Statistics

Adam Klepáč

October 2, 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.

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'. 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

Data

What Do We Mean By Data?

Data

Sets (called *inputs* and *outputs*) describing the studied system. There is typically only one set of inputs and possibly multiple sets of outputs.

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 coupled with the location of the junction.

An **output** is the number of traffic accidents in the given day on the given junction.

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**.

Both previous examples feature **discrete** data.

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

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.
  - 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 $f : \mathbb{R}^3 \to [0, \infty)$.

Visualizing Discrete Data

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:

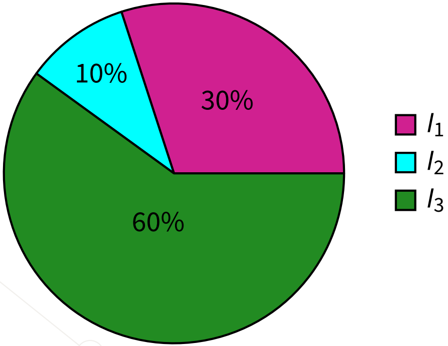
Input	1	2	3	4	5	6	7	8	9	10
Output	180	169	191	177	175	181	171	153	180	183

Pie Chart

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

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

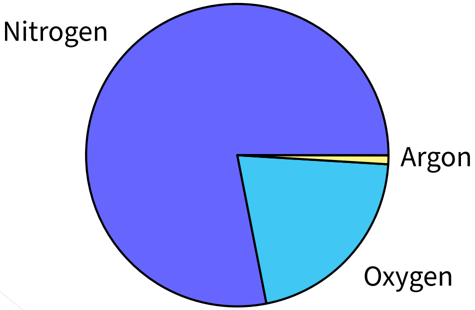
Pie chart of this data looks like this



PIE CHART – EXAMPLES

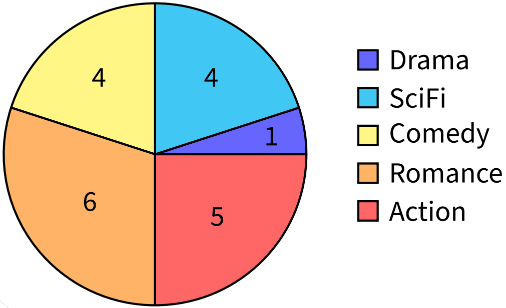
Pie charts are frequently used to represent compositions of chemicals.

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

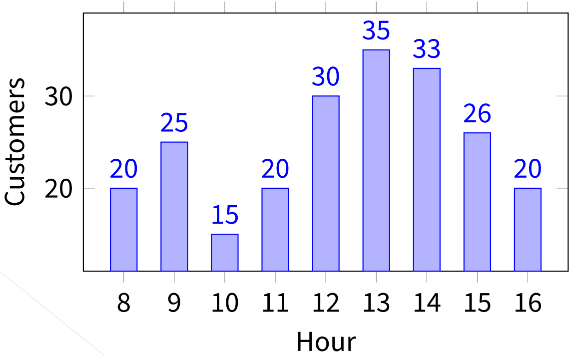
Usable basically for any discrete data.

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

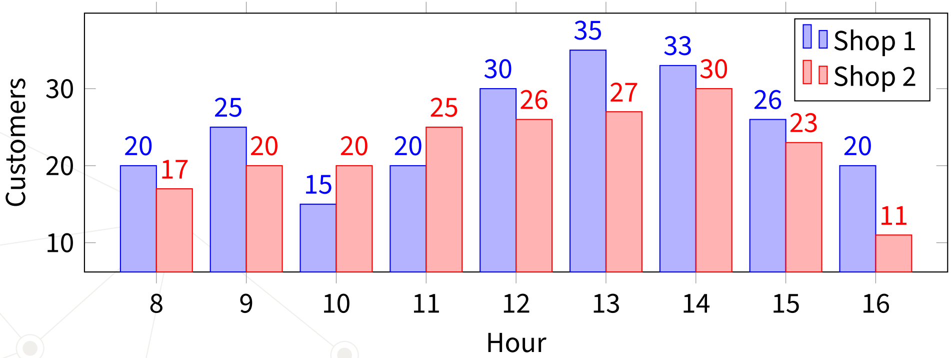
Suppose we count the number of customers in our shop over each hour. If we're open from 8 AM to 5 PM, a bar chart of such an experiment can look like this:



Bar Chart – Example

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

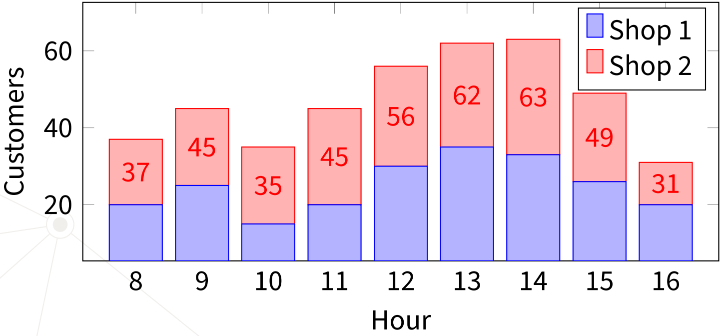
We can show both outputs in the same bar chart:



Bar Chart – Example

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

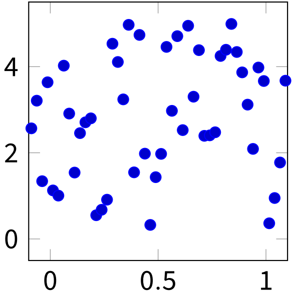
For example, if I wanted to know the **total** number of customers in both my shops, I could draw a chart like this:



Scatter Plot

Scatter plots are useful when studying 'random' data.

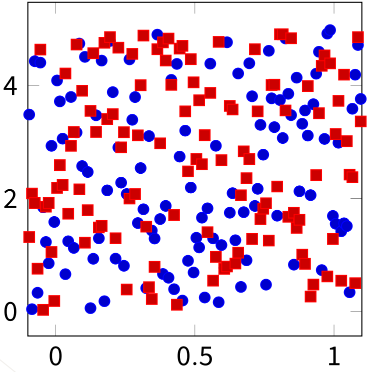
Something like the position of an air molecule in a box over time.



Here, the *x*-axis represents time (0 to 1s) and the *y*-axis represents one coordinate of the molecule (say the box is a 5x5x5 cube).

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.



Mean – Median – Deviation – Correlation

USEFUL VALUES

When studying data, there are **certain numerical values** which prove useful in predicting future behaviour.

Facts we wish to learn about our data include:

- the expected value of the next experiment (the **mean**),
- the ‘middle’ value of the outputs regardless of proportion (the **median**),
- the expected/observed measure of difference of observed values from the mean (the **deviation**),
- dependence on any other data (the **correlation**).

1

The Mean

Types of Mean

The **mean** in some sense is ‘the most probable’ next output based on the received data. What is ‘the most probable’ output however depends heavily on the type of experiment we are performing.

Types of Mean – Arithmetic Mean

Arithmetic Mean

The **arithmetic mean** is the sum of outputs divided by their number. If $x_1, \ldots, x_n$ are the outputs, their arithmetic mean (often denoted $\bar{x}$) is

$$\bar{x} := \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Most useful when dealing with data in **absolute** proportion.
Meaning that we're interested in '**how much**' is one output smaller/larger than another.

Arithmetic Mean – Example

The arithmetic mean is for most experiments the relevant one.

Consider for example an experiment tailored to determine the average height of a 15-year-old British male.

While comparing the heights of two people, we care about the **absolute** difference in centimetres.

For example, if this is our data

Input	1	2	3	4	5
Output	165	161	164	172	168,

we conclude that the expected height of a randomly chosen 15-year-old British male is

$$\frac{165 + 161 + 164 + 172 + 168}{5} = 166.$$

Types of Mean – Geometric Mean

Geometric Mean

The **geometric mean** is the n-th root of the product of n outputs. That is, if $x_1, \dots, x_n$ are the outputs, their geometric mean is

$$\bar{x} := \sqrt[n]{(x_1 \cdot x_2 \cdot \ldots \cdot x_n)}.$$

Most useful when dealing with data in **relative** proportion.

Meaning that we're interested in '**how many times**' is one output smaller/larger than another.

Geometric Mean – Example

Suppose we're comparing the increase in population in Asian countries since the year 2000.

When comparing two populations, we don't really care *by how much* they differ, but *how many times* is one larger than the other.

If this is our data

Input	India	China	Japan	South Korea	Mongolia	Taiwan
Output	1.328	1.118	0.991	1.100	1.366	1.078

then the expected increase in population in a randomly chosen Asian country is

$$\sqrt[6]{(1.328 \cdot 1.118 \cdot 0.991 \cdot 1.100 \cdot 1.366 \cdot 1.078)} = 1.156.$$

Types of Mean – Harmonic Mean

Harmonic Mean

The **harmonic mean** is the reciprocal of the sum of reciprocals divided by their number. If $x_1, \ldots, x_n$ are the outputs, their harmonic mean is

$$\bar{x} := \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}.$$

Most useful when dealing with **rates** and **ratios**.
Meaning when comparing outputs which are actually ratios of two numbers.

Harmonic Mean – Example

We study the speed of a train between individual stations.

If this is our data

Input	$1 \to 2$	$2 \to 3$	$3 \to 4$	$4 \to 5$	$5 \to 6$	$6 \to 7$
Output	65 km/h	52 km/h	71 km/h	60 km/h	62 km/h	53 km/h,

then the average speed of the train over the whole track is

$$\frac{6}{\frac{1}{65}+\frac{1}{52}+\frac{1}{71}+\frac{1}{60}+\frac{1}{62}+\frac{1}{53}} = 59.78\text{ km/h}.$$

Actually, here the arithmetic mean is 60.5 km/h which is not just an *inadequate* estimate, it's simply **the wrong answer!**

If you summed up all the distances between stations and divided them by the total time, you would get the **harmonic mean!**

2

The Median

The Median

Median

The **median** is the value which lies exactly in the middle of a dataset. It is essentially the value separating the lower and upper half of outputs. If $x_1, \ldots, x_n$ are the outputs **ordered from least to greatest**, the median is

$$\mathrm{median}(x) := \begin{cases} x_{(n+1)/2} & \text{if n is odd,} \\ \frac{x_{n/2}+x_{n/2+1}}{2} & \text{if n is even.} \end{cases}$$

The Median – Example

The median is useful when dealing with data manifesting radical extremes.

It is very often used in relationship to **location**.

For example, imagine we're trying to determine the distance from our measuring station to the hypocentre of an earthquake.

We can detect where the quake is strongest, giving us this data:

Input	1	2	3	4	5	6
Output	1 km	2 km	2 km	2 km	3 km	14 km

The median of this dataset is 2 km which is a much better estimate of a 'centre' than for example the arithmetic mean, being equal to 4, is.

Also, the mean and the median cannot be 'too far' apart and the median requires at most two values to calculate, making it a very resource efficient approximation of the mean.

3

The Deviation

Deviation

Deviation

Deviation is a measure of the difference between the observed outputs and the computed mean.

There are many types of deviations, we'll focus on two of them:

- the **standard deviation** (a measure of 'dispersion'),
- the **average absolute deviation** (a measure of actual 'difference').

A **very important distinction** is that the *standard deviation* concerns **future** measurements while the *average absolute deviation* concerns **past** measurements.

Types of Deviation – Standard Deviation

Standard Deviation

The **standard deviation** measures the dispersion of a set of values. Basically, it measures how likely the data is to concentrate around the mean. If $x_1, \ldots, x_n$ are the outputs and $\bar{x}$ is their **arithmetic** mean, then their standard deviation is

$$\sigma := \sqrt{\frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)}.$$

Standard Deviation – Example

Let us repeat the height experiment. We measured the heights of 5 15-year-old British males to try to determine the national average. This is the data:

Input	1	2	3	4	5
Output	165	161	164	172	168

We computed the arithmetic mean to be 166. This means that the standard deviation of this data is

$$\sigma = \sqrt{\frac{1}{5}((165-166)^2 + (161-166)^2 + (164-166)^2 + (172-166)^2 + (168-166)^2)}$$
$$= 3.742,$$

meaning we can expect most new values to concentrate 3.742 cm around 166 cm.

Types of Deviation – Average Absolute Deviation

Average Absolute Deviation

The **average absolute deviation** is the average of the absolute deviations from a chosen central point (typically the mean). If $x_1, \ldots, x_n$ are the outputs and $\bar{x}$ is the chosen central point, then the average absolute deviation of this dataset is

$$\frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}.$$

Average Absolute Deviation – Example

If we return to the height experiment yet again, we can calculate that the average absolute deviation of the data (with the central point being the arithmetic mean)

Input	1	2	3	4	5
Output	165	161	164	172	168

is

$$\frac{|165-166|+|161-166|+|164-166|+|172-166|+|168-166|}{5}=3.2,$$

meaning that the measured heights differ on average by 3.2 cm from the calculated arithmetic mean.

4

Correlation

Correlation

Correlation

Correlation or **dependence** is a relationship between two outputs of a dataset. A 'correlation' in many cases means some type of association.

Correlation is an number between -1 and 1. Intuitively, a

- negative correlation means that the two series of outputs **contradict** each other;
- zero correlation means that the two series of outputs are **unrelated**;
- positive correlation means that the two series of outputs **influence** each other.

Computing Correlation

Correlation Formula

If $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are two series of outputs for the same inputs with means $\bar{x}$ and $\bar{y}$, their correlation is

$$\mathrm{cor}(x, y) := \frac{(x_1 - \bar{x})(x_2 - \bar{x}) \cdots (x_n - \bar{x})(y_1 - \bar{y})(y_2 - \bar{y}) \cdots (y_n - \bar{y})}{\sqrt{(x_1 - \bar{x})^2 (x_2 - \bar{x})^2 \cdots (x_n - \bar{x})^2 (y_1 - \bar{y})^2 (y_2 - \bar{y})^2 \cdots (y_n - \bar{y})^2}}.$$

Interpreting Correlation – Table

A crude interpretation of correlation is given in the following table:

Coefficient	Strength	Type
-0.7 to -1	Very strong	Negative
-0.5 to -0.7	Strong	Negative
-0.3 to -0.5	Moderate	Negative
0 to -0.3	Weak	Negative
0 to 0.3	Weak	Positive
0.3 to 0.5	Moderate	Positive
0.5 to 0.7	Strong	Positive
0.7 to 1	Very strong	Positive

Interpreting Correlation – Chart

You can use a scatter plot to visualize two sets of outputs with the same inputs. Correlation tells you **how well you can approximate** this scatter plot by a straight line. For example, imagine you measure the height and weight of a sample of people and want to see if they correlate. You might get a scatter plot like this: