Data Distribution

Statistics Tutorial

Day 4

Prabesh Dhakal

2020 April 30

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

EXERCISE

We continue where we left off last week.
&
Do some R exercises.

DATA DISTRIBUTION

We talk about how data distribution.

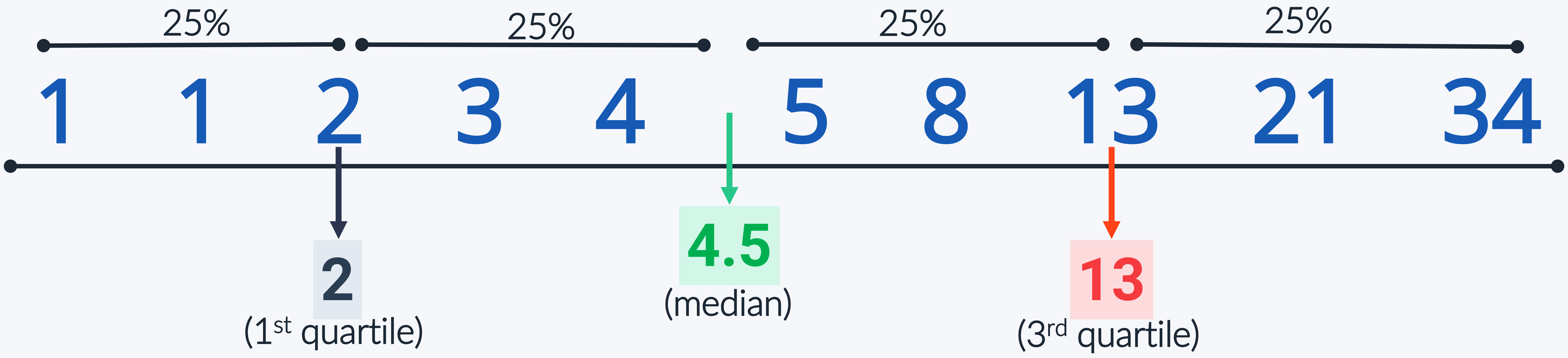
Q&A and Recap

Please ask if you have any questions now.

Otherwise, we can move on to the recap.

Note: you might want to grab a pen, paper & calculator for today’s session.

SUMMARIZING DATA



25% 25% 25% 25%

1 1 2 3 4 5 8 13 21 34

2

(1st quartile)

4.5

(median)

13

(3rd quartile)

$$mean\ (\bar{x}) = \frac{1}{N}\sum_{i=1}^{10} x_i = \frac{1+1+2+3+4+5+8+13+21+34}{10} = 9.7778$$

$$Range = l\ - s = 34\ - 1 = 33$$

$$outlier = \text{values that are} \begin{cases} < \ Q_1 - 1.5\ * IQR\ \ or \\ > \ Q_3 + 1.5\ * IQR \end{cases}$$

$$IQR\ = Q_3 - Q_1 = 13\ - 2 = 11$$

Exercise

Calculate variance and standard deviation.

Use R for simple data analysis.

🛠 CLASS EXERCISE - 1 🛠

Continue with calculation of variance. (pen and paper)

Do some R. (RStudio)

Data Distribution

Discuss different types of data distribution

Talk about normal distribution and why it is important

Box plots and Outliers

DISTRIBUTION OF THE DATA

1. **What?**
   - An arrangement of values of a variable showing their observed or theoretical frequency of occurrence
2. **Why?**
   - Shows how frequent each value is in a given data set
   - Enables us to get a better sense of the data than what just the numbers in the tables suggest
3. **How?**
   - *Bar charts / histograms / density plots / box plots*

PROBABILITY DISTRIBUTION

Probability distribution

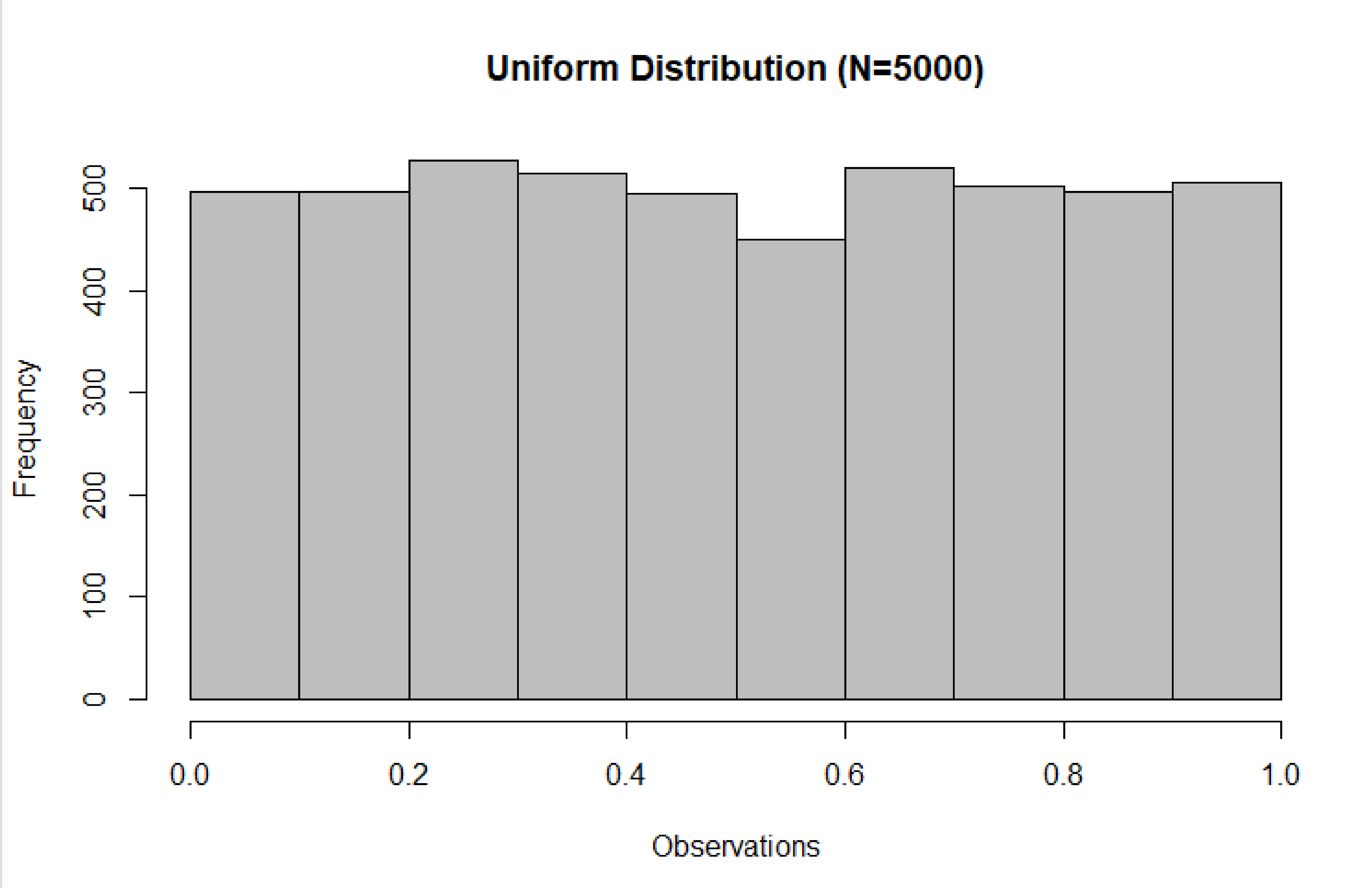
A statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.

Random Variable

A variable whose value is the outcome of a random event.

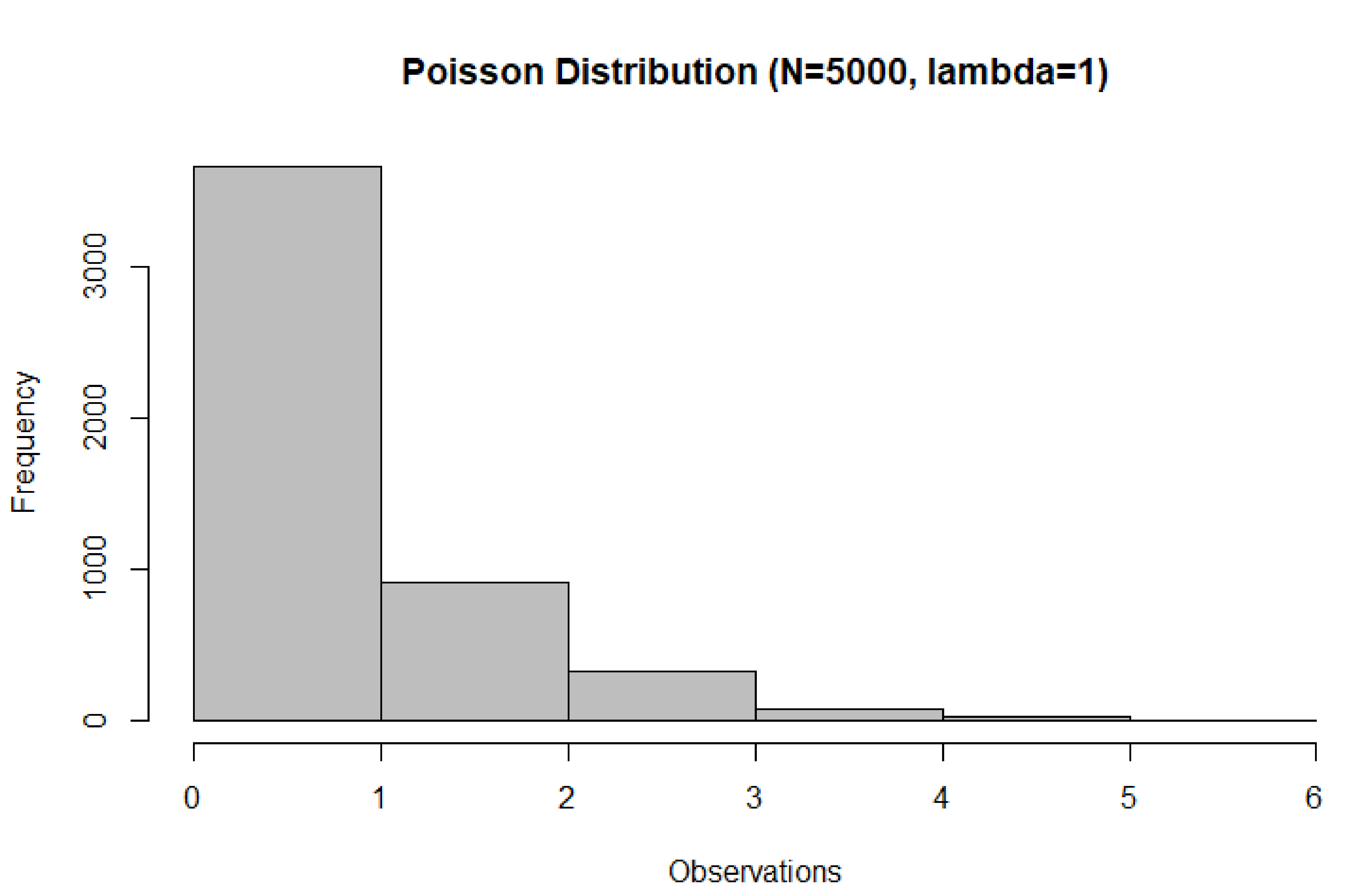
UNIFORM DISTRIBUTION

- Signify probability distribution with equally likely outcomes
- Looks (relatively) flat



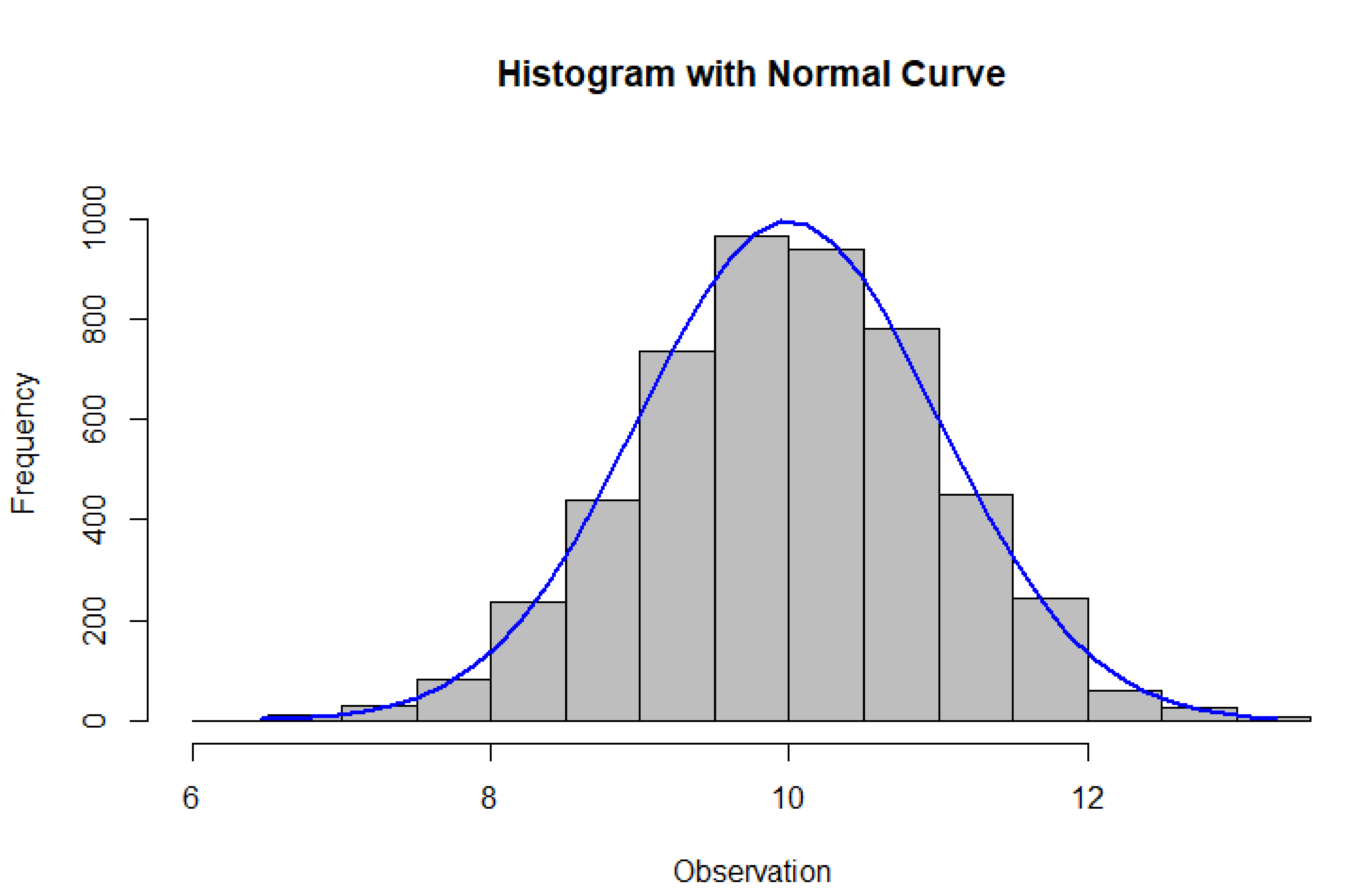
POISSON DISTRIBUTION

- expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time, or space since the last event



NORMAL DISTRIBUTION

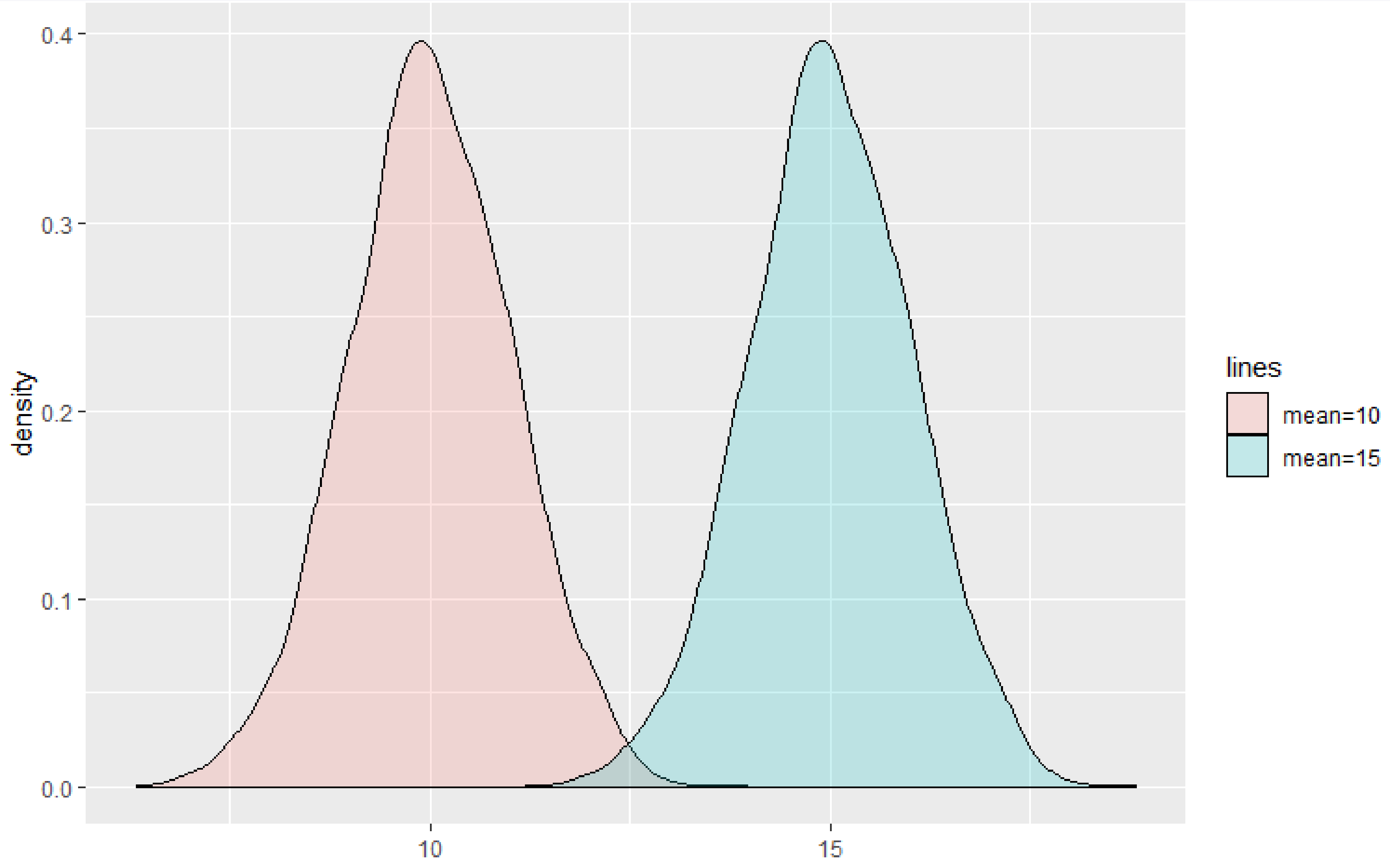
- Most values lies close to the mean
- Variance governs the spread of the values
- Symmetric, but can also be skewed



NORMAL DISTRIBUTION

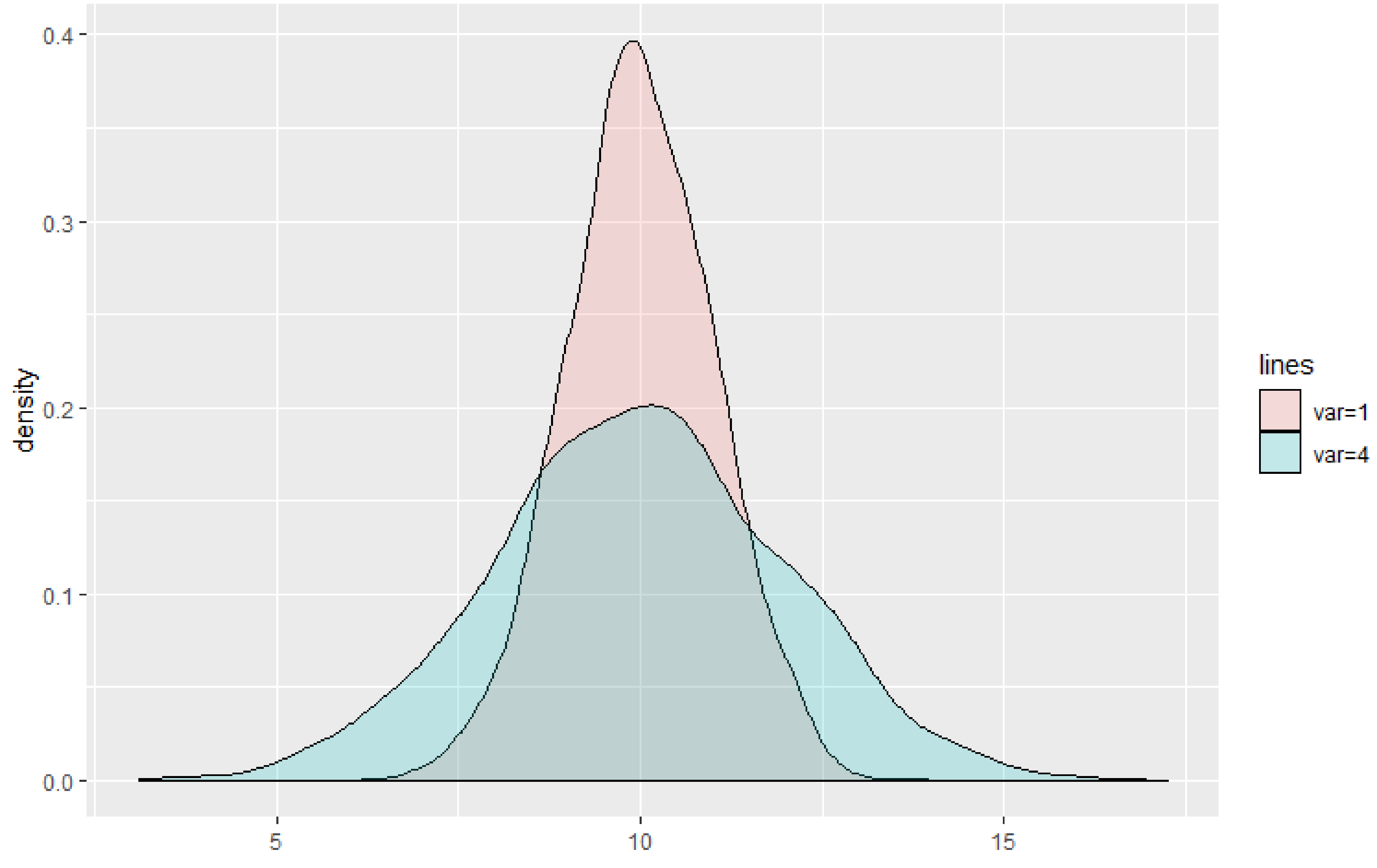
- Parameters: mean and variance

Difference in the mean



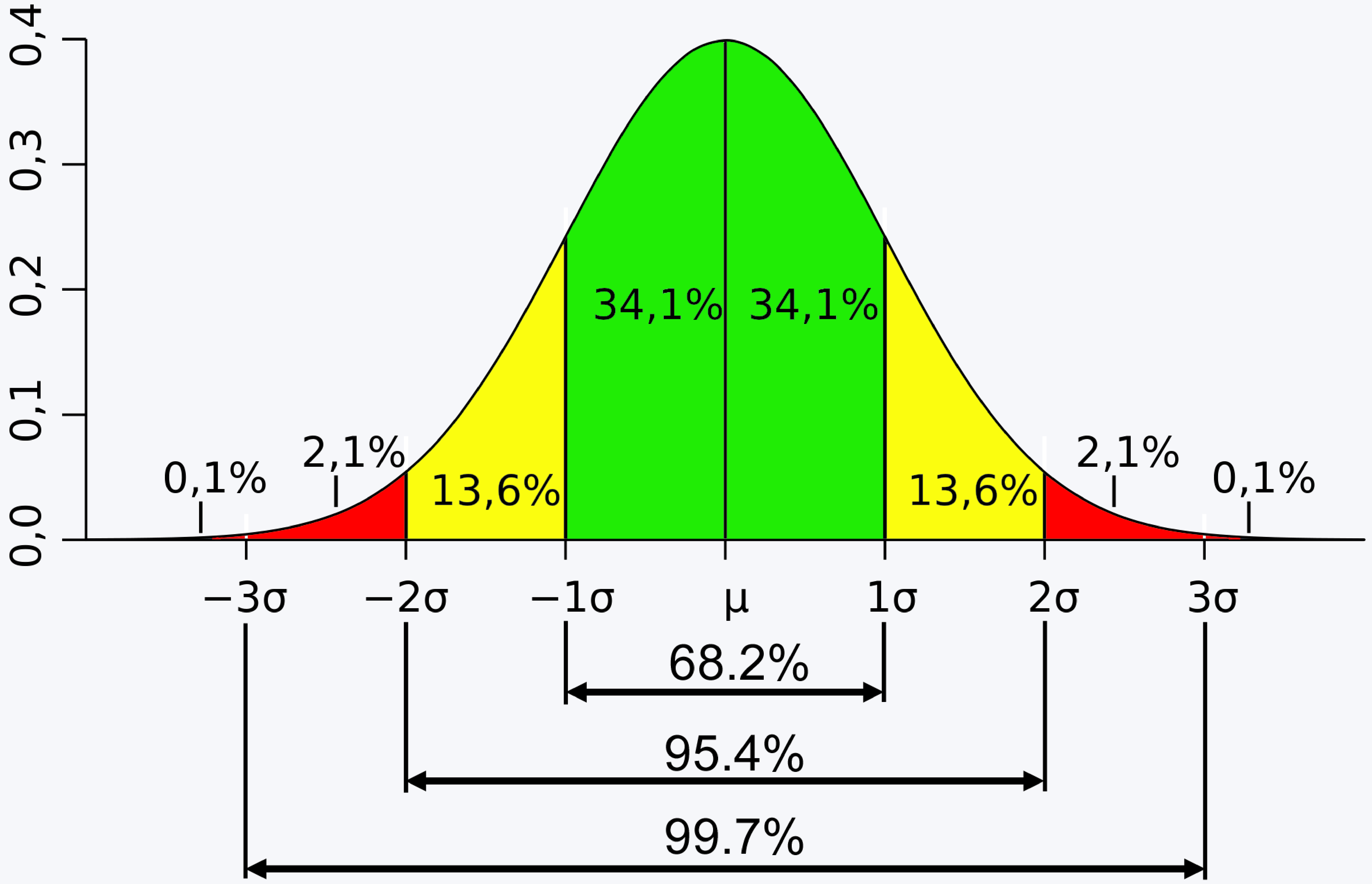
$$\bar{x} = \frac{\sum x}{N}$$

Difference in the variance



$$\sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{N}$$

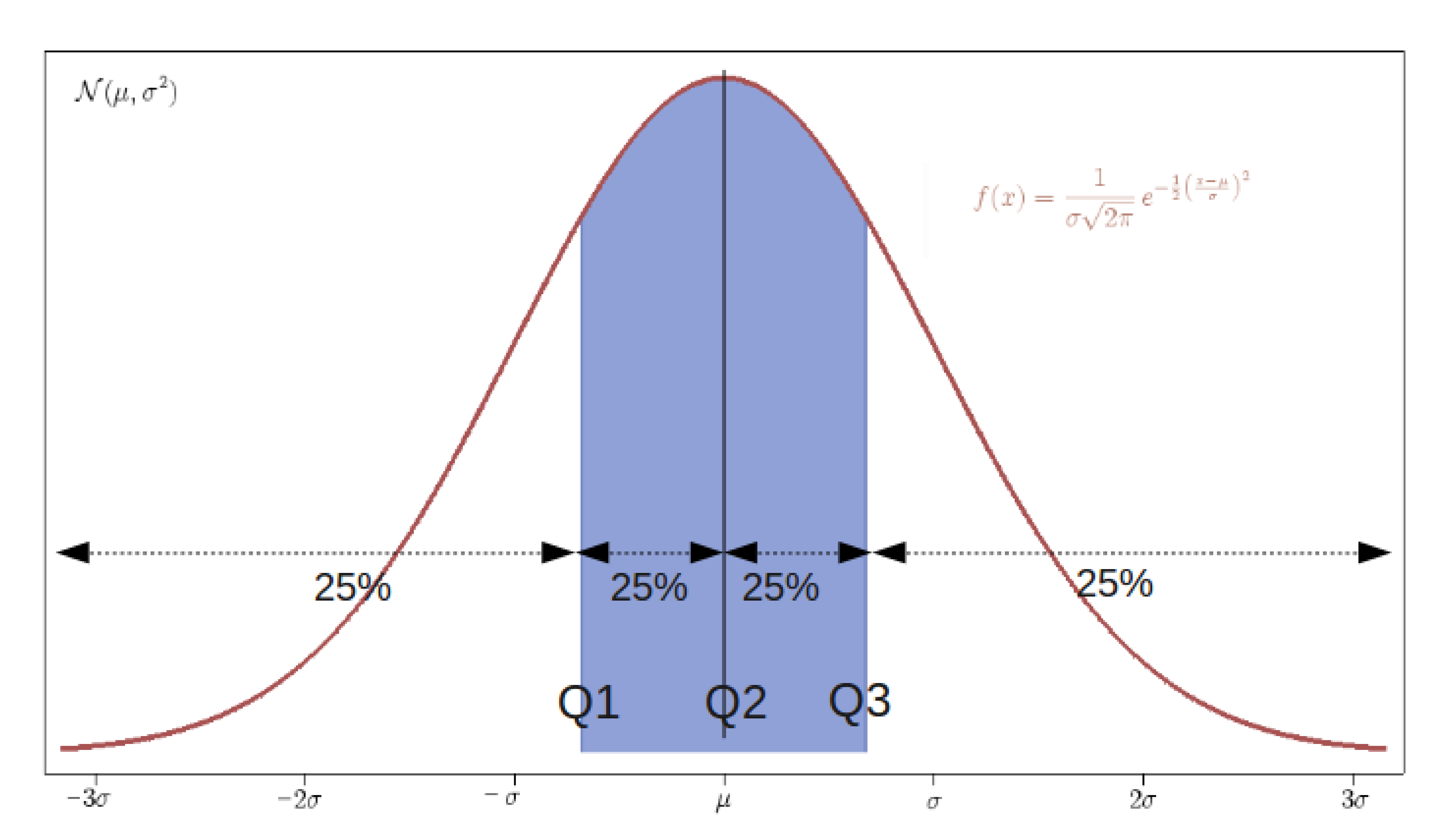
NORMAL DISTRIBUTION



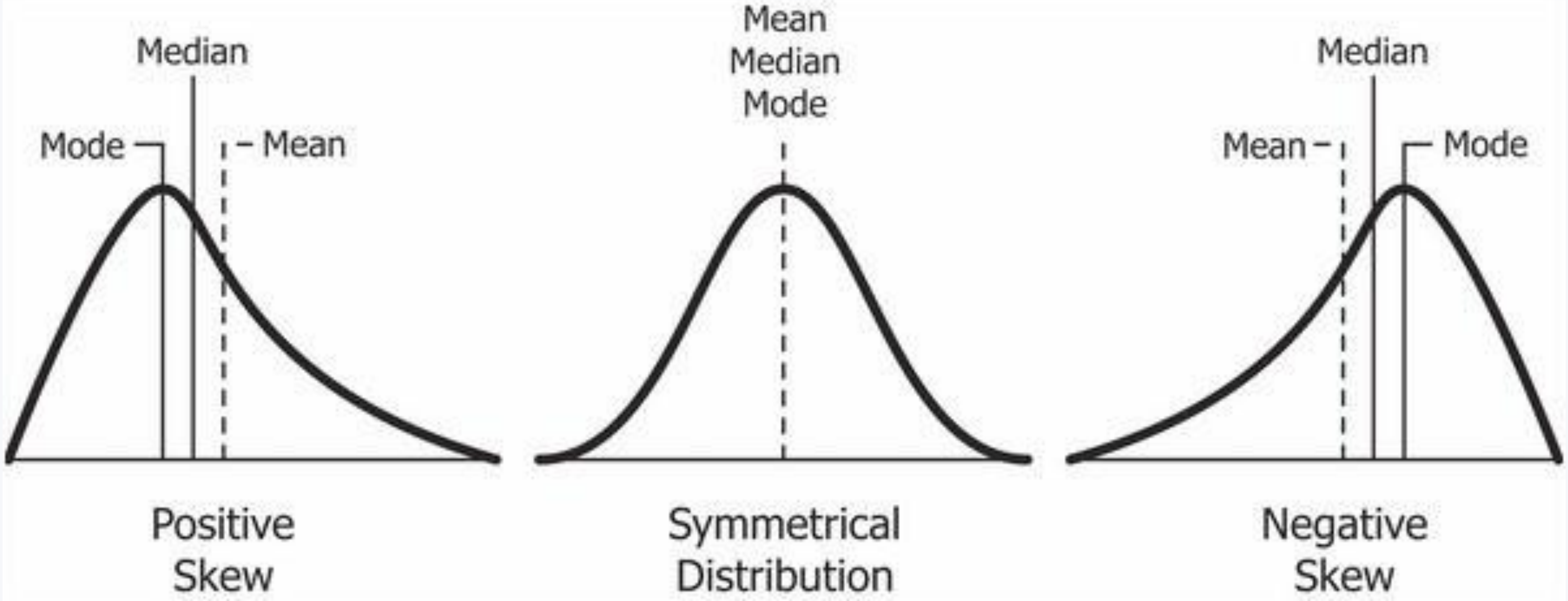
More on normal distribution: https://www.mathsisfun.com/data/standard-normal-distribution.html

NORMAL DISTRIBUTION

(IN CONTEXT OF QUARTILES)



SKEWED DISTRIBUTIONS



(or Right Skew)

(or Left Skew)

STANDARD NORMAL DISTRIBUTION

The Normal Distribution

Z-Score, Standardization, Standard Normal Distribution

Reading:
Standard Normal Distribution

PLAN FOR NEXT WEEK

That’s it for today! :-)

Next week, we are going to discuss:

- A little more on Normal Distribution
- Hypothesis Testing,
- Chi-squared Test

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`