Math 150 - Methods in Biostatistics - Homework 3

Ethan Ashby

Due: Friday, February 19, 2021

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=4, fig.width=6.5,
                      fig.align = "center")
library(tidyverse)
library(broom)
library(praise)
library(knitr)
library(kableExtra)
```

Assignment Summary (Goals)

- Understanding null hypotheses with respect to odds and proportions
- Testing via z-stat, Fisher, Chi-sq (note: if you did all for all scenarios, you'd like end up with the same conclusion each time!)
- Making conclusions about different study types
- (Decided against the Chi Square test, you are not responsible for it.)

Q1. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Annika got the part in the play! Good for her! Lian went roller skating twice today (and she can do a single-leg squat)! Annie is figuring her life out after graduation! Lian helped me with plot on 4d!

Q2. Chp 6, A23 Show that the null hypothesis $H_0 : p_1 = p_2$ is mathematically equivalent to the null hypothesis $H_0 : \theta_1/\theta_2 = 1$ where p represents the proportion successful and θ represents the odds of success for any two groups (labeled 1 and 2).

Let's begin by constructing a nice little 2x2 table to illustrate the equivalency:

Groups	Success	Failure
1	a	b
2	c	d

The null hypothesis $H_0 : p_1 = p_2$ means that the probability of success in group 1 is equal to the probability of success in group 2. Using our 2x2 table, we note that $H_0 : p_1 = p_2 \implies H_0 : \frac{a}{a+b} = \frac{c}{c+d}$. Using some algebraic manipulation:

$$p_1 = p_2 \implies \frac{a}{a+b} = \frac{c}{c+d} \implies \frac{a+b}{a} = \frac{c+d}{c}$$

$$\frac{a+b}{a} = \frac{c+d}{c} \implies 1 + \frac{b}{a} = 1 + \frac{d}{c}$$

$$1 + \frac{b}{a} = 1 + \frac{d}{c} \Longrightarrow \frac{b}{a} = \frac{d}{c}$$
$$\frac{b}{a} = \frac{d}{c} \Longrightarrow \frac{a}{b} = \frac{c}{d} \Longrightarrow \theta_1/\theta_2 = 1$$

Now in the words of the Cupid Shuffle guy, "REVERSE REVERSE" (i.e. we're going to prove the reverse):

$$\theta_1/\theta_2 = 1 \Longrightarrow \frac{a}{b} = \frac{c}{d} \Longrightarrow \frac{b}{a} = \frac{d}{c}$$
$$\Longrightarrow 1 + \frac{b}{a} = 1 + \frac{d}{c} \Longrightarrow \frac{a+b}{a} = \frac{c+d}{c}$$
$$\Longrightarrow p_1 = p_2$$

We arrive at our desired result by showing that the equality of proportions of success implies that the odds of success in the two groups are equal and vice versa.

Q3. Chp 6, E7 Cancer Cells: Testing for Homogeneity of Odds Use the data from Table 6.1 and define a benign cell as a success. Conduct a hypothesis test for the homogeneity of odds.

(a) State the null and alternative hypotheses.
(b) Calculate the odds ratio and the test statistic (the Z statistic!). See pgs 191-192 in your book.
(c) Provide the p-value and state your conclusions within the context of the study.

Let's get the data read in first

Shape	Malignant	Benign
Round	7	9
Concave	17	4

(a) The null hypothesis is that the odds of malignancy in the Concave group are equal to the odds of malignancy in the Round group. In other words: $H_0 : \frac{\theta_{\text{concave}}}{\theta_{\text{round}}} = 1$. The alternative hypothesis is that the odds of malignancy are higher in the Concave group than the Round group. In other words: $H_0 : \frac{\theta_{\text{concave}}}{\theta_{\text{round}}} > 1$.

(b) The odds ratio in a 2x2 table is calculated using the formula $\frac{a/b}{c/d} = \frac{17/4}{7/9} = 5.464$.

It is known that the natural log of the odds is approximately normal with standard deviation

$$\text{SD}(\ln(\hat{OR})) = \sqrt{\frac{1}{n_{\text{concave}}\,\hat{p}\,(1-\hat{p})} + \frac{1}{n_{\text{round}}\,\hat{p}\,(1-\hat{p})}} = \sqrt{\frac{1}{21\,(24/37)\,(13/37)} + \frac{1}{16\,(24/37)\,(13/37)}} = 0.695$$

The corresponding Z-statistic is $Z = \frac{\ln(5.464)-0}{\text{SD}(\ln(\hat{OR}))} = 2.44$

(c) Using the function `pnorm` to calculate the probability associated with our Z-statistic, I obtain a p-value of 0.0072823. This means that we can reject the null hypothesis, and conclude that the odds of malignancy in the Concave group are significantly higher than the odds of malignancy in the Round group.

Q4. Chp 6, E12 The Pill Scare: understanding relative risk reduction In October 1995, the United Kingdom Committee on Safety of Medicines (CSM) issued a warning to 190,000 general practitioners, pharmacists, and directors of public health about oral contraceptive pills containing gestodene or desogestrel. The warning, based on three unpublished epidemiological research studies, stated > "It is well known that the pill may rarely produce thrombosis (blood clots) involving veins of the legs. New evidence has become available indicating that the chance of thrombosis occurring in a vein increases about two-fold for some types of pills compared to others."

Table 6.15 provides data from one of the studies.

Pill	VT	No VT
Third Gen	127	249
Second Gen	132	402

Since the occurrence of venous thrombosis is very rare (1 in 7000 for people using the second generation pill), 259 subjects were selected who had thrombosis and 651 similar subjects (from hospitals and community) who did not have thrombosis. Then these subjects were classified by the type of contraceptive they used.

(a) Was either the explanatory (row) or the response (column) variable fixed before the study was conducted?

The columns were fixed in the study, since people were selected on the basis of whether they have Vein Thrombosis.

(b) Is this an example of an experiment or an observational study?

This is an observational study all the way! We are NOT allocating the treatment (which is a requirement for the experiment)!

(c) Is this a cross-classification, cohort, or case-control study?

Case-control, because the scientists selected people on the basis of **disease status** (outcome), rather than exposure.

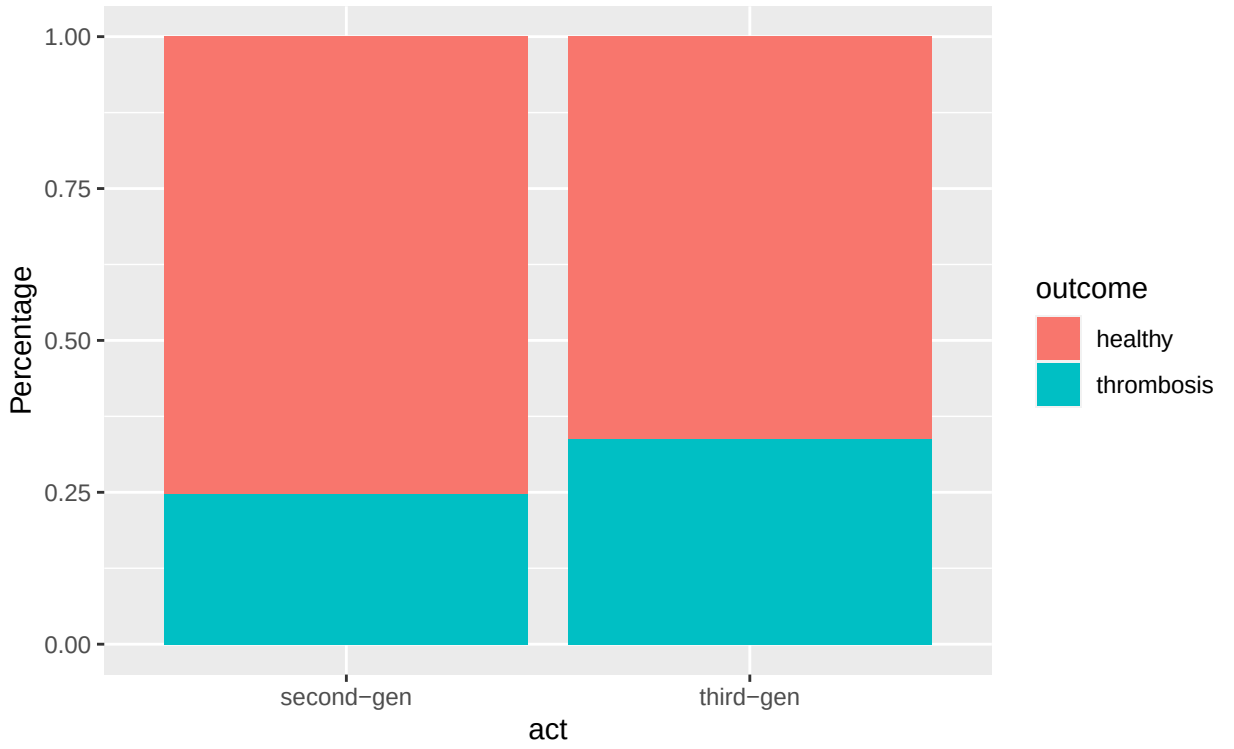
(d) Create a segmented bar chart for the data.

Thank you to Lian for helping me w/ the code!

```
library(reshape2)

thrombosis <- data.frame(act = c(rep("third-gen", 376), rep("second-gen", 534)),
                         outcome = c(rep("thrombosis", 127), rep("healthy", 249),
                                     rep("thrombosis", 132), rep("healthy", 402)))

thrombosis %>%
  ggplot(aes(x = act)) +
  geom_bar(aes(fill = outcome), position = "fill") +
  ylab("Percentage")
```



(e) Use a two-sided hypothesis and Fisher's exact test to determine if the type of contraceptive impacts the likelihood of thrombosis.

```
fisher.test(pill_df[1:2, 2:3])
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  pill_df[1:2, 2:3]
## p-value = 0.003577
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.149033 2.098289
## sample estimates:
## odds ratio
##   1.552484
```

Thus, the odds of ratio is significantly different from 1, so we conclude that the odds of getting vein thrombosis is significantly different between those who recieve third and second gen drugs.

Do you expect the researchers took care to collect a simple random sample of subjects?

I think this wasn't a random sample... indeed, they were sampling people according to disease status, which is inherently non-random. Unless there is a database of all people with/without vein thrombosis that the authors randomly sampled from, their sampling scheme was non-random.

What conclusions can be drawn?
This limits the generalizability of their study. While they observed a significant difference in odds in their study, they are unable to generalize this to the population due to the non-random sample.

The warning contained no numerical information other than the fact that the chance of blood clots was likely to double when birth control pills contained gestodene or desogestrel. This warning was widely publicized throughout the press, and evidence suggests that, as a result of this warning, many women ceased contraception altogether. Evidence shows a strong association between the warning and an increase in the number of unintended pregnancies and abortions (especially in women younger than 20 years old). This resulted in an estimated increase in cost of £ 21 million for maternity care and £ 4 to £ 6 million for abortion provision.

(f) Remember that the actual occurrence of venous thrombosis is only 1 in 7000.
If third generation pills double the chances of venous thrombosis, the likelihood of occurrence is still only 2 in 7000. Explain the difference between absolute risk reduction and relative risk reduction in this study.

Absolute risk reduction will be really small ($\frac{2}{7000} - \frac{1}{7000} = \frac{1}{7000}$), indicating that taking the second/third gen pills have small effects on your overall risk of venous thrombosis. The relative risk reduction will be larger ($\frac{1/7000}{2/7000} = 0.50$), meaning that taking the second generation pills reduces your risk of getting venous thrombosis by 50%! So... your overall risk of getting thrombosis is low no matter which pill you take, though the gen 3 pills double your (already small) chance of getting thrombosis.

(g) Death from venous thrombosis related to third generation pills is estimated to be 1 in 11 million, much lower than the probability of death resulting from pregnancy. In 2005, the lifetime risk of maternal death in developed countries was 1 in 7300. The CSM warning did suggest that patients see a doctor before altering their contraceptives; however, it appears that many women simply stopped taking any contraceptives. Write a brief statement (just 1-2 sentences!) to the press, general practitioners, pharmacists, and directors of public health about this study.

If you go off contraceptives and get pregnant, you are more than 1500 more likely to die in a maternal death event than from venous thrombosis caused by these third gen pills. You are more likely to be struck by lightning 21 times than to die by venous thrombosis related to third generation pills!

```
praise()
```

```
## [1] "You are extraordinary!"
```