Measures of dispersion and shape

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

Lecture 3: 17-18.10.2017

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Descriptive statistics

„Two different data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have different concentrations around the mean.

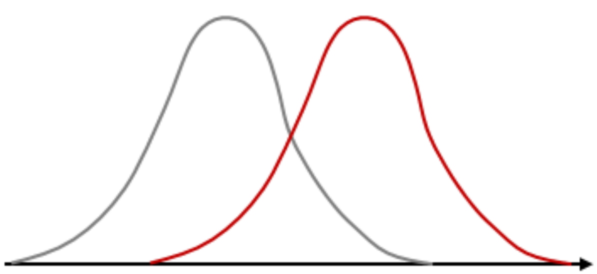
In this case, the location measures may not be adequate enough to describe the distribution of the data.”
(Heuann and Shalabh, 2016)

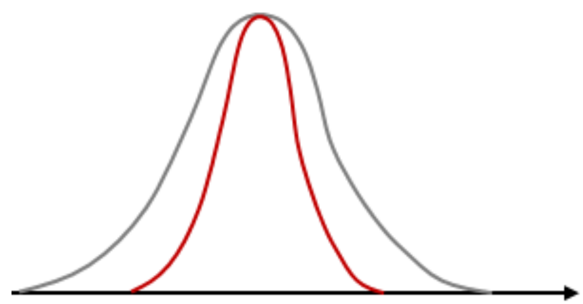
descriptive statistics

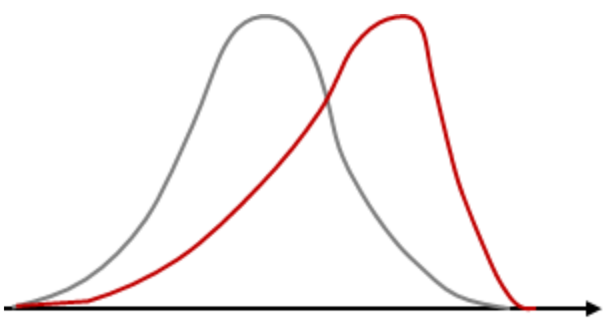
Measures of location (central tendency/position)

Measures of dispersion

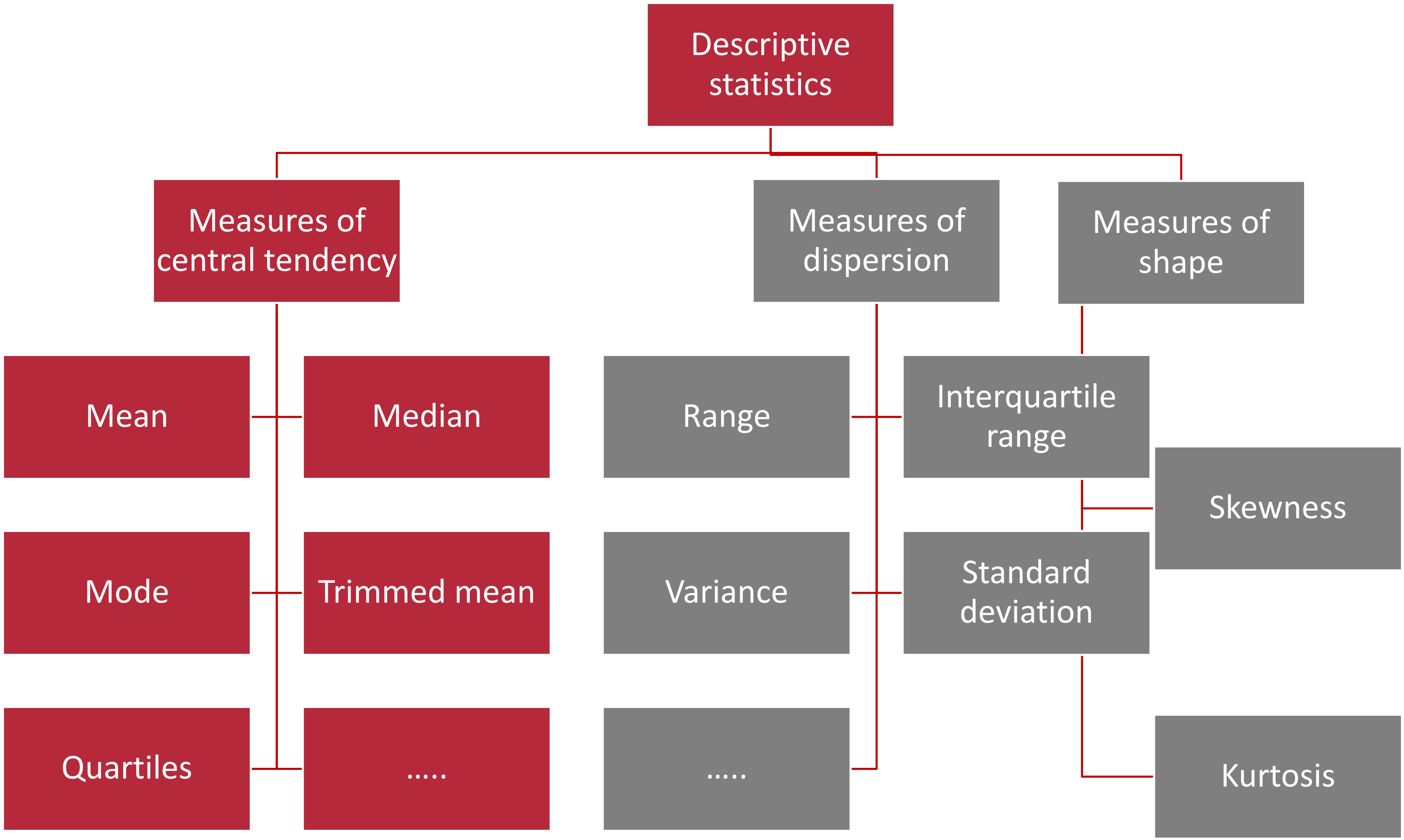
Measures of shape







Descriptive statistics



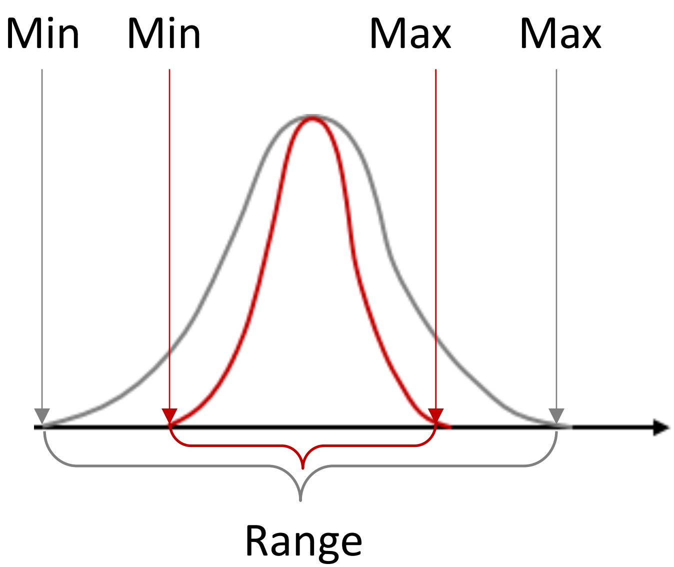
Range

- The range is defined as the difference between the maximum and the minimum value of the data:

$$R = \max(X) - \min(X)$$

- It is easy to calculate, but **it is highly sensitive to extreme values (outliers)!**

- Examples:



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

Min = 21

R=36-21=15

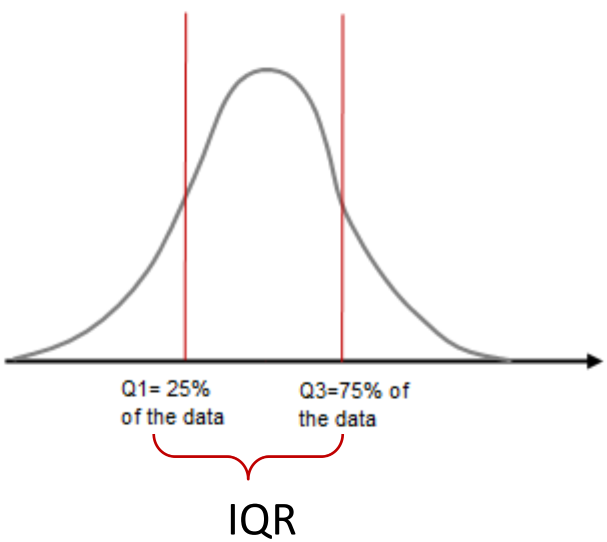
Max=36

Interquartile range

- The interquartile range is defined as a difference between the 3rd and the 1st quartile:

$$IQR = Q3 - Q1$$

- Examples:



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

$Q1 = 25$

$IQR = 35 - 25$

$Q3 = 35$

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

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

$$IQM = 35 - 25.5 = 9.5$$

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

Variance

- The variance measures how the data are spread out around the mean
- It takes into account the whole distribution
- The variance is calculated as:

Population Variance:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Sample Variance:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\text{-}1}$$

- Examples:

Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	21	-8	64
2	23	-6	36
3	25	-4	16
4	27	-2	4
5	28	-1	1
6	29	0	0
7	31	2	4
8	35	6	36
9	35	6	36
10	36	7	49
Mean	29	Sum	246
		Variation (sample)	27.33

Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	0	-27.8	772.84
2	23	-4.8	23.04
3	25	-2.8	7.84
4	27	-0.8	0.64
5	28	0.2	0.04
6	29	1.2	1.44
7	31	3.2	10.24
8	35	7.2	51.84
9	35	7.2	51.84
10	45	17.2	295.84
Mean	27.8	sum	1215.6
		Variation (sample)	135.067

Variance

- The variance measures how the data are spread out around the mean
- It takes into account the whole distribution
- The variance is calculated as:

Population Variance:
$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Sample Variance:
$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Examples:

Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	21	-8	64
2	23	-6	36
3	25	-4	16
4	27	-2	4
5	28	-1	1
6	29	0	0
7	31	2	4
8	35	6	36
9	35	6	36
10	36	7	49
Mean	29	Sum	246
		Variation (sample)	27.33

Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	0	-27.8	772.84
2	23	-4.8	23.04
3	25	-2.8	7.84
4	27	-0.8	0.64
5	28	0.2	0.04
6	29	1.2	1.44
7	31	3.2	10.24
8	35	7.2	51.84
9	35	7.2	51.84
10	45	17.2	295.84
Mean	27.8	sum	1215.6
		Variation (sample)	135.067

What would be the variation of the sample consisting of 10 observations, each equal to 20 (degrees)?

What is variance's unit of measurement?

Standard deviation

- **It has the same unit of measurement as the data**
- The standard deviation is the square root of the variance (population/sample):

Population sd:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Sample sd:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Examples:

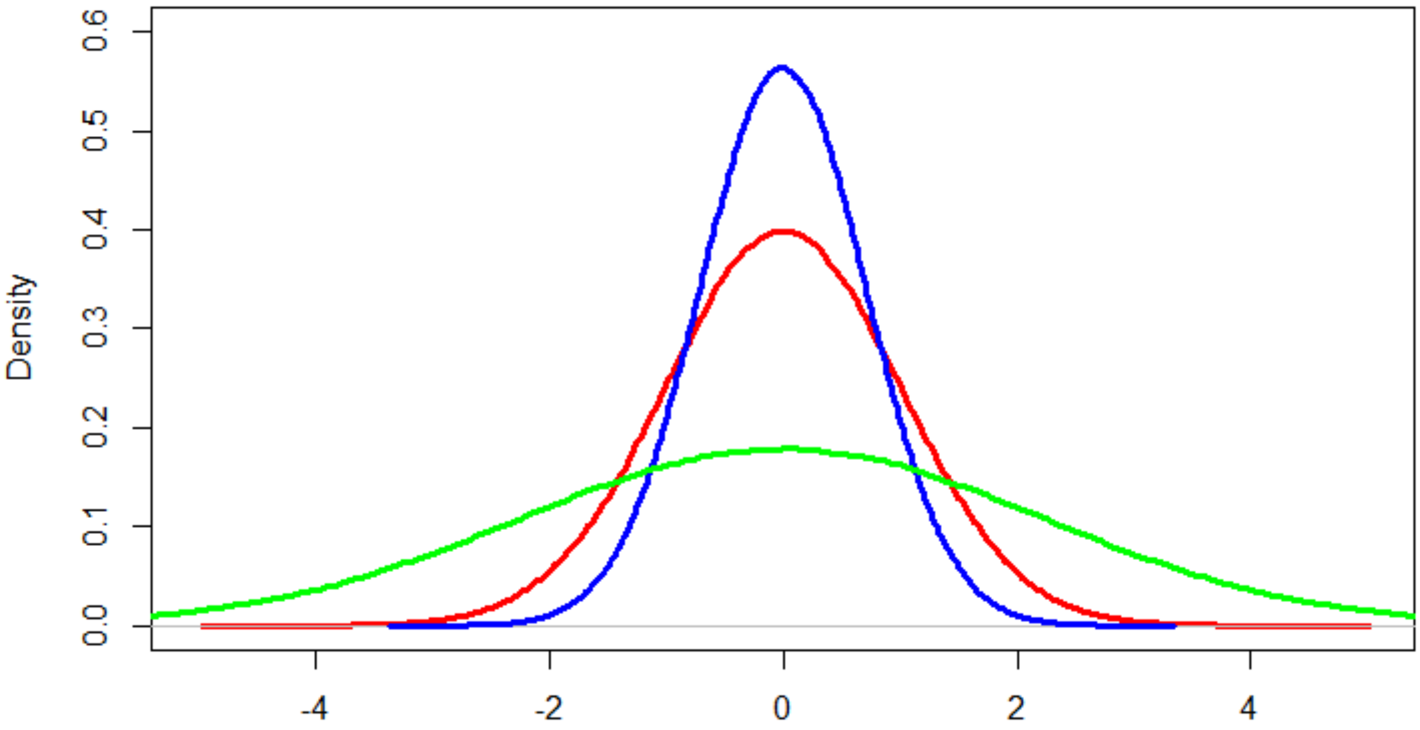
Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	21	-8	64
2	23	-6	36
3	25	-4	16
4	27	-2	4
5	28	-1	1
6	29	0	0
7	31	2	4
8	35	6	36
9	35	6	36
10	36	7	49
Mean	29	Sum	246
		Variation (sample)	27.33
		SD	5.23

Day	Temp.	$(x_i - x)$	$(x_i - x)^2$
1	0	-27.8	772.84
2	23	-4.8	23.04
3	25	-2.8	7.84
4	27	-0.8	0.64
5	28	0.2	0.04
6	29	1.2	1.44
7	31	3.2	10.24
8	35	7.2	51.84
9	35	7.2	51.84
10	45	17.2	295.84
Mean	27.8	sum	1215.6
		Variation (sample)	135.067
		SD	11.62

What would be the standard deviation of the sample consisting of 10 observations, each equal to 20 (degrees)?

Variance and standard deviation

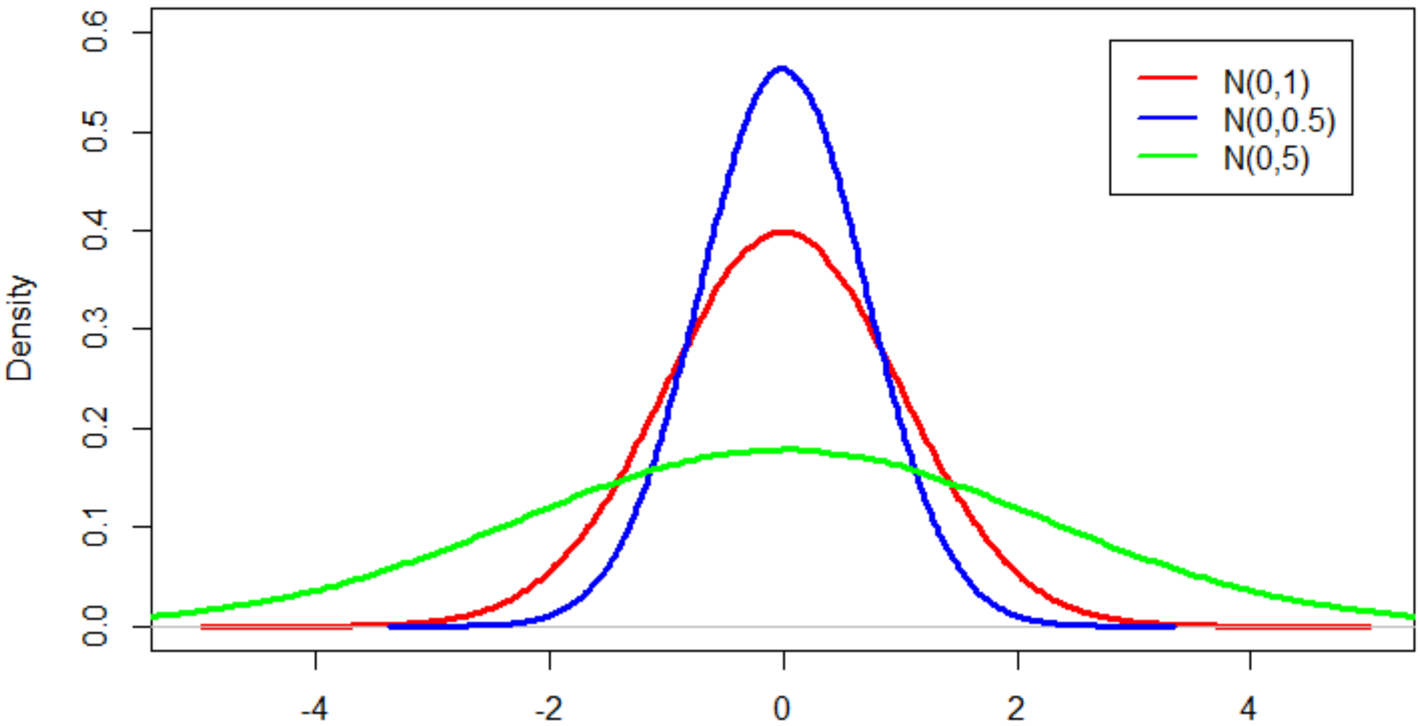
- Which sample has the greatest variance and standard deviation?



- Why?

Variance and standard deviation

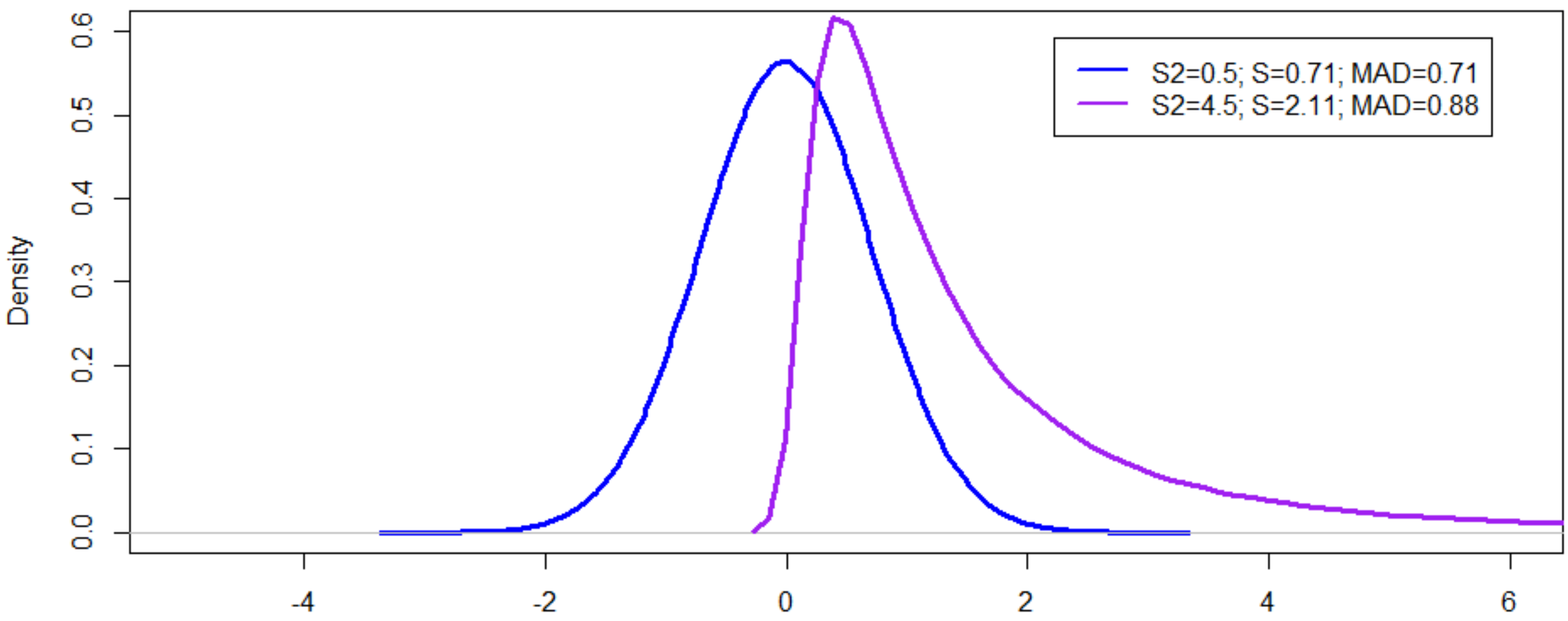
- Which sample has the greatest variance and standard deviation?



- Why?

Median Absolute Deviation

- The median absolute deviation (MAD) is also a measure of how spread the data are - but around the median.
- It is defined as: $MAD = median(|x_i - \tilde{x}_{0.5}|)$
- Compared to the variance and the standard deviation, MAD is less affected by extreme values and non-normality.



Coefficient of variation

- Sometimes we want to compare the variability in two samples which use different measurement units (e.g. prices in PLN and EUR)
- The **coefficient of variation** is a unit-free measure of dispersion and takes into account both the mean and the standard deviation
- It is defined as: $CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$
- Example:

Hotels in Warsaw	Price (in PLN)	Hotels in London	Price (in GBP)
Mariott	220	Indigo	110
Hilton	270	Park Grand London	100
Mercure	170	Crowne Plaza	125
Novotel	120	Strand Palace	120
Polonia	100	Double Tree	90
Intecontinental	180	Rosewood	320
Bristol	150	Rubens Palace	140
Metropol	190	Hilton	180
Holiday Inn	230	Holiday Inn	170
Sofitel	200	Picadilly London	175
MEAN	183	MEAN	153
STD	51.22	STD	66.72
CV	28%	CV	44%

Measure of dispersion

Exercise 1:

Use data on airbnb offers in Warsaw (Airbnb_Warsaw_July_2017.csv) and Vienna (Airbnb_Vienna_July_2017.csv) as for July 2017.

- Calculate the overall mean price and mean prices for various room types in both cities.
- For each city summarize the variability of the prices for various types of rooms using: range, interquartile range, variance, standard deviation, MAD.
- Identify the room type for which the variation in prices is the greatest.
- Compare the variation in prices of various room types in Warsaw and in Vienna.

Measures of shape

- To define measures of shape we first need to define the concept of moments
- Recall from probability theory:

The *k* ordinary moment (for discrete distribution) is defined as: $$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

The *k* central moment (for discrete distribution) is defined as: $$M_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k$$

- For k=1 we have: $m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i^1 = \bar{x}$ ← Mean

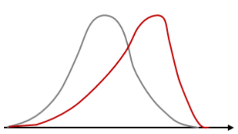
For k=2 we have: $M_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = S^2$ ← Variance

Measures of shape

- Measures of shape use higher order moments, i.e. 3rd and 4th central moments:

$$M_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

It is used to measure how **ASSYMETRIC/SKEWED** the distribution is.

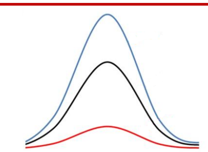


Coefficient of assymetry defined as:

$$\rho_{asym} = \frac{M_3}{s^3}$$

$$M_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

It is used to measure how **FLAT** the distribution is.



Kurtosis defined as:

$$\hat{\rho}_{kurtosis} = \frac{M_4}{s^4}$$

Coefficient of assymetry

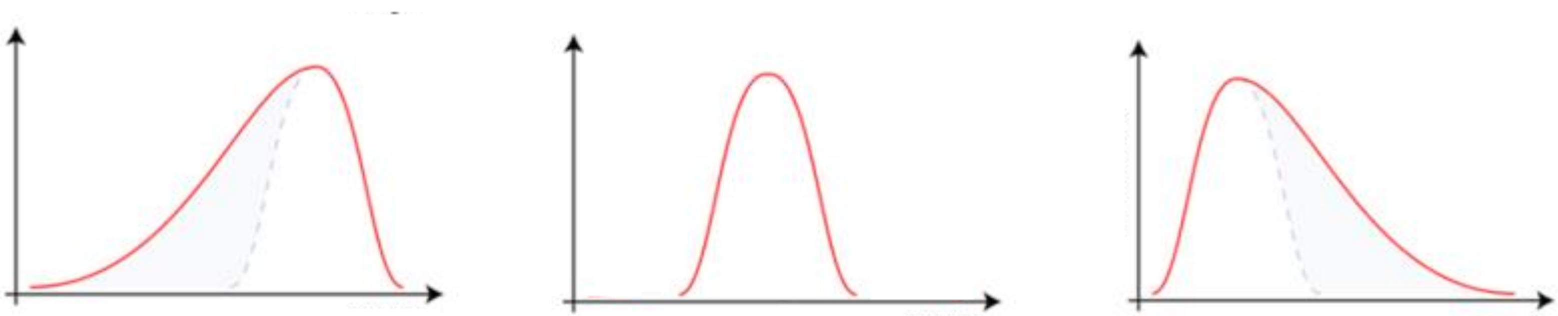
$$\rho_{asym} = \frac{M_3}{s^3}$$

Symmetric (normal) distribution

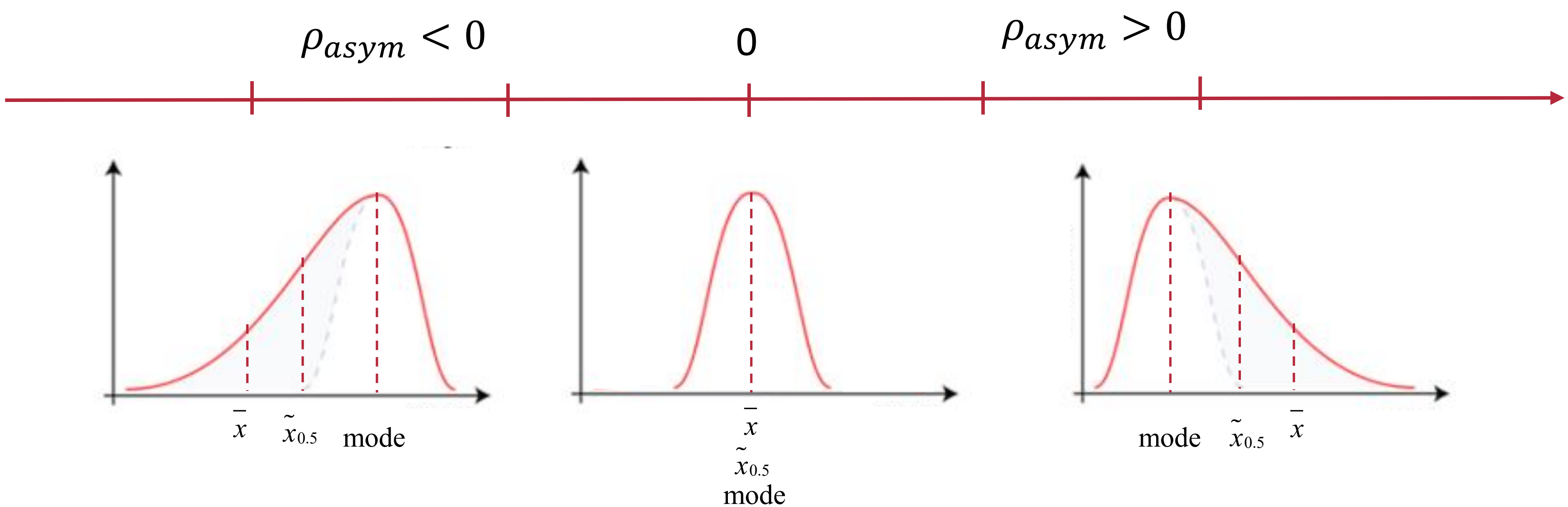
ρ_{asym}<0 ⇔ negative skeweness (long left tail)

0

ρ_{asym}>0 ⇔ positive skeweness (long right tail)



Coefficient of assymetry and its relations to mean, median and mode



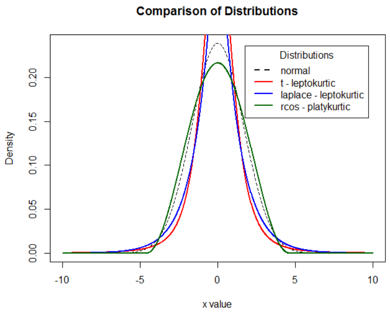
mean<median<mode

mean=median=mode

mode<median<mean

Kurtosis

$$\rho_{kurtosis} = \frac{M_4}{s^4}$$



- Kurtosis describes how „fat” are the tails of the distribution, i.e. the probability of the observation very distant from the average

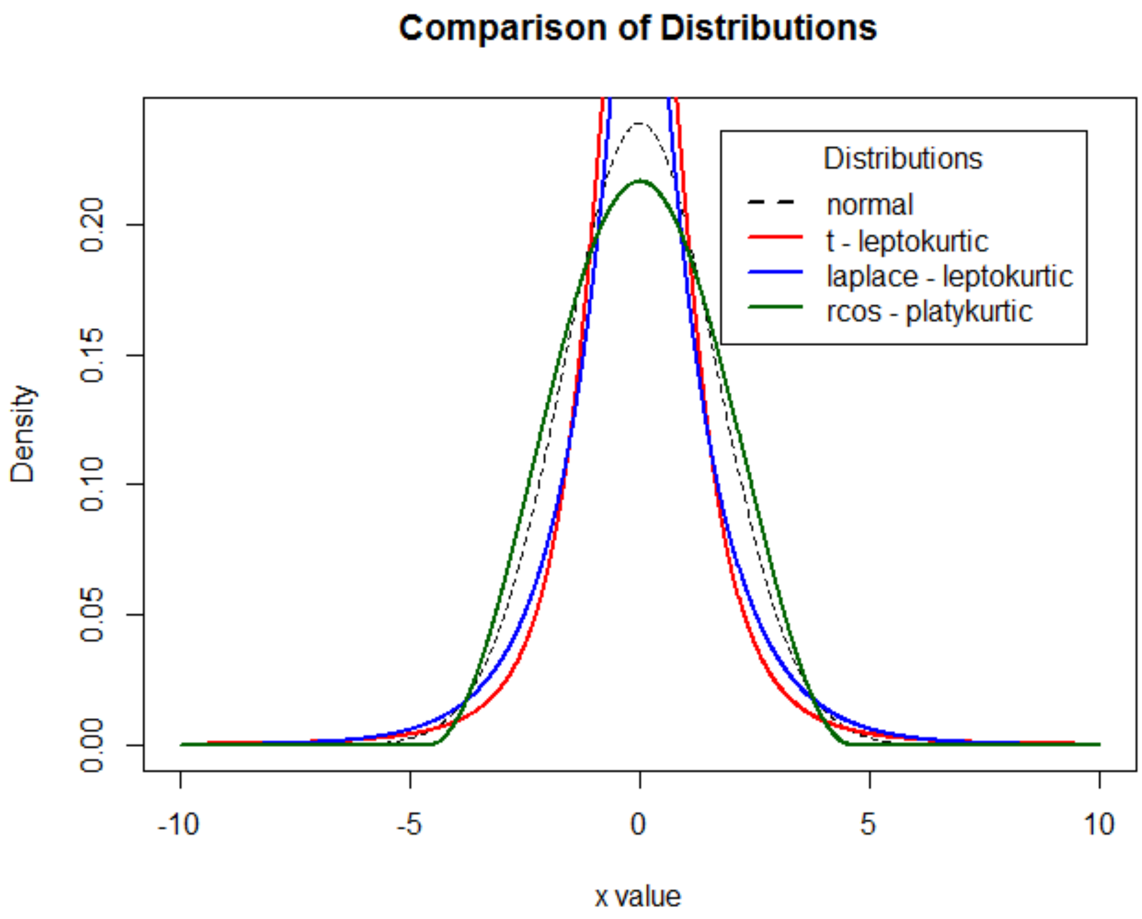
- If the tails of the distribution are thick we can expect that the are some unusual observations (outliers)

- For standard normal distribution $\rho_{kurtosis} = 3$

- Sometimes we define an **excess kurtosis** which makes the measure comparable to the standard normal distribution:

$$\rho_{excess\_kurtosis} = \rho_{kurtosis} - 3$$

Kurtosis - distribution

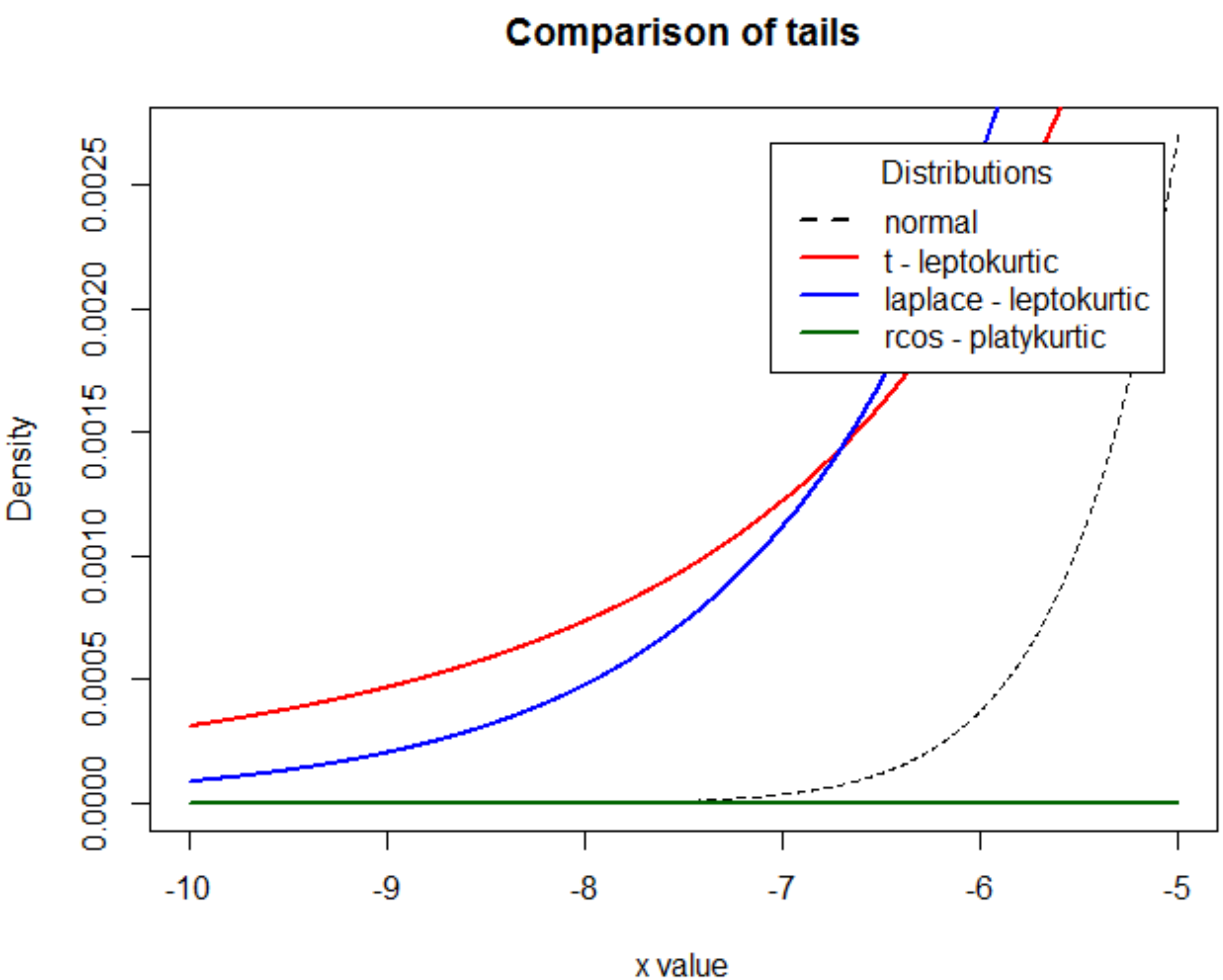


Positive excess kurtosis

Standard normal distribution
excess kurtosis = 0
kurtosis=3

Negative excess kurtosis

Kurtosis - tail



Positive excess kurtosis

Standard normal distribution
excess kurtosis = 0
kurtosis=3

Negative excess kurtosis

Measures of shape

Exercise 2:

Use data on airbnb offers in Warsaw (Airbnb_Warsaw_July_2017.csv).

- Calculate the mean, median and mode for prices of rooms in Warsaw.
- What can you say about the shape of the price distribution based on these measures?
- Verify your answer by calculating relevant measure of shape.

Measures of shape

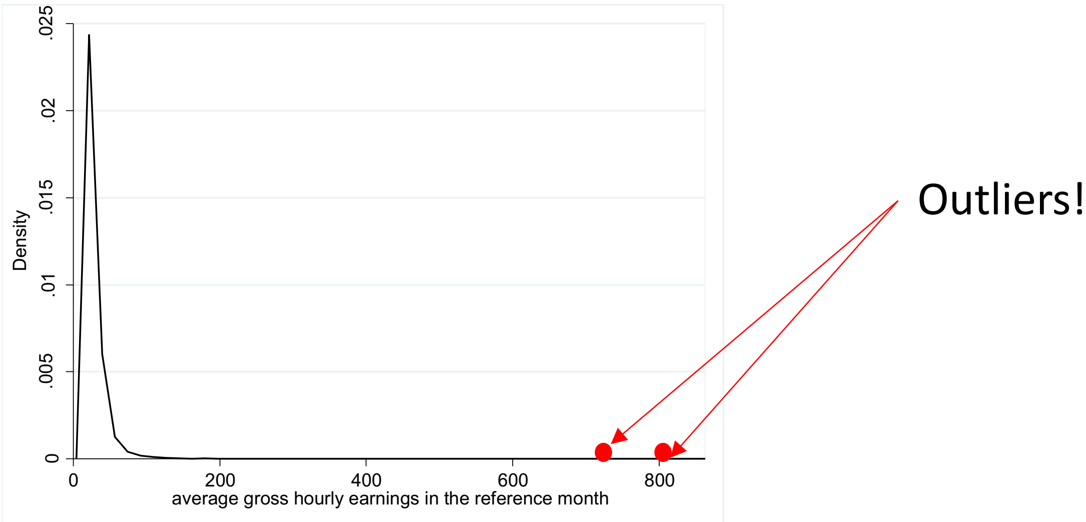
Exercise 3:

Use data on airbnb offers in Vienna (Airbnb_Viennna_July_2017.csv).

- Calculate mean, median and mode for the satisfaction of the travelers, who stayed with aribnb in Vienna.
- What can you say about the shape of the satisfaction distribution based on these measures?
- Once again verify your answer by calculating relevant measure of shape.

Outliers

- Outliers affect many of the descriptive statistics



- It is often useful to determine which observations can be considered as outliers

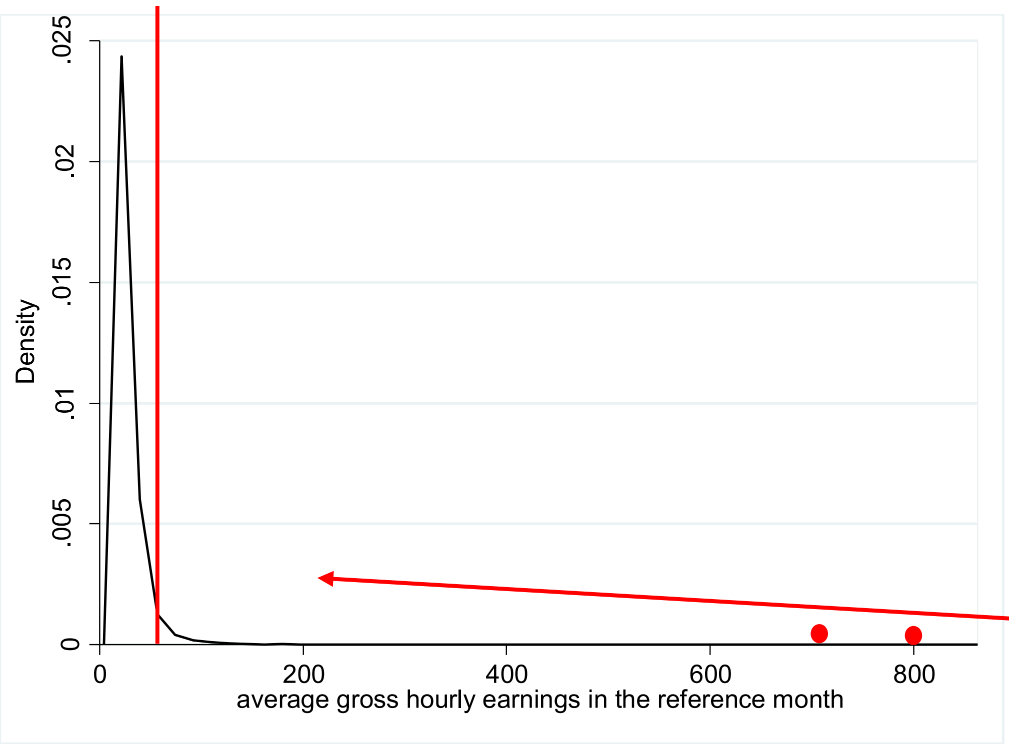
- There are several rules to determine outliers:
    - The IRQ rule
    - The Z-score rule
    - The modified Z-score rule

Outliers: The IRQ rule

- Calculate first quartile (Q1)
- Calculate third quartile (Q3)
- Calculate the interquartile range (IQR=Q3-Q1)
- Compute Q1-1.5 × IQR ➔ Any observation less than this is a potential outlier
- Compute Q3+1.5 × IQR ➔ Any observation more than this is a potential outlier

Outliers: The IRQ rule

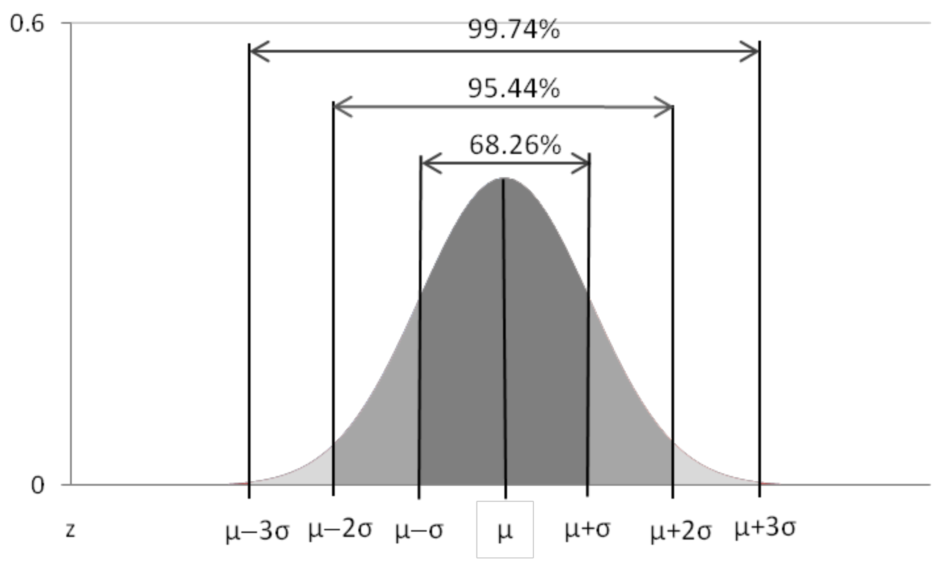
- Calculate first quartile (Q1)
- Calculate third quartile (Q3)
- Calculate the interquartile range (IQR=Q3-Q1)
- Compute Q1-1.5 × IQR ➔ Any observation less than this is a potential outlier.
- Compute Q3+1.5 × IQR ➔ Any observation more than this is a potential outlier.



- Q1=12 PLN
- Q3=27 PLN
- IQR=27-12=15
- Q1-1.5 × IQR = 12-22.5 = -10.5
  no outliers in the left tail of the distrubution
- Q3+1.5IQR = 27+22.5=49.5
  all obervations>49.5 are potential outliers!

Outliers: The Z-score

- The Z-score is calculated for each observation in the sample as: $Z = \dfrac{X - \overline{X}}{S}$
- It tells how many standard deviations away from the mean a given observation is located in the distribution
- For normal distribution we observe **that nearly all the observation (99.7%) are located 3 standard deviations away from the mean**



Source: http://simulationeducators.blogspot.com/

If a given observation has:
Z-score>2
Z-score>3
→ we may suspect it is a potential outlier

Outliers: The modified Z-score

- The Z-score may be misleading as the maximum Z-score is given by $(n-1)\sqrt{n}$
- It means that it may be problemetic for **small samples.**
- For small samples it is recommended to use a modified Z-score:

$$Z_{\text{modified}} = \frac{0.6745(x - \tilde{x}_{0.5})}{MAD}$$

- Iglewicz & Hoaglin (1993) recommend that any observation for which **modified Z-score>3.5** should be treated as a potential outlier

Outliers: The Z-score and modified Z-score

- Example:

Day	Temp.	Z-score	Modified Z-score
1	21	-1.59	-0.76
2	23	-1.21	-0.56
3	25	-0.83	-0.35
4	26	-0.64	-0.25
5	27	-0.44	-0.15
6	28	-0.25	-0.05
7	29	-0.06	0.05
8	31	0.32	0.25
9	35	1.08	0.66
10	35	1.08	0.66
11	36	1.27	0.76
12	36	1.27	0.76
Mean	29.33		
Variance	27.52		
Standard deviation	5.25		
Median	28.5		
MAD	6.67		

Day	Temp.	Z-score	Modified Z-score
1	21	-0.65	-0.76
2	23	-0.55	-0.56
3	25	-0.46	-0.35
4	26	-0.41	-0.25
5	27	-0.36	-0.15
6	28	-0.32	-0.05
7	29	-0.27	0.05
8	31	-0.17	0.25
9	35	0.02	0.66
10	35	0.02	0.66
11	36	0.06	0.76
12	100	3.09	7.23
Mean	34.67		
Variance	446.42		
Standard deviation	21.13		
Median	28.5		
MAD	6.67		

Outliers

Exercise 4:

Use data on airbnb offers in Warsaw (Airbnb_Warsaw_July_2017.csv).

Using three methods for identifying outliers that we covered during the class check data on room prices in Warsaw and determine potential outliers.

Functions in R

Measure	Function in R	Alternative function
Range	range()	max()-min()
Interquartile range	IQR()	quantile(, probs=c(0.75))-quantile(, probs=c(0.25))
Variance	var()	
Standard deviation	sd()	
Median absolute deviation	mad()	
Coefficient of variation	install.packages("RVAideMemoire") library(RVAideMemoire) cv()	[sd()/mean()]/100
Coefficient of assymetry	install.packages("e1071") library(e1071) skeweness()	
Kurtosis	install.packages("e1071") library(e1071) kurtosis()	
Z-score	scale(x,center=TRUE, scale=TRUE)	
Modified Z-score	install.packages("spatialEco") library(spatialEco) outliers()	

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

Boris Iglewicz and David Hoaglin (1993), "Volume 16: How to Detect and Handle Outliers", *The ASQC Basic References in Quality Control: Statistical Techniques*.

Thank you for your attention

Time for practice!