STA 304H1F-1003HF Fall 2019
Surveys, Sampling and Observational Data
Section: L0101

Professor Tounkara

Week 5 - Simple Random Sample, Inference

Simple Random Sampling: Inference

Estimation of parameters and estimation of accuracy

Common parameters of interest

Population Total	Population Mean	Population Variance
$\tau_y = \sum_1^N y_i$	$\mu_y = \frac{1}{N}\sum_1^N y_i$	$\sigma_y = \frac{1}{N}\sum_1^N (y_i - \mu_y)^2$

Population size: N known (not always)

Consider a SRS of size n: $y_1, y_2, \ldots, y_n$

Calculation from the sample

Sample Mean	Sample Variance
$\bar{y} = \frac{1}{n}\sum_1^n y_i$	$s^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y})^2$

Estimation of the population mean and total

We will adopt the estimators

Estimator of μ_y | **Estimator of τ_y**

$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_1^n y_i \quad \rightarrow \quad \hat{\tau} = N \times \hat{\mu} = N \times \bar{y} = \frac{N}{n} \times \sum_1^n y_i$ **(*)**

(or $\hat{\mu} = \frac{\hat{\tau}}{N}$)

(*): Because μ_y is related to τ_y by the equation $\tau_y/N = \mu_y$

Properties of estimators $\hat{\mu}$

Case I – simpler case: In SRS with replacement (repitition)

1. $\hat{\mu}$ is Unbiased Estimator of μ_y: $\mathbf{E}(\hat{\mu}) = \mu_y$

2. Variance of $\hat{\mu}$: $\mathbf{Var}(\hat{\mu}) = \frac{\sigma_y^2}{n}$

Case II – more complicated: In SRS without replacement (SRSWR),

1. $\hat{\mu}$ is Unbiased Estimator of μ_y: $\mathbf{E}(\hat{\mu}) = \mu_y$

2. Variance of $\hat{\mu}$: $\mathbf{Var}(\hat{\mu}) = \frac{N-n}{N-1} \times \frac{\sigma_y^2}{n} < \frac{\sigma_y^2}{n}$

Case I, proof

In sampling with replacement, $y_1, y_2, \ldots, y_n$ are independent and identically distributed, e.g. $\mathbf{E}(y_i) = \mu_y$ **and** $\mathbf{Var}(y_i) = \sigma_y^2$

Iidentically distributed:

$$\Longrightarrow \quad \boxed{P(y_i = y \mid y_{i-1}, y_{i-2}, \ldots, y_1) = P(y_i = y) = \frac{1}{N},\ i = 1, 2, \ldots, n}$$

Now: $\mathbf{E}(\bar{y})$? $\mathbf{Var}(\bar{y})$?

$$\begin{aligned}
\mathbf{E}(\bar{y}) &= \mathbf{E}\big[\frac{1}{n}(y_1 + y_2 + \cdots + y_n)\big] = \frac{1}{n}\mathbf{E}\big[(y_1 + y_2 + \cdots + y_n)\big] \\
&= \frac{1}{n}\big[\mathbf{E}(y_1) + \mathbf{E}(y_2) + \cdots + \mathbf{E}(y_n)\big] = \frac{1}{n} n\mu_y = \boxed{\mu_y} \\
\mathbf{Var}(\bar{y}) &= \mathbf{Var}\big[\frac{1}{n}(y_1 + y_2 + \cdots + y_n)\big] = \frac{1}{n^2}\mathbf{Var}\big[(y_1 + y_2 + \cdots + y_n)\big] \\
&= \frac{1}{n^2}\big[\mathbf{Var}(y_1) + \mathbf{Var}(y_2) + \cdots + \mathbf{Var}(y_n)\big] = \frac{1}{n^2} n\sigma_y^2 = \boxed{\frac{\sigma_y^2}{n}}
\end{aligned}$$

Case II, proof

Case II – more complicated case: SRSWR, $y_1, y_2, \ldots, y_n$ are dependent random variables.

e.g.: $\Longrightarrow$

$$\boxed{P(y_i = y) = \frac{1}{N},\ P(y_2 \mid y_1) = \frac{1}{N-1}, y \neq y_1, P(y_3 = y) = \frac{1}{N-3}, y \neq y_1, y_2}$$

. . . they still have the same marginal distribution

$$\boxed{P(y_i = y) = \frac{1}{N},\ y_1, y_2, \ldots, y_n}$$

Formal Proof:

$$P(y_2 = y) = P(y_1 \neq y) \times P(y_2 = y \mid y_1 \neq y) = \tfrac{N-1}{N} \times \tfrac{1}{N-1} = \boxed{\frac{1}{N}}$$

$$P(y_3 = y) = P(y_1 \neq y) \times P(y_2 \neq y \mid y_1 \neq y) \times P(y_3 = y \mid y_2 \neq y, y_1 \neq y)$$

$$= \tfrac{N-1}{N} \times \tfrac{N-2}{N-1} \times \tfrac{1}{N-2} = \boxed{\frac{1}{N}}, \ldots P(y_i = y) = \cdots = \boxed{\frac{1}{N}}$$

Case II, proof-continued

Theorem: In sampling without replacement (SRSWR),

$$\mathbf{E}(\hat{\mu}) = \mathbf{E}(\bar{y}) = \mu_y, \quad \mathbf{Var}(\hat{\mu}) = \mathbf{Var}(\bar{y}) = \frac{N-n}{N-1} \times \frac{\sigma_y^2}{n} < \frac{\sigma_y^2}{n}$$

Proof: $y_1, y_2, \ldots, y_n$ are dependent, but identically distributed and again $\mathbf{E}(y_i) = \mu_y$, and agin, as in case I, $\mathbf{E}(\bar{y}) = \mu_y$

remember, to find expectation of a sum, we don't need independence assumption. Variance is the main problem:

$\mathbf{Var}(\bar{y}) = \mathbf{Var}(\frac{1}{n}\sum_1^n y_i) = \frac{1}{n^2}[\sum_1^n \mathbf{Var}(y_i) + \sum_{i \neq j} \mathbf{Cov}(y_i, y_j)]$

$\mathbf{Var}(\bar{y}) = \frac{1}{n^2}[n\sigma_y^2 + \sum_{i \neq j} \mathbf{Cov}(y_i, y_j)] = \frac{\sigma_y^2}{n^2} + \frac{1}{n^2}\sum_{i \neq j} \mathbf{Cov}(y_i, y_j)$

To find $\mathbf{Cov}(y_i, y_j)$, we need joint distribution of (y_i, y_j).

Case II, proof-continued

They all have the same marginal joint distribution

$$\boxed{P(y_i = y^{'}, y_j = y^{''}) = \frac{1}{N(N-1)},\ y_i \neq y_j,\ i \neq j}$$

same argument as for one-dimensional case; there are $N(N-1)$ different pairs $(y^{'}, y^{''})$ that can be selected at any two given selections (y_i, y_j)

It implies $\mathbf{Cov}(y_i, y_j) = \mathbf{Cov}(y_1, y_2) = -\frac{\sigma_y^2}{N-1}$

$\mathbf{Var}(\bar{y}) = \frac{\sigma_y^2}{n^2} + \frac{1}{n^2}\sum_{i \neq j} \mathbf{Cov}(y_i, y_j)$

Since there are $(n(n-1)$ prairs of $(i,j), i \neq j$, we have that

$\mathbf{Var}(\bar{y}) = \frac{\sigma_y^2}{n^2} - \frac{1}{n^2} n(n-1)\frac{\sigma_y^2}{N-1} = \frac{\sigma_y^2}{n^2}\left(1 - \frac{n-1}{N-1}\right)\frac{N-n}{N-1} \times \frac{\sigma_y^2}{n}$

Estimation of the population Variance

We want to esitmate $\mathbf{Var}(\bar{y}) = \frac{\sigma_y^2}{n}\left(\frac{N-n}{N-1}\right)$ (**SRSWR**), and

$\mathbf{Var}(\bar{y}) = \frac{\sigma_y^2}{n}$ (**SRS**), which are linear functions of σ_y^2,

We consider the sample variance $S^2 = \frac{1}{n-1}\sum_1^n (y_i - \bar{y})^2$.

It can be shown that:

$$\mathbf{E}(S^2) = \begin{cases} \sigma_y^2, & \text{in SRS with replacment} \\ \frac{N}{N-1}\sigma_y^2, & \text{in SRS without replacement} \end{cases}$$

and then

$$\widehat{\sigma}_y^2 = \begin{cases} S^2, & \text{Unbiased in SRS with replacment} \\ \frac{N-1}{N}S^2, & \text{Unbiased in SRS without replacement} \end{cases}$$

So that $\mathbf{Var}(\bar{y})$ can be unbiasedly estimated from the sample by

$$\widehat{\mathbf{Var}}(\bar{y}) = \begin{cases} \frac{S^2}{n}, & \textbf{Unbiased in SRS with replacment} \\ \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, & \textbf{Unbiased in SRS without replacement} \end{cases}$$

Estimation of the mean and variance of the estimator

Conclusion: In SRS without replacement,

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{1}^{n} y_i \qquad \widehat{\mathbf{Var}}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

Note that both are unbiased esimators

A general note on estimation:

If $\eta = g(\theta)$ and $\hat{\theta}$ is an unbiased estimator of θ ($\mathbf{E}(\hat{\theta}) = \theta$),

then $\hat{\eta} = g(\hat{\theta})$ is not , in general an unbiased estimator of η, except when $g(\theta) = a\theta + b$- is a linear function of θ

Then $\mathbf{E}(\hat{\eta}) = \mathbf{E}(g(\hat{\theta})) = \mathbf{E}(a\hat{\theta} + b) = a\mathbf{E}(\hat{\theta}) + b$

In our case, Var$(\bar{y}) = g(\sigma_y^2) = a\sigma_y^2$, where $a = \dfrac{N-n}{N-1}\dfrac{1}{n}$, $b = 0$

Estimation of the population standard deviation

Using our methodology for $\sigma = \sqrt{\sigma^2} = g(\sigma)$, we have that

$$\boxed{\hat{\sigma} = g(\hat{\sigma}^2) = \sqrt{\tilde{S}^2}} = \tilde{S}$$

$\hat{\sigma} = \tilde{S}$ is a biased estimator of σ, but the bias is usually small.

Estimation of the standard error:

$$\boxed{\hat{\sigma}_{\bar{y}} = \widehat{\mathrm{SD}(\bar{y})} = \sqrt{\widehat{\mathbf{Var}}(\bar{y})} = S\sqrt{\frac{N-n}{nN}} \sim \frac{S}{\sqrt{n}}} \text{ for large N}$$

Note that $\hat{\sigma}_{\bar{y}}$ is biased too.

Estimation of the population total

From $\tau = N\mu$, we have that

Estimator of the population total τ:

$$\hat{\tau} = N\hat{\mu} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$$

Variance of $\hat{\tau}$ (theoritical formula for Var$(\hat{\tau})$):

$$\mathbf{Var}(\hat{\tau}) = \mathbf{Var}(N\bar{y}) = N^2\mathbf{Var}(\bar{y}) = N^2\frac{N-n}{N-1} \times \frac{\sigma^2}{n}$$

Estimated Variance for τ(unbiased estimator for Var$(\hat{\tau})$):

$$\widehat{\mathbf{Var}}(\hat{\tau}) = \widehat{\mathbf{Var}}(N\bar{y}) = N^2\widehat{\mathbf{Var}}(\bar{y}) = N^2\frac{N-n}{N} \times \frac{S^2}{n} = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n}$$

Estimated SD (also biased estimator):

$$\widehat{\sigma}_{\hat{\tau}} = \widehat{\mathrm{SD}}(\hat{\tau}) = \sqrt{\widehat{\mathbf{Var}}(\hat{\tau})} = N\sqrt{\widehat{\mathbf{Var}}(\bar{y})} = N\sqrt{\hat{\sigma}_{\bar{y}}} = N \times \widehat{\mathrm{SD}}(\hat{\mu}) = N \times \widehat{\mathrm{SD}}(\hat{\bar{y}})$$

Error bounds and confidence intervals

Error bound:

$$\mathbf{B}_\mu = 2\hat{\sigma}_{\hat{\mu}} = 2 \times \widehat{\mathrm{SD}}(\hat{\mu}) = 2 \times \widehat{\mathrm{SD}}(\bar{y})$$

$$\mathbf{B}_\tau = 2\hat{\sigma}_{\hat{\tau}} = 2 \times \widehat{\mathrm{SD}}(\hat{\tau}) = 2 \times \widehat{\mathrm{SD}}(N \times \bar{y}) = 2 \times N \times \widehat{\mathrm{SD}}(N\bar{y}) = N \times \mathbf{B}_\mu$$

Note: We do not distinguiss in notation between theoritical error bound $\mathbf{B}_\mu = 2 \times \mathbf{SD}(\hat{\mu})$ and estimated error bound $\mathbf{B}_\mu = 2 \times \widehat{\mathbf{SD}}(\hat{\mu})$

Confidence interval:

For μ: $\hat{\mu} \pm \mathbf{B}_\mu = \bar{y} \pm \mathbf{B}_\mu$ $\Longleftarrow$ **short notation** $\Downarrow$

For τ: $\hat{\tau} \pm \mathbf{B}_\tau = N \times \bar{y} \pm \mathbf{B}_\mu = N \times \bar{y} \pm N \times \mathbf{B}_\mu = N(\bar{y} \pm \mathbf{B}_\mu)$

For μ: $\left[\hat{\mu} - \mathbf{B}_\mu, \hat{\mu} + \mathbf{B}_\mu\right]$ $\Longleftarrow$ **Expanded notation**

$\Downarrow$

For τ: $\left[\hat{\tau} - \mathbf{B}_\tau, \hat{\tau} + \mathbf{B}_\tau\right] = \left[N(\hat{\mu} - \mathbf{B}_\mu), N(\hat{\mu} + \mathbf{B}_\mu)\right]$

example

Simple Random Sampling: Examples (I)

Population of Toronto neighbourhoods (I)

Example 1: Population consists of 140 Toronto neighbourhoods

- Variables of interest – population size for 2001 (x) and 2006 (y)
- Sampling frame – List of neighbourhoods; sampling units - neighbourhoods
- An SRS of 10 neighbourhoods selected and x and y observed

Sample obtained (observations)										
i	1	2	3	4	5	6	7	8	9	10
x	10.18	12.51	9.97	10.15	35.77	19.98	10.09	15.78	22.80	12.86
y	10.25	12.74	10.03	10.27	33.83	20.44	9.61	14.41	23.73	12.86

Calculation for y:

$$\bar{y} = \frac{1}{10}(10.25 + 12.74 + \cdots + 12.86) = 15.817$$

$$S^2 = \frac{1}{10-1}\sum_{1}^{10}(y_i - 15.817)^2 = 62.49158 = 62.49$$

$$S = \sqrt{S^2} = \sqrt{62.49158} = 7.905162 = 7.91$$

example

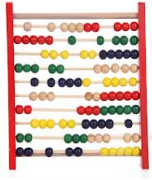
Simple Random Sampling: Examples (I)

Population of Toronto neighbourhoods (II)

Estimation of the average per neighbourhood:

$$\hat{\mu} = \bar{y} = 15.817 = 15.82 \; (= 15.817 \times 1{,}000 = 15{,}817 \text{ residents})$$

$$\hat{\sigma}_{\bar{y}} = \hat{SD}(\bar{y}) = \sqrt{\frac{140-10}{140}\frac{62.49158}{10}} = \sqrt{5.802789} = 2.408898$$



$$B_{\mu} = 2 \times 2.408898 = 4.817796 = 4.82$$

CI for μ: $\bar{y} \pm B_{\mu} = 15.82 \pm 4.82 = [11.00, 20.64]$

Estimation of the total number of residents in Toronto:

$$\hat{\tau} = N\bar{y} = 140 \times 15.817 = 2214.38 \; (= 2{,}214{,}380 \text{ residents})$$

$$\hat{\sigma}_{\hat{\tau}} = N \times \hat{SD}(\bar{y}) = 140 \times 2.408898 = 337.24572$$

$$B_{\tau} = 2 \times 337.24572 = 674.49144 = 674.49$$

CI for τ: $\hat{\tau} \pm B_{\tau} = 2214.38 \pm 674.49 = [1539.89, 2888.87]$

or, $[1{,}539{,}890; 2{,}888{,}870]$

example

Simple Random Sampling: Examples (I)

Population of Toronto neighbourhoods (III)

Standard deviation:

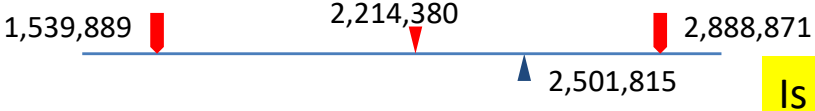
$$\hat{\sigma}^2 = \hat{\sigma}_y^2 = \tilde{S}^2 = \frac{N-1}{N} S^2 = \tilde{S}^2 = \frac{140-1}{140} 62.49158 = 62.04521$$

$$\hat{\sigma} = \hat{\sigma}_y = \tilde{S} = \sqrt{62.04521} = 7.87688 = 7.877$$

Summary of results:

	Sample (± B)	Actual
Average per nd	15,817 (± 4,818)	17,870
Standard deviation	7,877 (± ?)	8,431
Total population	2,214,380 (± 674,491)	2,501,815

Are the sample and actual results consistent?



Is the interval wide? 1) … 2) …

Exercise: 1) Complete calculation for 2001 population (variable x) .
2) Take another sample of size 10 and compare results.

example

Simple Random Sampling: Examples (II)

Science and Medicine Library: Collection of statistical books (I)

Goal of the study: Investigate some properties of the Collection (e.g., find the number of books in the collection, percentage of books with original cover, …)

Target population: Statistical books (Library classification call number QA273-280)

Sampling population: All shelves carrying statistical books

Sampling units: Shelves (why not books?)

Population size: 161 (as it was in 1998)

Variables (characteristics) of interest: Number of books per shelf (y), number of books with original cover per shelf (x)

original cover

rebounded books

Gerstein Science Information Centre

example

Simple Random Sampling: Examples (II)

Science and Medicine Library: Collection of statistical books (II)

Sampling frame: Floor plan, with enumeration of shelves

	1	2	3	4	5
1	1	7	13	19	25
2	2	8	14	20	26
3	3	9	15	21	27
4	4	10	16	22	28
5	5	11	17	23	29
6	6	12	18	24	30

	1	2	3	4	5	6
1	31	38	45	52	59	66
2	32	39	46	53	60	67
3	33	40	47	54	61	68
4	34	41	48	55	62	69
5	35	42	49	56	63	70
6	36	43	50	57	64	71
7	37	44	51	58	65	72

	1	2	3	4	5
1	73	79	85	91	97
2	74	80	86	92	98
3	75	81	87	93	99
4	76	82	88	94	100
5	77	83	89	95	101
6	78	84	90	96	102

	1	2	3	4
1	103	109	115	121
2	104	110	116	122
3	105	111	117	123
4	106	112	118	124
5	107	113	119	125
6	108	114	120	126

	1	2	3	4	5
1	127	134	141	148	155
2	128	135	142	149	156
3	129	136	143	150	157
4	130	137	144	151	158
5	131	138	145	152	159
6	132	139	146	153	160
7	133	140	147	154	161

Five blocks, 25 book cases, 161 shelves

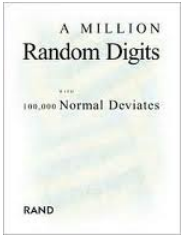
example

Simple Random Sampling: Examples (II)

Science and Medicine Library: Collection of statistical books (III)

Sample design: Simple Random Sampling

Sample size: 20



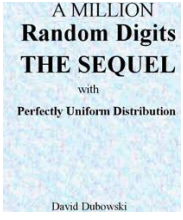


Table of Random Numbers: N = 161
(three digits, modulo 200)

098	300	098	910	331	258	741	625	755	839	390	691	207
98	**100**	~~98~~	**110**	**131**	**58**	**141**	**25**	**155**	**39**	~~190~~	**91**	**7**

841	287	496	374	576	086	507	935	195	339	948	528	532
41	**87**	**96**	~~172~~	~~176~~	**86**	**107**	**135**	~~195~~	**139**	**148**	**128**	-

Sample: 7, 25, 39, 41, 58, 86, 88, 91, 96, 98, 100, 107, 110, 128, 131, 135, 139, 141, 148, 155

Sample:
(first value y,
second value x)

	1	2	3	4	5
1		20, 9			23,10
2					
3					
4					
5					
6					

	1	2	3	4	5	6
1						
2		25,19				
3						
4		29,25				
5						
6						
7				27,19		

	1	2	3	4	5
1				32,18	
2			21,13		
3					
4			20,14		
5					
6				24,17	

	1	2	3	4
1				
2		40,22		
3				
4				
5	15,13			
6				

	1	2	3	4	5
1			13,5	21,8	17,10
2	16,4	19,8			
3					
4					
5	14,10				
6		18,8			
7					

example

Simple Random Sampling: Examples (II)

Science and Medicine Library: Collection of statistical books (IV)

Simple Random Sample: Population size N = 161. Sample size n = 20. Sample:

y	20	23	25	29	27	21	32	24	32	32	15	40	16	14	19	18	13	21	17	20
x	9	10	19	25	19	13	18	17	14	22	13	22	4	10	8	8	5	8	10	14

Calculation:

$\Sigma y = 458,\ \Sigma y^2 = 11474,\quad \Sigma x = 268,\ \Sigma x^2 = 4272,\ \Sigma xy = 6790,$

$\bar{y} = 22.9,\ S_y^2 = 51.8842,\ \bar{x} = 13.4,\ S_x^2 = 35.8316$

Estimation:

$$\hat{\mu}_y = \bar{y} = 22.9 \qquad \hat{\tau}_y = N\bar{y} = 161 \times 22.9 = 3686.9,$$

$$\hat{Var}(\bar{y}) = \frac{N-n}{N}\frac{S_y^2}{n} = \frac{161-20}{161} \times \frac{51.8842}{20} = 2.272$$

$$\hat{Var}(\hat{\tau}_y) = N^2 \hat{Var}(\bar{y}) = 161^2 \times 2.272$$

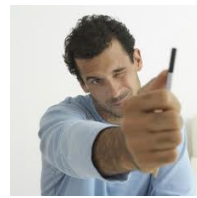
$$\hat{Sd}(\bar{y}) = 1.507 \qquad \hat{Sd}(\hat{\tau}) = 161 \times 1.507 = 242.675$$

CI for μ: $\bar{y} \pm B_\mu = 22.8 \pm 3.0 = [19.8, 25.8]$,

CI for τ: $\hat{\tau} \pm B_\tau = 3686.9 \pm 485.4 = [3191.6,\ 4172.3]$

Discussion

10% ?



You: $\hat{\mu}_x = 13.4,$

$\hat{\tau}_x = 2157.4,$

$\hat{Sd}(\hat{\mu}_x) = 1.253,$

$\hat{Sd}(\hat{\tau}_x) = 201.670$

Your read **example**

Simple Random Sampling: Examples (III)

Farms and cattle (I)

Population: 75,308 farms from a province

Variable: Number of cattle on farm (y)

Parameters to be estimated: Average number of cattle per farm (μ_y), total number of cattle in the province (τ_y), proportion of farms with at least 31 cattle (p), …

Sample: SRS of 2,072 farms (results given in a form of frequency distribution)

# of cattle y_j	Midpoint y_j'	# of farms f_j (frequency)
0	0	263
1-5	3	403
6-10	8	439
11-15	13	310
16-20	18	245
21-25	23	146
26-30	28	108
31-35	33	60
36-40	38	44
41 +	43	54
	Total	2072

Calculation (using midpoints):

$$\bar{y} = \sum y_j' f_j / n = (0 \times 263 + 3 \times 403 + \cdots + 4 \times 354) / 2072 = \mathbf{12.315}$$

$$S^2 = (\sum y_j'^2 f_j - n\bar{y}^2)/(n-1)$$

$$= (0^2 \times 263 + 3^2 \times 403 + \cdots + 4^2 \times 354 - 2072 \times 12.315^2) / 2071 = \mathbf{115.826}$$

Calculation (from complete sample):

$$\bar{y} = 12.492,\ S^2 = 133.342$$

Your read **example**

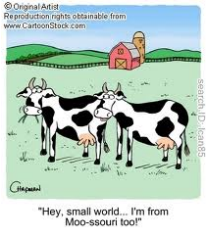
Simple Random Sampling: Examples (III)

Farms and cattle (II)

Estimation, mean:

$$\hat{\mu}_y = 12.315$$

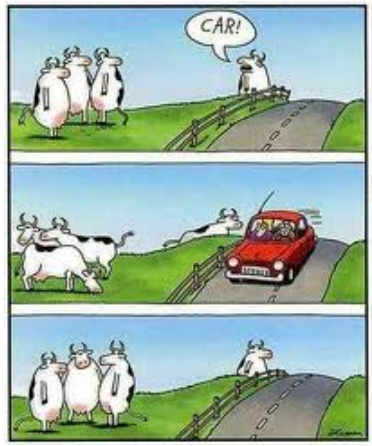
$$B_\mu = 2 \times 0.233 = 0.466$$



$$\hat{\sigma}_{\hat{\mu}} = \sqrt{\frac{N-n}{N}\frac{S_y^2}{n}} = \sqrt{\frac{S^2}{2072} \times \frac{75308-2072}{75308}} = \sqrt{0.05436} = 0.233$$

CI for μ: $\bar{y} \pm B_\mu = 12.315 \pm 0.466 = [11.85, 12.78]$

Estimation, total:



$$\hat{\tau}_y = N\eta_y = 75{,}308 \times 12.315 = 927{,}418$$

$$\hat{\sigma}_{\hat{\tau}} = N\hat{\sigma}_{\hat{\mu}} = 75{,}308 \times 0.233 = 17{,}546.8 \qquad B_\tau = 35{,}094$$

CI for τ: $\hat{\tau} \pm B_\tau = 927{,}418 \pm 35{,}094 = [892{,}324;\ 962{,}512]$

Your read example

Simple Random Sampling: Examples (III)

Farms and cattle (III)

Required error bound: If an error bound $B_\tau = 25{,}000$ is required, and from a *presample* only range 0 – 50 for y is known, then

$$\hat{\sigma} = (50-0)/4 = 12.5 \quad D = \left(B_\tau/2N\right)^2 = 0.027551$$

$$\hat{n} = \frac{N\hat{\sigma}^2}{ND+\hat{\sigma}^2} = \frac{75308\times(12.5)^2}{75308\times D+(12.5)^2} = 5274.117 = 5275$$

Proportion of farms with at least 31 cattle:

$$\hat{p} = \frac{a}{n} = \frac{60+44+54}{2072} = 0.07625 = 7.625\%$$

$$\hat{\sigma}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{2072-1}\times\frac{75308-2072}{75308}} = 0.00575 \quad B_p = 0.0115 = 1.15\%$$

CI for p: $\hat{p} \pm B_p = 7.625\% \pm 1.15\% = [6.475\%;\ 8.775\%]$

Number of farms with at least 31 cattle:

$$\hat{\tau}' = N\hat{p} = 75{,}308\times 0.07625 \ = \ 5{,}742 \ \text{(find CI for } \tau')$$

Simple Random Sampling: Some New Useful Notation

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2$$ Population variance

$$\tilde{\sigma}_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)^2$$ Adjusted population variance

$$\sigma_y^2 = \frac{N-1}{N}\tilde{\sigma}_y^2, \qquad \tilde{\sigma}_y^2 = \frac{N}{N-1}\sigma_y^2$$

$$= (1-f)\tilde{\sigma}_y^2/n$$

$$Var(\bar{y}) = \frac{N-n}{N-1}\frac{\sigma_y^2}{n} = \frac{N-n}{N}\frac{N}{N-1}\frac{\sigma_y^2}{n} = \frac{N-n}{N}\frac{\tilde{\sigma}_y^2}{n} = (1-\frac{n}{N})\frac{\tilde{\sigma}_y^2}{n}$$

A convenient and useful notation

$$E(S^2) = \frac{N}{N-1}\sigma^2 = \tilde{\sigma}^2 \text{ (SRS without replacement)}$$

$$\hat{\tilde{\sigma}}^2 = S^2$$ an unbiassed estimator of $\tilde{\sigma}^2$

$$\hat{Var}(\bar{y}) = \frac{N-n}{N}\frac{\hat{\tilde{\sigma}}_y^2}{n} = \frac{N-n}{N}\frac{S^2}{n}$$ an unbiassed estimator of $Var(\bar{y})$

$1-f$ Finite population correction (fpc)

$f = \frac{n}{N}$ Sampling fraction

Simple Random Sampling: Inference (VIII)

Selecting the sample size (I)

Estimation of μ and τ : $\hat{\mu} = \bar{y}$ and $\hat{\tau} = N\bar{y}$

Accuracy measured by B_μ and B_τ $\quad B_\mu = 2 \times SD(\bar{y}), B_\tau = N \times B_\mu$

Two criteria to select sample size n:

I Cost of sampling

II Given error bound (accuracy, standard error)

A wise man once said: "Never begin data collection without calculating the necessary sample size!"

Case I: Cost of sampling/total cost

= fixed cost + cost of actual sampling

→ Cost model – a function of sample size

Linear model: $\boxed{C = C(n) = c_0 + c_1 \times n}$ $\quad (C = C(n) = c_0 + c_1 \times \sqrt{n})$

c_0 – fixed cost, c_1 – cost of sampling one unit

fixed cost = preparation + inference +…

Simple Random Sampling: Inference (VIII)

Selecting the sample size (II)

Case I: Total cost given, find sample size *n*

- simple for SRS and linear model

$$n = \frac{C - c_0}{c_1} = \frac{C'}{c_1}, \quad C' = C - c_0$$

Example : $c_0 = \$20,\ c_1 = \$5,$

$$C = \$100 \Rightarrow n = \frac{100 - 20}{5} = 16$$

Note: *n* above is the *largest* value that can be obtained for that money (C). Then, the variance/error bound is the *smallest* that can be obtained for that money .

Case II: Given error bound (variance), find sample size *n*

We start with $B_\mu = 2 \times SD(\bar{y}) = 2\sqrt{Var(\bar{y})}$

$$Var(\bar{y}) = \left(\frac{B_\mu}{2}\right)^2 = D, \quad \frac{N-n}{N-1}\frac{\sigma_y^2}{n} = \frac{N-n}{N}\frac{\tilde{\sigma}_y^2}{n} = D$$

S^2

Simple Random Sampling: Inference (VIII)

Selecting the sample size (III)

Solve for n: $\tilde{\sigma}_y^2(N-n) = DNn, \quad \tilde{\sigma}_y^2 N = n(\tilde{\sigma}_y^2 + DN)$

$$n = \frac{N\tilde{\sigma}_y^2}{ND + \tilde{\sigma}_y^2} = \frac{N\sigma_y^2}{(N-1)D + \sigma_y^2}$$

often in class | book

Required sample size for given B_μ (D)

$$n = \frac{N\tilde{\sigma}_y^2}{ND + \tilde{\sigma}_y^2} = \frac{\tilde{\sigma}_y^2}{D + \frac{\tilde{\sigma}_y^2}{N}} \to \frac{\tilde{\sigma}_y^2}{D}, N \to \infty \qquad D = \left(\frac{B_\mu}{2}\right)^2$$

In practice, for large N: and it does not depend on N!

$$n \approx \frac{\tilde{\sigma}_y^2}{D} \approx \frac{\sigma_y^2}{D}$$

$$\frac{\sigma_y^2}{D} = \left(\frac{2\sigma_y}{B_\mu}\right)^2$$

Notice this and discuss!

Simple Random Sampling: Inference (VIII)

Selecting the sample size (IV)

Let $n_0 = \dfrac{\widetilde{\sigma}_y^2}{D}$

$$n = \frac{n_0}{1 + \dfrac{n_0}{N}} \leq n_0 = n_{max}$$

"Infinite population" case

For large *N*, or no good knowledge about *N*, use n_0

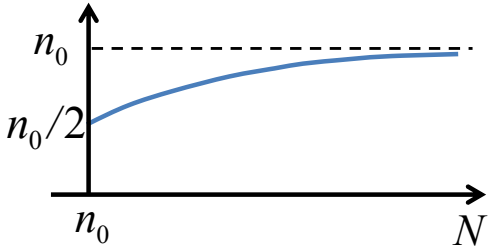
Common mistake laymen do: 10 times bigger population, 10 times bigger sample (Discussion: No, if …? Yes, if …?)

Some simple math: $n_0 = 1000$

N = 10,000	n = 910
N = 100,000	n = 991
N = 1,000,000	n = 999
N = 10,000,000	n = 1000

Remember

$$\frac{\sigma_y^2}{D} = \left(\frac{2\sigma_y}{B_\eta}\right)^2$$



$$N \geq 10n_0 \Rightarrow 0.91n_0 \leq n \leq n_0$$

Simple Random Sampling: Inference (VIII)

Selecting the sample size (V)

Practical problem: $\tilde{\sigma}_y^2$ (or σ_y^2) – usually unknown; we have to estimate it somehow beforehand

- Use a previous study, experience, educated guess, …
- Use a presample (usually more complex studies)
- Quick and dirty method: use the range of variable y

Using range: $\hat{\tilde{\sigma}}_y = \frac{Range}{4}$

Example: Test/exam mark range 50 – 100 points

$$\hat{\tilde{\sigma}} = \frac{100-50}{4} = 12.5$$

My experience in calculating SD from several exams: 12.0 – 14.5

Simple Random Sampling: Inference (VIII)

Selecting the sample size (VI)

Example: Toronto neighbourhoods population.

Absolute range: 6 – 52.5 (x 1,000)

$\hat{\tilde{\sigma}}_1 = \frac{52.5 - 6}{4} \approx 11.5$ Actual $\sigma \approx 8.5$ An overestimation

Out of 140, only 5 are between 35 and 52.5

$\hat{\tilde{\sigma}}_2 = \frac{35 - 6}{4} \approx 7.5$ A slight under estimation

$$\hat{n} = \frac{N\hat{\tilde{\sigma}}^2}{ND + \hat{\tilde{\sigma}}^2}$$

We use in actual calculation

You try some other cases

Simple Random Sampling: Inference (VIII)

Selecting the sample size (VII)

For total τ: $\hat{\tau} = N\bar{y}$, B_τ given:

We start with $B_\tau = 2N\sqrt{Var(\bar{y})}$ $\qquad Var(\bar{y}) = \left(\frac{B_\tau}{2N}\right)^2 = D,$

Same equation as for μ

$$n = \frac{N\tilde{\sigma}_y^2}{ND + \tilde{\sigma}_y^2} = \frac{N\sigma_y^2}{(N-1)D + \sigma_y^2} \begin{cases} D = D_\mu = \left(\frac{B_\mu}{2}\right)^2 & \text{for } \mu \\ D = D_\tau = \left(\frac{B_\tau}{2N}\right)^2 & \text{for } \tau \end{cases}$$

Simple Random Sampling: Inference (VIII)

Selecting the sample size (VIII)

Example: Toronto neighbourhoods population. Find sample size of an SRS to estimate the average neighbourhood size with an error bound of 3 (thousands).

Assume/guess/estimate σ ≈ 10.

$$n = \frac{140 \times 10^2}{140 \times (3/2)^2 + 10^2} = 33.73 = 34 \quad \text{(roundup!)}$$

$$D = (B_\mu/2)^2$$

If error bound 5 is required:

$$n = \frac{140 \times 10^2}{140 \times (5/2)^2 + 10^2} = 14.36 = 15 \quad \text{(roundup!)}$$

Compare:

For $\sigma \approx 8.5, B_\mu = 5,$

$n_{max} = 12$

If error bound 5 is required, but estimate σ ≈ 8.5 :

$$n = \frac{140 \times (8.5)^2}{140 \times (5/2)^2 + (8.5)^2} = 10.68 = 11$$

Compare: In our sample

$n = 10$ and $\hat{B}_\mu = 4.82$

Simple Random Sampling: Inference (VIII)

Selecting the sample size (IX)

Smaller error bound, bigger sample. What is the rule?

In our example: For σ ≈ 8.5

B_μ = 5 ⟹ n = 11

B_μ = 2.5 ⟹ n = 35

It is not 2 times larger, but more!

For large N, $n \approx \dfrac{\sigma_y^2}{D}$

Compare different error bounds

$$B_\mu : B_1 \text{ and } n_1, B_2 \text{ and } n_2; B_2 = B_1/k \Rightarrow n_2 = k^2 n_1$$

Rule: k times smaller error bound, k^2 time bigger sample!

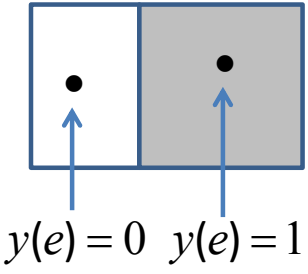
SRS: Population Proportion, Ch. 4.5 (I)

Introduction

Proportion of elements which poses certain property, or belong to certain specified group. Define variable ***y*** taking two values (binary variable):

$$y(e)=\begin{cases}0\,,\ e \text{ does not have the property}\\ 1\,,\ e \text{ has the property}\end{cases}$$

$$p=\frac{1}{N}\sum_{1}^{N}y(e_i)=\frac{M}{N}=\frac{\#\{\text{elements with the property}\}}{N}=\mu_y$$



$$\mu_y=\mu=0\times(1-p)+1\times p=p$$

$$\tau=\tau_y=N\mu=Np=M$$

Important!

SRS: Population Proportion (II)

Using sample: Estimating proportion

SRS of size n, a = # elements with the property in the sample

element	1	2	3	⋯	n - 1	n
sample	1	0	1	⋯	1	0
	yes	no	yes		yes	no

$$a = \sum_1^n y_i$$

count of ones

Estimator of *p*:

We only need to know a

$$\hat{p} = \bar{y} = \frac{1}{n}\sum_1^n y_i = \frac{a}{n}$$

relative frequency

unbiased estimators

$$\hat{M} = \hat{\tau} = N\hat{p} = N\frac{a}{n}$$

We apply our general theory about estimating μ and τ from SRS!

example

SRS: Population Proportion (II)

Example: Estimating proportion (I)

Left side **Me** Right side

1	2 F	3 F	4 F	5 M	6 F
13	14 M	15 F	16 F	17 F	18 F
25	26	27 M	28 M	29 F	30 F
37 F	38 F	39 F	40 M	41 F	42 F
49 M	50 M	51 F	52 F	53 F	54
61	62 M	63 M	64 F	65 F	66
73	74 M	75 F	76 F	77 F	78 M
85	86 M	87 F	88 F	89 F	90 M
97	98 M	99	100	101 F	102 M
109	110	111 M	112 M	113 M	114 M

7 F	8 F	9 M	10	11	12
19 F	20 F	21 F	22 M	23 M	24
31 F	32 F	33 M	34 M	35 M	36
43 F	44 F	45 F	46 M	47 M	48
55 M	56 F	57 F	58 F	59 M	60 M
67 M	68 M	69 F	70 F	71 F	72 M
79 F	80 F	81	82	83 M	84 M
91 M	92 M	93 F	94 F	95 F	96 F
103 F	104 M	105	106	107	108
115 M	116 M	117	118	119	120

Population: Class of 90 students

Task: Select 10 students at random and estimate proportion of girls

occupied

unoccupied

(F: 27 + 30 = 50)

example

SRS: Population Proportion (II)

Example: Estimating proportion (II)

Selecting an SRS of size 10: Use 3 digits from TRN (frame size *N*'=120), mod 200, start, e.g., at line 6, read first 3 out of 5.

TRN	779	069	110	427	277	534	186	706	906	150	219	818	443	428
read	~~179~~	69	~~110~~ un	27	77	~~134~~	~~186~~	~~106~~ un	~~106~~ un	~~150~~	19	18	43	28
TRN	995	729	564	699	988	310	711	187	440	488	632	210	106	129
read	~~195~~	~~129~~	~~164~~	99	~~188~~	~~110~~ un	111	~~187~~	40	88	stop			

Sample	18	19	27	28	40	43	69	77	88	111
Gender	F	F	M	M	M	F	F	F	F	M
y	1	1	0	0	0	1	1	1	1	0

← After looking at the class

Estimation: $a = 6 \Rightarrow \hat{p} = \frac{6}{10} = 0.6 = 60\%$, $\hat{\tau} = 90 \times \hat{p} = 90 \times 0.6 = 54$ girls

Exact value from the population : $p = \frac{M}{N} = \frac{50}{90} = 0.55 = 55\%$

Exact error of estimation : $|p - \hat{p}| = |0.55 - 0.60| = 0.05 = 5\%$

$\hat{Var}(\hat{p}) = ?$

example

SRS: Population Proportion (II)

Example: Estimating proportion (III)

What about boys?:

$$\hat{q} = 1 - \hat{p} = 1 - 0.6 = 0.4 = 40\%, \hat{\tau}_B = 90 \times \hat{q} = 90 \times 0.4 = 36 \text{boys}$$

Exact value from the population : $q = \frac{N - M}{N} = \frac{40}{90} = 0.45 = 45\%$

Exact error of estimation : $|q - \hat{q}| = |0.45 - 0.40| = 0.05 = 5\%$

example

SRS: Population Proportion (II)

Example: Estimating proportion (IV)

Left side **Me** Right side

1	2 F	3 F	4 F	5 M	6 F
13	14 M	15 F	16 F	17 F	18 F
25	26	27 M	28 M	29 F	30 F
37 F	38 F	39 F	40 M	41 F	42 F
49 M	50 M	51 F	52 F	53 F	54
61	62 M	63 M	64 F	65 F	66
73	74 M	75 F	76 F	77 F	78 M
85	86 M	87 F	88 F	89 F	90 M
97	98 M	99	100	101 F	102 M
109	110	111 M	112 M	113 M	114 M

7 F	8 F	9 M	10	11	12
19 F	20 F	21 F	22 M	23 M	24
31 F	32 F	33 M	34 M	35 M	36
43 F	44 F	45 F	46 M	47 M	48
55 M	56 F	57 F	58 F	59 M	60 M
67 M	68 M	69 F	70 F	71 F	72 M
79 F	80 F	81	82	83 M	84 M
91 M	92 M	93 F	94 F	95 F	96 F
103 F	104 M	105	106	107	108
115 M	116 M	117	118	119	120

Population: Class of 90 students
Sample: 10 selected students

occupied

unoccupied

occupied selected

unoccupied selected

proof

SRS: Population Proportion (II)

Using sample: Estimating variance (I)

Estimating $Var(\hat{p})$:

$$Var(\hat{p}) = \frac{N-n}{N-1}\frac{\sigma_y^2}{n} = \frac{N-n}{N}\frac{\tilde{\sigma}_y^2}{n} \qquad \sigma_y^2 = ?, \text{or } \tilde{\sigma}_y^2 = ?$$

$$\mu_y = p, y_i = \begin{cases} 0 \\ 1 \end{cases} \Rightarrow y_i^2 = \begin{cases} 0 \\ 1 \end{cases}$$

$$\sigma_y^2 = \frac{1}{N}\sum_i^N (y_i - \mu)^2 = \frac{1}{N}\sum_i^N y_i^2 - \mu^2$$

$$= \frac{1}{N}\sum_i^N y_i - p^2 = p - p^2 = p(1-p) = pq$$

$$\sigma_y^2 = p(1-p) = pq\,, \quad \tilde{\sigma}_y^2 = \frac{N}{N-1}\sigma_y^2 = \frac{N}{N-1}pq$$

$$\boxed{Var(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n}}$$

Variance, theoretical formula

(**Warning**: I might ask you to calculate the theoretical variance of $\hat{p}$)

proof

SRS: Population Proportion (II)

Using sample: Estimating variance (II)

A small class

Estimating $Var(\hat{p})$: $\hat{\mu} = \hat{p}$

$$\tilde{\sigma}_y^2 = S^2 = \frac{1}{n-1}\sum_i^n (y_i - \hat{p})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_i^n y_i^2 - \hat{p}^2\right]$$

$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_1^n y_i - \hat{p}^2\right] = \frac{n}{n-1}\left[\hat{p} - \hat{p}^2\right] = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

$$\hat{Var}(\hat{p}) = \frac{N-n}{N}\frac{\tilde{\sigma}_y^2}{n} = \frac{N-n}{N}\frac{S^2}{n} = \frac{N-n}{N}\frac{1}{n}\frac{n}{n-1}\hat{p}(1-\hat{p})$$

$$= \frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n-1} = \frac{N-n}{N}\frac{\hat{p}\hat{q}}{n-1}$$

an unbiased estimator

$$\boxed{\hat{Var}(\hat{p}) = \frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n-1} = \frac{N-n}{N}\frac{\hat{p}\hat{q}}{n-1}}$$

Variance, estimated

example

SRS: Population Proportion (II)

Using sample: Estimating variance (III)

Error bound: $B_p = 2\hat{SD}(\hat{p}) = 2\sqrt{\hat{Var}(\hat{p})}$

CI for p: $\hat{p} \pm B_p$

Example, continued (class of 90 students):

$$\hat{p} = 0.6 = 60\%, \hat{q} = 0.4 = 40\%$$

$$B_p = 2\sqrt{\hat{Var}(\hat{p})} = 2\sqrt{\frac{90-10}{90}\frac{0.6\times 0.4}{10-1}} = 2\sqrt{0.0237}$$

$$= 2\times 0.15396 = 0.3079 \approx 0.308 = 30.8\% \approx 31\%$$

keep more decimal places when calculating variance, it is usually a small value

CI for p: $\hat{p} \pm B_p = 0.60 \pm 0.308 = [0.292, 0.908] \approx [29\%, 91\%]$

Not great, interval is quite wide (confidence is ≈ 95%).

Don't complain, this is the best you can buy for that money you paid (for a sample of size 10)

SRS: Population Proportion (II)

Using sample: Estimating variance (IV)

Estimator of *q*: $\hat{q} = 1 - \hat{p} = \dfrac{n-a}{n}$

$$Var(\hat{q}) = Var(1-\hat{p}) = Var(\hat{p}) = \frac{N-n}{N-1}\frac{p(1-p)}{n} = \frac{N-n}{N-1}\frac{pq}{n}$$

$$\hat{Var}(\hat{q}) = \hat{Var}(\hat{p}) = \frac{N-n}{N}\frac{\hat{p}\hat{q}}{n-1}, \quad B_q = B_p$$

CI for q: $\hat{q} \pm B_q = 1 - \hat{p} \pm B_p = 1 - [\hat{p} \mp B_p]$

In our example, $\hat{q} = 1 - 0.6 = 0.4 = 40\%$

CI for q: $100\% - [91\%, 29\%] = [9\%, 71\%]$ (you do direct calculation)

SRS: Population Proportion (III)

Selecting the sample size (I)

Just apply general theory: remember, *p* is *μ* for simple variable

$$D = D_p = \left(\frac{B_p}{2}\right)^2 \qquad \boxed{n = \frac{N\sigma_y^2}{(N-1)D + \sigma_y^2} = \frac{Npq}{(N-1)D + pq}}$$

Find some *preestimate* or *guess* for *p* and use it to calculate *n*

In our example, for sample of $n = 10$ students $\Rightarrow B_p \approx 0.31$

What sample size should we use if we want $B_p = 0.2$?

Guess *p* ≈ 0.5 (girls and boys)

$$n = \frac{90 \times 0.5 \times 0.5}{89 \times 0.01 + 0.5 \times 0.5} = \frac{90 \times 0.25}{89 \times 0.01 + 0.25} = 19.73 \Rightarrow n = 20$$

$$D = \left(\frac{0.2}{2}\right)^2 = 0.01$$

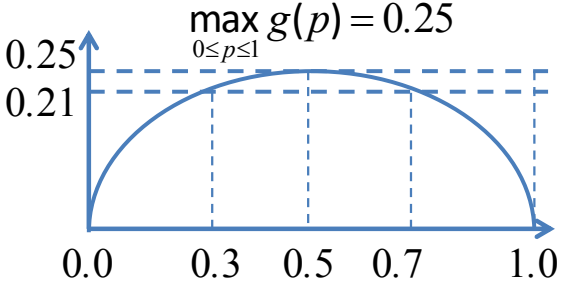
How good/bad is our calculation?

SRS: Population Proportion (III)

Selecting the sample size (II)

Look at *n* as a function of *p*(*1-p*):

$$n_p = \frac{Np(1-p)}{(N-1)D + p(1-p)} = \frac{Ng}{(N-1)D + g} \uparrow \text{ in } g = p(1-p)$$



n_p decreases when p decreases from 0.5 to 0.0, or increases from 0.5 to 1.0.

p	$p(1-p)$
0.5	0.25
0.4, 0.6	0.24
0.3, 0.7	0.21
0.2, 0.8	0.16

$$n_p \le n_{max} = \frac{N \times 0.25}{(N-1)D + 0.25}$$

How do we use this "mathematics"?

SRS: Population Proportion (III)

Your read

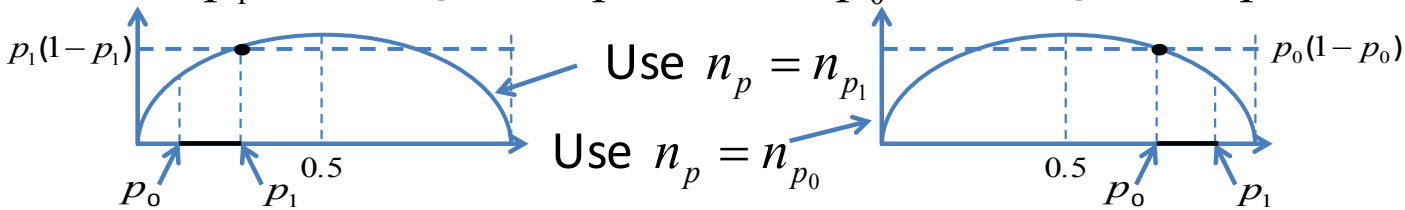
Selecting the sample size (III)

Two basic rules:

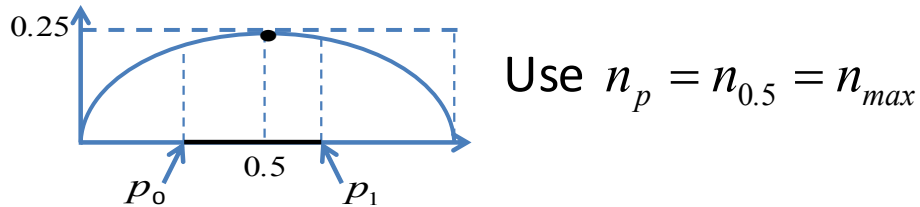
Rule A: If no information on *p* is available, use $n_p = n_{max}\ (p = 0.5)$

Rule B: If it is known/guessed that $p_0 \le p \le p_1$, for some p_0 and p_1 (e.g., $0.1 \le p \le 0.2$, or $0.3 \le p \le 0.6$), then

Case B1: $p_1 \le 0.5$ (e.g., $0.1 \le p \le 0.2$), or $p_0 \ge 0.5$ (e.g., $0.7 \le p \le 1.0$)



Case B2: $p_0 \le 0.5 \le p_1$ (e.g., $0.3 \le p \le 0.6$)



Use $n_p = n_{0.5} = n_{max}$

Information $p_0 \le 0.5 \le p_1$ does not help

example

SRS: Population Proportion (III)

Your read

Selecting the sample size (IV)

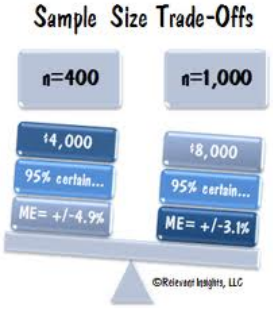
Example, with girls and boys: For $B_p = 0.2, D = \left(\frac{0.2}{2}\right)^2 = 0.01$ and no information $(p = 0.5)$

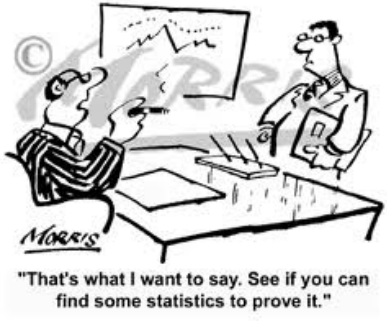
$$n \le n_{max} = \frac{90 \times 0.5 \times 0.5}{89 \times 0.01 + 0.5 \times 0.5} = 19.76 = 20$$

If (case B1) $p \le 0.1$ (unlikely for girls)

$$n \le n_{0.1} = \frac{90 \times 0.1 \times 0.9}{89 \times 0.01 + 0.1 \times 0.9} = 8.26 = 9$$

Twice smaller!





"That's what I want to say. See if you can find some statistics to prove it."

example

SRS: Population Proportion (III)

Ipsos Reid Error bound and sample size (VI)

Recent Ipsos polls: Ipsos Reid surveyed 1,004 Canadians between July 23 and July 30 with a 3.2 per cent margin of error, 19 times out of 20.
(= 95% confidence)

A: Only 36 per cent of Canadians believe federal health care will improve in the next two or three years, ...

B: Furthermore, the report found 38 per cent of Canadians are happy with the Harper government's health-care record.

1) Question: Why 3.2% margin of error (error bound)?

Answer: $$B_p \approx 2\sqrt{Var(\hat{p})} = 2\sqrt{\frac{p(1-p)}{n}} \le 2\sqrt{\frac{0.5\times 0.5}{n}}$$

$$= \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1004}} = 0.03\,16 \approx 3.2\%$$

Canadians, a large population

So, the estimate of any proportion will have 3.2% margin of error, at most.