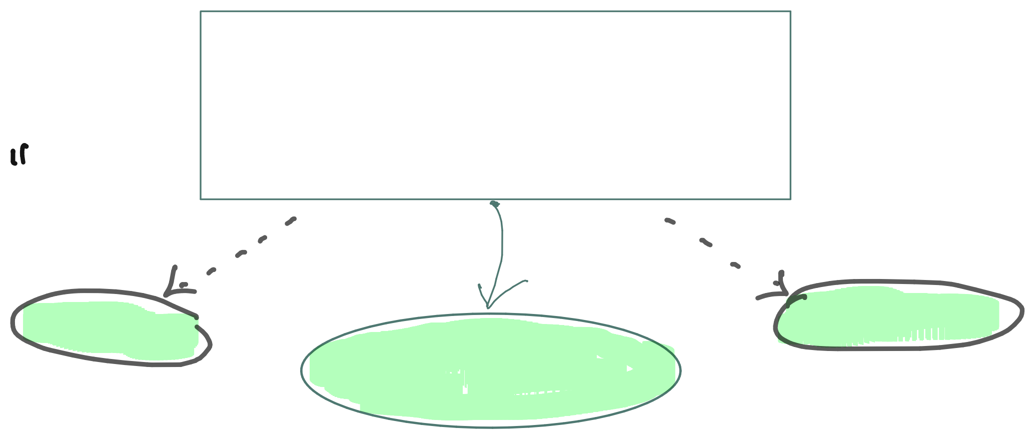
Module 3: Sampling Data



T6: Understanding Chance
T7: Chance Variability (The Box Model)
T8: Sample Surveys

LO6: Use the box model to describe chance & chance variability, including sample surveys & the CLT.

Parameters and Estimates

"The estimate is what the researcher knows"
"The parameter is what the researcher wants to know"

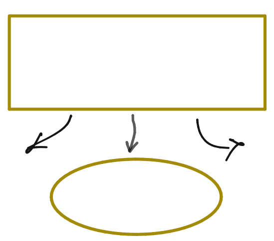
population eg/ census* → parameter (eg/ population mean)

↓

sample → estimate (eg sample mean)

* ideal, but many limitations

Selecting a Sample



Many types of bias:

- selection
- non-response
- interviewer
- measurement
  - question wording/order
  - recall bias
  - sensitive questions
  - lack of clarity
  - interview process

Warning: A larger sample does not reduce bias.

It can amplify bias!

Types of sampling

- multi-stage cluster
- quota
- convenience
- simple random

Unavoidable bias

estimate = parameter
+ bias
+ chance error ↑ Box Model

Box Model for Sample Surveys

Q: what % of DATA1001 students are International?

1 = International → $\boxed{1 \times p \quad 0 \times (1-p)}$ $\xrightarrow{n}$ 1,0,1...

mean = p = % of Internationals

$SD = (1-0)\sqrt{p(1-p)} = \sqrt{p(1-p)}$

Sum of Sample = # of Internationals

$EV = np$

$SE = \sqrt{n}\sqrt{p(1-p)}$

Mean of Sample = % of Internationals

$EV = p$

$SE = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Correction Factor for Samples

N

with out replacement

n

$$\underset{w/o}{SE} = \underset{with}{SE} \cdot f_{pc}$$

finite population correction

$$f_{pc} = \sqrt{\frac{N-n}{N-1}}$$

Bootstrapping = estimating the properties of the population, by using the properties of the sample.

population → sample

parameter = unknown

Eg//

$N = 100{,}000$

$n = 10$

$8 \times 1 \quad 2 \times 0$

so we estimate the population to have:

$$\frac{8}{10} \times 100{,}000 = 80{,}000 \text{ 1's}$$

$$\frac{2}{10} \times 100{,}000 = 20{,}000 \text{ 0's}$$

Confidence Intervals for p

unknown population parameter (EV)

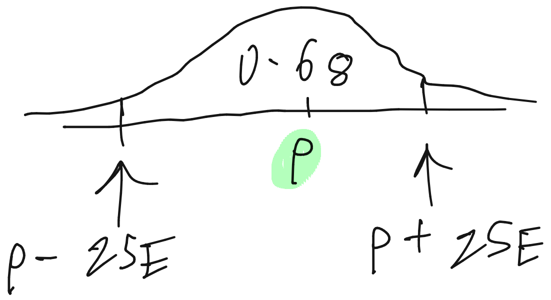
sample mean = $\hat{p}$

$$N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

68% CI for p : sample proportion $\pm$ 1. SE

95% CI for p : $\pm$ 2SE

99.7% CI for p : $\pm$ 3SE



Interpretation

✗ "The 95% CI for p contains p 95% of the time"

✓ "If we worked out 100 CIs for p, we would expect 95% to contain p"

✓ "The range of values of p_0 for which we would retain $H_0: p = p_0$, where $\alpha = 0.05$"

[see Module 4]

Eg//

$N = 100{,}000$

Bootstrap:

$0.8 \times 1 \quad 0.2 \times 0$

mean = p = ?

$$SD \doteq (1-0)\sqrt{0.8 \times 0.2} = 0.4$$

$n = 10$

$8 \times 1 \quad 2 \times 0$

sample proportion = 0.8

Approx. 95% CI for p :

$$0.8 \pm 1.0.4$$

$$(0.4,\ 1.2)$$

Summary of the Box Model

applied to a Sample Survey

The Box Model

size = N

population = “box” = ... *survey question*

$N_1 \times 1$ $\quad$ $(N - N_1) \times 0$

(proportion of 1's)

mean = $\frac{N_1}{N}$

SD = $(1-0)\sqrt{\frac{N_1}{N}\frac{(N-N_1)}{N}}$

n =

observed sample

observed sum =

observed mean = *(proportion)*

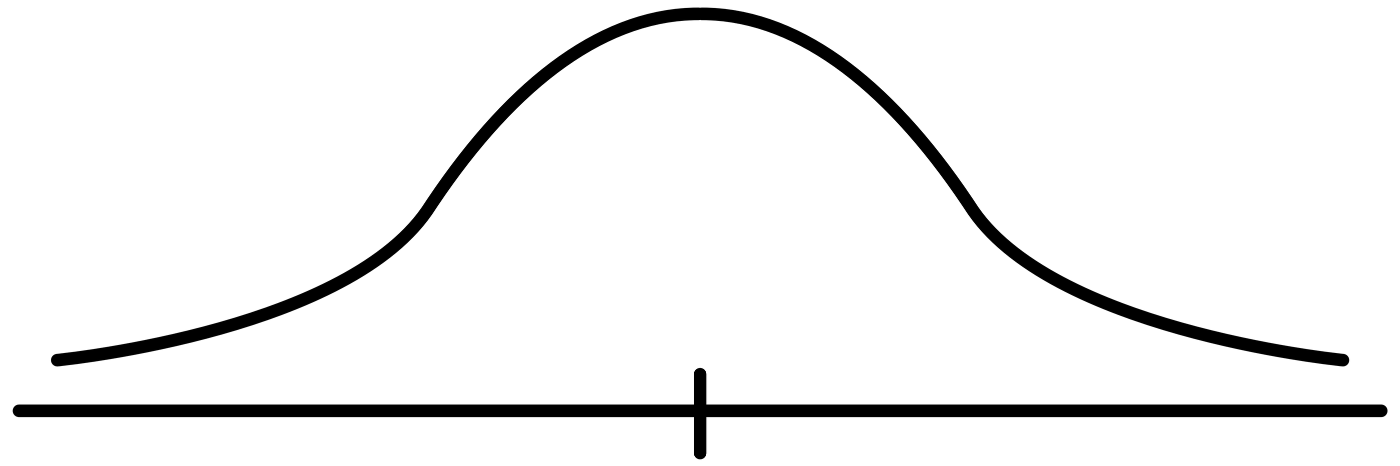
observed SD =

$*$ *with correction factor:*

$$SE^* = \sqrt{\frac{N-n}{N-1}} \cdot SE$$

Model of Sample

Sample Sum	EV = n mean	
	$*$ SE = $\sqrt{n}$ SD	
Sample Mean	EV = mean	
	$*$ SE = $\frac{SD}{\sqrt{n}}$	



Sample Sum / Mean

N ($\qquad$, $\qquad$)