Recitation, Week 3

Ye Wang

New York University

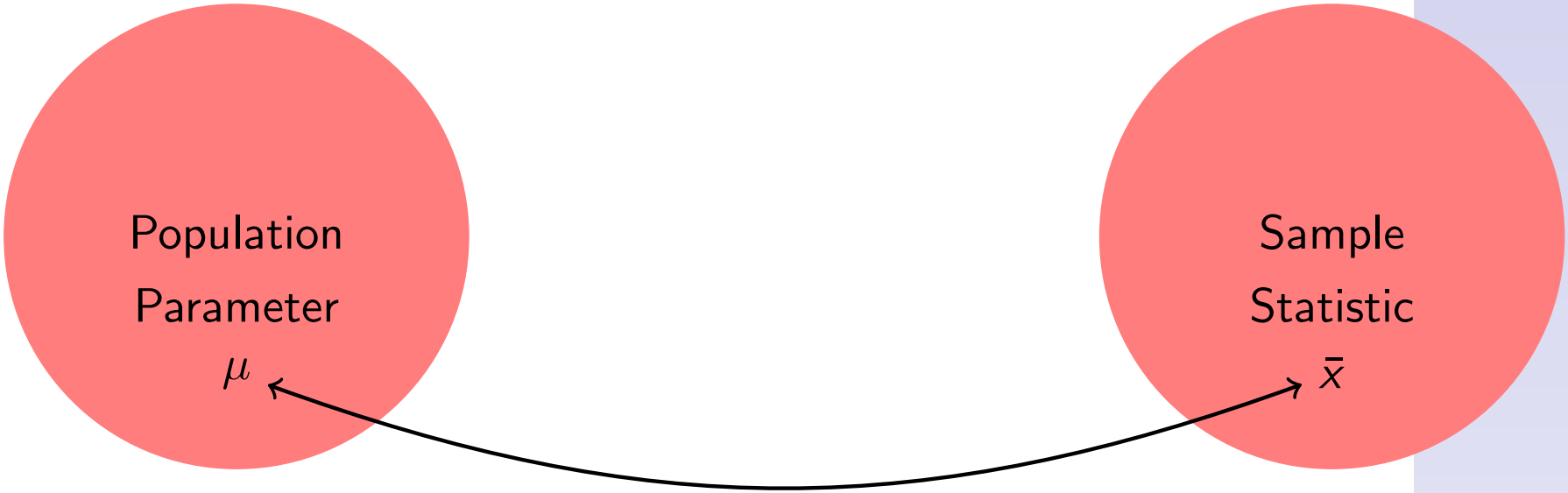
POL-850

Spring 2018

Outline

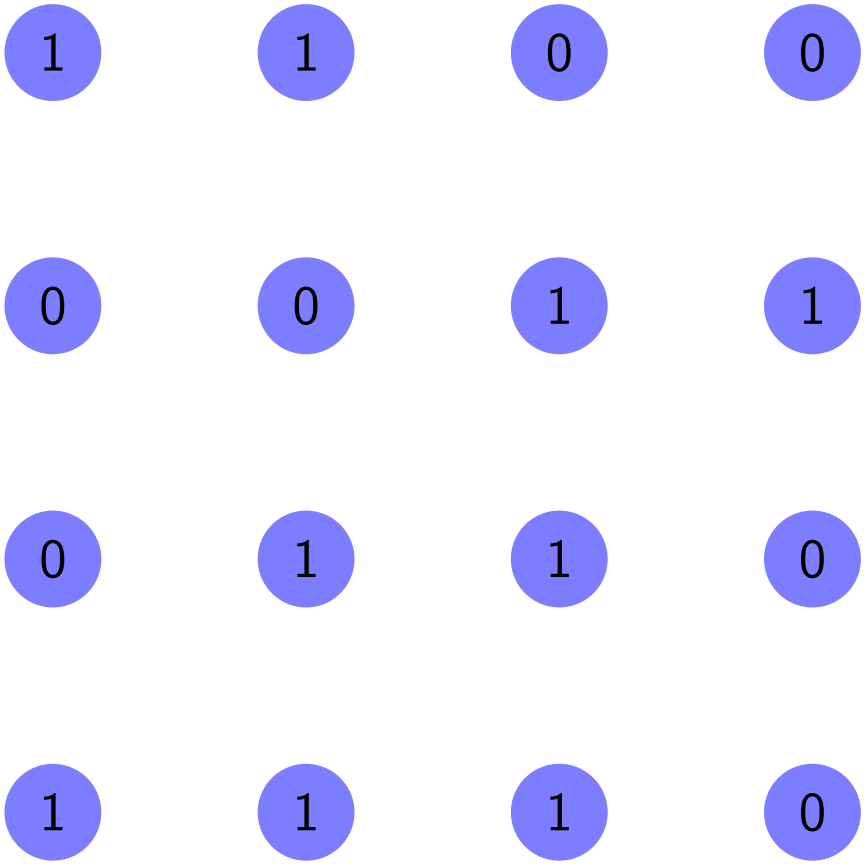
(1) Sampling
(2) Causality
(3) STATA session

What is sampling?



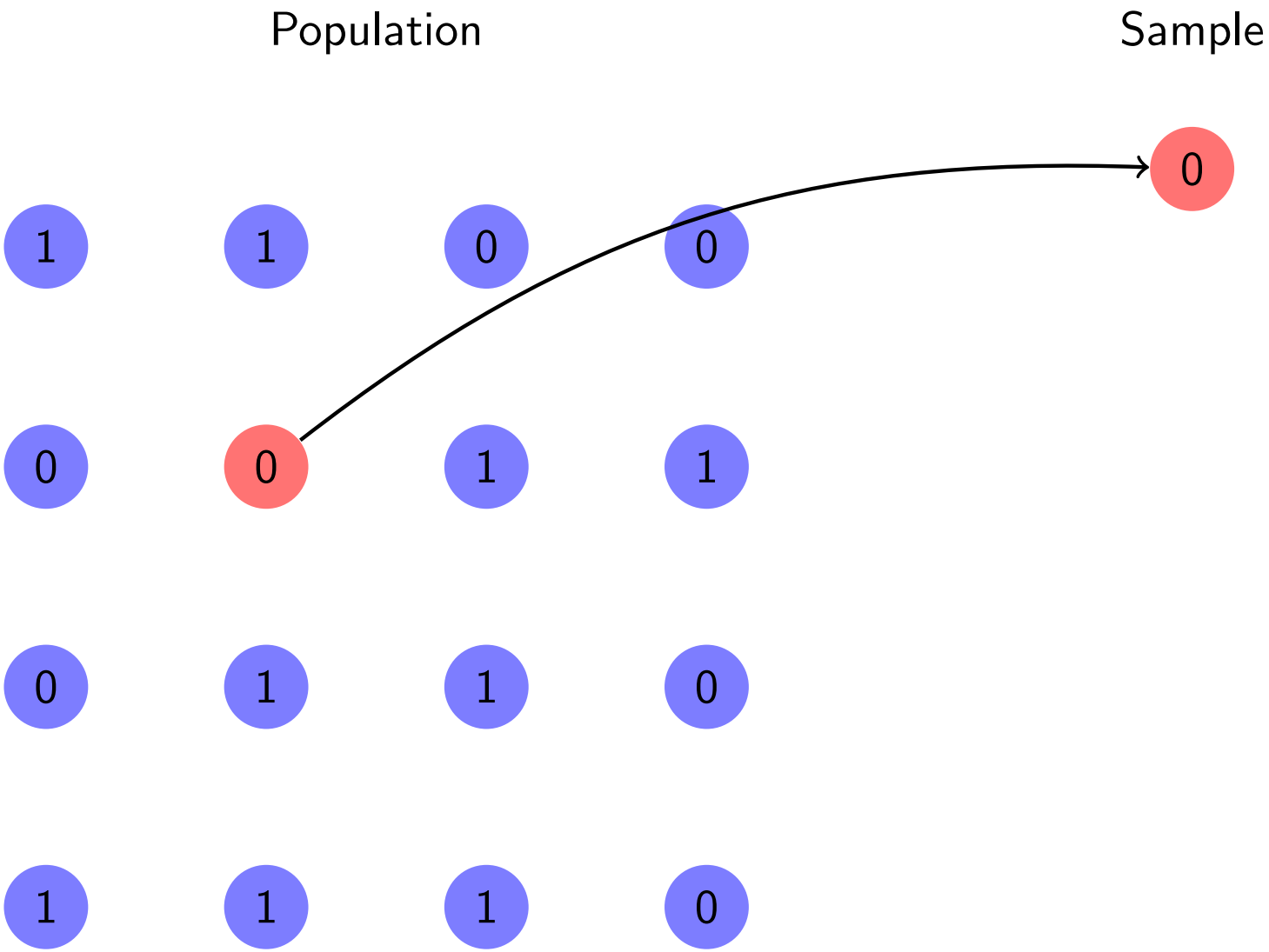
Random Sampling Error

Population

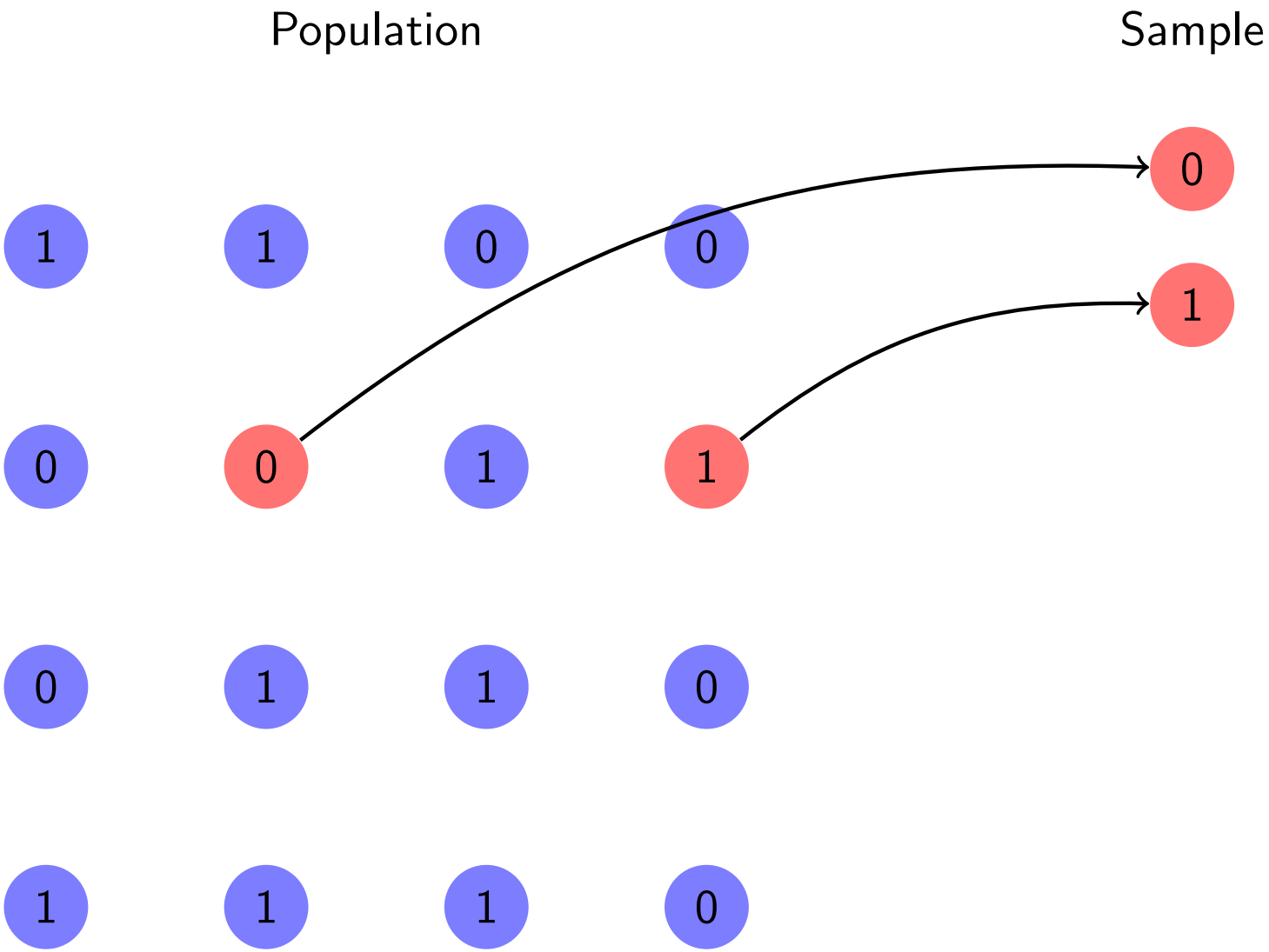


Sample

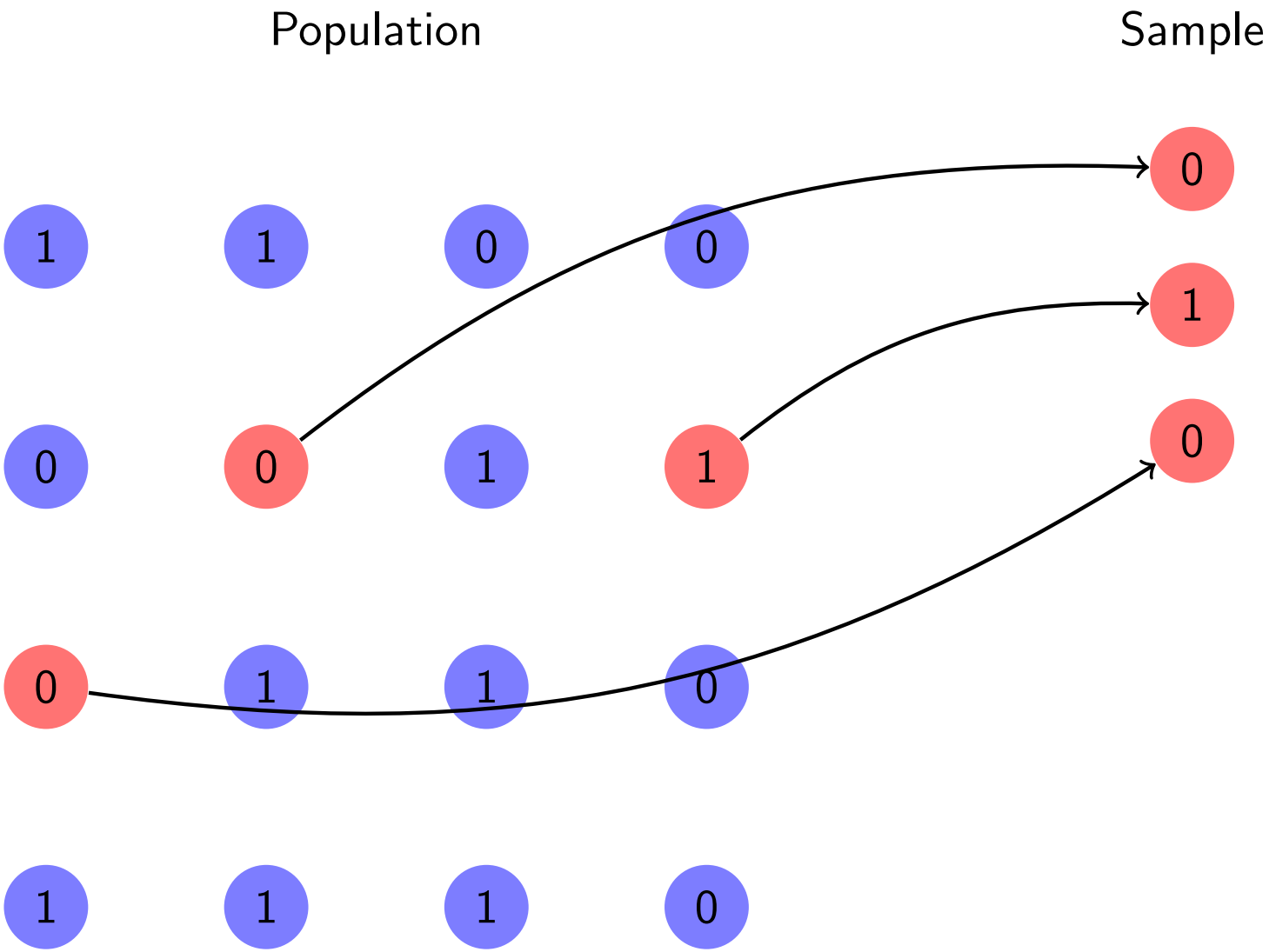
Random Sampling Error



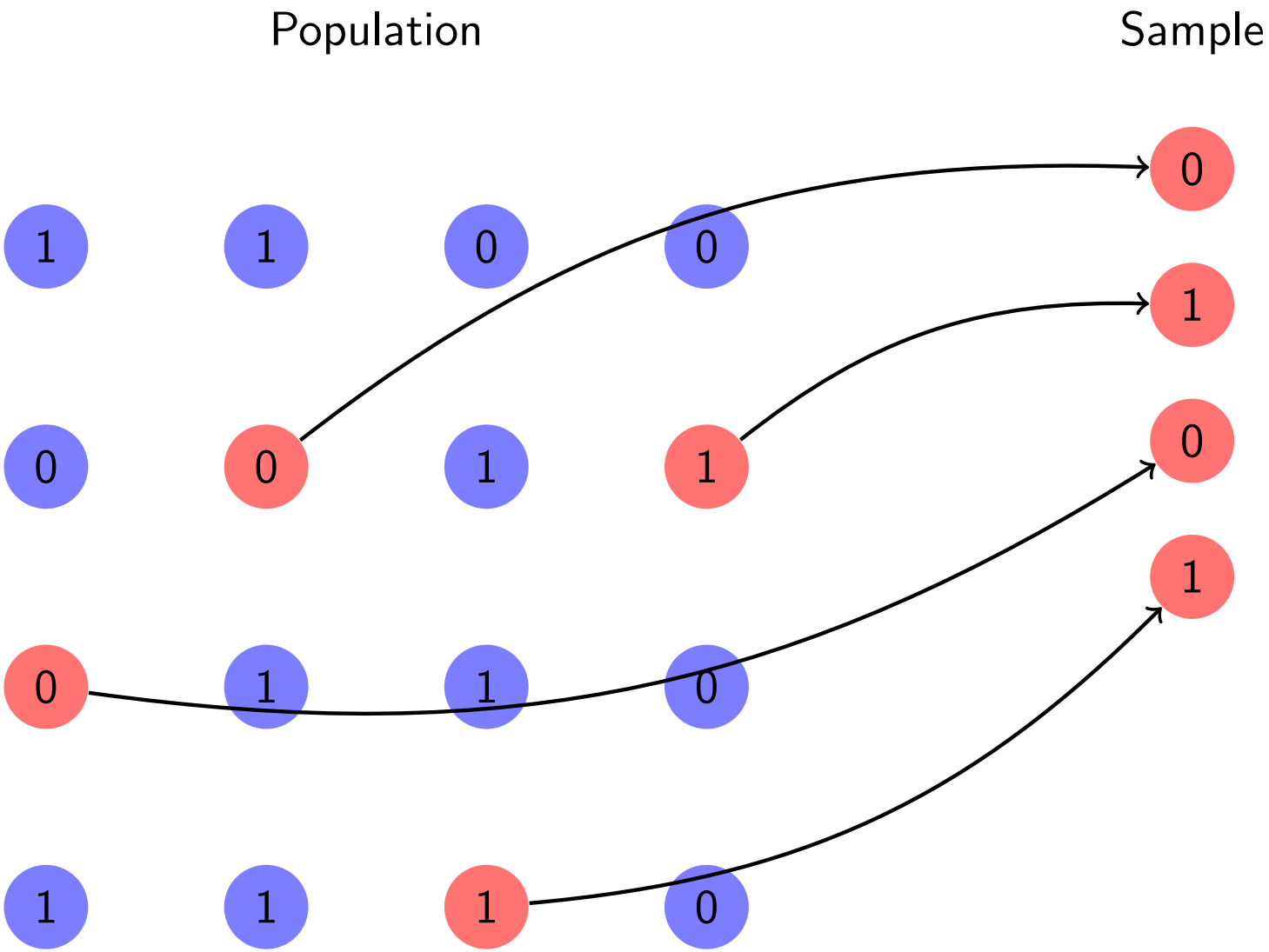
Random Sampling Error



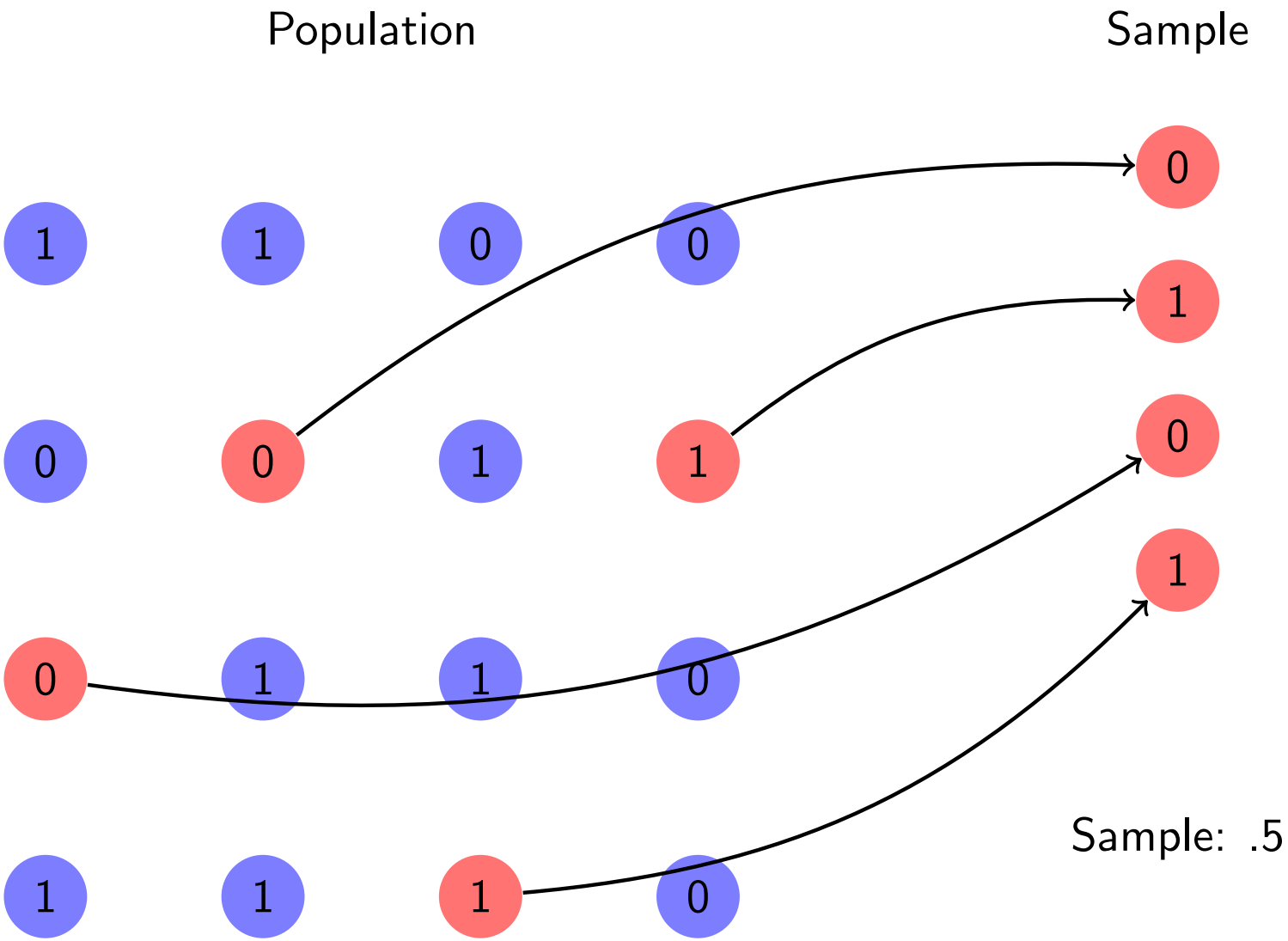
Random Sampling Error



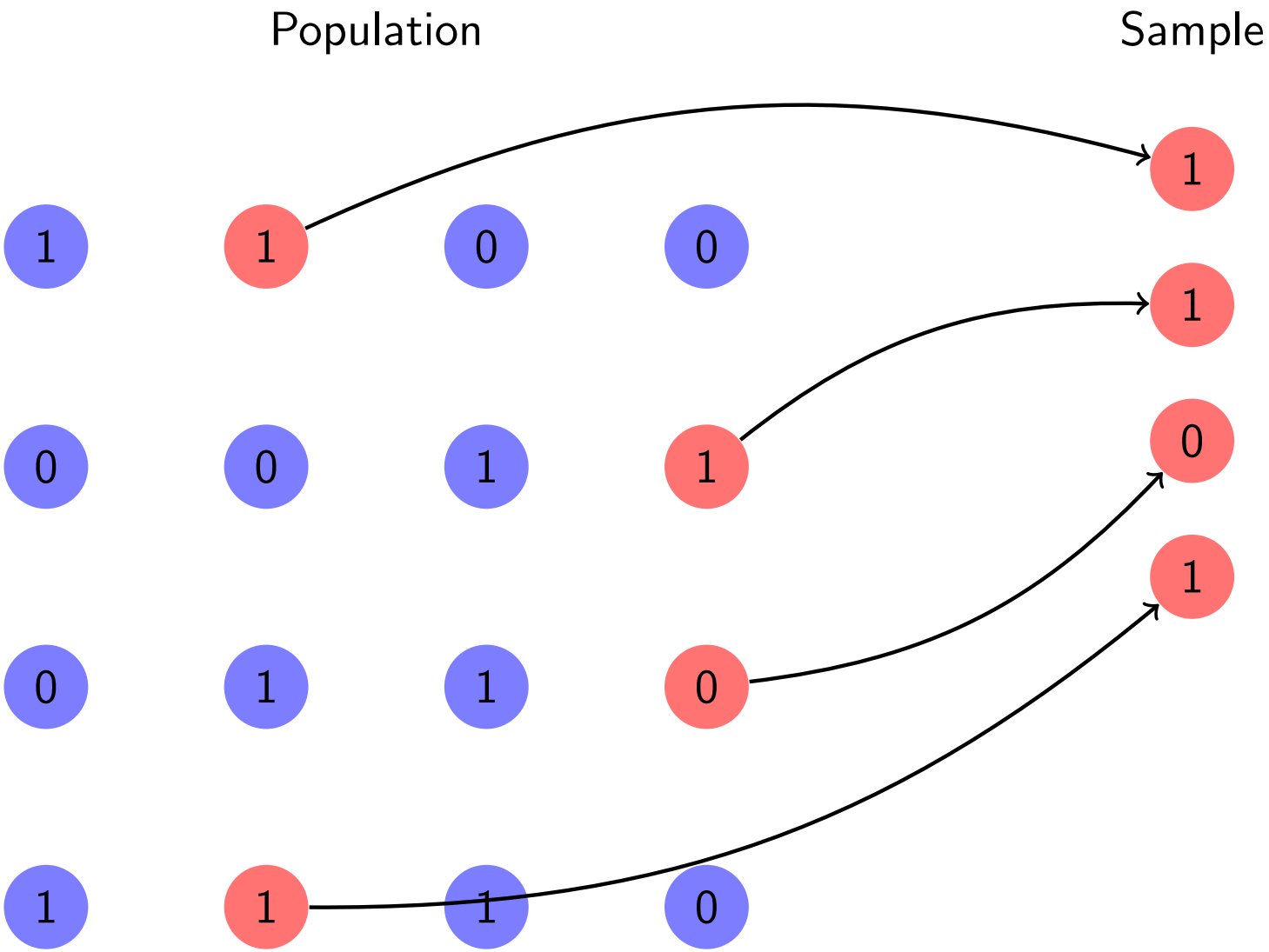
Random Sampling Error



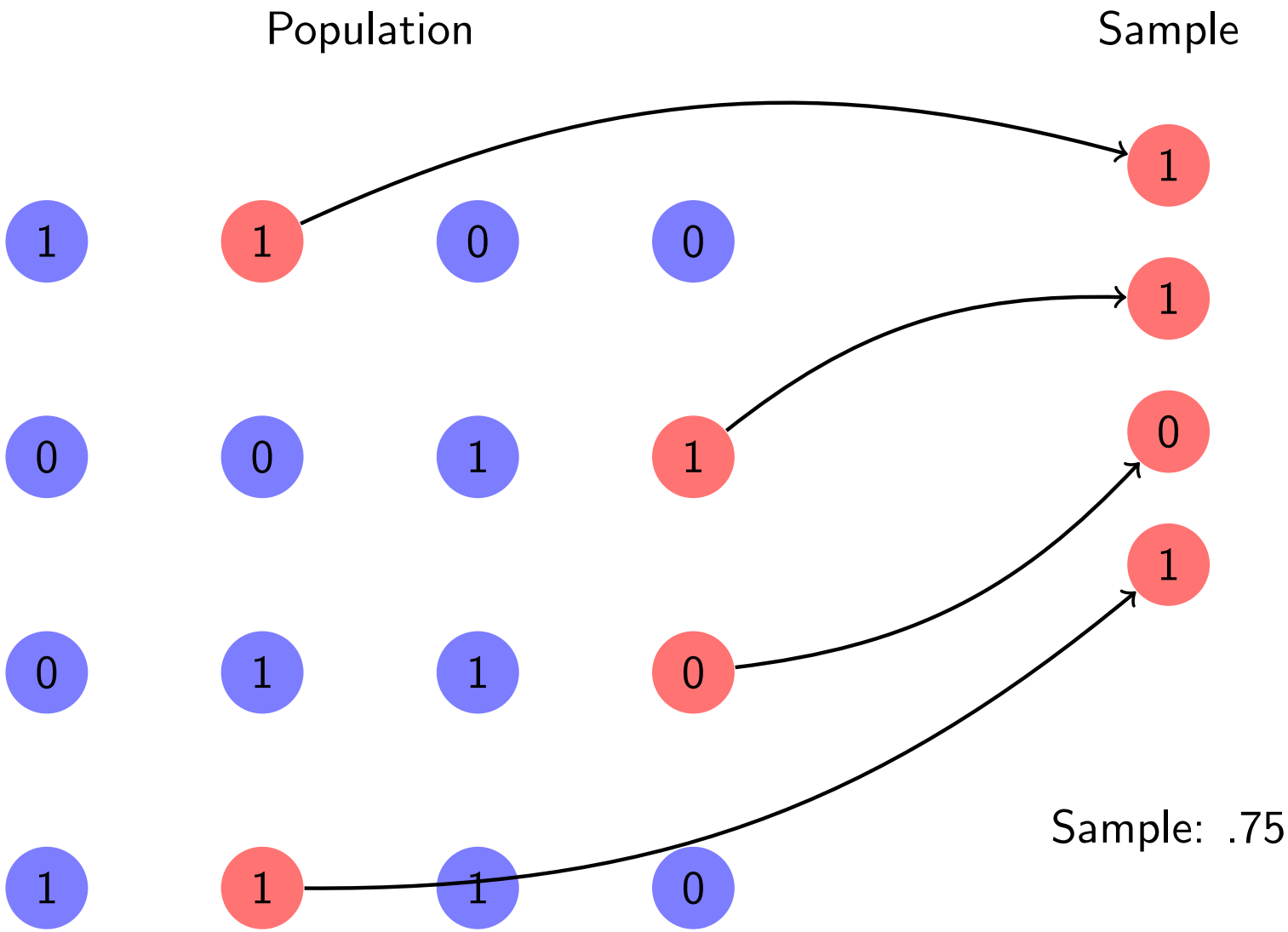
Random Sampling Error



Random Sampling Error



Random Sampling Error



Why does sampling work?

- Intuitively, sampling is random, thus the sample should be representative of the population (like a miniature)

Why does sampling work?

- Intuitively, sampling is random, thus the sample should be representative of the population (like a miniature)
- Mathematically, we have the law of large numbers:

Why does sampling work?

- Intuitively, sampling is random, thus the sample should be representative of the population (like a miniature)
- Mathematically, we have the law of large numbers:
- *As the sample size N goes to infinity, the sample average will converge to the population mean*

Why does sampling work?

- Intuitively, sampling is random, thus the sample should be representative of the population (like a miniature)
- Mathematically, we have the law of large numbers:
- *As the sample size N goes to infinity, the sample average will converge to the population mean*
- One of the most power theorems in statistics (first proved by Jakob Bernoulli)

Random Sampling Error

Note that we're interested in estimating a population parameter:

$$\text{Population Parameter} = \quad +$$

Random Sampling Error

Note that we're interested in estimating a population parameter:

$$\text{Population Parameter} = \text{Sample Statistic} +$$

Random Sampling Error

Note that we're interested in estimating a population parameter:

Population Parameter = Sample Statistic + Random Sampling Error

Random Sampling Error

$$\text{Random Sampling Error} = \frac{\text{component}}{\text{component}}$$

Random Sampling Error

$$\text{Random Sampling Error} = \frac{\text{variation in population component}}{\text{component}}$$

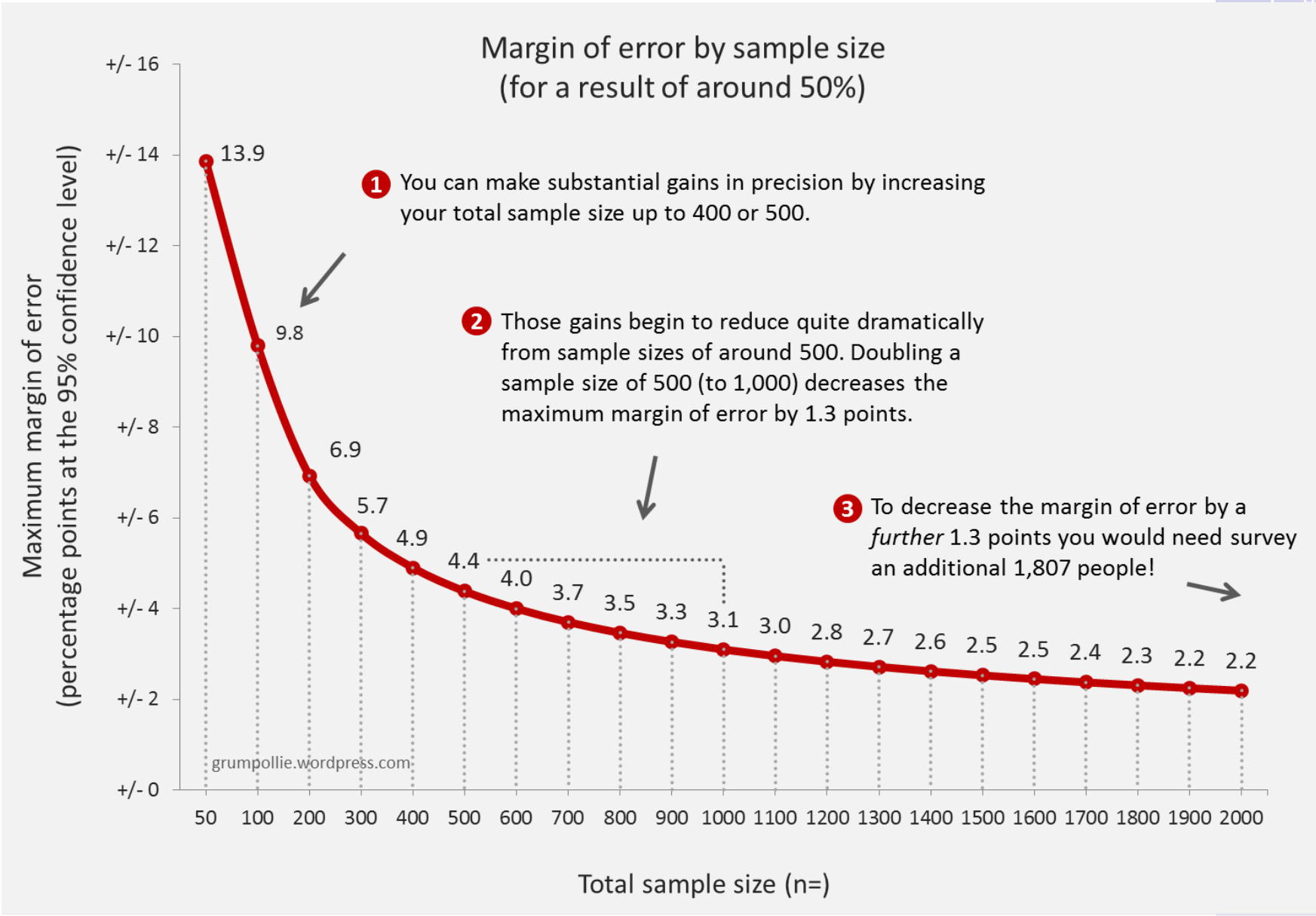
Has nothing to do with the population size!

Random Sampling Error

$$\text{Random Sampling Error} = \frac{\text{variation in population component}}{\text{sample size component}}$$

Has nothing to do with the population size!

Sample Size Improvement $= \sqrt{n}$



Group Exercise

- Suppose an enterprising politics student filled up two bathtubs with 1,000 marbles each. In bathtub 1, she used 1 red marble and 999 blue marbles. In bathtub 2, she used 500 of each color. If she took a sample of 50 marbles from both bathtubs, which would have less random sampling error? Why?

500 red, 500 blue	1 red, 999 blue

Concepts you need to know

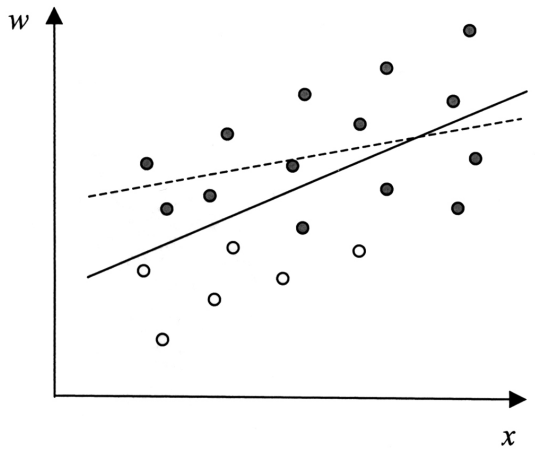
- Population
- Population parameter (true value; but unknown)
- Sample
- Sample statistic
- Estimate (statistical inference)

Problems in sampling

- Selection bias (sampling bias)

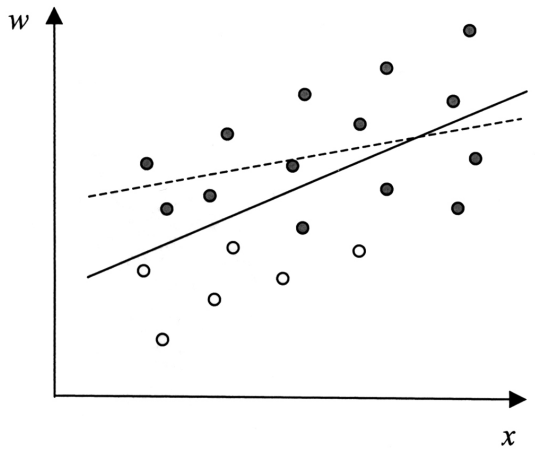
Problems in sampling

- Selection bias (sampling bias)



Problems in sampling

- Selection bias (sampling bias)



- Response bias
- Sampling frame

Sampling in the modern era

- Random sampling can still be expensive and not representative

Sampling in the modern era

- Random sampling can still be expensive and not representative
- Stratified sampling

Sampling in the modern era

- Random sampling can still be expensive and not representative
- Stratified sampling
- How to sample HIV carriers?

Sampling in the modern era

- Random sampling can still be expensive and not representative
- Stratified sampling
- How to sample HIV carriers? Snowball sampling

Sampling in the modern era

- Random sampling can still be expensive and not representative
- Stratified sampling
- How to sample HIV carriers? Snowball sampling
- Big Data: Why do we still need sampling?

CAUSALITY

Causality

Why do we care about it?

Causality

Why do we care about it?

- Social scientists care about causality
- Correlation (association) does not necessarily mean causation...
- But establishing causality is extremely difficult...

Causality

Suppose we see high correlation between X and Y and we conclude X causes Y. What's wrong with this reasoning?

Causality

Suppose we see high correlation between X and Y and we conclude X causes Y. What's wrong with this reasoning?

By and large, there could be two causes for a spurious relationship:

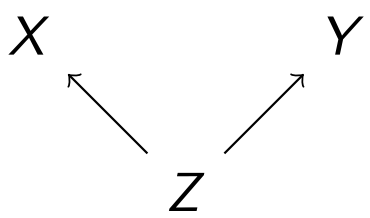
- Reverse causality
- **Confounders (confounding factors)**; which often unobserved

Causality

Suppose we see high correlation between X and Y and we conclude X causes Y. What's wrong with this reasoning?

By and large, there could be two causes for a spurious relationship:

- Reverse causality
- **Confounders (confounding factors)**; which often unobserved



Z is a confounding variable.

Group exercise

Discuss what could be a confounder which causes a spurious relationship between X and Y

- Education level and wage
  (at individual level)
- Average temperature and GDP per capita
  (at country level)
- Campaign spending and electoral victory
  (at the congressional district level)

How to Identify Causality?

- Ideally, we need a time machine
- "The fundamental problem of causal inference" (Holland, 1986)
- We have to rely on assumptions
- Scientific solutions vs. Statistical solutions

Stata Exercise

- Find out the list of people who are older than Cate
- Find out the list of people whose income is greater than $10,000
- Find out the information about Jake (without using the browser)
- Find out the mean of income
- Find out the minimum value of age