Sample Size for Testing Homogeneity of Two a Priori Dependent Binomial Populations Using the Bayesian Approach

ATHANASSIOS KATSIS[†]   a.katsis@aegean.gr

Department of Statistics and Actuarial Science, University of the Aegean, 68 Dikearhou Street, Athens, 116 36, GREECE

Abstract. This paper establishes new methodology for calculating the optimal sample size when a hypothesis test between two binomial populations is performed. The problem is addressed from the Bayesian point of view, with prior information expressed through a Dirichlet distribution. The approach of this paper sets an upper bound for the posterior risk and then chooses as 'optimum' the combined sample size for which the likelihood of the data does not satisfy this bound. The combined sample size is divided equally between the two binomials. Numerical examples are discussed for which the two proportions are equal to either a fixed or to a random value.

Keywords: Optimal sample size, Dependent prior proportions, Dirichlet distribution, Bayesian point of view.

1. Introduction

The size of the sample required to perform an experiment is a topic of particular interest in statistical theory. The researcher does not wish to compromise the validity of her/his findings by collecting too small a sample or unnecessarily consume the available resources on an excessively large sample. Furthermore, the issue of hypothesis testing between two binomial populations has a considerable range of practical applications. Such examples include a comparison of the proportions of male births in two human populations, testing the unemployment rates between inner city and suburban areas, or comparing the satisfaction levels amongst clients from two branches of a bank.

In the first case, the population proportions of male births may be tested against a fixed value p, usually obtained through a mathematical model under study. In the unemployment example, it would be reasonable to assume that the researcher is mostly interested in verifying that a certain

---

† Requests for reprints should be sent to Athanassios Katsis, Department of Statistics and Actuarial Science, University of the Aegean, 68 Dikearhou Street, Athens, 116 36, GREECE.

nationwide target level for unemployment is met rather than estimating any differences. These are essentially goodness-of-fit tests. In the last example, the satisfaction proportions of the bank's different branches are compared to the satisfaction rate of its headquarters, itself a random variable. In all of the above cases it is not unreasonable to expect that the prior knowledge for the population proportions depend on each other. The classical theory, through the use of the normal approximation to the binomial, utilizes the power function as the sole source of inference to derive the unknown sample size n.

In Bayesian methodology, posterior accuracy is employed to express experimental precision. DasGupta and Vidakovic in [4] quantify this approach in the context of one way ANOVA. In hypothesis testing a decision α_i is correct if $\theta \in \Theta_i$, which corresponds to H_i $(i = 0, 1)$. Under the "$0 - 1$" loss function: $L(\theta, \alpha_i) = \{ \begin{array}{l} 0, \theta \in \Theta_i \\ 1, \theta \in \Theta_j \end{array}$, $(i \neq j)$, the posterior risk is given by $\min\{P(H_0|y), P(H_1|y)\}$ where $y = (y_1, \ldots, y_n)$ denotes the data vector. It must not exceed a pre-specified bound, that is,

$$\min\{P(H_0|y), P(H_1|y)\} \leq \varepsilon \tag{1}$$

Furthermore, we must account for the fact that this is a pre-experimental procedure, hence the data are unknown beforehand. Thus if there are data y that do not satisfy (1), it must so happen with a small likelihood, that is,

$$P(y \in T^c) < \delta \tag{2}$$

where $K = T^c$ is the set of all data not satisfying (1), δ is a constant and the probability in (2) is calculated based on the marginal distribution of the data.

In the area of one binomial population, Adcock in [1], [2] and [3] suggested a tolerance region, which includes the parameters of a multinomial distribution with certain likelihood. Pham-Gia and Turkkan in [6] derived sample sizes for a binomial proportion by setting precision bounds on the posterior variance and Bayes risk while Joseph et al. in [5] and Rahme and Joseph in [7] fixed either the length or the probability of coverage for intervals to obtain the required sample sizes.

In this paper, we obtain the optimal sample size to compare two binomial populations when a priori the proportions are dependent through a Dirichlet distribution. The structure is as follows. In Section 2, we consider two cases: The first one is a test of hypothesis $H_0 : p_1 = p_2 = p$, where p is a specified constant, the other a test $H_0 : p_1 = p_2$. Based on (1) and (2) the optimal sample size is derived. In Section 3, the results for specific cases are presented.

2. Hypothesis Testing

Let Y_i, $i = 1, 2$ be binomial random variables with parameters n_i, p_i. Our goal is to compute the optimal sample size $n = n_1 + n_2$, for conducting the test $H_0 : p_1 = p_2 = p$ vs $H_1 : p_1 \neq p_2$. We shall consider the case where p is a fixed constant and also the case where p is not specified but follows a Beta distribution. Each binomial distribution will have an equal number of observations, that is, $n_i = \frac{n}{2}$. If n_i is not an integer, we can substitute it with $[n_i]$, which is the highest integer less than n_i. The prior probabilities of H_0 and H_1 are π_0 and π_1 respectively. The prior information on the proportions is summarized by a Dirichlet distribution with parameters $\lambda_i \geq 0$, $i = 1, 2, 3$, that is,

$$h(p_1, p_2) = \frac{\Gamma(\sum_{i=1}^{3} \lambda_i)}{\Pi_{i=1}^{3} \Gamma(\lambda_i)} p_1^{\lambda_1 - 1} p_2^{\lambda_2 - 1} (1 - p_1 - p_2)^{\lambda_3 - 1}$$

The formulation of the Dirichlet prior relies on $p_1 + p_2$ being no greater than 1. If this is not the case, we could always reparametrise in terms of $1 - p_1$ and $1 - p_2$. As expressed by (1) and (2), the posterior risk is bounded from above accounting for the fact that the data are unknown.

2.1. Case 1: p is fixed constant

2.1.1. Theoretical results

We shall first examine the case where the proportion p is constant. Under the null hypothesis the posterior density of p_1 and p_2 is given by:

$$g(p_1, p_2 | y) = \frac{\Pi_{i=1}^{2} \binom{n_i}{y_i} p^{y_1 + y_2} q^{n - y_1 - y_2} \pi_0}{m(y)}$$

Similarly, under the alternative hypothesis, the joint posterior density of p_1 and p_2 is given by:

$$g(p_1, p_2 | y) = \frac{\left(\Pi_{i=1}^{2} \binom{n_i}{y_i} p_i^{y_i} q_i^{n_i - y_i}\right) \frac{\Gamma(\sum_{i=1}^{3} \lambda_i)}{\Pi_{i=1}^{3} \Gamma(\lambda_i)} p_1^{\lambda_1 - 1} p_2^{\lambda_2 - 1} (1 - p_1 - p_2)^{\lambda_3 - 1} \pi_1}{m(y)}$$

where $q_i = 1 - p_i$, $q = 1 - p$ and

$$\begin{aligned}
m(y) &= \prod_{i=1}^{2} \binom{n_i}{y_i} p^{y_1 + y_2} q^{n - y_1 - y_2} \pi_0 + \prod_{i=1}^{2} \int_0^1 \binom{n_i}{y_i} p_i^{y_i} q_i^{n_i - y_i} \\
&\quad \frac{\Gamma(\sum_{i=1}^{3} \lambda_i)}{\Pi_{i=1}^{3} \Gamma(\lambda_i)} p_1^{\lambda_1 - 1} p_2^{\lambda_2 - 1} (1 - p_1 - p_2)^{\lambda_3 - 1} \pi_1 dp_i
\end{aligned}$$

where the Bayes factor B is given by:

$$\frac{p^{y_1+y_2}q^{n-y_1-y_2}\prod_{i=1}^{3}\Gamma(\lambda_i)}{\Gamma(\sum_{i=1}^{3}\lambda_i)\int_0^1\int_0^1\prod_{i=1}^{2}p_i^{y_i+\lambda_i-1}(1-p_i)^{n_i-y_i}(1-p_1-p_2)^{\lambda_3-1}dp_1dp_2} \quad (3)$$

To find the set $K = T^c$ where (1) is not satisfied, we note that

$$P(H_0|y) = \frac{(\pi_0/\pi_1)B}{1+(\pi_0/\pi_1)B}$$

Hence

$$\min\{P(H_0|y), P(H_1|y)\} = P(H_0|y)\mathbf{I}(B \leq \frac{\pi_1}{\pi_0}) + P(H_1|y)\mathbf{I}(B \geq \frac{\pi_1}{\pi_0})$$

where $\mathbf{I}(\cdot)$ is the indicator function that takes the value 1 if the condition inside the parenthesis is satisfied, 0 otherwise. So the set K is given by:

$$K = \{y : \frac{\epsilon}{1-\epsilon}\frac{\pi_1}{\pi_0} < B < \frac{1-\epsilon}{\epsilon}\frac{\pi_1}{\pi_0}\} \quad (4)$$

It is important that the probability of this set calculated with respect to the marginal distribution of the data converge to zero as the sample size increases to infinity.

THEOREM 1 *The probability of the set K with respect to the marginal distribution of the data approaches zero as n approaches infinity: $P(K) \to 0$, as $n \to \infty$.*

Proof: This is immediate if we prove that the Bayes factor B converges either to 0 or to ∞. The proof of the latter is presented in the Appendix.

2.1.2.   Sample size calculations

To derive the exact sample size we set $P(K) = \delta$, or in a more explicit form:

$$P\{\frac{\epsilon}{1-\epsilon}\frac{\pi_1}{\pi_0} < B < \frac{1-\epsilon}{\epsilon}\frac{\pi_1}{\pi_0}\} = \delta \quad (5)$$

Assuming that $\lambda_3 - 1$ is a positive integer (if not, it can be replaced by $[\lambda_3 - 1]$), Newton's binomial formula yields:

$$(1-p_1-p_2)^{\lambda_3-1} = \sum_{k=0}^{\lambda_3-1}\binom{\lambda_3-1}{k}(1-p_1)^k p_2^{\lambda_3-k-1}(-1)^{\lambda_3-k-1} \quad (6)$$

Based on (6), the double integral in the denominator of (3), is expressed in the following way:

$$
\begin{aligned}
F(y_1, y_2) &= \sum_{k=0}^{\lambda_3-1} \binom{\lambda_3-1}{k}(-1)^{\lambda_3-k-1} \int_0^1 p_1^{y_1+\lambda_1-1}(1-p_1)^{n_1+k-y_2}dp_1 \\
&\quad \int_0^1 p_2^{y_2+\lambda_2+\lambda_3-k-2}(1-p_2)^{n_2-y_2}dp_2 \\
&= \Gamma(y_1+\lambda_1)\Gamma(n_2-y_2+1)\sum_{k=0}^{\lambda_3-1}\binom{\lambda_3-1}{k}(-1)^{\lambda_3-k-1} \\
&\quad \frac{\Gamma(n_1+k-y_1+1)\Gamma(n_2-y_2+1)}{\Gamma(\lambda_1+n_1+k+1)\Gamma\lambda_2+\lambda_3+n_2-k}
\end{aligned}
\tag{7}
$$

Combining (3), (5) and (7), the optimal sample size is derived by solving the following equation for n.

$$
P(K) = P[W_1 < \frac{p^{y_1+y_2}}{q^{y_1+y_2}F(y_1, y_2)} < W_2] = \delta
$$

where $W_1 = \frac{\epsilon}{1-\epsilon}\frac{\pi_1}{\pi_0}\frac{\Gamma(\sum_{i=1}^3 \lambda_i)}{q^n \prod_{i=1}^3 \Gamma(\lambda_i)}$ and $W_2 = \frac{1-\epsilon}{\epsilon}\frac{\pi_1}{\pi_0}\frac{\Gamma(\sum_{i=1}^3 \lambda_i)}{q^n \prod_{i=1}^3 \Gamma(\lambda_i)}$

Alternatively, by multiplying with the necessary constants, the double integral in (3) could have been expressed as $E[\prod_{i=1}^2 (1-p_i)^{n_i-y_i}]$, the expected value of $(1-p_1)^{n_1-y_1}(1-p_2)^{n_2-y_2}$ with respect to the Dirichlet distribution with parameters $y_1+\lambda_1, y_2+\lambda_2, \lambda_3$. In the latter case, the equation yielding the sample size would have been the following:

$$
P(K) = P[W_1 < \frac{p^{y_1+y_2}\Gamma(y_1+y_2+\lambda_1+\lambda_2+\lambda_3)}{q^{y_1+y_2}\prod_{i=1}^2 \Gamma(y_i+\lambda_i)E[\prod_{i=1}^2(1-p_i)^{n_i-y_i}]} < W_2] = \delta
$$

2.2.   Case 2: p is a random variable

We shall turn our attention to the hypothesis testing of $H_0 : p_1 = p_2 = p$ where there is no special interest in the value p. Therefore a priori we have that $p \sim Beta(\alpha, \beta)$ and the prior information on the proportions p_i's is still described through a Dirichlet distribution with parameters $\lambda_i \geq 0$, $i = 1, 2, 3$. Similarly, the posterior odds are given by $\frac{P(H_0|y)}{P(H_1|y)} = B\frac{\pi_0}{\pi_1}$ where the Bayes factor B is given by:

$$
B = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y_1+y_2+\alpha)}{\Gamma(\alpha+\beta+n)}\Gamma(n+\beta-y_1-y_2)\prod_{i=1}^3 \Gamma(\lambda_i)}{\Gamma(\sum_{i=1}^3 \lambda_i)\int_0^1\int_0^1 \prod_{i=1}^2 p_i^{y_i+\lambda_i-1}(1-p_i)^{n_i-y_i}(1-p_1-p_2)^{\lambda_3-1}dp_1 dp_2}
$$

Using the same methodology as in the previous case we obtain the set $K = \{y : \frac{\epsilon}{1-\epsilon}\frac{\pi_1}{\pi_0} < B < \frac{1-\epsilon}{\epsilon}\frac{\pi_1}{\pi_0}\}$. Again, since as $n \to \infty$, B converges to either 0 or ∞, we establish that the marginal probability of this set converges to zero as the sample size increases. Like before, the optimal sample size is derived by solving the following equation for n.

$$P(K) = P[\frac{\epsilon}{1-\epsilon}\frac{\pi_1}{\pi_0}C < \frac{\Gamma(y_1 + y_2 + \alpha)\Gamma(n + \beta - y_1 - y_2)}{F(y_1, y_2)} < \frac{1-\epsilon}{\epsilon}\frac{\pi_1}{\pi_0}C]$$

where $C = \frac{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)\Gamma(\sum_{i=1}^{3})(\lambda_i)}{\Gamma(\alpha+\beta)\prod_{i=1}^{3}\Gamma(\lambda_i)}$.

3. Numerical Results

In this section we present the results of the previously developed methodology, using specific examples. For a given value of n, points of the random variables y_1 and y_2 are generated and the probability $P(K)$ is obtained. The algorithm is based on the calculation of the double integral presented in Section 2.1.2. The optimal sample size is derived when the condition $P(K) = \delta$ is satisfied. The program is written in SAS and is available upon request.

Numerical investigation has shown that for given values of the parameters, a proper value of n exists in all the cases that have been examined. However, for specific values of δ, n can be very small which is of no practical significance to the researcher.

Table 1 summarizes the results for the case when p is a specified constant. The parameters p and λ_i are given in the table. We have set $\pi_0 = 0.5$, $\epsilon = 0.1$ and $\delta = 0.3$. The value of δ represents a measure of the likelihood of the data. As δ increases, the researcher is less strict about the existence of "undesirable data" (that is, data not satisfying (1)) and therefore the optimal sample size is expected to decrease.

One important factor affecting the sample size is the distance between the prior mean values and the fixed value p. Generally, for a fixed value of the precision constant δ, a larger sample size is needed to detect any differences when the average prior proportions are closer to the specified constant p, than when they are further apart. The following cases illustrate the above argument. When p is equal to 0.2, the average prior values are equal to 0.2 for both binomial populations and the covariance between the two prior proportions takes the value -0.0019, the optimal sample size needed to reach the specified precision is 90. On the other hand, if p is equal to 0.1, 0.4, 0.5 or 0.6 and the average prior values and the covariance remain the same, the required sample size reduces to 76, 40, 18 or 12 respectively. This specific

Table 1. Sample sizes for the case where p is fixed. ($\epsilon = 0.10$, $\delta = 0.30$).

p	$E(p_1)$	$E(p_2)$	$Cov(p_1, p_2)$	λ_1	λ_2	λ_3	n
0.1	0.2	0.2	-0.0019	4	4	12	76
0.2	0.2	0.2	-0.0019	4	4	12	90
0.2	0.2	0.2	-0.0036	2	2	6	50
0.2	0.2	0.2	-0.0067	1	1	3	36
0.2	0.2	0.3	-0.0054	2	3	5	50
0.2	0.2	0.1	-0.00095	4	2	14	60
0.2	0.3	0.3	-0.0082	3	3	4	44
0.2	0.2	0.5	-0.0047	4	10	6	20
0.5	0.2	0.2	-0.0036	2	2	6	16
0.5	0.2	0.2	-0.0019	4	4	12	18
0.5	0.2	0.2	-0.0067	1	1	3	12
0.5	0.4	0.4	-0.0076	8	8	4	52
0.5	0.4	0.4	-0.0145	4	4	2	36
0.5	0.4	0.4	-0.027	2	2	1	22
0.8	0.4	0.4	-0.0145	8	8	4	10
0.6	0.2	0.2	-0.0019	4	4	12	12
0.6	0.4	0.4	-0.0145	4	4	2	24
0.6	0.4	0.4	-0.0076	8	8	4	28
0.6	0.4	0.4	-0.027	2	2	1	18
0.4	0.2	0.2	-0.0019	4	4	12	40
0.4	0.4	0.4	-0.0145	4	4	2	36
0.4	0.4	0.4	-0.0076	8	8	4	56
0.4	0.4	0.4	-0.027	2	2	1	24

trend is depicted in Figure 1 where the common expected value of p_1 and p_2 is defined as $p^\star$.

The same pattern holds for different values of $E(p_1)$ and $E(p_2)$. The smaller this distance is, the larger is the sample size required to distinguish any significant difference. Compare, for example, the cases when $E(p_1) = 0.2$, $E(p_2) = 0.3$ and $E(p_1) = 0.2$, $E(p_2) = 0.5$. The value of p remains unchanged at 0.2 and the covariance is approximately the same. In the former case, the required sample size is 50, whereas in the latter situation a considerably smaller sample, ($n = 20$), is needed .

Yet another factor that determines the sample size is the covariance between the prior proportions. A strong negative value indicates that a priori the proportions are negatively correlated. Therefore, when one of them

a priori assumes a large(small) value, the other proportion will have a small(large) prior value. Hence, a small sample is required to establish H_1. This is highlighted in the case of $p = 0.5$, $E(p_1) = E(p_2) = 0.4$. Three different sets of prior parameters yield the following values for the covariance: -0.0076, -0.0145 and -0.027. The respective sample sizes are 52, 36 and 22. This comes to no surprise since the increasing negative correlation between the two prior proportions enables the researcher to use an exceedingly smaller sample to detect any differences.

Table 2 provides the optimal sample sizes for the cases when the null hypothesis is of the form: $H_0 : p_1 = p_2$. The common value p is a noninformative Beta prior distribution with parameters $\alpha = \beta = 1$. In most cases, the sample sizes tend to be higher than in the previous case of a specified constant. This is evident in the case of $E(p_1) = E(p_2) = 0.2$ where, for the same sets of the prior parameters, the resulting sample sizes (40, 52 and 30) are consistently higher than in the case of $p = 0.5$ (16, 18 and 12 respectively). This should be expected since the prior specifications on p exhibit a higher degree of uncertainty thus requiring larger sample sizes.

Table 2. Sample sizes for the case where p is random. ($\epsilon = 0.10$, $\delta = 0.30$).

$E(p)$	$Var(p)$	$E(p_1)$	$E(p_2)$	$Cov(p_1, p_2)$	λ_1	λ_2	λ_3	n
0.5	0.008	0.2	0.2	-0.0036	2	2	6	40
0.5	0.008	0.2	0.2	-0.0019	4	4	12	52
0.5	0.008	0.2	0.2	-0.0067	1	1	3	30
0.5	0.008	0.4	0.2	-0.0145	4	4	2	32
0.5	0.008	0.4	0.2	-0.0076	8	8	4	46
0.5	0.008	0.4	0.2	-0.027	2	2	1	24
0.5	0.008	0.1	0.2	-0.00047	2	2	16	54
0.5	0.008	0.3	0.2	-0.0082	3	3	4	38
0.5	0.008	0.2	0.3	-0.0054	2	3	5	40
0.5	0.008	0.2	0.5	-0.009	2	5	3	18

The covariance of the prior proportions largely determines the optimal sample size. The stronger the two proportions are negatively correlated, the smaller the sample needed to detect a difference. The distance between the average proportions and p, as well as the distance between the average proportions, do not seem to be important factors in this case.

In summary, this methodology constitutes a Bayesian way of establishing the solution to the problem of sample size determination for hypothesis
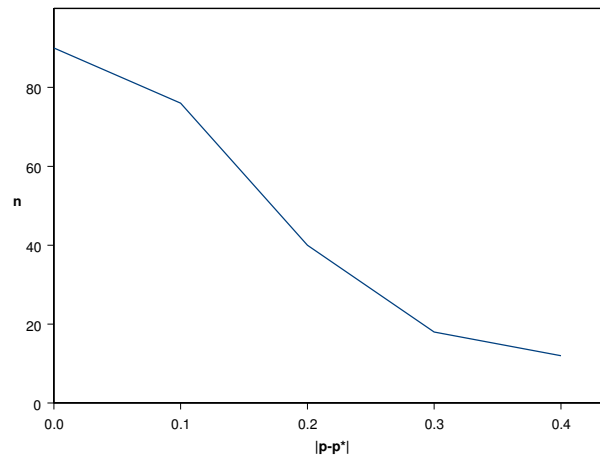
Figure 1. Optimal Sample size n versus the distance between various fixed p values and p^* ($\epsilon = 0.10$, $\delta = 0.30$, $Cov(p_1, p_2) = -0.0019$).

testing among two binomial proportions with dependent prior parameters. It can be applied when the comparison is against a fixed value (essentially a goodness-of-fit situation) or simply between two populations. The basic factors determining the sample size include the prior covariance and, in the case of a fixed p, the distance between the prior mean values and p.

Acknowledgments

We wish to thank the Area Editor and two referees for their very helpful comments that greatly improved the initial manuscript.

References

1. C.J. Adcock. A Bayesian approach to calculating sample sizes. *Statistician*, 36:155–159, 1987.
2. C.J. Adcock. A Bayesian approach to calculating sample sizes for multinomial sampling. *Statistician*, 37:433–439, 1988.
3. C.J. Adcock. Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling-some commments on the paper by Pham-Gia and Turkkan. *Statistician*, 41:399–404, 1992.

4. A. DasGupta and B. Vidakovic. Sample size problems in ANOVA: Bayesian point of view. *Journal of Statistical Planning and Inference*, 65:335–347, 1997.

5. L. Joseph, D. Wolfson and D.B. Berger. Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician*, 44:143–154, 1995.

6. T. Pham-Gia and N. Turkan. Sample size determination in Bayesian analysis. *Statistician*, 41:389–397, 1992.

7. E. Rahme and L. Joseph. Exact sample size determination for binomial experiments. *Journal of Statistical Planning and Inference*, 66:83–93, 1998.

8. G. Schwarz. A Sequential Student Test. *The Annals of Mathematical Statistics*, 42:1003–1009, 1971.

Appendix

Proof of the fact that B converges either to 0 or to ∞

It suffices to show that in general the posterior odds converge either to 0 or to ∞. We shall use a result by Schwarz in [8] adapted to our case. If $n \to \infty$, then

$$\ln P(H_0|y) = -n \ln \lambda_0 + C_0 \ln \ln n + O(1)$$

where $\lambda_0 = \frac{\max_\Omega f(y|\theta)}{\max_{\Theta_0} f(y|\theta)}$ and C_0 is a constant in terms of n and depends on the dimensions of Ω and Θ_0. Similarly for $P(H_1|y)$ we obtain the following:

$$\ln P(H_1|y) = -n \ln \lambda_1 + C_1 \ln \ln n + O(1)$$

Hence, the difference of the above logarithmic expressions is the natural logarithm of the posterior odds, that is,

$$\ln \frac{P(H_0|y)}{P(H_1|y)} = n \ln \frac{\max_{\Theta_0} f(y|\theta)}{\max_{\Theta_1} f(y|\theta)} + (C_0 - C_1) \ln \ln n + O(1) \qquad (A.1)$$

The limit of (A.1) is dictated by the limit of $n \ln \frac{\max_{\Theta_0} f(y|\theta)}{\max_{\Theta_1} f(y|\theta)}$. If $\ln \frac{\max_{\Theta_0} f(y|\theta)}{\max_{\Theta_1} f(y|\theta)} > 0$ (meaning that $\max_{\Theta_0} f(y|\theta) > \max_{\Theta_1} f(y|\theta)$), then $\ln \frac{P(H_0|y)}{P(H_1|y)} \to \infty$ which shows that the posterior odds (and hence B as well) converge to ∞. Similarly if $\ln \frac{\max_{\Theta_0} f(y|\theta)}{\max_{\Theta_1} f(y|\theta)} < 0$ (meaning that $\max_{\Theta_0} f(y|\theta) < \max_{\Theta_1} f(y|\theta)$), then $\ln \frac{P(H_0|y)}{P(H_1|y)} \to -\infty$. In this case the posterior odds and subsequently the Bayes factor converge to 0.