Quantitative Methods – I (Statistics)

A. Y. 2022-23

Prof. Lorenzo Cavallo

Chapter 3

Summarizing Data

Summarizing Data: Road Map

1. **Measures of position (for both population and samples)**

i. Mode, median, quartiles, percentiles;
ii. Simple mean, trimmed mean;
iii. Weighted mean (grouped data)
iv. Relationships among the measures of position

2. **Measures of dispersion (for both population and samples)**

i. Range, interquartile range;
ii. Variance, standard deviation, coefficient of variation

3. **Box-Plot**

Measures of Central Tendency (Position)

Mode

Value/category/class with the highest frequency

Median

Value of the observation(s) in the middle of the ranked data, where the middle position is $\frac{n+1}{2}$

Quartiles

Three values that divide the **ranked** data into four equal parts

Percentiles

Values that divide the **ranked** data into 100 equal parts

MEANS

Arithmetic mean/average

Sum of all values divided by number of observations

POPULATION

$$\mu = \frac{\sum x}{N}$$

SAMPLE

$$\bar{x} = \frac{\sum x}{n}$$

Geometric mean

The *n*th root of the product of all observations

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

Harmonic mean

The reciprocal of the arithmetic mean

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

What are the measures of central tendency?

A measure of central tendency (also referred to as measures of centre, central location or position) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

The three main measures of central tendency are:

1. the mode
2. the median (also with quartiles and percentiles)
3. the mean (arithmetic, geometric, trimmed, harmonic, etc.)

Each of these measures describes a different indication of the typical or central value in the distribution.

What is the Mode?

The mode is the most commonly occurring value in a distribution.

Advantage of the mode:

The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the mode:

The are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well.

Mode

Consider this dataset showing the retirement age of 11 people, in whole years:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

This table shows a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Mode

Plus: The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Minus: It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.

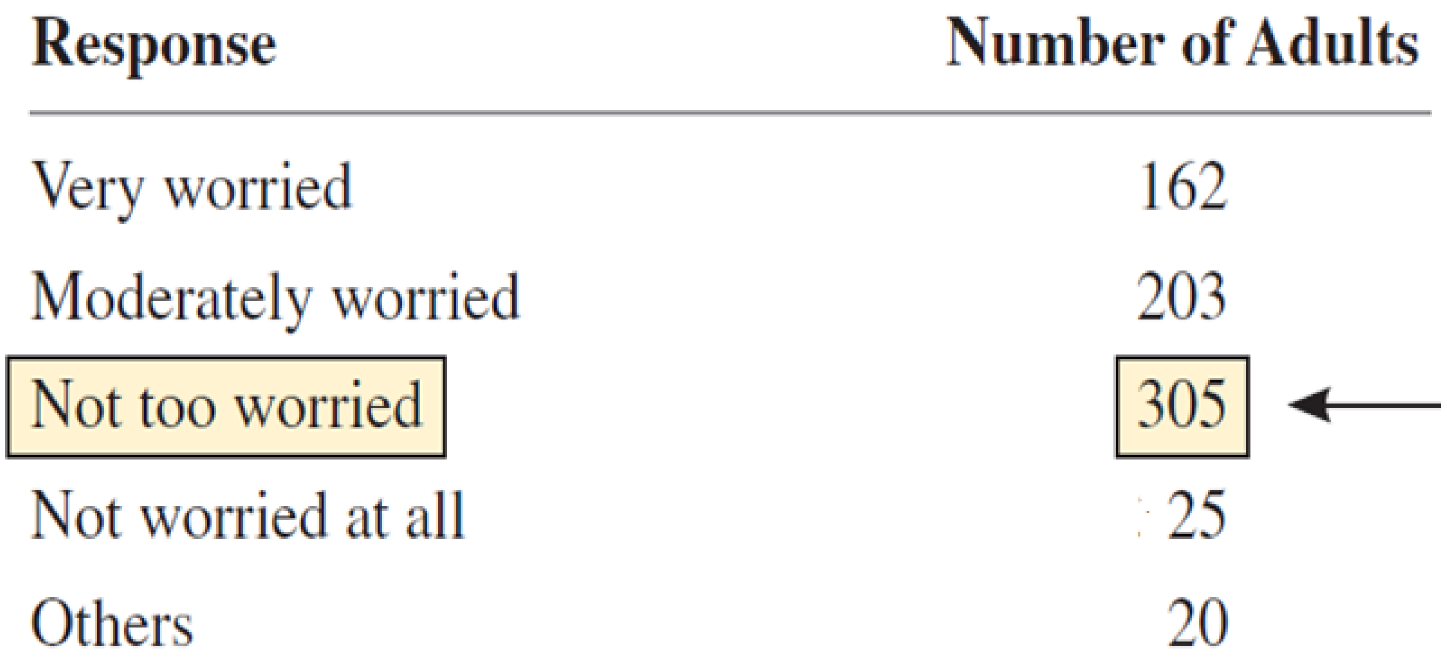
In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).

Mode

Mode: Value/category/class with the highest frequency

Plus: can be computed for ALL types of variables, included Qualitative ones.

Ex: mode for worries about reaching the end of the month… is «Not too worried» (and not 305!)



Response	Number of Adults
Very worried	162
Moderately worried	203
Not too worried	305
Not worried at all	25
Others	20

Mode

Mode: Value/category/class with the highest frequency

Plus: can be computed for ALL types of variables

Minus: a dataset can have 1, 2, 2+, or even…no mode!

Ex: The following data give the speeds (in miles per hour) of 8 cars that were stopped for speeding violations.

77 82 74 81 79 84 74 78

Mode? 74 ➔ UNIMODAL

Mode

Mode: Value/category/class with the highest frequency

Plus: can be computed for ALL types of variables

Minus: a dataset can have 1, 2, 2+, or even…no mode!

Ex: The following data give the speeds (in miles per hour) of 9 cars that were stopped for speeding violations.

77 82 74 81 79 84 74 78 77

Mode? 74 and 77 ➔ BIMODAL

Mode

Mode: Value/category/class with the highest frequency

Plus: can be computed for ALL types of variables

Minus: a dataset can have 1, 2, 2+, or even…no mode!

Ex: The following data give the speeds (in miles per hour) of 10 cars that were stopped for speeding violations.

77 82 74 81 79 84 74 78 77 81

Mode? 74 and 77 and 81 ➔ MULTIMODAL

Mode

Mode: Value/category/class with the highest frequency

Plus: can be computed for ALL types of variables

Minus: a dataset can have 1, 2, 2+, or even…no mode!

Ex: The following data give the speeds (in miles per hour) of 10 cars that were stopped for speeding violations.

77 82 74 81 79 84 85 78 87 91

Mode? ….➔ NO MODE!

What is the Median?

The median is the **middle value** in distribution when the values are arranged in ascending or descending order.

The median divides the distribution in half (there are 50% of observations on either side of the median value).

Advantage of the median:

The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median:

The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

Median

Value of the observation(s) in the middle of the ***ranked*** data, where the middle position is $\frac{n+1}{2}$

- **Meaning:** 50% of the observations in the dataset have value lower than the median, 50% have it higher

- **Plus**: not sensitive to outliers

- **Minus**: (i) uses only few obs (1 or 2) of the distribution,
  (ii) computation differs for *odd* or *even* number of obs.

Median: *odd* number of observations

Median: Value of the observations in the middle of the ***ranked*** data

If nr of obs is *odd*: only one middle observation, in position $\frac{n+1}{2}$

Ex: These are the speeds (in miles per hour) of 9 cars stopped for speeding violations.

77 82 74 81 79 84 74 78 77

To compute the median:

1. Rank the observations in ascending order

74 74 77 77 78 79 81 82 84

Median: ***odd*** number of observations

Median: Value of the observations in the middle of the ***ranked*** data

If nr of obs is *odd*: only one middle observation, in position $\frac{n+1}{2}$

Ex: These are the speeds (in miles per hour) of **9** cars stopped for speeding violations.

77 82 74 81 79 84 74 78 77

To compute the median:

1. Rank the observations in ascending order

   74 74 77 77 78 79 81 82 84

2. Find the one in the position $\frac{n+1}{2} = \frac{9+1}{2} = 5^{\text{th}}$ value → median is 78

Median: *even* number of observations

With *even* nr of obs: two middle observations (since $\frac{n+1}{2}$ is non integer)

Ex: These are the speeds (in miles per hour) of 10 cars stopped for speeding violations.

77 82 74 81 79 84 74 78 77 94

To compute the median:

1. Rank the observations in ascending order

74 74 77 77 78 79 81 82 84 94

Median: *even* number of observations

With *even* nr of obs: two middle observations (since $\frac{n+1}{2}$ is non integer)

Ex: These are the speeds (in miles per hour) of 10 cars stopped for speeding violations.

77 82 74 81 79 84 74 78 77 94

To compute the median:

1. Rank the observations in ascending order

   74 74 77 77 78 79 81 82 84 94

2. Find the one in the position $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$ ➔ median is between the 5th value and the 6th value …

Median: ***even*** number of observations

1. Rank the observations in ascending order

   74 74 77 77 78 79 81 82 84 94

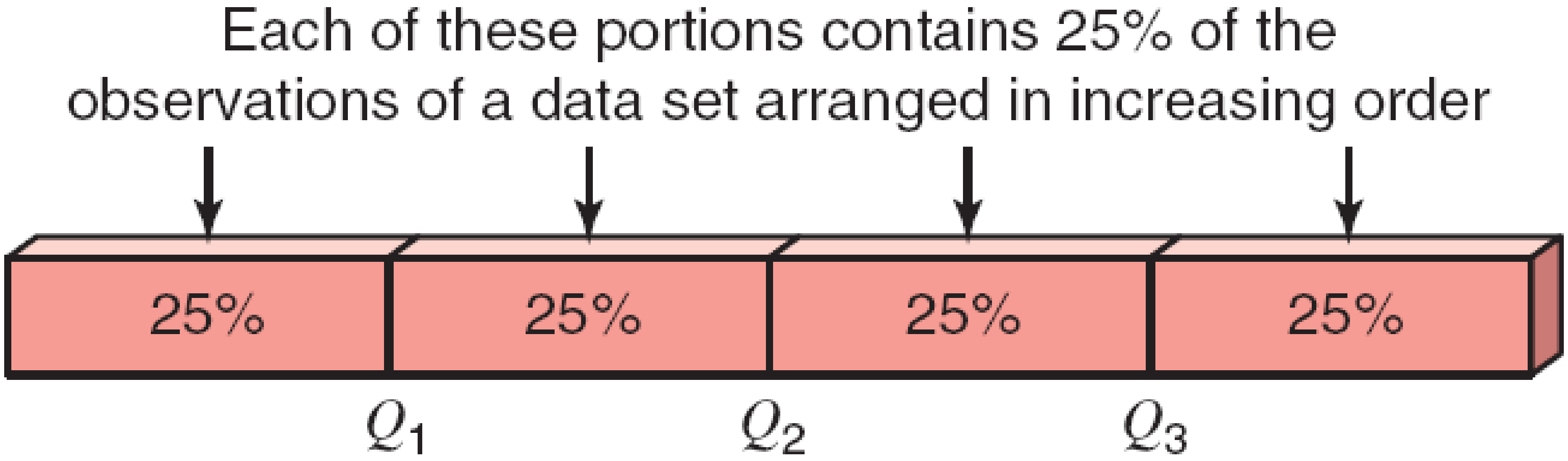
2. Find the one in the position $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$ ➔ median is between the 5th value and the 6th value

3. The 5th value is 78, the 6th value is 79 so the median is between these two values ➔ $\frac{5^{th}+6^{th}}{2} = \frac{78+79}{2} = 78.5$

Quartiles

In statistics, a quartile divides the number of data points into four parts, or quarters. The data must be ordered from smallest to largest to compute quartiles.

Three values that divide the *ranked* data into four equal parts.



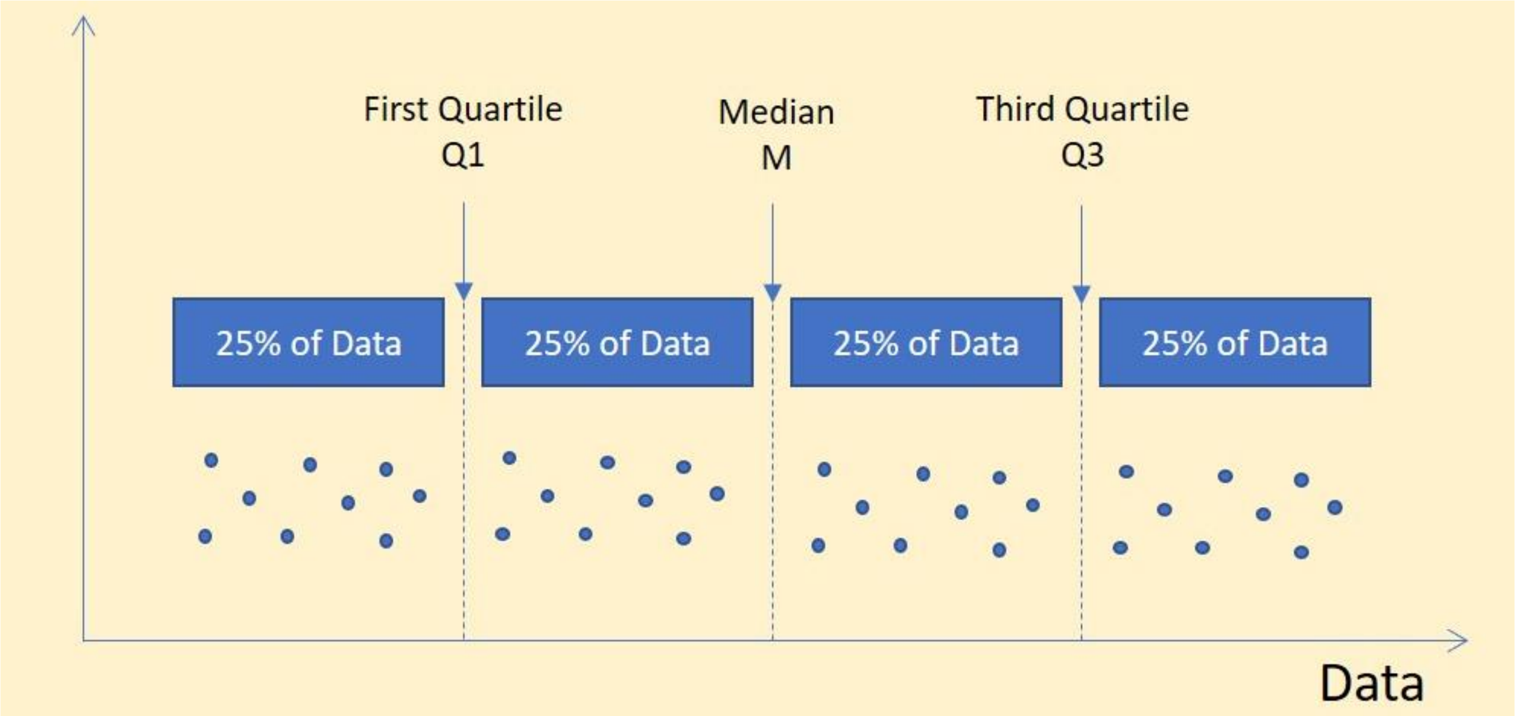
Each of these portions contains 25% of the observations of a data set arranged in increasing order

Quartiles

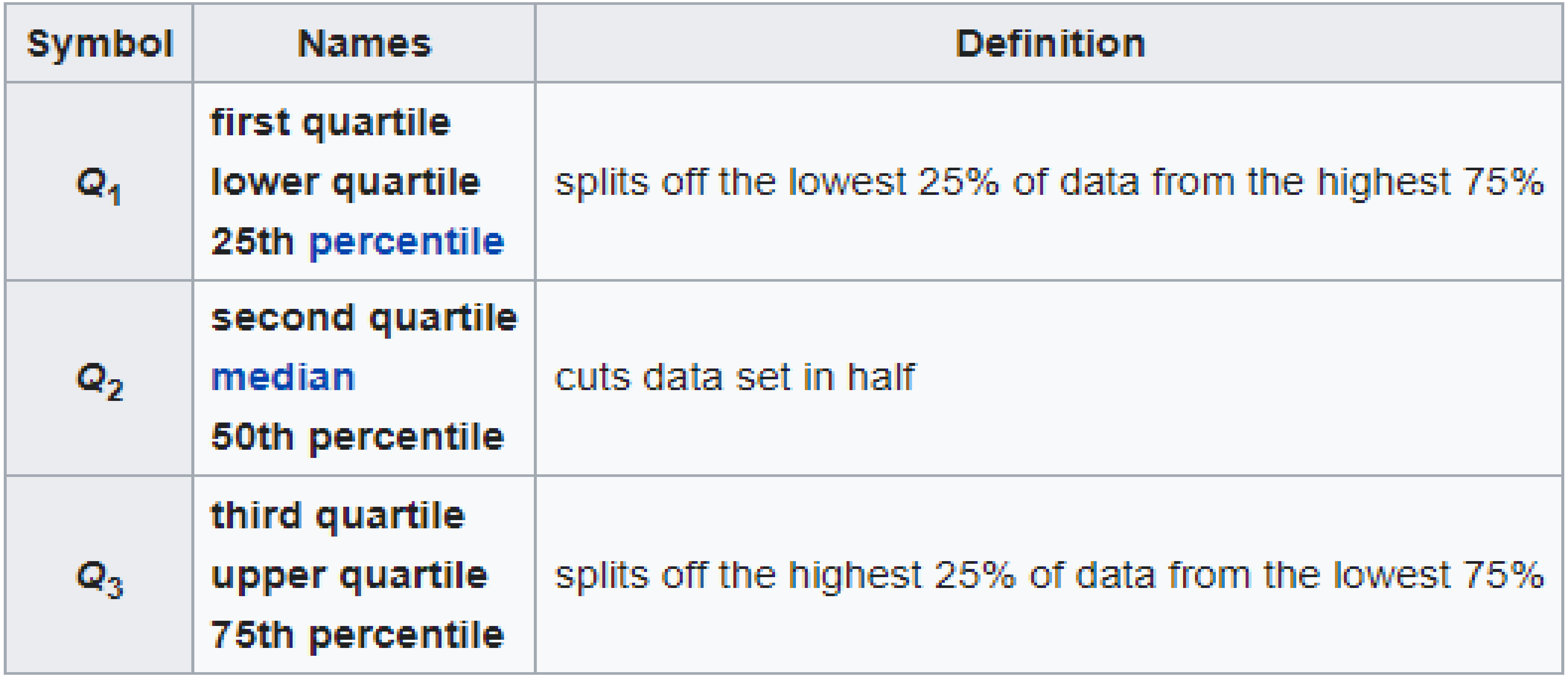
The **first quartile (Q1)** is defined as the middle number between the smallest number (minimum) and the median of the data set.

The **second quartile (Q2)** is the **median** of a data set.

The **third quartile (Q3)** is the middle value between the median and the highest value (maximum) of the data set).



Quartiles



Symbol	Names	Definition
Q_1	**first quartile** **lower quartile** **25th percentile**	splits off the lowest 25% of data from the highest 75%
Q_2	**second quartile** **median** **50th percentile**	cuts data set in half
Q_3	**third quartile** **upper quartile** **75th percentile**	splits off the highest 25% of data from the lowest 75%

Quartiles: Computing methods

Use the median to divide the ordered data set into two-halves.

- If there is an odd number of data points in the original ordered data set, **do not include the median** in either half.
- If there is an even number of data points in the original ordered data set, split this data set exactly in half.

The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.

Quartiles: Computing methods

Ex. Find the quartiles of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

26 24 30 57 40 25 32 27 44 52 50 55

Quartiles: Computing methods

Ex. Find the quartiles of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

26 24 30 57 40 25 32 27 44 52 50 55

Sort the data from smallest to largest

24, 25, 26, 27, 30, 32, 40, 44, 50, 52, 55, 57

n=12 so the position of the median is $\frac{n+1}{2} = \frac{12+1}{2} = 6.5$ ➔ median is between 6th (32) and 7th (40) value ➔ $\text{Median} = \frac{32+40}{2} = 36$

Quartiles: Computing methods

24, 25, 26, 27, 30, 32, 40, 44, 50, 52, 55, 57

There is an even number of data points in the original data set (n=12), split this data set exactly in half.

First half (values smaller than the Median=36):

24,25,26,27,30,32

Second half (values larger than the Median=36):

40,44,50,52,55,57

Quartiles: Computing methods

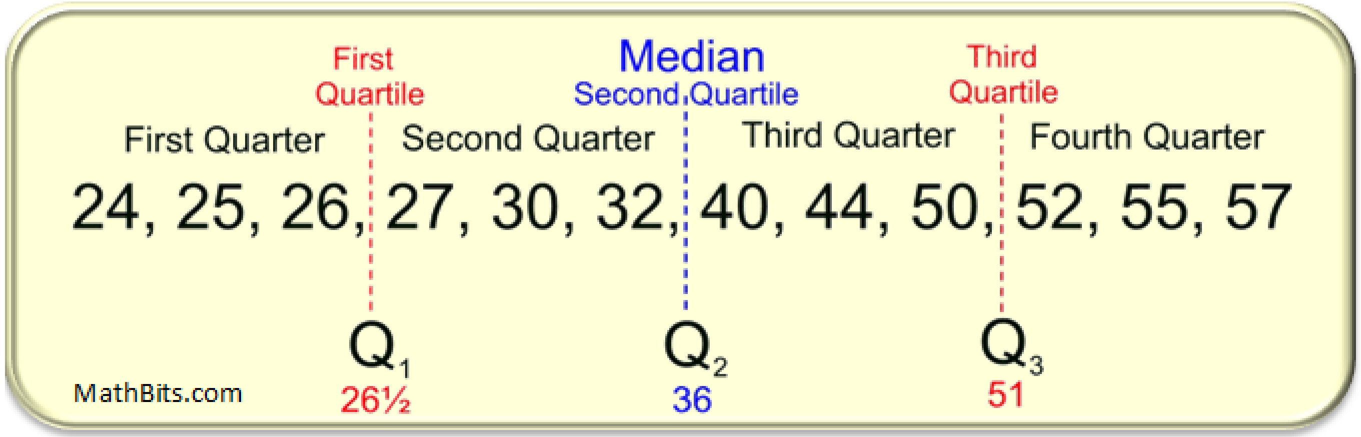
First half (values smaller than the Median=36):

24,25,26,27,30,32

Second half (values larger than the Median=36):

40,44,50,52,55,57

The Median of the 2 distributions (position $\frac{n+1}{2} = \frac{6+1}{2} = 3.5$) are the Q1 and the Q3



Quartiles: example

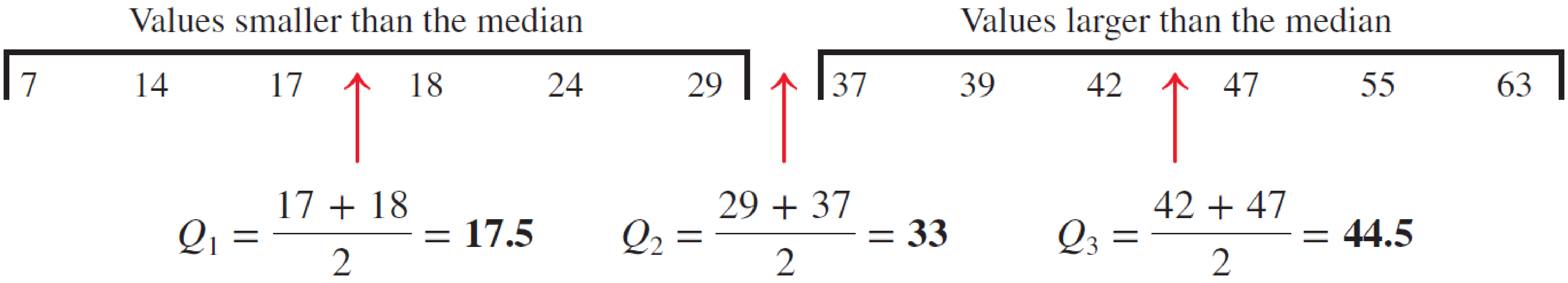
Find the quartiles of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

$$29 \quad 14 \quad 39 \quad 17 \quad 7 \quad 47 \quad 63 \quad 37 \quad 42 \quad 18 \quad 24 \quad 55$$

Quartiles: example

Find the quartiles of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55



Values smaller than the median

7 14 17 18 24 29

Values larger than the median

37 39 42 47 55 63

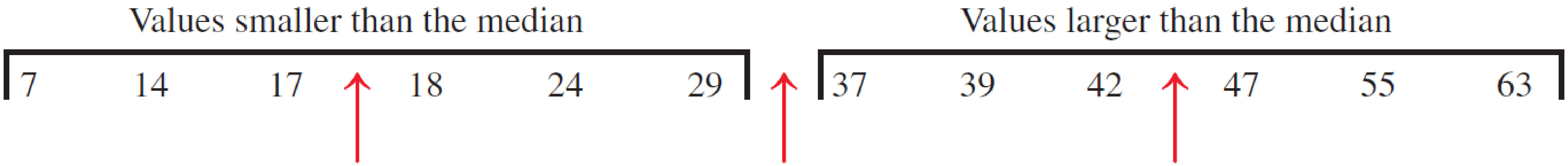
$$Q_1 = \frac{17+18}{2} = \mathbf{17.5} \qquad Q_2 = \frac{29+37}{2} = \mathbf{33} \qquad Q_3 = \frac{42+47}{2} = \mathbf{44.5}$$

Interpretation: 25% of the students commute for less (and 75% commute for more) than 17.5 minutes.

Quartiles: example

Find the quartiles of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55

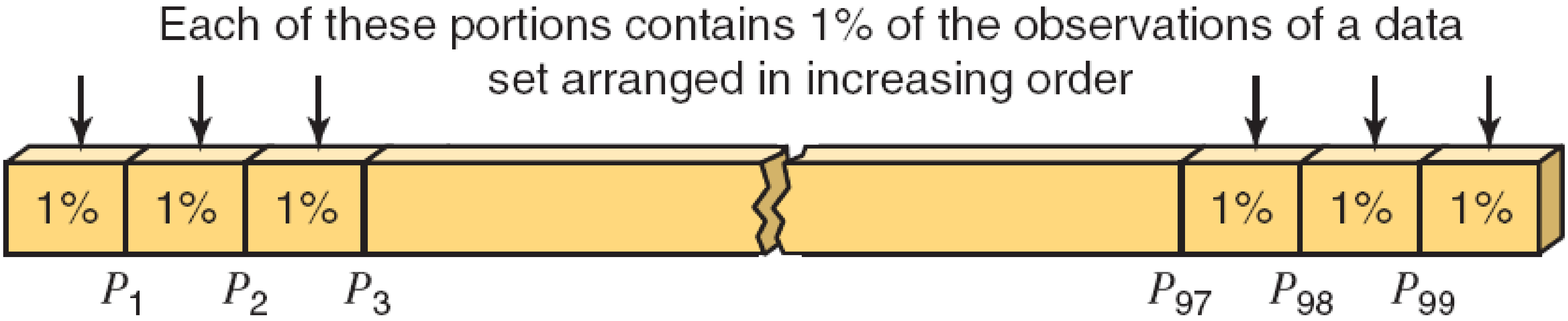


$$Q_1 = \frac{17+18}{2} = \mathbf{17.5} \qquad Q_2 = \frac{29+37}{2} = \mathbf{33} \qquad Q_3 = \frac{42+47}{2} = \mathbf{44.5}$$

Question: Where does the commuting time of 47 fall in relation to the three quartiles? It lies in the **top 25%** of the commuting times.

Percentiles

Values that divide the ***ranked*** data into **100** equal parts.


Each of these portions contains 1% of the observations of a data set arranged in increasing order
1%
1%
1%
1%
1%
1%
P_1
P_2
P_3
P_{97}
P_{98}
P_{99}

The *k*-th *percentile* is the value of the observation in the $\frac{k \times n}{100}$ position (rounded up), where *n* is the dataset size.

Which percentile is the median?

Percentiles: example

Find and interpret the 70th percentile of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55

- Since $\frac{k \times n}{100} = \frac{70 \times 12}{100} = 8.4$ (rounded up 9) ➔ the 70th percentile is the value is the 9th observation of the ***ranked*** dataset
- Sort the data

7 14 17 18 24 29 37 39 42 47 55 63

Percentiles: example

Find and interpret the 70th percentile of the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55

- Since $\frac{k \times n}{100} = \frac{70 \times 12}{100} = 8.4$ (rounded up 9) ➔ the 70th percentile is the value is the 9th observation of the ***ranked*** dataset ➔ 42 minutes.

- **Interpretation**: 70% of these 12 students commute for 42 minutes or less.

Percentiles: other examples

Given the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55

1. Find and interpret the 12th percentile

Percentiles: other examples

Step 1 ➔ sort the data

7 14 17 18 24 29 37 39 42 47 55 63

Step 2 ➔ $\frac{k\times n}{100} = \frac{12\times 12}{100} = 1.14$ (rounded up 2) ➔ the 12th percentile is the value of the 2nd observation of the ***ranked*** dataset ➔ 14 minutes.

Step 3 ➔ Interpretation: 12% of these 12 students commute for 14 minutes or less.

Percentiles: other examples

Given the following (one-way) commuting times (in minutes) from home to TV for 12 students:

29 14 39 17 7 47 63 37 42 18 24 55

1. Find and interpret the 60th percentile

Percentiles: other examples

Step 1 ➔ sort the data

7 14 17 18 24 29 37 39 42 47 55 63

Step 2 ➔ $\frac{k \times n}{100} = \frac{60 \times 12}{100} = 7.2$ (rounded up 8) ➔ the 60th percentile is the value of the 8th observation of the ***ranked*** dataset ➔ 39 minutes.

Step 3 ➔ Interpretation: 60% of these students commute for 39 minutes or less

What is the Mean?

The mean is the sum of the value of each observation in a dataset divided by the number of observations.

This is also known as the arithmetic average.

Looking at the retirement age distribution again:

54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

Mean

Advantage of the mean:

The mean can be used for both continuous and discrete numeric data.

Limitations of the mean:

The mean cannot be calculated for categorical data, as the values cannot be summed.

As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

The **population mean** is indicated by the Greek symbol μ (pronounced ‘mu’). When the mean is calculated on a sample distribution it is called **sample mean** indicated by the symbol $\bar{x}$ (pronounced x-bar).

Simple (Arithmetic) Mean

Mean/Average: sum of all values divided by number of observations

POPULATION

$$\mu = \frac{\sum x}{\mathrm{N}}$$

SAMPLE

$$\bar{x} = \frac{\sum x}{n}$$

Pro: most widely used measure of central tendency; univocal; uses all observations in the dataset

Con: only for quantitative variables; sensitive to outliers

Simple (Arithmetic) Mean: Example Population

Mean/Average: sum of all values divided by number of observations

POPULATION **SAMPLE**

$$\mu = \frac{\sum x}{\mathrm{N}} \qquad \bar{x} = \frac{\sum x}{n}$$

Ex: These are the number of books read in the last year by ***all*** 10 residents in a house

10 1 9 4 4 1 3 3 4 1

Mean? $\frac{\sum x}{N} = \frac{40}{10} = 4$

Simple (Arithmetic) Mean: Example sample

Mean/Average: sum of all values divided by number of observations

POPULATION **SAMPLE**

$$\mu = \frac{\sum x}{\mathrm{N}} \qquad \bar{x} = \frac{\sum x}{n}$$

Ex: These are the number of books read in the last year by ***a sample of*** 10 residents in New York

10 1 9 4 4 1 3 3 4 1

Mean? $\frac{\sum x}{n} = \frac{40}{10} = 4$

Simple (Arithmetic) Mean: sensitivity

Consider the mean for a sample of observations (same for population):

$$\bar{x} = \frac{\sum x}{n}$$

Ex: These are the number of books read in the last year by ***a sample of*** 10 residents in New York

40 1 9 4 4 1 3 3 4 1

Mean? $\frac{\sum x}{n} = \frac{70}{10} = \mathbf{7}$

Without the first observation the mean is $\frac{\sum x}{n} = \frac{30}{9} = \mathbf{3.33} \neq \frac{70}{10} = \mathbf{7}$

Trimmed Mean

Mean computed on a subset of observations, dropping one portion at each end of the **ranked** data.

For example, the 10% trimmed mean is obtained dropping 10% of observations at each end of the ranked data

Pros: univocal, less sensitive to outliers

Cons: only for quantitative variables; not clear which portion to drop

Trimmed Mean: example

Compute the 10% trimmed mean of the books read last year by a sample of 10 residents in New York

Ex: These are the **ranked** books read in the last year

Case I:	1	1	3	3	4	4	4	9	10
Case II:	1	1	3	3	4	4	4	9	**40**

Dropping 10% of observations (i.e. 1) at each end, we get $\frac{\sum x}{n} = \frac{28}{8} = 3.5$ in both cases

Simple (Arithmetic) Mean: Raw distribution

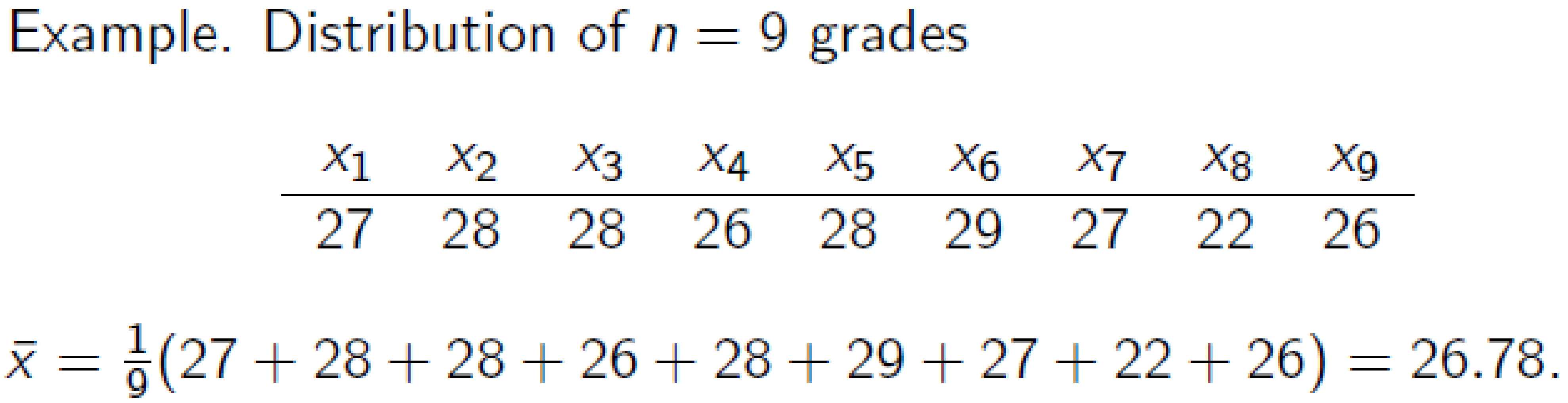
Raw distribution (or Unit distribution)

POPULATION

$$\mu = \frac{\sum x}{\mathrm{N}}$$

SAMPLE

$$\bar{x} = \frac{\sum x}{n}$$



Example. Distribution of $n = 9$ grades

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
27	28	28	26	28	29	27	22	26

$$\bar{x} = \tfrac{1}{9}(27 + 28 + 28 + 26 + 28 + 29 + 27 + 22 + 26) = 26.78.$$

Simple (Arithmetic) Mean: Frequency distribution

Frequency distribution

Sum of values times the frequencies, divided by number of obs.

POPULATION

$$\mu = \frac{\sum x_i f_i}{\mathrm{N}}$$

SAMPLE

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

where $x_i$are the values, and f_i the associated frequencies.

Weighted Mean (grouped data)

Sum of values weighted by frequencies, divided by number of obs

POPULATION

$$\mu = \frac{\sum x_i f_i}{\mathrm{N}}$$

SAMPLE

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

where $x_i$are the values, and f_i the associated frequencies.

Pro: most widely used measure of central tendency; univocal

Con: only for quantitative variables; sensitive to outliers

Weighted Mean (grouped data): example

The following table reports the prices and quantities of gas bought by ***all*** 4 drivers of a company in this week. Find the average price paid

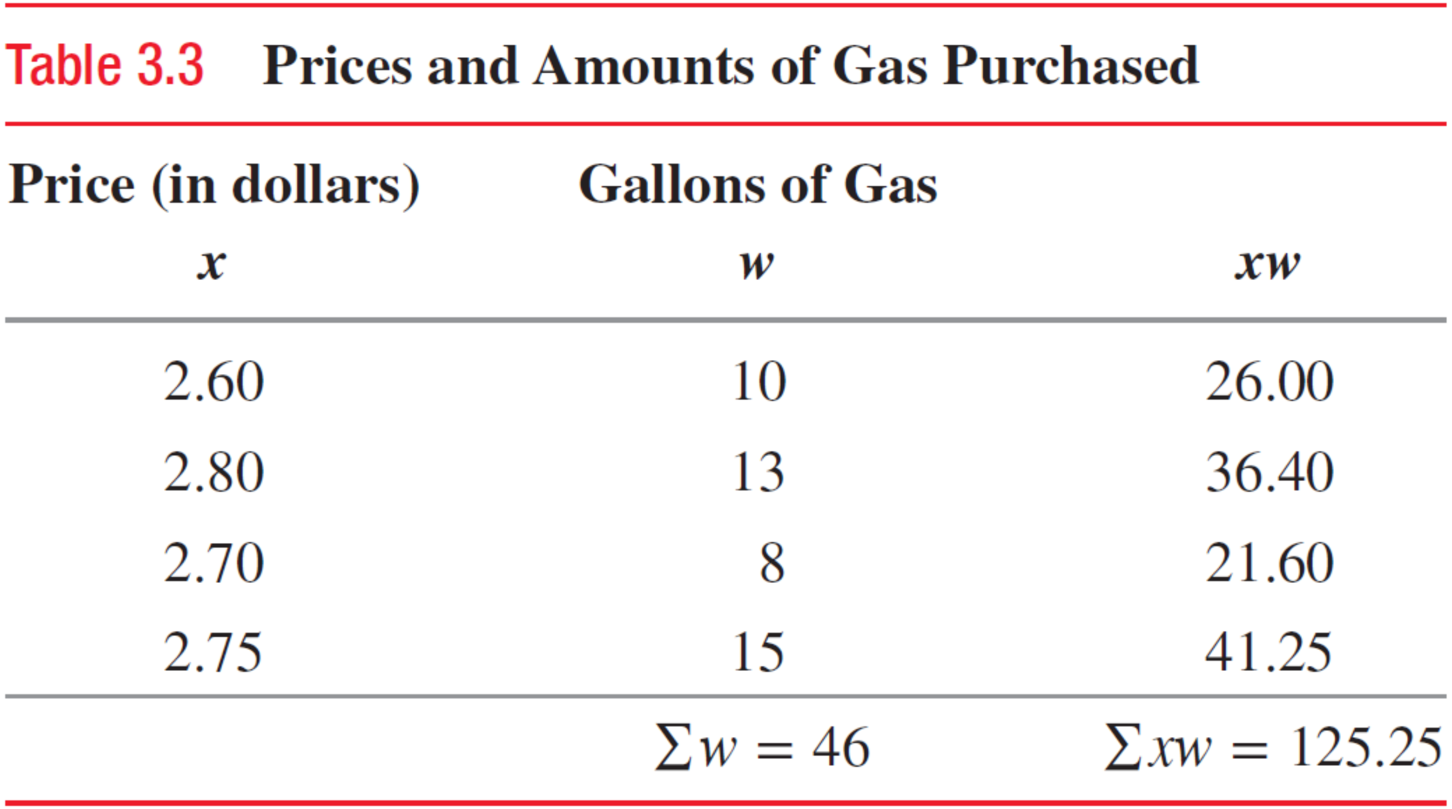


Table 3.3 **Prices and Amounts of Gas Purchased**

Price (in dollars) x	**Gallons of Gas** w	xw
2.60	10	26.00
2.80	13	36.40
2.70	8	21.60
2.75	15	41.25
	$\Sigma w = 46$	$\Sigma xw = 125.25$

➔The average price is

$$\mu = \frac{\sum x_i f_i}{\text{N}} = \frac{125.25}{46} = 2.72$$

Weighted Mean (grouped data): example

How would the solution change if the following table reported the prices and quantities of gas bought by ***a sample*** of 4 drivers of a company in this week?

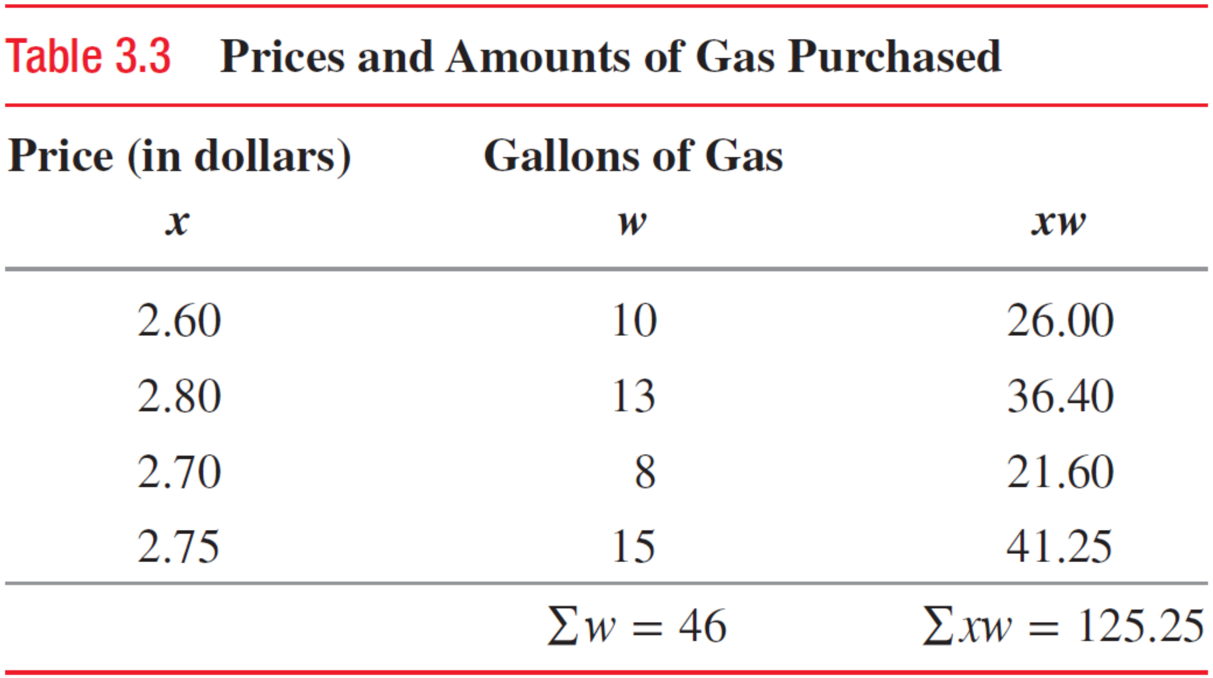


Table 3.3 **Prices and Amounts of Gas Purchased**

Price (in dollars) x	Gallons of Gas w	xw
2.60	10	26.00
2.80	13	36.40
2.70	8	21.60
2.75	15	41.25
	$\Sigma w = 46$	$\Sigma xw = 125.25$

➔The average price is

$$\bar{x} = \frac{\sum x_i f_i}{n} = \frac{125.25}{46} = 2.72$$

Note: your final mark before graduation is… a weighted average!

Weighted Mean (grouped data): relative frequencies

The relative frequencies are:

$$rf_i = \frac{f_i}{n}$$

Using relative frequencies instead of absolute frequencies

POPULATION

$$\mu = \sum x_i rf_i$$

SAMPLE

$$\bar{x} = \sum x_i rf_i$$

Weighted Mean (grouped data): relative frequencies

Ex. Distribution of houselds, by number of components, Italy. Source: Istat, 2001 Census.

nr. of components	f_i	rf_i
1 person	5,427,621	0.2489
2 persons	5,905,411	0.2708
3 persons	4,706,206	0.2158
4 persons	4,136,206	0.1896
5 persons	1,265,826	0.0580
6 persons	369,406	0.0169
Total	21,810,676	1.0000

$$\mu = \sum x_i r f_i = 1\times 0.2489 + 2\times 0.2708 + 3\times 0.2158 + 4\times 0.1896 + 5\times 0.0580 + 6\times 0.0169 = 2.6$$

Weighted Mean: Simpson's paradox

A trend that appears in different groups of data disappears (or even reverses!) when the groups are combined and the data aggregated!

Ex: in 1973, the University of California-Berkeley was sued for sex discrimination: they admitted 44% of male but only 36% of female applicants.

Yet, looking at the data disaggregated by department…the picture looked quite differently!

https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell

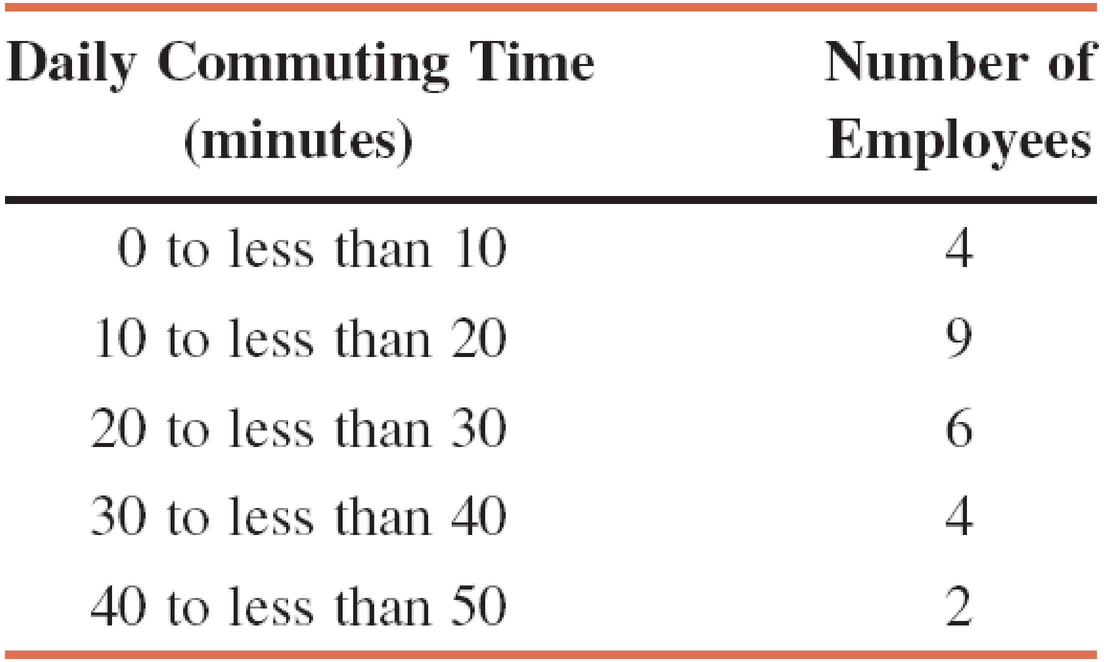
Mean for distributions in classes

Mean computed for data grouped in classes

$$\bar{x} = \frac{\sum m_i f_i}{N}$$

where $m_i$are the midpoints of class i, and f_i the associated frequencies.

Ex: find the average commuting time



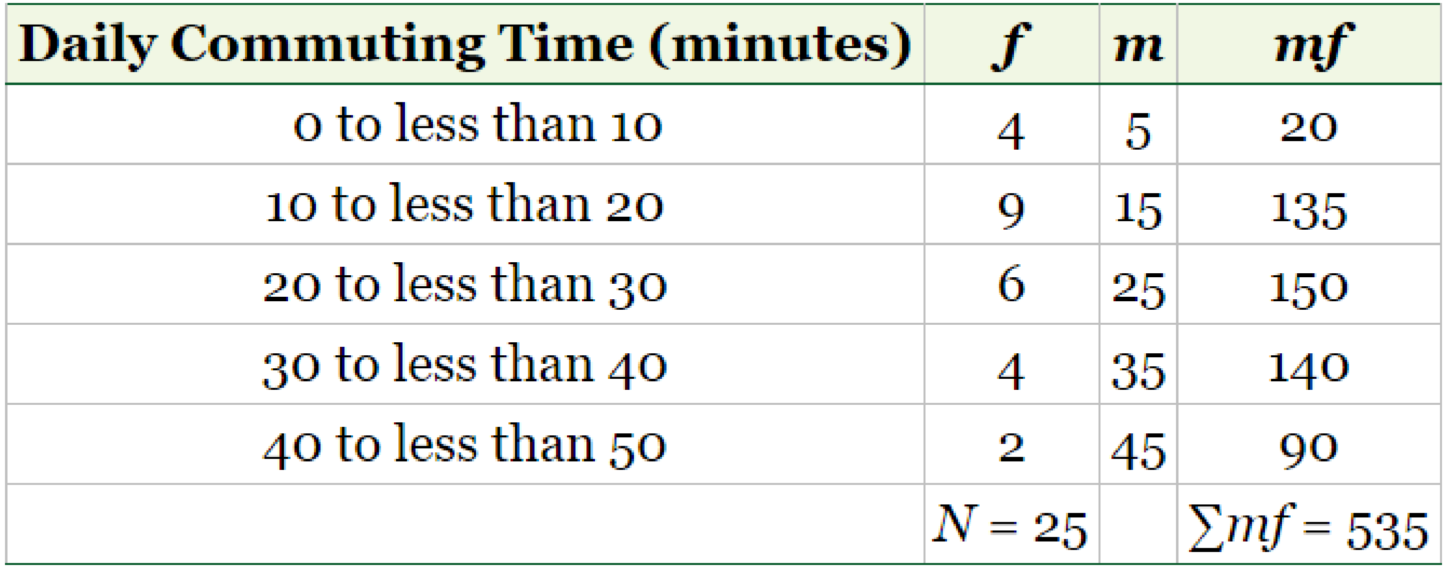
Daily Commuting Time (minutes)	Number of Employees
0 to less than 10	4
10 to less than 20	9
20 to less than 30	6
30 to less than 40	4
40 to less than 50	2

Mean for distributions in classes

Mean computed for data grouped in classes

$$\bar{x} = \frac{\sum m_i f_i}{N}$$

where $m_i$are the midpoints of class i, and f_i the associated frequencies.



Daily Commuting Time (minutes)	*f*	*m*	*mf*
0 to less than 10	4	5	20
10 to less than 20	9	15	135
20 to less than 30	6	25	150
30 to less than 40	4	35	140
40 to less than 50	2	45	90
	$N = 25$		$\sum mf = 535$

Average commuting time is $\bar{x} = \frac{535}{25} = 21.40$ minutes

Properties of the Mean

1. Internality

$$min(x_i) \leq \bar{x} \leq \max(x_i)$$

2. The sum of the observations is the mean times the nr. of obs.

$N\bar{x}$ =$\sum x_i$ (or, for grouped data, $N\bar{x}$ =$\sum x_i f_i$)

This derives directly from the definition of simple (or weighted) mean

3. The sum of the deviations of the x_i from the mean is zero

$\sum(x_i - \bar{x})$=0 (or, for grouped data, $\sum(x_i - \bar{x})\ f_i$=0)

Proof: $\sum x_i - \sum \bar{x} = n\bar{x}$-$n\bar{x} = 0$

Properties of the Mean

4. Linearity

If $y = a + bx$, then $\bar{y} = a + b\bar{x}$.

Proof:

$$\bar{y} = \frac{1}{N}\sum y_i = \frac{1}{N}\sum (a + bx_i) = \frac{1}{N}\sum a + \frac{1}{N}\sum bx_i = \frac{1}{N}Na + b\frac{1}{N}\sum x_i =$$

$$= a + b\bar{x}$$

Properties of the Mean

5. Least squares property

$\bar{x}$ is that value of c that minimizes $\sum (x_i - c)^2$

Proof: $\frac{\partial}{\partial c} \sum (x_i - c)^2 = 2 \sum (x_i - c)(-1)$

Setting this to 0 gets to

$2 \sum (x_i - c)(-1) = 0$ ➔ $\sum (x_i - c) \ = 0$

➔ $\sum x_i - \sum c = \sum x_i - Nc = 0$ ➔ $c = \frac{1}{N} \sum x_i = \bar{x}$

Properties of the Mean: check with an example

Compute the simple (and weighted mean) and check its properties:

5 3 5 3 9 5 0 5 0 3 5 5

Properties of the Mean: check with an example

Compute the simple (and weighted mean) and check its properties:

5 3 5 3 9 5 0 5 0 3 5 5

Simple mean is $\bar{x} = \frac{\sum x}{n} = \frac{48}{12}$ =4.

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{x_i f_i}$
0	2	$0 \times 2 = 0$
3	3	$3 \times 3 = 9$
5	6	30
9	1	9
Total	**12**	**48**

Weighted mean is $\bar{x} = \frac{\sum x_i f_i}{n} = \frac{48}{12} = 4$

Properties of the Mean: check with an example

1. Internality ➔

$$0 \leq 4 \leq 9$$

2. $N\bar{x} = \sum x_i$ ➔

$$N\bar{x} = 12 \times 4 = 48$$

$$\sum x_i = 5+3+5+3+9+5+0+5+0+3+5+5=48$$

Properties of the Mean: check with an example

3. $\sum (x_i - \bar{x})\, f_i = 0$

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{(x - \bar{x})\, f_i}$
0	2	$(0 - 4) \times 2 = -8$
3	3	$(3 - 4) \times 3 = -3$
5	6	$(5 - 4) \times 6 = 6$
9	1	$(9 - 4) \times 1 = 5$
Total	**12**	**0**

Properties of the Mean: check with an example

4. If $y = a + bx$, then $\bar{y} = a + b\bar{x}$.

Suppose $a = 2$ and $b = 1$ ➔ $y = 2 + x$ and hence $\bar{y} = 2 + \bar{x} = 6$

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{y_i = 2 + x}$	$\boldsymbol{y_i f_i}$
0	2	$2 + 0 = 2$	$2 \times 2 = 4$
3	3	$2 + 3 = 5$	$5 \times 3 =15$
5	6	$2 + 5 = 7$	42
9	1	$2 + 9 = 11$	11
Total	**12**		**72/12=6**

Properties of the Mean: check with an example

5. $\bar{x}$ is that value c that minimizes $\sum(x_i - c)^2$

$\boldsymbol{x_i}$	$\boldsymbol{f_i}$	$\boldsymbol{(x-\overline{x})^2 f_i}$	$\boldsymbol{(x-2)^2 f_i}$
0	2	$(0-4)^2 \times 2 = 32$	$(0-2)^2 \times 2 =8$
3	3	3	3
5	6	6	54
9	1	25	49
Total	**12**	**66**	**114**

Combined Mean

$$\bar{x} = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2 + \cdots + \bar{x}_k n_k}{n_1 + n_2 + \cdots + n_k}$$

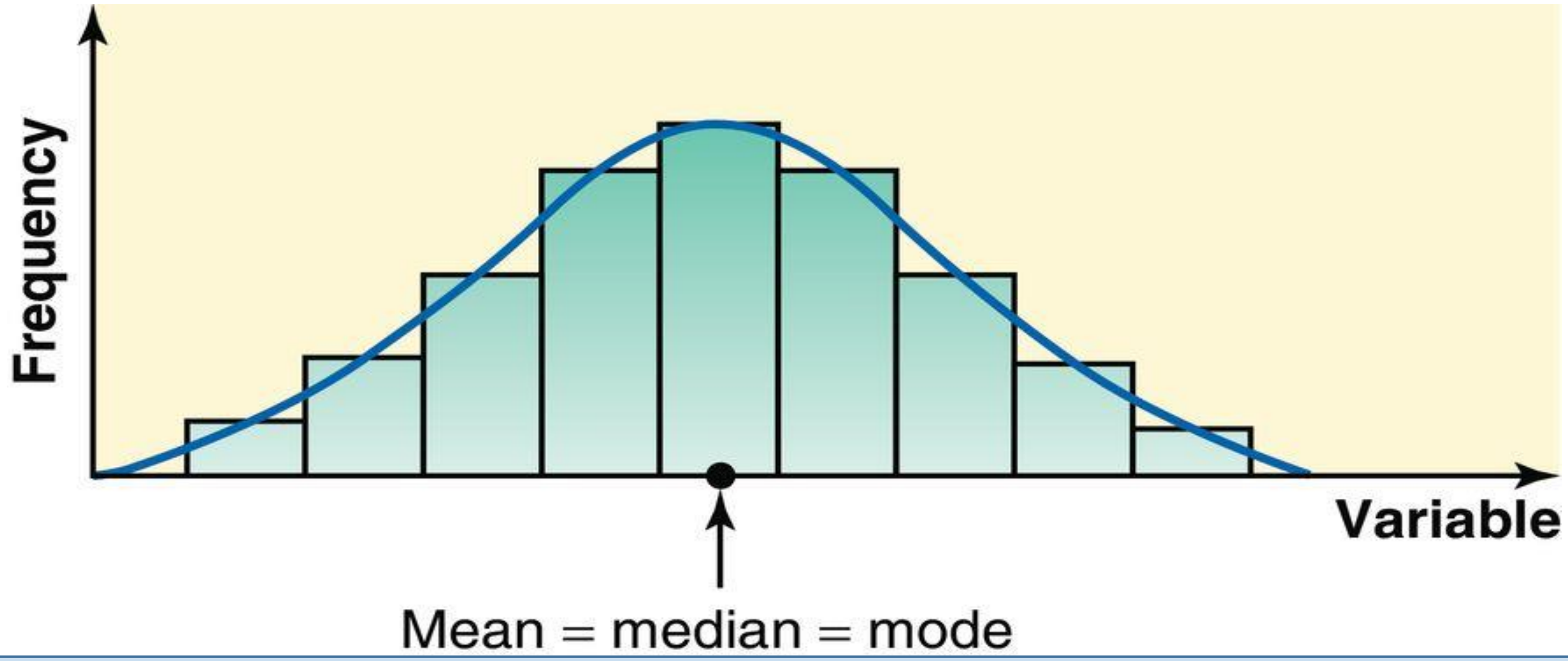
Ex: the average income of 200 people from Rome is 10.000€, while that of 100 living in Milan is of 20.000€. What is the average income considering people from both cities?

$$\bar{x} = \frac{10{,}000 \ \times 200 + 20{,}000 \times 100}{200 + 100} = 11{,}667€$$

Measures of position: comparison and relationship

- No measure is best overall
- A comparison between the mean and the median can give an idea about the shape of the histogram

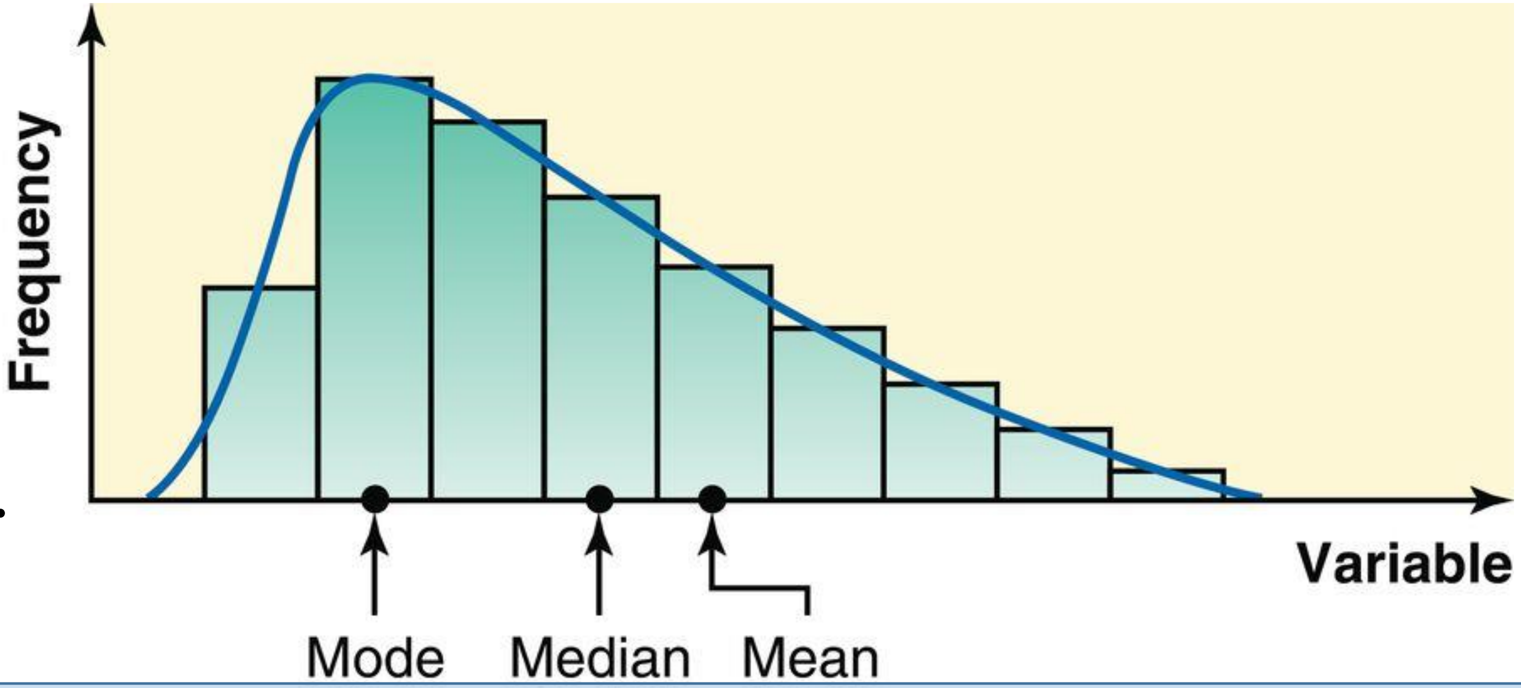
➔ Unimodal distribution with coinciding mode, mean and median is symmetric



Measures of position: comparison and relationship

- No measure is best overall
- A comparison between the mean and the median can give an idea about the shape of the histogram

➔Unimodal distribution with mode < median < mean is **asymmetric to the right**



The outliers

in the right tail

pull the mean to the right.

Measures of position: comparison and relationship

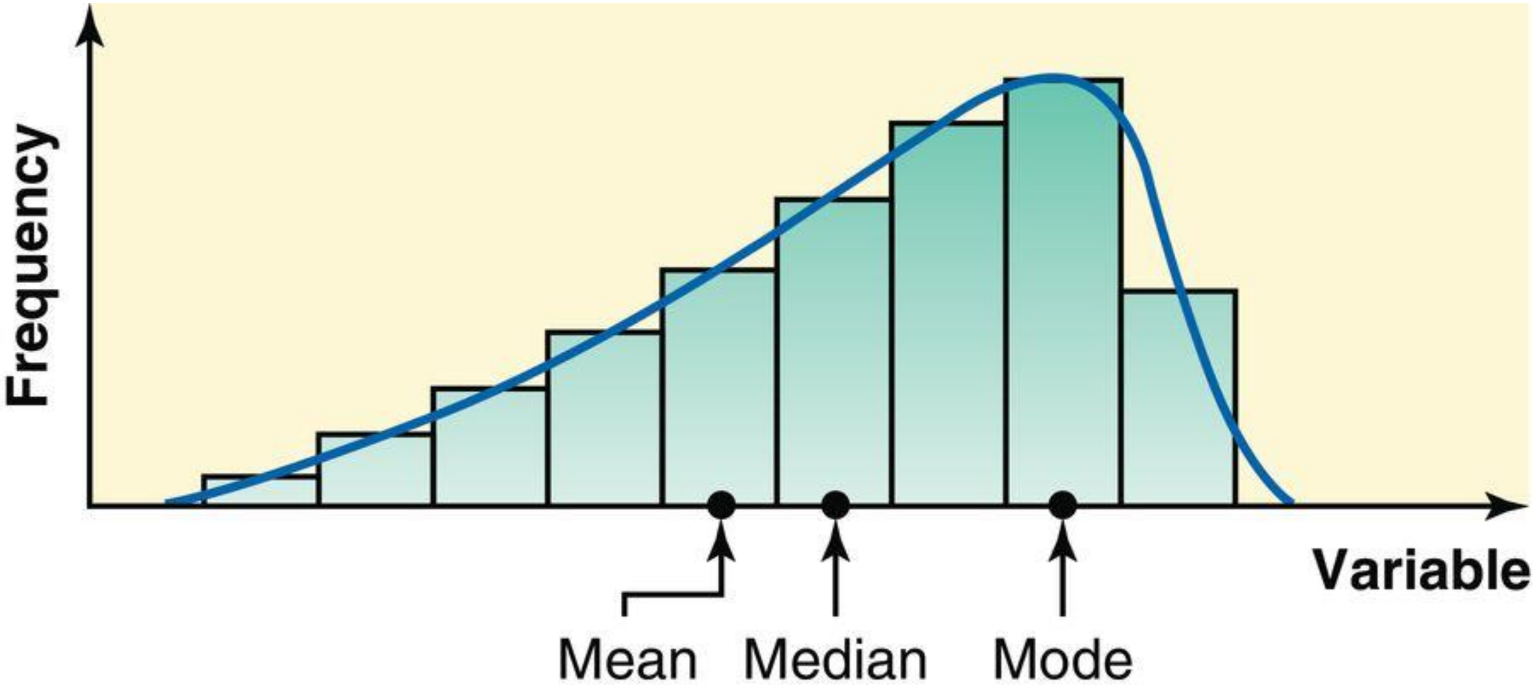
- No measure is best overall
- A comparison between the mean and the median can give an idea about the shape of the histogram

➔Unimodal distribution with mean < median < mode is **asymmetric to the left**

The outliers

in the left tail

pull the mean to the left.



Measures of dispersion

Range

It is obtained by taking the difference between the largest and the smallest values in a data set.

Range=Largest value−Smallest value

Interquartile Range

The difference between the third and the first quartiles

IQR = Q3 − Q1

Variance and Standard Deviation

The variance is the squared deviation of a variable from its mean.

The standard deviation is obtained by taking the positive square root of the variance.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N} \quad \text{and} \quad s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}} \quad \text{and} \quad s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Coefficient of Variation

The coefficient of variation, denoted by CV, expresses standard deviation as a percentage of the mean.

For population data : $\text{CV} = \frac{\sigma}{\mu} \times 100\%$

For sample data : $\text{CV} = \frac{s}{\bar{x}} \times 100\%$

Why do we measure dispersion?

Summarising the dataset can help us understand the data, especially when the dataset is large.

As discussed in the Measures of Central Tendency, the mode, median, and mean summarise the data into a single value that is typical or representative of all the values in the dataset, but this is only part of the 'picture' that summarises a dataset.

Measures of spread summarise the data in a way that shows how scattered the values are and how much they differ from the mean value.

What are measures of dispersion?

The measures of dispersion or spread describe how similar or varied the set of observed values are for a particular variable.

Measures of spread include:

1. **Range**
2. **Interquartile range**
3. **Variance and standard deviation**

The spread of the values can be measured **only** for quantitative data, as the variables are numeric and can be arranged into a logical order with a low end value and a high end value.

Measures of dispersion

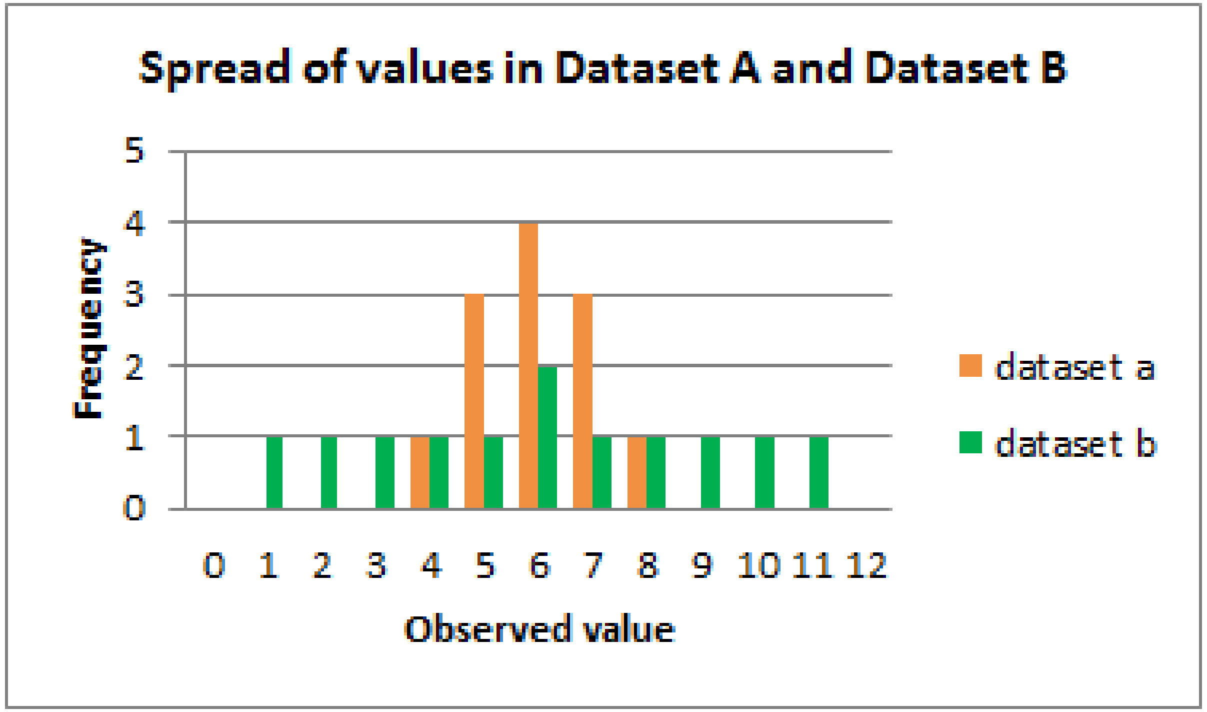
For example: Dataset A: 4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

Dataset B: 1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The mode (most frequent value), median and mean (arithmetic average) of both datasets is 6.

If we just looked at the measures of central tendency, we may assume that the datasets are the same.

However, if we look at the spread of the values in the graph, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data.



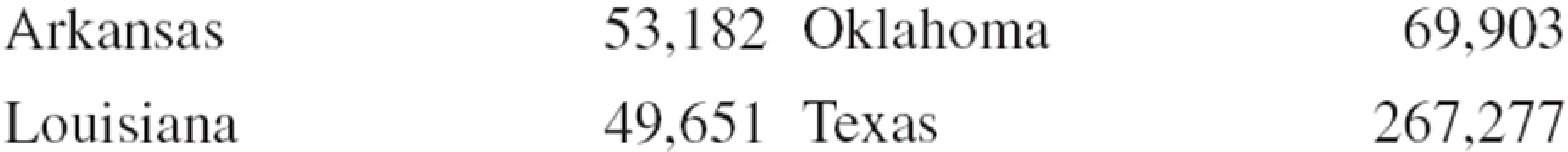
Measures of dispersion: Range

$$\text{Range} = \max(x_i) - \min(x_i)$$

Plus: simplest

Minus: only uses 2 obs out of a whole dataset; sensitive to outliers

Ex: these are the total areas (sq. miles) of 4 States of US:



Arkansas	53,182	Oklahoma	69,903
Louisiana	49,651	Texas	267,277

Range? $\text{Range} = \max(x_i) - \min(x_i) = 267{,}277 - 49{,}651 \ = \ 217{,}626$

Measures of dispersion: Interquartile Range (IQR)

The interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data.

The IQR may also be called the middle 50%.

It is defined as the difference between the 75th and 25th percentiles of the data.

$$\text{IQR} = Q3 - Q1$$

The IQR is an example of a trimmed estimator (less sensitive than the range).

It can be clearly visualized by the box on a Box-Plot (see below).

Measures of dispersion: Variance

It measure how clustered around the mean the values of dataset are : the higher σ^2, the more the values are spread over a relatively larger range around the mean.

POPULATION	SAMPLE
$\sigma^2 = \frac{1}{N}\sum_{i}^{N}(x_i - \mu)^2 \ \ or$ $\sigma^2 = \frac{\sum {x_i}^2}{\mathrm{N}} - \mu^2$	$s^2 = \frac{1}{n-1}\sum_{i}^{N}(x_i - \bar{x})^2 \ \ or$ $s^2 = \frac{\sum {x_i}^2}{\mathrm{n}-1} - \left(\frac{n-1}{\mathrm{n}}\right)\bar{x}^2$

VARIANCE CANNOT BE NEGATIVE!

Measures of dispersion: Variance

Pros: comprehensive measure

Problem1: its measurement unit is the **square** of measurement unit of the data

Problem2 : cannot be used to compare phenomena with different scales

Variance: Example Population

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \quad \text{or} \quad \sigma^2 = \frac{1}{N}\sum_i {x_i}^2 - \mu^2$$

Data 1	Data 2	Data 3
3	6	3
-3	-8	3
3	3	3
5	2	3
-3	-10	3
6	18	3
7	12	3
6	5	3

Variance Data 1?

Variance Data 2?

Variance Data 3?

Variance: Example Population (Data 1)

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$ or $$\sigma^2 = \frac{1}{N}\sum_i x_i^{\,2} - \mu^2$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$	$x_i^{\,2}$
3	0	0	9
-3	-6	36	9
3	0	0	9
5	2	4	25
-3	-6	36	9
6	3	9	36
7	4	16	49
6	3	9	36
24	**0**	**110**	**182**

Variance Data 1?

$$\mu = \frac{24}{8} = 3$$

$$\sigma^2 = \frac{110}{8} = 13.75$$ or

$$\sigma^2 = \frac{1}{8}182 \; - \; 3^2 = 22.75 - 9$$
$$= 13.75$$

Variance: Example Population (Data 2)

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$ or $$\sigma^2 = \frac{1}{N}\sum_i {x_i}^2 - \mu^2$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$	${x_i}^2$
6	2.5	6.25	36
-8	-11.5	132.25	64
3	-0.5	0.25	9
2	-1.5	2.25	4
-10	-13.5	182.25	100
18	14.5	210.25	324
12	8.5	72.25	144
5	1.5	2.25	25
28	**0**	**608**	**706**

Variance Data 1?

$$\mu = \frac{28}{8} = 3.5$$

$$\sigma^2 = \frac{608}{8} = 76 \text{ or}$$

$$\sigma^2 = \frac{1}{8}706 - 3.5^2 = 76$$

Variance: Example Population (Data 3)

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \quad \text{or} \quad \sigma^2 = \frac{1}{N}\sum_i {x_i}^2 - \mu^2$$

x_i	$x_i - \mu$	$(x_i - \mu)^2$	${x_i}^2$
3	0	0	9
3	0	0	9
3	0	0	9
3	0	0	9
3	0	0	9
3	0	0	9
3	0	0	9
3	0	0	9
24	**0**	**0**	**72**

Variance Data 1?

$$\mu = \frac{24}{8} = 3$$

$$\sigma^2 = \frac{0}{8} = 0 \quad \text{or}$$

$$\sigma^2 = \frac{1}{8}72 - 3^2 = 0$$

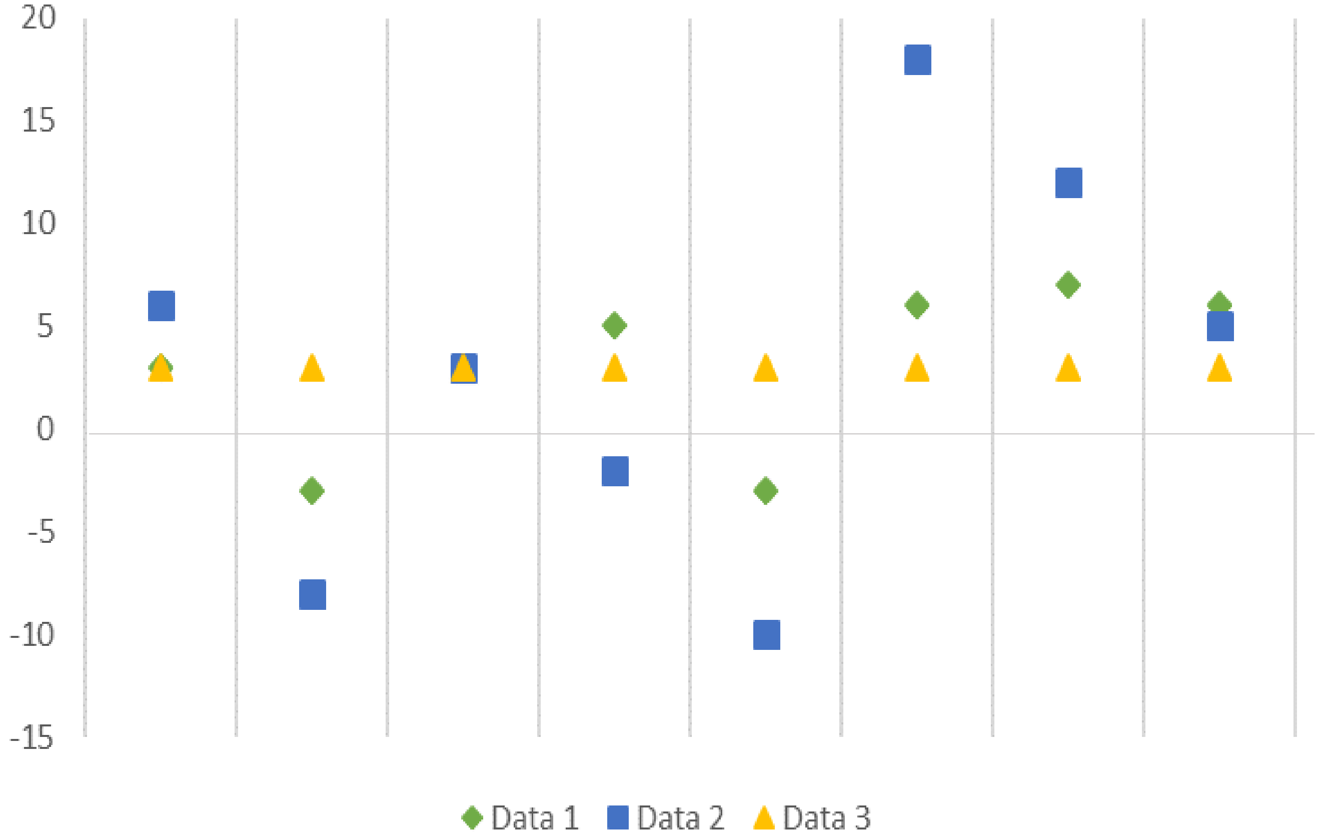
Variance: Example Population

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2 \quad \text{or} \quad \sigma^2 = \frac{1}{N}\sum_i x_i{}^2 - \mu^2$$

$\boldsymbol{Data\ 1}$	$\boldsymbol{x_i - \mu}$	$\boldsymbol{(x_i - \mu)^2}$	$\boldsymbol{x_i{}^2}$	$\boldsymbol{Data\ 2}$	$\boldsymbol{x_i - \mu}$	$\boldsymbol{(x_i - \mu)^2}$	$\boldsymbol{x_i{}^2}$	$\boldsymbol{Data\ 3}$	$\boldsymbol{x_i - \mu}$	$\boldsymbol{(x_i - \mu)^2}$	$\boldsymbol{x_i{}^2}$
3	0	0	9	**6**	2.5	6.25	36	**3**	0	0	9
-3	-6	36	9	**-8**	-11.5	132.25	64	**3**	0	0	9
3	0	0	9	**3**	-0.5	0.25	9	**3**	0	0	9
5	2	4	25	**2**	-1.5	2.25	4	**3**	0	0	9
-3	-6	36	9	**-10**	-13.5	182.25	100	**3**	0	0	9
6	3	9	36	**18**	14.5	210.25	324	**3**	0	0	9
7	4	16	49	**12**	8.5	72.25	144	**3**	0	0	9
6	3	9	36	**5**	1.5	2.25	25	**3**	0	0	9

Variance: Example Population

Data 1	Data 2	Data 3
3	6	3
-3	-8	3
3	3	3
5	2	3
-3	-10	3
6	18	3
7	12	3
6	5	3



Data 1: σ^2=13.75

Data 2: σ^2=76

Data 3: σ^2= 0

Variance: Example Sample (Data 1)

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \quad \text{or} \quad s^2 = \frac{\sum x_i{}^2}{\text{n}-1} - \left(\frac{n-1}{\text{n}}\right)\bar{x}^2$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$x_i{}^2$
3	0	0	9
-3	-6	36	9
3	0	0	9
5	2	4	25
-3	-6	36	9
6	3	9	36
7	4	16	49
6	3	9	36
24	**0**	**110**	**182**

➔ $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$

$= \frac{1}{8-1} 110 = 15.714$

or

➔ $s^2 = \frac{\sum x_i{}^2}{n-1} - \left(\frac{n-1}{n}\right)\bar{x}^2 =$

$\frac{182}{8-1} - \left(\frac{8-1}{8}\right)3^2 = 15.714$

Measures of dispersion: Standard Deviation

POPULATION	SAMPLE
$\sigma \quad = \sqrt{\sigma^2}$	$s \quad = \sqrt{s^2}$

It solves Problem1 of Variance

STANDARD DEVIATION CANNOT BE NEGATIVE!

Plus: comprehensive measure, **same** measurement unit of the data

Problem2 : cannot be used to compare phenomena with different scales

Standard Deviation: example Population

$$\sigma = \sqrt{\sigma^2}$$

Data 1	Data 2	Data 3
3	6	3
-3	-8	3
3	3	3
5	2	3
-3	-10	3
6	18	3
7	12	3
6	5	3

Standard Deviation Data 1?

Standard Deviation Data 2?

Standard Deviation Data 3?

Data 1: σ^2= 13.75 → $\sigma = 3.708$

Data 2: σ^2= 76 → $\sigma = 8.72$

Data 3: σ^2= 0 → $\sigma = 0$

Standard Deviation: example Sample

$$s = \sqrt{s^2}$$

Data 1	Data 2	Data 3
3	6	3
-3	-8	3
3	3	3
5	2	3
-3	-10	3
6	18	3
7	12	3
6	5	3

Standard Deviation Data 1?

Standard Deviation Data 2?

Standard Deviation Data 3?

Data 1: $s^2 = 15.714 \rightarrow s = 3.964$

Data 2: $s^2 = 86.857 \rightarrow s = 9.320$

Data 3: $s^2 = 0 \quad \rightarrow s = 0$

Measures of dispersion: Coefficient of Variation

To solve Problem1 (*measurement unit of the data*) and Problem2 (*compare phenomena with different scales*) of Variance (and Standard Deviation), we can use

POPULATION	SAMPLE
$\text{CV} = \frac{\sigma}{\mu}$	$\text{CV} = \frac{s}{\bar{x}}$

It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean

Note: it can be computed only for variables taking positive values

Coefficient of Variation: Example Population

Ex: this table reports the income and years of experience of all 6 workers in a SMB

	Monthly net income	Years of experience
Raul	1950	3
Luke	2600	7
Sally	1150	0
Marleen	3600	10
Bruce	2800	4
Kim	2300	6

➔

	Monthly net income	Years of experience
Mean	2400.000	5.000
Variance	569166.667	10.000
St. Dev.	754.431	3.162
CV	31.43%	63.25%

$$\mathrm{CV} = \frac{\sigma}{\mu} \qquad \mathrm{CV} = \frac{\sigma}{\mu}$$

Measures of dispersion: formulas for distributions

In case of absolute frequency distributions use

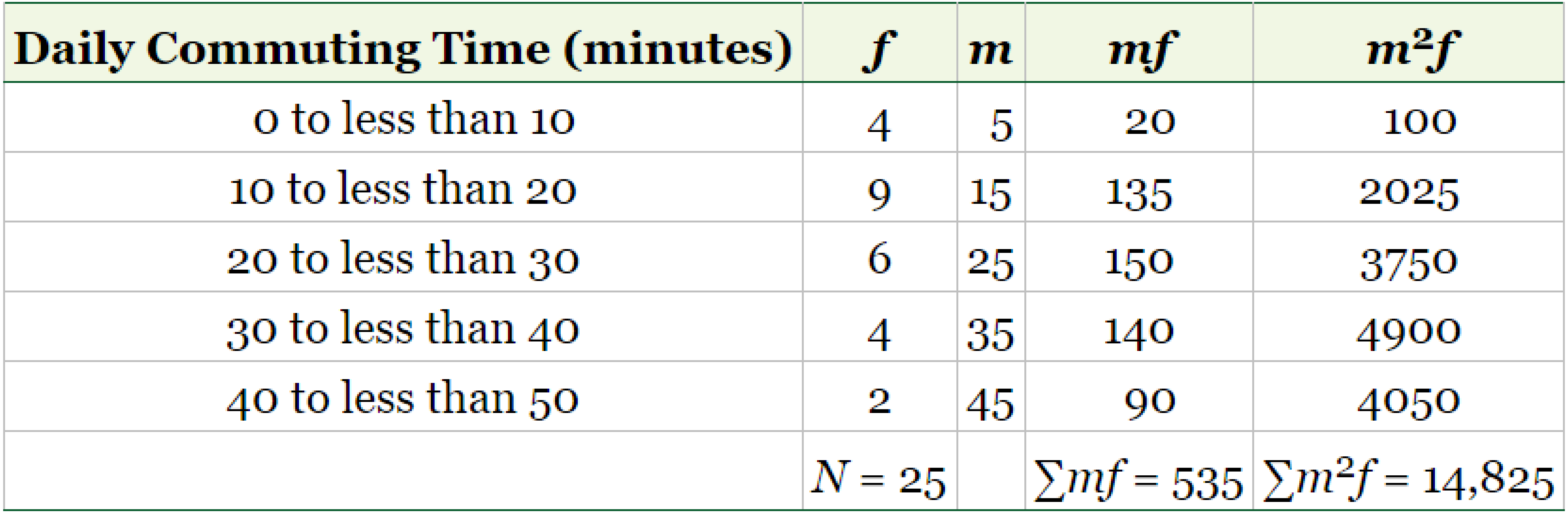
$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \bar{x})^2 f_i = \frac{1}{N}\sum_i {x_i}^2 f_i - \mu^2$$

In case of distributions in classes use

$$\sigma^2 = \frac{1}{N}\sum_i (m_i - \bar{x})^2 f_i = \frac{1}{N}\sum_i {m_i}^2 f_i - \mu^2$$

σ and CV are computed accordingly

Measures of dispersion: formulas for distributions



Daily Commuting Time (minutes)	f	m	mf	m^2f
0 to less than 10	4	5	20	100
10 to less than 20	9	15	135	2025
20 to less than 30	6	25	150	3750
30 to less than 40	4	35	140	4900
40 to less than 50	2	45	90	4050
	$N = 25$		$\sum mf = 535$	$\sum m^2f = 14{,}825$

➔ $\sigma^2 = \frac{14825}{25} - 21.4^2 = 135.04$ and $\text{CV} = 11.62$

Chebyshev’s Theorem

Provides the MINIMUM frequency with which a variable takes values within an interval around its mean

$$f(|x-\mu| \leq k\sigma) \geq 1-\frac{1}{k^2}$$

where: μ = mean, σ = standard deviation, and k = positive number

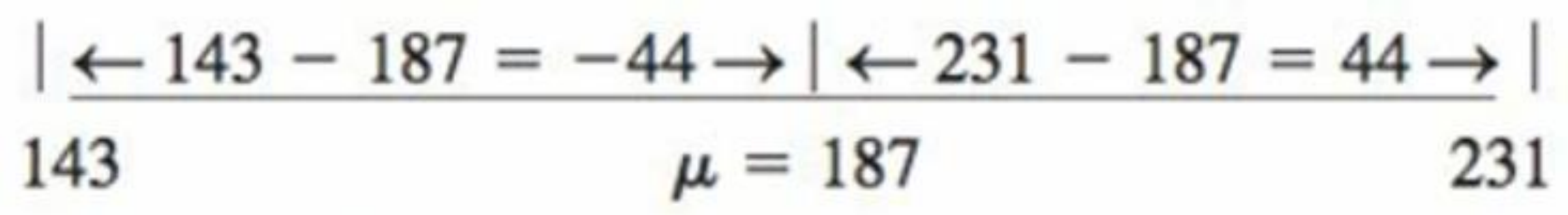
Plus: it only requires the knowledge of mean and variance

Minus: provides a lower bound, not the ACTUAL frequency

Chebyshev’s Theorem: example

The average systolic blood pressure for women is 187 mmHg with a standard deviation of 22. Using Chebyshev’s theorem, find at least what percentage of women have a systolic blood pressure between 143 and 231 mmHg.

We know that $\mu = 187$ and $\sigma = 22$. How much is k?



$|\leftarrow 143 - 187 = -44 \rightarrow | \leftarrow 231 - 187 = 44 \rightarrow |$

143 $\mu = 187$ 231

$$k = \frac{44}{22} = 2 \Rightarrow f(|x - 187| \leq 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

Chebyshev’s Theorem: another example

	Monthly net income	Years of experience
Raul	1950	3
Luke	2600	7
Sally	1150	0
Marleen	3600	10
Bruce	2800	4
Kim	2300	6

	Monthly net income	Years of experience
Mean	2400.000	5.000
Variance	569166.667	10.000
St. Dev.	754.431	3.162
CV	31.43%	63.25%

What is the minimum frequency of workers having an experience within at most 1.5 standard deviations around the average experience?

$$f(|x-\bar{x}| \leq 1.5\sigma) \geq 1 - \frac{1}{1.5^2} = 0.56$$

The actual frequency is 2/3

Box-Whiskers Plot (or Box-Plot)

Comprehensive graphical representation of a distribution.

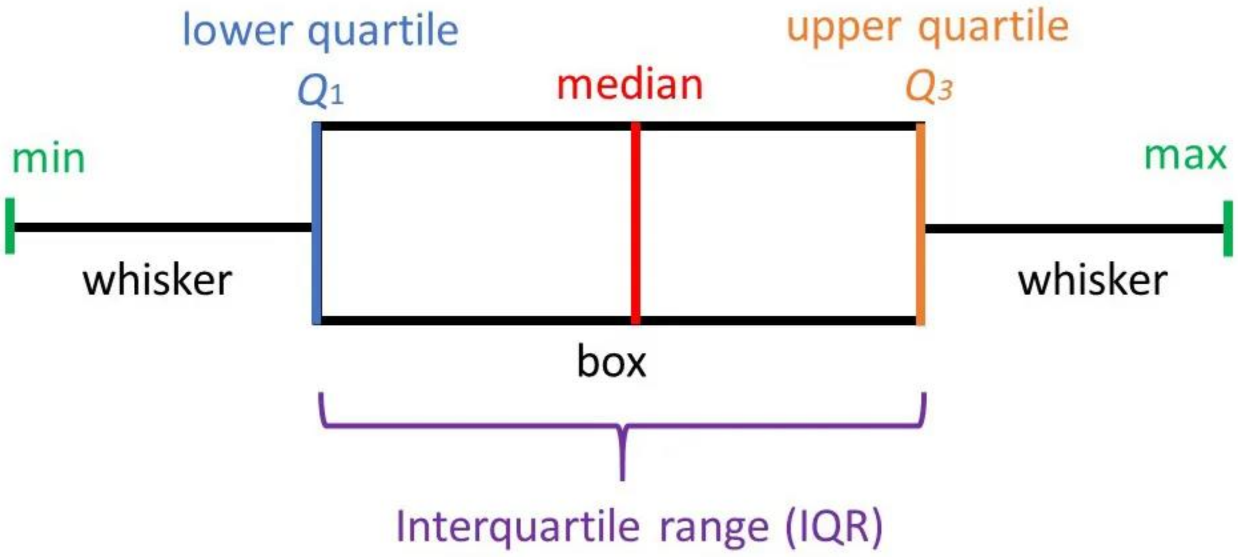
Indeed, it provides info on:

-position: median, Q1, and Q3 (lines)

-dispersion: interquartile range (box)

-shape of the distribution (whiskers)

-extreme values: outliers



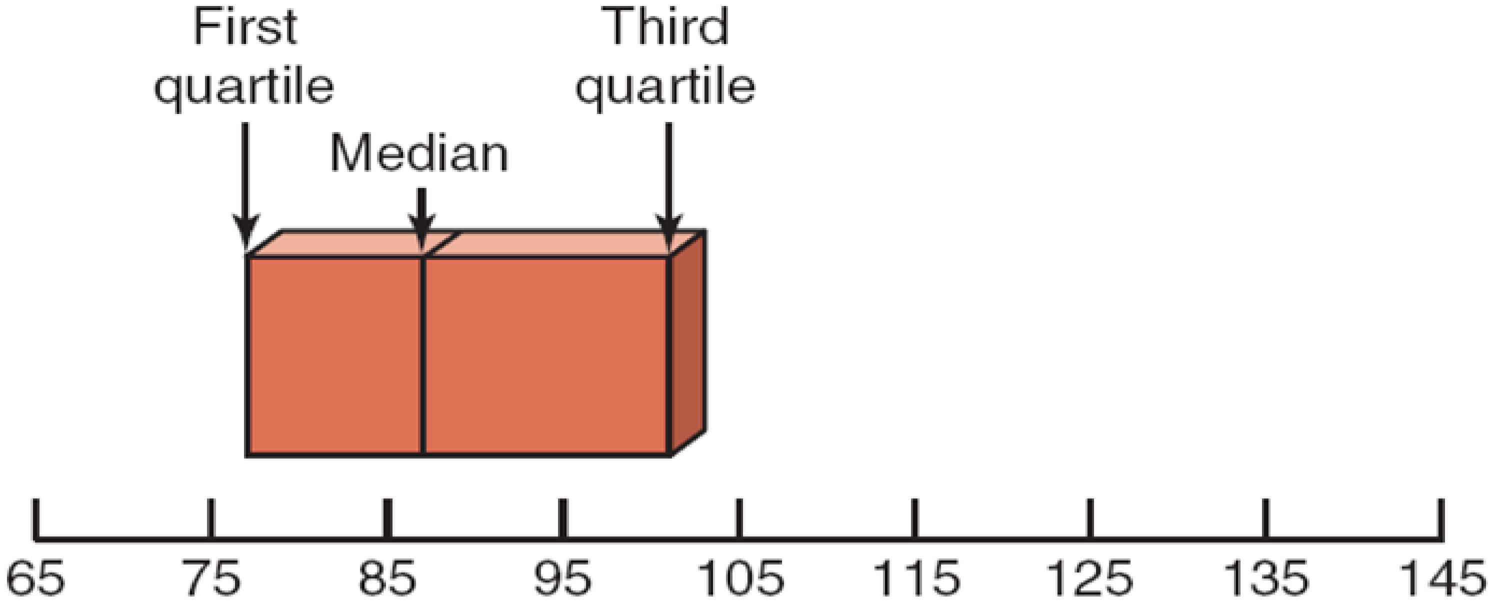
Box-Plot: how to construct it

Draw the Box-Plot for the following yearly incomes (thousand dollars) of 12 households:

75 69 84 112 74 104 81 90 94 144 79 98

Step 1: the lines ➔ Q1 = 77, Median = 87, Q3 = 101

Step 2: the box ➔ IQ= 24



Box-Plot: how to construct it

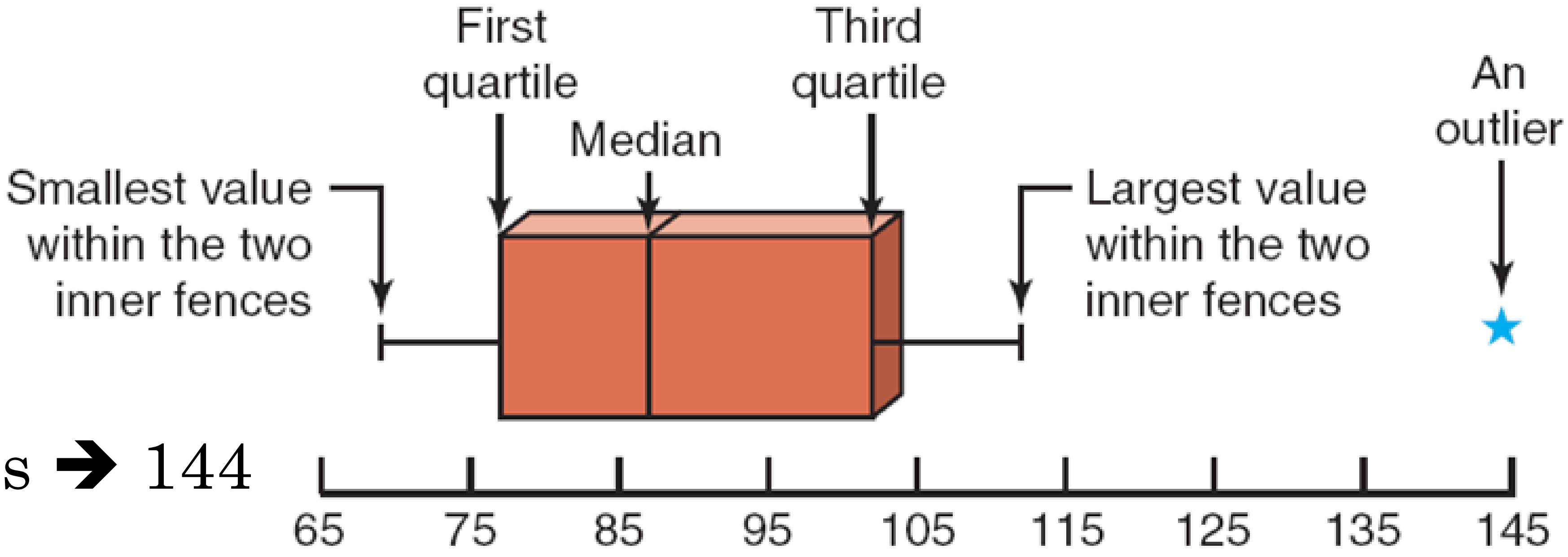
Step 3: whiskers ➔ *closest* values within lower and upper inner fences

Lower = Q1 – 1.5 x IQ = 41 ➔ 1st obs within 41 is 69 (no lower outliers)

Upper = Q3 +1.5 x IQ = 137 ➔ 1st obs within 137 is 112

Step 4: the outliers

Only observation outside inner fences is ➔ 144


First quartile
Third quartile
Median
An outlier
Smallest value within the two inner fences
Largest value within the two inner fences
65
75
85
95
105
115
125
135
145

Box-Plot: another example

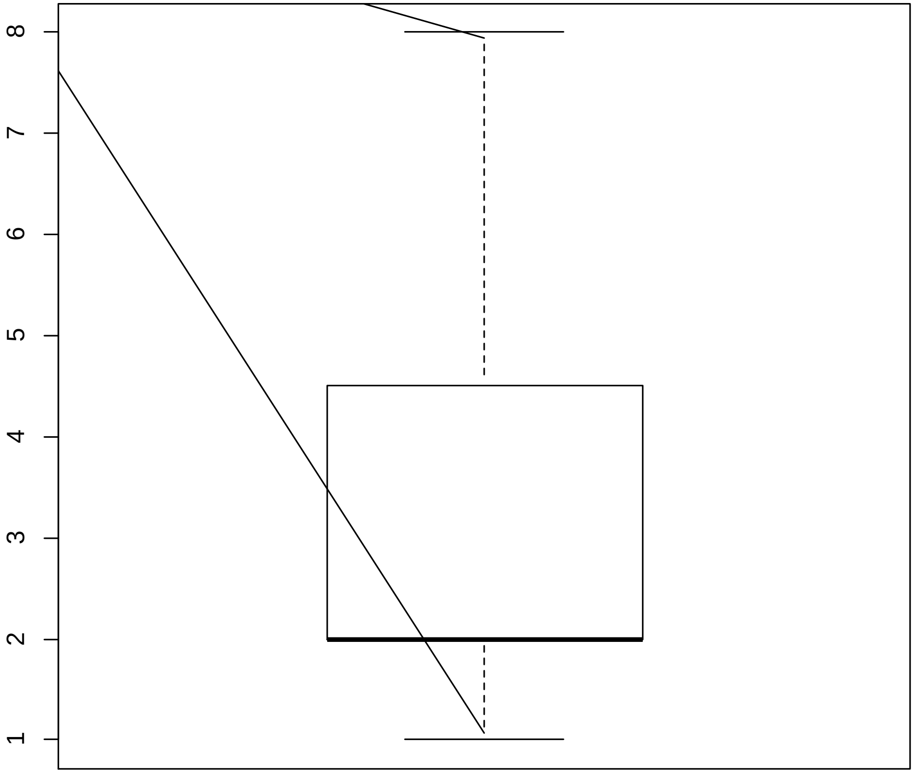
Draw the Box-Plot for the following distribution:

2, 8, 2, 8, 4, 2, 2, 1, 2, 5,2, 1

Box-Plot: another example

Draw the Box-Plot for the following distribution:

2, 8, 2, 8, 4, 2, 2, 1, 2, 5, 2, 1



Q1 = 2, Median = 2, Q3 = 4.5

Lower inner fence = 2 – 1.5 x 2.5= – 1.75

Upper inner fence = 4.5 + 1.5 x 2.5= 8.25

Box-Plot: another example

The Box-Plot for the following distribution is:

8, 2, 2, 5, 1, 19, 8, 2, 2