Statistics

Adam Klepáč

September 19, 2023

What Even Is Statistics?

Statistics

Statistics is a mathematical discipline concerned with predicting future state of a system based *solely* on its past behaviour.

What Even Is Statistics?

Statistics

Statistics is a mathematical discipline concerned with predicting future state of a system based *solely* on its past behaviour.

The collective information about a system's past state is called **data**.

What Even Is Statistics?

Statistics

Statistics is a mathematical discipline concerned with predicting future state of a system based *solely* on its past behaviour.

The collective information about a system's past state is called **data**.
It assigns **probabilities** to each possible future state of system based on data.

What Even Is Statistics?

Statistics

Statistics is a mathematical discipline concerned with predicting future state of a system based *solely* on its past behaviour.

The collective information about a system's past state is called **data**.
It assigns **probabilities** to each possible future state of system based on data.
It also assigns probabilities to the **possibility of wrong prediction**.

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’.

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’. We can ask two questions:

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’. We can ask two questions:

- What is the probability that the **next toss** will come out ‘heads’/‘tails’?

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’. We can ask two questions:

- What is the probability that the **next toss** will come out ‘heads’/‘tails’?

- Is this coin is **biased towards** ‘heads’/‘tails’ with *allowed probability of error* α?

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’. We can ask two questions:

- What is the probability that the **next toss** will come out ‘heads’/‘tails’?
  - We got 7 heads out of 10 tosses, so the probability for the next toss being heads is $7/10$.
- Is this coin is **biased towards** ‘heads’/‘tails’ with *allowed probability of error* α?

Example – Biased Coin?

We throw a coin 10 times with the following outcome:

$$\{H, H, H, T, H, T, H, H, H, T\},$$

H for ‘heads’, *T* for ‘tails’. We can ask two questions:

- What is the probability that the **next toss** will come out ‘heads’/‘tails’?
  - We got 7 heads out of 10 tosses, so the probability for the next toss being heads is $7/10$.
- Is this coin is **biased towards** ‘heads’/‘tails’ with *allowed probability of error* α?
  - **No**, for $\alpha = 0.05$.
  - **Yes**, for $\alpha = 0.2$.

Contents

Visualizing Discrete Data

Data

What Do We Mean By Data?

Data

Two sets (called *inputs* and *outputs*) describing the studied system.

Example – Junctions

For a year, we keep track of the number of traffic accidents per day on road junctions across the city to determine which should be first replaced by roundabouts.

Example – Junctions

For a year, we keep track of the number of traffic accidents per day on road junctions across the city to determine which should be first replaced by roundabouts.
An **input** is a day in a year.

Example – Junctions

For a year, we keep track of the number of traffic accidents per day on road junctions across the city to determine which should be first replaced by roundabouts.
An **input** is a day in a year.
An **output** is the number of traffic accidents in a given day.

Example – First Baby

We study the age that women bear children for the first time across Europe.

Example – First Baby

We study the age that women bear children for the first time across Europe.
An **input** would be a name of a European country.
An **output** is the average age of a first-time mother in that country.

1

Types of Data

Discrete Data vs. Continuous Data

Discrete Data

We call a data **discrete** if the set of *inputs* (and therefore also that of *outputs*) is **countable**.

Discrete Data vs. Continuous Data

Discrete Data

We call a data **discrete** if the set of *inputs* (and therefore also that of *outputs*) is **countable**.

Both previous examples feature **discrete** data.

- There are only *finitely many* junctions in a city.

Discrete Data vs. Continuous Data

Discrete Data

We call a data **discrete** if the set of *inputs* (and therefore also that of *outputs*) is **countable**.

Both previous examples feature **discrete** data.

- There are only *finitely many* junctions in a city.
- There are only *finitely many* countries on a continent.

Discrete Data vs. Continuous Data

Continuous Data

We call a data **continuous** if the set of inputs is **uncountable**. In this case, the data is actually a **function**: set of inputs $\rightarrow$ set of outputs.

Discrete Data vs. Continuous Data

Continuous Data

We call a data **continuous** if the set of inputs is **uncountable**. In this case, the data is actually a **function**: set of inputs $\rightarrow$ set of outputs.

More often than not, the inputs in a continuous data are **moments in time** or **coordinates in space**.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.
  - The data is a function $f : [0, 24] \to \mathbb{N}$.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.
  - The data is a function $f : [0, 24] \to \mathbb{N}$.
- Another example is the density of air per cubic meter.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.
  - The data is a function $f : [0, 24] \to \mathbb{N}$.
- Another example is the density of air per cubic meter.
  - Input: Coordinates of a unit cube in space.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.
  - The data is a function $f : [0, 24] \to \mathbb{N}$.
- Another example is the density of air per cubic meter.
  - Input: Coordinates of a unit cube in space.
  - Output: The combined weight of air molecules.

Continuous Data – Examples

- We study the number of trains in a railway station at any given time.
  - Input: time (of day);
  - Output: number of trains in the station.
  - The data is a function $f : [0, 24] \to \mathbb{N}$.
- Another example is the density of air per cubic meter.
  - Input: Coordinates of a unit cube in space.
  - Output: The combined weight of air molecules.
  - The data is a function $f : \mathbb{R}^3 \to \mathbb{R}$.

Visualizing Discrete Data

Tables

The simplest possible visualization.

Tables

The simplest possible visualization.

You simply write *inputs* into one row/column and *outputs* into the other.

Tables

The simplest possible visualization.

You simply write *inputs* into one row/column and *outputs* into the other.

For example, suppose you measure the height of 10 random people. You can visualize your experiment like this:

Tables

The simplest possible visualization.

You simply write *inputs* into one row/column and *outputs* into the other.

For example, suppose you measure the height of 10 random people. You can visualize your experiment like this:

Input	1	2	3	4	5	6	7	8	9	10
Output	180	169	191	177	175	181	171	153	180	183

Pie Chart

Only usable if your outputs **total a predetermined number**, typically *percentages*.

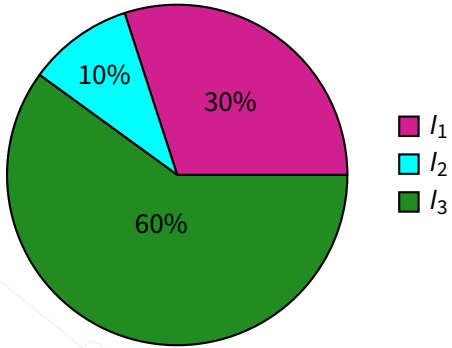
Pie Chart

Only usable if your outputs **total a predetermined number**, typically *percentages*.

Suppose we have three inputs – I_1, I_2 and I_3 – with three outputs – 30%, 10% and 60%.

Pie Chart

Only usable if your outputs **total a predetermined number**, typically *percentages*.
Suppose we have three inputs – I_1, I_2 and I_3 – with three outputs – 30%, 10% and 60%.
Pie chart of this data looks like this

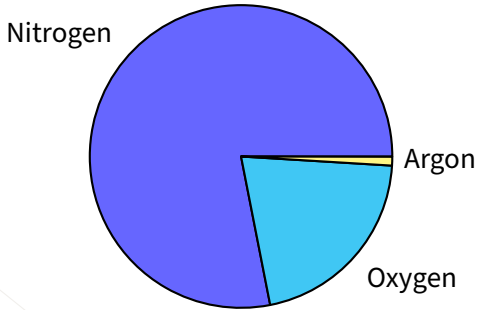


Pie Chart – Examples

Pie charts are frequently used to represent compositions of chemicals.

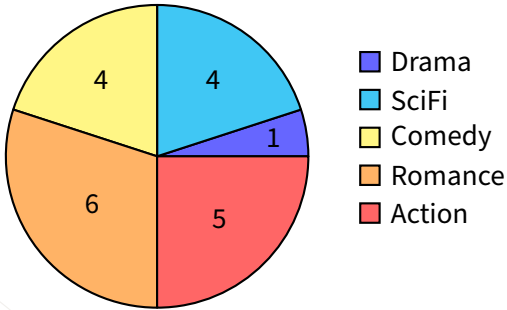
Pie Chart – Examples

Pie charts are frequently used to represent compositions of chemicals.
For instance, here is a pie chart of the composition of *air*.



Pie Chart – Examples

Favourite type of movie as determined by a survey.



Bar Chart

Usable basically for any data.

Bar Chart

Usable basically for any data.
Especially useful when your inputs are ordered and when you expect the data to follow a certain trend – it can be easily approximated by a polygonal curve.

Bar Chart

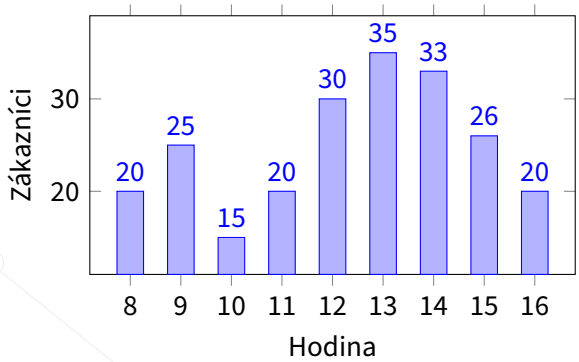
Usable basically for any data.
Especially useful when your inputs are ordered and when you expect the data to follow a certain trend – it can be easily approximated by a polygonal curve.
Also very good for comparing more outputs for the same inputs.

Bar Chart – Example

Suppose we count the number of customers in our shop over each hour. If we're open from 8 AM to 5 PM, a bar chart of such an experiment can look like this:

Bar Chart – Example

Suppose we count the number of customers in our shop over each hour. If we're open from 8 AM to 5 PM, a bar chart of such an experiment can look like this:



Bar Chart – Example

Let’s say we open another shop and want to compare how the two shops are doing at each hour.

Bar Chart – Example

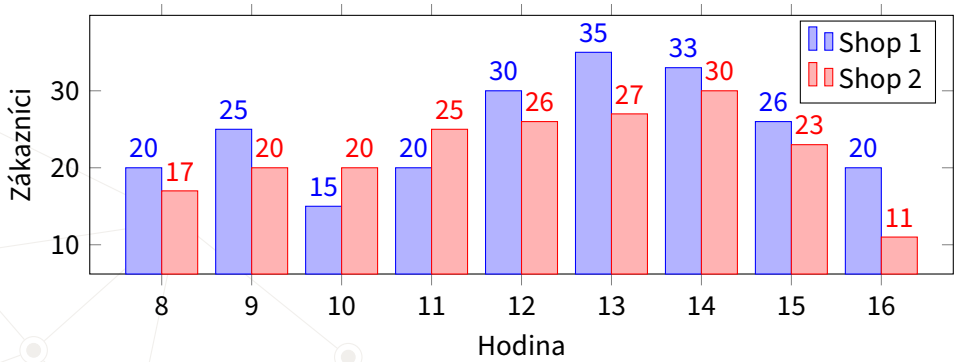
Let’s say we open another shop and want to compare how the two shops are doing at each hour.

We can draw both outputs into the same bar chart:

Bar Chart – Example

Let's say we open another shop and want to compare how the two shops are doing at each hour.

We can draw both outputs into the same bar chart:



Bar Chart – Example

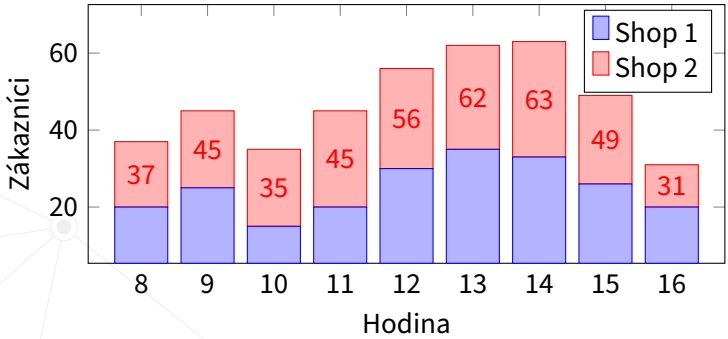
You can also use bar chart to stack outputs on top of each other.

Bar Chart – Example

You can also use bar chart to stack outputs on top of each other.
For example, if I wanted to know the **total** number of customers in both my shops, I could draw a chart like this:

Bar Chart – Example

You can also use bar chart to stack outputs on top of each other.
For example, if I wanted to know the **total** number of customers in both my shops, I could draw a chart like this:



Scatter Plot

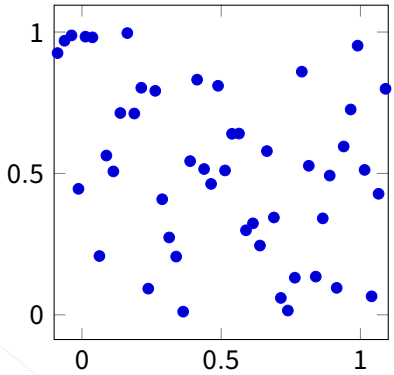
Scatter plots are useful when studying ‘random’ data.

Scatter Plot

Scatter plots are useful when studying ‘random’ data.
Something like the position of an air molecule in a box in time.

Scatter Plot

Scatter plots are useful when studying ‘random’ data.
Something like the position of an air molecule in a box in time.



Scatter Plot – Example

Of course, you can also display multiple outputs with the same inputs in a scatter plot.

Scatter Plot – Example

Of course, you can also display multiple outputs with the same inputs in a scatter plot. Let's add another air molecule.

Scatter Plot – Example

Of course, you can also display multiple outputs with the same inputs in a scatter plot. Let's add another air molecule.