Cheat Sheet: Chi-square test

Measurement and Evaluation of HCC Systems

Scenario

Use the chi-squre if you want to test the association between two nominal variables `var1` with *k* levels and `var2` with *r* levels in your dataset `data`.

Power analysis for linear regression

- Use Test family "χ^2 tests", "Goodness-of-fit tests: Contingency tables".
- A power analysis has four variables: Effect size, α (usually .05), power (usually .85), and *N*. If you know three of these, G*Power will calculate the fourth. Select the correct type of power analysis, based on the information you have, and what you want to find out.
- "Df" is the degrees of freedom of your test, usually (*k*–1)(*r*–1).
- By clicking on "Determine", you can compute the effect size *w* from the observed (p(H1)) and expected (p(H0)) frequencies.
- Click on "Calculate" to calculate the missing parameter.

Plotting a mosaic-plot

- Create a mosaic-plot:
  ```
  mosaicplot(table(data$var1, data$var2), shade=T)
  ```
  Cells that are shaded blue have a higher observed frequency than expected; cells that are shaded red have a lower observed frequency than expected.

Pre-testing assumptions

- Make sure that your observations are independent.
- At least 20% of the expected frequencies should be > 5. All expected frequencies should be > 1. You can get the expected frequencies from the chi-square test output.

Running the test

- Run the chi-square test by using the function `CrossTable` in the `gmodels` package:
  ```
  CrossTable(data$var1,data$var2,expected=T,fisher=T,sresid=T,format="SPSS")
  ```
- Interpret the numbers in each cell of the resulting table:
  - Line 1: observed count
  - Line 2: predicted count

    - Line 3: Standardized deviance
    - Line 4: Percentage of row total
    - Line 5: Percentage of column total
    - Line 6: Percentage of grand total
    - Line 7: Standardized residual
- Interpret the test results:
    - The chi-square test, *df*, and *p*-value: these test whether there is an association.
    - The chi-square test with Yates' correction adds 0.5 to each cell (works better for smaller datasets).
    - Fisher's exact test (only for 2x2 tables) is more precise than the chi-square test when the expected frequencies are too low (see Pre-testing assumptions); this test also has a confidence interval, and the two one-sided versions of this test are listed as well.
    - The minimum frequency and the cells with expected frequency < 5 can be used to test the assumptions (see Pre-testing assumptions).
- In 2x2 tables you can also present odds ratios. You can turn any table into a 2x2 table by subsetting (only looking at 2 levels of a certain variable) or collapsing (e.g. comparing a certain level against all other levels) the data.

Reporting

- Use the following format to report on a chi-square test (replace the full names (not just the variable names) of var1 and var2, and replace the xx'es with the actual numbers):
  "There was a significant association between [var1] and [var2] χ^2(x) = xx.xx, *p* = .xxx. The odds of [var1 level B] rather than [var1 level A] were x.xx times higher for [var2 level D] than for [var2 level C] (95% CI: [x.xx, x.xx]).