Estimating a Population Proportion with a Confidence Interval

Example – What is the proportion of ALL Canadians support legalizing marijuana?

We never know this number because it is almost impossible to collect the opinion from ALL Canadians.

Confidence Interval for a Population Proportion

A random sample of 200 Canadians is selected to estimate this proportion.

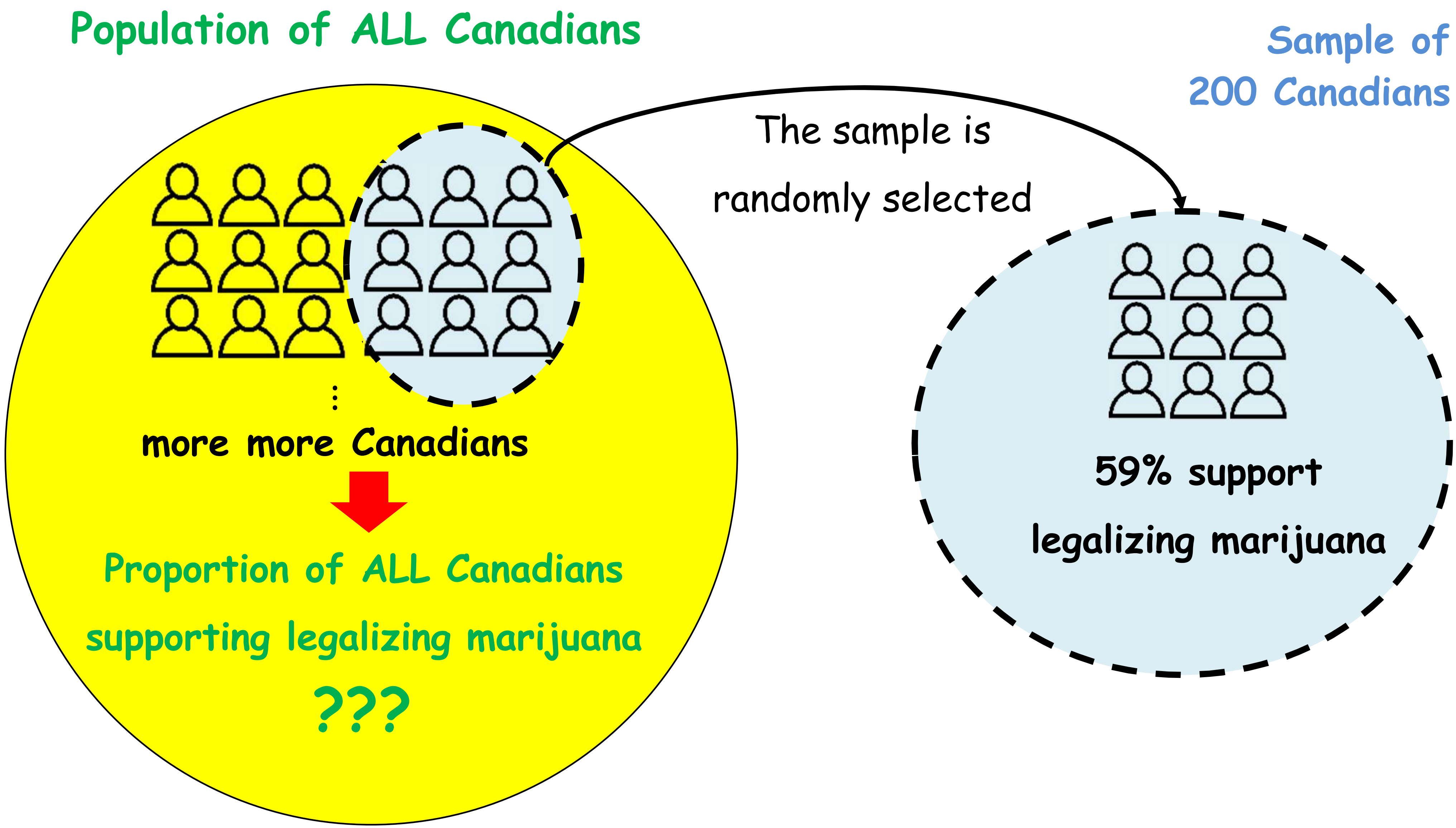
The opinions of the 200 Canadians are collected and saved in the file below:

http://mylinux.langara.bc.ca/~sli/Marijuana.csv

Question:

What percentage of the 200 Canadians support legalizing marijuana?

The objective of using the sample data is to estimate the **proportion of ALL Canadians** who support legalizing marijuana.



The simplest way to estimate **a population proportion (p)** is to use **a sample proportion ($\bar{p}$)**.

$$\bar{p} \xrightarrow{\textbf{Estimate}} p$$

Back to our example,

- In the **sample of 200 Canadians**, **59%** support legalizing marijuana.
- So, we estimate that about **59%** of **ALL Canadians** support legalizing marijuana

- However, it **DOES NOT** mean that **exactly** **59%** of **ALL Canadians** support legalizing marijuana
- It is because the **sample proportion** **"59%"**
  - is computed from the **sample of 200 students**,
  - **NOT** the **entire population of ALL Canadians**.
- Therefore, we prefer an **interval estimate**.
- What is an **Interval Estimate**?
- Here is the simple analogy.

How tall is this girl?

126 cm

- **Question:** Are you sure that she is **exactly** 126 cm?
- The answer is **NO** because
- there are many possible values such as 125.6 or 126.2 etc
- Please remember:
- All measurements contain some **uncertainty**.

In our example, the ruler's precision is about 1 cm.

126 cm

The uncertainty is about 1cm

126 cm

The uncertainty is about 1cm

Now, we take

$$\underbrace{\text{Estimate}}_{126} \pm \underbrace{\text{Uncertainty}}_{1} = \underbrace{\text{Interval Estimate}}_{(125,\ 127)}$$

An interval estimate gives a range of possible values of an unknown quantity (e.g. height of the girl)

- Similarly, when we use a sample proportion to estimate a population proportion,
- the estimate contains some uncertainty
- which can be quantified by the Margin of Error.
- Once we have the sample proportion and its margin of error (or uncertainty),
- we can take

$$\text{Sample proportion} \pm \text{Margin of Error} = \text{Interval Estimate}$$

- to construct an Interval Estimate
- that gives a range of possible values of a population proportion
- The Interval Estimate is formally called Confidence Interval.

The confidence interval for a population proportion is defined as

Sample proportion

$$\underbrace{\bar{p}}_{\text{Sample proportion}} \pm \underbrace{\underbrace{z_c}_{\text{critical value}} * \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}}_{\text{Margin of Error}}$$

Sample Size

Example - In a random sample of 200 Canadians, 59 percent support legalizing marijuana

Question – Estimate the **proportion of all Canadians** who support legalizing marijuana using a **95% confidence interval**.

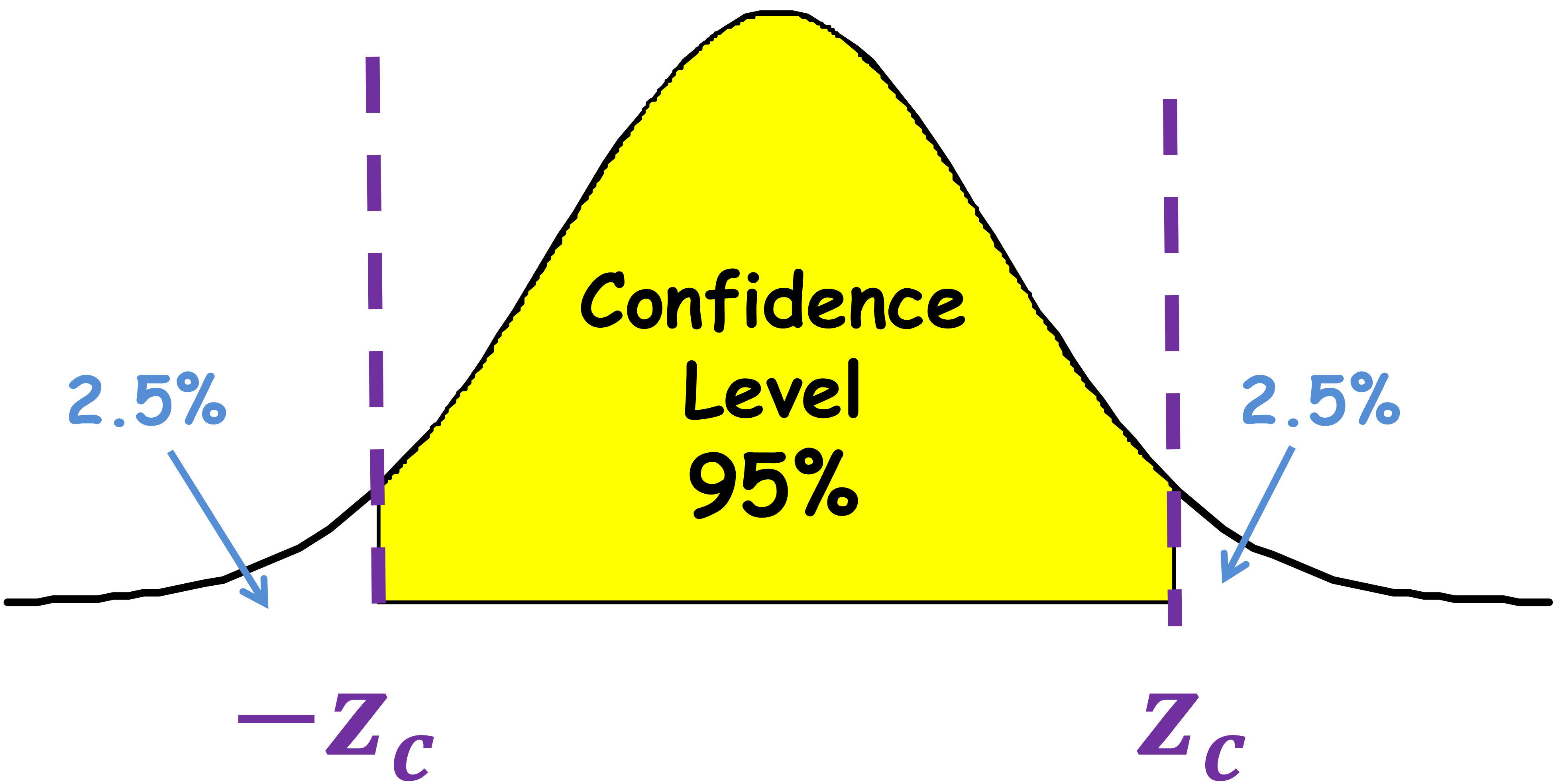
Sample proportion $= 0.59$

$$\bar{p} \pm \underbrace{z_c}_{\text{critical value} = ???} * \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

0.59
Sample proportion

Sample Size $= 200$

- To determine the z-critical value, first,
- we draw a **normal curve** and
- shade the area in the middle of the distribution.
- The size of the middle area is given by the confidence level (95% or 0.95).
- The **critical value** is the **z-score** (denoted by z_c) such that the middle area bounded by z_c is 0.95.



Area = 2.5% $\rightarrow$ 0.025

z	***0.05***	***0.06***	***0.07***	***0.08***
-1.8	0.032	0.0[illegible]1	0.031	0.030
-1.9		0.025	0.024	0.024

$$z = -1.96 \rightarrow \boxed{z_c = 1.96}$$

Sample proportion $= 0.59$

0.59 Sample proportion

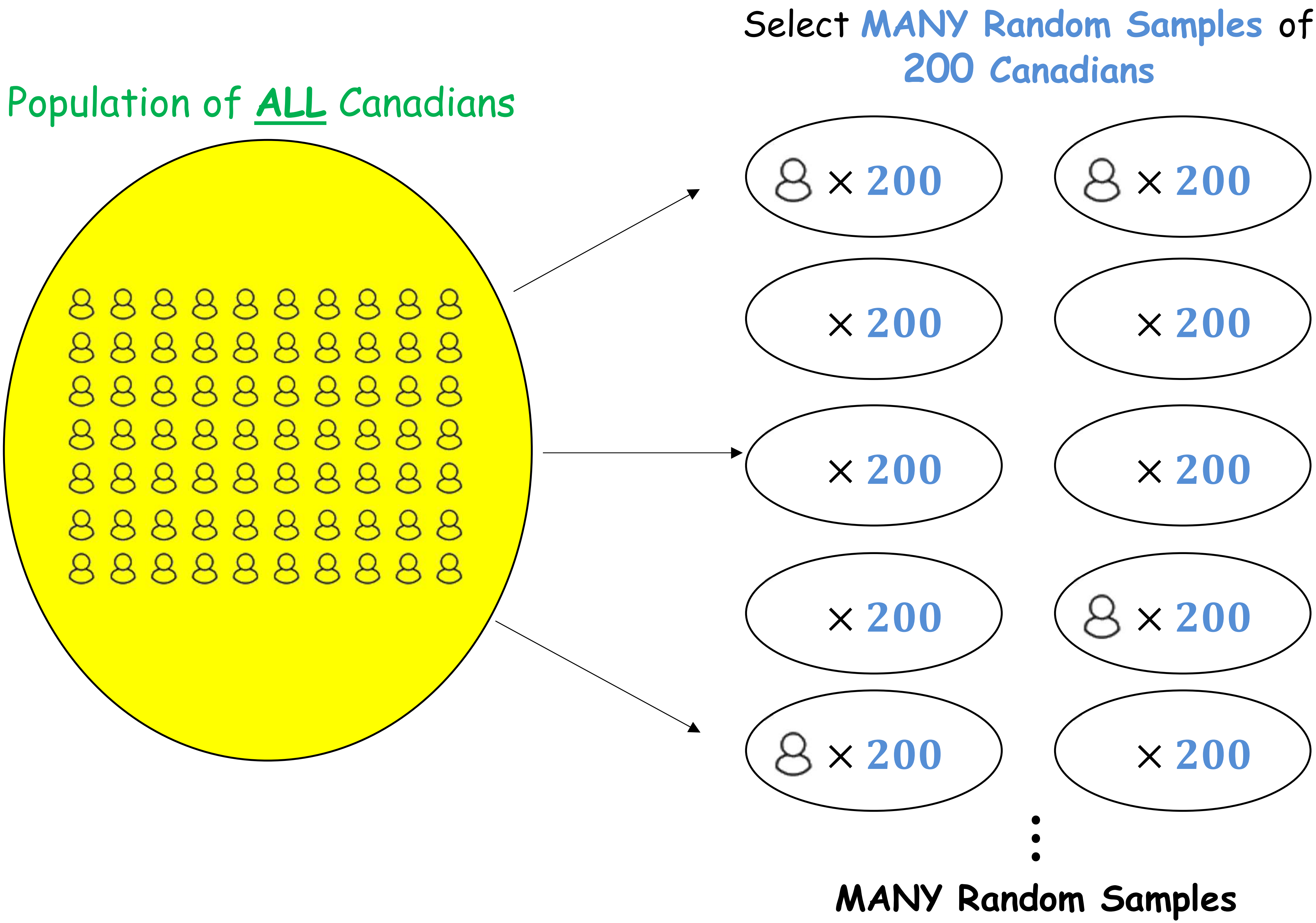
$$\bar{p} \pm \underbrace{z_c}_{\text{critical value} = 1.96} * \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

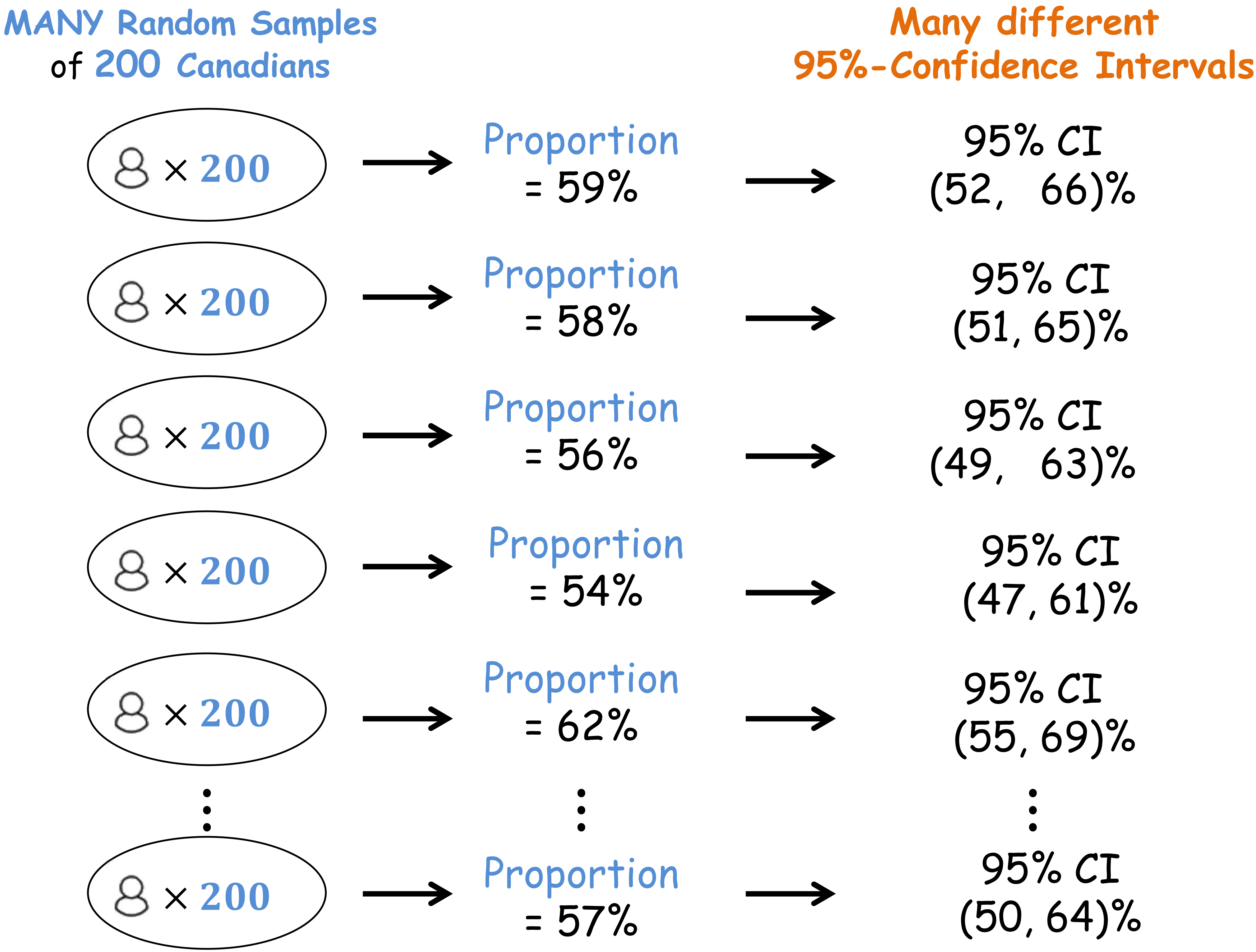
Sample Size $= 200$

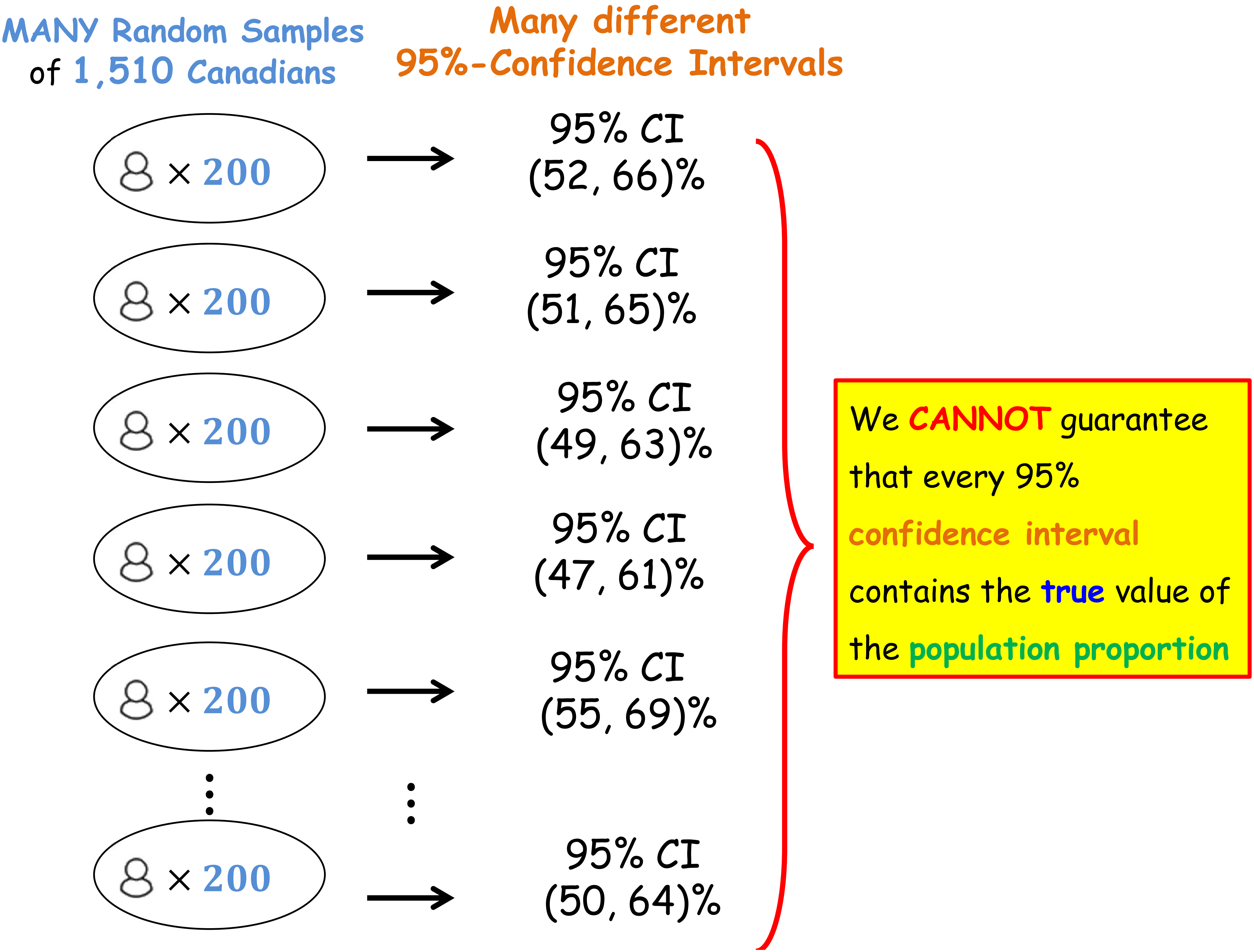
Interpret the Confidence Interval

- Although we don't know the **true value of the proportion of all Canadians** who support legalizing marijuana,
- from sample data, we are **95% confident**

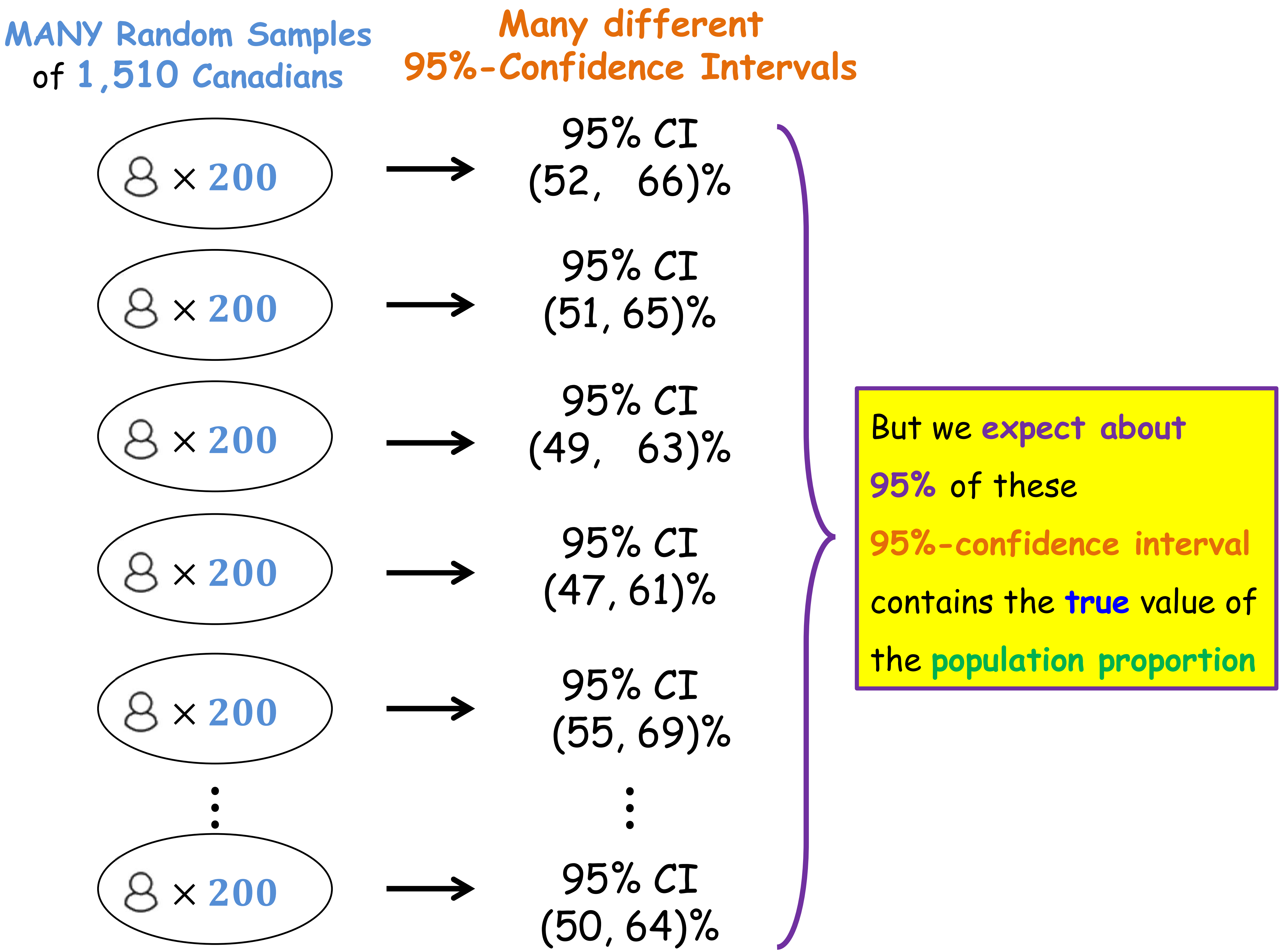
Interpreting Confidence Level (95%)




MANY Random Samples of 200 Canadians
Many different 95%-Confidence Intervals
× 200
Proportion = 59%
95% CI (52, 66)%
× 200
Proportion = 58%
95% CI (51, 65)%
× 200
Proportion = 56%
95% CI (49, 63)%
× 200
Proportion = 54%
95% CI (47, 61)%
× 200
Proportion = 62%
95% CI (55, 69)%
× 200
Proportion = 57%
95% CI (50, 64)%



MANY Random Samples of 1,510 Canadians
Many different 95%-Confidence Intervals
× 200
95% CI (52, 66)%
× 200
95% CI (51, 65)%
× 200
95% CI (49, 63)%
× 200
95% CI (47, 61)%
× 200
95% CI (55, 69)%
× 200
95% CI (50, 64)%
We CANNOT guarantee that every 95% confidence interval contains the true value of the population proportion


MANY Random Samples
of 1,510 Canadians
Many different
95%-Confidence Intervals
× 200
95% CI
(52, 66)%
× 200
95% CI
(51, 65)%
× 200
95% CI
(49, 63)%
× 200
95% CI
(47, 61)%
× 200
95% CI
(55, 69)%
× 200
95% CI
(50, 64)%
But we expect about
95% of these
95%-confidence interval
contains the true value of
the population proportion

In short,

- although we are **not** sure whether between **52% and 66%** of **ALL Canadians** who support legalizing marijuana
- at least we use a method to estimate a population proportion that gives the **correct results about 95% of times.**

Assumptions / Conditions Required for valid Confidence Interval for a Population Proportion

- First, **not all datasets can be used to estimate a population proportion with a confidence interval**.
- The data must satisfy certain conditions; otherwise, any conclusions drawn from the confidence interval will be **invalid**.
- To obtain **valid** conclusions from a confidence interval for a population proportion, the sample data must meet specific conditions.
- **What are these required conditions?**

Technically, it requires that the sample is sufficiently large. When we construct a confidence interval, the actual number of individuals in a random sample who

- fall into the category of interest and
- do not fall into the category of interest

are both at least 5.

In a random sample of 200 Canadians, there are

- 59% of 200 Canadians support legalizing marijuana

	Actual Number of Canadians	
Support		
Do not support		

Since both actual numbers are at least 5, the sample is sufficiently large.

Since large-sample condition is satisfied,

any conclusion drawn from the confidence interval are valid and can be trusted.