Week 8 8.2 Note

Lecture 25: Sample Surveys

---

1. Population and Samples

> **Population**: the *full amount of information* being studied, collected though a *census（普查）*.

> **Sample**: *part* of the population.

- Limitations of a Census:
    - Collecting every unit of a population:
        - Is hard
        - Is time-consuming
        - Costs money
        - Needs lots of resources

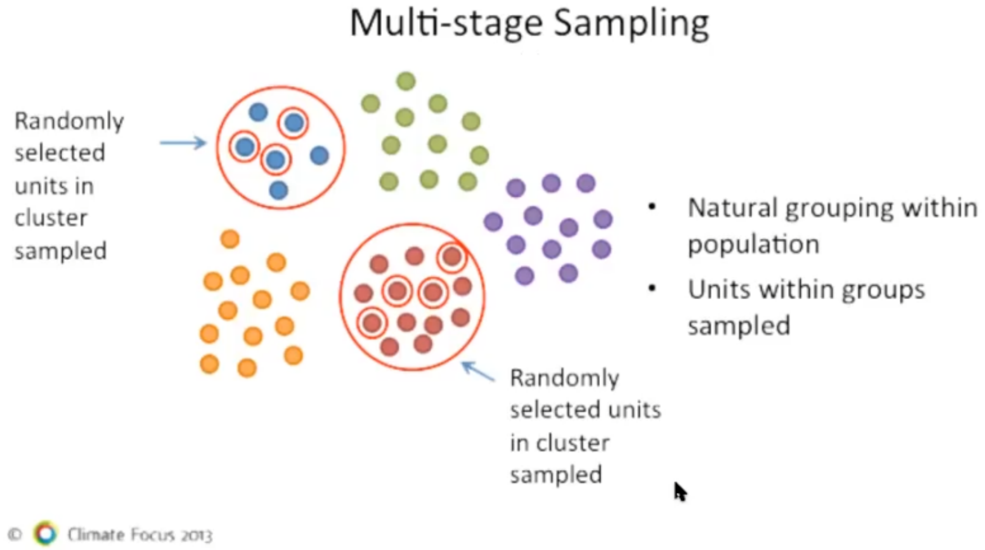
> **Parameter**: a *numerical fact* about the population which we are interested in.

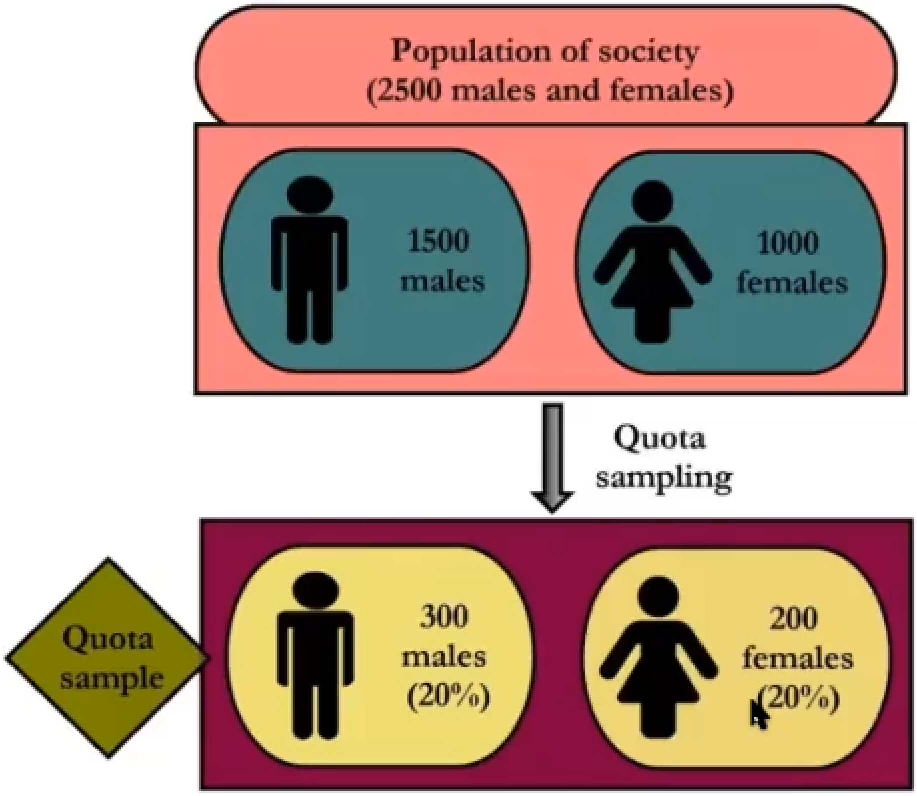
> **Estimate (statistics)**: a *calculation* of sample values which **best predicts** the parameter.

- Quote: "The **estimate** is *what the investigator knows*. Th **parameter** is *what the investigator wants to know*.

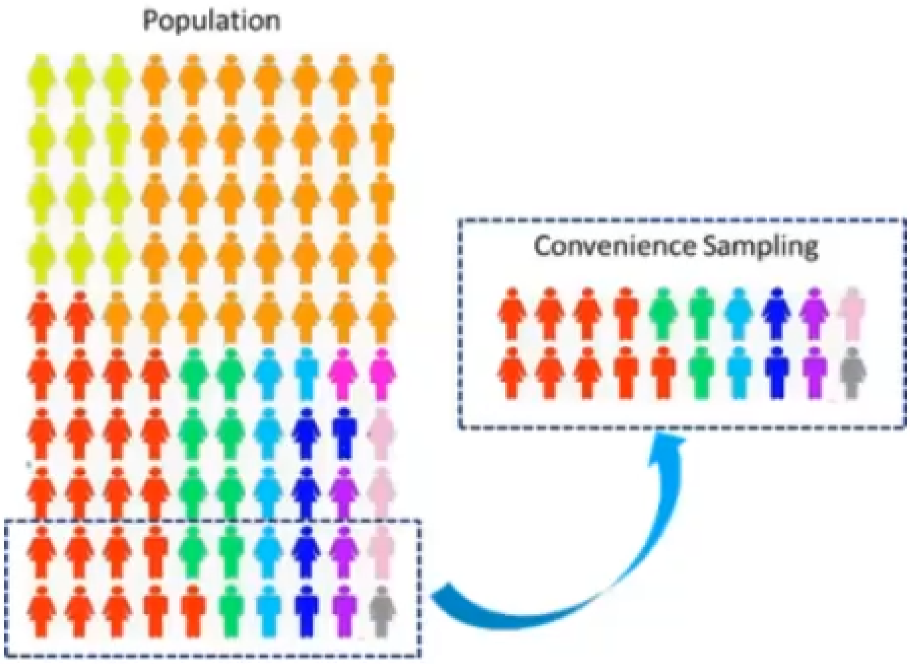
2. Sample Bias

- Common Types of Bias:
    1. **Selection Bias（选择偏差）**: A systematic tendency to *exclude or include one type of person* from the sample.
    2. **Non-response Bias（无反应偏差）**: Participants *fail to complete* surveys.
    3. **Interviewer Bias（访谈者偏差）**: A *distortion of response* related to the person questioning informants in research.
    4. **Measurement Bias**: The *form of the question* in the survey affects the response to the question.
    - Examples of measurement bias:
        - Bias in _**question wording and order**.
            - Example: Should a doctor be allowed to murder unborn children who can't defend themselves?
        - People **forget** details when **recalling**.
        - People may not tell the truth on **sensitive questions**.
            - Example: Do you use illegal drugs?
        - Question lacks **clarity**.
            - People may misinterpret on certain words or questions.
        - Attributes of the **interview process**.
            - Example: Start interviews at night.
- **Warning** about Bias and Sample Size:

    - **Taking a larger sample can amplify the bias** instead of reducing if a section process is biased, because it **repeats the mistake on a larger scale**!
- How to pick a good sample?
    - Multi-stage cluster sampling（多层整群抽样）：
        - A *probability sampling technique* which takes **samples in stages**, and individuals or clusers are **chosen at random** at each stage.
        - 

    - Quota Sampling（定额抽样）：
        - A *non-probability sampling technique* where the **assembled sample has the same proportions** of individuals **as the entire population** with respect to known characteristics, traits, or focused phenomenon.
        - This results in **unintentional bias** from the interviewers when they choose subjects to survey.
        - 

    - Convenience (Grab) Sampling（方便抽样）：
        - A *non-probability sampling technique* where subjects are selected because of their **convenient accessibility**.
        - Not recommended except testing a pilot (initial) survey.

        - 

- Unavoidable Bias:
    - Bias can happen even with a probability method determining the sample. For example: non-response bias.
    - We always have **chance error** because the sample is only part of the population.
    - Sampling & Non-sampling Error:
        - Estimate = Parameter + Bias + Chance Error
        - Estimate = Parameter + Non-sampling Error + Sampling Error
- Common Methods of Surveys:
    - Face-to-Face Interviews
    - Phone Interviews
    - Self-administered Surveys
    - Mail
    - Online

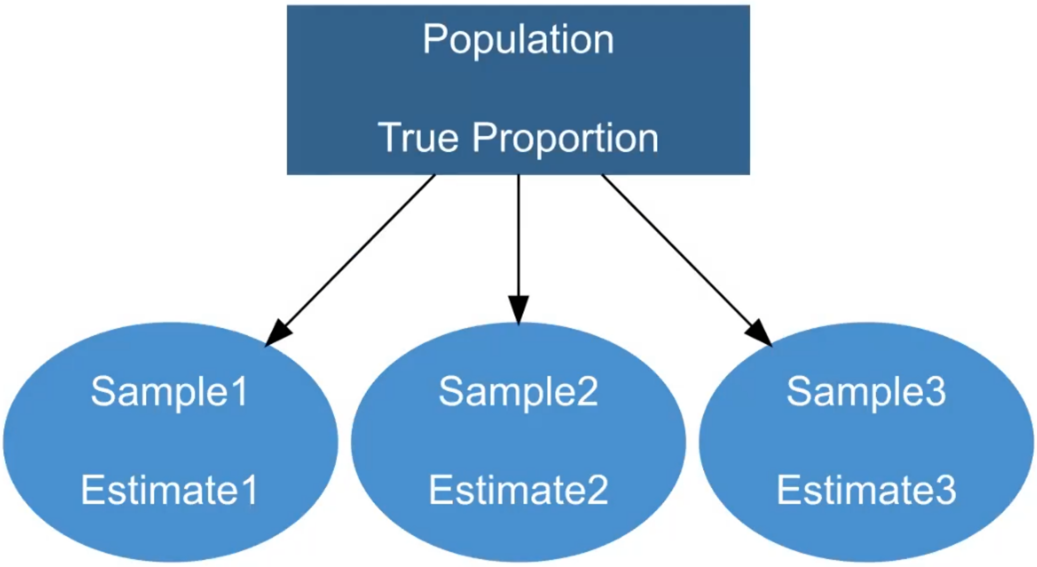
Lecture 26: The Box Model for Sample Surveys

---

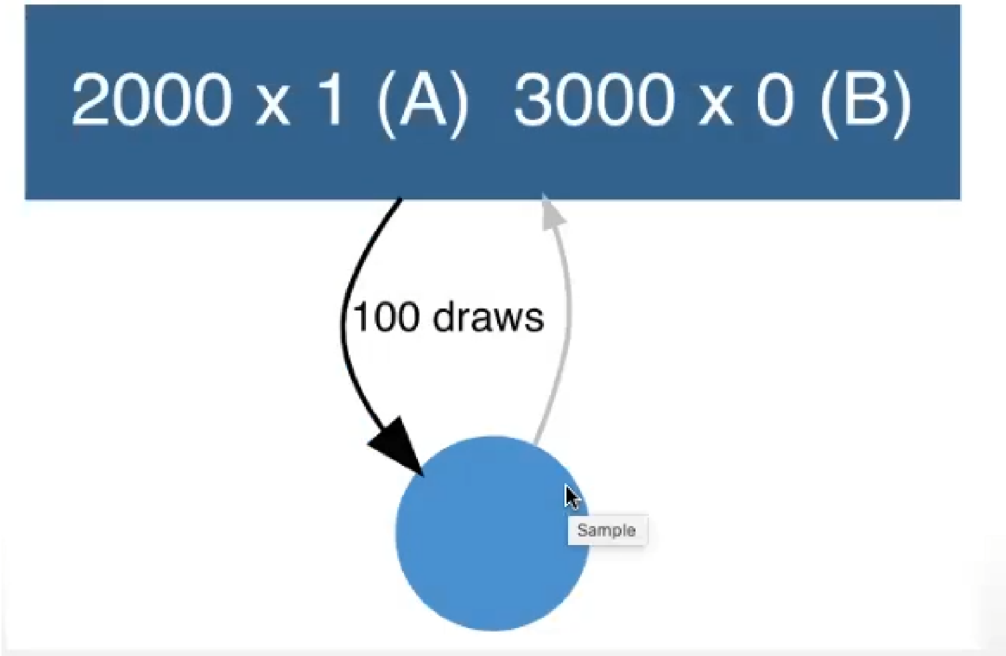
1. The Box Model: Modelling the Proportion (Mean) of a Sample

- Chance Errors in Sample Surveys:
    - We use **box model** to quantify *the likely size of the chance error* when estimating a proportion.
    - **Standard errors (SE)** measure *the variability across different samples* from the same population.

- Drawing a simple random sample:

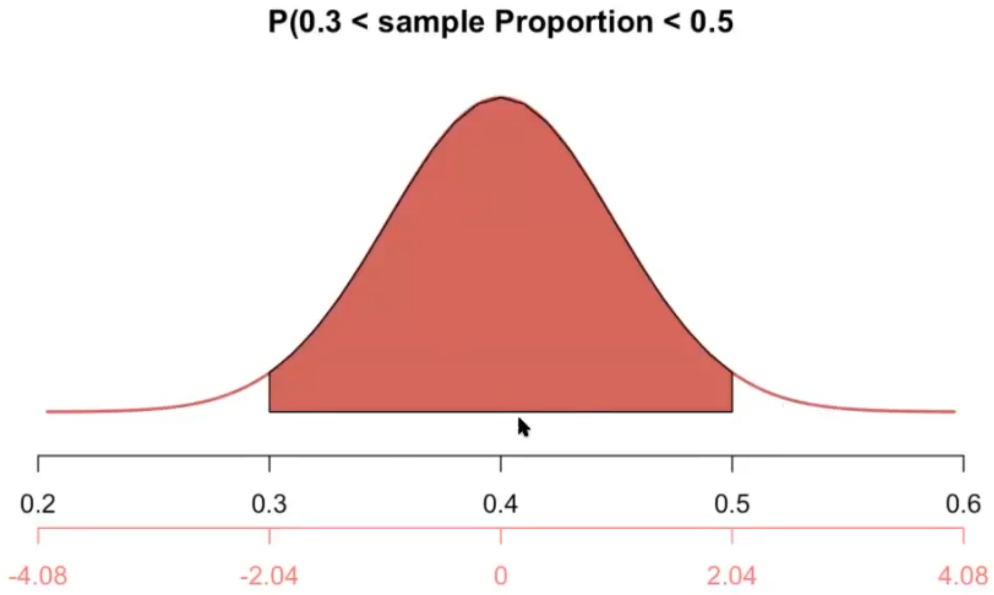


- Modelling Sampling by a Box Model:
  - Consider a Simple random sample of **100 draws from 5000 individuals**, where **2000 will vote A** and **3000 will vote B**.
  - We are interested in **the proportion of A voters**.
  - What is the chance that **the number of A voters is between 0.3 and 0.5**?
  - Steps:
    - Step 1: Draw a box model



    - Step 2: Calculate the mean and SD of the box
      - The mean is $\frac{2000\times1+3000\times0}{5000} = 0.4$.
      - The SD is $(1-0)\sqrt{2/5\times3/5} = \sqrt{6}/5 \approx 0.5$. [Note SE is rounded up here to simplify illustration.]
    - Step 3: 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 $\frac{0.5}{\sqrt{100}} = 0.05$.

- Step 4: 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).
- Step 5: Draw the normal curve



- In R:

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
```

- Step 6: Calculate the chance
  - The x values (data points) are 0.3 and 0.5.
  - Z scores (standard units) are approximately -2 and 2. (Z score = (x - EV)/SE)

- A similar solving process using **Sum**:
  - 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	sample size $\times$ mean $_{box}$	$\sqrt{\text{sample size}} \times \mathrm{SD}_{box}$
Proportion (Mean)	mean_{box}	$\frac{\mathrm{SD}_{box}}{\sqrt{\text{sample size}}}$

in R

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

Need multicon package in R

popsd( )

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

2. The Correction Factor

- What affects accuracy?
  - The SE is determined by **the absolute size of the sample** when sampling **with replacement**.
  - The SE will be decreased by increasing **the ratio of sample size to population size** when sampling **without replacement**. 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 (n) determines accuracy?

Why sample size determines accuracy

- Assume Box1 is size $\boxed{N_1}$ (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 $\text{EV}_{Proportion}$.

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

- Both boxes have the same chance error.

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

- 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:
  - **Sample surveys are drawn without replacement** so it's different to box model (drawn with replacement)!
  - We need to use **correction factor** to adjust SE from the box model to get the exact SE.
    - Correction factor (finite population correction):

Correction Factor

survey box

$$\text{SE}_{withoutreplacement} = \text{correction factor} \times \text{SE}_{withreplacement}$$

where

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

is equivalent to

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

  - if the population is a lot bigger than the sample, CF is almost 1.
    - Example:

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

Lecture 27: Bootstrapping & Confidence Intervals (Accuracy of Proportions)

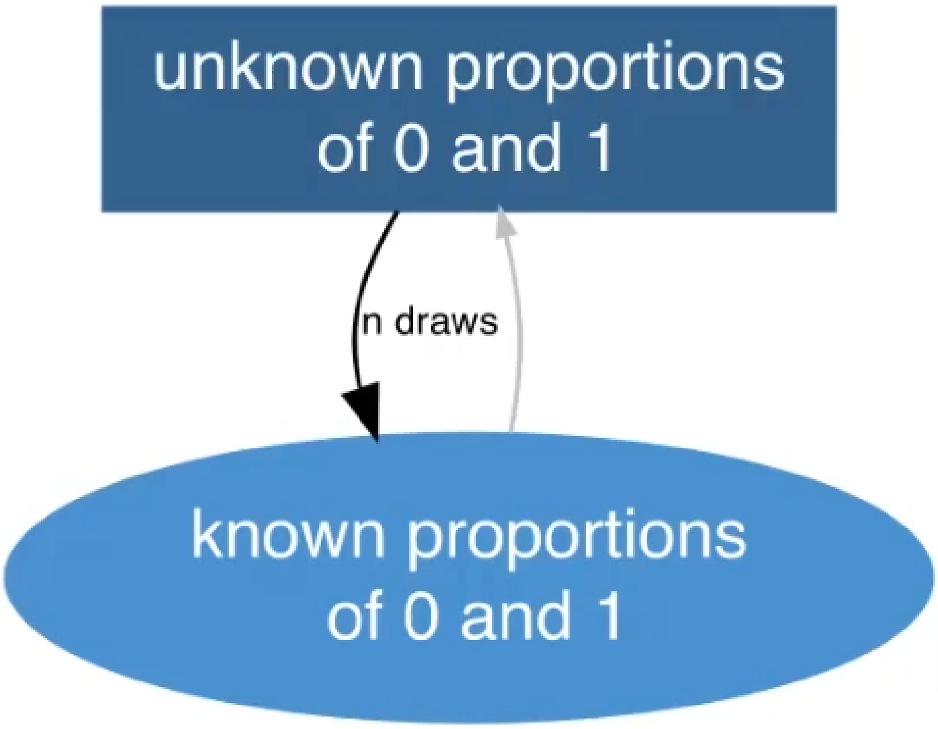
---

1. Estimating the Population proportion Using Bootstrapping

- The gap in information:
  - Previously we found:
    - The EV of the sample proportion is equal to the population proportion.

      $$EV_{proportion} = \text{mean}_{box} = \text{population proportion}$$
    - The chance error is related to the SE of the sample proportion.

      $$SE_{proportion} = \frac{SD_{box}}{\sqrt{\text{sample size}}}$$
    - However, the mean and SD of the box is unknown. The formulas above are useless in this case.



> Bootstrapping: **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.
  - Steps:
    - Step 1: Create an approximate box (we don't know the real box!) -- box 1 which has the same proportion of 0s and 1s as the sample.

- Step 2: 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}	$\frac{\text{SD}_{box1}}{\sqrt{\text{sample size}}}$

2. Confidence Interval

- Chance Error and Standard Error:
  - We have often taken the estimate of the chance error to be 1 unit of the SE.
  - The chance error, however, can be out by 2 or even 3 SEs. We can use **confidence intervals (CI)** to generalize.
- Confidence Intervals (CI)（置信区间）：
  - 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

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

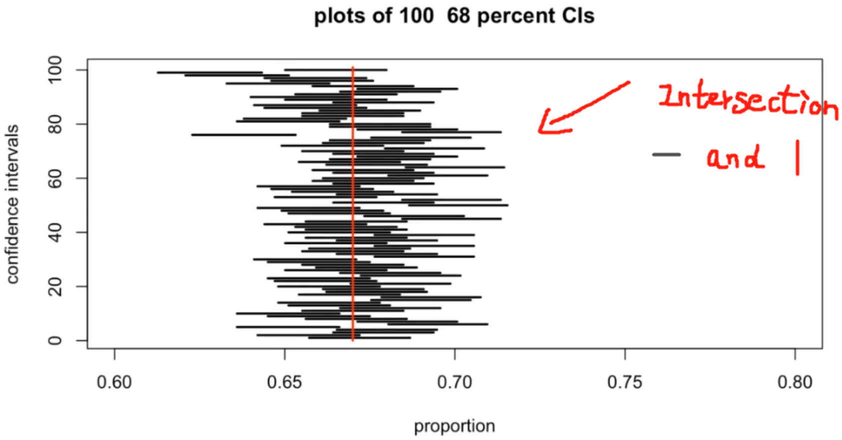
  - Interpreting CI:
    - For a 95% CI:
      - It is wrong to say "the probability that the interval contains the unknown parameter is 0.95." (cannot make conclusion on only 1 CI)
      - Rather, we say "**if we worked out a series of CIs for a series of samples, then 95% of the CIs would contain the unknown parameter**.

- Simulation:

  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
```