Back to the basics, Part2
Data exploration:
representing and testing data properties

Spyros Veronikis

Electrical and Electronics Engineer
Dept. of Archives and Library Sciences
Ionian University

spver@ionio.gr

http://dlib.ionio.gr/~spver/seminars/statistics/

DBIS
database & information systems group
ionian university

More information

Seminar Content

- Basic statistics concepts
- Data exploration
- Correlation
- Comparing two means
- Comparing several means (ANOVA)
- Non-parametric test
- Nominal data

Today's Content

- Statistical significance
- Parametric Data
- Histograms and Boxplots
- Descriptive statistics
- Correcting problems in the data
- Exploring groups of data
- Testing whether a distribution is normal
- Testing for homogeneity of variance
- Summary

Statistical significance

- Ronald Fisher and Muriel Bristol, The Lady tasting tea

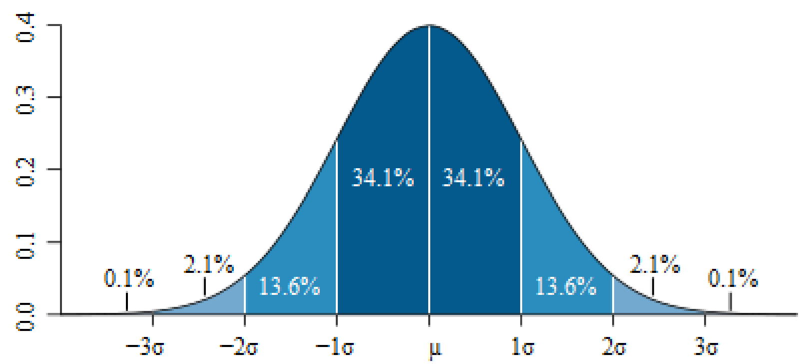
Given 6 cups of tea and milk (in 3 of which the tea was placed first and 3 had the milk added first)

“What is the probability that lady Bristol finds all 3 cups where tea was placed first?”		(Answer: 1/20= 0.05 ή 5%)

Fisher suggested that only when we are at least 95% certain that a result is genuine (not a chance finding) should we accept it as true, and therefore statistically significant.

Frequency distributions can be used to assess the probability. In a typical normal distribution, chance of z occuring more than:

- Z= 1.96 is 0.05 or 5%
- Z= 2.58 is 0.01 or 1%
- Z= 3.29 is 0.001 or 0,1%



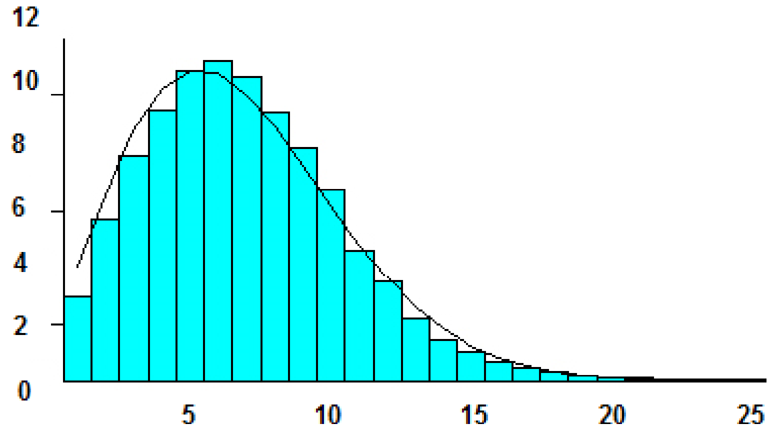
Parametric data

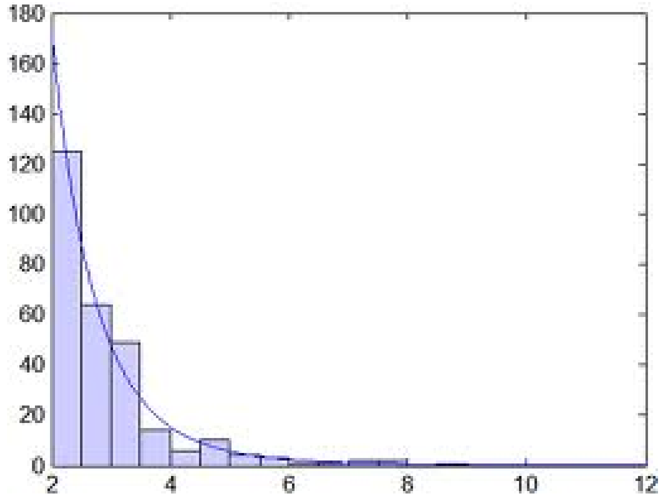
- Assumptions of parametric tests:
  1. Normally distributed data:
     - Checked with histograms and Kolmogorov-Smirnov or Shapiro-Wilk criterion
  2. Homogeneity of variance
     - In several groups: data in samples come from populations of same variance
     - In correlational designs: variance of one variable remains stable across all levels of other variable(s)
  3. Interval data
     - Arithmetic values, equal differences among successive points in measurement scale
  4. Independence
     - Data come from different participants who don't influence each other.

Only requirements 1 and 2 are tested by objective criteria (tests). Requirements 3 and 4 are tested by common sense.

Graphing and screening data

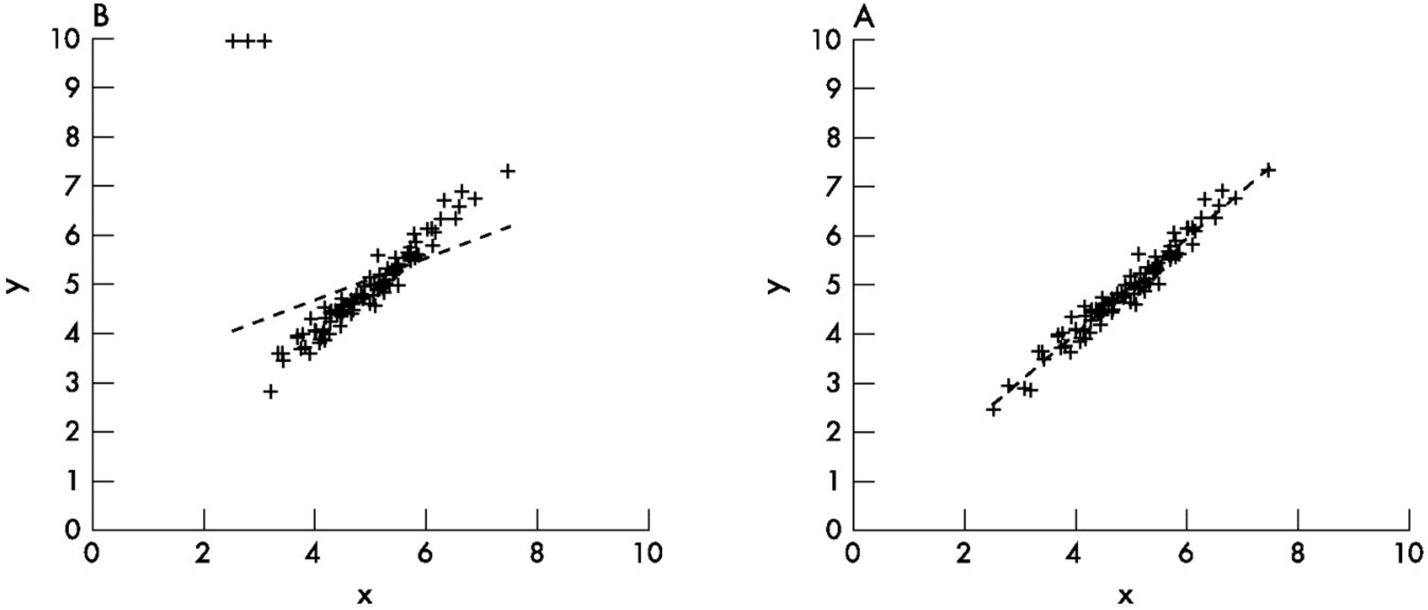
- **Histograms**
  - Show the number of times each recorded value occurs
  - The horizontal axis represents the levels of measurement of the variable
  - They make outliers easy to spot





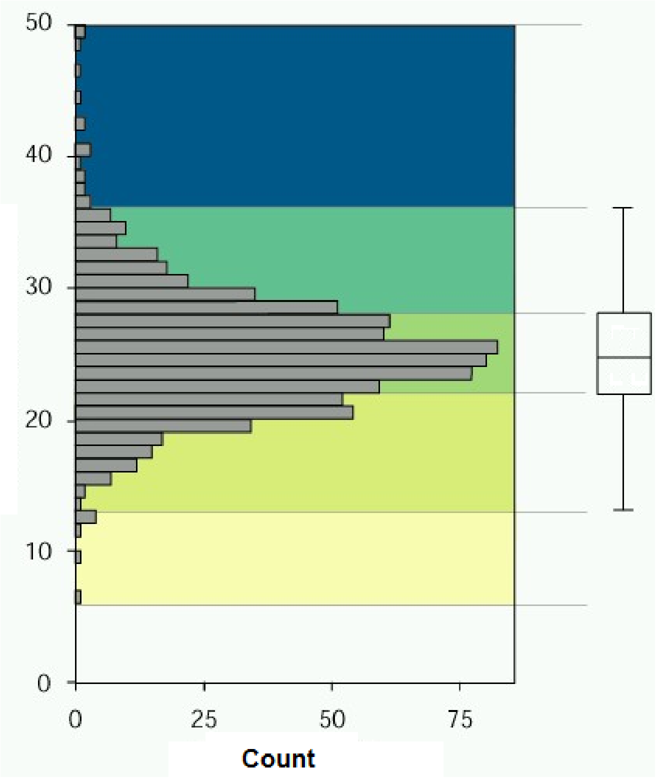
Outliers

- These are scores very different from the rest of the data
- They occur rarely
- They can bias the model we fit to the data
- They need to be identified and omitted from the dataset



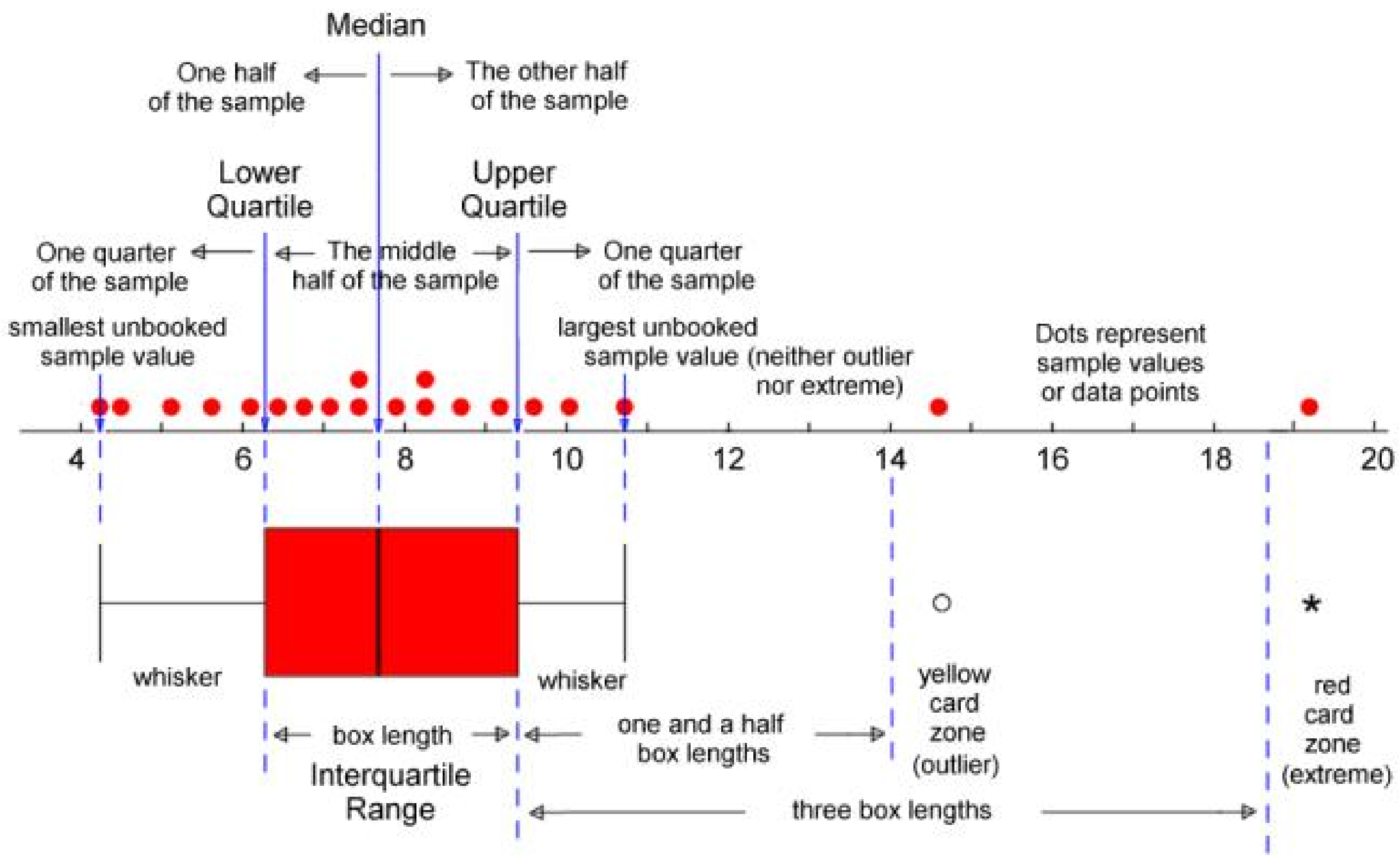
Graphing and screening data

- **Boxplots (box-whisker diagrams)**
  - Show the lowest and highest scores
  - Show the quartile (25%) ranges
  - Show the interquartile range (50%)
  - Show the ***median****
- Normal distribution has a symmetrical boxplot
- Skewed distributions don't
- Platykurtic distributions have wide boxplots



* The median of a list of numbers is the number that splits the dataset in half.
E.g. for dataset A={1,2,2,4,13,15,51} the median is 4 (the average is 12.57)

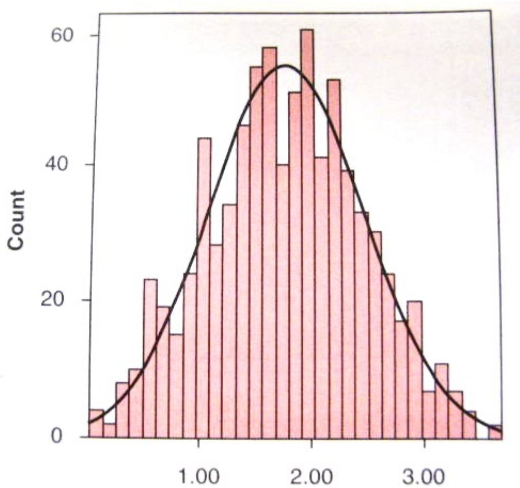
Graphing and screening data



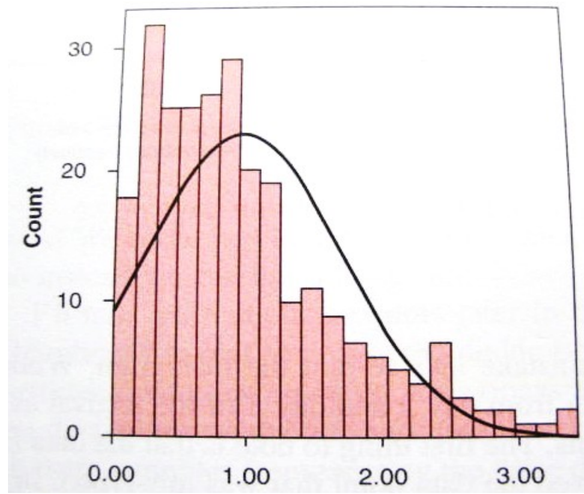
Descriptive statistics

- "An alternative search tool was provided to the library patrons and we assess its adoption by recording the number of hours spent on the previous tool before and after 2 demonstration sessions".
- Descriptive statistics
  - Mean, Standard error of mean
  - Median
  - Mode
  - Standard deviation
  - Variance
  - Skewness and std. Error
  - Kurtosis and std. Error
  - Range
  - Min, Max

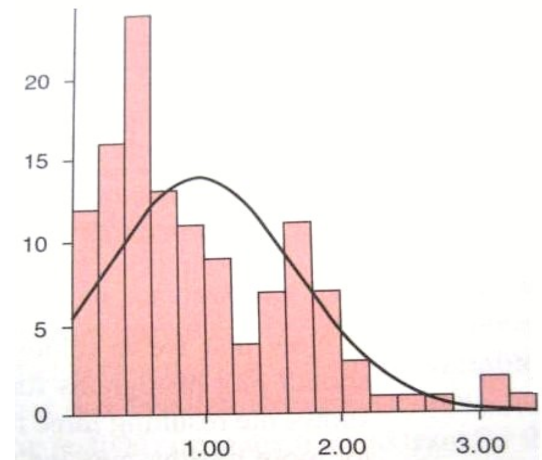
Descriptive statistics



First Period



Second Period



Third Period

		First period	Second period	Third period
N	Valid	810	264	123
	Missing	0	546	687
Mean		1,77	0,96	0,97
Std error of Mean		0,0244	0,0444	0,0640
Median		1,79	0,79	0,76
Mode		2,00	0,23	0,44
Std. Deviation		0,6935	0,7208	0,7103
Variance		0,481	0,519	0,504
Skewness		-0,004	1,095	1,033
Std. Error of skewness		0,086	0,150	0,218
Kurtosis		-0,410	0,822	0,732
Std. Error of kurtosis		0,172	0,299	0,433
Range		3,67	3,44	3,39
Minimum		0,02	0,00	0,02
Maximum		3,69	3,44	3,41

Skewness and kurtosis

- These are 0 for normal distribution
- z-values are more informative because they are standardized
  - Skewness: $z_{sk} = SK/SE_{sk}$
  - Kurtosis: $z_{kur} = KUR/SE_{kur}$

1st period: $z_{sk} = -.004/.086 = .047$, $z_{kur} = -2.38$

2nd period: $z_{sk} = 1.095/.150 = 7.30$, $z_{kur} = -2.75$

3rd period: $z_{sk} = 1.033/.218 = 4.73$, $z_{kur} = 1.69$

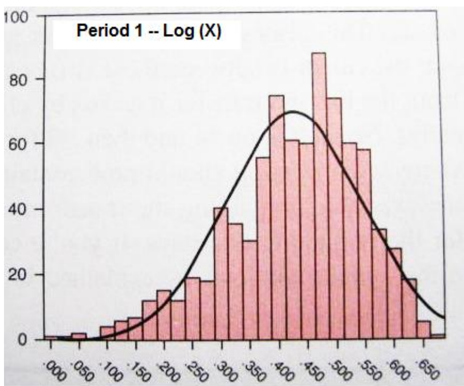
Correcting problems in data

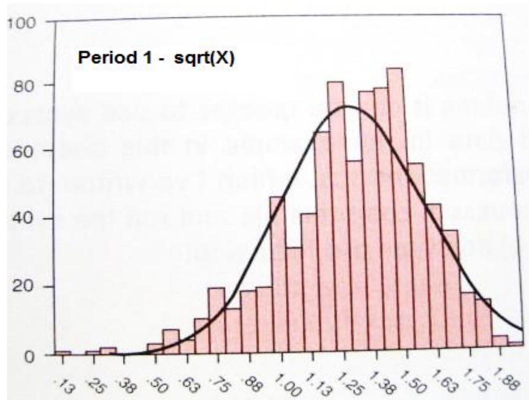
- How do we deal with problems in data (problems in distribution, outliers, missing values)
  - Remove problematic cases
  - Transform the data (X → Y= function(X))
- Transformations can correct distribution form, i.e., remove skewness by “smoothening” outliers
- Transform ALL data, of ALL variables that are going to be compared/related even if there are variables that aren't skewed. This isn't cheating!
- Do the statistic tests and analysis

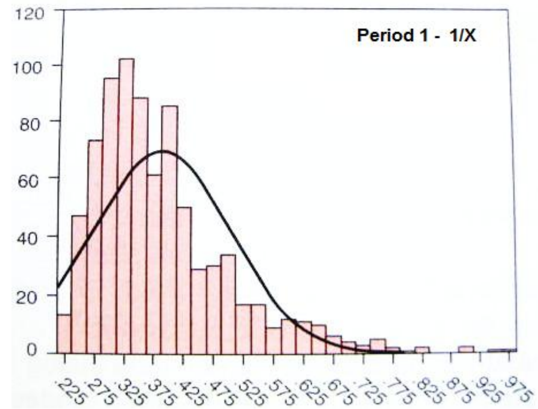
Correcting positively skewed data

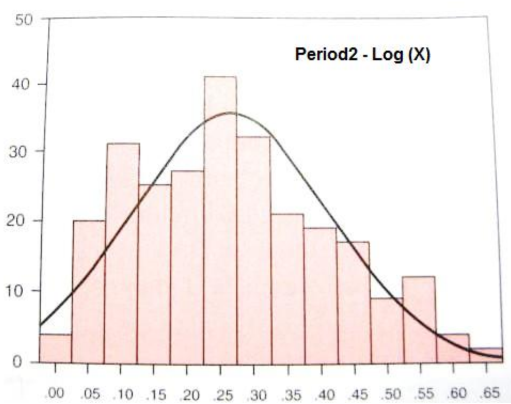
- Log transformation, $X_{tr} \rightarrow Log(X)$:
  - Reduces positive skew.
  - X must be greater than 0 (shifting might be needed)
- Square root transformation, $X_{tr} \rightarrow sqrt(X)$
  - Brings large scores closer to the center
  - X must be greater than 0 (shifting might be needed)
- Reciprocal transformation, $X_{tr} \rightarrow 1/X$
  - Reduces the impact of large scores
  - Reverses the scale. Therefore, prior to transformation we need to reverse scores ourselves (e.g. $Xrev \leftarrow HighestX - X$). Then $X_{tr} \rightarrow 1/Xrev$
- For negatively skewed data, reverse the scores prior to the above transformations.

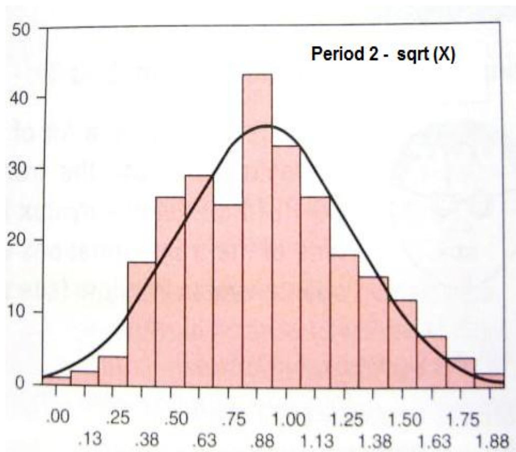
Transformed data

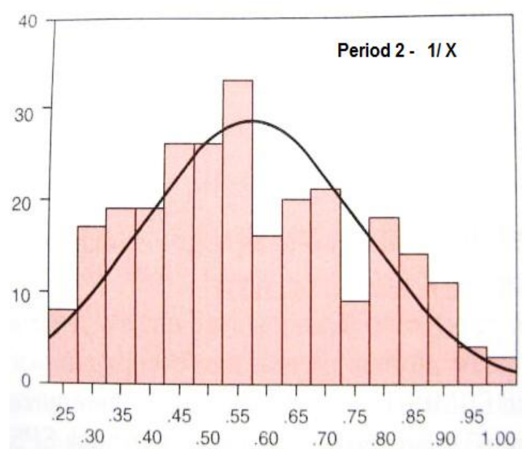












Explore groups of data

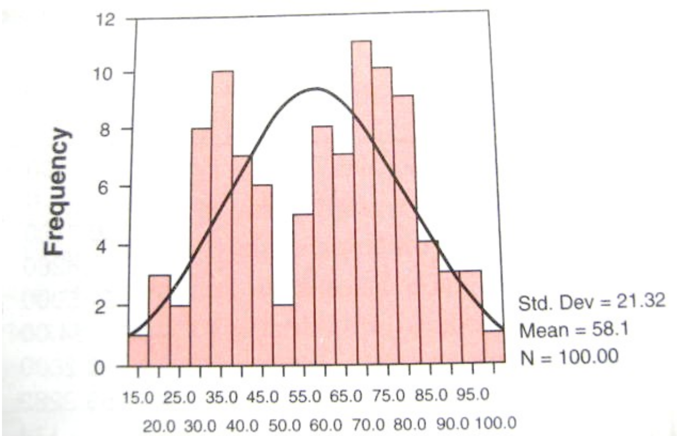
Example:

- A professor is recording the score of a project report delivered by his students. He also takes a note to:
  - The percentage of the given bibliography studied (biblio)
  - Their attendaces in the lectures (lectures)
  - The number of student collaborated to deliver the report (participants).
- He collects the same data a year later (from the students of the new class) and looks for differences in performance.

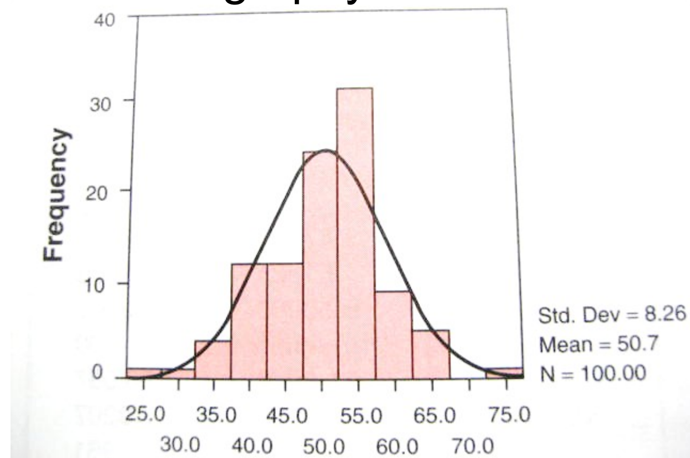
  - Score= *function*(biblio, lectures, participants, year)

Explore groups of data

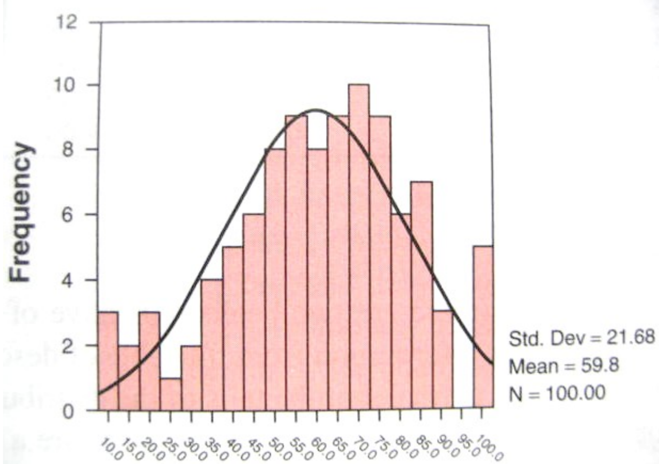
Project score



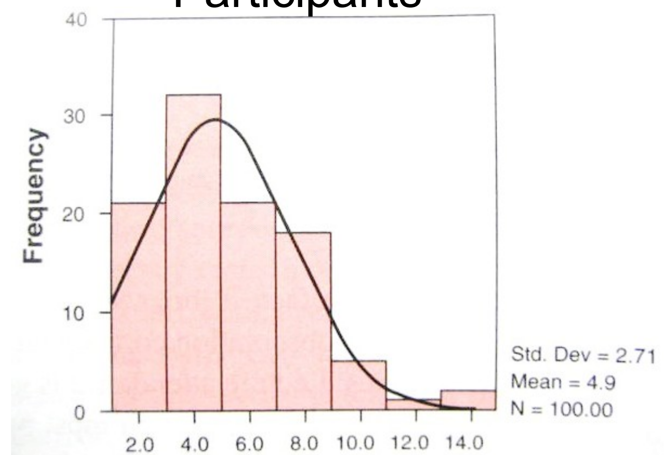
Bibliography studied



Lectures attended



Participants



DBIS
database & information systems group
ionian university

Explore groups of data

		Bibliography studied (%)	Score (%)	Lectures (%)	Participants
N	Valid	100	100	100	100
	Missing	0	0	0	0
Mean		50,7100	58,1000	59,7650	4,8500
Std error of Mean		0,8260	2,1316	2,1685	.2706
Median		51,5000	60,0000	62,0000	4,0000
Mode		54,00	72,00	48,50	4,00
Std. Deviation		8,2600	21,3156	21,6848	2,7057
Variance		68,2282	454,3535	470,2296	7,3207
Skewness		-0,174	-0,107	-0,422	0,961
Std. Error of skewness		0,241	0,241	0,241	0,241
Kurtosis		0,364	-1,105	-0,179	0,946
Std. Error of kurtosis		0,478	0,478	0,478	0,478
Range		46,00	84,00	92,00	13,00
Minimum		27,00	15,00	8,00	1,00
Maximum		73,00	99,00	100,00	14,00

Explore groups of data

Comments on overall descriptives (previous table):

- The distribution of score for both years seems bimodal (could be a difference in performance by year)
- We can compare scores of two years (because each one comes from a normal distribution), but we can compare the whole dataset of scores to another similar dataset.
- The participants' distribution might also be due to different collaborations among years.

- We ask for descriptives per year

Explore groups of data

Year 2010			
		Bibliography studied (%)	Participants
N	Valid	50	50
	Missing	0	0
Mean		40,1800	4,1200
Std error of Mean		1,7803	0,2922
Median		38,0000	4,0000
Mode		34,00	4,00
Std. Deviation		12,5880	2,0660
Variance		158,4771	4,2710
Skewness		0,309	0,512
Std. Error of skewness		0,337	0,337
Kurtosis		-0,567	-0,484
Std. Error of kurtosis		0,662	0,662
Range		51,00	8,00
Minimum		15,00	1,00
Maximum		66,00	9,00

Year 2011			
		Bibliography studied (%)	Participants
N	Valid	50	50
	Missing	0	0
Mean		76,0200	5,5800
Std error of Mean		1,4432	0,4343
Median		75,0000	5,0000
Mode		72,00	5,00
Std. Deviation		10,2050	3,0712
Variance		104,1420	9,4322
Skewness		0,272	0,793
Std. Error of skewness		0,337	0,337
Kurtosis		-0,264	0,260
Std. Error of kurtosis		0,662	0,662
Range		43,00	13,00
Minimum		56,00	1,00
Maximum		99,00	14,00

Testing normality of a distribution

- Normality is not assessed visually (i.e., it looks normal to me)
- We mathematically examine whether a given distribution as a whole **deviates** from a comparable normal distribution (having same mean and same standard deviation) .
- We use Kolmogorov-Smirnov and Shapiro-Wilk tests

  "Is the given distribution different than normal?"
- None significant test outcome (p>. 05) indicates similar distribution, therefore normality
- A difference (outcome) found significant ($p < 0.05$) shows non-normality

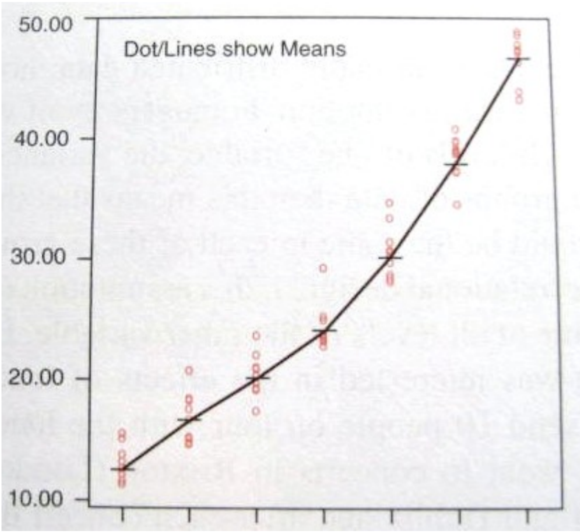
	Kolmogorov – Smirnov				Shapiro – Wilk		
	Statistic	df	Significance		Statistic	df	Significance
Score (%)	.102	100	.012		.961	100	.005
Participants	.153	100	.000		.924	100	.000

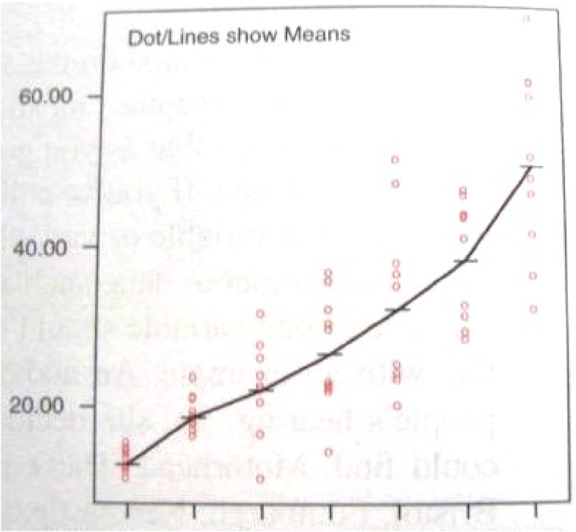
		Kolmogorov – Smirnov				Shapiro – Wilk		
		Statistic	df	Significance		Statistic	df	Significance
Score (%)	2010	.106	50	.200		.972	50	.283
	2011	.073	50	.200		.984	50	.715
Participants	2010	.183	50	.000		.941	50	.015
	2011	.155	50	.004		.932	50	.007

Testing for homogeneity of variance

- Homogeneity of variance is a requirement for parametric tests to be applied
- As you go through the levels of one variable, the variance of another variable must not change

  "For how many hours do you feel sick after an extra pint of beer?"





Testing for homogeneity of variance

- For groups of data we use Levene's test which reveals if there is homogeneity of variance.
- It test the hypothesis that “There is no difference between variances in the groups”, i.e., the difference between variances is zero.
  - If the test outcome (difference) is found significantly different from 0 (i.e., there is a difference and that is not a chance finding, $p<.05$) then we reject the null hypothesis and acknowledge heterogeneous variances.
  - If the test outcome is found non-significant (error probability $p>.05$) then we accept the null hypothesis and consider the group data to be of homogeneous variance.

	Levene statistic (d)	df1	df2	Significance
Score	2.584	1	98	0.111 > 0.05
Participants	7.368	1	98	0.008 < 0.05

Summary

- What we 've seen
  - Examine data properly before proceeding to analysis
  - Look at the data distribution
  - Spot any problems (e.g. outliers)
  - In case of non-normality try to transform data
  - When comparing data from different groups look at distributions within each group
  - Also test for homogeneity of variance

References

- Field, A. (2005). **Discovering Statistics Using SPSS**, 2nd ed., Sage Publications.
- Statsoft, Inc. (2011). **Electronic Statistics Textbook**. Tulsa, OK: Statsoft. WEB: http://www.statsoft.com/textbook/