

Categorical Data Analysis

Last Updated 2018-02-06 23:06:23

Contents

Difference of two proportions	1
Contingency tables	3
Measures of Association	3
Relative Risk	4
Odds Ratio	4
Confidence Intervals	5
Tests of Association	6
Goodness of Fit	6
Test of Independence	7
Test of Homogeneity	7
Pearsons' Chi-Square	8
Assumptions and Extensions	11

Spam Data

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.

```
email <- read.delim("https://norcalbiostat.netlify.com/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

Difference of two proportions

Let's review 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))
```

```
## # A tibble: 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
```

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

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.

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
```

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='ccc'))
```

% latex table generated in R 3.4.1 by xtable 1.8-2 package % Tue Feb 6 15:06:25 2018

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

Sometimes the cell contents are abbreviated as:

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

% latex table generated in R 3.4.1 by xtable 1.8-2 package % Tue Feb 6 15:06:25 2018

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

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.

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

The Relative Risk is calculated as

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

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:

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

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.

```
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

Tests of Association

There are three main tests of association for $r \times c$ 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 do 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 subgroup of the second characteristic.

$$H_0 :$$

$$p_{11} = p_{12} = \dots = p_{1c}$$

$$p_{21} = p_{22} = \dots = p_{2c}$$

$$\vdots$$

$$p_{r1} = p_{r2} = \dots = p_{rc}$$

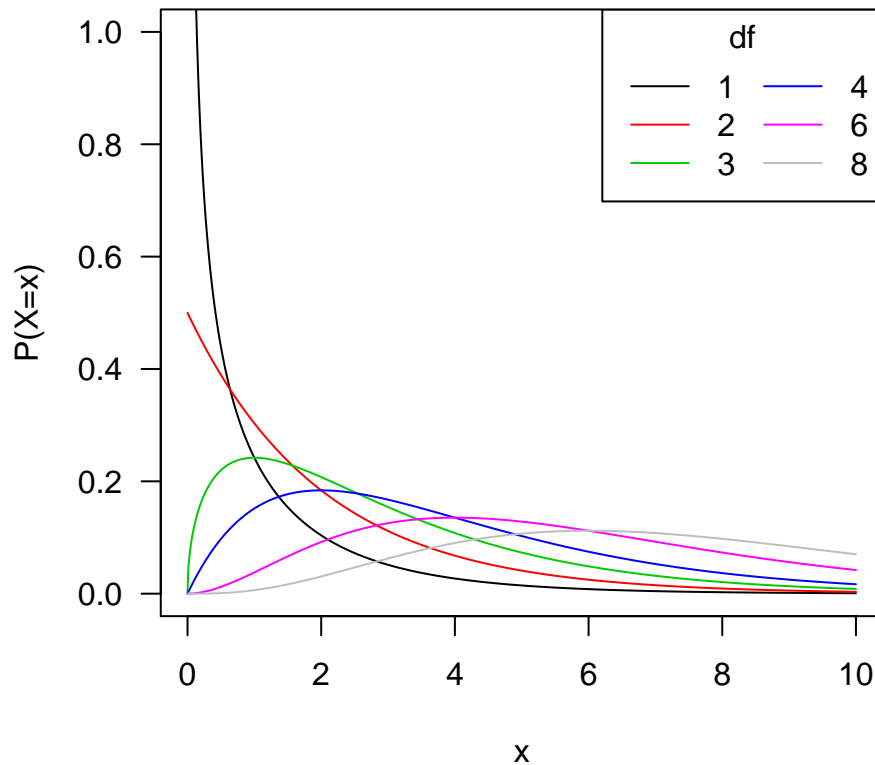
$$H_A : \text{At least one of the above statements is false.}$$

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

Chi-Squared Distribution

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

Chi-Squared PDF



- 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.

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
- $E_{ij} = Np_{i \cdot} p_{\cdot j}$ 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()`

```
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

```
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 matrices of O_{ij} , E_{ij} , the residuals and standardized residuals

```
cs.out$expected
```

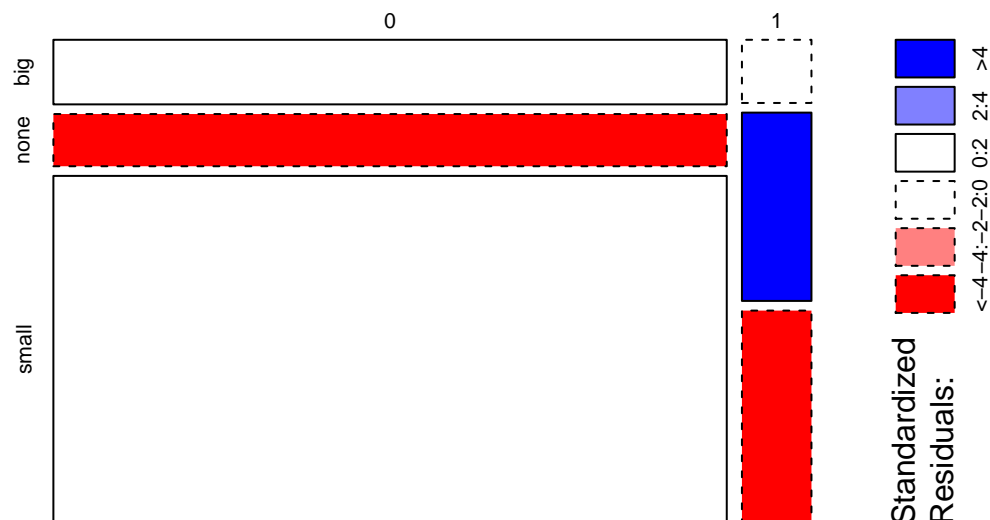
```
##
##          big      none      small
## 0 493.98878 497.61438 2562.3968
## 1  51.01122  51.38562  264.6032
```

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.

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

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