Lec 07: Categorical Data Analysis

MATH 456 - Spring 2016

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

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 reference 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.

Assigned Reading

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

Additional References

- Mike Marin: Relative Risk and Odds Ratio https://www.youtube.com/watch?v=V_YNPQoAyCc
- An **excellent** additional textbook reference: *Categorical Data Analysis* by Alan Agresti
- Using R for Biomedical Statistics http://a-little-book-of-r-for-biomedical-statistics.readthedocs.org/en/latest/src/biomedicalstats.html#
- Graphics
  - **New find!** Graphical Data Analysis with R http://www.gradaanwr.net/content/04-displaying-categorial-data/
  - My Data viz 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)/

Spam Data

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

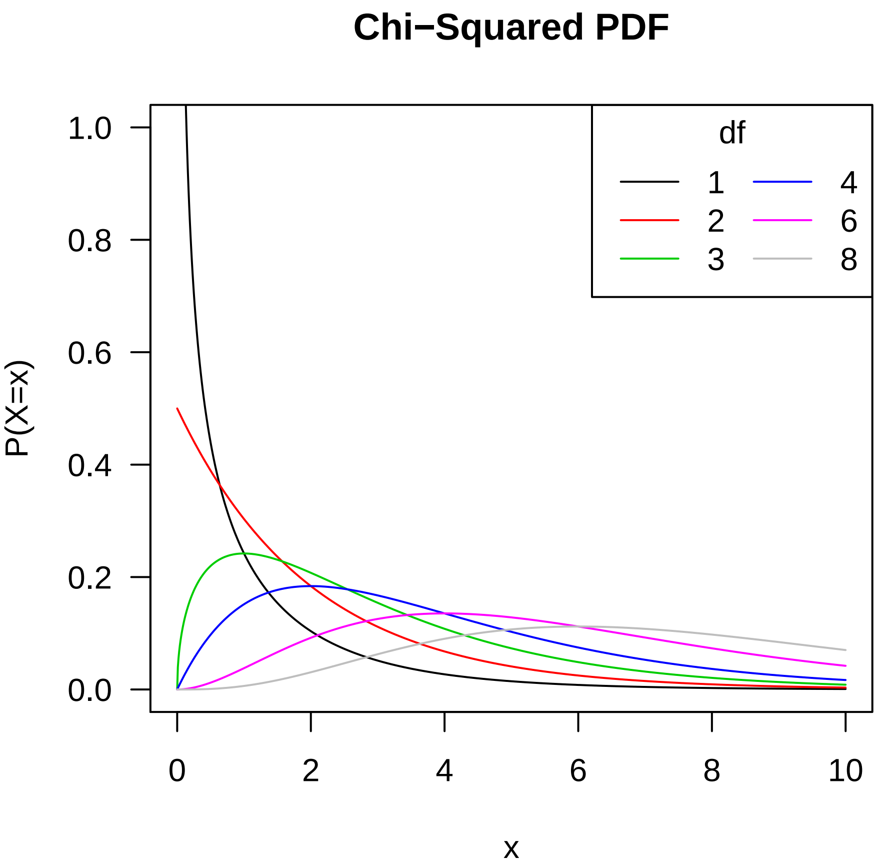
```r
email <- read.delim("C:/GitHub/MATH456/data/email.txt", header=TRUE, sep="\t")
email <- email %>% mutate(hasnum = ifelse(number %in% c("big", "small"), 1, 0))
```

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
- `hasnum`: 0/1 binary indicator for if the email contains any sized number

Chi-Squared Distribution

Much of categorical data analysis uses the χ^2 distribution.



- The shape is controlled by a degrees of freedom parameter (df)
- Is used in many statistical tests for categorical data.
- Is always positive (it's squared!)

    - High numbers result in low p-values
- Mathematically connected to many other distributions
    - Special case of the gamma distribution (One of the most commonly used statistical distributions)
    - The sample variance has a χ^2_{n-1} distribution.
    - The sum of k independent standard normal distributions has a χ^2_k distribution.
    - The ANOVA F-statistic is the ratio of two χ^2 distributions divided by their respective degrees of freedom.

[top]

Review: Single categorical variable

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 multiplying by 100.

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

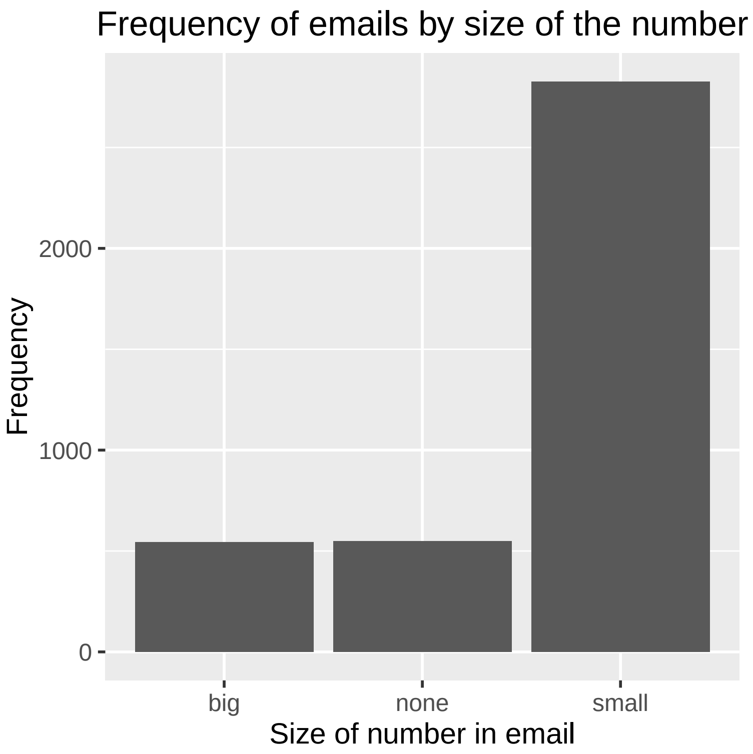
```
##
##   big  none small
##  13.9  14.0  72.1
```

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.

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.

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



Inference

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 the following:

- Read Open Intro: Chapter 6.1
- Complete 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: Do more than half of the spam emails contain numbers?

We are only interested in the emails that are flagged spam so we filter the data first.

```
spam <- email %>% filter(spam==1)
```

Using case-level data

Since we created `hasnum` as a binary indicator, we can calculate the sample proportion as the sample mean.

```
mean(spam$hasnum)
```

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

- Construct a 95% 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$hasnum, alternative = "two.sided", conf.level=.95)
```

```
##
##  One Sample t-test
##
## data:  spam$hasnum
## t = 23.141, df = 366, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5435275 0.6444834
## sample estimates:
## mean of x
## 0.5940054
```

A significant majority (59.4%, 95%CI: 54.3-64.4) of emails that are flagged spam contain numbers (p<.0001).

Using summary numbers only

If we only had summary statistics, for example from a published table,

```
addmargins(table(spam$hasnum))
```

```
##
##   0   1 Sum
## 149 218 367
```

we can still conduct a test to determine if this proportion is significantly more than half by calling `prop.test()`.

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

```
##
##  1-sample proportions test with continuity correction
##
## data:  218 out of 367, null probability 0.5
## X-squared = 12.599, df = 1, p-value = 0.0003859
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5416663 0.6443445
## sample estimates:
##         p
## 0.5940054
```

Significantly more than half (59%, 95%CI 54.2%-64.4%) of emails that were flagged as spam contain numbers (p<.0004).

[top]

Difference of two proportions

- OpenIntro Section 6.2

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

Ex: 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.

Example 1: Do numbers in emails affect rate of spam? (Case level data)

If we look at the rate of spam for emails with and without numbers, we see that 6% of emails with numbers are flagged as spam compared to 27% of emails without numbers are flagged as spam.

```
email %>% group_by(hasnum) %>% summarize(p.spam=round(mean(spam),2))
```

```
## Source: local data frame [2 x 2]
##
##   hasnum p.spam
##    (dbl)  (dbl)
## 1      0   0.27
## 2      1   0.06
```

This is such a large difference that we don't really *need* a statistical test to tell us that this difference is significant. But we will do so anyhow for examples sake.

1. **State the research question:** Are emails that contain numbers more likely to be spam?
2. **Define your parameters:**
   Let p_{nonum} be the proportion of emails *without* numbers that are flagged as spam.
   Let p_{hasnum} be the proportion of emails *with* numbers that are flagged as spam.
3. **Set up your statistical hypothesis:**
   $H_0 : p_{nonum} = p_{hasnum}$
   $H_A : p_{nonum} \neq p_{hasnum}$
4. **Check assumptions:** Use the pooled proportion $\hat{p}$ to check the success-failure condition.

```
p.hat <- mean(email$spam)
p.hat
```

```
## [1] 0.09359857
```

- $\hat{p} * n_{nonum}$ = `p.hat * sum(email$hasnum==0)` = 51.3856159
- $\hat{p} * n_{hasnum}$ = `p.hat * sum(email$hasnum==1)` = 315.6143841
- $(1 - \hat{p}) * n_{nonum}$ = `(1-p.hat)* sum(email$hasnum==0)` = 497.6143841

- $(1-\hat{p}) * n_{hasnum}$ = `(1-p.hat)* sum(email$hasnum==1)` = 3056.3856159

The success-failure condition is satisfied since all values are at least 10, and we can safely apply the normal model.

5. **Test the hypothesis** by calculating a test statistic and corresponding p-value. Interpret the results in context of the problem.

```
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:
##  0.1685303 0.2449747
## sample estimates:
## mean in group 0 mean in group 1
##      0.27140255      0.06465006
```

Significantly more emails with numbers were flagged as spam compared to emails without numbers (27.1% versus 6.4% , p<.0001).

Example 2: Are mammograms helpful? (Summary numbers only)

Read Open Intro: 6.2.3

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 equivalence of these two proportions. The assumptions for using the normal model for this test have been discussed in detail in the textbook.

```
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
```

The interval for the difference in proportions covers zero and the p-value for the test is 0.894, therefore the proportion of deaths due to breast cancer are equal in both groups. There is no indication from this data that mammograms in addition to regular breast cancer screening, change the risk of death compared to just the regular screening exams alone.

[top]

Contingency tables

- Both the explanatory and the response variables are categorical (Nominal or Ordinal)
- Tables representing all combinations of levels of explanatory and response variables
- A.k.a Two-way tables or *cross-tabs*
- Numbers in table represent Counts of the number of cases in each cell

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

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

- Row and column totals are called Marginal counts

```
addmargins(tab)
```

```
##
##        big none small  Sum
##   0    495  400  2659 3554
##   1     50  149   168  367
##   Sum  545  549  2827 3921
```

Cell percents

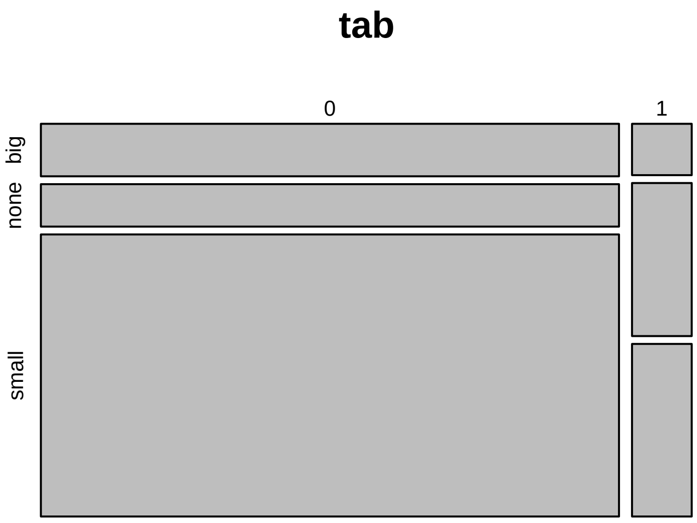
A simple `prop.table()` shows the cell percents. All cells add up to 1.

```
round(addmargins(prop.table(tab)),3)
```

```
##
##         big  none small   Sum
##   0   0.126 0.102 0.678 0.906
##   1   0.013 0.038 0.043 0.094
##   Sum 0.139 0.140 0.721 1.000
```

The `mosaicplot` is a visual representation of these percentages. The larger the box the larger the cell proportion.

```
mosaicplot(tab)
```



We will come back to the usefulness of mosaicplots in analyzing contingency tables later.

Row percents

Here, the percents add up to 1 across the rows. The reference group (denominator) is the row margin. This is used when you want to compare the distribution of percents within each column

Ex: How does the distribution of number sizes differ for spam vs non-spam emails?

```
row.pct <- round(prop.table(tab, 1)*100, 2)
addmargins(row.pct, 2)
```

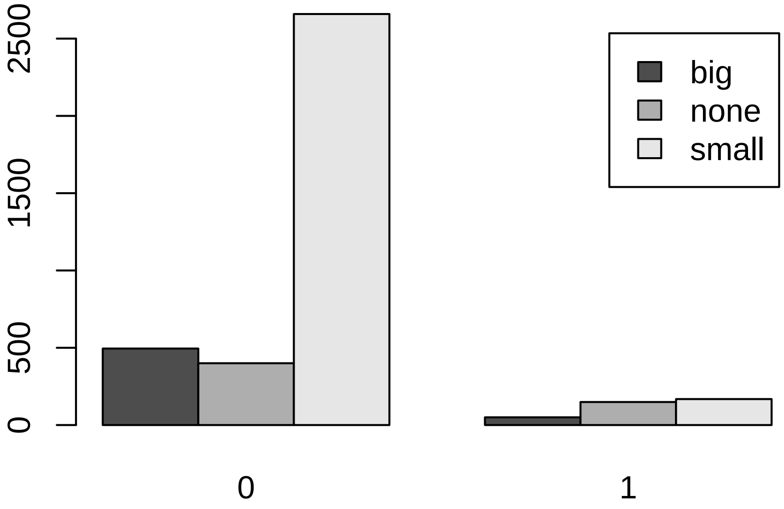
```
##
##        big   none  small    Sum
##   0  13.93  11.25  74.82 100.00
##   1  13.62  40.60  45.78 100.00
```

- 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 (note I had to transpose the table `t()` to plot the distribution of numbers within spam). Also you have to do more work to get a reasonable legend and axes.

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



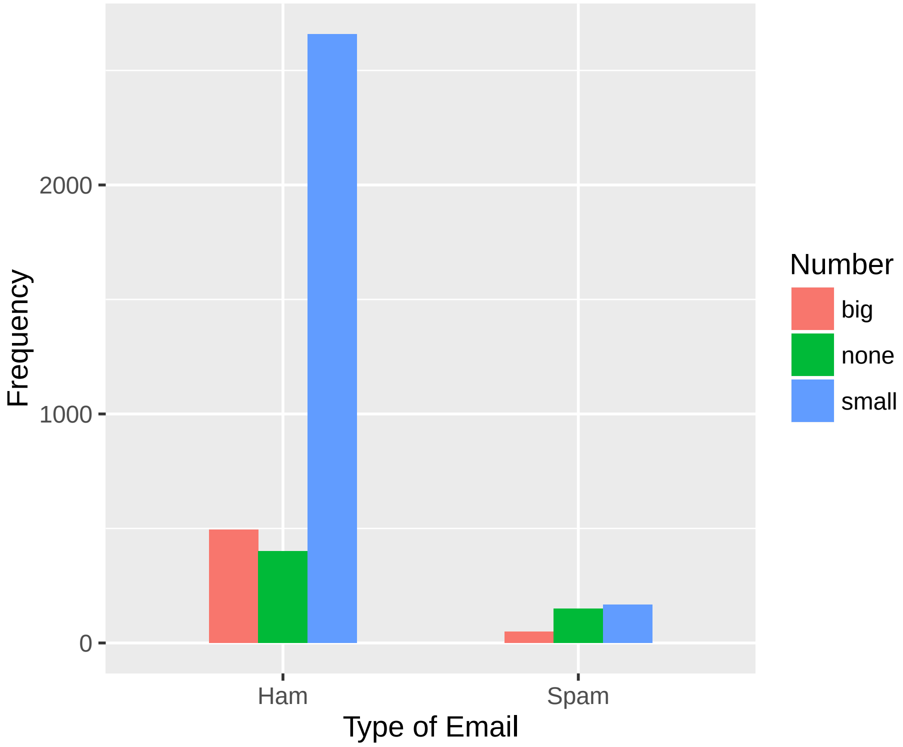
GGplot2

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 margin. This is used when you want to compare the distribution of the rows within each column.

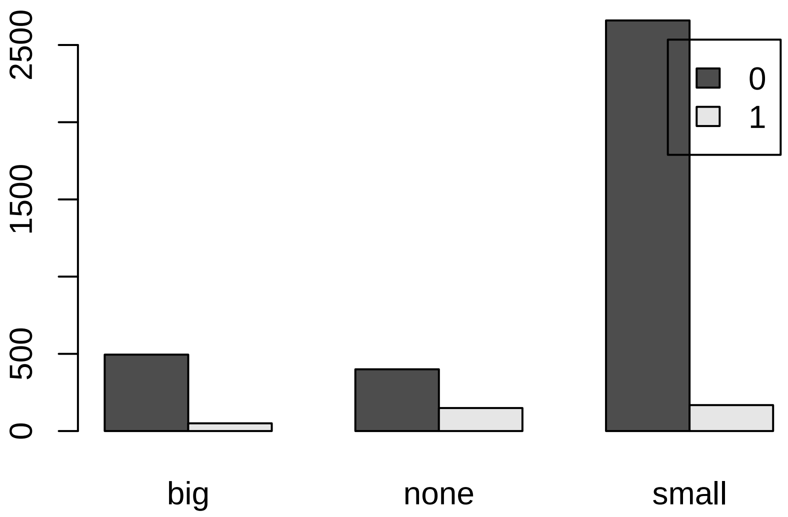
Ex: How does the distribution of spam differ across number sizes?

```
col.pct <- round(prop.table(tab, 2)*100, 2)
addmargins(col.pct,1)
```

```
##
##          big   none  small
##   0    90.83  72.86  94.06
##   1     9.17  27.14   5.94
##   Sum 100.00 100.00 100.00
```

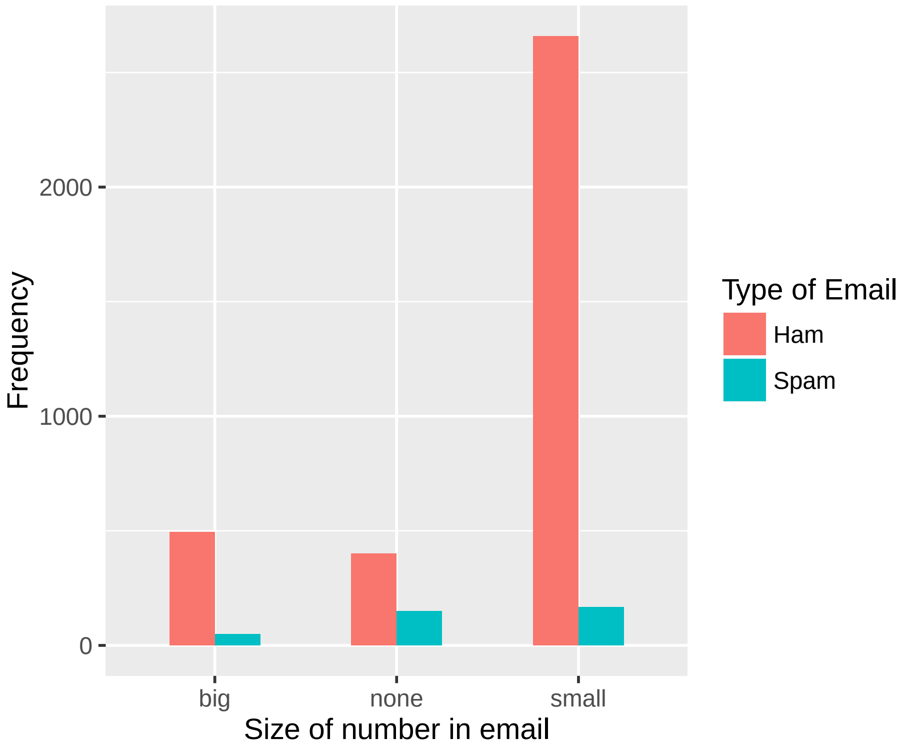
- 90.8% *of emails with big numbers* are not spam
- 27.1% *of emails with no numbers* are spam
- 94.1% *of emails with small numbers* are not spam

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



Once the data is melted, all you have to do is swap x and color variables to plot type of email within number size category.

```
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.

[top]

Measures of Association

We will consider two measures of association in this class.

- Relative Risk
- Odds Ratio

These both are calculated on a 2x2 contingency table similar to the following:

```
nnnn <- matrix(c("$n_{11}$", "$n_{12}$", "$n_{1.}$",
                 "$n_{21}$", "$n_{22}$", "$n_{2.}$",
                 "$n_{.1}$", "$n_{.2}$", "$n_{..}$"), nrow=3, byrow=TRUE,
          dimnames = list(c("Exposed", "Not-Exposed", "Total"), c("Diseased", "Not-Diseased", "Total"))
print(xtable(nnnn, align='cccc'), type='latex')
```

Sometimes the cell contents are abbreviated as:

	Diseased	Not-Diseased	Total
Exposed	$n_{11}$	$n_{12}$	$n_{1.}$
Not-Exposed	$n_{21}$	$n_{22}$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

```
abcd <- matrix(c("a", "b", "c", "d"), nrow=2,
          dimnames = list(c("Exposed", "Not-Exposed"), c("Diseased", "Not-Diseased")))
print(xtable(abcd, align='ccc'), type='latex')
```

	Diseased	Not-Diseased
Exposed	a	c
Not-Exposed	b	d

Watch this 6 minute Marin Stats Lecture: **https://www.youtube.com/watch?v=V_YNPQoAyCc**

Relative Risk

The **Relative Risk (RR)** or **Risk Ratio** is the ratio of the probability of an event occurring in an exposed group compared to the probability of an event occurring in a non-exposed group.

- Asymptotically approaches the OR for small probabilities.
- Often used in cohort studies and randomized control trials.

Consider sample proportions Diseases within Exposed and Non-exposed groups.

$$\hat{\pi}_1 = \frac{n_{11}}{n_{1.}} \qquad \text{and} \qquad \hat{\pi}_2 = \frac{n_{21}}{n_{2.}}$$

The Relative Risk is calculated as

$$RR = \frac{\hat{\pi}_1}{\hat{\pi}_2} \qquad \text{or} \qquad \frac{a/(a+b)}{c/(c+d)}$$

with variance

$$V = \frac{1-\hat{\pi}_1}{n_{11}} + \frac{1-\hat{\pi}_2}{n_{21}}$$

Odds Ratio

The **Odds Ratio (OR)** is a way to quantify how strongly the presence or absence of a characteristic affects the presence or absence of a second characteristic.

- Often used in case-control studies
- The main interpretable estimate generated from logistic regression

The **Odds of an event** is the probability it occurs divided by the probability it does not occur.

$$odds_1 = \frac{n_{11}/n_{1.}}{n_{12}/n_{1.}} = \frac{n_{11}}{n_{12}}$$

$$odds_2 = \frac{n_{21}/n_{2.}}{n_{22}/n_{2.}} = \frac{n_{21}}{n_{22}}$$

The **Odds Ratio** for group 1 compared to group 2 is the ratio of the two odds written above:

$$OR = \frac{odds_1}{odds_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \qquad \text{or} \qquad \frac{ad}{bc}$$

with variance $V = n_{11}^{-1} + n_{12}^{-1} + n_{21}^{-1} + n_{22}^{-1}$.

Confidence Intervals

Neither the Risk Ratio nor the Odds Ratio are linear functions, so a 95% CI for the population estimates are not your typical $\hat{\theta} \pm 1.96\sqrt{Var(\hat{\theta})}$.

Instead they are calculated as the point estimate $\hat{\theta}$ times e raised to the ±1.96 times the standard deviation of the estimate.

$$(\hat{\theta}e^{-1.96*\sqrt{V}}, \hat{\theta}e^{1.96*\sqrt{V}})$$

Example: Are emails with numbers in them more likely to be flagged as spam?

Reconsider the 2x2 table that compares emails flagged as spam to those containing numbers.

```
table(email$hasnum, email$spam, dnn=c("Has Number", "Spam"))
```

```
##             Spam
## Has Number    0    1
##           0  400  149
##           1 3154  218
```

Note that both the columns and rows are swapped when compared to the a/b/c/d format. For ease of interpretation I will recreate the table manually.

```
tab_sn <- matrix(c(149, 218, 400, 3154), nrow=2, byrow=T,
                 dimnames = list(c("Has Num", "No Num"), c("Spam", "Ham")))
tab_sn
```

```
##         Spam  Ham
## Has Num  149  218
## No Num   400 3154
```

Now I use the `epi.2by2` function contained in the `epiR` package to calculate the Odds Ratio, the Risk Ratio, and their respective confidence intervals.

```r
library(epiR)
epi.2by2(tab_sn)
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          149          218        367              40.6
## Exposed -          400         3154       3554              11.3
## Total              549         3372       3921              14.0
##                   Odds
## Exposed +       0.683
## Exposed -       0.127
## Total           0.163
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                               3.61 (3.09, 4.21)
## Odds ratio                                   5.39 (4.27, 6.80)
## Attrib risk *                                29.34 (24.21, 34.48)
## Attrib risk in population *                  2.75 (1.24, 4.25)
## Attrib fraction in exposed (%)               72.28 (67.65, 76.24)
## Attrib fraction in population (%)            19.62 (15.83, 23.23)
## -------------------------------------------------------------------
##  X2 test statistic: 237.889 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

- Emails containing numbers are 3.6 (3.09, 4.21) times as likely as emails without numbers to be flagged as spam.
- Emails containing numbers have 5.4 (4.27, 6.80) times the odds of being flagged as spam compared to emails without numbers in them.

Both intervals are greater than 1, therefore the event (spam) is statistically more likely to occur in the exposed group (has num) than in the control (no num) (p<.0001). The p-value for the Wald χ^2 test is <.0001.

- Mathematical reference for the Wald test Statistic http://www.statlect.com/Wald_test.htm

[top]

Tests of Association

There are three main tests of association for rxc contingency table.

- Test of Goodness of Fit
- Tests of Independence
- Test of Homogeneity

Notation

- r is the number of rows and indexed by i
- c is the number of columns and indexed by j.

Goodness of Fit

- OpenIntro Statistics: Chapter 6.3
- Tests whether a set of multinomial counts is distributed according to a theoretical set of population proportions.
- Does a set of categorical data come from a claimed distribution?
- Are the observed frequencies consistent with theory?

H_0: The data come from the claimed discrete distribution
H_A: The data to not come from the claimed discrete distribution.

Test of Independence

- OpenIntro Statistics: Chapter 6.4
- Determine whether two categorical variables are associated with one another in the population
  - Ex. Race and smoking, or education level and political affiliation.
- Data are collected at random from a population and the two categorical variables are observed on each unit.

$H_0 : p_{ij} = p_{i.}p_{.j}$
$H_A : p_{ij} \neq p_{i.}p_{.j}$

Test of Homogeneity

- A test of homogeneity tests whether two (or more) sets of multinomial counts come from different sets of population proportions.
- Does two or more sub-groups of a population share the same distribution of a single categorical variable?
  - Ex: Do people of different races have the same proportion of smokers?
  - Ex: Do different education levels have different proportions of Democrats, Republicans, and Independents?
- Data on one characteristic is collected from randomly sampling individuals within each subroup of the second characteristic.

H_0 :

$$\begin{gathered} p_{11} = p_{12} = \ldots = p_{1c} \\ p_{21} = p_{22} = \ldots = p_{2c} \\ \vdots \\ p_{r1} = p_{r2} = \ldots = p_{rc} \end{gathered}$$

H_A : At least one of the above statements is false.

All three tests use the **Pearsons' Chi-Square** test statistic.

Pearsons' Chi-Square

The chi-squared test statistic is the sum of the squared differences between the observed and expected values, divided by the expected value.

One way table

$$\chi^2 = \sum_{i=1}^{r} \frac{(O_i - E_i)^2}{E_i}$$

- O_i observed number of type i
- E_i expected number of type i. Equal to Np_i under the null hypothesis
- N is the total sample size
- df = r-1

Two way tables

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- O_{ij} observed number in cell ij
- $E_{ij} = Np_{i.}p_{.j}$ under the null hypothesis
- N is the total sample size
- df = (r-1)(c-1)

Watch this 3 minute Marin Stats Lecture: https://www.youtube.com/watch?v=POiHEJqmiC0

Conducting these tests in R.

- Test of equal or given proportions using `prop.test()`

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

```
##
##  3-sample test for equality of proportions without continuity
##  correction
##
## data:  table(email$number, email$spam)
## X-squared = 243.51, df = 2, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3
## 0.9082569 0.7285974 0.9405730
```

- Chi-squared contingency table tests and goodness-of-fit tests using `chisq.test()`. This function can take raw data as input

```r
chisq.test(email$number, email$spam)
```

```
##
##  Pearson's Chi-squared test
##
## data:  email$number and email$spam
## X-squared = 243.51, df = 2, p-value < 2.2e-16
```

or a table object.

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 243.51, df = 2, p-value < 2.2e-16
```

prop.test vs chisq.test()

```
pt.out <- prop.test(table(email$number, email$spam))
cs.out <- chisq.test(tab)
```

- Same calculated test statistic and p-value

```
c(pt.out$statistic, pt.out$p.value)
```

```
##    X-squared
## 2.435137e+02 1.323321e-53
```

```
c(cs.out$statistic, cs.out$p.value)
```

```
##    X-squared
## 2.435137e+02 1.323321e-53
```

- `prop.test`
  - has a similar output appearance to other hypothesis tests
  - shows sample proportions of outcome within each group
- `chisq.test`
  - stores the matricies of O_{ij}, E_{ij}, the residuals and standardized residuals

```
cs.out$expected
```

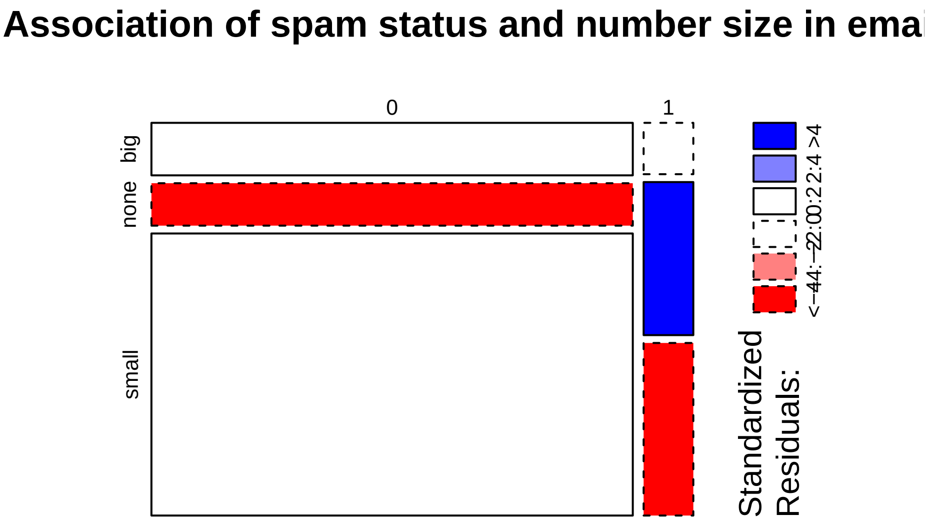
```
##
##           big      none     small
##   0 493.98878 497.61438 2562.3968
##   1  51.01122  51.38562  264.6032
```

Reminder: You can always look at the `names()` of model objects or the help file to see what information is available for post-analysis investigation.

Mosaicplots

- The Pearson χ^2 test statistic = Sum of squared residuals.
- A shaded mosaicplot shows the magnitude of the residuals.
  - Blue (positive residuals) = More frequent than expected
  - Red (negative residuals) = Less frequent than expected.

```r
mosaicplot(tab, shade=TRUE, main="Association of spam status and number size in emails")
```



There are more spam emails with no numbers, fewer Ham emails with no numbers, and fewer spam emails with small numbers than would be expected if these factors were independent.

- More information on mosaicplots - http://www.datavis.ca/online/mosaics/about.html

Assumptions and Extensions

- Simple random sample
- Adequate expected cell counts
  - At least 5 in all cells of a 2x2, or at least 80% of cells in a larger table.
  - NO cells with 0 cell count
- Observations are independent

If one or more of these assumptions are not satisfied, other methods may still be useful. * McNemar's Test for paired or correlated data * Fishers exact test for when cell sizes are small (<5-10) * Inter-rater reliability: Concordant and Discordant Pairs

[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 a two-sample t-test for proportions on the student `survey` data set contained in the `MASS` library, test if males smoke more than females. *Hint: check your output carefully and consider if the difference was taken in the desired direction*

2. Anti-tumor necrosis factor α (TNFα) drugs are a class of drugs that are commonly used to treat inflammatory conditions such as arthritis. However, these drugs tend to be associated with an increased risk of infectious complications. Bergstrom, et al (2004) conducted a study to test the hypothesis that patients on these drugs are at increased risk for coccidiomycosis (a fungal pneumonia). Calculate and interpret the RR with 95% confidence interval. Here are the tabular results.

```
tnf <- matrix(c(7, 240, 4, 734), nrow=2, byrow=TRUE,
          dimnames = list(c("TNF", "Other"), c("COC", "No COC")))
print(xtable(tnf, align='ccc', digits=0), type='latex')
```

	COC	No COC
TNF	7	240
Other	4	734

3. Using the provided bag of M&M's, test the hypothesis that the color distribution has not changed since reported in 2008. https://www.exeter.edu/documents/mandm.pdf

4. In July 2008 the US National Institutes of Health announced that it was stopping a clinical study early because of unexpected results. The study population consisted of HIV-infected women in sub-Saharan Africa who had been given single dose Nevaripine (a treatment for HIV) while giving birth, to prevent transmission of HIV to the infant. The study was a randomized comparison of continued treatment of a woman (after successful childbirth) with Nevaripine vs. Lopinavir, a second drug used to treat HIV. 240 women participated in the study; 120 were randomized to each of the two treatments. Twenty-four weeks after starting the study treatment, each woman was tested to determine if the HIV infection was becoming worse (an outcome called **virologic failure**). Twenty-six of the 120 women treated with Nevaripine experienced virologic failure, while 10 of the 120 women treated with the other drug experienced virologic failure. *(Lockman 2007)*

   a. Create a two-way table presenting the results of this study. Include both margins.
   b. State appropriate hypotheses to test for independence of treatment and virologic failure.
   c. Complete the hypothesis test and state an appropriate conclusion. (Reminder: verify any necessary conditions for the test.)

5. A 2010 survey asked 827 randomly sampled registered voters in California “Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college.

```
oil <- matrix(c(154, 132, 180, 126, 104, 131), nrow=3, byrow=T,
       dimnames = list(c("Support", "Oppose", "Do not know"), c("Grad", "Non-Grad")))
print(xtable(oil, align='ccc', digits=0), type='latex')
```

	Grad	Non-Grad
Support	154	132
Oppose	180	126
Do not know	104	131

a. What percent of college graduates and what percent of the non-college graduates in this sample do not know enough to have an opinion on drilling for oil and natural gas off the Coast of California?
b. Create an appropriate graphic to compare the distribution of opinions within college graduates and non-graduates.
c. Is this a test of homogeneity or independence? Justify your answer.

6. Consider only those participants who have an opinion on off shore drilling. Calculate the Odds Ratio for opposing off shore drilling for college grads compared to non-grads. Include a 99% confidence interval and interpret the results in context of the problem.