



# Sample Size Calculations for Biometric Identification Devices

Michael E. Schuckers  
schuckers@stlawu.edu



# Thank You

J. Phillips - We need to have the correct statistical intervals

V. Valencia – Need to develop sample size calculations

# Goals

1. Create appropriate methodology for confidence intervals

1a. Use that methodology to make sample size calculations

# Example

Binomial

CI:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample Size:

$$n = \left\lceil \left( \frac{z^*}{\varepsilon} \right)^2 p(1-p) \right\rceil$$

for fixed width  $\varepsilon$  of CI

# Notation

Testing  $n$  individuals  $m_i$  times.

$X_i$  = # of errors for  $i^{\text{th}}$  indiv.

$p_i = X_i/m_i$  = observed error rate for  $i^{\text{th}}$  indiv.

Assume  $m_i = m$  for all  $i$

Call error rate,  $\pi$  (not 3.14159)

Estimated error rate is

$$\hat{\pi} = \frac{\sum_{i=1}^m X_i}{nm}$$

# Model

Assume 1<sup>st</sup> two moments of  $X_i$  are known

$$E[X_i|m,\pi,\rho]=m\pi$$

$$\text{Var}[X_i|m,\pi,\rho]=m\pi(1-\pi)(1+(m-1)\rho)$$

Not assuming iid of attempts, assuming  $X_i$ 's are conditionally indep allowing for intra-individual correlation.

Not assuming a form for the distribution of  $p_i$ 's. Moore (1987) says first two moments sufficient.

# Beta-binomial (Schuckers, 2003)

Binomial

$$\hat{\pi} \pm z^* \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{mn}}$$

Beta-binomial

$$\begin{aligned} & \hat{\pi} \pm z^* \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{mn}} C \\ &= \hat{\pi} \pm z^* \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{mn} (1 + (m-1)\rho)} \end{aligned}$$

# Correlation with $\rho=0.01, \pi=0.1$

1=error

0=no error

i	$p_i$	1	2	3	4	5	6	7	8	9	10
1	0.3	0	0	0	0	0	0	1	1	1	0
2	0.0	0	0	0	0	0	0	0	0	0	0
3	0.3	1	1	0	0	0	0	1	0	0	0
4	0.0	0	0	0	0	0	0	0	0	0	0
5	0.1	0	0	0	0	0	0	0	1	0	0
6	0.1	0	1	0	0	0	0	0	0	0	0
7	0.0	0	0	0	0	0	0	0	0	0	0
8	0.1	0	0	0	1	0	0	0	0	0	0
9	0.1	0	0	0	0	0	1	0	0	0	0
10	0.0	0	0	0	0	0	0	0	0	0	0

# Correlation with $\rho=0.4$ , $\pi=0.1$

1=error

0=no error

i	$p_i$	1	2	3	4	5	6	7	8	9	10
1	0.1	0	0	0	0	0	0	1	0	0	0
2	0.0	0	0	0	0	0	0	0	0	0	0
3	0.0	0	0	0	0	0	0	0	0	0	0
4	0.1	0	0	0	0	1	0	0	0	0	0
5	0.8	1	1	0	1	1	0	1	1	1	1
6	0.0	0	0	0	0	0	0	0	0	0	0
7	0.0	0	0	0	0	0	0	0	0	0	0
8	0.0	0	0	0	0	0	0	0	0	0	0
9	0.0	0	0	0	0	0	0	0	0	0	0
10	0.0	0	0	0	0	0	0	0	0	0	0

# Suggested Values for $\rho$

Daugman (2003) paper in Pattern Recognition suggests values of  $\rho \approx 0.08, 0.001$

Recent work by S. Schuckers suggests  $\rho \approx 0.01$ .

# Notes on Beta-binomial

1. Parametric Approach
2. Reparametrizations of Best Practices Approach of Wayman & Mansfield (We use Lui, Cumberland and Kuo (1996))
3. Possible to get sample size calculations (Lui, 1991)
4. Doesn't perform well in practice

# Statistical Evaluation of Confidence Intervals

1. Start with distribution with known parameters
2. Generate data from that distribution
3. Use that data to create a  $100(1-\alpha)\%$  CI
4. Determine if the estimand (parameter) is inside the created CI
5. Repeat Steps 2 –4 many times
6. Determine % of times that CI ‘captures’ the parameter and call that coverage
7. Does coverage =  $100(1-\alpha)\%$

# Beta-binomial Coverage

Example: Beta-binomial 95% CI

Correlated binary (Oman & Zucker)

$n$	$m$	$\pi$	$\rho$	Coverage
2000	10	0.004	0.001	0.946
2000	10	0.004	0.01	0.950
2000	10	0.004	0.1	0.927
2000	10	0.004	0.4	0.914

# Logit Beta-binomial

General form of model assuming only 1<sup>st</sup> two moments  
(Moore 1987)

Transform data to another scale

$\text{Logit}(\pi) = \log(\pi / (1-\pi)) = \text{log-odds ratio}$

Approximate variance by Delta Method (Taylor Series)

Make CI, then transform back.

# Logit Transform

1. Create CI on logit scale

$$\text{logit}(\hat{\pi}) \pm z^* \sqrt{\frac{1 + (m-1)\hat{\Delta}}{\hat{\pi}(1-\hat{\pi})mn}}$$

2. Get bounds (L,U)

3. Transform back to original scale  
( $\text{logit}^{-1}(L)$ ,  $\text{logit}^{-1}(U)$ )

# Logit Beta-binomial Coverage

Example: Logit 95% CI

n	m	$\pi$	$\rho$	Coverage
2000	10	0.004	0.001	0.955
2000	10	0.004	0.01	0.957
2000	10	0.004	0.1	0.953
2000	10	0.004	0.4	0.950

# Sample Size

**Problem of solving for  $n$  and  $m$  simultaneously**

**Proposed method based on Logit CI (Schuckers, submitted):**

- 1. Fix  $m$  (the number of attempts/indiv.)**
- 2. Determine upper bound of  $\pi_{\max}$  (equate to  $\text{logit}^{-1}(U)$ )**
- 3. Estimate other quantities (significance level,  $\pi$ ,  $\rho$ )**
- 4. Solve for  $n$**
- 5. Repeat for other values of  $m$**

***Developed independently of Lui (1991) but same basic algorithm***

# Sample Size Calculation for 100(1- $\alpha$ )% CI

Necessary sample size given  $m$ ,  $\rho$ ,  $\pi$ ,  $\pi_{\max}$ ,  
and  $\alpha$

$$n = \left\lceil \frac{1 + (m - 1)\rho}{m(\pi(1 - \pi))} \left( \frac{z_{1-\alpha/2}}{\text{logit}(\pi_{\max}) - \text{logit}(\pi)} \right)^2 \right\rceil$$

# Implementation

We need to specify  $m$ ,  $\rho$ ,  $\pi$ ,  $\pi_{\max}$ , and  $\alpha$

$\pi$  is what we think the error rate is

$\pi_{\max}$  is what is the maximum allowable

$\rho$  is the intra-individual correlation

$m$  is the number of attempts per person

$(1-\alpha)*100\% = \text{confidence level}$

# Implementation

Suppose we want to make a 95% confidence interval for the error rate. We believe that the error rate is 0.001, but we want an upper bound on that interval of 0.002. We also believe that the intra-individual correlation rate is 0.1.

# Sample Size

Let

$$\rho=0.1$$

$$\pi=0.001$$

$$\pi_{\max}=0.002$$

$$100(1-\alpha)\% = 95\%$$

$m$	$n$
2	4390
5	2235
8	1696
10	1517
15	1277
20	1158

# Sample Size

Let

$$\rho=0.01$$

$$\pi=0.01$$

$$\pi_{\max}=0.02$$

$$100(1-\alpha)\% = 99\%$$

$m$	$n$
2	746
5	380
8	288
10	258
15	238
20	217

# Sample Size

Let

$$\rho=0.1$$

$$\pi=0.0005$$

$$\pi_{\max}=0.001$$

$$100(1-\alpha)\% = 95\%$$

$m$	$n$
2	8787
5	4474
8	3395
10	3036
15	2557
20	2317

# Final Comments

Could replace  $n$  by  $n^*(n^*-1)$  for FAR and  $m$  is the number of comparisons between each pair of indiv.

Further testing to compare methods seems to indicate that Logit BB is best methodology for this type of data.

Sample size calculation asymptotes (Anil Jain)

# Final Comments

Sample size Calculation implies

$$P(U > \pi_{\max}) = 0.50$$

Power-type calculation can improve this

Data available for better estimation of  $\rho$ .

Need to get handle on possible values.

# Sample Size Calculation for 100(1- $\alpha$ )% CI

Necessary sample size given  $m$ ,  $\rho$ ,  $\pi$ ,  $\pi_{\max}$ ,  
and  $\alpha$

$$n = \left\lceil \frac{1 + (m - 1)\rho}{m(\pi(1 - \pi))} \left( \frac{z_{1-\alpha/2}}{\text{logit}(\pi_{\max}) - \text{logit}(\pi)} \right)^2 \right\rceil$$