Homework 7 Solution

Charlotte Wickham

2017-10-27

1. Fisher's exact test & Chi-square test

Compare the performance and p-values for Fisher's exact test and the Chi-square test by simulation.

The following compares the p-values from the two tests in the scenario:

- $n_Y = 10, n_X = 15; p_Y = 0.5, p_X = 0.5$ (Null true)

```
n_sim <- 1000

n_y <- 10
n_x <- 15
p_y <- 0.5
p_x <- 0.5

# Generate sample data of (Y, G) form.
samples <- rerun(n_sim,
  y = c(rbinom(n = n_y, size = 1, prob = p_y),
        rbinom(n = n_x, size = 1, prob = p_x)),
  g = rep(c(0, 1), c(n_y, n_x))
)

# p-values for Chi-square test
chisq_p <- map_dbl(samples,
  ~ chisq.test(table(.x$y, .x$g), correct = FALSE)$p.value)

# p-values for Fisher Exact test
fisher_p <- map_dbl(samples,
  ~ fisher.test(table(.x$y, .x$g))$p.value)

# Rejection Rates at alpha = 0.05
c(chisq_rr = mean(chisq_p < 0.05), fisher_rr = mean(fisher_p < 0.05))

# p-value distributions
ggplot() +
 geom_histogram(aes(x = chisq_p))
ggplot() +
 geom_histogram(aes(x = fisher_p))

# Comparing p-values
ggplot() +
 geom_point(aes(x = chisq_p, y = fisher_p))
```

You should compare the tests in the following additional settings:

a. $n_Y = 100$, $n_X = 150$; $p_Y = 0.5$, $p_X = = 0.5$ (Null true)

b. $n_Y = 100$, $n_X = 150$; $p_Y = 0.25$, $p_X = 0.25$ (Null true)

c. $n_Y = 10$, $n_X = 15$; $p_Y = 0.5$, $p_X = 0.25$ (Null false)

d. $n_Y = 100$, $n_X = 150$; $p_Y = 0.5$, $p_X = 0.4$ (Null false)

e. $n_Y = 100$, $n_X = 150$; $p_Y = 0.4$, $p_X = 0.5$ (Null false)

Comment on the calibration and power of the two tests, and the extent to which they make the same decisions.

Solution: Change parameters in code as neccessary. Should generally see Chi-square and Fisher's give very similar p-values for large samples, Type I error rate a little worse for small samples, or p further from 0.5.

2. Log Odds Ratio test

a) In class we saw that the odds ratio for $Y_i = 1$ given a grouping variable G_i, is the same as the odds ratio for $G_i = 1$ given the outcome Y_i, that is

$$\frac{P(Y_i = 1 \mid G_i = 1)/P(Y_i = 0 \mid G_i = 1)}{P(Y_i = 1 \mid G_i = 0)/P(Y_i = 0 \mid G_i = 0)} = \frac{P(G_i = 1 \mid Y_i = 1)/P(G_i = 0 \mid Y_i = 1)}{P(G_i = 1 \mid Y_i = 0)/P(G_i = 0 \mid Y_i = 0)}$$

Use the properties of conditional probability to derive this fact.

Let A be the event $Y_i = 1$, then A^C is the event $Y_i = 0$. Similarly, let B be the event $G_i = 1$, then B^C is the event $G_i = 0$.

Need to show

$$\frac{P(A|B)/P(A^C|B)}{P(A|B^C)/P(A^C|B^C)} = \frac{P(B|A)/P(B^C|A)}{P(B|A^C)/P(B^C|A^C)}$$

$$\begin{aligned}
\frac{P(A|B)/P(A^C|B)}{P(A|B^C)/P(A^C|B^C)} &= \frac{\frac{P(AB)/P(B)}{P(A^CB)/P(B)}}{\frac{P(AB^C)/P(B^C)}{P(A^CB^C)/P(B^C)}} && \text{definition of conditional prob.} \\
&= \frac{\frac{P(AB)}{P(A^CB)}}{\frac{P(AB^C)}{P(A^CB^C)}} && \text{cancelling common factors} \\
&= \frac{\frac{P(AB)}{P(AB^C)}}{\frac{P(A^CB)}{P(A^CB^C)}} && \text{re-arranging terms} \\
&= \frac{\frac{P(AB)/P(A)}{P(AB^C)/P(A)}}{\frac{P(A^CB)/P(A^C)}{P(A^CB^C)/P(A^C)}} && \text{multiply top by } \frac{1/P(A)}{1/P(A)} \text{ and bottom by } \frac{1/P(A^C)}{1/P(A^C)} \\
&= \frac{P(B|A)/P(B^C|A)}{P(B|A^C)/P(B^C|A^C)} && \text{definition of conditional prob.}
\end{aligned}$$

Consider the following table from class:

	no	yes
cats	6	9
dogs	6	14

where

$$G_i = \begin{cases} 0, & \text{subject i prefers cats} \\ 1, & \text{subject i prefers dogs} \end{cases} \qquad Y_i = \begin{cases} 0, & \text{subject i didn't eat breakfast} \\ 1, & \text{subject i ate breakfast} \end{cases}$$

b) Find the sample odds ratio for $Y_i = 1$ given the grouping variable G_i. Interpret your answer in the context of the data.

$$\hat{\omega} = \frac{\frac{\hat{P}(\text{prefers dogs and ate breakfast})}{\hat{P}(\text{prefers dogs and didn't eat breakfast})}}{\frac{\hat{P}(\text{prefers cats and ate breakfast})}{\hat{P}(\text{prefers cats and didn't eat breakfast})}} = \frac{\frac{d}{c}}{\frac{b}{a}} = \frac{ad}{bc} = \frac{14(6)}{9(6)} = 1.56$$

The odds of eating breakfast when you prefer dogs is 1.56 times the odds of eating breakfast when you prefer cats.

c) Show that if we consider not eating breakfast as a success (i.e. $Y_i = 1$, if the subject **didn't** eat breakast), the sample odds ratio is the reciprical of that in b).

$$\hat{\omega^*} = \frac{\frac{\hat{P}(\text{prefers dogs and didn't eat breakfast})}{\hat{P}(\text{prefers dogs and ate breakfast})}}{\frac{\hat{P}(\text{prefers cats and didn't eat breakfast})}{\hat{P}(\text{prefers cats and ate breakfast})}} = \frac{\frac{c}{d}}{\frac{a}{b}} = \frac{bc}{ad} = \frac{9(6)}{14(6)} = 0.64 = \frac{1}{1.56}$$

d) Find the log of the sample odds ratio from b), and the variance for log of the sample odds ratio from b).

Some may use odds ratio from (c), grade either as correct

$\log(\hat{\omega}) = \log(1.56) = 0.44$

$$\text{Var}(\log(\hat{\omega})) = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right) = \left(\frac{1}{6} + \frac{1}{9} + \frac{1}{6} + \frac{1}{14}\right) = 0.52$$

e) Use your results from c) to construct a 95% confidence interval for the population log odds ratio, and population odds ratio.

95% CI for $\log(\omega)$

$$\log(\hat{\omega}) \pm 1.96\sqrt{\text{Var}(\log(\hat{\omega}))} = 0.44 \pm 1.96\sqrt{0.52} = (-0.28, 1.16)$$

95% CI for ω

$$(\exp(-0.28), \exp(1.16)) = (0.76, 3.19)$$

With 95% confidence the odds of eating breakfast when you prefer dogs is between 0.76 and 3.19 times the odds of eating breakfast when you prefer cats.

f) Based on your CI in d) would you reject or fail to reject the null hypothesis $H_0 : \omega = 1$ at the 5% level.

Since one is inside the above interval for the odds ratio, we would fail to reject the null hypothesis that the odds ratio is one.

3. Data Analysis

Consider the `brfss` data from previous homeworks. The variables `exercising` and `dieting` are the answers to the questions:

- `exercising`: Are you using physical activity or exercise to lose weight or keep from gaining weight?
- `dieting`: Are you eating either fewer calories or less fat to lose weight or keep from gaining weight?

a) Create the 2×2 contingency table with the response to `exercising` in the columns and response to `dieting` in the rows.

	No Exercise	Exercise
No Diet	263	398
Diet	468	1167

b) Which margins in the table are fixed by the study design?

This study is multinomial in design, the survey selects some number of people to participate and the cross-classifies them based on their answers to the diet and exercise questions, suggesting only a fixed total sample size, N.

However, you may notice there are a number of missing values! The number of respondents with non-missing responses for both these questions is unknown at the study outset, so technically no margins are fixed.

c) Use the data to estimate (make sure you also include sentences that present each estimate in context):

- the probability a US resident is exercising to lose weight

$$\hat{P}(\text{exercising}) = \frac{R_2}{N} = \frac{1565}{2296} = 0.68$$

It is estimated the probability a US resident is exercising to lose weight is 0.68.

- the probability a US resident is exercising to lose weight, given they are dieting to lose weight

$$\hat{P}(\text{exercising}|\text{dieting}) = \frac{d}{C_2} = \frac{1167}{1635} = 0.71$$

It is estimated the probability a US resident is exercising to lose weight, given they are dieting to lose weight is 0.71.

- the difference in the probability a US resident is exercising to lose weight between those that are also dieting and those that are not

$$\hat{P}(\text{exercising}|\text{dieting}) - \hat{P}(\text{exercising}|\text{not dieting}) = \frac{d}{C_2} - \frac{b}{C_1} = \frac{1167}{1635} - \frac{398}{661} = 0.11$$

It is estimated, that a US resident that is dieting is 11 *percentage points* more likely to be exercising than a US resident that is not dieting.

- the odds of a US resident exercising to lose weight, given they are dieting to lose weight

$$\frac{\hat{P}(\text{exercising}|\text{dieting})}{\hat{P}(\text{not exercising}|\text{dieting})} = \frac{\frac{d}{C_2}}{\frac{c}{C_2}} = \frac{d}{c} = \frac{1167}{468} = 2.49$$

It is estimated, the odds a US resident is exercising given they are dieting to lose weight is 2.49.

- the odds ratio of exercising to lose weight, between dieting and not dieting

$$\hat{\omega} = \frac{ad}{bc} = \frac{1167 \times 263}{398 \times 468} = 1.65$$

It is estimated that the odds a US resident is exercising for those that are dieting is 1.65 times the odds of exercising for those that aren't dieting.

d) Find a 95% confidence interval for the difference in the probability a US resident is exercising to lose weight between those that are also dieting and those that are not.

Use `prop.test()`

```
# reorder to put exercising in first column
brfss <- brfss %>%
  mutate(exercising = factor(exercising,
    levels = c("yes", "no")))
(prop_test <- prop.test(xtabs(~ dieting + exercising,
  data = brfss), correct = FALSE))
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  xtabs(~dieting + exercising, data = brfss)
## X-squared = 27.035, df = 1, p-value = 1.998e-07
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.15491370 -0.06837323
## sample estimates:
##    prop 1    prop 2
## 0.6021180 0.7137615
```

Or calculate directly

$$\hat{p}_Y - \hat{p}_X \pm 1.96\sqrt{\frac{\hat{p}_Y(1-\hat{p}_Y)}{n} + \frac{\hat{p}_X(1-\hat{p}_X)}{m}}$$

It is estimated, that a US resident that is dieting is between 6.8 *percentage points* and 15.5 *percentage points* more likely to be exercising than a US resident that is not dieting.

e) Is there an association between exercising to lose weight and dieting to lose weight? Conduct a Chi-square test and Fisher's Exact test. Would you expect the two tests to reach the same conclusion in this setting? Write a statistical summary.

With such a large sample, based on results from Q1 the two procedures should give similar results.

```
chisq.test(xtabs(~ dieting + exercising,
  data = brfss), correct = TRUE)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  xtabs(~dieting + exercising, data = brfss)
## X-squared = 26.523, df = 1, p-value = 2.604e-07
```

```
fisher.test(xtabs(~ dieting + exercising,
  data = brfss))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  xtabs(~dieting + exercising, data = brfss)
## p-value = 3.262e-07
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5002548 0.7369219
## sample estimates:
```

```
## odds ratio
##  0.6069934
```

And they do:

There is convincing evidence of an association between dieting and exercising to lose weight for US residents (Chi-square test of association, p-value < 0.0001).