Lab 12: Chi-square tests of association

STAT218

The purpose of this lab is to learn to implement tests of independence/association for two-way contingency tables:

* inference and residual analysis in tables
* extension to tables

Much of the focus of the lab activity is on the setting. You’ll replicate the examples from class using the asthma data from an NHANES subsample, and practice using the diabetes\_meds data.

The diabetes\_meds data comes from an observational study of 227,571 Medicare beneficiaries who initiated treatment with one of two diabetes medications. In the study, it was recorded whether each patient reported cardiovascular problems.

Applications in the case will both use the famuss dataset.

library(tidyverse)  
load('data/asthma.RData')  
load('data/diabetes\_meds.RData')  
load('data/famuss.RData')

### Inference for tables

#### Basic implementation and checking assumptions

The test of association/independence in two-way tables pertains to the hypotheses:

In words, the hypotheses are: (null) the row variable and column variable are independent; (alternative) there is an association between the row and column variables. This is fairly straightforward to implement in R using prop.test(...), which takes a contingency table as input:

# make a two-way table  
table(asthma$sex, asthma$asthma)

asthma no asthma  
 female 49 781  
 male 30 769

# pass to chisq.test  
table(asthma$sex, asthma$asthma) |>  
 chisq.test()

Pearson's Chi-squared test with Yates' continuity correction  
  
data: table(asthma$sex, asthma$asthma)  
X-squared = 3.6217, df = 1, p-value = 0.05703

The test (with Yates’ continuity correction) produces , so we’d fail to reject . The typical narrative style for the report is:

The data provide suggestive but insufficient evidence at the 5% significance level that sex and asthma prevalence are associated ( = 3.62 on 1 degree of freedom, *p* = 0.057).

|  |
| --- |
| Your turn 1 |
| Test whether diabetes medication is associated with cardiovascular problems. Carry out inference at the 5% significance level and report the result of the test in the narrative style above.  # test whether diabetes medication is associated with cardiovascular problems |

These inferences rely on the assumption that expected counts are all over 10. The expected counts can be inspected directly by storing the result of the test:

# store test result  
asthma.rslt <- table(asthma$sex, asthma$asthma) |>  
 chisq.test()  
  
# view expected counts to check assumptions  
asthma.rslt$expected

asthma no asthma  
 female 40.25169 789.7483  
 male 38.74831 760.2517

Each expected count exceeds ten, so the test is appropriate to use.

|  |
| --- |
| Your turn 2 |
| Check the expected counts for the inference you performed above to determine whether conditions for the Chi-square test are met.  # store test result  # view expected counts to check assumptions |

Luckily, prop.test(...) will print a warning if the counts are too small. For instance:

# example using a dataset with low counts  
openintro::malaria |> table() |> chisq.test()

Warning in chisq.test(table(openintro::malaria)): Chi-squared approximation may  
be incorrect

Pearson's Chi-squared test with Yates' continuity correction  
  
data: table(openintro::malaria)  
X-squared = 4.6561, df = 1, p-value = 0.03094

We can check that indeed the assumptions are not met by inspecting expected counts:

# all expected counts are under ten  
malaria.rslt <- openintro::malaria |> table() |> chisq.test()  
malaria.rslt$expected

outcome  
treatment infection no infection  
 placebo 3.3 2.7  
 vaccine 7.7 6.3

#### Residual analysis

If significant results (or suggestive evidence) is obtained from the Chi-square test, it is useful to be able to say which categories account for the association (or possible association).

In this context, residuals are standardized differences in expected and observed counts:

These convey information about which value combinations are ‘unusual’ under the assumption of independence between row and column variables.

* positive residual: larger-than-expected count
* negative residual: smaller-than-expected count

You can retrieve the residuals from the test and inspect for the largest (absolute) values:

# view residuals  
asthma.rslt$residual

asthma no asthma  
 female 1.3788982 -0.3113006  
 male -1.4053932 0.3172821

Here, the residuals suggest that asthma rates are higher than expected among women and lower than expected among men.

|  |
| --- |
| Your turn 3 |
| Check the residuals from your inference of whether diabetes medication is associated with cardiovascular problems. See if you can explain any inferred association.  # check residuals from your inference above; what explains the association? |

#### Measuring association by a difference in proportions

Once you’ve inferred an association between two categorical variables, it’s helpful to be able to provide a quantitative measure.

A difference in proportions is one potential measure, though not always possible to compute (*e.g.*, for outcome-based sampling). Here, since data are a random sample from the target population (U.S. adults), we can estimate the difference in the proportion of individuals with asthma between men and women:

# use prop.test to get estimates and confidence interval  
table(asthma$sex, asthma$asthma) |>  
 prop.test(conf.level = 0.90)

2-sample test for equality of proportions with continuity correction  
  
data: table(asthma$sex, asthma$asthma)  
X-squared = 3.6217, df = 1, p-value = 0.05703  
alternative hypothesis: two.sided  
90 percent confidence interval:  
 0.002841347 0.040137075  
sample estimates:  
 prop 1 prop 2   
0.05903614 0.03754693

For a report, we’d combine this with the inferred association as follows:

The data provide suggestive but insufficient evidence at the 5% significance level that sex and asthma prevalence are associated ( = 3.62 on 1 degree of freedom, *p* = 0.057). With 90% confidence, asthma prevalence is estimated to be between 0.28 and 4.01 percentage points higher among women compared with men, with a point estimate of 2.15 percentage points.

(A 90% interval is reported, not consistent with the significance level of the test, to point to the direction of possible association.)

|  |
| --- |
| Your turn 4 |
| Provide point and interval estimates for the difference in proportions of patients experiencing cardiovascular problems on each diabetes medication as a measure of association.  (Unlike the example above, your confidence level for the interval should match the significance level of your test, as usual.)  # estimate the difference in proportions of patients experiencing cardiovascular problems |

### Inference for tables

The implementation of the test is identical for larger tables, as are the means of obtaining expected counts and residuals from the test. The only difference is the check on assumptions. For tables, the following two conditions should be met:

1. All expected counts are at least 1
2. Roughly 80% or more of expected counts are at least 5

The commands below illustrate the implementation, assumption check, and residual analysis.

# perform test  
famuss.rslt <- table(famuss$race, famuss$genotype) |>  
 chisq.test()  
  
# check expected counts to assess assumptions  
famuss.rslt$expected

CC CT TT  
 African Am 7.850420 11.84370 7.305882  
 Asian 15.991597 24.12605 14.882353  
 Caucasian 135.783193 204.85210 126.364706  
 Hispanic 6.687395 10.08908 6.223529  
 Other 6.687395 10.08908 6.223529

# interpret test result  
famuss.rslt

Pearson's Chi-squared test  
  
data: table(famuss$race, famuss$genotype)  
X-squared = 19.4, df = 8, p-value = 0.01286

# residual analysis  
famuss.rslt$residuals

CC CT TT  
 African Am 2.90863193 -1.69802497 -0.85310170  
 Asian 1.25242978 -1.24720387 0.28971360  
 Caucasian -0.92538910 0.77888407 -0.03244366  
 Hispanic -1.03920927 -0.02804356 1.11294757  
 Other 0.12088363 0.28678513 -0.49045147

The interpretation of this result would be:

The data provide moderate evidence of an association between genotype and race for the ACTN3 gene ( = 19.4 on 8 degrees of freedom, *p* = 0.01286).

While there isn’t a standard style or language for the residual analysis, and appropriate interpretation would be:

The data further suggest that African American and Asian populations have higher-than-expected and lower-than-expected CC and CT genotype frequencies, respectively; and Hispanic populations have higher-than-expected and lower-than-expected TT and CC genotype frequencies, respectively.

As an aside, measures of association are trickier for tables; typically one would calculate a *set* of measures that capture pairwise comparisons.

|  |
| --- |
| Your turn 5 |
| Test for an association between genotype and sex.   * check the test assumptions * assuming they are met, interpret the test * if the test is significant, carry out a residual analysis to explain the inferred association   # perform test  # check expected counts to assess assumptions  # interpret test result  # residual analysis |

### Practice problems

1. [L7] The mammogram dataset contains observations from a 30-year study to investigate the effectiveness of mammograms versus a standard non-mammogram breast cancer exam on survival. The study was conducted in Canada with 89,835 female participants: during a 5-year screening period, each woman was randomized to either receive annual mammograms or standard physical exams for breast cancer; the study recorded the number of breast cancer deaths during a 25-year follow-up period in each group.
   1. Explain the hypotheses for the Chi-square test of association in words.
   2. Check assumptions for the test.
   3. If assumptions are met, test for association at the 5% significance level and report the result; if the test is significant, perform a residual analysis to explain the inferred association.
2. [L7] The smoking dataset contains observations from a retrospective case-control study in which smoking status was recorded for 86 lung cancer patients and 86 healthy patients.
   1. Construct a contingency table of the data.
   2. Which proportions are possible to estimate using data from this study?
   3. Check assumptions for the test of association.
   4. If assumptions are met, perform the test at the 5% level. Interpret the result in the usual narrative style.
   5. (Extra credit) Make a plot of the residuals that allows you to identify which group/outcome combinations differ from expectations under independence.
3. [L6] The gss data contains several demographic measurements for a random sample of 500 U.S. adults.
   1. Provide a point estimate for the proportion of U.S. adults with a college degree.
   2. Provide a 99% confidence interval for the proportion. Interpret the interval in context.
   3. Test whether at least three in ten U.S. adults have a college degree at the 5% significance level. If the result is significant, provide a corresponding lower confidence bound.
4. [LX] (Extra credit) Again using the gss data, test whether political party is associated with socioeconomic class.
   1. Check assumptions for the test.
   2. If appropriate, perform the test at the 1% significance level. Interpret the result in context, and if a significant association is found, perform a residual analysis to explain the inferred association.