Lec 03: Categorical Data Analysis

Last Updated 2018-10-22 21:33:04

Contents

Introduction

This set of lecture notes uses data on incoming emails for the first three months of 2012 for David Diez's (An Open Intro Statistics Textbook author) Gmail Account, early months of 2012. All personally identifiable information has been removed.

The research question of interest is > Does including numbers in an email increase or decrease the chance the email is flagged as spam?

or more generally,

> Is there an association between numbers in an email and the email being flagged as spam?

Data setup

Response Variable

The response variable is whether or not the email is flagged as spam. In the data set originally this variable is listed as

- `spam` 0/1 binary indicator if a an email is flagged as spam

For plotting, and for some analyses we need to have a categorical version of this variable. So we create a new variable `spam_cat` that is a categorical (factor) version of spam with levels Ham and Spam.

```
##
##          Ham Spam <NA>
##   0    3554    0    0
##   1       0  367    0
##   <NA>    0    0    0
```

Explanatory Variable

The explanatory variable `number` measures the size of the number contained in the email. It is a factor variable with three levels:

- `none`: No numbers
- `small`: Only values under 1 million
- `big`: A value of 1 million or more

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

It may not be the size that matters, but whether or not there any number at all. So let's make a binary version of this variable, both as numeric 0/1 and factor Yes/No.

- `hasnum`: 0/1 binary indicator for if the email contains any sized number
- `hasnum_at`: Categorical (factor) version of `hasnum`: Yes/No

```
##
##          big none small <NA>
##   No       0  549     0    0
##   Yes    545    0  2827    0
##   <NA>     0    0     0    0
```

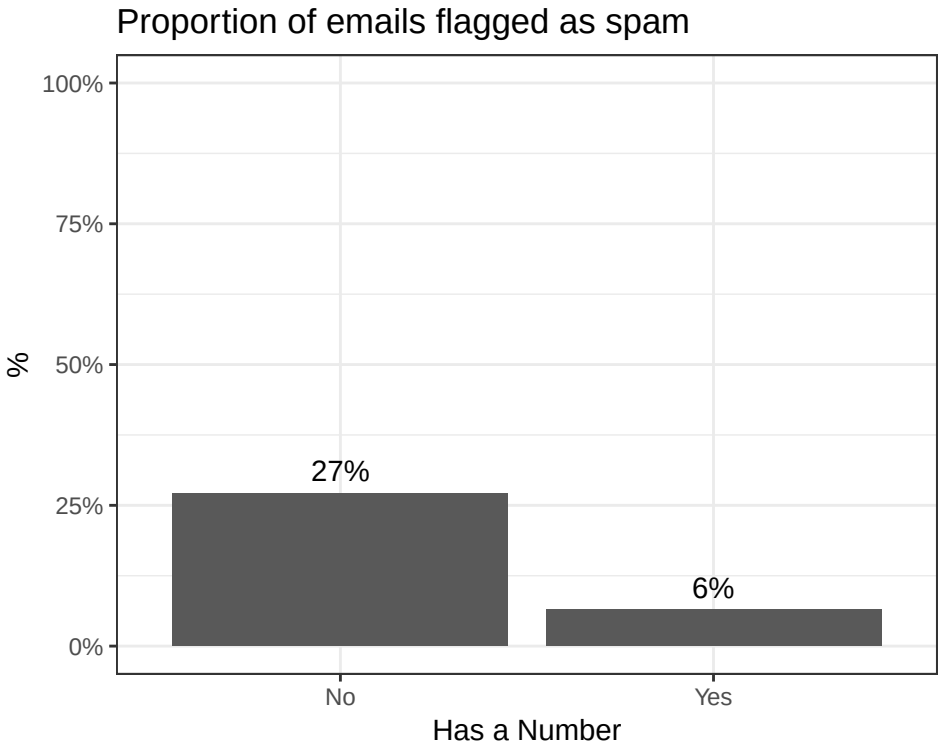
Exploratory Data Analsysis

Looking at the frequency and column percent tables we see that

- 27.1% (149/549) of emails without numbers are flagged as spam
- 6.5% (218/3372) of emails with numbers in them are flagged as spam.

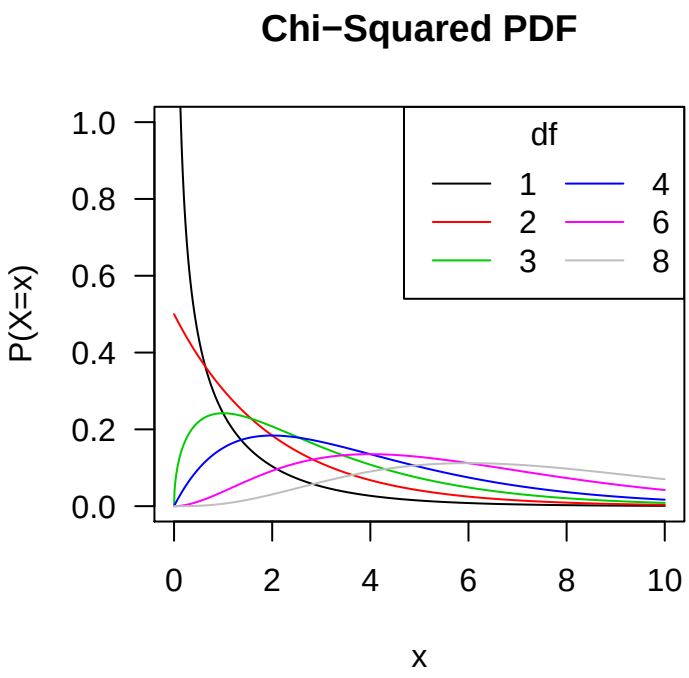
```
##       Has Number
## Spam     No  Yes  Sum
##   Ham   400 3154 3554
##   Spam  149  218  367
##   Sum   549 3372 3921
```

```
##       Has Number
## Spam     No  Yes
##   Ham  72.9 93.5
##   Spam 27.1  6.5
```



Chi-Squared Distribution

Much of categorical data analysis uses the χ^2 distribution. It is a probability distribution, but it has a different shape compared to the Normal. It looks a little bit like an F-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.

Difference of two proportions

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: $\hat{p}_1 - \hat{p}_2$.

Example 1: Do numbers in emails affect rate of spam?

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.

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 $\hat{p} = 367/3921 = .094$ to check the success-failure condition.

- $\hat{p} * n_{nonum}$ = `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.

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  email$spam_cat and email$hasnum_cat
## X-squared = 235.46, df = 1, p-value < 2.2e-16
```

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

Tests of Association

There are three main tests of association for rxc contingency table, where r is the number of rows and indexed by i, and c is the number of columns and indexed by j.

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

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 :

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

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)

Conducting these tests in R.

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

```
##
##  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()`.

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

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

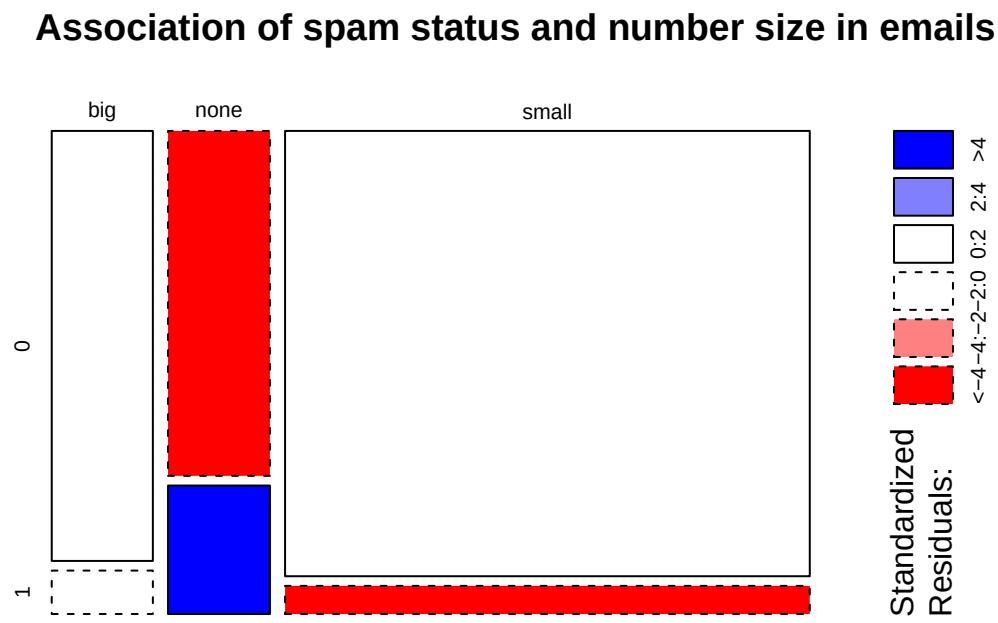
```
##             email$spam
## email$number         0         1
##        big    493.9888  51.01122
##        none   497.6144  51.38562
##        small 2562.3968 264.60316
```

Conducting these tests in SPSS

https://libguides.library.kent.edu/SPSS/ChiSquare

Mosaicplots

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



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