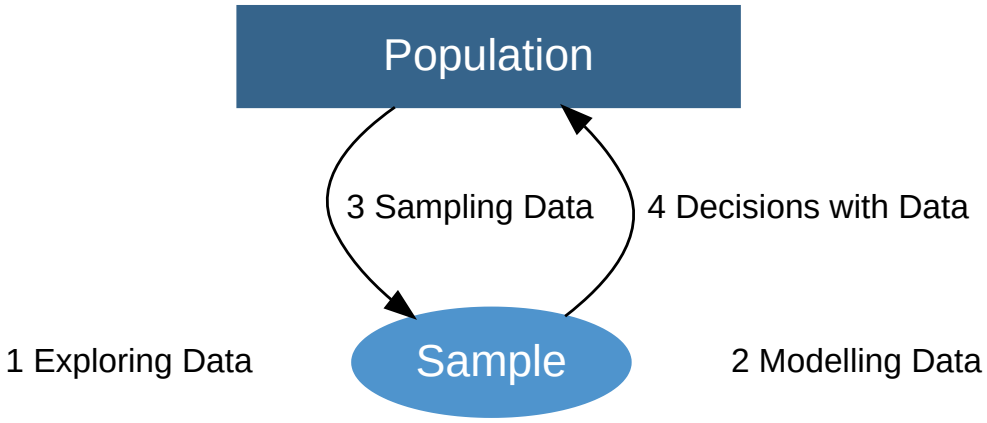
Bootstrapping & Confidence Intervals (Accuracy of Proportions)

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?

Accuracy of Proportions

Data Story

Same-Sex Marriage Polls

Same-Sex Marriage Polls

In the lead up to the plebiscite, the following three opinion polls were conducted in August 2017.

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%

Statistical Thinking

With the person next to you discuss:

- Could you have predicted the final result from these polls?
- See SMH visualisation

Estimating the population proportion using bootstrapping

The gap in information

We want to predict the population proportion from the sample proportion.

Previously, we found these results:

- The EV of the **Sample** Proportion is equal to the **population** proportion.

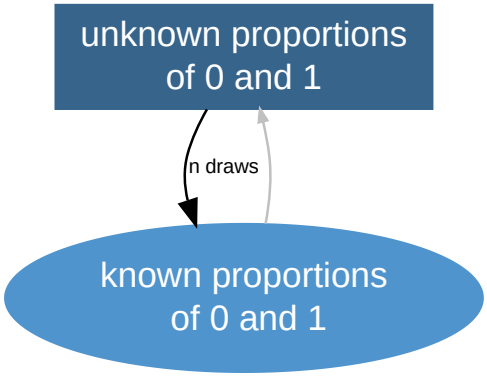
$$\text{EV}_{proportion} = \text{mean}_{box} = \text{population proportion}$$

- The chance error is related to the SE of the Sample Proportion, which is:

$$\text{SE}_{proportion} = \frac{\text{SD}_{box}}{\sqrt{\text{sample size}}}$$

What is the problem with using these formulas? What could be a solution?

Solution?


unknown proportions
of 0 and 1
n draws
known proportions
of 0 and 1

Bootstrapping

Bootstrapping

- Bootstrapping is estimating the properties of the population, by using the properties of a particular sample.
- When sampling from a 0-1 box, we replace the unknown proportion of 1's in the box (population) by the known proportion of 1's in a particular sample.

Bootstrapping the population proportion

Step1: Create an approximate box (box1) - which has the same proportion of 0s and 1s as the sample.

Step2: Use the Box Model

Focus in the Sample	EV	SE
Sum	$\text{sample size} \times \text{mean}_{box1}$	$\sqrt{\text{sample size}} \times \text{SD}_{box1}$
Proportion (Mean)	mean_{box1}	$\frac{\text{SD}_{box1}}{\sqrt{\text{sample size}}}$

Statistical Thinking

Consider the YouGov survey which gave support of 59%, opposed 33%, and undecided 8%, with a sample size of 1012.

- How would you estimate the proportion of support in the population?
- What do you need to assume?
- Using bootstrapping, find the chance that the proportion of support in the population is less than 0.625.

Bootstrapping assumptions

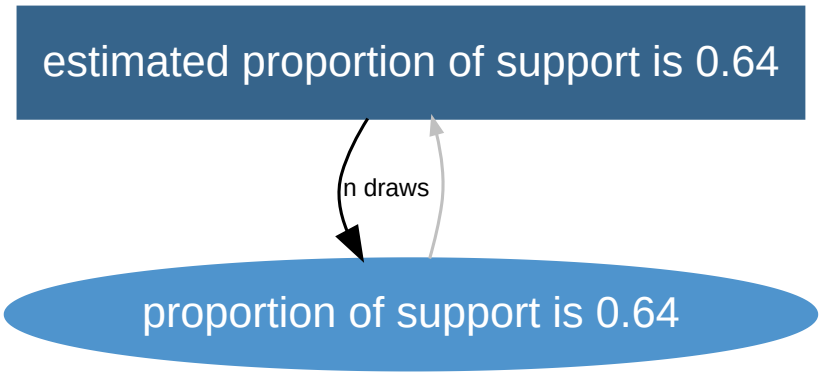
For the YouGov survey:

1. Assume the population proportion of support to be $\frac{59}{59+33}\% \approx 64\%$.

2. We assume the Australian population of eligible voters to be approximately 16,655,856 at the time of the poll.

- Note there is bias as only 79.5% voted.

Step1: Draw the (approximate) box model



Use the box model

- The estimate of the population proportion is the EV of the Sample Proportion = 0.64.
- The chance error is related to the SE of the Sample Proportion, which is $\frac{0.48}{\sqrt{1012}} = 0.015$.

```
# Using the short cut formula for SD of 0-1 box (with sample proportions)
(1 - 0) * sqrt(0.64 * 0.36)
```

```
## [1] 0.48
```

```
# Or: calculating SD of box with sample proportions direct
N = 16655856
box = c(rep(1, 0.64 * N), rep(0, 0.36 * N))
sd(box) * sqrt((N - 1)/N)
```

```
## [1] 0.48
```

Final Result

- We could estimate the proportion of Australians supporting the final plebiscite to be 0.64, with a chance error of around 0.015.

Note: We should adjust the SE by the finite population correction factor $\sqrt{(N-n)/(N-1)} = \sqrt{(16,655,856-1012)/(16,655,856-1)}$.

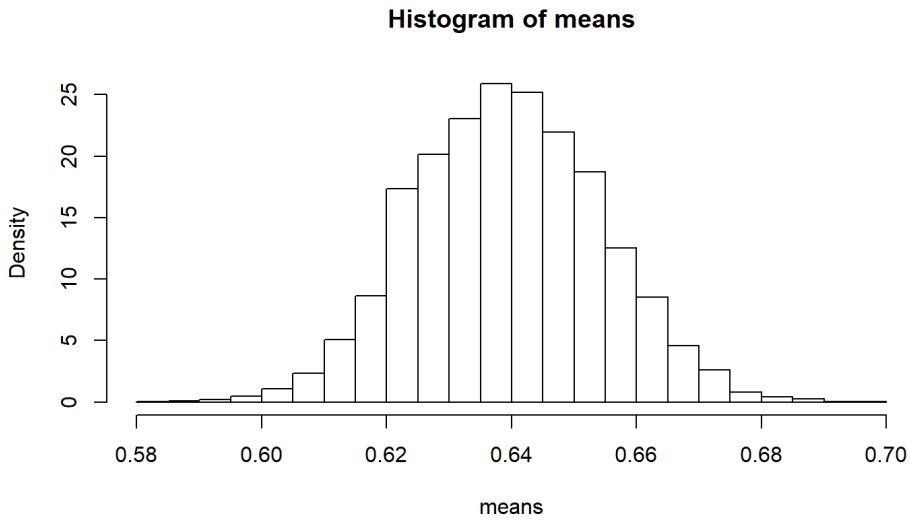
However, here the fpc is close to 1.

Bootstrapping a distribution

```
set.seed(123)
N = 16655856
box = c(rep(1, 0.64 * N), rep(0, 0.36 * N))
means = round(replicate(10000, mean(sample(box, 1012, rep = F))), 3)
cumsum(table(means))
```

```
## 0.584 0.587 0.588  0.59 0.591 0.592 0.593 0.594 0.595 0.596 0.597 0.598
##     1     2     5     6     8    10    12    15    16    19    25    30
## 0.599   0.6 0.601 0.602 0.603 0.604 0.605 0.606 0.607 0.608 0.609  0.61
##    36    40    48    51    65    78    93   114   133   154   181   209
## 0.611 0.612 0.613 0.614 0.615 0.616 0.617 0.618 0.619  0.62 0.621 0.622
##   246   297   335   388   462   540   620   709   791   895  1024  1142
## 0.623 0.624 0.625 0.626 0.627 0.628 0.629  0.63 0.631 0.632 0.633 0.634
##  1275  1443  1763  1935  2136  2332  2547  2770  2985  3216  3447  3681
## 0.635 0.636 0.637 0.638 0.639  0.64 0.641 0.642 0.643 0.644 0.645 0.646
##  3921  4185  4447  4682  4941  5216  5492  5753  5998  6242  6476  6697
## 0.647 0.648 0.649  0.65 0.651 0.652 0.653 0.654 0.655 0.656 0.657 0.658
##  6934  7141  7342  7574  7799  7989  8179  8369  8511  8647  8775  8896
## 0.659  0.66 0.661 0.662 0.663 0.664 0.665 0.666 0.667 0.668 0.669  0.67
##  9012  9138  9237  9335  9421  9486  9565  9622  9680  9732  9764  9794
## 0.671 0.672 0.673 0.674 0.675 0.676 0.677 0.678 0.679  0.68 0.681 0.682
##  9824  9859  9879  9906  9924  9937  9944  9951  9956  9964  9972  9975
## 0.683 0.684 0.685 0.686 0.687 0.688 0.689  0.69 0.693   0.7
##  9977  9983  9986  9989  9993  9994  9996  9998  9999 10000
```

```
hist(means, breaks = 20, freq = F)
```



Using the output, we can approximate
$P(Proportion < 0.625) = 1746/10000 \approx 0.17.$

Confidence Intervals

Chance Error and Standard Error

- Up to now, we have often taken the estimate of the chance error to be 1 unit of the SE. Hence, the proportion of Australians supporting the final plebiscite could be anywhere between 0.64 ± 0.015.
- However, the chance error could be out by as much as 2 or even 3 SEs.
- Now we generalise, using Confidence Intervals.

Confidence Intervals (CI)

Confidence Intervals for population proportion

68% confidence interval

$$\text{sample proportion} \pm 1 \times \text{SE}$$

95% confidence interval

$$\text{sample proportion} \pm 2 \times \text{SE}$$

99.7% confidence interval

$$\text{sample proportion} \pm 3 \times \text{SE}$$

Interpreting Confidence Intervals

For a 95% Confidence Interval:

- It is a mistake to say “the probability that the interval contains the unknown parameter is 0.95”.
- Rather, we say ’if we worked out a series of CIs for a series of samples, then 95% of the CIs would contain the unknown parameter."

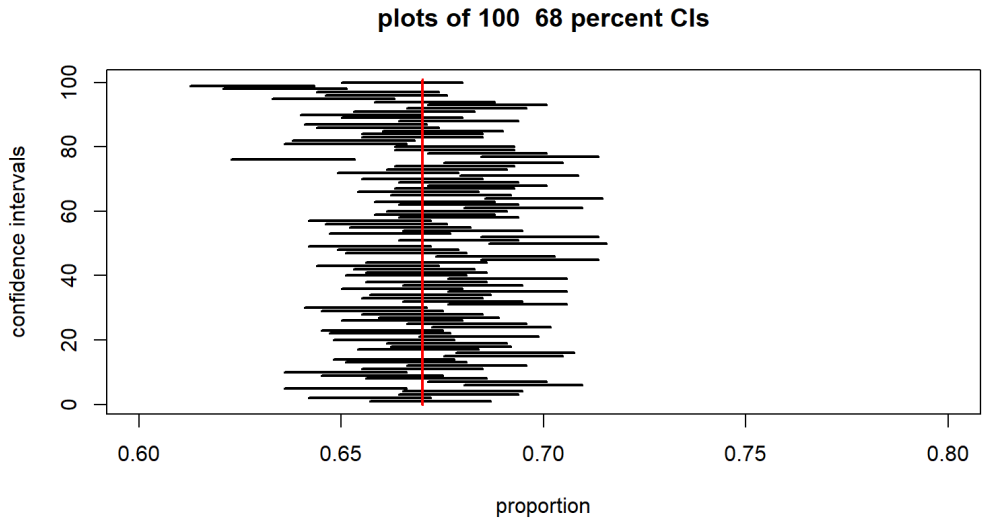
Visualising a series of CI

Simulating a series of CIs

Here we simulate the data story:

- Create a population of size 1000000, where the proportion of 1s (“Yes” votes) is 0.67.
- Draw a sample of size 1000 from the population, and calculate a 95% CI (black line) for the population proportion.
  - Repeat this sampling 100 times, forming 100 CIs.
  - Graph the 100 CIs.
- Draw a red line to represent the true population proportion (0.67) and check how many CIs fall inside and outside the red line.
  - We expect approximately 95% of CIs to “cover” the true proportion.

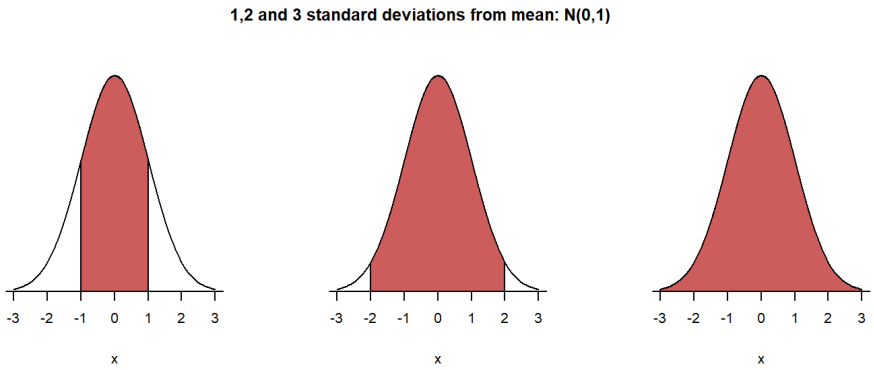
Note: Unless we draw without replacement, the fpc applies for the SE, though here it is very close to 1 for the sample survey.




```
## The true value of p was captured in the CI  72  percent of times
```

Justifying the CI formulas

- We assume the Sample Proportion (estimating the population proportion) follows a Normal distribution.
- Recall all Normal distributions satisfy the "68%-95%-99.7% Rule.



```
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

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

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

```
pnorm(3) - pnorm(-3)
```

```
## [1] 0.9973002
```

Statistical Thinking

- Using the YouGov survey, what is the 95% confidence interval for the support of same sex marriage plebiscite?
- What does it represent?

Methods of Sampling

- The methods described above, relate to simple random samples.
- In practise, sampling can be much more complex, requiring more sophisticated statistical reasoning and approaches.

Summary

We use bootstrapping to estimate the proportions within the population from the proportions within the sample. This allows us to work out the EV, the SE and confidence intervals.

Key Words

plebiscite, opinion polls, simple random sample

Further Practise

App:Brown

App:Kristoffer Magnusson

App:Rossman


Understanding Confidence Intervals: Statistics Help


Confidence Intervals - Introduction