Lecture 41 Monte Carlo Simulation, Bootstrapping And Permutation Test

BIO210 Biostatistics

Xi Chen

Spring, 2022

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

Monte Carlo Simulations

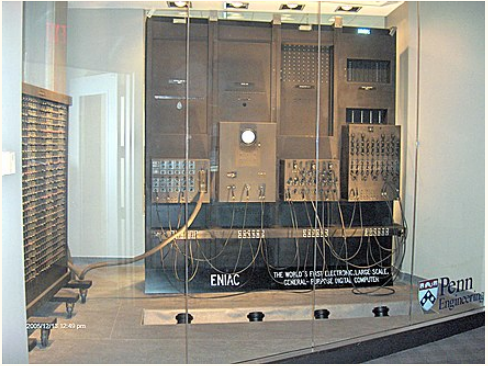
Casino de Monte-Carlo, picture taken on 26 Dec 2017.

A Little History About Monte Carlo Simulations

Stanislaw Ulam

John von Neumann

ENIAC

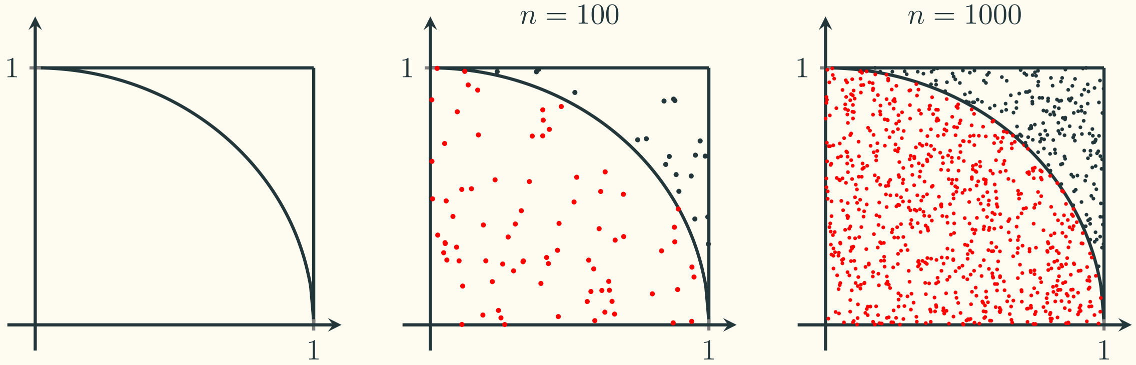


Code name: Monte Carlo

`https://en.wikipedia.org/wiki/Monte_Carlo_method`

A Monte Carlo Simulation To Calculate π

- **Monte Carlo Simulation:** a method of solving deterministic problems using a probabilistic analog.
  - An example to calculate π using Monte Carlo Simulation.

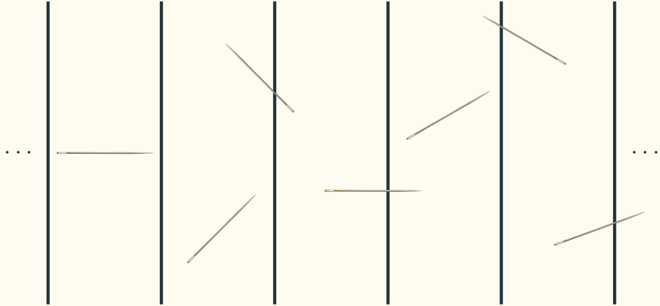


A Monte Carlo Simulation To Calculate π

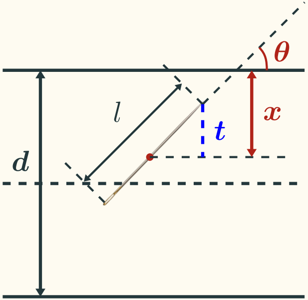
Number of dots	Estimated π
10	2.0
100	3.0
$1,000$	3.124
$10,000$	3.1276
$100,000$	3.14112
$1,000,000$	3.141772
$10,000,000$	3.14163332
$100,000,000$	3.141831323

Buffon's Needle

- First Posed by Georges-Louis Leclerc, Comte de Buffon in 1733, and reproduced with solution in 1777.
  - Suppose we have a floor made of parallel strips of wood, each the same width, and we drop a needle onto the floor. What is the probability that the needle will lie across a line between two strips?



Buffon's Needle



We have two random variables: $\boldsymbol{X}$, $\boldsymbol{\Theta}$ to describe the position of the needle:

$$0 \leqslant x \leqslant \frac{d}{2}$$

$$0 \leqslant \theta \leqslant \frac{\pi}{2}$$

Marginal PDF of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$: $f_{\boldsymbol{X}}(x) = \frac{2}{d}$, $f_{\boldsymbol{\Theta}}(\theta) = \frac{2}{\pi}$

Joint PDF of $\boldsymbol{X}$ and $\boldsymbol{\Theta}$: $f_{\boldsymbol{X},\boldsymbol{\Theta}}(x,\theta) = f_{\boldsymbol{X}}(x) f_{\boldsymbol{\Theta}}(\theta) = \frac{4}{\pi d}$

$A =\{$ the needle lies across a line between two strips$\}$

Event $A \Leftrightarrow$ red is shorter than blue $\Leftrightarrow 0 \leqslant \boldsymbol{x} \leqslant \frac{l \cdot \sin\theta}{2}$

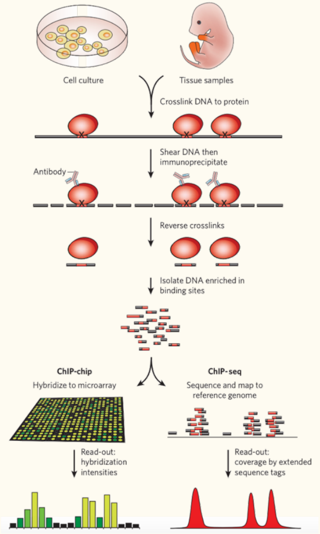
$$P(A) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{l \cdot \sin\theta}{2}} f_{\boldsymbol{X},\boldsymbol{\Theta}}(x,\theta) \mathrm{d}x\mathrm{d}\theta = \int_0^{\frac{\pi}{2}} \int_0^{\frac{l \cdot \sin\theta}{2}} \frac{4}{\pi d} \mathrm{d}x\mathrm{d}\theta$$

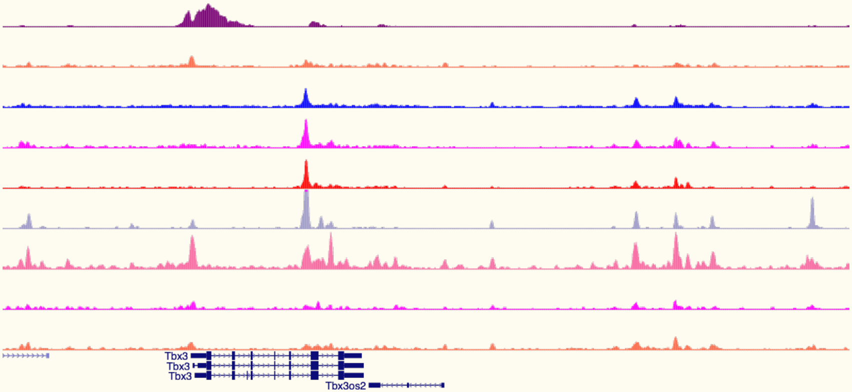
$$= \int_0^{\frac{\pi}{2}} \left[\frac{4}{\pi d} \cdot x\right]_0^{\frac{l \cdot \sin\theta}{2}} \mathrm{d}\theta = \int_0^{\frac{\pi}{2}} \frac{2l \cdot \sin\theta}{\pi d} \mathrm{d}\theta = \frac{2l}{\pi d} \int_0^{\frac{\pi}{2}} \sin\theta \mathrm{d}\theta = \frac{2l}{\pi d}$$

Buffon's Needle - Monte Carlo simulations to estimate π

Number of needles	Estimated π
10	3.333333
100	3.125
$1,000$	3.333333
$10,000$	3.22684737
$100,000$	3.13293023
$1,000,000$	3.14433768
$10,000,000$	3.14071042
$100,000,000$	3.14148011

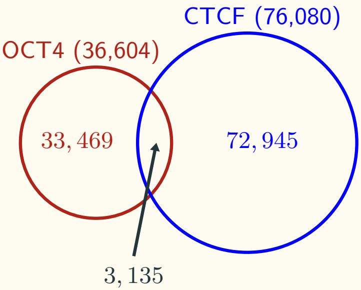
Overlap of Transcription Factors



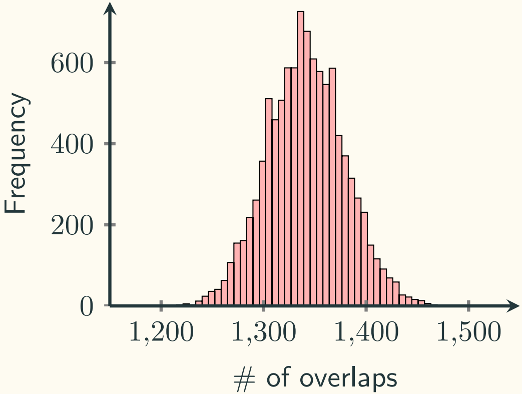


Overlap of OCT4 And CTCF In mESC

ChIP-seq in mESC



10,000 simulations



Bootstrapping

- Point/interval estimation of mean/median $etc.$ from a population with very little information.
- How? - Bootstrapping methods.

$$\left\{\begin{array}{l}\text{parametric bootstraps}\\ \text{nonparametric bootstraps}\\ \text{weighted bootstraps}\\ \dots\ \dots\end{array}\right.$$

Steps of Bootstrapping

1. Replace the population with the sample
2. Sample **with replacement** B times. B should be large, say 1,000
3. Compute sample means/medians each time, M_i
4. Obtain the approximate distribution of the sample mean/median

Bootstrapping Example

Original sample ($n = 25$)

16.66	13.58	2.27	9.96	13.11
6.4	11.33	10.54	10.02	9.13
12.17	16.02	5.17	15.14	11.14
12.23	4.32	10.68	17.42	4.6
0.11	1.28	11.33	21.92	15.62

sampling
with replacement
with the same sample size $n = 25$

Bootstrapping sample #1 ($n = 25$)

15.62	15.62	17.42	5.17	11.33
2.27	10.02	12.23	15.14	15.62
12.23	11.33	11.33	1.28	12.17
11.33	16.02	4.32	5.17	16.66
11.33	10.68	12.23	5.17	10.02

Bootstrapping sample #2 ($n = 25$)

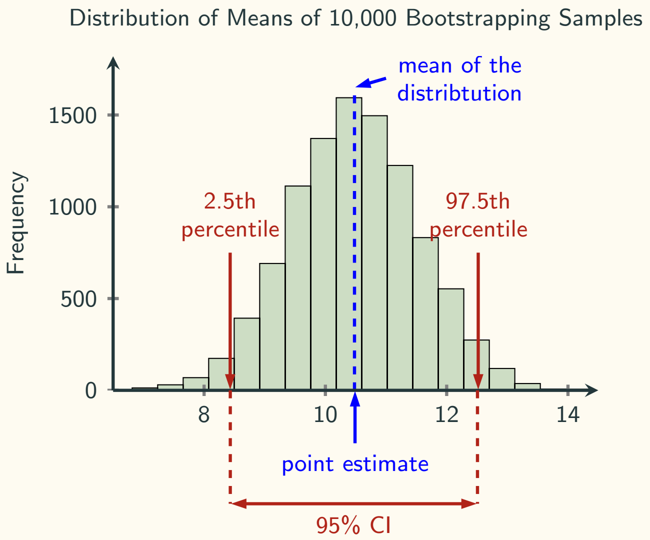
6.4	21.92	12.23	13.58	9.13
17.42	9.13	6.4	21.92	13.58
16.66	10.54	15.62	9.13	9.13
11.14	10.02	0.11	11.14	4.32
0.11	9.13	17.42	10.02	21.92

... ...

Bootstrapping sample #10,000 ($n = 25$)

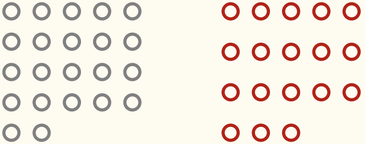
21.92	15.14	17.42	16.02	4.32
10.54	15.14	0.11	16.66	12.17
15.62	13.11	15.62	11.33	15.62
6.4	15.14	15.62	9.13	15.14
16.66	12.23	2.27	12.17	2.27

Bootstrapping - Point And Interval Estimation



Permutation Tests

Control (X) Treatment (Y)



$\bar{x} = 13.18$ $\bar{y} = 9.50$

$$H_0 : \mu_X - \mu_Y = 0$$
$$H_1 : \mu_X - \mu_Y \neq 0$$

Test statistic: $\bar{x} - \bar{y} = 3.68$

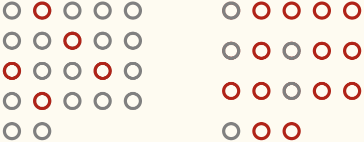
How to assess statistic significance?

Using permutation
(shuffle the group labels):

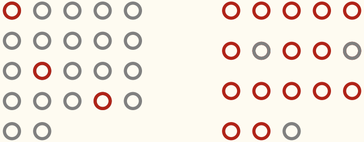
$$\binom{40}{22} = 113{,}380{,}261{,}800$$

Control (X) Treatment (Y)

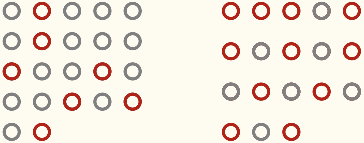
Permutation #1
$d_1 = \overline{x_1} - \overline{y_1}$



Permutation #1
$d_2 = \overline{x_2} - \overline{y_2}$

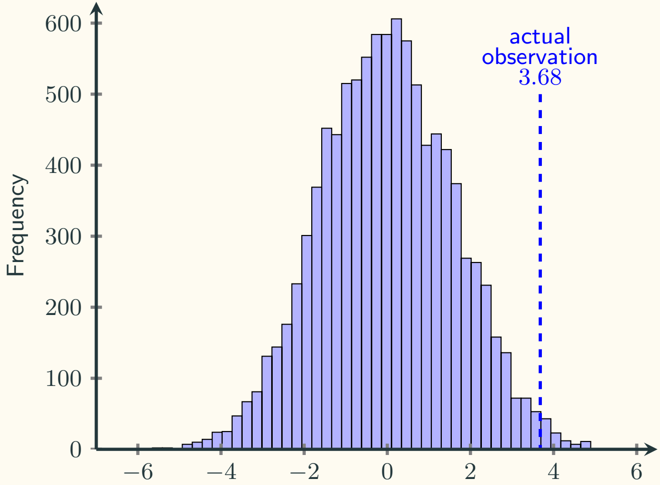


Permutation #1
$d_3 = \overline{x_3} - \overline{y_3}$



Permutation Tests - $p-$value Calculation

Distribution of 10,000 Differences (d_1 to d_{10000})



p-value: probability of seeing the observation or more extreme given that H_0 is true!

$$p_{\text{one-sided}} = \frac{\#\text{ of simulations} \geqslant 3.68}{\text{total }\#\text{ of simulations}}$$

$$p_{\text{two-sided}} = 2 \times p_{\text{one-sided}}$$

https://www.jwilber.me/permutationtest/

Why Simulation Works?

Law of Large Numbers

Proposed by Gerolama Cardano, proved by Jakob Bernoulli:

$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \ldots$ is an infinite sequence of i.i.d. random variables:

$$E[\boldsymbol{X}_1] = E[\boldsymbol{X}_2] = E[\boldsymbol{X}_3] = \cdots = \mu, \text{ and } \overline{\boldsymbol{X}_n} = \frac{\boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_3 + \cdots + \boldsymbol{X}_n}{n}$$

Then, we have:

$$\overline{\boldsymbol{X}_n} \text{ converges to } \mu \text{ when } n \to \infty$$