Statistics

Section: Basics

Exercises

1 Data types

1.1 Identify types of variables

Name	Id	Class 2006	Grade 2006	Age 2006	Year of birth	Hometown
Miller	731	13	1	23	1983	Bonn
Meyer	422	12	3	22	1984	Cologne
...	...	...	...	...	...	...
Smith	159	11	5	21	1985	Bonn

Which variables are categorical, which are ordinal, and which are quantitative?

	Categorical	Ordinal	Quantitative
Name			
Id			
Class 2006			
Grade 2006			
Age 2006			
Year of birth			
Hometown			

1.2 Identify types of variables

Company	Production key	Employees 2005 (x 1,000)	Turnover 2005 (Mill. $)	Global Operating 1=yes 0=no	Ranking within field of production
Shell	2111.4	123	1,234	1	1
Exxon	2111.5	213.4	1,567	1	2.5
...	...	...	...	...	...
Trimega	5121.1	54	0.345	0	5

Which variables are categorical, which are ordinal, and which are quantitative?

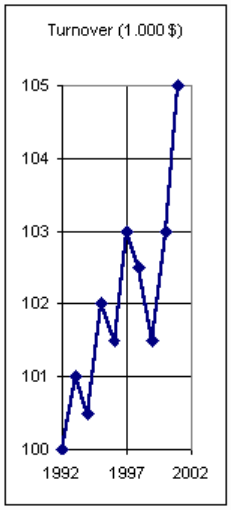
	Categorical	Ordinal	Quantitative
Company			
Production key			
Employees			
Turnover			
Global			
Ranking			

1.3 Meaningful sample statistics for different types of data

	Proportion	Mode	Median	Mean, Standard deviation, Skewness, Kurtosis
Categorical=Nominal				
Ordinal				
Quantitative=Continuous= Scale				

2 Graphs

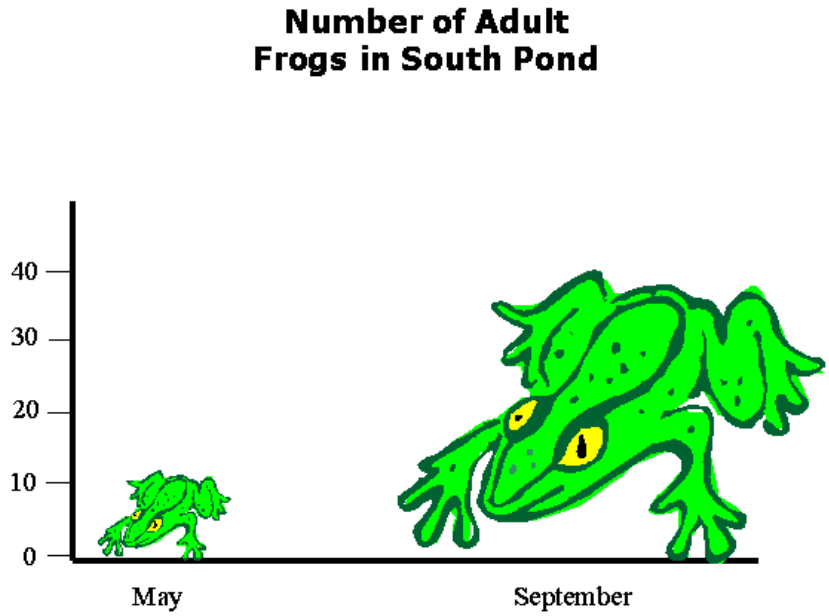
2.1 Does this graph indicate an impressive success? Why? Sketch your argument.



2.2 Can you use a pie chart to display these data? Why?

What sort of music do you like	Percent
Techno	5%
Pop	55%
Country	20%
Folk	20%
Classic	5%

2.3 Can you explain why the following graph[1] is not adequate?

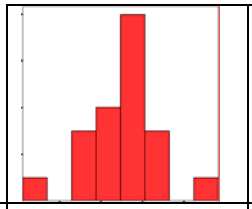
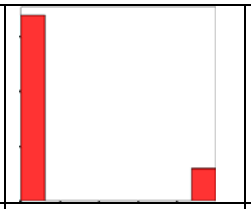
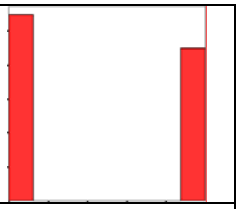


[1] "Display of statistic data", Retrieved November 10, 2006, from http://www.physics.csbsju.edu/stats/display.html .

3 Density shape (distribution of variables)

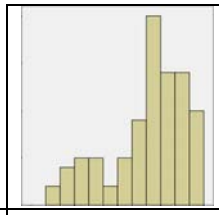
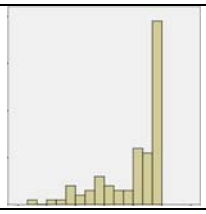
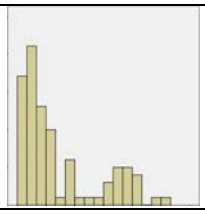
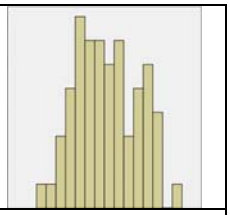
3.1 Hypothetical data from your class:

For 20 students "Gender (female=0, male=1)" "Height in cm" "Handed (right-handed = 0, left-handed=1)" are recorded.
Can you guess which histogram goes with each variable?
Why?

			
Gender			
Height <cm>			
Handed			

3.2 Data from the UNDP and World Bank development report 1997[2]

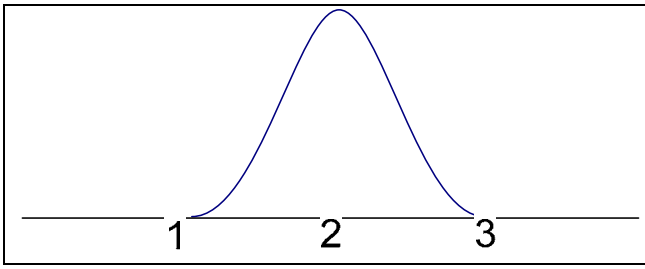
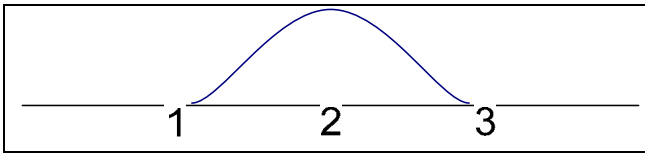
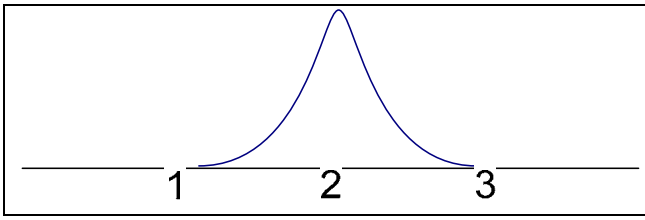
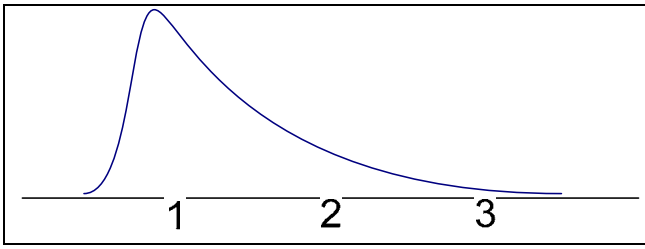
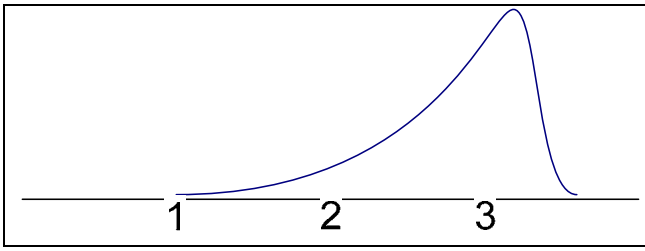
For 90 countries the "Gini coefficient" "Life expectancy at birth" "Literacy rate of adult" and "Gross domestic product" is recorded.
Can you guess which histogram goes with each variable?
Why?

				
Gross domestic product				
Life expectancy at birth				
Literacy rate of adult (0-100%)				
Gini coefficient (0-100%)				

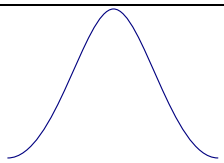
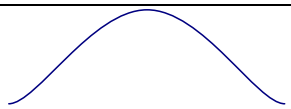
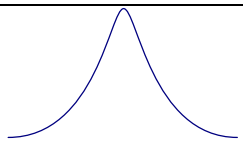
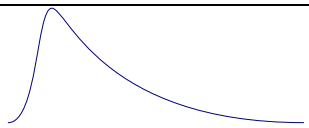
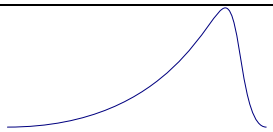
[2] Data compiled from Human Development Report, UNDP, 1997 and World Development Report, The World Bank, 1997

4 Density shape, parameter and sample statistic

4.1 Can you guess where Mean and Median of these idealized densities should be located?

	Mean	Median
		
		
		
		
		

4.2 Can you guess the skewness and kurtosis of these idealized densities?

	Skewness right (positively) >0 not skewed = 0 left (negatively) < 0	**Kurtosis** leptokurtic > 0 mesokurtic = 0 platykurtic < 0
		
		
		
		
		

4.3 Which sample statistics are included in the "five-number summary"?

Minimum	
Q1 (1. quartile) (25%)	
Mean	
Standard deviation	
Median = Q2 (2. quartile) (50%)	
Coefficient of variation	
Skewness	
Mode	
Kurtosis	
Q3 (3. quartile) (75%)	
Maximum	

4.4 Five-number summary vs. mean/standard deviation

Under what circumstances **should you prefer the "five-number summary**" rather than Mean/Standard deviation for adequately characterizing the distribution of a variable?

	adequate choice	
	Five-number summary	**Mean / Standard Deviation**
normal		
symmetric		
not symmetric		
many "outliers"		
positively skewed		
negatively skewed		

	Median	Mean
The typical value		
Robust against extreme values		
Used in most statistical tests		
Switch to sum of all values is easy		

4.5 Calculate mean and five-number summary[3]

Data : **10 , 20, 30, 40, 50, 60, 70**
Hint : sum = 280

Sample statistic	Mean	Std					
		22					

Data : **10 , 20, 30, 40, 50, 60, 7000**
Hint : sum = 7210

Sample statistic	Mean	Std					
		2633					

[3] "There are several rules for calculating quartiles…The differences are always small." Moore (2006)
Use Moore's rule to calculate Q1 and Q3 by :
determine the median to the right resp. the left of the median (excluding the median).

4.5 Can you give the "five-number summary" for the following sample (already ordered)?

Hint: use simplified formulas (see Moore)

Customer	X
Miller	-10
Meyer	-2
Smith	3
Brown	4
Jacobs	5
Will	6
Cliff	7
Burger	15
Breen	30

Name of statistic					
Value					

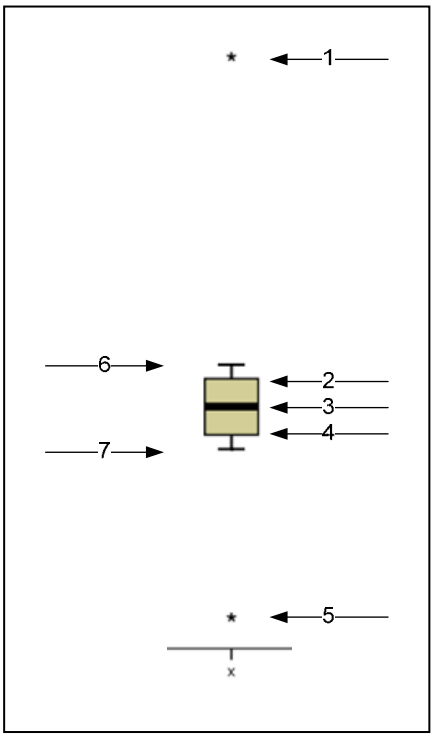
4.6 Identify box plot's characteristics

Data of 5.5 are shown in Figure 5.6

Can you identify names and values?

Number in Figure 5.6	Name of statistic	Value	Label if any
1			
2			
3			
4			
5			
6			
7			

Figure 5.6



4.7 IQR (definition)

The IQR = Interquartile range (Q3-Q1) is often used as an indicator for outliers (outsiders[4]) and extreme outliers (far out).

Which rules[5] are often (e.g. SPSS) used?

	Mark "correct"
An outlier is marked by a O (circle) accompanied by the name of the object	
An outlier is defined as 1.5 * IQR below or above Q1 resp. Q3	
An outlier is defined as 1.5 * IQR below or above the median	
An outlier is defined as 2 * IQR below or above Q1 resp. Q3	
An outlier is defined as 2 * IQR below or above the median	
An extreme outlier is marked by a ✶ (star) accompanied by the name of the object	
An extreme outlier is defined as 3 * IQR below or above Q1 resp. Q3	
An extreme outlier is defined as 3 * IQR below or above the median	
An extreme outlier is defined as 4 * IQR below or above Q1 resp. Q3	
An extreme outlier is defined as 4 * IQR below or above the median	

4.8 Box plot vs. Histogram

Main advantages of Box plots and Histograms

	Histogram	Boxplot
"Quick and dirty" test of significance between groups		
Compare location sample statistics between groups		
Compare spread sample statistics between groups		
Identify outliers (value and label by default)		
Identify a multimodal distribution		
Kurtosis of one variable		
Location of mean (by default)		
Location of median (by default)		
Location of modal (by default)		
Skewness of one variable		

[4] Although modern text books (e.g. Moore (2006)) defines such a value as an "outlier" the original notation was "outsider". Which seems more reasonable as "outlier" is a concept that would fail if the underlying distribution has values that appear quite naturally outside (e.g. Lognormal). An "extreme outlier" was called "far out" then (see Tukey (1977)).

[5] The "correct" rules which we refer to are those by John W. Tukey who introduced the concept of Box Plots in "Exploratory Data Analysis", Addison-Wesley, 1977

5 Exploratory data analysis with STATA or SPSS

5.1 Analyze GPA-Data[6] (GPA.xls)

Objective: Relationship between the students "self concept" and their academic performance

Data on 78 seventh-grade students in a rural Midwestern school.

id : Identification of students (11 students dropped out)
gpa : **g**rade **p**oint **a**verage
iq : Score on a standard IQ test
gender : 1 = Female 2=Male
concept: Self-concept test score (Piers-Harris Children's Self-Concept test)

Population center values (IQ=100; Self-Concept=60)
gender, gpa and iq were taken from school records.

Table 5.1
Data (extraction) on 78 seventh-grade students in a rural Midwestern school

id	gpa	iq	gender	concept
1	7.94	111	2	67
2	8.292	107	2	43
…	…	…	…	…
88	6.057	93	1	21
89	6.938	106	2	56

Explore data of the survey

Hint:

1. Import the EXCEL file into STATA or SPSS
2. Assign the correct type of data
3. Label variables and values
4. Make a boxplot and histogram of the quantitative variables
5. Make a frequency table and bar chart of the categorical variable
6. Make a five-number summary of categorical variables
7. Describe shape, center and shape
8. Identify outliers (outsiders)

[6] Moore (2006), p.39 (Data provided by Darlene Gordon, Purdue University)

6 Data validation

6.1 Benford's law

Three different studies presenting data drawn from a loan file of a certain bank.
Each study includes the monthly income of 1,000 customers in local currency.
Each study claims that data are randomly taken from all loans files of this bank.

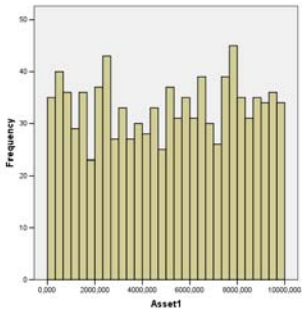
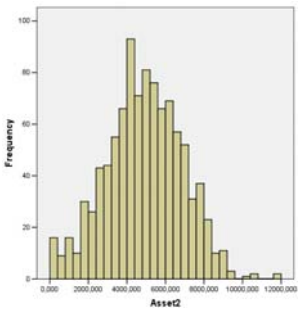
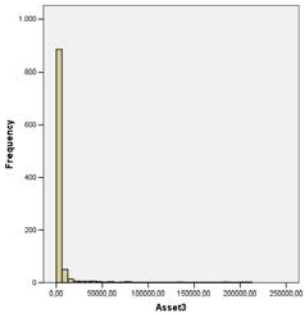
Can you detect, **which data set is most likely been manipulated**?
Why?

Extract from sampled data

Study 1	
Income	**First digit**
6456.00	6
8306.00	8
…	…
…	…
2309.00	2
4045.00	4
1343.00	1
526.00	5
406.00	4
2088.00	2

Study 2	
Income	**First digit**
7847.00	7
3491.00	3
…	…
…	…
1725.00	1
5946.00	5
6118.00	6
6004.00	6
2677.00	2
4224.00	4

Study 3	
Income	**First digit**
270.00	2
156.00	1
…	…
…	…
263.00	2
155.00	1
124.00	1
155.00	1
182.00	1
226.00	2

	Study 1	**Study 2**	**Study 3**
	Income	**Income**	**Income**
			
Mean	**5,100**	**4,900**	**5,000**
Std.	**2,900**	**1,900**	**19,900**
	First digit	**First digit**	**First digit**