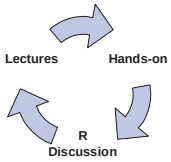
Course 02441: Applied Statistics and Statistical Software

Lasse Engbo Christiansen, IMM

Today's program: outline

- Tabular data – counts and proportions
- Hands-on exercises



Methods covered in the course

- Descriptive statistics — Day I
- Comparing treatment means (t-test and non-parametric tests) — Day I
- Multiple regression analysis — Day II
- Analysis of variance — Day III
- Analysis of proportions and counts — Day IV
- The general linear model — Day V

Tabular Data

- Tests of single proportions are generally based on the binomial distribution, $X \sim B(n,p)$
- For large samples this can be well approximated by the normal distribution

Example: Comparing two independent proportions

- Linus Pauling, recipient of Nobel Prizes in Chemistry and in Peace, advocated the use of vitamin C for preventing common cold.
- A Canadian study investigated this claim. 818 volunteers where randomly assigned into two groups – placebo and vitamin C.
- The experiment was double-blinded, meaning that neither the volunteers nor the physicians who examined the volunteers knew what treatment the volunteers received.

Example: Comparing two independent proportions

	Cold	No Cold	Totals
Placebo	335	76	411
Vitamin C	302	105	407
Totals	637	181	818

Example: Comparing two independent proportions

The test statistics is obtained by calculating the expected number of counts for each cell and then calculate

$$\chi^2 = \sum_{i,j} (Observed - Expected)^2 / Expected$$

The p-value is found by comparing this test statistic to a chi-squared distribution on (r-1)(c-1) degrees of freedom, where r is the number of rows and c is the number of columns in the table

Fisher's exact test

- If the number of observation is small the chi-square test may be invalid
- In such cases, Fisher's exact test can be used

RC tables of counts

- More general tables arise from counts of subjects falling into cross-classifications of several factors, each with many levels
- The test statistics is calculated as

$$\chi^2 = \sum_{i,j} (Observed - Expected)^2 / Expected$$

Example

- The table on the next slide shows the number of homicides of children by their parents, over 10 years in Canada, categorized according to the parent-offspring sex combination and according to age category

Example continued

	0-1 years	2-5 years	6-10 years	11-16 years	>16 years
Male/ male	24	21	21	29	104
Male/ female	17	27	10	14	47
Female/ male	53	21	19	9	8
Female/ female	50	27	5	4	15

RC tables of counts

If the p-value indicates dependence between the row and the column categories, then the this dependence can be further analyzed by considering the contributions (each term or cell) in the calculated test statistics

$$\chi^2 = \sum_{i,j} (Observed - Expected)^2 / Expected$$