Homework Chapter 11

Emily Maloney

March 24, 2019

```
library(tidyverse)
library(rethinking)
library(brms)
library(knitr)
library(bayesplot)
```

Chapter 11

Easy Problems

11E1

An ordered categorical variable is one in which the points on the scale correspond to levels while an unordered categorical variable's elements cannot be ranked against each other. An example of an ordered categorical variable is grades (A, B, C, D, F), and an example of an unordered categorical variable is sex (Male, Female, Intersex).

11E2

An ordered logistic regression employs a cumulative logit link, which means that it uses the log-cumulative-odds for each of the ordered outcomes - in other words, the log odds of that value or anything below it on the scale. An unordered logistic regression employs a logit link, which is just the log-odds for the specific outcome.

11E3

When count data are zero-inflated, using a model that ignores zero-inflation will tend to induce Type I error.

11E4

An example of a process that might produce over-dispersed counts would be the number of times someone talks in class, because everyone has a different threshold for speaking. An example of a process that might produce under-dispersed counts would be the number of children women have in Scandanavia, because I think the variance would be lower than the expected value.

Medium Problems

11M1

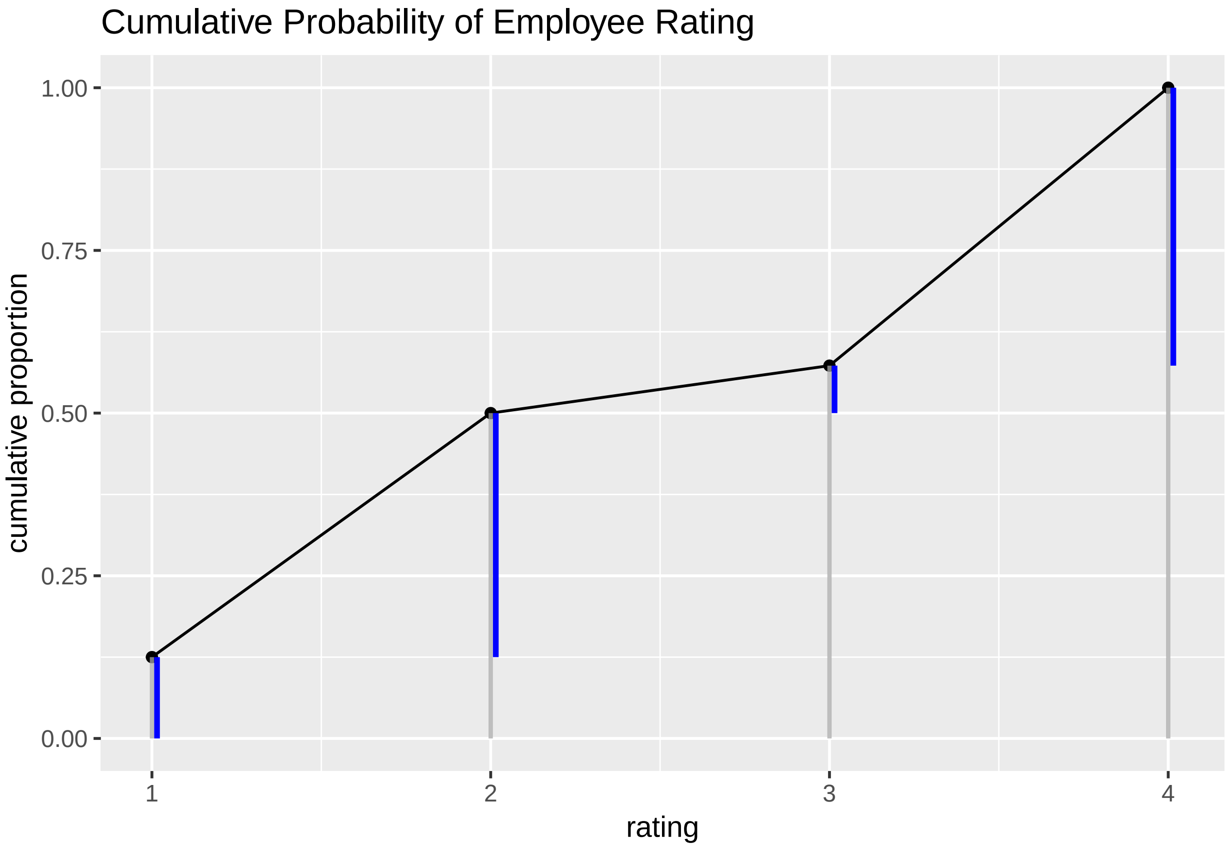
```
#make dataset with rating and frequency
d <- tibble(rating = seq(1, 4, by = 1),
            freq = c(12, 36, 7, 41))

#calculate log cum odds and the previous proportion
d <- d %>% mutate(total = sum(freq),
                  cum_sum = cumsum(freq),
                  cum_prop = cum_sum/total,
                  cum_odds = cum_prop/(1-cum_prop),
                  log_cum_odds = log(cum_odds),
                  prev_prop = lag(cum_prop, default = 0))

#make table of log_cum_odds
table <- d %>% select(rating, log_cum_odds)
kable(table)
```

rating	log_cum_odds
1	-1.9459101
2	0.0000000
3	0.2937611
4	Inf

The log-cumulative-odds for getting a rating 1 is -1.95. For getting a rating of 2, the log-cumulative-odds is 0, and for getting a rating of 3, the log-cumulative-odds is 0.29. The log-cumulative-odds for the rating 4 is infinite because the cumulative probability of the highest value on the scale is 1, so the odds are 1/(0), which is indefinite.

11M2

```
#making plot
ggplot(data = d, mapping = aes(x = rating, y = cum_prop)) +
  geom_line() +
  geom_point() +
  geom_linerange(mapping = aes(ymin = 0,
                               ymax = cum_prop),
                               alpha = 3/4,
                               color = "dark gray",
                               size = 0.75) +
  geom_linerange(mapping = aes(x = rating + .015,
                               ymin = prev_prop,
                               ymax = cum_prop),
                               color = "blue",
                               alpha = 1,
                               size = 1) +
  labs(title = "Cumulative Probability of Employee Rating",
       x = "rating", y = "cumulative proportion")
```



This plot shows the growing cumulative probability for each rating from 1 to 4 in the gray vertical lines. The blue lines are the discrete probability for each rating and represent the likelihood of each rating. As the plot shows, the ratings 2 and 4 have a much higher likelihood than the ratings 1 or 3.

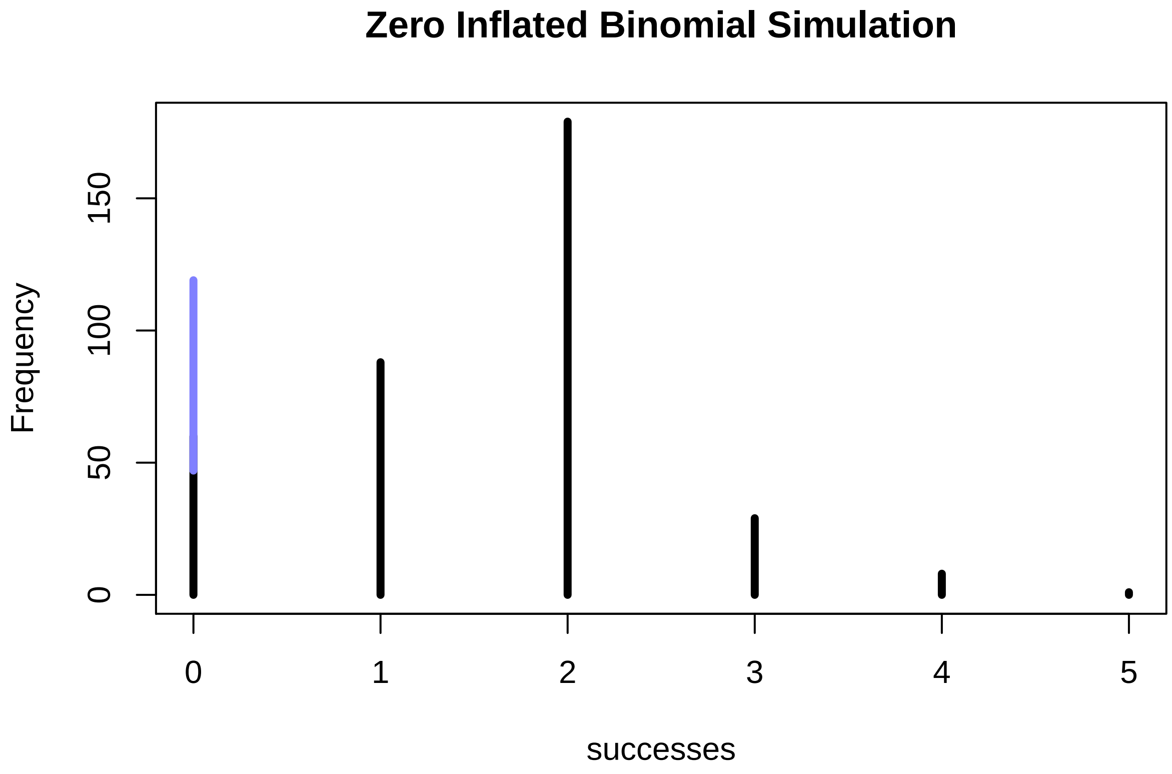
11M3

With probability d that secondary process produces 0, n equal to the number of trials, and p the probability of success, the likelihood of a zero result is:
$d + (1 - d)(1 - p)^n$
and the likelihood of a non-zero result is:
$(1 - d)(p^y)\binom{n}{y}(1 - p)^{(n-y)}$.

To construct this distribution, I simulated each of these processes and then plotted them together, to show how the secondary process inflates the number of zeros in the binomial distribution.

```r
#defining parameters
d <- 0.2
n <- 365
p <- 0.2

#simulate days w/no production
zero_2 <- rbinom(n, 1, d)

#simulate number completed
y <- (1-d)*rbinom(n, 10, p)
y <- as.data.frame(y)
```

```r
simplehist(y, xlab = "successes", lwd = 4)
zeros_second <- sum(zero_2)
zeros_first <- sum(y == 0 & zero_2 == 0)
zeros_total <- zeros_second + zeros_first
lines(c(0,0), c(zeros_first, zeros_total), lwd = 4, col = rangi2)
title(main = "Zero Inflated Binomial Simulation")
```



In this histogram, the black lines show the typical binomial distribution for the parameters n equal to 365 and p equal to 0.2, given a trial size of 10. The blue line at 0 represents the number of days in which there were no successes from a secondary process unrelated to the binomial. Including this process inflates the number of expected zeros to even be slightly more than the number of expected ones.