

4 Simple Random Sampling

4.1 Introduction

Let N be the population size.

Let n be the sample size.

Defn. 4.1 If a sample of size n is drawn from a population of size N such that every sample of size n has the same chance of being selected, the sampling procedure is called **simple random sampling**.

4.2 How to Draw a Simple Random Sample

Sampling **with** replacement implies the probability of an item being drawn remains $1/N$ regardless of whether or not the item had been drawn previously.

Sampling **without** replacement implies that once an item is sampled, the item cannot be sampled again.

Which is better, sampling *with* replacement, or sampling *without* replacement?

When are sampling *with* replacement and sampling *without* replacement almost equivalent mathematically?

Example: Consider a small population with $N = 4$, where the population is $\{1, 2, 3, 4\}$. Take a simple random sample without replacement of size 2.

What are the possible outcomes?

How many outcomes are there?

In terms of a general formula involving N and n , how many outcomes are there?

In the above example, what is the probability the sample will be $\{1,4\}$?

In terms of a general formula, what is the probability that a particular set of n outcomes will be sampled from a population of size N , in a simple random sample withOUT replacement?

Solve example 4.1 on p. 80 using R ; i.e., take a simple random sample of size $n = 20$ from $N = 1000$ patient records.

> # First sample withOUT replacement, which is typically done.

> # Next sample WITH replacement.

4.3 Estimation of a Population Mean and Total

Notation:

μ = population mean

τ = population total

How do we typically estimate a population mean, μ ?

What is $E\bar{Y}$?

Suppose we sample WITH replacement (i.e., all observations are independent).

What is $V(\bar{Y})$ (the variance of \bar{Y})?

Suppose we sample withOUT replacement, but $N \gg n$.

What is $V(\bar{Y})$?

Suppose it is NOT the case that $N \gg n$.

Example: Let μ be the average age of all 100 U.S. Senators.

Select sample size $n = 100$ withOUT replacement, and determine the sample mean age \bar{Y} .

What is $V(\bar{Y})$?

When sampling withOUT replacement,

$$V(\bar{Y}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

What is an unbiased estimator of μ ?

What is an unbiased estimator of σ^2 , when sampling WITH replacement?

When sampling WITH replacement, what is an unbiased estimator of $V(\bar{Y})$?

When sampling withOUT replacement, what is an unbiased estimator of $V(\bar{Y})$?

$$\hat{V}(\bar{Y}) = \frac{s^2}{n} \left(\frac{N-n}{N} \right) = \frac{s^2}{n} \left(1 - \frac{n}{N} \right)$$

The quantity $(N - n)/N$ is called the **finite population correction (fpc)**.

What is *fpc* when $N \gg n$?

Textbook (p. 84) suggests that *fpc* can be ignored if $(N - n)/N \geq 0.95$, or if $n \leq 0.05 N$.

Confidence intervals

Recall from Math 220 (or 318): What is the formula for a confidence interval on μ when independent observations are taken from an approximately normal population OR the sample size is large (with finite standard deviation)?

If many 95% confidence intervals on the mean μ are independently constructed, approximately what proportion of these 95% confidence intervals will contain the true population mean μ ?

Example:

- (a) Sample 5 independent observations from a $N(\mu = 70, \sigma = 10)$ population.
- (b) Using 'qt', 'mean' and 'sd', construct the 95% confidence interval on the population mean. Does your confidence interval contain the population mean?
- (c) Repeat part (b) using 't.test'.
- (d) Repeat part (b) using 'ci.t.test'.

- (e) Use ‘replicate’ with ‘ci.t.test’ to generate 100 confidence intervals (of level 95%) on the population mean, where each confidence interval is based on a simple random sample with replacement of size 5.
- (f) Use ‘plot.ci’ to plot these 100 confidence intervals, and draw a line corresponding to the population mean.
- (g) What proportion of your confidence intervals contain the population mean?
- (h) Repeat parts (f) and (g) using 10,000 confidence intervals.

□

Textbook uses 2, as the critical value for the confidence interval, rather than 1.96 or t_{n-1} .

Example: Consider the following population:

```
> pop = sqrt(1:1000)
```

```
> mu = mean(pop)
```

- (a) Sample 90 INDEPENDENT observations from this population.
- (b) Using ‘qt’, ‘mean’ and ‘sd’, construct the 95% confidence interval on the population mean. Does your confidence interval contain the population mean?

$$\bar{y} \pm t_{n-1} \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

(c) Construct the 95% confidence interval on the population mean using 't.test' and 'ci.t.test'.

(d) Repeat this sampling procedure 10,000 times to produce 10,000 95% confidence intervals on the population mean.

(e) What proportion of these confidence intervals really do contain the population mean?

□

Example: Consider the following population:

```
> pop = sqrt(1:1000)
```

```
> mu = mean(pop)
```

(a) Sample 90 observations withOUT replacement from this population.

(b) Using 'qt', 'mean' and 'sd', construct the 95% confidence interval on the population mean. Does your confidence interval contain the population mean?

$$\bar{y} \pm t_{n-1} \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

(c) Construct the 95% confidence interval on the population mean using 'ci.t.test'.

(d) Repeat this sampling procedure 10,000 times to produce 10,000 95% confidence intervals on the population mean.

(e) What proportion of these confidence intervals really do contain the population mean?

□

Bound on error of estimation

\bar{Y} estimates μ .

When sampling withOUT replacement, a 95% confidence interval on μ is approximately

$$\bar{y} \pm 2 s \sqrt{(1 - n/N)/n}$$

OR

$$\bar{y} \pm 2 \sqrt{\frac{s^2}{n} \left(\frac{N - n}{N} \right)}$$

The **bound on the error of estimation** when estimating μ is

$$B = 2 \sqrt{\hat{V}(\bar{Y})} = 2 \sqrt{\frac{s^2}{n} \left(\frac{N - n}{N} \right)}$$

Example: Take just **ONE** sample of 90 observations withOUT replacement from the population:

```
> pop = sqrt(1:1000)
```

```
> # What is the likelihood that the confidence interval will contain the true value  $\mu$ ?
```

```
> y = sample(pop, n)
```

```
> mean(y)
```

> # Estimate the bound on the error of estimation.

> N = length(pop)

A 95% confidence interval on μ is approximately:

$\text{mean}(y) \pm \text{the bound}$

Is the true value of μ in the above confidence interval?

□

More on sampling distributions

Example: SIMILAR TO EXERCISES #4.1 and #4.2, p. 103.

Consider the population $\{0, 1, 2, 3\}$.

(a) Determine the mean and variance of this population.

> pop = 0:3

> N = length(pop) # N is the population size.

(b) List all possible simple random samples (withOUT replacement) of size $n = 2$ that can be selected from the population $\{0, 1, 2, 3\}$, and list the sample mean \bar{Y} and sample variance s^2 .

(c) Determine the probability distribution of the sample mean \bar{Y} .

Is \bar{Y} approximately normal?

(d) Construct a *line graph* of the distribution of \bar{Y} .

```
> y.bar.pop = c(0.5, 1, 1.5, 1.5, 2, 2.5)
```

```
> y.bar = c(0.5, 1, 1.5, 2, 2.5)
```

```
> y.bar.probs = c(1/6, 1/6, 2/6, 1/6, 1/6)
```

(e) Using the probability distribution of \bar{Y} , compute $E\bar{Y}$.

$$E\bar{Y} = \sum_{\bar{y}} \bar{y} p(\bar{y})$$

(f) Using the probability distribution of \bar{Y} , compute $V(\bar{Y})$.

Is the answer $\sigma_y^2/n = 1.25/2 = 0.625$?

$$V(\bar{Y}) = E\bar{Y}^2 - (E\bar{Y})^2 = E\bar{Y}^2 - \mu_{\bar{Y}}^2$$

$$E\bar{Y}^2 = \sum_{\bar{y}} \bar{y}^2 p(\bar{y})$$

The formula for population variance of \bar{Y} , when sampling withOUT replacement, is (from p. 83):

$$\text{Var}(\bar{Y}) = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

(g) Using 10,000 simulations, simulate the population histogram of \bar{Y} .

```
> y.bar.sim = sample(y.bar, 1e4, replace=T, prob=y.bar.probs)
> y.bar.sim = sample(y.bar.pop, 1e4, replace=T)
> y.bar.sim = replicate( 1e4, mean( sample( pop, 2 ) ) ) # Replace is FALSE.
```

(h) Determine the *mean* of these 10,000 simulated values of \bar{Y} .

(i) Determine the *variance* of these 10,000 simulated values of \bar{Y} .

(j) Determine the probability distribution of the sample variance s^2 .

(k) Construct the *line graph* of the distribution of s^2 .

```
> s2.pop = c(0.5, 0.5, 0.5, 2, 2, 4.5)
> s2 = c(0.5, 2, 4.5)
```

```
> s2.probs = c(3/6, 2/6, 1/6)
```

(l) Compute Es^2 .

$$Es^2 = \sum_{s^2} s^2 p(s^2)$$

(m) Using 10,000 simulations, simulate the population histogram of s^2 .

```
> s2.sim = sample(s2, 1e4, replace=T, prob=s2.probs)
```

```
> s2.sim = sample(s2.pop, 1e4, replace=T)
```

```
> s2.sim = replicate( 1e4, var( sample( pop, 2 ) ) ) # Replace is FALSE.
```

(n) Determine the *mean* of these 10,000 simulated values of s^2 .

(o) Noting exercise 4.2, p. 103, and the biasedness of s^2 when sampling without replacement, p. 83, show (for this example) that

$$Es^2 = \frac{N}{N-1} \sigma^2.$$

□

More on population variance

Suppose that a and b are constants, and X is random with mean μ_x and variance σ_x^2 .

Let $Y = a + bX$

What is EY ?

What is σ_y^2 (the variance of Y)?

What is σ_y (the standard deviation of Y)?

Example:

> x = c(30, 40, 50, 60, 70)

> y = 300 - 2*x

> mean(x)

> mean(y)

For population size N , what is the population total τ in terms of the population mean μ ?

How should we estimate τ (when sampling WITH or withOUT replacement)?

When sampling withOUT replacement, what is an unbiased estimator of $V(\hat{\tau})$?

$$\begin{aligned}\hat{V}(\hat{\tau}) &= \hat{V}(N\bar{Y}) = N^2 \hat{V}(\bar{Y}) = N^2 \frac{s^2}{n} \left(\frac{N-n}{N} \right) \\ &= N^2 \frac{s^2}{n} \left(1 - \frac{n}{N} \right)\end{aligned}$$

The **bound on the error of estimation** when estimating τ is

$$\begin{aligned}B &= 2 \sqrt{\hat{V}(\hat{\tau})} = 2 \sqrt{\hat{V}(N\bar{Y})} = 2 \sqrt{N^2 V(\bar{Y})} \\ &= 2N \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}\end{aligned}$$

□

Example: Take just **ONE** sample of 90 observations with**OUT** replacement from the population:

```
> pop = sqrt(1:1000)
```

```
> # Compute the bound on the error of estimation.
```

A 95% confidence interval on τ is approximately:

$N * \text{mean}(y) \pm \text{the bound.}$

□

Homework C4.3.1:

Revisit unfair die problem (i.e., sample with replacement). Probabilities are

$p(1) = 0.3$, $p(2) = 0.4$, $p(3) = 0.2$, $p(4) = 0.1$.

(a) Sample 19 observations, and print them.

(b) Using ‘qt’, ‘mean’ and ‘sd’, construct the 95% confidence interval on the population mean. Does your confidence interval contain the population mean?

(c) Repeat part (b) using ‘t.test’.

(d) Repeat part (b) using ‘ci.t.test’.

(e) Use ‘replicate’ with ‘ci.t.test’ to generate 100 confidence intervals (of level 95%) on the population mean, where each confidence interval is based on a simple random sample with replacement of size 19.

(f) Use ‘plot.ci’ to plot these 100 confidence intervals, and draw a line corresponding to the population mean.

(g) What proportion of your confidence intervals contain the population mean?

(h) Repeat parts (f) and (g) using 10,000 confidence intervals.

End of Homework C4.3.1 □

4.4 Selecting the Sample Size for Estimating Population Means and Totals

Decide in advance how large of a sample should be taken.

What is the drawback of taking too *large* of a sample?

What is the drawback of taking too *small* of a sample?

For a simple random sample withOUT replacement and large n , recall that a 95% confidence interval on μ is (approximately)

$$\bar{Y} \pm 2 \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

OR $\bar{Y} \pm B$

$$\text{Hence, } B = 2 \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

Solve for n .

SAMPLE SIZE DETERMINATION regarding μ :

The sample size n required to estimate μ with a bound on the error of estimation B , when performing simple random sampling withOUT replacement, is:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2},$$

where $D = B^2/4$.

Solving for B , we obtain $B = 2\sqrt{(D)}$, so D is the pre-determined variance of \bar{Y} .

Example: Consider the (larger) population:

```
> pop = sqrt(1:1e5)
```

```
> N = length(pop)
```

In a preliminary sample, take a simple random sample withOUT replacement of size 100, to estimate σ^2 .

```
> n0 = 100 ; y0 = sample( pop, n0 ) ; var.hat = var(y0)
```

What sample size n_1 is needed to estimate μ with a bound on the error of estimation 3, in a simple random sample withOUT replacement?

$$n_1 = \frac{N\sigma^2}{(N-1)D + \sigma^2},$$

where $D = B^2/4$.

Take a simple random sample withOUT replacement of size n_1 , and construct a 95% confidence interval on μ .

What is the likelihood that the confidence interval will contain the true value μ ?

```
> fpc = 1-n1/N
```

```
> # Determine the confidence interval.
```

Is μ in the confidence interval?

What is the **error of estimation**?

Is this number smaller than $B = 3$?

□

SAMPLE SIZE DETERMINATION regarding τ :

The sample size n required to estimate τ with a bound on the error of estimation B , when performing simple random sampling withOUT replacement, is:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2},$$

where $D = B^2/(4N^2)$.

Solving for B , we obtain $B = 2N\sqrt{(D)}$, so D is again the pre-determined variance of \bar{Y} .

Example: Solve exercise 4.41 on p. 111. An auditor randomly samples 20 accounts receivable from the 500 accounts of a certain firm. The auditor lists the amount of each account and checks to see whether the underlying documents are in compliance with stated procedures.

(a) Read in the data.

(b) Estimate the *total* of the accounts receivable for the 500 accounts of the firm.

(c) Place a *bound* on the *error of estimation*, when using $\hat{\tau}$ to estimate τ .

The **bound on the error of estimation** when estimating τ is

$$B = 2 \sqrt{\hat{V}(\hat{\tau})} = 2 \sqrt{\hat{V}(N\bar{Y})} = 2 \sqrt{N^2 V(\bar{Y})} = 2N \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

(d) Construct a 95% confidence interval on τ , the population total.

Interpretation: We are 95% confident that the population total of the accounts lies between \$78,644.17 and \$118,455.85.

(e) What sample size n_1 is needed to estimate τ with a bound on the error of estimation \$8,000, in a simple random sample withOUT replacement?

Hence, we want $|\hat{\tau} - \tau| < \$8000$ with 95% confidence.

$$n_1 = \frac{N\sigma^2}{(N-1)D + \sigma^2},$$

where $D = B^2/(4 N^2)$.

> B = 8000

(f) Estimate the *mean* of the accounts receivable for the 500 accounts of the firm.

(g) Place a *bound* on the *error of estimation*, when using $\hat{\mu}$ to estimate μ .

The **bound on the error of estimation** when estimating μ is

$$B = 2 \sqrt{\hat{V}(\hat{\mu})} = 2 \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

(h) Construct a 95% confidence interval on μ , the population mean.

Interpretation: We are 95% confident that the population mean accounts lies between \$157.29 and \$236.91.

(i) Do you think that the *average* account receivable for the firm exceeds \$250?

(j) What sample size n_2 is needed to estimate μ with a bound on the error of estimation \$16, in a simple random sample withOUT replacement?

$$n_2 = \frac{N\sigma^2}{(N-1)D + \sigma^2}, \text{ where } D = B^2/4$$

□

4.5 Estimation of a Population Proportion

Example: Let p be the (unknown) population proportion of American adults who are Republicans.

Take a simple random sample of **100** adults in the United States, and inquire about political affiliation.

Suppose that 42 of them say they are Republicans.

What is your best estimate of p ?

From Math 220 or 318, what is an estimated standard deviation of \hat{p} ?

□

PROPORTION IS A SPECIAL CASE OF A MEAN

Example: Again, let p be the (unknown) population proportion of American adults who are Republicans.

Take a simple random sample of only **10** adults in the United States, and inquire about political affiliation.

Suppose the responses are the following:

{Dem, Rep, Dem, Dem, Rep, Ind, Dem, Ind, Rep, Rep}.

What is the sample proportion (\hat{p}) of Republicans?

Convert the above responses to 1 if Republican, and 0 if not Republican.

Hence, the data become:

{0, 1, 0, 0, 1, 0, 0, 0, 1, 1}

Compute the mean, \bar{Y} .

□

For large sample sizes, \hat{p} is approximately normally distributed, by the Central Limit Theorem.

A commonly used rule for “large sample sizes” is $np \geq 10$ and $n(1 - p) \geq 10$.

Example: Graph 10,000 values of \hat{p} , where \hat{p} is the sample proportion based on $n = 10$ and $p = 0.7$. Repeat for $n = 20$ and $n = 30$.

```
> linegraph( replicate( 1e4, mean( sample( 0:1, 10, replace=T, prob=c(0.3, 0.7) ) ) ), F
)
```

□

Estimated variance of \hat{p} :

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

Bound on the error of estimation:

$$2 \sqrt{\hat{V}(\hat{p})} = 2 \sqrt{\frac{\hat{p}\hat{q}}{n-1} \left(\frac{N-n}{N} \right)}$$

Exercise 4.42, p. 112: An auditor randomly samples 20 accounts receivable from the 500 accounts of a certain firm. The auditor lists the amount of each account and checks to see whether the underlying documents are in compliance with stated procedures.

(a) Estimate p , the proportion of the firm's accounts that fail to comply with stated procedures.

```
> y.mat = scan2( "EXER4.41.DAT", T )
```

```
> compliance = y.mat[ , 3]
```

```
> p.hat = mean(compliance=="N")
```

(b) Place a bound on the error of estimation.

(c) Construct a 95% confidence interval on p , the population proportion of the firm's accounts that fail to comply with stated procedures.

Interpretation: We are 95% confident that the population proportion of the firm's accounts that fail to comply with stated procedures lies between 9.4% and 50.6%.

(d) Do you think the proportion of accounts that comply with stated procedures exceeds 80%? Why?

□

SAMPLE SIZE DETERMINATION regarding p :

The sample size n required to estimate p with a bound on the error of estimation B , when performing simple random sampling withOUT replacement, is:

$$n = \frac{Npq}{(N-1)D + pq},$$

where $D = B^2/4$ and $q = 1 - p$.

Note: Since the variance of a Bernoulli random variable (i.e., consisting of only zeros and ones) is pq , we simply replace σ^2 by pq in the formula for sample size determination regarding μ to obtain the above formula.

How do we select the sample size if we do not have an initial estimate of p ?

Example: *Continue with exercise 4.42 on p. 112.* An auditor randomly samples 20 accounts receivable from the 500 accounts of a certain firm. The auditor lists the amount of each account and checks to see whether the underlying documents are in compliance with stated procedures.

(e) What sample size is needed to estimate the population proportion of the firm's accounts that fail to comply with stated procedures, with a bound on the error of estimation $B = 7\%$?

□

Example: Often, large national polling organizations quote a **margin of error** of 3 percentage points.

The **margin of error** is another name for the **bound on the error of estimation**.

What sample size is needed to estimate the population proportion of U. S. adults who call themselves Democrats with a **margin of error** of 3%?

$$n = \frac{Npq}{(N-1)D + pq},$$

where $D = B^2/4$ and $q = 1 - p$.

What number should we use for p in the above formula?

How large is N ?

□

Homework C4.5.1: For each of the following values of n and p , construct a line graph of 10,000 values of \hat{p} , and explain whether or not the distribution of \hat{p} is approximately normal.

- (a) $n = 50$ and $p = 0.99$
- (b) $n = 50$ and $p = 0.7$
- (c) $n = 50$ and $p = 0.5$
- (d) $n = 100$ and $p = 0.02$
- (e) $n = 20$ and $p = 0.55$
- (f) $n = 30$ and $p = 0.4$

End of Homework C4.5.1 □

4.6 Comparing Estimates

ESTIMATE THE DIFFERENCE OF PROPORTIONS, $p_1 - p_2$

Example: Suppose p_1 is the population proportion of Republicans, and suppose p_2 is the population proportion of Democrats.

Suppose we want to:

- ⊙ estimate $p_1 - p_2$,
- ⊙ place a bound on the error of estimation of $p_1 - p_2$, and
- ⊙ construct a 95% confidence interval on $p_1 - p_2$.

How should we estimate $p_1 - p_2$?

If \hat{p}_1 and \hat{p}_2 are dependent, then

$$\begin{aligned} V(\hat{p}_1 - \hat{p}_2) &= \text{var}(\hat{p}_1) + \text{var}(\hat{p}_2) - 2 \text{cov}(\hat{p}_1, \hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n} + 2 \frac{p_1 p_2}{n} \end{aligned}$$

Therefore,

$$\hat{V}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{n} + 2 \frac{\hat{p}_1 \hat{p}_2}{n}$$

Bound on the error of estimation is $B = 2 \sqrt{\hat{V}(\hat{p}_1 - \hat{p}_2)}$, and a 95% confidence interval on $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2 \pm B$. \square

Example: Now, suppose p_1 is the population proportion of *males* who are Republicans, and let p_2 is the population proportion of *females* who are Republicans.

Let n_1 be the sample size of *males*, and let n_2 be the sample size of *females*.

Let \hat{p}_1 and \hat{p}_2 be the sample proportions.

Are \hat{p}_1 and \hat{p}_2 independent?

We still use $\hat{p}_1 - \hat{p}_2$ to estimate $p_1 - p_2$.

The new estimate of the variance becomes

$$\hat{V}(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

\square

See example 4.11, pp. 99-100, for both independent and dependent examples involving estimating $p_1 - p_2$.

ESTIMATE THE DIFFERENCE OF MEANS, $\mu_1 - \mu_2$, for independent samples

Example: Suppose μ_1 is the mean income among Virginians, and μ_2 is the mean income among North Carolinians.

Assume that the sample mean incomes of the two states are independent.

How should we estimate $\mu_1 - \mu_2$?

How should we estimate the VARIANCE of $\bar{Y}_1 - \bar{Y}_2$?

Bound on the error of estimation is $B = 2 \sqrt{\hat{V}(\bar{Y}_1 - \bar{Y}_2)}$, and a 95% confidence interval on $\mu_1 - \mu_2$ is $\bar{Y}_1 - \bar{Y}_2 \pm B$.

See example 4.10, pp. 96-99, for example involving estimating $\mu_1 - \mu_2$.