RESEARCH METHODOLOGY

Olga Konnikova

Ass.Prof. of Marketing Department, Saint-Petersburg State University of Economics

PhD in Economics

E-mail: Konnikova.o@unecon.ru

AGENDA

1. Quantitative research in Management: methodology. Introduction to IBM SPSS – September 6
2. **Data visualization. Descriptive statistics. Cross-tabulating (Contingency tables) – September 13, October 11**
3. Analysis of variance (dispersion analysis)
4. Correlation and regression analysis
5. Cluster analysis
6. Summary

DESCRIBING DATA: *«FIRST SIGHT ON THE DATA»*

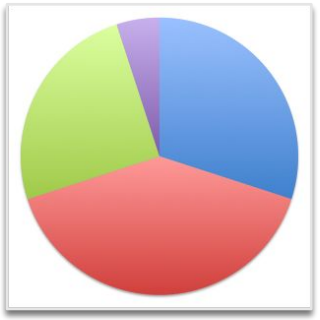
Graphical description

- E.g., histograms (to identify outlines – «*выбросы*»)

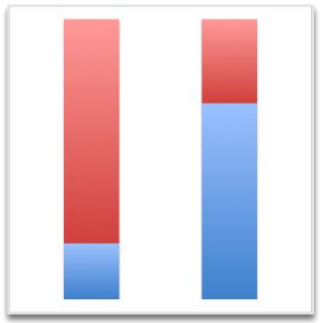
Numerical descriptive measures

- Median, mode
- Range, Minimum, Maximum
- Mean, Standard deviation
- …

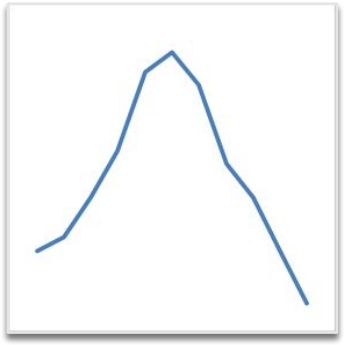
GRAPHICAL DESCRIPTION



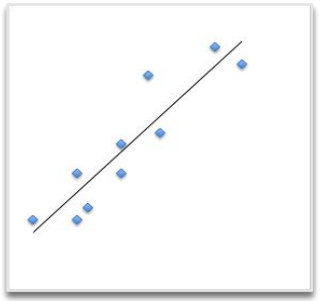
pie chart

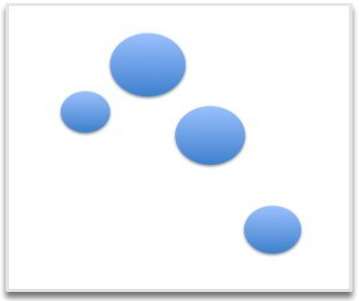


bar
charts



line (graph) – used for showing the tendency *(through time!)*

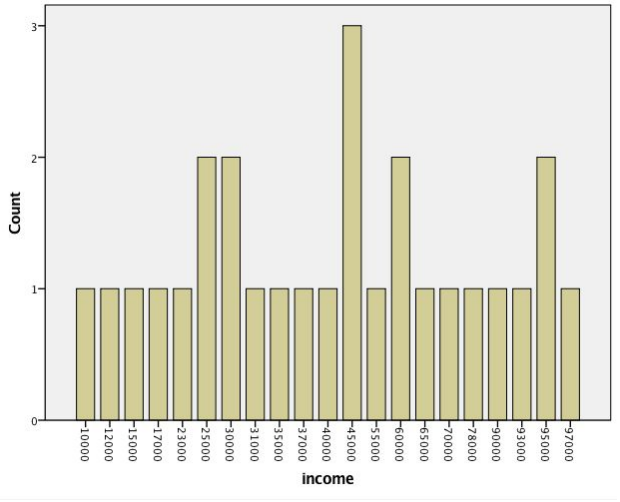


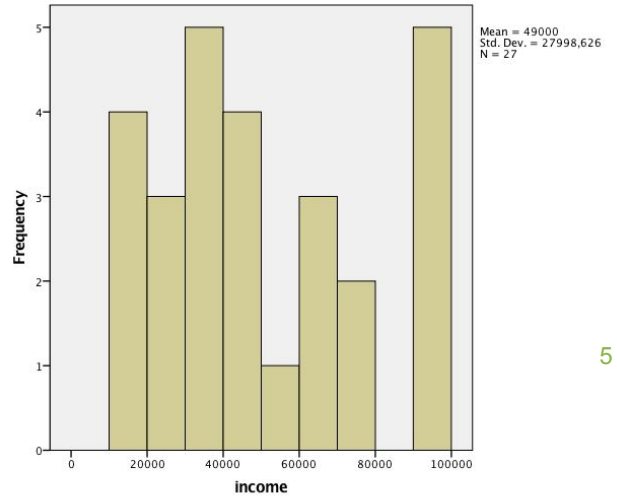


scatterplots and bubbles - used for comparison of ***two*** variables

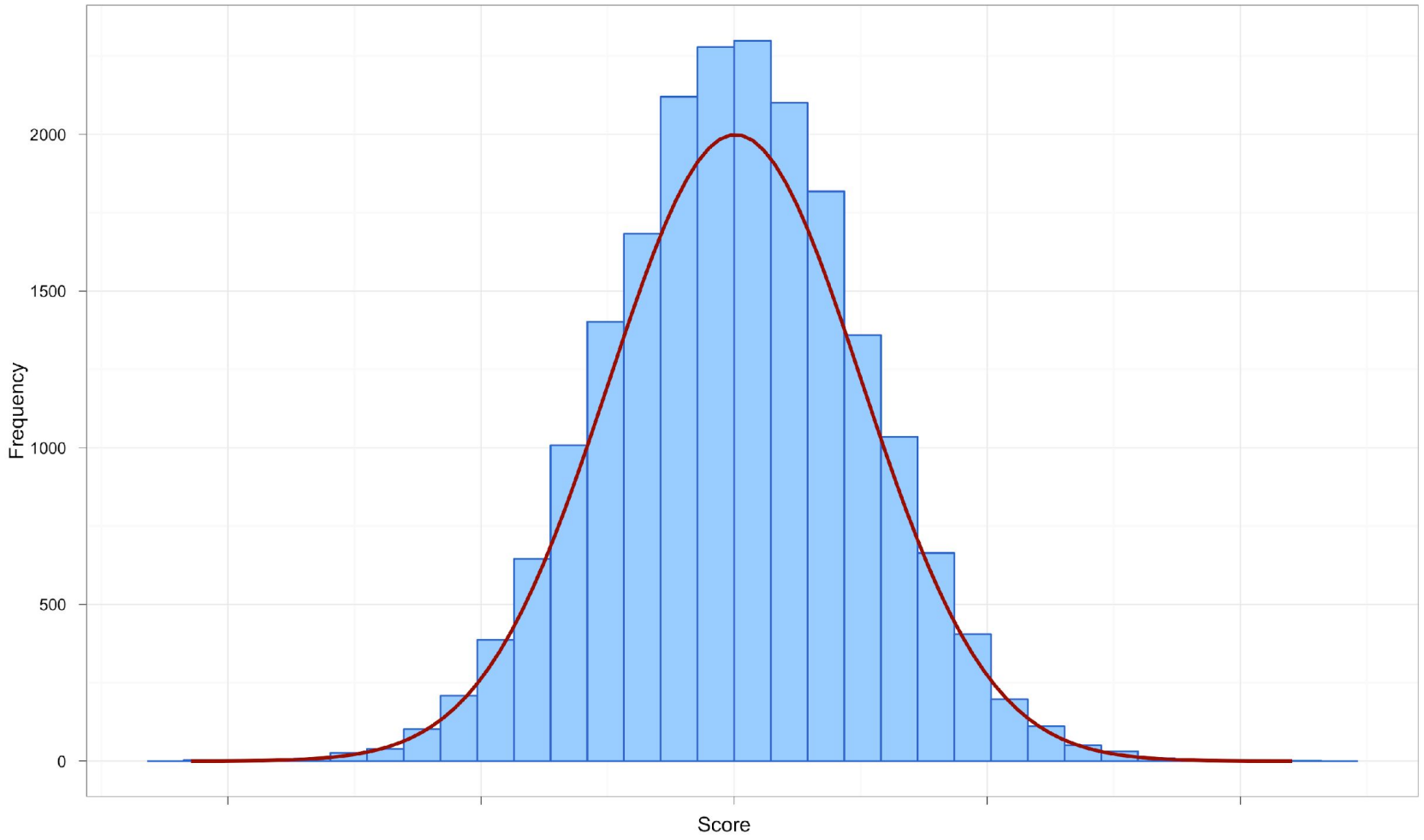
GRAPHICAL DESCRIPTION: HISTOGRAM

- Histograms are used for graphical representation of quantitative scaled variables
- Histograms show the comparison of not the values of the observation but the ***frequency*** of values
- For this purpose, histogram automatically divides values of the observation into certain intervals for the convenience of interpretation
- **Histogram** - a graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set





THE NORMAL DISTRIBUTION



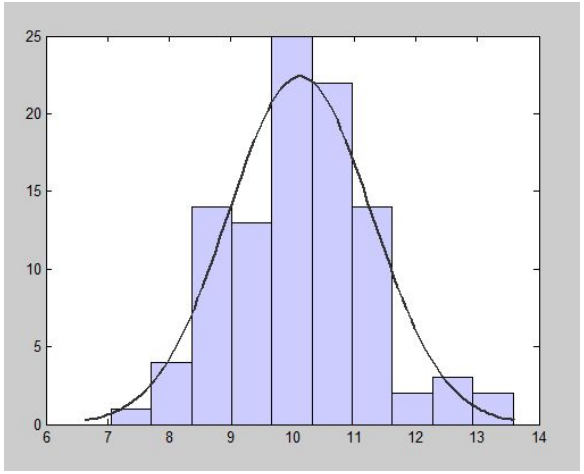
GRAPHICAL DESCRIPTION: HISTOGRAMS AND NORMAL DISTRIBUTION

The 'Normal' distribution

- Bell *(«колокол»)* shaped
- Symmetrical around the center
- No outlnine cases

TEST OF NORMALITY:
HOW TO TEST IF THE DATA IS NORMALLY DISTRIBUTED?

1st way: To look at the histogram (*Graphs – Legacy Dialogs – Histogram / Tick "Display normal curve"*)



TEST OF NORMALITY:
HOW TO TEST IF THE DATA IS NORMALLY DISTRIBUTED?

2nd way: To conduct Kolmogorov-Smirnov ***OR*** Shapiro-Wilk test of normality

- We use Kolmogorov-Smirnov criterion if we have large sample (more than 60 observations)
- We use Shapiro-Wilk criterion if we have small sample (less than 60 observations)

TEST OF NORMALITY IN SPSS

Analyze – Descriptive Statistics – Explore / Plots / Tick “Normality plots with tests”

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Spendings for FC per month	,118	25	,200*	,967	25	,564

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

TEST OF NORMALITY: CONDUCTION

- H_0: sample is not normally distributed
- H_1: sample is normally distributed

- We fix ***significance level (α), e.g. 5%***
- *We can calculate **p-value** in SPSS (we conduct the appropriate test procedure)*

- If ***p-value>α*** than we accept main hypothesis $\boldsymbol{H_0}$
- If ***p-value<α*** than we accept alternative hypothesis $\boldsymbol{H_1}$

WHY NORMAL DISTRIBUTION IS IMPORTANT ?

TYPE OF VARIABLE		INDEPENDENT VARIABLE	
		Quantitative scale	**Nominal / Ordinary scale**
DEPENDENT VARIABLE	**Quantitative scale**	Correlation and regression analysis	Analysis of variance (dispersion analysis)
	Nominal / Ordinary scale	Discriminant analysis	Cross-tabulating (Contingency tables)

Some types of data analysis are appropriate only for normally distributed variables or closed to them

How to make data more normally distributed?

DESCRIBING DATA: *«FIRST SIGHT ON THE DATA»*

Graphical description

- E.g., histograms (to identify outlines – «*выбросы*»)

Numerical descriptive measures

- Median, mode
- Range, Minimum, Maximum
- Mean, Standard deviation
- …

DESCRIPTIVE STATISTICS

Analysis of the basic statistical parameters in order to get acquainted with the data, to reveal its features, to correct the hypotheses.

Descriptive statistics is carried out in different ways depending on which scale the variables are measured in:

- Nominal
- Ordinal
- Quantitative

DESCRIPTIVE STATISTICS: MAIN INDICATORS

- Mode «мода»
- Median «медиана»
- Range «размах»
- Minimum
- Maximum
- Mean (=average) «среднее»
- Standard deviation «стандартное отклонение»

DESCRIPTIVE STATISTICS: THE MODE

- **Mode** – the most frequent observation, typical observation, represents most frequent category

Category *e.g. some brand*	Number of Observations
A	57
B	38
C	86
D	45
E	119
F	42

DESCRIPTIVE STATISTICS: THE MODE

Mode

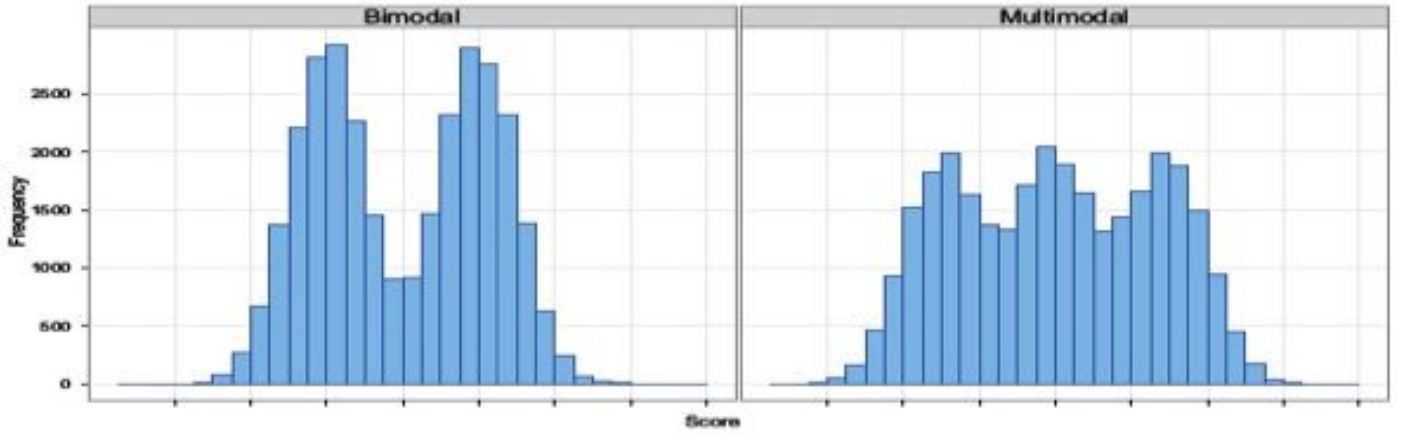
- The most frequent score

Bimodal

- Having two modes

Multimodal

- Having several modes



DESCRIPTIVE STATISTICS: THE MEDIAN

Median – the value that is in the middle: half of the observations are higher than median and half of the observations are lower than median

The median is the middle score when scores are ordered:

- ***Ex. 1.*** *Median(15,27,14,18,21) = Median(14,15,**18**,21,27) = 18*
- ***Ex. 2****. Median(15,27,14,18) = Median(14,**15,18**,27) = (15+18)/2 = 16,5*

Category	Number of Observations
A	57
B	38
C	86
D	45
E	119
F	42

DESCRIPTIVE STATISTICS: RANGE, MINIMUM, MAXIMUM

Range

- The smallest / lowest score (minimum) subtracted from the largest / highest score (maximum)

Category	Number of Observations
A	57
B	38
C	86
D	45
E	119
F	42

DESCRIPTIVE STATISTICS: THE MEAN

Mean

- The sum of scores divided by number of scores

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

Category	Number of Observations
A	57
B	38
C	86
D	45
E	119
F	42

DESCRIPTIVE STATISTICS: STANDARD DEVIATION

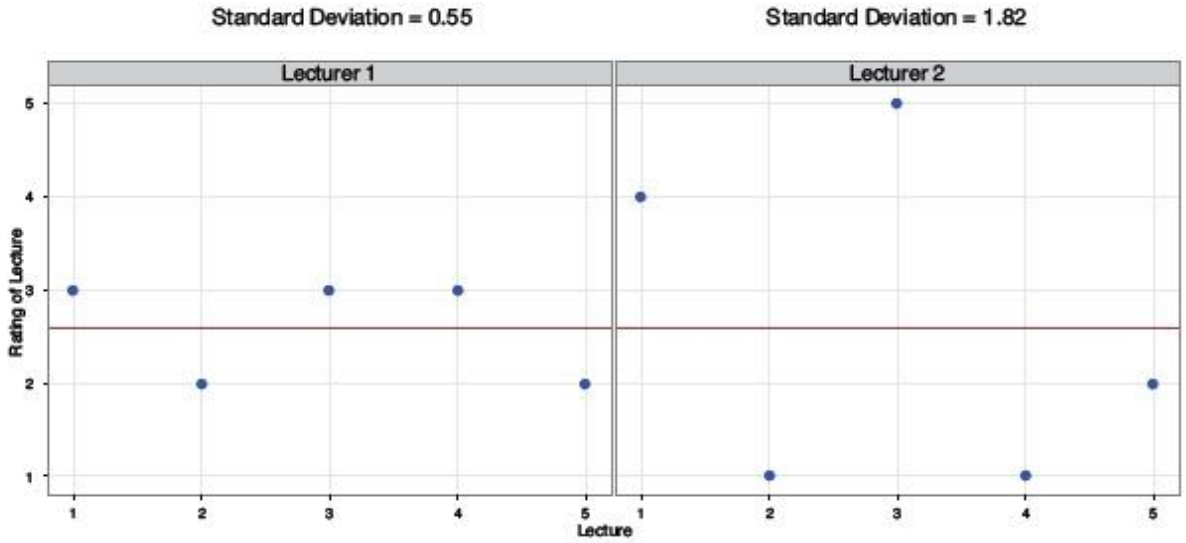
Standard deviation

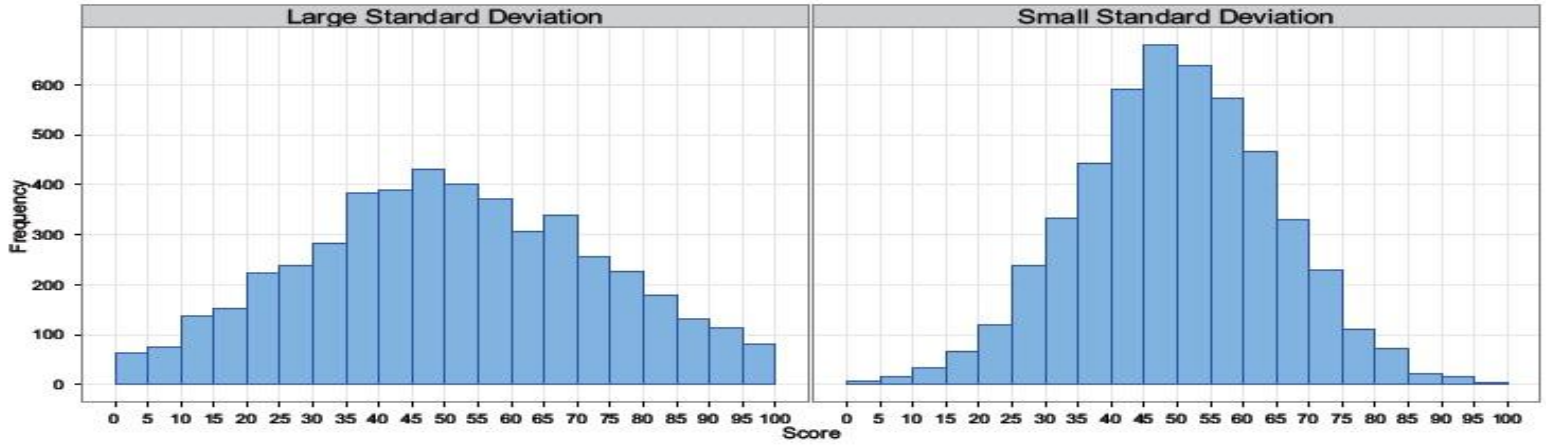
- the most common indicator of the dispersion of values of a random variable with respect to its mathematical expectation (in most cases the mathematical expectation = the mean)

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

Category	Number of Observations
A	57
B	38
C	86
D	45
E	119
F	42

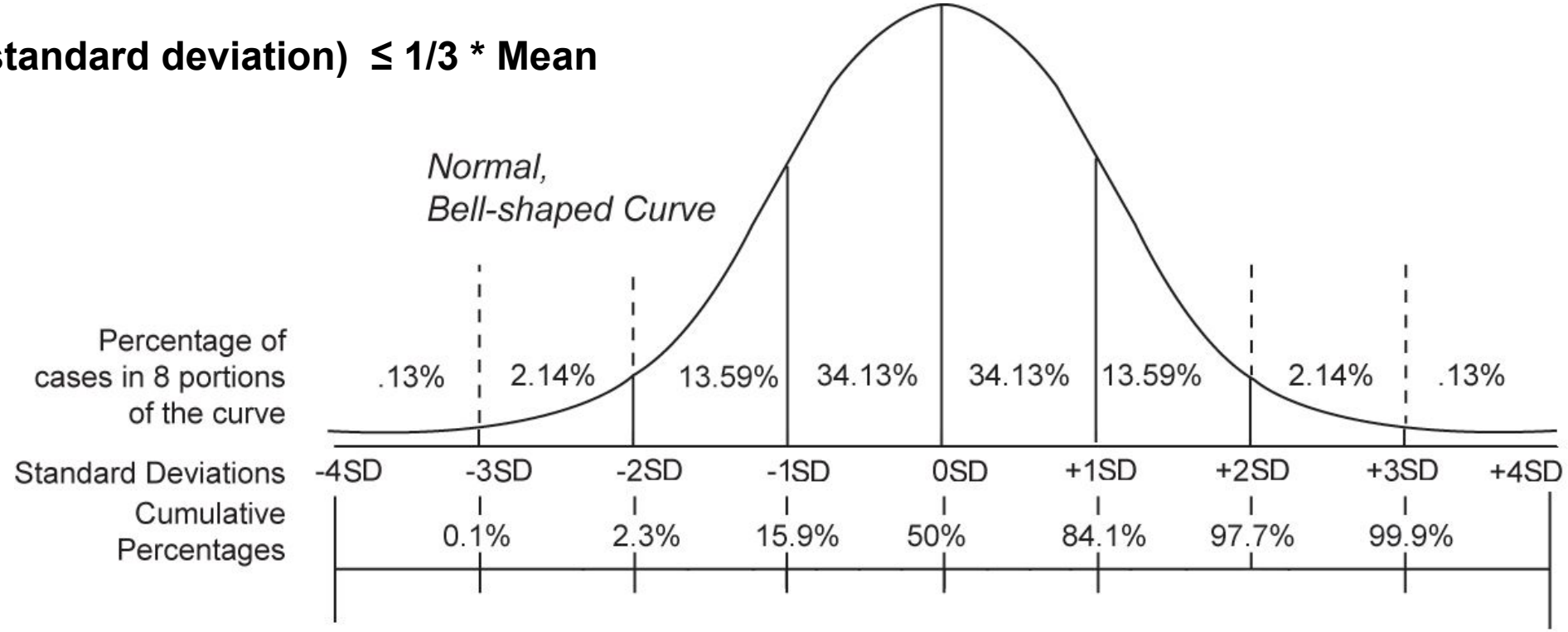
DESCRIPTIVE STATISTICS: STANDARD DEVIATION





STANDARD DEVIATION AND NORMAL DISTRIBUTION

SD (standard deviation) ≤ 1/3 * Mean



DESCRIPTIVE STATISTICS IN SPSS

Analyze – Descriptive statistics – ***Frequencies***

OR

Analyze – Descriptive statistics – ***Descriptives***

Example №1:

- Calculate the mode for “gender” variable. Interpret the results.
- Calculate the median for “education” variable. Interpret the results.
- Calculate the mean, standard deviation, range, minimum, maximum for “income” variable in two ways in SPSS. Interpret the results.

DESCRIPTIVE STATISTICS FOR VARIABLES IN DIFFERENT SCALES

- ***Nominal*** – mode
- ***Ordinal*** – mode + median, mean, standard deviation
- ***Quantitative*** (***Scale***) – mode, median, mean, standard deviation + range, minimum, maximum

CROSS-TABULATING (CONTINGENCY TABLES)

CROSS-TABULATING

- Contingency tables (or cross tables) are usually constructed in the case when two qualitative (nominal or ordinal) variables are analyzed and there is a question about the influence of one of them on the other.
- Contingency tables (or cross tables) allow to prove a hypothesis about the relationship between two qualities (= two qualitative variables).
- Contingency tables (or cross tables) is a means of visualizing the joint distribution of two variables. The general format of a contingency table is a group statistical table. In its rows, the values of one variable are located, and the values of another variable are displayed in columns.

THE EXAMPLE OF USING CROSS-TABULATING FOR SEGMENTING THE MARKET

Customer	Number of visits a week	Age	Income, rub.	Education
1	2	39	> 60 000	bachelor
2	1	63	20 000-39 000	bachelor
3	4	24	20 000-39 000	master
4	7	21	< 20 000	master
5	6	26	40 000-60 000	bachelor
…				

Marketing research of coffee shop customers (fragment)

Number of visits a week	Age					Sum
	20 and less	21-29	30-39	40-49	50 and more	
1 and less	10%	5%	15%	30%	40%	100%
2-3	5%	20%	35%	25%	15%	100%
4-5	15%	35%	25%	20%	5%	100%
6 and more	10%	40%	30%	15%	5%	100%

Contingency table for frequency of visits to a coffee shop with the age of customers

THE EXAMPLE OF USING CROSS-TABULATING FOR SEGMENTING THE MARKET

Customer	Number of visits a week	Age	Income, rub.	Education
1	2	39	> 60 000	bachelor
2	1	63	20 000-39 000	bachelor
3	4	24	20 000-39 000	master
4	7	21	< 20 000	master
5	6	26	40 000-60 000	bachelor
…				

Marketing research of coffee shop customers (fragment)

Number of visits a week	Age					Sum
	20 and less	21-29	30-39	40-49	50 and more	
1 and less	10%	5%	15%	30%	40%	100%
2-3	5%	20%	**35%**	**25%**	15%	100%
4-5	15%	**35%**	25%	20%	5%	100%
6 and more	10%	**40%**	30%	15%	5%	100%

Contingency table for frequency of visits to a coffee shop with the age of customers

CONTINGENCY TABLES: VISUALIZATION

- Put the **independent** variable on ***columns*** and the **dependent** variable on ***rows***
- Percentages are usually more informative, but always report the row/column sums so that the counts can be reconstructed

CHI-SQUARE TEST

- *Pearson Chi-Square test* is a nonparametric method that allows to check the presence or absence of a relationship between two qualitative variables

- H_0: there is no connection between variables
- H_1: there is connection between variables

- If ***Sig.>0.05*** than we accept main hypothesis $\boldsymbol{H_0}$
- If ***Sig.<0.05*** than we accept alternative hypothesis $\boldsymbol{H_1}$

EXAMPLE №2: CROSS-TABULATING

Is there any connection between family status and the fact of keeping any diet?

H_0: There is no connection between family status and the fact of keeping any diet
H_0: People who are married and who are not married keep the diet with the same frequency.

H_1: There is connection between family status and the fact of keeping any diet
H_1: People who are married keep the diet less frequently than those who are not married

CROSS-TABULATING IN SPSS

Analyze – Descriptive statistics – Crosstabs

1. Choose dependent and independent variables, identify the types of scales they are measured in, formulate main and alternative hypothesis
2. Look at the cross tab (make different variants in numbers and in percentage).
3. Perform the analysis in SPSS once again (in Statistics tip Chi-square). Check the hypothesis about the relationship between variables by checking Significance of the Chi-Square test. Make conclusions.

WHAT TO DO WITH THE QUANTITATIVE DATA?..

TASK №2

Example №1 or 1-1:

- Build possible graphs for this dataset (choosing the most appropriate chart for each variable) + two charts of comparisons between them
- Estimate the descriptive statistics for this dataset (choosing the most appropriate indicators of descriptive statistics for each variable)
- Formulate 3 hypotheses that can be tested using the cross-tabulating method. Verify hypotheses by making necessary calculations *(* use a quantitative variable in at least 1 hypothesis)*
- Make some conclusions about the data

All results should be presented on one .doc file