Lec 3

Biostatistic

Exploring data for errors:

Before running the statistical analysis, we need to explore the data to make sure that there are no data entry errors.

This can be done using many techniques:

• **Check the range (minimum and maximum)**

Are there any incorrect extreme values? Are they consistent with other data values?

• **Check the frequency distribution for categorical variables**

Are there any typing mistakes or unusual codes or groups?

• **Check the missing values**

Are they really not available? Or we just forgot them during data entry?

• **Checking the consistency of data**

For example, a man can't be pregnant, disease duration can't be larger than age, and diastolic blood pressure can't be larger than systolic blood pressure.

• Graphically checking the data

A histogram or a boxplot for a single numeric variable, and a scatterplot for two related variables as weight and waist circumference may be helpful to explore possible errors.

Descriptive statistics

It is important to learn how to describe our data and present those correctly using numbers (in the proper table format) or graphs.

The first table in most scientific research papers shows descriptive statistics of the study subjects.

As in the following table:

Table 1. Baseline Characteristics of the Study Participants*

Characteristics	Active Treatment (n = 817)	Placebo (n = 813)
Age, mean (SD), y	42.1 (9.0)	42.4 (9.1)
Sex		
Men	440	440
Women	377	373
Daily smoking	203 (25)	198 (24)
Alcohol use	194 (24)	160 (20)
Dyspepsia symptoms	417 (51)	409 (50)
Dietary intake ≥2 times/wk		
Green tea	205 (25)	181 (22)
Preserved vegetables	144 (18)	132 (16)
Salty fish	364 (45)	372 (46)
Fish sauce	172 (21)	241 (30)
Fruit	112 (14)	83 (10)
Fresh vegetables	275 (34)	253 (31)
Histopathologic test results		
Chronic active gastritis	485 (59.4)	503 (61.9)
Gastric atrophy	72 (8.8)	57 (7.0)
Intestinal metaplasia	243 (29.7)	234 (28.8)
Gastric dysplasia	4 (0.5)	5 (0.6)
Unclassified†	13 (1.6)	14 (1.7)

*Data are expressed as No. (%) of participants unless otherwise indicated.
†Histology slides were uninterpretable or no definite conclusions could be drawn.

There are different ways of numerically describing data based on the type of the variable.

1- Describing categorical variables

Categorical variables such as sex, smoking status, and disease severity are described as:

- • **Frequencies (numbers):** which is the number of participants in each category, as the number of males and the number of females?
- • **Relative frequencies (percentages):** which is the percentage of participants in each category?
  - **For example**:
  - If you have 200 participants, 120 are males and 80 are females.
  - We can express the frequencies and percentages as follows:
  - Males: 120 (60%)
  - Females: 80 (40%)

- Percentages can be calculated easily by dividing the number of that category by the total
- Number and multiplying it by 100.
- For the males, it is:
- 120
- 200
- × 100 = 60% .
- **2- Describing numeric variables**
- Numerical variables are usually described using two numbers, one represents the center
- of the data (**central tendency**), and the other represents the spread of the data
- (**dispersion**).
- **• Measures of central tendency**
- The most common measures for the center of the data are the mean, median, and mode.
- **a- Mean**
- • The mean of a variable can be computed as the sum of the observed values
- divided by the number of observations.
- For example: if we want to calculate the mean for the age of 7 children;
- 7,5,6,8,2,9,3
- it will be:
- 7+5+6+8+2+9+3
- 7
- =
- 40
- 7
- = 5.71 years.
- • The mean is easily affected by extreme values.
- If we add one adult whose age is 64 years to this group and try to calculate the
- mean again it will be:
- 7+5+6+8+2+9+3+64
- 8
- =

- 104
- 8
- = 13 years.
- • We can see that the mean age has changed obviously from 5.71 to 13 years by
- adding only one value and that the new value (13) is even larger than the age of
- all the 7 children. The mean here is not a good representative of our data.

• The mean is also called the average or the arithmetic mean

b- Median

• The median is the point at the center of the data, where half of the values are above, and half are below it.

• To calculate the median, we first arrange (order) our data from the smallest value to the largest value. Then, the median is the value in the middle.

For example: if we want to calculate the median for the age of 7 children mentioned above; 7,5,6,8,2,9,3

First, we order the data:

2,3,5,6,7,8,9

Then it is obvious that the center of it is the number 6, where 3 values are below, and 3 values are above it:

2,3,5,⑥,7,8,9

So, the median= 6 years

• What if we try to add the adult with the age of 64 years old?

Then the ordered data will be:

2,3,5,6,7,8,9,64

Here, we can't see one value in the middle with half the values above and half below it. In this case, we will take the average of the two values in the middle:

2,3,5,⑥,⑦,8,9,64

So, the median = 6+72=6.5 years

As we notice, the median didn't change much when that extreme value was added.

- **c- Mode**

- • Simply, the mode is the most frequently occurring value in the dataset.

So , if you have a data set like: 2,3,5,6,7,8,9,64,3,4,5,3
Then the mode is 3.

- • It can be also calculated for categorical variables as it depends only on the frequency of each value.
- • The mode can be more than one value; if two values have the same highest frequency, then, both are the modes, and data is called bimodal.

The mode is rarely reported in scientific research.

	advantage	disadvantages
Mean	Uses all data values Algebraically defined	Distorted by outliers Distorted by skewed data
Median	Not distorted by outliers Not distorted by skewed data	Ignores most of the information Not algebraically defined
Mode	Easily determined for categorical data	Ignores most of the information Not algebraically defined

The five-number summary
If we arrange our values from lowest to highest and choose five points on the arranged data to divide the variable into 4 quarters, those five points (numbers) will be:

- • **The minimum value**
- • **The maximum value**
- • **The median:** which is the point at the center of the data where half of the values above and half are below it.
- • **The first quartile (lower quartile):** where 25% of the data are below it, it is the center point for the lower half of the data. It is also called the 25th percentile.
- • **The third quartile (upper quartile):** where 75% of the data are below it, it is the center point for the upper half of the data. It is also called the 75th percentile.

If you have the following values for a variable:

8,10,10,10,12,14,15,15,18,23,25,27
The five-number summary will be:
Min: 8 **Q1:** 10 Median: 14.5 **Q3:** 21.75 **Max:** 27
It is easily calculated using computer software. The following is an SPSS output:

N	Valid	12
	Missing	0
Minimum		8
Maximum		27
Percentiles	25	10.00
	50	14.50
	75	21.75

The example is graphically presented in a graph called the box-plot as follows:

- Measures of dispersion

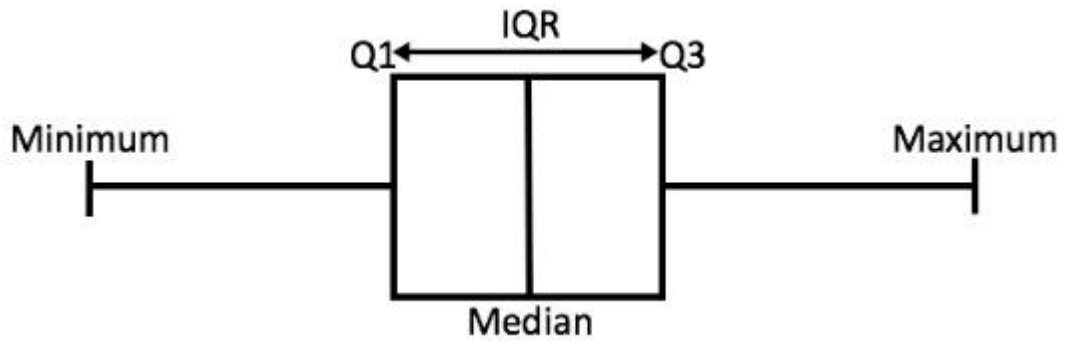
The most commonly used measures of dispersion (spread of the data) are range, inter-quartile range, variance, and standard deviation.

- **a- Range:**
- • The range is simply the difference between the largest and smallest values.

If you have the following values for the age variable: 8,10,10,10,12,14,15,15,18,23,25,27

- • The lowest value is 8, the highest value is 27, so the range is 27-8= 19 years.
- • It is obvious that the range is affected by any extreme values.
- • Adding one adult aged 64 to this group will increase the range significantly. The range becomes 64-8= 56 years.

- **b- Inter-quartile range (IQR):**
- • The inter-quartile range is simply the difference between the upper quartile and the lower quartile = Q3-Q1
- • It represents the middle 50% of the data, where 25% of the data are below it, and 25% are above it.



For those values: 8,10,10,10,12,14,15,15,18,23,25,27
Q1=10, Q3= 21.75

The IQR = Q3-Q1 = 21.75-10 = 11.75
The IQR is not calculated using the minimum or the maximum values, so it is not affected by extreme values.

c- Variance

• The variance is a measure of spread that takes all data points in the calculation. It represents the distance of all data points from the mean.

• We calculate it as in the following steps:

1- Calculate the mean.

2- Calculate the difference between each data point and the mean, then square it (not to have negative values).

3- Sum all the squared differences calculated in step 2.

4- Divide this sum by the number of observations -1 (n-1)

This is the variance; it is in square units (as we squared the difference!).
This means that if the mean height in m, then the variance is in m2

Example:

We have a group of 7 children, and their age in years is; 7,5,6,8,4,9,3, let's calculate the variance.

1- Calculate the **mean**: $7+5+6+8+4+9+37=427=6$ years.

2- Calculate the **difference** between each data point and the mean, then **square** it. $(7-6)^2$, $(5-6)^2$, $(6-6)^2$, $(8-6)^2$, $(4-6)^2$, $(9-6)^2$, $(3-6)^2$

= 12,-12,02,22,-22,32,-32
=1,1,0,4,4,9,9

3- **Sum** all the squared differences =28

4- **Divide** this sum by the number of datapoints -1 (n-1)

$s^2=287-1=$ 4.67 $years^2$
So, the variance = 4.67 $years^2$
But the interpretation of variance of age with a squared unit ($years^2$) is not easy to understand.
So, we take the square root of the variance to have the standard deviation (s), which is now of the same unit as the mean.
$s=\sqrt{s^2} = \sqrt{4.67}$ = 2.16 years

d- Standard deviation

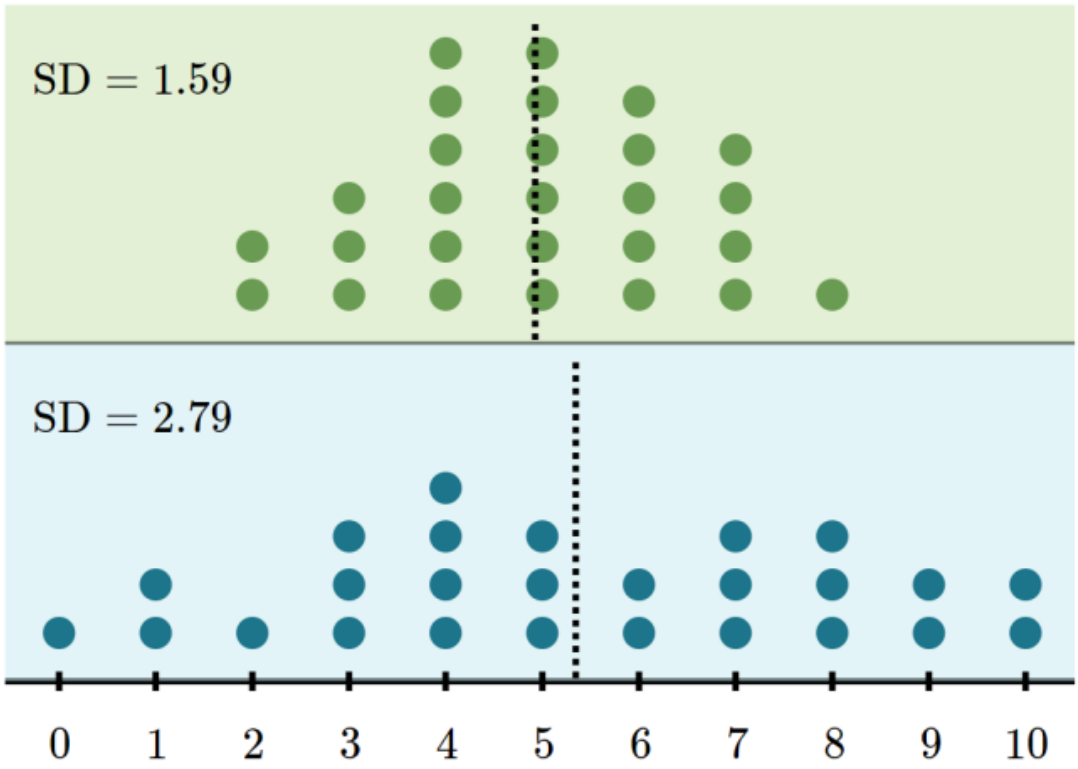
• The standard deviation is a measure of spread that represents the average distance of the data values from their mean.

• It is calculated as the square root of the variance that has been calculated before.

$s=\sqrt{s^2}$

In the previous example, the variance = 4.67 $years^2$
So, the standard deviation, $s=\sqrt{s^2} = \sqrt{4.67}$ = 2.16 years
If the data values are widely spread, the average distance of the values from their mean will be large, and the standard deviation will be large.
If the values are narrowly spread, this average distance will be small, and the standard deviation will be small.

The figure below shows how the spread of data affects the value of the standard deviation.

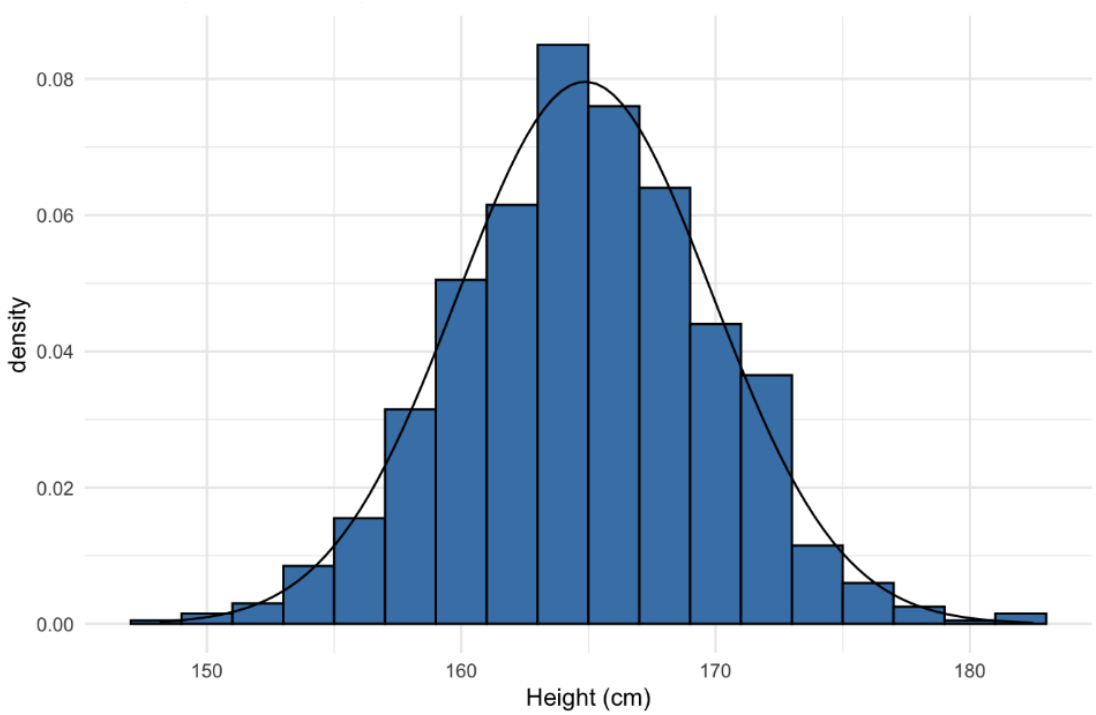


Combining measures of central tendency and measures of dispersion:

When summarizing a numerical variable, we present it using two measures; one for central tendency and one for dispersion.

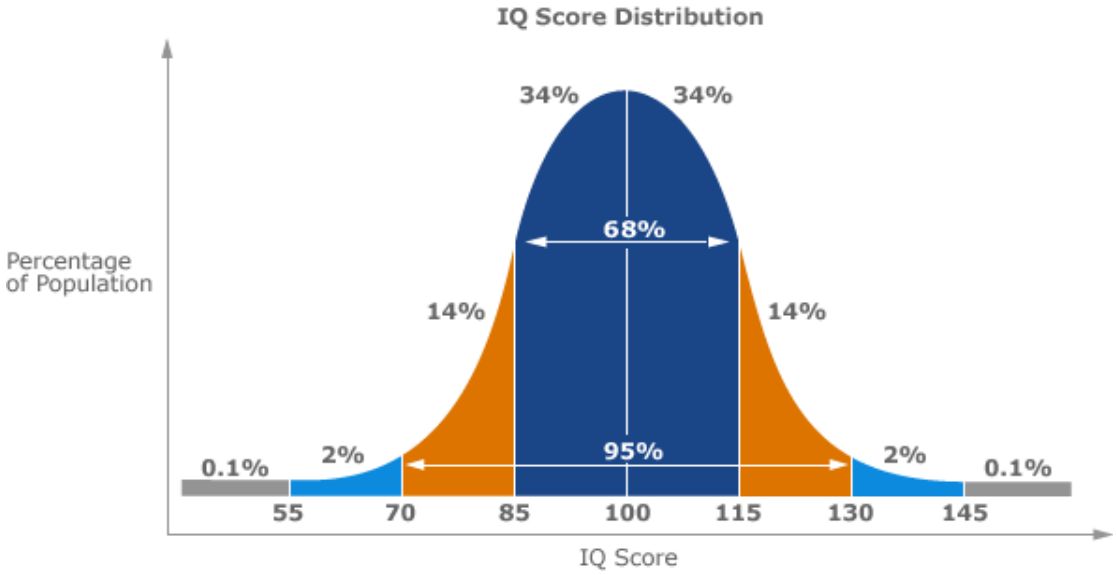
- For the normally distributed data, we use the mean and standard deviation.
- for the non-normally distributed data, we use the median and inter-quartile range (IQR).

What is normally distributed data?

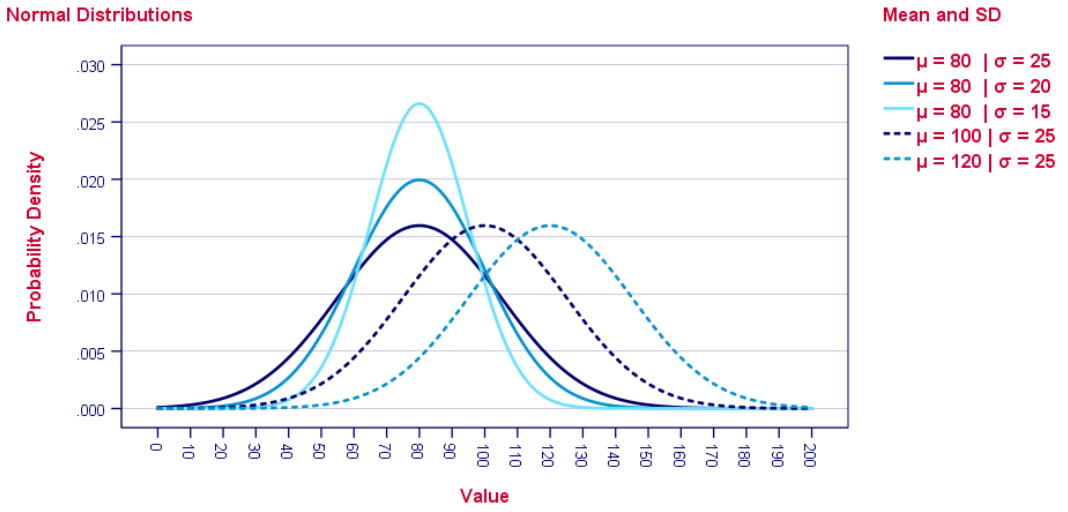


Normally distributed variables are common in biological measurements and have the following characteristics:

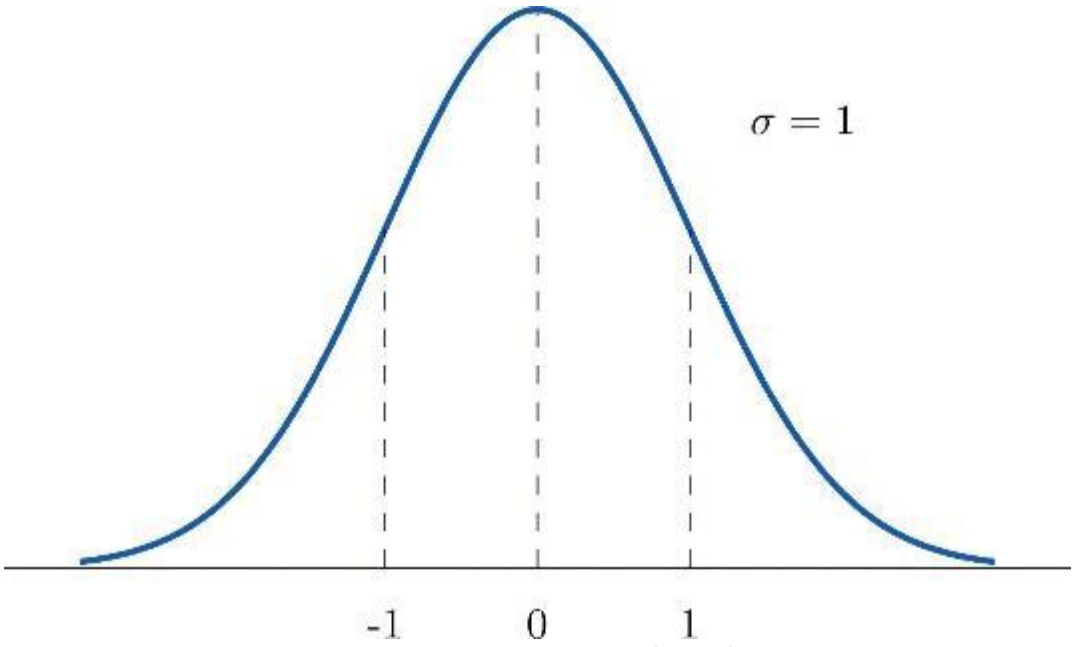
- • Symmetric around the mean.
- • The mean, median, and mode of a normal distribution are equal.
- • Normal distributions are denser in the center and less dense in the tails (bell shape).
- • 50% of values less than the mean and 50% greater than the mean
- • ormal distributions are defined by two parameters, the mean (μ) and the standard deviation (σ).
- 68% of the area of a normal distribution is within one standard deviation of the mean.
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.
- Approximately 99.7% of the area of a normal distribution is within three standard deviations of the mean.



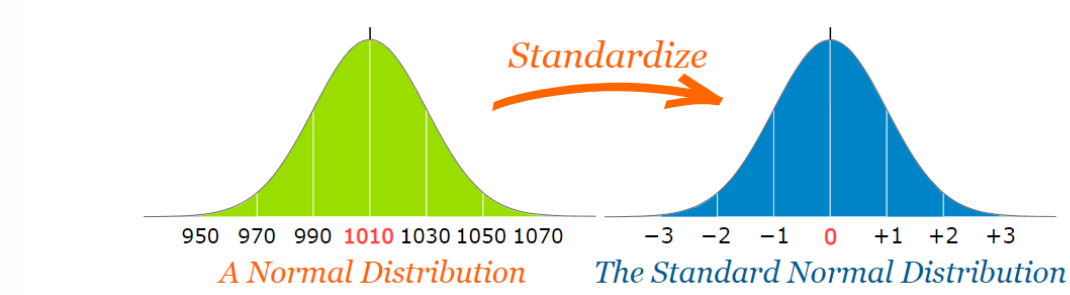
Examples for normally distributed data: height, blood pressure, Q, …
The following graph represents normal distributions with different means and standard deviations:



The normal distribution with mean μ = 0 and standard deviation σ = 1 is called the **standard normal distribution.**

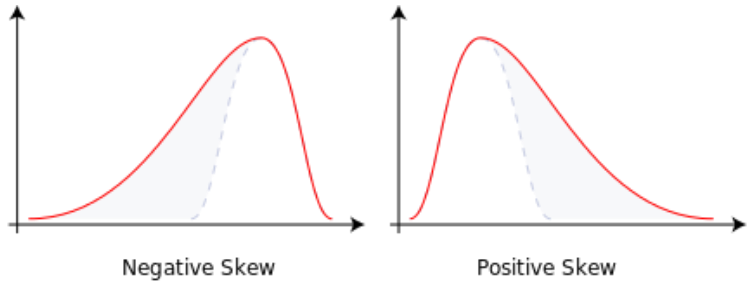


Any normal distribution values can be standardized (transferred to a standardized Z value)



Examples of non-normally distributed data:
Data can be "skewed", meaning it tends to have a long tail on one side or the other.
Positive skew is when the long tail is on the positive side and is skewed to the **right**.

Negative skew is when the long tail is on the negative side and is skewed to the **left**.



Note that the mean is nearer to the tail (it is affected by the extreme values).