Statistik I - Exercise session 3
26.5.2014 & 2.6.2014

Info

- Classroom: SPA1 220
- Time: Mondays, 16:15 - 17:45
- in English
- Assignments on webpage (lvb>staff>PB)

Contact: Petra Burdejova
petra.burdejova@hu-berlin.de
Office: SPA1 R400 (upon agreement)

Schedule:

Date	Week	Exercises
28.04.14	E1	1-2, 1-3 (even), 1-10
05.05.14	E1	1-2, 1-3 (even), 1-10
12.05.14	E2	1-20, 1-22, 1-32
19.05.14	E2	1-20, 1-22, 1-32
26.05.14	E3	1-80, 1-83, (1-98)
02.06.14	E3	1-80, 1-83, (1-98)
09.06.14	–	–
16.06.14	E4	2-4, 2-14, 3-1, 3-7, (3-11)
23.06.14	E5	TBA
30.06.14	E5	TBA
07.07.14	E6	TBA
14.07.14	E6	TBA

Review

- week 5 & week 6
- Slides: Descriptive Statistics (cca 76-120)

Note:
! Notation from lecture: Having n observation $x_1, \ldots, x_n$ with k different values $x_1, \ldots, x_k$

Mean absolute deviation about median (MAD median)

$$d = \frac{1}{n} \cdot \sum_{j=1}^{k} |x_j - c| \cdot h(x_j) = \sum_{j=1}^{k} |x_j - c| \cdot f(x_j),$$

where $c = x_{0,5}$ or $c = \overline{x}$

Sample variance

$$\begin{aligned} s^2 &= \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \overline{x}^2 \\ &= \frac{1}{n} \cdot \sum_{j=1}^{k} (x_j - \overline{x})^2 \cdot h(x_j) = \sum_{j=1}^{k} (x_j - \overline{x})^2 \cdot f(x_j) \end{aligned}$$

Standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n} \cdot \sum_{j=1}^{k} (x_j - \overline{x})^2 \cdot h(x_j)} = \sqrt{\sum_{j=1}^{k} (x_j - \overline{x})^2 \cdot f(x_j)}$$

Contingency tables

X \ Y	y_1	$\cdots$	y_r	marg.distr
x_1				$\vdots$
$\vdots$		h_{ij}		$h_{i\bullet}$
x_m				$\vdots$
marg.distr	$\cdots$	$h_{\bullet j}$	$\cdots$	n

Joint distribution

Absolute frequency $h(x_i, y_j) = h_{ij}$
Relative frequency $f(x_i, y_j) = f_{ij} = \frac{h_{ij}}{n}$

Marginal distribution

Absolute frequency for X $h_{i\bullet} = \sum_{j=1}^{r} h_{ij} \quad i = 1, \ldots, m$
Relative frequency for X $f_{i\bullet} = \sum_{j=1}^{r} f_{ij} \quad i = 1, \ldots, m$
Absolute frequency for Y $h_{\bullet j} = \sum_{i=1}^{m} h_{ij} \quad j = 1, \ldots, r$
Relative frequency for Y $f_{\bullet j} = \sum_{i=1}^{m} f_{ij} \quad j = 1, \ldots, r$

Conditional distribution Conditional relative frequency of X given Y $f(x_i|y_j) = \frac{f_{ij}}{f_{\bullet j}} = \frac{h_{ij}}{h_{\bullet j}}$

Independence Two variable X and Y are independent if $f(x_i, y_j) = f(x_i) \cdot f(y_j)$ for all x_i, y_j

Empirical covariance

$$\begin{aligned} s_{xy} &= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{r}(x_i-\overline{x})(y_j-\overline{y})\,h_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{r}(x_i-\overline{x})(y_j-\overline{y})\,f_{ij} \\ &= \left\{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{r}x_i y_j h_{ij}\right\} - \overline{x}\,\overline{y} = \left\{\sum_{i=1}^{m}\sum_{j=1}^{r}x_i y_j f_{ij}\right\} - \overline{x}\,\overline{y} \end{aligned}$$

Bravais-Pearson correlation coefficient

$$\begin{aligned} r_{xy} = r_{yx} &= \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2 \cdot \sum_{i=1}^{n}(y_i-\overline{y})^2}} \qquad \text{with } -1 \le r_{xy} \le +1 \\ &= \frac{n\cdot\sum_{i=1}^{n}x_i\cdot y_i - \sum_{i=1}^{n}x_i \cdot \sum_{i=1}^{n}y_i}{\sqrt{\left\{n\cdot\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right\}\left\{n\cdot\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right\}}} \end{aligned}$$

Spearman's rank correlation coefficient $r_s = 1 - \dfrac{6\cdot\sum_{i=1}^{n}d_i^2}{n\cdot(n^2-1)}$ with $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ and $-1 \le r_s \le +1$

Kendall' rank correlation coefficient $\tau = \dfrac{P-Q}{P+Q}$ with $-1 \le \tau \le +1$

P number of pairs of observations where $x_i < x_j$ and $y_i < y_j$
Q number of pairs of observations where $x_i < x_j$ and $y_i > y_j$

Quadratic Contingency (χ^2-coefficient)

$$K^2 = \sum_{i=1}^{m}\sum_{j=1}^{r}\frac{(h_{ij} - \frac{1}{n}\cdot h_{i\bullet}\cdot h_{\bullet j})^2}{\frac{1}{n}\cdot h_{i\bullet}\cdot h_{\bullet j}} = n\cdot\sum_{i=1}^{m}\sum_{j=1}^{r}\frac{(f_{ij} - f_{i\bullet}\cdot f_{\bullet j})^2}{f_{i\bullet}\cdot f_{\bullet j}}$$

Contingency coefficient und corrected contingency coefficient

$$C = \sqrt{\frac{K^2}{n+K^2}} \qquad C_{corr} = C\cdot\sqrt{\frac{C^*}{C^*-1}} \qquad C^* = \min\{\text{number of rows, number of columns}\}$$

Exercises

Exercise 1-80

A survey of 300 viewers from 2 types of sport events (tennis and football) gave the following result: 52 people visit tennis frequently and football rarely, 62 people rarely tennis and football frequently, 118 people visit both frequently, and 68 people both rarely.

100 of viewers are over 30 years old. 24 of them visit tennis frequently and football rarely, 14 rarely tennis and often football, 6 often both and 56 rarely both.

From the people under 30: 28 visit tennis frequently and football rarely, 48 rarely tennis and frequently football, 112 both frequently and 12 both rarely.

a) for all viewers,

b) for viewers 30+ years old,

c) for viewers under 30,

has to be tested.

d) Evaluate the previous results a) - c).

Exercise 1-80

Student recorded the outside temperature X (in Celsius) and the duration of his way to university Y (in minutes).

x_i	-20	-10	0	10	20
y_i	60	40	35	20	20

How strong is the correlation between these two features?

Exercise 1-98

The following coefficients were calculated by an economist with the help of statistical program for all variables in data set:

a) mean

b) Bravais.-Pearson corr. coefficient

c) geometric mean

d) interquartile range

e) Kendalls rank correlation coefficient τ

f) Corr. contingency coefficient

g) covariance

h) variance

i) median

j) mode

k) Quadratic contingency χ^2

l) span

m) Spearman's rank correlation coefficient

n) standard deviation

o) variance

p) none of the coefficients

Help the economist to pick up the coefficients:

1. that represent the robust measure of central tendency,
2. that can be used as a measure of association between two nominal scaled variables,
3. that represent the robust measure of dispersion for a metric variable,
4. that remain unchanged for metric variables under linear transformation (i.e. y = a + xb).