Inference for Categorical Data

Categorical Variables

Reading

- Textbook sections 6.3, 6.4
- OpenIntro Slides:
  - χ^2 test of goodness-of-fit: http://www.openintro.org/redirect.php?go=gdoc_os3_slides_6-3&referrer=os3_pdf
  - χ^2 test of independence: http://www.openintro.org/redirect.php?go=gdoc_os3_slides_6-4&referrer=os3_pdf
- Recommended Exercise: 6.39, 6.40, 6.41, 6.42, 6.45, 6.47, 6.48

Examples of Analyses with Categorical Data

Let’s say we are doing a survey where we ask people what was their first pet.

If we ask 10 people, and our responses are: dog, cat, dog, hamster, goldish, dog, dog, ferret, cat, no pet

Then, how do we calculate average??

Examples of Analyses with Categorical Data

Let's say we are doing a survey where we ask people what was their first pet.

If we ask 10 people, and our responses are: dog, cat, dog, hamster, goldish, dog, dog, ferret, cat, no pet

Then, how do we calculate average??

We can't!

Examples of Analyses with Categorical Data

Let’s say we are doing a survey where we ask people what was their first pet.

If we ask 10 people, and our responses are: dog, cat, dog, hamster, goldish, dog, dog, ferret, cat, no pet

Then, how do we calculate average??

We can’t!

Can we still do analysis??

Examples of Analyses with Categorical Data

Let's say we are doing a survey where we ask people what was their first pet.

If we ask 10 people, and our responses are: dog, cat, dog, hamster, goldish, dog, dog, ferret, cat, no pet

Then, how do we calculate average??

We can't!

Can we still do analysis??

Yep!!!!!! Can use χ^2 tests

Contingency Table Recap

Contingency Tables

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

Before we discuss the test, recall what a contingency table is.

The following contingency table shows:

-count of people using Air France or US Air

-count of service quality score of *poor, fair, good* or *excellent* for each person's ariline

Recall Calculating Probabilities

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

First define the following terms

-n_{ij} = count in i^{th} row, j^{th} column

-$n_{i.}$ = total for i^{th} row

-$n_{.j}$ = total for j^{th} col

-$n_{..}$ = total count

-e.g. $n_{11} = 27, n_{1.} = 63, n_{.1} = 66,$ $n_{..} = 190$

Recall Calculating Probabilities

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

Probability randomly selected person flew Air France?

$$= P(\text{Air France}) = \frac{\text{\# of Air France Ratings}}{\text{total \# of people}}$$

$$= \frac{66}{190} = \frac{n_{\cdot 1}}{n_{\cdot\cdot}}$$

Probability randomly selected rated their service quality as excellent?

$$= P(\text{Excellent}) = \frac{\text{\# of Excellent Ratings}}{\text{total \# of people}}$$

$$= \frac{63}{190} = \frac{n_{1\cdot}}{n_{\cdot\cdot}}$$

Recall Calculating Probabilities

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

Probability randomly selected person flew Air France and rated it excellent?

$$= P(\text{Air France and Excellent})$$
$$= \frac{\text{\# of Excellent Air France Ratings}}{\text{total \# of people}} = \frac{27}{190} = \frac{n_{11}}{n_{..}}$$

Recall Calculating Probabilities

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

Note that if service quality is *independent* of airline, then

$$P(\text{Air France and Excellent}) \\ = P(\text{Air France}) * P(\text{Excellent})$$

Here we had

$$P(\text{Air France and Excellent}) = 0.142 \\ \neq 0.115 = P(\text{Air France})P(\text{Excellent})$$

Does this mean Service Quality is *independent* of airline?

Recall Calculating Probabilities

	Air France	US Air	Row Totals
excellent	27	36	63
fair	10	19	29
good	24	55	79
poor	5	14	19
Col Totals	66	124	190

Note that if service quality is *independent* of airline, then

$$P(\text{Air France and Excellent}) \\ = P(\text{Air France}) * P(\text{Excellent})$$

Here we had

$$P(\text{Air France and Excellent}) = 0.142 \\ \neq 0.115 = P(\text{Air France})P(\text{Excellent})$$

Does this mean service quality is *independent* of airline?

Nope!

Estimating Count

Since the probabilities calculated on the last slide were estimates of the true probabilities, we cannot yet conclude that service quality and airline are indpendent.

Estimating Count

Since the probabilities calculated on the last slide were estimates of the true probabilities, we cannot yet conclude that service quality and airline are indpendent.

Let $\hat{E}_{ij} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \frac{n_{\cdot j}}{n_{\cdot\cdot}} n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

This is the expected count of the ij^{th} cell, **if the variables are independent**.

Estimating Count

Since the probabilities calculated on the last slide were estimates of the true probabilities, we cannot yet conclude that service quality and airline are indpendent.

Let $\hat{E}_{ij} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \frac{n_{\cdot j}}{n_{\cdot\cdot}} n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

This is the expected count of the ij^{th} cell, **if the variables are independent**.

Therefore we can reject the independence assumption if the cell counts, n_{ij} are *far away* from the expected cell counts under independence $\hat{E}_{ij}$

Estimating Count

Since the probabilities calculated on the last slide were estimates of the true probabilities, we cannot yet conclude that service quality and airline are indpendent.

Let $\hat{E}_{ij} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \frac{n_{\cdot j}}{n_{\cdot\cdot}} n_{\cdot\cdot} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

This is the expected count of the ij^{th} cell, **if the variables are independent**.

Therefore we can reject the independence assumption if the cell counts, n_{ij} are *far away* from the expected cell counts under independence $\hat{E}_{ij}$

This is the intuition for Pearson's Chi-Squared test.

Comparing Observed vs Expected Table

Observed Table

	Air France	US Air
excellent	27	36
fair	10	19
good	24	55
poor	5	14

Expected Table

	Air France	US Air
excellent	$\hat{E}_{11} = \frac{n_{1\cdot}n_{\cdot 1}}{n_{\cdot\cdot}}$	$\frac{n_{1\cdot}n_{\cdot 2}}{n_{\cdot\cdot}}$
fair	$\frac{n_{2\cdot}n_{\cdot 1}}{n_{\cdot\cdot}}$	$\frac{n_{2\cdot}n_{\cdot 2}}{n_{\cdot\cdot}}$
good	$\frac{n_{3\cdot}n_{\cdot 1}}{n_{\cdot\cdot}}$	$\frac{n_{3\cdot}n_{\cdot 2}}{n_{\cdot\cdot}}$
poor	$\frac{n_{4\cdot}n_{\cdot 1}}{n_{\cdot\cdot}}$	$\frac{n_{4\cdot}n_{\cdot 2}}{n_{\cdot\cdot}}$

Comparing Observed vs Expected Table

Observed Table

	Air France	US Air
excellent	27	36
fair	10	19
good	24	55
poor	5	14

Expected Table

	Air France	US Air
excellent	21.9	41.1
fair	10.1	18.9
good	27.4	51.6
poor	6.6	12.4

Chi-Square Test for Independence

Hypotheses

H_0: There is no association between the two variables (independence)

H_a: The variables are associated (service quality depends on airline)

Test Statistic

For r rows, and c columns, the test statistic is

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \overset{H_0}{\sim} \chi^2_{df}$$

where $df = (r-1)(c-1)$

Why this degrees of freedom?

Pvalue for this test

Since we know the distribution of X^2 under the null, and we know the larger X^2 is, the more in favor of the alternative our statistic is, we can cancluate our pvalue as:

$$pvalue = P(X^2 > \chi^2_{(r-1)(c-1)})$$

In R

```
#if X2 is your chi square statistic
#r is number of rows, c number of columns
1-pchisq(X2, df = (r-1)*(c-1))
```

Running everything in R

```
t(table(data))
```

```
##              airline
## quality       Air France US Air
##   excellent           27     36
##   fair                10     19
##   good                24     55
##   poor                 5     14
```

```
chisq.test(airline, quality)
```

```
##
##  Pearson's Chi-squared test
##
## data:  airline and quality
## X-squared = 3.0891, df = 3, p-value = 0.3781
```

Conclusion: Since the pvalue > 0.05, at the 5% significance level we do not have evidence of dependence of in service quality and airline.

Quick Hair Color vs Iris Color Example

```
HairEyeColor[,,1]
```

```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

```
chisq.test(HairEyeColor[,,1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 1]
## X-squared = 41.28, df = 9, p-value = 4.447e-06
```

What can we conclude here?

Quick Hair Color vs Iris Color Example

```
HairEyeColor[,,1]
```

```
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
```

```
chisq.test(HairEyeColor[,,1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 1]
## X-squared = 41.28, df = 9, p-value = 4.447e-06
```

What can we conclude here?

Eye Color and Hair Color is dependent at the 5% significance level.

Required Assumptions

-Each cell has at least 5 observations

-More than 2 degrees of freedom

-Independent observations

Chi-Square test for Goodness of Fit

Goodness of Fit

What if instead we have one categorical random variable?

For example, in the original example about “first pet”, we might think 30% of people had a dog, 30% had a cat, 30% had other, and 10% had none.

Can we test this?

Goodness of Fit

What if instead we have one categorical random variable?

For example, in the original example about “first pet”, we might think 30% of people had a dog, 30% had a cat, 30% had other, and 10% had none.

Can we test this?

You bet we can!

Goodness of Fit

If we have a distribution in mind, we can get the expected number of observations in a category by

$E_i = n * P($ in category i $)$.

Then we can calculate our test statistic as:

$$X^2 = \sum_{i=1}^{r} \frac{(n_i - E_i)^2}{E_i} \sim \chi^2_{r-1}$$

We're still interested in if our observed count is close to the expected count.

(Why $r - 1$ degrees of freedom?)

Goodness of Fit Example

According to the 2000 census, the education level of adult residents in New York breaks down as follows:

No HS diploma	HS grad	Some college	Assoc or BA	Grad degree
20.8%	27.8%	16.8%	22.8%	11.8%

A state grand jury appoints 64 persons. This jury has the following composition

No HS diploma	HS grad	Some college	Assoc or BA	Grad degree
3	10	16	20	15

Is the education level of the state grand jury representative of that in the general population?

Goodness of Fit Example in R

```
table.ed
```

```
## Education
##   Assoc or BA           Grad       HS Grad No HS Diploma  Some Collage
##            20             15            10             3            16
```

```
chisq.test(table.ed, p = c(0.228, 0.118, 0.278, 0.208, 0.168))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table.ed
## X-squared = 23.312, df = 4, p-value = 0.0001097
```

Requirements to run this test are the same as for independence testing!