Lec 07: Categorical Data Analysis

MATH 456 - Spring 2016

Navbar: [Home] [Schedule] [Data] [Week 10 Overview] [HW Info] [Google Group]

Assigned Reading & additional references

- OpenIntro Lab on categorical data [HTML]
- OpenIntro Statistics Free PDF Textbook Chapter 6 (6.1-6.4)
- Afifi Chapter 12 (Logistic Regression)

Additional references

- Data vis tutorial http://norcalbiostat.github.io/R-Bootcamp/labs/Data_Visualization_Tutorial_Full.html
- Cookbook for R http://www.cookbook-r.com/Graphs/Bar_and_line_graphs_(ggplot2)/

Introduction

Up until now we have been analyzing continuous outcomes. We will now turn our focus to methods to analyze categorical data. This will allow us to answer questions like the following:

- What proportion of the American public approves of the job the Supreme Court is doing?
- The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

The methods you learned in previous classes will be useful in these settings. For example, sample proportions are well characterized by a nearly normal distribution when certain conditions are satisfied, making it possible to employ the usual CI and HT tools. In other instances, such as those with contingency tables or when sample size conditions are not met, we will use a different distribution, though the core ideas remain the same.

Be sure to referece the categorical data sections of the Data Visualization Tutorial or the Cookbook for R for more information and guidance on how to plot categorical data correctly.

Spam Data

This set of lecture notes uses the `spam` data set.

```
email <- read.delim("C:/Dropbox/CSUC/Data/email.txt", header=TRUE, sep="\t")
```

Two categorical variables of current interest are

- `spam` (0/1 binary indicator if a an email is flagged as spam). Converted into a Ham/Spam factor variable.
- `number` categorical variable describing the size of the numbers contained in the email.
    - `none`: No numbers
    - `small`: Only values under 1 million
    - `big`: A value of 1 million or more

Review: Inference on a single proportion

This section is considered review and will not be covered directly in class. If you have never analyzed proportions in R or it has been a while since you took MATH 315 it is advised that you do not skip this section.

Inference for a single proportion

- Open Intro: Chapter 6.1
- Complete the OpenIntro Lab on categorical data [HTML]

You must have the `openintro` package installed & loaded to have access to the custom `inference()` function used in the lab.

Example: Examining the frequency of number sizes in emails

One-way Tables

A table for a single variable is called a *frequency table.* The values displayed represent the number of records in the data set that fall into that particular category.

```
table(email$number)
```

```
##
##   big  none small
##   545   549  2827
```

If we replaced the counts with percentages or proportions, the table would be called a *relative frequency table.*

```
prop.table(table(email$number))
```

```
##
##       big      none     small
## 0.1389952 0.1400153 0.7209895
```

We make this output more human readable as percentages by rounding the results and multplying by 100.

```
round(prop.table(table(email$number))*100,2)
```

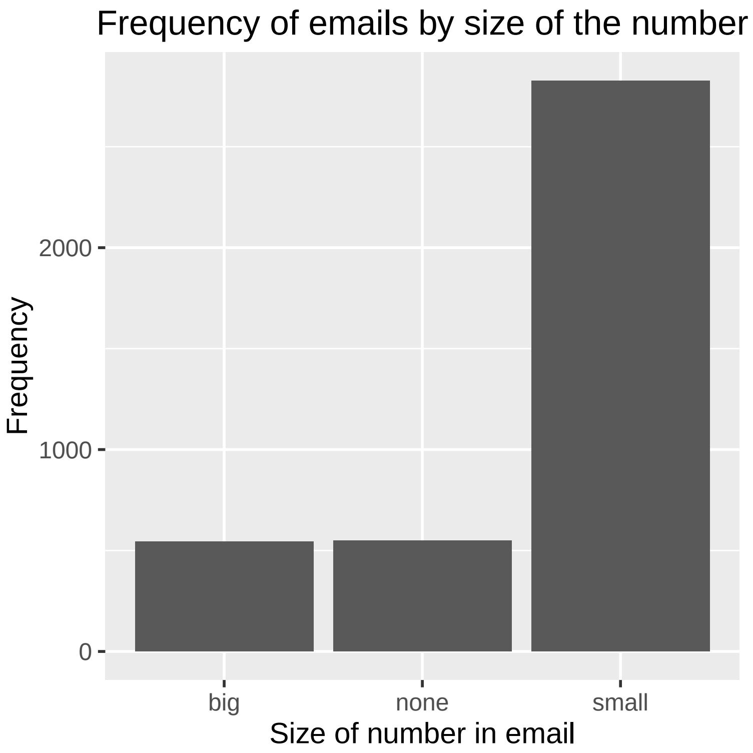
```
##
##   big  none small
##  13.9  14.0  72.1
```

Out of the 3921 emails in this data set, 545 (13.9%) contain big numbers, 2827 (72.1%) contain small numbers, and 549 (14.0%) contain no numbers.

One variable barcharts

The most common method to display frequencies in a graphical manner is with a *bar chart* (aka barplot or bar graph). One bar per distinct category with the height of the bar representing the frequency (or percent) of the data that fall into that category.

```r
ggplot(email, aes(number)) + geom_bar() + ylab("Frequency") + xlab("Size of number in email") +
  ggtitle("Frequency of emails by size of the number")
```



Research Question: Do more than half of the emails flagged as spam contain numbers?

We are only interested in the emails that are flagged spam, and that contain any number, so let's do a little data cleaning first.

```r
spam.num <- email %>%
              filter(spam==1) %>%
              mutate(hasnum = ifelse(number %in% c("big", "small"), 1, 0))
# Confirm recode
table(spam.num$number, spam.num$hasnum, useNA="always")
```

```
##
##             0   1 <NA>
##   big       0  50    0
##   none    149   0    0
##   small     0 168    0
##   <NA>      0   0    0
```

- What is the proportion of spam emails that contain numbers? Since we created `hasnum` as a binary indicator, we can calculate the sample proportion as the sample mean.

```r
mean(spam.num$hasnum)
```

```
## [1] 0.5940054
```

- Construct a 99% CI for the proportion of spam emails with numbers. Since this proportion follows the normal model we can conduct a t.test.

```
t.test(spam.num$hasnum, alternative = "two.sided", conf.level=.99)
```

```
##
##  One Sample t-test
##
## data:  spam.num$hasnum
## t = 23.141, df = 366, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  0.5275391 0.6604718
## sample estimates:
## mean of x
## 0.5940054
```

Over half (59.4%, 95%CI: 52.7, 66.0) of emails that are flagged spam contain numbers (p<.0001). This is statistically significantly larger than half.

[top]

Inference for two categorical variables

Two way (two variable) tables

```
tab <- table(email$spam, email$number)
tab
```

```
##
##      big none small
##   0  495  400  2659
##   1   50  149   168
```

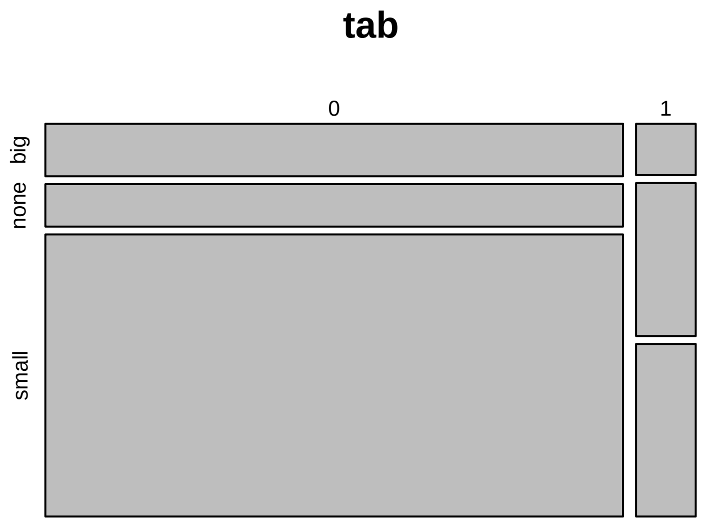
Cell percents

A simple `prop.table()` shows the cell percents. All cells add up to 1. The `mosaicplot` is a visual representation of these percentages. The larger the box the larger the proportion.

```
prop.table(tab)
```

```
##
##            big       none      small
##   0 0.12624331 0.10201479 0.67814333
##   1 0.01275185 0.03800051 0.04284621
```

```
mosaicplot(tab)
```



We will come back to these mosaicplots later.

Row percents

Here, the percents add up to 1 across the rows. The reference group (denominator) is the row category.

```
round(prop.table(tab, 1)*100, 2)
```

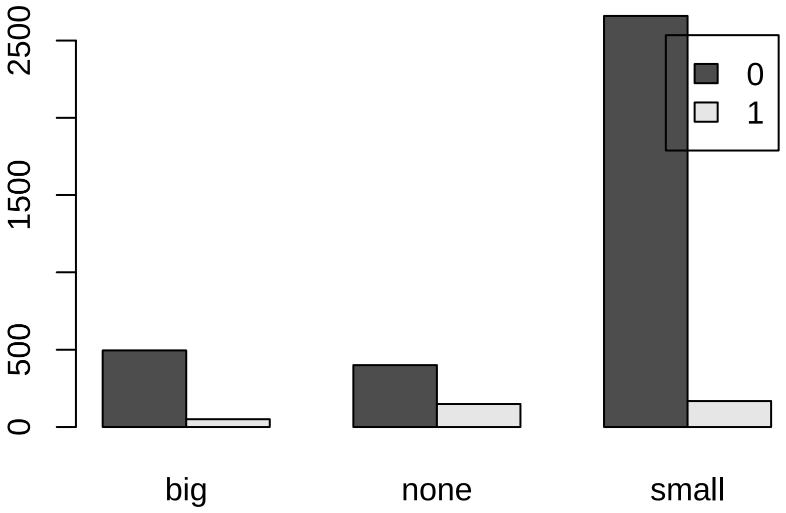
```
##
##       big  none small
##   0 13.93 11.25 74.82
##   1 13.62 40.60 45.78
```

- 13.9% *of non-spam emails* contain big numbers
- 40.6% *of spam emails* contain no numbers
- 74.8% *of non-spam emails* contain small numbers.

Base graphics

Does "ok", but it depends on how you set up your table (rows vs columns), and you have to do more work to get a reasonable legend and y axis.

```
barplot(tab, beside=TRUE, legend=rownames(tab))
```



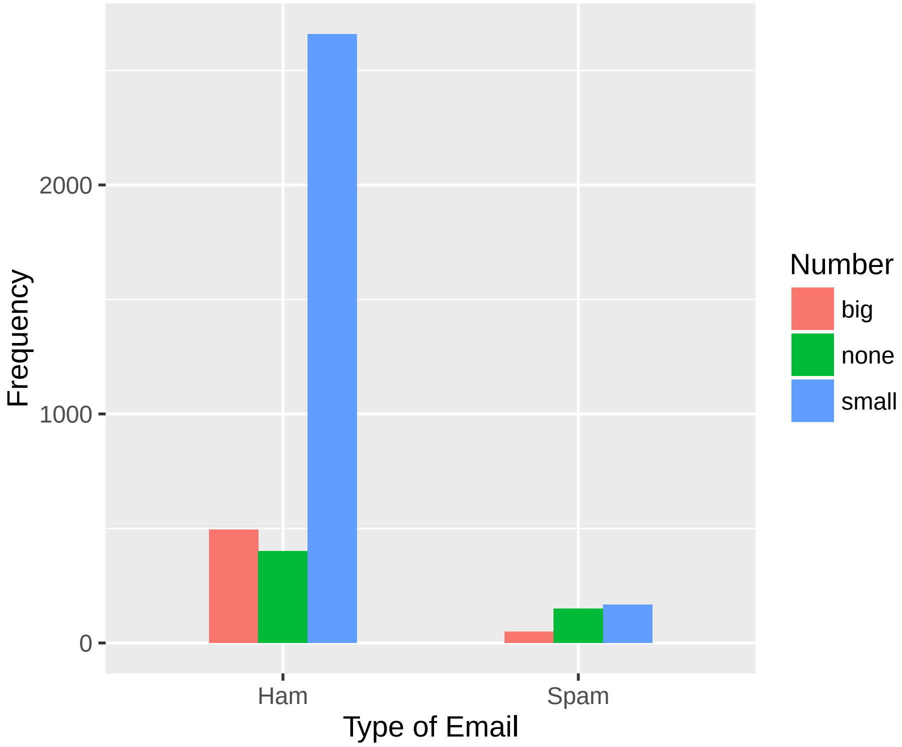
GGplot2

Requires the data to be aggregated first, then specify the heights of the bars as a new variable.

```
library(reshape2)
melted_tab <- melt(tab)
colnames(melted_tab) <- c("Spam", "Number", "count")
melted_tab
```

```
##   Spam Number count
## 1    0    big   495
## 2    1    big    50
## 3    0   none   400
## 4    1   none   149
## 5    0  small  2659
## 6    1  small   168
```

```
ggplot(melted_tab, aes(x = factor(Spam), y= count, fill = Number)) +
  geom_bar(stat="identity", width=.5, position = "dodge")  +
  scale_x_discrete("Type of Email", labels=c("Ham", "Spam")) +
  ylab("Frequency")
```



Column percents

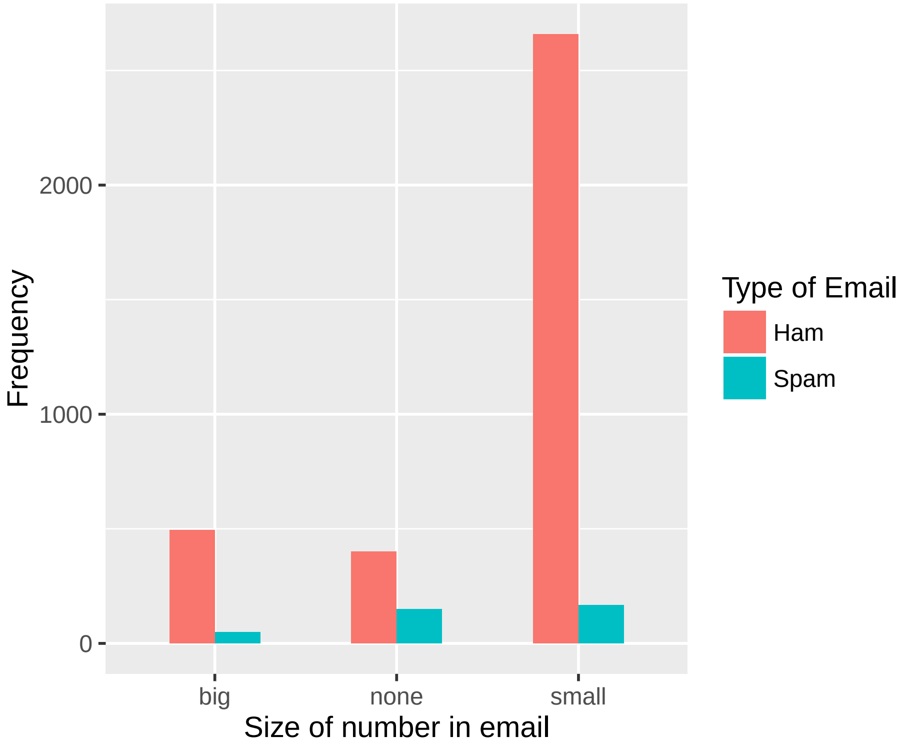
Here, the percents add up to 1 down the columns. The reference group (denominator) is the column category.

```
round(prop.table(tab, 2)*100, 2)
```

```
##
##       big  none small
##   0 90.83 72.86 94.06
##   1  9.17 27.14  5.94
```

- 90.8% *of emails with big numbers* are not spam
- 27.1% *of emails with no numbers* are spam
- 94.1% *of emails with small numbers* are not spam

```
ggplot(melted_tab, aes(x = Number, y= count, fill = factor(Spam))) +
  geom_bar(stat="identity", width=.5, position = "dodge")  +
  xlab("Size of number in email") + ylab("Frequency") +
  scale_fill_discrete("Type of Email", labels=c("Ham", "Spam"))
```



Summary / Take home message

It is very important to be clear as to what comparison you (or the researcher) are interested in making. Sometimes both directions are equally important, but often there is one primary direction that is of interest.

Difference of two proportions

- OpenIntro Section 6.2

Now let's consider comparisons of proportions in two independent samples.

Example: Comparison of proportions of head injuries sustained in auto accidents by passengers wearing seat belts to those not wearing seat belts. You may have already guessed the form of the estimate: $\hat{p}_1 - \hat{p}_2$.

We are not going to go in depth into the calculations for the test statistic for a test of the difference in proportions. The OpenIntro textbook explains the assumptions and equations very well. Instead we are going to see how to use R to perform these calculations for us.

Since the sample proportion can be calculated as the mean of a binary indicator variable, we can use the same `t.test` function in R to conduct a hypothesis test and create a confidence interval.

Tests of proportions using case level data

Do emails that contain numbers more likely to be spam?

```
email <- email %>% mutate(hasnum =ifelse(number %in% c("big", "small"), 0, 1))
t.test(spam~hasnum, data=email)
```

```
##
##  Welch Two Sample t-test
##
## data:  spam by hasnum
## t = -10.623, df = 603.6, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2449747 -0.1685303
## sample estimates:
## mean in group 0 mean in group 1
##      0.06465006      0.27140255
```

Tests of proportions in R using summary numbers only

Sometimes you only have a summary table of numbers to go on, not the raw data. We can use the `prop.test()` function in R to conduct a test of equal proportions.

This function calculates a test statistic and a p-value using a χ^2 distribution.

- The χ^2 distribution is mathematically related to many other distributions.
  - Specific of note here is that a Z^2 distribution has a χ^2_1 distribution.
  - Other relationships that are very useful:
    * https://onlinecourses.science.psu.edu/stat414/node/154
    * https://en.wikipedia.org/wiki/Chi-squared_distribution#Relation_to_other_distributions

Example 1: Spam vs Ham

Recall that 218 out of 367 (59%) of all emails that were flagged spam contained numbers.

```
table((email$number !="none"), email$spam, dnn=c("Contain numbers", "Spam"))
```

```
##                  Spam
## Contain numbers    0    1
##           FALSE  400  149
##           TRUE  3154  218
```

```
round(prop.table(table((email$number !="none"), email$spam, dnn=c("Contain numbers", "Spam")), 2)*100,2
```

```
##                  Spam
## Contain numbers     0     1
##           FALSE 11.25 40.60
##           TRUE  88.75 59.40
```

We can test if that proportion is significantly more than half by calling `prop.test()`.

```
prop.test(x=218, n=367, p=.5, alternative="greater")
```

```
##
##  1-sample proportions test with continuity correction
##
```

```
## data:  218 out of 367, null probability 0.5
## X-squared = 12.599, df = 1, p-value = 0.0001929
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5499273 1.0000000
## sample estimates:
##         p
## 0.5940054
```

Example 2: Are mammograms helpful?

Test whether there was a difference in breast cancer deaths in the mammogram and control groups. By entering in x and n as vectors we can test equivelance of these two proportions. The assumptions for using the normal model for this test have been discussed in detail in the textbook.

```r
prop.test(x=c(500, 505), n=c(44925, 44910))
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  c(500, 505) out of c(44925, 44910)
## X-squared = 0.01748, df = 1, p-value = 0.8948
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.001512853  0.001282751
## sample estimates:
##     prop 1     prop 2
## 0.01112966 0.01124471
```

[top]

Chi-Squared tests for contingency tables.

prop.test(table(email$number, email$spam))

mosaicplot(tab, shade=TRUE)

Testing Equality of proportions

Testing Goodness of Fit

Testing Independence

Comments on Fishers Exact test

[top]

Logistic Regression

Going further

When your outcome has more than one level and you want to build a regression model to assess the impact a specific variable (or set of variables) has on the levels of this outcome variable, you would need to turn to more generalized linear models such as: * Multinomial distribution for a nominal outcome - http://www.ats.ucla.edu/stat/r/dae/mlogit.htm * Ordinal logistic regression

[top]

On Your Own

For all hypothesis tests you must:

- Clearly state what your null and alternative hypothesis.

- Discuss your assumptions
- Show your R code to conduct the test and any data management required prior to the test.
- Write the conclusion of the statistical test in context of the problem.

On Your Own

1. Using the `survey` data set contained in the `MASS` library, recode the exercise variable into a dichotomous indicator of any exercise. Test the hypothesis that more than 75% of the students in the sample exercise at some level.
2. Using the student `survey` data where the variable `smoker` is a binary indicator of smoking, test if males smoke more than females. *Hint: check your output carefully and consider if the difference was taken in the desired direction*