UNIVERSITY OF WARSAW
Faculty of Economic Sciences

STATISTICS & ECONOMETRICS
LECTURE 1

Marcin Chlebus
mchlebus@wne.uw.edu.pl

INTRODUCTION

University of Warsaw
Faculty of Economic Sciences

STATISTICAL OUTPUTS

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

The New York Times
Friday, October 17, 2014

Business Day
Markets

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

Search | Global | DealBook | Markets | Economy | Ener

MARKETS OVERVIEW | U.S. MARKETS | WORLD MARKETS | FUNDS | CURRENCIES | COMMODITIES | BONDS | CONSUMER RATES

Markets Overview

POWERED BY THOMSON REUTERS

U.S. Markets »

S.&P. 500 | DOW | NASDAQ

1 day | 5 days | 1 month | 1 year

1,862.76 +0.27 +0.01%

At close 10/16/2014

Market Summary

At Close 10/16/2014: Among individual stocks, the two top percentage gainers in the S.&P. 500 were United Rentals Inc. and Denbury Resources Inc.

World Markets »

AMERICAS | EUROPE | ASIA/PACIFIC | MIDDLE EAST/AFRICA

Key Indexes	At 5:41 AM ET At least 15-minute delay	Change	% change	1 month	1 year	Low 52-week High
FTSE 100 BRITAIN	6,259.36	+63.45	+1.02%	−7.69%	−4.82%	
DAX GERMANY	8,732.85	+149.95	+1.75%	−9.61%	−0.90%	
CAC 40 FRANCE	3,990.05	+71.43	+1.82%	−9.96%	−5.89%	
FTSE Eurofirst 300 EUROPE	1,266.17	+20.39	+1.64%	−8.59%	−0.15%	

Coming Market Holidays

http://markets.on.nytimes.com/research/markets/overview/overview.asp

SPORT FOOTBALL

Home | Football | Formula 1 | Cricket | Rugby U | Tennis | Golf | Athletics | Cycling | All Sport

Premier League > Results | Fixtures | Table | Live Scores | All Teams | Leagues & Cups

Football Tables

Share

Win Draw Loss

Barclays Premier League Table

Position	Team	P	W	D	L	F	A	GD	Pts	Last 10 games	
1	Chelsea	7	6	1	0	21	7	14	19		Report
2	Man City	7	4	2	1	14	7	7	14		Report
3	Southampton	7	4	1	2	11	5	6	13		Report
4	Man Utd	7	3	2	2	13	10	3	11		Report
5	Swansea	7	3	2	2	10	8	2	11		Report
6	Tottenham	7	3	2	2	9	7	2	11		Report
7	West Ham	7	3	1	3	12	10	2	10		Report
8	Arsenal	7	2	4	1	11	9	2	10		Report
9	Liverpool	7	3	1	3	10	10	0	10		Report
10	Aston Villa	7	3	1	3	4	9	-5	10		Report
11	Hull	7	2	3	2	11	11	0	9		Report
12	Leicester	7	2	3	2	11	12	-1	9		Report
13	Sunderland	7	1	5	1	8	7	1	8	Win 1 - 0 v Newcastle 29th September 2014	Report
14	West Brom	7	2	2	3	8	9	-1	8		Report
15	Crystal Palace	7	2	2	3	10	12	-2	8		Report
16	Stoke	7	2	2	3	6	8	-2	8		Report
17	Everton	7	1	3	3	13	16	-3	6		Report
18	Newcastle	7	0	4	3	7	14	-7	4		Report
19	Burnley	7	0	4	3	3	10	-7	4		Report
20	QPR	7	1	1	5	4	15	-11	4		Report

Last updated 11 days ago

http://www.bbc.com/sport/football/premier-league/table

LEARNIG FROM DATA

recordid	pgssyear	weight	voiev49	region8	size	hompop	adults	fepol
1	1992r	.51775	warszawskie	centralny	m 10-24tys	jedna (resp)	jedno	zgadzam się
2	1992r	.81117	warszawskie	centralny	m 10-24tys	dwie osoby	dwoje	zgadzam się
3	1992r	.84094	warszawskie	centralny	m 10-24tys	dwie osoby	dwoje	nie zgadzam się
4	1992r	.81117	warszawskie	centralny	m 10-24tys	cztery osoby	dwoje	nie zgadzam się
5	1992r	1.21675	warszawskie	centralny	m 10-24tys	piec osób	troje	nie zgadzam się
6	1992r	.51775	warszawskie	centralny	m 10-24tys	jedna (resp)	jedno	zgadzam się
7	1992r	.74129	warszawskie	centralny	m 10-24tys	cztery osoby	dwoje	zgadzam się
8	1992r	.51775	warszawskie	centralny	m 10-24tys	jedna (resp)	jedno	zgadzam się
9	1992r	1.21352	warszawskie	centralny	m 10-24tys	dwie osoby	dwoje	zgadzam się
10	1992r	.60676	warszawskie	centralny	m 10-24tys	jedna (resp)	jedno	zgadzam się
11	1992r	1.11193	warszawskie	centralny	m 10-24tys	cztery osoby	troje	nie zgadzam się
12	1992r	.94985	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	nie zgadzam się
13	1992r	.8494	warszawskie	centralny	m 500 + tys	trzy osoby	dwoje	zgadzam się
14	1992r	.74568	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	nie jestem pewien/a
15	1992r	.94985	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	zgadzam się
16	1992r	.74568	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	zgadzam się
17	1992r	.74568	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	zgadzam się
18	1992r	2.26464	warszawskie	centralny	m 500 + tys	cztery osoby	czworo	zgadzam się
19	1992r	.94985	warszawskie	centralny	m 500 + tys	trzy osoby	dwoje	zgadzam się
20	1992r	1.11853	warszawskie	centralny	m 500 + tys	trzy osoby	troje	zgadzam się
21	1992r	.94985	warszawskie	centralny	m 500 + tys	dwie osoby	dwoje	nie zgadzam się

Statistics is the science of **designing studies or experiments, collecting data and modeling/analyzing data** for the purpose of **decision making and scientific discovery** when the **available information is both limited and variable**. That is, statistics is the science of ***Learning from Data***.

DATA AND STATISTICS

- Def. DATA means groups of information that represent the qualitative or quantitative attributes of a variable or set of variables. Data are typically the results of measurements.
- DEF. Statistics is the science of making effective use of numerical data relating to groups of individuals or experiments. It deals with all aspects of this, including not only the collection, analysis and interpretation of such data, but also the planning of the collection of data, in terms of the design of surveys and experiments. (Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, OUP.)

University of Warsaw
Faculty of Economic Sciences

WHY STATISTICS IS IMPORTANT

UNDERSTANDING OF INFORMATION

- Need to know how to evaluate published numerical facts (commercials, polls – sampling issue)

WORK EXPECTATION

- your profession or employment may require you to interpret the results of sampling (surveys or experimentation) or to employ statistical methods of analysis to make inferences in your work.

STATISTICS MISUNDERSTANDING

- Misunderstandings of statistical results can lead to major errors by government policymakers, medicalworkers, and consumers of this information.

STATISTICS MISUNDERSTANDING

Reasons of misunderstanding statistics

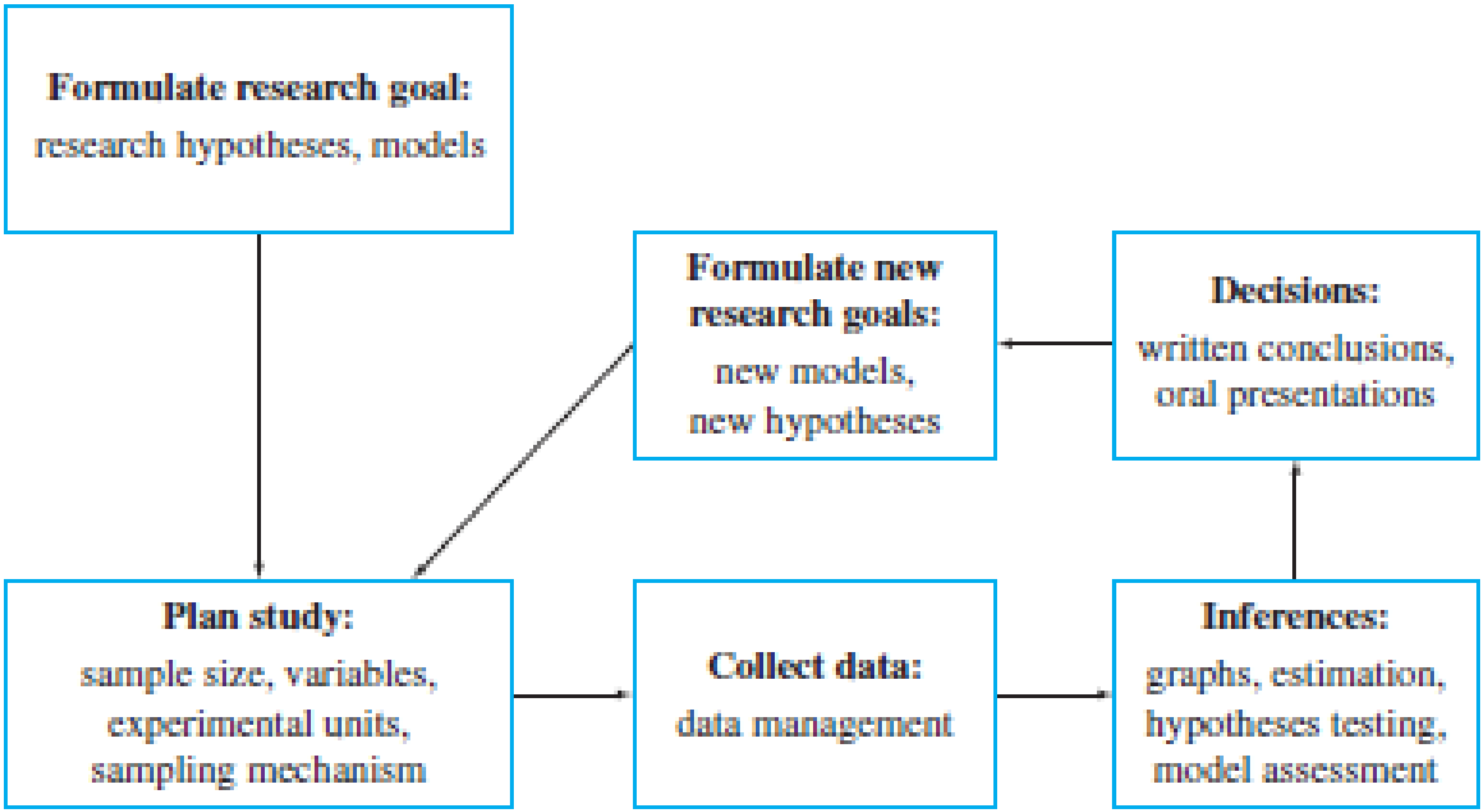
1. Causation issue (Ice-cream consumption vs. Refreshment drinks consumption)
2. Statistically vs. practically significant findings (Difference between height of people born in different months)
3. Size of the sample (sample size not large enough)
4. Bias caused by data collection options (selection of sample group, the way in which questions are phrased)
5. Probability versus conditional probability (Probability of win Oscar)
6. Role of degree of variability in interpreting what is a “normal” occurrence (Average vs. Interval).

"There are three kinds of lies:
lies,
damned lies
and statistics.„

IMPORTANCE OF FAIRNESS

- Even true and correctly calculated statistics can be misused in order to support false thesis.
- Possible abuse of statistics:
  - Use of ‘improperly’ selected statistical methods,
  - ‘Data adjustments’
  - Underlying some data and ignoring other
- Data analysis is/should be objective
  - Should represent statistical measures/approaches that fit, in the best possible way, a given problem (data and research question)
- Data interpretation is usually subjective
  - Therefore it should be performed in honest, neutral and clear way.

Scientific Method



Ott, R. Longnecker, M. (2010) An Introduction to Statistical Methods and Data Analysis, Cengage Learning

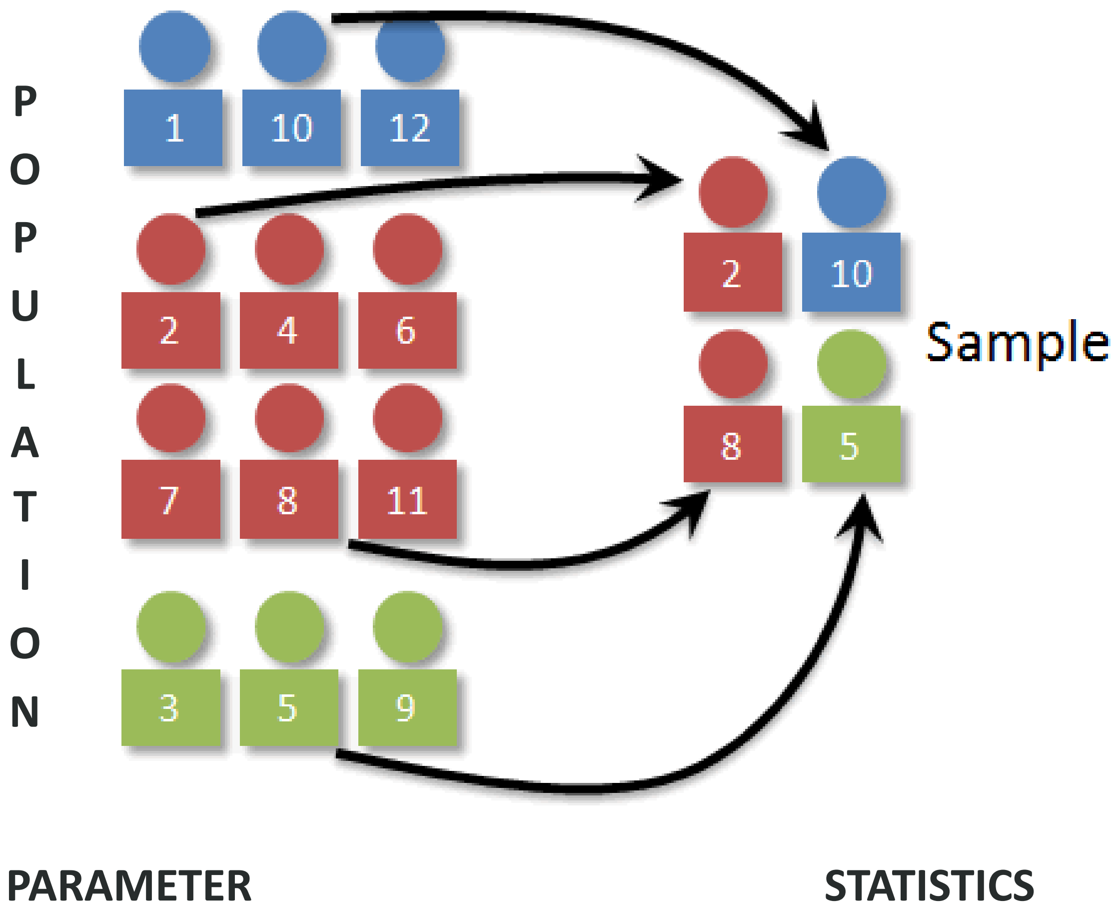
REFERENCE

- SUGGESTED REFERENCE:
  - CHAPTER 1: Ott, R. Longnecker, M. (2010) An Introduction to Statistical Methods and Data Analysis, Cengage Learning , 6th Edition.

University of Warsaw
Faculty of Economic Sciences

POPULATION VS. SAMPLE
DESCRIPTIVE STATISTICS DEFINITION
TYPE OF DATA

Population vs. Sample



REPRESENTATIVNESS OF A SAMPLE ISSUE

- **Samples used in statistical tests that do not represent the population adequately can give reliable results but with little relevance to the population that it came from**

http://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

DEFINITIONS

Population vs. sample

- A **population is the collection of *all outcomes, responses, measurements, or*** counts that are of interest.
  - Population of our S&E group
- A **sample is a subset, or part, of a population.**
  - **4 students from** our S&E group

Parameter vs. statistics

- A **parameter is a numerical description of a *population characteristic.***
- A **statistic is a numerical description of a *sample characteristic.***

DESCRIPTIVE STATISTICS

- **Two branches of statistics:**
  - **Descriptive statistics is the branch of statistics that involves the organization,** summarization, and display of data.
  - **Inferential statistics is the branch of statistics that involves using a sample to** draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

- **The main use of the descriptive statistics is:**
  - to ‘feel’ the data
  - assess quality of the data

TYPES OF DATA

Constant vs. Variable

- DATA
  - Numerical/non-numerical
    - Qualitative
    - Quantitative
  - Level of measurement
    - Nominal (Color of eyes)
    - Ordinal (Educational level)
    - Interval (Temperature)
    - Ratio (Money)
  - Continuity
    - Continous (Money)
    - Categorical (Education level)

University of Warsaw
Faculty of Economic Sciences

REFERENCE

- SUGGESTED REFERENCE:
  - CHAPTER 1: Larson, R. Farber, E. Farber, B. (2011), Elementary Statistics: Picturing the World, Addison Wesley, 5th Edition.

University of Warsaw
Faculty of Economic Sciences

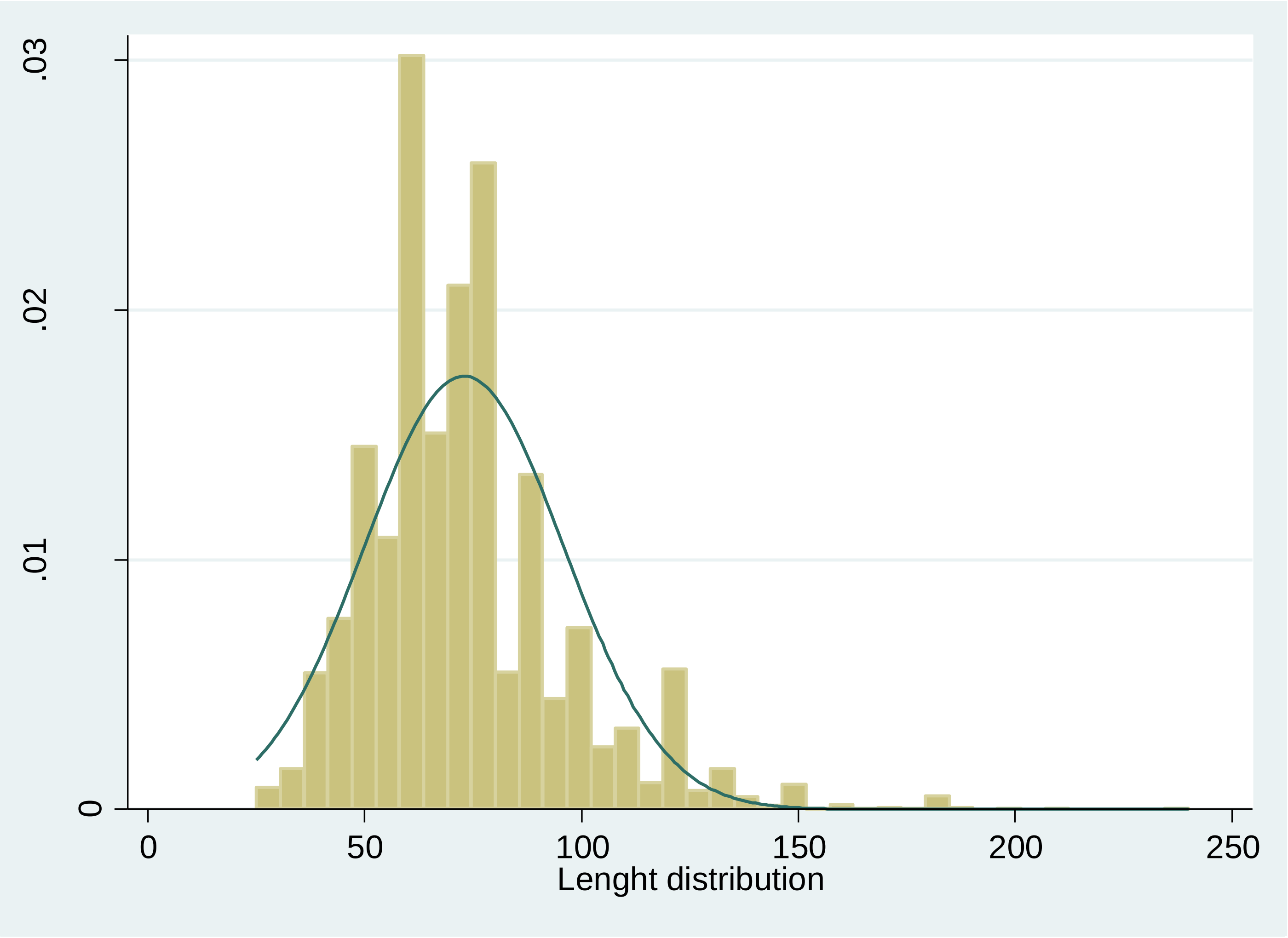
DESCRIPTIVE STATISTICS

DATA TO BE ANALYSED

On the Faraway Stock Exchange 10 shares are listed. Below table presents closing prices on 30/09/2014.

#	Company	Price
1	Abas	103
2	Berton	102
3	Coporin	94
4	Delia	96
5	Ertocon	100
6	Figure	104
7	Gravy	98
8	Hipotonic	105
9	Ixi	93
10	Jot	100

FREQUENCY PLOT (HISTOGRAM)

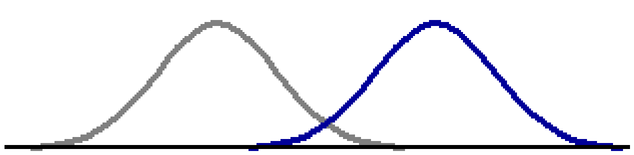


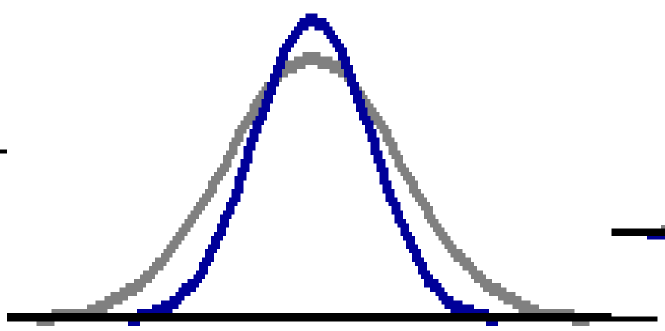
University of Warsaw
Faculty of Economic Sciences

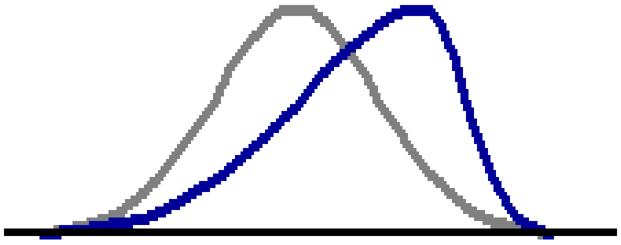
DESCRIPTIVE STATISTICS

descriptive statistics

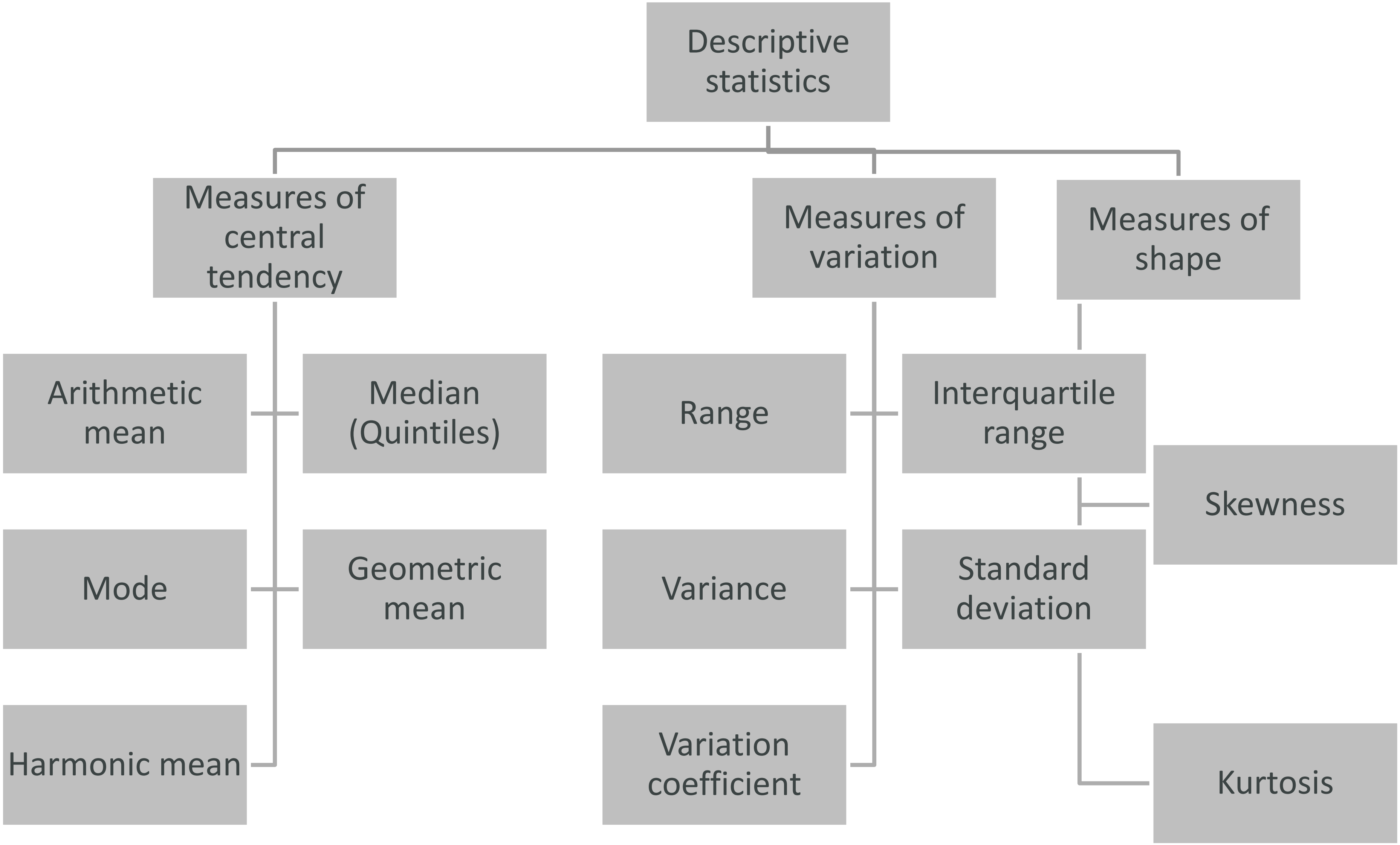
- Measures of central tendency (position)
- Measures of dispersion
- Measures of shape







DESCRIPTIVE STATISTICS - DETAILS



University of Warsaw
Faculty of Economic Sciences

ARITHMETIC MEAN

#	Company	Price
1	Abas	103
2	Berton	102
3	Coporin	94
4	Delia	96
5	Ertocon	100
6	Figure	104
7	Gravy	98
8	Hipotonic	105
9	Ixi	93
10	Jot	100 /1000

observed values

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

sample size

$$\overline{X} = \frac{103+102+94+96+100+104+98+105+93+100}{10} = 99.5$$

OUTLIER SENSITIVITY

$$\overline{X} = \frac{103+102+94+96+100+104+98+105+93+1000}{10} = 189.5$$

WEIGHTED MEAN

#	Company	Price change	Number of assets in portfolio	Weights
1	Abas	5	1000	0.25
2	Berton	7	3000	0.75

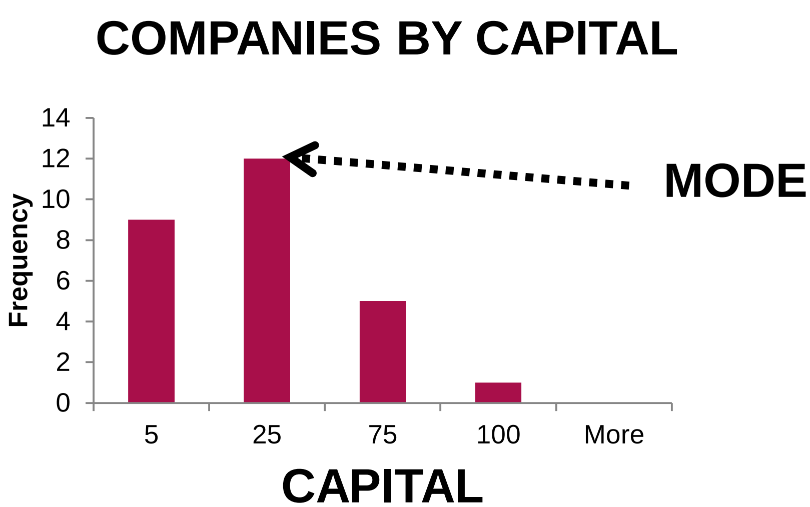
$$\overline{X}_w = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} = \frac{0.25*5+0.75*7}{0.25+0.75} = 1.25+5.25 = 6.5$$

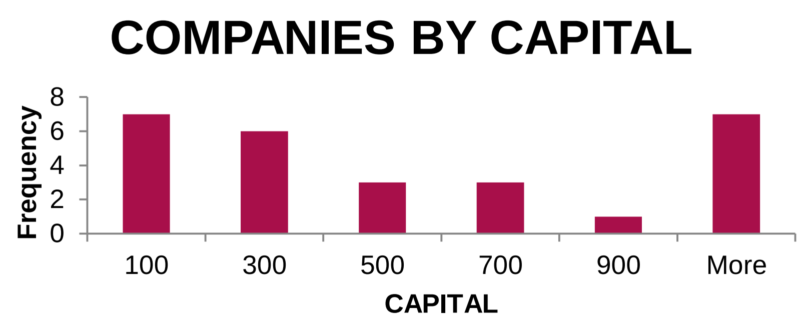
$$\overline{X} = \frac{5+7}{2} = 6$$

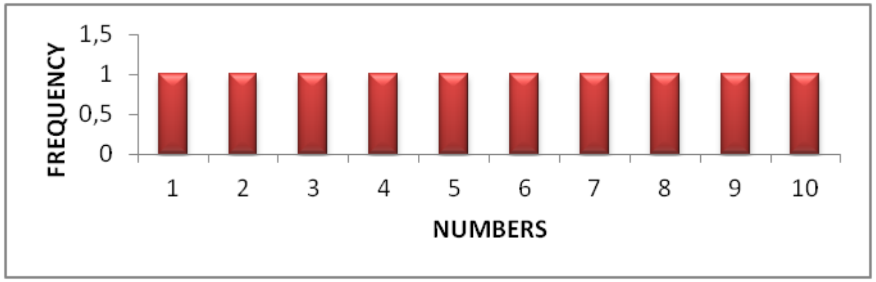
University of Warsaw
Faculty of Economic Sciences

MODE

- It is the most common value in the sample
  - Not influenced by outliers
  - Qualitative and quantitative variables
  - Problems
    - Sometimes mode does not exist;
    - More than one mode







University of Warsaw
Faculty of Economic Sciences

MEDIAN

#	Company	Price
1	Ixi	93
2	Coporin	94
3	Delia	96
4	Gravy	98
5	Ertocon	100
6	Jot	100
7	Berton	102
8	Abas	103
9	Figure	104
10	Hipotonic	105/1000

- Median is the middle value;
  - At least half of the values are greater or equal to and half of the values are smaller or equal than the median;
  - Outliers do not influence median
  - There is always only one median (uniqueness)
- Calculation process:
  - Sorting data in ascending order
  - Calculation of Median position
    - If even number of observation than median is something between two midlle values (mean)
  - Checking the value of the Median

University of Warsaw
Faculty of Economic Sciences

MEDIAN CALCULATION

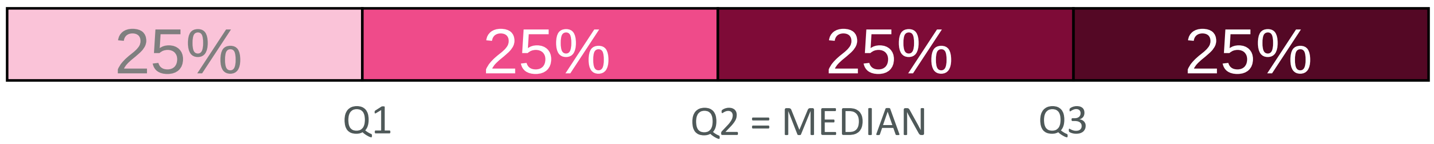
#	Company	Price
1	Ixi	93
2	Coporin	94
3	Delia	96
4	Gravy	98
5	Ertocon	100
6	Jot	100
7	Berton	102
8	Abas	103
9	Figure	104
10	Hipotonic	105/1000

$$median\_position = \frac{n+1}{2} = \frac{10+1}{2} = 5,5$$

MEDIAN

$$median = \frac{100+100}{2} = 100$$

QUARTILES



- Quartiles divide observations into 4 groups (so called quartile groups) which are separated by 3 quartiles (Q1, Q2, Q3).
  - The first (lower) quartile(Q1) is such a value that at least 25% of observations is below or equal to this number.
  - Second quartile (Q2) is the same as median (50% of observations are smaller or equal, 50% are bigger or equal),
  - At most 25% of observations is greater than the third (upper) quartile (Q3);

QUARTILES CALCULATION

#	Company	Price
1	Ixi	93
2	Coporin	94
3	Delia	96
4	Gravy	98
5	Ertocon	100
6	Jot	100
7	Berton	102
8	Abas	103
9	Figure	104
10	Hipotonic	105/1000

$$Q1\_position = \frac{(n+1)}{4} = \frac{(10+1)}{4} = 2.75$$

$$Q2\_position = \frac{2(n+1)}{4} = \frac{2(10+1)}{4} = 5,5$$

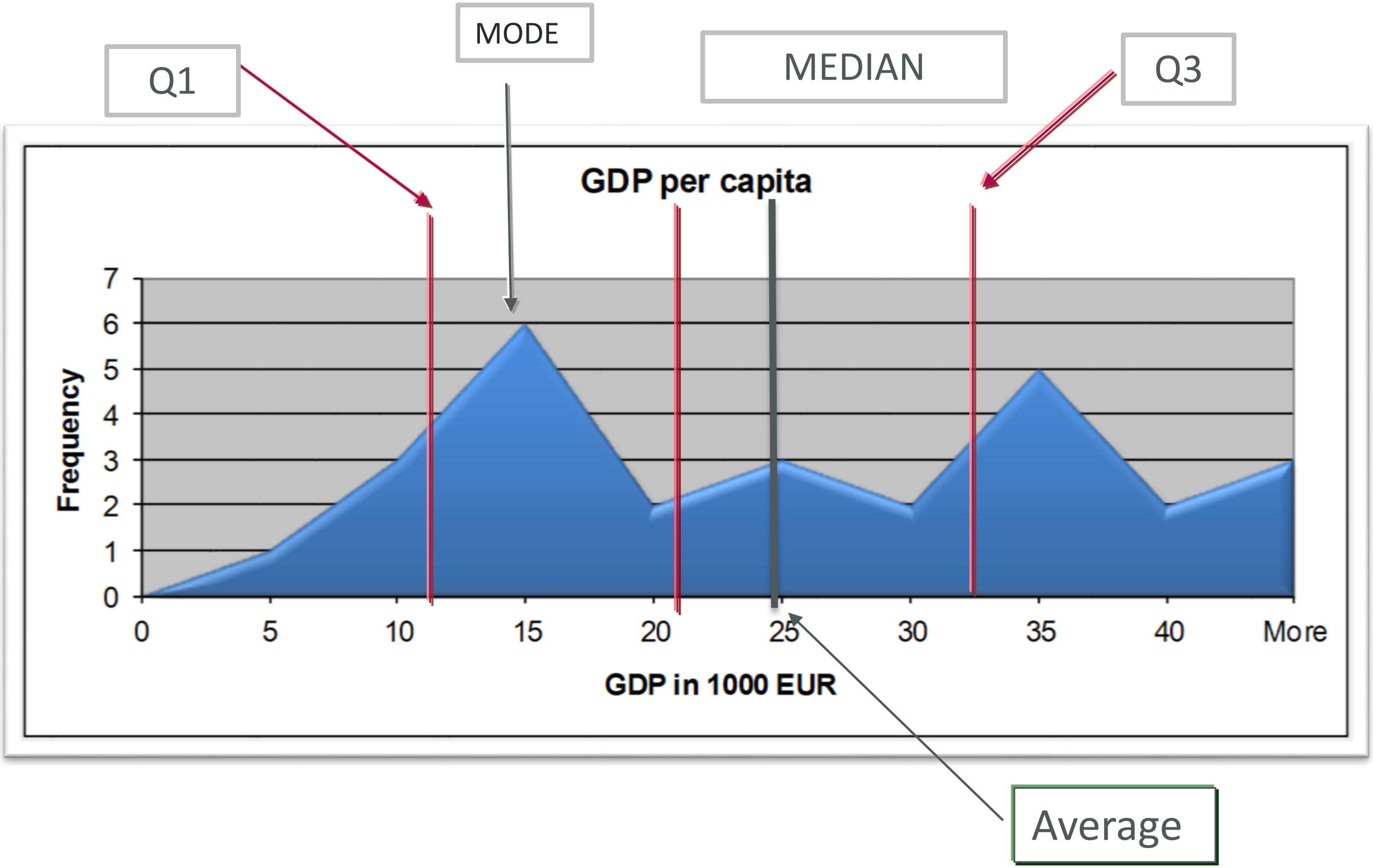
$$Q3\_position = \frac{3(n+1)}{4} = \frac{3(10+1)}{4} = 8,25$$

$$Q1 = \frac{94+96}{2} = 95$$

$$Q2(median) = \frac{100+100}{2} = 100$$

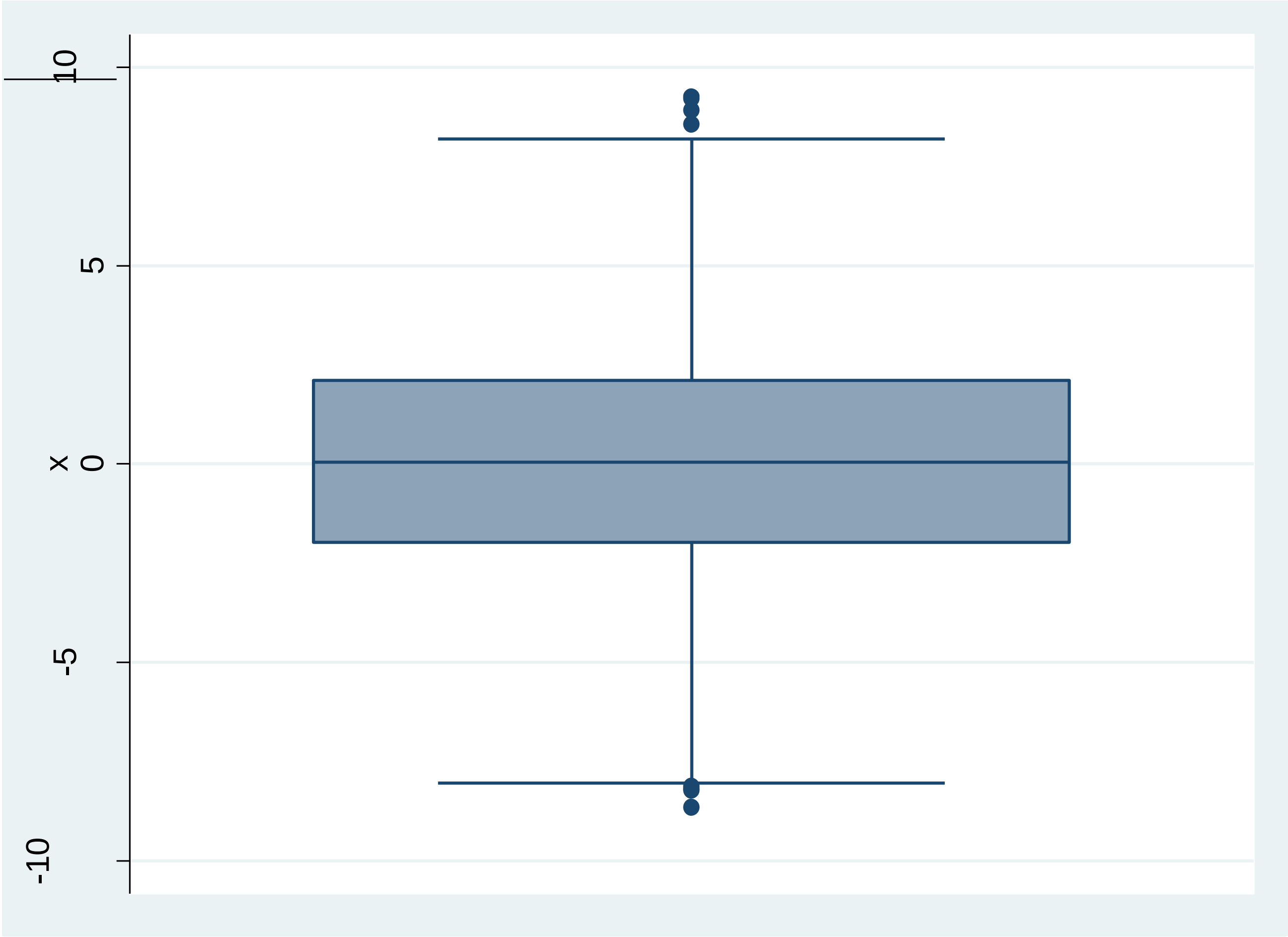
$$Q3 = \frac{103+104}{2} = 103.5$$

DESCRIPTIVE STATISTICS – SUM UP



University of Warsaw
Faculty of Economic Sciences

BOX PLOT



REFERENCE

- SUGGESTED REFERENCE:
  - CHAPTER 3: Ott, R. Longnecker, M. (2010) An Introduction to Statistical Methods and Data Analysis, Cengage Learning , 6th Edition.
  - CHAPTER 2: Larson, R. Farber, E. Farber, B. (2011), Elementary Statistics: Picturing the World, Addison Wesley, 5th Edition.

CATEGORICAL DATA

- What to do if we have information which is not numeric?
- In example we did a survey and ask: How do you like public transportation in Warsaw?
  - Possible answers:
    - Very Good
    - Good
    - So So
    - Bad
    - Very Bad

University of Warsaw
Faculty of Economic Sciences

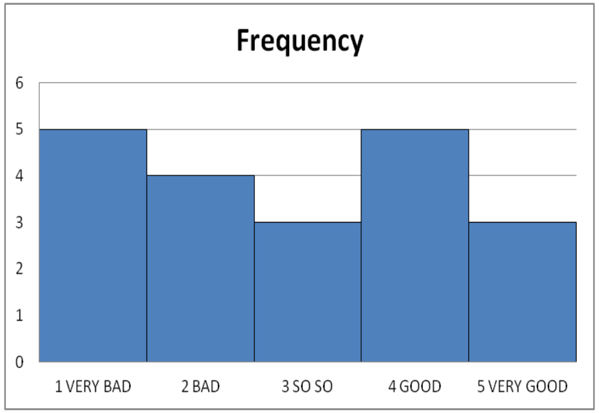
CATEGORICAL DATA

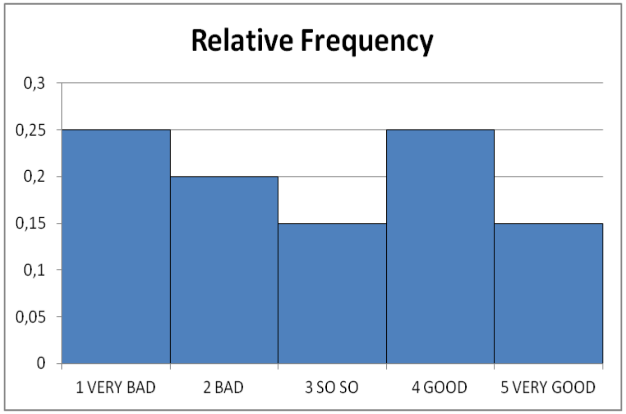
Person	Answers
1	5 VERY GOOD
2	1 VERY BAD
3	4 GOOD
4	4 GOOD
5	1 VERY BAD
6	4 GOOD
7	1 VERY BAD
8	4 GOOD
9	1 VERY BAD
10	5 VERY GOOD
11	3 SO SO
12	1 VERY BAD
13	3 SO SO
14	2 BAD
15	4 GOOD
16	2 BAD
17	2 BAD
18	5 VERY GOOD
19	2 BAD
20	3 SO SO

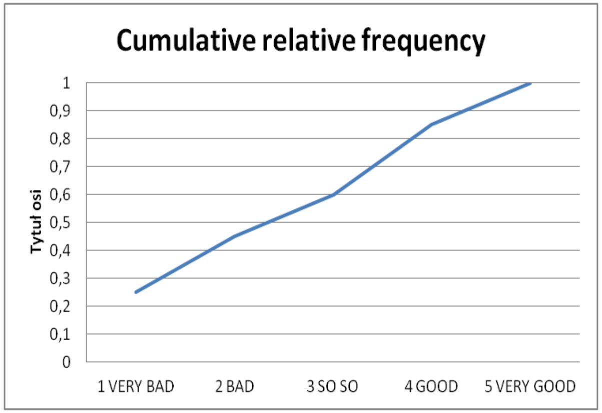
- For the categorical data average may be not the best measure.
- In these cases the frequency tables should be derived (frequency or contingency table)

CATEGORICAL DATA

	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
1 VERY BAD	5	0,25	5	0,25
2 BAD	4	0,2	9	0,45
3 SO SO	3	0,15	12	0,6
4 GOOD	5	0,25	17	0,85
5 VERY GOOD	3	0,15	20	1
ALL ANSWERS	20	1	20	1







CONTINGENCY TABLE

Person	Answers	Gender
1	5 VERY GOOD	1 Men
2	1 VERY BAD	2 Women
3	4 GOOD	1 Men
4	4 GOOD	2 Women
5	1 VERY BAD	1 Men
6	4 GOOD	2 Women
7	1 VERY BAD	1 Men
8	4 GOOD	2 Women
9	1 VERY BAD	1 Men
10	5 VERY GOOD	2 Women
11	3 SO SO	1 Men
12	1 VERY BAD	2 Women
13	3 SO SO	1 Men
14	2 BAD	2 Women
15	4 GOOD	1 Men
16	2 BAD	2 Women
17	2 BAD	1 Men
18	5 VERY GOOD	2 Women
19	2 BAD	1 Men
20	3 SO SO	2 Women

	1 VERY BAD	2 BAD	3 SO SO	4 GOOD	5 VERY GOOD	TOTAL
1 Men	3	2	2	2	1	10
2 Women	2	2	1	3	2	10
TOTAL	5	4	3	5	3	20

Difficult to present on chart, but contingency table is enough.

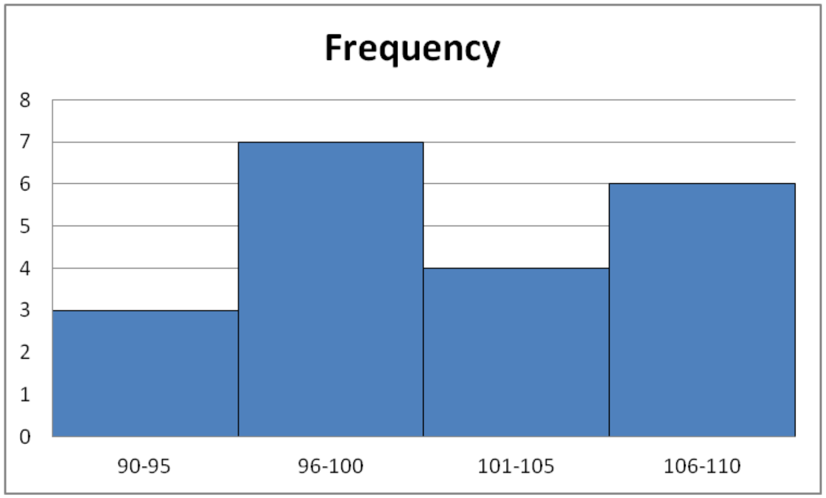
HISTOGRAM

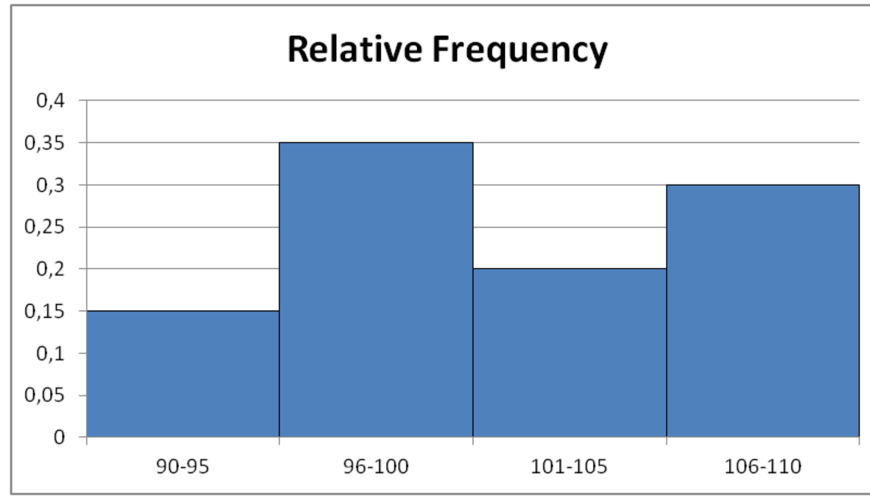
Share	Price
1	100
2	107
3	104
4	108
5	110
6	90
7	102
8	107
9	109
10	96
11	104
12	99
13	100
14	109
15	100
16	96
17	92
18	97
19	93
20	104

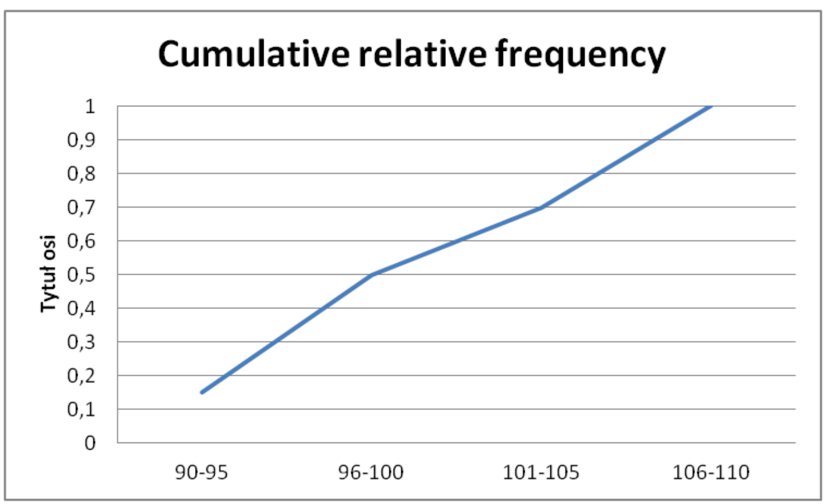
	MAX	MIN	RANGE	WIDTH OF BIN
RANGE	110	90	20	5
NUMBER OF BINS	4			

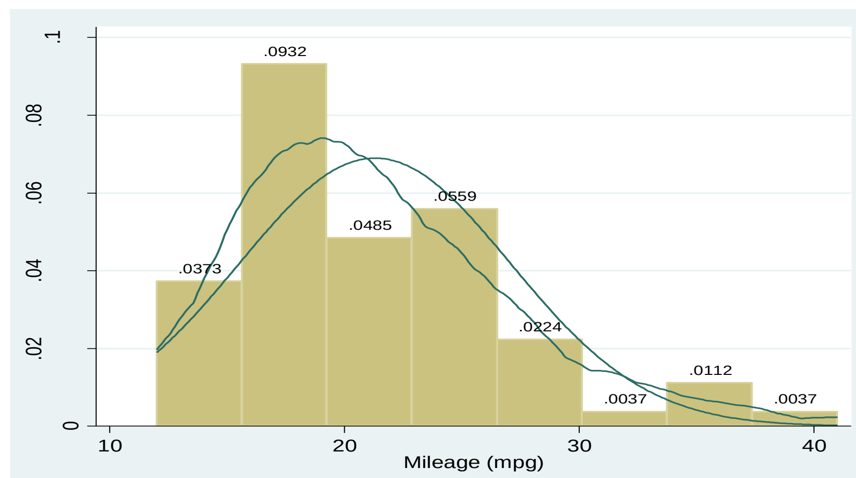
Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency	
90-95	3	0,15	3	0,15
96-100	7	0,35	10	0,5
101-105	4	0,2	14	0,7
106-110	6	0,3	20	1
	20			

HISTOGRAM


Frequency
8
7
6
5
4
3
2
1
0
90-95
96-100
101-105
106-110


Relative Frequency
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
90-95
96-100
101-105
106-110

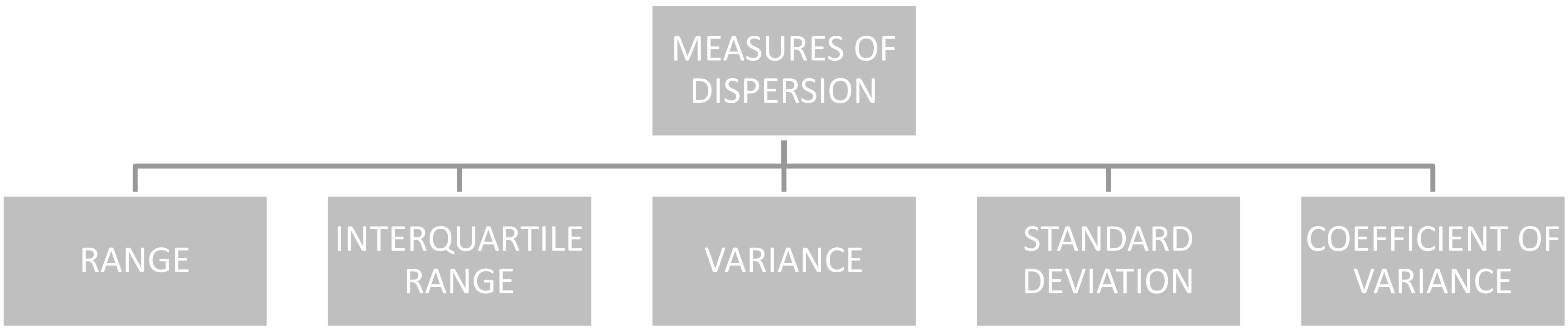

Cumulative relative frequency
Tytuł osi
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
90-95
96-100
101-105
106-110


0
.02
.04
.06
.08
.1
.0373
.0932
.0485
.0559
.0224
.0037
.0112
.0037
10
20
30
40
Mileage (mpg)

REFERENCE

- SUGGESTED REFERENCE:
  - CHAPTER 2.1: Larson, R. Farber, E. Farber, B. (2011), Elementary Statistics: Picturing the World, Addison Wesley, 5th Edition.

MEASURES OF DISPERSION



Statistical dispersion (also called statistical variability or variation) is variability or spread in a variable or a probability distribution

RANGE

#	Company	Price
1	Abas	103
2	Berton	102
3	Coporin	94
4	Delia	96
5	Ertocon	100
6	Figure	104
7	Gravy	98
8	Hipotonic	105
9	Ixi	93
10	Jot	100 /1000

$$RANGE =$$

$$= \max(obs\_val - \max(obs\_val) =$$

$$= 105 - 93 = 12$$

Range imperfections:

- Does not take into account how data distribution (add 5 obs equal to 105 – the same range)
- Sensitive to the presence of atypical observations (outliers)

INTERQUARTILE RANGE

#	Company	Price
1	Ixi	93
2	Coporin	94
3	Delia	96
4	Gravy	98
5	Ertocon	100
6	Jot	100
7	Berton	102
8	Abas	103
9	Figure	104
10	Hipotonic	105/1000

$$Q1\_position = \frac{(n+1)}{4} = \frac{(10+1)}{4} = 2.75$$

$$Q3\_position = \frac{3(n+1)}{4} = \frac{3(10+1)}{4} = 8{,}25$$

$$Q1 = \frac{94+96}{2} = 95$$

$$Q3 = \frac{103+104}{2} = 103.5$$

$$Interquartile\_range = Q3 - Q1 =$$

$$= 8{,}25 - 2{,}75 = 5{,}5$$

University of Warsaw
Faculty of Economic Sciences

VARIANCE

#	Company	Price	Deviation	Deviation Squared
1	Abas	103	3,5	12,25
2	Berton	102	2,5	6,25
3	Coporin	94	-5,5	30,25
4	Delia	96	-3,5	12,25
5	Ertocon	100	0,5	0,25
6	Figure	104	4,5	20,25
7	Gravy	98	-1,5	2,25
8	Hipotonic	105	5,5	30,25
9	Ixi	93	-6,5	42,25
10	Jot	100	0,5	0,25
	Mean	99,5	Variance	15,65
			Variance_pop	15,65
			Variance_sam	17,39

$$\overline{X} = 99{,}5$$

Population variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}$$

Sample variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

STANDARD DEVIATION

Mean	99,5	Variance	15,65	Standard Deviation	3,96
		Variance Pop	15,65	Standard Deviation Pop	3,956
		Variance Sample	17,39	Standard Deviation Sample	4,170

Population variance

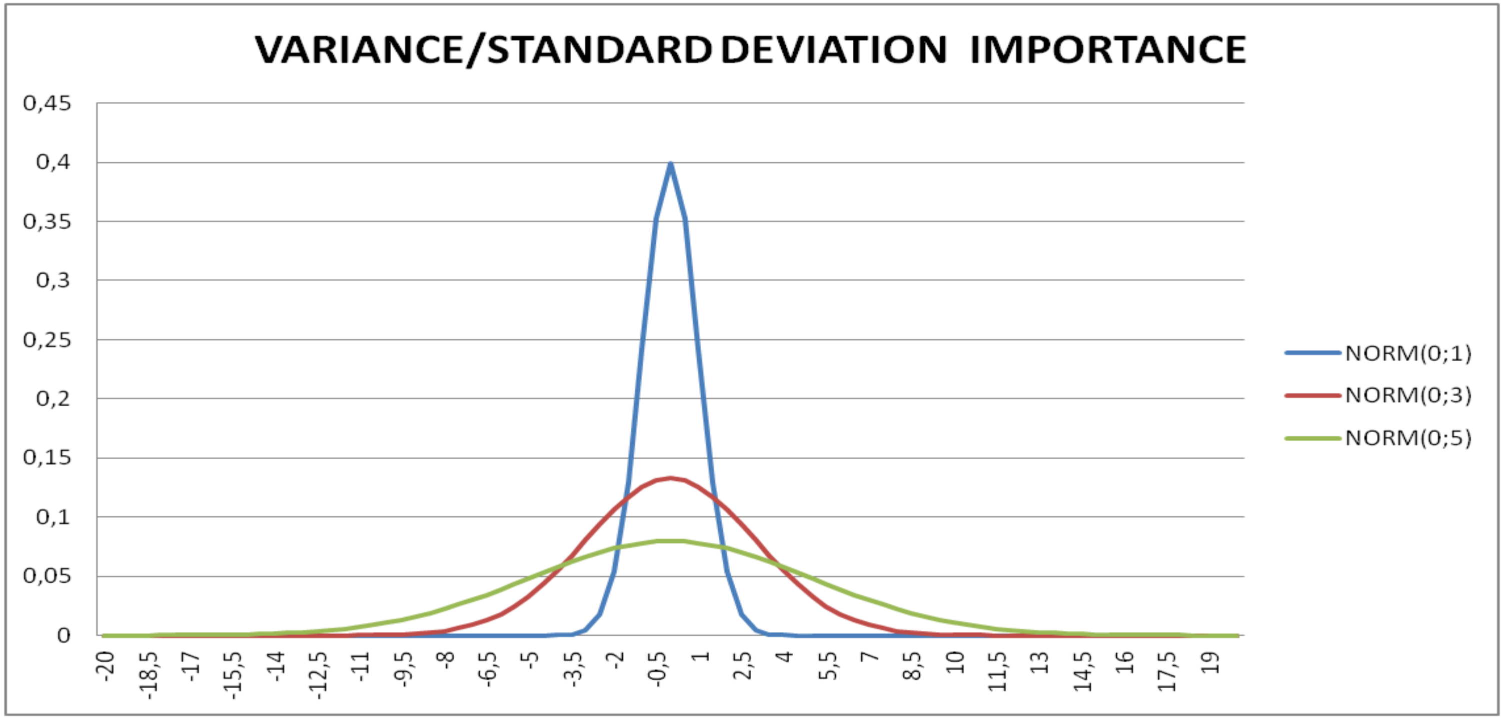
$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}}$$

Sample variance

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1}}$$

MAIN ADVANTAGE: THE SAME UNIT AS ANALYSED VARIABLE

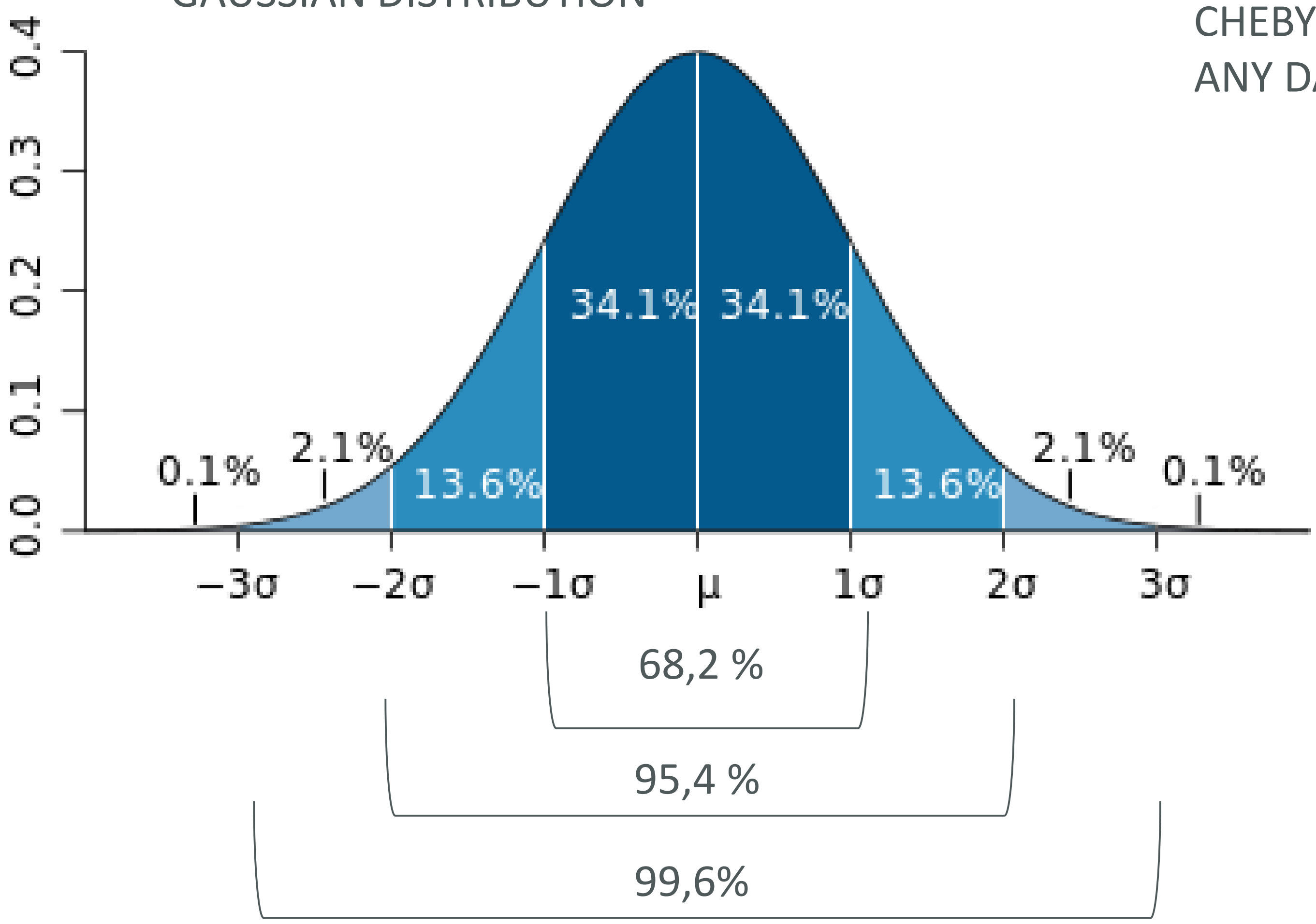
IMPORTANCE OF VARIANCE AND STANDARD DEVIATION



- All data are taken into account in these calculations
- Values with higher distance from the mean have bigger impact to variance (due to square operation)

3 – SIGMA RULE (EMPIRICAL RULE)

GAUSSIAN DISTRIBUTION



CHEBYCHEV TEOREM (FOR k>1)
ANY DATA SET

$$1 - \frac{1}{k^2}$$

k	% OF DATA
2	75,00%
3	88,89%
4	93,75%
5	96,00%
6	97,22%
7	97,96%
8	98,44%
9	98,77%
10	99,00%

http://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg

University of Warsaw
Faculty of Economic Sciences

OUTLIERS – Z-SCORE

#	Company	Price	Deviation	Deviation Squared	Z-score	Z>2
1	Abas	103	3,5	12,25	0,884730244	0
2	Berton	102	2,5	6,25	0,631950174	0
3	Coporin	94	-5,5	30,25	-1,390290383	0
4	Delia	96	-3,5	12,25	-0,884730244	0
5	Ertocon	100	0,5	0,25	0,126390035	0
6	Figure	104	4,5	20,25	1,137510313	0
7	Gravy	98	-1,5	2,25	-0,379170104	0
8	Hipotonic	105	5,5	30,25	1,390290383	0
9	Ixi	93	-6,5	42,25	-1,643070452	0
10	Jot	100/ 1000	0,5	0,25	0,126390035	0
	Mean	99,5	Variance	15,65	Standard Deviation	3,956008089
			Variance Pop	15,65	Standard Deviation Pop	3,956
			Variance Sample	17,39	Standard Deviation Sample	4,170
					3-sigma value	87,63197573
						111,3680243

$$Z = \frac{X - \overline{X}}{S}$$

OUTLIER:

IF Z>2

IF Z>3

COEFFICIENT OF VARIANCE

TIME	Company	Price	Company	Price
1	Abas	103	Berton	53
2	Abas	102	Berton	52
3	Abas	94	Berton	44
4	Abas	96	Berton	46
5	Abas	100	Berton	50
6	Abas	104	Berton	54
7	Abas	98	Berton	48
8	Abas	105	Berton	55
9	Abas	93	Berton	43
10	Abas	100	Berton	50
	STD	3,96		3,96
	MEAN	99,5		49,5
	COEFFICIENT OF VARIANCE	3,98%		7,99%

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Standard deviation of data must always be understood in the context of the mean

The coefficient of variation is a dimensionless number.

MEASURES OF SHAPE

MEASURES OF SHAPE

- SKEWNESS
- KURTOSIS

MOMENTS

- Except central tendency & dispersion: shape of the distribution can be considered.
  - In order to be able to find it, we should first introduce the concept of moments.
- We can distinguish the following moments:
  - ordinary/raw moments
  - central moments

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

$$M_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^k$$

SKEWNESS

- For symmetric distributions, all central moments of odd orders are equal to 0;
- Coefficient of asymmetry (skewness) - third standardized moment

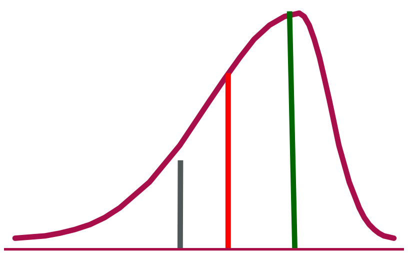
$$\rho_{asym} = \frac{M_3}{s^3}$$

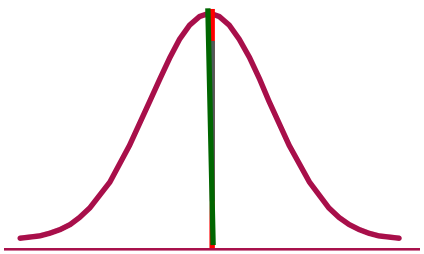
$$\hat{\rho}_{asym} = \frac{M_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}$$

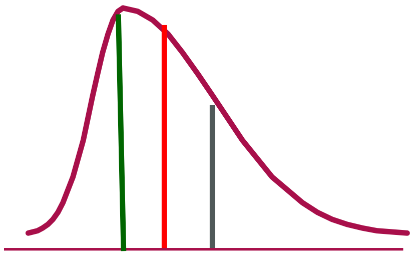
University of Warsaw
Faculty of Economic Sciences

SKEWNESS - ASSYMETRY

SKEWNESS < 0	SKEWNESS = 0	SKEWNESS > 0
NEGATIVE SKEW		POSITIVE SKEW
LEFT-SKEWED	SYMETRIC	*RIGHT-SKEWED*
mean< median < mode	mean = median = mode	mean > median > mode







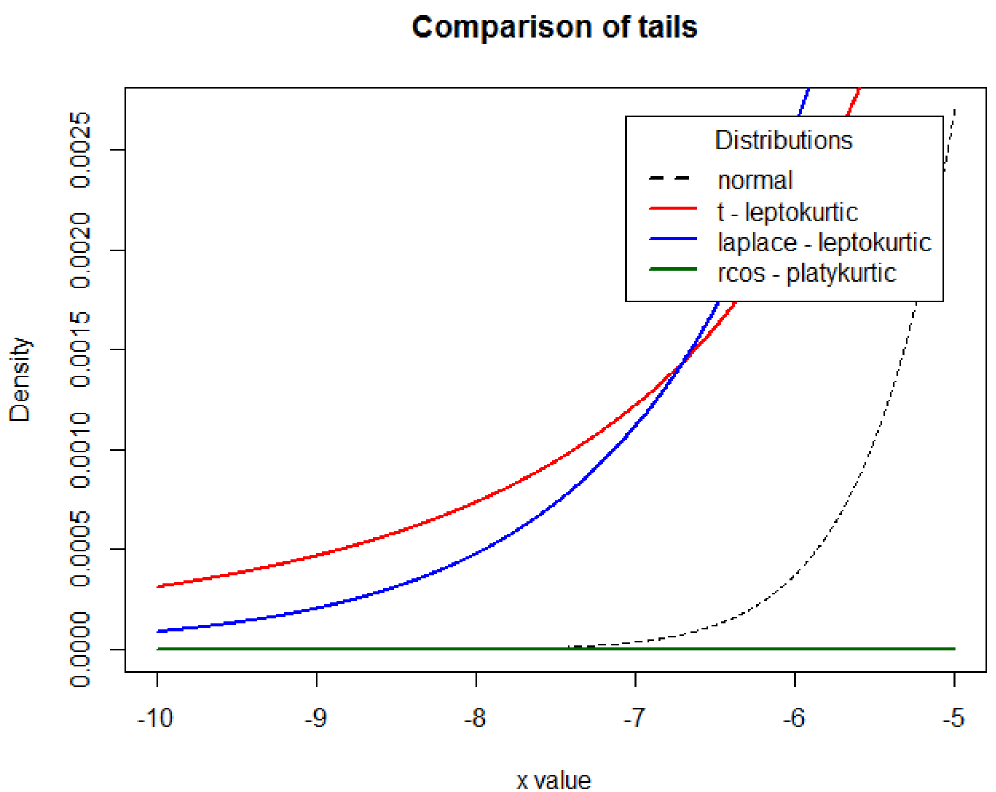
TAIL DISTRIBUTION - KURTOSIS

- We say that the distribution has a fat tail, if a large part of the mass is in the tail.
- If the distribution has a thick tail, we can more likely expect outliers.
- Kurtosis - describes the thickness of the "tail" distribution, i.e. the probability of observation very distant from the average;

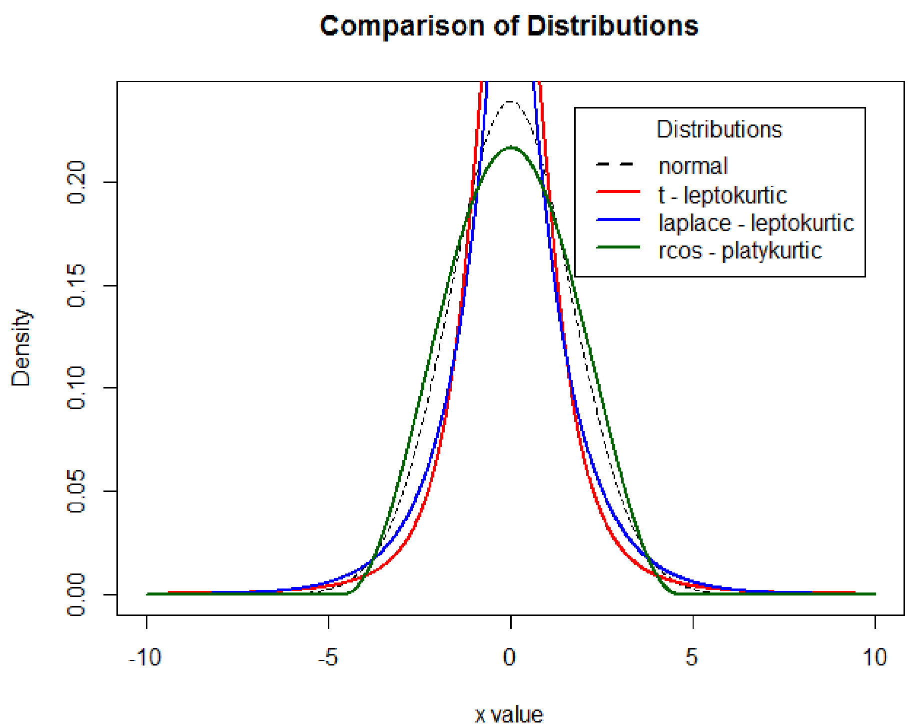
$$\hat{\rho}_{kurtosis} = \frac{M_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

$$\hat{\rho}_{excess\_kurtosis} = \frac{M_4}{s^4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3$$

KURTOSIS



Comparison of tails
Distributions
normal
t - leptokurtic
laplace - leptokurtic
rcos - platykurtic
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
-10
-9
-8
-7
-6
-5
x value



Comparison of Distributions
Distributions
normal
t - leptokurtic
laplace - leptokurtic
rcos - platykurtic
Density
0.00
0.05
0.10
0.15
0.20
-10
-5
0
5
10
x value

University of Warsaw
Faculty of Economic Sciences

PROBABILITY, RANDOM VARIABLES, R.V. DISTRIBUTIONS

University of Warsaw
Faculty of Economic Sciences

Probability definitions

- Classical interpretation of probability
  - Each possible distinct = outcome;
  - Event = collection of outcomes.

$$P(E) = \frac{N_E}{N}$$

- Relative frequency concept of probability;
  - Empirical approach to probability.

$$P(E) = \frac{n_E}{n}$$

  - Repetition of experiment n times (a lot of times)
    - Law of large numbers: $\lim_{n \to \infty} P(\hat{E}) = P(E)$
- Personal or subjective probability

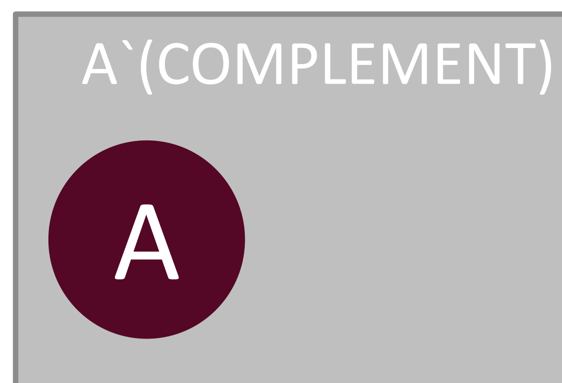
Mutual exclusive events properties

- For mutually exclusive events:

$$0 \le P(A) \le 1$$

$$P(A \cup B) = P(A) + P(B)$$

$$P(A) + P(A`) = 1$$



Conditional vs. unconditional probability

- Probability of event A given event B

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

- Probability of intersection

$$P(A\cap B)=P(A)P(B|A)=P(B)P(A|B)$$

- Independency of events

$$P(A|B)=\frac{P(A\cap B)}{P(B)}=P(A) \quad P(B|A)=\frac{P(A\cap B)}{P(A)}=P(B)$$

- Probability of union

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

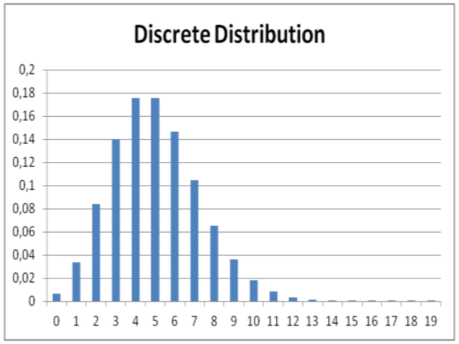
University of Warsaw
Faculty of Economic Sciences

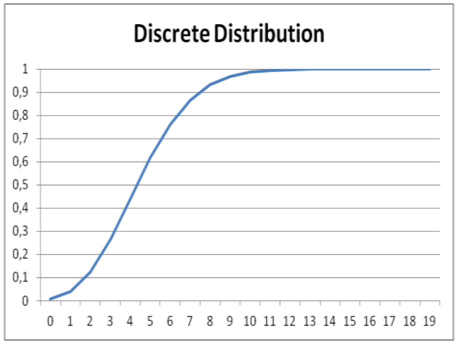
RANDOM VARIABLE

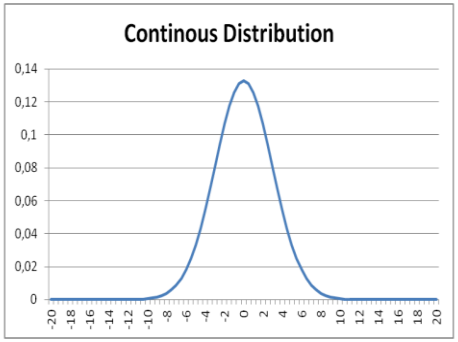
- Random variable X
  - variable that can take a set of possible different values each with an associated probability
  - Realization of random variable X (x_i) is a possible outcome of an probability experiment.
- **Discrete random variable:** countable number of possible outcomes
- **Continuous random variable :** uncountable number of possible outcomes

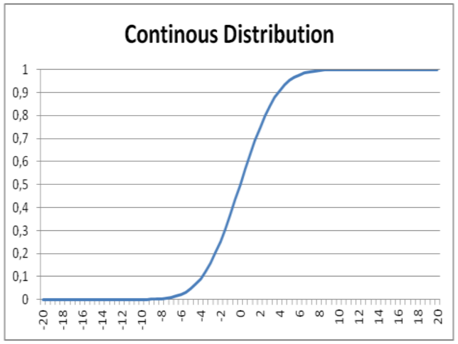
PDF & CDF

- *Relative frequencies of outcomes* generate a distribution the **probability distribution of *RV.***
  - ***Probability distributions differ for discrete and continuous random***









University of Warsaw
Faculty of Economic Sciences

DISCRETE PROBABILITY DISTRIBUTION

- DPD must satisfy following conditions

$$0 \leq P(x_i) \leq 1, \forall x_i$$

$$\sum P(x_i) = 1$$

- Property of DPD

$$P(x_i \cup x_j) = P(x_i) + P(x_j)$$

- Discrete RV measures

$$\mu = \sum xP(x) \qquad \sigma^2 = \sum (x-\mu)^2 P(x)$$

University of Warsaw
Faculty of Economic Sciences

BINOMIAL DISTRIBUTION

- Probability of exactly x number of successes in a sequence of N independent yes/no experiments, each of which yields success with probability p
- Random variable is derived from the set of natural numbers (including 0)

$$P(x;N,p)=\frac{N!}{x!(N-x)!}p^x(1-p)^{N-x}$$

$$F(x;N,p)=\sum_{i=0}^{x}\frac{N!}{x_i!(N-x_i)!}p^{x_i}(1-p)^{N-x_i}$$

BINOMIAL DISTRIBUTION

- Properties: $\mu = Np \quad \sigma^2 = Np(1-p) \quad \mu \geq \sigma^2$
  - Sum of 2 binomial rv with p:

$$X \to B(N, p), Y \to B(M, p)$$

$$Z = X + Y \to B(N + M, p)$$

  - For large N:

$$B(N, p) \approx Poisson(Np)$$

  - For N→∞, p→0 and constant Np=λ

$$B(N, p) \approx N(Np, Np(1-p))$$

POISSON DISTRIBUTION

- Probability of **x** events occurrence within one unit of time (month, year)
- Random variable is derived from the set of natural numbers (including 0)
- Assumptions:
  - Mean value of number of events within one unit of time is constant and equal to λ,
  - Probability of event occurrence is independent form time that last form last occurrence

$$P(x;\lambda)=\frac{\lambda^x e^{-\lambda}}{x!} \qquad F(x;\lambda)=e^{-\lambda}\sum_{i=0}^{x}\frac{\lambda^i}{i!}$$

- Properties: $\mu=\sigma^2=\lambda$

$$Z=\sum_{i=1}^{n}X\rightarrow Poisson(\lambda)\Leftrightarrow Z\rightarrow Poisson(\sum_{i=1}^{n}\lambda_i)$$

NEGATIVE BINOMIAL DISTRIBUTION

- In each trial the probability of success is p and of failure is (1 – p). We are observing a sequence of trials until a predefined number y of failures has occurred. Then the random number of successes we have seen, x, will have the negative binomial distribution
- Variable from the set of Natural numbers. (without 0)
- Properties:

$$f(k; r, p) = P(X = k) = \binom{k+r-1}{k} p^k (1-p)^r$$

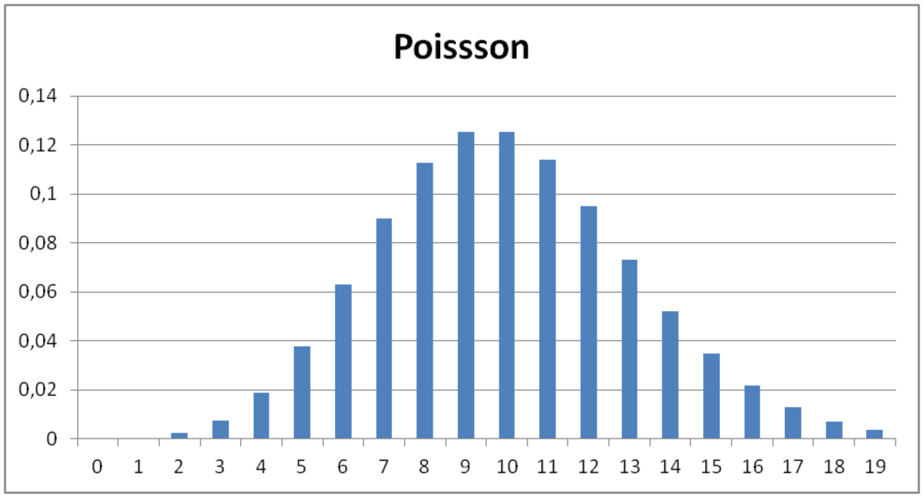
$$\mu \leq \sigma^2$$

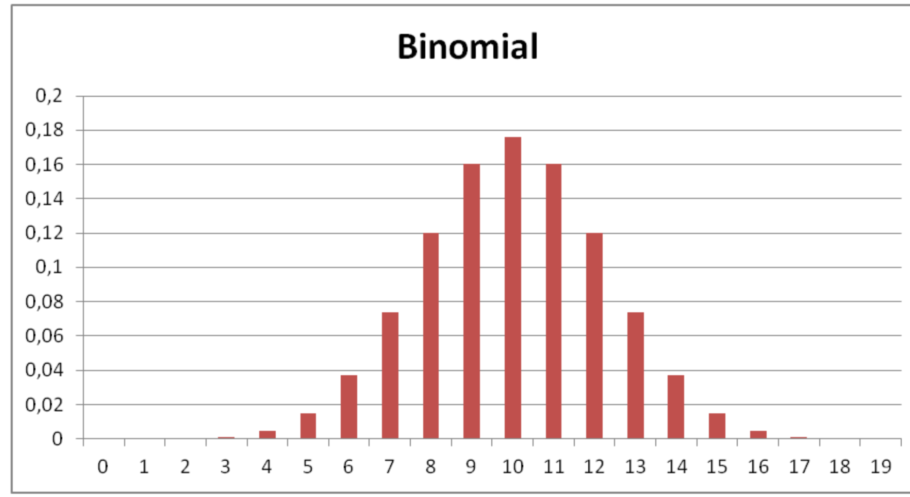
- When r→∞, p→0 and constant $r\left(\frac{p}{1-p}\right)$

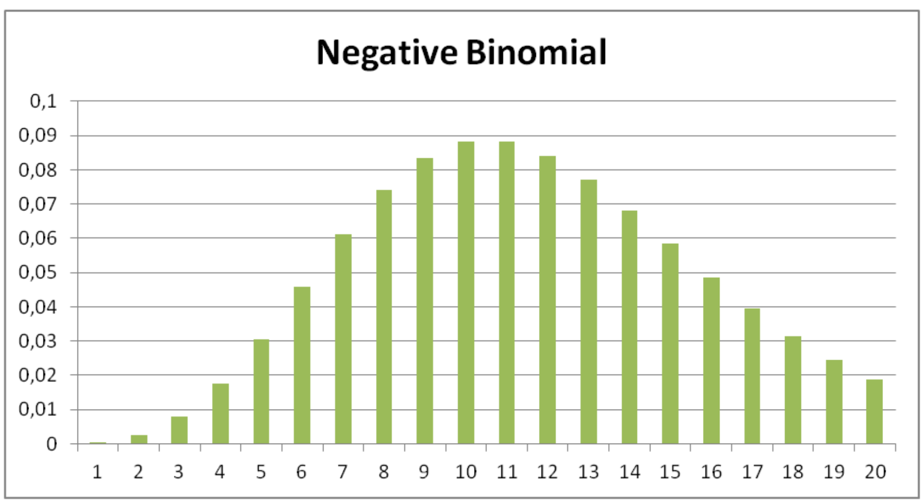
$$X \to NB(r, p) \approx Poisson\left(r\frac{p}{1-p}\right)$$

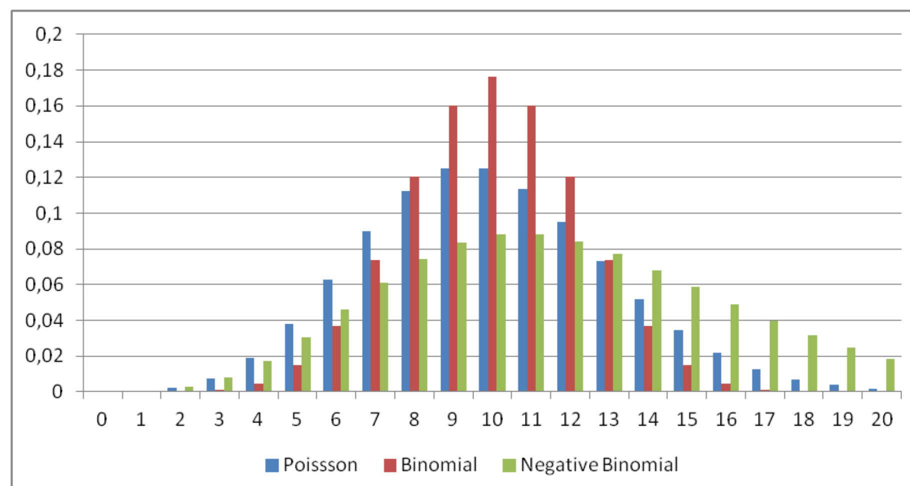
University of Warsaw
Faculty of Economic Sciences

DISCRETE PDF COMPARISON









University of Warsaw
Faculty of Economic Sciences

GEOMETRIC DISTRIBUTION

- Geometric distribution determines probability of first success occurrence in *k-th* trial when probability of success is *p* (Alternatively: probability of k failures before first success)

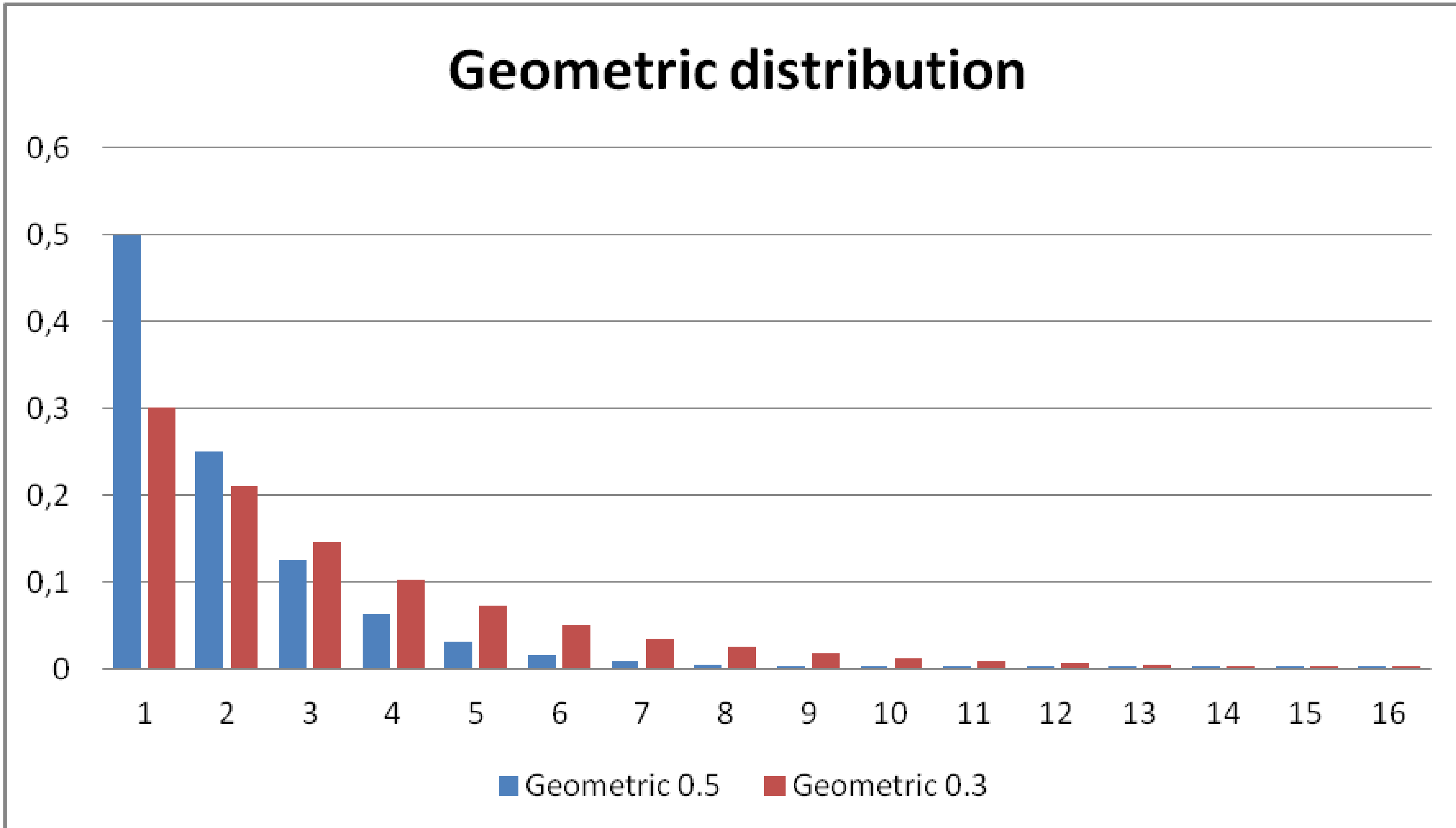
$$P(k;p) = (1-p)^{k-1}p \qquad F(k;p) = \sum_{i=1}^{k}(1-p)^{i-1}p$$

- Properties: $\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2}$
  - Geometric function is discrete equivalent of exponential distribution.
  - Geometric function is memory-less. It means that conditional probability of the first success occurrence at moment *k+t do not depend on number t* trials made before. $P(k+t \mid t; p) = P(k;p)$
  - Sum of *r* r.v. from geometric distribution with *p* probability of success comes form negative binomial distribution with *r & p* parameters

$$Z = \sum_{i=1}^{n} X \to Geometric(1-p) \Leftrightarrow Z \to NegBin(r,p)$$

University of Warsaw
Faculty of Economic Sciences

GEOMETRIC DISTRIBUTION



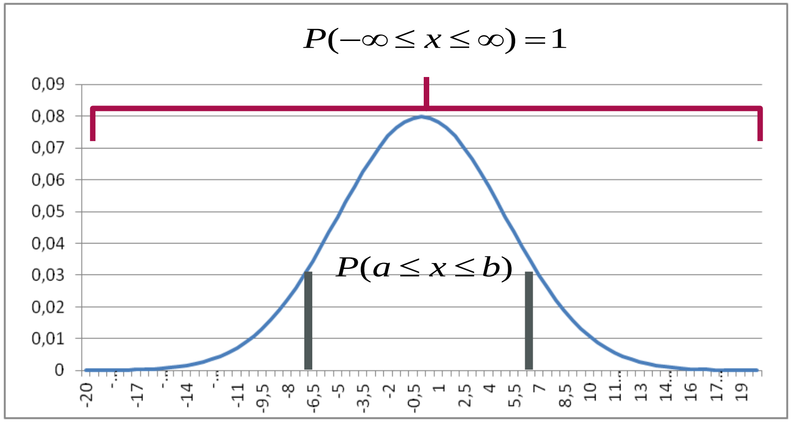
CONTINUOUS DISTRIBUTION

- Continuous variables → possible values form the whole interval or range (i.e. dollar amount of return from some investment)
- Infinitely number of possible outcomes
- Assumptions:

$$P(x = x_i) = f(x = x_i) = 0, \forall x$$

$$\int f(x) = 1$$

$$P(a \le x \le b) = \int_a^b f(x)$$



NORMAL (GAUSSIAN) DISTRIBUTION

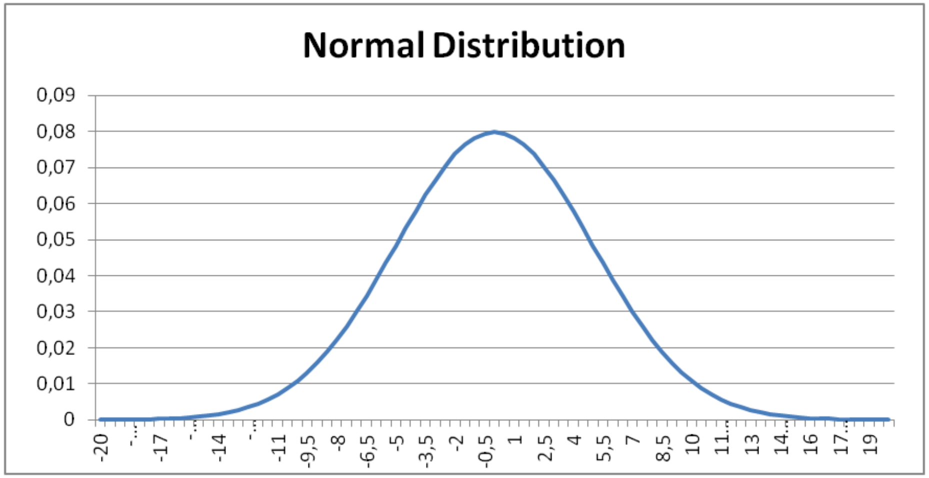
- In probability theory, the normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution
  - Very useful because of central limit theorem
  - PDF & CDF

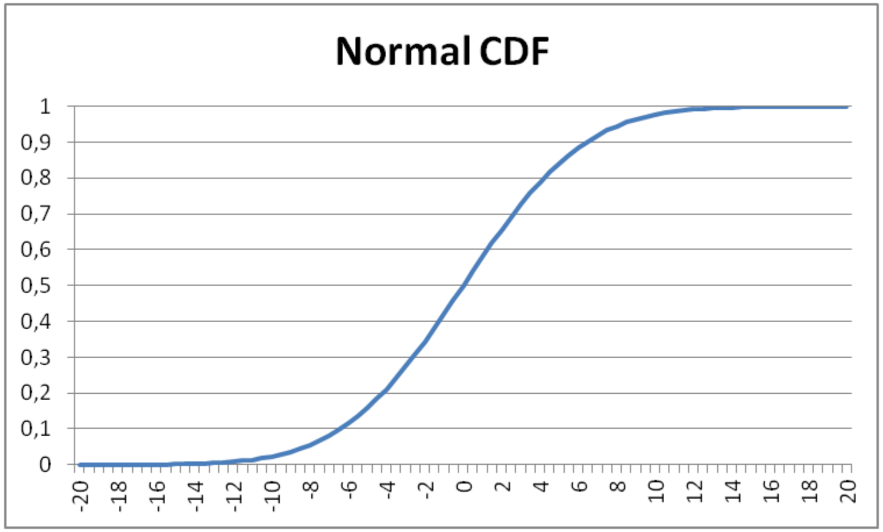
$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \Phi(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$

- Properties:
  - Mean & Variance finite
  - Mean = Median = Mode
  - Skewness = 0, Kurtosis =3, Ex. Kurtosis = 0
  - Unimodal
- Standardization:

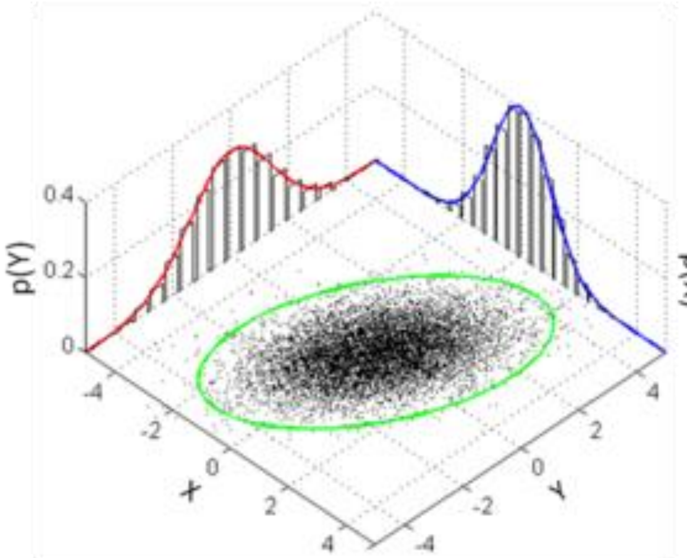
$$X\rightarrow N(\mu,\sigma) \qquad Z=\frac{X-\mu}{\sigma}\rightarrow N(0,1)$$

NORMAL (GAUSSIAN) DISTRIBUTION





MULTIVARIATE



NORMAL DISTRIBUTION & OTHERS

- If X_1, X_2, …, X_n are independent normal random variables with μ and σ:

$$X_1 + \ldots + X_n \rightarrow N(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$$

- If X_1, X_2, …, X_n are independent standard normal random variables:

$$X_1^2 + \ldots + X_n^2 \rightarrow \mathrm{X}_n^2$$

- If X_1, X_2, …, X_n are independent normal random variables with μ and σ:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}} \rightarrow t_{n-1}$$

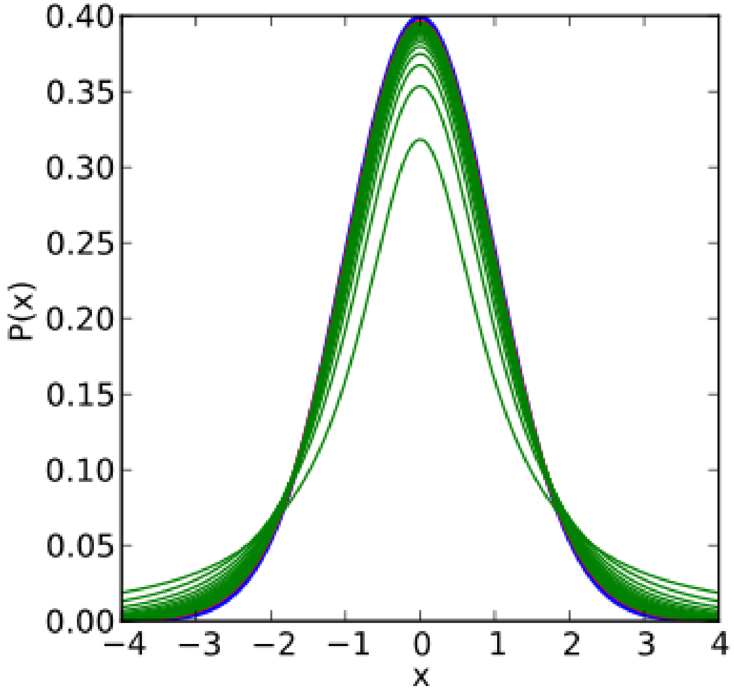
- If X_1, X_2, …, X_n & Y_1, Y_2, …, Y_n , independent standard normal random variables:

$$F = \frac{(X_1^2 + X_2^2 + \ldots + X_n^2)}{(Y_1^2 + Y_2^2 + \ldots + Y_m^2)} \rightarrow F_{n,m}$$

STUDENT'S T DISTRIBUTION

- Student's t-distribution: family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

$$t_{\infty} = N(0,1)$$

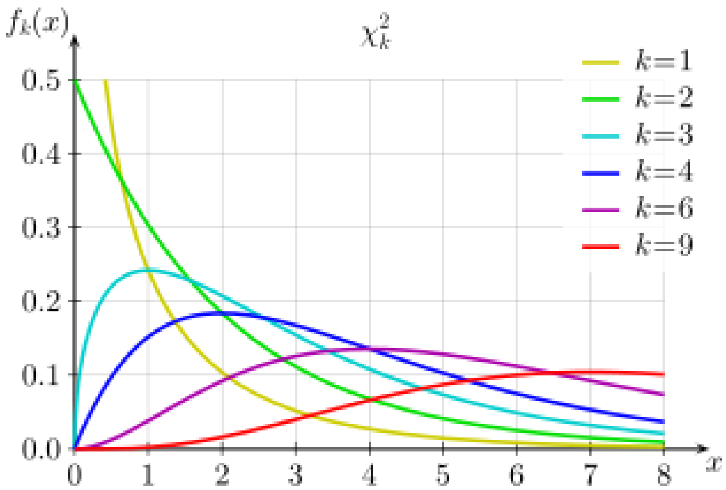


Source: Wikipedia

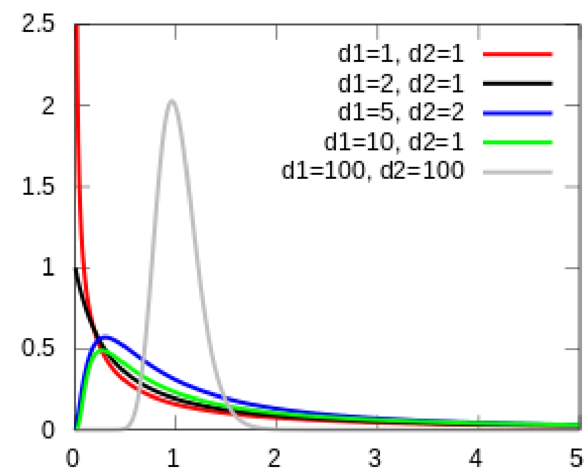
CHI² & F DISTRIBUTIONS

University of Warsaw
Faculty of Economic Sciences

- Chi² distribution: one of the most widely used probability distributions in inferential statistics (goodness of fit test, independence etc.)



- F distribution: used in statistics inference (analysis of variance, significance test)



Source: Wikipedia

CENTRAL LIMIT THEOREM

- Sampling distribution → PDF of a sample statistics that is formed when samples of size n are repeatedly taken from a population
  - Standard deviation of the sampling distribution of the sample means is called the standard error of the means

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- If samples of size n (n>=30) are drawn from any population with a μ & σ then sampling distribution of sample means approx Normal Distribution:

$$\bar{X} \to N(\mu_{\bar{x}} = \mu, \sigma^2_{\bar{x}} = \frac{\sigma^2}{n})$$

LOG NORMAL DISTRIBUTION

- Continuous variable in the range (0, ∞).

$$X \rightarrow LogNormal \Leftrightarrow Y = \ln(X) \rightarrow Normal$$

$$Y \rightarrow Normal \Leftrightarrow e^{Y} \rightarrow LogNormal$$

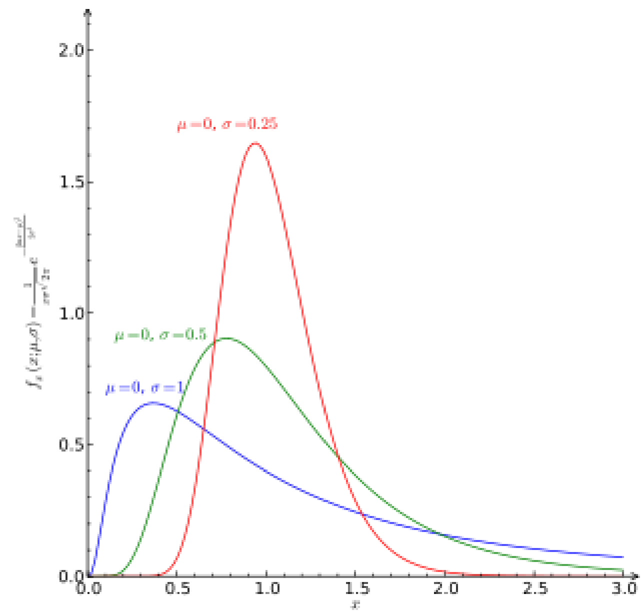
$$f(x;\mu;\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(\frac{(\ln x - \mu)^2}{2\sigma^2} \right)$$

$$F(x;\mu;\sigma) = \Phi\left(\frac{\ln x - \mu}{\sigma} \right)$$

$$E[X] = e^{\mu + \frac{1}{2}\sigma^2}$$

$$Var[X] = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}$$

- Product of many independent random variables from the same distribution with finite mean and variance have log-normal distribution (analogy to CLT).



Source: Wikipedia

GAMMA DISTRIBUTION

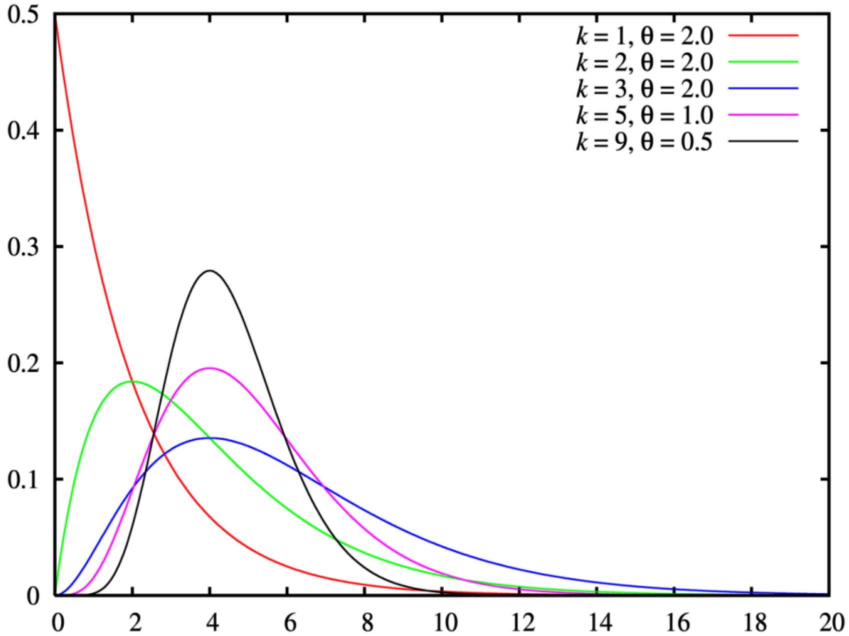
- Continuous variable *x>0*
- Shape parameter *k>0*
- Scaling parameter *Θ>0*
- *Special cases:*

$$X \to Gamma(1,\lambda) \Leftrightarrow X \to Exp(\lambda)$$

$$X \to Gamma(v/2,2) \Leftrightarrow X \to \mathrm{X}^2(v)$$

$$f(x;k;\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}$$

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$



University of Warsaw
Faculty of Economic Sciences

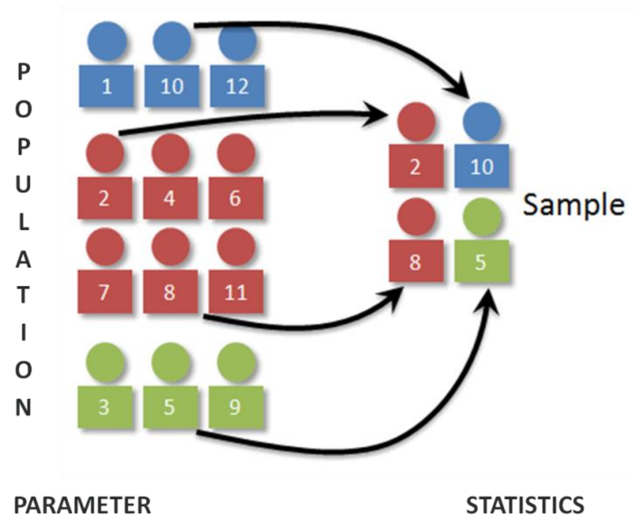
STATISTICAL INFERENCE

Inference
- Historical information
- Scenarios / Expert judgment
- Upcoming changes

Prediction
- Abundance of data
- Inconsistency
- Overwhelming

Decision
- Correct or not, the most probably the best decision

STATISTICAL INFERENCE



http://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

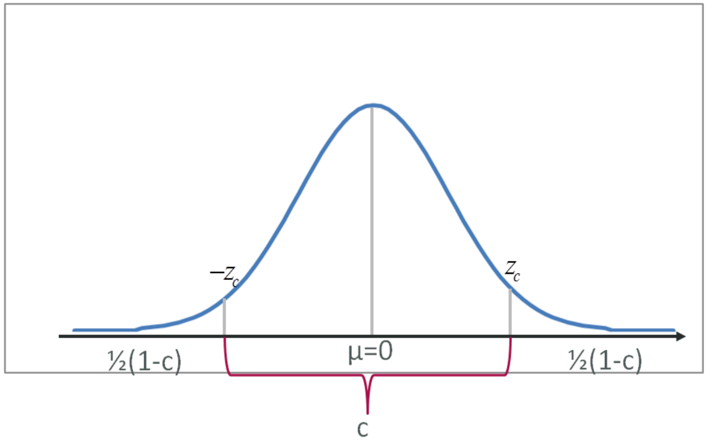
POPULATION PARAMETERS:
- μ – mean
- M - median
- σ – standard deviation
- π - proportion

INFERENCE ABOUT PARAMETERS:
- Estimation
- Hypothesis testing about parameter value

University of Warsaw
Faculty of Economic Sciences

POINT ESTIMATE & CONFIDENCE INTERVAL

- Point estimate
  - Single value estimate for a population parameter(for example sample mean)
- Interval estimate
  - Interval (range) of possible values of an unknown population parameter
- Level of confidence
  - The probability that the interval estimate contains the population parameter
- Margin of error E for given c
  - The greatest possible distance between the point estimate and the value of the parameter
  - For normal distribution (assumption σ, but for n>=30 sample std *may* be used)

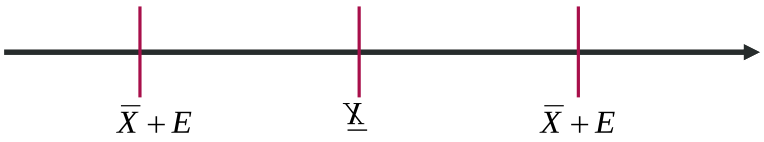


$$E = z_c \frac{\sigma}{\sqrt{n}}$$

$$n = \left(\frac{z_c \sigma}{E} \right)^2$$

c-CONFIDENCE INTERVAL FOR μ

$$\overline{X} - E < \mu < \overline{X} + E$$



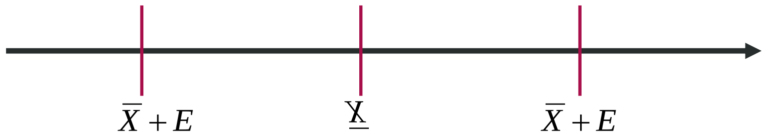
University of Warsaw
Faculty of Economic Sciences

CONFIDENCE INTERVALS FOR MEAN FOR SMALL SAMPLE SIZE AND

- When distribution of a random variable is approximately normal, but the sample size n is smaller then 30, instead of statistics z, statistics t may be calculated, which comes from t-distribution with n-1 degrees of freedom

c-CONFIDENCE INTERVAL FOR μ

$$\bar{X} - E < \mu < \bar{X} + E$$



$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \to t_{n-1}$$

$$E = t_{c,n-1} \frac{s}{\sqrt{n}}$$

CONFIDENCE INTERVALS FOR PROPORTION

- Point estimate for p (X/N)
- If binomial distribution may be approximated by normal distribution (np.>=5 and nq>=5)

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\hat{q}}{n}}} \to N(0,1)$$

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

University of Warsaw
Faculty of Economic Sciences

CONVIDENCE INTERVALS FOR VARIANCE AND STD

- Statistics

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \rightarrow X^2_{n-1}$$

- Confidence interval for variance

$$\frac{(n-1)s^2}{X_R^{\ 2}} < \sigma^2 < \frac{(n-1)s^2}{X_L^{\ 2}}$$

- Confidence interval for standard deviation

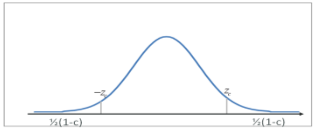
$$\sqrt{\frac{(n-1)s^2}{X_R^{\ 2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{X_L^{\ 2}}}$$

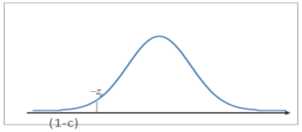
University of Warsaw
Faculty of Economic Sciences

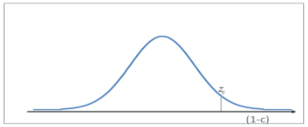
HYPOTHESIS TESTING

- Hypothesis test
  - Process that use sample statistics to test a claim about the value of a population parameter
  - Process of hypothesis testing
    - Stating a claim
      - Mean value of number of clients that would come to the shop within a day is equal to 100
    - Appropriate test choice (according to population parameter and to available data)
      - We have information from 100 days and average sample is equal to 95, and standard deviation equal to 10 (z statistics)
    - Null and Alternative hypothesis definition

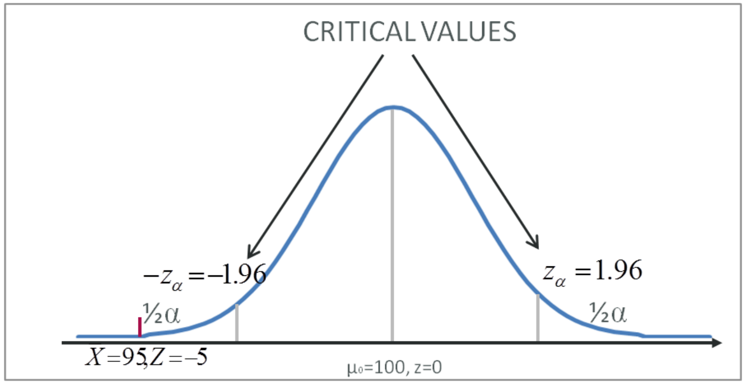
$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \qquad \begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$







- Choice Level of significance α
  - Level of significance α defines how big part of the distribution which is true under the null hypothesis should be out of acceptance level – I type error definition
  - α=5%



$$z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{95 - 100}{\frac{10}{\sqrt{100}}} = -5 \rightarrow N(0,1)$$

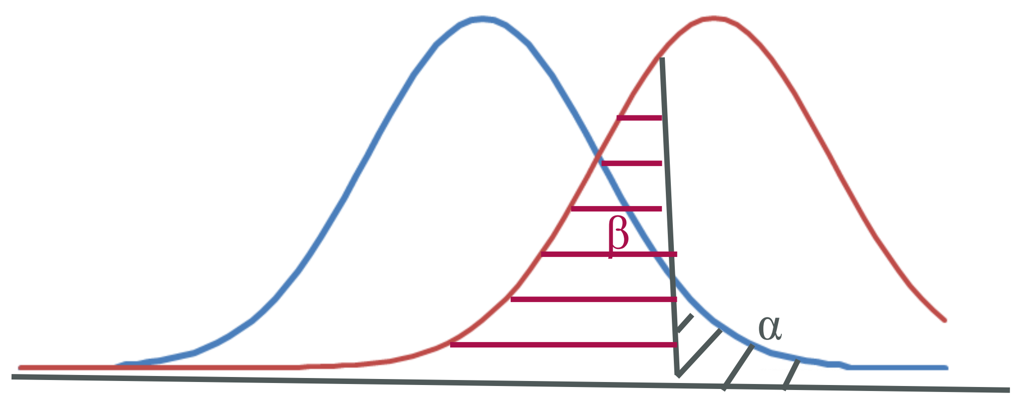
WE MAY:

- REJECT NULL HYPOTHESIS
- FAIL TO REJECT NULL HYPOTHESIS

TYPE I & II ERROR

- TYPE I: Null hypothesis is rejected when its true
- TYPE II: Null hypothesis is not rejected when it is false

	TRUTH OF H_0	
DECISION	H_0 IS TRUTH	H_0 IS FALSE
DO NOT REJECT H0	CORRECT DECISION	TYPE II ERROR (β)
REJECT H0	TYPE I ERROR (α)	CORRECT DECISION



P-VALUE CONCEPT

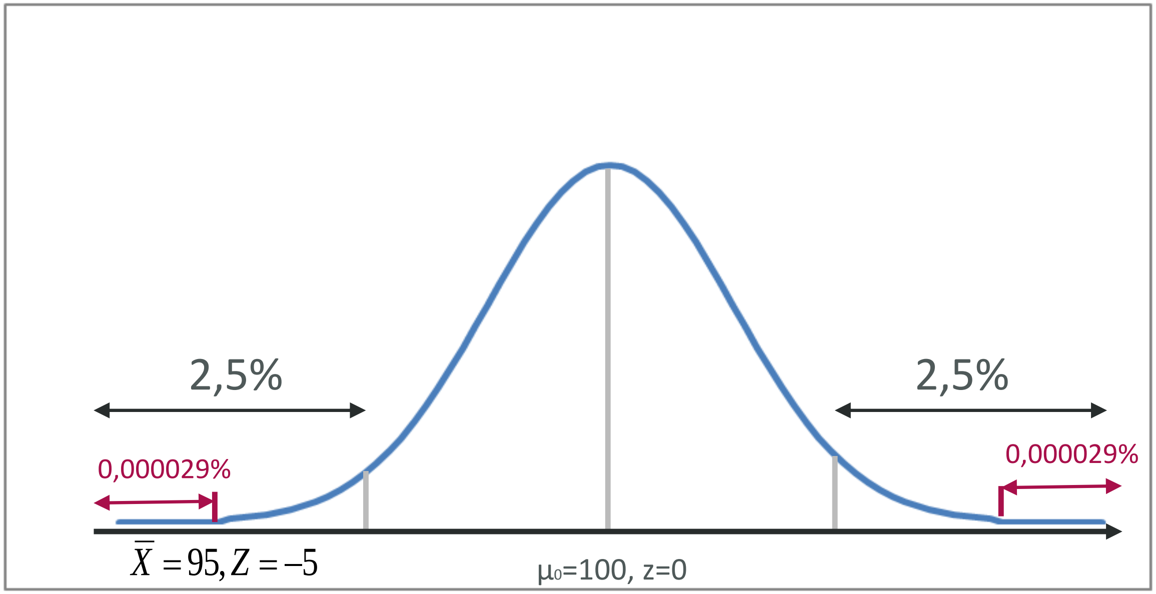
- STANDARD HYPOTHESIS CONCEPT:
  - Sample statistics value versus critical values
- BUT: THERE IS ANOTHER APPROACH:
  - p-value analysis
    - P-value → probability of obtaining sample statistics as extreme as or even more extreme than observed sample statistics (in absolute term)
      - If p-value <= significance level → reject H0
      - If p-value > significance level → fail to reject H0
    - BENEFITS: We do not have to know precisely value of test statistics and critical value (simplicity of interpretation)

P-VALUE EXAMPLE

$$p-value = P(-5 < z \cup z > 5) = P(-5 < z) + P(z > 5) =$$

$$= P(-5 < z) + 1 - P(z < 5) = \Phi(-5) + 1 - \Phi(5) =$$

$$= 0{,}00000029 + 1 - 0{,}999999713 = 0{,}00000057$$

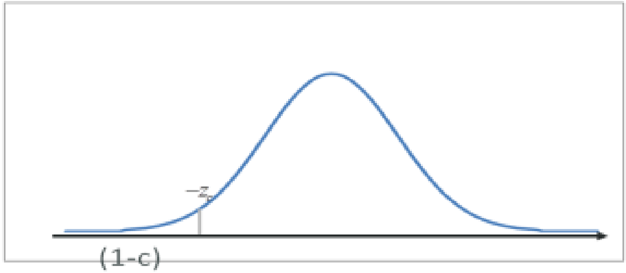


TEST FOR MEAN (n>30 or known σ)

- In auto racing, a pit stop is where a racing vehicle stops for new tires, fuel, repairs, and other mechanical adjustments. The efficiency of a pit crew that makes these adjustments can affect the outcome of a race. A pit crew claims that its mean pit stop time is less than 13 seconds. A random selection of 32 pit stop times has a sample mean of 12.9 seconds and a standard deviation of 0.19 second. Is there enough evidence to support the claim on the α=0.01
  - CLAIM: μ<13
  - DATA: N=32, X_{AVG}=12.9, STD=0.019
  - TEST STATISTICS?
  - HYPOTHESIS: H0: μ>=13, H1: μ<13 (CLAIM)
  - SIGNIFICANCE: α=0.01

$$\begin{cases} H_0 : \mu \geq 13 \\ H_1 : \mu < 13 \end{cases}$$

$$z = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \rightarrow N(0,1) = \frac{12{,}9 - 13}{\frac{0{,}19}{\sqrt{32}}} = \frac{-0{,}01}{0{,}0336} = -2{,}98$$

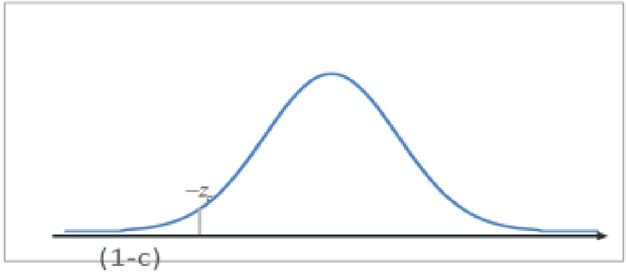


$$\Phi^{-1}{}_{0,01} = -2{,}33 \qquad -2{,}98 < -2{,}33 \rightarrow REJECTION\_H0$$

TEST FOR MEAN (n<30, unknown σ)

- Government assessed that average price of bucket of goods in different shops is at least 150 PLN. We suspect that this claim is incorrect and basing on a sample of 10 shops we have checked that average cost is equal to 146 PLN with standard deviation equal to 7 PLN. Is there enough evidence to reject government assessment on α=0.05?
  - CLAIM: μ>=150
  - DATA: N=10, X_{AVG}=146, STD=7
  - TEST STATISTICS?
  - HYPOTHESIS: H0: μ>=150 (CLAIM), H1: μ<150
  - SIGNIFICANCE: α=0.05

$$\begin{cases} H_0 : \mu \geq 150 \\ H_1 : \mu < 150 \end{cases}$$



$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \rightarrow t_{n-1} = \frac{146 - 150}{\frac{7}{\sqrt{10}}} = \frac{-4}{2.214} = -1{,}81$$

$$t^{-1}_{9;0.05} = -1{,}83 \qquad -1{,}81 > -1{,}83 \rightarrow FAIL\_REJECTION\_H0$$

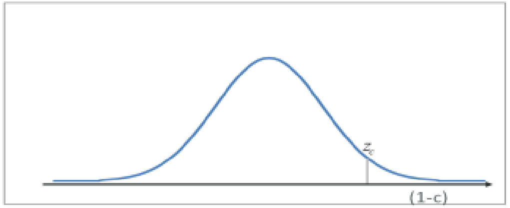
University of Warsaw
Faculty of Economic Sciences

TEST FOR PROPORTION

- Government assessed that in Poland more than 25% adults lives for less then 30 PLN a day. In a random sample of 100 adults 28% of responders said that they live for less than 30 PLN a day. At the significance level α=0.05, is there enough evidence to support the government statement?
  - CLAIM: π>25%
  - DATA: N=100, $\hat{p}$=0.28
  - TEST STATISTICS? (CHECK: $N\hat{p} = 100 * 0{,}28 = 28$)
  - HYPOTHESIS: H0: π<=25%, H1: π>25% (CLAIM)
  - SIGNIFICANCE: α=0.05

$$\begin{cases} H_0 : \pi \le 25\% \\ H_1 : \pi > 25\% \end{cases}$$

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \rightarrow N(0,1) = \frac{0.28 - 0.25}{\sqrt{\frac{0.28 * 0.72}{100}}} = \frac{0.03}{0.045} = 0{,}69$$



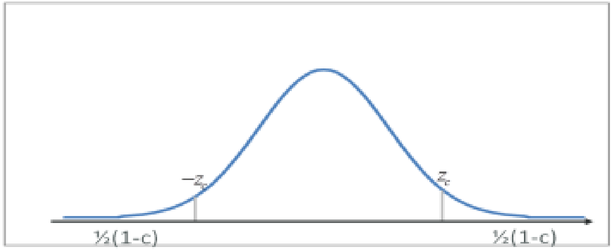
$$\Phi^{-1}{}_{0,95} = 1.64 \qquad 0{,}69 < 1{,}64 \rightarrow FAIL\_REJECTION\_H0$$

University of Warsaw
Faculty of Economic Sciences

TEST FOR EQUAL VARIANCES/STD

- Advisor from the company that produces bulbs told you that theirs bulbs works almost the same long. He told you that the standard deviation is equal to 8 hours. You have prepared experiment on 50 bulbs and it has turned out that standard deviation was equal to 9.5 hours. Is there enough evidence to support he claim at the 5% level?
  - CLAIM: σ=8
  - DATA: N=50, s=9.5
  - TEST STATISTICS?
  - HYPOTHESIS: H0: σ=8 (CLAIM) , H1: σ<>8
  - SIGNIFICANCE: α=0.05

$$\begin{cases} H_0 : \sigma = 8 \\ H_1 : \sigma \neq 8 \end{cases}$$



$$\mathrm{X}^2 = \frac{(n-1)s^2}{\sigma^2} \rightarrow \mathrm{X}^2{}_{n-1} = \frac{49 * 9.5^2}{8^2} = \frac{539}{64} = 69.0977$$

$$\mathrm{X}_L{}^2 = \mathrm{X}^2{}_{0.025;9} = 31.5549$$

$$\mathrm{X}_R{}^2 = \mathrm{X}^2{}_{0.975;9} = 70.2224$$

$$31.56 < 69.098 < 70.22$$

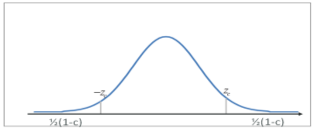
$$\rightarrow FAIL\_REJECTION\_H0$$

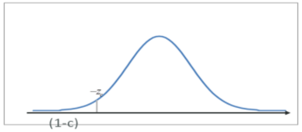
University of Warsaw
Faculty of Economic Sciences

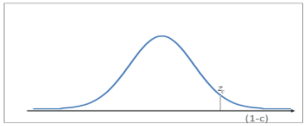
TESTS FOR TWO POPULATIONS

- Independent vs dependent samples
  - Independent samples – sample selected from the one population are not related to the sample selected from the second population
  - Dependent samples – each member of one sample corresponds to a member of the other sample (paired sample or matched sample)
- Hypothesis test
  - Process of hypothesis testing
    - Stating a claim
      - Mean value of number of clients that would come to the shop A within a day is equal to mean in shop B
    - Appropriate test choice (according to population parameter and to available data)
    - Null and Alternative hypothesis definition

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \qquad \begin{cases} H_0 : \mu_1 \geq \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases} \qquad \begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$







    - Choice Level of significance α

University of Warsaw
Faculty of Economic Sciences

TEST FOR DIFFERENCE OF TWO MEANS (n>30 or known σ)

- SAMPLES MUST BE RANDOMLY SELECTED
- SAMPLES MUST BE INDEPENDENT
- EACH SAMPLE SIZE >30 OR KNOWN σ

- Manager form the bank claims that there is a difference in the mean credit card debts of clients from big cities and rest of the country. A results of a random survey of 400 clients from big cities and the rest of country are investigated. The two samples are independent. Results for the big cities was MEAN= 50.000 PLN with STD=10.000 PLN and for the results for the rest of the country was MEAN= 48.000 PLN with STD=8.000 PLN. Perform test on 10% significance level.

$$z=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\rightarrow N(0,1)=\frac{(50000-48000)-0}{\sqrt{\frac{10000^2}{400}+\frac{8000^2}{400}}}=\frac{2000}{640{,}31}=1{,}5617$$

$$\Phi^{-1}{}_{0,05}=-1{,}64$$

$$\Phi^{-1}{}_{0,95}=1{,}64$$

$$1{,}56<1{,}64\rightarrow FAIL\_TO\_REJECT\_H0$$

University of Warsaw
Faculty of Economic Sciences

TEST FOR DIFFERENCE OF TWO MEANS (DIFFERENT VARIANCES)

- SAMPLES MUST BE RANDOMLY SELECTED
- SAMPLES MUST BE INDEPENDENT
- EACH POPULATION MUST HAVE A NORMAL DISTRIBUTION

- Previous case, but:
   1. **20 results for every case**. Big cities (MEAN= 50.000 PLN, STD=10.000 PLN), rest of the country (MEAN= 48.000 PLN, STD=8.000 PLN).

$$t=\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\rightarrow t_{\min(n_1-1,n_2-1)}=\frac{(50000-48000)-0}{\sqrt{\frac{10000^2}{20}+\frac{8000^2}{20}}}=$$

$$=\frac{2000}{2863.564}=0.698$$

$$t^{-1}{}_{19;0,05}=-1.73$$

$$t^{-1}{}_{19;0,95}=1.73$$

$$0,69<1,73\rightarrow FTRH0$$

University of Warsaw
Faculty of Economic Sciences

TEST FOR DIFFERENCE OF TWO MEANS (EQUAL VARIANCES)

- SAMPLES MUST BE RANDOMLY SELECTED
- SAMPLES MUST BE INDEPENDENT
- EACH POPULATION MUST HAVE A NORMAL DISTRIBUTION

- Previous case, but:
  1. **20 results for every case**. Big cities (MEAN= 50.000 PLN, STD=10.000 PLN), rest of the country (MEAN= 48.000 PLN, **STD=10.000 PLN**).

$$t=\frac{\left(\bar{X}_1-\bar{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\rightarrow t_{n_1+n_2-2}$$

$$=\frac{(50000-48000)-0}{\sqrt{\frac{100000000*19+100000000*19}{38}}\sqrt{\frac{1}{20}+\frac{1}{20}}}=$$

$$=\frac{2000}{2236.068}=0.791$$

$$t^{-1}{}_{38;0,05}=-1.69$$

$$t^{-1}{}_{38;0,95}=1.69$$

$$0,791<1.69\rightarrow FTRH0$$

TEST FOR DIFFERENCE OF TWO MEANS (PAIRED DATA)

- EACH POPULATION MUST HAVE A NORMAL DISTRIBUTION

- Quality of new control process is assessed in bank. 10 entities were analyzed. For each of entity total value of operational risk events were calculated before and after introduction of new control process. Claim is that introduction of new control process reduce average cost of operational risk events by 2000 PLN . Data for the case are presented below. Perform a test on α=1%.

$$t = \frac{\overline{X}_D - \mu_D}{\frac{s_D}{\sqrt{n}}} \rightarrow t_{n-1} = \frac{\left(-4889.9 - (-2000)\right)}{\frac{4268.345}{\sqrt{10}}} =$$

$$= \frac{-2889.9}{1349.769} = -2.141$$

$$t^{-1}{}_{9;0,01} = -2.821$$

$$-2,88 < -2,141 \rightarrow RH0$$

University of Warsaw
Faculty of Economic Sciences

TEST FOR DIFFERENCE OF TWO PROPORTIONS

- SAMPLES MUST BE RANDOMLY SELECTED
- SAMPLES MUST BE INDEPENDENT
- EACH SAMPLE SIZE Np & Nq >5

- A study of 150 randomly selected shares listed on the WSE and 200 randomly selected shares listed on the NYSE shows that 86% of the shares from the WSE and 74% of the shares from the NYSE went up on 12/10/2014. At α=0.05 can you reject the claim that the proportion of shares that went up is the same for shares from the WSE and shares from the NYSE?

$$z=\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}\rightarrow N(0,1)=\frac{(0,86-0,74)-0}{\sqrt{0.79*(1-0.79)*\left(\frac{1}{150}+\frac{1}{200}\right)}}=2.73$$

$$\bar{p}=\frac{x_1+x_2}{n_1+n_2}$$

$$\Phi^{-1}{}_{0,025}=-1,96$$

$$\Phi^{-1}{}_{0,975}=1,96$$

$$2,73>1,96\rightarrow RH0$$

TEST FOR TWO VARIANCES

- SAMPLES MUST BE RANDOMLY SELECTED
- SAMPLES MUST BE INDEPENDENT
- EACH SAMPLE COME FROM NORMAL DISTIRBUTION

- A call center manager is creating a system to decrease the variance of the time client waits before its call is taken. Under the old system sample of 100 clients have variance of 100 and under the new system a random sample of 100 clients had a variance of 64. At the level of significance equal to 5% is there enough evidence to start a new process?

$$F = \frac{s_1^2}{s_2^2} \to F(n_1 - 1, n_2 - 1) = \frac{100}{64} = 1{,}5625 \qquad s_1^2 \geq s_2^2$$

$$F^{-1}{}_{99,99;0,95} = 1{,}39 \qquad 1{,}5625 > 1{,}39 \to RH0$$

WHAT IF NORMAL ASSUMPTION IS NOT MET?

- Normality tests
  - Jarque – Bera test
  - Shapiro – Wilk
  - Shapiro – Francia

$$JB = \frac{n-k+1}{6}\left(S^2 + \frac{1}{4}(C-3)^2\right)$$

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}, \qquad C = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2},$$

Wicoxon & Mann – Whitney for unpaired data

Operational risk losses	GROUP ID	After Control	GROUP ID
530	1	423	2
209	1	109	2
974	1	483	2
190	1	425	2
796	1	662	2
927	1	254	2
266	1	758	2
917	1	447	2
946	1	513	2
911	1	649	2
974	1	911	2
360	1	478	2
442	1	834	2
928	1	684	2
350	1	298	2
841	1	872	2
624	1	816	2
827	1	468	2
785	1	508	2

$$T = \sum_{i=1}^{n_1} R_{1i}$$

$$U = T - \frac{n_1(n_1+1)}{2}$$

University of Warsaw
Faculty of Economic Sciences

Wicoxon & Mann – Whitney for unpaired data

JOINT		
109	2	1
190	1	2
209	1	3
254	2	4
266	1	5
298	2	6
350	1	7
360	1	8
423	2	9
425	2	10
442	1	11
447	2	12
468	2	13
478	2	14
483	2	15
508	2	16
513	2	17
530	1	18
624	1	19
649	2	20
662	2	21
684	2	22
758	2	23
785	1	24
796	1	25
816	2	26
827	1	27
834	2	28
841	1	29
872	2	30
911	1	31
911	2	32
917	1	33
927	1	34
928	1	35
946	1	36
974	1	37
974	1	38

All ranks	Expected 1	Expected 2
741	370,5	370,5
	Observed 1	Observed 2
T	422	319
U	232	129
n1n2	361	

Not easy distribution of statistic (normal approximations are used)

University of Warsaw
Faculty of Economic Sciences

Wilcoxon matched-pairs signed-ranks test (paired data)

- Wilcoxon matched-pairs signed-ranks test
  - Null hypothesis: both distributions are the same
  - Alterantive: both distributions are different

$$W = \sum_{i=1}^{N_r} [\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$$

Operational risk losses	After Control	Difference	Abs. Value
530	423	-107	107
209	109	-100	100
974	483	-491	491
190	425	235	235
796	662	-134	134
927	254	-673	673
266	758	492	492
917	447	-470	470
946	513	-433	433
911	649	-262	262
974	911	-63	63
360	478	118	118
442	834	392	392
928	684	-244	244
350	298	-52	52
841	872	31	31
624	816	192	192
827	468	-359	359
785	508	-277	277

University of Warsaw
Faculty of Economic Sciences

Wilcoxon matched-pairs signed-ranks test (paired data)

$$W = \sum_{i=1}^{N_r} [\operatorname{sgn}(x_{2,i} - x_{1,i}) \cdot R_i]$$

Not easy distribution of statistic (normal approximations are used)

			Ordered Value			
Operational risk losses	After Control	Difference	Abs. Value	Sign	Rank	
841	872	31	31	1	1	1
350	298	-52	52	-1	2	-2
974	911	-63	63	-1	3	-3
209	109	-100	100	-1	4	-4
530	423	-107	107	-1	5	-5
360	478	118	118	1	6	6
796	662	-134	134	-1	7	-7
624	816	192	192	1	8	8
190	425	235	235	1	9	9
928	684	-244	244	-1	10	-10
911	649	-262	262	-1	11	-11
785	508	-277	277	-1	12	-12
827	468	-359	359	-1	13	-13
442	834	392	392	1	14	14
946	513	-433	433	-1	15	-15
917	447	-470	470	-1	16	-16
974	483	-491	491	-1	17	-17
266	758	492	492	1	18	18
927	254	-673	673	-1	19	-19
					W Statistc	-78

ANOVA (ANALYSIS OF VARIANCE)

- SAMPLES MUST BE RANDOMLY SELECTED APPROX. FROM NORMAL DISTRIBUTION
- SAMPLES MUST BE INDEPENDENT
- EACH POPULATION MUST HAVE THE SAME VARIANCES

- One-way (one factor) analysis of variance is a hypothesis-testing technique that is used to compare the means more than 2 samples.
  - Often called ANOVA

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \\ H_1 : \exists \mu_i \neq \mu_j \end{cases}$$

$$Statistic = \frac{VarianceBetweenSamples}{VarianceWithinSamples} \rightarrow F(k-1, N-k)$$

- Multiple-way analysis of variance (MANOVA) → more than 1 factor.
  - Hypothesis for each factor & interactions

University of Warsaw
Faculty of Economic Sciences

STATISTICS & ECONOMETRICS

Introduction to Econometrics

University of Warsaw
Faculty of Economic Sciences

Definition

- Econometrics → ‘measurement in economics’.
  - The origins of econometrics are rooted in economics.
  - Main techniques are also important in:
    - Finance
    - Sociology
    - Psychology
    - Demography
    - Medicine
    - Etc.

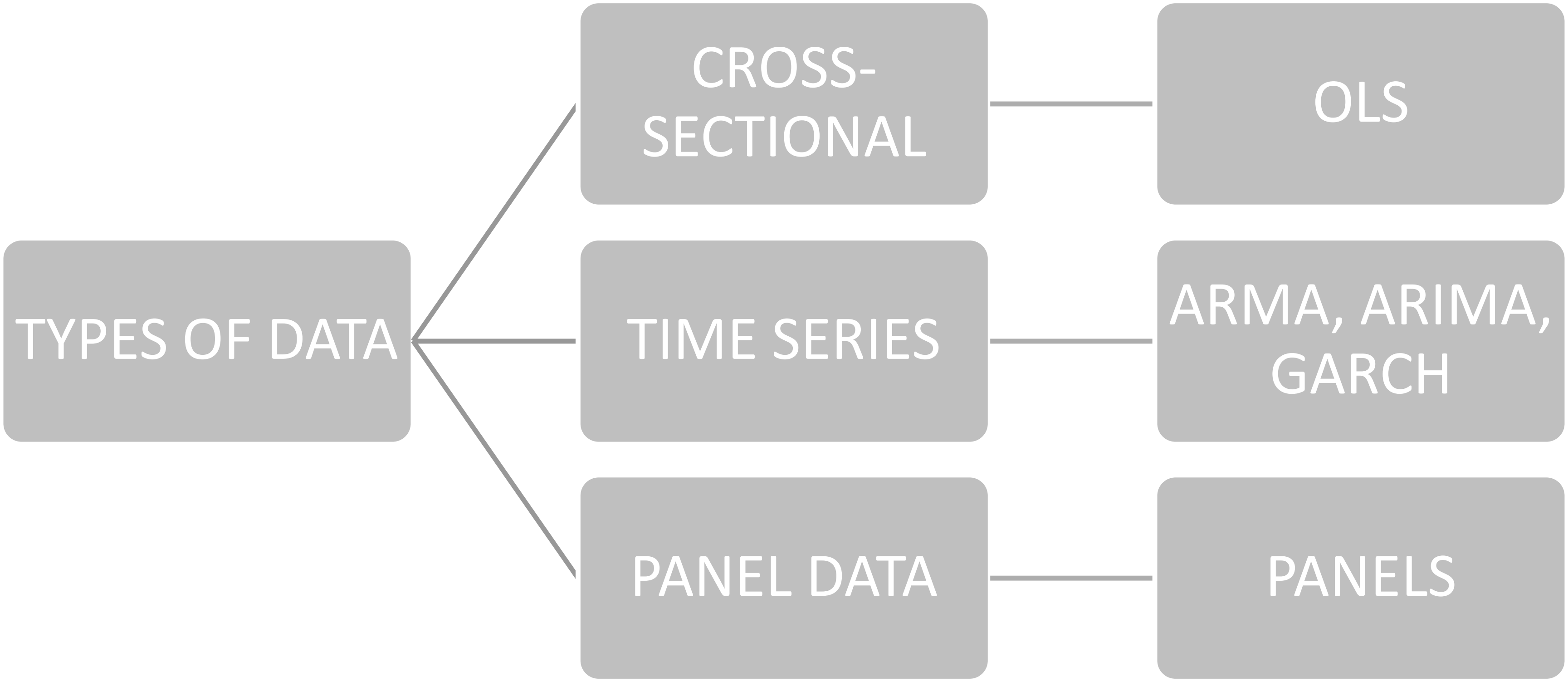
- In other words: *application of statistical techniques to problems in economics (finance etc.).*

University of Warsaw
Faculty of Economic Sciences

EXAMPLES OF USAGE IN FINANCE

TESTING THEORIES IN FINANCE	• Financial markets efficiency (weak-form)
DETERMINING ASSET PRICES OR RETURNS	• Calculating Options Prices
TESTING HYPOTHESES CONCERNING THE RELATIONSHIPS BETWEEN VARIABLES	• Explaining the determinants of bond credit ratings used by the ratings agencies • Modeling long-term relationships between prices and exchange rates
FORECASTING FUTURE VALUES OF FINANCIAL VARIABLES AND FOR FINANCIAL DECISION-MAKING	• Measuring and forecasting the volatility of bond returns • Forecasting the correlation between the stock indices of two countries

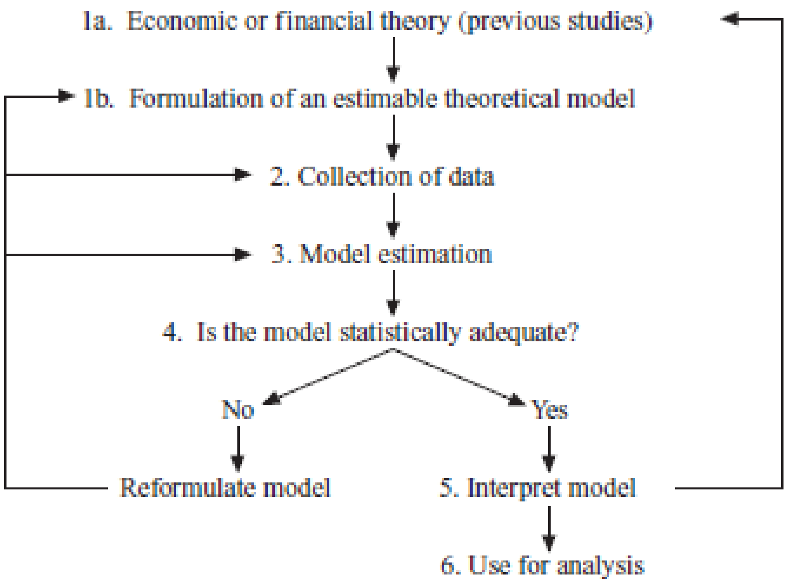
TYPES OF DATA



University of Warsaw
Faculty of Economic Sciences

STEPS OF FORMULATING THE MODEL

- *Step 1a and 1b:* ***general statement of the problem***
  - Formulation of theoretical model or intuition from financial theory about relation between variables
  - Should present a sufficiently good approximation
- *Step 2:* ***collection of data relevant to the model***
  - Existing databases, questionnaire etc.
- *Step 3:* ***choice of estimation method relevant to the model proposed in step 1***
  - OLS, TS model, Panel?
- *Step 4:* ***statistical evaluation of the model***
  - What assumptions were required to estimate the parameters of the model optimally?
  - Were these assumptions satisfied by the data or the model?
  - Also, does the model adequately describe the data?
- *Step 5:* ***evaluation of the model from a theoretical perspective***
  - Are the parameter estimates of the sizes and signs relevant to theory or intuition?
- *Step 6:* ***use of model***

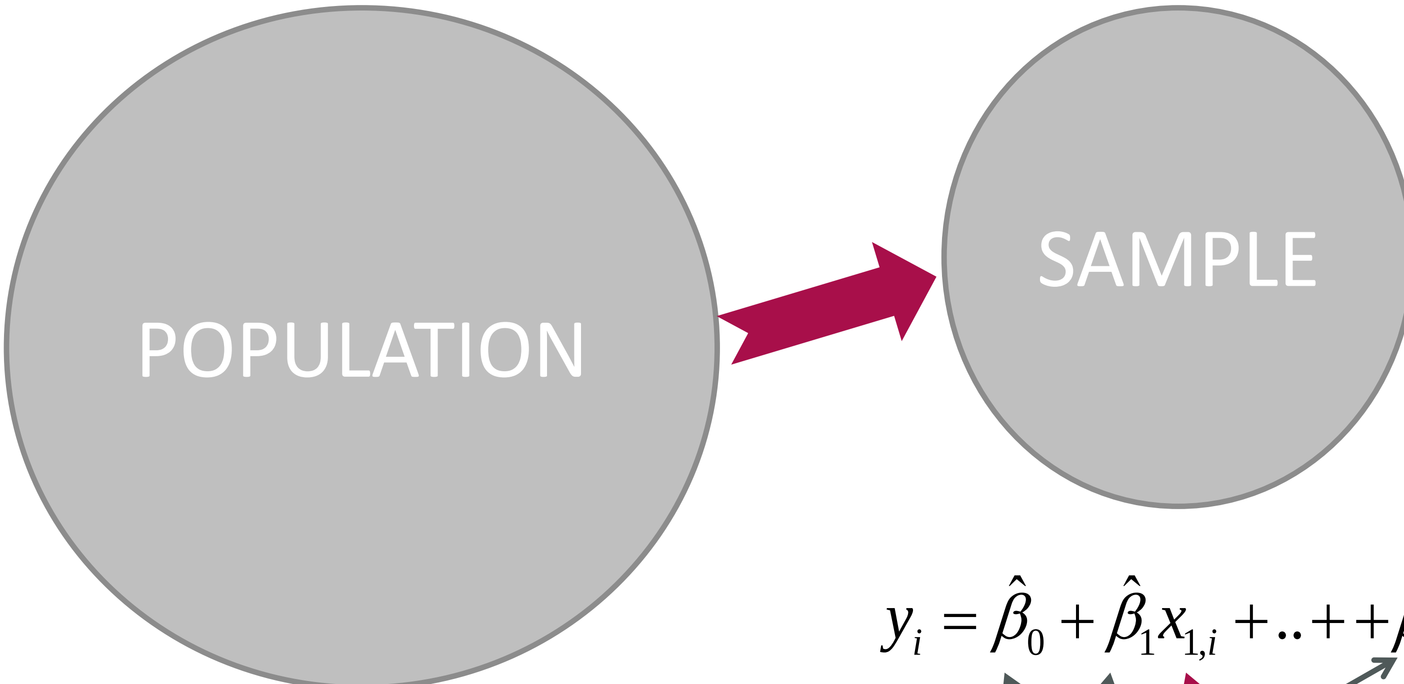


Process of building a robust empirical model is an iterative and it is certainly not an exact science.

Final preferred model could be very different from the one originally proposed and from the other researchers

Brooks, Ch. (2008), Introductory Econometrics for Finance, Cambridge University Press, 2nd edition.

CLRM



$$y_i = \beta_0 + \beta_1 x_{1,i} + .. + + \beta_n x_{n,i} + \varepsilon_i$$

dependent variable

parameters

independent variables

error term

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + .. + + \hat{\beta}_n x_{n,i} + e_i$$

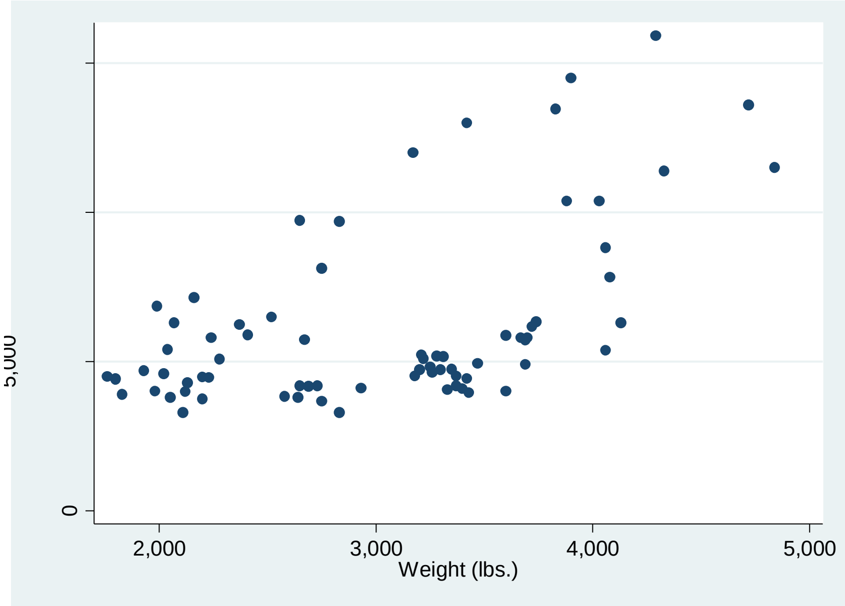
dependent variable

estimates

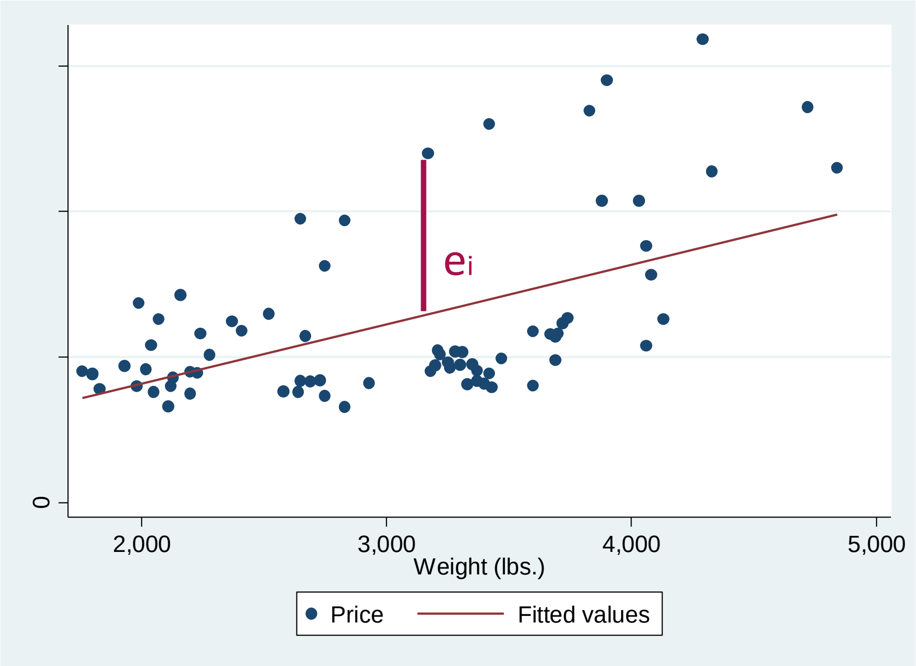
independent variables

residual

OLS



$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + e_i$$



$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y - \hat{y})^2 =$$

$$= \sum_{i=1}^{N} (y - \hat{\beta}_0 + \hat{\beta}_1 x_{1,i})^2$$

$$\hat{\beta} = (X'X)^{-1} X'y$$

University of Warsaw
Faculty of Economic Sciences

OLS EXAMPLE

Source	SS	df	MS
Model	184233937	1	184233937
Residual	450831459	72	6261548.04
Total	635065396	73	8699525.97

Number of obs	=	74
F(1, 72)	=	29.42
Prob > F	=	0.0000
R-squared	=	0.2901
Adj R-squared	=	0.2802
Root MSE	=	2502.3

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	2.044063	.3768341	5.42	0.000	1.292857	2.795268
_cons	-6.707353	1174.43	-0.01	0.995	-2347.89	2334.475

University of Warsaw
Faculty of Economic Sciences

CLRM ASSUMPTIONS

1. Linear equation

$$y_i = \beta_0 + \beta_1 x_{1,i} + .. + + \beta_n x_{n,i} + \varepsilon_i$$

2. Non-random independent variables
3. Expected value of error term is equal to 0

$$E(\varepsilon_i) = 0, \forall i$$

4. Covariance between error terms for any two observations is equal to 0

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$$

5. Homoscedasticity of the error term

$$Var(\varepsilon_i) = \sigma^2 < \infty, \forall i$$

University of Warsaw
Faculty of Economic Sciences

OLS PROPERTIES IN CLRM

- For CLRM OLS estimator is named as Best Linear Unbiased Estimators (BLUE)
  - Unbiased - on average, the actual values of *estimators* will be equal to their true values
  - Best - means that the OLS estimator has minimum variance among the class of linear unbiased estimators(Gauss--Markov theorem)
- PROPERTIES
  - ***Consistency*** $$\lim_{n\to\infty} P(|\hat{\beta}-\beta|>\delta)=0, \forall\delta>0$$
  - ***Unbiasedness*** $$E(\hat{\beta}_i)=\beta_i, \forall i$$
    - On average, the estimated values for the coefficients will be equal to their true values. That is, there is no systematic overestimation or underestimation of the true coefficients.
  - ***Efficiency***
    - Estimator is said to be efficient if no other estimator has a smaller variance.
    - If the estimator is 'best', the uncertainty associated with estimation will be minimized for the class of linear unbiased estimators.

MODEL SPECIFICATIONS – MODELS CONVERTED TO LINEAR

- LINEAR MODEL

$$y_i = \beta_0 + \beta_1 x_{1,i} + .. + \beta_n x_{n,i} + \varepsilon_i \rightarrow y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + .. + \hat{\beta}_n x_{n,i} + e_i$$

- EXPONENTIAL MODELS

$$y_i = B_0 x_{1,i}^{\beta_1} * .. * x_{n,i}^{\beta_n} \exp(\varepsilon_i) \rightarrow \ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_{1,i}) + .. + \hat{\beta}_n \ln(x_{n,i}) + e_i$$

$$y_i = \exp(\beta_0 + \beta_1 x_{1,i} + .. + \beta_n x_{n,i} + \varepsilon_i) \rightarrow \ln(y_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + .. + \hat{\beta}_n x_{n,i} + e_i$$

TREATMENT OF DIFFERENT TYPES OF VARIABLES

- CONTINUOUS (PRICE, WEIGHT ETC.)
  - AS IT IS
  - TRANSFORMATION
- BINARY (TWO POSSIBLE OUTCOMES – GENDER)
  - AS IT IS
- DISCRETE
  - BINARIZATION → BASE LEVEL (OMITTING)
- INTERACTIONS
  - BINARYxBINARY, BINARYxCONTINUOUS, CONTINUOUSxCONTINUOUS
- POLYNOMIALS

INTERPRETATION OF ESTIMATES FOR DIFFERENT TYPES OF VARIABLES

- CONTINUOUS (PRICE, WEIGHT ETC.)
  - y vs x → y will change by $\hat{\beta}_1$ when x increase by 1
  - ln(y) vs ln(x) → y will change by $\hat{\beta}_1$ % when x increase by 1%
  - ln(y) vs x → y will change by $\hat{\beta}_1$ *100% when x increase by 1
- BINARY (TWO POSSIBLE OUTCOMES – GENDER)
  - DIFFERENCE IN EXPECTED y FOR BINARY GROUP
- DISCRETE
  - DIFFERENCE IN EXPECTED y BETWEEN ANALYZED GROUP AND BASE GROUP
- INTERACTIONS
  - BINARYxCONTINOUS – DIFFERENCE IN RELATION BETWEEN INDEPENDENT AND DEPENDENT VARIABLE WRT BINARY GROUPS
- POLYNOMIALS
  - JOINT INTERPRETATION

INTERPRETATION EXAMPLE

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  7,    61) =    9.22
       Model |   296575650     7    42367950           Prob > F      =  0.0000
    Residual |   280221309    61  4593791.95           R-squared     =  0.5142
-------------+------------------------------           Adj R-squared =  0.4584
       Total |   576796959    68  8482308.22           Root MSE      =  2143.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.160401   .4699196     6.73   0.000     2.220739    4.100064
     foreign |  -889.5453   3541.729    -0.25   0.803     -7971.67    6192.579
             |
    foreign#|
    c.weight |
          1  |   1.915014   1.489782     1.29   0.204     -1.06399    4.894018
             |
       rep78 |
          2  |   601.1732   1698.628     0.35   0.725    -2795.444    3997.791
          3  |   939.8501   1574.434     0.60   0.553    -2208.426    4088.126
          4  |   564.7018   1642.707     0.34   0.732    -2720.093    3849.496
          5  |   767.1887   1768.874     0.43   0.666    -2769.894    4304.272
             |
       _cons |  -5232.744   2102.146    -2.49   0.016    -9436.245   -1029.243
------------------------------------------------------------------------------
```

t & F tests

- t test for a simple hypothesis

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \to t_{N-K} \qquad \begin{cases} H_0 : \hat{\beta} = 0 \\ H_1 : \hat{\beta} \neq 0 \end{cases}$$

- F test for a multiple hypothesis

$$F = \frac{\left(e'_R e_R - e'e\right)/g}{e'e/(N-K)} \to F(g, N-K) \qquad \begin{cases} H_0 : \hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_K = 0 \\ H_1 : \exists \hat{\beta}_i \neq 0 \end{cases}$$

DIFFERENT HYPOTHESIS MIGHT BE CHECKED

University of Warsaw
Faculty of Economic Sciences

t & F tests EXAMPLE

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  7,    61) =    9.22
       Model |  296575650     7   42367950             Prob > F      =  0.0000
    Residual |  280221309    61  4593791.95            R-squared     =  0.5142
-------------+------------------------------           Adj R-squared =  0.4584
       Total |  576796959    68  8482308.22            Root MSE      =  2143.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.160401   .4699196     6.73   0.000     2.220739    4.100064
     foreign |  -889.5453   3541.729    -0.25   0.803     -7971.67    6192.579
             |
    foreign#|
    c.weight |
         1   |   1.915014   1.489782     1.29   0.204     -1.06399    4.894018
             |
       rep78 |
         2   |   601.1732   1698.628     0.35   0.725    -2795.444    3997.791
         3   |   939.8501   1574.434     0.60   0.553    -2208.426    4088.126
         4   |   564.7018   1642.707     0.34   0.732    -2720.093    3849.496
         5   |   767.1887   1768.874     0.43   0.666    -2769.894    4304.272
             |
       _cons |  -5232.744   2102.146    -2.49   0.016    -9436.245   -1029.243
------------------------------------------------------------------------------
```

TSS DECOMPOSITION

- Sum of squares – measure of variation of the variable around its mean
- Decomposition of sum of squares (model with constant)

$$TSS = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = \quad ESS = \sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2 + \quad RSS = \sum_{i=1}^{N}\left(e_i\right)^2$$

Total variation — Explained variation — Unexplained variation

- TSS may be decomposed into explained and unexplained part of the variation

University of Warsaw
Faculty of Economic Sciences

R2 – GOODNESS OF FIT STATISTICS

- R2 describes how big part of dependent variable variation may be explained by the variation of independent variables

$$R^2 = \frac{ESS}{TSS}$$

- Main drawback: R2 increases with number of variables no matter how good they are.
- Adjustment:

$$R^2{}_{ADJ} = 1 - \frac{N-1}{N-K}(1-R^2)$$

University of Warsaw
Faculty of Economic Sciences

R2 EXAMPLE

```
      Source |       SS       df       MS              Number of obs =      69
-------------+------------------------------           F(  7,    61) =    9.22
       Model |   296575650     7    42367950           Prob > F      =  0.0000
    Residual |   280221309    61  4593791.95           R-squared     =  0.5142
-------------+------------------------------           Adj R-squared =  0.4584
       Total |   576796959    68  8482308.22           Root MSE      =  2143.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.160401   .4699196     6.73   0.000     2.220739    4.100064
     foreign |  -889.5453   3541.729    -0.25   0.803     -7971.67    6192.579
             |
    foreign#|
    c.weight |
          1  |   1.915014   1.489782     1.29   0.204     -1.06399    4.894018
             |
       rep78 |
          2  |   601.1732   1698.628     0.35   0.725    -2795.444    3997.791
          3  |   939.8501   1574.434     0.60   0.553    -2208.426    4088.126
          4  |   564.7018   1642.707     0.34   0.732    -2720.093    3849.496
          5  |   767.1887   1768.874     0.43   0.666    -2769.894    4304.272
             |
       _cons |  -5232.744   2102.146    -2.49   0.016    -9436.245   -1029.243
------------------------------------------------------------------------------
```

DIAGNOSTICS

- Basic diagnostics
  - Ramsey test for correct model specification
    - H0: model has no omitted variables
    - Two versions (powers of fitted values or powers of independent variables)
    - Solution: Looking for additional variables
  - Jarque-Bera test for normality:
    - H0: error term comes from normal distribution
    - Solution: sample size
  - Breusch-Godfrey test for autocorrelation
    - H0:there is no autocorrelation in error terms
    - Soulution: Robust estimator of Variance-Covariance matrix
  - Breusch-Pagan test for homoscedasticity
    - H0:there is no heteroscedasticity in error terms
    - Solution: Robust estimator of Variance-Covariance matrix
  - VIF statistics analysis
    - When the predictors are highly correlated there may be a significant change in the regression coefficients if you add or delete an independent variable.
    - The estimated standard errors of the fitted coefficients are inflated --> the estimated coefficients may not be statistically significant even though a statistical relation exists between the dependent and independent variables.
    - Rules of thumb applied to the VIF:
      - The largest VIF is greater than 10 (30).

MLE ESTIMATOR

- Sample of $x_{1,\dots,}$ x_n i.i.d. random variables
- Density function:

$$f(x_1 \mid \theta)$$

- Joint density function:

$$f(x_1, x_2, \dots, x_n \mid \theta) = f(x_1 \mid \theta) * f(x_2 \mid \theta) * \dots * f(x_n \mid \theta)$$

- Maximum likelihood estimator:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} f(x_1, x_2, \dots, x_n \mid \theta)$$

University of Warsaw
Faculty of Economic Sciences

LOGIT/PROBIT MODEL

$$y_i^* = \boldsymbol{\beta} * \boldsymbol{X}_i + \varepsilon$$

$$y_i = \begin{cases} 1 & if\ y_i^* > 0 \\ 0 & otherwise \end{cases}$$

where:
y_i^* – latent variable,
β – parameter,
X_i – independent variable,
ε – random error,
y_i – observable result of the phenomenon.

LOGIT – error term comes form logistic distribution
PROBIT – error term comes form normal distribution

University of Warsaw
Faculty of Economic Sciences

MONTE CARLO SIMULATION

- Instead of analytical formulation, simulation is performed
- The Monte Carlo method
  – numerical method for statistical simulation based on numbers of random realizations
- Case:
  – Calculation of possible OpRisk loss.
    - For typical year (9/10) OpRisk loss comes from N(50000$,5000$)
    - For extreme year (1/10) LN(11,0.1)
    - What is the expected loss?