LECTURER: TAI LE QUY

INTRODUCTION TO DATA SCIENCE

INTRODUCTION TO DATA SCIENCE
TOPIC OUTLINE

Introduction to Data Science	1
Data	2
Data Science in Business	3
Statistics	4
Machine Learning	5
Summary session	6

UNIT 4

STATISTICS

STUDY GOALS

- Identify the importance of statistics in data science.
- Know about probability and its relation to the prediction model’s outputs.
- Learn about conditional probability and the probability density function.
- Understand the different probability distributions.
- Know the Bayesian statistics.

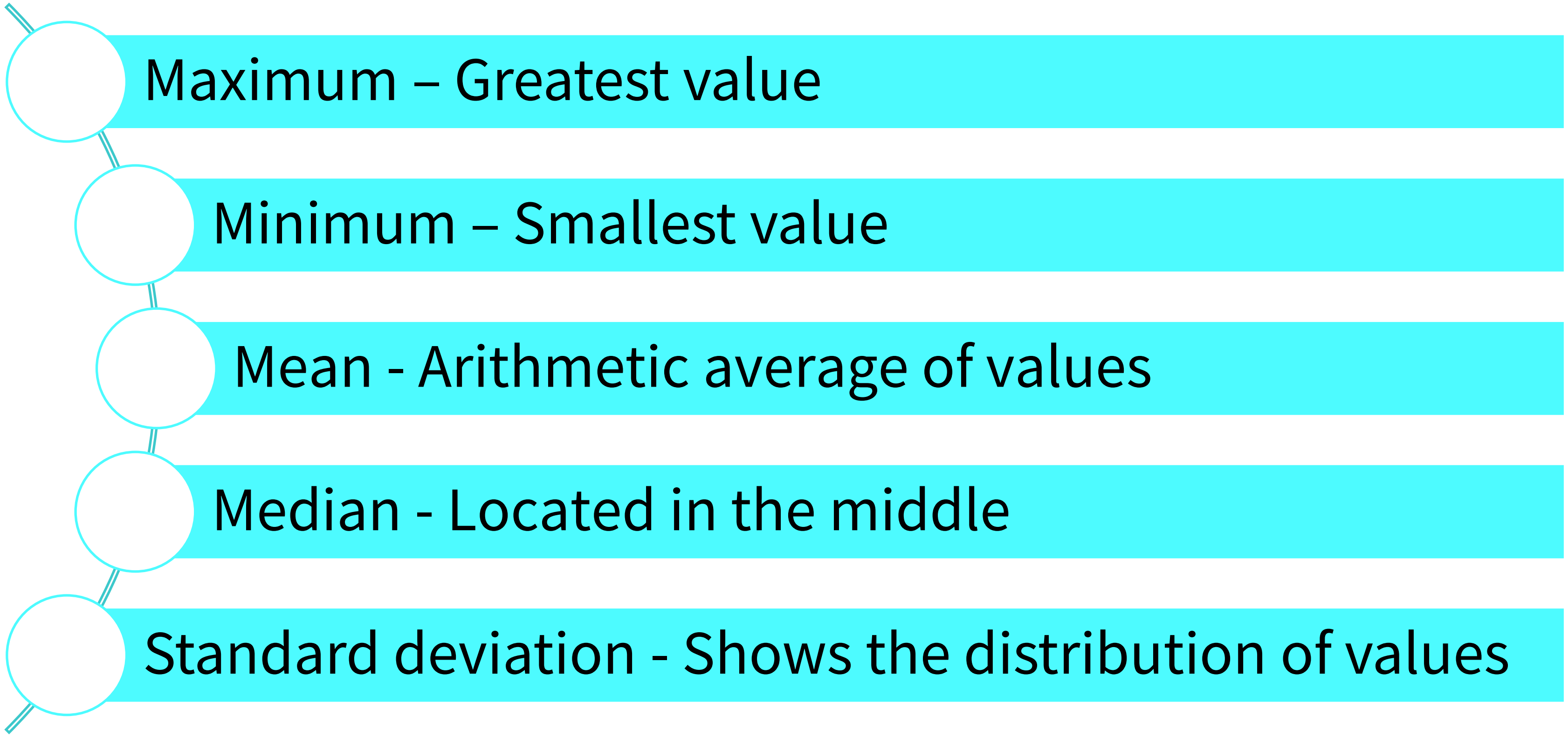
EXPLAIN SIMPLY

- What is the role of statistics in Data Science?
- What are the importance statistical parameters?
- What are the important statistical concepts?

Statistics can help to …

- … extract the main properties of a dataset.
- … summarize the observations.
- … reduce a large dataset to smaller statistics.
- … consider the likelihood of possible events.
- … describe almost all realistic systems.

Important statistical parameters



- Maximum – Greatest value
- Minimum – Smallest value
- Mean - Arithmetic average of values
- Median - Located in the middle
- Standard deviation - Shows the distribution of values

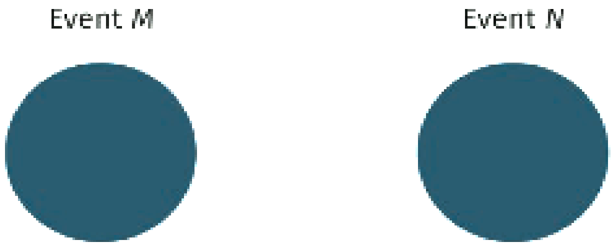
- **Probability** – The likelihood that an event will happen
  - $0 \leq P \leq 1$
  - $P = 0$: It is impossible for the event to occur.
  - $P = 1$: The event will definitely occur.

- **Probability theory** – Core theory for many Data Science techniques

Source of the text: Zöller (2020).

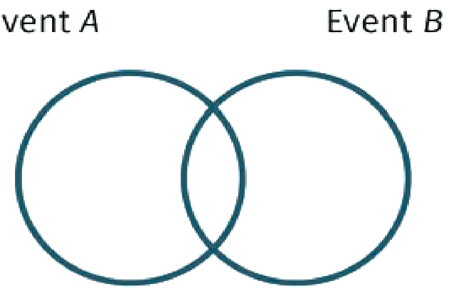
Mutually exclusive events:

events cannot occur at the same time



Multi independent events:

events can occur simultaneously without affecting each other



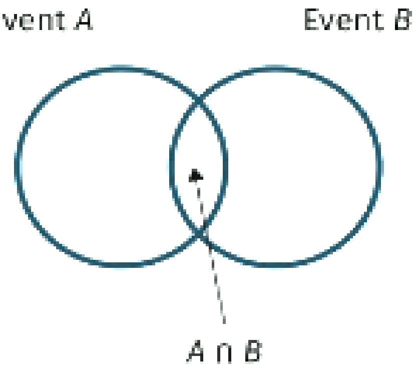
$P(A \; and \; B) = P(A \cap B) = P(A) \cdot P(B);$

$P(A \; or \; B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Multi conditional probability:

events are correlated

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



Source of the images: Zöller, 2020.

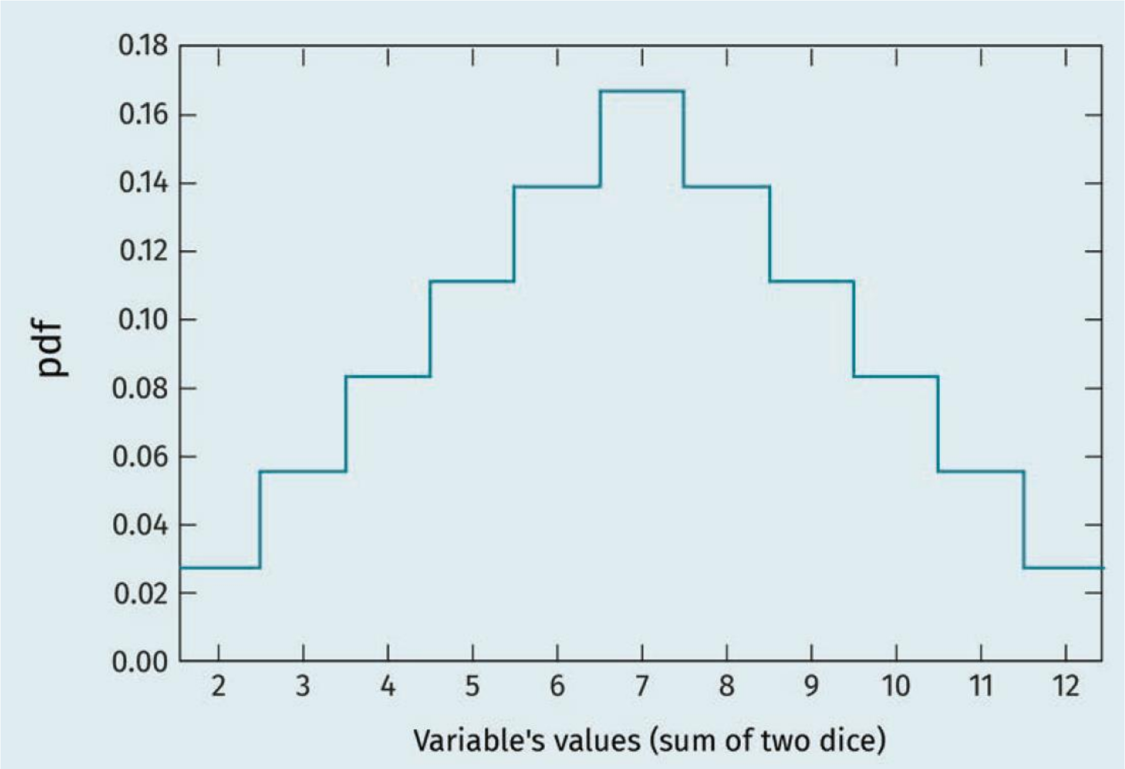
Probability distribution:

- A random variable can take on a given set of values.
- The **occurrence** of each of these values has a certain **probability**.

Probability distribution function

maps outcomes with their respective probability

- **X-axis**: possible values of the variable
- **Y-axis**: probability of each value

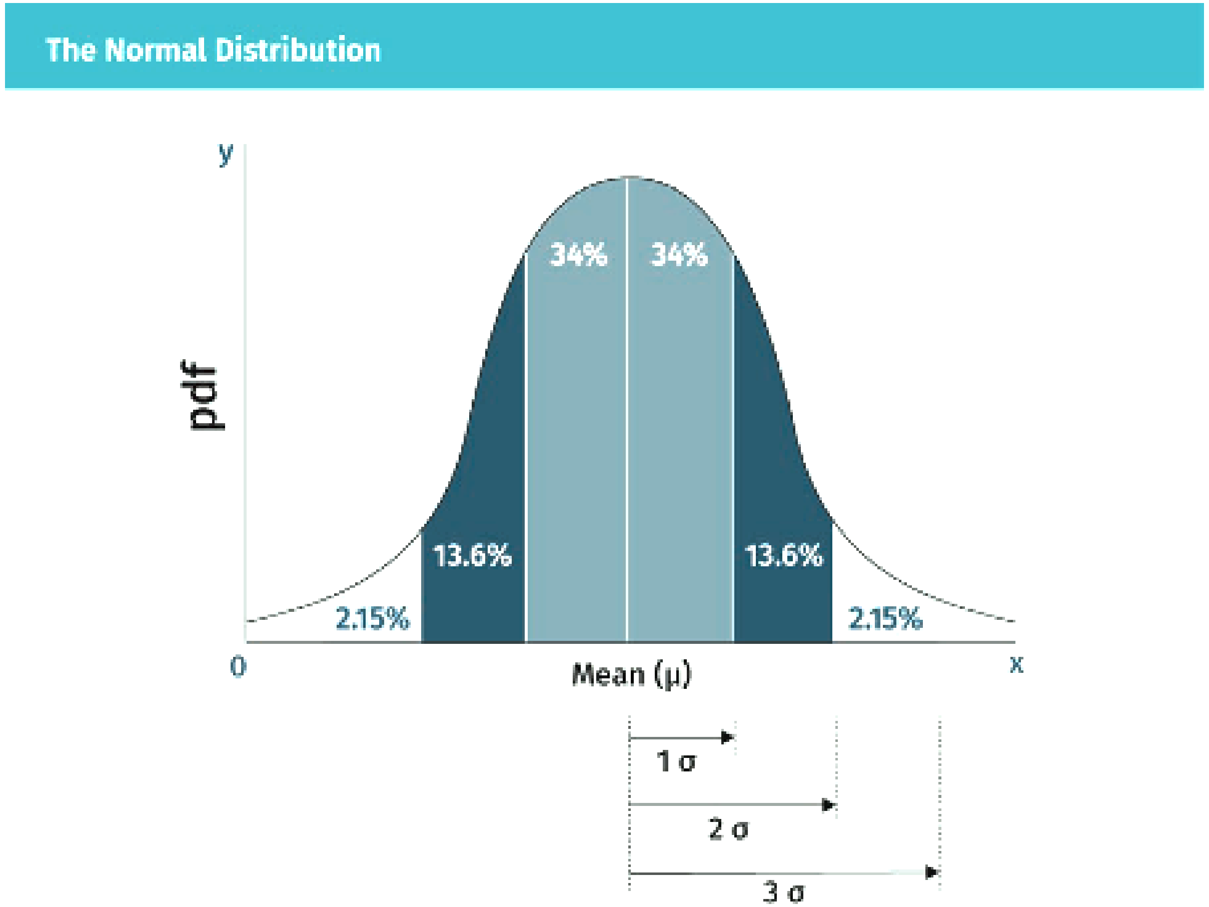


Source of the image: : Zöller, 2020.

Normal distribution

- has a bell-shaped curve
- has a symmetrical distribution around the mean value
  - $1\sigma \sim 68\%$
  - $2\sigma \sim 95\%$
  - $1\sigma \sim 99{,}7\%$

Example: Performance assessment of an organization’s employees

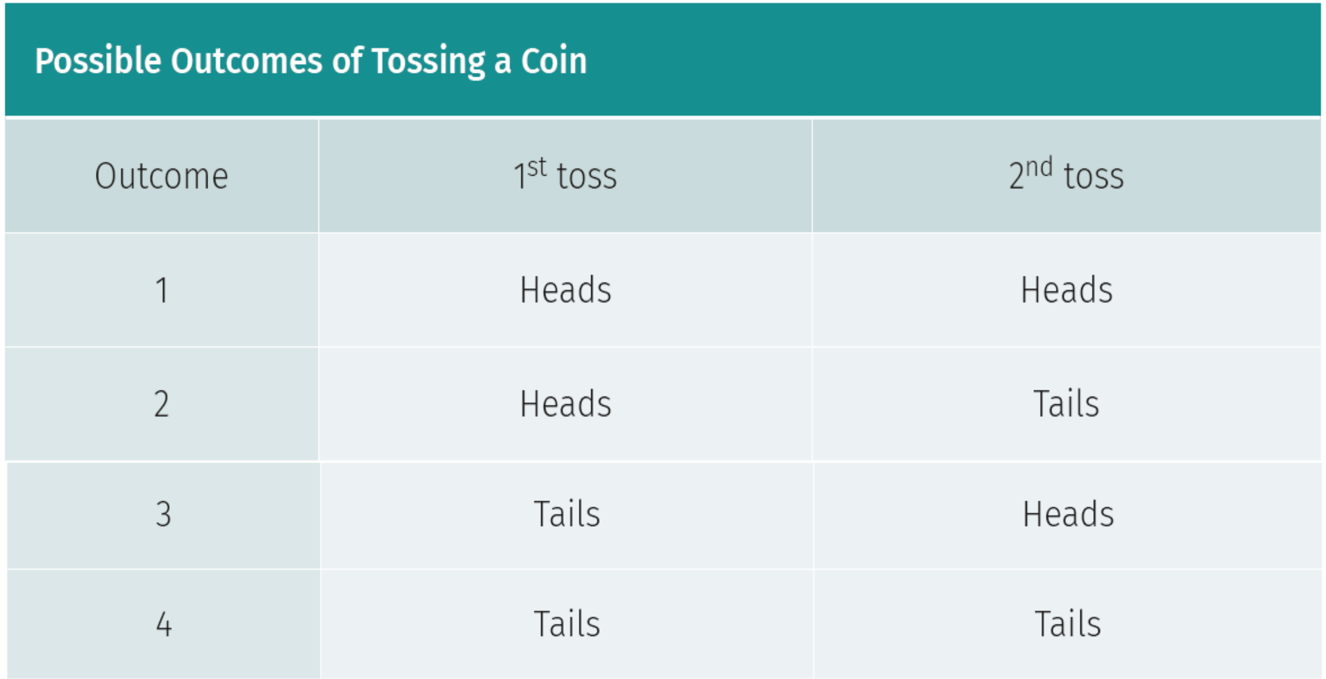


Source of the image: Zöller, 2020.

Binomial distribution

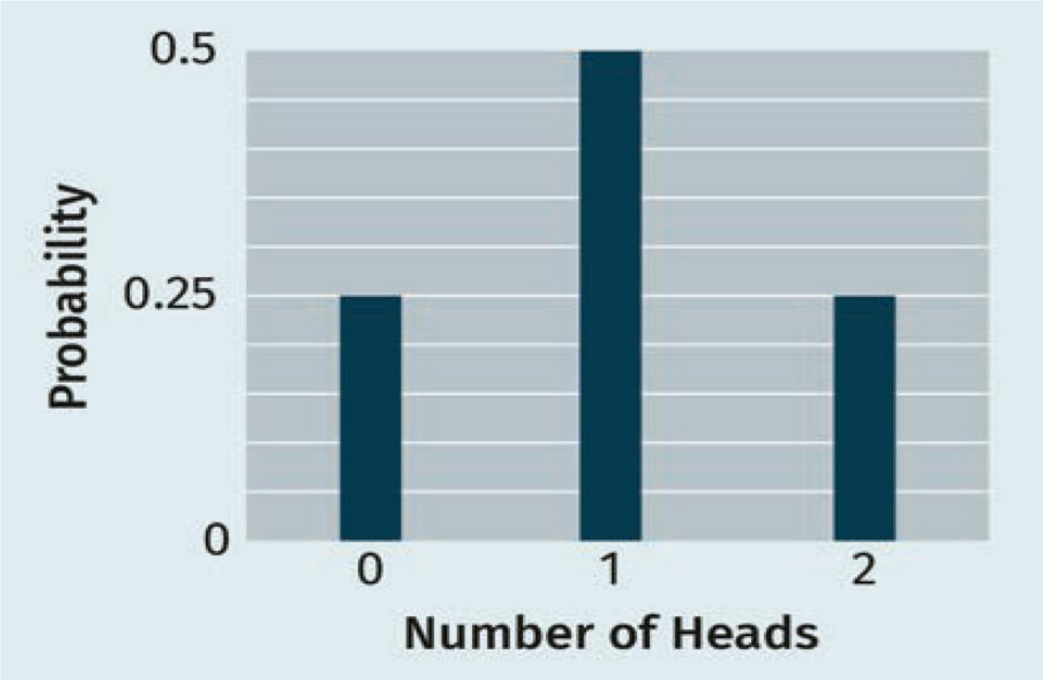
The probability distribution of the **number of successes** in a sequence of independent trials that each can be described by a **binary random** variable.

Example: tossing a coin twice



Possible Outcomes of Tossing a Coin

Outcome	1st toss	2nd toss
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails



Source of the image: Zöller, 2020.

Poisson distribution

The probability of a given number of independent events occurring in a fixed time interval

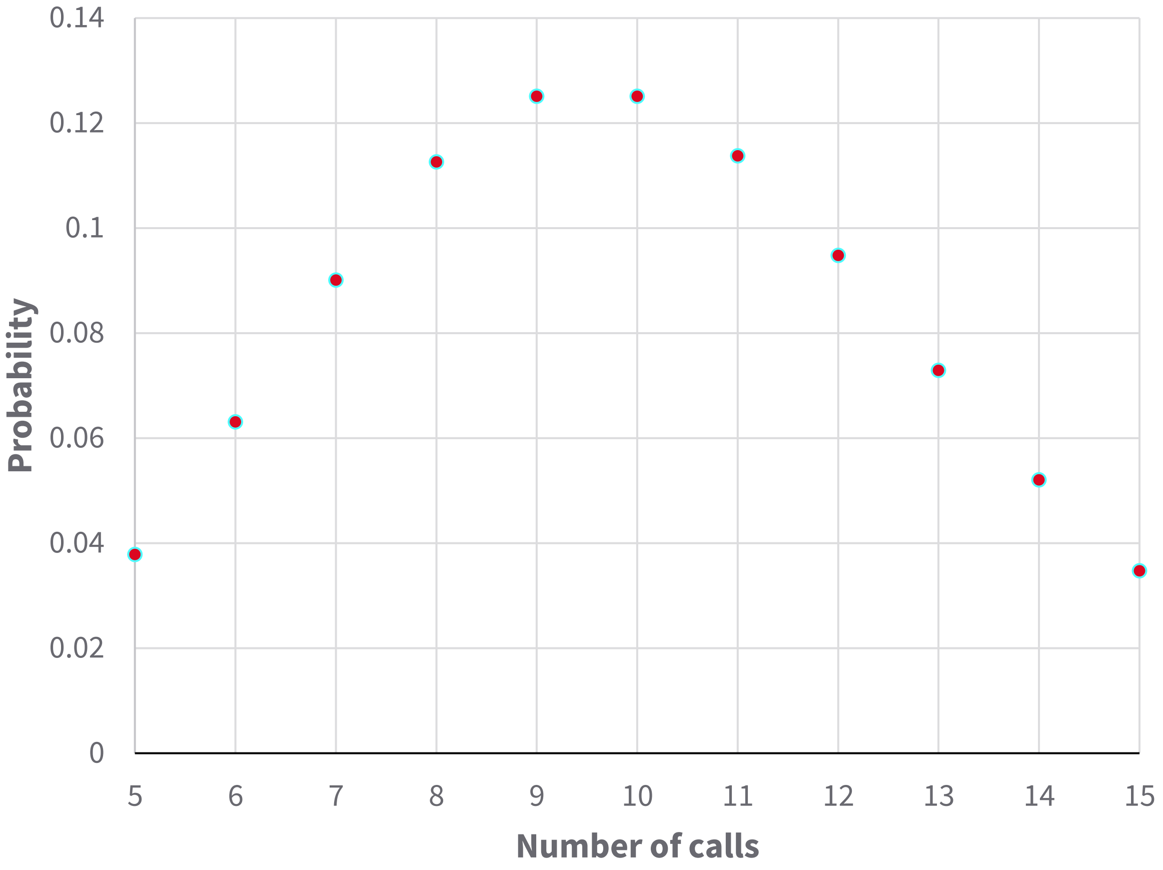
$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Where:

μ – the mean number of occurrences

x – the required number of occurrences

Example: The probability that a call center will receive exactly n calls on a given day.



Source of the image: Zöller, 2020.

Bayerian statistics interprets probabilities **as expectation of belief.**

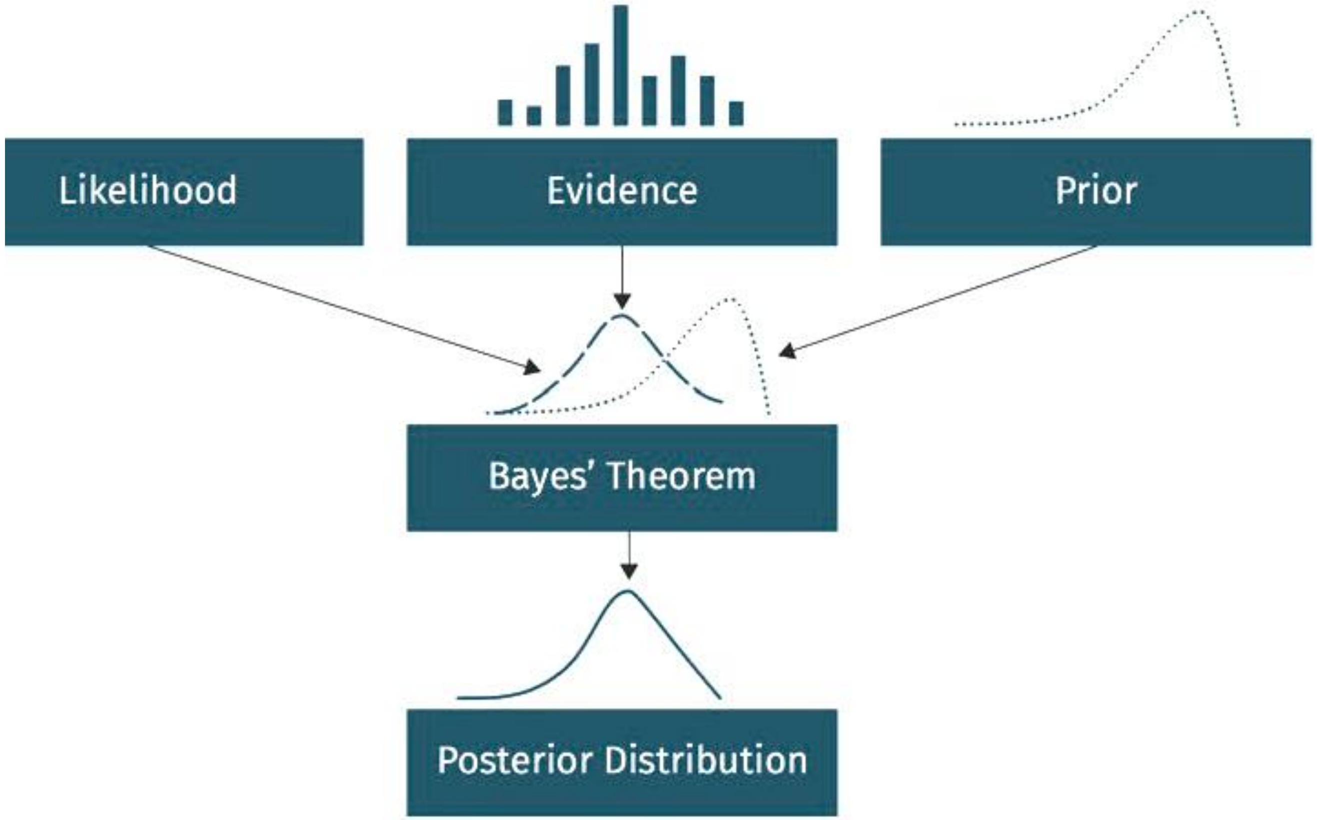
Conditional probability **equation**:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:

$P(A|B)$ is the **posterior** belief of the event A after observing the **evidence** B.

Example: Drug test analysis



Source of the image: Zöller, 2020.

STUDY GOALS

- Identify the importance of statistics in data science.
- Know about probability and its relation to the prediction model’s outputs.
- Learn about conditional probability and the probability density function.
- Understand the different probability distributions.
- Know the Bayesian statistics.

SESSION 4

TRANSFER TASK

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.
The results will be discussed in plenary.

TRANSFER TASK

Given

The values of 5 categories are measured as follows: 2, 6, 9, 18, 20.

Questions

- What are the values of the statistical parameters: max, min, median, mean, and standard deviation?
- Draw the chart to visualize the results.

LEARNING CONTROL QUESTIONS

1. Which of the following is a true statement?

   a) Mean is the arithmetic average of values and median is the maximum value.

   b) Mean is the arithmetic average of values and median is the minimum value.

   c) Mean is the arithmetic average of values and median is the value positioned in the middle.

   d) Median is the arithmetic average of values and mean is the value positioned in the middle.

LEARNING CONTROL QUESTIONS

2. Which probability distribution has a bell-shaped curve?

a) Normal distribution

b) Binomial distribution

c) Poisson distribution

d) None of them

LEARNING CONTROL QUESTIONS

3. Which of the following is a true statement?

   a) Bayesian statistics interprets probabilities as frequencies of occurrence
   b) Bayesian statistics interprets probabilities as an expectation of belief
   c) Bayesian statistics does not interpret probabilities
   d) none of the above

LIST OF SOURCES

Zöller, T. (2020). *Course Book – Introduction to Data Science*. IU International University of Applied Science.