

Sample Size Determination for dichotomous variables-an example from forensic science

Dimitrios Mavridis
University of Edinburgh

Colin Aitken
University of Edinburgh

June 6, 2008

Outline

- An example from the forensic science where sampling is required
- Binomial Sampling
- Prior elicitation using power priors, informed by historical data
- Sequential Probability Ratio Test
- Use of predictive distributions for determining the appropriate sample size for testing a specific hypothesis.

Forensic problems requiring sampling

- A consignment of m pills suspected to be illicit (drugs)
- Sentence imposed is dependent (in most jurisdictions) on the type and quantity of illicit drugs
- Reasons for resorting to sampling of n units
 1. Save of financial, time and manpower resources
 2. Avoid exposure of forensic scientists to chemical material
 3. Analyzing a unit eventually destroys it. Some evidence may be presented in court or given to the defence to conduct their own analysis

Arbitrary sampling

- Investigate 5%, 10% ... of the total seizure.
- Investigate $n = \sqrt{m}$.
- Investigate $20 + 0.1 \times (m - 20)$ units.
- $n = 1$.

United Nations Drug Control Program

- if $m < 10$ then $n=m$.
- if $10 < m < 100$ then $n=10$.
- if $m > 100$ then $n=\sqrt{m}$.

Bernoulli trials

- A Bernoulli trial is an experiment with two possible outcomes ('failure'=0, 'success'=1)
- Examples of Bernoulli trials : flipping of a coin, gender, defective or non-defective product, illicit or licit substance
- $f(y) = \theta^y(1 - \theta)^{1-y}$, $y = 0, 1$, θ =probability of 'success'.

Binomial sampling

- Repeated independent Bernoulli trials y_1, \dots, y_n (probability of 'success' remains constant).

- Let x be the number of successes in n independent trials, $x = \sum_{i=1}^n y_i$.

- We say that x follows a binomial distribution

$$P(x \text{ 'successes' in } n \text{ Bernoulli trials}) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}.$$

- Mean of the binomial distribution, $E(x) = n\theta$.

- Variance of the binomial distribution, $V(x) = n\theta(1 - \theta)$.

- An unbiased estimate of θ , $\hat{\theta} = \frac{x}{n}$.

Classical criteria for SSD

- The sample proportion $\hat{\theta}$ is assumed to be normally distributed
$$\hat{\theta} \sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$
- The quantity $\frac{\hat{\theta}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \sim N(0, 1)$
- A $(1 - \alpha)\%$ confidence interval for θ is given by $\theta \pm \Phi^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{\theta(1-\theta)}{n}}$
- $\Phi^{-1}(\alpha)$ is the inverse of the standard normal cumulative function at point α , or alternatively the point to the left of which there are $\alpha\%$ of the observations.

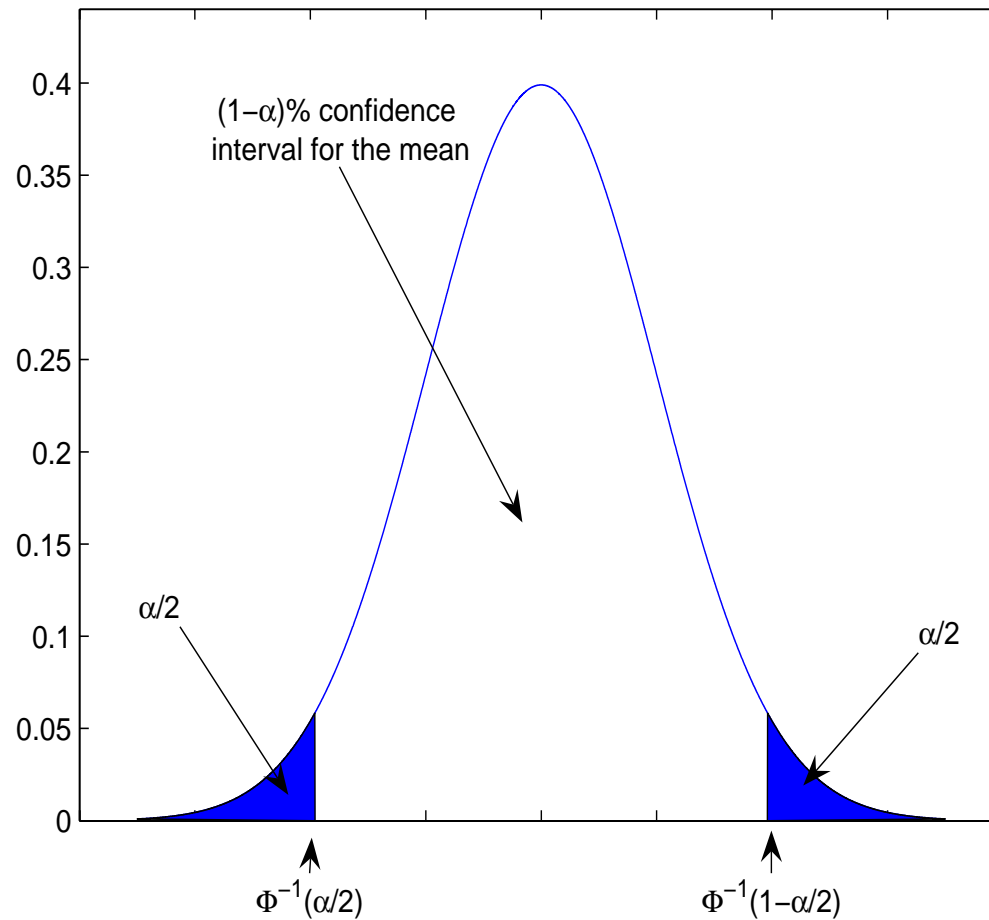


Figure 1: Probability density plot of a standard normal distribution with a $(1 - \alpha)\%$ confidence interval.

- The smallest sample size n is determined so that

$$Pr(|\hat{\theta} - \theta| \leq d) \geq 1 - \alpha$$

- $\Phi^{-1}\left(\frac{\alpha}{2}\right) \sqrt{\frac{\theta(1-\theta)}{n}} \leq d \Leftrightarrow n \geq \left(\Phi^{-1}\left(\frac{\alpha}{2}\right)\right)^2 \frac{\theta(1-\theta)}{d^2}$
- Use of the worst possible scenario $\theta = 0.5$ leads to

$$n \geq \frac{\left(\Phi^{-1}\left(\frac{\alpha}{2}\right)\right)^2}{4d^2}$$

- θ is not known in advance.

Table 1: Sample size required for various probabilities of success, $\alpha=0.05$ and $d=0.05$

θ	n
0.001	2
0.01	16
0.1	139
0.3	323
0.5	385
0.7	323
\vdots	\vdots

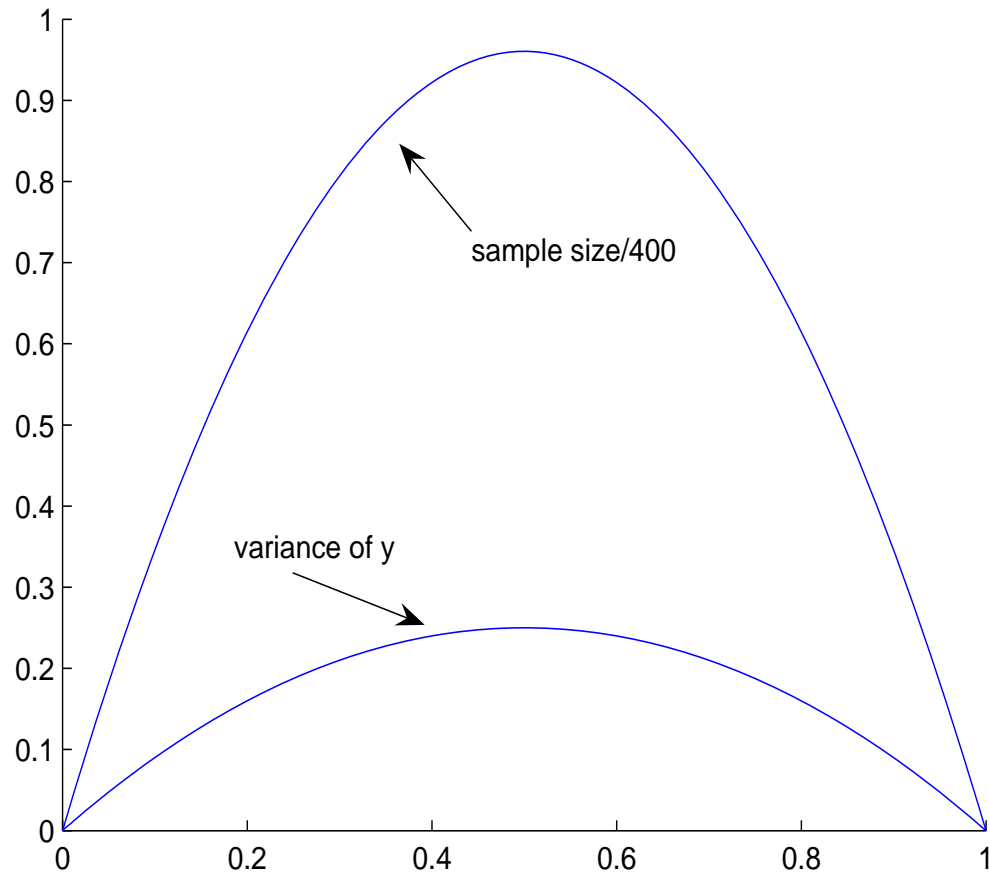


Figure 2: Sample size required for various θ 's and variance of y , $\theta \times (1 - \theta)$

SSD when hypotheses testing is considered

Usually we want to test a hypothesis of the kind $H_0 : \theta = \theta_0$ versus an alternative $H_1 : \theta = \theta_1$.

There are two errors that may occur

- Type I error, $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$
- Type II error, $\beta = P(\text{accept } H_0 \mid H_1 \text{ is true})$

power of a test is denoted by $1 - \beta = P(\text{reject } H_0 \mid H_1 \text{ is true})$

Determine the minimum sample size such that for fixed type I error the probability of correctly rejecting the null hypothesis is sufficiently large

- For $n > 20$, a $(1 - \alpha)\%$ confidence interval for θ is given by
$$(L_b, U_b) = (\theta - \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\theta(1-\theta)}{n}}, \theta + \Phi^{-1}(\frac{\alpha}{2})\sqrt{\frac{\theta(1-\theta)}{n}})$$

- power = $1 - \beta = P(\theta > U_b \cup \theta < L_b \mid \theta = \theta_1)$

- if $\theta_0=0.2$, $\theta_1=0.3$, $\alpha = 0.05$ and $n=20$, then $U_b=0.347$ and $P(\theta > 0.347 \mid \theta = 0.3) = 0.3229$

- Lead to excessive sample sizes, does not hold for $n < 20$.

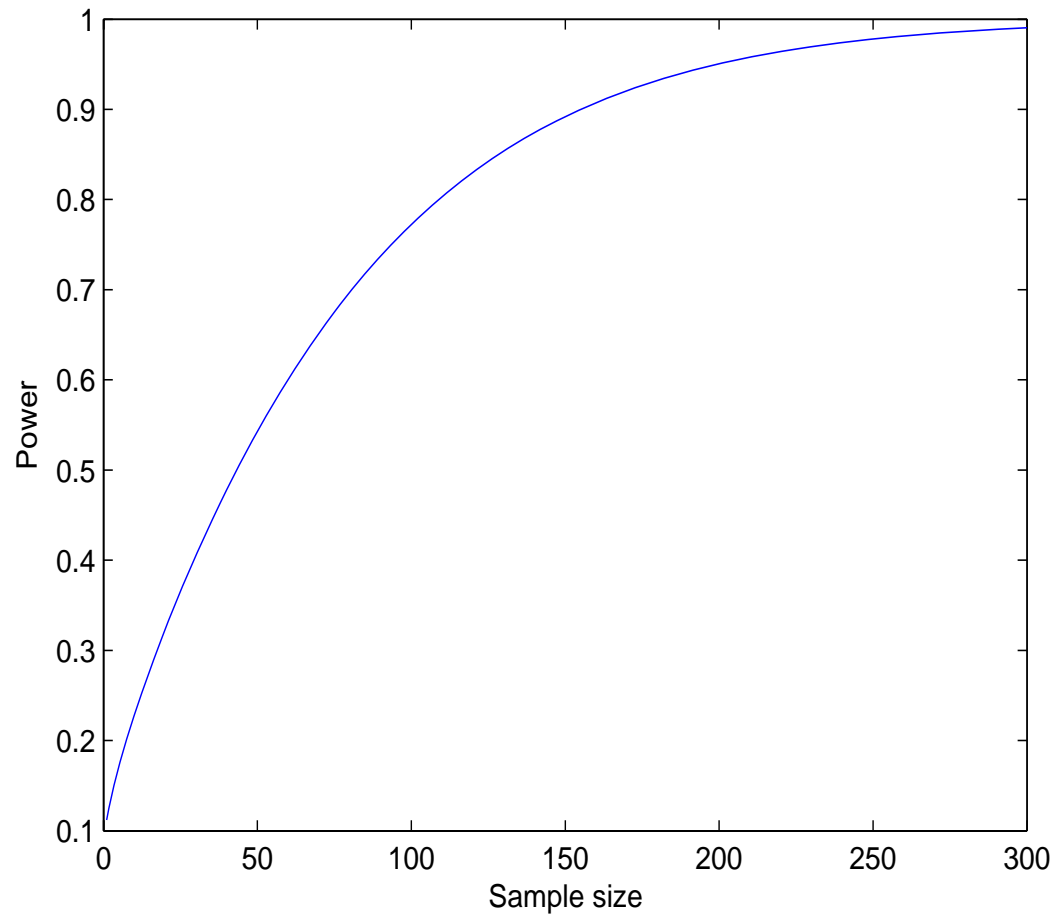


Figure 3: Power of the test for various sample sizes ($\theta_0=0.2$, $\theta_1=0.3$, $\alpha = 0.05$).

Bayes Theorem

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

$$f(\theta | x) = \frac{f(\theta, x)}{f(x)} = \frac{f(x | \theta)f(\theta)}{f(x)} = \frac{f(x | \theta)f(\theta)}{\int f(x | \theta)f(\theta)d\theta} \propto f(x | \theta)f(\theta)$$

posterior \propto *likelihood* \times *prior*

- If the resulting posterior distribution is of the same class with the prior distribution then the prior is said to be a conjugate prior

Beta distribution

- A continuous probability distribution defined on the interval $[0,1]$

$$f(x | \nu_1, \nu_2) = \frac{\Gamma(\nu_1 + \nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} x^{\nu_1-1} (1-x)^{\nu_2-1}$$

where Γ is the gamma function $\Gamma(t) = (t-1)!$ and $\nu_1, \nu_2 > 0$

- $E(x) = \frac{\nu_1}{\nu_1 + \nu_2}$
- $V(x) = \frac{\nu_1 \times \nu_2}{(\nu_1 + \nu_2)^2 \times (\nu_1 + \nu_2 + 1)}$

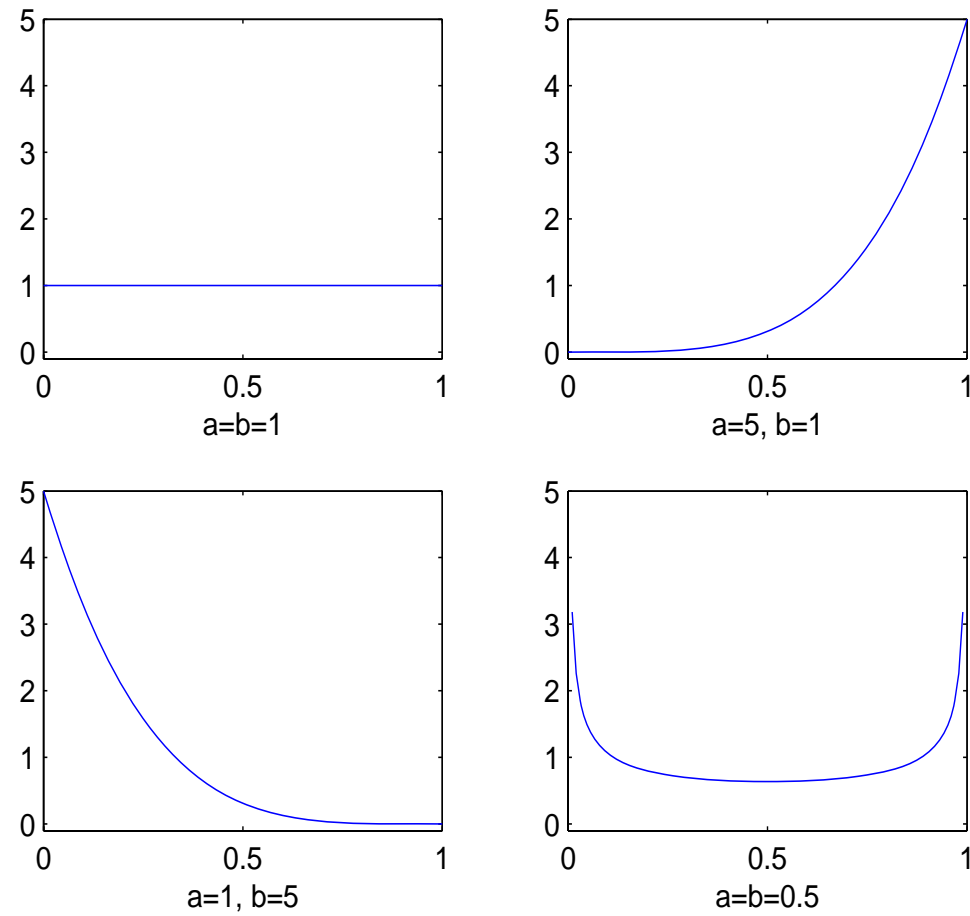


Figure 4: Probability density function of a Beta distribution for various parameter values

Modeling uncertainty using a conjugate prior for the binomial distribution

$$x \mid \theta \sim \text{Bin}(n, \theta)$$

$$\theta \sim B(\nu_1, \nu_2) \Rightarrow f(\theta \mid \nu_1, \nu_2) \propto \theta^{\nu_1-1} (1 - \theta)^{\nu_2-1} \quad 0 < \theta < 1$$

$$f(\theta \mid x, n, \nu_1, \nu_2) \propto \theta^{\nu_1+x-1} (1 - \theta)^{n-x+\nu_2-1}.$$

$$\theta \mid x \sim B(\nu_1 + x, n - x + \nu_2)$$

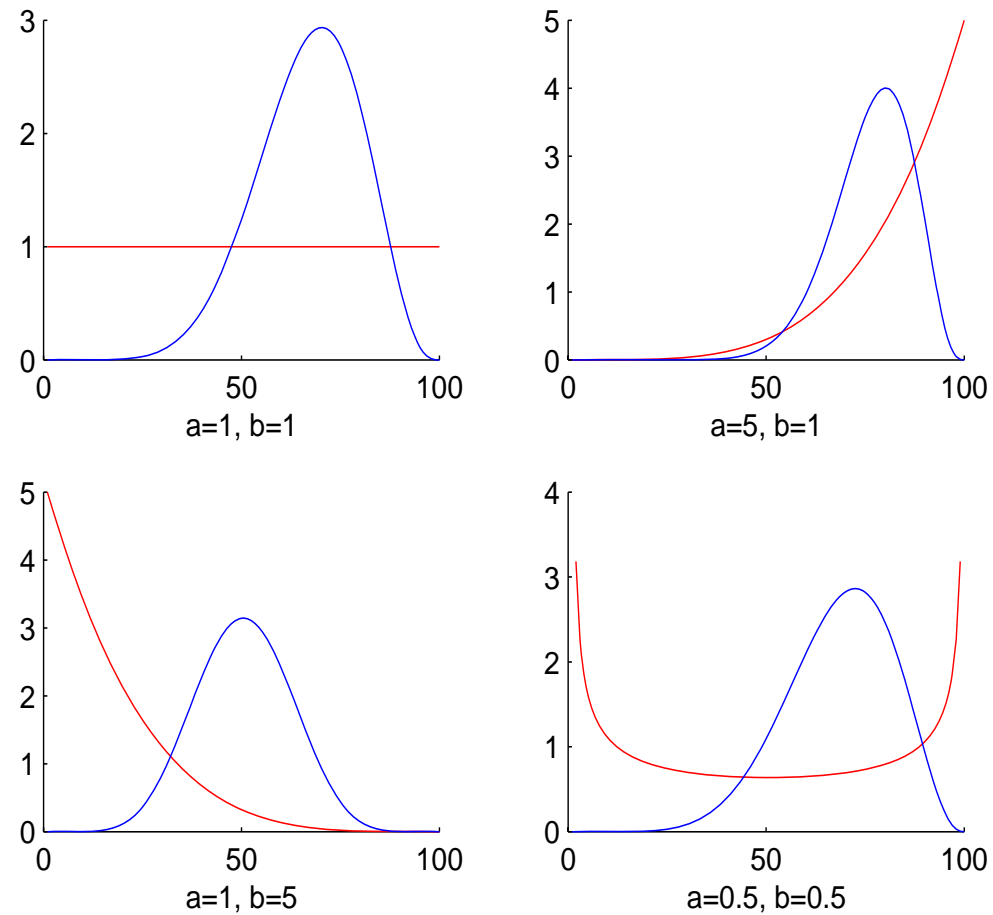


Figure 5: Probability density function of a Beta posterior for $x=7$, $n=10$ and various parameter values, red lines denote the Beta priors that were given in each case.

Use of power priors to extract information from historical data

- Ibrahim and Chen (2000) incorporated information from previous studies to form a suitable prior for a current study.
- The same likelihood should be used among different studies for drawing inference for the parameters of interest

$$f^p(\theta | D_0) \propto L(\theta; D_0)^{\alpha_0} f(\theta)$$

- data from a previous similar study are denoted by D_0
- L denotes the likelihood
- α_0 is a coefficient weighting the effect of historical data on the current study

Suppose we have a seizure that was captured under circumstances similar to a previous seizure D_0 in which out of the n_0 pills, x_0 were found to be illicit.

The likelihood from the historical data is $L(\theta; D_0) = \binom{n_0}{x_0} \theta^{x_0} (1 - \theta)^{n_0 - x_0}$

A Beta prior for θ is $f(\theta) \propto \theta^{\nu_1 - 1} (1 - \theta)^{\nu_2 - 1}$

The design power prior for a binomial parameter turns out to be a beta-binomial with parameters $(a_0 x_0 + \nu_1, a_0(n_0 - x_0) + \nu_2)$.

$$\begin{aligned} f^p(\theta | D_0) \propto L(\theta; D_0)^{a_0} f(\theta) &\propto \left(\binom{n_0}{x_0} \theta^{x_0} (1 - \theta)^{n_0 - x_0} \right)^{a_0} \theta^{\nu_1 - 1} (1 - \theta)^{\nu_2 - 1} \\ &\propto \theta^{a_0 x_0 + \nu_1 - 1} (1 - \theta)^{(n_0 - x_0) a_0 + \nu_2 - 1} \end{aligned}$$

Simulation-based approach for SSD

- Draw θ^* 's from the power prior distribution $f^p(\theta \mid D_0)$, $B(a_0x_0 + \nu_1, a_0(n_0 - x_0) + \nu_2)$.
- Draw, for each θ^* , a sample x^* of size n from the sampling distribution $f(x \mid \theta^*)$, (binomial distribution).
- Compute $T(x_n^*)$ (a function of the posterior distribution) for each of the generated samples We choose to estimate the average posterior variance of θ $E[\text{var}(\theta \mid x_n)] \leq \epsilon$, $\epsilon > 0$

Consider the case where we have a consignment of pills that is seized under similar circumstances to previous seizure of 25 pills in which all pills were found to be illicit ($n_0 = x_0 = 25$). Suppose also that a $B(1, 1)$ is considered.

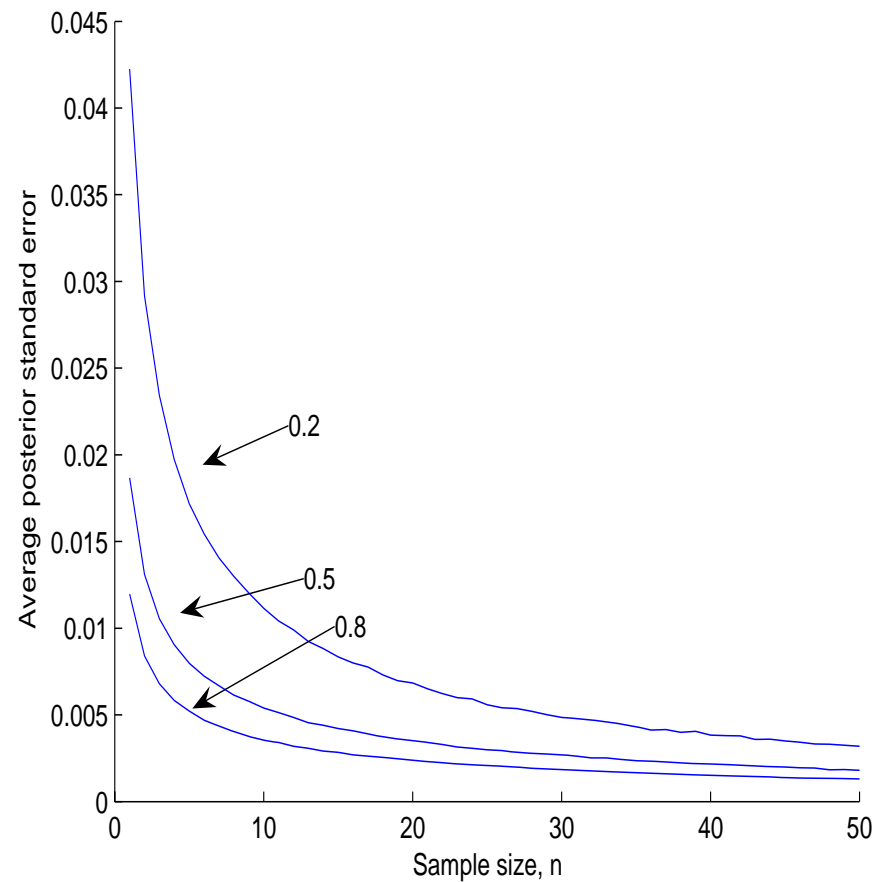


Figure 6: Plot of average posterior standard error versus sample size for different weights attached to historical data

Criterion for Sample Size Calculations for Proportions with Binary Responses

Determine the sample size such that we are $100p\%$ certain that at least $100l\%$ of a consignment contains drugs when all n units in the sample contain illicit drugs

when $p = 0.95$ and $l = 0.5$, the criterion can be written mathematically as

$$Pr(\theta > 0.5 \mid \nu_1 + n, \nu_2) = \frac{\int_{0.5}^1 \theta^{n+\nu_1-1} (1-\theta)^{\nu_2-1} d\theta}{B(n+\nu_1, \nu_2)} \geq 0.95.$$

- Assumes all units are of the same kind

Sequential Analysis

Sequential analysis is a statistical analysis where the sample size is not fixed in advance. Data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed

We might end up with a much smaller sample size saving time and financial resources

Analysis is heavily dependent on collected data, hypothesis being tested and stopping rule used

A two-sided sequential criterion

Suppose that there are two competing propositions $H_0 : \theta \leq \theta_\ell$ and $H_1 : \theta \geq \theta_u$, ($\theta_u > \theta_\ell$). These two propositions are tested sequentially.

Sampling is stopped, either when

$$Pr(\theta < \theta_\ell \mid \nu_1 + x_i, \nu_2 + i - x_i) \Leftrightarrow \frac{\int_0^{\theta_\ell} \theta^{\nu_1 + x_i - 1} (1 - \theta)^{\nu_2 + i - x_i - 1} d\theta}{B(\nu_1 + x_i, \nu_2 + i - x_i)} \geq p_1,$$

or when

$$Pr(\theta > \theta_u \mid \nu_1 + x_i, \nu_2 + i - x_i) \Leftrightarrow \frac{\int_{\theta_u}^1 \theta^{\nu_1 + x_i - 1} (1 - \theta)^{\nu_2 + i - x_i - 1} d\theta}{B(\nu_1 + x_i, \nu_2 + i - x_i)} \geq p_2.$$

Sequential Probability Ratio Test (SPRT)

- Consider a simple hypothesis $H_0 : \theta \leq \theta_0$ against an alternative $H_1 : \theta \geq \theta_1$
- The Likelihood Ratio $LR = \log \frac{L(\theta_1, \mathbf{x})}{L(\theta_0, \mathbf{x})}$ is computed after an observation is collected

Two courses of action

- Stop sampling
 1. Accept hypothesis H_0 if $LR \leq B$
 2. Accept hypothesis H_1 if $LR \geq A$
- Continue sampling if $B \leq LR \leq A$

- We stop sampling when $B \leq \frac{f(\mathbf{x}, \boldsymbol{\theta}_2)}{f(\mathbf{x}, \boldsymbol{\theta}_1)} \leq A$
- The test controls the probabilities of type I and type II errors (α and β)
- The likelihood is conditioned so that $P(H_0 | H_0) \geq 1 - \alpha$ and $P(H_0 | H_1) \leq \beta$

Table 2: Probabilities of accepting a certain hypothesis

	$LR > A$ (accept H_1)	$LR < B$ (accept H_0)
H_1 is correct	α	$1 - \alpha$
H_2 is correct	$1 - \beta$	β

Accept H_1 if

$$f(x | \theta_1) > Af(x | \theta_0) \Leftrightarrow 1 - \beta > A\alpha \text{ when } H_1 \text{ is correct, so } A < \frac{1-\beta}{\alpha}$$

Similarly, $B > \frac{\beta}{1-\alpha}$

- $\frac{1-\beta}{\alpha}$ is a lower limit for A , $\frac{\beta}{1-\alpha}$ is an upper limit for B

$$B \leq \frac{f(x | \theta_2)}{f(x | \theta_1)} \leq A,$$

$$\frac{\beta}{1-\alpha} \leq \frac{\binom{n}{x}(1-\theta_2)^{n-x}\theta_2^x}{\binom{n}{x}(1-\theta_1)^{n-x}\theta_1^x} \leq \frac{1-\beta}{\alpha},$$

$$\log \frac{\beta}{1-\alpha} \leq n \log \frac{1-\theta_2}{1-\theta_1} + x \log \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} \leq \log \frac{1-\beta}{\alpha},$$

For a binomial population

1. accept H_0 , if $x \leq k_1 + \lambda n$;
2. accept H_1 , if $x \geq k_2 + \lambda n$;
3. continue sampling if $k_1 + \lambda n \leq x \leq k_2 + \lambda n$

$$\text{where } k_1 = \frac{\ln \frac{\beta}{1-\alpha}}{\ln \frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}}, \quad k_2 = \frac{\ln \frac{1-\beta}{\alpha}}{\ln \frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}} \quad \text{and} \quad \lambda = \frac{\ln \frac{1-\theta_1}{1-\theta_0}}{\ln \frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}}$$

Example

- Suppose that we want to test the hypothesis $H_0 : \theta \leq 0.2$ and $H_1 : \theta \geq 0.6$;
- $\alpha = 0.01, \beta = 0.1$.
- $k_1 = -1.3, k_2 = 2.8$ and $\lambda = 0.4$.

The two parallel lines represent the lower and upper thresholds ($k_1 + \lambda n, k_2 + \lambda n$) are

$$k_1 + \lambda n = -1.3 + 0.4n,$$

$$k_2 + \lambda n = 2.8 + 0.4n.$$

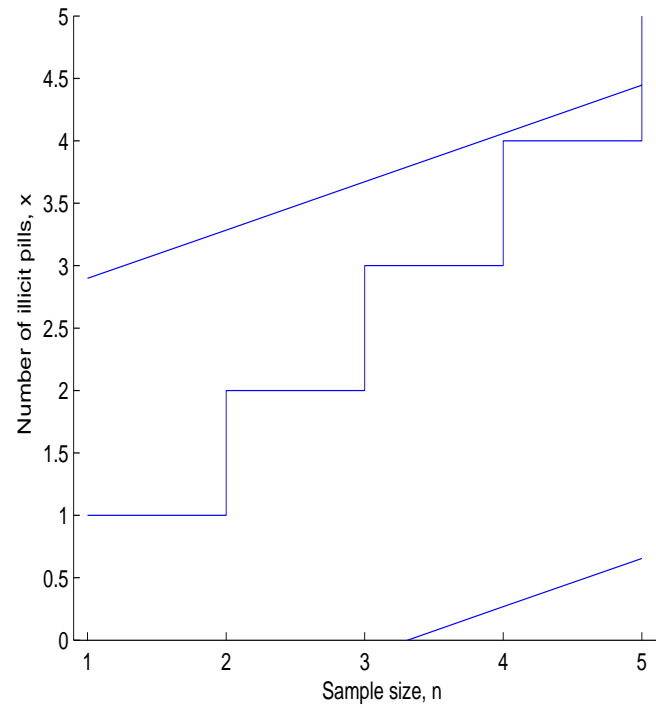


Figure 7: Monitoring the SPRT. Solid lines represent the lower and upper thresholds of the procedure

Use of posterior predictive distribution to determine the sample size for testing a specific hypothesis

• x are the 'successes' the units inspected at a given time, y are the 'successes' in the units not inspected at a given time, θ is the proportion of 'successful' units

$$\bullet \theta \mid x \sim B(\nu_1 + x, n - x + \nu_2)$$

$$\bullet p(y \mid x) = \int p(y \mid \theta)p(\theta \mid x)d\theta$$

$$\bullet p(y \mid \theta) = \binom{m-n}{y} \theta^y (1 - \theta)^{m-n-y}$$

$$\bullet p(y \mid x) = \binom{m-n}{y} \frac{B(\nu_1+x+y, m-x-y+\nu_2)}{B(\nu_1+x, n-x+\nu_2)} \propto B(a + x + y, m - x - y + b)$$

The normal approximation to the beta-binomial distribution is used

$$y | x \sim N \left(\frac{(m - n)(x + \nu_1)}{n + 2}, \frac{(m - n)(x + \nu_1)(n - x + \nu_2)(m + a + \nu_2)}{(n + \nu_1 + \nu_2)^2(n + \nu_1 + \nu_2 + 1)} \right)$$

Sample size is determined as the minimum sample size for which either the length of the confidence interval is below a specified threshold

$$\frac{x + \mu - \Phi^{-1}(1 - \frac{\alpha}{2})\sigma}{m} \quad \text{---} \quad \frac{x + \mu + \Phi^{-1}(1 - \frac{\alpha}{2})\sigma}{m}$$

or the probability that the proportion lies within a certain interval $((-\infty, c]$ or $[c, +\infty))$ is less than a specified threshold.

$$Pr(y \leq cm - x_n) \geq 1 - \alpha \text{ or } Pr(y \geq cm - x_n) \geq 1 - \alpha, \quad 0 \leq c \leq 1.$$

Suppose a sample of 6 units ($n=6$) from a seizure of $m = 5000$ pills is taken and there are six successes (*i.e.*, the number x of illicit pills equals the sample size 6).

- choose $c=0.6$.
- Stop sampling if $Pr(y \geq 0.6m - x_n) \geq 1 - \alpha$.
- Under the most conservative scenario $y = y_\alpha = \mu + \Phi^{-1}(\alpha)\sigma$
- Therefore, stop sampling if $\mu + \Phi^{-1}(\alpha)\sigma \geq cm - x_n$.
- Or when $\frac{\mu + \Phi^{-1}(\alpha)\sigma + x_n}{m} \geq c$.

- The beta prior is taken to be $B(1, 1)$.
- The mean μ of the posterior beta-binomial distribution is 4369.75
- The variance σ^2 is 550.978².
- The significance level α is taken to be 0.01 so that $\Phi^{-1}(\alpha) = -2.3263$ and $\Phi^{-1}(1 - \alpha) = 2.3263$.
- Then $y_\alpha = \mu + \Phi^{-1}(\alpha)\sigma = 3088.01$ and $\frac{x+y_\alpha}{m} = 0.62 > 0.6$, *i.e.*, the probability that the true proportion of illicit pills is greater than 0.62 is 0.99

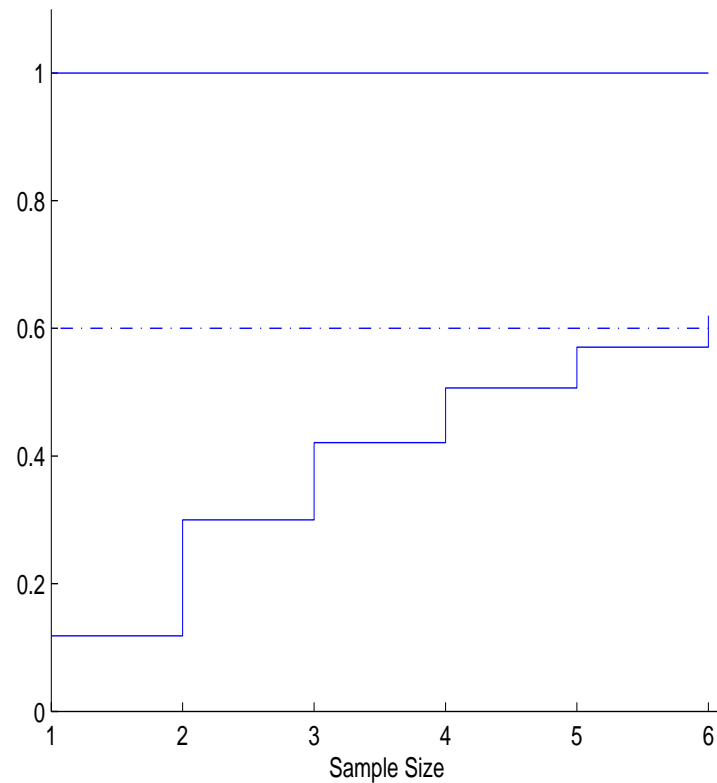


Figure 8: Monitoring the lower and upper bounds of the total seizure based on the predictive distribution of units not inspected (the dotted line represents the threshold considered).

Conclusions

- Sample size required is significantly reduced if prior knowledge is taken into account.

- Methods to acquire prior knowledge

1. Prior beliefs

2. Historical data, obtained under similar circumstances with that at hand.

3. Sequential methods