

# Probability Sampling and Confidence Intervals

UBCO MDS — DATA 543





Office hours on **Thursday** this week!

Assignment #1 posted.

Quiz #1 + #2 will be closed book.

→ Formula sheet.

Key concepts for today:

- Estimate vs. Estimator
- Bias, precision, accuracy

	HIGH ACCURACY	MEDIUM ACCURACY	LOW ACCURACY
HIGH PRECISION	BARACK OBAMA WAS PRESIDENT FOR 70,128 HOURS	BARACK OBAMA WEIGHS AS MUCH AS 17.082 CATS	BARACK OBAMA IS 70.128 FEET TALL
MEDIUM PRECISION	MOST CATS HAVE 4 LEGS	BARACK OBAMA IS 6'1"	BARACK OBAMA HAS 4 LEGS
LOW PRECISION	MOST CATS HAVE LEGS	BARACK OBAMA HAS FEWER LEGS THAN YOUR CAT	BARACK OBAMA'S CAT HAS HUNDREDS OF LEGS

- Estimate of  $\text{Var}(\tilde{\mu})$ ,  $t_{pc}$
- Critical values for confidence intervals.
- **Confidence intervals for  $\mu$ ,  $\sigma$ ,  $\pi$**
- **Sample size calculations**

# Introduction

- Last class we introduced some non-probability and probability methods for selecting a sample.
- While probability samples are harder to take than a convenience sample, say, they provide information that can be used to assess the precision of statistics calculated from the sample.
- Today we focus on parameter estimates for some of these probability sampling techniques.
- More specifically, we will be investigating the following for different sampling estimators:
  - Expectation and variance of the estimator
  - Approximate distribution of the estimator
  - Confidence Intervals

# Recall Notation:

- Let  $U = \{1, 2, \dots, N\}$  denote a finite population
- $s$ : our sample is a subset of the population ( $s \subset U$ ).
- $n$ : the sample size.
- $y_i$ : the value of the response variate for unit  $i$ .

A *sampling protocol* or *sampling design* is the mechanism by which we choose our samples.

- The design is determined by assigning the probability  $P(s)$  to each possible sample  $s$ .
- Let  $\mathcal{D}$  be the set of samples  $s$  with  $P(s) > 0$
- $\sum_{s \in \mathcal{D}} P(s) = 1$

The *inclusion probability*,  $\pi_i = P(i \in s)$ , for unit  $i$  is the probability that unit  $i$  is in the selected sample  $s$ .

# Parameters of Interest

- After a sample  $s = (j_1, j_2, \dots, j_n)$  is chosen, the values  $y_i$  for  $i = 1, \dots, n$  is collected, where  $y$  is the variable of interest.
- Assuming a successful gathering, then the survey data is at hand is  $d = ((j_1, y_1), \dots, (j_n, y_n))$  aka  $d = \{s, y_i \mid i \in s\}$ .
- An objective of sample surveys is to estimate some population quantity or *population parameter*. Some common objectives:

Population average  $\mu = \frac{1}{N} \sum_{i=1}^N y_i$

Population total  $\tau = \sum_{i=1}^N y_i$

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

# Some Remarks

- $\mu$  is a population parameter, i.e. it is an unknown<sup>1</sup> constant.
- The sample mean:  $\hat{\mu} = \frac{1}{n} \sum_{i \in s} y_i$  serves as an **estimate** for the population parameter  $\mu$ , where  $s$  is a random sample with  $P(S = s)$  defined by our sampling protocol.
- Subtly different to an estimate is an **estimator**:  $\tilde{\mu} = \frac{1}{n} \sum_{i \in S} y_i$ .
- Note the difference: estimates are *fixed values*, estimators are *random variables* which depends on the sample. Defined this way, we can compute the expected value of an estimator.

---

<sup>1</sup>in a practical application setting

# Example: Grocery store

Consider four grocery stores:  $A$ ,  $B$ ,  $C$ , and  $D$  for which we have the record of the number of sales for a particular item.

Grocery Store	$A$	$B$	$C$	$D$
Sale	4	1	3	2

Consider this our population so that  $U = \{A, B, C, D\}$ ,  $N = 4$  and we want to take a  $n = 2$  sample to estimate the population mean sale value. All the possible samples and estimates are summarized below:

*order doesn't matter*

Sample	First element	Second element	Sample mean
$s_1$	$A$	$B$	2.5 → estimate
$s_2$	$A$	$C$	3.5
$s_3$	$A$	$D$	3.0
$s_4$	$B$	$C$	2.0
$s_5$	$B$	$D$	1.5
$s_6$	$C$	$D$	2.5

*All possible samples  $n=2$*

*estimator*

# Example: Grocery store

- Recall that in probability sampling, each possible sample  $S$  from the population has a known probability  $P(S)$  of being chosen, and the probabilities of the possible samples sum to 1. For instance,

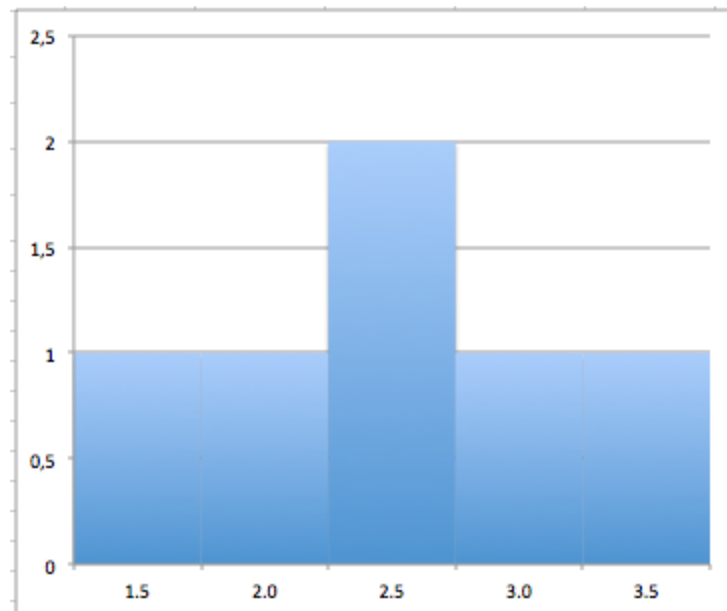
	$\mathcal{D} = \text{All possible samples } S$					
$S$	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$
$P(S)$	1/6	1/6	1/6	1/6	1/6	1/6

- One way to select the sample would be to place six labeled balls in a box with the labels 1 to 6. Now choose one ball at random; if a ball labeled 6 is chosen, then  $s_6$  is the sample.
- The above design is an SRSWOR. Each unit is in exactly 3 of the possible samples, so  $\pi_i = 3/6 = 1/2$  for  $i = A, B, C, D$ .



# Example: Grocery store

The *sampling distribution* defines the distribution of different values of the statistic obtained by the process of taking all possible samples from the population. For our example, the sampling distribution can be expressed either in table or graph:



$\hat{\mu}$	1.5	2.0	2.5	3.0	3.5
$P(\tilde{\mu} = \hat{\mu})$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Note that the true population parameter  $\mu = \frac{4+1+3+2}{4} = 2.5$

# Probability Sampling

- In reality, our sample will (almost certainly) differ from the target population, so we will have uncertainty about the population parameter we're interested in.
- If we have a statistical model for how we've sampled units, we can estimate this uncertainty in the form of confidence intervals and hypothesis tests.
- Most results in sampling rely on the sampling distribution of a statistic, where a *statistic*  $t$  is some function of the sample data where  $\underline{y} = (y_1, \dots, y_n)$ :

$$t(d) = t(s, \underline{y})$$

- A sampling distribution is an example of a discrete probability distribution.

# Example: Grocery store

The *expected value* of the statistic is the weighted average of the possible sample values of the statistic, weighted by the probability that particular value of the statistic would occur.

$$\begin{aligned} \hat{\mu} &= \text{estimate} \\ \tilde{\mu} &= \text{estimator} \\ E[\tilde{\mu}] &= \sum_{s \in \mathcal{D}} \bar{y}(s) P(S = s) \quad \text{where } \bar{y}(s) = \sum_{i \in s} y_i \\ &= \sum \hat{\mu} P(\tilde{\mu} = \hat{\mu}) \\ &= \frac{1}{6}1.5 + \frac{1}{6}2 + \frac{2}{6}2.5 + \frac{1}{6}3 + \frac{1}{6}3.5 = 2.5 \end{aligned}$$

The *estimation bias* of the estimator  $\tilde{\mu}$  is  $\text{Bias}[\tilde{\mu}] = E[\tilde{\mu}] - \mu$

$$\begin{aligned} \text{Bias}[\tilde{\mu}] &= E[\tilde{\mu}] - \mu \\ &= 2.5 - 2.5 = 0 \end{aligned}$$

For this example, our estimator  $\tilde{\mu}$  is said to be *unbiased*.

# Properties of Estimators

- Is the property of being unbiased enough to say that it is a “good” estimator?
- For instance, we might look at the variance of the sampling distribution of  $\tilde{\mu}$

$$\begin{aligned} V[\tilde{\mu}] &= E[(\tilde{\mu} - E[\tilde{\mu}])^2] = \sum_{s \in \mathcal{D}} P(S = s) [\bar{y}(s) - E[\tilde{\mu}]]^2 \\ &= \frac{1}{6}(1.5 - 2.5)^2 + \frac{1}{6}(2 - 2.5)^2 + \frac{2}{6}(2.5 - 2.5)^2 + \\ &\quad \frac{1}{6}(3 - 2.5)^2 + \frac{1}{6}(3.5 - 2.5)^2 = 0.417 \end{aligned}$$

Because we sometimes use biased estimators, we often use the **mean squared error (MSE)** rather than variance to measure the accuracy of an estimator. **Accuracy** is how close the estimate is to the true value.

$$\begin{aligned}
 \text{MSE}[\tilde{\mu}] &= E[(\tilde{\mu} - \mu)^2] \\
 &= E[(\tilde{\mu} - E[\tilde{\mu}] + E[\tilde{\mu}] - \mu)^2] \quad (\text{add } 0) \\
 &= E[(\tilde{\mu} - E[\tilde{\mu}])^2] + (E[\tilde{\mu}] - \mu)^2 + 2E[(\tilde{\mu} - E[\tilde{\mu}])(E[\tilde{\mu}] - \mu)] \\
 &= V[\tilde{\mu}] + \text{Bias}[\tilde{\mu}]^2
 \end{aligned}$$

*Note:  $\theta$  is a general parameter population*

- An estimator  $\tilde{\theta}$  of  $\theta$  is **unbiased** if  $E(\tilde{\theta}) = \theta$
- **precise** if  $V(\tilde{\theta}) = E[(\tilde{\theta} - E[\tilde{\theta}])^2]$  is small, and
- **accurate** if  $\text{MSE}[\tilde{\theta}] = E[(\tilde{\theta} - \theta)^2]$  is small.



### FIGURE 2.3

Unbiased, precise, and accurate archers. Archer A is unbiased—the average position of all arrows is at the bull's-eye. Archer B is precise but not unbiased—all arrows are close together but systematically away from the bull's-eye. Archer C is accurate—all arrows are close together and near the center of the target.

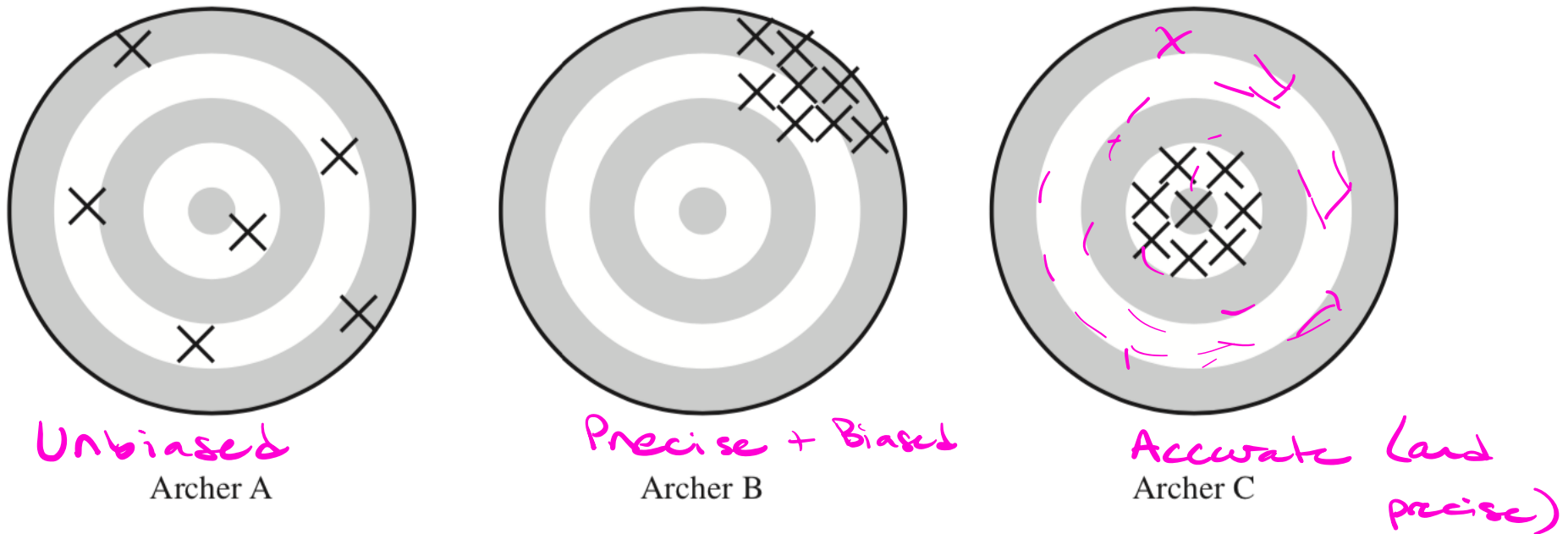


Figure: Source of image: Lohr (2009)

A badly biased estimator may be precise but it will not be accurate; while precision measures how close estimates from different samples are to each other.

- A single probability sample is not guaranteed to be representative of the population with regard to the characteristics of interest, we can quantify how likely it is that our sample is a “good” one.
- The notion is the same as that of confidence intervals: We do not know whether the particular 95% confidence interval we construct for the mean contains the true value of the mean.
- We do know, however, that if the assumptions for the confidence interval procedure are valid and if we repeat the procedure over and over again, we can expect 95% of the resulting confidence intervals to contain the true value of the mean.

# Probability Sampling

Today we consider protocols where the sample size  $n$  is fixed so that the only subsets with  $n$  units have positive probability.

## Major advantage

If we understand the probabilistic mechanism we've used to form our sample, we can assess the sample error mathematically.

## SRSWOR vs. SRSWR

SRSWOR is more efficient than SRSWR. When  $N$  is very large and  $n$  is small, SRSWOR and SRSWR will be very close to each other.

# SRSWOR

- With a finite population  $U = \{1, 2, \dots, N\}$  and fixed sample size fixed at  $n$ , there are  $\binom{N}{n}$  possible samples, each equally likely.
- Sampling protocol:  $P(S = s) = 1/\binom{N}{n} = \frac{n!(N-n)!}{N!}$  if  $s$  has  $n$  distinct elements  $n$ , otherwise  $P(S = s) = 0$ .
- The probability that the  $i$ th unit appears in the sample is  $\pi_i = n/N$  (Lohr, 2009, Section 2.8)

Under SRSWOR

$$E(\tilde{\mu}) = \mu, \quad \text{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

- Hence  $\tilde{\mu}$  is an unbiased estimator for the  $\mu$  in SRSWOR.

# Some Definitions

A couple of related definitions:

- $f = \frac{n}{N}$  is the *sampling fraction*.
- $1 - f = 1 - \frac{n}{N}$  is the *finite population correction factor* (fpc).

Therefore, we can write  $Var(\tilde{\mu}) = (1 - f) \frac{\sigma^2}{n}$

- The larger the sampling fraction  $n/N$ , the more information we have about the population and thus the smaller the variance.
- For a census, the fpc, and hence  $Var(\tilde{\mu})$ , is 0.
- When the sampling fraction  $n/N$  is small, the variance of the estimator is dominated by  $\sigma^2/n$
- In many cases, samples are taken from extremely large populations, so the fpc is approximately 1.



# Variance estimator

Problem: The population variance  $\sigma^2$  is in general unknown.

Solution: We therefore estimate it by the sample variance:

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \tilde{\mu})^2$$

We can show<sup>2</sup> that  $\tilde{\sigma}^2$  is an unbiased estimator for  $\sigma^2$ , i.e.  $E[\tilde{\sigma}^2] = \sigma^2$ .

An unbiased estimator for  $V(\tilde{\mu})$  is

$$\widetilde{V(\tilde{\mu})} = \left(1 - \frac{n}{N}\right) \frac{\tilde{\sigma}^2}{n}$$

---

<sup>2</sup>(Lohr, 2009, Section 2.8)

# Variance Estimate

An estimate of  $Var(\tilde{\mu})$  is then

$$\widehat{Var}(\tilde{\mu}) = \left(1 - \frac{n}{N}\right) \frac{\hat{\sigma}^2}{n}$$

For a specific sample and estimate  $\hat{\mu}$  we define

- the *standard error* is the square root of the estimated variance of a statistic
- In this case, is calculated by

$$\begin{aligned} \text{s.e.}(\hat{\mu}) &= SE(\hat{\mu}) = \sqrt{\widehat{Var}(\tilde{\mu})} \\ &= \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{\sigma}^2}{n}} = \sqrt{(1 - f) \frac{\hat{\sigma}^2}{n}} = \hat{\sigma} \sqrt{\frac{(1 - f)}{n}} \end{aligned}$$

# Confidence Intervals

We've shown:

$$E[\tilde{\mu}] = \mu \quad \text{Var}(\tilde{\mu}) = (1 - f) \frac{\sigma^2}{n}$$

The central limit theorem tells us if  $N$ ,  $n$  and  $N - n$  are 'large', then

$$\frac{\tilde{\mu} - \mu}{\sqrt{(1 - f) \frac{\tilde{\sigma}^2}{n}}} \sim N(0, 1)$$

A large sample  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is therefore

$$\hat{\mu} \pm c \times \text{s.e.}(\hat{\mu})$$

where  $c$  is chosen such that  $P(|z| \leq c) = 1 - \alpha$  or  $P(|z| > c) = \alpha$  for a standard normal distribution  $Z \sim N(0, 1)$ .

# Confidence Intervals

*critical value (determines the confidence level)*

Reminder:  $c$  values for commonly used confidence levels:

- 99%: 2.576 because  $P(z \leq 2.576) = P(z \geq -2.576) = 0.995$
- 95%: 1.960 because  $P(z \leq 1.960) = P(z \geq -1.960) = 0.975$
- 90%: 1.645 because  $P(z \leq 1.645) = P(z \geq -1.645) = 0.950$

Recall that  $c$  can be determined in R using:

```
> # for a 99% CI:
> c(qnorm((1-0.99)/2), qnorm((1-0.99)/2, lower.tail = FALSE))
[1] -2.575829  2.575829
>
> # for a 95% CI:
> c(qnorm((1-0.95)/2), qnorm((1-0.95)/2, lower.tail = FALSE))
[1] -1.959964  1.959964
>
> # for a 90% CI:
> c(qnorm((1-0.90)/2), qnorm((1-0.90)/2, lower.tail = FALSE))
[1] -1.644854  1.644854
```

# Comment

- In practice, we often find  $c$  using the  $t$ -distribution instead of the standard normal.
- Hence  $c = t_{\alpha/2, n-1}$ , the  $(1 - \alpha/2)^{th}$  percentile of a  $t$  distribution with  $n - 1$  degrees of freedom instead of  $z_{\alpha/2}$ .
- Note that for large samples,  $t_{\alpha/2, n-1} \approx z_{\alpha/2}$ , however, in smaller samples, using  $t_{\alpha/2, n-1}$  instead of  $z_{\alpha/2}$  produces *wider* CIs.

A small sample  $100(1 - \alpha)\%$  confidence interval for the population mean  $\mu$  is therefore

$$\hat{\mu} \pm t_{\alpha/2, n-1} \times \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{\sigma}^2}{n}}$$



# Confidence Interval: Example

## Example 1

I have the final grades of 116 students from one of my previous courses. I want to estimate the average grade received in the class, but can only be bothered to retrieve 10 exam papers from the department office. Weirdly, I *can* be bothered to select the exam papers via simple random sampling without replacement, and find the following scores:

87 78 81 90 61 63 63 73 72 84

# Confidence Interval: Example

What is a 95% confidence interval for the average grade for the whole class? Use:  $\hat{\mu} = 75.2$ ,  $\hat{\sigma}^2 = 110.2$

We need:  $\hat{\mu} \pm c \times \sqrt{(1 - \frac{n}{N}) \frac{\hat{\sigma}^2}{n}}$  We have:

- $n = 10$ ,  $N = 116$  ;  $\hat{\mu} = 75.2$ ,  $\hat{\sigma}^2 = 110.2$   $c = 1.96$

Plugging in, we get:

$$\begin{aligned}\hat{\mu} \pm c \times \sqrt{(1 - \frac{n}{N}) \frac{\hat{\sigma}^2}{n}} &= 75.2 \pm 1.96 \times \sqrt{(1 - \frac{10}{116}) \frac{110.2}{10}} \\ &= 75.2 \pm 1.96 \times 3.173 \\ &= 75.2 \pm \underline{6.220}\end{aligned}$$

Margin of Error

Giving a 95% CI of:  $(\underline{69.0}, \underline{81.4})$   
lower higher

# Confidence Interval: Example


A word of warning about **rounding errors**:

- The exact sample variance was 110.1778. While we used a rounded version of a statistic (110.2 in this case) for ease of viewing we should be careful about rounding errors creeping into your calculations elsewhere in your work!

For instance, as will be seen in the associated R code, the CI we get if we are careful about error is: [68.98102, 81.41898]

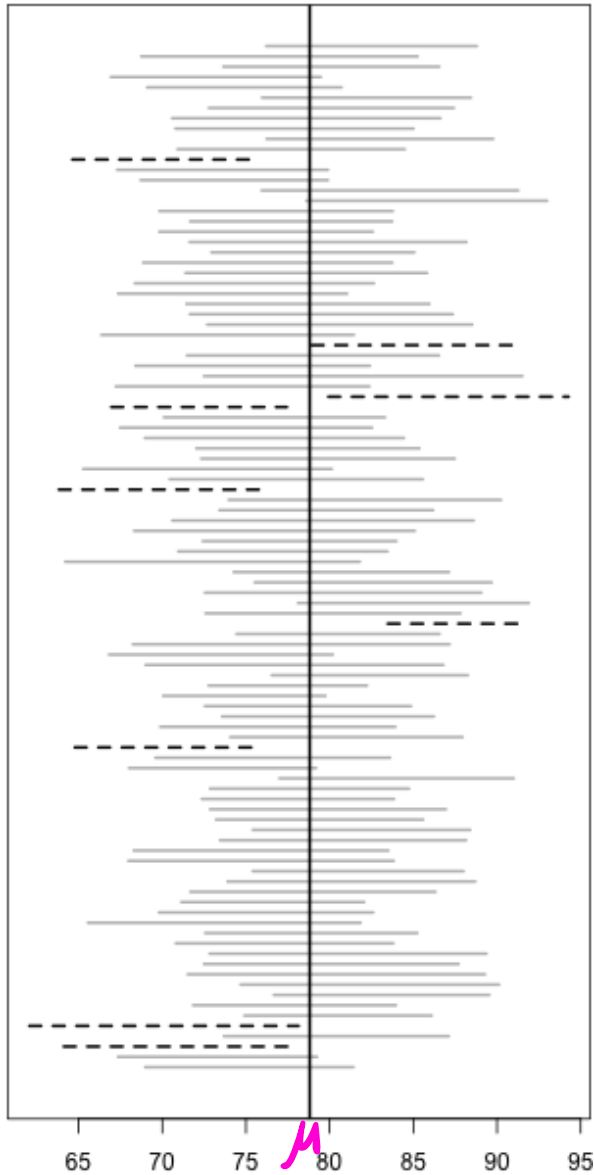
# Confidence Interval interpretation

**Interpretation:** “If we were to repeat the process of taking samples and calculating 95% confidence intervals, we’d expect 95% of confidence intervals to contain the true population average  $\mu$ .”

**Warning:** a 95% confidence interval does not mean there’s a 95% chance the true value  $\mu$  lies in the interval! Remember:  $\mu$  is a *fixed number*. It doesn’t make sense to talk about the probability it either is, or isn’t, in a confidence interval with probability 1! 

“We are \_\_\_ % confident the interval  
from lower to higher captures the population  
parameter.”

# Confidence Intervals: Sampling distribution



- Here we have generated 100 confidence intervals by taking a SRSWOR with  $n = 10$  and then calculated the confidence interval.
- Nine CIs do not cover the true parameter value.

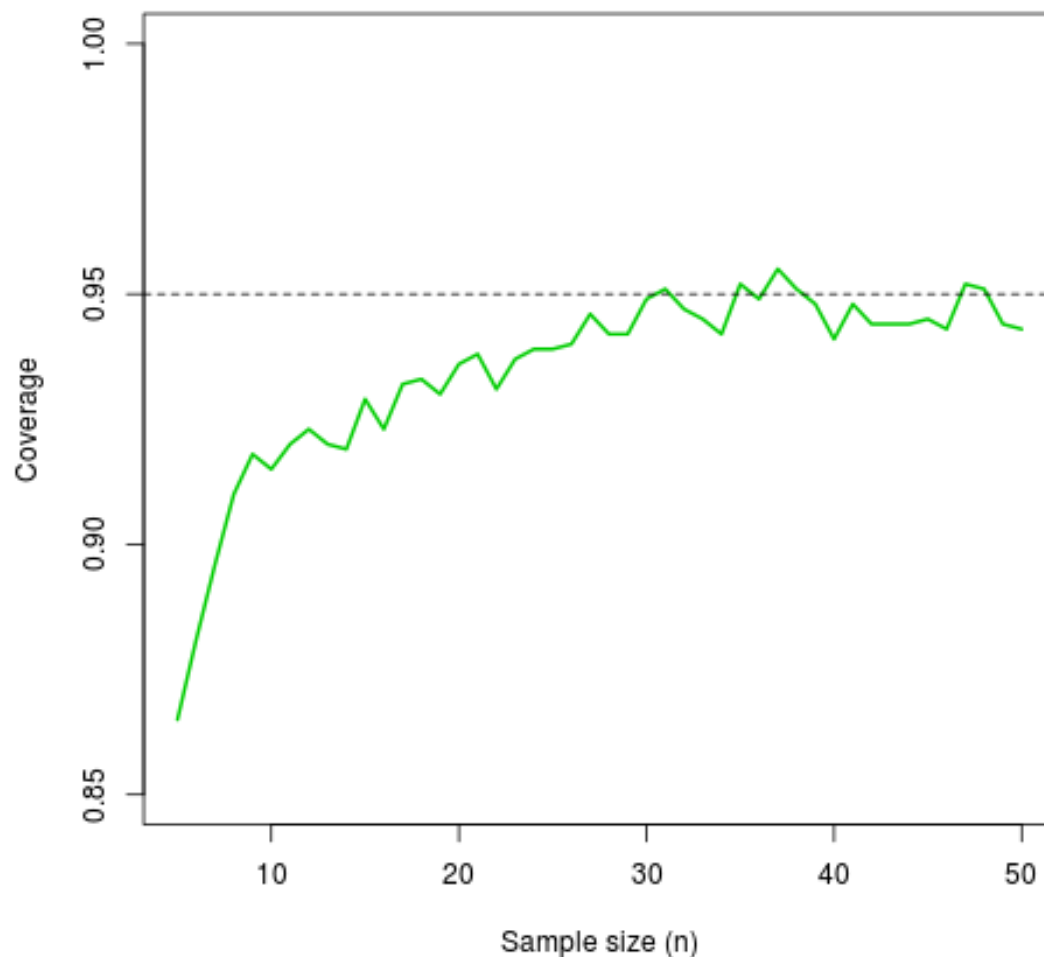


# Confidence Intervals: Sampling distribution

- CI calculations ultimately rely on the central limit theorem. In particular, that  $n$ ,  $N$  and  $N - n$  are 'large'.
- A 'rule of thumb' is that these should all be at least equal to 30.
- In the grades example,  $n = 10$ ,  $N = 116$  and  $N - n = 106$ , so we'd have concerns.
- We can investigate the impact by taking a large number of samples and calculating a confidence interval for each.
- We can then look at the **coverage**: we count the proportion of confidence intervals contain the population average ( $\mu$ ).
- For  $n = 10$  only about 91% of confidence intervals contain the true average!

# Coverage Probability while varying $n$

Our coverage should be around 95% (or 0.95), which happens around the sample size of  $n = 30$ .



# Population Total

- Sometimes we may be interested in a population *total* rather than an average.
- e.g. rather than the average price of items in a store room, we want to know the total value of items in a store room.
- We can derive the necessary formulas to calculate a confidence interval for the total value using our formulas for the average.
- The results apply to the estimation of a population total,  $\tau$ , quite easily since

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

N.B. in the above  $\bar{y}$  refers to the population mean  $= \frac{1}{N} \sum_{i=1}^N y_i$ . We might use  $\bar{y}_U$  to emphasize that this is based on the population.

# Confidence Interval: Population Total

- True value:  $\tau = N\mu$
- Sample estimate:  $\hat{\tau} = N\hat{\mu}$
- Sample estimator:  $\tilde{\tau} = N\tilde{\mu}$
- $E[\tilde{\tau}] = E[N\tilde{\mu}] = NE[\tilde{\mu}]$
- $Var(\tilde{\tau}) = Var(N\tilde{\mu}) = N^2 Var(\tilde{\mu})$
- $s.e.(\hat{\tau}) = N \times s.e.(\hat{\mu})$

A  $100(1 - \alpha)\%$  CI for  $\tau$  is:

$$\hat{\tau} \pm c \times s.e.(\hat{\tau}) = N\hat{\mu} \pm c * N \times s.e.(\hat{\mu})$$

# Confidence Interval: Population Proportion

- We might also be interested in a population proportion.
- e.g. what proportion of students scored over 90?
- If we define a variable such that  $y_i = 1$  if the unit has that property, and  $y_i = 0$  otherwise, then the population proportion,  $\pi$ , is the attribute of interest.

$$\pi = \frac{1}{N} \sum_{i \in U} y_i = \frac{M}{N} = \bar{y}_U$$

i.e. the number of units with the characteristic in the population,  $M = \sum_{i \in U} y_i$ , divided by  $N$ , the total number of units in the population.

- $\pi$  is just a population average so all our previous theory applies!

# Confidence Interval: Population Proportion

See (Lohr, 2009, Example 2.6, Section 2.3)

- True value:  $\pi = \frac{1}{N} \sum_{i \in U} y_i = \frac{M}{N} = \bar{y}_U$
- Sample estimate:  $\hat{\pi} = \frac{1}{n} \sum_{i \in s} y_i = \bar{y}_s$
- Sample estimator:  $\tilde{\pi} = \frac{1}{n} \sum_{i \in S} y_i = \bar{y}_S$
- $E[\tilde{\pi}] = \pi$  (unbiased)

For the response  $y_i$ , taking on values 0 or 1,

$$\sigma^2 = \frac{N}{N-1} \pi(1 - \pi)$$

Since  $Var(\tilde{\mu}) = (1 - \frac{n}{N}) \frac{\sigma^2}{n}$  this implies that

$$Var(\tilde{\pi}) = \left( \frac{N-n}{N-1} \right) \frac{\pi(1-\pi)}{n}.$$

# Confidence Interval: Population Proportion

Again, since the population variance  $\sigma^2$  depends on the values for the entire population, which is in general unknown, we estimate it by the sample variance:

$$\hat{\sigma}^2 = \frac{n}{n-1} \hat{\pi}(1 - \hat{\pi})$$

An estimate of  $Var(\hat{\pi})$  is then

$$\widehat{Var}(\hat{\pi}) = \left(1 - \frac{n}{N}\right) \frac{\hat{\pi}(1 - \hat{\pi})}{n-1}$$

A  $100(1 - \alpha)\%$  CI for  $\tau$  is:

$$\hat{\pi} \pm c \times \text{s.e.}(\hat{\pi}) = \hat{\pi} \pm c \times \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{\pi}(1 - \hat{\pi})}{n-1}}$$

# Sample size calculation

- Noticed that the standard errors for the confidence intervals we consider are inversely proportional to our sample size  $n$ .
- That is, since  $n$  appears in the denominator, the larger our sample size the less our plus/minus term (aka the margin of error) in the CI becomes.
- Let's consider the margin of error for the CI for a large sample:

$$e = z_{\alpha/2} \sqrt{1 - \frac{n}{N}} \frac{\sigma}{\sqrt{n}}$$

- Factors that affect the error:
  - $\alpha$
  - $\sigma$
  - $n$



# Sample size calculation

- Only the investigators in the study can say how much precision is needed. For the population mean:

$$P(|\mu - \tilde{\mu}| < e) = 1 - \alpha$$

where  $e$  is the *margin of error*. In the above,

- $\alpha$ ,  $e$ , settled by the investigator
- $\sigma$  is unknown
- If the fpc is approximately 1, as is the case most of the time, we can rearrange the margin of error equation to solve for  $n_0$ :

$$e = z_{\alpha/2} \frac{\sigma}{\sqrt{n_0}} \quad \implies \quad n_0 = z_{\alpha/2}^2 \frac{\sigma^2}{e^2}$$

- In cases where the fpc is small (i.e. when  $n$  is large compared with the population size) we would make the fpc adjustment:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{z_{\alpha/2}^2 \sigma^2}{e^2 + \frac{z_{\alpha/2}^2 \sigma^2}{N}}$$

# Sample size calculation

- We will need a value for  $\sigma$  to determine a sample size, however, that is usually unknown. In that case:
  - A pilot sample can provide information and guidance for the design of the main survey.
  - Use previous studies or data available in the literature
  - If nothing else is available, guess the variance
- In surveys on proportions, for large populations,  $\sigma^2 \approx \pi(1 - \pi)$ , which is maximized when  $\pi = 1/2$ .
- Hence using  $n_0 = 1.962/(4e^2)$  will result in a 95% CI with width at most  $2e$ .

## Example 2 (Example 2.11 from Lohr (2009))

Suppose we want to estimate the proportion of recipes in the *Better Homes & Gardens New Cook Book* that do not involve animal products. We plan to take an SRSWOR of the  $N = 1251$  test kitchen-tested recipes, and want to use a 95% CI with margin of error 0.03. What sample size should we take?

$$n_0 = (1.96) \frac{\frac{1}{2}(1 - \frac{1}{2})}{0.03^2} \approx 1067$$

Since the sample size is large compared to  $N = 1251$ , we use the correction:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1067}{1 + \frac{1067}{1251}} = 576$$

N.B. if  $n_0 \geq N$  we require to take a census with  $n = N$ .

# SRSWR

- SRSWR is near-identical to SRSWOR.
- Only difference: at every step we pick from the full set of individuals in the frame with probability  $1/N$ .
- Units can therefore appear more than once in the sample.
- If  $N$  is large, then SRSWOR and SRSWR are similar.

We estimate the population average with the sample average

$$\hat{\mu} = \bar{y}.$$

Similarly, the sample variance is as before:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \hat{\mu})^2.$$

# SRSWR

Under SRSWR

$$E[\tilde{\mu}] = \mu \quad \text{and} \quad \text{Var}(\tilde{\mu}) = \left(1 - \frac{1}{N}\right) \frac{\sigma^2}{n}$$

So SRSWR is also unbiased for the mean!

# SRSWR vs SRSWOR

For sampling without replacement we have  $\tilde{\mu}_{SRSWOR}$  and  $\tilde{\mu}_{SRSWR}$  for with replacement. We have

$$Var(\tilde{\mu}_{SRSWOR}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}$$

$$Var(\tilde{\mu}_{SRSWR}) = \left(1 - \frac{1}{N}\right) \frac{\sigma^2}{n}$$

The variance of the mean estimator is larger for SRSWR than for SRSWOR since  $\frac{n}{N} > \frac{1}{N}$  and as long as  $n \geq 2$

We say that SRSWOR is **more efficient** than SRSWR: our resulting estimator is more precise (less uncertainty).

# When to use SRSWOR

Simple random samples are usually easy to design and easy to analyze. But they are not the best design to use in the following situations:

- To find causal relationships, we'll need to do an experiment instead of a survey.
- Expensive to do for a list of observation units, eg. cannot construct a sampling frame of individual mosquitoes in an area. (in this case, perhaps cluster sampling should be used)
- You may have additional information that can be used to design a more cost-effective sampling scheme, eg. strata consisting of terrain likely to have high/low mosquito densities.

# When to use SRSWOR

- When little extra information is available that can be used when designing the survey, eg. sampling university students' names from a list (no info on year, major), then simple random sampling is probably the best probability sampling strategy.
- Note that systematic sampling can be used as a proxy for simple random sampling, when no list of the population exists or when the list is in roughly random order.
  - If a list of students is ordered by randomly generated student identification numbers, systematic sampling will probably obtain a sample that will behave much like an SRSWOR.

## References for this lecture:

Lohr, S. (2009), *Sampling: design and analysis*, Nelson Education.