MATH 3339
Statistics for the Sciences
Chapter 12: Analysis of Categorical Data

Wendy Wang
wwang60@central.uh.edu

Lecture 19 - 3339

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outline

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Mars Inc. claims that they produce M&Ms with the following distributions:

expected values

Brown	30%	Red	20%	Yellow	20%
Orange	10%	Green	10%	Blue	10%

A bag of M&Ms was randomly selected from the grocery store shelf, and the color counts were:

Brown	14	Red	14	Yellow	5
Orange	7	Green	6	Blue	10

We want to know if the distribution of color the same as the manufacturer's claim.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Goodness-of-fit Test

- This is a test to see how well on sample proportions of categories "match-up" with the known population proportions.
- The Chi-square goodness-of-fit test extends inference on proportions to more than two proportions by enabling us to determine if a particular population distribution has changed from a specified form.
- Hypotheses:
  - H_0: The proportions are the same as what is claimed.
  - H_a: At least one proportion is different than what is claimed.

  This would be better in context of the problem. For example in our M&Ms test;
  - H_0: The distribution of candy colors is as the manufacturer claims.
  - H_a: The distribution of candy colors is not what the manufacturer claims.

Chi-Square Test

Test Statistic: Called the **chi-square statistic** is a measure of how much the observed cell counts diverge from the expected cell counts. To calculate for each problem you will make a table with the following headings:

Sample

Observed Counts (O)	Expected Counts (E)	$\frac{(O-E)^2}{E}$

The sum of the third column is called the Chi-square test statistic, χ^2.

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \sim \chi^2(df)$$

Where expected counts = total count $\times$ proportion of each category.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Chi-square of M&Ms

Color	Observed Counts (O)	Proportions	Expected Counts (E)	$\frac{(O-E)^2}{E}$
Brown	14	0.3	56 x 0.3 = 16.8	$\frac{(14-16.8)^2}{16.8}$
Red	14	0.2	56 x 0.2 = 11.2	$\frac{(14-11.2)^2}{11.2}$
Yellow	5	0.2	11.2	
Orange	7	0.1	5.6	
Green	6	0.1	5.6	
Blue	10	0.1	5.6	

sum: 14+14+...
= 56

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Chi-square

χ^2-test-statistic $= \sum \frac{(O-E)^2}{E} = 8.4345$

- Chi-square distributions have only positive values and are skewed right.
- This has a degrees of freedom which is $n-1$.
- As the degrees of freedom increases it become more like a Normal distribution.
- The total area under the χ^2 curve is 1.
- To find area under the curve
  - Table provided
  - In R: 1 - pchisq(x,df)

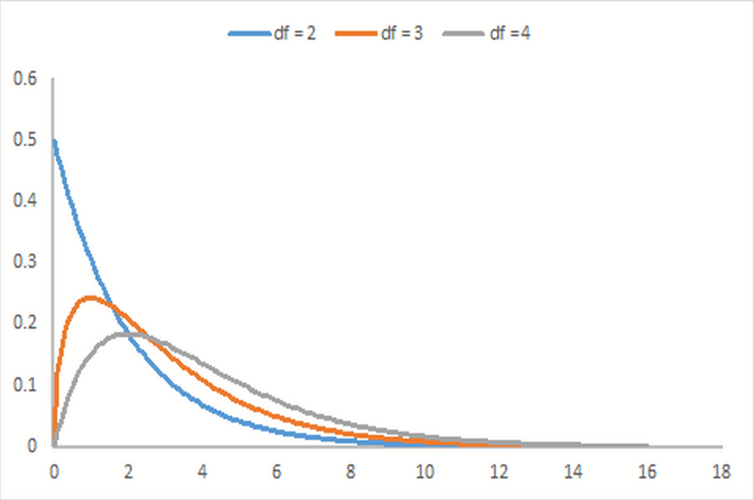
P-value $= P(\chi^2 = 8.4345)$

$= 1 - \text{pchisq}(8.4345, 5)$

$= 0.1339$

df = category - 1

Chi-Square



Assumptions for a Chi-Square Goodness-of-fit Test

1. The sample must be an SRS from the populations of interest.
2. The population size is at least 10 times the size of the sample.
3. All expected cell counts must be at least 5.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Is the manufacturers claim correct?

P-value = 0.1339 → F to regect H_0.

There is no evidiena that the diff. of colors of M&Ms is different from what the manufeturer claims.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Using R

- chisq.test(c(list of observed values),correct = FALSE, p = c(list of proportions))
- If we are not given a list of proportions then p = 1/n and that is a default for R so we do not need to give that information.

```
> chisq.test(c(14,14,5,7,6,10),p=c(.3,.2,.2,
.1,.1,.1))

Chi-squared test for given probabilities

data:  c(14, 14, 5, 7, 6, 10)
X-squared = 8.4345, df = 5, p-value = 0.1339
```

• list sample to proportion.
• order matters

Zodiac Signs

Does your zodiac sign determine how successful you will be in later life? *Fortune* magazine collected the zodiac signs of 256 heads of the largest 400 companies. The following are the number of births for each sign:

Sign	Births
Aries	23
Taurus	20
Gemini	18
Cancer	23
Leo	20
Virgo	19
Libra	18
Scorpio	21
Sagittarius	19
Capricorn	22
Aquarius	24
Pisces	29

$H_0: p = \frac{1}{12}$ for all zodiac signs

H_a: at least one proportion is different from the others.

"O"	"E"	$\frac{(O-E)^2}{E}$
	$\frac{256}{12}$	

chisq.test(c(23,20,18,...))

$\chi^2 = 5.0938$, df = 11,
pvalue = 0.9265

pvalue = 1 − pchisq(5.0938, 11)
0.9265 → FTR

From: *Intro Stats*, De Veaux, Velleman, Bock. 2nd Edition, Pearson, pg 604.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

The following table shows three different airlines **row variable** and the number of delayed or on-time flights **column variable** from lightstats.com.

	Delayed	On-time	Total
American	112	843	955
Southwest	114	1416	1530
United	61	896	957
Total	287	3155	3442

- Does on-time performance depend on airline?
- We will use a significance test to answer this question.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Significance Tests For Two-Way Tables

1. The assumptions necessary for the test to be valid are:
   a. The observations constitutes a simple random sample from the population of interest, and
   b. The expected counts are at least 5 for each cell of the table.

2. Hypotheses
   - Null hypothesis: There is no association (independence) between the row variable and column variable.
   - Alternative hypothesis: There is an association (dependence) between the row variable and column variable.
   - In the previous example:

$$H_0 : \text{Airline and on-time performance are independent.}$$
$$H_A : \text{On-time performance depends on airline.}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic: Called the **chi-square statistic** is a measure of how much the observed cell counts in a two-way table diverge from the expected cell counts. To calculate.

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Where "observed" represents an observed sample count, and "expected" is calculated by

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

The sum is over all $r \times c$ cells in the table. Where r is the number of rows and c is the number of columns.

Significance Tests For Two-Way Tables

If H_0 is true, the chi-square statistic X^2 has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. Where r = number of rows and c = number of columns.

4. The P-value for the chi-square test is $P(\chi^2 \geq X^2)$. Given that all of the expected cell counts be 5 or more.
5. Decision: If P-value is less than α level of significance, we reject H_0. Otherwise we fail to reject H_0.
6. Conclusion: In context of the problem.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

The following table shows three different airlines **row variable** and the number of delayed or on-time flights **column variable** from flightstats.com.

	Delayed	On-time	Total
American	112	843	955
Southwest	114	1416	1530
United	61	896	957
Total	287	3155	3442

Does on-time performance depend on airline?

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times 287}{3442}=79.6296$	$\frac{955\times 3155}{3442}=875.3704$	955
Southwest	$\frac{1530\times 287}{3442}=127.5741$	$\frac{1530\times 3155}{3442}=1402.4259$	1530
United	$\frac{957\times 287}{3442}=79.7963$	$\frac{957\times 3155}{3442}=877.20367$	957
Total	287	3155	3442

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times287}{3442}=79.6296$	$\frac{955\times3155}{3442}=875.3704$	955
Southwest	$\frac{1530\times287}{3442}=127.5741$	$\frac{1530\times3155}{3442}=1402.4259$	1530
United	$\frac{957\times287}{3442}=79.7963$	$\frac{957\times3155}{3442}=877.20367$	957
Total	287	3155	3442

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times287}{3442} = 79.6296$	$\frac{955\times3155}{3442} = 875.3704$	955
Southwest	$\frac{1530\times287}{3442} = 127.5741$	$\frac{1530\times3155}{3442} = 1402.4259$	1530
United	$\frac{957\times287}{3442} = 79.7963$	$\frac{957\times3155}{3442} = 877.20367$	957
Total	287	3155	3442

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times 287}{3442} = 79.6296$	$\frac{955\times 3155}{3442} = 875.3704$	955
Southwest	$\frac{1530\times 287}{3442} = 127.5741$	$\frac{1530\times 3155}{3442} = 1402.4259$	1530
United	$\frac{957\times 287}{3442} = 79.7963$	$\frac{957\times 3155}{3442} = 877.20367$	957
Total	287	3155	3442

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times 287}{3442} = 79.6296$	$\frac{955\times 3155}{3442} = 875.3704$	955
Southwest	$\frac{1530\times 287}{3442} = 127.5741$	$\frac{1530\times 3155}{3442} = 1402.4259$	1530
United	$\frac{957\times 287}{3442} = 79.7963$	$\frac{957\times 3155}{3442} = 877.20367$	957
Total	287	3155	3442

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Expected cell counts

The following table gives the expected cell count.

	Delayed	On-time	Total
American	$\frac{955\times287}{3442}=79.6296$	$\frac{955\times3155}{3442}=875.3704$	955
Southwest	$\frac{1530\times287}{3442}=127.5741$	$\frac{1530\times3155}{3442}=1402.4259$	1530
United	$\frac{957\times287}{3442}=79.7963$	$\frac{957\times3155}{3442}=877.20367$	957
Total	287	3155	3442

Significance Test of Two-Way Table Example

1. Assumptions: SRS, All of the expected cell counts are greater than 5.
2. Hypothesis:

$$\begin{aligned} H_0 &: \text{Airline and on-time performance are independent.} \\ H_A &: \text{On-time performance depends on airline.} \end{aligned}$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

3. Test Statistic

The following table gives us the chi-square contribution for each cell, $\frac{(O-E)^2}{E}$.

	Delayed	On-time
American	$\frac{(112-79.6296)^2}{79.6296} = 13.159$	$\frac{(843-875.3704)^2}{875.3704} = 1.197$
Southwest	$\frac{(114-127.5741)^2}{127.5741} = 1.4443$	$\frac{(1416-1402.4259)^2}{1402.4259} = 0.1314$
United	$\frac{(61-79.7963)^2}{79.7963} = 4.428$	$\frac{(896-877.20367)^2}{877.20367} = 0.4028$

Test statistic:

$$X^2 = 13.159 + 1.197 + 1.4443 + 0.1314 + 4.428 + 0.4028 = 20.7625$$

- The P-value for the chi-square test is $P(\chi^2 \geq X^2)$. With $df = (r-1)(c-1)$ where r = # of rows and c = # of columns.

- In our airline example r = 3, c = 2, df = (3 - 1)(2 -1) = 2.

- For our airline example, P-value = $P(\chi^2 \geq 20.7625) = 1 - pchisq(\underline{20.7625}, 2) = 0.000031$

test statistic

df = (r-1)×(c-1)
(3-1)×(2-1)

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

5. Decision

- **Reject** H_0 if the P-value $\leq \alpha$.
- **Fail** to reject H_0 if the $P - \text{value} > \alpha$.
- In our airplane example, $P - \text{value} < 0.0001$ so we **reject** the null hypothesis.

6. Conclusion

- If H_0 is **rejected** then there is a dependence between the row variable and the column variable.
- If H_0 is not rejected then there is no association.
- In our airplane example, we **reject** the null hypothesis. Thus we conclude that on-time status **depends** on airline.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Chi-square Test Using R

1 variable test on final

1. Input the data as a matrix.
2. R-code: chisq.test(matrix name,correction=FALSE)

```
> airline<-matrix(c(112,114,61,843,1416,896),nrow=3,ncol=2)
> chisq.test(airline,correct = FALSE)

Pearson's Chi-squared test

data:  airline
X-squared = 20.762, df = 2, p-value =3.102e-05
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Understanding Dependence

- By itself, the chi-square test determines only whether the data provide evidence of a relationship between the two variables. If the result is significant, one can go on to identify the source of that relationship by finding the cells of the table that contribute most to the χ^2 value (i.e. those cells with the biggest discrepancy between the observed and expected counts) and by noting whether the observed count falls above or below the observed count in those cells.
- To get these "Chi-square contribution" values in R use `residuals(chisq.test(matrix,correction=FALSE))^2`.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Eating Out

A survey was conducted in five countries. The following table is based on 1,000 respondents in each country that said they eat out once a week or more (yes) or not (no). $c=5$

$df=(r-1)*(c-1)=4$

$r=2$

	Country				
Eat out	Germany	France	UK	Greece	US
Yes	100	120	280	390	570
No	900	880	720	610	430

At the 0.05 level of significance, determine whether there is a significant difference in the proportion of people who eat out at least once a week in the various countries.

R Output

$\sum \frac{(O-E)^2}{E}$

```
> eat<-matrix(c(100,900,120,880,280,720,390,610,570,430),nrow = 2
,ncol = 5)
> eat
     [,1] [,2] [,3] [,4] [,5]
[1,]  100  120  280  390  570
[2,]  900  880  720  610  430
> chisq.test(eat,correct = FALSE)

Pearson's Chi-squared test

data:  eat
X-squared = 742.4, df = 4, p-value < 2.2e-16

> residuals(chisq.test(eat,correct = FALSE))^2
         [,1]      [,2]      [,3]     [,4]     [,5]
[1,] 126.2466 101.31507 0.4931507 32.89041 264.6712
[2,]  52.0678  41.78531 0.2033898 13.56497 109.1582
```

5%

$\times 10^{-16}$

R H_0

Reject the null hypothesis

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS