Optimal Sample Size for Two Binomials

CHENGCHANG LI, PETER MÜELLER & GIOVANNI PARMIGIANI

Institute of Statistics and Decision Sciences, Duke University

93-A25

Summary

In an experiment aims at comparing two success probabilities $\theta_1$ and $\theta_2$ in two binomial distributions, to choose the optimal sample size is a primary consideration in our design. In this paper, we derive the procedure of choosing the optimal sample size in three different type of utility functions – quadratic loss, hypothesis test power and experiment information introduced by Lindley (1956). Independent prior distributions for the $\theta_1$ and $\theta_2$ is assumed in this paper.
1. Introduction

In an experiment for comparing two treatments with dichotomous responses, the goal is to get information about the two success probabilities \( \theta_1 \) and \( \theta_2 \), and compare \( \theta_1 \) and \( \theta_2 \). Assume we have same sample size \( n \) for both treatments. Sufficient statistics for \( \theta_1 \) and \( \theta_2 \) are the number of success \( x_1 \) and \( x_2 \) in the two treatment with distribution

\[
p(x_i|\theta_i, n) = C^n x_i \theta_i^{x_i} (1 - \theta_i)^{n-x_i}, \quad i = 1, 2
\]

To choose the optimal sample size is our primary consideration in our design. Depends on circumstances, we might be interested in the difference of \( \theta_1 \) and \( \theta_2 \), or whether \( \theta_1 \) and \( \theta_2 \) are close enough (say \(|\theta_1 - \theta_2| < \delta\)), or the experiment information (Lindley 1956). We will derive the procedures of choosing optimal sample sizes in these three cases in the following sections.

In this paper, we assume \( \theta_1 \) and \( \theta_2 \) have independent beta\((a_1, b_1)\) and beta\((a_2, b_2)\) prior distributions. Then their posterior distributions are independent too, and the posterior distribution of \( \theta_i \) is beta\((a_i + x_i, n + b_i - x_i)\), \( i = 1, 2 \).

The marginal distribution of \( x_i \) is

\[
m(x_i) = \int_0^1 p(\theta_i, x_i) d\theta_i
\]

\[
= \int_0^1 p(\theta_i) p(x_i|\theta_i) d\theta_i
\]

\[
= \int_0^1 \theta_i^{x_i-1} (1 - \theta_i)^{b_i-1} \frac{C^n x_i \theta_i^{x_i} (1 - \theta_i)^{n-x_i} d\theta_i}{B(a_i, b_i)}
\]

\[
= \frac{C^n x_i B(a_i + x_i, n + b_i - x_i)}{B(a_i, b_i)}
\]

\[
= \frac{B(a_i + x_i, n + b_i - x_i)}{(n + 1)B(x_i + 1, n - x_i + 1)B(a_i, b_i)}
\]
2. Quadratic Loss

In this section, we are interested in the difference of $\theta_1$ and $\theta_2$: $\theta_1 - \theta_2$. We use posterior mean $\hat{\theta}_i = \frac{\bar{x}_i + a_i}{n + b_i + a_i}$ of $\theta_i$ to estimate $\theta_i$, $i = 1, 2$. $\hat{\theta}_1 - \hat{\theta}_2$ is used to estimate $\theta_1 - \theta_2$.

We consider quadratic loss here, the loss function is

$$L(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2) = [\hat{\theta}_1 - \hat{\theta}_2 - (\theta_1 - \theta_2)]^2$$

let $c$ be the cost for each observation, then the utility of the experiment is $-L(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2) - 2nc$

The expected utility of the experiment is

$$U(n) = \mathbb{E}_x \mathbb{E}_{\theta|x}(-L(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2)) - 2nc$$

$$= -\mathbb{E}_x[\mathbb{E}_{\theta|x_1}(\hat{\theta}_1 - \theta_1)^2 + \mathbb{E}_{\theta|x_2}(\hat{\theta}_2 - \theta_2)^2] - 2nc$$

$$= -\mathbb{E}_{x_1}[\frac{(a_1 + x_1)(n + b_1 - x_1)}{(n + a_1 + b_1)^2(n + a_1 + b_1 + 1)}] - \mathbb{E}_{x_2}[\frac{(a_2 + x_2)(n + b_2 - x_2)}{(n + a_2 + b_2)^2(n + a_2 + b_2 + 1)}] - 2nc$$

$$= - \sum_{i=1}^{2} \sum_{x=0}^{n} \frac{(a_i + x)(n + b_i - x)C_{a_i}^n B(a_i + x, n + b_i - x)}{(n + a_i + b_i)^2(n + a_i + b_i + 1)B(a_i, b_i)} - 2nc$$

$$= - \frac{1}{n + 1} \sum_{i=1}^{2} \sum_{x=0}^{n} \frac{(a_i + x)(n + b_i - x)B(a_i + x, n + b_i - x)}{(n + a_i + b_i)^2(n + a_i + b_i + 1)B(a_i, b_i)B(x + 1, n - x + 1)} - 2nc$$

for $n = 0$

$$U(0) = - \sum_{i=1}^{2} \frac{a_i b_i B(a_i, b_i)}{(a_i + b_i)^2(a_i + b_i + 1)B(a_i, b_i)B(1, 1)}$$
\[
\sum_{i=1}^{2} \frac{a_i b_i}{(a_i + b_i)^2 (a_i + b_i + 1)}
\]

for \( n = 1 \)

\[
U(1) = \frac{1}{2} \sum_{i=1}^{2} \sum_{x=0}^{1} \frac{(a_i + x)(1 + b_i - x)B(a_i + x, 1 + b_i - x)}{(1 + a_i + b_i)^2 (1 + a_i + b_i + 1)B(a_i, b_i)B(x + 1, 1 - x + 1)} - 2c
\]

\[
= \frac{1}{2} \sum_{i=1}^{2} \left[ \frac{a_i (b_i + 1) B(a_i, b_i + 1)}{(a_i + b_i + 1)^2 (a_i + b_i + 2)B(a_i, b_i)B(1,2)} + \frac{(a_i + 1)b_i B(a_i + 1, b_i)}{(a_i + b_i + 1)^2 (a_i + b_i + 2)B(a_i, b_i)B(2,1)} \right] - 2c
\]

\[
= \sum_{i=1}^{2} \frac{a_i b_i}{(a_i + b_i)(a_i + b_i + 1)} - 2c
\]

\( n \) that maximizes \( U(n) \) would be the optimal sample size.

Figure 1 illustrates the relation of \( U(n) \) and \( n \) for two sets of parameter values.

The solid line corresponds to parameter value \( a_1 = 3, \ b_1 = 3, \ a_2 = 4, \ b_2 = 8 \) and sample cost \( c = .0001 \), the optimal sample size is \( n = 37 \).

The dotted line corresponds to parameter value \( a_1 = 3, \ b_1 = 5, \ a_2 = 6, \ b_2 = 4 \) and sample cost \( c = 0.00005 \), the optimal sample size is \( n = 56 \).

3. Hypothesis Test Power

In this section, we are interested in testing whether \( \theta_1 \) and \( \theta_2 \) are close.

Hypothesis Test. \( H_0 : |\theta_1 - \theta_2| \leq \delta \) vs. \( H_1 : |\theta_1 - \theta_2| > \delta \)

Our criteria is to accept or reject the hypothesis according to the posterior distribution of \( \theta_1 \) and \( \theta_2 \).

if \( P(|\theta_1 - \theta_2| < \delta | x) > 1/2 \) then accept \( H_0 \)
if $P(|\theta_1 - \theta_2| < \delta | x) \leq 1/2$ then reject $H_0$

we assume that

if $P(|\theta_1 - \theta_2| < \delta | x) > 1/2$ then $P(H_0 \text{ is true}) = P(|\theta_1 - \theta_2| < \delta | x)$

if $P(|\theta_1 - \theta_2| < \delta | x) \leq 1/2$ then $P(H_1 \text{ is true}) = 1 - P(|\theta_1 - \theta_2| < \delta | x)$

Let "Test Power" $P(n)$ be the mean probability that we make correct decision when the sample size for each treatment is $n$, that is $P(n) = E_x(P(\text{correct\,decision}|x))$. It is very complicated to calculate $P(n)$ analytically. We fit a logistic model for $P(n)$,

$$P(n) = \frac{1}{2} + \frac{1}{2} \cdot \frac{e^{\alpha + \beta n}}{1 + e^{\alpha + \beta n}}$$

$P(0) = \frac{1}{2} + \frac{1}{2} \cdot \frac{e^\alpha}{1 + e^\alpha}$ is the power without any observation. We can calculate it from the prior distribution, $P(0) = P(|\theta_1 - \theta_2| < \delta)$ Then we can get

$$\alpha = \ln\left(\frac{2P(0) - 1}{2 - 2P(0)}\right)$$

For getting the coefficient $\beta$, We need to generate test powers at some simulated points for some different $n$

1. For each $n$ in 5, 10, 15, ..., $N$, $N$ is a prechosen number.
2. Generate $\hat{\theta}_1$, $\hat{\theta}_2$ from prior distribution $p(\theta_1)$ and $p(\theta_2)$.
3. Generate $x_1$, $x_2$ from distribution $p(x_1|\hat{\theta}_1, n)$ and $p(x_2|\hat{\theta}_2, n)$
4. Calculate test power at $(x_1, x_2, n)$
5. repeat step 2-4 $M$ times, $M$ is a prechosen number
6. Average the $M$ calculated power to get a generated power $\hat{P}(n)$
We can get the coefficient $\beta$ by fitting the logistic model to data \( \{ n, \hat{P}(n), n = 5, 10, 15, \ldots, N \} \).

The utility function is

\[
U(n) = P(n) - 2nc
\]
\[
= \frac{1}{2} + \frac{1}{2} \cdot \frac{e^{\alpha + \beta n}}{1 + e^{\alpha + \beta n}} - 2nc
\]
\[
= 1 - \frac{1}{2} \cdot \frac{1}{1 + e^{\alpha + \beta n}} - 2nc
\]

\[
U'(n) = \frac{1}{2} \frac{\beta e^{\alpha + \beta n}}{(1 + e^{\alpha + \beta n})^2} - 2c
\]

Solve \( U'(n) = 0 \) with respect to \( n \), we get

\[
\hat{n} = \frac{\ln(\frac{1}{8} \sqrt{\frac{4}{2^2} - \frac{16\beta}{c} + \frac{\beta}{8c} - 1}) - \alpha}{\beta}
\]

\[
U''(\hat{n}) < 0
\]

So \( \hat{n} \) maximizes \( U(n) \), it is the optimal sample size.

Figure 2 illustrates the relation of \( n \) and power \( P(n) \) for two sets of parameter values. Note that the points are generated points, and the lines are the logit fit curves.

The •'s and the solid curve correspond to parameter \( a_1 = 3, b_1 = 3, a_2 = 4, b_2 = 8 \), sample cost \( c = .0005, \delta = 0.1 \) and \( M = 50 \), the optimal sample size is \( n = 101 \).

The △'s and the dotted curve correspond to parameter \( a_1 = 3, b_1 = 5, a_2 = 6, b_2 = 4 \) sample cost \( c = 0.0004, \delta = 0.1 \) and \( M = 50 \), the optimal sample size is \( n = 89 \).

4. Lindley's Information

In this section, we consider the measure of the information provided by an experiment proposed
by Lindley (1956). Lindley proposed that the expected information provided by an experiment \( \mathcal{E} \) before performing it is

\[
I(\mathcal{E}) = E \left\{ \log \left( \frac{p(\theta|x)}{p(\theta)} \right) \right\}
\]

where expectations are taken with respect to the joint distribution of \( \theta \) and \( x \).

Parmigiani and Berry (1993) use this measure of information to a single binary data experiment, and get observed information

\[
A_i(x_i, n) = \log \frac{B(a_i, b_i)}{B(a_i + x_i, b_i + n - x_i)} + x_i [\psi(a_i + x_i) - \psi(a_i + b_i + n)]
\]

\[
+ (n_i - x_i) [\psi(b_i + n - x_i) - \psi(a_i + b_i + n)]
\]

\( i = 1, 2 \)

The prior distributions of \( \theta_1 \) and \( \theta_2 \) are independent, so are their posterior distributions. So the observed information \( A(x, n) \) is the summation of the observed information for individual treatment.

\[
A(x, n) = A_1(x_1, n) + A_2(x_2, n)
\]

Note that for integer \( m \), \( \psi(m) = \sum_{k=1}^{m-1} \frac{1}{k} - C \), where \( C \) is Euler’s constant. So, when \( a_1, b_1, a_2 \) and \( b_2 \) are all integers, we have

\[
A_i(x_i, n) = \log \frac{B(a_i, b_i)}{B(a_i + x_i, b_i + n - x_i)} - x_i \sum_{k=a_i+x_i}^{a_i+b_i+n-1} \frac{1}{k}
\]

\[
-(n - x_i) \sum_{k=b_i+n-x_i}^{a_i+b_i+n-1} \frac{1}{k}
\]

Expected information is the expectation of \( A(x, n) \) with respect to the marginal distribution of \( x \).

\[
E_x A(x, n) = E_{x_1} A_1(x_1, n) + E_{x_2} A_2(x_2, n)
\]
\[
= \sum_{i=1}^{2} \sum_{k=0}^{n} A_i(k, n) P(x_i = k)
= \frac{1}{n+1} \sum_{i=1}^{2} \sum_{k=0}^{n} \frac{A_i(k, n) B(a_i + k, n + b_i - k)}{B(a_i, b_i) B(k + 1, n - k + 1)}
\]

The utility function is

\[
U(n) = E(x, n) - 2nc
= \frac{1}{n+1} \sum_{i=1}^{2} \sum_{k=0}^{n} \frac{A_i(k, n) B(a_i + k, n + b_i - k)}{B(a_i, b_i) B(k + 1, n - k + 1)} - 2nc
\]

the \( n \) that maximizes \( U(n) \) is the optimal sample size.

Figure 3 illustrates the relation of \( U(n) \) and \( n \) for two sets of parameter values.

The solid curve corresponds to \( a_1 = 3, b_1 = 3, a_2 = 4, b_2 = 8 \) and sample cost \( c = 0.0001 \), the optimal sample size is \( n = 45 \)

The dotted curve corresponds to \( a_1 = 3, b_1 = 5, a_2 = 6, b_2 = 4 \) and sample cost \( c = 0.00005 \), the optimal sample size is \( n = 66 \).

References


Parmigiani, G. and Berry, D.A. (1993) Applications of Lindley Information Measure to the Design of Clinical Experiments
Figure 1: Utility as a function of sample size $n$
Figure 2: Test power as a function of sample size $n$
Figure 3: Utility as a function of sample size $n$