Lecture 6

Empirical Research Methods
IN4304

Data preparation methods

TU Delft

Previous lecture

participant-observation and non-participant observation
Does the observer act as a member of the group?

Sampling strategies
Event sampling, state sampling, interval sampling, time sampling

Coding scheme
- Focused
- Objective
- No context-dependent
- Explicitly defined
- Exhaustive
- Mutually exclusive
- Easy to record

Inter-observer agreement (Cohen's Kappa)
$K = (P_0 - P_c)/(1-P_c)$

TU Delft

Today

- Data coding
- Measures of central tendency
- Measures of variability
- Explorative Data Analysis
- Data cleaning
- Normal distribution
- Standardizing data
- Confidence intervals

TU Delft

Learning outcomes of this lecture

After today's lecture you should be able :
- to enter your data into SPSS
- to examine the central tendency of your data set
- to understand the output of some explorative data analysis techniques
- to reflect up on reasons for data cleaning
- to explain the importance of normal distribution

TU Delft

Data coding

- Coding is needed to process the data answers of questionnaire or observation

How much do you like this lecture?

(tick one answer)

Very much	**O**	**(code = 1)**
Quite a lot	**O**	**(code = 2)**
Moderately	**O**	**(code = 3)**
Not very much	**O**	**(code = 4)**
Not at all	**O**	**(code = 5)**
Not applicable	**O**	**(code 900)**
No answer (missing value)		***(code = 999)***

TU Delft

Entering data into SPSS

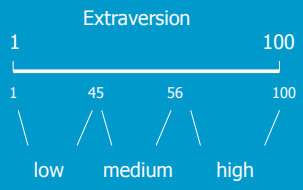
- Give variable a clear name and label
- Specify data type (string, number, date etc)
- Specify values code
- Specify missing value code
- Specify level of variable (nominal, ordinal, interval)

SPSS Demo

TU Delft

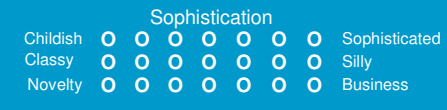
Changing data from one level to another

- Changing data from interval to ordinal level



- SPSS recode function

Direction scale



Code1 : 1 2 3 4 5 6 7
Code2: -3 -2 -1 0 1 2 3

- Classy-Silly needs to be reversed
- For example:
  - code1 : 8-value
  - Code2 : -1*value
- SPSS command: *Compute*

Measures of central tendency

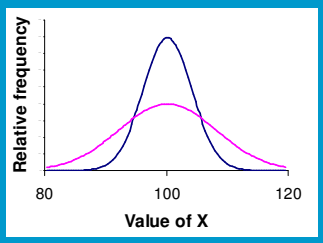
- Interval level: Mean
- Ordinal level: Median
- Nominal level: Mode

What is the best measure to represent:
1 3 3 3 4 7 8 8 9 10 11 11 12 13 13 15 16 17 17 19 10000
Mean: 485.71
Median: 11
Mode : 3
5% trimmed mean: 10.47

Measures of variability

To express the distribution of the data set.

Mean is 100. The spread is different.



Range

Range of a distribution is the difference between the highest and the lowest value in a set of numbers

Range = Max-Min

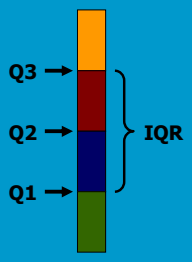
For example

8 9 10 2 4 3 100 12 80

Range = 100 - 2 = 98

Inter Quartile Range

- A quartile is any of the three values which divide the sorted data set into four equal parts
- First quartile (Q1) = cuts off lowest 25% of data = 25th percentile
- Second quartile (Q2) = Median = cuts data set in half = 50th percentile
- Third quartile (Q3) = cuts off highest 25% of data, or lowest 75% = 75th percentile
- Inter Quartile Range (IQR) = Q3-Q1
- The IQR is a more stable statistic than the range, as it is less effected by extreme values.



Standard Deviation (SD)

Standard Deviation measures the spread from the mean. You can think of it as the mean deviation from the mean

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

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

Population mean: μ
Sample mean: $\bar{x}$
Population SD: σ
Sample SD: s

Explorative Data Analysis (SDA)

Examine your data before you apply statistical analysis on it.

Use techniques such as

- Box plot
- stem and leaf plot
- histogram

SPSS – explorative command

Stemplot (stem-and-leaf plot)

Number of times a week that a website was visited Stem-and-Leaf Plot

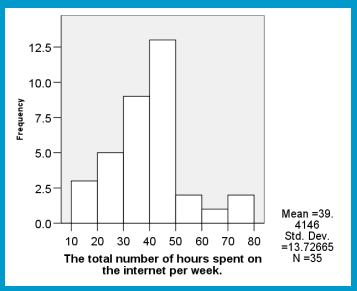
Frequency	Stem &	Leaf
11.00	0 .	00012233447
2.00	1 .	22
3.00	2 .	057
7.00	3 .	1357888
6.00	4 .	234569
1.00	5 .	0
5.00	6 .	25689
4.00	7 .	2359
2.00	8 .	69

Stem width: 100
Each leaf: 1 case(s)

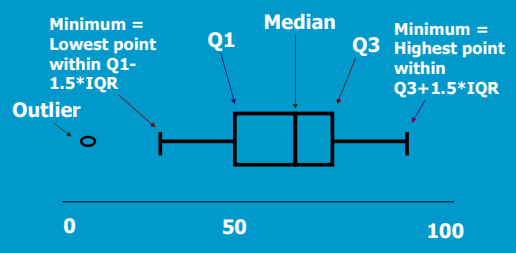
Weekno	Hits
43	3
41	5
44	9
42	12
37	25
36	28
38	34
39	35
35	44
40	46
14	72
22	121

Histogram

- Gives an overview of shape of the distribution
- SPSS – graph - histogram



Box plot



SPSS – Explorative analysis or graph command

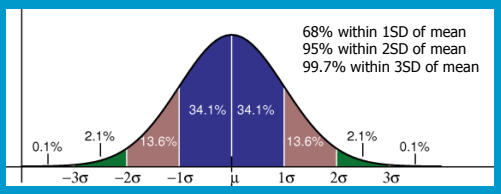
Outliers

- A case whose value is very different from most others
- Can bias statistics such as mean
- What should you do with outliers?
  - Exclude them?
  - Investigate them?
  - Include them in the analysis?

Data cleaning

- Reason to remove cases
  - Observation was distorted, e.g. did not understand task, questionnaire, purposely messing up your experiment…, error in measuring instrument
  - Outliers, but only statistical reason is questionable
  - Check your data for unlikely data points

Normal distribution



Why normal distribution important?

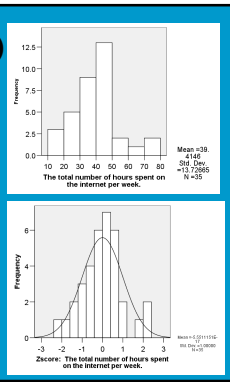
1. Gives a good model of data for some real world data sets (e.g. large populations)
2. Good approximation of results of random outcomes (e.g. throwing a dice many times)
3. Large number of inferential statistics are based on normal distributions

(Moore and McCabe, 2006)

Standardizing data (1)

- Transforming data into z-score, makes it possible to compare data from different unit of measurement

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \bar{x}}{s}$$



Standardizing data (2)

Repeated measure design – focus on relative difference

- *z*-scores to overcome individual use of questionnaire scale e.g.
  - People that only use the extremes
  - People that never use the extremes
  - People that only give high marks

For each individual calculate mean and sd across the his/her questions

Calculate *z*-score for value an individual gave to question based on this persons mean and sd of his/her questions

Unlikely
1 extremely
2 quite
3 slightly
4 neither
5 slightly
6 quite
7 extremely
Likely

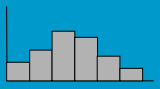
Standardizing data (3) – Example

Participant 1 answered:
Q1=1 Q2=3 Q3=7 Q4=7
z-score would be:
Q1= -1.17 Q2=-0.50 Q3=0.83 Q4=0.83

Participant 2 answered:
Q1=4 Q2=5 Q3=6 Q4=6
z-score would be:
Q1=-1.31 Q2=-0.26 Q3=0.78 Q4=0.78

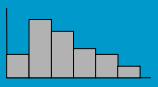
Unlikely
1 extremely
2 quite
3 slightly
4 neither
5 slightly
6 quite
7 extremely
Likely

Medium, mean and distribution shape

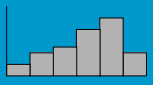


Symmetrical (median = mean)

Notice that for the sample
2 8 110 2
The median = 5
The mean = 30.5



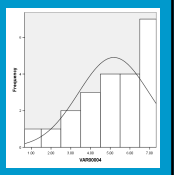
Positively skewed or skewed to the right (median < mean)



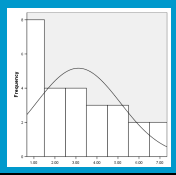
Negatively skewed or skewed to the left (median > mean)

Skewness

Measure
Skewness = 0 --> symmetrical distribution
Skewness > 0 --> positively skewed
Skewness < 0 --> negatively skewed



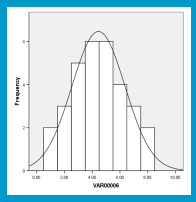
Ceiling effects



Floor effect

Check for normality (1)

Look at the histogram and the projected normal curve



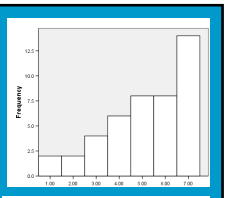
Rule of the thumb for serious deviation from normality

- skewness (or kurtosis) statistic > 2*Std Error (i.e. sd/sqrt(n))
- Difference between mean and median > 0.5*SD

(Coolican, 2004)

Check for normality (2)

- Kolmogorov-Smirnov Normality test



	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
VAR00004	.178	44	.001

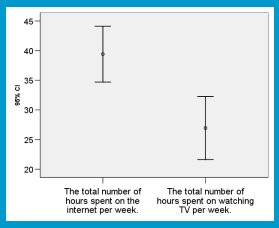
a. Lilliefors Significance Correction

Probability that this sample was obtained from normal distributed population

Remember:
- it is easier to show deviation from normality with large samples
- It does not tell about the size of deviation

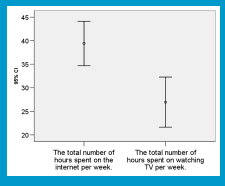
Confidence intervals

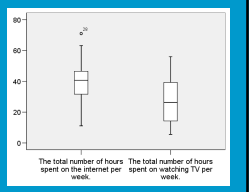
- If we assume the sample is taken from a normal distribution it is possible to make a **prediction** of the interval in which the population mean might be with e.g. 95% certainty.



Class Questions

- What is the difference between box plot and confidence interval?





Summary

Data Entry
Code data, direction scales, clear name, missing data, level

Central tendency
Mean, median, mode, IOR, SD

Explorative data analysis
- Boxplot
- Histogram
- Stem and leaf plot

Normal distribution
1. Gives a good model of data for some real world data sets (e.g. large populations)
2. Good approximation of results of random outcomes (e.g. throwing a dice many times)
3. Large number of inferential statistics are based on normal distributions

Check for Normality if you want to use parametric statistics

This week in practicum

Explorative Data Analysis
- Box-plot
- Scatter plot
- Stem-and-leaf plot

TUDelft

Next time

Quantitative Data analysis I – Differences
- **Hypothesis testing**
- **Chi-square test, *t*-test**
- **(M)ANOVA**
- **Mann-Whitney U-test**
- **(Robson ch. 13)**

TUDelft

References

- Coolican, H., (2004). *Research methods and statistics in psychology* (4th ed). London, UK: Hodder Arnold.
- Moore D.S, and McCabe, G.P. (2006) *Statistiek in de praktijk; theorieboek* (5 ed) Den Haag: The Netherlands, Academic Service.

TUDelft