Measures of location

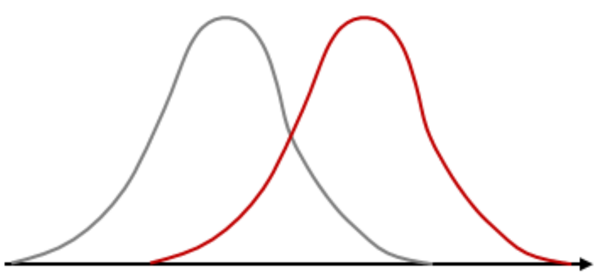
Marcin Chlebus, Ewa Cukrowska-Torzewska
Faculty of Economic Sciences
University of Warsaw

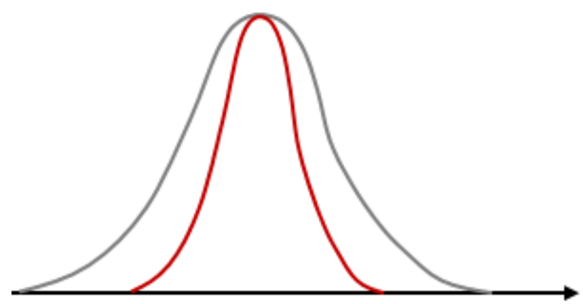
Lecture 2: 10-11.10.2017

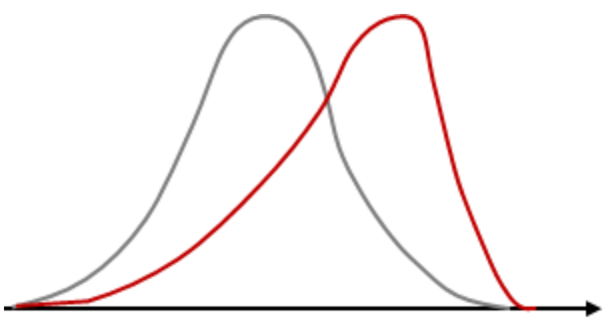
Descriptive statistics

Descriptive statistics

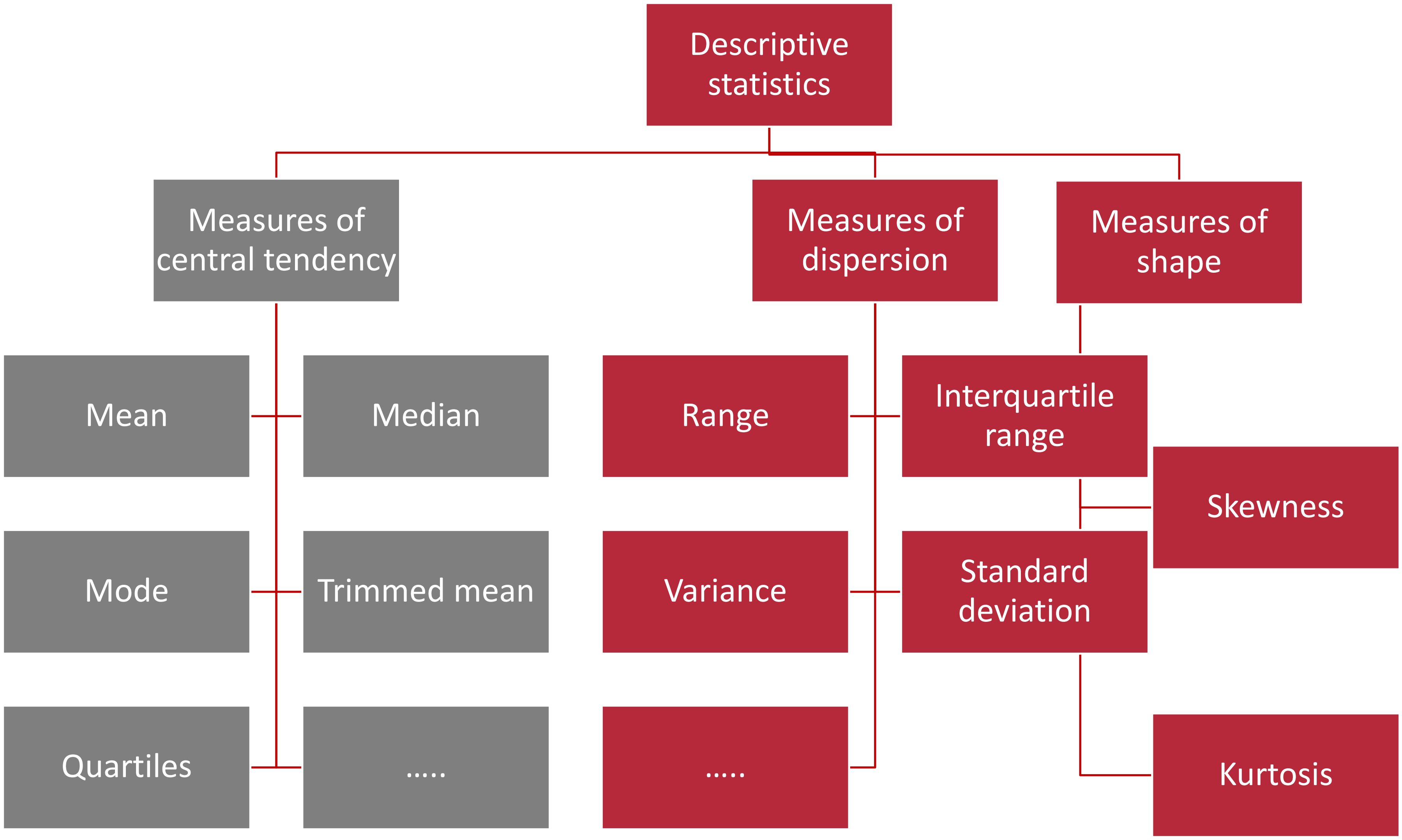
- Measures of location (central tendency/position)
- Measures of dispersion
- Measures of shape







Descriptive statistics

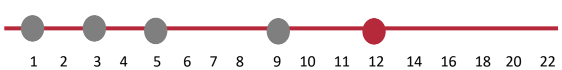


Arithmetic mean

- The **arithmetic mean** is one of the most intuitive measures of central tendency.
- It is often simply referred to as „the mean” or „the average”

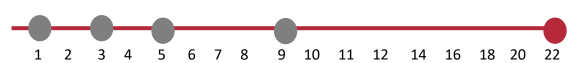
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- **The measure is sensitive to extreme values (outliers)!**



Sample 1: 1,3,5,9,12

$$\bar{x} = \frac{(1+3+5+9+12)}{5} = 6$$



Sample 2: 1,3,5,9,22

$$\bar{x} = \frac{(1+3+5+9+22)}{5} = 8$$

Properties of the arithmetic mean

- The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^{n}(x_i-\bar{x})=\sum_{i=1}^{n}x_i-n\bar{x}=n\bar{x}-n\bar{x}=0$$

- For linear transformation of the form $y_i=a+bx_i$, where *a* and *b* are known constants, it holds that:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}(a+bx_i)=\frac{1}{n}\sum_{i=1}^{n}a+\frac{b}{n}\sum_{i=1}^{n}x_i=a+b\bar{x}$$

- **Caution! The mean is not equal to the mean of the means:**

e.g.:

$$\frac{1+2+4+5+8+10+12}{7}=6$$

$$\frac{1}{3}\left(\frac{1+2+4}{3}+\frac{5+8}{2}+\frac{10+12}{2}\right)=\frac{1}{3}(2.3+6.5+11)=6.61$$

Mean for grouped data (weighted mean)

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j$$

Absolute frequencies

Relative frequencies

The mid-value of the *j* th class interval (group)

Example:

Age	<20	(20-35]	(35-50]	(50-100]
Absolute frequencies	20	25	40	55
Relative frequencies	20/140	25/140	40/140	55/140

$$\bar{x} = \frac{1}{140}(20*10 + 25*27.5 + 40*42.5 + 55*75) =$$

$$\frac{20}{140}*10 + \frac{25}{140}*27.5 + \frac{40}{140}*42.5 + \frac{55}{140}*75 = 47.95$$

Weighted mean

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

Weights

Example:

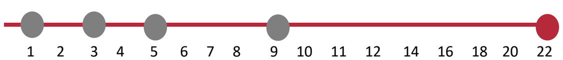
Goods in consumer's basket	Number of goods	Price of a good	Weights
Good A	200	25	0.2
Good B	300	15	0.3
Good C	500	10	0.5

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} = \frac{0.2*25+0.3*15+0.5*10}{0.2+0.3+0.5} = 14.5$$

Trimmed mean (truncated mean)

- It is the arithmetic mean value computed with a specified percentage of values removed from each tail to eliminate the highest and lowest outliers and extreme values.
- For small samples a specific number of observations that represent extreme cases (e.g. 1) rather than a percentage, is simply dropped when calculating the mean.

- In our example:



Sample 2: 1,3,5,9,22

$$\bar{x} = \frac{(1+3+5+9+22)}{5} = 8$$

We would drop the extreme value of 22 to get the trimmed mean equal to:

$$\bar{x} = \frac{(1+3+5+9)}{4} = 4.5$$

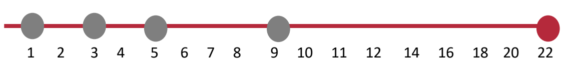
We could also calculate the trimmed 20% mean to get:

$$\bar{x} = \frac{(3+5+9)}{3} = 5.7$$

Winsorized mean

- The Winsorized mean (named after the biostatistician C P Winsor) is similar to the trimmed mean, but instead of dropping extreme values they are simply replaced with the most extreme remaining values.

- In our example:



Sample 2: 1,3,5,9,22

$$\bar{x} = \frac{(1+3+5+9+22)}{5} = 8$$

We would drop the extreme value of 22 and replace it with 9 to get the winsorized mean equal to get:

$$\bar{x} = \frac{(1+3+5+9+9)}{5} = 5.4$$

We could also calculate the 20% winsorized mean to get:

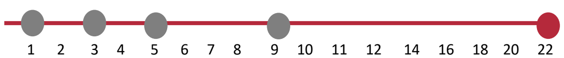
$$\bar{x} = \frac{(3+3+5+9+9)}{5} = 5.8$$

Harmonic mean

$$\bar{x}_H = \frac{n}{\sum_{1=1}^{n} \frac{1}{x_i}}$$

- It is used when there are few very large/small values
- It is often applied to averaging rates of speed.
- Intuition: 5 machines – we produce 1,3,5,9,22 notebooks per hour, we produce the same amount of notebooks for every machine (i.e. 1). We spent 1,68 hours to have 5. If we have 1 machines producing 2,96 notebooks per hour, we would have 5 notebooks in this time.

In our example:



$$\bar{x} = \frac{(1+3+5+9+22)}{5} = 8$$

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{5}{\frac{1}{1}+\frac{1}{3}+\frac{1}{5}+\frac{1}{9}+\frac{1}{22}} = 2.96$$

Or for a series of 8 speeds of: 20,40,60,100,120,180,200,10000

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{8}{\frac{1}{20}+\frac{1}{40}+\frac{1}{60}+\frac{1}{100}+\frac{1}{120}+\frac{1}{180}+\frac{1}{200}+\frac{1}{10000}} = 66.3$$

Geometric mean

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

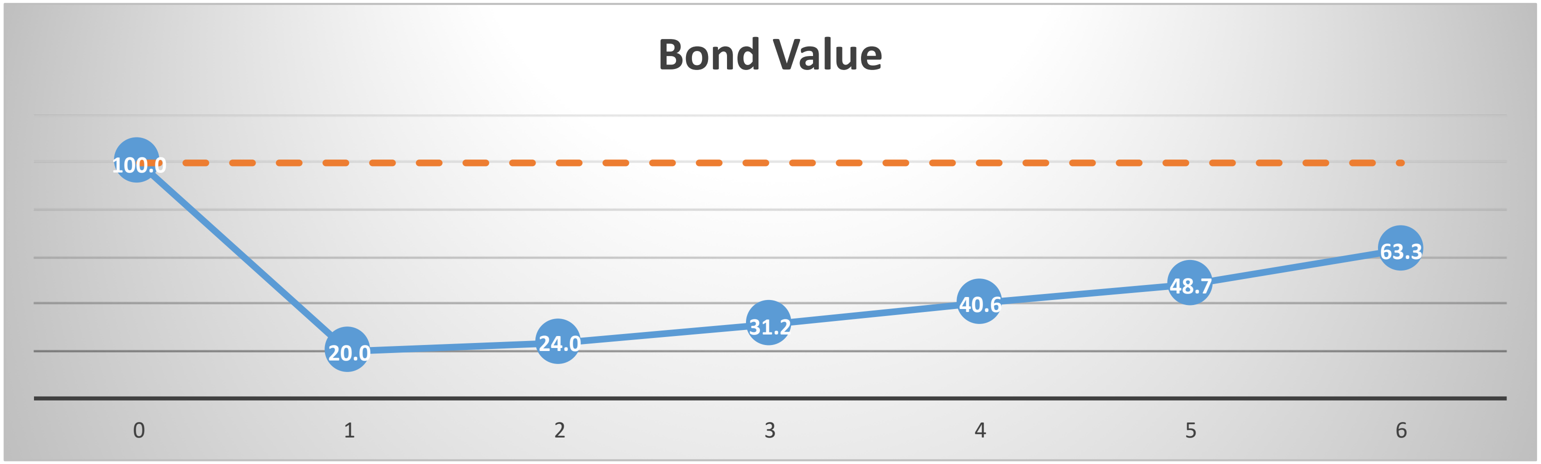
- It is used when the variable is log-normally distributed
- It is often applied to calculating average rate of change

An example:

The annual rates of return from a bond A are: -0.80, 0.20, 0.30, 0.30, 0.20, 0.30
The respective rates of change are: 0.20, 1.20, 1.30, 1.30, 1.20, 1.30

$$\bar{x} = \frac{(0.20+1.20+1.30+1.30+1.20+1.30)}{6} = 1.0833$$

$$\bar{x}_G = \sqrt[6]{0.20*1.20*1.30*1.30*1.20*1.30} = 0.926$$



Means

Exercise 1:

Use data on Apple stocks (Apple.csv) and calculate rate of change for each year. Then calculate:

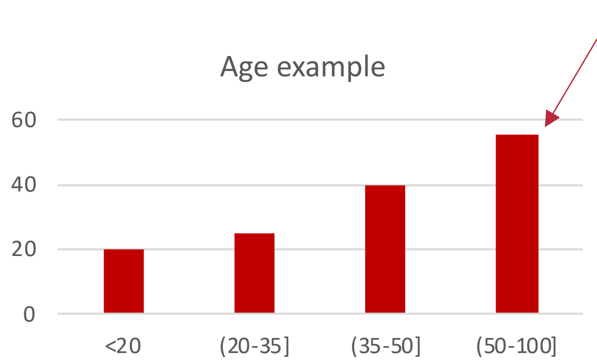
- Airthemitic mean
- Geometric mean
- Weighted mean
- Trimmed mean
- Winsorized mean

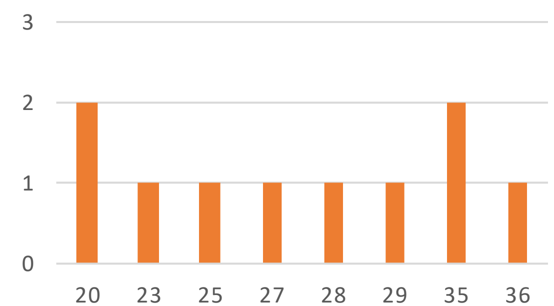
Mode

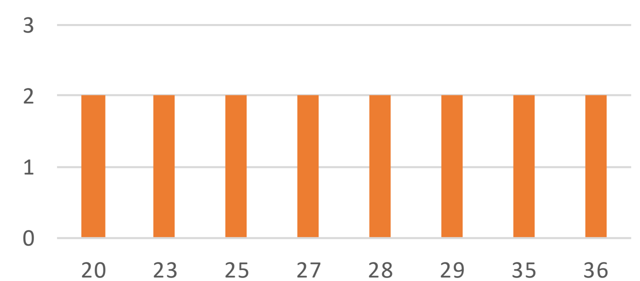
- The mode is the value that occurs most frequently
- **It is not influenced by outliers**
- It can be applied to quantitaive and qualitative variables

- Sometimes there is more than one mode
- Sometimes the mode does not exist

Mode

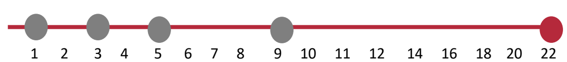






Mid-range

- It is the arithmetic mean of the maximum and minimum values in a dataset
- **It is highly sensitive to outliers** (it only takes into account the two most extreme values from a sample).
- In our example:



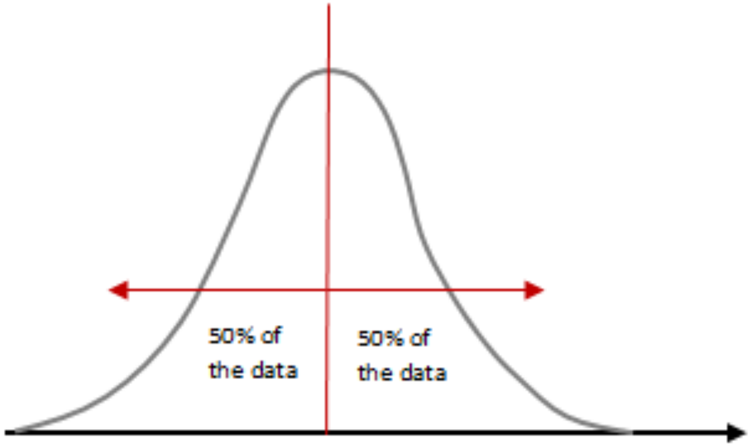
Sample 2: 1,3,5,9,22

$$\bar{x} = \frac{(1+3+5+9+22)}{5} = 8$$

The mid-range is $\frac{(1+22)}{2} = 11.5$

Median

- Median is the middle value; we will denote it as $\tilde{x}_{0.5}$
- It is the value which divides the observations into two equal parts such that at least 50% of the values are greater than or equal to the median and at least 50% of the values are less than or equal to the median.
- In terms of the empirical cumulative distribution function the median satisfies: $F(\tilde{x}_{0.5}) = 0.5$
- **Outliers do not influence median**
- There is always only one median (uniqueness)



Median

- To calculate median „by hand" we need to sort the data in an ascending order

- Then calculate the median as:
  - When n is odd: $\tilde{x}_{0.5} = x_{((n+1)/2)}$
  - When n is even: $\tilde{x}_{0.5} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$

- Example:

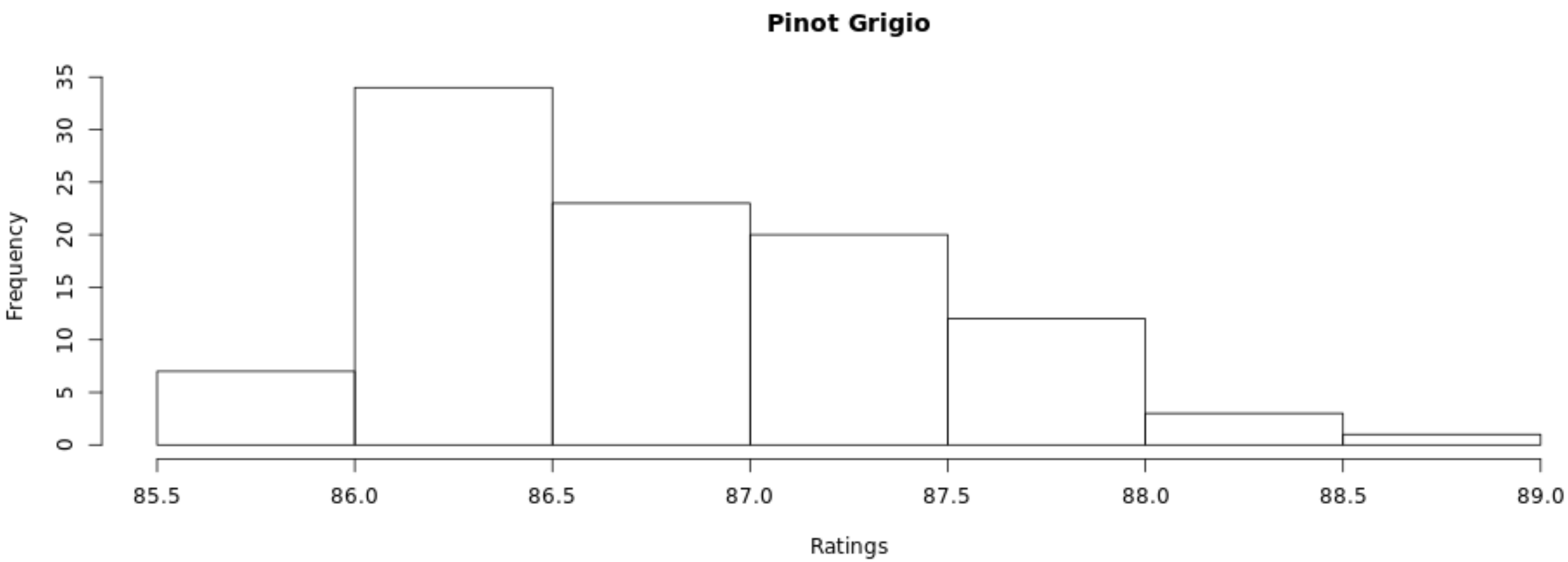
Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35
10	36

$$\tilde{x}_{0.5} = \frac{1}{2}(x_5 + x_6) = \frac{1}{2}(28+29) = 28.5$$

Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35

$$\tilde{x}_{0.5} = x_5 = 28$$

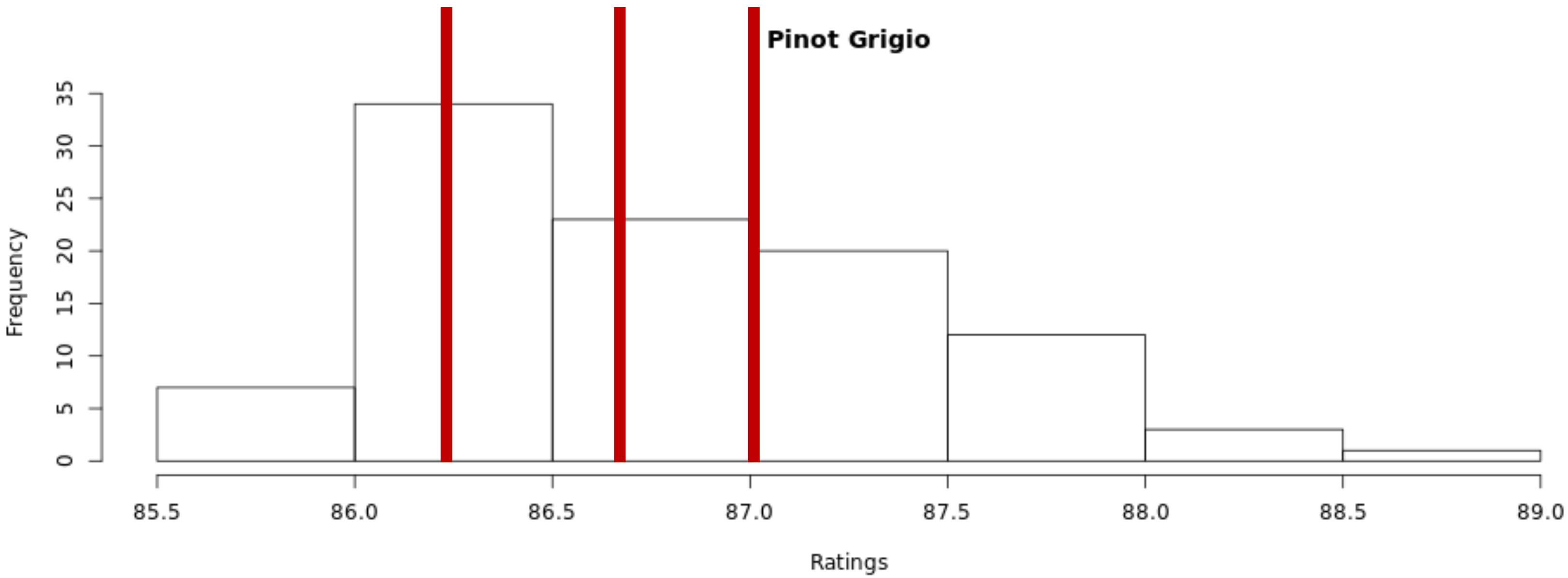
Mean, median, mode



By looking at the graph can you guess what is the relationship between the mean, median, and mode of the Pinot Grigio ratings distribution displayed here?

Source:https://campus.datacamp.com

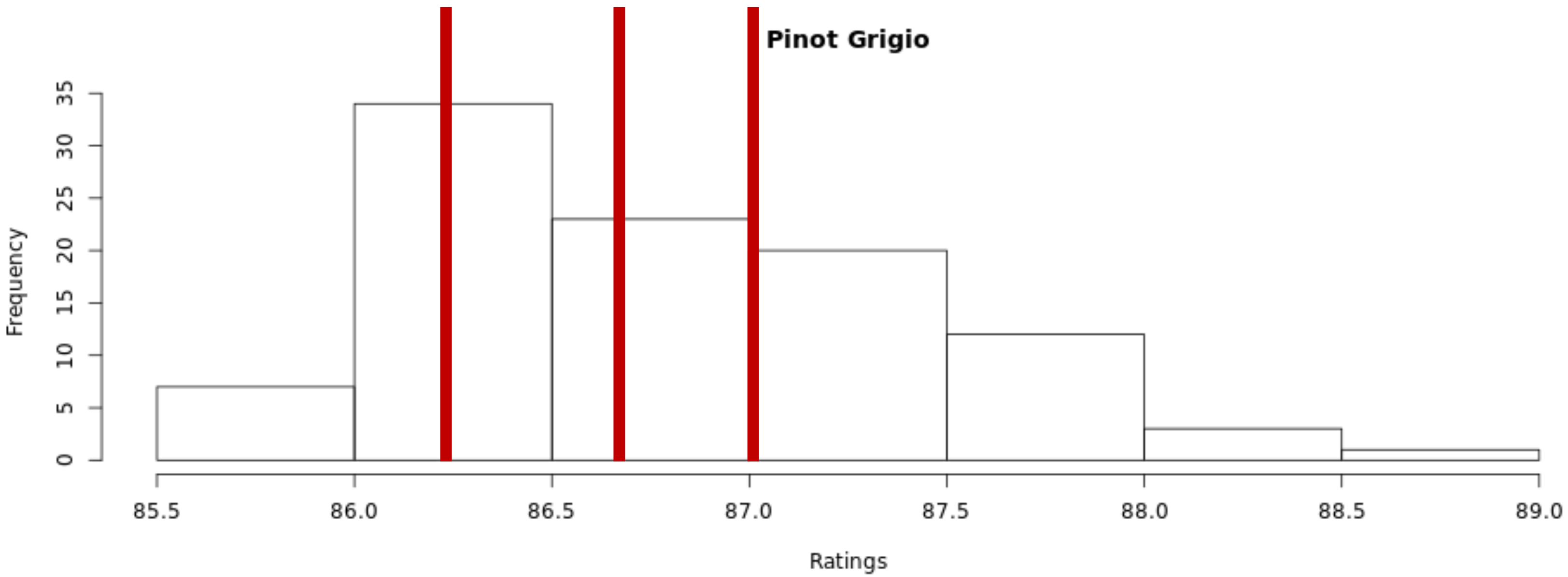
Mean, median, mode



By looking at the graph can you guess what is the relationship between the mean, median, and mode of the Pinot Grigio ratings distribution displayed here?

Source:https://campus.datacamp.com

Mean, median, mode



By looking at the graph can you guess what is the relationship between the mean, median, and mode of the Pinot Grigio ratings distribution displayed here?

Source:https://campus.datacamp.com

mode < median < mean

Mean, median, mode

Exercise 2:

Create 1000 random sampling from binomial distribution with n=10 and p=0.6.

Calculate mean, median and mode of your sample.

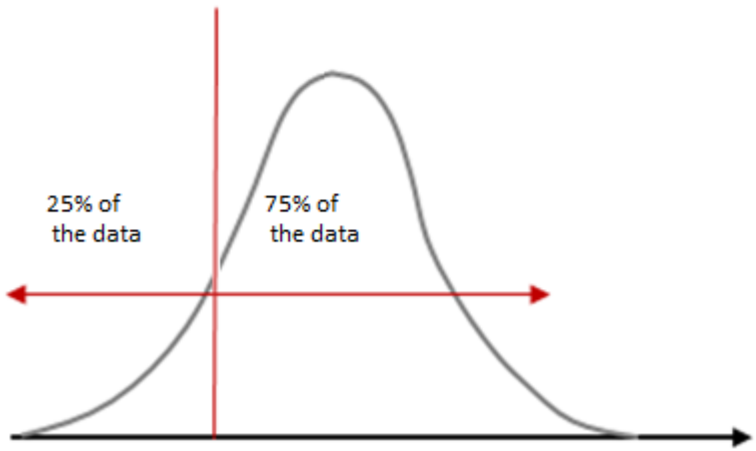
Exercise 3:

Create 1000 random sampling from log-normal distribution.
Calculate mean and median of your sample.

Interpret the relations between the measures - what do they imply?

Quantiles

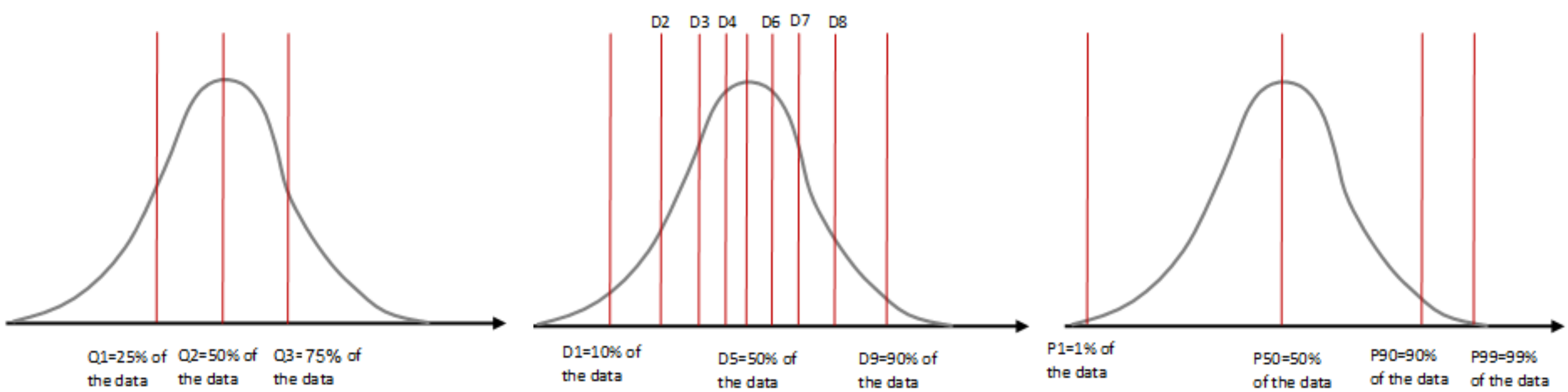
- Quantiles are a generalization of the median idea
- Median splits the data into two equal parts; quantiles split the data into other proportions
- Let's denote a numer from 0 to 1 as α
- The (α *100)% quantile is denoted as $\tilde{x}_\alpha$ and it is defined as the value that divides the data in proportions of (α *100)% and ((1-α) *100)% such that at least (α *100)% values are less than or equal to the quantile and at least ((1-α) *100)% values are greater than or equal to the quantile.



Quartiles, deciles, percentiles

- For specific value of α quantiles have different names:

Name	Proportions in which data are split	α
Quartiles (Q)	4	0.25; 0.5; 0.75
Deciles (D)	10	0.1; 0.2; 0.3;…;0.9
Percentiles (P)	100	0.01;0.02;….;0.99



Quartiles, deciles, percentiles

- Quartiles, deciles and percentiles are calculated mannually in a similar manner to median

- Example:

Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35
10	36

$$Q1 = 25$$

$$Q2 = \frac{1}{2}(28 + 29) = 28.5$$

$$Q3 = 35$$

Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35
10	36

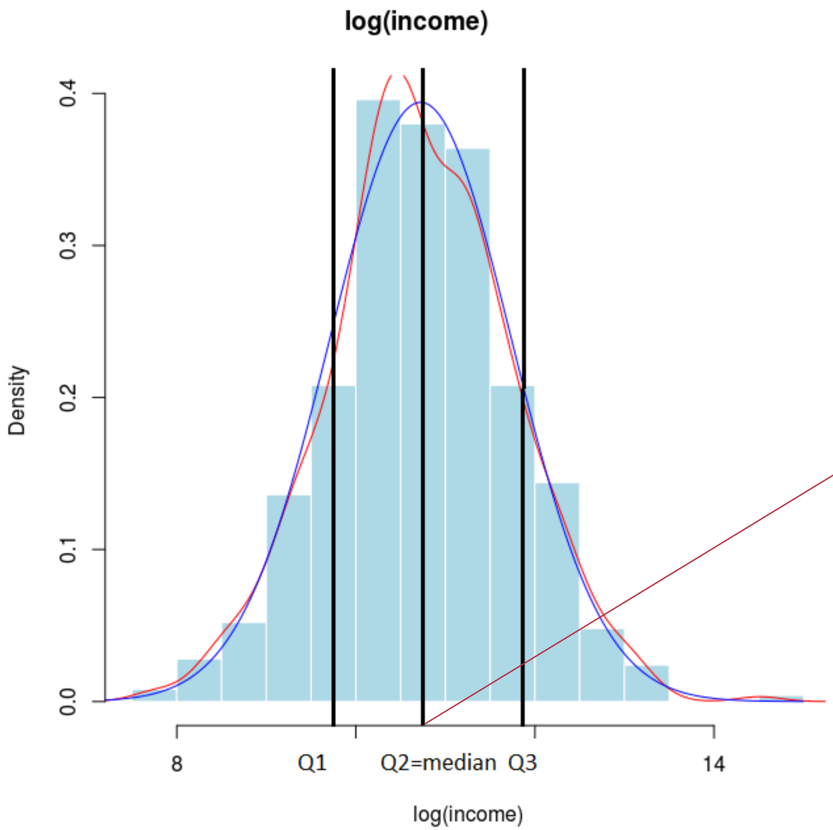
$$D1 = \frac{1}{2}(21 + 23) = 22$$

$$D2 = \frac{1}{2}(23 + 25) = 24$$

$$D5 = \frac{1}{2}(28 + 29) = 2.5$$

$$D9 = \frac{1}{2}(35 + 36) = 35.5$$

Quartiles, deciles, percentiles



Which decile is this??

Which percentile is this??

Quartiles, deciles, percentiles

Exercise 4:

Use data on wages from the NLSY dataset for the US (data for 2010); (NLSY_EDA_class.csv).

- Calculate mean and median wages by sex, by race and by education and interpret the values
- Calculate the value of the 1st, 2nd and 3rd quartile, 1st and 9th decile and 1st, 90th, 99th percentile for full sample and by sex. Interpret the values.

Trimean

- It is sometimes referred to as Tukey's trimean aftern John Tukey - its inventor (1977)

- It is defined as the weighted average of the median and upper and lower quartiles:

$$TM = \frac{Q1 + 2 * Q2 + Q3}{4}$$

- Unlike median it also utilizes information on the first and the third quartiles, which makes it more likely to be representative for the data.

- Example:

Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35
10	36

$$\mathrm{Q1} = 25$$

$$\mathrm{Q2} = \frac{1}{2}(28 + 29) = 28.5$$

$$\mathrm{Q3} = 35$$

$$\mathrm{TM} = \frac{1}{4}(25 + 2 * 28.5 + 35) = 29.25$$

Midmean / Interquartile mean

- It is the mean of the middle 50% of the data
- By dropping from the calculations 25% of extreme values from above and below of the disrtibution, the measure is more resistant to outliers than arithmetic mean
- Example:

We have n=12, so in each quartile there are 3 observations and in the interquartile there are 6 observations.

Day	Temperature
1	21
2	23
3	25
4	26
5	27
6	28
7	29
8	31
9	35
10	35
11	36
12	36

$$Q1 = \frac{1}{2}(25 + 26) = 25.5$$

$$IQM = \frac{1}{6}(26 + 27 + 28 + 29 + 31 + 35) = 29.(3)$$

$$Q3 = \frac{1}{2}(35 + 35) = 35$$

We have n=10, so in each quartile there are 2.5 observations and in the interquartile there are 5 observations: 4 that contribute in 100% and 2 that make up 1 remaining observation, i.e. contribute in 50% each.

Day	Temperature
1	21
2	23
3	25
4	27
5	28
6	29
7	31
8	35
9	35
10	36

$$Q1 = 25$$

$$IQM = \frac{1}{5}(27 + 28 + 29 + 31 + 0.5 * 25 + 0.5 * 35) = 29$$

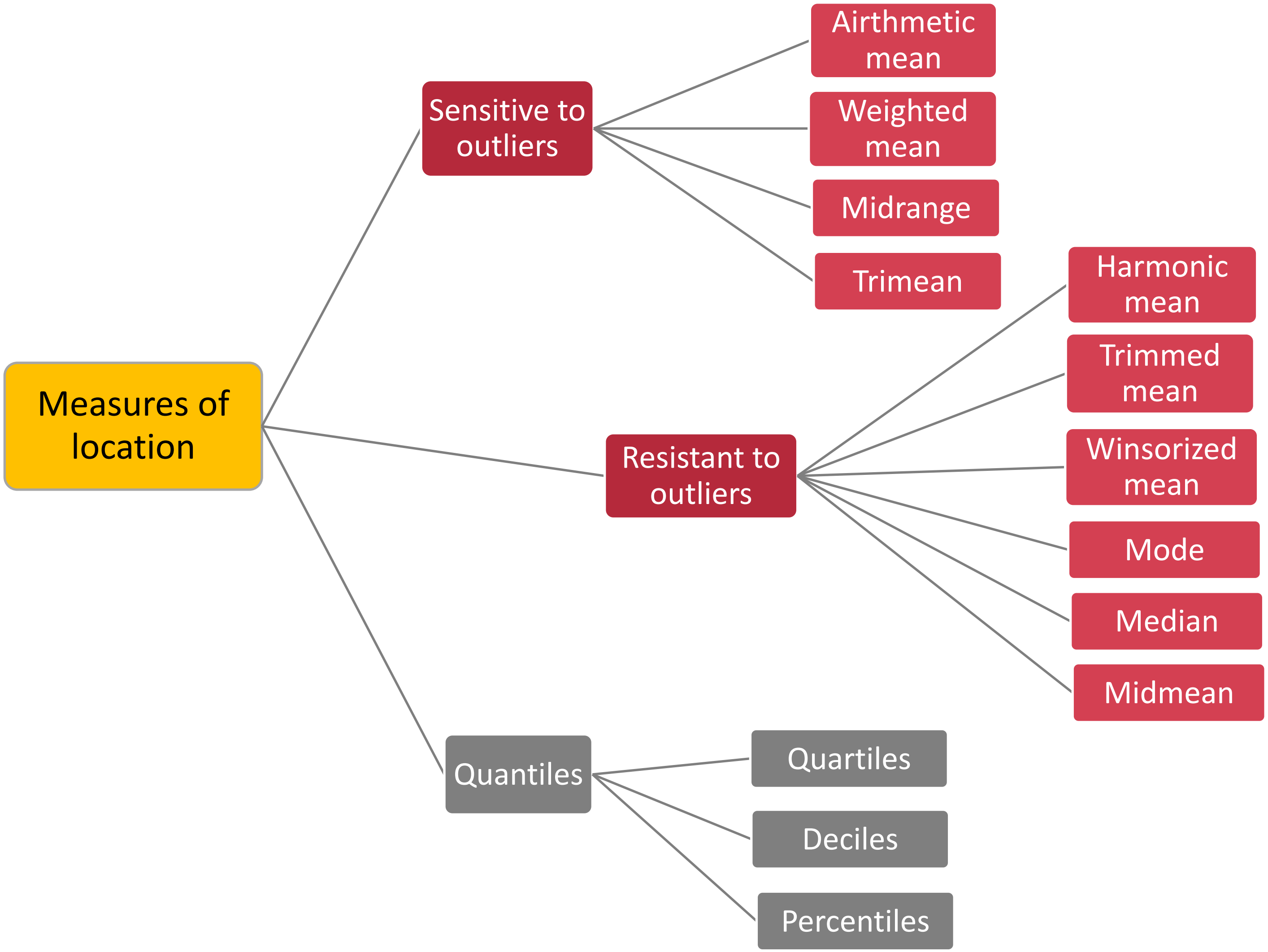
$$Q3 = 35$$

Midmean / Interquartile mean

Exercise 5:

Calculate the interquartile mean for the following temperature data: 28, 43, 32, 18, 7, 15, 22, 23, 29, 9, 11, 16.

Review and summary



Measures of location in R

Measure	Function in R	Alternative function
Arithmetic mean	mean()	
Harmonic mean	1/(mean(1/()))	library(psych) harmonic.mean()
Geometric mean	prod()^(1/length())	library(psych) geometric.mean()
Weighted mean	weighted.mean()	
Midrange	(min()+max())/2	
Trimean	TMH() – it is modified and based on hinges not quartiles	
Trimmed mean	mean(, trim=)	
Winsorized mean	winsor.mean	
Mode	table( ) or: names(sort(-table()))[1]	For continous data: d <- density(x) d$x[d$y==max(d$y)]
Median	median()	
Midmean	Calculate step-by-step using quantiles with specified type=2	
Quartiles	quantile(, probs=c(0.25, 0.5, 0.75)	
Deciles	quantile(, probs=c(0.1,..., 0.9)	
Percentiles	quantile(, probs=c(0.01, ..., 0.99)	

Bibliography

Christian Heumann, Michael Schomaker Shalabh „Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R”, Springer 2016: Chapter 3.2.

http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/ for helpful examples

Thank you for your attention

Time for practice!