Introduction to Probability and Statistics

Sajan Kumar (01/15/2021)

Mark Twain: “There are three kinds of lies: lies, damned lies, and statistics”

Outline

- Probability and statistics
- Statistical thinking
- Data types and their distributions
- Summarizing data
- Correlation between random variables
- Central limit theorem
- Hypothesis testing
- Confidence interval
- Bootstrap method

Use of statistics in Machine learning

- Exploratory data analysis (Regression or classification problem)
- Summary statistics and relation between different variables
- Data cleaning (outlier detection, filling missing values)
- Data transformation (Standardizing data, scaling, log transformations)
- Resampling methods (imbalance classes)
- Model selection (hypothesis testing)

Probabilistic thinking

- The thinking which not only human species but most of species on this planet are used to.
- Probability is a quantitative measure of uncertainty about a uncertain process.
- Historically, it starts with solving the gambling problem.
- In Indian context, I think Mahabharata is the perfect example of using probabilistic thinking.
- Objective (throwing a fair coin) and subjective (whether it will rain or not) probability

Statistical thinking

When you translate a complex problem in relatively simple terms that not only capture the essential aspects of the problem, but also provide us how uncertain we are about our knowledge

Three main things statistics can do

- ✓ Describe: For example, mean, median, maximum, minimum and many more….
- ✓ Decide: Help us to take decision based on data, For example, Did you detect a source of gamma-ray in sky?
- ✓ Predict: Make prediction about future events based on previous data points or knowledge

Probability Vs Statistics

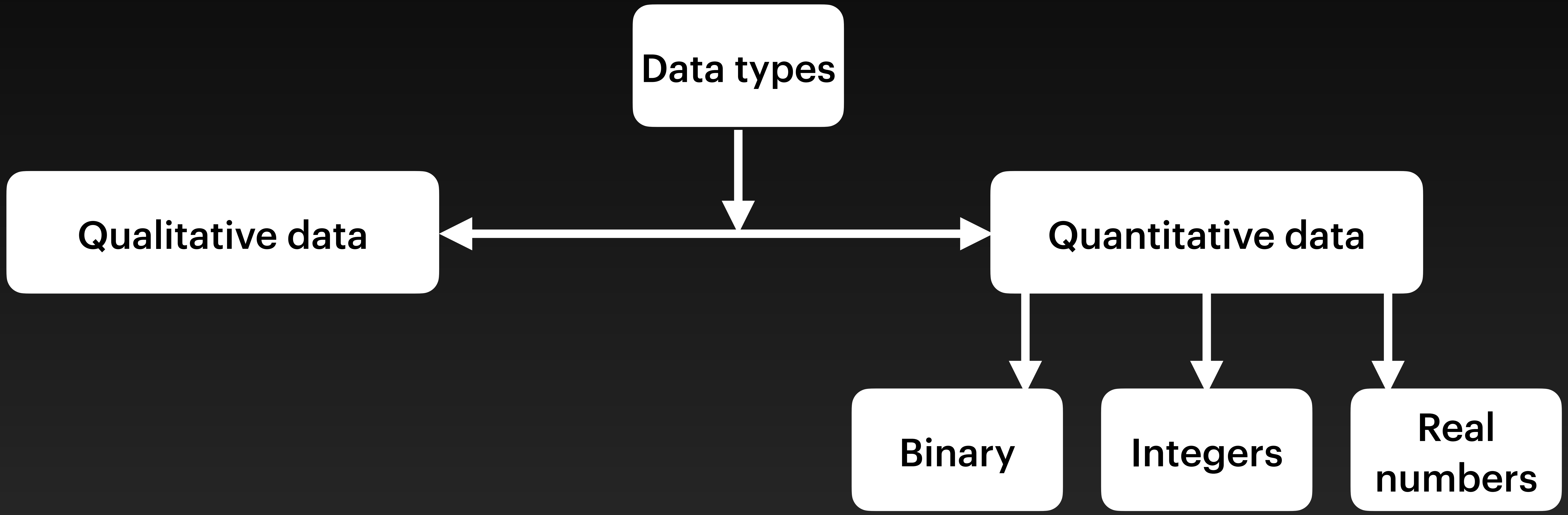
Probability

- In probability, you start with a known model
- Model parameters are fixed, data is random
- Predict the likelihood of data for future events
- For example, If I say that I have a fair coin with probability of head 0.5, how will my data look like when I throw my coin many times
- Random process is known, try to find the future outcome (data)

Statistics

- In statistics, you start with a data
- Data is fixed (i.e. outcome is known)
- Try to draw conclusions (inference) about the model parameters from this data
- For example, after tossing a coin for 100 times, I got 70 heads and 30 tails
- Question we can ask is: is our coin fair or not?
- Outcome is known, try to understand unknown random process

Data types



Data types

Qualitative data

Quantitative data

Binary

Integers

Real numbers

In short, we can divide data into discrete and continuous type

Random data distributions

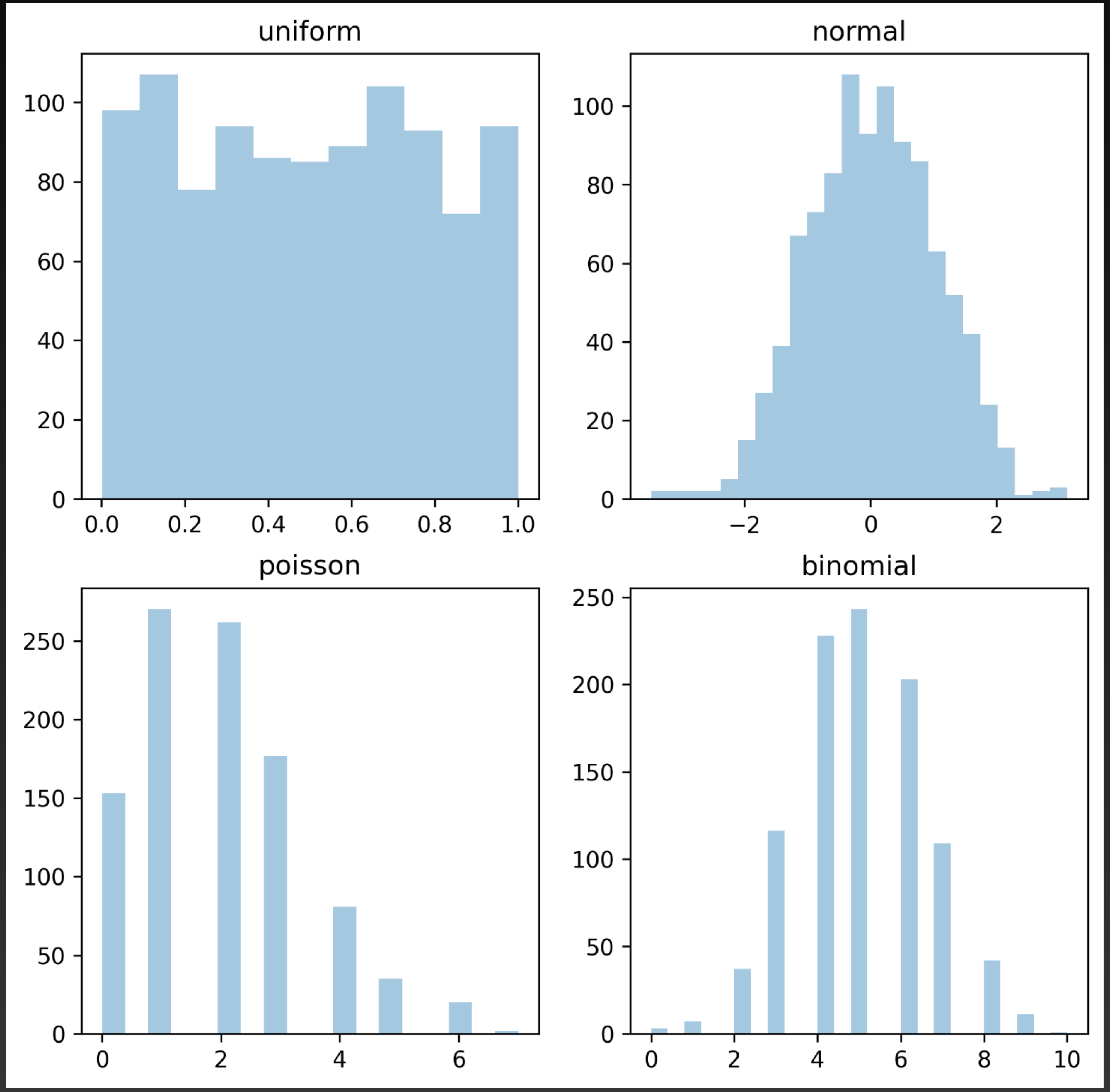
Random variable is a quantity of interest whose true value is unknown, however, the variable can be characterized by certain probability distribution function

Normal distribution

$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Poisson distribution

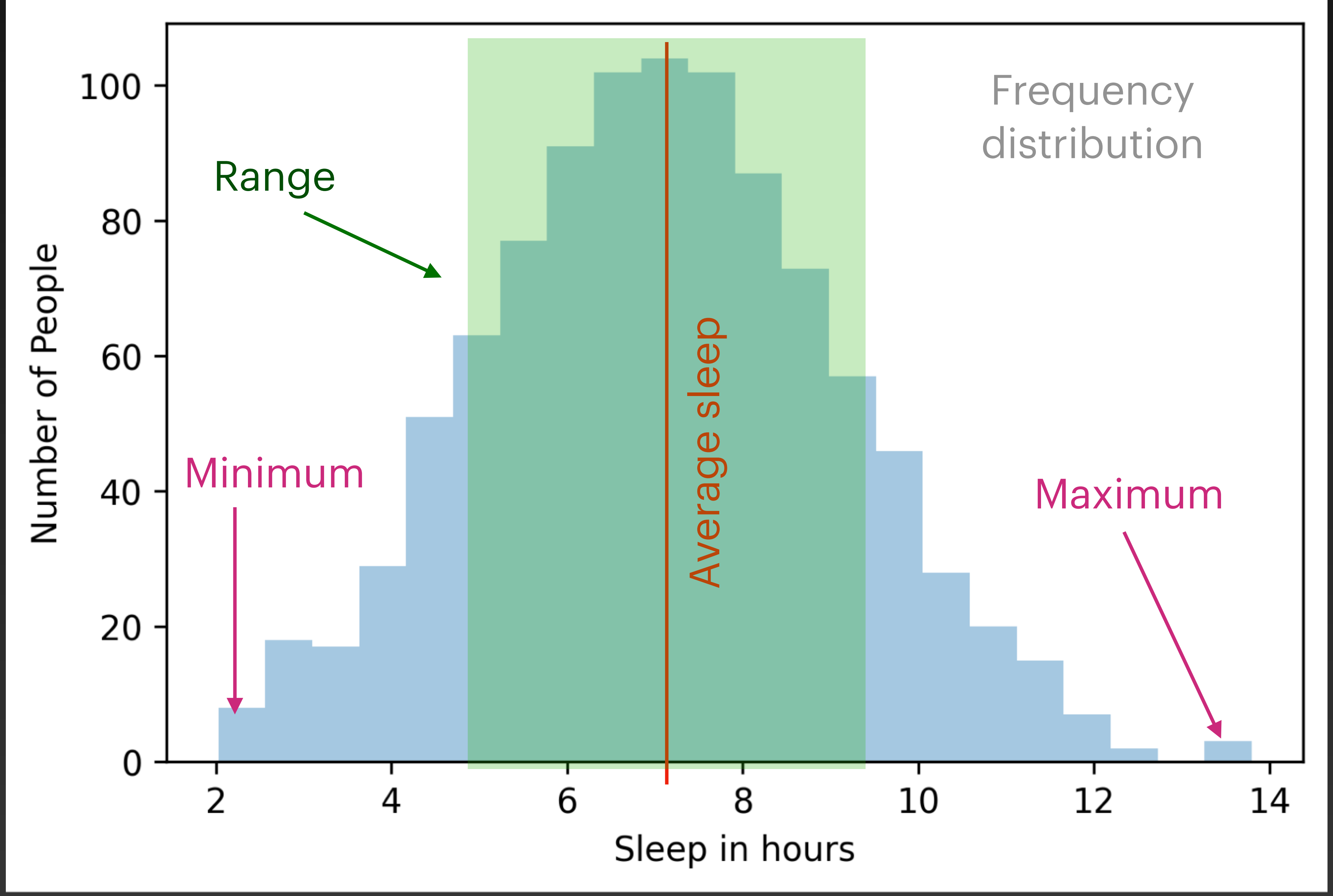
$$\mathrm{P(X = x)} \quad = \quad \mathrm{e}^{-\lambda}\frac{\lambda^x}{x!}$$



Summarizing data

“The art of throwing away the information to understand the world better can be called summarizing data”

Most important trait of human thinking is doing generalization which in other words is summarizing data

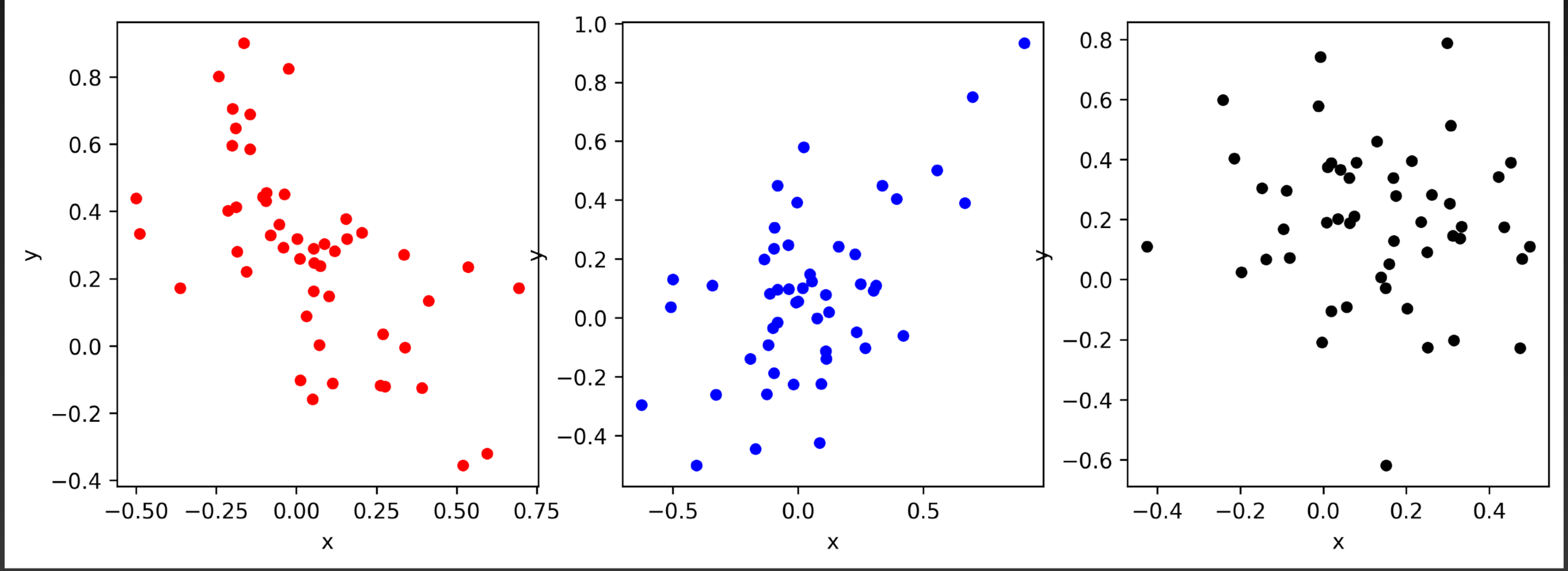


Data

Person	Sleep time
Rohan	7.2 h
Manoj	8.1 h
Saroj	5 h
Kohli	4 h
..	...

Correlation between random variables

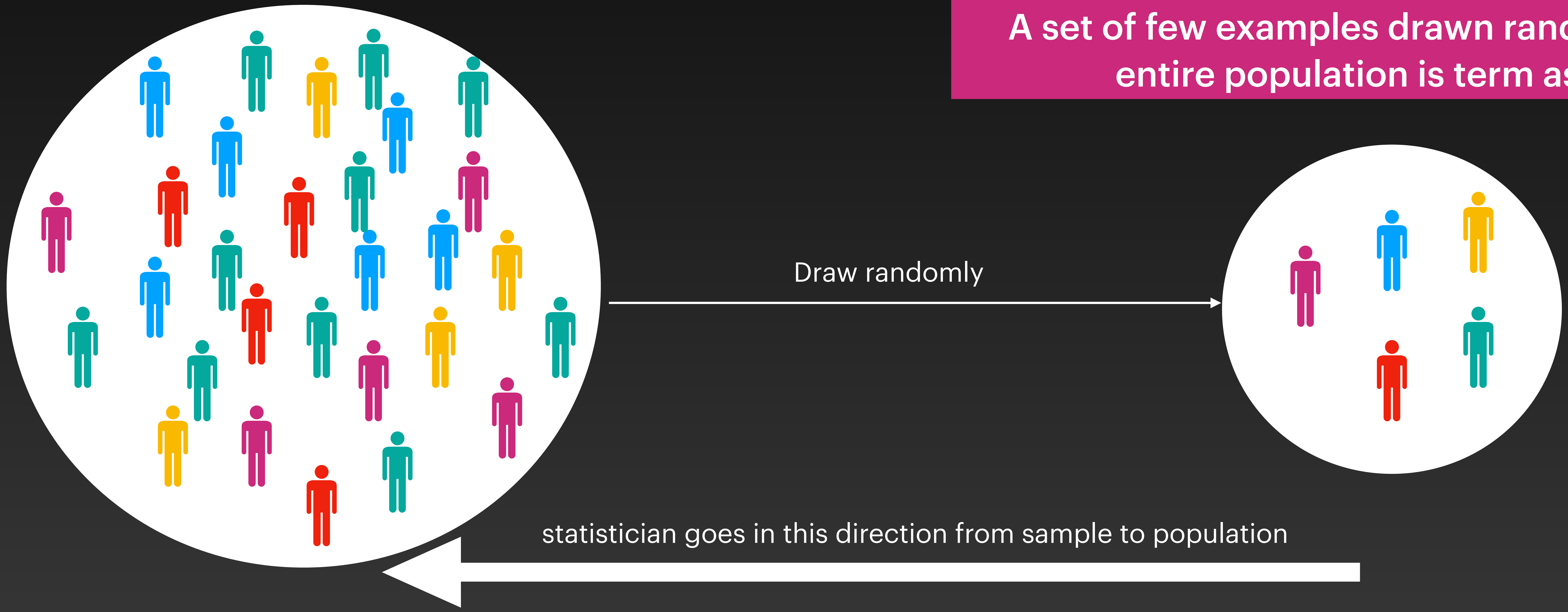
- The statistical relationship between two variables is referred to as correlation
- A correlation could be positive, negative or zero



Population and sample

Population is the entire group about which we want to draw some conclusions

A set of few examples drawn randomly from the entire population is term as sample

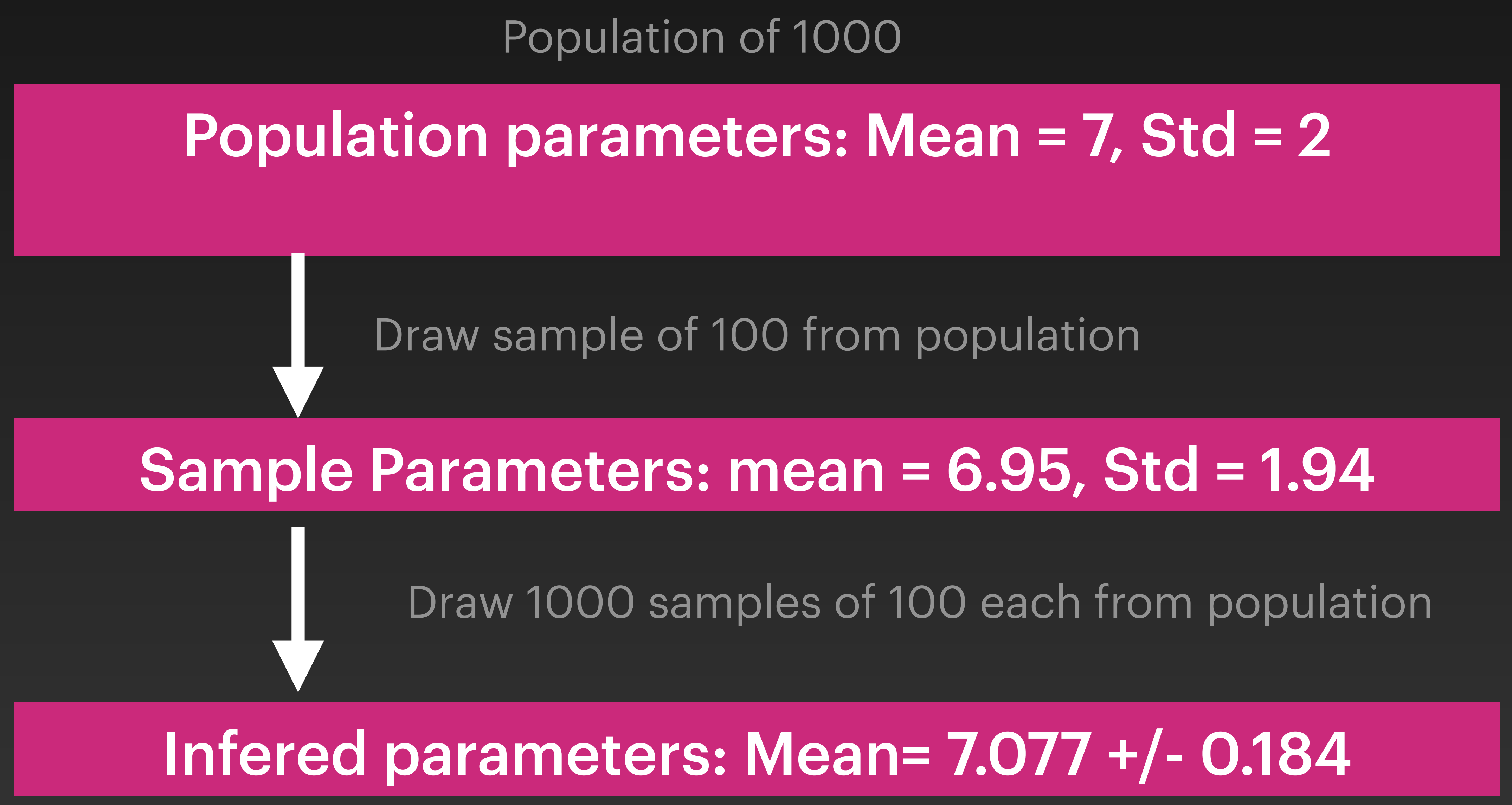


Draw randomly

statistician goes in this direction from sample to population

Sampling error (Standard error of mean)

- ✓ **Sampling error:** No matter how good our sample is, there is always a difference between the sample parameter and population parameter
- ✓ **Statistical fluctuations:** within in each different sample of measurements, no matter even you perform the experiment under the same conditions, the sample statistics will be different



Population of 1000

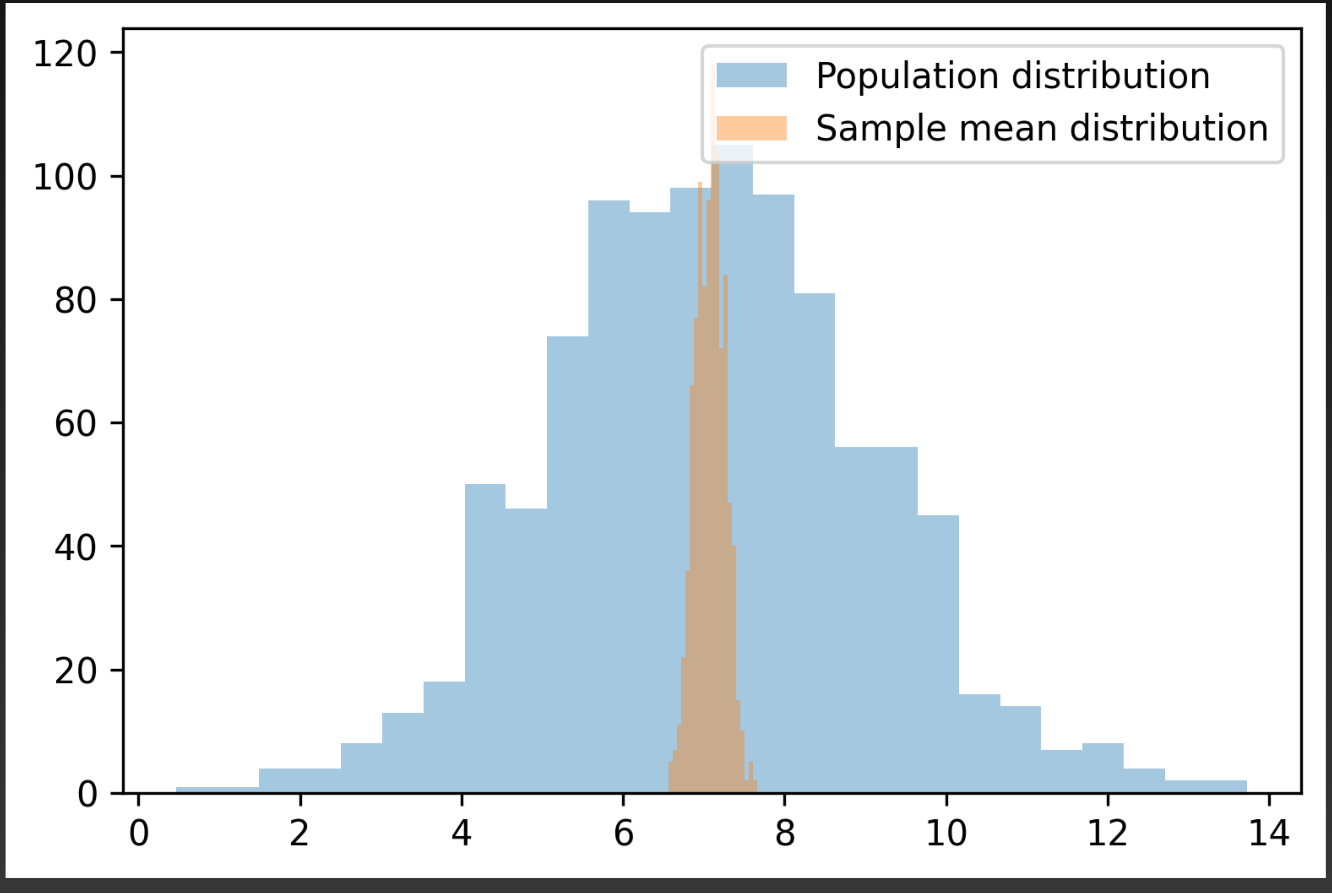
Population parameters: Mean = 7, Std = 2

Draw sample of 100 from population

Sample Parameters: mean = 6.95, Std = 1.94

Draw 1000 samples of 100 each from population

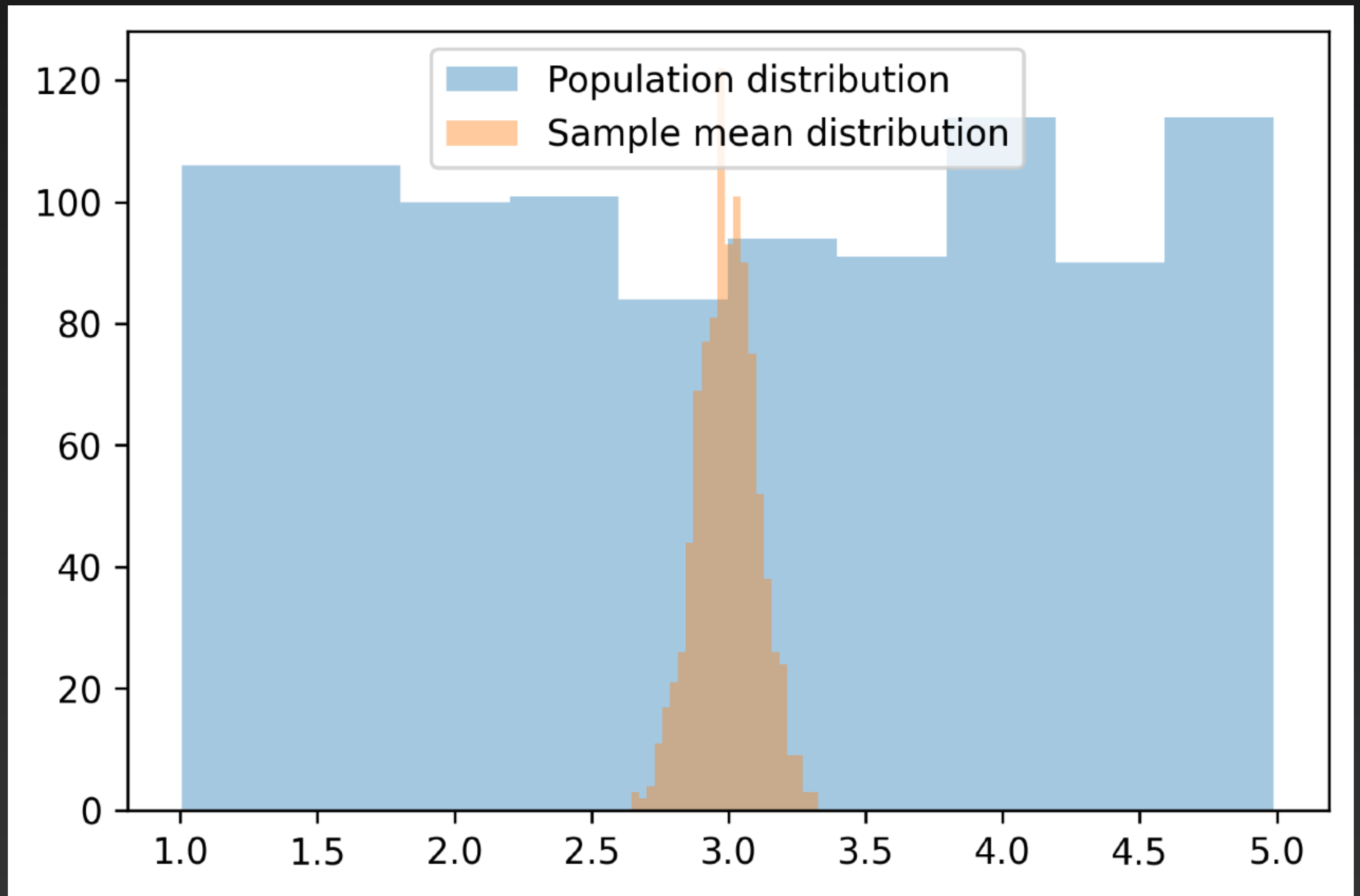
Infered parameters: Mean= 7.077 +/- 0.184

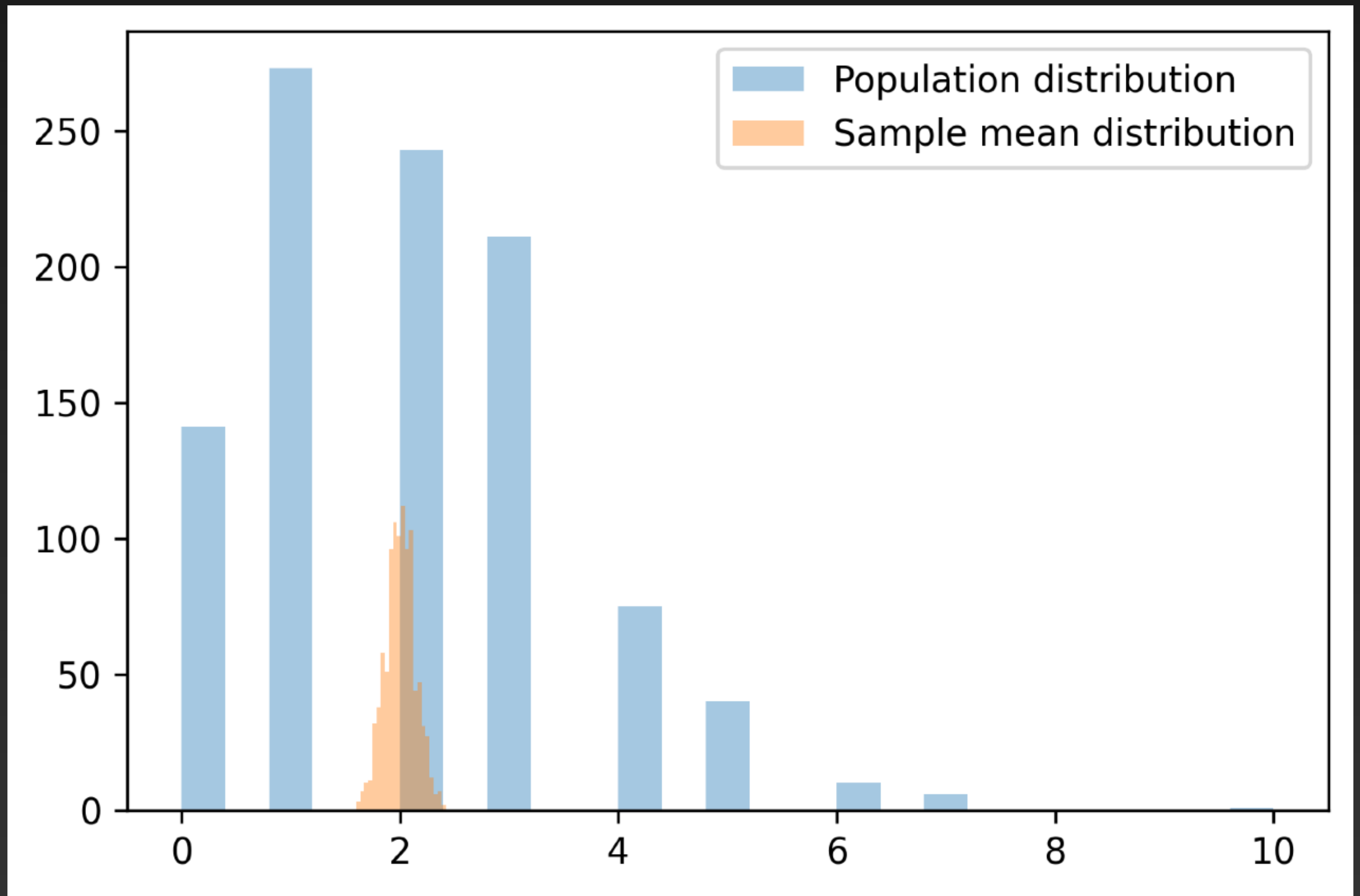


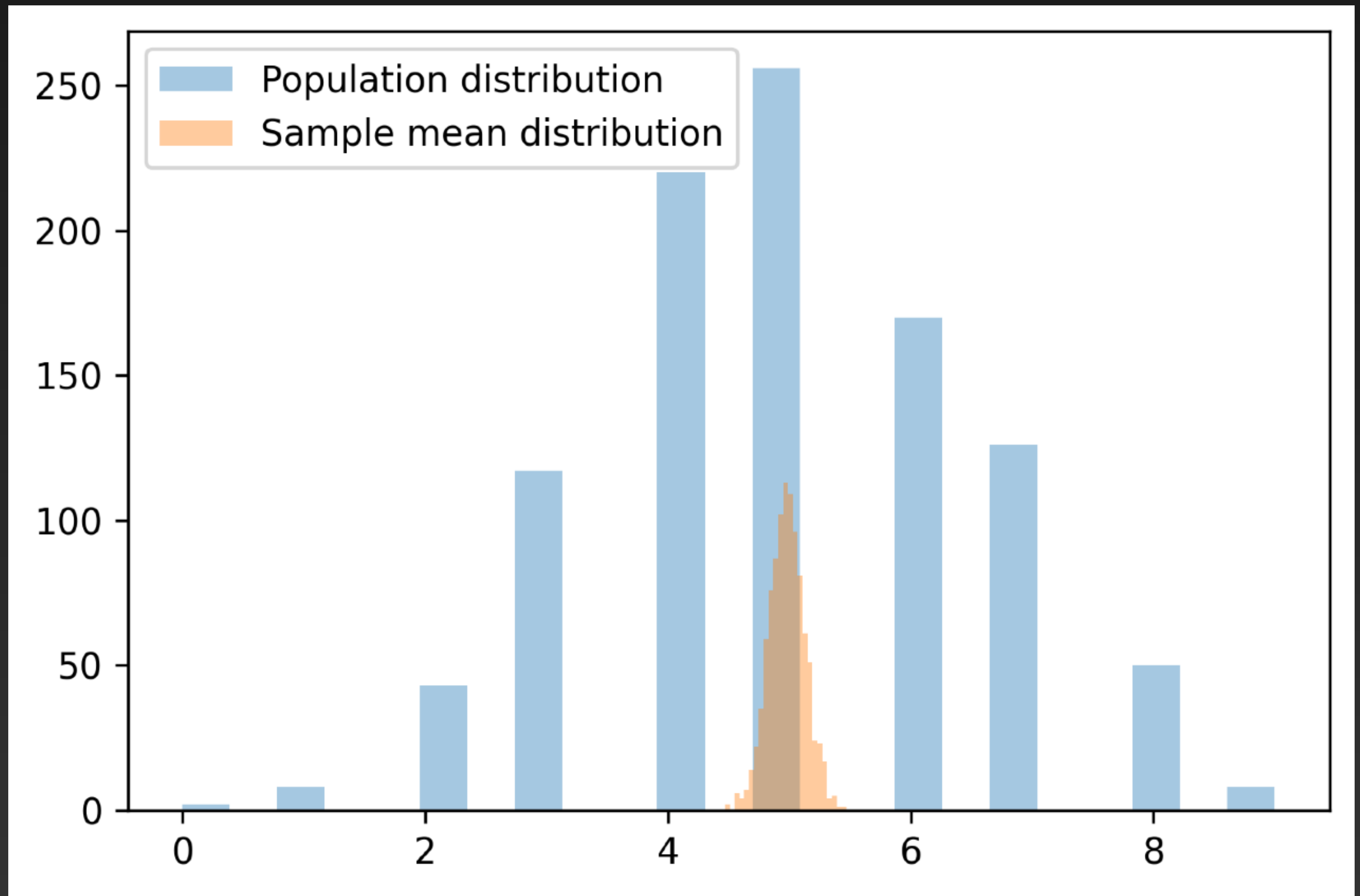
Central limit theorem (CLT)

As the sample size becomes larger, the sampling distribution of mean is approximated by normal distribution, even if the data in each sample is not normally distributed

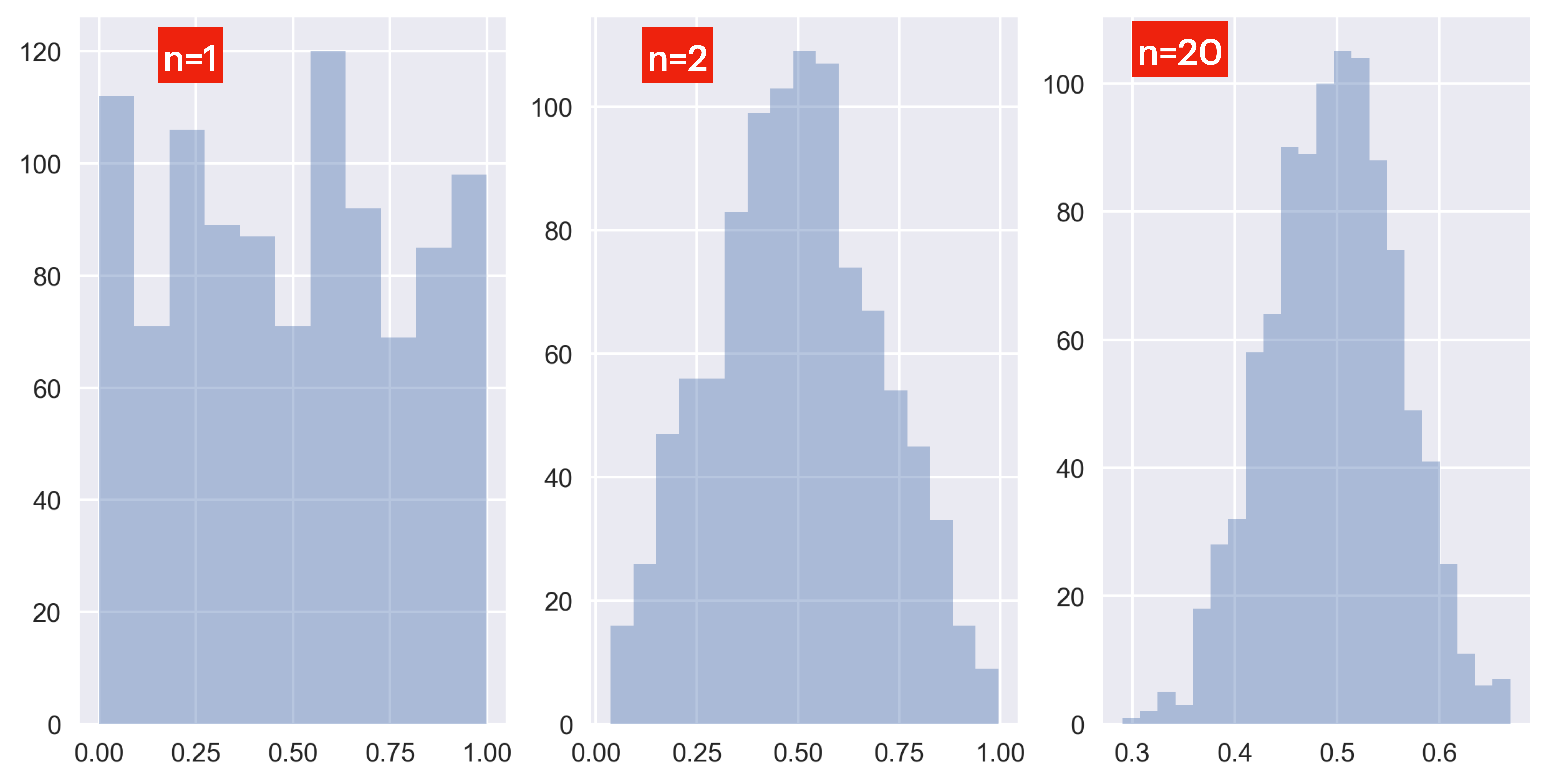
Generally, a sample of 30 or more is considered okay for CLT to hold







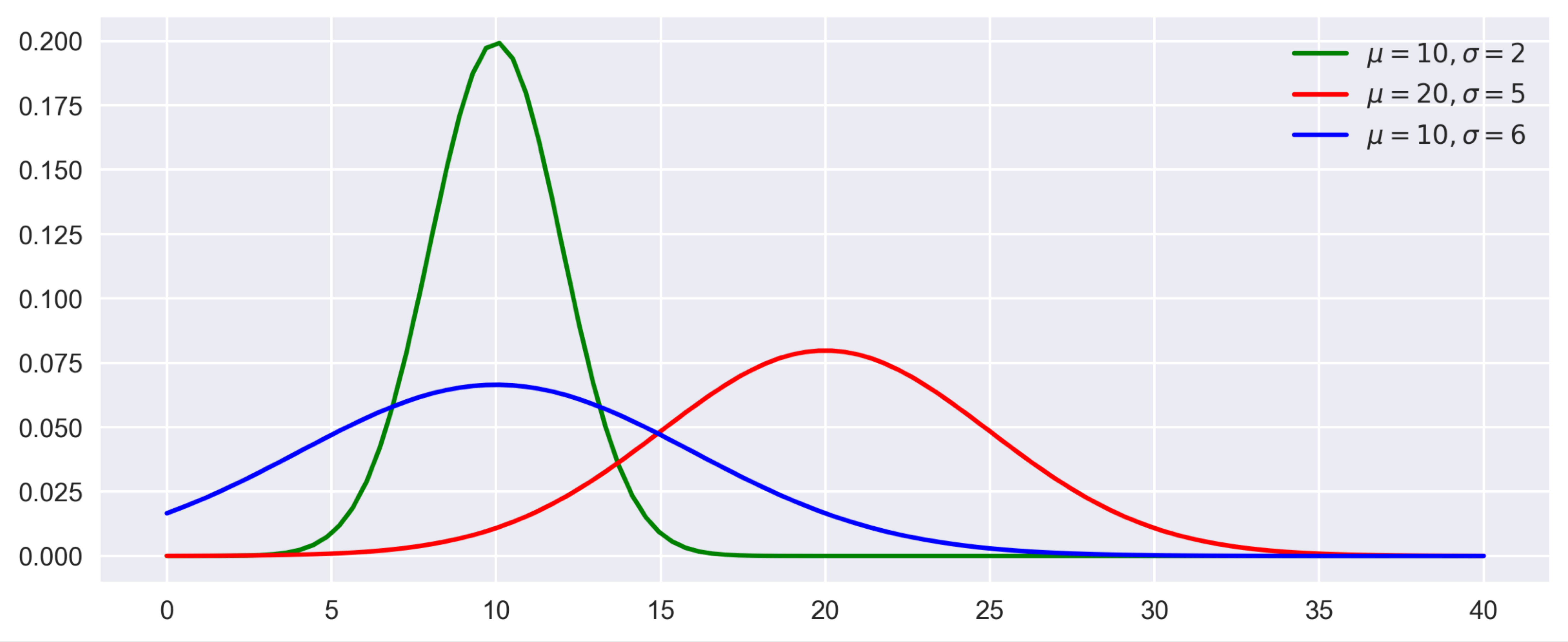
Example of CLT for uniform distribution



Normal distribution

The standard normal random variable is a continuous variable whose probability density function is given by:

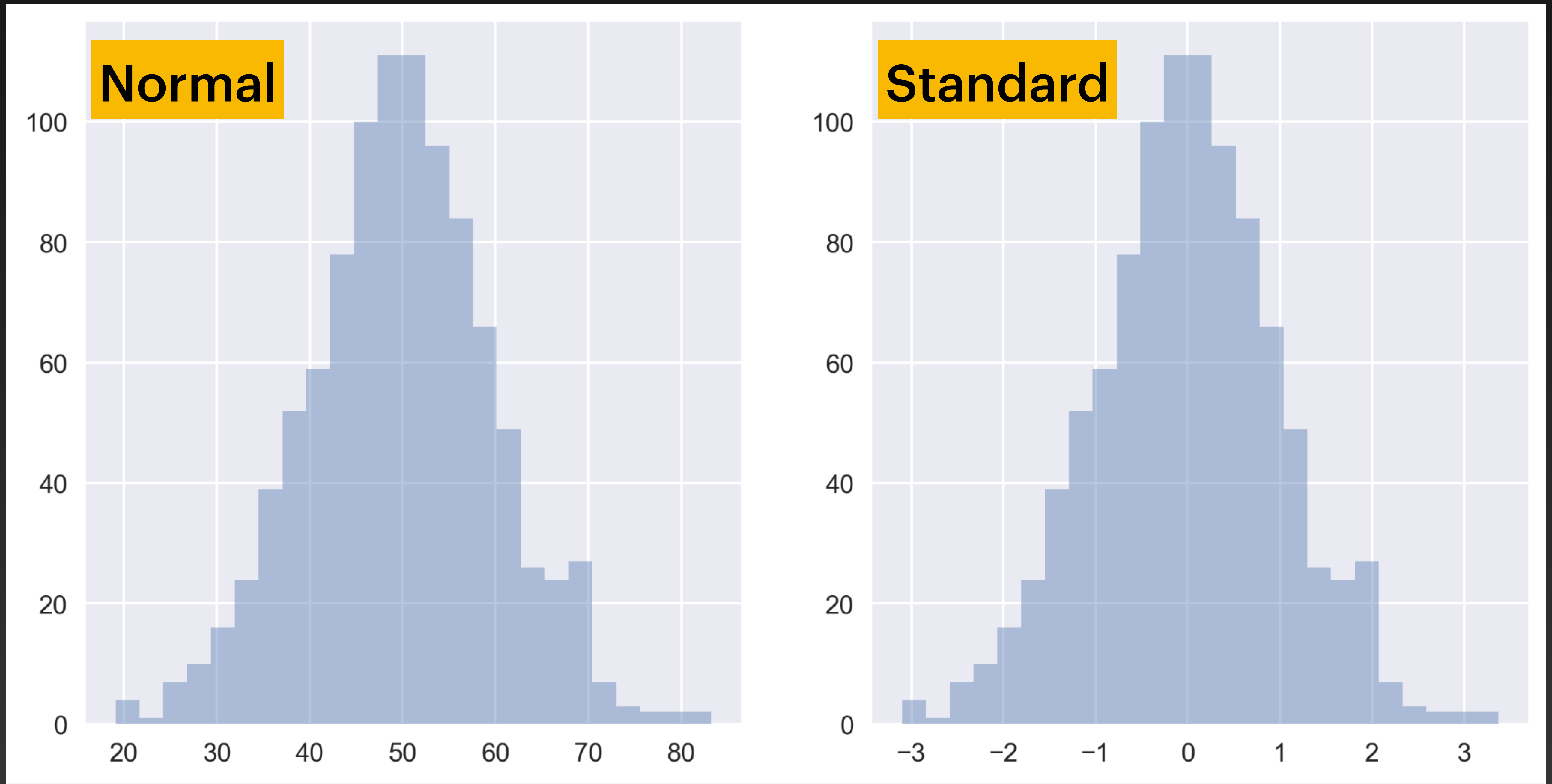
$$P(y)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$



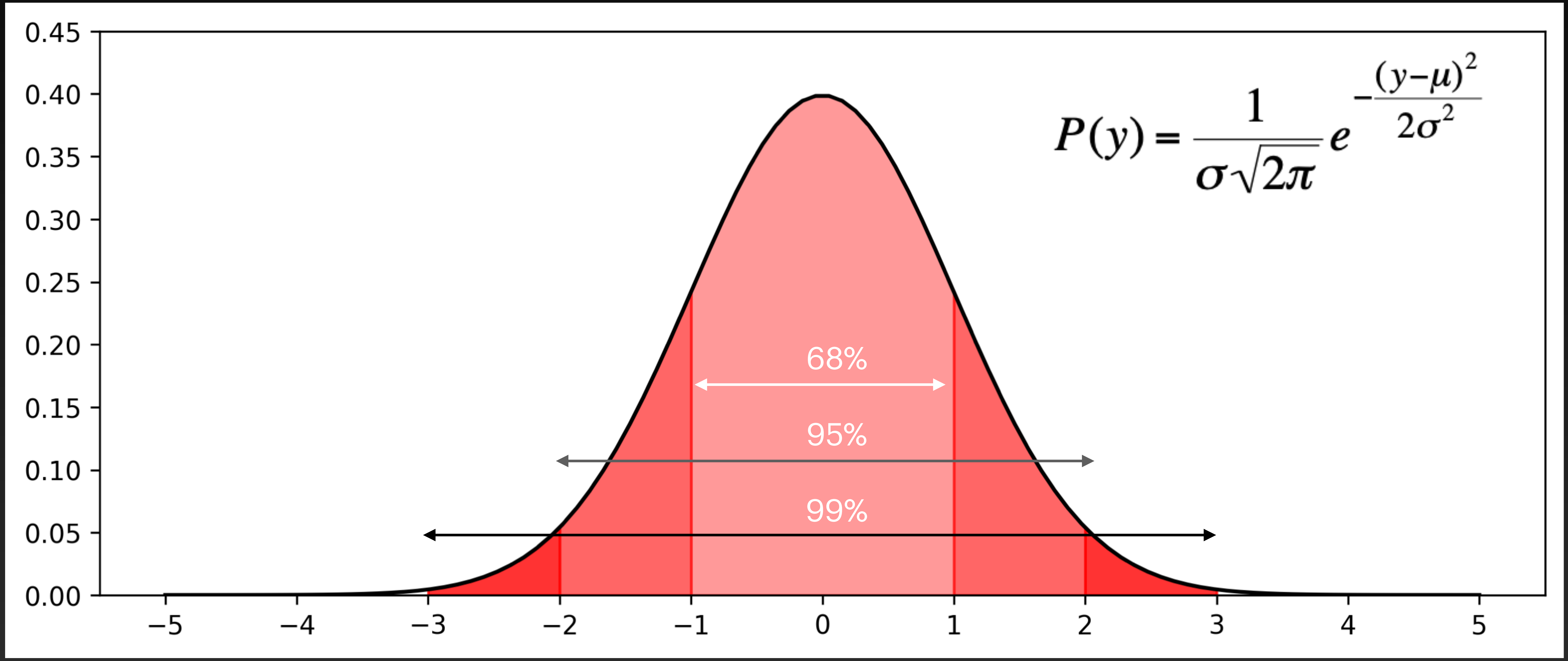
Standard normal distribution (SND)

The standard normal distribution is a distribution with mean 0 and standard distribution 1

$Z = (X - \mu) / \sigma$



Applications of SND and CLT



$$P(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Example: Flip a fair coin 100 times, estimate the probability to get more than 60 heads

Expectd mean = 50 and std = 5

P(H > 60) = P((60-50)/5) = P(Z > 2) ~ 0.023

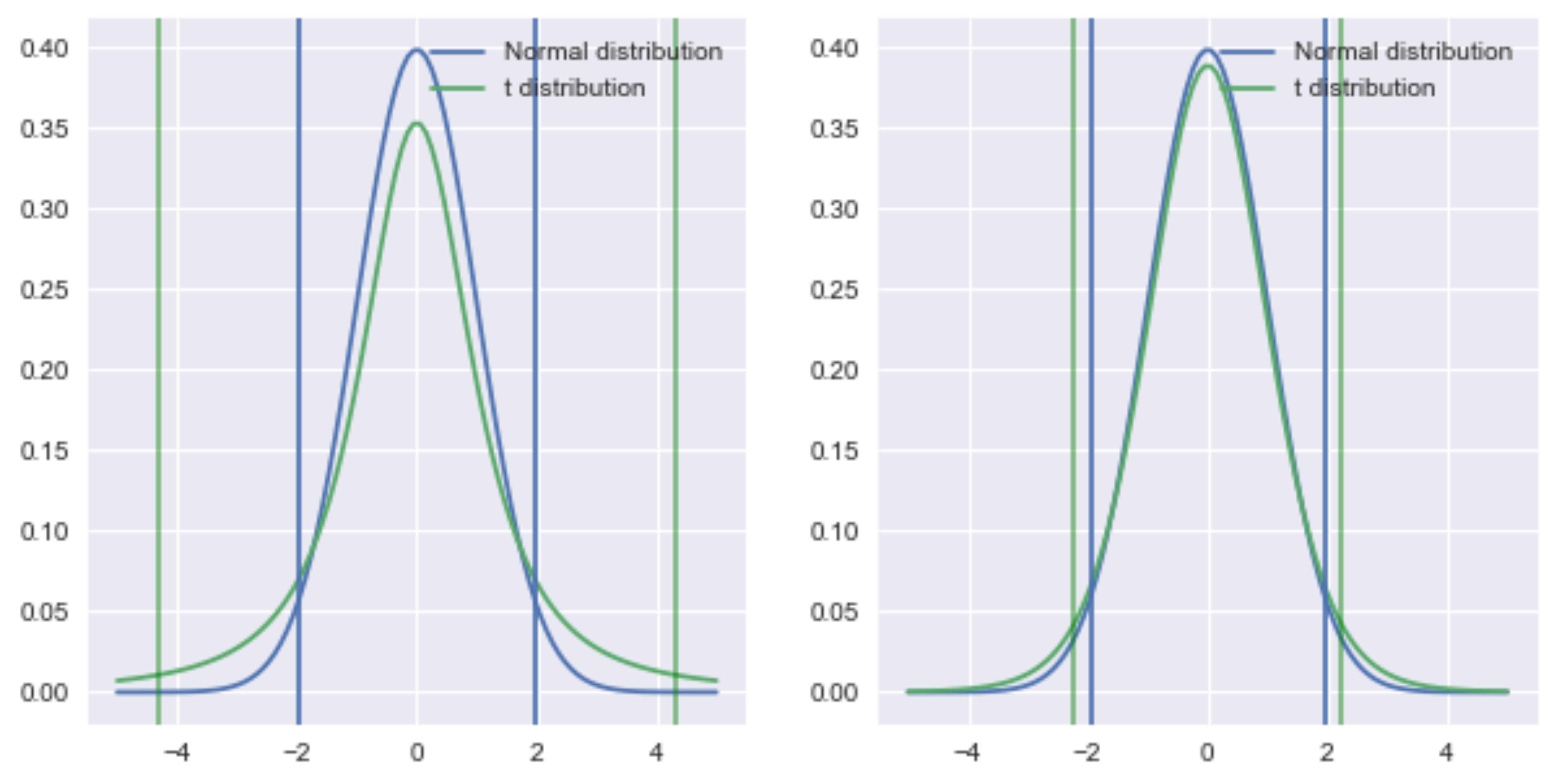
Confidence interval

Confidence interval is defined as the range of values estimated from sample which contain the true unknown parameter of the population

if mean = 122, sample_size=25, sigma=20
95% confidence interval [114.16, 129.84]

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

If sigma is known, use normal distribution to estimate confidence interval
If sigma is unknown, use t-distribution



Hypothesis testing

Hypothesis testing is the process of examining whether the measurement of given statistics such as mean, variance, is consistent with its theoretical distribution

Example 1 : If I toss a coin 100 times and get 70 heads and 30 coins, I might ask the question whether my coin is fair or not?, and at what confidence level (quantitatively) I can answer this question

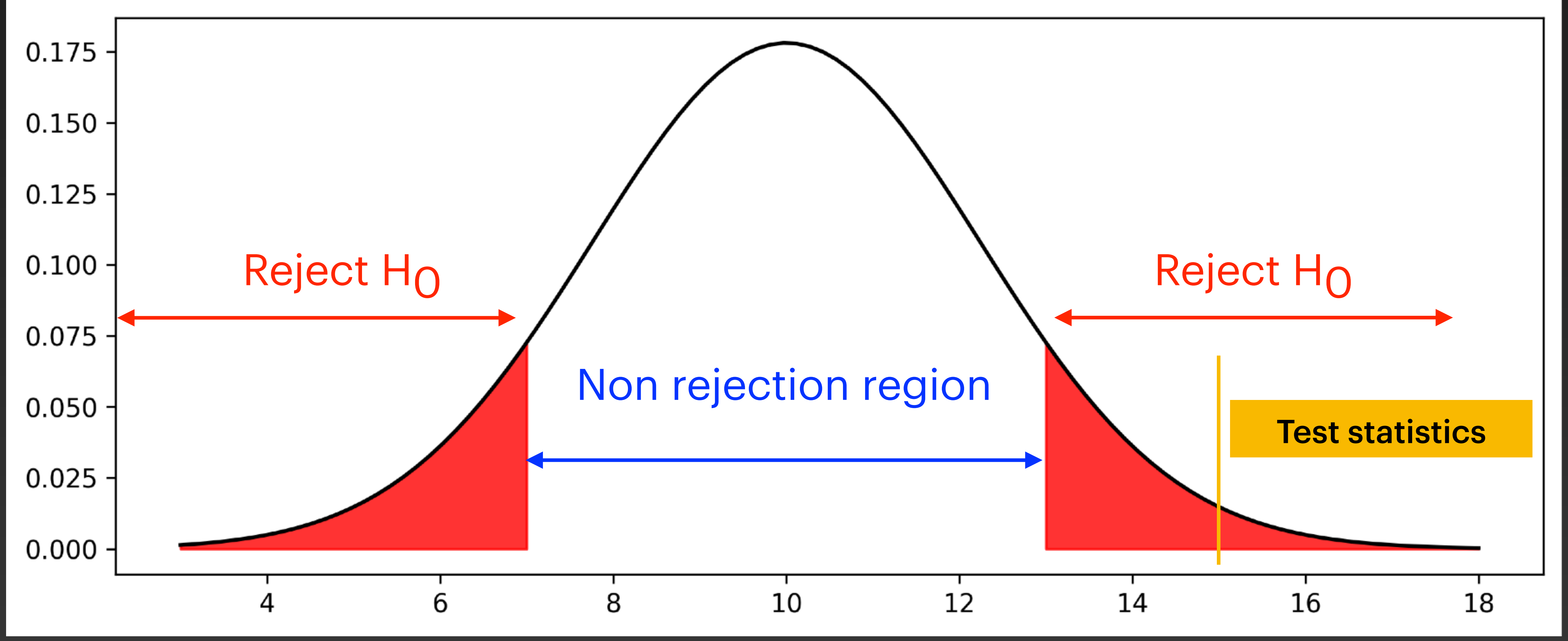
Example 2 : lets say I want to quantify if doing physical exercise reduce the risk of heart diseases, I can use the hypothesis testing framework

Example 3 : Lets say I am a quality control inspector and go to a pipe factory. Factory claim that each pipe has a length of 10 cm. If I want to verify their claim, I can use hypothesis framework. For example, I can collect samples of their pipes and create a statistics to verify their claim.

Ingredients of hypothesis testing

1. Begin with a hypothesis to test, called null hypothesis : H_0
2. Determine the test statistics to use for null hypothesis. For example, sample mean can be a test statistics
3. Collect sample and calculate test statistics
4. Determine the probability (p-value) for getting the test statistics under the assumption that null hypothesis is true
5. Based on the P-value, we can decide whether to accept or reject the null hypothesis

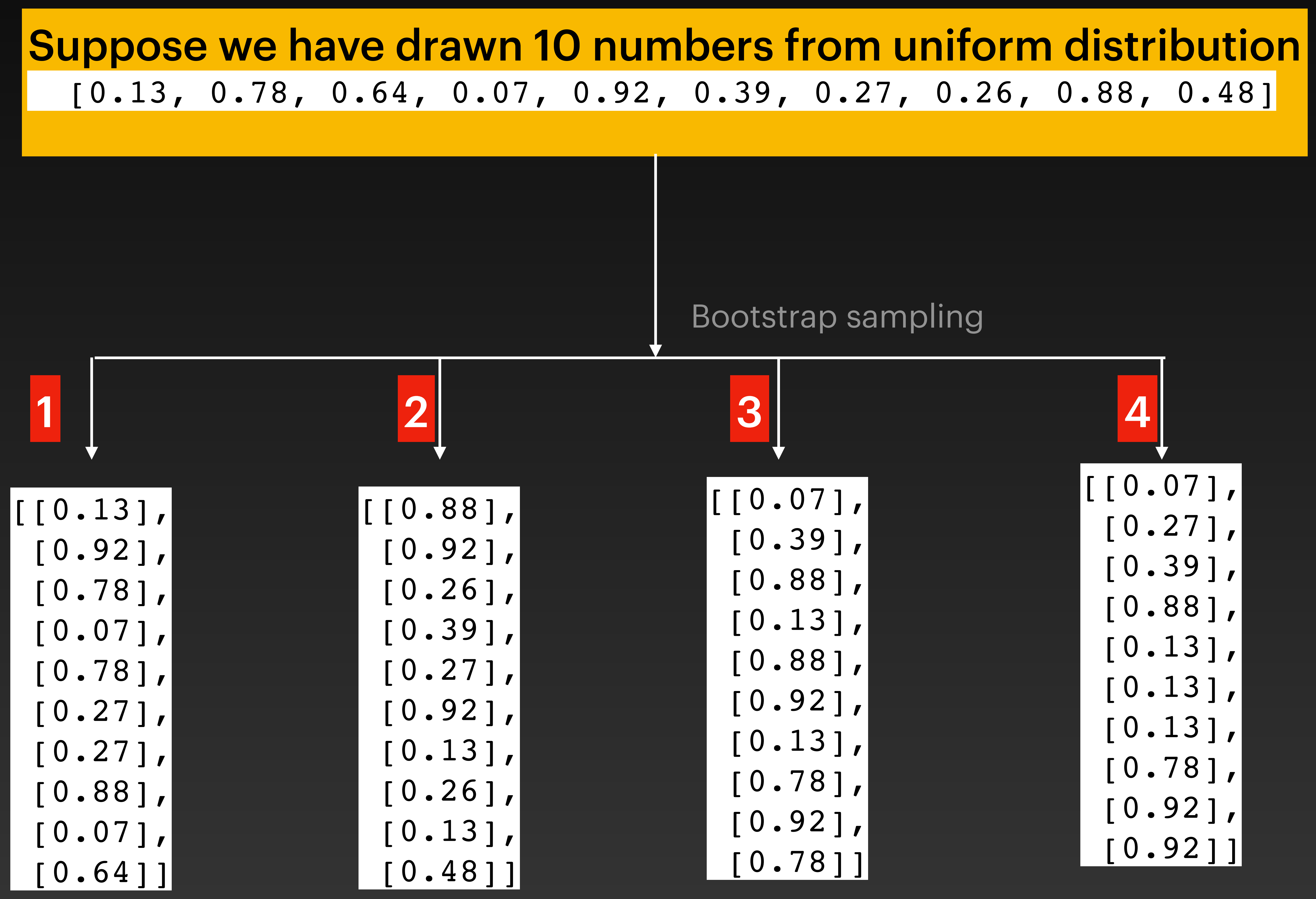
Example: Tossed a coin 20 times and get 15 heads
Is our coin fair or not?



Bootstrap method

- Bootstrap method is a statistical technique to compute statistics from “sample data” by sampling from the same data
- Only feasible due to modern computing power
- Confidence interval for sample mean is calculated as
- $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$, this is true if we know the distribution
- If data is drawn completely from an unknown distribution, then what?
- Bootstrap method will help

Bootstrap example



Suppose we have drawn 10 numbers from uniform distribution

```
[0.13, 0.78, 0.64, 0.07, 0.92, 0.39, 0.27, 0.26, 0.88, 0.48]
```

Bootstrap sampling

1

```
[[0.13],
 [0.92],
 [0.78],
 [0.07],
 [0.78],
 [0.27],
 [0.27],
 [0.88],
 [0.07],
 [0.64]]
```

2

```
[[0.88],
 [0.92],
 [0.26],
 [0.39],
 [0.27],
 [0.92],
 [0.13],
 [0.26],
 [0.13],
 [0.48]]
```

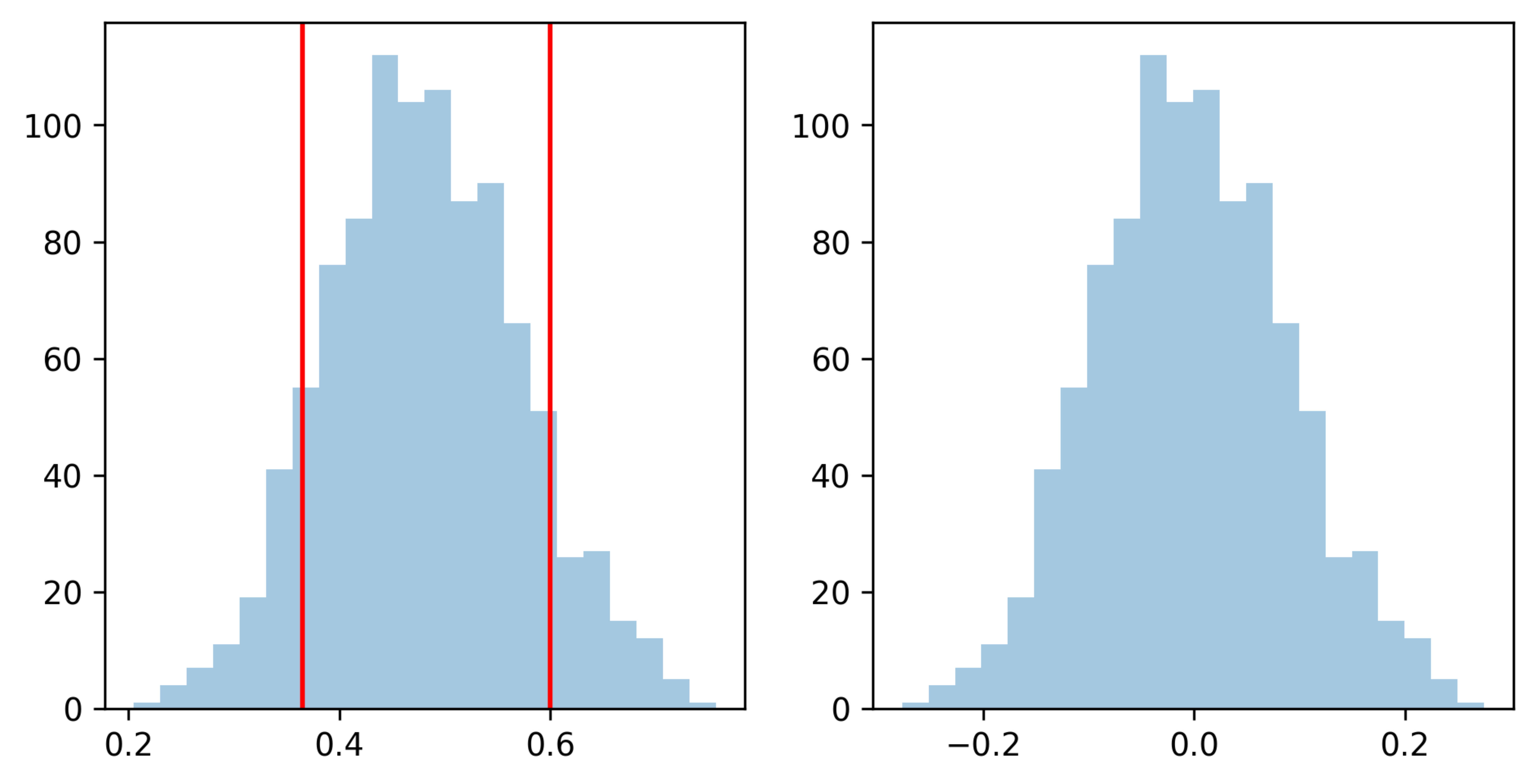
3

```
[[0.07],
 [0.39],
 [0.88],
 [0.13],
 [0.88],
 [0.92],
 [0.13],
 [0.78],
 [0.92],
 [0.78]]
```

4

```
[[0.07],
 [0.27],
 [0.39],
 [0.88],
 [0.13],
 [0.13],
 [0.13],
 [0.78],
 [0.92],
 [0.92]]
```

Bootstrap confidence level



```
80% confidence interval [0.365, 0.600]
```