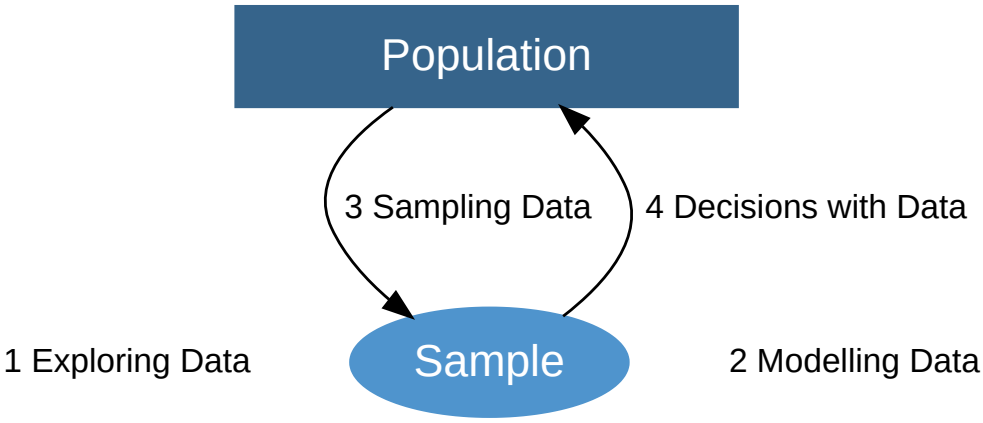
The Box Model for Sample Surveys

Sampling Data | Sample Surveys

Unit Overview



Module3 Sampling Data

Understanding Chance

What is chance?

Chance Variability

How can we model chance variability by a box model?

Sample Surveys

How can we model the chance variability in sample surveys?

The Box Model for Sample Surveys

Data Story

Same-Sex Marriage

Same-Sex Marriage

- Between September 12 and November 7 2017 Australians participated in a postal plebiscite on same-sex marriage.

- In the lead up to the plebiscite a number of prominant opinion polls conducted phone surveys in order to estimate the proportion of the Australian population who supported same-sex marriage.

Same-Sex Marriage

- A summary of three opinion polls conducted in August 2017 are summarised below.

Date	Firm	Support	Oppose	Undecided
17-22 August 2017	Essential	57%	32%	11%
17-21 August 2017	YouGov	59%	33%	8%
17-20 August 2017	Newspoll	63%	30%	7%

- Note 57% is equivalent to a proportion of 0.57.

Statistical Thinking

With the person next to you discuss:

- Why do the different opinion polling firms have different percentages for “Support”, “Oppose” and “Undecided”?
- Are the populations the same for each of the opinion polls?

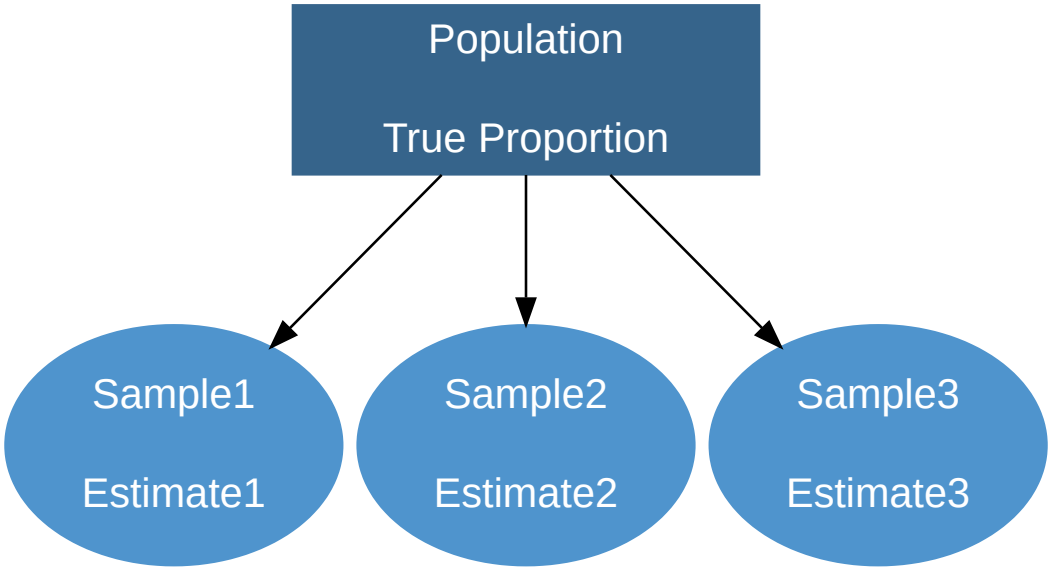
The Box Model: Modelling the Proportion (Mean) of a Sample

Chance Errors in Sample Surveys

- Sample surveys involve chance error, because each sample is just one possible draw from the population.
- Here we use the Box Model to quantify the likely size of the chance error when estimating a proportion using simple random sampling. Standard Errors (SE) measure the variability across different samples from the same population.
- Note that the **Proportion** of a sample survey based on a 1-0 box model, is a special case of the **Mean** of the sample.

	EV	SE
Sum of the Sample	n mean	$\sqrt{n}$ SD
Mean of the Sample	mean	SD / $\sqrt{n}$

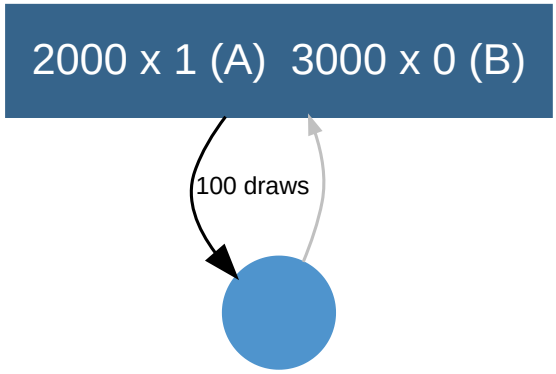
Drawing a Simple Random Sample



Modelling Sampling by a Box Model

- Consider a simple random sample of 100 draws from a population of 5000 individuals, where 2000 will vote A and and 3000 will vote B.
- We are interested in the **Proportion** of A voters in the sample.
- What is the chance that the number of A voters is between 0.3 and 0.5?

Step1: Draw the box model



Step2: Calculate the mean and SD of the box

- The mean is $\frac{2000\times1+3000\times0}{5000} = 0.4$.
- The SD is $(1-0)\sqrt{2/5 \times 3/5} = \sqrt{6}/5 \approx 0.5$. [Note SE is rounded up here to simplify illustration.]

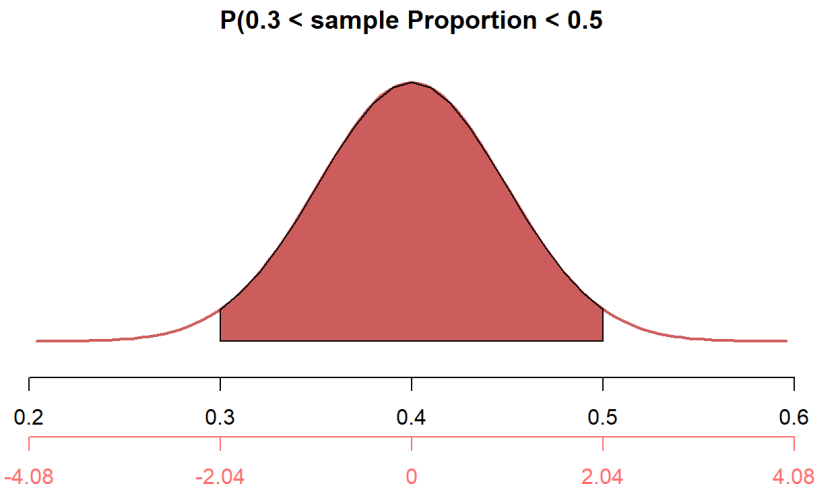
Step3: Calculate the EV and SE of the Proportion (Mean) of the Sample

- The EV of the Proportion of the draws is 0.4.
- The SE of the Proportion of the draws is $\frac{0.5}{\sqrt{100}} = 0.05$.

Step4: Conclusion

- We would expect a Sample Proportion of 0.4 (EV) with SE 0.05.
- This means, it would not be unusual to get the proportion of A voters between $0.4 \pm 2 \times 0.05$ or even $0.4 \pm 3 \times 0.05$ (assuming a Normal curve).

Step5: Draw the Normal curve



Step6: Calculate the chance

- The x values are 0.3 and 0.5.
- The z scores are approximately $\frac{0.3-0.4}{0.05} = -2$ and $\frac{0.5-0.4}{0.05} = 2$.
- So we expect the Proportion to be between 0.3 and 0.5 about 95% of the time.

In R

```
box = c(0, 0, 0, 1, 1)
# Or box = c(rep(0, 3), rep(1, 2))

c(mean(box), popsd(box)/sqrt(100))
```

```
## [1] 0.40000000 0.04898979
```

```
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

```
pnorm(0.5, 0.4, 0.05) - pnorm(0.3, 0.4, 0.05)
```

```
## [1] 0.9544997
```

```
pbinom(50, 100, 2/5) - pbinom(30, 100, 2/5)
```

```
## [1] 0.9584555
```

- Note the effect of the sample size n on the SE:
  - $\mathrm{SE}_{sum} = \sqrt{n} \times \mathrm{SD}_{box}$
  - $\mathrm{SE}_{proportion} = \frac{\mathrm{SD}_{box}}{\sqrt{n}}$.
- This is an equivalent problem: What is the chance that the **number** of A voters is between 0.3 and 0.5? We model the **Sum** of the Sample.

```
box = c(0, 0, 0, 1, 1)

c(100 * mean(box), sqrt(100) * popsd(box))
```

```
## [1] 40.000000  4.898979
```

```
pnorm(50, 40, 5) - pnorm(30, 40, 5)
```

```
## [1] 0.9544997
```

Summary of Sample Survey

Focus in the Sample	EV	SE
Sum	$\text{sample size} \times \text{mean}_{box}$	$\sqrt{\text{sample size}} \times \text{SD}_{box}$
Proportion (Mean)	mean_{box}	$\frac{\text{SD}_{box}}{\sqrt{\text{sample size}}}$

in R

	EV	SE
Sum	`n*mean(box)`	`sqrt(n)*popsd(box)`
Proportion	`mean(box)`	`popsd(box)/sqrt(n)`

where `n` = size of sample (number of draws from the box).

The correction factor

Statistical Thinking

With the person next to you discuss, which of the 2 polls would be more accurate.

- State poll on the voting preference of 1 million of the 7.5 million people in NSW.
- National poll on the voting preference of 1 million of the 24 million people in Australia.

What affects accuracy?

- When sampling with replacement, the SE is determined by the absolute **size of the sample**.
- When sampling without replacement, the SE will be decreased by increasing the ratio of **sample size** to **population size**, as when a higher proportion of the population is sampled, the variability will decrease.
- When the sample is only a small part of the population, the size of the population has almost no effect on the SE of the estimate.

Why sample size determines accuracy

- Assume Box1 is size N_1 (large) and Box2 is size N_2 (much smaller).
- Assume Box1 and Box2 both have 50% 0's and 50% 1's (modelled by 0 and 1).
- Assume we sample n draws from each box with replacement.
- Both boxes have the same mean 0.5 and SD 0.5.
- Both boxes have the same $\mathrm{EV}_{Proportion}$.

$$\mathrm{EV}_{Box1} = \mathrm{EV}_{Box2} = 0.5$$

- Both boxes have the same chance error.

$$\mathrm{SE}_{Box1} = \frac{0.5}{\sqrt{n}} = \mathrm{SE}_{Box2}$$

- Hence both boxes have the same accuracy in estimating the population proportion. Drawing with replacement, the box (0,1) is equivalent to (0,0,1,1) etc.

Drawing without replacement

- In practise, most sample surveys are drawn **without** replacement (ie the same person can't be interviewed twice).
- Strictly, this is a different context to the box model, which assumes draws **with** replacement.
- Hence, we need to slightly adjust the SE from the box model, to get the exact SE.
- Hoewever, if the size of the population is a lot bigger than the sample, then the correction factor is almost 1.

The correction factor (finite population correction)

Correction Factor

$$\text{SE}_{without replacement} = \text{correction factor} \times \text{SE}_{with replacement}$$

where

$$\text{correction factor} = \sqrt{\frac{\text{number of tickets-number of draws}}{\text{number of tickets -1}}}$$

$$\text{correction factor} = \sqrt{\frac{\text{population size-sample size}}{\text{population size -1}}}$$

The correction factor

Suppose that the sample size is fixed at 2,500. The table below summarises the correction factor (to 5 dp) for different population sizes.

Population size	Correction factor
5,000	0.70718
10,000	0.86607
100,000	0.98743
500,000	0.99750
1,000,000	0.99875
12,500,000	0.99990

Summary

We can model a simple survey by a box model with 0 and 1 in the relevant proportions.

Focus in the Sample	EV	SE
Sum	$\text{sample size} \times \text{mean}_{box}$	$\sqrt{\text{sample size}} \times \text{SD}_{box}$
Proportion (Mean)	mean_{box}	$\frac{\text{SD}_{box}}{\sqrt{\text{sample size}}}$

in R

	EV	SE
Sum	`n*mean(box)`	`sqrt(n)*popsd(box)`
Proportion	`mean(box)`	`popsd(box)/sqrt(n)`

where `n` = size of sample (number of draws from the box).

- To calculate the chance of getting a proportion in a certain range, we convert to standard units and use the Normal curve.

- If we draw without replacement, then strictly the SE should be adjusted by the correction factor.

$$\text{correction factor} = \sqrt{\frac{\text{population size-sample size}}{\text{population size -1}}}$$

- However, for large population size compared to sample size, the correction factor is almost 1.

Key Words

plebiscite, opinion polls, simple random sample, correction factor