculations for the ALC and ACC using an incomplete beta prior distribution.** $p_0 = 0.1$.

ALC — — — $\mathbf{0 < p_1 < p_0}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi < 1$	risk ratio, $R < 1$
(3, 3, 0.5, .95)	$\bar{n} = 133$ $\text{std}_{\bar{n}} = 0.543$	$\bar{n} = 107$ $\text{std}_{\bar{n}} = 0.731$
ALC — — — $\mathbf{p_0 < p_1 < 1}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi > 1$	risk ratio, $R > 1$
$(3, 12, 1.0, .95)$	$\bar{n} = 901$ $\text{std}_{\bar{n}} = 1.875$	$\bar{n} = 324$ $\text{std}_{\bar{n}} = 0.547$
$(6, 4, 1.0, .95)$		$\bar{n} = 580$ $\text{std}_{\bar{n}} = 0.563$
ACC — — — $\mathbf{0 < p_1 < p_0}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi < 1$	risk ratio, $R < 1$
$(3, 3, 0.5, .95)$	$\bar{n} = 187$ $\text{std}_{\bar{n}} = 0.526$	$\bar{n} = 164$ $\text{std}_{\bar{n}} = 0.651$
ACC — — — $\mathbf{p_0 < p_1 < 1}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi > 1$	risk ratio, $R > 1$
$(3, 12, 1.0, .95)$	$\bar{n} = 2068$ $\text{std}_{\bar{n}} = 12.925$	$\bar{n} = 409$ $\text{std}_{\bar{n}} = 0.737$
$(6, 4, 1.0, .95)$		$\bar{n} = 615$ $\text{std}_{\bar{n}} = 0.943$

Table 3.7: **Monte Carlo-based sample size calculations for the ALC and ACC using the generalized beta prior distribution.** $p_0 = 0.1$.

ALC — — — $\mathbf{0 < p_1 < p_0}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi < 1$	risk ratio, $R < 1$
$(3, 3, 0.5, .95)$	$\bar{n} = 131$ $\text{std}_{\bar{n}} = 0.149$	$\bar{n} = 131$ $\text{std}_{\bar{n}} = 0.233$
ALC — — — $\mathbf{p_0 < p_1 < 1}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi > 1$	risk ratio, $R > 1$
$(3, 12, 1.0, .95)$	$\bar{n} = 1012$ $\text{std}_{\bar{n}} = 2.675$	$\bar{n} = 265$ $\text{std}_{\bar{n}} = 0.373$
$(6, 4, 1.0, .95)$		$\bar{n} = 311$ $\text{std}_{\bar{n}} = 0.359$
ACC — — — $\mathbf{0 < p_1 < p_0}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi < 1$	risk ratio, $R < 1$
$(3, 3, 0.5, .95)$	$\bar{n} = 135$ $\text{std}_{\bar{n}} = 0.249$	$\bar{n} = 134$ $\text{std}_{\bar{n}} = 0.233$
ACC — — — $\mathbf{p_0 < p_1 < 1}$		
$(a,\ b,\ l,\ 1-\varepsilon)$	odds ratio, $\psi > 1$	risk ratio, $R > 1$
$(3, 12, 1.0, .95)$	$\bar{n} = 1240$ $\text{std}_{\bar{n}} = 5.707$	$\bar{n} = 272$ $\text{std}_{\bar{n}} = 0.359$
$(6, 4, 1.0, .95)$		$\bar{n} = 315$ $\text{std}_{\bar{n}} = 0.850$

3.3.2 Results and applications

To illustrate our methods, consider the problem of determining the sample sizes in the context of the restricted model for estimating the relative risk and odds ratio with the **ALC** and **ACC**. Let (a, b) represent the prior distribution parameters in the models in equations 3.3.1 and 3.3.11, $p_0 = 0.1$ the probability of success among control (case-control study) or non-exposed (cohort study), and let $1 - \alpha = 0.95$ be the desired coverage level. As usual, in cohort settings, the parameter investigated is the risk ratio and in case-control settings, the parameter is the odds ratio. Let $\bar{n}$ represent the average of 10 Monte Carlo estimated sample sizes. For each of the two Monte Carlo steps involved in the computation of criteria functions, we simulated m=M=2000 observations, using the algorithm described in subsection 3.2.5.

Tables 3.6 and 3.7 provides results. Three sets of prior parameters were chosen, namely, (3, 3), (3, 12), and (6, 4), representing typical cases where p_1 equally takes values near 0 and 1, where p_1 more often takes values near 1 than 0, and vice-versa, respectively. The mean and standard deviation of the 10 trials are provided in these tables along with their mean, which is highlighted in blue. The empty cells in Table 3.6 and 3.7 are cases for which the sample sizes are larger than 20,000. It is clear from Tables 3.6 and 3.7 that the sample sizes based on case-control study for estimating the odds ratio are larger than the sample sizes based on cohort for estimating the relative risk. This behavior is marked when $(a, b) = (6, 4)$. The sample size based on the generalized beta prior distribution (see Table 3.7) under **ALC** and **ACC** are similar. For the incomplete Beta (see Table 3.6), there is a marked gap between the sample sizes provided by **ALC** and **ACC**.

3.4 Conclusion

In this chapter, we addressed the sample size problem for inference in exposure-only and case-only settings under the Bayesian paradigm for each of the five criteria $\mathbf{ALC}_k$, $\mathbf{ACC}_k$, **WOC**, **MLOC** and **MCOC**. The three parameters of main interest in epidemiological studies, namely, $R = \frac{p_1}{p_0}$, $\psi_e = \frac{p_1(1-p_0)}{p_0(1-p_1)}$, and $\phi_e = \log(\psi_e)$ are studied, where $p_0 < 1$ is known. We studied both HPD and equal-tailed intervals. HPD intervals are optimal in the sense that they lead to the smallest possible sample sizes under any criterion. One advantage to using equal-tailed intervals is that they are simpler to compute and therefore computationally efficient. When there are no restrictions on p_1, we developed three approaches: exact, approximate, and Monte Carlo-based. We derived sample size formulae for $\mathbf{ALC}_k$, **WOC**, and **MLOC**, and also discuss the approximate linear relationship between, for instance, $\frac{1}{\text{alc}_k^2(n,a,b)}$ and n. This linear relation is exploited to reduce Monte Carlo errors by fitting a regression equation to the Monte Carlo sample.

When imposing conditions of the type $0 < p_1 < p_0$ or $p_0 < p_1 < 1$ on p_1, we investigate two families of prior distributions that can be used to model these restrictions. We derived the posterior distributions of R, ψ_e, and ϕ_e arising from these prior and showed how one can simulate observations from these posteriors. Numerous tables of sample sizes are provided in appendix J.

We have established a simple, but extremely important, relation between the one sample problem and both exposure-only and case-only designs. As a consequence, not only have we solved the question of sample size determination for exposure-only and case-only designs, we have also presented an alternative solution to the sample size calculation problem for estimating a single proportion p_1 extending the work by Joseph et al., 1995. Here we looked at the odds, $\omega = \frac{p_1}{1-p_1}$, and the log-odds $\phi = \log(\omega)$, parameters often of interest in areas such quality control (Berger, 1998).

In the next chapter we derive Bayesian interval-based sample size methods for estimating risk and odds ratios for the two sample problem, where p_0 is also unknown. Four major designs in epidemiologic are studied, namely, cohort, case-control, cross-sectional and matched designs. Most of the ideas developed in this chapter carry over again in the next chapter.

Chapter 4

Bayesian interval-based sample size determination for estimating risk and odds ratios for two sample problems

In this chapter, we investigate the problem of sample size determination in case-control, cohort, matched, and cross-sectional designs. Let D and E again represent the disease and exposure status, as described in section 2.3. Define $p_0 = \Pr(D = 1|E = 0)$ and $p_1 = \Pr(D = 1|E = 1)$, the conditional probabilities of disease among both exposed and non-exposed subjects in a cohort setting, and let $p'_0 = \Pr(E = 1|D = 0)$ and $p'_1 = \Pr(E = 1|D = 1)$, the conditional exposure probabilities among diseased and non-diseased subjects in a case-control setting. In cohort studies, we investigate the relative risk, $R = \dfrac{p_1}{p_0}$ as well as $\log(R)$, while in case-control studies we investigate the exposure odds ratio, $\psi_e = \dfrac{p'_1(1-p'_0)}{p'_0(1-p'_1)}$ and $\log(\psi_e)$. Let $p_{11} = \Pr(D=1, E=1)$, $p_{10} = \Pr(D=1, E=0)$, $p_{01} = \Pr(D=0, E=1)$, and $p_{00} = \Pr(D=0, E=0)$ be the cell probabilities in the setting of a cross-sectional study. In cross-sectional studies, we investigate four parameters, the risk ratio, $R = \dfrac{p_{11}(p_{10}+p_{00})}{p_{10}(p_{11}+p_{01})}$, the odds ratio, $\psi = \dfrac{p_{11}p_{00}}{p_{10}p_{01}}$, along with $\log(R)$ and $\phi = \log(\psi)$. Finally, let p'_{11}, p'_{10}, p'_{01}, and p'_{00} be the probabilities of the pairs $(+,+)$, $(+,-)$, $(-,+)$ and $(-,-)$, respectively (see Table 2.2) in a pair-matched setting. In a pair-matched analysis, we investigate the exposure odds ratio, $\psi'_e = \dfrac{p'_{10}}{p'_{01}}$ and $\phi'_e = \log(\psi'_e)$.

In contrast to the previous chapter, we do not assume that any of the proportions involved here are a priori exactly known. We consider the sample size problem in two different contexts. The first context, the unrestricted model, places no restrictions on any of the "success" probabilities defined above. In cohort and case-control studies, we define a second scenario, the restricted model, which deals with cases where we know a priori that $0 < p_0 < p_1 < 1$ or $0 < p_1 < p_0 < 1$ (cohort studies) and $0 < p'_0 < p'_1 < 1$ or $0 < p'_1 < p'_0 < 1$ (case-control studies). In other words, we know a priori that $R < 1$ or $R > 1$ and $\psi_e > 1$ or $\psi_e < 1$. The restrictions on the pairs (p_0, p_1) and (p'_0, p'_1) can also be seen as a way to break the usual hypothesized independence between columns and rows of Table 2.1. To our knowledge, no one has previously considered sample size calculations for interval estimation from this viewpoint. This allowance for dependence allows for wider applicability of our results than currently available results in the literature.

Throughout this thesis, for all the designs investigated, we find sample sizes for both HPD and equal-tailed intervals. Unfortunately, in the two sample problem, in contrast to the one sample problem, we do not develop exact or third order approximation approaches because of complications dues to the presence of a nuisance parameter. We are left with only three options: a sample size formula approach, a "straight" Monte Carlo approach, and a regression-based Monte Carlo approach. We derive sample size formulae for the unrestricted model where an explicit expression of posterior variances is possible. In practice, however, a Monte-Carlo approach is frequently the only option for the two-sample problem. Since Monte-Carlo approaches have been described extensively in chapter 3, we only give a short outline of the main steps in this chapter.

We also consider the choice of the optimal ratio of controls per case, $g = \frac{n_0}{n_1}$, in case-control studies or the ratio of non-exposed to exposed, $g = \frac{m_0}{m_1}$, in cohort studies, in the sense that the total sample size, $N = n_0 + n_1 = (g+1)n_1(g)$ or $N = m_0 + m_1 = (g+1)m_1(g)$, respectively, is minimized over $0 < g < \infty$, while still attaining the desired estimation precision. The optimal sample size N then corresponds to the overall minimal sample size over the set $(n_1, n_0) \in \mathbb{N}$ or $(m_1, m_0) \in \mathbb{N}$. The combined problem of sample size and optimality of g are easily addressed by minimizing the cost $C = c_1 n_1(g) + c_0 n_0(g) = (g+r)c_0 n_1(g)$ in case-control studies, where c_1 and c_0 represent the cost per case and per control, and where $r = \frac{c_1}{c_0}$ is the cost ratio of cases to controls. Similarly for cohort studies.

This chapter can be divided into two parts, according to the sample size criteria used. Section 4.1 discusses the five criteria $\mathbf{ALC}_k, \mathbf{ACC}_k$, **WOC**, **MLOC**, and **MCOC**. Section 4.2 addresses a new set of Bayesian sample size criteria that average over nuisance parameters. Section 4.1 is divided in four subsections: cohort studies, case-control studies, cross-sectional studies, and pair-matched studies. The subsections concerning cohort and case-control studies are also divided in two: the unrestricted model and the restricted model. In general, irrespective of the design, the main steps for the unrestricted cases are:

- State the required posterior densities.
- Prove the unimodality of the posterior density.
- State the required predictive distribution.
- Using the unimodality of the posterior density, derive approximate sample size formulae and discuss the optimality of g.
- For cohort and case-control studies, extend the problem of sample size to the problem of reducing cost. We also derive approximate cost formulae and again discuss the optimality of g in this context.
- Give a sketch of the Monte Carlo approach to sample size calculation.

For the restricted cases, the main steps are:

- State the model for the restrictions, and find the posterior density.

- Prove the unimodality of the posterior density.
- State the required predictive distribution.
- Derive algorithms to simulate observations from the posterior distribution for the four prior-likelihood models under investigation.
- Using the unimodality of the posterior density and the simulated observations from the posterior distribution, carry out a Monte Carlo approach to sample size calculation.

The first set of sample size criteria considered are the **ALC**$_k$, **ACC**$_k$, **WOC** **MLOC**, and **MCOC**.

4.1 Sample size calculations based on the ALC$_k$, ACC$_k$, WOC, MLOC, and MCOC

4.1.1 Cohort studies

We again use the notation $\mathtt{T} = (\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d})$, first introduced in section 3.3 to represent the data in Table 2.1. Cohort studies are mainly concerned with estimating the risk ratio $R = \frac{p_1}{p_0}$. Let $\mathtt{T} = (\mathtt{a}, \mathtt{b}, \mathtt{m}_1 - \mathtt{a}, \mathtt{m}_0 - \mathtt{b})$ denote an observed table from a cohort study. In addition to independent sampling both between and within the exposed and non-exposed groups, we assume that $p_1 = \Pr(D = 1|E = 1)$ and $p_0 = \Pr(D = 1|E = 0)$ are a priori independent with

$$\begin{array}{lll} a \sim \mathbf{Bin}(m_1, p_1) & \text{given } E = 1 \text{ and } p_1, & p_1 \sim \mathbf{Be}(a', c'), \text{ and} \\ b \sim \mathbf{Bin}(m_0, p_0) & \text{given } E = 0 \text{ and } p_0, & p_0 \sim \mathbf{Be}(b', d'). \end{array}$$

Let $\mathtt{T}' = \left(\mathsf{a}', \mathsf{b}', \mathsf{c}', \mathsf{d}'\right)$ denote the table of prior parameters. Let g be the ratio of non-exposed to exposed subjects such that $m_0 = gm_1$. The constant g is an integral part of the design, and is selected either optimally, as will be discussed in the sequel, or by more practical considerations. The combination of the prior-likelihood tables, $\mathtt{T}'$ and $\mathtt{T}$, leads to a posterior table, $\mathtt{T}'' = \left(\mathsf{a}'', \mathsf{b}'', \mathsf{c}'', \mathsf{d}''\right)$ where $a'' = a + a'$, $b'' = b + b'$, $c'' = m_1 + c' - a$, $d'' = m_0 + d' - b$. $N = m_0 + m_1$ is the total sample size. We investigate two models for the computation of sample size. The first model does not place any restrictions on the values of p_1 and p_0. The second model assumes that either $0 < p_1 < p_0$ or $p_0 < p_1 < 1$, leading to $R < 1$ or $R > 1$, respectively.

4.1.1.1 Case when p_1 and p_0 are unrestricted

Under this model, the posterior density of the relative risk, first derived in chapter 2 (see equation (2.4.5)), R, is

$$p_R(R \,|\, \mathtt{T}'') = \begin{cases} \dfrac{R^{a''-1}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1}\left[1 - R \cdot Z\right]^{c''-1} dZ, & \\ \hfill 0 \le R < 1 & \\ \\ \dfrac{R^{-(b''+1)}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1}\left[1 - \dfrac{1}{R} \cdot Z\right]^{d''-1} dZ, & \\ \hfill R \ge 1 & \end{cases} \tag{4.1.1}$$

where $K = \mathbf{B}(a'', c'') \times \mathbf{B}(b'', d'')$, and the posterior density of the logarithm of the relative risk, $\log(R)$, is

$$p_{\log(R)}(R \,|\, \mathrm{T}'') = \begin{cases} \dfrac{e^{a''R}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1}\left[1 - e^{R} \cdot Z\right]^{c''-1} dZ, \\ \hfill R < 0 \\ \\ \dfrac{e^{-b''R}}{K} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1}\left[1 - e^{-R} \cdot Z\right]^{d''-1} dZ, \\ \hfill R \geq 0. \end{cases} \tag{4.1.2}$$

Unimodality All of the Monte Carlo based algorithms used for the computation of HPD intervals require unimodality of the posterior density. We prove below when $c'', d'' > 1$ that the posterior density of R is unimodal and the posterior density of $\log(R)$ is strongly unimodal. It is clear that R is not strongly unimodal since strongly unimodal densities possess all their k-th moments finite, for all k, whereas the k-th moment of R

$$\mathbf{E}(R^k|\mathrm{T}'') = \frac{\mathbf{B}(a + a' + k, m_1 - a + c')\,\mathbf{B}(b + b' - k, m_0 - b + d')}{\mathbf{B}(a + a', m_1 - a + c')\,\mathbf{B}(b + b', m_0 - b + d')} \tag{4.1.3}$$

is only defined for $0 \leq k < b + b'$.

Lemma 4.1.1. *If $p \sim Be(\alpha, \beta)$ with $\alpha > 0$ and $\beta \geq 1$, then $\log(p)$ is strongly unimodal.*

Proof. The density of $\log(p)$ is

$$f_{\log(p)}(z) = \frac{e^{\alpha z}(1 - e^z)^{\beta - 1}}{\mathbf{B}(\alpha, \beta)}, \qquad z < 0,$$

so that $\log\big(f_{\log(p)}(z)\big) = \alpha z + (\beta - 1)\log(1 - e^z) - \log \mathbf{B}(\alpha, \beta)$ and $\dfrac{\partial \log\big(f_{\log(p)}(z)\big)}{\partial z} =$

$\alpha - \dfrac{(\beta-1)e^z}{1-e^z}$. Thus $\dfrac{\partial^2 \log\big(f_{\log(p)}(z)\big)}{\partial z^2} = -\dfrac{(\beta-1)e^z}{(1-e^z)^2}$ is non-positive when $\beta \geq 1$. □

Theorem 4.1.2. *Let $p_1 \sim Be(\alpha_1, \beta_1)$ and $p_0 \sim Be(\alpha_0, \beta_0)$ be two independent random variables with $\alpha_1, \alpha_0 > 0$ and $\beta_1, \beta_0 \geq 1$. Define $\varphi = \dfrac{p_1}{p_0}$. Then*

(i) $\log(\varphi)$ is strongly unimodal, and, therefore, unimodal.

(ii) φ is unimodal.

Proof. (i) Lemma 4.1.1 combined with Proposition 2.7.1 implies that $\log(\varphi) = \log\left(\dfrac{p_1}{p_0}\right) = \log(p_1) - \log(p_0)$ is strongly unimodal, since it is the difference of two independent strongly unimodal random variables.

(ii) Lemma 3.3.1 guarantees the unimodality of $\varphi = \dfrac{p_1}{p_0} = e^{\log(\varphi)}$.

□

It is well known that the posterior distribution of $p_1 \sim \mathbf{Be}(a'', c'')$ and the posterior distribution of $p_0 \sim \mathbf{Be}(b'', d'')$, are independent. Under the assumption that $c'', d'' \geq 1$, a corollary to Theorem 4.1.2 is that $\log(R)$ is strongly unimodal and R is unimodal. A similar statement holds for $R_1 = \dfrac{p_0}{p_1}$. For the other variables $R_2 = \dfrac{1-p_1}{1-p_0}$ and $R_3 = \dfrac{1-p_0}{1-p_1}$ we have unimodality when $a'', b'' \geq 1$. A consequence of the unimodality of R is that $p_R(R \mid \mathrm{T}'')$ is decreasing when $a'' \leq 1$.

The pre-posterior predictive distribution Also of primary importance in the computation of sample size, is the pre-posterior predictive distribution.

In the cohort setting, we have

$$p(a,b|m_1,m_0,a',b',c',d') = \binom{m_1}{a}\frac{\mathrm{B}(a'',c'')}{\mathrm{B}(a',c')} \times \binom{m_0}{b}\frac{\mathrm{B}(b'',d'')}{\mathrm{B}(b',d')}, \quad (4.1.4)$$

for $a = 0,\cdots,m_1,$ and $b = 0,\cdots,m_0$ which is the distribution of two independent Beta-Binomial random variables.

Note 4.1.1. *Since $m_0 = gm_1$, m_1 tends to ∞ as m_0 tends to ∞. This together with the independence shown in equation* (4.1.4) *and Theorem 3.2.6 applied to each $\frac{a}{m_1}, \frac{b}{m_0}$, implies that we have that $Y_{m_1} = \left(\frac{a}{m_1}, \frac{b}{m_0}\right) \longrightarrow^d$ (p_1, p_0). This result is used implicitly in the proof of Corollary 4.1.3.*

We are ready to derive approximate sample size formulae for the **ALC**$_2$ and the **ALC**.

Approximate sample size formulae for the ALC$_2$ and ALC: Let

$$l_R = l_R(m_1, gm_1, a'', b'', c'', d'') = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(R)},$$

be the first order approximation of the length of an HPD or equal-tailed interval, where $\mathbf{Var}(R) = \mathbf{Var}(R|\,\mathrm{T}'')$. We impose the condition $b' > 2$ to ensure the existence of all posterior variances for any $a = 0,\ldots,m_1$ and $b = 0,\ldots,gm_1$.

It is very difficult to directly derive approximate sample size formulae for the **ALC** using the exact pre-posterior predictive distribution. Fortunately, a corollary to Theorem 3.2.6 again suggests a solution to this problem.

Corollary 4.1.3. *For $a', c', d' > 0$, and $b' > \max(2, 3k/2)$,*

$$\lim_{m_1 \longrightarrow \infty} \frac{\sqrt{m_1}\Big\{E_{X_{m_1}}[l_R^k]\Big\}^{1/k}}{2z_{1-\alpha/2}} = \int_0^1 \int_0^1 \left[\frac{x(1-y)}{g} + (1-x)y\right]^{k/2} \times \frac{x^{a'+k/2-1}(1-x)^{c'-1}}{B(a',c')} \frac{y^{b'-3k/2-1}(1-y)^{d'-1}}{B(b',d')}\, dxdy. \tag{4.1.5}$$

Proof. Suppose k is even. Define $Y_{m_1} = \left(\frac{a}{m_1}, \frac{b}{gm_1}\right)$ and let $\mathcal{F}_{m_1}$ be the set of points (x, y) where the mass function of Y_{m_1} is positive. After some algebraic manipulations of the posterior variance of R, $\mathbf{Var}(R)$ (use equation (4.1.3) to derive $\mathbf{Var}(R)$), and setting $a = m_1 x$, $c = m_1(1-x)$, $b = gm_1 y$, and $d = gm_1(1-y)$ one can show that

$$\begin{aligned}\mathbf{Var}(R) &= \frac{(m_1 x + a')(gm_1 + b' + d' - 1)}{(m_1 + a' + c')^2(m_1 + a' + c' + 1)(gm_1 y + b' - 1)^2(gm_1 y + b' - 2)} \times \\ &\quad \Big\{(m_1 x + gm_1 y + a' + b' - 1)(gm_1 + b' + d' - 2)(m_1 + a' + c') - \\ &\quad (m_1 + gm_1 + a' + b' + c' + d' - 1)(m_1 x + a')(gm_1 y + b' - 2)\Big\}.\end{aligned}$$

Taking the limit as $m_1 \longrightarrow \infty$, we have

$$\lim_{m_1 \longrightarrow \infty} m_1^{k/2}\Big\{\mathbf{Var}(R)\Big\}^{k/2} = \frac{x^{k/2}}{y^{3k/2}}\left[\frac{x(1-y)}{g} + y(1-x)\right]^{k/2}.$$

The limiting function is neither continuous nor bounded due to the term, $y^{3k/2}$ in the denominator. To remove this indeterminacy, we need the decomposition

$$\begin{aligned}p_{Y_{m_1}}(x, y | m_1, gm_1, a', b', c', d') &= c_{m_1}\left\{\prod_{i=1}^{3k/2}\left(\frac{b'-i}{gm_1} + y\right)\right\} \\ &\qquad p_{Y_{m_1}}(x, y | m_1, gm_1, a', b' - 3k/2, c', d'),\end{aligned}$$

where $c_{m_1} = (gm_1)^{3k/2}\frac{\Gamma(gm_1 + b' + d' - 3k/2)}{\Gamma(gm_1 + b' + d')}\frac{\mathrm{B}(b' - 3k/2, d')}{\mathrm{B}(b', d')}$. The idea is to obtain a term that contains the factor $y^{3k/2}$ in its limit as $m_1 \longrightarrow \infty$, so that this term cancels the factor $y^{3k/2}$ in $\lim_{m_1} m_1^{k/2}\Big\{\mathbf{Var}(R)\Big\}^{k/2}$. Here this term is $\prod_{i=1}^{3k/2}\Big(\frac{b'-i}{gm_1}+y\Big)$. Define

$$h_{m_1}(x,y) = m_1^{k/2}\left\{\prod_{i=1}^{3k/2}\Big(\frac{b'-i}{gm_1}+y\Big)\right\}\{\mathbf{Var}(R)\}^{k/2},$$

where h_{m_1} is a sequence of continuous and uniformly bounded functions by construction, and

$$\lim_{m_1\longrightarrow\infty} h_{m_1}(x,y) = x^{k/2}\Big[\frac{x(1-y)}{g}+y(1-x)\Big]^{k/2}.$$

If $b' > 3k/2$, we have

$$\begin{aligned}\lim_{m_1\longrightarrow\infty}\frac{\sqrt{m_1}\Big\{\mathbf{E}_{X_{m_1}}[l_R^k]\Big\}^{1/k}}{2z_{1-\alpha/2}} &= c_{m_1}\sum_{(x,y)\in\mathcal{F}_{m_1}} m_1^{k/2}\left\{\prod_{i=1}^{3k/2}\Big(\frac{b'-i}{gm_1}+y\Big)\right\}\{\mathbf{Var}(R)\}^{k/2}\\ &\qquad p_{Y_{m_1}}(x,y|m_1,gm_1,a',b'-3k/2,c',d')\\ &= c_{m_1}\mathbf{E}[h_{m_1}(Z_{m_1})],\end{aligned}$$

where $Z_{m_1} \sim p_{Y_{m_1}}(x,y|m_1,gm_1,a',b'-3k/2,c',d')$. We have $Z_{m_1} \longrightarrow^d (p_1,p_0)$, where $p_1 \sim \mathbf{Be}(a',b'-3k/2)$ is independent of $p_0 \sim \mathbf{Be}(c',d')$.

Theorems 3.2.7 and 3.2.6 imply that

$$\begin{aligned}\lim_{m_1}\mathbf{E}[h_{m_1}(Z_{m_1})] &= \int_0^1\int_0^1 h(x,y)\frac{x^{a'-1}(1-x)^{c'-1}}{\mathbf{B}(a',c')}\,\frac{y^{b'-3k/2-1}(1-y)^{d'-1}}{\mathbf{B}(b'-3k/2,d')},\\ &= \int_0^1\int_0^1\Big[\frac{x(1-y)}{g}+x(1-x)y\Big]^{k/2}\times\\ &\qquad\frac{x^{a'+k/2-1}(1-x)^{c'-1}}{\mathbf{B}(a',c')}\,\frac{y^{b'-3k/2-1}(1-y)^{d'-1}}{\mathbf{B}(b'-3k/2,d')}\,dxdy,\end{aligned}$$

which completes the proof for k even since $\lim_{m_1} c_{m_1} = \frac{\mathbf{B}(b'-3k/2,d')}{\mathbf{B}(b',d')}$.

Similarly, for k odd, we define

$$h_{m_1}(x,y) = m_1^{k/2}\left\{\prod_{i=1}^{3k/2+1/2}\Big(\frac{b'-i}{gm_1}+y\Big)\right\}\{\mathbf{Var}(R)\}^{k/2},$$

to obtain

$$\lim_{m_1 \longrightarrow \infty} h_{m_1}(x, y) = x^{k/2} y^{1/2} \Big[\frac{x(1-y)}{g} + y(1-x)\Big]^{k/2}.$$

Next, defining $Z_{m_1} \sim p_{Y_{m_1}}(x, y|m_1, gm_1, a', b' - 3k/2 - 1/2, c', d')$ we obtain $Z_{m_1} \longrightarrow_d \dfrac{x^{a'-1}(1-x)^{c'-1}}{\mathbf{B}(a', c')} \dfrac{y^{b'-3k/2-1/2-1}(1-y)^{d'-1}}{\mathbf{B}(b' - 3k/2 - 1/2, d')}$, and the result follows.

□

The idea behind obtaining an approximate sample size is to solve the equation $\Big\{\mathbf{E}_{X_{m_1}}[l_R^k]\Big\}^{1/k} = l$ in m_1 using the limiting result in Corollary 4.1.3. Thus an approximate sample size for $N_R(g) = m_1 + m_0 = (g+1)m_1(g)$ given g, when estimating R under the $\mathbf{ALC}_k$ when $b' > \max(2, 3k/2)$, is

$$\begin{aligned} N_R(g) \quad &= \quad (g+1)\frac{4z^2_{1-\alpha/2}}{l^2}\bigg\{\int_0^1 \int_0^1 \bigg[\frac{x(1-y)}{g} + (1-x)y\bigg]^{k/2} \times \\ &\quad \frac{x^{a'+k/2-1}(1-x)^{c'-1}}{\mathbf{B}(a', c')} \frac{y^{b'-3k/2-1}(1-y)^{d'-1}}{\mathbf{B}(b', d')}\, dxdy\bigg\}^{2/k} - a' - b' - c' - d'. \end{aligned} \tag{4.1.6}$$

Remark 4.1.1. Note that the term $a' + b' + c' + d'$ which represents the extra sample size provided by the prior distribution, was adjusted for after solving for m_1 as indicated above.

When $b' - 3k/2 > 1$, the integrand in the expression (4.1.6) is well-behaved and its integral can be computed with most integral subroutines, for example, the **ISML** subroutine DTWODQ. To prevent overflow, one first evaluates the logarithm of all the terms in the integrand, adds them and exponentiates the final value. Alternatively, one could also use a Monte Carlo integration technique to compute $N_R(g)$ by generating n pairs (x_i, y_i), $i =$

$1, \cdots, n$ from two independent Beta distributions: $x_i \sim \mathbf{Be}(a' + k/2, c' - 1)$ and $y_i \sim \mathbf{Be}(b' - 3k/2, d')$ and approximate the integral in (4.1.6) as

$$\begin{aligned} N_R(g) \quad \approx \quad & (g+1)\,\frac{4z^2_{1-\alpha/2}}{l^2}\left\{\frac{\mathrm{B}(a'+k/2,\, c')\,\mathrm{B}(b'-3k/2,\, d')}{\mathrm{B}(a',\, c')\,\mathrm{B}(b',\, d')}\times\right. \\ & \left.\frac{1}{n}\sum_{i=0}^{n}\left[\frac{x_i(1-y_i)}{g}+(1-x_i)y_i\right]^{k/2}\right\}^{2/k} - a' - b' - c' - d'. \end{aligned}$$

Figure 4.1.3 was generated using the subroutine DTWODQ. The prior and pre-specified parameters were $(a', b', c', d', 1-\alpha, l) = (3.0, 3.0, 3.0, 3.0, .95, .50)$. The overall minimal sample size was $N = 590$ and the corresponding optimal g is seen to be $g_{opt} = 1.60$. The common choice, $g = 1$, corresponding to an equal number of exposed and non-exposed, leads to $N = 616$, and therefore $g = 1$ is nearly optimal.

Approximate formula for the minimal cost problem: A natural extension of the sample size estimation problem is that of minimizing cost for a given precision of estimation in case-control and cohort designs. Denote by c_1 and c_0, the unit cost, respectively, per case and per control in the case-control design, and by $r = \frac{c_1}{c_0}$, the cost ratio of cases to controls. We seek the couple $(g, m_1(g))$ that minimizes the total cost $C = c_1 m_1(g) + c_0 m_0(g) = (g+r)c_0 m_1(g)$, under the constraint that for each g the pair $(m_0(g), m_1(g))$ is the solution of the $\mathbf{ALC}_k$ problem, for instance. More general and non-traditional approaches to incorporate costs into design issues are the object of the papers by Lindley, 1997, Bernardo, 1997, Stallard, 1998, Pezeshk and Gittins, 1999, Gittins and Pezeshk, 2000a, and Gittins and Pezeshk, 2000b. These approaches are based on eliciting a loss, a gain, or utility function, for

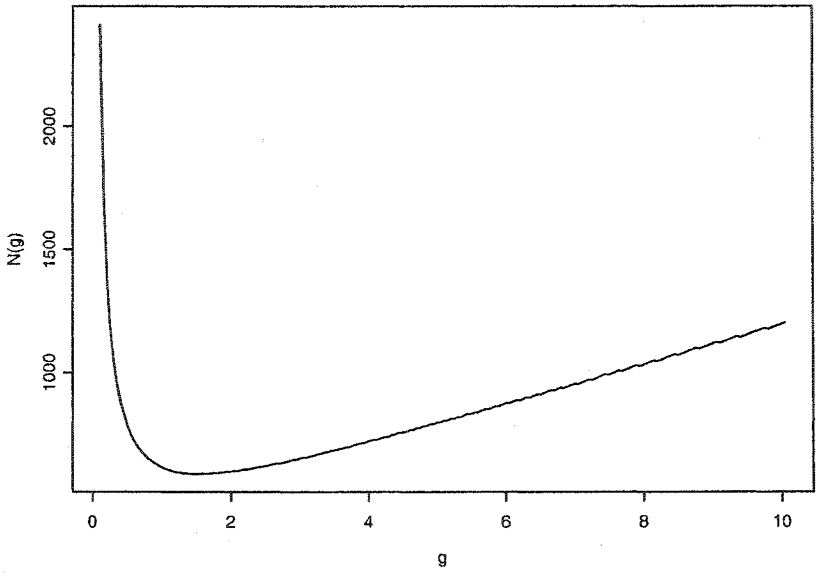


Figure 4.1.1: Graph of $N_R(g)$ against $g = .10, .15, \cdots, 10.00$ for the risk ratio with $(a', b', c', d', 1-\alpha, l) = (3.0, 3.0, 3.0, 3.0, .95, .50)$. The optimal ratio is $g_{opt} = 1.60$ with a corresponding sample size of $N_R(g_{opt}) = 590$.

example expressing financial costs of the treatment, potential profits from a marketing company, or a public health benefit. These methods are often called "fully Bayesian" or "decision theoretic" approaches.

The minimal cost $C_R(g) = c_0 m_0(g) + c_1 m_1(g) = (g+r)c_0 m_1(g)$ for R given g under the constraint that $m_1(g)$ satisfies the $\mathbf{ALC}_k$ is, from (4.1.6)

$$\begin{aligned} C_R(g) &= c_0(g+r)\frac{4z^2_{1-\alpha/2}}{l^2}\left\{\int_0^1\int_0^1\left[\frac{x(1-y)}{g}+(1-x)y\right]^{k/2}\times\right. \\ &\quad \left.\frac{x^{a'+k/2-1}(1-x)^{c'-1}}{\mathbf{B}(a',c')}\,\frac{y^{b'-3k/2-1}(1-y)^{d'-1}}{\mathbf{B}(b',d')}\,dxdy\right\}^{2/k} - \\ &\quad c_1(a'+c') - c_0(b'+d'), \end{aligned} \tag{4.1.7}$$

where $r = \dfrac{c_1}{c_0}$. Note that when $c_0 = c_1 = 1$ we have $C_R(g) = N_R(g)$. Thus the sample size problem is a subset of the cost problem. For $k = 2$ and $a', c', d' > 0$, $b' > 3$, the minimal cost is approximated by

$$\begin{aligned} C_R(g) &= c_0(r+g)\frac{4z^2_{1-\alpha/2}}{l^2} \times \\ &\left[\frac{\mathrm{B}(a'+2,c')\,\mathrm{B}(b'-3,d'+1)}{g\,\mathrm{B}(a',c')\,\mathrm{B}(b',d')} + \frac{\mathrm{B}(a'+1,c'+1)\,\mathrm{B}(b'-2,d')}{\mathrm{B}(a',c')\,\mathrm{B}(b',d')}\right] - \\ &c_1(a'+c') - c_0(b'+d'). \end{aligned} \tag{4.1.8}$$

A simple derivative of $C_R(g)$ in equation (4.1.8) with respect to g shows that the optimal ratio, g, is

$$g_R = \sqrt{r\frac{\mathrm{B}(a'+2,c')\,\mathrm{B}(b'-3,d'+1)}{\mathrm{B}(a'+1,c'+1)\,\mathrm{B}(b'-2,d')}} = \sqrt{r\frac{(a'+1)d'}{(b'-3)c'}}. \tag{4.1.9}$$

In the context of frequentist hypothesis testing, Gail et al., 1976 demonstrated that the optimal ratio g follows what he called the square root rule, that is $g = \sqrt{\eta r}$, $\eta = \dfrac{p_1 q_0}{p_0 q_1} =$ for a cohort study. Here we also have a square root rule $g_R = \sqrt{r\,\eta_R}$ when $b' > 3$ with $\eta_R = \dfrac{(a'+1)d'}{(b'-3)c'}$, although in a Bayesian credibility interval context.

Similarly, for $\log(R)$, the minimal cost $C_{\log(R)}(g) = c_0 m_0(g) + c_1 m_1(g) = (g+r)c_0 m_1(g)$ given g under the constraint that $m_1(g)$ satisfies the **ALC**$_k$

$$\begin{aligned} C_{\log(R)}(g) &= c_0(g+r)\frac{4z^2_{1-\alpha/2}}{l^2}\left\{\int_0^1\int_0^1\left[\frac{x(1-y)}{g} + (1-x)y\right]^{k/2} \times \right. \\ &\left.\frac{x^{a'-k/2-1}(1-x)^{c'-1}}{\mathrm{B}(a',c')}\,\frac{y^{b'-k/2-1}(1-y)^{d'-1}}{\mathrm{B}(b',d')}\,dxdy\right\}^{2/k} - \\ &c_1(a'+c') - c_0(b'+d'). \end{aligned} \tag{4.1.10}$$

Equation (4.1.10) which is not formally proved here, is based on the fact that $\lim\limits_{m_1 \longrightarrow \infty} m_1 \mathrm{Var}(\log(R)|\,\mathrm{T}'') = \dfrac{1}{xy}\left[\dfrac{x(1-y)}{g} + (1-x)y\right]$, when one sets $a'' =$

m_1x, $c'' = m_1(1-x)+c$, $b = gm_1y+b'$, $d'' = gm_1(1-y)+d'$ in the expression for $\mathbf{Var}(\log(R)|\,\mathrm{T}'')$ (see appendix E for the derivation of $\mathbf{Var}(\log(R)|\,\mathrm{T}''))$. For $\log(R)$ when $k = 2$, we have $g_{\log(R)} = \sqrt{r\,\eta_{\log(R)}}$ with $\eta_{\log(R)} = \dfrac{(a'-1)d'}{(b'-1)c'}$. We observed empirically that the sample size formula obtained using equation (4.1.10) when $c_0 = c_1$ is an excellent approximation of the exact $\mathbf{ALC}_k$ sample size when $a', b' > \dfrac{(k+1)}{2}$ for both HPD and equal-tailed intervals. Therefore, we encourage the use of this formula, even if costs are not a factor.

So far, we have only derived approximate sample size and minimum cost formulae for R when $b' > 3k/2$, and $\log(R)$ when $a', b' > 3k/2$, under the $\mathbf{ALC}_k$. For the other four sample size criteria, we must rely on both the "straight" and regression-based Monte Carlo approaches (fully described in chapter 3) to find the required sample sizes. Below we sketch briefly the main steps of the "straight" Monte Carlo simulations for the sample size problem for R. A similar algorithm can be derived for $\log(R)$. Since the regression-based approach for the two sample problem is no different from that of the one sample problem, it will not be discussed. We implemented all these approaches and algorithms and all others in this chapter using Visual Fortran 6.1 (Compaq) that includes most **ISML libraries**.

Sketch of the Monte Carlo simulation approach to determine sample sizes when estimating R under $\mathbf{ALC}_k$: For each step in the bisectional search over $n_1(g)$, one performs the following.

- Simulate $p_1^i \sim \mathbf{Be}(a', c')$ and $p_0^i \sim \mathbf{Be}(b', d')$, $i = 1, \cdots, m$.
- For each pair (p_1^i, p_0^i), simulate two independents observations $a_i \sim \mathbf{Bin}(n_1(g), p_1^i)$ and $b_i \sim \mathbf{Bin}((g+1)n_1(g), p_0^i)$.
- For each i, simulate $p_1^j \sim \mathbf{Be}(a' + a_i, n_1(g) - a_i + c')$ and $p_0^j \sim \mathbf{Be}(b_i + b', (g+1)n_1(g) - d_i + d')$ and set $R_j = \dfrac{p_1^j}{p_0^j}$, $j = 1, \cdots, M$. Use the observations R_j to estimate the length l_i of the `HPD` interval given $1-\alpha$, using Algorithm 4 in subsection 2.7.2.3. (below, we describe methods for evaluating $p_R(R_j|\,\mathrm{T}'')$, an important step in Algorithm 4).
- Compute $\mathrm{alc}_k(n_1(g), a', b', c', d') \approx \left(\dfrac{1}{m}\displaystyle\sum_{i=1}^{m} l_i^k\right)^{1/k}$.

One important step in the above straight Monte Carlo approach is the evaluation of the posterior density of R at the simulated R_j's. We now present various ways to compute $p_R(R|\,\mathrm{T}'')$. Similar ideas may be applied for $p_{\log(R)}(R|\,\mathrm{T}'')$.

Computation of the approximate the posterior densities of R and $\log(R)$: We have three options to evaluate $p_R(R|\,\mathrm{T}'')$ given by equation (4.1.1). The first option is based on using the univariate integral subroutine DQDAG (Piessens et al., 1983) from the **ISML library** with the integrand

$$g^\star(R, z) = \begin{cases} z^{a''+b''-1}(1-z)^{d''-1}\,\dfrac{R^{a''}}{K}\Big[1 - R\cdot z\Big]^{c''-1}, & 0 \le R < 1 \\ z^{a''+b''-1}(1-z)^{c''-1}\,\dfrac{R^{-b''}}{K}\left[1 - \dfrac{1}{R}\cdot z\right]^{d''-1}, & R \ge 1, \end{cases}$$

where $K = \mathbf{B}(a'', c'') \times \mathbf{B}(b'', d'')$, rather than the natural integrand

$$g^{\star\star}(R, z) = \begin{cases} z^{a''+b''-1}(1-z)^{d''-1}\,\dfrac{R^{a''-1}}{K}\Big[1 - R\cdot z\Big]^{c''-1}, & 0 \le R < 1 \\ z^{a''+b''-1}(1-z)^{c''-1}\,\dfrac{R^{-b''+1}}{K}\left[1 - \dfrac{1}{R}\cdot z\right]^{d''-1}, & R \ge 1. \end{cases}$$

The advantage of using $g^\star$ is that

$$\begin{aligned}\lim_{R\to 0}\int_0^1 g^\star(R,z)\,dz &= \lim_{R\to 0} p_{\log(R)}(\log(R)|\,\mathrm{T}'') = 0\\ \lim_{R\to\infty}\int_0^1 g^\star(R,z)\,dz &= \lim_{R\to\infty} p_{\log(R)}(\log(R)|\,\mathrm{T}'') = 0,\end{aligned}$$

which together with the unimodality of $p_{\log(R)}(R|\,\mathrm{T}'')$ ensures that $\int_0^1 g^\star(R,z)\,dz$ is uniformly bounded. This would not be the case for $g^{\star\star}$ when $a < 1$ as $R \longrightarrow 0$.

Note that one can recover $p_R(R|\,\mathrm{T}'')$ when the integrand $g^\star$ is used by using the relationship $p_R(R|\,\mathrm{T}'') = \dfrac{1}{R}\displaystyle\int_0^1 g^\star(R,z)\,dz$.

The second option relies on the use of Monte Carlo integration methods. Let $Y \sim \mathbf{Be}(b'', d'')$ and $Z \sim \mathbf{Be}(a'', c'')$. Indeed, when $R \le 1$,

$$p_R(R|\,\mathrm{T}'') \propto \mathbf{E}_Y\left[(RY)^{a''-1}(1-RY)^{c''-1}\right] \approx \sum_{j=1}^{m}(RY_i)^{a''-1}(1-RY_i)^{c''-1},$$

where Y_i, $i = 1,\cdots,m$ are m random observations. When $R > 1$,

$$p_R(R|\,\mathrm{T}'') \propto \mathbf{E}_Z\left[\left(\frac{Z}{R}\right)^{b''+1}\left(1-\frac{Z}{R}\right)\right] \approx \sum_{j=1}^{m}\left(\frac{Z_i}{R}\right)^{b''+1}\left(1-\frac{Z_i}{R}\right)^{d''-1},$$

where Z_i, $i = 1,\cdots,m$ are m random observations.

The third option is based on Horner's algorithm (**Press et al., 1990**) to compute the sums in equation (4.1.11) when c' and d' are integers,

$$p_R(R\,|\,\mathrm{T}'') = \begin{cases} \dfrac{R^{a''-1}}{K}\displaystyle\sum_{j=0}^{c''-1}\binom{c''-1}{j}\mathbf{B}(a''+b''+j, c''+d''-1-j)\,(1-R)^j, \\ \hfill 0 < R \le 1, \\ \dfrac{R^{-(b''+1)}}{K}\displaystyle\sum_{j=0}^{d''-1}\binom{d''-1}{j}\mathbf{B}(a''+b''+j, c''+d''-1-j)\left(1-\dfrac{1}{R}\right)^j, \\ \hfill R > 1. \end{cases} \tag{4.1.11}$$

Equation (4.1.11) is easily derived by expanding the polynomial $(1-Rz)^{c''-1} = \left[1 - z + z(1-R)\right]^{c''-1}$ and $\left(1-\frac{z}{R}\right)^{d''-1} = \left[1-z+z\left(1-\frac{1}{R}\right)\right]^{d''-1}$ in equation (4.1.1). A similar expansion for $\frac{R}{R+1}$ was given by Aitchison and Bacon-Shone, 1981. The implementation of Horner algorithm uses integer inputs for a', b', c', d' although it is sufficient that b', d' be integers. In practice, the use of Horner algorithm for our sample size calculations is by far the fastest of the methods. For this reason, we encourage researchers to round their prior parameters to the closest integers to take advantage of Horner's algorithm to determine the required sample size, as long as rounding results in minor changes to the prior moments of R. This is clearly the case for large input parameters a', b', c', d' where rounding results in minor loss of accuracy in the required sample size.

4.1.1.2 Case when p_1 and p_0 are restricted

The discussion so far has been confined to the scenario where p_1 and p_0 are unrestricted. We now retrace our steps and apply them to the restricted case. When it is known a priori that $p_1 > p_0$ or $p_1 < p_0$, two Bayesian prior distributions have been suggested in the literature. The first uses a family of conjugate priors, the bivariate incomplete Beta distribution (Smith, 1975; Tsutakawa and Lin, 1986)

$$f_{(p_1,p_0)} \propto p_1^{a'-1}(1-p_1)^{c'-1}p_0^{b'-1}(1-p_0)^{d'-1} \qquad (4.1.12)$$

where either $0 < p_1 < p_0 < 1$ or $0 < p_0 < p_1 < 1$. The prior-likelihood combination yields the following posterior distributions for R:

$$p_R(R\,|\,\mathrm{T}'') \propto R^{a''-1} \int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1}\,[1-R\cdot Z]^{c''-1}\,dZ, \qquad 0 < R < 1, \tag{4.1.13}$$

when $0 < p_1 < p_0 < 1$, and

$$p_R(R\,|\,\mathrm{T}'') \propto R^{-(b''+1)} \int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1}\left[1-\frac{1}{R}\cdot Z\right]^{d''-1} dZ, \qquad R > 1, \tag{4.1.14}$$

when $0 < p_0 < p_1 < 1$.

The second family of prior distributions is the Generalized Beta density family used by Franck et al., 1988 in the context of a randomized trial. Here one assumes the hierarchical model $p_0 \sim \mathbf{Be}(b', d')$ and $p_1|\,p_0 \sim \mathbf{Be}(a', c', 0, p_0)$ for $0 < p_1 < p_0 < 1$, or $p_1 \sim \mathbf{Be}(a', c')$ and $p_0|\,p_1 \sim \mathbf{Be}(b', d', 0, p_1)$ for $0 < p_0 < p_1 < 1$. The combination of these prior-likelihood models yields the following posterior distributions:

$$p_R(r|\,\mathrm{T}'') \propto r^{a''-1}(1-r)^{c'-1} \int_0^1 p_0^{a+b''-1}(1-p_0)^{d''-1}(1-rp_0)^{c}\,dp_0, \quad 0 < r < 1, \tag{4.1.15}$$

when $R < 1$ and

$$p_{1/R}(r|\,\mathrm{T}'') \propto r^{b''-1}(1-r)^{d'-1} \int_0^1 p_1^{a''+b-1}(1-p_1)^{c''-1}(1-rp_1)^{d}\,dp_1, \quad 0 < r < 1, \tag{4.1.16}$$

when $R > 1$.

Unimodality:

Proposition 4.1.4. *The distributions given by equation 4.1.15 and 4.1.16 are unimodal when $c', d'' \geq 1$ and $d', c'' \geq 1$, respectively.*

Proof. We first derive the posterior distribution of $\log(R)$ resulting from equation 4.1.15. We have

$$p_{\log(R)}(r|\,\mathrm{T}'') \propto e^{a''r}(1-e^r)^{c'-1}\int_0^1 p_0^{a+b''-1}(1-p_0)^{d''-1}(1-e^r p_0)^c\,dp_0, \quad -\infty < r < 0.$$

We recognize that the posterior distribution of $\log(R)$ can be written as the product of two strongly unimodal distribution when $c' \geq 1$, and, therefore, is strongly unimodal. Indeed,

$$p_{\log(R)}(r|\,\mathrm{T}'') \propto p_1(r) \times p_2(r),$$

where

$$\begin{aligned} p_1(r) &\propto e^{(a'-\varepsilon)r}(1-e^r)^{c'-1}, \quad -\infty < r < 0, \quad c' \geq 1, \quad 0 < \varepsilon < \min(a', b'') \\ p_2(r) &\propto e^{(a+\varepsilon)r}\int_0^1 p_0^{a+b''-1}(1-p_0)^{d''-1}(1-e^r p_0)^c\,dp_0, \quad -\infty < r < 0, \quad d'' \geq 1. \end{aligned}$$

The strong unimodality of $p_1(r)$ follows from Lemma 4.1.1, while the strong unimodality of $p_2(r)$ is a result of Theorem 4.1.2. Since $p_2(r)$ is proportional to the expression for the posterior distribution of $\dfrac{p_1}{p_0}$ where $p_1 \sim \mathbf{Be}(a + \varepsilon,\, c+1)$ and $p_0 \sim \mathbf{Be}(b'' - \varepsilon,\, d'')$, $0 < \varepsilon < \min(a', b'')$ when $0 < r < 1$, and the result follows.

Again, a straight application of Lemma 3.3.1 confirms that the random variable R is unimodal. Similarly for equation 4.1.16. □

We now present various algorithms to simulate observations from the distributions in equations (4.1.13), (4.1.14), (4.1.15), and (4.1.16). These have not previously appeared in the literature. We saw in subsection 4.1.1.1 that the simulation of observations from the posterior distribution is an important step in the Monte Carlo-based sample size approach to the determination of sample size.

Generating observations from the distributions in equations (4.1.13), (4.1.14), (4.1.15) **and** (4.1.16)**:** There are four cases to consider depending on the prior information that is used.

Case 1: It is easily seen from equation (4.1.15) that

$$\begin{aligned} p_R(r|\,\mathrm{T}'') &\propto r^{a''-1}(1-r)^{c'-1}\int_0^1 p_0^{a+b''-1}(1-p_0)^{d''-1}(1-rp_0)^c\,dp_0, \quad 0<r<1, \\ &= r^{a''-1}(1-r)^{c'-1}\int_0^1 p_0^{a+b''-1}(1-p_0)^{d''-1}\sum_{j=0}^{c}\frac{(1-p_0)^{c-j}\,p_0^j\,(1-r)^j}{(c+1)\mathbf{B}(j+1,c+1-j)}\,dp_0, \\ &= \sum_{j=0}^{c}\frac{\mathbf{B}(a+b''+j,d''+c-j)}{(c+1)\mathbf{B}(j+1,c+1-j)}\,r^{a''-1}\,(1-r)^{c'+j-1}. \end{aligned}$$

Thus the posterior density of R can be written as a finite mixture of $c+1$ independent Beta distributions:

$$p_R(r|\,\mathrm{T}'') = \sum_{j=0}^{c} w_j\,\frac{r^{a''-1}\,(1-r)^{c'+j-1}}{\mathbf{Be}(a'',c'+j)}, \tag{4.1.17}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{c} v_k}$ and $v_j = \dfrac{\mathbf{B}(a+b''+j,d''+c-j)\,\mathbf{B}(a'',c'+j)}{\mathbf{B}(j+1,c+1-j)}$.

Case 2: The case $R>1$ in equation (4.1.16) is dealt with by considering $\dfrac{1}{R}$. We have

$$p_{1/R}(r|\,\mathrm{T}'') \propto r^{b''-1}(1-r)^{d'-1}\int_0^1 p_1^{a''+b-1}(1-p_1)^{c''-1}(1-rp_1)^d\,dp_1, \quad 0<r<1.$$

Thus the posterior density of $\dfrac{1}{R}$ can be written as a finite mixture of independent Beta distribution:

$$p_{1/R}(r|\,\mathrm{T}'') = \sum_{j=0}^{d} w_j \, \frac{r^{b''-1}\,(1-r)^{d'+j-1}}{\mathrm{Be}(b'', d'+j)}, \tag{4.1.18}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{d} v_k}$ and $v_j = \dfrac{\mathrm{B}(a''+b+j, c''+d-j)\,\mathrm{B}(b'', d'+j)}{\mathrm{B}(j+1, d+1-j)}$.

Case 3: For the posterior distribution in equation (4.1.13), $p_R(r\,|\,\mathrm{T}'') \propto r^{a''-1} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{d''-1}\Big[1 - r\cdot Z\Big]^{c''-1}\,dZ,\quad R < 1$, when c' is an integer, we have the decomposition:

$$p_R(r\,|\,\mathrm{T}'') = \sum_{j=0}^{c''-1} w_j \, \frac{r^{a''-1}\,(1-r)^{j}}{\mathrm{Be}(a'', j+1)} \tag{4.1.19}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{x_n} v_k}$ and $v_j = \dfrac{\mathrm{B}(a''+b''+j,\, c''+d''-j-1)\,\mathrm{B}(a'',\, j+1)}{\mathrm{B}(j+1,\, c''-j)}$.
When c' is not an integer, one technique is to generate a pair of observations $(p_0^j,\, \psi_j)$ according to equation (4.1.41), where one first switches b'' and c''. Finally, set $R_j = \dfrac{\psi_j}{1 - p_0^j + p_0^j\psi_j}$.

Case 4: The case $R > 1$ in equation (4.1.14) is dealt with by considering $\dfrac{1}{R}$. We have $p_{1/R}(r\,|\,\mathrm{T}'') \propto r^{b''-1} \displaystyle\int_0^1 Z^{a''+b''-1}(1-Z)^{c''-1}\,[1 - r\cdot Z]^{d''-1}\,dZ$, $R > 1$. Thus, we have the decomposition when d' is an integer:

$$p_{1/R}(r\,|\,\mathrm{T}'') = \sum_{j=0}^{d''-1} w_j \, \frac{r^{b''-1}\,(1-r)^{j}}{\mathrm{Be}(b'', j+1)} \tag{4.1.20}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{x_n} v_k}$ and $v_j = \dfrac{\mathrm{B}(a''+b''+j,\, c''+d''-j-1)\,\mathrm{B}(b'',\, j+1)}{\mathrm{B}(j+1,\, d''-j)}$.
When d' is not an integer, one technique is again to generate a pair $\left(p_1^j,\, \dfrac{1}{\psi_j}\right)$

according to equation (4.1.42), where one first switches b'' and c''. Finally, set $R_j = p_1^j + (1 - p_1^j)\psi_j$.

Sample size calculations for the case when p_1 and p_0 are restricted must be entirely Monte Carlo simulations-based. Below, we sketch very quickly the main steps of the straight Monte Carlo approach when using the prior model $p_1 \sim \mathbf{Be}(a', c')$ and $p_0 | p_1 \sim \mathbf{Be}(b', d', 0, p_1)$. The three techniques used to evaluate the posterior distribution of R when p_1 and p_0 are unrestricted carry over to the distributions in equations (4.1.13), (4.1.14), (4.1.15) and (4.1.16). Unlike the unrestricted case, the Horner algorithm always applies, since the constant c, in the term $(1 - p_0 R)^c$, is always an integer.

Sketch of the Monte Carlo simulation approach to determine sample size when estimating $R > 1$ under MLOC: The following algorithm applies when the model in equation is adopted. For each step in the bisectional search over $n_1(g)$, one performs the following.

- Simulate $p_1^i \sim \mathbf{Be}(a',\, c')$ and $p_0^i \sim \mathbf{Be}(b',\, d'),\ i = 1, \cdots, m$.

- For each pair $(p_1^i,\, p_0^i)$, simulate two independent observations $a_i \sim \mathbf{Bin}(n_1(g),\, p_1^i)$ and $b_i \sim \mathbf{Bin}((g+1)n_1(g),\, p_0^i\, p_1^i)$.

- For each i, simulate $\dfrac{1}{R_j}$, $j = 1, \cdots, M$ from equation (4.1.18) with $a = a_i$, $c = n_1(g) - a_i$, $b = b_i$, and $d = (g+1)n_1(g) - b_i$. Use the observations R_j to estimate the length l_i of the HPD interval given $1 - \alpha$ using Algorithm 4 in subsection 2.7.2.3.

- Compute $\text{mloc}(n_1(g), a', b', c', d') \approx \text{med}_{1\leq i \leq n}\, l_i$.

Similar Monte Carlo algorithms can be constructed for all other criteria.

To close this subsection on cohort studies, we would like to mention that all the methods and results obtained apply directly to any design where the intention is to compare incidence proportions from two independent groups, including randomized controlled trials. In such studies, the parameters $R-1$, $1-R$, and $1-\dfrac{1}{R}$ are referred to as the risk attributable to the exposure, vaccine efficacy (preventive fraction), and the attributable proportion irrespectively, depending on the context. We explain quickly how sample size computations for these functions of R can be obtained. Indeed, Theorem 3.2.1 guarantees that the optimal sample sizes for estimating $0 < R < 1$ and $0 < 1 - R < 1$ are equal when $0 < p_1 < p_0 < 1$. Similarly, a straightforward application of Theorem 3.2.1 implies that the optimal sample size computations for AP_e is equal to that of $0 < \dfrac{1}{R} < 1$ when $0 < p_0 < p_1 < 1$. We derive the optimal sample size for estimating $\dfrac{1}{R}$ by permuting the prior parameters between exposures and disease. In other words, if (a', b', c', d') are the prior parameters, one should input the prior parameters (c', d', a', b') in the program computing the optimal sample size for R.

4.1.1.3 An example

The following is an adaptation of Jolson et al., 1992. Bone marrow ablation is one of the few effective treatments for some type of cancers. The drug cytarabine, a potent bone marrow suppressant given by injection for chemotherapy, is available in two forms: generic and innovator. Although the generic drug is good and less expensive, it was believed to be associated with a higher risk of cerebellar toxicity than the innovator drug. This drawback led to a momentary interruption in its use. To study the relation between cytarabine and cerebellar toxicity, a retrospective cohort study could be undertaken. Two groups of patients would be selected: those that had taken the generic drug (generic+) and the those that had taken the original drug (generic-). Let m_0 and m_1 proposed respective sample sizes of these two groups. Each patient would be classified as having experienced some cerebellar toxicities (toxicity+) or not (toxicity-). Prior data collected from a pilot study on cytarabine and cerebellar toxicity consisting of $m_1' = 25$ generic drug user and $m_0' = 34$ non-generic drug user are given in Table 4.1. Suppose, reasonably, that we are to use the relative risk to determine whether the generic drug is associated with an increase in cerebellar toxicity. More informatively we would anticipate computing a credible interval for the relative risk. We are thus led to the design problem:

Table 4.1: **Table of prior information.**

	generic+	generic-
toxicity+	11	3
toxicity-	14	31
Total	25	34

What sample size is required to estimate R with a posterior credible interval of specified length when considering the ALC and ACC?

Table 4.2: **Sample size calculations for the ALC and ACC when the unrestricted model is used.** $1 - \alpha = 0.95$, $l = 5.0$, **and** $m = M = 2000$.

g	ALC		ACC
	equation (4.1.6)	Monte Carlo	Monte Carlo
1.0	$N = 1608$ $n_1 = 806$ $n_0 = 797$	$\overline{N} = 1412$ $\sigma_{\overline{N}} = 16.242$ $\overline{n}_1 = 706$ $\overline{n}_0 = 706$	$\overline{N} = 2882$ $\sigma_{\overline{N}} = 32.610$ $\overline{n}_1 = 1441$ $\overline{n}_0 = 1441$
2.0	$N = 1258$ $n_1 = 414$ $n_0 = 844$	$\overline{N} = 1143$ $\sigma_{\overline{N}} = 18.225$ $\overline{n}_1 = 381$ $\overline{n}_0 = 762$	$\overline{N} = 2284$ $\sigma_{\overline{N}} = 27.356$ $\overline{n}_1 = 761$ $\overline{n}_0 = 1523$
3.0	$N = 1177$ $n_1 = 284$ $n_0 = 893$	$\overline{N} = 1116$ $\sigma_{\overline{N}} = 16.049$ $\overline{n}_1 = 279$ $\overline{n}_0 = 837$	$\overline{N} = 2146$ $\sigma_{\overline{N}} = 38.317$ $\overline{n}_1 = 537$ $\overline{n}_0 = 1609$
4.0	$N = 1156$ $n_1 = 218$ $n_0 = 938$	$\overline{N} = 1060$ $\sigma_{\overline{N}} = 18.415$ $\overline{n}_1 = 212$ $\overline{n}_0 = 848$	$\overline{N} = 2020$ $\sigma_{\overline{N}} = 27.250$ $\overline{n}_1 = 404$ $\overline{n}_0 = 1616$
5.0	$N = 1165$ $n_1 = 179$ $n_0 = 986$	$\overline{N} = 1054$ $\sigma_{\overline{N}} = 13.920$ $\overline{n}_1 = 176$ $\overline{n}_0 = 878$	$\overline{N} = 2023$ $\sigma_{\overline{N}} = 27.556$ $\overline{n}_1 = 337$ $\overline{n}_0 = 1686$

Given a length and prior information based on the above information, the sample size formulae in equation (4.1.6) given in this thesis can be used to find the sample size. Column 2 of Table 4.2 recaps the values obtained when $1 - \alpha = 0.95$ and $l = 5.0$ for $g = 1.0$, 2.0, 3.0, 4.0, and 5.0. The **ALC** optimal ratio of non-exposed to exposed is $g = 4.0$. An alternative approach to equation (4.1.6) is described in column 3 of Table 4.2. This Monte Carlo-based approach suggests that equation (4.1.6) consistently overestimates the true sample size. For the **ACC**, we also resort to the Monte Carlo approach. Column 3 of Table 4.2 gives a good summary of 10 estimated Monte Carlo sample sizes. $\overline{N}$ represents the average of the 10 estimated Monte Carlo sample sizes and $\sigma_{\overline{N}}$ the standard error associated with $\overline{N}$.

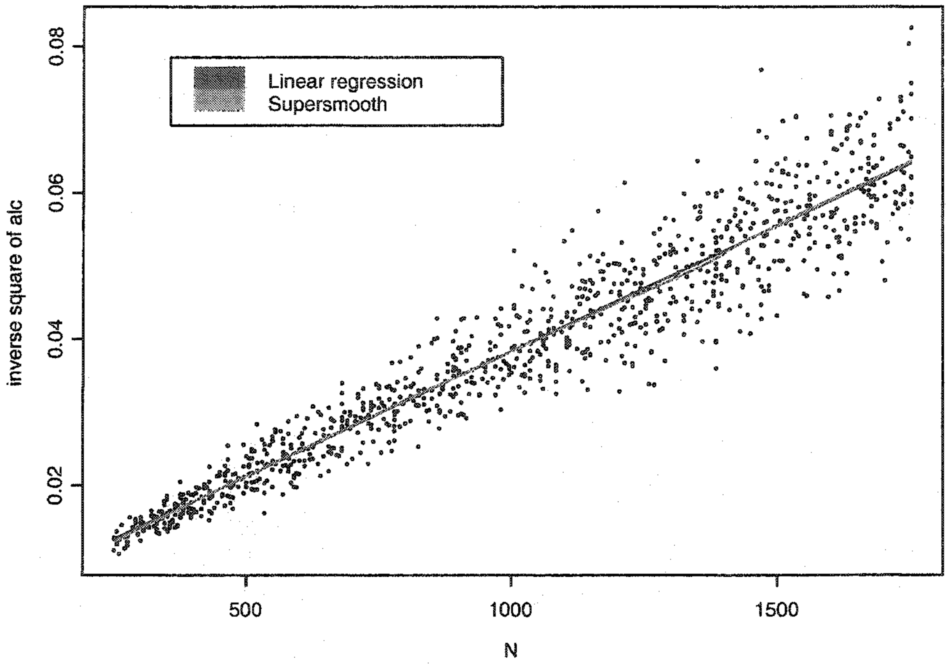


Figure 4.1.2: Graphs of the Monte Carlo pairs $\left(N, \dfrac{1}{\widetilde{\text{alc}}^2(N, 11, 3, 14, 31, .95)}\right)$ when $g = 4$.

We next discuss how a regression-based Monte Carlo approach can be used to improve our estimate of the **ALC** sample size. The equation used in the regression-based Monte Carlo approach is $\dfrac{1}{\widetilde{\text{alc}}^2(N, 11, 3, 14, 31, 1-\alpha)} = e_1 + e_2 N$, where $\widetilde{\text{alc}}(N, 11, 3, 14, 31, 1-\alpha)$ is a Monte Carlo estimate of the criterion function for the **ALC** with the given parameter specifications. The regression-based Monte approach proceeds as follows.

- First, choose J random integers, N_j, $j = 1, \cdots, J$, in a well-defined interval and compute $\widetilde{\text{alc}}(N_j, 11, 3, 14, 31, 1-\alpha)$ for each $1 \leq j \leq J$.

- Next, estimate e_1 and e_2 using a least squares procedure, and solve the equation $\hat{e}_1 + \hat{e}_2 N = \frac{1}{l^2}$ in N to determine the required sample size.

Using $J = 1000$ and $m = M = 500$, the estimated regression-based Monte Carlo sample size are $N = 1412, 1119, 1068, 1050$, and 1056 for $g =$ 1, 2, 3, 4, and 5, respectively. These adjusted estimates suggest that the previous Monte Carlo approach is very stable.

4.1.2 Case-control studies

We again use the notation $\mathtt{T} = \left(\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d}\right)$ first introduced in section 3.3 to represent Table 2.1. Case-control studies are mainly concerned with the odds ratio $\psi_e = \dfrac{p_1'(1-p_0')}{p_0'(1-p_1')}$. Let $\mathtt{T} = \left(\mathsf{a}, \mathsf{n}_1 - \mathsf{a}, \mathsf{c}, \mathsf{n}_0 - \mathsf{c}\right)$ denote results from a case-control study. Most Bayesian analyses of case-control studies assume an independent sampling both between and within disease and non-disease subjects, so that $p_1' = \Pr(E = 1|D = 1)$ and $p_0' = \Pr(E = 1|D = 0)$ are independent. We shall further assume that

$$\begin{array}{lll} a \sim \mathbf{Bin}(n_1, p_1') & \text{given } D = 1, & p_1' \sim \mathbf{Be}(a', b'), \text{ and} \\ c \sim \mathbf{Bin}(n_0, p_0') & \text{given } D = 0, & p_0' \sim \mathbf{Be}(c', d'). \end{array}$$

Let $\mathtt{T}' = \left(\mathsf{a}', \mathsf{b}', \mathsf{c}', \mathsf{d}'\right)$ denote the table of prior parameters. We also assume that there is a constant, g, possibly unknown such that $n_0 = gn_1$. The constant g is an integral part of the design, and is selected either optimally, as will be discussed in the sequel, or by more practical considerations. The combination of the prior-likelihood Tables, $\mathtt{T}'$ and $\mathtt{T}$, leads to a posterior

Table, $\mathrm{T}'' = \left(\mathrm{a}'', \mathrm{b}'', \mathrm{c}'', \mathrm{d}''\right)$ where $a'' = a+a'$, $b'' = n_1 - a + b'$, $c'' = c + c'$, $d'' = n_0 + d' - c$. $N = n_0 + n_1$ is the total sample size. Again, two models are investigated here for the computation of sample sizes. The first model does not assume any restrictions on p_1' and p_0'. The second model assumes that $p_1' < p_0'$ or $p_1' > p_0'$, leading to $\psi_e < 1$ or $\psi_e > 1$, respectively.

We proceed in analogous fashion to our approach for cohort studies. The main steps for both the unrestricted cases are:

- State the posterior density for ψ_e and ϕ_e.
- Show the unimodality of the posterior density of ψ_e and ϕ_e.
- Using the unimodality of the posterior density, derive approximate sample size and approximate cost formulae, and discuss the optimality of g.
- Give a sketch of the Monte Carlo approach to sample size calculation.

For the restricted cases, the main steps are:

- State the model for restrictions and find the posterior density for ψ_e.
- Show the unimodality of the posterior density of ψ_e.
- Derive algorithms to simulate observations from the posterior distribution of ψ_e for the four prior-likelihood models under study.
- Using unimodality of the posterior density and the simulated observations from the posterior distribution, carry out a Monte Carlo approach to sample size calculation.

4.1.2.1 Case when p_1 and p_0 are unrestricted

Following equation (2.4.12) in chapter 2, the posterior density of the odds ratio, $\psi_e = \dfrac{p_1' q_0'}{p_0' q_1'}$, in the case-control design is

$$p_{\psi_e}(\psi \,|\, \mathrm{T}'') \;=\; \begin{cases} \dfrac{\psi^{a''-1}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,\psi)^{a''+b''}}\,dy, & 0<\psi<1, \\[3ex] \dfrac{\psi^{-(c''+1)}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{\left(1-y+\dfrac{y}{\psi}\right)^{c''+d''}}\,dy, & \psi \geq 1, \end{cases} \tag{4.1.21}$$

where $C = \mathbf{B}(a'', b'')\mathbf{B}(c'', d'')$. The posterior density of the log-odds ratio, $\phi_e = \log(\psi_e)$, is given by

$$p_{\phi_e}(\phi \,|\, \mathrm{T}'') \;=\; \begin{cases} \dfrac{e^{a''\phi}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,e^{\phi})^{a''+b''}}\,dy, & -\infty<\phi<0 \\[3ex] \dfrac{e^{-c''\phi}}{C} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,e^{-\phi})^{c''+d''}}\,dy, & \phi \geq 0. \end{cases} \tag{4.1.22}$$

We now prove the unimodality of the ψ_e and ϕ_e using a straight application of Theorem 4.1.5. Again here ψ_e cannot be a strongly unimodal since the k-th moments for ψ_e,

$$\mathbf{E}(\psi_e^k \,|\, \mathrm{T}'') = \frac{\mathbf{B}(a''+k, b''-k)\,\mathbf{B}(c''-k, d''+k)}{\mathbf{B}(a'', b'')\,\mathbf{B}(c'', d'')}, \qquad 0 \leq k \leq \min(b'', c'') \tag{4.1.23}$$

are not defined for all values of k (Marshall, 1988).

Unimodality of the posterior density of ψ_e and $\log(\psi_e)$:

Theorem 4.1.5. *Let $p_1' \sim \boldsymbol{Be}(\alpha_1, \beta_1)$ and $p_0' \sim \boldsymbol{Be}(\alpha_2, \beta_2)$ be two independent random variables with $\alpha_1, \alpha_0, \beta_1, \beta_0 > 0$. Define $\rho = \dfrac{p_1'(1-p_0')}{p_0'(1-p_1')}$. Then* *(i)* $\log(\rho)$ *is strongly unimodal (and therefore unimodal).*

(ii) ρ *is unimodal.*

Proof. (i) We have shown in subsection 3.2.3 that the random variables $\log\left(\frac{p'_1}{1-p'_1}\right)$ and $\log\left(\frac{p'_0}{1-p'_0}\right)$ are strongly unimodal. Consequently, $\log(\rho) = \log\left(\frac{p'_1}{1-p'_1}\right) - \log\left(\frac{p'_0}{1-p'_0}\right)$ is strongly unimodal as the difference of two independent strongly unimodal random variables.

(ii) Again Lemma 3.3.1 implies that $\rho = e^{\log(\rho)}$ is unimodal.

□

In similar fashion, the inverse of ψ_e, namely $\psi_1 = \frac{p'_0(1-p'_1)}{p'_1(1-p'_0)}$ is also unimodal.

The pre-posterior predictive distribution: The other distribution of interest in Bayesian sample size criteria, the pre-posterior predictive distribution can easily be shown to have mass function

$$p(a, c|n_1, n_0, a', b', c', d') = \binom{n_1}{a}\frac{\mathrm{B}(a'', b'')}{\mathrm{B}(a', b')} \times \binom{n_0}{c}\frac{\mathrm{B}(c'', d'')}{\mathrm{B}(c', d')} \qquad (4.1.24)$$

$a = 0, \cdots, n_1$ and $c = 0, \cdots, n_0$ as the distribution of two independent Beta-Binomial random variables.

Note 4.1.2. *Since* $n_0 = gn_1$, n_1 *tends to* ∞ *as* n_0 *tends to* ∞. *This together with the independence shown in equation* (4.1.4) *and Theorem 3.2.6 applied to each* $\frac{a}{n_1}, \frac{c}{n_0}$, *implies that we have that* $Y_{n_1} = \left(\frac{a}{n_1}, \frac{c}{n_0}\right) \longrightarrow^d (p'_1, p'_0)$. *This result is used implicitly in the proof of Corollary 4.1.3.*

We now derive approximate sample sizes for the $\mathbf{ALC}_k$.

Approximate sample size formulae for ψ_e under the ALC$_k$: Since the first order approximation of the lengths of an HPD or of an equal-tailed interval are the same, we denote them by

$$l_{\psi_e} = l_{\psi_e}(n_1, gn_1, a'', b'', c'', d'') = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(\psi_e)},$$

where $\mathbf{Var}(\psi_e) = \mathbf{Var}(\psi_e | \mathrm{T}'')$. We impose the condition $b', c' > 2$ to ensure the existence of all posterior variances for any $b = 0, \ldots, n_1$ and $c = 0, \ldots, gn_1$.

It is very difficult to directly derive approximate sample size formulae for the **ALC** using the exact pre-posterior distribution. Fortunately, a corollary to Theorem 3.2.6 suggests a solution to this problem.

Corollary 4.1.6. *For $a', d' > 0$ and $b', c' > \max(2, 3k/2)$,*

$$\begin{aligned}\lim_{n_1 \longrightarrow \infty} \frac{\sqrt{n_1}\left\{E_{X_{n_1}}[l^k_{\psi_e}]\right\}^{1/k}}{2z_{1-\alpha/2}} &= \left\{\int_0^1 \int_0^1 \left[\frac{x(1-x)}{g} + y(1-y)\right]^{k/2} \times \right. \\ &\quad \frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{B(a', b')} \times \\ &\quad \left. \frac{y^{c'-3k/2-1}(1-y)^{d'+k/2-1}}{B(c', d')}\, dx\, dy \right\}^{1/k} \qquad (4.1.25)\end{aligned}$$

Proof. Define $Y_{n_1} = \left(\dfrac{a}{n_1}, \dfrac{c}{gn_1}\right)$ and let $\mathcal{F}_{n_1}$ be the set of points (x, y) where the mass function of Y_{n_1} is positive. After some algebraic manipulations and setting $a = n_1 x$, $b = n_1(1-x)$, $c = gn_1 y$, and $d = gn_1(1-y)$, we have

$$\begin{aligned}\mathbf{Var}(\psi_e) &= \frac{\Big(n_1 x + a'\Big)\Big(gn_1(1-y) + d'\Big)}{\Big(n_1(1-x) + b'\Big)^2\Big(n_1(1-x) + b' - 2\Big)\Big(gn_1 y + c' - 1\Big)^2\Big(gn_1 y + c' - 2\Big)} \times \\ &\quad \Big\{\Big(n_1 x + a' + 1\Big)\Big(gn_1(1-y) + d' + 1\Big)\Big(n_1(1-x) + gn_1 y + b' + c' - 3\Big) - \\ &\quad \Big(n_1(1-x) + b' - 2\Big)\Big(gn_1 y + c' - 2\Big)\Big(n_1 x + gn_1(1-y) + a' + d' - 1\Big)\Big\}.\end{aligned}$$

It is easily seen that

$$\lim_{n_1\longrightarrow\infty} n_1^{k/2}\Big\{\mathbf{Var}(\psi)\Big\}^{k/2} = \frac{x^{k/2}(1-y)^{k/2}}{(1-x)^{3k/2}y^{3k/2}}\left[\frac{x(1-x)}{g}+y(1-y)\right]^{k/2} \quad (4.1.26)$$

The idea now is to decompose the distribution $p_{X_{n_1}}(x,y|n_1,gn_0,a',b',c',d')$ in a way that eliminates the indeterminacy due to the denominator of equation 4.1.26. For this, we decompose the gamma functions, $\Gamma(n_1x+a')$, $\Gamma\Big(n_1(1-x)+b'\Big)$, $\Gamma(gn_1y+c')$ and $\Gamma\Big(gn_1(1-y)+b'\Big)$ to take care of the indeterminacy created by the terms x, $(1-x)$, y and $(1-y)$ in the denominator of equation 4.1.26, respectively. The rest of the proof is similar to the one in Corollary 4.1.3. □

In order to obtain an approximate sample size formula, one needs to solve the approximate equation $\Big\{\mathbf{E}_{X_{m_1}}[l_R^k]\Big\}^{1/k} = l$ using the result in Corollary 4.1.6. Thus an approximate sample size $N_{\psi_e}(g)$ for the $\mathbf{ALC}_k$ given g when $a',d'>0$ and $b',c'>\max(2,3k/2)$ after adjusting for the extra sample size $a'+b'+c'+d'$ provided by our prior information, is

$$\begin{aligned} N_{\psi_e}(g) \;=\; & (g+1)\frac{4z_{1-\alpha/2}^2}{l^2}\left\{\int_0^1\int_0^1\left[\frac{x(1-x)}{g}+y(1-y)\right]^{k/2}\times\right. \\ & \left.\frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{\mathbf{B}(a',b')}\,\frac{y^{c'-3k/2-1}(1-y)^{d'+k/2-1}}{\mathbf{B}(c',d')}\,dx\,dy\right\}^{2/k} \\ & -a'-b'-c'-d'. \end{aligned} \qquad (4.1.27)$$

To evaluate the integral in equation (4.1.27), one would again either use the **ISML** subroutine, DTWODQ or a Monte Carlo integration technique. We do not discuss the subject further as the paragraph following equation (4.1.6) discusses the use of both techniques to compute the approximate optimal sample size in cohort studies, $N_R(g)$ in detail.

Figure 4.1.2 represents the computation of $N_{\psi_e}(g)$ for various values of g en route to determining the optimal control to case ratio, g_{opt}, and the overall minimal sample size, $N_{\psi_e}(g_{opt})$. The prior and pre-specified parameters were $(a', b', c', d', 1-\alpha, l) = (3.0, 3.0, 3.0, 3.0, .95, 2.0)$. We obtained the following optimal ratios, $g_{opt} = 0.975, 0.99, 1.0, 1.015, 1.025$ with a corresponding sample size of $N_{\psi_e}(g_{opt}) = 472$. The disparate values for g_{opt} are most likely an artifact of the discreteness of N_{ψ_e}.

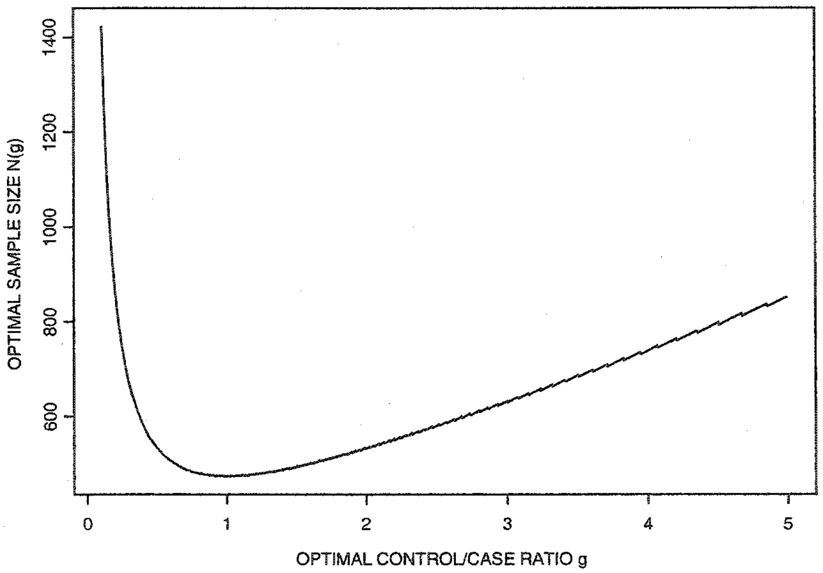


Figure 4.1.3: Graph of $N_{\psi_e}(g)$ against $g = .05, .10, \cdots, 5.00$ for the odds ratio with $(a', b', c', d', 1-\alpha, l) = (3.0, 3.0, 3.0, 3.0, .95, 2.0)$. The optimal ratios are $g_{opt} = 0.975, 0.99, 1.0, 1.015, 1.025$ with a corresponding sample size of $N_{\psi_e}(g_{opt}) = 472$.

Table 4.3 displays the required sample sizes with $l = 2.0$ and various control to case ratios, g, and various prior parameters values (a', b', c', d', g) when considering **ALC** using equation (4.1.27). The estimate $\widehat{\psi_e}$ corresponds to the prior mean of ψ_e. It appears that the prior mean of ψ_e strongly influences the required sample size. The closer $\widehat{\psi_e}$ is to 1, the smaller the sample size. For all the prior parameters chosen, we note that $g = 1$ is an optimal solution for the choice of g. This is not an isolated observation. Indeed, it is easily seen that $N_{\psi_e}(g) = N_{\psi_e}\left(\frac{1}{g}\right)$ when $b' = c'$ and $d' = a'$ or when $k = \frac{b'-d'}{2} = \frac{c'-a'}{2}$, which means that it does not matter whether exposure is beneficial or harmful, as far as sample size is concerned. Therefore, if there is a unique optimal ratio g, this unique solution is $g = 1$ because $\lim_{g \longrightarrow -\infty} N_{\psi_e}(g) = \lim_{g \longrightarrow \infty} N_{\psi_e}(g) = \infty$.

Approximate formulae for the minimal cost problem: More generally, if c_1 represents the cost per case and c_0 the cost per control, the minimal cost $C_{\psi_e}(g) = c_0 n_0(g) + c_1 n_1(g) = (g + r)c_0 n_1(g)$ under the constraint that $n_1(g)$ satisfies the **ALC**$_k$ when $a', d' > 0$ and $b', c' > \max(2, 3k/2)$. Then it follows as before, that

$$\begin{aligned} C_{\psi_e}(g) &= c_0(g+r)\frac{4z^2_{1-\alpha/2}}{l^2}\left\{\int_0^1\int_0^1\left[\frac{x(1-x)}{g}+y(1-y)\right]^{k/2}\times\right. \\ &\quad \left.\frac{x^{a'+k/2-1}(1-x)^{b'-3k/2-1}}{\mathrm{B}(a',b')}\,\frac{y^{c'-3k/2-1}(1-y)^{d'+k/2-1}}{\mathrm{B}(c',d')}\,dxdy\right\}^{2/k} - \\ &\quad c_1(a'+b') - c_0(c'+d') \end{aligned} \tag{4.1.28}$$

Table 4.3: **Table of sample sizes for estimating ψ with an interval of length 2.0 for various parameters values (a', b', c', d', g).**

alc$(3,3,3,3)$				**alc**$(3,4,5,6)$			
$\widehat{\psi}_e = 2.25$				$\widehat{\psi}_e = 1.5$			
g	n_1	n_0	N	g	n_1	n_0	N
2.00	175	356	531	2.00	57	117	174
1.72	175	324	509	1.72	61	106	167
1.54	193	303	496	1.54	64	99	163
1.00	236	236	472	1.00	79	75	154
alc$(4,3,3,4)$				**alc**$(3,4,4,3)$			
$\widehat{\psi}_e = 4.00$				$\widehat{\psi}_e = 1.00$			
g	n_1	n_0	N	g	n_1	n_0	N
2.00	627	1261	1888	2.00	22	51	71
1.72	662	1144	1806	1.72	23	45	68
1.54	692	1070	1762	1.54	25	43	68
1.00	841	841	1682	1.00	31	31	62

when $a', d' > 0$ and $b', c' > \max(2, 3k/2)$ and where $r = \frac{c_1}{c_0}$. For $k = 2$ and $a', d' > 0$, $b', c' > 3$, the minimal cost $C_{\psi_e}(g)$ is

$$\begin{aligned} C_{\psi_e}(g) \quad = \quad & c_0(r+g)\frac{4z^2_{1-\alpha/2}}{l^2}\left[\frac{\mathbf{B}(a'+2,b'-2)\,\mathbf{B}(c'-3,d'+1)}{g\,\mathbf{B}(a',b')\,\mathbf{B}(c',d')} + \right. \\ & \left. \frac{\mathbf{B}(a'+1,b'-3)\,\mathbf{B}(c'-2,d'+2)}{\mathbf{B}(a',b')\,\mathbf{B}(c',d')}\right] - c_1(a'+b') - c_0(c'+d'). \end{aligned} \tag{4.1.29}$$

Differentiating $C_{\psi_e}(g)$ in equation 4.1.29 with respect to g shows that the optimal ratio g_{ψ_e} is attained at

$$g_{\psi_e} = \sqrt{r\frac{\mathbf{B}(a'+2,b'-2)\,\mathbf{B}(c'-3,d'+1)}{\mathbf{B}(a'+1,b'-3)\,\mathbf{B}(c'-2,d'+2)}} = \sqrt{r\eta_{\psi_e}}, \tag{4.1.30}$$

when $b', c' > 3$ and where $\eta_{\psi_e} = \frac{(a'+1)(b'-3)(c'+d'-2)(c'+d'-1)}{(d'+1)(c'-3)(a'+b'-2)(a'+b'-1)}$. The optimal ratio g_{ψ_e} again satisfies the square root rule by Gail et al. (1976).

Similarly, the minimal cost $C_{\phi_e}(g)$ given g for $\phi_e = \log(\psi_e)$ under the constraint that $n_1(g)$ satisfies the $\mathbf{ALC}_k$ is

$$\begin{aligned} C_{\phi_e}(g) \quad = \quad & c_0(g+r)\frac{4z^2_{1-\alpha/2}}{l^2}\left\{\int_0^1\int_0^1\left[\frac{x(1-x)}{g} + y(1-y)\right]^{k/2} \times \right. \\ & \left. \frac{x^{a'+k/2-1}(1-x)^{b'-k/2-1}}{\mathbf{B}(a',b')}\,\frac{y^{c'-k/2-1}(1-y)^{d'+k/2-1}}{\mathbf{B}(c',d')}\,dxdy\right\}^{2/k} \\ & -c_1(a'+b') - c_0(c'+d'), \end{aligned} \tag{4.1.31}$$

when $a', d' > 0$ and $b', c' > k/2$. Equation (4.1.31) is based on an expression for $\mathbf{Var}(\phi_e|\,\mathrm{T}''))$ derived by Maritz, 1989. One finds that

$$\begin{aligned} \lim_{n_1 \longrightarrow \infty} n_1 \mathbf{Var}(\phi_e|\,\mathrm{T}'')) \quad &= \quad \frac{1}{x} + \frac{1}{1-x} + \frac{1}{y} + \frac{1}{1-y} \\ &= \quad \frac{1}{x(1-x)y(1-y)}\left[\frac{x(1-x)}{g} + y(1-y)\right] \end{aligned}$$

with $a'' = n_1x + a', b'' = n_1(1-x) + b', c'' = gn_1y + c', d'' = gn_1(1-y) + d'$. Again, the square root applies for ϕ_e when $k = 2$ and $b', c' > 1$, with

$$\eta_{\phi_e} = \frac{\mathrm{B}(a'+2, b')\,\mathrm{B}(c'-1, d'+1)}{\mathrm{B}(a'+1, b'-1)\,\mathrm{B}(c', d'+2)}. \tag{4.1.32}$$

So far, we have derived approximate sample size formulae under the $\mathbf{ALC}_k$ for ψ_e when $b', c' > 3k/2$ and for $\log(R)$ when $a', b', c', d' > 3k/2$. For the other four sample size criteria and the cases where there is no closed form expression for $\mathbf{ALC}_k$, we rely on both the straight and regression-based Monte Carlo approaches to find the required sample sizes. Below, we briefly sketch the main steps of the Monte Carlo approach for estimating ψ_e. A similar algorithm can be derived for ϕ_e.

Sketch of the Monte Carlo simulation approach to determining the optimal sample size when estimating ψ_e under $\mathbf{ALC}_k$: For each step in the bisectional search over $n_1(g)$, one performs the following.

- Simulate $p_1^i \sim \mathbf{Be}(a', b')$ and $p_0^i \sim \mathbf{Be}(c', d')$, $i = 1, \cdots, m$.

- For each pair (p_1^i, p_0^i), simulate two independent observations $a_i \sim \mathbf{Bin}(n_1(g), p_1^i)$ and $c_i \sim \mathbf{Bin}((g+1)n_1(g), p_0^i)$.

- For each i, simulate $p_1^j \sim \mathbf{Be}(a' + a_i, n_1(g) - a_i + b')$ and $p_0^j \sim \mathbf{Be}(c_i + b', (g+1)n_1(g) - c_i + d')$ and set $\psi_j = \dfrac{p_1^j(1-p_0^j)}{p_0^j(1-p_1^j)}$, $j = 1, \cdots, M$. Use the observations ψ_j to estimate the length l_i of the HPD interval given $1-\alpha$ using Algorithm 4 in subsection 2.7.2.3. (below, we describe methods for evaluating $p_{\psi_e}(\psi_j \mid \mathrm{T}'')$, an important step in Algorithm 4).

- Compute $\text{alc}_k(n_1(g), a', b', c', d') \approx \left(\frac{1}{m}\sum_{i=1}^{m} l_i^k\right)^{1/k}$.

Below we describe three techniques for the computation of the posterior distribution of ψ_e and ϕ_e, an essential step of the Monte Carlo approaches to finding sample sizes. As mentioned before, these are needed in the above algorithm.

Computation of the approximate posterior density of ψ_e and ϕ_e: We have three options to compute $p_{\phi_e}(\phi|\,\mathrm{T}'')$. The first option for the computation of $p_{\phi_e}(\phi|\,\mathrm{T}'')$ is to use the integral subroutine DQDAGS (Piessens et al., 1983) from the **ISML library** in combination with the integrand

$$g(\phi,\, y) \;=\; \begin{cases} \dfrac{e^{a''\phi}}{C}\,\dfrac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,e^{\phi})^{a''+b''}}, & d'' \geq a'' \quad \phi < 0, \\[2ex] \dfrac{e^{d''\phi}}{C}\,\dfrac{y^{b''+d''-1}(1-y)^{a''+c''-1}}{(1-y+y\,e^{\phi})^{c''+d''}}, & d'' < a'' \quad \phi < 0, \\[2ex] \dfrac{e^{-b''\phi}}{C}\,\dfrac{y^{b''+d''-1}(1-y)^{a''+c''-1}}{(1-y+y\,e^{-\phi})^{a''+b''}}, & c'' \geq b'' \quad \phi \geq 0, \\[2ex] \dfrac{e^{-c''\phi}}{C}\,\dfrac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,e^{-\phi})^{c''+d''}}, & c'' < b'' \quad \phi \geq 0, \end{cases} \tag{4.1.33}$$

where $C = \mathbf{B}(a'', b'')\mathbf{B}(c'', d'')$. This integrand was obtained using various identities on hypergeometric functions (Zhang, 1996). The reason for the above four part decomposition can be seen for instance by looking at the second row of equation (4.1.33). Indeed,

$$\lim_{\phi \longrightarrow -\infty} \frac{y^{b''+d''-1}(1-y)^{a''+c''-1}}{(1-y+y\,e^{\phi})^{c''+d''}} = y^{b''+d''-1}(1-y)^{a''-d''-1}$$

if $d'' < a''$. Therefore, y-values near 1 would inflate the integral if $d'' < a''$, increasing the computation time. Unfortunately, this technique does not deal explicitly with the case of $d'' = a''$, although in practice $g(\phi, y)$ works properly, possibly because $\lim_{R\to-\infty} p_{\phi_e}(\phi|\,\mathrm{T}'') = Re^{a''R}$. The main reason for using all four rows of $g(\phi, y)$ instead of only the first and fourth rows is computation efficiency.

The second option is based on Monte Carlo integration methods. Indeed, $p_{\phi_e}(\phi|\,\mathrm{T}'') \propto \frac{1}{m}\sum_{j=1}^{m}\left(\frac{Y_i\, e^{\phi}}{1-Y_i+Y_i\, e^{\phi}}\right)^{a''}\left(\frac{1-Y_i}{1-Y_i+Y_i\, e^{\phi}}\right)^{b''}$ when $\phi < 0$ and $p_{\phi_e}(\phi|\mathrm{T}'') \propto \frac{1}{m}\sum_{j=1}^{m}\left(\frac{Z_i\, e^{-\phi}}{1-Z_i+Z_i\, e^{-\phi}}\right)^{c''}\left(\frac{1-Z_i}{1-Z_i+Z_i\, e^{-\phi}}\right)^{d''}$ when $\phi \geq 0$, where Y_i and Z_i are m random observations from the $\mathbf{Be}(c'', d'')$ and $\mathbf{Be}(a'', b'')$, respectively.

The third option, discussed previously in chapter 2 (see equation (2.4.19)), is based on expanding the posterior density of ϕ_e into hypergeometric series when $d'' \geq a''$ as follows:

$$p_{\phi_e}(\phi|\,\mathrm{T}'') \;=\; \frac{e^{a''\phi}}{K}\sum_{j=0}^{\infty}\frac{\Gamma(a''+b''+j)}{\Gamma(a''+b'')}\,\frac{\Gamma(a''+c''+j)}{\Gamma(a''+c'')}\,\frac{\Gamma(N'')}{\Gamma(N''+j)}\,\frac{(1-e^{\phi})^j}{j!}, \qquad \phi < 0,$$

where $K = \frac{\mathbf{B}(a'', b'')\,\mathbf{B}(c'', d'')}{\mathbf{B}(a''+c'', b''+d'')}$ and $N'' = a''+b''+c''+d''$. It is easily seen that $\frac{\Gamma(a''+b''+j)}{\Gamma(j+1)} \leq (a''+b''+j-1)$ and $\frac{\Gamma(a''+c''+j)}{\Gamma(N''+j)} \leq \frac{1}{(a''+c''+j)^{d''+b''}}$. Therefore, the rate of convergence of the expansion of $p_{\phi_e}(\phi|T'')$ in series is $j^{-(d''-a''+1)}$ when $d'' \geq a''$ and $\phi < 0$. A similar rate was obtained by Hora and Kelley, 1983 for the cumulative distribution function of ψ_e.

Similar approaches can be derived for $p_{\psi_e}(\psi|\,\mathrm{T}'')$.

In our experience and for the problem of sample size, the Monte Carlo approach was the overall most efficient.

4.1.2.2 Case when p_1' and p_0' are restricted

Here we use the same prior-likelihood models as in subsection 4.1.1.2.

The first prior distribution for (p_1', p_0') when either $0 < p_1' < p_0' < 1$ or $0 < p_0' < p_1' < 1$ is

$$f_{(p_1',p_0')}(p_1', p_0' | \mathrm{T}'') \propto p_1'^{a'-1}(1-p_1')^{b'-1} p_0'^{c'-1}(1-p_0')^{d'-1}. \tag{4.1.34}$$

The resulting posterior distributions for ψ_e are

$$p_{\psi_e}(\psi \,|\, \mathrm{T}'') \propto \psi^{a''-1} \int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+y\,\psi)^{a''+b''}}\, dy, \qquad 0 < \psi < 1, \tag{4.1.35}$$

when $0 < p_1' < p_0' < 1$ and

$$p_{\psi_e}(\psi \,|\, \mathrm{T}'') \propto \psi^{-(c''+1)} \int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{\left(1-y+\dfrac{y}{\psi}\right)^{c''+d''}}\, dy, \qquad \psi > 1, \tag{4.1.36}$$

when $0 < p_0' < p_1' < 1$.

The second family of prior distributions is the family of Generalized Beta density. We assume the hierarchical model $p_0' \sim \mathbf{Be}(c', d')$ and $p_1' | p_0' \sim \mathbf{Be}(a', b', 0, p_0')$ for $0 < p_1' < p_0' < 1$ and $p_1' \sim \mathbf{Be}(a', b')$ and $p_0' | p_1' \sim \mathbf{Be}(b', d', 0, p_1')$ for $0 < p_0' < p_1' < 1$. These prior-likelihood combinations yield the posterior densities

$$p_{\psi_e}(\psi | \mathrm{T}'') \propto \psi^{a''-1}(1-\psi)^{b'-1} \int_0^1 \frac{p_0'^{a+c''-1}(1-p_0')^{b''+d''-1}}{(1-p_0'+p_0'\psi)^{a''+b''}}\, dp_0', \quad 0 < \psi < 1, \tag{4.1.37}$$

when $\psi_e < 1$ and

$$p_{1/\psi_e}(\psi|\,\mathrm{T}'') \propto \psi^{c''-1}(1-\psi)^{d'-1}\int_0^1 \frac{p_1'^{a''+c-1}(1-p_1')^{b''+d''-1}}{(1-p_1'+p_1'\psi)^{c''+d''}}\,dp_1', \quad 0<\psi<1, \tag{4.1.38}$$

when $\psi_e > 1$.

Proposition 4.1.7. *The distributions given by equations 4.1.37 and 4.1.38 are unimodal when* $b' \geq 1$, $a' < c'' + d''$ *and* $d' \geq 1$, $c' < a'' + b''$, *respectively.*

Proof. Using equation (4.1.37), the posterior density of $\log(\psi_e)$ is given by

$$p_{\log(\psi_e)}(\psi|\,\mathrm{T}'') \propto e^{a''\psi}(1-e^{\psi})^{b'-1}\int_0^1 \frac{p_0'^{a+c''-1}(1-p_0')^{b''+d''-1}}{(1-p_0'+p_0'e^{\psi})^{a''+b''}}\,dp_0', \quad -\infty<\psi<0.$$

We recognize again that the posterior density of $\log(\psi_e)$ can be written as the product of two strongly unimodal distributions, and therefore is strongly unimodal. Indeed,

$$p_{\log(\psi_e)}(\psi|\,\mathrm{T}'') \propto p_3(\psi) \times p_4(\psi),$$

where

$$\begin{aligned}
p_3(\psi) &\propto e^{(a'-\varepsilon_1)r}(1-e^{r})^{b'-1}, \qquad -\infty<r<0, \qquad b'>1,\\
p_4(\psi) &\propto \propto e^{(a+\varepsilon_1)\psi}\int_0^1 \frac{p_0'^{a+c''-1}(1-p_0')^{b''+d''-1}}{(1-p_0'+p_0'e^{\psi})^{a''+b''}}\,dp_0', \quad -\infty<r<0.
\end{aligned}$$

$p_3(r)$ is strongly unimodal as $\dfrac{\partial^2\log(p_3(r))}{\partial r^2} = -\dfrac{(b'-1)e^r}{(1-e^r)^2} \leq 0$ irrespective of ε_1. The strong unimodality of $p_4(r)$ is a result of Lemma 4.1.1, since $p_4(r)$ is proportional to the expression of posterior density of $\psi = \dfrac{p_1(1-p_0)}{p_0(1-p_1)}$ when $0 < \psi < 1$, where $p_1 \sim \mathbf{Be}(a+\varepsilon_1,\, b''+\varepsilon_2)$ and $p_0 \sim \mathbf{Be}(c''-\varepsilon_1,\, d''-\varepsilon_2)$ with $\varepsilon_1+\varepsilon_2 = a'$, $0<\varepsilon_1<c''$, $0<\varepsilon_2<d''$. Such a pair $(\varepsilon_1, \varepsilon_2)$ exists

anytime $a' < c'' + d'' = m_0 + b' + d'$. A simple graph of regions delimited by the line $\varepsilon_1 + \varepsilon_2 = a'$, $\varepsilon_1 + \varepsilon_2 < b'' + c''$, $0 < \varepsilon_1 < c''$, $0 < \varepsilon_2 < d''$ confirms the existence of such pairs.

A straightforward application of Lemma 3.3.1 confirms the corollary that the random variable ψ_e is unimodal. Similarly for equation (4.1.38). □

Although, we have not shown the unimodality of the posterior density in (4.1.37) for $a' \geq m_0 + b' + d'$, in practice, often $a' < m_0 + b' + d'$ is often the case. When the condition $a' < m_0 + b' + d'$ is not met and when unimodality is a strongly desired property, we suggest to increase or decrease the ratio of controls per case, g, until one satisfies the constraint $a' < m_0 + b' + d'$. This imposition of unimodality for HPD intervals and its consequent computational advantage (for equal-tailed intervals, unimodality is not required) comes at the expense of a possibly suboptimal choice of g. Similarly for equation (4.1.38).

We propose various algorithms for simulating observations from the posterior distributions implied by the expressions (4.1.35), (4.1.36), (4.1.37), and (4.1.38). The process of simulating observations from the posterior distributions is an important step of Monte Carlo approaches. For the restricted case, Monte Carlo approaches are the only choice to sample size calculation.

Generating observations from (4.1.35), (4.1.36), (4.1.37), **and** (4.1.38)**:**

There are four cases to consider, depending on the prior information that is used.

Case 1: To simulate an observation ψ_i from (4.1.37), we first simulate a pair of independent observations $(p_0^i,\, R_i)$ using the mixture

$$\begin{aligned} f_{p_0,R}(p_0,R|\,\mathrm{T}'') &\propto p_0^{a+c''-1}(1-p_0)^{d''-1}R^{a''-1}(1-R)^{b'-1}(1-Rp_0)^{b}, \\ &= \sum_{j=0}^{b} w_j\, \frac{p_0^{a+c''+j-1}(1-p_0)^{d''+b-j-1}}{\mathbf{B}(a+c''+j,\, d''+b-j)}\, \frac{R^{a''-1}(1-R)^{b'+j-1}}{\mathbf{B}(a'',\, b'+j)}, \\ &\qquad 0<p_0,\ R<1, \qquad (4.1.39) \end{aligned}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{b} v_k}$ and $v_j = \dfrac{\mathbf{B}(a+c''+j,\, d''+b-j)\,\mathbf{B}(a'',\, b'+j)}{\mathbf{B}(j+1,\, b+1-j)}$. Then, solving for $p_1^i = R_i p_0^i$, we are able to compute $\psi_i = \dfrac{p_1^i(1-p_0^i)}{p_0^i(1-p_1^i)} = \dfrac{(1-p_0^i)R_i}{1-p_0^i R_i}$.

Case 2: The case $\psi_e > 1$ in equation (4.1.38) is dealt with by considering $\dfrac{1}{\psi_e}$. Similar to case 1, the density implied by (4.1.38) uses the mixture

$$\begin{aligned} f_{p_1,\frac{1}{R}}(p_1,R|\,\mathrm{T}'') &= \sum_{j=0}^{d} w_j\, \frac{p_1^{a''+c+j-1}(1-p_0)^{b''+d-j-1}}{\mathbf{B}(a''+c+j,\, b''+d-j)}\, \frac{R^{c''-1}(1-R)^{d'+j-1}}{\mathbf{B}(c'',\, d'+j)}, \\ &\qquad 0<p_0,\ r<1, \qquad (4.1.40) \end{aligned}$$

with $w_j = \dfrac{v_j}{\sum_{k=0}^{d} v_k}$ and $v_j = \dfrac{\mathbf{B}(a''+c+j,\, b''+d-j)\,\mathbf{B}(c'',\, d'+j)}{\mathbf{B}(j+1,\, d+1-j)}$. Finally, for each simulated pair $\left(p_1^i,\, \dfrac{1}{R_i}\right)$, one computes $p_0^i = \dfrac{1}{R_i}p_1^i$ and $\psi_i = \dfrac{p_1^i(1-p_0^i)}{p_0^i(1-p_1^i)} = \dfrac{1-p_1^i\dfrac{1}{R_i}}{(1-p_1^i)\dfrac{1}{R_i}}$.

Case 3: For $p_{\psi_e}(\psi\,|\,\mathrm{T}'') \propto \psi^{a''-1}\displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+\psi y)^{a''+b''}}\,dy$, $0<\psi<1$, we have the infinite decomposition,

$$\begin{aligned} f_{p_0,\psi_e}(p_0,\psi\,|\,\mathrm{T}'') &\propto \psi^{a''-1}\frac{p_0^{a''+c''-1}(1-p_0)^{b''+d''-1}}{(1-p_0+\psi p_0)^{a''+b''}\,dy}, \qquad 0<p_0,\ \psi<1, \\ &= \sum_{j=0}^{\infty} w_j\, \frac{p_0^{a''+c''+j-1}(1-p_0)^{b''+d''-1}}{\mathbf{B}(a''+c''+j,\, b''+d'')}\, \frac{\psi^{a''-1}(1-\psi)^{j}}{\mathbf{B}(a'',\, j+1)} \end{aligned}$$

(4.1.41)

with $w_j = \dfrac{v_j}{\sum_{k=0}^{\infty} v_k}$ and $v_j = \dfrac{\Gamma(a''+b''+j)\,\Gamma(a''+c''+j)}{\Gamma(N''+j)\,\Gamma(a''+1+j)}$.

Case 4: The case $\psi_e > 1$ in equation (4.1.36) is again dealt with by considering $\dfrac{1}{\psi_e}$. For $p_{\psi_e}(\psi \,|\, \mathrm{T}'') \propto \psi^{-(c''+1)} \displaystyle\int_0^1 \frac{y^{a''+c''-1}(1-y)^{b''+d''-1}}{(1-y+\psi y)^{c''+d''}}\, dy, \quad \psi \geq 1$, we have the infinite decomposition

$$\begin{aligned} f_{p_1,\,\frac{1}{\psi_e}}(p_1, \psi \,|\, \mathrm{T}'') \;&\propto\; \psi^{c''-1}\, \frac{p_1^{a''+c''-1}(1-p_1)^{b''+d''-1}}{(1-p_1+\psi p_1)^{c''+d''}\, dy}, \qquad 0 < p_1,\, \psi < 1, \\ &= \sum_{j=0}^{\infty} w_j\, \frac{p_1^{a''+c''+j-1}(1-p_1)^{b''+d''-1}}{\mathrm{B}(a''+c''+j,\, b''+d'')}\, \frac{\psi^{c''-1}(1-\psi)^j}{\mathrm{B}(c'',\, j+1)} \end{aligned} \tag{4.1.42}$$

where $w_j = \dfrac{v_j}{\sum_{k=0}^{\infty} v_k}$ and $v_j = \dfrac{\Gamma(c''+d''+j)\,\Gamma(a''+c''+j)}{\Gamma(N''+j)\,\Gamma(c''+1+j)}$. These series with leading terms w_j from (4.1.41) and (4.1.42) converge at the rate of $j^{-(d''+1)}$ and $j^{-(b''+1)}$, respectively. Thus, when $d' \geq 3$ and $c' \geq 3$, a mixture using the first 2000 terms, will provide an excellent approximation.

The decompositions in equations (4.1.41) and (4.1.42) are used to simulate an observation from the posterior distribution of the relative risk, R, as discussed in subsection 4.1.1.2.

Having described how to simulate observations from various joint posterior distributions, we are ready to describe an algorithm that provides a Monte Carlo approach to sample size determination for ψ_e. Below we sketch the main steps of a Monte Carlo approach when using the prior model $p'_0 \sim \mathbf{Be}(c', d')$ and $p'_1 \sim \mathbf{Be}(a', b', 0, p'_0)$ given p'_0 first, followed by the algorithm for the model $f_{(p'_1, p'_0)}(p'_1, p'_0 \,|\, \mathrm{T}'') \propto p_1'^{a'-1}(1-p'_1)^{b'-1} p_0'^{c'-1}(1-p'_0)^{d'-1},\ 0 < p'_0 < p'_1$.

Sketch of the Monte Carlo simulation approach to determining the optimal sample size when estimating ψ under ACC: For each step in the bisectional search over $n_1(g)$ based on the prior model $p_0' \sim \mathbf{Be}(c', d')$ and $p_1' \sim \mathbf{Be}(a', b', 0, p_0')$ given p_0', one performs the following:

- Simulate $p_1^i \sim \mathbf{Be}(a',\, b')$ and $p_0^i \sim \mathbf{Be}(c',\, d')$, $i = 1, \cdots, m$.

- For each pair $(p_1^i,\, p_0^i)$, simulate two independents observations $a_i \sim \mathbf{Bin}(n_1(g),\, p_0^i\, p_1^i)$ and $c_i \sim \mathbf{Bin}((g+1)n_1(g),\, p_0^i)$.

- For each i, simulate the pairs $(R_j,\, p_0^j)$, $j = 1, \cdots, M$ from equation (4.1.39), where $a = a_i$, $b = n_1(g) - a_i$, $c = c_i$, and $d = (g+1)n_1(g) - c_i$, respectively. Use the observations $\psi_j = \dfrac{(1-p_0^j)R_j}{1 - R_j\, p_0^j}$ to estimate the coverage α_i of the HPD interval given l using Algorithm 3 in subsection 2.7.2.3.

- Compute $\mathrm{acc}(n_1(g), a', b', c', d') \approx \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \alpha_i$.

For the model $f_{(p_1',p_0')}(p_1', p_0' |\, \mathrm{T}'') \propto p_1'^{a'-1}(1-p_1')^{b'-1} p_0'^{c'-1}(1-p_0')^{d'-1}$, $0 < p_0' < p_1'$, we have the following steps:

- Simulate $\left(p_1^i,\, \dfrac{1}{\psi_e^i}\right)$, $i = 1, \cdots, m$ using (4.1.42), where $a'' = a'$, $b'' = b'$, $c'' = c'$ and $d'' = d'$.

- For each pair $(p_1^i,\, \psi_e^i)$, simulate two independent observations $a_i \sim \mathbf{Bin}((g+1)n_1(g),\, p_1^i)$ and $c_i \sim \mathbf{Bin}\left(n_1(g),\, \dfrac{p_1^i}{p_1^i + (1-p_1^i)\psi_e^i}\right)$.

- For each i, simulate $\dfrac{1}{\psi_e^j}$, $j = 1, \cdots, M$ using (4.1.42) where $a'' = a_i + a'$, $b'' = n_1(g) - a_i + b'$, $c'' = c_i + c'$ and $d'' = (g+1)n_1(g) - c_i + d'$.

Use the observations ψ_e^j to estimate the length l_i of the HPD interval given $1-\alpha$ using algorithm 4 in subsection 2.7.2.3.

- Compute $\text{alc}_k(n_1(g), a', b', c', d') \approx \left(\frac{1}{m}\sum_{i=1}^{m} l_i^k\right)^{1/k}$.

Subsection 4.1.2.3 provides an illustration of the regression-based Monte Carlo approach.

4.1.2.3 An example

Example : A case-control of thalidomide intake in regards with birth defects After observing that many women with malformed children at birth had taken thalidomide tablets during their pregnancy, it was decided to conduct a case-control study to determine whether there was an association between intake of thalidomide and the occurrence of birth defects among the children of mothers who had taken this drug. A group of n_1 children with birth defects (cases) would be compared to a group of n_0 children without birth defects (controls). It would then be ascertained for each mother whether or not she had taken thalidomide during her pregnancy. Let p_1' and p_0' represents the proportion of mothers who took thalidomide during their pregnancy among the cases and the controls, respectively. Suppose a pilot study reveals that out of 10 children with birth defects and 50 without birth defects, the number of the mothers that took thalidomide during their pregnancy is 5 and 10 respectively.

What should the sample sizes of the main study, n_1 and n_0 be in order to ensure that a 95 per cent posterior credible interval of for the odds ratio

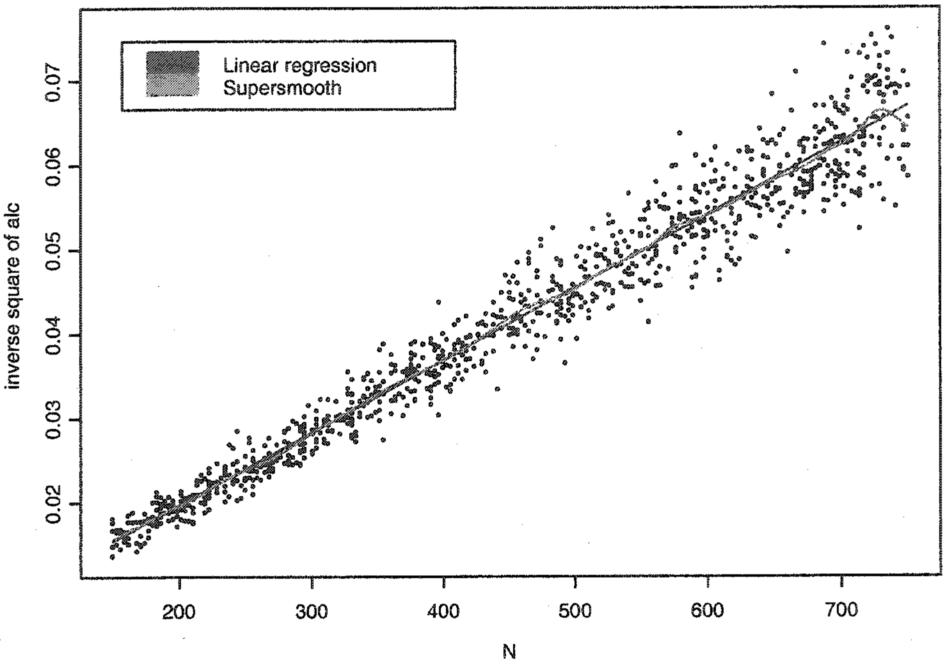


Figure 4.1.4: Graphs of the Monte Carlo pairs $\left(N, \dfrac{1}{\widetilde{\text{alc}}^2(N,5,5,10,40,.95)}\right)$.

$\psi_e = \dfrac{p_1'(1-p_0')}{(1-p_1')p_0'}$, would have a length no larger than 5.0 on average? Thus the criteria to be used here is the **ALC**. For this study, the posterior credible intervals chosen is the equal-tailed interval. It was decided to take 2 controls per case, i.e. $g = 2$. The total sample size provided by equation (4.1.27) is $N = 420$ with $n_1 = 150$ and $n_0 = 270$. The smallest admissible sample size for this setting using, again, (4.1.27), was attained at $g_{\text{opt}} = 1.25$ with $N = 393$, $n_1 = 191$, and $n_0 = 202$. Clearly $g = 2.0$ is nearly optimal. With the Monte Carlo approach based on taking $m = M = 5000$, we obtain $\overline{N} = 437$ with $\overline{n}_1 = 146$ and $\overline{n}_0 = 291$ using the average of 10 estimated sample sizes, with a standard error for $\overline{N}$ of 2.766. This suggests that the sample size provided by formulae (4.1.27) underestimates the true sample size. This behavior was confirmed by a regression-based Monte Carlo approach which resulted in a sample size of 436 using on $m = M = 1000$ and 1000 simulated values for the criterion function. The data are plotted in figure 4.1.3.

If it were known a priori that thalidomide intake was associated with an increase in birth defects, the posterior density in equation (4.1.36) could be used for the analysis. In this situation, we resort to the Monte Carlo approach to determine the required sample size. The corresponding sample size when $g = 2.0$ is $\overline{N} = 87$ with $\overline{n}_1 = 29$ and $\overline{n}_0 = 58$, with a standard error for $\overline{N} = 0.458$. This example illustrates that using the prior information that $\psi_e > 1$ can result in tremendous savings (about 80% here).

4.1.3 Cross-sectional studies

As discussed in chapter 2, the prior-likelihood model of interest for cross-sectional studies is the Dirichlet-Multinomial. We again use the notation $\mathtt{T} = \left(\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\right)$, first introduced in section 3.3 to represent the data in Table 2.1. Let $\mathtt{T} = \left(\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\right)$ again denote a realization of the multinomial distribution and let $\mathtt{T}' = \left(\mathtt{a}', \mathtt{b}', \mathtt{c}', \mathtt{d}'\right)$ denote the Dirichlet prior distribution parameters, respectively. The combination of these two prior-likelihood Tables, $\mathtt{T}'$ and $\mathtt{T}$, leads to a posterior table $\mathtt{T}'' = \left(\mathtt{a}'', \mathtt{b}'', \mathtt{c}'', \mathtt{d}''\right)$, where $a'' = a + a'$, $b'' = b + b'$, $c'' = c + c$, and $d'' = d + d'$. We consider the problem of determining the required sample size $N = a+b+c+d$ for estimating the risk ratio $R = \dfrac{p_{11}(p_{10} + p_{00})}{p_{10}(p_{11} + p_{01})}$ and the odds ratio $\psi = \dfrac{p_{11}p_{00}}{p_{10}p_{01}}$ under the criteria

ALC$_k$, ACC$_k$, WOC, MLOC and MCOC.

It was shown in subsection 2.4.2 that the posterior distributions of R and ψ in the cross-sectional setting are the same as the posterior distributions of R in the cohort design and the posterior distribution of ψ_e in the case-control designs, respectively. A similar remark applies to the parameters $\log(R)$ and $\log(\psi_e)$. Using these observations, we avoid repetitions in the following section.

In the sequel, the following notation is used for convenience.

Notation 4.1.1. *Let*

$$\binom{N}{a_1\, a_2\, \cdots a_j} = \frac{N!}{a_1!\, a_2!\, \cdots\, a_j!\, (N - a_1 - a_1 - \cdots - a_j)!}$$

be the multivariate binomial coefficient and let

$$B(a_1, a_2, \cdots, a_j) = \frac{\Gamma(a_1)\Gamma(a_2)\cdots\Gamma(a_j)}{\Gamma(a_1 + a_2 + \cdots + a_j)}$$

be the multivariate Beta function.

We give below the expression of the pre-posterior predictive distribution, which plays an important role in the computation of sample sizes. This expression is used in subsection 4.1.3.1 to derive approximate sample size formulae.

4.1.3.1 The pre-posterior predictive distribution

The pre-posterior predictive distribution, derived by Maritz, 1989, is

$$\begin{aligned} p_G(a,b,c,d|\, N,a',b',c',d') &= \frac{\Gamma(A)\Gamma(N+1)}{\Gamma(N+A)}\frac{\Gamma(a+a')}{\Gamma(a+1)\Gamma(a')}\frac{\Gamma(b+b')}{\Gamma(b+1)\Gamma(b')}\times \\ &\quad \frac{\Gamma(c+c')}{\Gamma(c+1)\Gamma(c')}\frac{\Gamma(d+d')}{\Gamma(d+1)\Gamma(d')}, \\ &= \binom{N}{a\; b\; c\; d}\frac{\mathbf{B}(a+a',\, b+b',\, c+c',\, d+d')}{\mathbf{B}(a',\, b',c',d')}, \\ & \qquad a+b+c+d=N, \qquad (4.1.43) \end{aligned}$$

with $A=a'+b'+c'+d'$. It is clear that $\left(\frac{a}{N},\frac{b}{N},\frac{c}{N},\frac{d}{N}\right) \longrightarrow^d (p_{11},p_{10},p_{01},p_{00})$ as $N \longrightarrow \infty$.

If we reconsider the parameterization (a,c,n_1) where the random variable n_1 is $a+b$, a straightforward change of variable in Equation 4.1.43 leads to

$$\begin{aligned} p_G(n_1,a,c|\, N,a',b',c',d') = \binom{N}{n_1}\frac{\mathbf{B}(n_1+a'+b',\, N+c'+d'-n_1)}{\mathbf{B}(a'+b',\, c'+d')}\times \\ \binom{n_1}{a}\frac{\mathbf{B}(a+a',\, n_1+b'-a)}{\mathbf{B}(a',\, b')}\times\binom{n_0}{c}\frac{\mathbf{B}(c+c',\, n_0+d'-c)}{\mathbf{B}(c',\, d')}, \end{aligned}$$

$$n_1=0,1,\cdots,N, \quad a=0,1,\cdots,n_1, \text{ and } c=0,1,\cdots,n_0, \quad n_0=N-n_1, \qquad (4.1.44)$$

whose form is suitable for simulating (n_1,a,c). This is discussed in the sequel.

We can now clearly see that the cross-sectional design involves one more hierarchical step on (n_1,n_0) compared to the case-control study where (n_1,n_0) are instead fixed by design.

For R, the decomposition based on (a,b,m_1) with $m_1=a+c$ can be used:

$$p_G(m_1, a, b|\, N, a', b', c', d') = \binom{N}{m_1} \frac{\mathrm{B}(m_1 + a' + c',\, N + b' + d' - m_1)}{\mathrm{B}(a' + c',\, b' + d')} \times$$
$$\binom{m_1}{a} \frac{\mathrm{B}(a + a',\, m_1 + c' - a)}{\mathrm{B}(a',\, c')} \times \binom{m_0}{b} \frac{\mathrm{B}(b + b',\, m_0 + d' - b)}{\mathrm{B}(b',\, d')},$$
$$m_1 = 0, 1, \cdots, N, \quad a = 0, 1, \cdots, m_1, \text{ and } b = 0, 1, \cdots, m_0, \quad m_0 = N - m_1. \tag{4.1.45}$$

Again, we can clearly see that the cross-sectional study involves one more hierarchical step on (m_1, m_0) when compared to the cohort study, where (m_1, m_0) are fixed by design.

The decompositions in equations (4.1.44) and (4.1.45) are especially useful for the simulation of observations from the pre-posterior predictive distribution. The process of simulating observations from the pre-posterior predictive distribution is an integral part of any Monte Carlo approach to sample size calculation. For example, to simulate an observation using equation 4.1.44, we have the following three step algorithm .

1. simulate n_1 from $\mathcal{BB}(N,\, a' + b',\, c' + d')$.
2. simulate a from $\mathcal{BB}(n_1,\, a',\, b')$.
3. simulate c from $\mathcal{BB}(n_0,\, c',\, d')$, and so on.

Now the calculation of the sample size itself is Monte-Carlo based, since averaging over three nested loops is not computationally efficient. Only a slight modification of the Monte Carlo algorithms in subsections 4.1.1.1 and 4.1.2.1 are necessary. For example, when estimating R, the first two steps are replaced by

- simulate $\pi_i \sim \mathbf{Be}(a' + b',\, c' + d')$ and $n_1^i \sim \mathbf{Bin}(N,\, \pi_i)$.
- simulate $p_1^i \sim \mathbf{Be}(a',\, c')$ and $a_i \sim \mathbf{Bin}(n_1^i,\, p_1^i)$.
- simulate $p_0^i \sim \mathbf{Be}(b',\, d')$ and $b_i \sim \mathbf{Bin}(N - n_1^i,\, p_0^i)$.

4.1.3.2 Approximate sample size formulae for the ALC$_k$

We are ready to derive approximate sample sizes for the risk ratio and the odds ratio for a cross-sectional study.

Odds ratio ψ and log-odds ratio $\phi = \log(\psi)$: The corollary to Theorem 3.2.6 below, will be used later to derive approximate **ALC**$_k$ sample sizes. Because the proof of corollary is similar to that of Corollary 4.1.3, we give only a quick outline.

Corollary 4.1.8. *For $b', c' > \max(2, 3k/2)$, $k \geq 1$,*

$$\lim_{N \longrightarrow \infty} \frac{\sqrt{N + N'}\Big\{E_G[l_\psi^k]\Big\}^{1/k}}{2z_{1-\alpha/2}} = \left\{\int_0^1 \int_0^1 \int_0^1 \left[\frac{wx(1-x)}{2} + (1-w)y(1-y)\right]^{k/2} \times x^{a'+k/2-1}(1-v)^{b'-3k/2-1}\, y^{c'-3k/2-1}(1-y)^{d'+k/2-1} \times \frac{w^{a'+b'-k/2-1}(1-w)^{c'+d'-k/2-1}}{B(a',\, b',\, c',\, d')}\, dx\, dy\, dw \right\}^{1/k}, \tag{4.1.46}$$

where E_G in an expectation over the pre-posterior predictive distribution in equation (4.1.43)*, $N = a + b + c + d$, and $N = a' + b' + c' + d'$.*

Proof. If one replaces $a, b, c,$ and d by $Np_{11}, Np_{10}, Np_{01},$ and Np_{00} in the expression for the posterior variance of ψ_e derived from the k-th posterior moments given by equation (4.1.23), and takes the limit as N goes to infinity, we have

$$\lim_{N\longrightarrow\infty}\left\{N''\,\mathbf{Var}(\psi)\right\}^{k/2} = \left\{\frac{p_{11}p_{00}(p_{10}p_{01}p_{00}+p_{11}p_{01}p_{00}+p_{11}p_{10}p_{01}+p_{11}p_{10}p_{00})}{p_{10}^3p_{01}^3}\right\}^{k/2}, \tag{4.1.47}$$

where $N'' = N + N'$. The idea again is to decompose the form of p_G in the first row of equation (4.1.43) in order to eliminate the indeterminacy due to the denominator in the right hand side of equation (4.1.47), as done in the proof for Corollary 4.1.3. One then obtains the existence of continuous and uniformly bounded sequence of functions h_n such that

$$(N+N')^{k/2}\mathbf{E}_G\Big[\mathbf{Var}(\psi)^{k/2}\Big] = c_N\mathbf{E}\Big[h_N(Z_N)\Big],$$

where $Z_N \longrightarrow^d \text{Dirichlet}(a', b'-3k/2, c'-3k/3, d')$, $\lim_N c_N = 1$, and

$$\begin{aligned}\lim_N h_N(p_{11},p_{10},p_{01},p_{00}) &= \Big\{p_{10}p_{01}p_{00}+p_{11}p_{01}p_{00}+p_{11}p_{10}p_{01}+p_{11}p_{10}p_{00}\Big\}^{k/2}p_{11}^{k/2}p_{00}^{k/2},\\ &= h(p_{11},p_{10},p_{01},p_{00}).\end{aligned}$$

Consequently,

$$\lim_{N\longrightarrow\infty}(N+N')^{k/2}\mathbf{E}_G\Big[\mathbf{Var}(\psi)^{k/2}\Big] = \iiint_{\mathcal{S}} h(p_{11},p_{10},p_{01},p_{00})\times \frac{p_{11}^{a'-1}p_{10}^{b'-3k/2-1}p_{01}^{c'-3k/2-1}p_{00}^{d'-1}}{\mathbf{B}(a',b',c',d')}\,dp_{11}\,dp_{10}\,dp_{01}\,dp_{00},$$

where $\mathcal{S} = \Big\{(p_{11},p_{10},p_{01},p_{00}),\ p_{11}+p_{10}+p_{01}+p_{00} = 1\Big\}$. Set $w = p_{00} + p_{10}$, $x = \dfrac{p_{11}}{p_{11}+p_{10}}$ and $y = \dfrac{p_{01}}{p_{10}+p_{00}}$. This change of variables implies that $p_{11} = wx$, $p_{10} = w(1-x)$, $p_{01} = (1-w)y$, and $p_{00} = (1-w)(1-y)$ with a

Jacobian, $J = w(1-w)$. We then have

$$\lim_{N \longrightarrow \infty} \sqrt{N+N'}\mathbf{E}_G\Big[\mathrm{Var}(\psi)^{k/2}\Big] = \int_0^1 \int_1^0 \int_0^1 \Big[wx(1-x) + (1-w)y(1-y)\Big]^{k/2} \\ x^{a'+k/2-1}(1-x)^{b'-3k/2-1}\; y^{c'-3k/2}(1-y)^{d'+k/2-1} \times \\ \frac{w^{a'+b'-k/2-1}(1-w)^{c'+d'-k/2-1}}{\mathbf{B}(a',\, b',\, c',\, d')}\, dx\, dy\, dw.$$

□

In order to obtain an approximate sample size formula, one needs to solve the approximate equation $\Big\{\mathbf{E}_G[l_\psi^k]\Big\}^{1/k} = l$ using the result in Corollary 4.1.8. Thus an approximate sample size based on the $\mathbf{ALC}_k$ for estimating ψ_e when $b', c' > \max(2, 3k/2)$ is

$$N_\psi = \frac{4z^2_{1-\alpha/2}}{l^2}\Bigg\{\int_0^1 \int_1^0 \int_0^1 \Big[wx(1-x) + (1-w)y(1-y)\Big]^{k/2} \times \\ x^{a'+k/2-1}(1-x)^{b'-3k/2-1}\; y^{c'-3k/2}(1-y)^{d'+k/2-1} \times \\ \frac{w^{a'+b'-k/2-1}(1-w)^{c'+d'-k/2-1}}{\mathbf{B}(a',\, b',\, c',\, d')}\, dx\, dy\, dw\Bigg\}^{2/k} - a' - b' - c' - d'. \tag{4.1.48}$$

Similarly, an approximate sample size based on the $\mathbf{ALC}_k$ for estimating $\phi = \log(\psi)$ when $a', b', c', d' > \max(1, k/2)$ is

$$N_\phi = \frac{4z^2_{1-\alpha/2}}{l^2}\Bigg\{\int_0^1 \Bigg\{\int_0^1 \int_0^1 \Big[wx(1-x) + (1-w)y(1-y)\Big]^{k/2} \\ x^{a'-k/2-1}(1-x)^{b'-k/2-1}\; y^{c'-k/2}(1-y)^{d'-k/2-1}\, dx\, dy\Bigg\} \times \\ \frac{w^{a'+b'-k/2-1}(1-w)^{c'+d'-k/2-1}}{\mathbf{B}(a',\, b',\, c',\, d')}\, dw\Bigg\}^{2/k} - a' - b' - c' - d'. \tag{4.1.49}$$

Note that the sample size formulae in equations (4.1.48) and (4.1.49) can be computed using integral subroutine DQAND from the **ISML library** (which computes any hyper-rectangle integral up to 20 arguments) when $b', c' > \frac{3k}{2} + 1$. Again, Monte Carlo integration methods based on generating t independent Beta distributions $x \sim \mathbf{Be}(a' + k/2,\, b' - 3k/2)$, $y \sim \mathbf{Be}(c' - 3k/2,\, d' + k/2)$, $w \sim \mathbf{Be}(a' + b' - k/2,\, c' + d' - k/2)$ and $x \sim \mathbf{Be}(a' - k/2,\, b' - k/2)$, $y \sim \mathbf{Be}(c' - k/2,\, d' - k/2)$, $w \sim \mathbf{Be}(a' + b' - k/2,\, c' + d' - k/2)$, respectively, appear to be the best of both approaches to compute equations (4.1.48) and (4.1.49). Monte Carlo integration is faster, always feasible when $b', c' > \frac{3k}{2}$ and can reach good accuracy when setting $t = 50000$.

Risk ratio R and log-risk ratio $\log(R)$: The corollary to Theorem 3.2.6 below, will be used later to derive approximate **ALC**$_k$ sample sizes. Again, we provide only a sketch of the proof.

Corollary 4.1.9. *For $a', c', d' > 0$, $a' + c' > k/2$, and $b' > 3k/2$,*

$$\lim_{N \longrightarrow \infty} \frac{\sqrt{N + N'}\left\{\boldsymbol{E}_G[l_R^k]\right\}^{1/k}}{2z_{1-\alpha/2}} = \left\{\int_0^1 \int_0^1 \int_0^1 \left[\frac{vx(1-y)}{2} + (1-v)(1-x)y\right]^{k/2} \times \right. \\ x^{a'+k/2-1}(1-x)^{c'-1}\; y^{b'-3k/2-1}(1-y)^{d'-1} \times \\ \left. \frac{v^{a'+c'-k/2-1}(1-v)^{b'+d'-k/2-1}}{\boldsymbol{B}(a',\, b',\, c',\, d')}\, dx\, dy\, dv \right\}^{1/k}, \tag{4.1.50}$$

where $\boldsymbol{E}_G$ in an expectation over the pre-posterior predictive distribution in equation (4.1.43)*, $N = a + b + c + d$, and $N = a' + b' + c' + d'$.*

Proof. Suppose that k is even. We have shown that $\left(\frac{a}{N}, \frac{b}{N}, \frac{c}{N}, \frac{d}{N}\right) \longrightarrow^d (p_{11}, p_{10}, p_{01}, p_{00})$. Define $g(p_{11}, p_{10}, p_{01}, p_{00}) = \left(p_{11} + p_{01}, \frac{p_{11}}{p_{11} + p_{01}}, \frac{p_{10}}{p_{10} + p_{00}}\right)$. Then

$$\left(\frac{m_1}{N}, \frac{a}{m_1}, \frac{b}{N - m_1}\right) = g\left(\frac{a}{N}, \frac{b}{N}, \frac{c}{N}, \frac{d}{N}\right) \longrightarrow^d g(p_{11}, p_{10}, p_{01}, p_{00}).$$

An application of Proposition 2.4.1 shows that the random variables $p_{11} + p_{01}$, $\dfrac{p_{11}}{p_{11}+p_{01}}$, and $\dfrac{p_{10}}{p_{10}+p_{00}}$ are independently distributed with $p_{11}+p_{01} \sim \mathbf{Be}(a'+c', b'+d')$, $\dfrac{p_{11}}{p_{11}+p_{01}} \sim \mathbf{Be}(a', c')$, and $\dfrac{p_{10}}{p_{10}+p_{00}} \sim \mathbf{Be}(b', d')$. Note that the form of the posterior predictive distribution that is used here is $p_G(m_1, a, b|\, N, a', b', c', d')$ in equation 4.1.45. Now set $m_1 = Nv$, $m_0 = N(1-v)$, $a = Nvx$, and $b = N(1-v)y$ in the expression for the posterior variance of R derived from the k-th posterior moments given by equation (4.1.3), and take the limit as N goes to infinity. We have

$$\begin{aligned}\lim_{N\longrightarrow\infty}\Big\{(N+N')\mathbf{Var}(R)\Big\}^{k/2} &= \frac{x^{k/2}}{v^{k/2}(1-v)^{k/2}y^{3k/2}}\Big[y - yx - vy + vx\Big]^{k/2},\\ &= \frac{x^{k/2}}{v^{k/2}(1-v)^{k/2}y^{3k/2}}\Big[vx(1-y) + (1-v)(1-x)\Big]^{k/2}.\end{aligned} \tag{4.1.51}$$

The trick once again is to decompose the form of $p_G(m_1, a, b|\, N, a', b', c', d')$ in a way similar to the proof for Corollary 4.1.3. This decomposition should contain $p_G(m_1, a, b|\, N, a'-k_1, b'-3k/2, c'-k_2, d')$, where $k_1 \leq a'$ and $k_2 \leq c'$ are two integers such that $k_1 + k_2 = k/2$. Once this is done, one obtains

$$E_G\Big\{(N+N')\mathbf{Var}(R)\Big\}^{k/2} = c_N \mathbf{E}[h_N(Z_N)],$$

where $\lim\limits_{N\longrightarrow\infty} c_N = \dfrac{\mathbf{B}(a'-k_1,\, b'-3k/2,\, c'-k_2,\, d')}{\mathbf{B}(a',\, b',\, c',\, d')}$, h_N is a sequence of uniformly bounded functions with

$$\begin{aligned}\lim_{N} h_N(x, y, v) &= x^{k_1+k/2}(1-x)^{k_2}\Big[vx(1-y) + (1-v)(1-x)y\Big]^{k/2}\\ &= h(x, y, v),\end{aligned}$$

and $Z_N \longrightarrow^{d} \left(p_{11}+p_{01}, \dfrac{p_{11}}{p_{11}+p_{01}}, \dfrac{p_{10}}{p_{10}+p_{00}}\right)$, where the parameters a', b', c' and d' are replaced by $a'-k_1, b'-3k/2, c'-k_2$ and d', respectively. In other words, the random variables $p_{11}+p_{01}$, $\dfrac{p_{11}}{p_{11}+p_{01}}$, and $\dfrac{p_{10}}{p_{10}+p_{00}}$ are independently distributed with $p_{11}+p_{01} \sim \mathbf{Be}(a'+c'-k/2, b'+d'-3k/2)$, $\dfrac{p_{11}}{p_{11}+p_{01}} \sim \mathbf{Be}(a'-k_1, c'-k_2)$, and $\dfrac{p_{10}}{p_{10}+p_{00}} \sim \mathbf{Be}(b'-3k/2, d')$. An application of 3.2.7 and 3.2.6 leads to

$$\begin{aligned}\lim_{N\longrightarrow\infty} E_G\Big\{N\mathbf{Var}(R)\Big\}^{k/2} \;&=\; \int_0^1\int_0^1\int_0^1 \Big[vx(1-y)+(1-v)(1-x)y\Big]^{k/2}\times\\ &\quad x^{a'+k/2-1}(1-x)^{c'-1}y^{b'-3k/2-1}(1-y)^{d'-1}\times\\ &\quad \frac{v^{a'+c'-k/2-1}(1-v)^{b'+d'-k/2-1}}{\mathbf{B}(a',\,b',\,c',\,d')}\,dx\,dy\,dv,\end{aligned} \tag{4.1.52}$$

which completes the proof when k is even. Similarly when k is odd. □

In order to obtain an approximate sample size formula, one needs to solve the approximate equation $\Big\{\mathbf{E}_G[l_R^k]\Big\}^{1/k} = l$ using the result in Corollary 4.1.9. Thus an approximate sample size based on the $\mathbf{ALC}_k$ for estimating R when $a', c', d' > 0$, $a'+c' > k/2$, and $b' > \max(2, 3k/2)$ is

$$\begin{aligned}N_R \;&=\; \frac{4z^2_{1-\alpha/2}}{l^2}\Bigg\{\int_0^1\int_0^1\int_0^1 \Big[vx(1-y)+(1-v)(1-x)\Big]^{k/2}\times\\ &\quad x^{a'+k/2-1}(1-x)^{c'-1}y^{b'-3k/2-1}(1-y)^{d'-1}\times\\ &\quad \frac{v^{a'+c'-k/2-1}(1-v)^{b'+d'-k/2-1}}{\mathbf{B}(a',\,b',\,c',\,d')}\,dx\,dy\,dv\Bigg\}^{2/k} - a'-b'-c'-d'.\end{aligned} \tag{4.1.53}$$

Similarly, an approximate sample size based on the $\mathbf{ALC}_k$ for estimating

$\log(R)$ when $c', d' > 0$, $a', b' > \max(1, k/2)$ is

$$\begin{aligned} N_{\log(R)} &= \frac{4z^2_{1-\alpha/2}}{l^2} \Biggl\{ \int_0^1 \Biggl\{ \int_0^1 \int_0^1 \Bigl[vx(1-y) + (1-v)(1-x)y \Bigr]^{k/2} \times \\ &\quad x^{a'-k/2-1}(1-x)^{c'-1}\, y^{b'-k/2-1}(1-y)^{d'-1} dx\, dy \Biggr\} \times \\ &\quad \frac{v^{a'+c'-k/2-1}(1-v)^{b'+d'-k/2-1}}{\mathrm{B}(a',\, b',\, c',\, d')}\, dv \Biggr\}^{2/k} - a' - b' - c' - d'. \end{aligned} \tag{4.1.54}$$

4.1.3.3 An example

Recently, case-only designs have been used in the field of genetic epidemiology to assess gene-environment interaction effects. We briefly discuss these designs below and place one of our sample size problems in the context of such designs.

The case-only design as a two sample problem We first introduce the traditional 2 × 2 × 2 or 2 × 4 case-control design for inference on gene-environment interaction. Let E be a dichotomous variable representing the presence or absence of exposure, let G be an inherited dichotomous variable indicating susceptibility genotype, and let D be the status variable for the disease under investigation. The values $E = 0$ and $E = 1$ will denote absence or presence of exposure, respectively. Similarly for G. As in any case-control study, we start by collecting N cases and M controls. Suppose that appropriate controls are available and that there are no confounding effects. Each case and each control selected is classified into one of the four combination groups of E and G. Table 4.4 serves as a summary of the collected information on disease, exposure, and genotype susceptibility along

Table 4.4: **Generic 2 × 4 table for gene-environment interaction analysis in case-control settings.**

Exposure	Susceptibility genotype	Cases	Control	Odds ratios	Estimates
0	0	A_{00}	B_{00}	1.0	
0	1	A_{01}	B_{01}	$\psi_G = \frac{p_{01}q_{00}}{p_{00}q_{01}}$	$\widehat{\psi}_G = \frac{A_{01}B_{00}}{A_{00}B_{01}}$
1	0	A_{10}	B_{10}	$\psi_E = \frac{p_{10}q_{00}}{p_{00}q_{10}}$	$\widehat{\psi}_E = \frac{A_{10}B_{00}}{A_{00}B_{10}}$
1	1	A_{11}	B_{11}	$\psi_{EG} = \frac{p_{11}q_{00}}{p_{00}q_{11}}$	$\widehat{\psi}_{EG} = \frac{A_{11}B_{00}}{A_{00}B_{11}}$
		N	M		

with the parameters of interest. Cases and controls in the first row of Table 4.4 form the reference group. Let $p_{ij} = P(G = i,\, E = j \mid D = 1)$, and $q_{ij} = P(G = i,\, E = j \mid D = 0)$, $i, j = 0, 1$ be the cell probabilities given disease status. Denote by ψ_G, ψ_E, and ψ_{GE}, the ratio of the odds of disease for $G = 1$ and $E = 0$, $G = 0$ and $E = 1$, and $G = 1$ and $E = 1$ relative to the reference group (G=0, E=0). By definition,

$$\begin{aligned}
\psi_G &= \frac{P(D=1 \mid G=1, E=0)}{P(D=1 \mid G=0, E=0)} \Bigg/ \frac{P(D=0 \mid G=1, E=0)}{P(D=0 \mid G=0, E=0)} = \frac{p_{01}\, q_{00}}{p_{00}\, q_{01}} \\
\psi_E &= \frac{P(D=1 \mid G=0, E=1)}{P(D=1 \mid G=0, E=0)} \Bigg/ \frac{P(D=0 \mid G=0, E=1)}{P(D=0 \mid G=0, E=0)} = \frac{p_{10}\, q_{00}}{p_{00}\, q_{10}} \\
\psi_{GE} &= \frac{P(D=1 \mid G=1, E=1)}{P(D=1 \mid G=0, E=0)} \Bigg/ \frac{P(D=0 \mid G=1, E=1)}{P(D=0 \mid G=0, E=0)} = \frac{p_{11}\, q_{00}}{p_{00}\, q_{11}}.
\end{aligned}$$

One way to measure the influence of gene-environment interaction on disease in the population is to compute the synergy index I_{GE},

$$I_{GE} = \frac{\psi_{GE}}{\psi_G\, \psi_E} = \frac{p_{11}\, p_{00}}{p_{01}\, p_{10}} \Bigg/ \frac{q_{11}\, q_{00}}{q_{01}\, q_{10}}.$$

More importantly, Piegorsch et al., 1994 shows that I_{GE} is equivalent to the interaction parameter between genotype and environment under a logistic regression model. Now the estimation of I_{GE} requires information on the controls, which would routinely be obtained in a traditional case-control study. However, under certain assumptions it is possible to avoid collecting information on controls when making inference about I_{GE}.

Suppose now that there is strong theoretical or empirical justification for assuming that genotype and exposure occur independently in the population. Under this independence assumption and assuming that the disease is rare, Piegorsch et al., 1994 show that

$$I_{GE} \approx \frac{P(E=1|\,G=1, D=1)\,P(E=0|\,G=0, D=1)}{P(E=0|\,G=1, D=1)\,P(E=1|\,G=0, D=1)} = \frac{p_{11}\,p_{00}}{p_{01}\,p_{10}}. \quad (4.1.55)$$

Approximation (4.1.55) implies that we can estimate the I_{GE} with a case-series only as all the parameters in the right hand side of (4.1.55) can be estimated without information on controls. Also shown by these authors and others (Yang et al., 1997, Umbach and Weinberg, 1997, Schmidt and Schaid, 1999, Albert et al., 2001), is that the estimation of the gene-environment interaction with a case-series offers greater precision than that provided by the traditional approach. Moreover, the case-series study is more economical, and, more importantly, using case-series avoids the difficult problem of validation for the control group. These papers discuss the drawback of case-only studies, especially when the independence assumption is violated.

Clearly, the above case-only design mimics that of the cross-sectional sampling design discussed in section 4.1.4 and, therefore our Bayesian sample size approaches for estimating the odds and log-odds discussed there can be used for sample size problems when estimating I_{GE} and $\log(I_{GE})$. This would, of course, also be case if q_{00}, q_{10}, q_{01} and q_{11} were known exactly in the traditional approach.

4.1.4 Matched analysis

In matched analyses first introduced in subsection 2.3.4, the parameters of interest are the exposure odds ratio, ψ_e', in case-control setting, and the disease odds ratio, ψ_d', in the cohort setting. Since both ψ_e' and ψ_d' are the ratio of two proportions under the assumption that the relative risk is constant over the level of the covariates (see subsection 2.3.4), it is sufficient to study ψ_e'. We again use the notation $\mathtt{T}$, first introduced in section 2.3 for Table 2.2. For the matched pair analyses, the prior/likelihood model that we use is the Dirichlet/multinomial model. Denote by $\mathtt{T} = \left(\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\right)$ a realization of the Multinomial distribution and by $\mathtt{T}' = \left(\mathtt{a}', \mathtt{b}', \mathtt{c}', \mathtt{d}'\right)$ the parameters of the Dirichlet prior distribution underlying the matched analysis. The combination of the prior-likelihood tables $\mathtt{T}$ and $\mathtt{T}'$ leads to a posterior table $\mathtt{T}'' = \left(\mathtt{a}'', \mathtt{b}'', \mathtt{c}'', \mathtt{d}''\right)$ where $a'' = a + a'$, $b'' = b + b'$, $c'' = c + c$, and $d'' = d + d'$, where $N = a + b + c + d$ is the sample size. Under this model, the posterior density of $\psi_e' = \dfrac{p_{10}'}{p_{01}'}$ is

$$p_{\psi_e'}(\psi|\,\mathtt{T}'') = \frac{1}{\mathbf{Be}(b'',\, c'')}\,\frac{\psi^{b''-1}}{(1+\psi)^{b''+c''}}, \tag{4.1.56}$$

whereas the posterior density of $\phi_e' = \log\left(\dfrac{p_{10}'}{p_{01}'}\right)$ is

$$p_{\phi_e'}(\phi|\,\mathtt{T}'') = \frac{1}{\mathbf{Be}(b'',\, c'')}\,\frac{e^{b''\phi}}{(1+e^{\phi})^{b''+c''}}. \tag{4.1.57}$$

Since b'' and c'' are the only parameters that play a role in the computation of the posterior distributions of ψ'_e and ϕ'_e, we can summarize the table $\mathtt{T}''$ further, simply as $\mathtt{T}'' = \left(\mathtt{b}+\mathtt{b}', \mathtt{m}-\mathtt{b}+\mathtt{c}', \mathtt{N}-\mathtt{m}+\mathtt{a}'+\mathtt{d}'\right)$, where $m = b+c$.

The densities in equations 4.1.56 and 4.1.57 have been studied in chapter 3 as the densities of the odds parameter, ω and the log-odds ϕ. Therefore, all techniques developed for ω and ϕ apply here, and, in particular, the results for approximate lengths of HPD and equal-tailed intervals.

We now derive the pre-posterior predictive distribution for matched studies. We use this form for the pre-posterior predictive distribution in our computation of Bayesian sample sizes for matched problems.

4.1.4.1 The pre-posterior predictive distribution

We begin with a proposition that allows us to derive a convenient form of the pre-posterior predictive distribution (see equation (4.1.61) below)

The prior distribution

$$(p'_{11}, p'_{10}, p'_{01}, p'_{00}) \sim \text{Dirichlet } (a', b', c', d')$$

implies the marginal prior distribution

$$(p'_{10}, p'_{01}, p'_{00}) \sim \text{Dirichlet } (b', c', a'+d').$$

Proposition 4.1.10. *Let*

$$(p_{10}, p_{01}, p_{00}) \sim \textit{Dirichlet } (b', c', a'+d'), \tag{4.1.58}$$

where $p_{10}, p_{01}, p_{00} > 0$ *and* $p_{10}+p_{01}+p_{00} = 1$. *Let* $\pi = p_{10}+p_{01}$ *and* $\theta = \dfrac{p_{10}}{p_{10}+p_{01}} = \dfrac{\psi'_e}{\psi'_e+1}$. *Then* π *and* θ *are independent random variables* *with*

$$\begin{aligned} \pi &\sim Be(b'+c', a'+d') \\ \theta &\sim Be(b', c'). \end{aligned} \tag{4.1.59}$$

Proof. We have $p_{10} = \pi\theta$ and $p_{01} = (1-\theta)\pi$. This transformation has a Jacobian $J = \pi$, $0 < \pi < 1$, and $\theta < 1$. Thus,

$$f(\pi, \theta) \propto \pi^{b'+c'-1}(1-\pi)^{a'+d'-1}\,\theta^{b'-1}(1-\theta)^{c'-1}.$$

□

It is well known (Royston, 1993) that $(a, b, c, d) \sim \text{Multinomial}(N, p_{11}, p_{10}, p_{01}, p_{00})$ also implies that

$$\begin{aligned} m = b + c &\sim \mathbf{Bin}(N,\, \pi) \\ b|m &\sim \mathbf{Bin}(m,\, \theta). \end{aligned} \tag{4.1.60}$$

Equations (4.1.59) and (4.1.60) yield the following decomposition of the pre-posterior predictive distribution of the random variable (m, b) as the product of two Beta-Binomial distributions,

$$\begin{aligned} p(m, b|N, a', b', c', d') &= p_m(m|\,N, b'+c', a'+d')\,p_b(b|\,m, b', c') \\ &= \binom{N}{m}\frac{\mathbf{B}(b'+c'+m,\, N+a'+d'-m)}{\mathbf{B}(b'+c',\, a'+d')} \times \\ &\qquad \binom{m}{b}\frac{\mathbf{B}(b'+b,\, m+c'-b)}{\mathbf{B}(b',\, c')}, \\ &\qquad m = 0, \cdots, N \text{ and } b = 0, \cdots, m. \end{aligned} \tag{4.1.61}$$

We use this form for the pre-posterior predictive distribution in our computation of Bayesian sample sizes for matched problems. Equation (4.1.61) also provides a simple two-step algorithm for generating observations from $p(m, b|N, a', b', c', d')$:

1. Simulate m from $\mathcal{BB}(N,\ b'+c',\ a'+d')$.

2. Simulate b from $\mathcal{BB}(m,\ b',\ c')$.

For the matched problem, the five main sample size criteria are related to the following criteria functions.

$$
\begin{aligned}
\mathrm{acc}_k(N,a',b',c',d') &= \left(\sum_{m=0}^{N}\sum_{b=0}^{m}\left\{\int_{\theta\in\mathtt{HPD}(l)} p(\theta|\,T'')\,d\theta\right\}^k p_{(m,b)}(m,b|\,N,a',b',c',d')\right)^{1/k}, \\
\mathrm{alc}_k(N,a',b',c',d') &= \left(\sum_{m=0}^{N}\sum_{b=0}^{m}\left\{\int_{\theta\in\mathtt{HPD}(1-\alpha)} d\theta\right\}^k p_{(m,b)}(m,b|\,N,a',b',c',d')\right)^{1/k}, \\
\mathrm{woc}(N,a',b',c',d') &= \sup_{0\le b\le m\le N}\left\{\int_{\theta\in\mathtt{HPD}(1-\alpha)} d\theta\right\}, \\
\mathrm{mcoc}(N,a',b',c',d') &= \mathrm{med}_{0\le b\le m\le N}\left\{\int_{\theta\in\mathtt{HPD}(l)} p(\theta|\,T'')\,d\theta\right\}, \\
\mathrm{mloc}(N,a',b',c',d') &= \mathrm{med}_{0\le b\le m\le N}\left\{\int_{\theta\in\mathtt{HPD}(1-\alpha)} d\theta\right\},
\end{aligned}
$$

where θ is either ψ_e' or either ϕ_e'. In the sequel, we simply use the notation $\mathrm{alc}(N,a',b',c',d')$ and $\mathrm{acc}(N,a',b',c',d')$ for $k=1$.

Having derived the pre-posterior predictive distribution, we can now derive the approximate sample size formulae for the $\mathbf{ALC}_k$, $k=1,2$. Depending on the accuracy required, these estimates can either be used by themselves, or they can serve as the starting values for more accurate estimates based on Monte Carlo simulations.

4.1.4.2 Approximate sample sizes for the ALC_k, $k = 1, 2$

Approximate sample size formula for the ALC: The following corollary to Theorem 3.2.4 on the limiting distribution of the pre-posterior predictive distribution, that is needed in the sequel, considerably simplifies the computational burden when assessing the approximate sample size.

Corollary 4.1.11.

$$\left(\frac{b}{N}, \frac{c}{N}, \frac{N-b-c}{N}\right) \longrightarrow^{d} \textit{Dirichlet}\ (b',\, c',\, a'+d'), \qquad \textit{as } N \longrightarrow \infty.$$

Remark 4.1.2. Corollary 4.1.11 allows us to replace m by $N\pi$, b by $N\pi\theta$ and c by $N\pi(1-\theta)$ when N is large. Thus, $\mathrm{alc}_k(N, a', b', c', d')$ may be replaced by

$$\begin{aligned}\widetilde{\mathrm{alc}}_k(N, a', b', c', d') \ = \ & \left(\int_0^1 \int_0^1 \left\{\int_{\psi \in \mathtt{HPD}(N,\pi,\theta,1-\alpha)} d\psi\right\}^k \frac{\theta^{b'-1}(1-\theta)^{c'-1}}{\mathbf{B}(b',\, c')}\right. \\ & \left. \frac{\pi^{b'+c'-1}(1-\pi)^{a'+d'-1}}{\mathbf{B}(b'+c',\, a'+d')}\, d\theta\, d\pi\right)^{1/k},\end{aligned}$$

where $\mathtt{HPD}(N, \pi, \theta, 1-\alpha)$ represents the HPD interval of coverage $1-\alpha$ for ψ_e' from the distribution

$$f_{\psi_e'}(\psi) = \frac{1}{\mathbf{Be}(N\pi\theta + b',\, N\pi(1-\theta) + c')} \frac{\psi^{N\pi\theta+b'-1}}{(1+\psi)^{N\pi+b'+c'}}.$$

The pre-posterior predictive distribution of $X_N = (b, c, N-b-c)$ can be rewritten as

$$p_{X_N}(b, c|N, b', c', a'+d') = \binom{N}{b\ c} \frac{\mathbf{B}(b+b',\, c+c',\, b+c+b'+c',\, N+a'+d'-b-c)}{\mathbf{B}(b',\, c',\, a'+d')}, \tag{4.1.62}$$

where $0 \leq b+c \leq N$.

Let $l_{\psi_e'}(b,c) = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(\psi_e'|\,T'')}$ be the first order approximation of the HPD interval for ψ_e'. Let $N' = a' + b' + c' + d'$ be the prior sample size.

The corollary to Theorem 3.2.6 below will be used later to derive approximate sample size when considering the alc_k.

Corollary 4.1.12. *For $c' > 2$,*

$$\lim_N \sqrt{N+N'}\frac{\boldsymbol{E}_{X_N}[l_{\psi_e'}(b,c)]}{2z_{1-\alpha/2}} = \frac{\boldsymbol{B}(b'+1/2,\, c'-3/2)\,\boldsymbol{B}(b'+c'-1/2,\, a'+d')}{\boldsymbol{B}(b',\, c',\, a'+d')}. \tag{4.1.63}$$

Proof. Let $Y_N = \left(\frac{b}{N}, \frac{c}{N}, \frac{N-b-c}{N}\right)$ and $\mathcal{F}_N$ be the set of pairs (b,c) where the mass of Y_N is positive. We have

$$\begin{aligned}
\sqrt{N+N'}\mathbf{E}_{X_N}[\sqrt{\mathbf{Var}(\psi_e'|\,\mathrm{T}'')}] &= \sqrt{N+N'}\sum_{(x,y)\in\mathcal{F}_N}\sqrt{\frac{(Nx+Ny+b'+c'-1)(Nx+b')}{(Ny+c'-1)^2(Ny+c'-2)}}\times \\
&\qquad p_{Y_N}(x,y|N,a',b',c',d') \\
&= \frac{\mathbf{B}(b',\, c'-2,\, a'+d')}{\mathbf{B}(b',\, c',\, a'+d')}\sqrt{N+N'}c_N \;\times \\
&\qquad \sum_{(x,y)\in\mathcal{F}_N} h_N(x,y)p_{Y_N}(x,y|N,b',c'-2,a'+d'), \\
&= \frac{\mathbf{B}(b',\, c'-2,\, a'+d')}{\mathbf{B}(b',\, c',\, a'+d')}\sqrt{N+N'}c_N\ \mathbf{E}[V_n],
\end{aligned}$$

where the functions

$$h_N(x,y) = \sqrt{\frac{(Nx+Ny+b'+c'-1)(Nx+b')(Ny+c'-2)}{(N+N')^3}},$$

are uniformly bounded and continuous, $c_N = \frac{(N+N')^{3/2}}{(N+N'-1)(N+N-2)}$, $\lim_N \sqrt{N+N'}c_N = 1$, and $V_N = h_N(Z_N)$ with $p_{Z_N}(x,y) = p_{Y_N}(x,y|N,b',c'-2,a'+d')$, $(x,y)\in\mathcal{F}_N$ and $Z_N \longrightarrow^d \text{Dirichlet}(b',c'-2,a'+d')$. We also have

$\lim_N f_N(x,y) = \sqrt{(x+y)xy}$. Theorems 3.2.7 and 3.2.6 imply that

$$\begin{aligned}
\lim_N \mathbf{E}[V_n] &= \iint_S \sqrt{(x+y)xy}\, \frac{x^{b'-1}y^{c'-3}(1-x-y)^{a'+d'-1}}{\mathrm{B}(b',\, c'-2,\, a'+d')}\, dx\, dy \\
&= \frac{1}{\mathrm{B}(b', c'-2, a'+d')} \left\{ \int_0^1 \theta^{b'+1/2-1}(1-\theta)^{c'-3/2-1}\, d\theta \right\} \times \\
&\quad \left\{ \int_0^1 \pi^{b'+c'-1/2-1}(1-\pi)^{a'+d'-1} \right\}, \\
&= \frac{\mathrm{B}(b'+1/2,\, c'-3/2)\,\mathrm{B}(b'+c'-1/2,\, a'+d')}{\mathrm{B}(b',\, c'-2,\, a'+d')},
\end{aligned}$$

where $S = \{(x,y) : 0 < x+y < 1\}$. Thus

$$\lim_N \sqrt{N+N'}\frac{\mathbf{E}_{X_N}[l_{\psi'_e}(b,c)]}{2z_{1-\alpha/2}} = \frac{\mathrm{B}(b'+1/2,\, c'-3/2)\,\mathrm{B}(b'+c'-1/2,\, a'+d')}{\mathrm{B}(b',\, c',\, a'+d')}.$$

□

Therefore, an approximate sample size for estimating ψ'_e using the **ALC** is

$$N_{\psi'_e} = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(b'+1/2,\, c'-3/2)\,\mathrm{B}(b'+c'-1/2,\, a'+d')}{\mathrm{B}(b',\, c',\, a'+d')}\right)^2 - a' - b' - c' - d'. \tag{4.1.64}$$

In general, Corollary 4.1.12 can be generalized to incorporate $k = 2$ as follows.

Corollary 4.1.13. *For $c' > 3k/2$,*

$$\lim_N \sqrt{N+N'}\frac{\left\{E_{X_N}[l^k_{\psi'_e}(b,c)]\right\}^{1/k}}{2z_{1-\alpha/2}} = \left\{\frac{B(b'+k/2,\, c'-3k/2)\, B(b'+c'-k/2,\, a'+d')}{B(b',\, c',\, a'+d')}\right\}^{1/k}. \tag{4.1.65}$$

Proof. Use the same decomposition and devices exploited in the proof of Corollary B.0.7 to eliminate the indeterminacy due to the denominator $(Ny + c'-1)^k(Ny+c'-2)^{k/2}$, and then proceed as in the proof of Corollary 4.1.12.

□

When $c' > 3k/2$, Corollary 4.1.13 gives an approximate sample size for ψ'_e under the $\mathbf{ALC}_k$, of

$$N_{\psi'_e} = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(b'+k/2,\, c'-3k/2)\,\mathrm{B}(b'+c'-k/2,\, a'+d')}{\mathrm{B}(b',\, c',\, a'+d')}\right)^{2/k} - a' - b' - c' - d'. \tag{4.1.66}$$

Similarly, define $l_{\phi'_e}(b,c) = 2z_{1-\alpha/2}\sqrt{\mathbf{Var}(\phi'_e|\,\mathrm{T}'')}$ be the first order approximation of the HPD interval for ϕ'_e. Then

Corollary 4.1.14. *For $b', c' > k/2$,*

$$\lim_N \sqrt{N+N'}\frac{\left\{E_{X_N}[l^k_{\phi'_e}(b,c)]\right\}^{1/k}}{2z_{1-\alpha/2}} = \left\{\frac{B(b'-k/2,\, c'-k/2)\,B(b'+c'-k/2,\, a'+d')}{B(b',\, c',\, a'+d')}\right\}^{1/k}. \tag{4.1.67}$$

Therefore, an approximate sample size for $\log(\phi'_e)$ under the $\mathbf{ALC}_k$ when $b', c' > k/2$ is

$$N_{\phi'_e} = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(b'-k/2,\, c'-k/2)\,\mathrm{B}(b'+c'-k/2, a'+d')}{\mathrm{B}(b',\, c',\, a'+d')}\right)^{2/k} - a' - b' - c' - d'. \tag{4.1.68}$$

Remark 4.1.3 below is a digression pointing out two similarities and one difference between Bayesian and frequentist sample size calculations as well as a natural linkage between the one sample problem and the two sample matched analysis.

Remark 4.1.3.

- Royston, 1993 has shown that the exact unconditional power function is an average of the conditional power over the distribution of the number of discordant pairs m. In the sense that the criterion function for the $\mathbf{ALC}_k$ and $\mathbf{ACC}_k$ are also averages of the conditional criterion functions over m, there is an analogy between Royston, 1993 frequentist albeit power-based approach and ours.

- Schlesselman, 1982, Parker and Bregman, 1986, Connett et al., 1987 Royston, 1993, and Julious and Campbell, 1998 point out that the overall sample size, N, for estimating ψ'_e can be approximated by $N = \frac{M}{\pi}$, where M is the sample size for estimating the odds, ψ, from the one sample problem (representing in matched studies, the number of discordant pairs, m) and $\pi = p'_{10} + p'_{01}$, the probability that a matched pairs is discordant. A similar relation applies to the Bayesian paradigm. Indeed, it is clear that equations (4.1.66) can be written as

$$N_{\psi'_e} + a' + b' + c' + d' = \left(\frac{\mathbf{B}(b' + c' - k/2,\, a' + d')}{\mathbf{B}(b' + c',\, a' + d')}\right)^{2/k} (M_\psi + b' + c'), \tag{4.1.69}$$

where $M_\psi = 4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathbf{B}(b' + k/2,\, c' - 3k/2)}{\mathbf{B}(b',\, c')}\right)^{2/k} - b' - c'$ is the sample size of the odds. More specifically, when $k = 2$ and $b' > 3$, we have $N_{\psi'_e} + a' + b' + c' + d' = \frac{M_\psi + b' + c'}{\pi_D}$, where $\frac{1}{\pi_D} = \mathbf{E}\left(\frac{1}{\pi}\right) = \frac{\mathbf{B}(b' + c',\, a' + d')}{\mathbf{B}(b' + c' - 1,\, a' + d')} = \frac{b' + c' - 1}{a' + b' + c' + d' - 1}$, with $\pi \sim \mathbf{Be}(b' + c',\, a' + d')$. A similar relation holds for ϕ'_e in the Bayesian context. As a consequence, the table of **ALC** sample sizes for the one sample problem in appendix J can easily be used to construct tables of **ALC** sample sizes for the two sample problem. We show in section 4.2 that similar relations occur with Bayesian criteria that average over nuisance parameters.

- In Bayesian approaches, since the **ALC**$_k$ approximate sample for estimating ψ_e and ϕ_e are not the same unless $c' - b' = 2k$, it matters whether exposure is beneficial or not. This contrast with the frequentist analysis (Connett et al., 1987, Royston, 1993, Julious and Campbell, 1998) where it is irrelevant whether the exposure is beneficial or not.

4.1.4.3 Computational challenges

In addition to the computational challenges discussed in chapter 3 for computing an HPD interval for the odds and log-odds, our algorithms in the matched pair settings require that HPD intervals be evaluated for all $(N+1)(N+2)/2$ mass points (m,b), $m = 0, \cdots, N$ and $b = 0, \cdots, m$, given N. In evaluating the criterion function, we therefore need to store a vector of HPD lengths or coverages and another vector of pre-posterior predictive weights for the **ALC** or **ACC**, both of size $(N+1)(N+2)/2$, at each intermediate step N. Since storage size increases with N, it is not computationally efficient to store the entire vectors. A natural alternative is to use a loop for $m = 0, \cdots, N$ and, for each m, to create a vector of length $m+1$ to store the lengths of the credible intervals of interest; that is, first average over the conditional distribution of b given m. Unfortunately, this idea can be implemented only under the $\mathbf{ALC}_k$, $\mathbf{ACC}_k$, and **WCOC** since $\sum_{(m,b)} = \sum_m \sum_b$ and $\sup_{(m,b)} = \sup_m \sup_b$, respectively, but not under the **MLOC** and **MCOC**. For the latter two criteria, sample sizes are based on Monte-Carlo estimates, as shown below.

Sketch of the Monte Carlo simulation approach to determine the optimal sample size when estimating ψ_e' using the MCOC: For each step in the bisectional search over N, one performs the following, sequence of steps.

- Simulate $\pi_i \sim \mathbf{Be}(b' + c',\, a' + d')$ and $\theta_i \sim \mathbf{Be}(b', c'),\ i = 1, \cdots, m$.

- For each pair $(\pi_i,\, \theta_i)$, simulate two independent observations $m_i \sim \mathbf{Bin}(N,\, \pi_i)$ and $b_i \sim \mathbf{Bin}(m_i,\, \theta_i)$ given m_i.

- For each i, simulate $p_j \sim \mathbf{Be}(b_i + b',\, m_i - b_i + c'),\ j = 1, \cdots, M$. Use the observations $\psi_j = \dfrac{p_j}{1 - p_j}$ to estimate the coverage α_i of the HPD interval given its pre-specified lenght l using algorithm 3 in subsection 2.7.2.3.

- Compute $\text{mcoc}(N, a', b', c', d') \approx \text{med}_{1 \leq i \leq m}\, \alpha_i$.

4.1.4.4 Additional results on linearity

Although we have written a program to exactly compute the criterion functions $\mathbf{ALC}_k(N, a', b', c', d')$, $\mathbf{ACC}_k(N, a', b', c', d')$, and $\mathbf{WOC}(N, a', b', c', d')$ for ψ'_e and ϕ'_e for $N = 1, \cdots, M$, we do not suggest using this algorithm unless M is less than 500. Indeed, it takes about 9 hours for $M = 250$ on a 600Mhz machine. Figure 3.1.3, which is based on the approximation of the HPD intervals in subsection 3.2.4.2, valid when $b' > 1$, suggests that all the observations made about the odds and log-odds in chapter 3 might be applicable here, especially the linear relation. It took about 15 hours to obtain the data used to plot Figure 3.1.3.

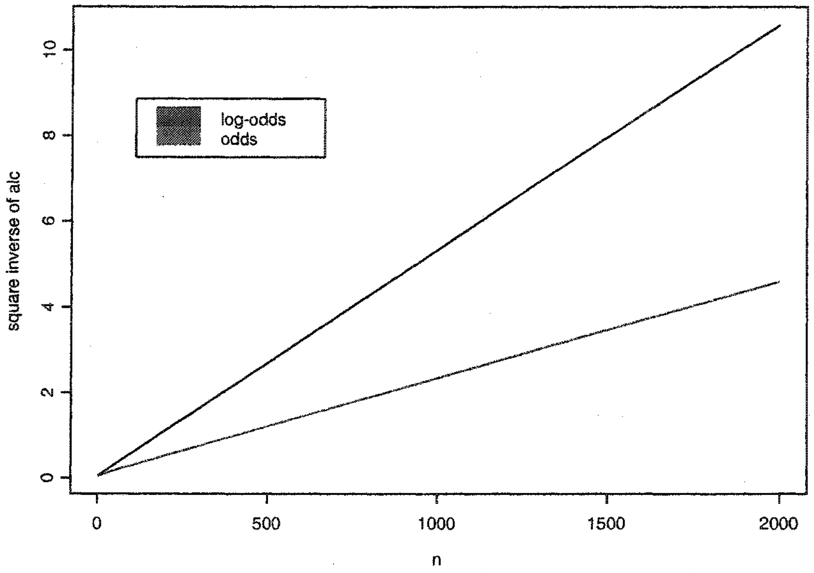


Figure 4.1.5: Graph of $\dfrac{1}{\text{alc}^2(n,3,3,3,3)}$ and $\dfrac{1}{\text{alc}^2(n,2,2,2,2)}$ for the odds and the log-odds as a function of n, respectively.

4.1.4.5 An example

The following example is an adaption of an example in Royston, 1993.

A certain drug to prevent or reduce nausea and vomiting is to be compared with a placebo in the treatment of intestinal obstruction due to terminal malignancy. The allocation to active drug or placebo is randomized. To avoid possible confounding effects, each patient from the drug group is matched with a patient from the placebo group on age, diet, and gender. Denote by 0 and 1 the absence and presence of nausea and/or vomiting. Let p_{10} be the probability that the treated patient in a pair responds to drug and the other does not respond to the placebo, within 24 hours following treatment and similarly, define p_{01} with the opposite results for the pair. Let $\psi = \dfrac{p_{10}}{p_{01}}$ be the odds ratio. A previous informal study suggests that the prior information available is equivalent to the data in Table 4.5.

Table 4.5: **Prior information**

Placebo	Drug	
	presence (1)	absence (0)
presence (1)	3	8
absence (0)	3	6

Assuming a researcher is interested in a sample size that guarantees an HPD interval of length l with a specified average coverage probability, the **ACC** would be chosen. Here the sample size is the total number of matched pairs to be included in the study. When $1 - \alpha = .90$ and $l = 2.0$, the corresponding true sample size is $N = 740$. With the Monte Carlo approach with $m = M = 2000$, we obtain an estimate $\bar{N} = 738$ using the average of 10 estimated sample sizes. The corresponding standard deviation for $\bar{N}$ is 10.084.

Up to now, we have been dealing with sample size criteria that do not explicitly involve nuisance parameters. Below we present some Bayesian sample size criteria that average explicitly over nuisance parameters and apply them to cohort and case-control designs.

4.2 Bayesian sample size criteria that average over nuisance parameters

It is clear that the posterior densities of R and $\log(R)$ given in equations 4.1.21 and 4.1.22 (resp. ψ_e and $\phi_e = \log(\psi_e)$ given by the equations 4.1.21 and 4.1.22) depend on the nuisance parameter $p = p_0$ (resp. $p = p_0'$). Another way to explicitly account for the presence of a nuisance parameter p in Bayesian sample size determination problems is to integrate over the prior distribution for p_0. That is, to seek for the minimal n that satisfies, for instance,

$$\left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_0^1\int_{R\in\mathrm{HPD}(\mathbf{x}_n,n,l,p)} p_R(R|\,\mathbf{x}_n,p)\,p_p(p|\,\mathbf{x}_n)\,dR\,dp\right\}^k p_{X_n}(\mathbf{x}_n)\,dx\right)^{1/k} \geq 1-\alpha, \tag{4.2.1}$$

or

$$\left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\int_0^1\left\{\int_{R\in\mathrm{HPD}(\mathbf{x}_n,n,l,p)} p_R(R|\,\mathbf{x}_n,p)\,dR\right\}^k p_p(p|\,\mathbf{x}_n)\,p_{X_n}(\mathbf{x}_n)\,dp\,dx\right)^{1/k} \geq 1-\alpha, \tag{4.2.2}$$

for the $\mathbf{ACC}_k$ and

$$\left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\left\{\int_0^1\int_{R\in\mathrm{HPD}(\mathbf{x}_n,n,1-\alpha,p)} p_p(p|\,\mathbf{x}_n)\,dR\,dp\right\}^k p_{X_n}(\mathbf{x}_n)\,dx\right)^{1/k} \leq l, \tag{4.2.3}$$

or

$$\left(\int_{\mathbf{x}_n\in\mathcal{X}_n}\int_0^1\left\{\int_{R\in\mathrm{HPD}(\mathbf{x}_n,n,1-\alpha,p)} dR\right\}^k p_p(p|\,\mathbf{x}_n)\,p_{X_n}(\mathbf{x}_n)\,dp\,dx\right)^{1/k} \leq l, \tag{4.2.4}$$

for the $\mathbf{ALC}_k$. Equations (4.2.1) and (4.2.2) (resp. (4.2.3) and (4.2.4)) are equivalent when $k = 1$ and were suggested by Joseph et al., 1995. These two different ways of generalizing the $\mathbf{ALC}_k$ and the $\mathbf{ACC}_k$ add to the richness of Bayesian sample size calculation criteria where the same notion can be looked at from various viewpoints. An equivalent worst outcome criterion can be defined by seeking the minimal n such that

$$\inf_{\mathrm{x}_n \in \mathcal{X}_n} \left\{ \int_0^1 \int_{R \in \mathtt{HPD}(\mathrm{x}_n, n, l, p)} p_R(R|\,\mathrm{x}_n, p)\, p_p(p|\,\mathrm{x}_n)\, dR\, dp \right\} \geq 1 - \alpha, \quad (4.2.5)$$

or

$$\inf_{\mathrm{x}_n \in \mathcal{X}_n} \left(\int_0^1 \left\{ \int_{R \in \mathtt{HPD}(\mathrm{x}_n, n, l, p)} p_R(R|\,\mathrm{x}_n, p)\, dR \right\}^k p_p(p|\,\mathrm{x}_n)\ dp \right)^{1/k} \geq 1 - \alpha. \quad (4.2.6)$$

Notice that here the averaging is carried out over the nuisance parameter p only after the HPD lengths or coverages have been computed. Similar extensions can be developed for the **MCOC** and **MLOC**.

We discuss below how the sample sizes based on the $\mathbf{ALC}_k$ in equations 4.2.1 and 4.2.2 can be applied to cohort and case-control studies, as well as how they are related to one sample problems that were solved exactly, approximately, or through Monte Carlo simulation in chapter 3.

4.2.1 Cohort studies

Here we use the prior-likelihood model from 4.1.1.1. Recall that we have shown that the pre-posterior predictive distribution in cohort studies is

$$\begin{aligned} p(a, b|m_1, m_0, a', b', c', d') &= \binom{m_1}{a} \frac{\mathbf{B}(a'', c'')}{\mathbf{B}(a', c')} \times \binom{m_0}{b} \frac{\mathbf{B}(b'', d'')}{\mathbf{B}(b', d')} \\ &= p_a(a|m_1, a', c') \times p_b(b|m_0, b', d'), \end{aligned}$$

where $a = 0, \cdots, m_1$ and $b = 0, \cdots, m_0$ and $m_0 = (g+1)m_1$, as the product of two independent Beta-Binomial distributions denoted simply by p_a and p_b.

4.2.1.1 Application of equation (4.2.1)

Notation 4.2.1.

- *By $alc_k(m_1, m_0, a', b', c', d')$ or simply alc_k, we mean the criterion function associated with $\boldsymbol{ALC_k}$ in the two sample problem while by $alc'_k(p_1, m_1, a', c')$ or simply alc'_k, we mean the criterion function associated $\boldsymbol{ALC_k}$ in the one sample problem for estimating the single proportion p_1 (see chapter 3).*

- *The notation* HPD(R, a'', c'', m_1, p_0) *represents the* HPD *intervals of coverage $1-\alpha$ for $R = \dfrac{p_1}{p_0}$ given p_0, i.e* HPD *intervals based on the posterior distribution*

$$p_R(R) = \frac{p_0^{a''}\, R^{a''-1}\, (1 - Rp_0)^{c''-1}}{B(a'',\, c'')}, \qquad 0 < R < \frac{1}{p_0}.$$

We have shown that the length of HPD(R, a'', c'', m_1, p_0) *is equal to the length of* HPD(p_1, a'', c'', m_1) *divided by p_0, as a straightforward application of Theorem 3.2.1. This result will be used in the derivation of equation* (4.2.7) *below.*

- *The posterior distributions of p_1 and p_0 are: $p_1 \sim \boldsymbol{Be}(a'', c'')$ and $p_0 \sim \boldsymbol{Be}(b'', d'')$.*

When $b' > 1$, we have

$$\begin{aligned}
\mathrm{alc}_k^k &= \sum_{a=0}^{m_1}\sum_{b=0}^{m_0}\left\{\int_0^1\int_{R\in\mathrm{HPD}(R,a'',c'',m_1;p_0)} p_{p_0}(p_0|\, m_0, b'', d'')\, dR\, dp_0\right\}^k p_a\, p_b \\
&= \sum_{a=0}^{m_1}\sum_{b=0}^{m_0}\left\{\int_0^1\int_{p_1\in\mathrm{HPD}(p_1,a'',c'',m_1)} \frac{p_{p_0}(p_0|\, m_0, b'', d'')}{p_0}\, dp_1\, dp_0\right\}^k p_a\, p_b \\
&= \left[\sum_{a=0}^{m_1}\left\{\int_{R\in\mathrm{HPD}(p_1,a'',c'',m_1)} dp_1\right\}^k p_a\right] \times \sum_{b=0}^{m_0}\left\{\frac{\mathrm{B}(b''-1,\, d'')}{\mathrm{B}(b'',\, d'')}\right\}^k p_b \\
&= \left[\sum_{b=0}^{m_0}\left\{\frac{\mathrm{B}(b''-1,\, d'')}{\mathrm{B}(b'',\, d'')}\right\}^k p_b\right] \times \mathrm{alc}_k'^k(p_1, m_1, a', c'). \qquad (4.2.7)
\end{aligned}$$

Thus

$$\mathrm{alc}_k(m_1, m_0, a', b', c', d') = \left[\sum_{b=0}^{m_0}\left\{\frac{\mathrm{B}(b''-1,\, d'')}{\mathrm{B}(b'',\, d'')}\right\}^k p_b\right]^{1/k} \times \mathrm{alc}_k'(m_1, a', c'),$$

which shows that the two sample problem is related to the one sample problem in a multiplicative manner. This relation seems to be the result of the assumed independence between exposure and non-exposure groups. Thus, the derivation of the sample sizes for the criterion in equation (4.2.1) requires only a slight modification of the programs used in the one sample problem for estimating p_1 (see chapter 3).

4.2.2 Application of equation (4.2.2)

Using the same steps used to derive equation (4.2.7), one obtains

$$\mathrm{alc}_k(m_1, m_0, a', b', c', d') = \left\{\frac{\mathrm{B}(b'-k,\, d')}{\mathrm{B}(b',\, d')}\right\}^{1/k} \times \mathrm{alc}_k'(m_1, a', c'), \qquad b' > k.$$

In other words, if $m_1(R, a', b', c', d')$ and $n_{p_1}(, a', c')$ are the required sample sizes for the exposed group when estimating R in the two sample problem and p_1 in the one sample problem, then we have

$$m_1(R, a', b', c', d') = \left\{\frac{\mathrm{B}(b'-k,\, d')}{\mathrm{B}(b',\, d')}\right\}^{2/k} n_{p_1}(a', c'), \quad b' > k \qquad (4.2.8)$$

irrespective of g and $m_0 = (g+1)m_1$. As a consequence, everything done in the one sample problem applies again. One such application leads to the approximate sample size formula

$$n = \left\{\frac{\mathrm{B}(b'-k,\, d')}{\mathrm{B}(b',\, d')}\right\}^{2/k} \left\{4\frac{z^2_{1-\alpha/2}}{l^2}\left(\frac{\mathrm{B}(a'+k/2, c'+k/2)}{\mathrm{B}(a', c')}\right)^{2/k} - a' - c'\right\}, \quad b' > k. \tag{4.2.9}$$

4.2.3 Case-control studies

4.2.3.1 Application of equation (4.2.1)

It is easily seen that

$$\mathrm{alc}_k(n_1, n_0, a', b', c', d') = \left[\sum_{c=0}^{m_0}\left\{\frac{\mathrm{B}(c''-1,\, d''+1)}{\mathrm{B}(b'',\, d'')}\right\}^k p_c\right]^{1/k} \times \mathrm{alc}'_k(n_1, a', b'),$$

for estimating the exposure odds ratio ψ_e, and

$$\mathrm{alc}_k(n_1, n_0, a', b', c', d') = \mathrm{alc}'_k(n_1, a', c'),$$

for estimating $\phi_e = \log(\psi_e)$, irrespective of n_0, b', d'. Sample size methods for ψ_e in case-control studies are therefore easily derived by a slight modification of those relating to $\mathbf{ALC}'_k$.

4.2.3.2 Application of equation (4.2.2)

It is easily seen that

$$n_1(\psi_e, a', b', c', d') = \left\{\frac{\mathrm{B}(c'-k,\, d'+k)}{\mathrm{B}(c',\, d')}\right\}^{2/k} n_\psi(a', b'), \quad b' > k, \tag{4.2.10}$$

irrespective of g, where $n_\psi(a', b')$ is the sample size based on the $\mathbf{ALC}_k$ for estimating $\psi = \dfrac{p_1}{1-p_1}$. For ϕ_e,

$$n_1(\phi_e, a', b', c', d') = n_\phi(a', b'), \tag{4.2.11}$$

with $\phi = \log(\psi)$. As a consequence, everything done in the one sample problem applies here as well.

There remains much work to do on the comparison between the methods used.

4.3 Summary

In this chapter, we address the sample size problem for inference about the main parameters of cohort, case-control, cross-sectional and pair-matched studies. In addition, in cohort and case-control studies, we extend our results to the problem of minimizing cost. We consider several criteria consisting of $\mathbf{ALC}_k$, $\mathbf{ACC}_k$, **WLOC**, **MLOC** and **MCOC**, and numerous prior/likelihood models including the restricted model. We derive sample size formulae for the $\mathbf{ALC}_k$ in all four designs when we make no specific restrictions of the parameters under investigation. We present numerous algorithms to simulate random variables from the posterior distributions. We also describe various ways to approximate posterior densities, an important step to estimate an HPD interval. Our algorithms for simulating observations from the posterior distribution of R and ψ_e are fast and efficient because they rely on simulating observations from a finite mixture of beta densities. Finally we give numerous sketches of the Monte Carlo approach to sample size computation. User-friendly software is being developed for most of the sample size methods discussed in this thesis.

Chapter 5

Conclusion

In this concluding chapter, we summarize the contributions in this thesis to the Bayesian sample size problem and discuss some topics for future research.

5.1 Contributions

In this thesis, we examine the problem of Bayesian sample size determination for estimating key parameters in 2 × 2 tables arising from four major epidemiological designs: cohort, case-control, cross-sectional and matched pairs. Our Bayesian approach to these sample size problems, is, we believe, the first attempted.

We mainly examine what we believe are practical Bayesian sample size criteria: $\mathbf{ALC}_k$, $\mathbf{ACC}_k$, **WOC**, **MLOC**, and **MCOC**

We summarize below some key results for Bayesian sample size determination when estimating odds ratio using HPD intervals in case-control studies under the $\mathbf{ALC}_k$. This summary, we hope conveys the flavour of this thesis.

We begin with the single sample problem, point out that this leads to an analysis of sample size problems for case-series or case-only studies, and then extend the methods to more general two-sample problems.

In the one sample problem, we present three solutions: an exact solution, a solution based on a third order approximation to HPD interval lengths, and a Monte Carlo-based solution. The methods lead to slightly different sample sizes, often undistinguishable when the prior parameters are large. Two sample size problems are mostly solved via a Monte Carlo simulation approach, since exact solutions are computationally inefficient while third order approximations, when the posterior distribution depends on nuisance parameters, are yet to be obtained. In both the one sample and two sample problems, we also discuss a regression-based approach used to reduce large variability very typical of Monte Carlo approaches to sample size computation. We show here that the criterion function of the $\mathbf{ALC}_k$ often converges to zero at the rate of $\frac{1}{\sqrt{n}}$. We derive sample size formulae despite the absence of a closed form formula for HPD interval lengths. Since no tractable forms for the criterion functions can be derived for the two sample problem, we base our analysis on an asymptotic form for the pre-posterior predictive distribution. We derive this asymptotic distribution in this thesis.

We present novel results on sample size calculations where the odds ratio, ψ_e, is restricted a priori, to $\psi_e > 1$ or $\psi_e < 1$.

Although in the two sample problem we present sample size formulae for the $\mathbf{ALC}_k$, some caution should be exercised if one of the prior parameters is small, say less than 10. We suggest that Monte Carlo techniques be used in this case. Using both the sample formula and Monte Carlo appraoches whenever possible is strongly advised since they are complementary. A large difference between these approaches indicates that we should employ the regression-based approach using various combinations of $m \geq 500$ and $M \geq 500$. This approach, in turn, often requires some form of monitoring of the final sample size until there is rough agreement for different m and M values.

We extend the sample size problem to that of minimizing cost in the two sample problem for cohort and case-control studies. Although our cost function is rather simple, it is intuitive, and similar to the cost function used in many frequentist analyses. We also discuss the important question of the optimality of the ratio of controls per case.

In the two sample problem, we address the computational challenges that occur in the computation of the posterior distribution when it depends on a nuisance parameter, and, we describe various algorithms to simulate observations from the posterior distribution. For all designs and all prior/likelihood models used here, we prove the unimodality of the posterior distribution, crucial to many of our algorithms and derivations.

We provide numerous illustrative sample size tables.

Many of the diverse results obtained in this thesis apply to other settings. For example, some sample size results for cohort studies apply to randomized trials which can be viewed as cohort studies. We show how our sample size derivation techniques for cross-sectional studies can be used to find sample size for the $2 \times 2 \times 2$ tables arising from case-control studies for estimating gene-exposure interaction, under independence between genotype and exposure. Such $2 \times 2 \times 2$ designs are currently an active research area in pharmacogenetic and pharmacoepidemiology (Ashby et al., 1998 and Weiss et al., 2001).

Although in many cases, Bayesian and frequentist approaches produce sample size formulae that are similar, Bayesian approaches have the advantage of allowing one to incorporate restrictions on the parameters of interest through the prior distribution, and fully account for all uncertainty inherent in the problem.

A user-friendly Windows-based Bayesian software program for sample size calculations that implements all the methods and criteria along the various designs discussed in chapters 3 and 4 is currently under construction. All programs have been written in Visual Fortran 6.1 (Compaq).

5.2 Future research projects

There are several avenues of research that this thesis have opened up. Some of these are listed here, recognizing that some of these are projects rather than new research ideas.

- **Project 1:** We propose to produce user-friendly software for Bayesian sample calculation for both the one sample problem (case-series and exposure-series) and two sample problems (cohort, case-control, cross-sectional, and matched studies). Although in this thesis we do not discuss the cases where one knows that $R > 1$ or $R < 1$ in a cross-sectional analysis, these cases will be covered by our software. Similarly for $\psi_e > 1$ and $\psi_e < 1$ in the matched and cross-sectional studies. (See Appendix F for a description of the prior/likelihood model used).

- **Project 2:** We propose to develope results on S-optimal criteria (see Appendix H for a brief description of these criteria) and connect these results to certain of those obtained when using Bayesian criteria based on HPD regions.

- **Project 3:** In Appendix G, we state a conjecture that is used to derive Bayesian sample size formulae in more general settings than those developed here. We plan to prove this conjecture.

- **Project 4:** It would be important to derive optimal sample sizes for $R * K$ tables in case-control and cohort settings including 2×4 case-control studies for estimating gene-environment interactions when the independence assumption between susceptibility genotype and exposure is violated.

- **Project 5:** In practice, most case-control and cohort studies incorporate covariates in their analysis. Therefore, it is very important to extend our Bayesian sample size methods to allow for covariate modelling.

Appendix A

Third Order Approximations

In this appendix, we discuss third order approximations to the left and right tails of HPD and equal-tailed intervals for p_1, ω, and ϕ.

Proportion

The essence of Theorem A.0.1 is that by introducing a vague prior/likelihood combination, we obtain the same posterior distribution as for the $\mathbf{Be}(a,\, b)/\mathbf{Bin}(n,\, p_1)$ combination.

Theorem A.0.1. *Under the assumptions that $p_1 \sim \boldsymbol{Be}(0,\, 0)$ and $\mathrm{x}_n|p_1 \sim \boldsymbol{Bin}(n + a + b,\, p_1)$, the posterior density of p_1 is given by the expression in equation* (3.2.9) *when $a \leq \mathrm{x}_n \leq a + n$, respectively.*

The proof is straightforward.

The important point here is that although this new prior-likelihood combination yields the same posterior as obtained under the original model $p_1 \sim \mathbf{Be}(a, b)$ and $\mathrm{x}_n|p_1 \sim \mathbf{Bin}(n, p_1)$, the difficulties referred to before are alleviated since all troublesome terms are now well defined.

The Corollary below to Theorem A.0.1 provides alternatives to the expressions in subsection 2.7.3 that are more amenable to our purposes.

Corollary A.0.2. *Under the prior/likelihood model,* $p_1 \sim \boldsymbol{Be}(0,\,0)$ *and* $\mathrm{x}_n|\,p_1 \sim \boldsymbol{Bin}(n+a+b,\,p)$*, we have*

$$\{-Nw_2\}^{1/2} = N\sqrt{v_1(\mathrm{x}_n)}, \quad \frac{1}{2}w_3(-w_2)^{-3/2} = \left\{\frac{v_2(\mathrm{x}_n)-2}{N}\right\}^{1/2},$$
$$\frac{1}{6}w_4(-w_2)^{-2} = -\frac{v_2(\mathrm{x}_n)-1}{N}, \quad \frac{1}{2}\{-Nw_2\}^{-1}\frac{d^2\log p(\widehat{\theta})}{d\widehat{\theta}^2} = \frac{v_2(\mathrm{x}_n)}{2N},$$
$$\{-Nw_2\}^{-1/2}\frac{d\log p(\widehat{\theta})}{d\widehat{\theta}} = -\left\{\frac{v_2(\mathrm{x}_n)-2}{N}\right\}^{1/2},$$
$$\frac{1}{2}N^{-1/2}w_3(-w_2)^{-2}\frac{d\log p(\widehat{\theta})}{d\widehat{\theta}} = -\frac{v_2(\mathrm{x}_n)-2}{N}, \tag{A.0.1}$$

where $N = n{+}a{+}b$, $v_1(\mathrm{x}_n) = \dfrac{1}{\mathrm{x}_n+a} + \dfrac{1}{n+b-\mathrm{x}_n}$ *and* $v_2(\mathrm{x}_n) = \dfrac{n+b-\mathrm{x}_n}{\mathrm{x}_n+a} + \dfrac{\mathrm{x}_n+a}{n+b-\mathrm{x}_n}$, $\mathrm{x}_n = 0, 1, \cdots, n.$

Odds

Theorem A.0.3. *Under the assumptions that* $p_1 \sim \boldsymbol{Be}(0,\,0)$ *and* $\mathrm{x}_n|p_1 \sim \boldsymbol{Bin}(n+a+b,\,p_1)$*, the posterior density of* ω *is given by the expression in equation* (3.2.10) *when* $a \le \mathrm{x}_n \le a+n$*, respectively.*

Corollary A.0.4. *Under the prior/likelihood model,* $p_1 \sim \boldsymbol{Be}(0,\,0)$ *and* $\mathrm{x}_n|\,p_1 \sim \boldsymbol{Bin}(n+a+b,\,p)$*, we have*

$$\{-Nw_2\}^{-1/2} = \left\{\frac{N(a+\mathrm{x}_n)}{(n+b-\mathrm{x}_n)^3}\right\}^{1/2} = v_3(\mathrm{x}_n), \quad \frac{1}{2}w_3(-w_2)^{-3/2} = \left\{\frac{v_2(\mathrm{x}_n)+2}{n+b-\mathrm{x}_n}\right\}^{1/2},$$
$$\frac{1}{6}w_4(-w_2)^{-2} = -\frac{v_2(\mathrm{x}_n)+1}{n+b-\mathrm{x}_n}, \quad \frac{1}{2}\{-Nw_2\}^{-1}\frac{d^2\log p(\widehat{\theta})}{d\widehat{\theta}^2} = \frac{1}{2}v_1(\mathrm{x}_n),$$
$$\{-Nw_2\}^{-1/2}\frac{d\log p(\widehat{\theta})}{d\widehat{\theta}} = -\sqrt{v_1(\mathrm{x}_n)},$$
$$\frac{1}{2}N^{-1/2}w_3(-w_2)^{-2}\frac{d\log p(\widehat{\theta})}{d\widehat{\theta}} = -v_1(\mathrm{x}_n) - \frac{1}{n+b-\mathrm{x}_n}, \tag{A.0.2}$$

where $N = n + a + b$, $v_1(\mathbf{x}_n) = \dfrac{1}{\mathbf{x}_n + a} + \dfrac{1}{n + b - \mathbf{x}_n}$ *and* $v_2(\mathbf{x}_n) = \dfrac{N}{\mathbf{x}_n + a} + \dfrac{\mathbf{x}_n + a}{N}$, $\mathbf{x}_n = 0, 1, \cdots, n$.

Log-odds

Theorem A.0.5. *Under the assumptions that* $p_1 \sim \boldsymbol{Be}(0, 0)$ *and* $\mathbf{x}_n | p_1 \sim \boldsymbol{Bin}(n + a + b, p_1)$*, the posterior density of* ω *is given by the expression in equation* (3.2.10) *when* $a \leq \mathbf{x}_n \leq a + n$.

Corollary A.0.6.

$$\{-Nw_2\}^{-1/2} = \sqrt{v_1(\mathbf{x}_n)}, \quad w_3^2(-w_2)^{-3} = \frac{v_2(\mathbf{x}_n) - 2}{N},$$
$$w_4(-w_2)^{-2} = -\frac{v_2(\mathbf{x}_n) - 4}{N}, \quad \frac{d \log p(\widehat{\theta})}{d\widehat{\theta}} = \frac{d^2 \log p(\widehat{\theta})}{d\widehat{\theta}^2} = 0,$$

where $N = n+a+b$, $v_1(\mathbf{x}_n) = \dfrac{1}{\mathbf{x}_n + a} + \dfrac{1}{n + b - \mathbf{x}_n}$ *and* $v_2(\mathbf{x}_n) = \dfrac{n + b - \mathbf{x}_n}{\mathbf{x}_n + a} + \dfrac{\mathbf{x}_n + a}{n + b - \mathbf{x}_n}$.

Appendix B

Corollary to Theorems 3.2.6

The corollary below which is a corollary to Theorem 3.2.6 is used in subsection 3.2.4.3 to derive approximate $\mathbf{ALC}_k$ sample sizes for ω.

Corollary B.0.7. *For $b > 3k/2$ and $k \geq 2$,*

$$\lim_n \sqrt{n+a+b}\Big\{ \boldsymbol{E}_{X_n}[l_\omega^k(X_n, n, a, b)]\Big\}^{1/k} = 2z_{1-\alpha/2}\left\{\frac{B(a+k/2, b-3k/2)}{B(a,b)}\right\}^{1/k}.$$

Proof. Let $Y_n = \dfrac{X_n}{n}$, $\mathcal{F}_n = \{0, \frac{1}{n}, \cdots, \frac{n-1}{n}, 1\}$, k an even number, and $h(y) = y^{k/2}$. We have

$$\begin{aligned}
N^{k/2}\frac{\mathbb{E}_{X_n}[l_\omega^k(X_n, n, a, b)]}{2^k z_{1-\alpha/2}^k} &= \sum_{y\in\mathcal{F}_n} \frac{N^{k/2}(N-1)^{k/2}(a+ny)^{k/2}}{(n+b-ny-1)^k(n+b-ny-2)^{k/2}}\, p_{Y_n}(y|n,a,b) \\
&= c_n \frac{\mathbf{B}(a, b-3k/2)}{\mathbf{B}(a,b)} \sum_{y\in\mathcal{F}_n} h_n(y)\, p_{Y_n}(y|n, a, b-3k/2), \\
&= c_n \frac{\mathbf{B}(a, b-3k/2)}{\mathbf{B}(a,b)} E[h_n(Z_n)],
\end{aligned}$$

where $nZ_n \sim \mathcal{BB}(n, a, b{-}3k/2)$, $c_n = \dfrac{(n+a+b-1)^{k/2}N^{k/2}n^{k/2}\Gamma(n+a+b-3k/2)}{\Gamma(n+a+b)}$ with $\lim_n c_n = 1$, $h_n(y) = \left(\dfrac{a}{n}+y\right)^{k/2} \dfrac{\prod_{i=1}^{3k/2}\left(\frac{b-i}{n}+1-y\right)}{\left(\frac{b-1}{n}+1-y\right)^k\left(\frac{b-2}{n}+1-y\right)^{k/2}}$. Note that the decomposition of $p_{Y_n}(y|n,a,b)$ in terms of $p_{Y_n}(y|n,a,b-3k/2)$ is just a technique to handle the indeterminacy due to the denominator of the $k/2$-th power of the posterior variances. The fact that we need $3k/2$ terms can be seen in $\lim_n n^{k/2}\Big\{\mathbf{Var}_\omega(X_n,n,a,b)\Big\}^{k/2} = \dfrac{x^{k/2}}{(1-x)^{3k/2}}$. We have $\|g_n(y)\| \leq (a+1)^{k/2}$, $\lim_n h_n(y) = h(y)$, and $Z_n \longrightarrow^d \mathbf{Be}(a,b-3k/2)$. An application of Theorems 3.2.7 and 3.2.7 implies that

$$\begin{aligned}\lim_n E[h_n(Z_n)] &= \int_0^1 h(y)\frac{y^{a-1}(1-y)^{b-3k/2-1}}{\mathbf{B}(a,b-3k/2)}\,dy\\ &= \frac{\mathbf{B}(a+k/2,b-3k/2)}{\mathbf{B}(a,b-3k/2)}.\end{aligned}$$

Thus

$$\begin{aligned}\lim_n n^{k/2}\frac{\mathbf{E}_{X_n}[l_\omega^k(X_n,n,a,b)]}{2^k z_{1-\alpha/2}^k} &= \frac{\mathbf{B}(a,b-3k/2)}{\mathbf{B}(a,b)}\frac{\mathbf{B}(a+k/2,b-3k/2)}{\mathbf{B}(a,b-3k/2)},\\ &= \frac{\mathbf{B}(a+k/2,b-3k/2)}{\mathbf{B}(a,b)}.\end{aligned}$$

Similarly for k odds where

$$h_n(y) = \left(\frac{a}{n}+y\right)^{k/2}\frac{\prod_{i=1}^{3k/2+1/2}\left(\frac{b-i}{n}+1-y\right)}{\left(\frac{b-1}{n}+1-y\right)^k\left(\frac{b-2}{n}+1-y\right)^{k/2}}$$

and $h(y) = y^k(1-y)^{1/2}$. □

Appendix C

Higher Order Terms

It is often important to include higher order terms in the expansion of the criterion function $\mathrm{alc}_k(n, a, b)$, $k = 1, 2$. This can be done using the limiting result $\frac{\mathbf{X_n}}{\mathbf{n}} \rightarrow^{\mathbf{d}} \mathbf{p}$, where $p \sim \mathbf{Be}(a, b)$.

C.1 Proportions

We have

$$\begin{aligned} \mathbf{E}_{X_n}\left[\mathbf{Var}_{p_1}(X_n, n, a, b)\right] &\simeq \int_0^1 \mathbf{Var}_p(np, n, a, b)\, \frac{p^{a-1}(1-p)^{b-1}}{\mathbf{B}(a, b)}\, dp \\ &= \int_0^1 \frac{(a+np)(n+b-np)}{N^2(N+1)}\, \frac{p^{a-1}(1-p)^{b-1}}{\mathbf{B}(a, b)}\, dp \\ &= \frac{c_{p_1}(a, b)}{N} + \frac{nc_{p_1}(a, b)}{N^2(N+1)}. \qquad \text{(C.1.1)} \end{aligned}$$

The major difference between equations (3.2.25) and (C.1.1) is that the discrete points $\mathbf{x}_n$ are replaced by continuous points of the type np, where $0 < p < 1$.

We have

$$\begin{aligned}\mathbf{E}_{X_n}[l_{p_1}(X_n,n,a,b)] &\simeq 2z_{1-\alpha/2}\int_0^1 \sqrt{\frac{(a+np)(b+n(1-p))}{N^2(N+1)}}\,\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp\\ &= c_n\int_0^1 \sqrt{\left(p+\frac{a}{n}\right)\left(1-p+\frac{b}{n}\right)}\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp,\end{aligned} \tag{C.1.2}$$

where $c_n = \dfrac{2nz_{1-\alpha/2}}{N\sqrt{N+1}}$. It is easily seen that

$$\left(p+\frac{a}{n}\right)\left(1-p+\frac{b}{n}\right) = p(1-p) + \frac{1}{n}\left\{a(1-p)+bp+\frac{ab}{n}\right\}$$

which, together with

$$\sqrt{a+x} = a^{1/2} + \frac{1}{2}\,a^{-1/2}\,\frac{x}{1!} - \frac{1}{4}\,a^{-3/2}\,\frac{x^2}{2!} + \cdots, \tag{C.1.3}$$

yields

$$\begin{aligned}\sqrt{\left(p+\frac{a}{n}\right)\left(1-p+\frac{b}{n}\right)} &\simeq p^{1/2}(1-p)^{1/2} + \frac{a}{2n}p^{-1/2}(1-p)^{1/2} +\\ &\quad \frac{b}{2n}p^{1/2}(1-p)^{-1/2} + \frac{ab}{2n^2}p^{-1/2}(1-p)^{-1/2} -\\ &\quad \frac{a^2}{8n^2}p^{-3/2}(1-p)^{1/2} - \frac{b^2}{8n^2}p^{1/2}(1-p)^{-3/2}.\end{aligned} \tag{C.1.4}$$

Using equation (C.1.4) together with the integral (C.1.2) gives

$$\begin{aligned}\frac{\mathbf{B}(a,b)}{c_n}\mathbf{E}_{X_n}[l_p(X_n,n,a,b)] &\simeq \mathbf{B}\left(a+\frac{1}{2},b+\frac{1}{2}\right) + \frac{a}{2n}\mathbf{B}\left(a-\frac{1}{2},b+\frac{1}{2}\right) +\\ &\quad \frac{b}{2n}\mathbf{B}\left(a+\frac{1}{2},b-\frac{1}{2}\right) + \frac{ab}{2n^2}\mathbf{B}\left(a-\frac{1}{2},b-\frac{1}{2}\right)\\ &\quad -\frac{a^2}{8n^2}\mathbf{B}\left(a-\frac{3}{2},b+\frac{1}{2}\right) - \frac{b^2}{8n^2}\mathbf{B}\left(a+\frac{1}{2},b-\frac{3}{2}\right).\end{aligned} \tag{C.1.5}$$

C.2 Odds

We have

$$\begin{aligned}
\mathbf{E}_{X_n}\left[\mathbf{Var}_\omega(X_n,n,a,b)\right] &\simeq \int_0^1 \mathbf{Var}_\omega(np,n,a,b)\,\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp,\\
&\simeq (N-1)\int_0^1 \frac{(a+np)}{(n+b-1-np)^3}\,\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp,\\
&\simeq \frac{n(N-1)}{(n+b-1)^3}\int_0^1 \frac{a_n+p}{(1-p)^3}\,\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp,\\
&= \frac{n(N-1)}{(n+b-1)^3}\left\{\frac{a_n\mathbf{Be}(a,b-3)}{\mathbf{Be}(a,b)}+\frac{\mathbf{Be}(a+1,b-3)}{\mathbf{Be}(a,b)}\right\},\\
&\simeq \frac{c_\omega(a,b)}{n+b-1}+2(a-1)\frac{c_\omega(a,b)}{(n+b-1)^2}, \qquad \text{(C.2.1)}
\end{aligned}$$

where $a_n=\dfrac{a}{n}$.

As far as the $E_{X_n}[l_\omega(X_n,n,a,b)]$ is concerned, we have

$$\begin{aligned}
\mathbf{E}_{X_n}[l_\omega(X_n,n,a,b)] &\simeq 2z_{1-\alpha/2}\sqrt{N-1}\int_0^1\sqrt{\frac{a+np}{(n+b-1-np)^3}}\,\frac{p^{a-1}\,(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp,\\
&\simeq 2z_{1-\alpha/2}\frac{\sqrt{n(N-1)}}{(n+b-1)^{3/2}}\int_0^1\sqrt{\frac{p}{(1-p)^3}+\frac{a_n}{(1-p)^3}}\,\frac{p^{a-1}\,(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp.
\end{aligned}$$
$$\text{(C.2.2)}$$

Using equation (C.1.3), we have

$$\sqrt{\frac{p}{(1-p)^3}+\frac{a_n}{(1-p)^3}}\simeq\frac{p^{1/2}}{(1-p)^{3/2}}+\frac{a_n}{2p^{1/2}(1-p)^{3/2}}-\frac{a_n^2}{8p^{3/2}(1-p)^{3/2}}.$$
$$\text{(C.2.3)}$$

Therefore,

$$\begin{aligned}
\mathbf{E}_{X_n}[l_\omega(X_n,n,a,b)] &\simeq 2z_{1-\alpha/2}\sqrt{\frac{n(N-1)}{(n+b-1)^3}}\left\{\frac{\mathbf{Be}(a+1/2,b-3/2)}{\mathbf{Be}(a,b)}+\right.\\
&\left.\frac{\mathbf{Be}(a-1/2,b-3/2)}{\mathbf{Be}(a,b)}\,\frac{a}{2n}-\frac{\mathbf{Be}(a-3/2,b-3/2)}{\mathbf{Be}(a,b)}\,\frac{a^2}{8n^2}\right\},
\end{aligned}$$
$$\text{(C.2.4)}$$

C.3 Log-odds

We have

$$\frac{1}{(a+np-1)} + \frac{1}{(n+b-np-1)} \simeq \frac{1}{np(1-p)} - \frac{1}{n^2}\left\{\frac{a-1}{p^2} + \frac{b-1}{(1-p)^2}\right\}.$$

Therefore,

$$\begin{aligned}
\mathbf{E}_{X_n}\left[\mathbf{Var}_\phi(X_n, n, a, b)\right] &= \int_0^1 \frac{N-2}{(a+np-1)(n+b-np-1)} \frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dp \\
&\simeq \frac{1}{n}\frac{\mathbf{Be}(a-1,b-1)}{\mathbf{Be}(a,b)} - \frac{1}{n^2}\frac{(a-1)\mathbf{Be}(a-2,b)}{\mathbf{Be}(a,b)} - \\
&\quad \frac{1}{n^2}\frac{(b-1)\mathbf{Be}(a,b-2)}{\mathbf{Be}(a,b)}, \\
&= \frac{c_\phi}{n} - \frac{1}{n^2}\frac{(a+b-1)(a+b-2)(a+b-4)}{(a-2)(b-2)}. \qquad \text{(C.3.1)}
\end{aligned}$$

As to $E_{X_n}[l_\phi(\mathbf{x}_n, n, a, b)]$, note that

$$\begin{aligned}
\sqrt{\frac{1}{(a+np-1)} + \frac{1}{(n+b-np-1)}} &\simeq \frac{1}{n^{1/2}p^{1/2}(1-p)^{1/2}} - \\
&\quad \frac{1}{2n^{3/2}}\left[\frac{(a-1)(1-p)^{1/2}}{p^{3/2}} + \frac{(b-1)p^{1/2}}{(1-p)^{3/2}}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbf{E}_{X_n}[l_\phi(X_n, n, a, b)] &= 2z_{1-\alpha/2}\int_0^1 \sqrt{\frac{N-2}{(a+np-1)(n+b-np-1)}}\,\frac{p^{a-1}(1-p)^{b-1}}{\mathbf{Be}(a,b)}\,dx \\
&\simeq 2z_{1-\alpha/2}\frac{1}{n^{1/2}}\left\{\frac{\mathbf{Be}(a-1/2,b-1/2)}{\mathbf{Be}(a,b)} - \frac{a-1}{n}\frac{\mathbf{Be}(a-3/2,b+1/2)}{\mathbf{Be}(a,b)}\right. \\
&\quad \left. -\frac{b-1}{n}\frac{\mathbf{Be}(a+1/2,b-3/2)}{\mathbf{Be}(a,b)}\right\}. \qquad \text{(C.3.2)}
\end{aligned}$$

Appendix D

The MLOC

In this appendix, we derive approximate sample size formula for the **MLOC**. Define where $z = z_{1-\alpha/2}$ and $N = n + a + b$.

D.1 The Odds, ψ

We have shown in subsection 3.2.4.3 that the posterior variance of the odds ψ, $\mathbf{Var}_\psi(x) = \dfrac{(a+x)(N-1)}{(n+b-x-1)^2(n+b-x-2)}$, is an increasing function of x. Therefore, the median of the posterior variances over $x = 0, \ldots, n$ is attained at $x_\psi = n/2$ with $\mathbf{Var}_\psi(x_\psi) = 4\dfrac{(2a+n)(N-1)}{(n+2b-2)^2(n+2b-4)}$. The **MLOC** solve approximately the equation $\mathbf{Var}_\psi(x_\psi) = \text{med}_{0\leq x\leq n}\mathbf{Var}_\psi(x) = \dfrac{l^2}{z^2}$. Solving this system and expanding the solution as a Taylor series of order 4 with Maples yields the approximate sample size

$$n_\psi = 16\frac{z^2}{l^2} + 7 - 5b + 3a + o\left(\frac{l^2}{z^2}\right). \tag{D.1.1}$$

D.2 The Proportion, p_1

The posterior variance of p_1 is $\mathbf{Var}_{p_1}(x) = \dfrac{(a+x)(n+b-x)}{N^2(N+1)}$. The median of these posterior variances over $x = 0, \ldots, n$ is attained approximately at $x_{p_1} = \dfrac{n+2(b-a)}{2}$ with $\mathbf{Var}_{p_1}(x_{p_1}) = \dfrac{(N+a+b)(N-a-b)}{16N^2(N+1)}$. We solve the system $\mathbf{Var}_{p_1}(x_{p_1}) = \text{med}_{0 \leq x \leq n} \mathbf{Var}_{p_1}(x) = \dfrac{l^2}{z^2}$ in n. An expansion of the solution n_{p_1} as a Taylor series yields the approximate sample size

$$n_{p_1} = \frac{3}{4}\frac{z^2}{l^2} - 1 - \frac{1}{3}(a+b) + o\left(\frac{l^2}{z^2}\right). \qquad \text{(D.2.1)}$$

Recall that in the frequentist context, $n_{\widehat{p}_1} = 4\hat{p}_1(1-\widehat{p}_1)\dfrac{z^2}{l^2}$. By setting $\widehat{p}_1 = \dfrac{1}{4}$, one has $n_{\widehat{p}_1} + \dfrac{1}{3}(a+b) \cong \dfrac{3}{4}\dfrac{z^2}{l^2}$, which shows again a strong link between Bayesian and frequentist sample size for proportions.

D.3 The log-odds ϕ

The posterior variance ϕ being $\mathbf{Var}_\phi(x) = \dfrac{(N-2)}{(a+x-1)(n+b-x-1)}$ implies that the median of these posterior variances over $x = 0, \ldots, n$ is attained again approximately at $x_\phi = \dfrac{n+2(b-a)}{2}$ with $\mathbf{Var}_\phi(x_\phi) = \dfrac{16(N-2)}{(N+a+b-4)(3N-a-b+4)}$. This leads to the approximate sample size

$$n_\phi = \frac{64}{3}\frac{z^2}{l^2} + \frac{10}{3} - \frac{5}{3}(a+b) + o\left(\frac{l^2}{z^2}\right). \qquad \text{(D.3.1)}$$

Appendix E

Variance of $\log(R)$

Derivation of the posterior variances of $\log(R)$**:** We have $\log(\mathbf{B}(a,c)) = \log(\Gamma(a)) + \log(\Gamma(c)) - \log(\Gamma(a+c))$. Therefore

$$\begin{aligned}\frac{\partial \log(\mathbf{B}(a,c))}{\partial a} &= \int_0^1 \log(u)\frac{u^{a-1}(1-u)^{c-1}}{\mathbf{B}(a,c)}\,du \\ &= \mathbf{E}[u] \;=\; \Psi(a)-\Psi(a+c)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 \log(\mathbf{B}(a,c))}{\partial a^2} &= \int_0^1 [\log(u)]^2\frac{u^{a-1}(1-u)^{c-1}}{\mathbf{B}(a,c)}\,du - \left\{\int_0^1 \log(u)\frac{u^{a-1}(1-u)^{c-1}}{\mathbf{B}(a,c)}\,du\right\}^2 \\ &= \mathbf{Var}[\log(u)] \;=\; \Psi'(a)-\Psi'(a+c)\end{aligned}$$

where $u \sim \mathbf{Be}(a,c)$ and $\Psi(x) = \dfrac{\Gamma'(x)}{\Gamma(x)}$. An expansion for the trigamma function $\Psi'(x)$ is given in subsection 3.2.4.3. Thus

$$\begin{aligned}\mathbf{Var}(\log(R)\,|\,\mathrm{T}'')) &= \mathbf{Var}(\log(p_1)\,|\,a'',c'') + \mathbf{Var}(\log(p_0)\,|\,b'',d'') \\ &= \Psi'(a'') - \Psi'(a''+c'') + \Psi'(b'') - \Psi'(b''+d'') \\ &= \Psi'(a'') + \Psi'(b'') - \Psi'(m_1+a'+c') - \Psi'(m_0+b'+d').\end{aligned}$$

Now set $a'' = m_1x$, $c'' = m_1(1-x) + c$, $b = gm_1y + b'$, $d'' = gm_1(1-y) + d'$. Then $\lim_{m_1 \longrightarrow \infty} m_1 \mathbf{Var}(\log(R)|\,\mathrm{T}'') = \frac{1}{xy}\left[\frac{x(1-y)}{g} + (1-x)y\right]$ since $\lim_{x \longrightarrow \infty} x\Psi'(x) = 1$.

Since the trigamma function is a decreasing function, $\max_{(a,b)} \Psi'(a'') + \Psi'(b'') = \Psi'(a') + \Psi'(b')$. Therefore, it does not make sense to apply the **WOC** to $\log(R)$.

Appendix F

Restricted Models for R

In this appendix, we first present a model that can be used to account for the case when the risk ratio, $R = \dfrac{p_{11}(p_{10}+p_{00})}{p_{10}(p_{11}+p_{01})}$, is restricted to either $R < 1$ or $R > 1$, in a cross-sectional analysis. A similar model can easily be derived when estimating the odds ratio $\psi = \dfrac{p_{11}p_{00}}{p_{10}p_{01}}$ in a cross-sectional analysis. We then present two other models that can be used for a restricted matched pair analysis.

F.1 Restricted model for estimating ψ_e in a cross-sectional analysis

Let p_{11}, p_{10}, p_{01}, and p_{00} be the cell probabilities of success in cross-sectional analysis as defined in Chapter 3. Let $\pi = p_{11} + p_{01}$, $p_1 = \dfrac{p_{11}}{\pi}$ and $p_0 = \dfrac{p_{10}}{1-\pi}$. In A standard cross-sectional analysis for estimating R, we assume a Dirichlet-Multinomial prior/likelihood model. Proposition 2.4.1 implies that the Dirichlet-Multinomial model is equivalent to:

$$\pi \sim \mathbf{Be}(a'+c',\, b'+d'), \qquad p_1 \sim \mathbf{Be}(a',\, c'), \qquad p_0 \sim \mathbf{Be}(b',\, d'),$$

$$m_1 \sim \mathbf{Bin}(N,\, \pi), \qquad a \sim \mathbf{Bin}(m_1,\, p_1), \quad \text{and} \quad b \sim \mathbf{Bin}(N-m_1,\, p_0).$$

We modify this second form of the prior/likelihood model to incorporate restrictions of the type $R < 1$ and $R > 1$. Clearly, for the restriction $R > 1$, the model below can be used

$$\pi \sim \mathbf{Be}(a'+c',\, b'+d'), \quad f(p_1, p_0) \propto p_1^{a'-1}(1-p_1)^{c'-1} p_0^{b'-1}(1-p_0)^{d'-1}, \quad p_1 > p_0,$$

$$m_1 \sim \mathbf{Bin}(N,\, \pi), \qquad a \sim \mathbf{Bin}(m_1,\, p_1), \qquad b \sim \mathbf{Bin}(N-m_1,\, p_0),$$

or

$$\pi \sim \mathbf{Be}(a'+c',\, b'+d'), \qquad p_1 \sim \mathbf{Be}(a',\, c'), \qquad p_0 \sim \mathbf{Be}(b',\, d',\, 0,\, p_1),$$

$$m_1 \sim \mathbf{Bin}(N,\, \pi), \qquad a \sim \mathbf{Bin}(m_1,\, p_1), \qquad b \sim \mathbf{Bin}(N-m_1,\, p_0),$$

where $R = \dfrac{p_{11}(p_{10}+p_{00})}{p_{10}(p_{11}+p_{01})} = \dfrac{p_1}{p_0}$. The resulting posterior distribution of ψ_e is the one discussed in subsection 4.1.2.2. Indeed, all the four prior models discussed in subsections 4.1.1.2 and 4.1.2.2 can be used. Similar models are used for the case $\psi_e < 1$.

F.2 Restricted model for estimating ψ_e in matched analysis

Let $p'_{11}, p'_{10}, p'_{01}$, and p'_{00} be the cell probabilities of success in a matched case-control analysis and suppose that the odds ratio, $\psi'_e = \dfrac{p'_{10}}{p'_{00}} < 1$, as defined in Chapter 3. For a matched analysis, the standard model is equivalent to

$$p'_{11}+p'_{00}\sim \mathbf{Be}(a'+d',\, b'+c'), \qquad \frac{p'_{10}}{p'_{11}+p'_{00}}\sim \mathbf{Be}(b',\, c')$$
$$b+c\sim \mathbf{Bin}(N,\, p'_{11}+p'_{00}),\quad b\sim \mathbf{Bin}\left(n_1,\, \frac{p'_{10}}{p'_{11}+p'_{00}}\right).$$

Therefore, the following models are used when $\psi'_e < 1$:

- **Model 1:**

$$p'_{11}+p'_{00}\sim \mathbf{Be}(a'+d',\, b'+c'), \quad \frac{p'_{10}}{p'_{11}+p'_{00}}\sim \mathbf{Be}(b',\, c',\, 0.0,\, 0.5,)$$
$$b+c\sim \mathbf{Bin}(N,\, p'_{11}+p'_{00}), \quad \text{and } b\sim \mathbf{Bin}\left(n_1,\, \frac{p'_{10}}{p'_{11}+p'_{00}}\right).$$

- **Model 2:**

$$p'_{11}+p'_{00}\sim \mathbf{Be}(a'+d',\, b'+c'), \quad \frac{p'_{10}}{p'_{11}+p'_{00}}\sim \text{IBeta}(b',\, c',\, 0.0,\, 0.5),$$
$$b+c\sim \mathbf{Bin}(N,\, p'_{11}+p'_{00}), \quad \text{and } b\sim \mathbf{Bin}\left(n_1,\, \frac{p'_{10}}{p'_{11}+p'_{00}}\right).$$

When $\psi'_e > 1$, one uses:

- **Model 3:**

$$p'_{11}+p'_{00}\sim \mathbf{Be}(a'+d',\, b'+c'), \quad \frac{p'_{10}}{p'_{11}+p'_{00}}\sim \mathbf{Be}(b',\, c',\, 0.5,\, 1.0),$$
$$b+c\sim \mathbf{Bin}(N,\, p'_{11}+p'_{00}), \quad \text{and } b\sim \mathbf{Bin}\left(n_1,\, \frac{p'_{10}}{p'_{11}+p'_{00}}\right).$$

- **Model 4:**

$$p'_{11}+p'_{00}\sim \mathbf{Be}(a'+d',\, b'+c'), \quad \frac{p'_{10}}{p'_{11}+p'_{00}}\sim \text{IBeta}(b',\, c',\, 0.5,\, 1.0),$$
$$b+c\sim \mathbf{Bin}(N,\, p'_{11}+p'_{00}), \quad \text{and } b\sim \mathbf{Bin}\left(n_1,\, \frac{p'_{10}}{p'_{11}+p'_{00}}\right).$$

Appendix G

A Conjecture on a Limiting Result

The Corollaries 3.2.8, 3.2.9, 3.2.10, B.0.7, 4.1.3, 4.1.6, 4.1.11, 4.1.8, 4.1.9, 4.1.12, and 4.1.13 suggest the following general result,

Conjecture G.0.1. *Let $X_1, \cdots, X_n$ be n exchangeable random variables such that $X_i|\theta \sim f_X(x|\theta)\, i = 1, \cdots, n$, $\theta \sim f(\theta)$. Suppose that $\boldsymbol{E}(\theta|\, x)$ for all x exists. Then there exists two constants $0 < \lambda \leq \frac{1}{2}$ and $c_k > 0$ such that*

$$\lim_{n \longrightarrow \infty} n^{\lambda} alc_k(n, 1-\alpha) = c_k.$$

Let $\mathbf{Var}(x)$ be the posterior variances of θ. Furthermore, if we assume that all posterior distributions can be approximated by a normal distribution then

$$c_k \approx 2z_{1-\alpha/2} \left(\int_{\Theta} \lim_{n \longrightarrow \infty} \{n\mathbf{Var}(n\theta)\}^{k/2} f(\theta) d\theta \right)^{1/k}$$

if the integral exists. In that case $\lambda = \frac{1}{2}$. This would lead to an approximate $\boldsymbol{ALC}_k$ sample size of $\left(\frac{c_k}{l}\right)^{1/\lambda}$.

Appendix H

The $S-$ ACC and $S-$ ALC

The S-average coverage criterion (ACC$_S$) and the S-average length criterion (ALC$_S$): This subsection describes a third family of measures that we are planning to investigate, that are S-average coverage criterion and S-average length criterion. As argued by Berger (1985, p.144), different measures of size other than interval length can be addressed. For instance, let $s(\theta)$ be a non-negative function and C be a credible set, and define

$$S(C) = \int_C s(\theta)\, d\theta, \tag{H.0.1}$$

a general measure of size. The problem of sample size determination then becomes to find the minimum n such that

$$\mathbf{ALC}_S \; : \int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \min_C \int_{\theta \in C(\mathbf{x}_n, n, 1-\alpha)} s(\theta)\, d\theta \right\} p_{X_n}(\mathbf{x}_n)\, dx \;\leq\; l, \tag{H.0.2}$$

given

$$\int_{C(\mathbf{x}_n, n, 1-\alpha)} p(\theta | \, \mathbf{x}_n)\, d\theta \,\geq\, 1-\alpha,$$

or

$$\mathbf{ACC}_S \; : \int_{\mathbf{x}_n \in \mathcal{X}_n} \left\{ \max_C \int_{\theta \in C(\mathbf{x}_n, n, l)} p(\theta | \, x_n) \, d\theta \right\} p_{X_n}(\mathbf{x}_n) \, dx \; \geq \; 1 - \alpha, \tag{H.0.3}$$

given

$$\int_{C(\mathbf{x}_n, n, 1-\alpha)} s(\theta) d\theta \; \leq \; l.$$

Remark H.0.1. When $s(\theta) \equiv 1$, we are dealing with the regular length used for $\mathbf{ALC}_k$ or $\mathbf{ACC}_k$.

Although the notion of S-optimal sets is not new, it does not seem to have been explored in the context of sample size calculations. Under mild conditions, the set $C(\mathbf{x}_n, n, l)$ that maximizes the posterior coverage or this measure of length is an HPD type regions where $p(\theta | \, \mathbf{x}_n)$ is substituted for $\frac{p(\theta | \, \mathbf{x}_n)}{s(\theta)}$ (Berger, 1985).

Joseph and al. (1997) have used the likelihood-based intervals for the computation of sample size and their method is known under the name of mixed Bayesian/likelihood criteria. The reason for that is one wants to average over the predictive data, but not use Bayesian methods for inference. It is easily seen that mixed Bayesian/likelihood criteria are special case $\mathbf{ALC}_S$ and $\mathbf{ACC}_S$ corresponding to $s(\theta) = p(\theta)$ where $p(\theta)$ is equal to the prior distribution.

Another interest for S-optimal arises as follows. The $\mathbf{ALC}_k$ and $\mathbf{ACC}_k$ have been criticized because HPD are not invariant under parameterization. Let $T(\theta)$ be a transformation on θ. By setting $s(\theta)$ to be equal to the jacobian of the transformation $\theta \mapsto T(\theta)$, one obtains HPD regions that preserve "optimality" under parameterization.

Appendix I

IMSL Subroutines

- **DUMINF:** minimizes a function of N variables using a quasi-Newton method and a finite-difference gradient (Denis and Schnabel, 1983; Gill and Murray, 1976).

- **DBCONF:** minimizes a function of N variables subjects to bounds using a quasi-Newton method and a finite-difference gradient (Denis and Schnabel, 1983; Gill and Murray, 1976).

- **DUMPOL:** minimizes a function of N variables using a direct search polytope algorithm (Nelder and Murray, 1965; Gill and Wright, 1981).

- **DNEQNF:** solves a system of nonlinear equations using a modified Powell hybrid algorithm and a finite-difference approximation to the Jacobian (More et al., 1980)

- **DQDAG:** integrates a function using a globally adaptive scheme based on Gauss-Kronrod rules (Piessens et al., 1983).

- **DQDAGS:** integrates (which may have endpoint singularities) a function using a globally adaptive scheme based on Gauss-Kronrod rules (Piessens et al., 1983).

- **DTWODQ:** computes a two dimensional-interated integral.

- **DQAND:** computes any hyper-rectangle integral up to 20 arguments.

Appendix J

Sample Size Tables for Selected Cases

Table J.1: **Table of sample sizes for p_1 with $(a, b) = (1, 1)$.**

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	.50	4	5	5	5	5	5		6	5	6	6	6	6
	.40	8	9	9	9	8	9		10	9	10	10	10	10
	.30	16	17	18	17	16	17		18	18	19	19	19	19
	.25	24	25	26	25	24	25		27	27	28	28	28	28
	.20	39	40	41	39	39	40		42	44	45	45	45	45
	.15	71	73	73	73	71	73	73	74	80	81	81	81	81
	.10	164	166	166	165	164	166	166	167	183	184	184	184	184
	.05	665	666	667	666	665	666	666	668	739	740	740	740	740
	.02	4169	4171	4172	4171	4169	4171	4171	4173	4631	4632	4632	4632	4632
	.01	16686	16688	16889	16688	16686	16688		16688	18533		18534	18534	18534
.95	.50	7	8	8	8	7	8		9	8	9	9	9	9
	.40	12	13	14	13	12	13		14	15	15	15	15	15
	.30	23	25	25	25	23	25		26	28	28	29	28	28
	.25	35	36	37	36	35	36	37	38	42	42	42	42	42
	.20	56	58	58	58	56	58	58	59	66	67	67	67	67
	.15	102	104	105	104	102	104	104	105	120	121	121	121	121
	.10	234	235	236	235	234	235	236	237	274	275	275	275	275
	.05	945	946	947	946	945	946	946	948	1105	1105	1106	1105	1106
	.02	5921	5922	5923	5923	5921	5922	5923	5924	6921	6921	6921	6921	6922
	.01	23693	23694	23696	23695	23693	23694	23695	23696	27691		27691	27692	27692
.99	.50	13	14	15	15	12	13		15	17	17	18	17	17
	.40	22	23	25	24	21	23		24	28	29	30	29	29
	.30	42	43	45	44	41	43	43	44	53	54	55	54	54
	.25	62	63	65	64	61	63	63	64	79	79	80	79	79
	.20	98	100	102	101	98	100	100	101	125	125	127	125	126
	.15	178	179	181	180	178	179	180	181	225	225	227	226	226
	.10	405	407	409	408	405	407	407	409	512	512	513	512	512
	.05	1633	1635	1637	1636	1633	1635	1635	1637	2058	2059	2060	2059	2059
	.02	10228	10229	10231	10230	10228	10229	10230	10231	12885	12885	12887	12885	12885
	.01	40923	40925	40927	40926	40923	40927	40925	40927			51554	51553	51553

Table J.2: **Table of sample sizes for p_1 with $(a,b) = (1,1)$.**

coverage	length	WLOC						MLOC					
		Exact		**1st order**		**3rd order**		**Exact**		**1st order**		**3rd order**	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	exact	formula	HPD	equal
.90	.50	8	8	8	9	8	8	5	6	6	7	6	6
	.40	15	15	14	15	15	15	9	10	10	10	10	10
	.30	28	28	28	29	28	28	10	10	10	12	10	10
	.25	41	41	41	42	41	41	21	21	21	21	21	21
	.20	65	65	65	66	65	65	49	49	49	50	49	49
	.15	118	118	118	119	118	118	88	89	89	89	88	89
	.10	268	268	268	269	268	268	201	201	201	201	201	201
	.05	1080	1080	1080	1081	1080	1080	809	810	809	810	809	810
	.02	6762	6762	6762	6762	6762	6762	5070	5070	5070	5070	5070	5070
	.01			27053	27054	27053	27053			20290	20290	20289	20290
.95	.50	12	12	13	14	12	12	9	9	10	10	9	9
	.40	21	21	21	23	21	21	14	15	17	17	14	15
	.30	40	40	40	41	40	40	29	30	30	31	29	29
	.25	59	59	59	60	59	59	42	44	45	45	42	44
	.20	93	93	94	95	93	93	69	70	70	71	69	70
	.15	168	168	168	169	168	168	125	126	126	127	125	126
	.10	381	381	382	383	381	381	285	285	286	287	285	286
	.05	1534	1534	1534	1535	1534	1534	1149	1149	1150	1151	1149	1150
	.02	9601		9601	9602	9601	9601	7200	7201	7201	7202	7200	7201
	.01	38412		38412	38413	38412	38412			28809	28810	28809	28809
.99	.50	22	22	24	25	22	23	15	17	18	19	14	17
	.40	37	37	39	40	37	37	26	27	29	30	26	27
	.30	69	69	71	72	69	69	50	52	53	54	50	51
	.25	102	102	104	105	102	102	75	77	78	78	75	77
	.20	162	162	163	164	162	162	121	121	121	123	121	121
	.15	291	291	292	293	291	291	217	217	219	220	217	217
	.10	659	659	661	662	659	659	493	494	496	496	493	494
	.05	2650	2650	2651	2652	2650	2650	1986	1986	1989	1989	1986	1986
	.02	16585		16585	16586	16583	16583			12438	12439	12437	12437
	.01	66346		66346	66347	66345	66345			49761	49761	49757	49758

Table J.3: **Table of sample sizes for p_1 with $(a, b) = (5, 5)$.**

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1 - \alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	.50	1	1	1	1	1	1	1	1	1	1	1	1	1
	.40	6	6	6	6	6	6	6	6	6	6	6	6	6
	.30	18	18	18	18	18	18	18	18	18	18	18	18	19
	.25	30	30	30	30	30	30	30	31	30	30	30	31	31
	.20	52	52	52	52	52	52	53	53	52	52	52	53	53
	.15	99	99	99	99	99	99	100	100	100	100	100	101	101
	.10	235	235	235	235	235	235	236	237	236	237	236	237	237
	.05	970	970	970	970	970	970	971	971	974	975	974	975	976
	.02	6112	6112	6112	6112	6112	6112	6113	6113	6140	6141	6141	6141	6142
	.01	24475	24475	24475	24475	24475	24475	24476	24477			24591	24592	24592
.95	.50	4	4	4	4	4	4	4	4	4	4	4	4	4
	.40	12	12	12	12	12	12	12	12	12	12	12	12	12
	.30	29	29	29	29	29	29	29	29	29	29	29	29	30
	.25	46	46	46	46	46	46	46	47	46	46	47	47	47
	.20	77	77	78	77	77	77	78	78	78	78	78	78	78
	.15	144	145	145	145	144	145	145	146	146	146	146	147	147
	.10	338	338	338	338	338	338	339	339	341	341	341	342	342
	.05	1381	1381	1381	1381	1381	1381	1382	1382	1394	1394	1394	1395	1395
	.02	8681	8681	8682	8682	8681	8681	8682	8682	8764	8764	8785	8765	8765
	.01		34755	34755	34755	34755	34755	34756	34756			35088	35088	35088
.99	.50	13	13	15	15	12	13	13	13	13	13	15	13	13
	.40	26	26	28	28	26	26	27	27	27	27	28	27	27
	.30	55	55	57	57	55	55	56	56	56	56	58	57	57
	.25	85	85	87	87	85	85	86	86	86	86	88	87	87
	.20	139	139	141	141	139	139	140	140	141	141	143	142	142
	.15	255	256	257	257	255	256	257	257	260	260	262	261	261
	.10	589	589	591	591	589	589	590	590	600	600	602	601	601
	.05	2390	2391	2392	2392	2390	2391	2391	2392	2436	2436	2438	2437	2437
	.02	15000	15000	15002	15002	15000	15000	15001	15001			15288	15287	15287
	.01			60036	60035	60034	60034	60035	60035			61182	61181	61881

Table J.4: **Table of sample sizes for p_1 with $(a,b) = (5,5)$.**

		WLOC						MLOC					
coverage	length	Exact		1st order		3rd order		Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	exact	formula	HPD	equal
.90	.50	1	1	1	1	1	1	1	1	1	4	1	1
	.40	7	7	6	7	7	7	6	6	6	9	6	6
	.30	20	20	20	21	20	20	17	17	17	19	17	17
	.25	33	33	33	34	33	33	26	26	26	29	26	26
	.20	57	57	57	58	57	57	45	45	45	47	45	45
	.15	110	110	110	111	110	110	85	85	85	89	85	85
	.10	260	260	260	261	260	260	197	198	198	199	197	198
	.05	1072	1072	1072	1073	1072	1072	806	806	806	808	806	806
	.02	6754	6754	6753	6754	6754	6754	5068	5069	5069	5069	5068	5068
	.01			27045	27046	27045	27045			20286	20288	20286	20286
.95	.50	4	4	5	6	4	14	4	4	5	8	4	4
	.40	13	13	13	15	13	13	11	11	11	14	11	11
	.30	32	32	32	33	32	32	25	26	26	28	25	26
	.25	51	51	51	52	51	51	40	41	41	42	40	41
	.20	85	85	86	87	85	85	69	66	66	68	66	66
	.15	160	160	160	161	160	160	122	122	122	124	122	122
	.10	373	373	374	375	373	373	282	282	283	284	282	282
	.05	1526	1526	1526	1527	1526	1526	1146	1147	1149	1149	1146	1147
	.02	9593	9593	9593	9594	9593	9593	7197	7198	7198	7199	7197	7198
	.01			38404	38404	38404	38404			28806	28807	28805	28806
.99	.50	14	14	16	17	14	14	15	17	18	19	14	17
	.40	29	29	31	32	29	29	12	12	14	16	11	12
	.30	61	61	63	64	61	61	23	24	25	27	23	23
	.25	94	94	96	97	94	94	73	73	78	76	73	73
	.20	154	154	155	156	154	154	117	118	119	121	117	118
	.15	283	283	284	285	283	283	214	214	217	217	214	214
	.10	651	651	653	654	651	651	490	491	493	494	490	491
	.05	2642	2642	2643	2644	2642	2642	1984	1985	1986	1987	1984	1985
	.02			16577	16578	16575	16575	12434	12434	12437	12437	12433	12434
	.01			66338	66339	66337	66336			49757	49758	49754	49755

Table J.5: **Table of sample sizes for p_1 with** $(a, b) = (10, 10)$.

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	.50	1	1	1	1	1	1	1	1	1	1	1	1	1
	.40	1	1	1	1	1	1	1	1	1	1	1	1	1
	.30	9	9	9	9	9	9	10	10	9	9	9	10	10
	.25	22	22	22	22	22	22	22	22	22	22	22	22	22
	.20	45	45	45	45	45	45	46	46	45	45	45	46	46
	.15	95	95	95	95	95	95	96	96	95	95	95	96	96
	.10	238	238	238	238	238	238	239	239	238	238	238	239	239
	.05	1010	1010	1010	1010	1010	1010	1011	1011	1011	1011	1011	1012	1012
	.02	6415	6415	6415	6417	6415	6415	6416	6416	6421	6422	6422	6422	6423
	.01	25717	25717	25717	25717	25717	25717	25718	25718			25744	25745	25745
.95	.50	1	1	1	1	1	1	1	1	1	1	1	1	1
	.40	3	3	3	3	3	3	3	3	3	3	3	3	3
	.30	21	21	21	21	21	21	21	21	21	21	21	21	21
	.25	38	39	39	39	38	39	39	39	39	39	39	39	39
	.20	71	71	72	72	71	71	72	72	72	72	72	72	72
	.15	142	142	143	143	142	142	143	143	143	143	143	143	144
	.10	345	346	346	346	345	346	346	346	346	346	347	347	347
	.05	1442	1442	1442	1442	1442	1442	1443	1443	1445	1445	1446	1446	1446
	.02	9115	9116	9116	9116	9115	9116	9117	9117		9137	9138	9138	9138
	.01		36522	36522	36522	36522	36522	36523	36523			36610	36610	36610
.99	.50	4	4	6	6	4	4	4	4	4	4	6	4	4
	.40	18	18	20	20	18	18	19	19	18	18	20	19	19
	.30	49	49	51	51	49	49	50	50	49	49	51	50	50
	.25	80	80	82	81	80	80	80	81	80	80	82	81	81
	.20	136	137	138	138	136	136	137	137	137	137	139	138	138
	.15	259	259	261	261	259	259	260	260	260	261	262	261	261
	.10	610	610	612	612	610	610	611	611	613	613	615	614	614
	.05	2503	2503	2505	2505	2503	2503	2504	2504	2516	2516	2518	2517	2517
	.02	15757	15757	15759	15759	15757	15757	15758	15758			15842	15841	15841
	.01			63095	63095	63093	63093	63094	63094			63429	63428	63428

Table J.6: **Table of sample sizes for p_1 with $(a, b) = (10, 10)$.**

		WLOC						MLOC					
coverage	length	Exact		1st order		3rd order		Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	exact	formula	HPD	equal
.90	.50	1	1	1	7	1	1	1	1	1	1	1	1
	.40	1	1	1	1	1	1	1	1	1	6	1	1
	.30	10	10	10	11	10	10	9	9	9	15	9	9
	.25	23	23	23	24	23	23	20	20	20	25	20	20
	.20	47	47	47	48	47	47	39	39	39	44	39	39
	.15	100	100	100	101	100	100	80	80	81	83	80	80
	.10	250	250	250	251	250	250	193	194	194	196	193	194
	.05	1062	1062	1062	1063	1062	1062	802	803	804	804	802	803
	.02	6744	6744	6743	6744	6744	6744	5065	5065	5065	5066	5065	5065
	.01			27035	27036	27035	27035			20284	20284	20282	20284
.95	.50	1	1	1	1	1	1	1	1	1	4	1	1
	.40	3	3	3	5	3	3	3	3	3	11	3	3
	.30	22	22	22	23	22	22	19	19	19	25	19	19
	.25	41	41	41	42	41	41	34	34	34	32	34	34
	.20	75	75	76	77	75	75	61	61	61	65	61	61
	.15	150	150	150	151	150	150	118	118	118	121	118	118
	.10	363	363	364	365	363	363	278	278	279	281	278	278
	.05	1516	1516	1516	1517	1516	1516	1143	1144	1145	1145	1143	1144
	.02	9583	9583	9593	9584	9583	9583	7194	7194	7194	7196	7194	7194
	.01			38394	38395	38394	38394			28802	2884	28802	28802
.99	.50	4	4	6	7	4	4	4	4	6	13	4	4
	.40	19	19	21	22	19	19	17	17	18	24	17	17
	.30	51	51	53	54	51	51	42	42	45	48	42	42
	.25	84	84	86	87	84	84	67	68	69	72	67	68
	.20	144	144	145	146	144	144	113	114	114	117	113	113
	.15	273	273	274	275	273	273	210	210	213	214	210	210
	.10	641	641	643	644	641	641	486	487	489	490	486	487
	.05	2632	2632	2633	2634	2632	2632	1981	1981	1982	1983	1981	1981
	.02			16567	16568	16565	16565	12430	12430	12433	12433	12430	12430
	.01			66328	66329	66327	66327			49754	49755	49751	49753

Table J.7: **Table of sample sizes for ω with $(a,b) = (3,3)$.**

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	20	27	34	31	19	26	11	18	24	30	37	15	23
	1.5	47	56	63	60	46	54	35	43	53	60	67	45	52
	1.0	127	137	144	141	126	136	112	121	137	144	151	129	136
	.8	209	219	227	224	208	218	192	202	222	229	236	214	221
	.6	386	397	405	402	385	396	367	378	406	413	420	398	405
	.5	565	576	584	581	564	576	545	556	591	597	604	583	589
	.4	894	906	915	911	894	906	873	885	931	938	944	923	929
	.3	1607	1619	1628	1624	1606	1619	1584	1596	1666	1672	1679	1657	1664
	.2	3643	3656	3665	3662	3643	3656	3619	3932	3765	3771	3778	3757	3763
	.1	14643	14657	14666	14663	14643	14657	14618	14631	15100	15107	15144	15092	15099
.95	2.0	39	46	50	47	37	44	28	35	80	89	97	66	75
	1.5	78	87	90	87	77	85	64	73	156	165	173	143	152
	1.0	193	203	206	203	192	202	176	185	374	382	390	361	369
	.8	309	320	323	320	309	319	291	301	594	602	610	581	589
	.6	562	574	576	573	561	573	542	553	1069	1078	1086	1056	1065
	.5	816	828	831	828	816	827	795	806	1547	1556	1564	1534	1543
	.4	1284	1297	1299	1296	1284	1296	1262	1274	2428	2436	2444	2415	2423
	.3	2296	2309	2312	2309	2296	2309	2273	2286	4329	4338	4346	4316	4325
	.2	5189	5202	5205	5201	5188	5202	5164	5177	9763	9771	9779	9750	9758
	.1	20808	20822	20824	20821	20808	20822	20782	20796			39119	39090	39098
.99	2.0	87	95	88	82	84	91	72	80	656	670	683	626	641
	1.5	157	166	158	154	154	163	140	148	1189	1204	1217	1160	1175
	1.0	357	368	357	354	355	365	337	347	2714	2728	2741	2685	2699
	.8	559	571	560	557	557	568	538	549	4257	4271	4284	4228	4243
	.6	996	1009	997	994	995	1007	974	985	7591	7606	7618	7562	7577
	.5	1436	1449	1437	1433	1435	1447	1413	1425	10994	10959	10972	10916	10930
	.4	2245	2258	2246	2243	2244	2257	2221	2234	17118		17145	17089	17103
	.3	3994	4007	3995	3991	3993	4006	3969	3982	30455	30455	30482	30426	30440
	.2	8990	9004	8991	8987	8990	9003	8964	8978			68587	68531	68546
	.1			35969	35966	35968	35982	35942	35956			274357	274302	274357

Table J.8: **Table of sample sizes for ω with $(a, b) = (3, 3)$.**

coverage	length	MLOC					
		Exact		**1st order**		**3rd order**	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	6	10	12	12	19	20
	1.5	15	18	20	21	30	32
	1.0	38	42	45	45	57	59
	.8	63	67	69	69	82	85
	.6	115	119	122	122	136	140
	.5	168	172	175	176	190	193
	.4	266	270	272	272	282	292
	.3	476	480	482	482	499	503
	.2	1077	1081	1084	1084	1101	1105
	.1	4324	4328	4330	4330	4348	4352
.95	2.0	12	16	16	17	26	27
	1.5	24	28	28	29	40	42
	1.0	58	62	63	63	77	80
	.8	93	97	97	97	113	116
	.6	168	171	172	172	189	192
	.5	243	247	247	247	265	268
	.4	381	385	386	386	404	407
	.3	680	684	684	684	703	707
	.2	1533	1537	1538	1538	1557	1561
	.1	6143	6147	6148	6148	6167	6171
.99	2.0	27	30	28	28	41	43
	1.5	48	51	51	49	64	67
	1.0	107	111	108	108	127	130
	.8	167	171	167	167	188	191
	.6	296	300	296	296	318	321
	.5	426	429	426	426	448	452
	.4	664	668	665	665	687	691
	.3	1180	1184	1181	1181	1204	1208
	.2	2655	2659	2655	2655	2679	2683
	.1	10617	10621	10617	10617	10641	10645

Table J.9: **Table of sample sizes for ω with $(a, b) = (5, 5)$.**

		ALC								ACC				
coverage	length	Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	4	9	13	10	3	8	1	6	5	11	16	3	7
	1.5	19	25	28	26	19	24	14	19	24	29	34	17	23
	1.0	63	69	73	70	63	69	56	62	75	81	86	68	73
	.8	108	114	118	115	108	114	100	106	128	133	138	120	126
	.6	205	211	215	213	205	211	197	202	240	246	251	233	238
	.5	303	309	313	310	303	309	294	300	354	359	363	346	352
	.4	483	489	492	490	483	489	474	480	563	568	573	555	560
	.3	871	877	881	879	871	877	862	868	1014	1019	1024	1006	1011
	.2	1982	1988	1992	1989	1982	1988	1972	1978	2302	2308	2313	2295	2300
	.1	7978	7984	7988	7986	7976	7984	7969	7975	9261	9267	9272	9253	9259
.95	2.0	15	20	21	19	14	19	11	15	27	34	39	16	23
	1.5	37	42	43	41	36	41	30	36	65	71	76	53	60
	1.0	99	105	106	104	99	105	92	97	171	178	183	159	166
	.8	163	169	170	168	163	169	155	161	279	285	290	267	274
	.6	301	307	308	306	301	307	292	298	512	518	523	500	506
	.5	440	446	447	445	440	446	431	437	746	752	757	734	740
	.4	695	701	702	700	695	701	686	692	1177	1183	1188	1165	1171
	.3	1247	1253	1254	1252	1247	1253	1238	1244	2108	2114	2119	2096	2103
	.2	2824	2830	2831	2828	2824	2830	2814	2820	4768	4775	4780	4757	4763
	.1	11328	11334	11344	11342	11337	11343	11328	11334	19134		19145	19121	19129
.99	2.0	41	46	42	40	40	45	35	40	168	177	182	143	153
	1.5	79	85	80	78	79	84	72	77	321	329	335	297	306
	1.0	188	194	189	187	188	193	180	185	757	766	771	734	742
	.8	299	304	299	297	298	304	290	295	1199	1208	1213	1175	1184
	.6	537	543	537	535	537	543	528	534	2153	2162	2167	2130	2139
	.5	777	782	777	775	776	782	767	773	3113	3122	3127	3090	3099
	.4	1218	1224	1218	1216	1218	1224	1208	1214	4880	4889	4894	4857	4866
	.3	2171	2177	2171	2169	2171	2177	2161	2167	8698	8707	8712	8675	8684
	.2	4894	4900	4894	4892	4894	4900	4884	4890	19606		19620	19583	19591
	.1	19599	19605	19599	19597	19599	19605	19589	19595			78522	78485	78493

Table J.10: **Table of sample sizes for ω with $(a,b) = (5,5)$.**

coverage	length	MLOC					
		Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	1	6	8	8	15	16
	1.5	11	14	16	17	26	28
	1.0	34	38	41	41	53	55
	.8	59	63	65	65	78	81
	.6	111	115	118	118	132	136
	.5	164	168	171	171	186	189
	.4	262	266	268	268	284	288
	.3	472	476	478	478	495	499
	.2	1073	1077	1084	1080	1097	1101
	.1	4320	4324	4326	4326	4344	4348
.95	2.0	8	12	12	13	22	23
	1.5	20	24	24	25	36	38
	1.0	54	58	59	59	73	76
	.8	89	93	93	94	109	112
	.6	164	167	168	168	185	188
	.5	239	243	243	243	261	264
	.4	377	381	382	382	400	403
	.3	676	680	680	680	699	703
	.2	1528		1534	1534	1553	1557
	.1	6139	6143	6144	6144	6163	6167
.99	2.0	23	26	19	24	37	39
	1.5	44	47	44	45	60	63
	1.0	103	107	104	104	123	126
	.8	163	167	163	163	184	187
	.6	292	296	292	292	314	317
	.5	422	425	422	422	444	448
	.4	660	664	661	661	683	687
	.3	1176	1180	1167	1171	1200	1204
	.2	2651	2655	2651	2651	2675	2676
	.1	10613	10617	10613	10613	10637	10641

Table J.11: **Table of sample sizes for ω with $(a, b) = (10, 10)$.**

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	1	1	1	1	1	1	1	1	1	1	1	1	1
	1.5	1	1	8	6	1	5	1	4	1	6	9	1	4
	1.0	32	37	40	37	32	36	28	33	37	41	45	31	36
	.8	64	69	72	70	64	69	59	64	73	77	81	71	71
	.6	133	138	141	139	133	138	128	132	150	155	159	143	148
	.5	203	208	211	208	203	208	197	202	228	233	237	221	225
	.4	331	336	339	337	331	336	325	330	372	377	380	364	369
	.3	608	613	616	614	608	613	602	607	682	687	690	674	679
	.2	1400	1405	1407	1405	1400	1405	1396	1393	1568	1573	1577	1560	1565
	.1	5674	5679	5682	5680	5674	5679	5668	5673	6354	6359	6363	6346	6351
.95	2.0	1	2	3	1	1	1	1	1	1	1	1	1	1
	1.5	13	17	19	16	13	17	11	15	19	24	27	13	17
	1.0	58	63	63	61	58	62	54	58	81	86	89	70	76
	.8	104	108	109	107	103	108	103	108	143	148	151	132	137
	.6	202	206	207	205	202	206	196	201	278	283	285	266	271
	.5	301	305	306	304	301	305	295	299	413	418	421	401	406
	.4	483	487	488	586	483	487	477	481	662	667	669	650	655
	.3	876	881	882	880	876	881	870	875	1199	1204	1207	1187	1192
	.2	2000	2005	2006	2004	2000	2005	1996	1998	2735	2740	2743	2723	2728
	.1	8069	8074	8075	8073	8069	8074	8063	8068	11028		11036	11016	11021
.99	2.0	17	21	17	15	16	20	15	15	39	45	45	21	28
	1.5	44	48	45	43	44	48	40	41	98	104	105	77	84
	1.0	122	126	122	120	122	126	116	121	266	272	273	245	251
	.8	200	205	201	199	200	205	195	199	437	443	444	415	422
	.6	370	375	371	369	370	375	364	369	804	810	811	783	789
	.5	541	546	542	540	541	546	535	539	1174	1180	1181	1153	1159
	.4	856	860	856	854	856	860	849	854	1855	1861	1862	1834	1840
	.3	1535	1540	1536	1533	1535	1540	1529	1533	3325	3331	3333	3304	3310
	.2	3476	3481	3477	3475	3476	3481	3470	3474	7527	7533	7534	7506	7512
	.1	13952	13957	13959	13957	13959	13964	13952	13957			30224	30196	30202

Table J.12: **Table of sample sizes for ω with $(a, b) = (10, 10)$.**

coverage	length	MLOC					
		Exact		**1st order**		**3rd order**	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	1	1	1	1	5	6
	1.5	1	4	6	7	16	18
	1.0	24	28	31	31	43	47
	.8	49	53	55	55	68	71
	.6	101	105	108	108	122	126
	.5	154	158	161	161	176	179
	.4	252	256	258	258	274	278
	.3	462	466	468	468	485	489
	.2	1063	1067	1070	1070	1087	1091
	.1	4310	4314	4316	4316	4334	4338
.95	2.0	1	1	1	1	12	13
	1.5	10	14	14	15	26	28
	1.0	44	48	49	49	63	66
	.8	79	83	83	84	99	102
	.6	154	157	158	158	175	178
	.5	229	233	233	231	251	254
	.4	367	371	372	372	390	393
	.3	666	670	670	670	689	693
	.2	1519	1523	1524	1524	1543	1547
	.1	6128	6133	6134	6134	6153	6157
.99	2.0	13	16	14	14	27	29
	1.5	34	37	34	35	50	53
	1.0	93	93	94	94	113	116
	.8	153	157	153	153	174	177
	.6	282	286	282	282	304	307
	.5	412	415	412	412	434	438
	.4	650	654	651	651	673	677
	.3	1166	1170	1167	1167	1190	1194
	.2		2645	2641	2641	2665	2669
	.1	10603	10607	10603	10603	10627	10631

Table J.13: **Table of sample sizes for ϕ with** $(a, b) = (2, 2)$.

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	11	11	11	12	11	11	10	10	11	11	12	10	10
	1.5	22	22	23	23	22	22	21	21	22	22	24	21	21
	1.0	55	56	56	57	55	55	53	54	56	56	58	54	54
	0.8	89	89	90	90	89	89	87	87	89	89	92	87	87
	.6	162	162	163	163	162	162	159	160	162	162	166	160	160
	.5	235	235	236	237	235	236	233	233	236	236	240	233	234
	.4	371	371	371	372	371	371	367	368	371	371	375	368	369
	.3	663	663	663	664	663	663	660	660	663	663	667	660	661
	.2	1497	1497	1498	1499	1497	1497	1493	1494	1497	1497	1502	1495	1495
	.1	6003	6003	6004	6005	6003	6003	5999	6000	6002	6002	6007	5999	6000
.95	2.0	17	17	17	18	17	17	16	16	18	18	19	16	16
	1.5	34	34	34	34	34	34	32	32	35	35	37	33	33
	1.0	81	81	81	82	81	81	79	79	85	85	88	82	82
	0.8	129	129	129	130	129	129	126	127	135	135	139	132	132
	.6	232	233	233	233	232	233	230	230	244	244	248	241	241
	.5	337	337	337	338	337	337	334	334	354	354	358	350	351
	.4	529	529	529	530	529	529	525	526	555	555	560	552	552
	.3	943	944	943	944	943	944	940	940	991	991	996	988	988
	.2	2128	2128	2128	2129	2128	2128	2124	2125	2236	2236	2242	2232	2233
	.1	8526	8526	8526	8527	8526	8526	8522	8522	8959	8959	8965	8955	8956
.99	2.0	33	34	33	33	33	33	32	32	40	40	42	35	36
	1.5	62	62	61	62	62	62	60	60	76	76	79	70	71
	1.0	144	144	143	144	144	144	141	141	180	180	186	173	174
	0.8	227	227	226	227	227	227	224	224	286	286	293	278	279
	.6	406	406	405	406	406	406	402	403	516	516	524	507	508
	.5	586	586	585	586	586	586	582	583	746	746	756	737	738
	.4	917	918	916	917	917	918	914	914	1171	1171	1181	1162	1163
	.3	1633	1634	1633	1634	1633	1634	1630	1630	2088	2088	2099	2079	2080
	.2	3680	3680	3679	3680	3680	3680	3676	3676	4709	4709	4721	4700	4701
	.1	14730	13731	14729	14730	14730	14730	14726	14727	18862	18862	18875	18852	18854

Table J.14: **Table of sample sizes for ϕ with $(a, b) = (2, 2)$.**

coverage	length	MLOC					
		Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	11	11	11	12	11	11
	1.5	22	22	22	23	22	22
	1.0	54	54	55	55	54	54
	.8	87	87	87	87	87	87
	.6	156	156	156	157	156	156
	.5	227	227	227	228	227	227
	.4	356	356	356	358	356	356
	.3	637	637	637	638	637	637
	.2	1439	1439	1439	1440	1439	1439
	.1	5768	5768	5768	5769	5768	5768
.95	2.0	16	16	16	10	16	16
	1.5	32	32	32	34	32	32
	1.0	79	79	79	79	79	79
	.8	124	124	124	125	124	124
	.6	224	224	224	225	224	224
	.5	324	324	324	325	324	324
	.4	508	508	508	509	508	508
	.3	907	907	907	908	907	907
	.2	2044	2044	2044	2046	2044	2044
	.1	8191	8191	8191	8192	8191	8191
.99	2.0	32	32	32	35	32	32
	1.5	60	60	59	60	60	60
	1.0	139	139	138	139	139	139
	.8	219	219	217	216	219	219
	.6	391	391	389	390	391	391
	.5	563	563	563	563	563	563
	.4	881	882	880	884	881	882
	.3	1570	1570	1568	1570	1570	1570
	.2	3535	3536	3535	3536	3535	3536
	.1		14151	14151	14152	14151	14151

Table J.15: Table of sample sizes for ϕ with $(a, b) = (5, 5)$.

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	2	2	1	1	2	2	2	2	1	1	3	1	1
	1.5	12	12	12	12	12	12	11	11	12	12	13	11	11
	1.0	39	39	39	39	39	39	38	38	39	39	40	38	38
	0.8	66	66	66	66	66	65	65	65	66	66	68	65	65
	.6	125	125	125	135	125	125	123	123	125	125	127	124	124
	.5	184	184	184	184	184	184	182	182	184	184	186	183	183
	.4	292	293	293	293	292	293	291	291	293	293	296	292	292
	.3	528	528	528	528	528	528	526	526	529	529	531	527	527
	.2	1199	1199	1199	1200	1199	1199	1198	1198	1202	1202	1205	1201	1201
	.1	4827	4827	4827	4827	4827	4827	4825	4825	4838	4838	4841	4837	4837
.95	2.0	8	8	8	8	8	8	7	8	8	8	8	7	7
	1.5	21	21	21	21	21	21	20	20	21	21	22	20	20
	1.0	59	59	59	59	59	59	58	58	60	60	61	58	58
	0.8	98	98	98	98	98	98	97	97	99	99	100	97	97
	.6	181	181	181	181	181	181	180	180	183	183	185	181	181
	.5	265	265	265	265	265	265	264	264	267	267	270	266	266
	.4	420	420	420	420	420	420	418	418	423	423	426	422	422
	.3	753	754	753	754	753	754	752	752	760	760	763	759	759
	.2	1707	1707	1707	1707	1707	1707	1706	1706	1723	1723	1725	1721	1721
	.1	6858	6858	6857	6858	6858	6858	6856	6856	6921	6921	6924	6919	6919
.99	2.0	21	21	20	20	21	21	20	20	21	21	22	20	20
	1.5	44	44	43	43	44	44	43	43	45	45	46	43	43
	1.0	110	110	109	109	110	110	108	108	112	112	114	111	111
	0.8	176	176	176	176	176	176	175	175	181	181	183	179	179
	.6	320	321	320	320	320	321	319	319	330	330	332	328	328
	.5	465	465	465	465	465	465	464	464	479	479	481	477	477
	.4	732	732	732	732	732	732	731	731	754	754	757	752	752
	.3	1309	1309	1309	1308	1309	1309	1307	1307	1349	1349	1351	1347	1347
	.2	2956	2956	2956	2956	2956	2956	2955	2955	3047	3047	3050	3045	3045
	.1	11852	11852	11851	11851	11852	11852	11850	11850	12219	12219	12222	12217	12217

Table J.16: **Table of sample sizes for ϕ with** $(a, b) = (5, 5)$.

coverage	length	MLOC					
		Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	3	3	3	2	3	3
	1.5	12	12	12	13	12	12
	1.0	44	44	44	45	44	44
	.8	76	76	76	77	76	76
	.6	147	147	147	147	147	147
	.5	216	216	216	218	216	216
	.4	347	347	347	348	347	347
	.3	627	627	628	628	627	627
	.2	1428	1428	1428	1430	1428	1428
	.1	5759	5759	5758	5759	5758	5759
.95	2.0	8	8	8	8	8	8
	1.5	23	23	23	24	23	23
	1.0	68	68	68	69	68	68
	.8	115	115	115	115	115	115
	.6	215	215	215	215	215	215
	.5	315	315	315	315	315	315
	.4	499	499	499	499	499	499
	.3	896	896	896	898	896	896
	.2	2035	2035	2035	2036	2035	2035
	.1			8181	8182	8181	8181
.99	2.0	23	23	23	23	23	23
	1.5	51	51	49	50	51	51
	1.0	128	128	128	129	128	128
	.8	208	208	207	208	208	208
	.6	380	380	379	380	380	380
	.5	552	553	562	553	552	553
	.4	871	872	871	872	871	872
	.3	1560	1560	1559	1560	1560	1560
	.2	3525	3526	3524	3526	3525	3526
	.1			14140	14142	14141	14141

Table J.17: **Table of sample sizes for ϕ with $(a, b) = (10, 10)$.**

coverage	length	ALC								ACC				
		Exact		1st order		3rd order		limiting		Exact		1st order	limiting	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal	HPD	equal	HPD	equal	exact	HPD	equal
.90	2.0	1	1	1	1	1	1	1	1	1	1	1	1	1
	1.5	1	1	1	1	1	1	1	1	1	1	1	1	1
	1.0	26	26	26	26	26	26	25	25	26	26	27	25	25
	0.8	52	52	52	52	52	52	51	51	52	52	53	51	51
	.6	107	107	107	107	107	107	106	106	107	107	109	106	106
	.5	163	163	163	163	163	163	162	162	163	163	165	162	162
	.4	265	266	266	266	266	266	264	264	266	266	268	265	265
	.3	487	487	487	487	487	487	486	486	488	488	490	487	487
	.2	1121	1121	1121	1121	1121	1121	1120	1120	1122	1122	1124	1121	1121
	.1	4543	4543	4543	4543	4543	4543	4542	4542	4546	4546	4549	4545	4545
.95	2.0	1	1	1	1	1	1	1	1	1	1	1	1	1
	1.5	9	9	9	9	9	9	9	9	9	9	10	9	9
	1.0	45	45	45	45	45	45	45	45	45	45	47	45	45
	0.8	82	82	82	82	82	82	81	81	82	82	84	81	81
	.6	161	161	160	160	161	161	159	160	161	161	163	160	160
	.5	240	240	240	240	240	240	239	239	240	240	242	239	239
	.4	385	385	385	385	385	385	384	384	386	386	388	385	385
	.3	700	700	700	700	700	700	699	699	702	702	704	701	701
	.2	1600	1600	1600	1600	1600	1600	1659	1659	1604	1604	1606	1603	1603
	.1	6459	6459	6459	6459	6459	6459	6458	6458	6474	6474	6476	6473	6473
.99	2.0	9	9	8	8	9	9	9	9	9	9	9	9	9
	1.5	31	31	30	30	31	31	30	30	31	31	32	30	30
	1.0	93	93	92	92	93	93	92	92	93	93	95	92	92
	0.8	156	156	155	155	156	156	155	155	157	157	158	156	156
	.6	292	292	291	291	292	292	291	291	294	294	295	292	292
	.5	429	429	428	428	429	429	428	428	431	431	433	430	430
	.4	680	680	680	680	680	680	679	679	685	685	686	683	683
	.3	1224	1224	1224	1224	1224	1224	1223	1223	1232	1232	1234	1231	1231
	.2	2778	2778	2778	2778	2778	2778	2777	2777	2797	2797	2799	2796	2796
	.1	11170	11170	11170	11170	11170	11170	11169	11169	11246	11246	11248	11245	11245

Table J.18: **Table of sample sizes for ϕ with $(a, b) = (10, 10)$.**

		MLOC					
coverage	length	Exact		1st order		3rd order	
$1-\alpha$	l	HPD	equal	exact	formula	HPD	equal
.90	2.0	1	1	1	1	1	1
	1.5	1	1	1	1	1	1
	1.0	29	29	29	28	29	29
	.8	60	60	61	61	60	60
	.6	131	131	131	131	131	131
	.5	200	200	200	201	200	200
	.4	331	331	331	331	331	331
	.3	611	611	611	612	611	611
	.2	1412	1412	1412	1413	1412	1412
	.1	5740	5741	5741	5742	5740	5741
.95	2.0	1	1	1	1	1	1
	1.5	10	10	10	10	10	10
	1.0	53	53	52	52	53	53
	.8	99	99	99	99	99	99
	.6	198	198	198	198	198	198
	.5	298	298	298	298	298	298
	.4	483	483	483	483	483	483
	.3	880	880	880	881	880	880
	.2	2019	2019	2019	2019	2019	2019
	.1	8164	8164	8164	8166	8164	8164
.99	2.0	9	9	9	6	9	9
	1.5	35	35	35	33	35	35
	1.0	112	112	112	112	112	112
	.8	192	192	191	192	192	192
	.6	364	364	363	364	364	364
	.5	536	536	536	537	536	536
	.4	855	855	855	855	855	855
	.3	1543	1543	1543	1543	1543	1543
	.2	3508	3508	3508	3509	3508	3508
	.1			14124	14125	14124	14124

Bibliography

[Abdelbasit and Plackett, 1983] Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, 78:90–98.

[Adcock, 1987] Adcock, C. J. (1987). A Bayesian approach to calculating sample sizes for multinomial sampling (corr: V37 p239). *The Statistician*, 36:155–159.

[Adcock, 1988] Adcock, C. J. (1988). A Bayesian approach to calculating sample sizes. *The Statistician*, 37:433–439.

[Adcock, 1992] Adcock, C. J. (1992). Comment on "sample size determination in Bayesian analysis". *The Statistician*, 41:399–404.

[Adcock, 1993] Adcock, C. J. (1993). An improved Bayesian procedure for calculating sample sizes in multinomial sampling. *The Statistician*, 42:91–95.

[Adcock, 1995] Adcock, C. J. (1995). The Bayesian approach to determination of sample sizes – Some comments on the paper by Joseph, Wolfson and du Berger. *The Statistician*, 44:155–161.

[Adcock, 1997] Adcock, C. J. (1997). Sample size determination: A review. *The Statistician*, 46:261–283.

[Agresti, 1999] Agresti, A. (1999). On logit confidence intervals for the odds ratio with small samples. *Biometrics*, 55:597–602.

[Aitchison and Bacon-Shone, 1981] Aitchison, J. and Bacon-Shone, J. (1981). Bayesian risk ratio analysis. *The American Statistician*, 35:254–257.

[Albert et al., 2001] Albert, P. S., Ratnasinghe, D., Tangrea, J., and Wacholder, S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *The American Journal of Epidemiology*, 154:687–693.

[Ashby et al., 1993] Ashby, D., Hutton, J. L., and McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *The Statistician*, 42:385–397.

[Ashby et al., 1998] Ashby, D., Smyth, R. L., and Brown, P. J. (1998). Statistical issues in pharmacoepidemiological case-control studies. *Statistics in Medicine*, 17:1839–1850.

[Atkinson and Donev, 1992] Atkinson, A. C. and Donev, A. N. (1992). *Optimum Experimental Designs.* Oxford University Press.

[Bedrick et al., 1997] Bedrick, E. J., Christensen, R., and Johnson, W. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, 51:211–218.

[Berger, 1985] Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer-Verlag.

[Bernardo, 1979] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (c/r p128-147). *Journal of the Royal Statistical Society, Series B, Methodological*, 41:113–128.

[Bernardo, 1997] Bernardo, J. M. (1997). Statistical inference as a decision problem: the choice of sample size. *The Statistician*, 46:151–153.

[Bernardo and Ramón, 1998] Bernardo, J. M. and Ramón, J. M. (1998). An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *The Statistician*, 47:101–135.

[Billingsley, 1995] Billingsley, P. (1995). *Probability and Measure (Third Edition).* Wiley.

[Blackwelder, 1993] Blackwelder, W. C. (1993). Sample size and power for prospective analysis of relative risk. *Statistics in Medicine*, 12:691–698.

[Box and Tiao, 1992] Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis (Classics Edition).* Wiley.

[Breslow and Day, 1980] Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-control Studies.* World Health Organization.

[Breslow and Day, 1988] Breslow, N. E. and Day, N. E. E. (1988). *Statistical Methods in Cancer Research Volume II: The Design and Analysis of Cohort Studies.* Oxford University Press.

[Brooks, 1987] Brooks, R. J. (1987). Optimal allocation for bayesian inference about an odds ratio. *Biometrika*, 74:196–199.

[Brooks, 1998] Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician*, 47:69–100.

[Carlin, 1992] Carlin, J. B. (1992). Meta-analysis for 2×2 tables: A Bayesian approach. *Statistics in Medicine*, 11:141–158.

[Carnahan, 1989] Carnahan, J. V. (1989). Maximum likelihood estimation for the 4-parameter beta distribution. *Communications in Statistics, Part B – Simulation and Computation*, 18:513–536.

[Casagrande et al., 1978] Casagrande, J. T., Pike, M. C., and Smith, P. G. (1978). An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, 34:483–486.

[Cesana et al., 2001] Cesana, B. M., Reina, G., and Marubina, E. (2001). Sample size for testing a proportion in clinical trials: A "two-step" procedure combining power and confidence interval expected width. *The American Statistician*, 55:288–292.

[Chaloner and Duncan, 1987] Chaloner, K. and Duncan, G. T. (1987). Some properties of the Dirichlet-multinomial distribution and its use in prior elicitation. *Communications in Statistics, Part A – Theory and Methods*, 16:511–523.

[Chaloner and Larntz, 1989] Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21:191–208.

[Chaloner and Verdinelli, 1995] Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, 10:273–304.

[Chaloner and Duncan, 1983] Chaloner, K. M. and Duncan, G. T. (1983). Assessment of a beta prior distribution: Pm elicitation. *The Statistician*, 32:174–180.

[Connett et al., 1987] Connett, J. E., Smith, J. A., and McHugh, R. B. (1987). Sample size and power for pair-matched case-control studies. *Statistics in Medicine*, 6:53–59.

[Connor, 1987] Connor, R. J. (1987). Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43:207–211.

[Cornfield, 1951] Cornfield, J. (1951). A method of estimating comparative rates from clinical data. applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, 11:1269–1275.

[Davies et al., 1998] Davies, H. T. O., Crombie, I. K., and Tavakoli, M. (1998). When can odds ratios mislead? *BMJ*, 316:989–991.

[Denis and Schnabel, 1983] Denis, J. E. J. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.

[Devroye, 1986] Devroye, L. (1986). *Non-uniform Random Variate Generation.* Springer-Verlag.

[Dharmadhikari and Joag-dev, 1988] Dharmadhikari, S. and Joag-dev, K. (1988). *Unimodality, Convexity, and Applications.* Academic.

[Dupont, 1988] Dupont, W. D. (1988). Power calculations for matched case-control studies. *Biometrics*, 44:1157–1168.

[Ejigou, 1996] Ejigou, A. (1996). Power and sample size for matched case-control studies. *Biometrics*, 52(3):925–933.

[Ejigou and McHugh, 1977] Ejigou, A. and McHugh, R. (1977). Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics*, 33:552–556.

[Fleiss, 1973] Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions.* Wiley.

[Fleiss, 1988] Fleiss, J. L. (1988). Confidence intervals for the odds ratio in case-control study: the state of the art. *Journal of Chronic Disease*, 32:69–77.

[Fleiss et al., 1980] Fleiss, J. L., Tytun, A., and Ury, H. K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, 36:343–346.

[Franck et al., 1988] Franck, W. E., Hewett, J. E., Islam, M. Z., Wang, S. J., and Cooperstock, M. (1988). A Bayesian analysis suitable for classroom presentation. *The American Statistician*, 42:75–77.

[Gail et al., 1976] Gail, M., Williams, R., Byar, D. P., and Brown, C. (1976). How many controls? *Journal of Chronic Diseases*, 29:723–732.

[Gardner and Altman, 1986] Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than p-values: Estimation rather than hypothesis testing. *British Medical Journal*, 292:746–750.

[Gart and Nam, 1988] Gart, J. J. and Nam, J.-m. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness (c/r: V46 p269-272). *Biometrics*, 44:323–338.

[Gelman et al., 1995] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman & Hall.

[Gill and Wright, 1981] Gill, Philip E., M. W. and Wright, M. (1981). *Practical Optimization*. Academic Press.

[Gill and Murray, 1976] Gill, P. E. and Murray, W. (1976). Minimization subjects to bounds on the variable, npl. Technical report, Report NAC 72, National Physical Laboratory, England.

[Gittins and Pezeshk, 2000a] Gittins, J. and Pezeshk, H. (2000a). A behavioral bayes method for determining the size of a clinical trial. *Drug Information Journal*, 34:355–363.

[Gittins and Pezeshk, 2000b] Gittins, J. and Pezeshk, H. (2000b). How large should a clinical trial be. *The Statistician*, 49:177–187.

[Goodman and Berlin, 1994] Goodman, S. N. and Berlin, J. A. (1994). The use of predicted confidence interval when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, 121:200–206.

[Gordis, 1996] Gordis, L. (1996). *Epidemiology.* W. B. Saunders Company.

[Greenland, 1999] Greenland, S. (1999). A unified approach to the analysis of case-distribution (case-only) studies. *Statistics in Medicine*, 18:1–15.

[Greenland and Thomas, 1982] Greenland, S. and Thomas, D. C. (1982). On the need for the rare disease assumption in case-control studies. *American Journal of Epidemiology*, 116:547–553.

[Hashemi et al., 1997] Hashemi, L., Nandram, B., and Goldberg, R. (1997). Bayesian analysis for a single 2 × 2 table. *Statistics in Medicine*, 16:1311–1328.

[Hoenig and Heisey, 2001] Hoenig, J. M. and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculation for data analysis. *The American Statistician*, 55:19–24.

[Hogue et al., 1986] Hogue, C. J. R., Gaylor, D. W., and Schulz, K. F. (1986). The case-exposure study: A further explication and response to a critique. *American Journal of Epidemiology*, 124:877–883.

[Hora and Kelley, 1983] Hora, S. C. and Kelley, G. D. (1983). Bayesian inference on the odds and risk ratios. *Communications in Statistics, Part A – Theory and Methods*, 12:725–738.

[Jeffreys, 1961] Jeffreys, H. (1961). *Theory of Probability, third edition.* Oxford University Press.

[Jewell, 1984] Jewell, N. P. (1984). Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics*, 40:421–435.

[Jewell, 1986] Jewell, N. P. (1986). On the bias of commonly used measures of association for 2×2 tables (c/r: V45 p1030-1032). *Biometrics*, 42:351–358.

[Johnson et al., 1994] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions. Volume 1 (Second Edition).* Wiley-Interscience.

[Jolson et al., 1992] Jolson, H. M., Bosco, L., Bufton, M. G., Gerstman, B. B., Rinsler, S. S., Williams, E., Flynn, B., Simmons, W. D., V., S. B., and Faich, G. A. (1992). Clustering of adverse drug events: Analysis of risk factors for cerebellar toxicity with high-dose cytarabine. *Journal of the National Cancer Institute*, 84:500–505.

[Joseph and Bélisle, 1997] Joseph, L. and Bélisle, P. (1997). Bayesian sample size determination for normal means and differences between normal means. *The Statistician*, 46:209–226.

[Joseph et al., 1997] Joseph, L., Du Berger, R., and Bélisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16:769–781.

[Joseph et al., 1995] Joseph, L., Wolfson, D. B., and Berger, R. d. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*, 44:143–154.

[Julious and Campbell, 1998] Julious, S. A. and Campbell, M. J. (1998). Sample size calculations for paired or matched ordinal data. *Statistics in Medicine*, 17:1635–1642.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

[Kass and Wasserman, 1996] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules (corr: 1998v93 p412). *Journal of the American Statistical Association*, 91:1343–1370.

[Katsis and Toman, 1999] Katsis, A. and Toman, B. (1999). Bayesian sample size calculations for binomial experiments. *Journal of Statistical Planning and Inference*, 81:349–362.

[Kupper et al., 1981] Kupper, L. L., Karon, J. M., Kleinbaum, D. G., Morgenstern, H., and Lewis, D. K. (1981). Matching in epidemiologic studies: Validity and efficiency considerations. *Biometrics*, 37:271–291.

[Langholz and Clayton, 1994] Langholz, B. and Clayton, D. (1994). Sampling strategies in nested case-control studies. *Environmental Health Perspectives*, 102:47–51.

[Langholz and Thomas, 1990] Langholz, B. and Thomas, D. C. (1990). Nested case-control and case-cohort methods of sampling from a cohort: A critical comparison (corr: V131 p578). *American Journal of Epidemiology*, 131:169–176.

[Langholz and Thomas, 1991] Langholz, B. and Thomas, D. C. (1991). Efficiency of cohort sampling designs: Some surprising results. *Biometrics*, 47:1563–1571.

[Latorre, 1982] Latorre, G. (1982). The exact posterior distribution of the cross-ratio of a 2×2 contingency table. *Journal of Statistical Computation and Simulation*, 16:19–24.

[Leemis and Trivedi, 1996] Leemis, L. M. and Trivedi, K. S. (1996). A comparison of approximate interval estimators for the Bernoulli parameter. *The American Statistician*, 50:63–68.

[Lemeshow et al., 1981] Lemeshow, S., Hosmer, D. W., and Stewart, J. P. (1981). A comparison of sample size determination methods in the two group trial where the underlying disease is rare. *Communications in Statistics, Part B – Simulation and Computation*, 10:437–449.

[Lenth, 2001] Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55:187–193.

[Leonard, 1972] Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika*, 59:581–589.

[Leonard, 1975] Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 37:23–37.

[Leonard, 1977] Leonard, T. (1977). Bayesian simultaneous estimation for several multinomial distributions. *Communications in Statistics, Part A – Theory and Methods*, 6:619–630.

[Liddell et al., 1977] Liddell, F. D. K., McDonald, J. C., and Thomas, D. C. (1977). Methods of cohort analysis: Appraisal by application to asbestos mining (c/r: P483-491). *Journal of the Royal Statistical Society, Series A, General*, 140:469–483.

[Lindley, 1956] Lindley, D. V. (1956). On the measure of information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005.

[Lindley, 1964] Lindley, D. V. (1964). The bayesian analysis of contingency tables. *Annals of Mathematical Statistics*, 35:1622–1643.

[Lindley, 1997] Lindley, D. V. (1997). The choice of sample size (disc: P139-166). *The Statistician*, 46:129–138.

[Lubin and Gail, 1984] Lubin, J. H. and Gail, M. H. (1984). Biased selection of controls for case-control analyses of cohort studies. *Biometrics*, 40:63–75.

[Mantel, 1973] Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics*, 29:479–486.

[Maritz, 1989] Maritz, J. S. (1989). Empirical Bayes estimation of the log odds ratio in 2×2 contingency tables. *Communications in Statistics, Part A – Theory and Methods*, 18:3215–3233.

[Marshall, 1988] Marshall, R. J. (1988). Bayesian analysis of case-control studies (c/r: V8 p1023-1024). *Statistics in Medicine*, 7:1223–1230.

[McCulloch, 1988] McCulloch, R. E. (1988). Information and the likelihood function in exponential families. *The American Statistician*, 42:73–75.

[Meydrech and Kupper, 1978] Meydrech, E. F. and Kupper, L. L. (1978). Cost considerations and sample size requirements in cohort and case-control studies. *American Journal of Epidemiology*, 107:201–205.

[Miettinen, 1976] Miettinen, O. S. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103:226–235.

[More et al., 1980] More, J., Burton, G., and Keneth, H. (1980). User guide for minpack1. Technical report, Argonne National Labs Report ANL-80-74, Argonne, Illinois.

[Mukerjee and Dey, 1993] Mukerjee, R. and Dey, D. K. (1993). Frequentist validity of posterior quantiles in the presence of a nuisance parameter: Higher order asymptotics. *Biometrika*, 80:499–505.

[Müller, 1998] Müller, P. (1998). Simulation based optimal design. *Bayesian Statistics 6*, 90:1322–1330.

[Müller and Parmigiani, 1995] Müller, P. and Parmigiani, G. (1995). Optimal design via curve fitting of Monte Carlo experiments. *Journal of the American Statistical Association*, 90:1322–1330.

[Nam and Fears, 1992a] Nam, J.-M. and Fears, T. R. (1992a). Control sample size when cases are given in constant ratio stratum-matched case-control studies. *Statistics in Medicine*, 11:1759–1766.

[Nam and Fears, 1992b] Nam, J.-M. and Fears, T. R. (1992b). Optimum sample size determination in stratified case-control studies with cost considerations. *Statistics in Medicine*, 11:547–556.

[Nelder and Murray, 1965] Nelder, J. A. and Murray, R. (1965). A simplex method for function minimization. *Computer Journal*, 7:308–313.

[Nurminen and Mutanen, 1987] Nurminen, M. and Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scandinavian Journal of Statistics*, 14:67–77.

[O'Neill, 1984] O'Neill, R. T. (1984). Sample sizes for estimation of the odds ratio in unmatched case-control studies. *American Journal of Epidemiology*, 120:145–153.

[Parker and Bregman, 1986] Parker, R. A. and Bregman, D. J. (1986). Sample size for individually matched case-control studies. *Biometrics*, 42:919–926.

[Peers, 1968] Peers, H. W. (1968). Confidence properties of Bayesian interval estimates. *Journal of the Royal Statistical Society, Series B, Methodological*, 30:535–544.

[Pezeshk and Gittins, 1999] Pezeshk, H. and Gittins, J. (1999). Sample size determination in clinical trials. *Student*, 3:19–26.

[Pham-Gia and Turkkan, 1992] Pham-Gia, T. and Turkkan, N. (1992). Sample size determination in Bayesian analysis (disc: P399-404). *The Statistician*, 41:389–397.

[Piegorsch et al., 1994] Piegorsch, W. W., Weinberg, C. R., and A., T. J. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13:153–162.

[Piessens et al., 1983] Piessens, R. E., deDoncher Kapenga, C. W. U., and Kahaner, D. K. (1983). *QUADPACK*. Springer-Verlag.

[Prentice and Breslow, 1978] Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, 65:153–158.

[Robins et al., 1989] Robins, J. M., Prentice, R. L., and Blevins, D. (1989). Designs for synthetic case-control studies in open cohorts. *Biometrics*, 45:1103–1116.

[Rosenberg et al., 2000] Rosenberg, L., Joseph, L., and Alan, B. (2000). *Surgical Arithmetic: Epidemiological Statistical and Outcome-Based Approach to Surgical Practice. 1st Edition.* Landes Bioscience.

[Ross, 1996] Ross, S. M. (1996). Bayesians should not resample a prior sample to learn about the posterior. *The American Statistician*, 50:116–116.

[Rothman, 1986] Rothman, K. J. (1986). *Epidemiology: an Introduction. 1st Edition.* Oxford University Press.

[Rothman and Greenland, 1998] Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology. 2nd Edition.* Lippincott Williams and Wilkins Publishers.

[Royston, 1993] Royston, P. (1993). Exact conditional and unconditional sample size for pair-matched studies with binary outcome: A practical guide. *Statistics in Medicine*, 12:699–712.

[Sato, 1992a] Sato, T. (1992a). Estimation of a common risk ratio in stratified case-cohort studies. *Statistics in Medicine*, 11:1599–1605.

[Sato, 1992b] Sato, T. (1992b). Maximum likelihood estimation of the risk ratio in case-cohort studies. *Biometrics*, 48:1215–1221.

[Satten and Kupper, 1990] Satten, G. A. and Kupper, L. L. (1990). Sample size requirements for interval estimation of the odds ratio. *American Journal of Epidemiology*, 131:177–184.

[Schlesselman, 1982] Schlesselman, J. J. (1982). *Case-control Studies: Design, Conduct, Analysis.* Oxford University Press.

[Schmidt and Schaid, 1999] Schmidt, S. and Schaid, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *American Journal of Epidemiology*, 150:878–885.

[Schork and Williams, 1980] Schork, M. A. and Williams, G. W. (1980). Number of observations required for the comparison of two correlated proportions. *Communications in Statistics, Part B – Simulation and Computation*, 9:349–357.

[Severini, 1991] Severini, T. A. (1991). On the relationship between Bayesian and non-bayesian interval estimates. *Journal of the Royal Statistical Society, Series B, Methodological*, 53:611–618.

[Shao, 1989] Shao, J. (1989). Monte Carlo approximations in Bayesian decision theory. *Journal of the American Statistical Association*, 84:727–732.

[Silvey, 1980] Silvey, S. D. (1980). *Optimal Design: An Introduction to the Theory for Parameter.* Chapman & Hall.

[Smith, 1975] Smith, A. F. M. (1975). A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62:407–416.

[Spiegelhalter et al., 1994] Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials (disc: P387-416). *Journal of the Royal Statistical Society, Series A, General*, 157:357–387.

[Srivastava and Wu, 1993] Srivastava, M. S. and Wu, Y. (1993). Local efficiency of moment estimators in beta-binomial model. *Communications in Statistics, Part A – Theory and Methods*, 22:2471–2490.

[Stallard, 1998] Stallard, N. (1998). Sample size determination for phase Ii clinical trials based on Bayesian decision theory. *Biometrics*, 54:279–294.

[Thomas, 1977] Thomas, D. C. (1977). Addendum to "methods of cohort analysis: Appraisal by application to asbestos mining". *Journal of the Royal Statistical Society, Series A, General*, 140:483–485.

[Tsutakawa and Lin, 1986] Tsutakawa, R. K. and Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, 51:251–267.

[Umbach and Weinberg, 1997] Umbach, D. M. and Weinberg, C. R. (1997). Designing and analysing case-control studies to exploit independence of genotype and exposure. *Statistics in Medicine*, 16:1731–1743.

[Walker, 1976] Walker, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, 32:829–849.

[Walker, 1977] Walker, S. D. (1977). Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *American Journal of Epidemiology*, 105:387–397.

[Walker and Cook, 1991] Walker, S. D. and Cook, R. J. (1991). A comparison of several point estimators of the odds ratio in a single 2 by 2 contigency table. *Biometrics*, 47:795–811.

[Walters, 1997] Walters, D. E. (1997). Confidence limits and log-odds ratios. *The Statistician*, 46:433–438.

[Weinberg and Wacholder, 1990] Weinberg, C. R. and Wacholder, S. (1990). The design and analysis of case-control studies with biased sampling. *Biometrics*, 46:963–975.

[Weiss et al., 2001] Weiss, S., Silverman, E., and Palmer, L. (2001). Guess editorial: Case-control association studies in pharmacogenetics. *The Pharmacogenomics Journal*, 1:157–158.

[Welch and Peers, 1963] Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *Statistical Methods in Medical Research*, 4:311–337.

[Wickramaratne, 1995] Wickramaratne, P. J. (1995). Sample size determination in epidemiologic studies. *Statistical Methods in Medical Research*, 4:311–337.

[Wilkins and Sinks, 1990] Wilkins, J. R. and Sinks, T. (1990). Parental occupation and intracranial neoplasms of chilhood: Results of a case-control interview study. *American Journal of Epidemiology*, 132:275–291.

[Yang et al., 1997] Yang, Q., Khoury, M. J., and Flanders, W. D. (1997). Sample size requirements in case-only designs to detect gene-environment interaction. *American Journal of Epidemiology*, 146:713–720.

[Zelen and Parker, 1986] Zelen, M. and Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statistics in Medicine*, 5:261–269.

[Zhang, 1996] Zhang, S. J. J. (1996). *Computation of Special Functions.* John Wiley.

[Zocchi and Atkinson, 1999] Zocchi, S. S. and Atkinson, A. C. (1999). Optimum experimental designs for multinomial logistic models. *Biometrics*, 55:437–444.