Simulation

Yujiao Li

2019-03-18

<u>**Simulate**</u> an experiment of tossing a fair coin 30 times, **P("Head") = ?**

Toss a coin,
it produces: ["H"] or ["T"]

⇩

Toss 30 times,
it produces: ["H", "T" ,…, "T"]

⇩

Calculate

$$P = \frac{\#\, Head}{30}$$

Procedure

Toss a coin,
it produces: ["H"] or ["T"]

⇩

Toss 30 times,
it produces: ["H", "T" ,…, "T"]

⇩

Calculate
$$P = \frac{\#\ Head}{30}$$

Write R function
with equal probability of "H" or "T"

⇩

Use function to generate data
by R loop 30 times

⇩

Calculate
by R devision function

What & Why

What

- Conducting experiments based on the model

Why

- Understanding the behavior of the system
- Evaluating various strategies for the operation of the system

Learning aims:

How to simulate data to:

- Test your statistical intuition.
- Generate random numbers given distributioin.
- Experiment for inferential stastistics (Estimation/Test)

1. Basic R functions

- Sampling

  sample()

  set.seed()

- Replicating

  for() { }

  replicate()

  sapply()

Syntax

- sample(x, size, replace = FALSE, prob = NULL)
- set.seed(2)
- for (i in 1:100) { x[i] }
- replicate(n, expr)
- sapply(X, FUN, ...)

Exercise 1

(1) Simulate an experiment of rolling a die 100 times and plot histogram of outcomes

(2) Replicate above trials 30 times to estimate the probability of showing "6"

2. Genrating random numbers

Normal distribution: N(0,1)

```
> rnorm(n, mean = 0, sd = 1)
```

Uniform distribution: U[0,1]

```
> runif(n, min=0, max=1)
```

Poisson distribution: Poisson($\lambda = 3$)

```
> rpois(n, lambda = 3)
```

Exercise 2

- Generate 50 numbers $\sim \mathrm{N}(10, 5)$
- Generate 100 numbers $\sim \mathrm{Poission}(\lambda = 50)$

- Generate 100 pair of (x, y) satisfying:

$x \sim Unif(-10, 10),\ \ y = 3x + \varepsilon$ and $\varepsilon \sim \mathrm{N}(0,4)$

Summary of R functions

Generate random numbers

- sample()
- rnorm()
- runif()
- rpois()

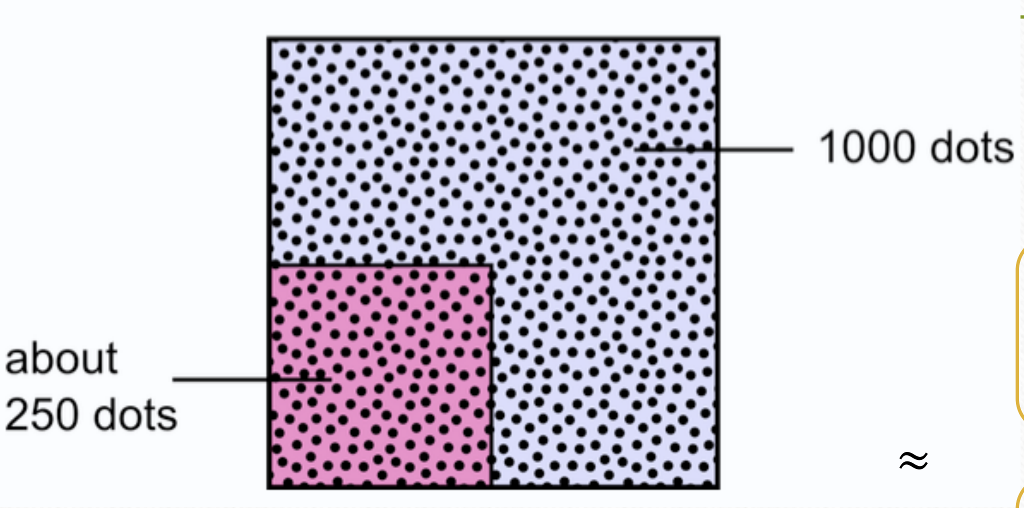
Relicate

- replicate()
- sapply()

Exercise 3

Design simulation to estimate $\pi = ?$

Probability of falling into dark purple area?

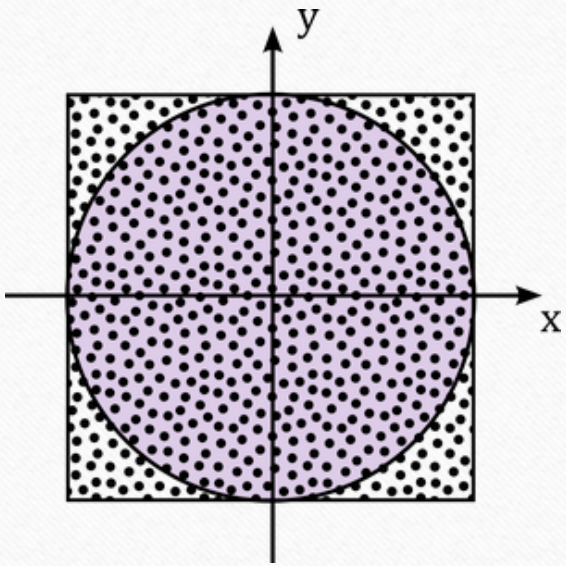


$\approx$

$$\text{TheoryRatio} = \frac{\text{DARK area}}{\text{ALL area}} = \frac{1\times1}{2\times2}$$

$$\text{SimulationRatio} = \frac{\textbf{\# of points} \text{ in DARK area}}{\textbf{\# of points} \text{ in ALL area}}$$

Probability of falling into dark purple area?



$$\pi = 4 * \frac{A_{circle}}{A_{square}}$$

$$\text{TheoryRatio} = \frac{\text{DARK area}}{\text{ALL area}} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}$$

$\approx$

$$\text{SimulationRatio} = \frac{\textbf{\# of points} \text{ in DARK area}}{\textbf{\# of points} \text{ in ALL area}}$$

3. Bootstrap

Question

1. What is averaged sleeping hours of all students at University?

   (i.e. interval estimation of $\bar{y}$)

2. How many hours of sleeping will lose if one add another course?

(i.e. interval estimation of regression coefficients a and b

for $y = a + bx$)

Sleeping data

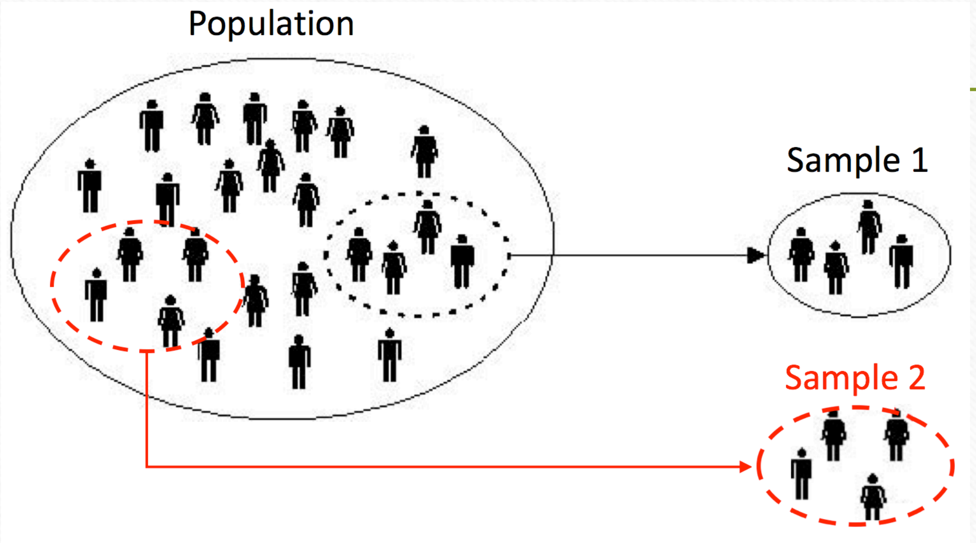
ID	name	Number of courses (x)	Sleeping hours (y)
1	Tom	4	9
2	Jerry	3	7
…	…	…	…
32	Yujiao	0	12

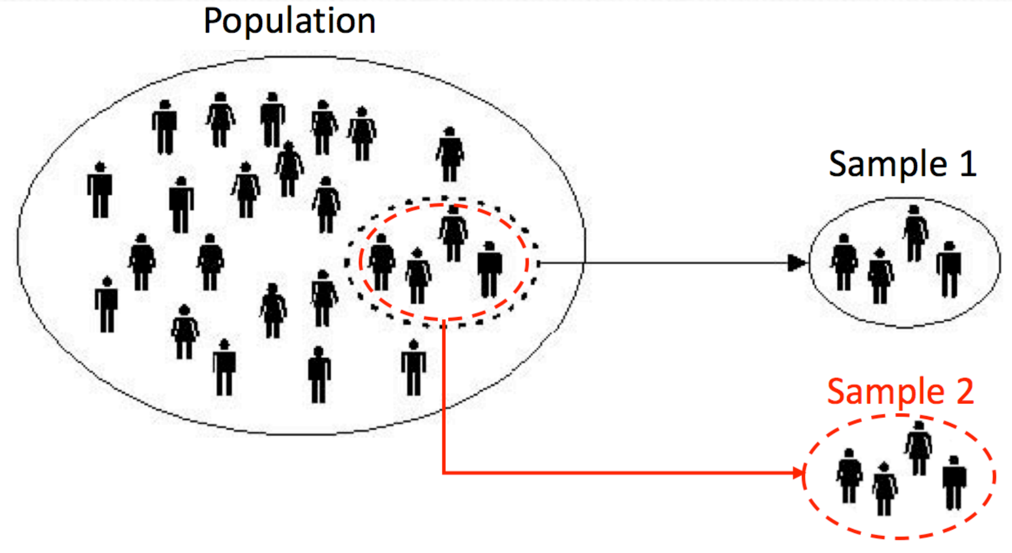
Solution: Bootrstrap

Bootstrapping is a approach to statistical inference based on building a sampling distribution for a statistic by resampling with replacement from the original data.

'Bootstrapping' means 'pulling oneself up by one's bootstraps'
– in this case, using the sample data as a population from which repeated samples are drawn.
(Efron, 1979)

Interval Estimation v.s. Bootstrap Interval Estimation





Procedure

Original data

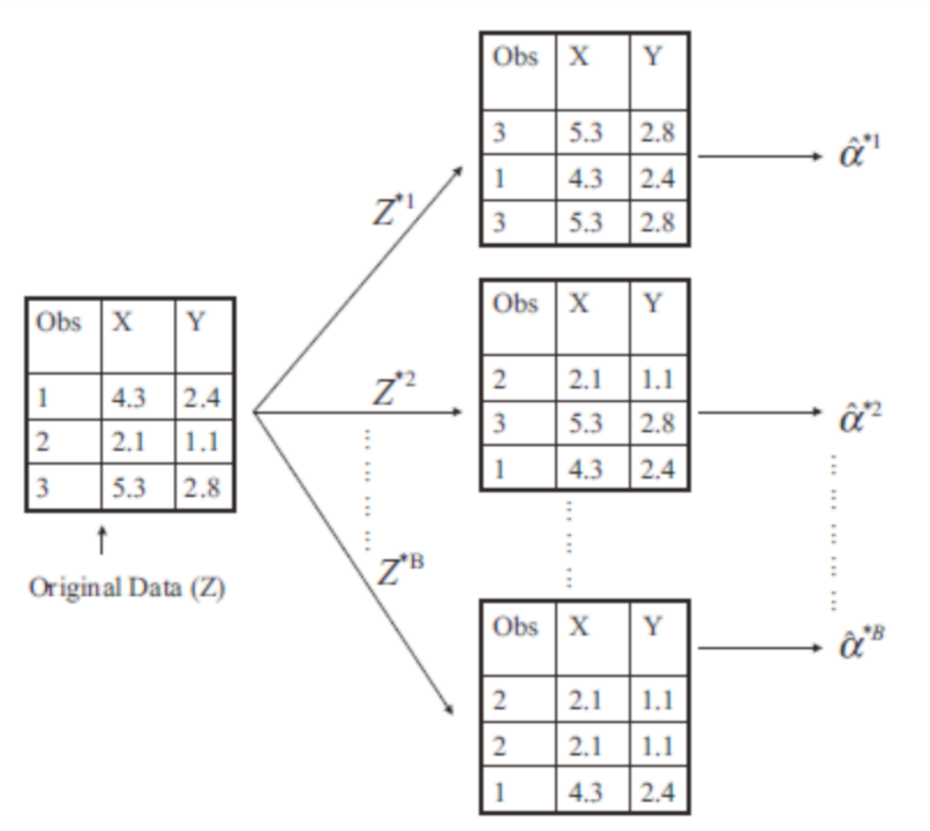
	Sleeping hours (y)
$\mathbf{y_1}$	**4**
$\mathbf{y_2}$	**6**
$\mathbf{y_3}$	**8**
$\bar{\mathbf{y}}$	**6**

Sampled data for bootstrapping

	Sample 1	Sample 2	…	Sample 200
$y_{(1)}$	y_1: 4	y_3: 8		y_3: 8
$y_{(2)}$	y_1: 4	y_2: 6		y_3: 8
$y_{(3)}$	y_2: 6	y_3: 8		y_1: 4
Mean	$(4+4+6)/3 = 4.7$	$(8+6+8)/3 = 7.3$		$(8+8+8)/3 = 8$

$[4.7, 7.3, \ldots, 8]_{1\times 200} \longrightarrow$ Interval estimation

Note: for simplicity, we assume orinigal data have only 3 observations instead of 29.



- Sampling with replacement from original dataset.
- Estimate statistic (coefficient α) from every bootstrap sample.
- Calculate sample distribution of α

Exercise 4

Use R to solve above bootstrapping question

Other applications

- Interval estimation of mean value
- Interval estimation of two samples' mean difference
- Interval estimation of parameters in linear regression

4. Placebo Test

Is treatment useful?

H_0: Treatment group is NOT different with Control group.

H_1: Treatment group is different with Control group.

ID	Treatment	Health score
1	**0**	10
2	**0**	20
3	**0**	40
4	**1**	30
5	**1**	40
6	**1**	90

Treatment Effects = mean_1 – mean_0

$$= \frac{30+40+90}{3} - \frac{10+20+40}{3}$$

$$= 30$$

Is 30 big enough to reject H_0?

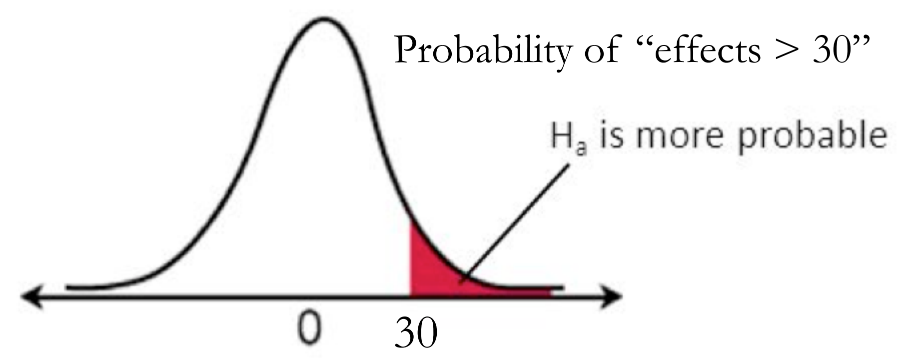
Reject H_0 if :

$$\frac{\text{Observed effect}}{\text{given treatment}} \gg \frac{\text{Expected effect}}{\text{given No treatment}}$$

$\Downarrow$

Known result: 30

$\Downarrow$



Distribution of effects without treatment?

Idea: "Blind" which treatment patients are getting.

Resampling without replacement

ID	Treat Original	Treat Sample 1	Health score
1	0	0	10
2	0	1	20
3	0	1	40
4	1	0	30
5	1	0	40
6	1	1	90

Treatment Effects_sample_1 = mean_1 – mean_0

$$= \frac{20+40+90}{3} - \frac{10+40+30}{3}$$

$$= 23.3$$

Treatment Effects_sample_2 = …

…

Treatment Effects_sample_k = …

Exercise 5

Use R to solve above placebo test