MA256 Lesson 14 - Multiple Proportions (8.1-8.3)

Review: Comparing two proportions.

Hypotheses (in symbols):

H_0 :

H_a :

Strength of Evidence:

Confidence Interval:

Validity conditions:

Multiple-proportions: Similar to our methods used when considering two-proportions, but:

-
-

Hypothesis test:

H_0 :

H_0 :

Validity conditions:

Chi-square (not "chai-squared"):

Binary response variable:

General (more than 2 categories for response variable):

Calculate the p-value for Chi-square distribution:

Conserving Hotel Towels

STEP 1. Ask a research question.

If you have stayed at a hotel recently, you may have noticed how guests are increasingly encouraged to practice conservation habits by not having their towels washed every day-instead hanging them back on the rack to be reused. Are some "encouragement strategies" more effective than others?

STEP 2. Design a study and collect data.

A recent study (Goldstein et al., 2008) conducted a randomized experiment to investigate how different phrasings on signs placed on bathroom towel racks impacted guests' towel reuse behavior. In particular, the researchers were interested in evaluating how messages that communicated different types of social norms impacted towel reuse. One week prior to a guest staying in the room, rooms at a particular hotel were randomly assigned to receive one of the following five messages on a sign hung on the towel bar in the room:

Message 1: HELP SAVE THE ENVIRONMENT. You can show your respect for nature and help save the environment by reusing your towels during your stay. (No social norm)
Message 2: JOIN YOUR FELLOW GUESTS IN HELPING TO SAVE THE ENVIRONMENT. In a study conducted in fall 2003, 75% of the guests participated in our new resource savings program by using their towels more than once. You can join your fellow guests in this program to help save the environment by reusing your towels during your stay. (Guest identity norm)
Message 3: JOIN YOUR FELLOW GUESTS IN HELPING TO SAVE THE ENVIRONMENT. In a study conducted in fall 2003, 75% of the guests who stayed in this room (#xxx) participated in our new resource savings program by using their towels more than once. You can join your fellow guests in this program to help save the environment by reusing your towels during your stay. (Same room norm)
Message 4: JOIN YOUR FELLOW CITIZENS IN HELPING TO SAVE THE ENVIRONMENT. In a study conducted in fall 2003, 75% of the guests participated in our new resource savings program by using their towels more than once. You can join your fellow citizens in this program to help save the environment by reusing your towels during your stay. (Citizens norm)
Message 5: JOIN THE MEN AND WOMEN WHO ARE HELPING TO SAVE THE ENVIRONMENT. In a study conducted in fall 2003, 76% of the women and 74% of the men participated in our new resource savings program by using their towels more than once. You can join the other men and women in this program to help save the environment by reusing your towels during your stay." (Gender identity norm)

Data were collected on 1,595 instances of potential towel reuse. For each of the 1,595 instances, room attendants recorded whether the hotel guest reused their towels.

1. In your own words, state the research question of interest in this study.

2. Identify the explanatory and response variables in this study, along with their types (categorical or quantitative). For the explanatory variable, indicate how many categories it involves if necessary.

Explanatory:

Response:

3. Did the study design make use of any pairing of subjects or are the subjects who receive the different signs independent of each other?

4. State the appropriate null and alternative hypotheses for testing the researchers' conjecture using both words and symbols.

H_0 :

H_a :

STEP 3. Explore the data.

5. Read in the `Towels.csv` file. Use the `count()`, `pivot_wider()`, and `adorn_totals()` functions to create a table showing the numbers of instances of each message type and the reuse status. Also, create a stacked barchart showing the proportions of each message type.

```
# library(tidyverse)
# library(janitor)
# towels <- read_csv("https://raw.githubusercontent.com/jkstarling/MA256/main/data/Towels.csv")
#
# towel.tab <- towels %>%
#    count(Message,Use_case) %>%
#    pivot_wider(names_from = Message, values_from = n) %>%
#    adorn_totals(c("row", "col"))
# towel.tab
# # Plot
# towels %>% ggplot(aes(x=Message, fill= Use_case)) +
#    geom_bar(position= position_fill())
```

STEP 4. Draw Inferences beyond the data.

6. Apply the 3S strategy:

    1) Statistic: Report and use the mean group difference for this study. (see §8.1, p.463)

```
# p.c <-   145 / 334
# p.ge <- XXX / XXX
# p.gu <- XXX / XXX
# p.n <-   XXX / XXX
# p.r <-   XXX / XXX
#
# # make vector with the absolute value of the differences
# gdiffs <- c(abs(p.c - p.ge), abs(p.c - p.gu), abs(p.c - p.n), abs(p.c - p.r),
#             ....)
# # calculate the mean of the absolute value of the diffs
# mean(gdiffs)
#
# # OR
# reuse <- towel.tab[2,2:6]
# tot <- towel.tab[3,2:6]
# p.all <- unlist(unname(reuse/tot))
#
# mysum <- c()
# for (ii in 1:4){
#    i= ii + 1
#    mysum <- append(mysum, abs(p.all[ii] - p.all[i:5]))
# }
# mean(mysum)
```

    2) Simulate: How would you simulate the towel use under the null hypothesis? Conduct 1000 replications and
       report a p-value.

```
# # define function to calculate mean group differences
# get.mean.group.diffs <- function(m.vec, u.vec){
#    mtab <- table(u.vec, m.vec)
#    p.all <- mtab['reuse',] / colSums(mtab)
#    mysum <- c()
#    for (ii in 1:4){
#       i= ii + 1
#       mysum <- append(mysum, abs(p.all[ii] - p.all[i:5]))
#    }
#    mean(mysum)
```

```
#    }
#
# set.seed(256)
#
# M <- 1000
# gdif.mean <- 0.0554887
#
# # duplicate vectors
# mess <- XXXX
# usecase <- XXXX
#
# # create df to record values
# RES <- data.frame(res = rep(NA, M))
#
# for(rep in 1:M){
#    use.shuf <- sample(XXXXX)
#    RES$res[rep] <- get.mean.group.diffs(XXX, XXX)
# }
#
# # plot results
# RES %>% ggplot(aes(x=res)) +
#    geom_histogram() +
#    geom_vline(xintercept = gdif.mean, color="red")
#
# # estimate p-value
# sum(RES$res >= gdif.mean) / M
```

3) Strength of evidence: Based on the p-value, evaluate the strength of evidence provided by the data against the null hypothesis. Summarize your conclusion in the context of this study.

Another statistic: Chi-square statistic

7. Determine the overall proportion, $\hat{p}$, of rooms that reused towels.

```
# phat <- XXXX / XXXX
# phat
```

8. We will calculate the standard error of a group proportion by computing $\sqrt{\hat{p}(1-\hat{p})/n_i}$ where is the overall proportion found in #7 and n_i is the sample size for the ith group. Find the standard error of the group proportion for each of the five groups.

```
# n.c <-   XXXX
# n.ge <- XXXX
# n.gu <- XXXX
# n.n <-   XXXX
# n.r <-   XXXX
#
# se.c <- sqrt(phat * (1-phat) / n.c)
# se.ge <- sXXXX
# se.gu <- XXXX
# se.n <- XXXX
# se.r <- XXXX
#
```

4

```
# # OR
# se.vec <- sqrt(phat * (1-phat) / tot)
```

9. Create a standardized z-statistic for each conditional proportion compared to this overall proportion. To do this take each of the five group proportions, subtract the overall proportion (#7), and then divide the result by the standard deviation you found for that group in #8.

```
# phat <- XXXX / XXXX
# zs <- c((p.c - phat) / se.c,
#           (p.ge - phat) / se.ge,
#           (p.gu - phat) / se.gu,
#           (p.n - phat) / se.n,
#           (p.r - phat) / se.r)
# zs
# # OR
# zs <- (c(p.c, p.ge, p.gu, p.n, p.r) - phat) / c(se.c, se.ge, se.gu, se.n, se.r)
# # OR
# zs <- unlist((p.all - phat) / se.vec)
```

10. Which of the five standardized statistics are larger than **2** or smaller than **-2**, suggesting strong evidence that the group proportion is significantly different than the overall proportion?

11. Square each of these values and sum them to calculate the chi-square statistic.

```
# chsq <- sum(xxx)
# chsq
```

12. What kinds of values (large, small, positive, negative) will the chi-square statistic have if the null hypothesis is true? What kinds of values will provide evidence against the null hypothesis?

13. Calculate the p-value.

```
# k <- XXXX
# XXXXX
```

14. How does the p-value based on the χ^2 statistic compare to the one based on the Mean Group Diff statistic? Are they similar? Is your conclusion about strength of evidence provided by the data against the null hypothesis similar?

15. Do we meet the validity conditions for a chi-squared test with these data? Justify your answer.

16. Calculate the χ^2 statistic using the general form of the chi-square statistic (e.g. the `Observed` and `Expected` method). (Hint: See p.482 for help.)

```
# towel.tab
# p.not <- XXX / XXX
# p.reuse <- XXX / XXX
# observed.data <- data.frame(towel.tab[XXX:XXX, XXX:XXX])
# observed.data
# expected.data <- observed.data
```

```
# expected.data[1,1] <- XXX
# expected.data[1,2] <- XXX
# expected.data[1,3] <- XXX
# expected.data[1,4] <- XXX
# expected.data[1,5] <- XXX
# expected.data[2,1] <- XXX
# expected.data[2,2] <- XXX
# expected.data[2,3] <- XXX
# expected.data[2,4] <- XXX
# expected.data[2,5] <- XXX
# expected.data
#
# mychisq <- sum((observed.data - expected.data)^2 / expected.data)
#
# # OR
# chisq.test(observed.data)
```

17. Calculate the 95% CI(s) for difference in proportions. a) How many intervals are produced? b) Which conditions are significantly different from each other? How are you deciding? c)For one of the intervals you identified in part (b), write a one-sentence interpretation of the interval, being very clear what is supposed to be captured inside the interval and which condition has a more positive effect on towel use and by how much.

```
# ci.2props <- function(p1, p2, n1, n2){
#    se <- sqrt((p1*(1-p1)/n1) + (p2*(1-p2)/n2))
#    diff <- p1 - p2
#    c(diff - 2 * se, diff + 2 * se)
# }
#
# ci.2props(p.c, p.ge, n.c, n.ge)
# .... # do it 9 x more times...
```

STEP 5: Summarize conculsions.

17. Is this an observational study or a randomized experiment?

18. Explain how the conclusions of this study are limited by conducting the study at a single hotel.

STEP 6. Look back and ahead.

19. Summarize your conclusion for your analysis of the data from this study. Include a recommendation for the hotel manager about whether it matters what message is used, which message should be used, and how large of an impact the best message would be expected to have. Think about the study and its design. Were there any limitations you feel need to be addressed? Did the design match the research question? Can you suggest a better sampling method of hotel guests-a better way to collect data on towel use? Share some ideas for a logical next study.

Nearsightedness and NighLights REDUX

Recall Example 4.1, which described a study investigating whether there is a relationship between use of night lights in a child's room before age 2 and the child's eyesight condition a few years later. In Chapter 4, we presented a two-way table of counts from the study examining whether the child was nearsighted as the response variable:

	Darkness	Night light	Room light
Nearsighted	18	78	41
Not nearsighted	154	154	34
Total	172	232	75

The conditional proportions of nearsightedness in the three lighting groups are 0.105, 0.336, and 0.547, respectively, which suggest that the more light used in the child's room, the more likely the child is to become nearsighted. Now we know how to calculate a standardized statistic and estimate a p-value to measure the strength of evidence provided by these sample data for the conjecture that eyesight is associated with the type of lighting used in the child's room.

1. **State null and alternative hypotheses in terms of population proportions.**

2. **State null and alternative hypotheses in terms of association between variables.**

3. **Are the validity conditions necessary for a chi-square test met for these data? Explain.**

4. **To find out how many children we expect to be in the (darkness, nearsighted) cell of the table, answer (a)–(d):**

   a) Find the overall proportion of children who were nearsighted in the sample.

```
# p.near <- XXX / XXX
```

   b) How many children in the sample slept in darkness?

```
# slept.in.darkness <- XXX
```

   c) Multiply these two numbers together to find the "expected" number of children who would be nearsighted in the "slept in darkness" group if the null hypothesis were true (the probability of nearsightedness is the same in each condition).

```
# p.near * slept.in.darkness
```

   d) How different is this expected cell count from what was observed for that cell? What does this suggest about the strength of evidence against the null hypothesis?

5. **A similar calculation is performed to find the expected cell count for each of the six cells of the table. The chi-square statistic is obtained by squaring the differences in the observed and expected counts, dividing by the expected count (a form of standardizing), and then summing the six values. This can be thought of as (O - E)2/E, where O = observed count and E = expected count. Calculate the χ^2 statistic using this method.**

```
# nv1 <- data.frame(vision = c(rep("near", 3), rep("notnear", 3)),
#               light = c("dark", "nlight", "rlight", "dark", "nlight", "rlight"),
#               count = c(18,78,41,154,154,35))
```

6. The original research study actually classified the response variable into three categories of eyesight, as shown in the following table:

	Darkness	Night light	Room light
Nearsighted	114	115	22
Farsighted	40	39	12

7. Create a segmented bar graph to summarize this table. Discuss what the graph reveals, including what this graph tells you that we did not learn from the initial table showing only nearsighted vs. not nearsighted.

```
# nv2 <- data.frame(vision = c(rep("near", 3), rep("normal", 3), rep("far", 3)),
#               light = c("dark", "nlight", "rlight", "dark", "nlight", "rlight", "dark", "nlight", "r
#               counts = c(18,78,41,114,115,22,40,39,12) )
#
# nv2 %>% ggplot(aes(x = XXX, y = XXX, fill = factor(XXX))) + geom_bar(position="fill", stat="identity")
```

8. With this new table, do you need to change your null and alternative hypotheses from #2?

9. Are the validity conditions for the chi-square test met for this table?

10. Calculate the χ^2 statistic. Hint: See p.482 for help.