D1EAD – Análise Estatística para Ciência de Dados 2021.1

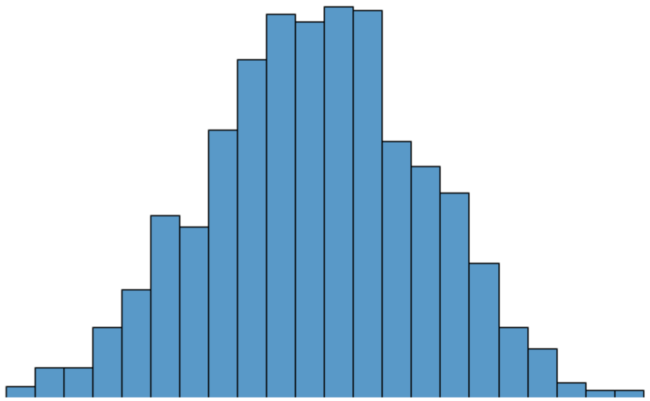
Data Distributions (Part 4)

Prof. Ricardo Sovat
sovat@ifsp.edu.br
Prof. Samuel Martins (Samuka)
samuel.martins@ifsp.edu.br

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas

Problems with Traditional Confidence Intervals

- Assumptions about **distribution** or **sample size:**
  - Normal distribution
  - Sample size is **large enough (central limit theorem)**
    - What is a large sample for a specific problem?
  - **Population standard deviation σ is known**
    - Otherwise, we **approximate** it from the **sample standard deviation s**
  - Calculating the **standard error** for some statistcs can be **difficult**
    - E.x: Estimate the range between the 80th to 90th percentiles



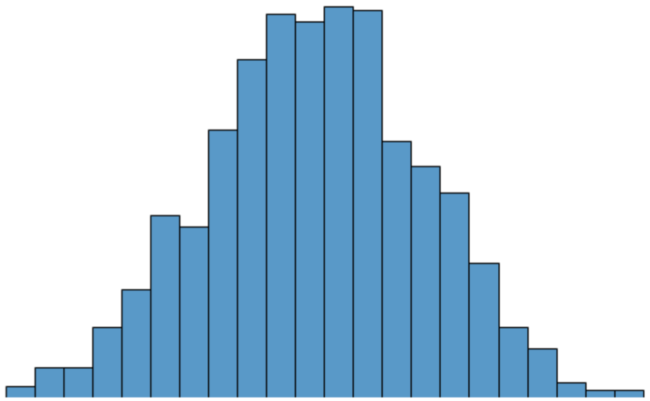
sampling distribution of the mean

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Problems with Traditional Confidence Intervals

- Assumptions about **distribution** or **sample size:**
  - Normal distribution
  - Sample size is **large enough (central limit theorem)**
    - What is a large sample for a specific problem?
  - **Population standard deviation σ** is **known**
    - Otherwise, we **approximate** it from the **sample standard deviation s**
  - Calculating the **standard error** for some statistcs can be **difficult**
    - E.x: Estimate the range between the 80th to 90th percentiles



sampling distribution of the mean

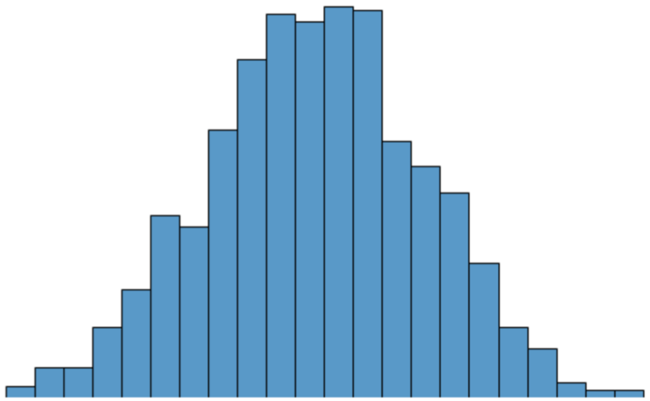
$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We can estimate population parameters **without these assumptions**.

Problems with Traditional Confidence Intervals

- Assumptions about **distribution** or **sample size:**
  - Normal distribution
  - Sample size is **large enough (central limit theorem)**
    - What is a large sample for a specific problem?
  - **Population standard deviation σ** is **known**
    - Otherwise, we **approximate** it from the **sample standard deviation s**
  - Calculating the **standard error** for some statistcs can be **difficult**
    - E.x: Estimate the range between the 80th to 90th percentiles



sampling distribution of the mean

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

$$\mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We can estimate population parameters **without these assumptions**.

Bootstrap

Resampling

Resampling

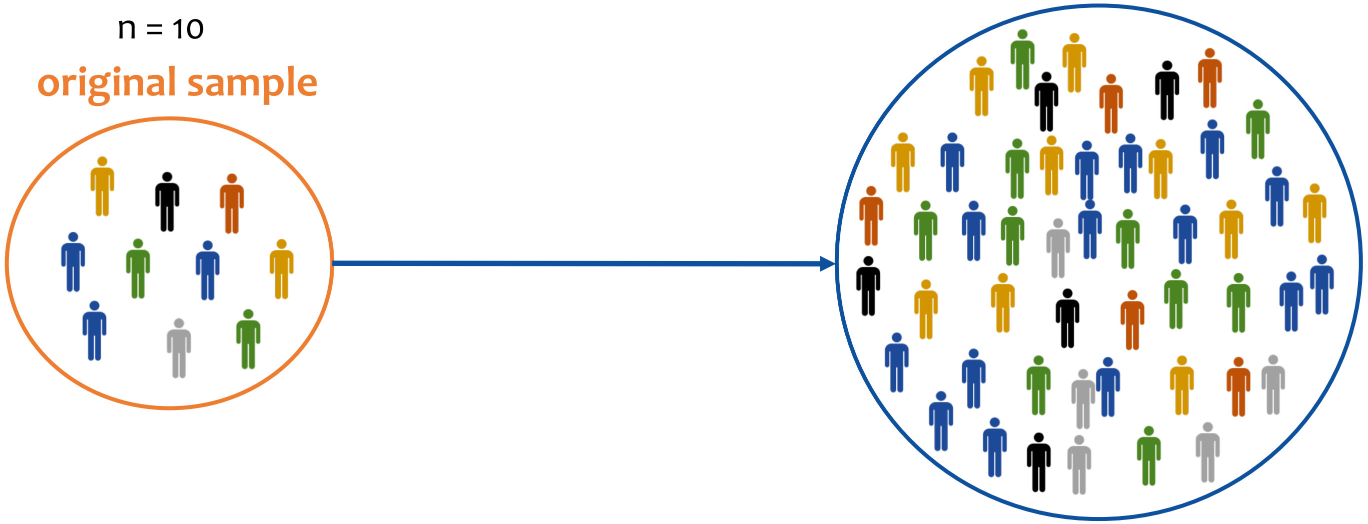
n = 10

original sample

Resampling

n = 10

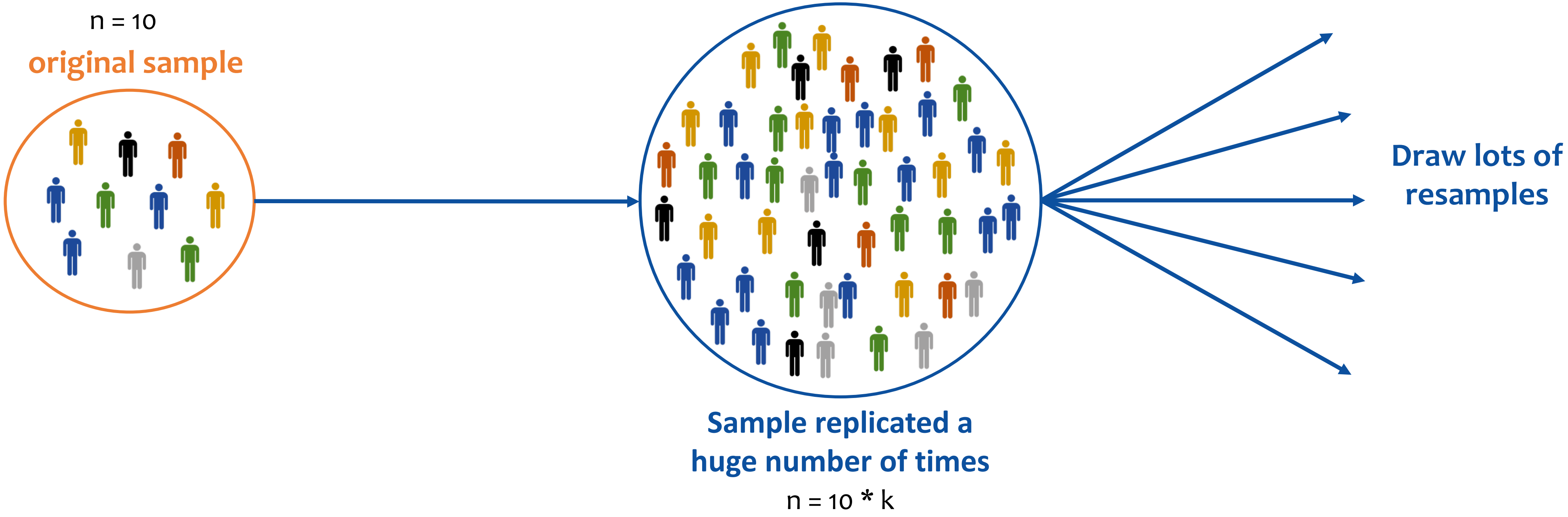
original sample



Sample replicated a huge number of times

n = 10 * k

Resampling


n = 10
original sample
Sample replicated a
huge number of times
n = 10 * k
Draw lots of
resamples

Resampling: Faster and Simpler Alternative

n = 10

original sample

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

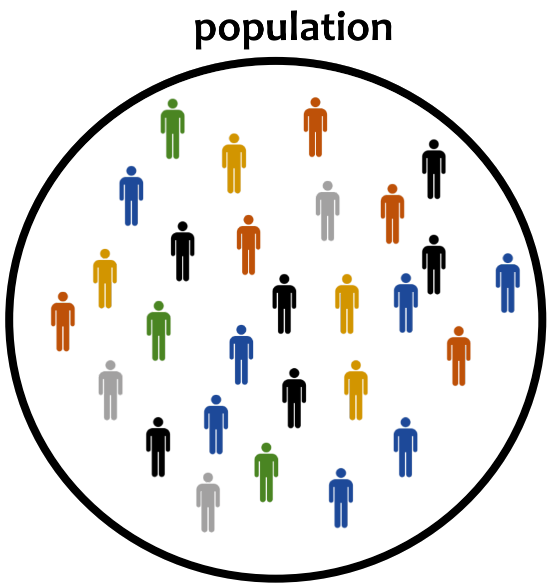
Resampling: Faster and Simpler Alternative

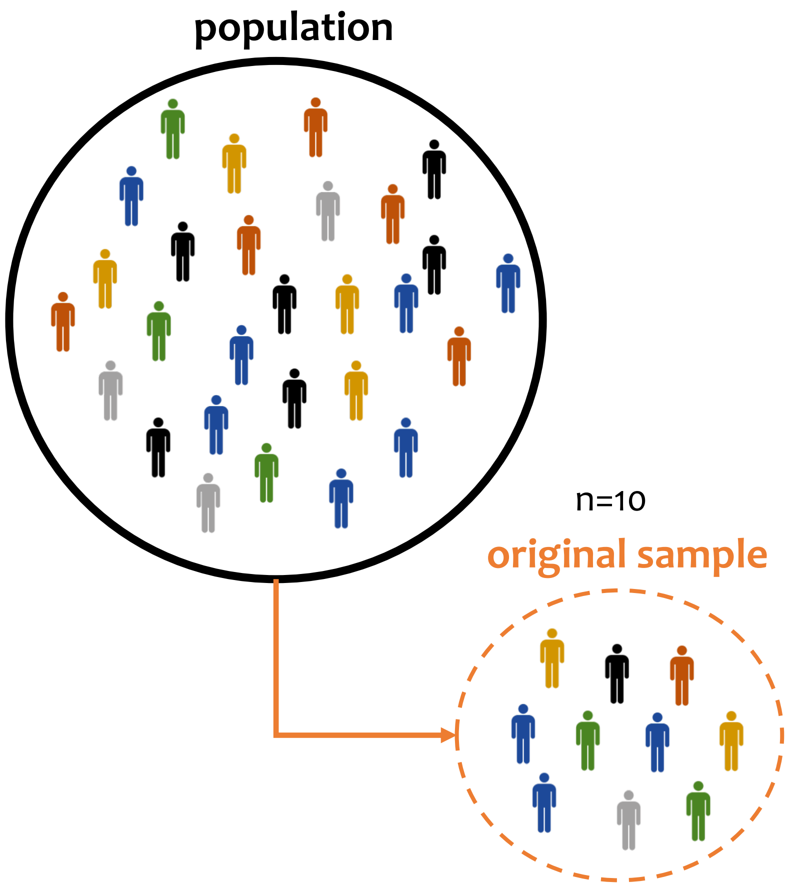
Resampling: Faster and Simpler Alternative

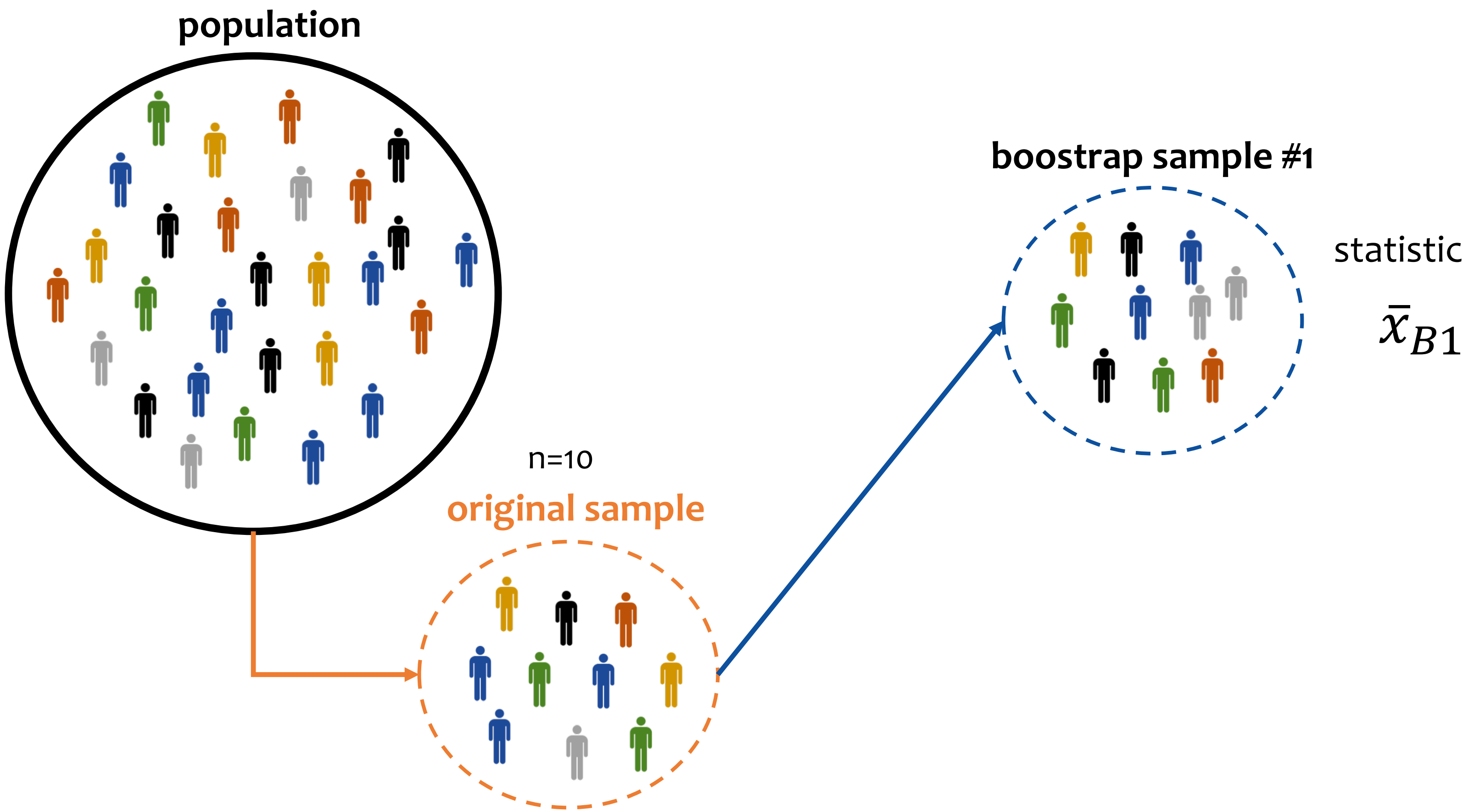
Resampling: Faster and Simpler Alternative

Resampling: Faster and Simpler Alternative

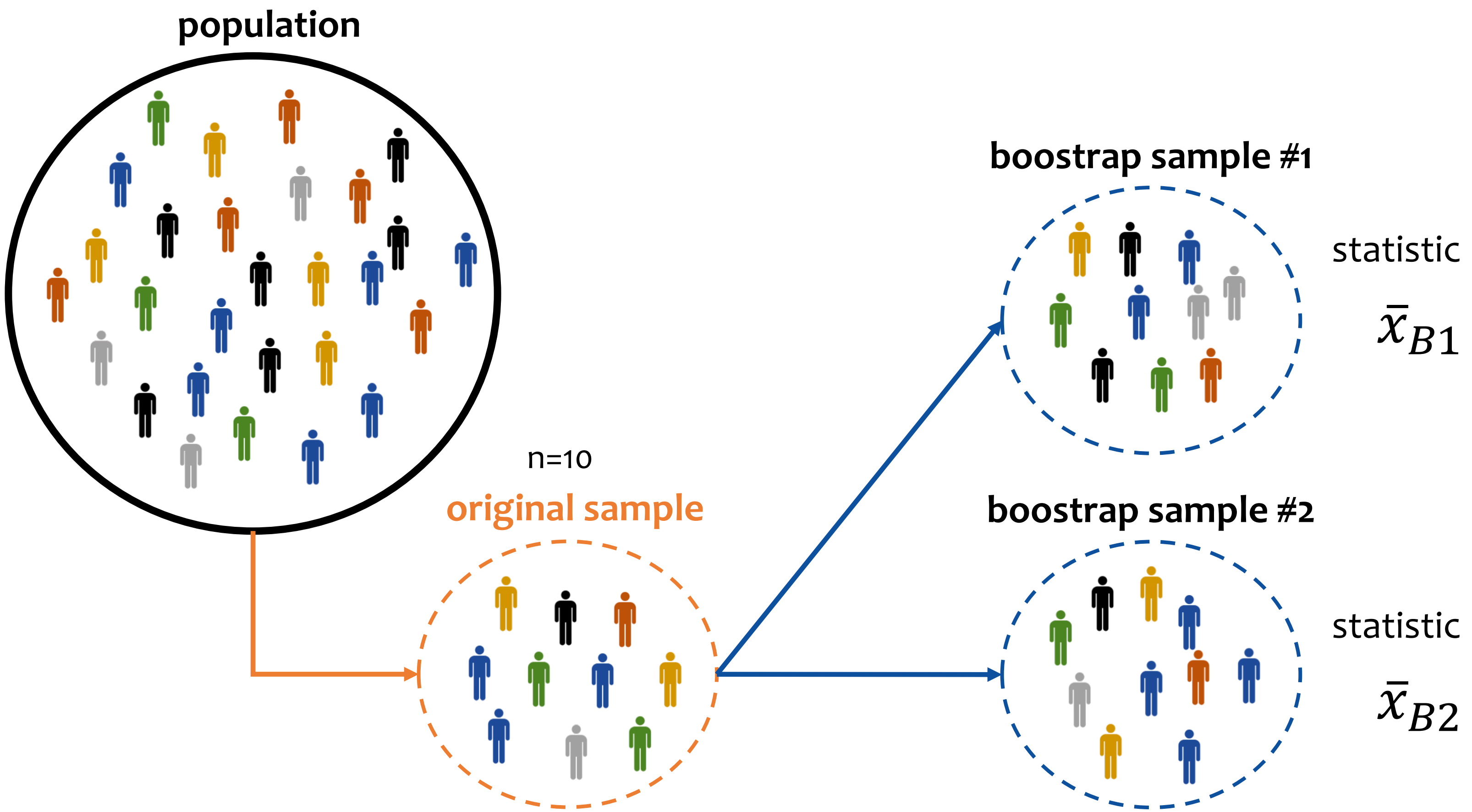
Bootstrap


population


population
n=10
original sample

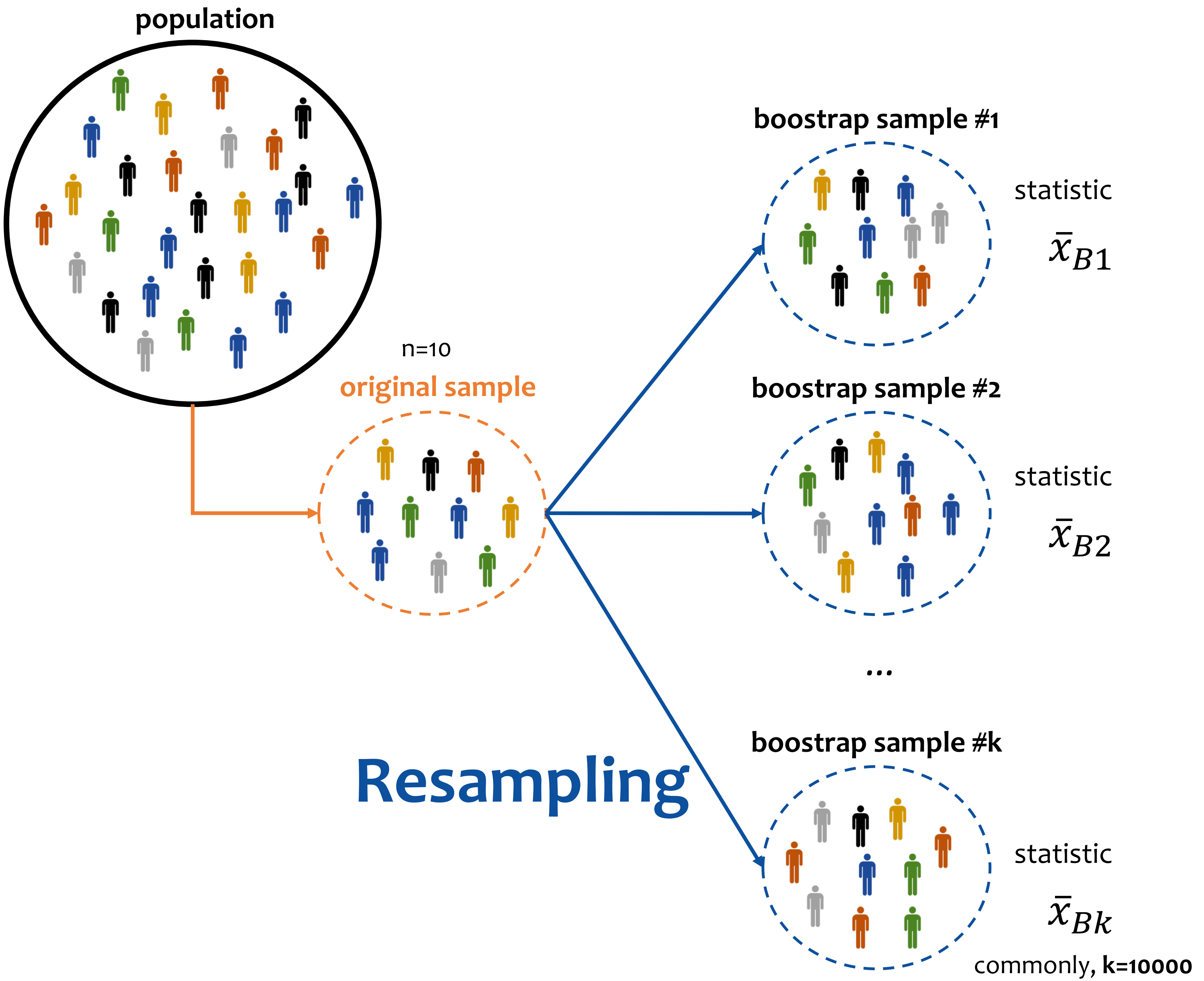

population
n=10
original sample
boostrap sample #1
statistic
$\bar{x}_{B1}$

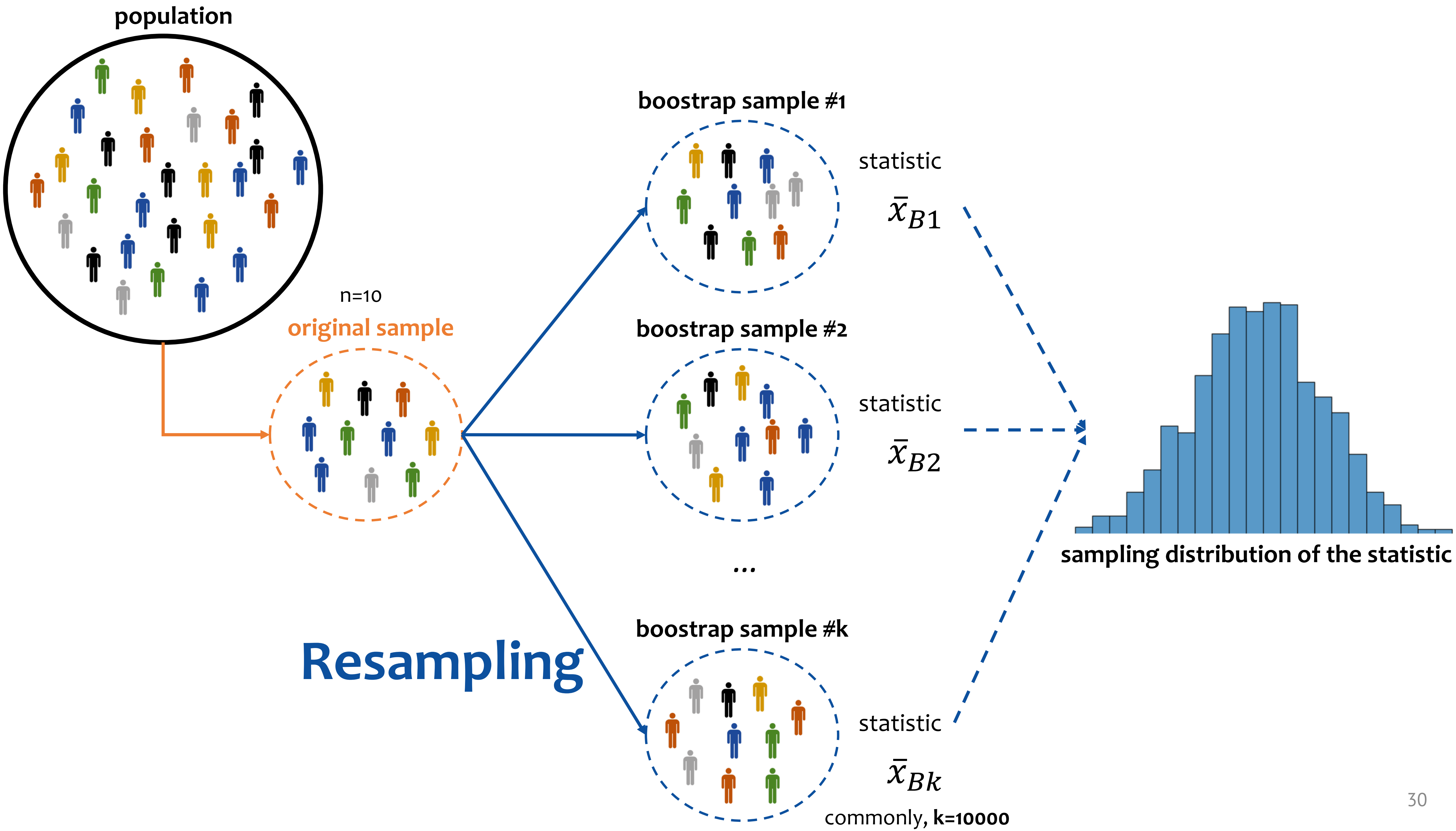
Resampling


population
boostrap sample #1
statistic
$\bar{x}_{B1}$
n=10
original sample
boostrap sample #2
statistic
$\bar{x}_{B2}$

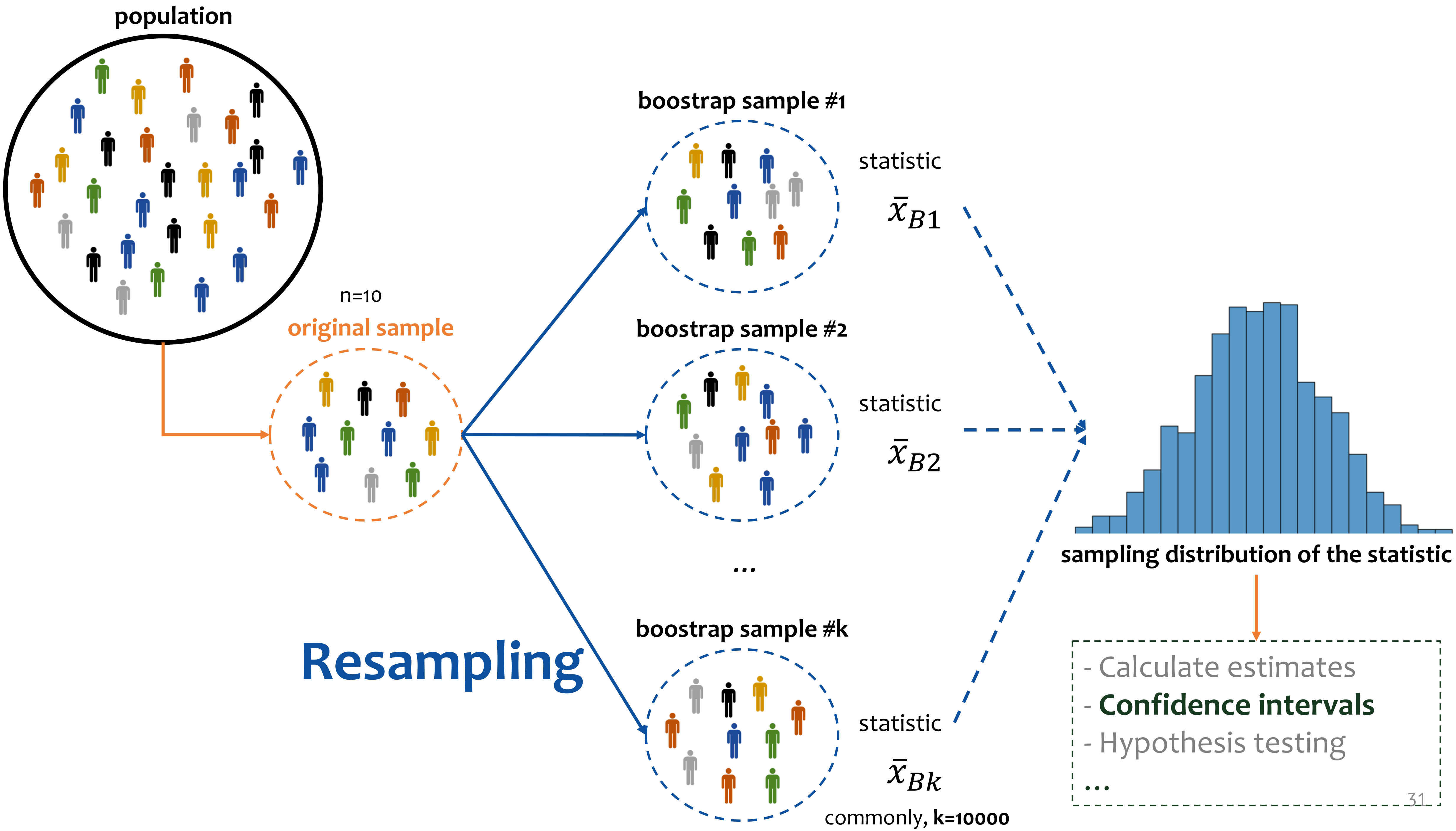
Resampling

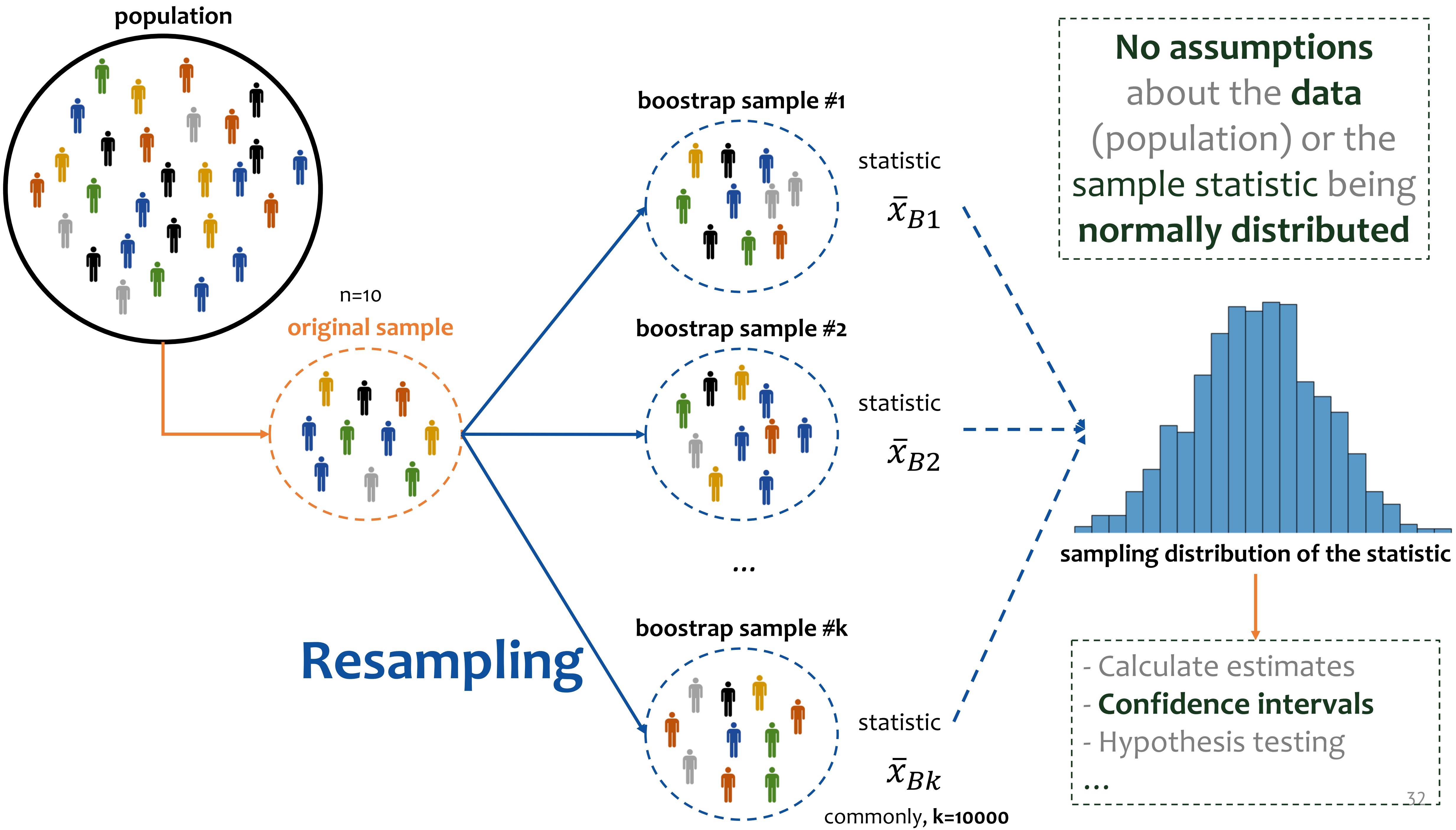
Resampling


population
n=10
original sample
boostrap sample #1
statistic
$\bar{x}_{B1}$
boostrap sample #2
statistic
$\bar{x}_{B2}$
...
boostrap sample #k
statistic
$\bar{x}_{Bk}$
commonly, **k=10000**



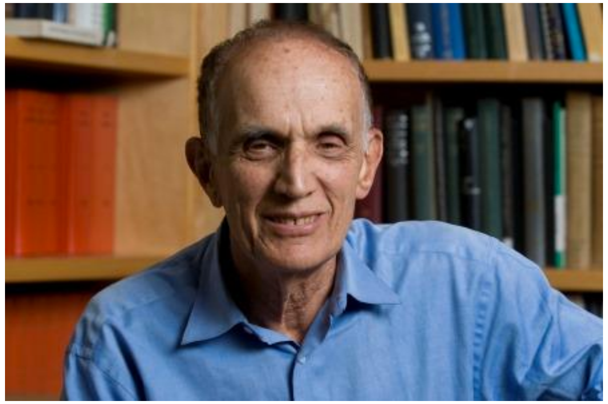
population
n=10
original sample
boostrap sample #1
statistic
$\bar{x}_{B1}$
boostrap sample #2
statistic
$\bar{x}_{B2}$
...
boostrap sample #k
statistic
$\bar{x}_{Bk}$
commonly, **k=10000**
sampling distribution of the statistic
Resampling


population
n=10
original sample
boostrap sample #1
statistic
$\bar{x}_{B1}$
boostrap sample #2
statistic
$\bar{x}_{B2}$
...
boostrap sample #k
statistic
$\bar{x}_{Bk}$
commonly, **k=10000**
Resampling
sampling distribution of the statistic
- Calculate estimates
- **Confidence intervals**
- Hypothesis testing
...


population
n=10
original sample
Resampling
boostrap sample #1
boostrap sample #2
...
boostrap sample #k
statistic
$\bar{x}_{B1}$
statistic
$\bar{x}_{B2}$
statistic
$\bar{x}_{Bk}$
commonly, k=10000
No assumptions about the data (population) or the sample statistic being normally distributed
sampling distribution of the statistic
- Calculate estimates
- Confidence intervals
- Hypothesis testing
...

Bootstrap (1979)

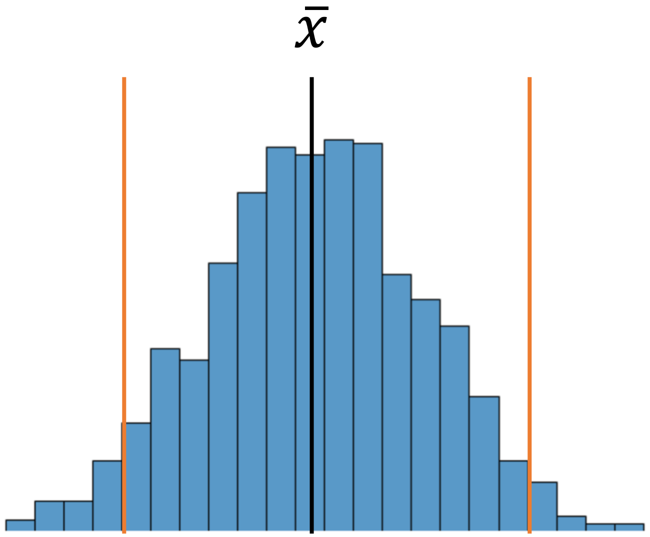
"The **bootstrap** is rarely the star of statistics, but it is **the best supporting actor**"



Bradley Efron

Bootstrap Confidence Interval

1. Get a sample **S** from the population
2. Repeat k times (~10000 times):
   1. Generate a **bootstrap sample** by **resampling S**
   2. Calculate the desired **statistic** for the bootstrap sample
3. Build the **sampling distribution** for the statistic
4. Compute the **interval around the mean** with the **concentration of c% observations/values**
   1. c% is the **confidence level = 1 – α**
   2. The interval consists of the **α/2 percentile** and **(1 - α/2) percentile**
   3. Thus, just sort the statistics and return the values of theses percentiles



sampling distribution of the statistic

Exercise

Given a dataset from stroke patients, we want to study their **mean glucose level**.

Provide a **95% bootstrap confidence intervals** for sample sizes of **100** and **1000**.