Practical "Analysis of categorical data"
P3

M. Hauptmann

Exercise 1 Eye and hair color have been recorded for 762 children from 2 different regions in data set `color.sav`. For each combination of hair color (1=dark, 2=medium, 3=fair, 4=black, 5=red), eye color (1=blue, 2=green, 3=brown) and region (1, 2), variable `count` represents the frequency of such children.

- Create contingency tables showing the numbers of children by hair color and eye color, overall and separately for each region.
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs and fill in the form.
- Does the distribution of hair color in each region differ from 30% fair, 12% red, 30% medium, 25% dark, and 3% black? Which test is appropriate? Report the p-value and, if less than .05, describe how the observed distribution differs from the specified distribution.
  SPSS tip: In order to limit analyses to region 1, click Data – Select Cases and fill in the form. Then click Analyze – Nonparametric Tests – Legacy Dialogs – Chi-square and provide the specified distribution as expected values.
- In the contingency table of hair color by eye color for both regions combined, evaluate whether both variables are independent. Report the appropriate test, its p-value and, if significant, describe the nature of the association. Perform the same test separately for each region. How do the region-specific results relate to the overall result?
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs and fill in the form.

Exercise 2 The mussel data set `mussel.sav` includes the allele frequencies at the Lap locus in the mussel *Mytilus trossulus* on the Oregon coast (McDonald and Siebenaller, *Evolution* 1989). At four estuaries (variable `location`), samples were taken from inside the estuary and from marine habitat outside the estuary (variable `habitat`). There were 3 common alleles and a couple of rare alleles, here grouped into 94 and non-94 alleles. For each combination of variables, variable `count` represents the frequency of such observations.

- There is a smaller proportion of 94 alleles in the estuarine location of each estuary when compared with the marine location – is this difference significant when accounting for location? Which test can be used to answer the question?
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs, fill in the form and request the Cochran-Mantel-Haenszel test after clicking on Statistics...

Exercise 3 Data set `wheeze.sav` contains data on body mass index (BMI, 1=underweight, 2=normal, 3=overweight, 4=obese) and wheezing after exercise (0=no, 1=yes) from a cross-sectional survey among 4,010 children aged 13–14 yrs in Brazil (Cassol et al., *Jornal de Pediatria* 2005). Variable `count` indicates the frequency of observations.

- Does the prevalence of wheezing differ between the 4 BMI groups? Does the prevalence of wheezing increase with increasing BMI group? Which tests answer those questions?
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs, fill in the form and request Chi-square after clicking on Statistics...

Exercise 4 The drug toxicity data set `drugtox.sav` includes data on patients treated with 4 doses of a drug (in mg) and 4 degrees of toxicity: 1=mild, 2=moderate, 3=severe, and 4=drug death (Hoyle: Statistical strategies for small sample research 1999). Variable `count` indicates the frequency of patients for a given dose and toxicity.

- Make a table of toxicity by dose group and describe the observed data. Which tests are appropriate for evaluating a possible association between drug dose and toxicity? Perform those and interpret the results. Note that data outside the mild toxicity category are sparse and it therefore appears prudent to request exact p-values based on Monte Carlo simulations.
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs, fill in the form and request Chi-square after clicking on Statistics...

- Someone makes the case that drug death should be considered catastrophic and orders of magnitude more serious than severe toxicity and suggests to use a score of 10,000 for drug death. Perform the test again with the modified score for the highest category. How do the results differ from the standard scores 1, 2, 3, 4?
  SPSS tip: Create a new variable with 10,000 as the maximum value by clicking Transform – Compute Variable, fill in the form and run the Chi-square test as above on the new variable.

Exercise 5 Richard Feynman, member of the presidential commission to investigate the explosion of the Challenger space shuttle, suspected that the low temperature at take-off caused the O-rings to fail. The space-shuttle O-ring data set `oring.sav` contains data on temperature and O-ring failures from 24 previous space shuttle flights (Feynmann: Why do we care what other people think, 1988).

- How much evidence do those data provide for Feynman's hypothesis? Perform a Kruskal-Wallis test and describe the results.
  SPSS tip: Click Analyze – Nonparametric Tests – Legacy Dialogs – K Independent Samples, fill in the form and make sure the Kruskal-Wallis test is requested.

- Alternatively, one could perform a linear-by-linear association test with the number of O-ring incidents as row scores and take-off temperatures as column scores. Compare the result of this test with a standard Pearson chi-square test which ignores the ordering. Which test is more powerful?
  SPSS tip: Click Analyze – Descriptive Statistics – Crosstabs, fill in the form and request Chi-square after clicking on Statistics...

Exercise 6 The cytomegalovirus data set `cmv.sav` contains data on a comparison of formaldehyde and acetone fixations for the cytomegalovirus antigenemia assay. Variables `AC_detected` and `FA_detected` indicate whether CMV was detected (0=no, 1=yes) by acetone (AC) and formaldehyde (FA), respectively. Variable `count` indicates the frequency of samples per group. In 405 samples, cytomegalovirus was detected 36 times (8.8%) using formaldehyde and 22 times (5.4%) using acetone (Perez et al., *J Clin Microbiol* 1995).

	Formaldehyde		
Acetone	Detected	Not detected	
Detected	18	4	22/405=5.4%
Not detected	18	365	383
	36/405=8.8%	369	405

- Evaluate whether the proportion of detection is significantly greater for formaldehyde compared with acetone. Which test is appropriate?
  SPSS tip: Tell SPSS about the frequency with which each record occurred by clicking Data – Weight Cases... and selecting the appropriate variable. Then click Analyze – Descriptive Statistics – Crosstabs, fill in the form and make sure McNemar is requested after clicking on Statistics...

- Calculate McNemar's test using the formula presented in the lecture, which only uses the numbers of discordant pairs.