Chi-squared Tests of Independence

Stacey Hancock

2/10/2022

There are two popular large-sample test statistics used for testing independence between two categorical variables: the *Pearson chi-squared statistic*,

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \exp_{ij})^2}{\exp_{ij}},$$

and the *likelihood ratio test statistic*,

$$G^2 = 2\sum_{i,j} n_{ij}\log\left(\frac{n_{ij}}{\exp_{ij}}\right)$$

In each case, $\exp_{ij}$ is called the "expected count" for the (i, j)th cell, and is computed by

$$\exp_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n}.$$

This is the value we would expect to see in the (i, j)th cell *if the two variables were independent*. Why? It's the value that would make the distribution of conditional proportions identical for each row (or column).

Each of these test statistics is a summary statistic that attempts to measure how far away the observed counts are from what we'd expect to see under the null hypothesis of independence, summarized into a single value. In an $I \times J$ table, for large samples (e.g., at least 5 in each cell), each of these statistics has an approximate χ^2 distribution with degrees of freedom $(I-1) \times (J-1)$ when H_0 holds.

Example 1: Swedish Fish Consumption and Prostate Cancer

Data input as a table

Medical researchers followed 6272 Swedish men for 30 years to see if there was an association between the amount of fish in their diet and prostate cancer ("Fatty Fish Consumption and Risk of Prostate Cancer," Lancet, June 2001).

Here are the data (in a 2x2 table):

```
fish <- matrix(c(110,2420,2769,507,
                 14,201,209,42),
               nrow = 4, ncol = 2,
               dimnames = list(fish_consumption = c("never_seldom","small","moderate","large"),
                               prostate_cancer = c("no","yes")))
fish
```

```
##                prostate_cancer
## fish_consumption   no yes
##     never_seldom  110  14
##     small        2420 201
##     moderate     2769 209
##     large         507  42
```

4 x 2 → DF: (4-1)(2-1) = 3

Since we have a large sample (at least 5 in each cell), we are going to use a chi-squared test of independence to test the null hypothesis that fish consumption level and incidence of prostate cancer are independent. First, let's do the calculations "by hand".

Check expected cell counts below

We can calculate the expected counts using a bit of matrix algebra. Specifically, calculate the outer product between the vector of row sums and the vector of column sums.

```
# Row sums
rowSums(fish)
```

```
## never_seldom        small     moderate        large
##          124         2621         2978          549
```

```
# Column sums
colSums(fish)
```

```
##   no  yes
## 5806  466
```

```
# Sample size
sum(fish)
```

```
## [1] 6272
```

```
# Expected counts = (row total) x (col total)/6272
# Calculate using an outer product of row and col sums
exp_counts <- outer(rowSums(fish), colSums(fish))/sum(fish)
```

Stacey – discuss more

Now that we have our observed and expected counts, we can calculate our chi-squared test statistic and the p-value. With four rows and two columns, our degrees of freedom are $(4-1) \times (2-1) = 3$.

```
test_stat <- sum((fish - exp_counts)^2/(exp_counts))
test_stat
```

```
## [1] 3.677281
```

```
pval <- pchisq(test_stat, df = 3, lower.tail=FALSE)
pval
```

```
## [1] 0.2984868
```

With this large of a p-value, we have little evidence of an association between fish consumption level and incidence of prostate cancer among Swedish men similar to those recruited for the study.

Now, let's check our calculations using the `chisq.test` function. If we assign the output of this function to an object, we can then extract many features of the test out of this object.

```
fish_test <- chisq.test(fish, correct=FALSE)
attributes(fish_test)
```

```
## $names
## [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
## [7] "expected"  "residuals" "stdres"
##
## $class
## [1] "htest"
```

The code below displays (in the following order):

- the default test output
- observed counts
- expected counts
- chi-squared test statistic
- p-value
- residuals
- standardized residuals

```
fish_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  fish
## X-squared = 3.6773, df = 3, p-value = 0.2985
```

```
fish_test$observed
```

```
##                 prostate_cancer
## fish_consumption   no yes
##     never_seldom  110  14
##     small        2420 201
##     moderate     2769 209
##     large         507  42
```

```
fish_test$expected
```

```
##                prostate_cancer
## fish_consumption        no       yes
##     never_seldom  114.7870   9.21301
##     small        2426.2637 194.73629
##     moderate     2756.7392 221.26084
##     large         508.2101  40.78986
```

e_{ij} ↑ row i, column j

```
fish_test$statistic
```

```
## X-squared
##  3.677281
```

```
fish_test$p.value
```

```
## [1] 0.2984868
```

```
fish_test$residuals
```

```
##                prostate_cancer
## fish_consumption          no        yes
##     never_seldom -0.44680309  1.5771091
##     small        -0.12716358  0.4488573
##     moderate      0.23351912 -0.8242672
##     large        -0.05368019  0.1894784
```

Pearson residuals → Pearson χ^2-test

```
fish_test$stdres
```

```
##                prostate_cancer
## fish_consumption         no        yes
##     never_seldom -1.6556264  1.6556264
##     small        -0.6114627  0.6114627
##     moderate      1.1821532 -1.1821532
##     large        -0.2061652  0.2061652
```

"Standardized residuals" = Deviance residuals → LRT

The residuals are the Pearson residuals:

$$\frac{\text{observed} - \text{expected}}{\sqrt{\text{expected}}}$$

Note that the sum of the squared residuals is equal to the chi-squared test statistic.

```
fish_test$statistic
```

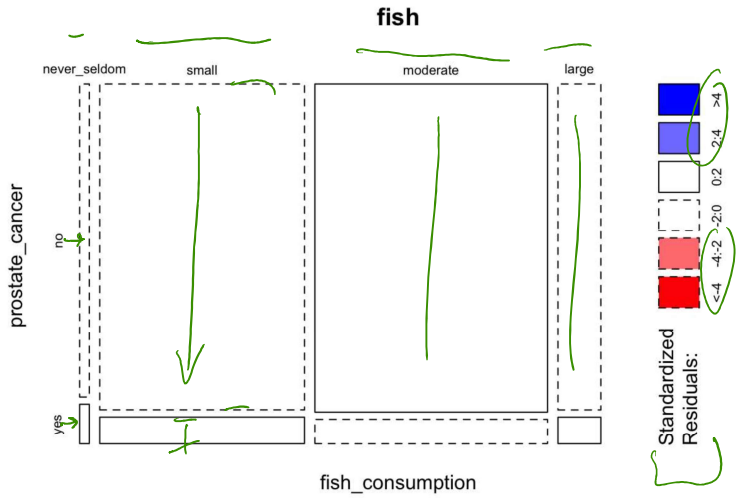
```
## X-squared
##  3.677281
```

```
sum(fish_test$residuals^2)
```

```
## [1] 3.677281
```

The standardized residuals are calculated as given in the formula (2.5) in Section 2.4.5 of the Agresti textbook. For large samples, under H_0, these standardized residuals have an approximate standard normal distribution. Thus, standardized residuals beyond -2 or 2 indicate lack of fit with the independence assumption. Visually, this can be seen in this mosaic plot:

```
mosaicplot(fish, shade=TRUE)
```



I think this defaults to the Pearson residuals not Deviance residuals (but labels don't change for shading).

For example, the Never/Seldom-No Prostate Cancer cell was slightly lower than expected, and the Moderate-No Prostate Cancer was slightly larger than expected. However, since every standardized residual in this table was between -2.0 and 2.0, we don't see much departure from what we'd expect to see if the two variables were independent.

We can also compute the likelihood ratio test statistic.

```
lrt_test_stat <- 2*sum(fish * log(fish/exp_counts))
lrt_test_stat
```

```
## [1] 3.345571
```

```
pchisq(lrt_test_stat, df = 3, lower.tail=FALSE)
```

```
## [1] 0.3413502
```

With the large sample size, the values of the chi-squared test statistic X^2 and the likelihood ratio test-statistic G^2 are similar, and they give the same conclusion.

Example 2: Nightlights and Nearsightedness

Data input as a data.frame

A survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 12 had a higher incidence of nearsightedness (myopia) later in childhood (*Sacramento Bee*, May 13, 1999, pp. A1, A18). (Taken from Example 2.2 in Utts and Heckard, 5th ed.)

Import the raw data into R:

```
eyesight <- read.csv("http://www.math.montana.edu/shancock/data/Nighlights_Nearsightedness.csv")
# Re-order ordinal factors (since R orders alphabetically)
eyesight$SleptWith = factor(eyesight$SleptWith,
                            levels = c("Darkness",
                                       "Nightlight",
                                       "Full Light" ))
```

The object `eyesight` should have appeared in your RStudio Environment. Click on it to view the data set.

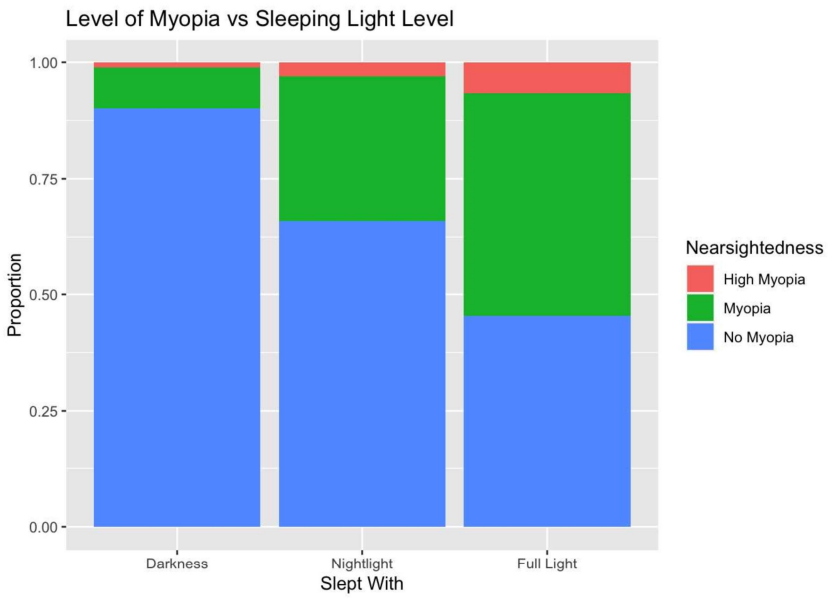
A two-way table summarizing these data can be created using `xtabs`.

```
xtabs(~ SleptWith + Nearsightedness, data = eyesight)
```

```
##             Nearsightedness
## SleptWith    High Myopia Myopia No Myopia
##   Darkness             2     15       155
##   Nightlight           7     72       153
##   Full Light           5     36        34
```

First, let's visualize the data with a bar plot. Note that the following code requires the `dplyr` and `ggplot2` packages from the `tidyverse`, which should have been loaded at the beginning of your .Rmd file.

```
eyesight %>%
    ggplot(aes(x = SleptWith, fill = Nearsightedness)) +
    geom_bar(position = position_fill()) +
    labs(
        title = "Level of Myopia vs Sleeping Light Level",
        x = "Slept With", y = "Proportion"
    )
```



We can do a chi-squared test using the raw data frame with the following syntax.

```
chisq.test(eyesight$SleptWith, eyesight$Nearsightedness)
```

```
## Warning in chisq.test(eyesight$SleptWith, eyesight$Nearsightedness): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  eyesight$SleptWith and eyesight$Nearsightedness
## X-squared = 58.374, df = 4, p-value = 6.368e-12
```

Note the warning! Why might the chi-squared approximation be incorrect in this case? Hint: Look at the counts in each cell. What other method would be more appropriate?

Fisher's Exact Test can be extended to two-way tables of larger dimensions than 2×2 using a multivariate extension of the hypergeometric distribution.

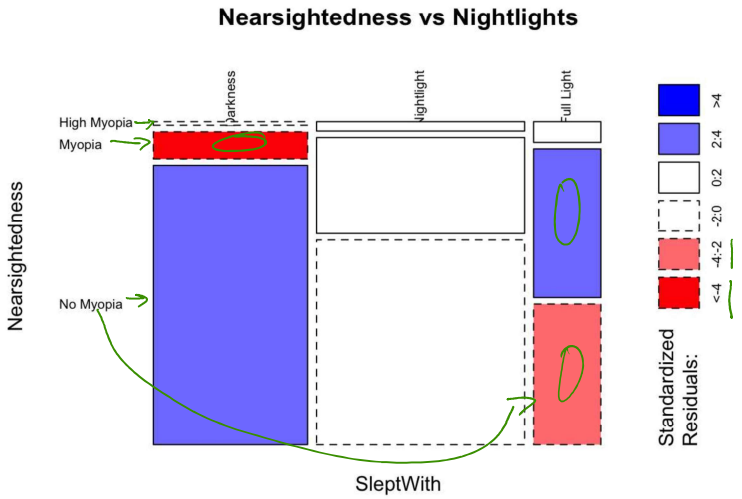
```
fisher.test(eyesight$SleptWith, eyesight$Nearsightedness)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  eyesight$SleptWith and eyesight$Nearsightedness
## p-value = 3.06e-13
## alternative hypothesis: two.sided
```

We see that there is strong evidence that the level of light slept with as a child has an association with development of myopia later in life.

Even though the assumptions are violated for the chi-squared test, the residuals still provide us with information on where the dependence between the two variables is strongest.

```
mosaicplot(~ SleptWith + Nearsightedness,
           data = eyesight, shade = TRUE, las = 2,
           main = "Nearsightedness vs Nightlights")
```



We see that there are a lot fewer subjects in the Myopia-Darkness cell than expected under the assumption of independence, and a lot more subjects in the Myopia-Full Light cell than expected. Thus, it seems that the probability of developing Myopia increases with the level of light.