Contingency tables and independence

April 28, 2019

Applied STAtistics group at AAU

Department of Mathematical Sciences

Aalborg University

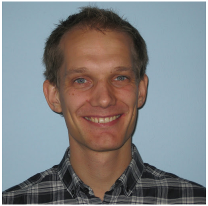
AALBORG UNIVERSITY
DENMARK

Introduction

Outline of session:

- Contingency tables
- Independence and expected table counts

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



A contingency table

- We consider the dataset `popularKids`, where we study **association** between 2 **qualitative variables** (factors): `Goals` and `Urban.Rural`.
- Based on a sample we make a cross tabulation of the factors and we get a so-called **contingency table** (krydstabel).

	Grades	Popular	Sports	Total
Rural	57	50	42	149
Suburban	87	42	22	151
Urban	103	49	26	178
Total	247	141	90	478

A conditional distribution

- Another representation of data is the percent-wise distribution of `Goals` for each level of `Urban.Rural`, i.e. the sum in each row of the table is 100 (up to rounding):

	Grades	Popular	Sports	Sum
Rural	38.3	33.6	28.2	100.1
Suburban	57.6	27.8	14.6	100.0
Urban	57.9	27.5	14.6	100.0

- Here we will talk about the **conditional distribution** of `Goals` given `Urban.Rural`.
- An important question could be:
  - Are the goals of the kids different when they come from urban, suburban or rural areas? I.e. are the rows in the table significantly different?
- There is (almost) no difference between urban and suburban, but it looks like rural is different.

Independence

- Recall, that two factors are **independent**, when there is no difference between the population's distributions of one factor given the other.
- Otherwise the factors are said to be **dependent**.
- If we e.g. have the following conditional **population distributions** of `Goals` given `Urban.Rural`:

	Grades	Popular	Sports
Rural	50	30	20
Suburban	50	30	20
Urban	50	30	20

- Then the factors `Goals` and `Urban.Rural` are independent.
- We take a sample and "measure" the factors F_1 and F_2. E.g. `Goals` and `Urban.Rural` for a random child.
- The hypothesis of interest today is:

$$H_0 : F_1 \text{ and } F_2 \text{ are independent}, \qquad H_a : F_1 \text{ and } F_2 \text{ are dependent}.$$

Independence for the data

- Our best guess of the distribution of `Goals` is the relative frequencies in the sample:

Grades	Popular	Sports
51.7	29.5	18.8

- If we assume independence, then this is also a guess of the conditional distributions of `Goals` given `Urban.Rural`.
- The corresponding expected counts in the sample are then:

	Grades	Popular	Sports	Sum
Rural	77.0 (51.7%)	44.0 (29.5%)	28.1 (18.8%)	149.0 (100%)
Suburban	78.0 (51.7%)	44.5 (29.5%)	28.4 (18.8%)	151.0 (100%)
Urban	92.0 (51.7%)	52.5 (29.5%)	33.5 (18.8%)	178.0 (100%)
Sum	247.0 (51.7%)	141.0 (29.5%)	90.0 (18.8%)	478.0 (100%)

Calculation of expected table

	Grades	Popular	Sports	Sum
Rural	77.0 (51.7%)	44.0 (29.5%)	28.1 (18.8%)	149.0 (100%)
Suburban	78.0 (51.7%)	44.5 (29.5%)	28.4 (18.8%)	151.0 (100%)
Urban	92.0 (51.7%)	52.5 (29.5%)	33.5 (18.8%)	178.0 (100%)
Sum	247.0 (51.7%)	141.0 (29.5%)	90.0 (18.8%)	478.0 (100%)

- We note that
  - The relative frequency for a given column is columnTotal divided by tableTotal. For example `Grades`, which is $\frac{247}{478} = 51.7\%$.
  - The expected value in a given cell in the table is then the cell's relative column frequency multiplied by the cell's rowTotal. For example `Rural` and `Grades`: $149 \times 51.7\% = 77.0$.
- This can be summarized to:
  - The expected value in a cell is the product of the cell's rowTotal and columnTotal divided by tableTotal.