Sampling Theory and Methods

Spring 2008

C. L. Williams

Chapter 3 Simple Random Sampling

Outline

1 Chapter 3

Simple Random Sample(SRS)-Total ways

Consider the physicians data. Recall there were 25 physicians with the number of visits for each of the physicians recorded. Suppose we wish to construct all possible samples of size $n=5$. We have

$$T = \frac{25!}{(5!)\,(25-5!)}$$

total samples of size 5.

SRS-Probabilty not being included

In the simple random sampling scheme we can establish the probability of being selected in a sample:

$$\begin{aligned}
\text{Probabilty not being included} &= \frac{\binom{N-1}{n}}{\binom{N}{n}} \\
&= \frac{\frac{N-1!}{(N-1-n)!n!}}{\frac{N!}{(N-n)!n!}} \\
&= \frac{(N-n)}{N}
\end{aligned}$$

SRS-Probabilty being included

$$
\begin{aligned}
&= 1 - \left(\frac{N-n}{N}\right) \\
&= \left(\frac{n}{N}\right).
\end{aligned}
$$

We have to take into consideration these "likelihoods" of being included or not included in the sample in constructing estimates of our population parameters.

Some key things to remember

Sample statistics-Total, mean and proportion

- **Sample Total:** The sample total is generally denoted by x_T and is the sum of the values over all elements in the sample:

$$x_T = \sum_{i=1}^{n} x_i$$

- **Sample Mean:** The sample mean generally denoted by $\bar{x}$ and is the sum of the values in the sample divided by the sample size:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Sample Proportion:** The sample proportion of a dichotomous characteristic is generally denoted by p_x is given by

$$p_x = \frac{x_t}{n}$$

Sample statistics

- **Sample Variance and standard deviation:** The sample variance s_x^2 is given by:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- **Dichotomous**

$$s_x^2 = \frac{np_x(1-p_x)}{n-1}$$

and for large sample sizes (> 20) an approximation can be used

$$s_x^2 = p_x(1-p_x)$$

- **Sample standard deviation**

$$s_x = \sqrt{p_x(1-p_x)}$$

Estimation of Population Characteristics

An estimate of the population total X_T can be obtained from the sample total x_T as given

$$x'_T = \left(\frac{N}{n}\right)(x)$$

An estimate σ_x^2 of the population variance σ_x^2 is given by

$$\hat{\sigma}_x^2 = \left(\frac{N-1}{N}\right)(s_x^2)$$

If the number of elements N in the population large, (N-1)/N:

$$\hat{\sigma}_x^2 \approx s_x^2$$

Total Estimates

So in determining estimates of the population total under simple random sampling we would construct our estimates by:

$$\begin{aligned} x_t' &= \frac{N\sum_{i=1}^{n} x_i}{n} \\ \widehat{Var(x_t')} &= N^2\left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\ \widehat{SE(x_t')} &= N\sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right) \end{aligned}$$

Note the fpc factor. Intuitively, we make this correction because with small populations the greater the sampling fraction $\left(\frac{n}{N}\right)$, (ie. the more likely an element will be included) the more information we have about the population and hence the smaller the variability.

Mean Estimates

$$\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
\widehat{Var(\overline{x})} &= \left(\frac{N-n}{N}\right)\left(\frac{s_x^2}{n}\right) \\
\widehat{SE(\overline{x})} &= \sqrt{\frac{N-n}{N}}\left(\frac{s_x}{\sqrt{n}}\right)
\end{aligned}$$

Proportion Estimates

$$
\begin{aligned}
p_y &= \frac{\sum_{i=1}^{n} y_i}{n} \\
\widehat{Var}(p_y) &= \left(\frac{N-n}{N}\right)\frac{p_y(1-p_y)}{n-1} \\
\widehat{SE}(p_y) &= \sqrt{\left(\frac{N-n}{N}\right)}\sqrt{\frac{p_y(1-p_y)}{n-1}}
\end{aligned}
$$

Simple Random Sampling

Schools in Sample	Total_i	Schools in Sample	Total_i
1,2,3	24	2,3,4	22
1,2,4	24	2,3,5	30
1,2,5	32	2,3,6	32
1,2,6	34	2,4,5	30
1.3,4	20	2,4,6	32
1,3,5	28	2,5,6	40
1,3.6	30	3,4,5	26
1,4,5	28	3,4,6	28
1,4.6	30	3,5,6	36
1,5,6	38	4,5,6	36

Total_i	f_i	$\pi_i = \frac{f_i}{T}$
20	1	.05
22	1	.05
24	2	.10
26	1	.05
28	3	.15
30	4	.20
32	3	.15
34	1	.05
36	2	.10
38	1	.05
40	1	.05
Total	20	1.00

$$\begin{aligned} E(x_t') &= 20(0.05) + 22(0.05) + 24(0.10) + 26(0.05) \\ &+ 28(0.15) + 30(0.20) + 32(0.15) + 34(0.05) \\ &+ 36(0.10) + 38(0.05) + 40(0.05) \\ &= 30. \end{aligned}$$

The variance $Var(x_t')$ of the sampling distribution of x_t' is

$$\begin{aligned} Var(x_t') &= (20-30)^2(0.05) + (22-30)^2(0.05) + (24-30)^2(0.10) \\ &+ (26-30)^2(0.05) + (28-30)^2(0.15) + (30-30)^2(0.20) \\ &+ (32-30)^2(0.15) + (34-30)^2(0.05) + (36-30)^2(0.10) \\ &+ (38-30)^2(0.05) + (40-30)^2(0.05) \\ &= 26.4. \end{aligned}$$

Thus, we see that for simple random sampling, the estimated population total, x_t', is an unbiased estimate of the population total X_T. The standard error of x_t' given by the equation above is directly proportional to σ_{X_T}, the standard deviation of the distribution of X_T the population total, in the population, and inversely proportional to the square root of the sample size, n.

The standard error also depends on the square root of the factor

$$\frac{(N-n)}{(N-1)},$$

which is known as the finite population correction, and is often denoted *fpc*.

We can obtain some insight into the role played by the *fpc* by examining its value for a hypothetical population containing N = 10,000 elements and for sample sizes as given in Table 3.3. From this table we see that if the sample size, n, is very much less than the population size, N, then the *fpc* is very close to unity and thus will have very little influence on the numerical value of the standard error $\mathrm{SE}(x_t')$ of the estimated total, x_t'. On the other hand, as n gets closer to N, the *fpc* decreases in magnitude and thus will cause a reduction in the value of $\mathrm{SE}(x_t')$

fpc

can be re-written as

$$\begin{aligned}\sqrt{fpc} &= \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{N}{N-1}} \times \sqrt{1-\frac{n}{N}}\end{aligned}$$

so for increasing sample sizes...

Sample Size,n	fpc $= \sqrt{\frac{N-n}{N-1}}$
1	1.0000
10	.9995
100	.9950
500	.9747
1000	.9487
5000	.7071
9000	.3162

Total Estimates

$$
\begin{aligned}
x_t' &= \frac{N \sum_{i=1}^{n} x_{t_i}}{n} \\
Var(x_t') &= N^2 \left(\frac{N-n}{N-1} \right) \left(\frac{\sigma_{X_T}^2}{n} \right) \\
SE(x_t') &= N \sqrt{\frac{N-n}{N-1}} \left(\frac{\sigma_x}{\sqrt{n}} \right)
\end{aligned}
$$

Mean Estimates

$$\begin{aligned} \overline{x} &= \frac{\sum_{i=1}^{n} x_{t_i}}{n} \\ Var(\overline{x}) &= \left(\frac{N-n}{N-1}\right)\left(\frac{\sigma_{x_t}^2}{n}\right) \\ SE(\overline{x}) &= \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma_x}{\sqrt{n}}\right) \end{aligned}$$

Proportion Estimates

$$
\begin{aligned}
p_y &= \frac{\sum_{i=1}^{n} y_i}{n} \\
Var\left(p_y\right) &= \left(\frac{N-n}{N}\right)\frac{P_y\left(1-P_y\right)}{n-1} \\
SE\left(p_y\right) &= \sqrt{\left(\frac{N-n}{N-1}\right)}\sqrt{\frac{P_y\left(1-P_y\right)}{n}}
\end{aligned}
$$

Coefficients of Variation

We define the coefficient of variation $CV\left(\widehat{d}\right)$ of an estimate $\widehat{d}$ of a population parameter d as its standard error SE(d) divided by the true value d of the parameter being estimated.

$$CV\left(\widehat{d}\right) = \frac{Var\left(\widehat{d}\right)}{\widehat{d}}$$

The square of the coefficient of variation $CV^2\left(\widehat{d}\right)$ is a measure of the relative variation of a estimate. That is the variation relative to the estimate.

Define

$$CV(PT) = \frac{\sigma_{PT}}{\overline{X}}$$

then

$$CV(x_t') = \left(\frac{CV(PT)}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

$$CV(\overline{x}) = \left(\frac{CV(PT)}{\sqrt{n}}\right)\sqrt{\frac{N-n}{N-1}}$$

$$CV(p_y) = \left(\frac{1-P_y}{\sqrt{nP_y}}\right)\sqrt{\frac{N-n}{N-1}}$$

Reliability of Estimates

The standard error of an estimate is a measure of the sampling variability of the estimate over all possible samples. Under the assumption that measurement error is nonexistent or negligible, the reliability of an estimate can be judged by the size of the standard error; the larger the standard error, the lower is the reliability of the estimate (see Section 2.4). for reasonably large values of n (say, greater than 20), distributions that are close to the normal or Gaussian distribution, then we can use normal theory to obtain approximate confidence intervals for the unknown population parameters being estimated. For example, approximate 100(1 - α)% confidence intervals for the population total are given by

$$x_t' \quad \pm \quad z_{\alpha/2}(N)\sqrt{\frac{N-n}{N}}\left(\frac{s_{pt}}{\sqrt{n}}\right)$$

$$\overline{x} \quad \pm \quad z_{\alpha/2}\sqrt{\frac{N-n}{N}}\left(\frac{s_{pt}}{\sqrt{n}}\right)$$

Text Example

$$\begin{aligned} \text{sample total} &= 44 \\ s_{pt} &= 3.48 \\ x_t' &= \left(\frac{25}{9}\right)(44) \\ &= 122.22 \end{aligned}$$

so that the confidence interval is given by:

$$\begin{aligned} 122.22 \quad &\pm \quad 1.96(25)\sqrt{\frac{25-9}{25}}\left(\frac{3.48}{\sqrt{9}}\right) \\ &\rightarrow \quad 122.22 \pm 45.47 \\ &\rightarrow \quad (76.75, 167.69) \end{aligned}$$

Note that the true population total, $X_T = 127$, is covered by this confidence interval. These 95% confidence intervals have the following usual interpretation: if we were to repeatedly sample n elements from this population according to the same sampling plan, and if, for each sample, confidence intervals were calculated, 95% of such confidence intervals would include the true unknown population parameter.

If the variable has a nearly symmetric distribution and the sample size is not small, then the confidence coefficients expressed in the confidence intervals will be approximately correct. If the data are badly skewed, however, and the sample size is small, the confidence coefficients may be misleading (Exercise 3.1 illustrates the situation using the data in Table 2.1).

Indicator functions

We have previously discussed the idea of constructing a dichotomous variable that would "indicate" the presence or the absence of a condition or attribute. We can use this same idea a s means to create subdomains of population for which we can do similar sampling techniques. We recall the "indicator" function:

$$Y = \begin{cases} 1 \\ 0 \end{cases}$$

Illustrative text example

Population: 6 families on one city block
Variable: Out of pocket medical expense

Family	Race	Out-of-Pocket Medical Expense (dollars)
1	W	500
2	B	350
3	B	430
4	W	280
5	W	170
6	B	50

Population characteristics for Z/Y

$$\frac{Z}{Y} = \frac{\$830}{3}$$

Family	Race	Out-of-Pocket Medical Expense (dollars X_i)	Y_i	Z_i
1	W	500	0	0
2	B	350	1	350
3	B	430	1	430
4	W	280	0	0
5	W	170	0	0
6	B	50	1	50

$$Y_i = \begin{cases} 1 & \text{African american family} \\ 0 & \text{Caucasian family} \end{cases}$$

Given the sampling distibution

Sample Elements	z	y	z/y^*
1,2,3,4	780	2	390
1,2,3,5	780	2	390
1,2,3,6	830	3	276.67
1,2,4,5	350	1	350
1,2,4,6	400	2	200
1,2,5,6	400	2	200
1,3,4,5	430	1	430
1,3,4,6	480	2	240
1,3.5,6	480	2	240
1.4,5,6	50	1	50
2,3,4,5	780	2	390
2,3,4,6	830	3	276.67
2,3,5,6	830	3	276.67
2,4,5,6	400	2	200
3,4,5,6	480	2	240

Using complementary indicator function

Family	Race	Out-of-Pocket Medical Expense (dollars X_i)	Y_i	Z_i
1	W	500	1	500
2	B	350	0	0
3	B	430	0	0
4	W	280	1	280
5	W	170	1	170
6	B	50	0	0

$$Y_i = \begin{cases} 1 & \text{Caucasian family} \\ 0 & \text{African american family} \end{cases}$$

Complementary sampling distribution

Sample Elements	z	y	z/y*
1,2,3,4	780	2	390
1,2,3,5	670	2	335
1,2,3,6	500	1	500
1,2,4,5	950	3	316.667
1,2,4,6	780	2	390
1,2,5,6	670	2	335
1,3,4,5	950	3	316.667
1,3,4,6	780	2	390
1,3,5,6	670	2	335
1,4,5,6	950	3	316.667
2,3,4,5	450	2	225
2,3,4,6	280	1	280
2,3,5,6	170	1	170
2,4,5,6	450	2	225
3,4,5,6	450	2	225

Similarly

We can show that $E\left(\frac{z}{y}\right)$=316.67 and $SE\left(\frac{z}{y}\right)$=81.165

This approach of subdomains analysis is a special case or ratio estimation that we we discuss in greater detail in chapter 7. In this special case the ratio estimates are unbiased when the denominator of the estimator (Z/Y) is the count of the elementary units considered in the subdomain of interest, and if simple random sampling is used.

A exact expression for the standard error of an estimated mean for the subdomain is unattainable, but for large value of the expected number in the subdomain an approximate standard error is given by:

$$SE\left(\frac{z}{y}\right) = \left[\frac{\sigma_z}{\sqrt{E(y)}}\right] \times \sqrt{\frac{Y - E(y)}{Y - 1}}$$

where

$$\sigma_z = \sqrt{\left[\frac{\sum_{i=1}^{Y} \left(Z_i - \overline{Z}\right)^2}{Y}\right]}$$

Standard Error Estimate

$$\widehat{SE\left(\frac{z}{y_t}\right)} = \left[\frac{\widehat{\sigma_z}}{\sqrt{y_t}}\right] \times \sqrt{\frac{y_t' - y_t}{y_t' - 1}}$$

How large of a sample is needed

Illustrative Text Example-Setup

Recall for the forced vital capacity problem that we had a sample of size forty(40) from a population of size 1200. In this case each of the individuals randomly selected "represented" 30 of the factory workers (including themselves). An interesting question should now be addressed. How was the sample of size 40 selected? It is also important to note that in order to determine a sample size a level of reliability must be established. We already know that the larger the sample the more reliable the estimates are, but the validity of the sample is based solely on the process of obtaining the sample.

How large of a sample is needed?

Illustrative Text Example-Setup

For example, suppose that from a hospital admitting 20,000 patients annually, a survey of hospital patients is to be taken for the purpose of determining the proportion of the 20,000 patients that received optimal care as defined by specified standards. The quality care review committee planning the survey may feel that some remedial action should be taken if fewer than 80% of the patients are receiving optimal care.

- N=20,000.
- Take action if $<$80% receiving optimal care.

- In this instance, the committee would be concerned about overestimates of the true proportion, but would probably not be too concerned if the estimated proportion were 80% when the true proportion were 75%.
- The statistician might formulate this by saying that the user would like to be "virtually certain" that the estimated proportion differs from the true proportion by no more than 100[(80-75)/75]% or 6.67% of the true proportion.
- Notice that bias=(80-75) and we're considering a "relative bias". That is, bias relative to the true proportion.

$$
\begin{aligned}
3 \times SE\left(p_y\right) &= 3 \times \sqrt{\frac{P_y\left(1-P_y\right)}{n}} \sqrt{\frac{N-n}{N-1}} \\
&\rightarrow 3 \times SE\left(p_y\right) \leq 0.0667 P_y \\
&\rightarrow 3 \times \sqrt{\frac{P_y\left(1-P_y\right)}{n}} \sqrt{\frac{N-n}{N-1}} \leq 0.0667 P_y
\end{aligned}
$$

or

$$
n \geq \frac{9 N P_y\left(1-P_y\right)}{(N-1)(0.667)^2 P_y^2+9 P_y\left(1-P_y\right)}
$$

So setting P_y=0.80 and N=20,000

$$\begin{aligned} n &\geq \frac{9(20,000)(0.80)\,(0.20)}{(19,999)\,(0.0667)^2(0.80)^2 + 9(0.80)\,(0.20)} \\ n &\geq 493.295 \text{ or } 494. \end{aligned}$$

See Box 3.5 for the Exact and approximate Sample size Required under simple random sampling

	Exact	Approximate
Total x_t'	$n \geq \frac{z^2 N (CV_x)^2}{z^2 (CV_x)^2 + (N-1)\epsilon^2}$	$n \geq \frac{z^2 (CV_x)^2}{\epsilon^2}$
Mean $\bar{x}$	$n \geq \frac{z^2 N (CV_x)^2}{z^2 (CV_x)^2 + (N-1)\epsilon^2}$	$n \geq \frac{z^2 (CV_x)^2}{\epsilon^2}$

Text Illustrative Example

A sample survey of retail pharmacies is to be conducted in a state that contains 2500 pharmacies. The purpose of the survey is to estimate the average retail price of 20 tablets of a commonly used vasodilator drug. An estimate is needed that is within 10 % of the true value of the average retail price in the state. A list of all pharmacies is available and a simple random sample is to be taken from the list. A phone survey of 20 of N = 1000 pharmacies in another state showed an average price of \$7.00 for 20 tablets with a standard deviation of \$1.40.

$$
\begin{aligned}
CV(\overline{x})^2 &= \frac{[(N-1)/N]s_x^2}{\overline{x}^2} \\
&= \frac{[999/1000](1.4)^2}{(7.00)^2} \\
&= 0.04
\end{aligned}
$$

with a tiny $\epsilon = 0.1$ and N=2500. Using the exact formula from Box 3.5

$$\begin{aligned} n &= \frac{9(2500)(0.04)}{9(0.04) + 2499(0.1)^2} \\ &= 35.6 \\ &\approx 36. \end{aligned}$$