Statistics

Adam Klepáč

September 21, 2023

Mean – Median – Deviation – Correlation

Useful Values

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

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:

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**),

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**),

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 measure of difference of observed values from the mean (the **deviation**),

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 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 ‘the most probable’ next output based on the received data.

Types of Mean

The **mean** in some sense '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

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

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.

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.

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.

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.

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 $x_1, \ldots, x_n$ are the outputs, their geometric mean is

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

Types of Mean – Geometric Mean

Geometric Mean

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

Types of Mean – Geometric Mean

Geometric Mean

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

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 *how many times* is one larger than the other.

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 *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

This means that 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

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

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**.

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.

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 across the whole track is

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

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 across 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!**

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 across 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!**

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, 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 we deal with data featuring radical extremes.

The Median – Example

The median is useful when we deal with data featuring radical extremes.

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

The Median – Example

The median is useful when we deal with data featuring 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.

The Median – Example

The median is useful when we deal with data featuring 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 – Example

The median is useful when we deal with data featuring 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.

The Median – Example

The median is useful when we deal with data featuring 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.

Deviation

Deviation

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

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:

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'),

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').

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

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

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

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

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.