Quantitative Methods of Decision Making

Slide presentations

Jerzy Mycielski

CMT

2008

Using statistical techniques in business

- The role of statistics is to collect, summarize and analyze the data
- Two branches of statistics:
  - Descriptive statistics
    - Describes the collections of objects (e.g. persons, products, firms) with respect to their characteristics
  - Inferential statistics
    - Techniques which make possible to make conclusions about large collection of objects (all Poles, all employees in the firm) on the basis of small portion of this collection
- The first part of this lecture will describe the methods of descriptive statistics and the second part will cover the inferential statistics

Population and sample

- *Population* is the collection of objects which is of interest of a given investigation
  - population can be finite (e.g. population of firms in Poland) or infinite (e.g. possible values of a price index)
- *Sample* is the part of the population which is used in investigation
  - sample is always finite
- *Sample frame* is the list of all the population members

Random sample

- Simple random sample is sample selected in such a way that each element of the population has equal chance of being selected.
- We say that sample is selected *without* replacement if an element of the population can only be selected once to the sample
- Sampling is almost always done without replacement

Relationship between probability and statistics

Definition

Parameter is a descriptive measure computed from or used to describe the population

- Parameter is nonrandom
- Value of parameter is our object of interest
- However, it is too difficult/costly to collect data in order to calculate the value parameter

Relationship between probability and statistics

Definition

Statistic is a descriptive measure computed from or used to describe the sample

- But: statistic is random as sample is randomly chosen
- Statistic is either used to estimate (approximate) the parameter or to make inference about the properties of the parameter
- As statistic is random it is necessary to use the laws of probability to investigate properties of statistic

Data sources

- Three main sources of data for the researchers and managers:
  - sample surveys
  - designed experiments
  - routine operation
- With sample surveys and designed experiments we collect exactly the data which is needed but they are usually costly
- Routine operation data does not often include the information of interest and it is difficult to gather
- Almost always, the construction of the database is the most costly part of the statistical investigation

Constructing statistical tables

Grouped data

- Ordered array: observations ordered according to their values
- *class intervals*: contiguous, nonoverlapping and exhaustive
  - usually class intervals are of equal width
- *grouped data*: frequency of the occurrence in class intervals

Constructing statistical tables

Definition (Frequency distribution)

Frequency distribution is any device which shows the values of the variable together with the frequency of occurrence of the values

- *cumulative frequency distribution* function: cumulated frequencies from the first class interval through the preceding interval, inclusive
- *relative frequencies*: proportion of observations within certain class interval
- Using relative frequencies we may construct *cumulative relative frequency distribution*
- It is important to define class intervals in such a way to obtain sufficient number of observations in each interval
- To obtain the sufficient numbers of observations in class intervals it is sometimes necessary to define the class intervals with unequal width

Example: Analyzing employment structure

- September survey of the wage structure classified by profession
- Year 2004, Poland
- Population: firms, organizations and individuals employing 10 persons or more
- Exceptions: individual farms, NG0s, political parties, trade unions
- Sample frame: REGON (National register of economic entities in Poland)
- Sampling: all the entities are obliged to fill questionnaires but on the firm level the employed are sampled
- Number of employed included in the sample depends on employment in the entity but not in the linear way
- Sample is not fully random as probability of being included in the sample depends on entity size

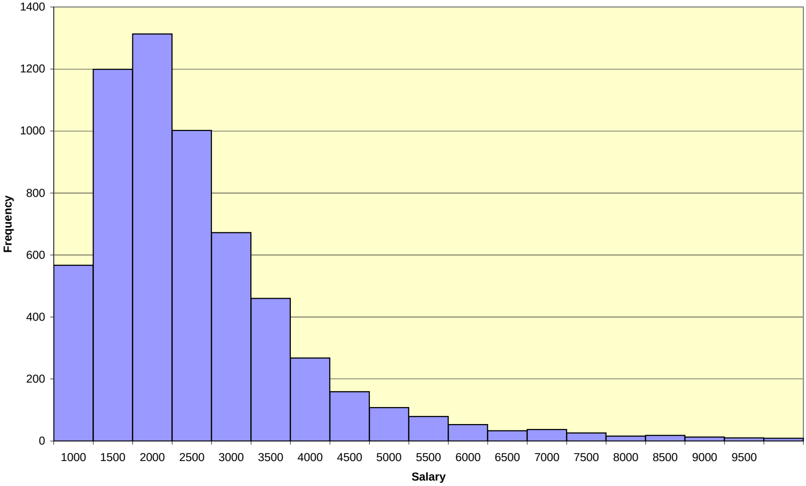
Constructing statistical tables

Histogram - grouped data table

Lower		Upper	*Frequency*	Relative frequency	Cumulative frequency
0	-	1000	567	0.093	0.093
1000	-	1500	1199	0.197	0.290
1500	-	2000	1313	0.216	0.506
2000	-	2500	1002	0.165	0.670
2500	-	3000	672	0.110	0.781
3000	-	3500	460	0.076	0.856
3500	-	4000	268	0.044	0.900
4000	-	4500	159	0.026	0.926
4500	-	5000	108	0.018	0.944
5000	-	5500	79	0.013	0.957
5500	-	6000	53	0.009	0.966
6000	-	6500	33	0.005	0.971
6500	-	7000	37	0.006	0.977
7000	-	7500	26	0.004	0.982
7500	-	8000	16	0.003	0.984
8000	-	8500	18	0.003	0.987
8500	-	9000	13	0.002	0.989
9000	-	9500	10	0.002	0.991
9500	-	10000	9	0.001	0.992
10000	-		46	0.008	1.000
Total			6088		

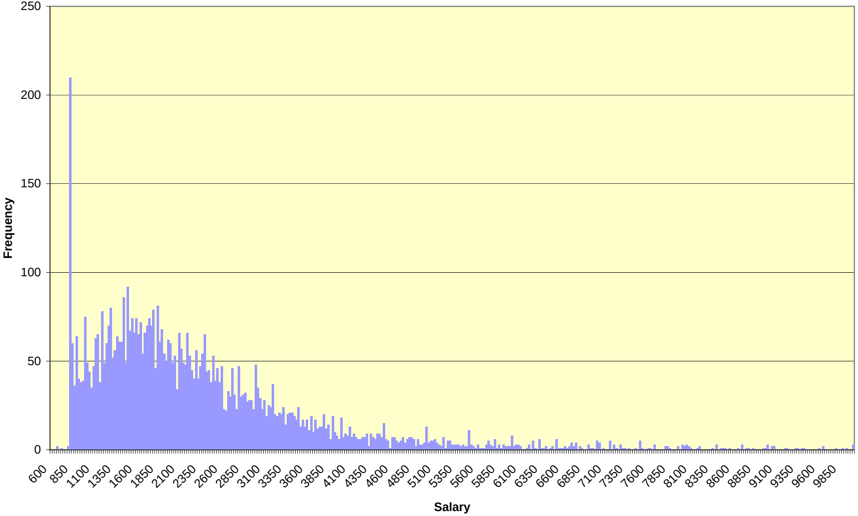
Constructing statistical graphs

Histogram - frequencies



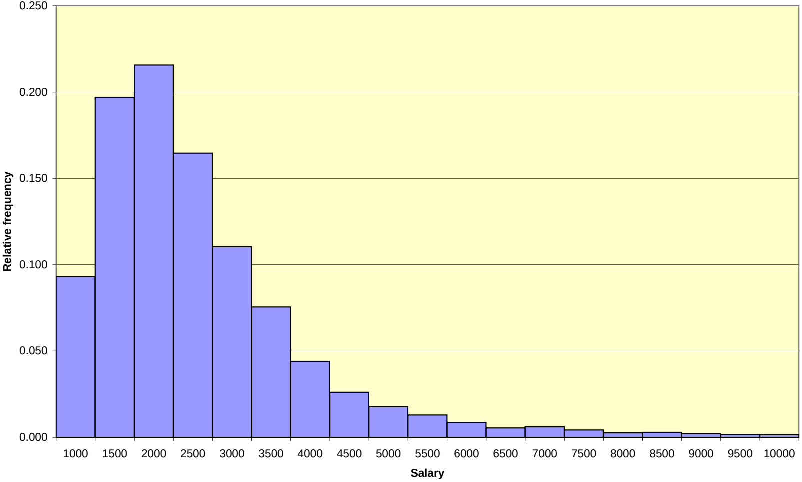
Constructing statistical graphs

Histogram - frequencies



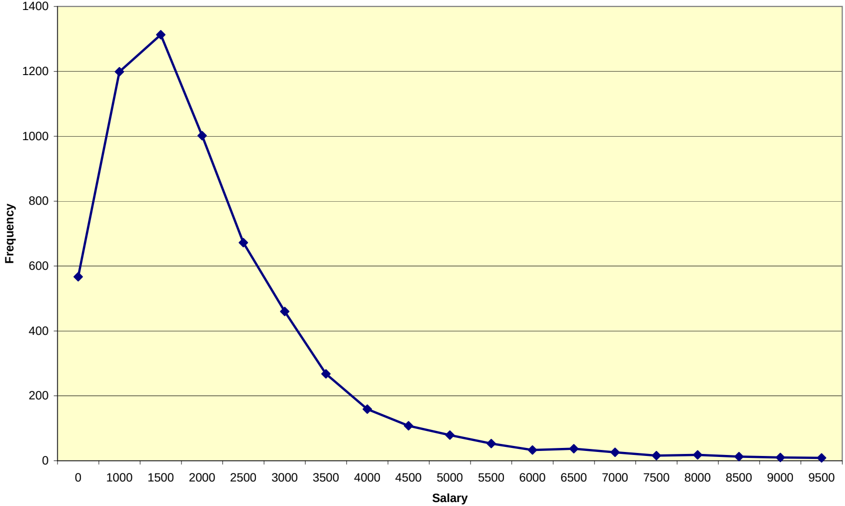
Constructing statistical graphs

Histogram - relative frequencies

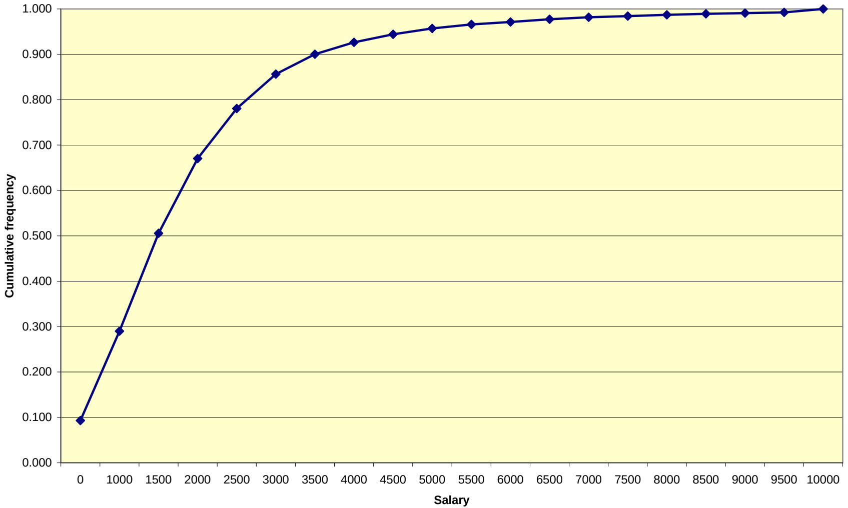


Constructing statistical graphs

Histogram - frequency polygon

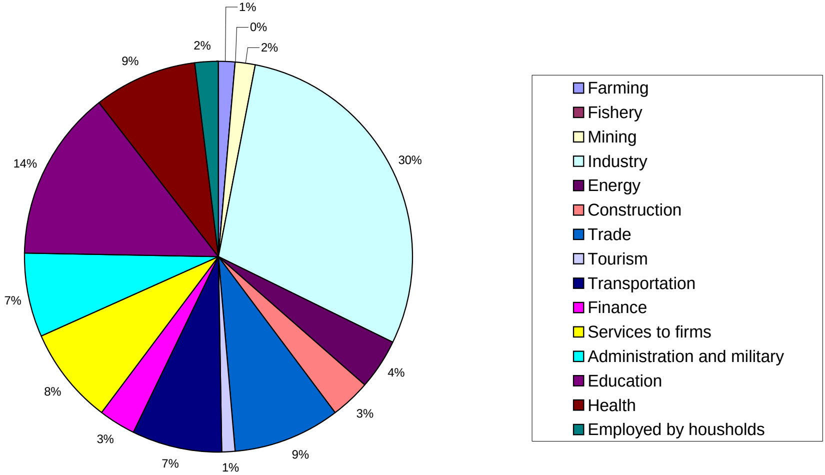


Constructing statistical graphs
Histogram - ogive



Constructing statistical graphs

Pie charts



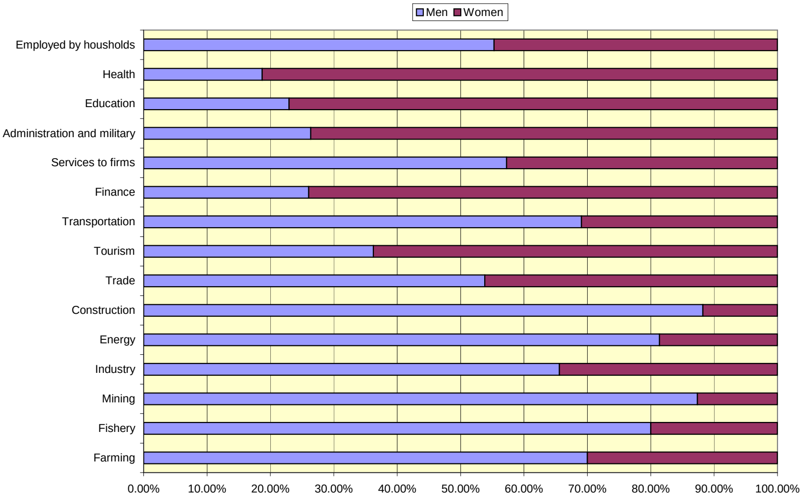
Constructing statistical graphs

Pivot table

	Man	Women
Farming	70.00%	30.00%
Fishery	80.00%	20.00%
Mining	87.38%	12.62%
Industry	65.59%	34.41%
Energy	81.40%	18.60%
Construction	88.24%	11.76%
Trade	53.82%	46.18%
Tourism	36.23%	63.77%
Transportation	69.08%	30.92%
Finance	26.04%	73.96%
Services to firms	57.26%	42.74%
Administration and military	26.35%	73.65%
Education	22.94%	77.06%
Health	18.69%	81.31%
Employed by housholds	55.28%	44.72%

Constructing statistical graphs

Bar charts



Measures of central tendency

Definition (Sample mean)

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Symbol $\sum_{i=1}^{n}$ means "summation from $i = 1$ to $i = n$)

- for a given sample, the value of the mean is unique
- the sum of deviations of observations from the sample mean is equal to zero
- mean is affected by the magnitude of each observation
- mean is additive: the mean of the sum of two characteristics is equal to the sum of means of these characteristics
- Note: sample mean is also called arithmetic average

Definition (Sample median)

Median is the value above which lie the half of the values of observation

- for a given sample median can always be calculated
- if the number of observation is even that it is calculated as a mean of two observations
- the median is not affected by the magnitude of the extreme observations
- median can also be used to characterize the qualitative data
- median is not additive

Measures of central tendency

Definition (Sample mode)

Mode for ungrouped discrete data is the value that occurs most frequently

- for some samples mode does not exist (e.g. all values for observations different)
- it can happen that mode is not unique
- mode is not additive

Example: Analyzing wages of employees

	Mean	Median	Mode
All	2425.3	1986.7	824.0
Men	2634.0	2124.5	824.0
Women	2204.1	1855.0	824.0

Dispersion

- By dispersion we mean the degree to which values in a set vary around their mean
- Other terms for the same concept are variation, scatter, spread
- When values in a set are concentrated around the mean we say that the dispersion is small

Measures of dispersion

Definition (Range)

Range is defined as the difference between the largest and the smallest values in a data set

- The range is usually unsatisfactory measure of dispersion as it is determined only by two most extreme values in the dataset.
- Notice that mean deviation is always equal to zero (see properties of the mean)
- Negative and positive deviation should be treated the same

Measures of dispersion

Definition (Mean absolute deviation)

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

- The mean absolute deviation is an intuitive measure of variation but it is not popular because of troublesome mathematical properties

Definition (Sample variance)

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

- Properties of the variance
  - variance of $y_i = ax_i$ is equal to $s_y^2 = a^2 s_x$ so that the change of units of x results in change of variance which is proportional to the square of a

Measures of dispersion

Definition (Sample standard deviation)

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

- The main advantage of the standard deviation over variance is that for $y_i = ax_i$, the standard deviation $s_y = as_x$.

Measures of dispersion

Definition (Coefficient of variation)

$$c_v = \frac{s_x}{\overline{x}}$$

- Range, mean absolute deviation, variance and standard deviation all depend on the units
- Therefore these measures cannot be to compare dispersion of the characteristics expressed in different units
- As coefficient of variation for $y_i = ax_i$ and x_i are equal the coefficient of variation is dimensionless number
- Therefore it can be used for comparisons of dispersion of variables
- Coefficient of variation should not be used if the mean $\overline{x}$ is close zero

Example: Analyzing the wages of employees

	Max	Min	Range	Variance	Mean absolute deviation	Standard deviation	Coefficient of variation
All	46841.0	659.2	46181.8	3735495.2	1096.8	1932.7	0.80
Men	46841.0	776.6	46064.4	5352659.8	1256.0	2313.6	0.88
Women	21057.3	659.2	20398.1	1926973.5	912.9	1388.2	0.63

Descriptive measures for grouped data

- In the case of grouped data we only know that the observation in a class interval but we do not know the exact value
- Class mark x_i is the midpoint of the interval
- Frequency f_i is the number of observations in the interval
- Sample mean

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$$

- Sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 f_i}{n-1}$$

- Standard deviation

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

Gross domestic product
Production side

- *gross output*: sum of outputs of all the sectors in economy
- *intermediate consumption*: sum of all the products used in production of output
- *gross value added* = gross output-intermediate consumption

$$\left.\begin{array}{l}\text{gross value added}\\\text{intermediate consumption}\end{array}\right\}\text{gross output}$$

- *Taxes and subsides*: *indirect taxes* levied on products
- *gross domestic product (GDP)* = gross value added+taxes-subsidies
- GDP measures the total production of *final goods* in an economy

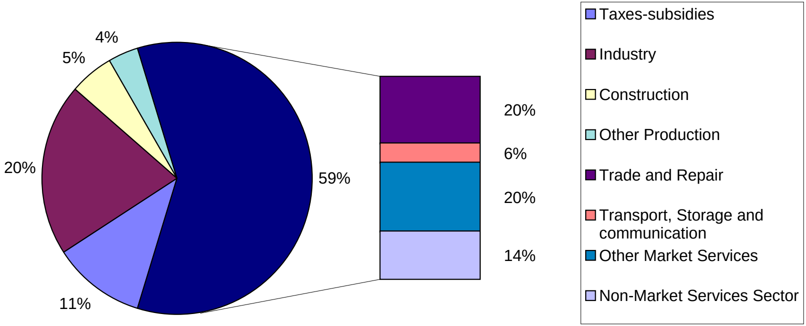
$$\left.\begin{array}{l}\text{gross value added}\\\text{taxes-subsidies}\end{array}\right\}\text{GDP}$$

Example: Analysing Polish Gross Domestic Product

Data - year 2008, 4 quarter

Final Consumption Expenditure	247839,3
Individual Consumption	192669,5
Public Consumption Expenditure	52401
NPISH	2768,8
Gross Capital Formation	54228
Gross Fixed Capital Formation	45348,5
Changes in Inventories	8879,5
Domestic Uses	302067,3
Exports	126996,6
Import	135010,1
Trade balance	-8013,5
Gross Domestic Product	294053,8
Gross Value Added	261666,5
Taxes-subsidies	32387,3
Industry	60961,5
Construction	15036
Other Production	11112,3
Trade and Repair	58359,6
Transport, Storage and communicatio	16454,9
Other Market Services	58895,4
Non-Market Services Sector	40846,8

Example: Analysing Polish Gross Domestic Product
Production side



- *private consumption:* purchases of market products, value of imputed rents for dwellings occupied by owners, etc.
- *public consumption:* value of services in education, culture and national heritage, health care, public administration, national defence, scientific and research activity, etc. provided by the government
- *NPISH*: Non-Profit Institutions Serving Households
- *final consumption* = private consumption+public consumption+NPISH

$$\left.\begin{array}{l}\text{private consumption}\\ \text{public consumption}\\ \text{NPISH}\end{array}\right\}\text{final consumption}$$

- *gross fixed capital formation*: outlays on tangible and intangible fixed assets
- *changes in inventories*: changes in inventories of raw materials, work-in-progress production, and final goods
- *gross capital formation* = gross fixed capital formation+changes in inventories

$$\left.\begin{array}{l}\text{gross fixed capital formation}\\ \text{changes in inventories}\end{array}\right\}\text{gross capital formation}$$

Gross domestic product

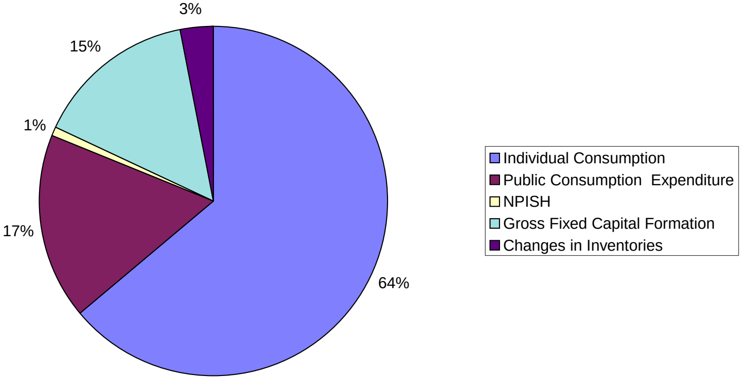
Expenditure side - domestic demand and trade balance

- *domestic demand* = final consumption+gross capital formation
- *foreign trade balance* = exports-imports
- *gross domestic product* = domestic demand-foreign trade balance

$$\left.\begin{array}{l}\left.\begin{array}{l}\text{final consumption}\\ \text{gross capital formation}\end{array}\right\}\text{domestic demand}\\ \qquad\qquad\text{foreign trade balance}\end{array}\right\}\text{GDP}$$

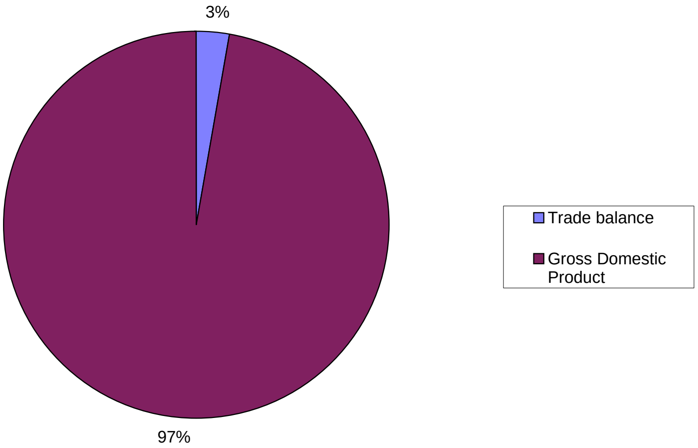
Example: Analysing Polish Gross Domestic Product

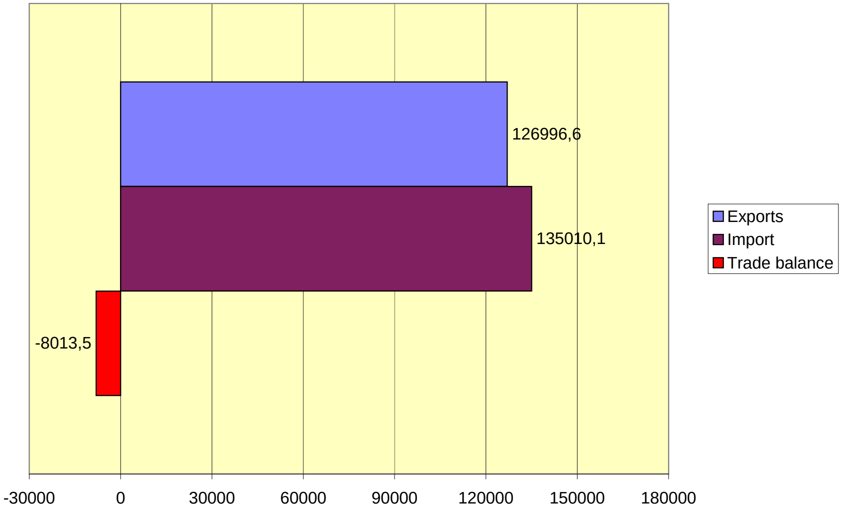
Expenditure side - domestic demand



Example: Analysing Polish Gross Domestic Product

Expenditure side - domestic demand and GDP





126996,6
135010,1
-8013,5
Exports
Import
Trade balance
-30000
0
30000
60000
90000
120000
150000
180000

- Aggregated expenditure in time for items $i = 1, 2, .., k$ is:

$$E = \sum_{i=1}^{k} p_i q_i = \sum_{i=1}^{k} E_i$$

  where p_i represents price of good i, q_i is the quantity of good i bought, and E_i is the expenditure for good i in time t

- An obvious measure of the change of price of good i it the ratio of price its price in time $t = 0$ and $t = 1$

$$P_i = \frac{p_i^*}{p_i}$$

- But, how to measure the change of prices for an aggregated expenditure E?

Price indexes

Constant quantities

- Assume that for all i the the quantities sold in time $t = 0$ and $t = 1$ are the same
- Then the change of prices can be measured as follows

$$P = \frac{\sum_{i=1}^{k} p_{i,1} q_i}{\sum_{i=1}^{k} p_{i,0} q_i} = \frac{\sum_{i=1}^{k} \frac{p_{i,1}}{p_{i,0}} q_i p_{i,0}}{\sum_{i=1}^{k} p_{i,0} q_i} = \sum_{i=1}^{k} P_i \frac{E_i}{E} = \sum_{i=1}^{k} P_i w_i$$

  where $w_i = \frac{E_i}{E}$ is a share of expenditure for good i
- P is then the prices index calculated as a weighted average with weight equal to shares in expenditure
- In practice, the quantities are seldom constant across periods
- The price indexes for aggregate expenditure is calculated with the expenditure pattern from initial or final period

Price indexes

Definition (Lespeyres index)

$$P_L = \frac{\sum_{i=1}^{k} p_{i,1} q_{i,0}}{\sum_{i=1}^{k} p_{i,0} q_{i,0}} = \sum_{i=1}^{k} P_{i,1} w_{i,0}$$

where $w_{i,0} = \frac{E_{i,0}}{E_0}$ is a share of expenditure for good i in time $t = 0$

- Lespeyres index is calculated as a weighted average of price change of individual good with weight equal to shares of this goods in expenditure in *initial period*
- Lespeyres can be interpreted as ratio of the cost of *basket of goods from initial period* bought at prices from final period to the cost of the same basket bought at prices from initial periods

Price indexes

Definition (Paasche index)

$$P_P = \frac{\sum_{i=1}^{k} p_{i,1} q_{i,1}}{\sum_{i=1}^{k} p_{i,0} q_{i,1}} = \sum_{i=1}^{k} P_i w_{i,1}$$

where $w_{i,1} = \frac{E_{i,1}}{E_1}$ is a share of expenditure for good i in time $t = 1$ calculated at prices from time $t = 0$.

- Paasche index is the prices index calculated as a weighted average of price change of individual good with weight equal to shares of this goods in expenditure in *final period*
- Paasche index can be interpreted as ratio of the cost of *basket of goods from final period* bought at prices from final period to the cost of the same basket bought at prices from initial periods

Price indexes – comparison

- The price index should measure how much more/less money we need to maintain the same utility level
- The basket of goods bought is changing over time as consumers react to price changes
- Quantities bought for goods which become relatively more expensive are decreasing
- Then Lespeyres index overstate inflation because it does not take into account the possibility of quantity adjustments
- For similar reason Paasche index understate inflation.
- Lespeyers index is much more popular that Paasche index as it is easier to collect data on prices then on quantities

Price indexes

The consumer price index (CPI) and producer price index (PPI)

- Two types of data used: price data and weight data
- Price data collected from sample of sales outlets
- Weight data taken from household data surveys
- CPI is fixed weight index but seldom a true Lespeyers index as weights are sampled less frequently that prices
- Producer price index is changes of prices domestic producers receive for their products
- This index is now less important as the share of production in GDP is decreasing

- Nominal GDP is GDP calculated at current prices
- Real GDP is defined as GDP at prices from the base period
- GDP deflator is equal to

$$\text{Deflator} = \frac{\text{Nominal GDP}}{\text{Real GDP}}$$

- If base year price level is $t-1$ then

$$\text{Deflator}_t = \frac{\frac{\text{Nominal GDP}_t}{\text{Nominal GDP}_{t-1}}}{\frac{\text{Real GDP}_t}{\text{Nominal GDP}_{t-1}}} = \frac{\text{Nominal GDP growth}}{\text{Real GDP growth}}$$

- GDP deflator is measuring how much of the rise of GDP is caused by changes in prices
- GDP deflator can be used to transform data GDP in nominal terms into GDP in real terms

$$\text{Real GDP growth} = \frac{\text{Nominal GDP growth}}{\text{Deflator}}$$

Price indexes

Example: GDP deflator year 2007

Nominal GDP 2006	**1060031.4**
Nominal GDP 2007	**1175266.3**
Nominal growth GDP	**110.9**
Real growth	**106.7**
GDP deflator	**103.9**

$$\text{Nominal GDP growth} = \frac{1060031.4}{1175266.3} \times 100\% = 110.9\%$$

$$\text{GDP deflator} = \frac{110.9}{106.7} = 103.9$$

- Corresponding period of previous year=100 (quarter to quarter, month to month)
- Chain index: previous period=100 (quarter to previous quarter, month to previous month)
- On the basis of chain index it is possible to approximate inflation between two arbitrary periods
- Index of inflation (no weight changes)

$$P_{t/t-k} = \frac{p_t}{p_{t-k}} = \frac{p_t}{p_{t-1}} \frac{p_{t-1}}{p_{t-2}} \cdots \frac{p_{t-k-1}}{p_{t-k}}$$
$$= P_{t/t-1} P_{t-1/t-2} \cdots P_{t-k-1/t-k}$$

- This kind of index is known as chained index of inflation

Example: Constructing CPI deflator from inflation data. Deflating the employee wages

Month	i/(i-12)	i/(i-1)	CPI deflator (2006 XII = 100%)	Wages in enterprise sector	Real wages in enterprise sector, base 2006 XII
2007 I	101.6	100.4	100.4	2663.6	2652.9
2007 II	101.9	100.3	100.7	2687.5	2668.8
2007 III	102.5	100.5	101.2	2852.7	2818.8
2007 IV	102.3	100.5	101.7	2786.3	2739.4
2007 V	102.3	100.5	102.2	2776.9	2716.6
2007 VI	102.6	100.0	102.2	2869.7	2807.4
2007 VII	102.3	99.7	101.9	2893.7	2839.4
2007 VIII	101.5	99.6	101.5	2886.0	2843.2
2007 IX	102.3	100.8	102.3	2858.8	2794.1
2007 X	103.0	100.6	102.9	2951.7	2867.6
2007 XI	103.6	100.7	103.7	3092.0	2983.1
2007 XII	104.0	100.3	104.0	3246.0	3122.3
2008 I	104.0	100.7	104.7	2969.7	2836.6
2008 II	104.2	100.4	105.1	3032.7	2885.3
2008 III	104.1	100.4	105.5	3144.4	2979.7
2008 IV	104.0	100.4	106.0	3137.7	2961.5
2008 V	104.4	100.8	106.8	3069.4	2874.0
2008 VI	104.6	100.2	107.0	3215.3	3004.6
2008 VII	104.8	100.0	107.0	3229.0	3017.4
2008 VIII	104.8	99.6	106.6	3165.1	2969.6
2008 IX	104.5	100.3	106.9	3171.7	2966.8

- CPI deflator in February 2007

$$\frac{100.4\% \times 100.3\%}{100\%} = 100.7\%$$

- Wage in February expressed in prices from December 2006

$$\frac{2687.5 \text{ zł}}{100.7\%} \times 100\% = 2668.8 \text{ zł}$$

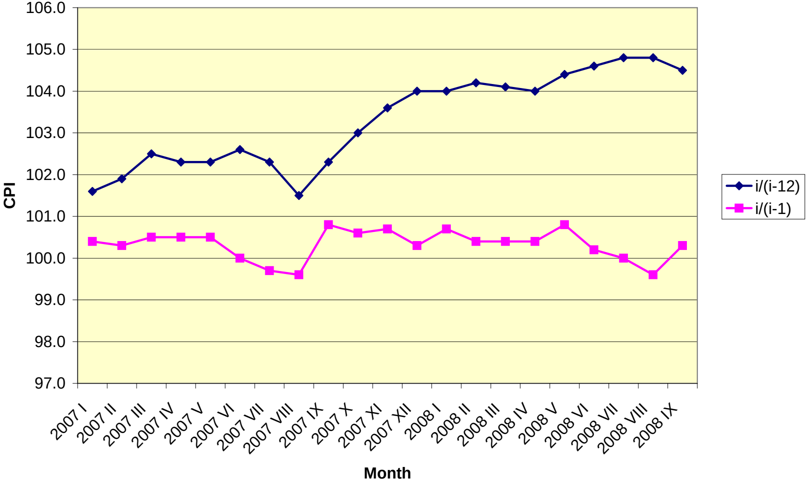
- CPI deflator in March 2007

$$\frac{100.7\% \times 100.5\%}{100\%} = 101.2\%$$

- Wage expressed in prices from December 2006

$$\frac{2852.7 \text{ zł}}{101.2\%} \times 100\% = 2818.8 \text{ zł}$$

Example: CPI in Poland



Time series

- Time series is the sequence of data point measured at successive times
- Time series analysis comprises of methods which
  - can be used to uncover the properties of the time series
  - can be used to forecast the future values of the series
- Base assumption: observations close which are close in time are more closely related than observations further apart

Time series

Definition (Simple moving average)

$$s_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}$$

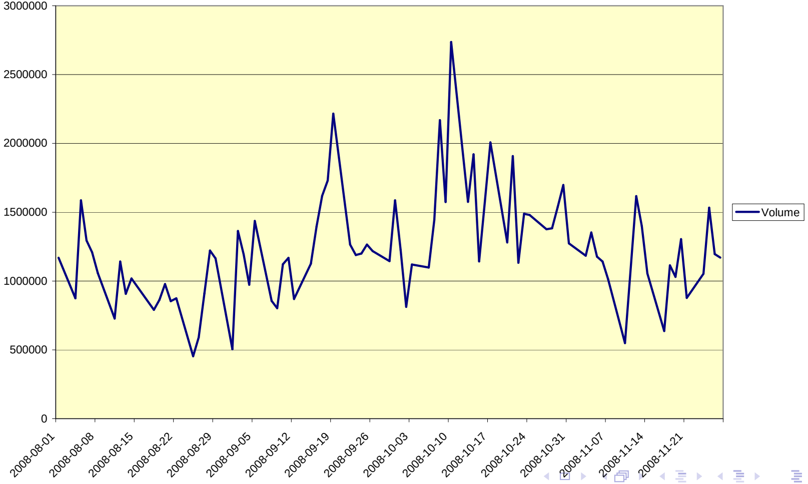
- Notice that

$$s_t = \frac{x_t + x_{t-1} + \ldots + x_{t-k+1}}{k} = s_{t-1} + \frac{x_t - x_{t-k}}{k}$$

- Choice of k is arbitrary - the larger is k, the more smooth is the series
- For smaller k, s_t is more responsive to changes in the series, for larger k, s_t is more smooth
- Notice that for first $k-1$ values of the original series it is not possible to calculate s_t

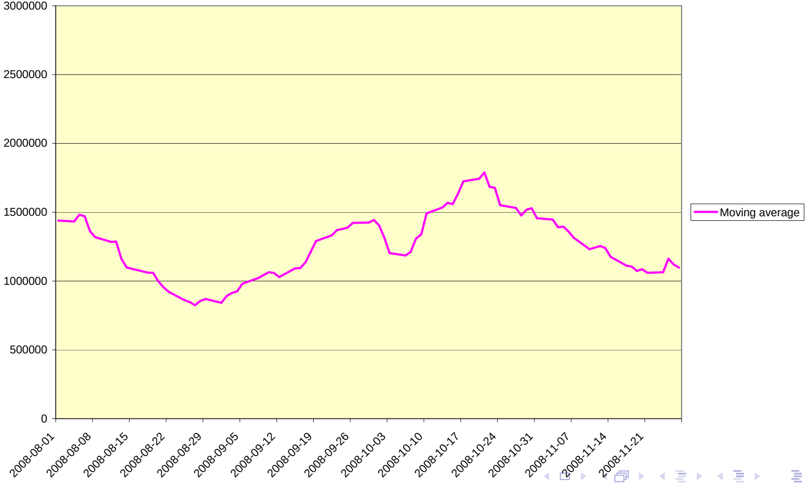
Example: Smoothing the WIG volume index with moving average

Raw data



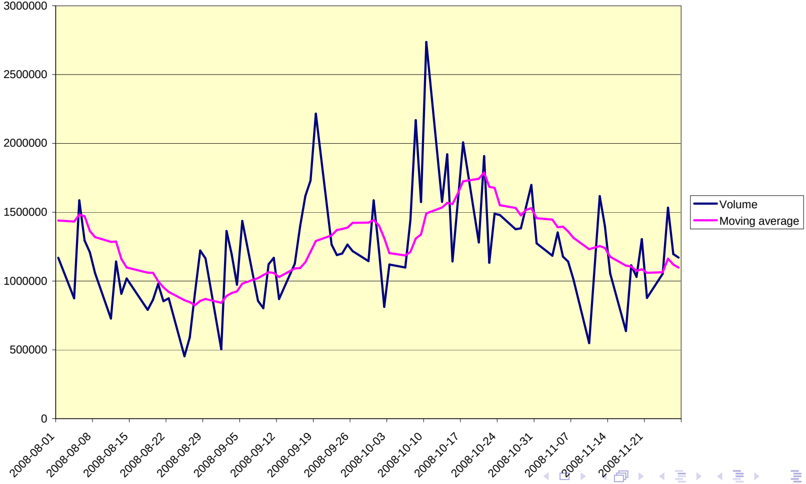
Example: Smoothing the WIG volume index with moving average

Moving average k=10



Example: Smoothing the WIG volume index with moving average

Raw data, moving average k=10



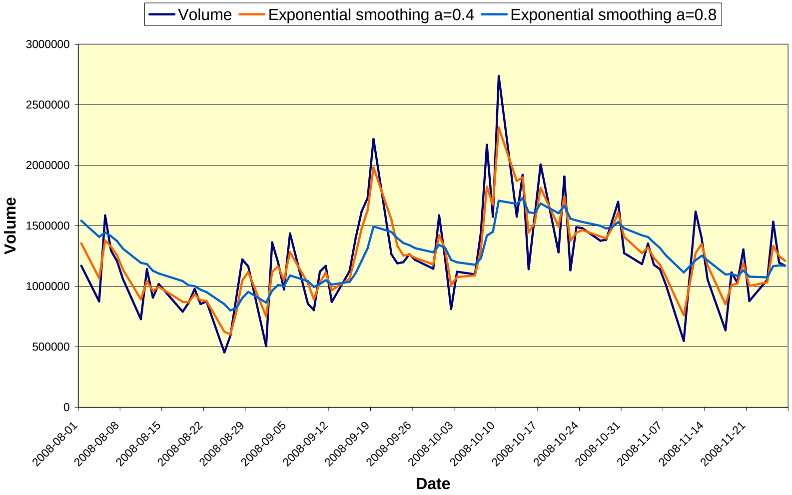
Definition (Exponential smoothing)

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1-\alpha)\, s_{t-1} = s_{t-1} + \alpha \left(x_t - s_{t-1}\right)$$

- The higher is α the more smooth is the smoothed series
- The choice of α is often arbitrary
- Statistical techniques can be used to find optimal value of α by estimation of $ARIMA(0,1,1)$ model

Example: Exponential smoothing of the WIG volume index

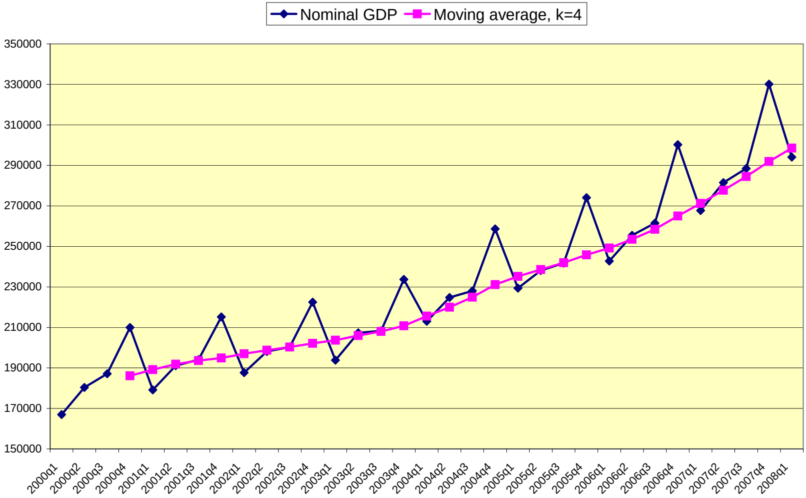
Raw data, exponential smoothing



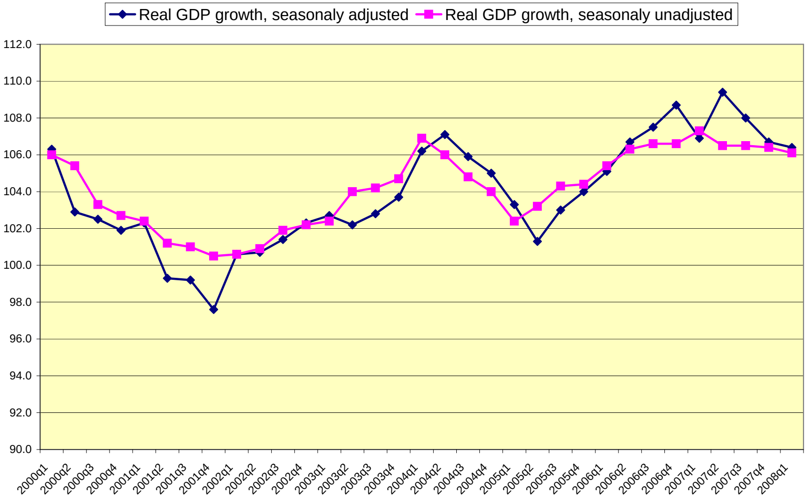
- Seasonality means that in the series we observe periodic fluctuations
- Seasonality is very common in economic time series
- Usually seasonality is either related seasons of the year of the time of the day
- Seasonal adjustment are used in order to remove seasonal effects to better reveal non-seasonal features
- Seasonality can also be used for better forecasting the future value of the series
- Statistical offices most often used for seasonal adjustments either X11 (US) and Tramo/Seats (EU)
- These methods are also adjusting time series for the variation of the number of workdays in a month/quarter

Example: Nominal GDP in Poland

Raw Q/Q data and moving average



Example: Comparing the seasonally adjusted and unadjusted real GDP growth Q/Q data for Poland



- if there is N mutually exclusive and equally likely possibilities of occurrence of an event and if m of these possibilities have characteristic E the probability of E is equal to

$$\Pr(E) = \frac{m}{N}$$

Example

Possible number of pips for a cube dice are $1, 2, 3, 4, 5, 6$, so $N = 6$. If pips are equally likely to obtain than probability of obtaining e.g. 2 pips is equal to $\frac{1}{6}$. The probability of obtaining the even number of pips is equal to $\frac{3}{6} = \frac{1}{2}$

- suppose that some that some process is repeated N times and m of resulting events have characteristics E. For N large, probability of E is approximately equal to

$$\Pr(E) = \frac{m}{N}$$

Example

When we say that the probability of obtaining even number of pips for a dice is equal to $\frac{1}{2}$ we mean that for a large number of rolls for about half of them we obtain an even number of pips.

- Subjective interpretation of probability
    - probability is the measure of the confidence in the truth of certain proposition
- Subjective interpretation of the probability is useful for events which nature is unknown an which cannot be repeated

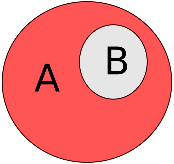
Example

When we say that the probability of depression in Poland in the next year is equal to $\frac{1}{10}$ we it means that we are not really confident that this event will place

- The interpretation of probability does not influence its properties

Set

- Set is a collection of objects
- If all the elements of set B belong to set A we say that B is a subset of A

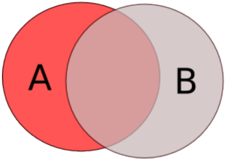


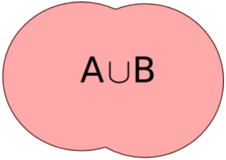
- Empty set is denoted as $\varnothing$

Example

$A = \{1, 2, 3, 4, 5, 6\}$ and $B = \{2, 4, 6\}$. B is a subset of A

- Set which consists of all the elements which are in set A *or* in set B is denoted as $A \cup B$
  - $A \cup B$ is called the *union* of sets A and B

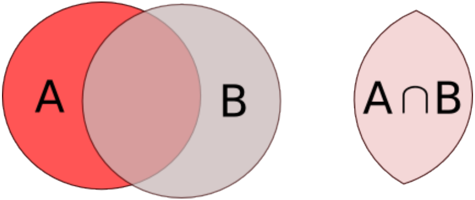




Example

$A = \{1, 3, 5\}$ and $B = \{2, 3, 6\}$, $A \cup B = \{1, 2, 3, 5, 6\}$

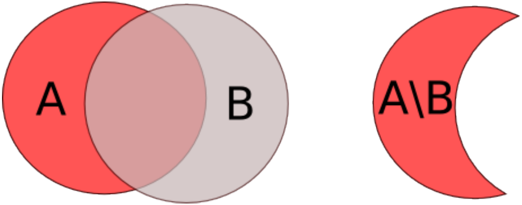
- Set which consists of all the elements which are in set A *and* in set B is denoted as $A \cap B$
  - $A \cap B$ is called the *intersection* of sets A and B



Example

$A = \{1, 3, 5\}$ and $B = \{2, 3, 6\}$, $A \cap B = 3$

- Set which consists of all the elements which are in set A *and not* in set B is denoted as $A \setminus B$
  - $A \setminus B$ is called the *difference* of sets A and B



Example

$A = \{1, 3, 5\}$ and $B = \{2, 3, 6\}$, $A \setminus B = \{1, 5\}$

Fundamental properties (axioms) of probability

- *Outcome* is a possible result of a process of interest
- Set of all possible outcomes is called the *sample space* an denoted as Ω
- *Event* is a subset of sample space

1. Probability of an event E is between 0 and 1

$$0 \leq \Pr(E) \leq 1$$

2. Probability of occurrence of one out of all possible outcomes (probability space) is equal to 1

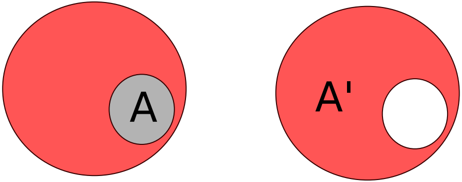
$$\Pr(\Omega) = 1$$

3. If events E_1, E_2 are mutually exclusive $(E_1 \cap E_2 = \varnothing)$ then the probability of occurrence of either E_1 or E_2 is equal to

$$\Pr(E_i \cup E_j) = \Pr(E_i) + \Pr(E_j)$$

Complementary events

- Complementary event consist of all the outcomes which can occur if A does not happen
- Event complementary to A is denoted as A'



- Probability that event A *does not* occur is equal to

$$\Pr\left(A'\right) = 1 - \Pr\left(A\right)$$

Complementary events

Example

The probability of obtaining an odd number of pips when rolling a dice is equal to one minus probability of obtaining even number of pips:

$$\Pr(\text{Odd}) = 1 - \Pr(\text{Even})$$

.

Probability of union (addition rule)

- Probability that event A or event B occurs is calculated is equal to

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

Example

What is the probability of obtaining the number of pips which is even or divisible by 3? Denote by $A = \{2,4,6\}$ and $B = \{3,6\}$ then

$$\Pr(A \cup B) = \Pr(\{2,3,4,6\}) = \frac{4}{6}$$

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \Pr(\{2,4,6\}) + \Pr(\{3,6\}) - \Pr(\{6\}) \\ &= \frac{1}{2} + \frac{1}{3} - \frac{1}{6} = \frac{4}{6}\end{aligned}$$

Independent events

Definition (Independent events)

Random events A and B are independent if

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

Example

Assume that two rolls of a dice are independent. What is the probability of obtaining 6 twice?

Denote probability of obtaining 6 in the first roll as A and in the second roll B. Assume that $\Pr(A) = \Pr(B) = \frac{1}{6}$ then

$$\Pr(A \cap B) = \Pr(A)\Pr(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Conditional probability

Definition (conditional probability)

If we know that random event B have taken place than probability of random event A conditional on this information is called conditional probability of A given B and is equal to

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

- Notice that for independent event A and B

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

- So: the information about an independent event B does not influence our assessment of the probability of event A

Example

We know that the number of pips obtained in a roll is even. What is the probability of obtaining $1, 2, 6$ pips in this roll conditional on this knowledge?

$$\begin{aligned}\Pr(\{1,2,6\} \mid \{2,4,6\}) &= \frac{\Pr(\{1,2,6\} \cap \{2,4,6\})}{\Pr(\{2,4,6\})} \\ &= \frac{\Pr(\{2,6\})}{\Pr(\{2,4,6\})} = \frac{2/6}{3/6} = \frac{2}{3}\end{aligned}$$

Random variable

- Random variable is a variable which value depends on a random event
- The value of a random variable can not be predicted with certainty
- A random variable can be:
  - qualitative: the value such variable have no quantitative interpretation but is coding same attribute (e.g. sex, occupation, place of residence).
  - quantitative
    - discrete: such random variable have integer values or can be transformed into variable with integer values (e.g. number of children, number of visits in a shop)
    - continuous: can take any real value (e.g. spending for food, profit/loss of a firm)
- We will denote the random variables with capital letters and the values of random variables by lowercase letters
- So: $\Pr(X = x)$ denotes the probability of the event that random variable X is equal to x

Independent random variables

- Random variables X and Y are independent if probability of an event that $X = x$ and $Y = y$ is given by

$$\Pr(X = x \cup Y = y) = \Pr(X = x) \Pr(Y = y)$$

for all possible values of y and x

Example

The results of two rolls of the dice can be considered independent if the probabilities of the events are looking as follows

	1	2	3	4	5	6
1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$

Expectation

Definition (Expected value)

For discrete random variable X that is taking values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ respectively, the expected value is equal to

$$\mathrm{E}(X) = \sum_{i=1}^{n} x_i p_i$$

- Notice that as $\sum_{i=1}^{n} p_i = 1$ then the expected value is the weighted average with weights equal to probabilities
- Expected value is the *population mean* of the random variable
- The expected value can be interpreted (under some conditions) as what you expect to be an average value for X calculated for large number of observations
- For any nonrandom number a expected value of $y = a + bX$ is

$$\mathrm{E}(Y) = \mathrm{E}(a + bX) = a + b\,\mathrm{E}(X)$$

Example

What is the expected number of pips for a dice roll?

$$\mathrm{E}(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2}$$

What is the expected value of the number of pips multiplied by 2 plus 1?

$$\mathrm{E}(2X + 1) = 8$$

- Notice that expected value of X can be equal to value which can not be observed for X

Variance

Definition (Variance)

For discrete random variable X that is taking values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ respectively, the variance is equal to

$$\mathrm{Var}(X) = \sum_{i=1}^{n} [x_i - \mathrm{E}(X)]^2 p_i$$

- Notice that for any nonrandom numbers a, b variance of $y = a + bX$ is

$$\mathrm{Var}(Y) = \mathrm{Var}(a + bX) = b^2\,\mathrm{E}(X)$$

- Standard deviation of the random variable is equal to $\sqrt{\mathrm{Var}(X)}$
- Variance of a random variable can be taught of as the *population variance* of random variable

Example

What is the population variance of the number of pips for a dice roll?

$$\begin{aligned}\mathrm{Var}(X) &= \left(1-\frac{7}{2}\right)^2 \times \frac{1}{6} + \left(2-\frac{7}{2}\right)^2 \times \frac{1}{6} + \left(3-\frac{7}{2}\right)^2 \times \frac{1}{6} \\ &+ \left(4-\frac{7}{2}\right)^2 \times \frac{1}{6} + \left(5-\frac{7}{2}\right)^2 \times \frac{1}{6} + \left(6-\frac{7}{2}\right)^2 \times \frac{1}{6} \\ &= \frac{35}{12}\end{aligned}$$

What is the variance of the number of pips multiplied by 2 plus 1?

$$\mathrm{Var}(2X+1) = 4\,\mathrm{Var}(X) = \frac{35}{3}$$

Expectation of the sum of random variables

- The expected value of the sum is equal to the sum of expected values

$$\mathrm{E}\,(aX + bY) = a\,\mathrm{E}\,(X) + b\,\mathrm{E}\,(Y)$$

- This property also holds for a number of variables larger then 2

Example

What is the expected return from package of assets containing 0.2 of asset X and 0.8 of asset Y?

$$\mathrm{E}\,(0.4 \times X + 0.6 \times \, Y) = 0.4\,\mathrm{E}\,(X) + 0.6\,\mathrm{E}\,(Y)$$

Variance of the sum of independent random variables

- The variance of the sum of independent random is equal to the sum variances
$$\operatorname{Var}(aX + bY) = a^2\,\mathrm{E}(X) + b^2\,\mathrm{E}(Y)$$
- This property also holds for a number of variables larger then 2

Example

What is the variance and standard deviation of return from package of assets containing 0.2 of asset X and 0.8 of asset Y assuming that X and Y are independent and have the same variance σ^2?

$$\operatorname{Var}(0.4 \times X + 0.6 \times Y) = 0.4^2 \operatorname{Var}(X) + 0.6^2 \operatorname{Var}(Y) = 0.16\sigma^2 + 0.36\sigma^2 = 0.$$

$$\sqrt{0.4 \times X + 0.6 \times Y} = \sqrt{\operatorname{Var}(0.52 \times \sigma^2)} = 0.72\sigma^2$$

Notice that the standard deviation of the portfolio is smaller that standard deviation of each of the assets being included in the portfolio

Probability distribution of discrete random variables

Definition

Probability distribution of a discrete random variable is a table, function or graph which specifies the all the possible values of the random variable, along with their respective probabilities

Example

Probability distribution of number of pips being the result of the dice roll can be specified as follows

Value of X	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Definition

Cumulative distribution function (cdf) of the random variable is given by function $F(x) = \Pr(X \leq x)$

- Cdf is equal to probability that the random variable X is smaller or equal to x
- Cdf is equivalent to probability distribution as it is possible to calculate the probability of all the events on the basis of Cdf

Example

Probability distribution of number of pips being the result of the dice roll can be specified as follows

Value of X	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{5}{6}$	1

Probability distribution of discrete random variables

Bernoulli distribution

- Bernoulli distribution is the distribution of a variable X which assumes only two values 1 and 0 with probabilities p and $q = 1 - p$
- Usually the variable in question is a qualitative variable and its values are related to some mutually exclusive outcomes
- These outcomes can be related to success or failure of some action, product being not defective or defective etc.
- The expected values of X is

$$\mathrm{E}(X) = 1 \times p + 0 \times q = p$$

- The variance of X is

$$\mathrm{Var}(X) = (1-p)^2 p + (0-p)^2 \times q = q^2 p + p^2 q = (q+p)\, pq = pq$$

Example

What is the expected value and variance of a random variable X which takes value 1 if the number of pips for the dice roll is equal to 1 or 2 and zero otherwise?

- The Bernoulli process is the sequence of independent trials with outcomes coded into random variables X_i having Bernoulli distribution
- The number of trials ended with success $(X_i = 1)$ can be calculated as $Y = \sum_{i=1}^{n} X_i$
- That expected number successes (Y) is equal to

$$\mathrm{E}(Y) = \mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$$

- As trials are assumed to be independent the variance of Y is equal to

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = npq$$

Example

We know that the probability of our product being defective is $\frac{1}{100}$ and that defects are independent. What is the expected value and the variance of the number of defective products among 200 produced?
Define the random variable X_i which is equal to 1 if product i is defective $X_i = 1$ equal to 0 if product is not defective. The number of products which are defective is equal to $Y = \sum_{i=1}^{200} X_i$. The expected number of defective products is equal to

$$\mathrm{E}\left(Y\right) = 200 \times \frac{1}{100} = 2$$

Using the assumption that defects are independent we obtain the variance

$$\mathrm{Var}\left(Y\right) = 200 \times \frac{1}{100}\frac{99}{100} = \frac{99}{50}$$

Definition (Binomial distribution)

$$\Pr(X = k) = \binom{n}{k} p^k q^{n-k}$$

- Binomial distribution gives the probability of the number of successes in Bernoulli process

Example

Calculate the probability that the number of defective products among 5 product is smaller or equal to 2 if the probability of defect is equal to $\frac{1}{4}$ and defects are independent.

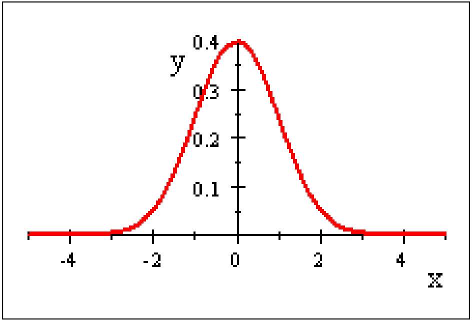
$$\binom{5}{0}\left(\frac{3}{4}\right)^5 + \binom{5}{1}\left(\frac{3}{4}\right)^4 \frac{1}{4} + \binom{5}{2}\left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right)^2 = \frac{243}{1024} + \frac{405}{1024} + \frac{135}{512} = \frac{459}{512}$$

- Density function of a continuous random variable can be thought as an analogue of the relative frequency function
- However, density function can not usually be interpreted as probability of event $\Pr(X = x)$
- For continuous random variable probability of event that $X = x$ is equal to zero
- Density function is of X is often as $f(x)$
- So the density function can be understood as intensity of probability in a given interval.

- Cumulative distribution function of a continuous random variable can be thought as an analogue of cumulative relative frequency function
- As cumulative distribution function of a discrete variable the cdf for continuous variable is define as $F(x) = \Pr(X \leq x)$
- Some properties of $F(x)$
  - $F(x)$ is nondecreasing for all x
  - $F(x)$ goes to 0 for x going to minus infinity
  - $F(x)$ goes to 1 for x going to infinity
  - probability of event $X > x$ can calculated as follows

$$\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - F(x)$$

- The normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called standard normal distribution



Normal density, $\mu = 0$, $\sigma^2 = 1$

- Normal distribution is the most important distribution in statistics

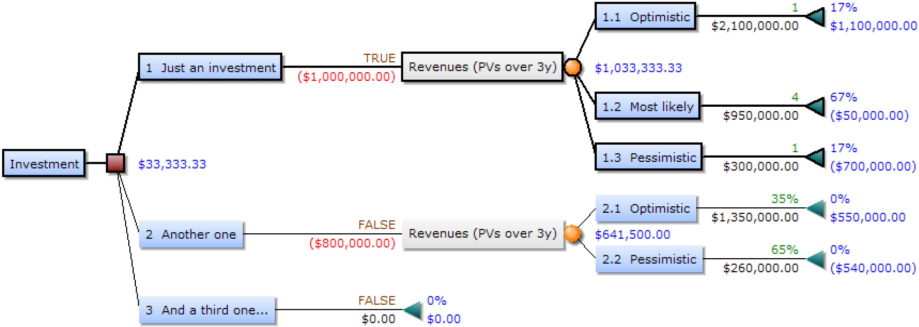
- Normal distribution is symmetric
- The shape of normal distribution is uniquely determined by its expected value μ and variance σ^2
- The sum of variables with normal distribution has also normal distribution
- Applications: for large number of observations the distribution of the sample mean can be approximated with normal distribution

Decision theory

- Deals with the problem of identifying best decisions
- A large part of the theory is concentrated on taking decisions under uncertainty
- The simpler part of the theory is based on assumptions that the decision maker have perfect knowledge about
  - possible outcomes
  - payoffs related to outcomes
  - probabilities of outcomes
- These assumptions are not very realistic

Decision tree

- The decision tree is a graphical representations of the our knowledge about reality



- The decision taken: the one which maximises expected PV from the project
- Is it a valid criterion?

Objective variables

- Objective variables are variables which are maximised/minimised when taking optimal decision
- Simplest case: one objective variable optimized (e.g. profit, utility)
- Typical problem in decision making: how to maximise profit but to minimise the risk?

Payoff table

- Payoff is the value of the objective variable for a given outcome
- Payoff table is a table of payoffs for all possible outcomes for all possible decisions
- For decisions taken under uncertainty the payoff table should also specify the probabilities of outcomes

Expected payoff

- Expected payoff is the expected value of the objective variable for outcomes of a given decision
- The expected payoff from the decision can easily be calculated on the basis of the payoff table

Example: credit scoring

- Choice of the optimal threshold for credit scoring decisions
- Bank has estimated p_i which gives the probability of credit of 10000 zł being paid off for client i
- p_i estimate is based on characteristics of the client i
- Profit if the credit was paid is 2000 zł, average loss if the credit is a default is 10000.
- How the payoff table looks like?
- What is the minimum acceptable value of p_i if bank maximises it expected profit?

Example: credit scoring

Solution

- Payoff table

	success	failure
probability	p_i	$1 - p_i$
accepted	2000	-10000
rejected	0	0

- Expected payoff from accepting
  $a = 2000 \times p_i - 10000 \times (1 - p_i) = -10000 + 12000 \times p_i$
- Expected payoff from rejecting $b = 0 \times p_i - 0 \times (1 - p_i) = 0$
- If $a > b$ application is accepted:

$$-10000 + 12000 \times p_i > 0$$

$$p_i > \frac{5}{6} \approx 0.83$$

- The minimum probability of success for accepted applications 0.83

Utility and risk

- What should be the criterion when we are taking decisions?
- Expected payoff - but then what about the risk?

Example: calculating the maximum acceptable price of insurance policy

- Why people want buy insurance and how it is possible that it is possible to buy insurance?
- Assume that utility function is of the form $U = \sqrt{x}$.
- Say that an individual is considering of insuring his car.
- His car has value of 40000$ and he believe that with probability 10% he can have in a given year an accident reducing the value of the car to 10000$
- Following outcomes are possible

	good outcome	bad outcome
probability	$\frac{9}{10}$	$\frac{1}{10}$
payoff	40000	10000

Example: calculating the maximum acceptable price of insurance policy

Customer side

- Expected utility of payoffs is $\frac{9}{10} \times \sqrt{40000} + \frac{1}{10} \times \sqrt{10000} = 190.0$
- Say that the cost of policy providing full coverage is v.
- What is the maximum amount individual will agree to pay for the policy?
- As the policy provides full coverage, individual is sure that he will have $40000 - v$
- Now calculate for what v the utility of this amount of money is equal to expected utility when not insured

$$190 = \sqrt{40000 - v}$$

$$v = 40000 - 190^2 = 3900$$

- So the maximum the individual will pay for insurance is 3900

Example: calculating the maximum acceptable price of insurance policy

Insurer side

- From the point of view of the insurer the payoff table is the following:

	good outcome	bad outcome
probability	$\frac{9}{10}$	$\frac{1}{10}$
payoff	v	$v - 30000$

- Assume that insurance firm can diversify the risk by selling a lot of insurance policies
- In this risk of insurer is close to zero, and the his expected profit made on the policy is equal to

$$\mathrm{E}\left(\text{Profit}\right) = \frac{9}{10}v + \frac{1}{10}\left(v - 30000\right) = v - 3000$$

- This implies that for all the prices of the insurance policy in between 3000 and 3900 is beneficial both for the insurer and the individual.

Sampling distributions

Definition

Sampling distribution is the distribution of the values of some statistic computed from randomly drawn samples of the same size

Law of large numbers

- The larger is the sample the sample size, the smaller is the size of the variance of the mean
- the smaller is the variance the higher is the probability that the deviation of the sample mean from the population mean is larger than a given value

Theorem (Law of Large Numbers)

For N going to infinity the probability that the value of sample mean is close to population mean goes to one.

The central limit theorem

Theorem (Central Limit theorem)

For large number of observations in the random sample the distribution of the sample mean is close to normal distribution with expected value equal to μ and variance equal to $\frac{\sigma^2}{n}$

- In practice we often replace the population variance σ^2 with sample sample variance ${s^2}_x$
- Approximating sampling distribution with normal distribution we obtain much more precise albeit less robust estimates of the probability

The meaning of “Large sample”

- Large sample is a sample large enough that CLT works
- Is is said that sample with size larger than 30 are big enough to CLT work reasonably well

Example: binomial distribution

- We know that the probability of product to be defective is $p = 0.05$. Calculate the probability that in the random sample containing 100 products we obtain the estimate of p which is smaller then 0.02.
- Estimate of p is given by $\hat{p} = \frac{k}{n}$
- Distribution of $\hat{p}$ is given by binomial distribution with values of k divided by n.
- Exact number: we obtain the estimate of 0.02 for three case $k = 0$, $k = 1$ and $k = 2$. Probability of this event is equal to

$$\binom{100}{0} \times 0.05^0 \times (0.95)^{100} + \binom{100}{1} \times 0.05^1 \times (0.95)^{99} + \binom{100}{2} \times 0.05^2 \times (0.95)^{98} = 0.118\,26$$

Example: binomial distribution

- Normal approximation:
  - expected value $\mu = 0.05$,
  - standard deviation $\sigma = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.05\times0.95}{100}} = 0.02179$
- Standardization $z = \frac{0.02-0.05}{0.02179} = -1.3768$
- Continuity correction $z = \frac{0.025-0.05}{0.02179} = -1.1473$
- Approximated probability
  - without continuity correction $\Phi(-1.3768) = 0.08428$
  - with continuity correction $\Phi(-1.1473) = 0.12562$

Distribution of the sample proportion

- Sample mean for the estimate of sample proportion is given by the mean of n variables with Bernoulli distribution

$$\hat{p} = \frac{k}{n} = \frac{\sum_{i=1}^{N} x_i}{n} = \overline{x}$$

- Variance of $\hat{p}$ is equal $\sigma_p^2 = \frac{pq}{n}$
- Assume that CLT works
- Variable $\hat{p}$ has approximately normal distribution with expected value $\mu = p$ and variance $\sigma_p^2 = \frac{pq}{n}$

Estimator

Definition (Estimator)

Estimator is a statistic which is designed to estimate (approximate) an unknown sample parameter

Example

We do not know before the election the proportion of the people who will vote for politician A. We collect the a sample of answers and calculate the share of answers of people who declare they will vote for A. This share is our estimate of unknown population parameter and the procedure itself defines the estimator.

Confidence intervals

- To provide exact definition of precision of an estimate of population parameter.
- Intuitively high precision means that with high confidence we believe that an estimate is not deviating much from the value of estimate parameter
- We have to specify what we mean by "high confidence" and "not deviating much"

Definition

Confidence interval is an interval which is containing true value of the parameter with a given probability

- The probability specified when defining the confidence interval is called confidence level and is usual denoted as $1 - \alpha$

- Probabilistic interpretation: $100 \times (1 - \alpha)\%$ of the intervals calculated on the basis of the large number of samples contain the true value of the population parameter.
- Practical interpretation: with $100 \times (1 - \alpha)\%$ confidence we believe that the confidence interval contains the true value of the population parameter

- Usually we use the symmetric confidence intervals of the form:

$$\Pr\left(\overline{x} - z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}}, \overline{x} + z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}}\right) = 1 - \alpha$$

- $z_{1-\frac{\alpha}{2}}$ is called reliability coefficient
- $z_{1-\frac{\alpha}{2}}\sigma_{\overline{x}}$ is the precision of the estimate
- Denote as $\Phi^{-1}(\alpha)$ the inverted normal cdf - this function gives such x for which $\Pr(X < x) = \alpha$
- x is normally distributed, standard deviation is known and equal to σ.
  - $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$, $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.
- x is normally distributed, standard deviation not known.
  - $\widehat{\sigma}_{\overline{x}} = \frac{s_x}{\sqrt{n}}$, $z_{1-\frac{\alpha}{2}} = F^{-1}\left(1 - \frac{\alpha}{2}\right)$ where F is the cdf of t-student distribution with $n-1$ degrees of freedom
- x is not normally distributed, standard deviation not known, sample large
  - $\sigma_{\overline{x}} = \frac{s_x}{\sqrt{n}}$, $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

- We already know that the sampling distribution of the mean $\hat{p} = \frac{k}{n}$ for variables with Bernoulli distribution has:
  - expected value p
  - variance $\frac{pq}{n} = \frac{p(1-p)}{n}$
- For large sample variance of $\widehat{p}$ can be estimated as $s_p^2 = \frac{\widehat{p}(1-\widehat{p})}{n}$ and standard deviation as $s_p = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
- For large sample we are using normal approximation
- Confidence interval is then of the form $\hat{p} \pm z_{1-\frac{\alpha}{2}} s_p$

Selecting the sample

- When selecting the sample th following should be taken into account
  - precision of estimates needed - sample size
  - cost of selecting the sample with a given size
  - sampling method - sample frame
  - representativness of the sample

Determing the sample size fo estimating means

- In order to determine the sample size we have to specify
  - precision d
  - confidence level α
- We also need some preliminary estimate of the standard error s_x
- Usually we obtain such an estimate from pilot survey - small survey done before the main survey
- The sample size can then be determined as follows

$$d = z_{1-\frac{\alpha}{2}} \frac{s_x}{\sqrt{n}}$$

- So

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}}{d} \right)^2$$

- Notice: the higher is the needed precision and confidence level the begger sample we need

Determing the sample size for estimating proportions

- In the case of estimating proportions we also have to specify d and α
- We should have the preliminary estimate of p
- This estimate is usually formulated on the basis of pilot survey
- Making use of the formula derived before and the formula for varinace in this case we obtain

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}}{d} \right)^2 \widehat{p} \left(1 - \widehat{p} \right)$$

Statistical inference – hypothesis testing

General considerations

- In many real problems we have to make a decision on the basis of information from the sample
- Usually this decision is based on some feature of the sample
- The statement of the feature in question is called *null hypothesis* and denoted as H_0

Example

A drug can only be accepted if it can be shown that it is effective in curing some disease. So it producer of the drug is required by law to demonstrate that by the ill people given the drug significantly improved in comparison to control group of ill people who were not given the drug. The null hypothesis in this case can be formulated as follows: there is no difference between the state of health of the people who were given the drug and the ones who were not given it. The task of the drug company is to show that the null hypothesis is false!

- When deriving the sampling distribution of statistics we assume that null hypothesis is true
- Apart from null hypothesis we also define the alternative hypothesis

Example

Denote the productivity in factories A,B as μ_A and μ_B. Manager is checking whether factory A is more productive that factory B. He is formulating his null hypothesis as $H_0 : \mu_A = \mu_B$ and his alternative as $H_1 : \mu_A > \mu_B$

Statistical inference – hypothesis testing

Type I and Type II errors

- When testing the hypothesis and making decision we can make two errors: type I error and type II error.
- Probability of type I error is called significance level or size of the test
- significance level is usually denoted as α
- we say that H_0 can be rejected at high significance level if it can be rejected for very small α
- We control the probability of type I error by setting the significance level
- We cannot control the probability of type II error (power of the test)

Decision H_0 true H_1 true H_0 true OK type I error H_1 true type II error OK

- When choosing the significance level we have to take into account that the higher is the significance level (the smaller the type I error), the higher is probability of type II error.
- Conventional significance levels used in statistics are 0.1%, 1%, 5%, 10%
- But this are only conventions!
- Choosing the significance level for a test on which the decision is based you should take into account the payoffs table - loses related to type I, type II erors and and gains associated with correct decisions.

- Assume that statistics Z used for testing have sampling distributions under H_0 is known
- Decision rule traditionally was that H_0 is rejected if statistics $Z > z$
- z is called critical value
- Consider the probability that statistics Z is larger than some value z if H_0 is true (probability of type I error)
- This probability is equal to $Pr(Z > z) = 1 - Pr(z < z) = 1 - F(z)$ where $F(z)$ is the cdf of sampling distribution of the test statistics
- For a significance level exogenously given, critical value can be calculate as $1 - F(z) = 1 - \alpha$ and then $z = F^{-1}(\alpha)$
- A more modern approach is to calculate $F(z)$ (which is called p-value) and to compare it with α.
- The decision rule which is equivalent to the previous one is the following: reject H_0 if p-value is (smaller) that the assumed significance level α

- In the case of testing $H_0 : \mu_0$ there are three possible versions of alternative hypothesis:
  - $H_1 : \mu = 0$
  - $H_1 : \mu > 0$
  - $H_1 : \mu < 0$
- The first version of the H_1 results in two sided test
- The second two versions of H_1 results in one sided test.
- The choice depends on research question or decision context

Testing hypothesis about the mean – unknown population variance

- We use the sample mean in order to verify a hypothesis about the mean in population
- The simplest case $H_0 : \mu = \mu^*$, $H_1 : \mu \neq \mu^*$
- If σ^2 is not known and x is normally distributedthan test statistic is:

$$t = \frac{\overline{x} - \mu^*}{s_x / \sqrt{n}}$$

  and t has t-student distribution with $N - 1$ degrees of freedom
- If σ^2 is unknown, and x is not normally distributed but sample is large than test statistic is:

$$z = \frac{\overline{x} - \mu^*}{s_x / \sqrt{n}}$$

  and have approximately standard normal distribution

Testing hypothesis about the mean – unknown population variance

- If σ_1^2, σ_2^2 unknown and x_1, x_2 are normally distributed:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_1^2}{n_1+n_2-2}$, and t has t-student distribution with $n_1 + n_2 - 2$ degrees of freedom

- If σ_1^2, σ_2^2 unknown unknown, and x_1, x_2 are not normally distributed but sample is large than test statistic is:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_x^2}{n_2}}}$$

and have approximately standard normal distribution

Testing hypothesis about the population proportion

- Null hypothesis $H_0 : p = p^*$, $H_1 : p \neq p^*$
- The null hupothesis can also be of the form $H_1 : p < p^*$, $H_1 : p > p^*$
- Statistics
$$z = \frac{\widehat{p} - p^*}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}$$
- If sample is large, the test statistics is approximately normally distributed

Relation between hypothesis testing and interval estimation

- Hypothesis testing and interval estimation have much in common
- In the case of interval estimation you construct an interval which covers the population parameter with given probability.
- In the case of hypothesis testing we assume the value of the parameter and check wheter our estimate is in the acceptance region.

- Total variation in the sample:

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - \overline{\overline{x}}\right)^2$$

- $\overline{\overline{x}}$ is the overall for all observations in all subgroups

$$\overline{\overline{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n}$$

- Variation which cannot be explained with the differences among means across subgroup
- This varion is also called within variation:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \overline{x}_j)^2$$

- $\overline{x}_j$ is the mean for subgroup j

$$\overline{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

- Variation which can be explained with the differencess among means across subgroup
- This sum of
- This varion is also called between variation:

$$SSA = \sum_{j=1}^{k} n_j \left(\overline{x}_j - \overline{\overline{x}}\right)^2$$

	SS	DF	MS	F
Between	SSA	$k-1$	$\frac{SSA}{k-1}$	$\frac{MSA}{MSE}$
Within	SSE	$n-k$	$\frac{SSE}{n-k}$	
Total	SST	$n-1$	$\frac{SST}{n-1}$	

One-way analysis of variance

- The null hyphothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

- Alternative hypothesis: not all the subgroup means are the same
- H_0 rejected if $F = \frac{MSA}{MSE}$ larger than critical value from F distribution with k and $n - k - 1$ degrees of fredom

Example: predicting the price of apartment

- Sample covariance of variables X and Y defined as

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y})}{n-1}$$

- Sample correlation coefficient is defined as

$$\widehat{\rho}_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Covariance and correlation coefficient are measures of dependence between variables
- Correlation coefficient is always between -1 and 1
- If sample covariance or correlation coefficient is positive it means that we tend to observe that observations for X and Y tends to deviate in the same direction from the mean.
- In the case of negative correlation the variables tend to deviate from the mean in opposite directions
- If the correlation coefficient is positive we say that variables are positively correlated, if it is negative we say that they are negatively correlated

Definition (Covariance)

$$\mathrm{Cov}(X, Y) = \mathrm{E}\{[X - \mathrm{E}(X)][Y - \mathrm{E}(Y)]\}$$

Definition (Correlation coefficient)

$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

- The above formulas are defining the population variance and population correlation coefficient
- The interpretation and properties are similar to the sample analogues

Distribution of two or more random variables

Covariance and correlation coefficients, special cases

- Special cases:
  - if $\rho_{xy} = 1$ then $X = Y$, perfect positive correlation
  - if $\rho_{xy} = 0$ variables are not correlated
  - if $\rho_{xy} = 1$ then $X = Y$, perfect negative correlation
- But: independence implies the lack of correlation but the lack of correlation does not imply independence
- Independence is a stronger property

- Explained variable is the variable which is to be explained by the model
- Explanatory variable is the variable is which is explaining the behavior of the expained variable
- Regression is the dependence of the expected value of the explained variable on explanatory variable
- In simple regression the dependence between the explained variable y and explanatory variable x is of the linear form

$$y_i = \alpha + \beta x_i + \varepsilon$$

- ε is the error term or unexplaned devations of y_i from the regression line
- Estimators of population parameters α and β are choosen in such a way to minimize devations of observations from the estimated regression line.

- The estimators of unknown population parameters α and β of simple regression can be shown to be equal to
  - estimator of β

$$a = \frac{s_{xy}}{s_x}$$

  - estimator of α

$$b = \overline{y} - b_1\overline{x}$$

- Using estimates of α and β we can formulate the prediction of y_i given our simple model of dependence
- Prediction is given by
$$\widehat{y} = a + bx$$
- Prediction of y_i can be formulated for observations in the sample or out of the sample

- Coefficient of determinantion R^2 is defined as

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

- Important property of R^2:

$$0 \leq R^2 \leq 1$$

- It can be intepreted as the percent of total variation of explained variable which is explained by explanatory variable x
- For simple linear regresion it can be shown that

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \widehat{\rho}^2$$

Significance tests for explanatory variable

- The hypotesis which is most frequently test in the context of linear regression is the hypotesis $H_0 : \beta = 0$
- This hypothesis can be intepreted as the hypothesis of the lack of the dependence between y and x
- The test statistics used for testing H_0 in simple is the following:

$$t = \frac{s_{y|x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

where

$$s_{y|x} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n-2}$$

Sometimes it is easier to use the formula:

$$t = \sqrt{\frac{(n-2)\,R^2}{1-R^2}}$$

- It has the t-Student distribution with $n-2$ degrees of freedom