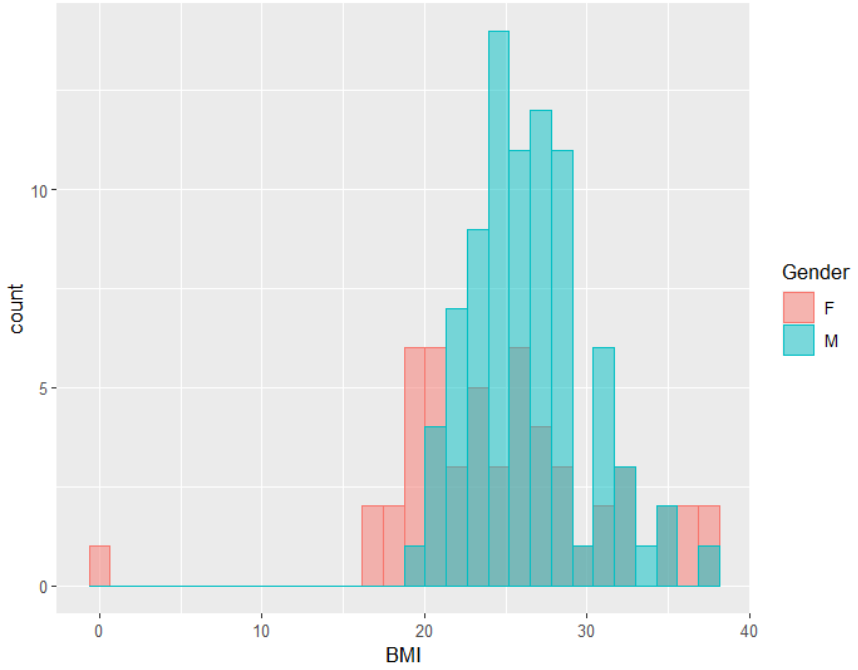
Measures of location and position of the frequency distribution

Measures of location

The frequency distribution typically centers on specific values, which may vary among different groups.



A **measure of location** of the frequency distribution is a **statistic** that indicates the central tendency or typical value of the data set.

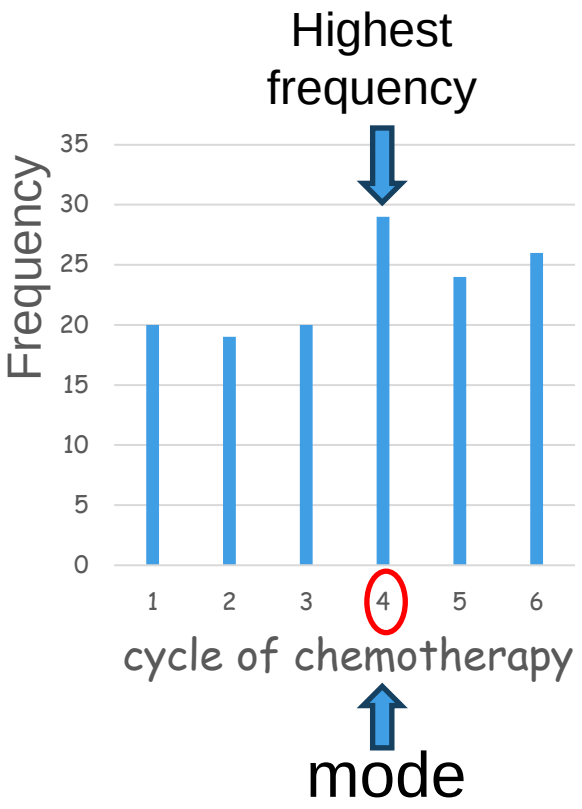
These measures provide information about where the "center" of the distribution lies and can help summarize the overall pattern of the data.

Common measures of location include the mean, median, and mode.

Mode

The mode of a frequency distribution is the value of the considered variable that appears most frequently (i.e., has the highest frequency) in the data.

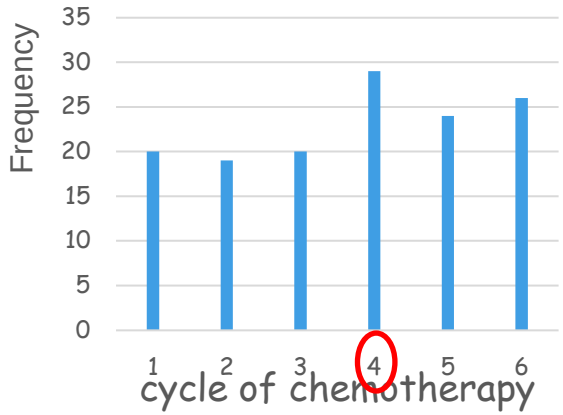
It represents the peak of the distribution so it is the most common value in the dataset.

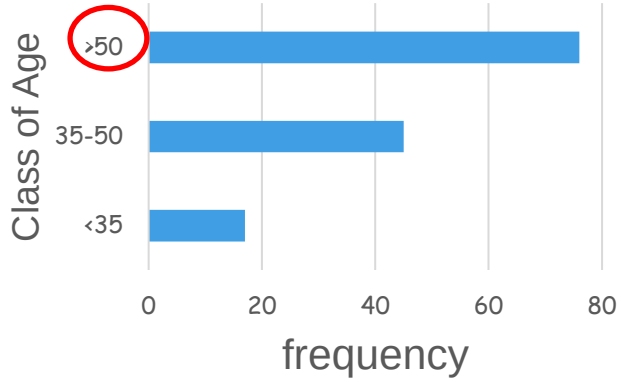


The mode is a measure of the most typical or prevalent value in the dataset.

Mode cont'd

It can be calculated for any kind of variable: numerical/categorical/nominal



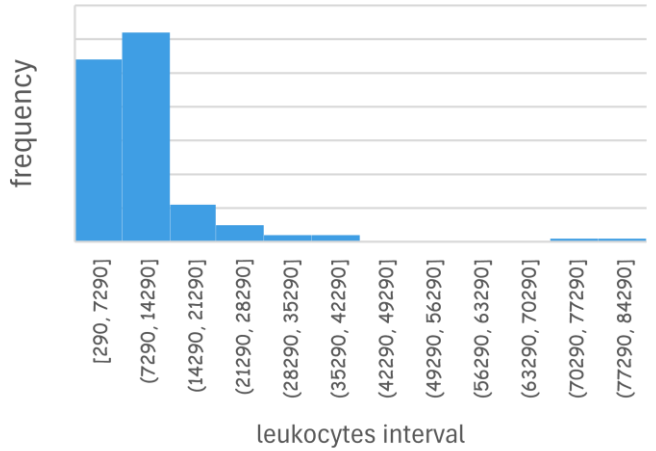


In case of a continuous variable, it is necessary to categorise it.

Data are grouped into intervals.

The mode is calculated with respect to the histogram.

The mode of a histogram is the interval with the highest frequency density.

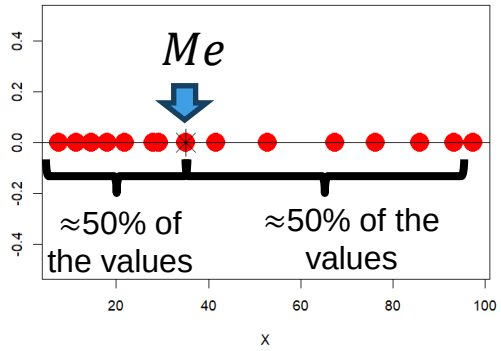


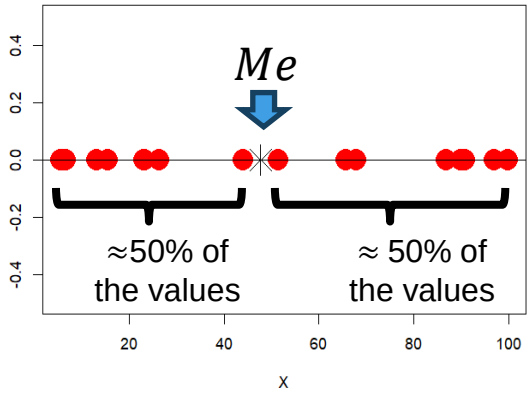
Median

The median is a measure of central tendency that represents the *middle value* of a data. In other words, it is the value that separates the higher half from the lower half of the values of X.

To calculate the median, it is necessary to arrange the X values in ascending order.

Then, if the sample size is **odd**, identify the value Me that lies in the middle of the sample values of X.

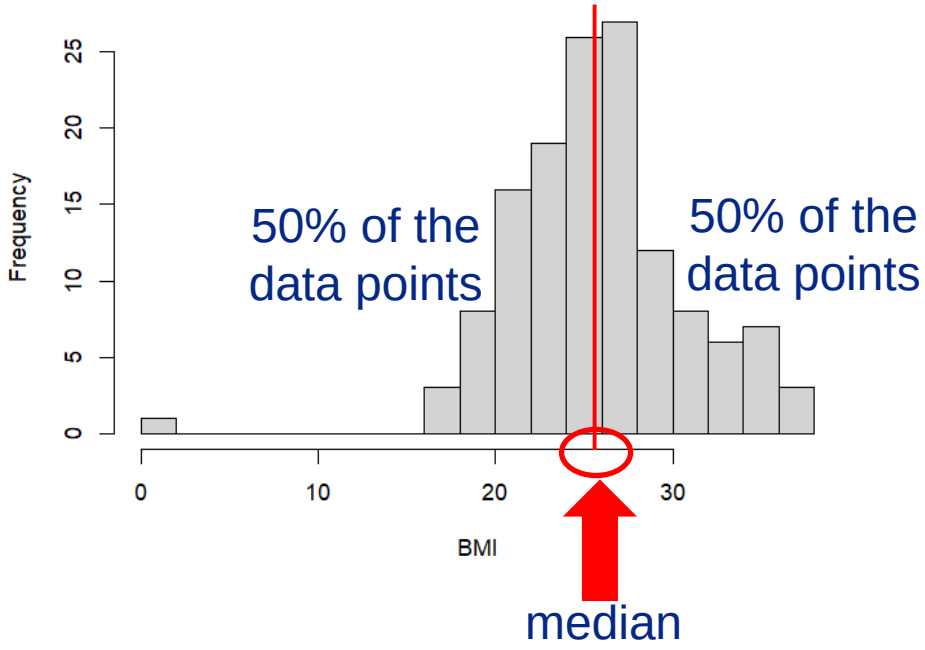




If the sample size is **even**, the semi-sum of the two central values of X represents the median.

Median cont'd

In other words, the median Me is the values that separates (roughly) 50% of the units of the sample lying below it from the rest of the data that are above it



The median can be calculated for categorical variables as well. In this case, the *median class* of the variable is obtained.

It is necessary, however, that the variable be measured on an **ordinal scale**

Mean

Let X be a numerical variable and let $x_1, \ldots x_n$ be the values it takes in the data.

The (arithmetic) mean of $x_1, \ldots x_n$ is defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$

If X takes 5 values (1, 0, –1, 1.5, 0) we get

$$\bar{X} = \frac{1}{5}(1 + 0 + (-1) + 1.5 + 0) = 0.3$$

The power mean

Let X be a numerical variable and let $x_1, \dots x_n$ be the values it takes in the data.

The power mean of $x_1, \dots x_n$ is defined by

$$M^r = \sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} x_i^r} = \sqrt[r]{\frac{1}{n}(x_1^r + \cdots + x_n^r)}$$

$r = 1$ => *the mean*

$r = 2$ => the *quadratic mean* $M^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$

The quantity $M_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ is known as the *second moment* of X

Weighted Mean

Let X be a numerical variable and let $x_1, \ldots x_n$ be the values it takes in the data.

The weighted mean of $x_1, \ldots x_n$ is defined by

$$\bar{X} = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + \cdots + w_n x_n$$

$w_1, \ldots, w_n$ are known as *weights* and must satisfy

- $w_i \geq 0$
- $\sum_{i=1}^{n} w_i = 1$

Idea: instead of treating all values equally, the weighted mean gives more importance (or weight) to certain values over other.

Weighted Mean: examples

- Those of you who graduated in Italy surely remember that the average score is obtained by weighting the marks with the number of credits (ECTS) awarded for each exam.
- If you need to calculate the average GDP of EU countries Republic of Cyprus and Germany cannot have the same relevance (need to weight using population or other measures).
- Suppose a clinical trial is conducted to compare the effectiveness of three different treatments. Patients are randomly assigned to one of the 3 treatment groups. Assume that the number of patients in each treatment group varies due to different recruitment rates or dropout rates. To calculate an appropriate average of effectiveness we need to adjust (to weight) for the number of patients in each treatment group.

Weighted Mean: examples cont'd

If data came in classes, we would need to adjust for the different number of cases in each class.

Number of deaths for cancer in the Veneto Region (2007 SER)

Age class $c_j - c_{j-1}$	Mid-point x_j	Number of deaths n_j	Weights w_j
0	0.5	1	0.0001
01-14	7.5	21	0.0015
15-29	22.5	39	0.0028
30-44	37	260	0.0190
45-64	54.5	2503	0.1825
65-74	69.5	3628	0.2646
75+	80	7261	0.5295
		13713	1

$$\text{Mid-point} = c_{j-1} + \frac{c_j - c_{j-1}}{2}$$

$$\text{w}_\text{j} = \frac{n_j}{\sum_{j=1}^{k} n_j} \quad \text{weight of class j}$$

Average age at death $\frac{1}{7}\sum_{i=1}^{7} w_i x_i = 71.5$ years

Geometric mean

Let X be a numerical variable and let $x_1, \dots x_n$ be the values it takes in the data.

Assume that $x_i > 0$ for any $i = 1, \dots, n$.

The geometric mean of $x_1, \dots x_n$ is defined by

$$G = \sqrt[n]{x_1 \times \cdots \times x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

where $x_1 \times \cdots \times x_n = \prod_{i=1}^{n} x_i$.

The geometric mean is commonly used when dealing with positive quantities that multiply together over time as a percentage, such as growth rates or rates of change.

Geometric mean cont'd

Example. A strain of bacteria increases its population by 8.75%, 26.19% and 100% growing from 10000 to 27446 units over three days.

The growth rates of the three days are 1.0875, 1.2619 and 2.

The geometric mean of the rates is ≈**1.4001**.

This is the compound rate, i.e., the daily rate we need to pass from 10000 at day 1 to 27445 at day 3.

In this case the arithmetic mean of the rates, **1.45,** does not satisfy this.

Note: 1.4001 < 1.45. The relationship $G < \bar{X}$ always holds true.

Note: $\log G = \frac{1}{n}\log(x_1 \times \cdots \times x_n) = \frac{1}{n}\sum_{i=1}^{n} \log x_i \;\; = \bar{Y}$ where $Y = \log X$ so $G = e^{\frac{1}{n}\sum_{i=1}^{n} \log x_i}$

Measures of position

A measure of position is a statistical summary that indicates the relative location of a specific value within a dataset.

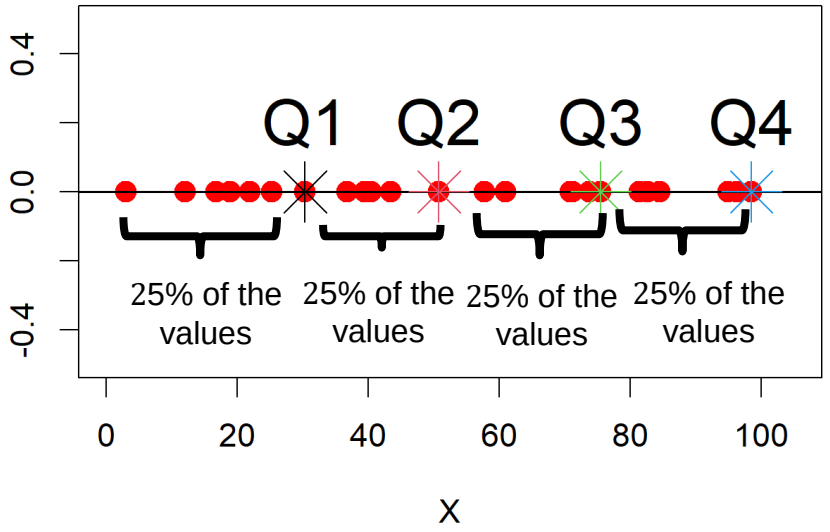
These measures provide insights into the distribution and spread of the data, allowing for comparisons and making it easier to understand the relative position of specific values within a dataset, i.e. where individual data points fall within the distribution of the data.

Common measures of position include percentiles, quartiles, and deciles.

Quartiles

Quartiles divide the dataset into four equal parts, each containing 25% of the data.

To calculate the quartiles, it is necessary to arrange the values in ascending order.



The first quartile (Q1) is the value below which 25% of the data falls, the second quartile (Q2) is the median, and the third quartile (Q3) is the value below which 75% of the data falls.

Quartiles cont'd

Let X be a numerical variable and let $x_1, \dots, x_n$ be the values it takes in the data.

Let $x_{(1)}, \dots, x_{(n)}$ be the values arranged in ascending order.

$x_{(1)}, \dots, x_{(n)}$ are called *ordered statistics* and their position r is called the *rank*.

x_1	x_2	x_3	x_4	x_5
17	8	9	11	15
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
8	9	11	15	17
r_1	r_2	r_3	r_4	r_5
5	1	2	3	4

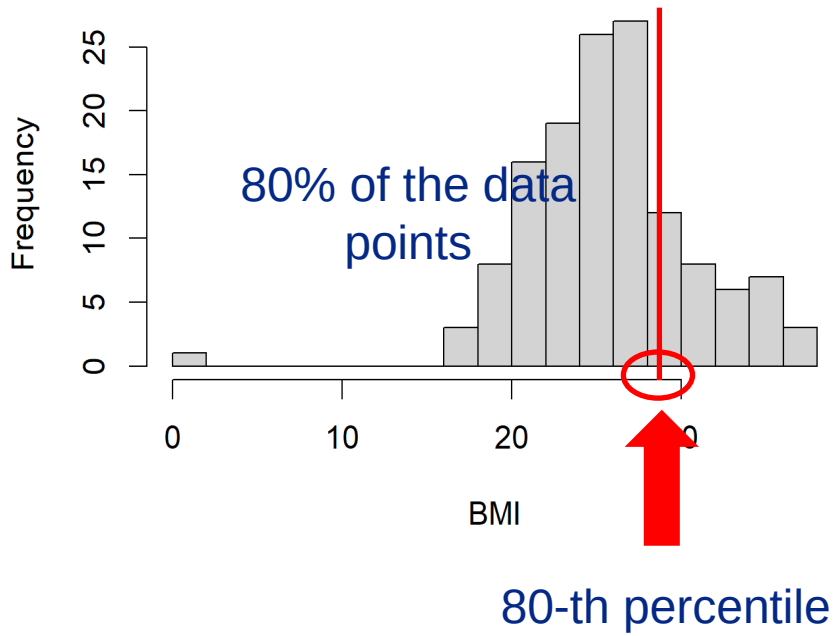
So $x_{(1)}$, the minimum of the observed data has rank $r = 1$ and $x_{(n)}$, the maximum of the observed data has rank $r = n$.

If position h in the ordered set is such that $h - 1 \leq n \times \frac{l}{4} \leq h$ the $l - th$ quartile is

$$Q_l = \begin{cases} x_{(h)} & if \quad n\frac{l}{4} > h - 1 \\ \frac{x_{(h)} + x_{(h-1)}}{2} & if \quad n\frac{l}{4} = h - 1 \end{cases}$$

Percentiles

A *percentile* is the value of X that identifies the percentage of data points that fall below a certain value. For example, the 80th percentile is the value below which 80% of the data points fall.



Deciles divide the dataset into ten equal parts, each containing 10% of the data. The first decile (D1) is the value below which 10% of the data falls, the second decile (D2) is the value below which 20% of the data falls, and so on.