Chi-squared test: Residuals

April 26, 2019

Applied STAtistics group at AAU

Department of Mathematical Sciences

Aalborg University

AALBORG UNIVERSITY

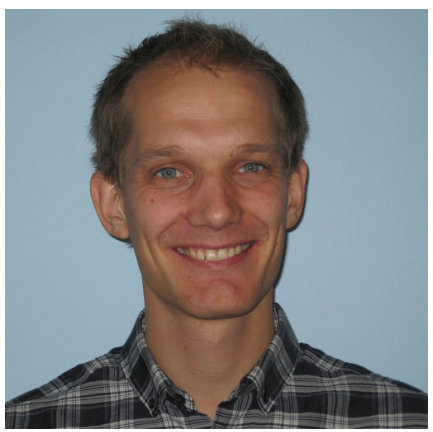
DENMARK

Introduction

Outline of session:

- Residuals for a Chi-squared test in general
- Residuals for our data example

Lecturer for this session is Ege Rubak, Dept. of Math. Sciences, AAU



Residual analysis

- If we reject the hypothesis of independence it can be of interest to identify the significant deviations.
- In a given cell in the table, $f_o - f_e$ is the deviation between data and the expected values under the null hypothesis.
- We assume that $f_e \geq 5$.
- If H_0 is true, then the standard error of $f_o - f_e$ is given by

$$se = \sqrt{f_e(1 - \text{rowProportion})(1 - \text{columnProportion})}$$

- The corresponding z-score

$$z = \frac{f_o - f_e}{se}$$

should in 95% of the cells be between ± 2. Values above 3 or below -3 should rarely appear.

Residual analysis for example data

	Grades	Popular	Sports
Rural	57	50	42
Suburban	87	42	22
Urban	103	49	26

	Grades	Popular	Sports
Rural	77.0	44.0	28.1
Suburban	78.0	44.5	28.4
Urban	92.0	52.5	33.5

- In popKids table cell `Rural and Grade` we got $f_e = 77.0$ and $f_o = 57$. Here columnProportion$= 51.7\%$ and rowProportion$= 31.2\%$ (not shown in table).
- We can then calculate

$$z = \frac{57 - 77}{\sqrt{77(1 - 0.517)(1 - 0.312)}} = -3.95.$$

- Compared to the null hypothesis there are way too few rural kids who find grades important.

Residual analysis for example data

- In summary: The standardized residuals allow for cell-by-cell (f_e vs f_o) comparision:

	Grades	Popular	Sports
Rural	-3.951	1.31	3.523
Suburban	1.767	-0.5484	-1.619
Urban	2.087	-0.7274	-1.819