STAT011 Statistical Methods I

Lecture 20 Chi-squared Test

Lu Chen
Swarthmore College
4/9/2019

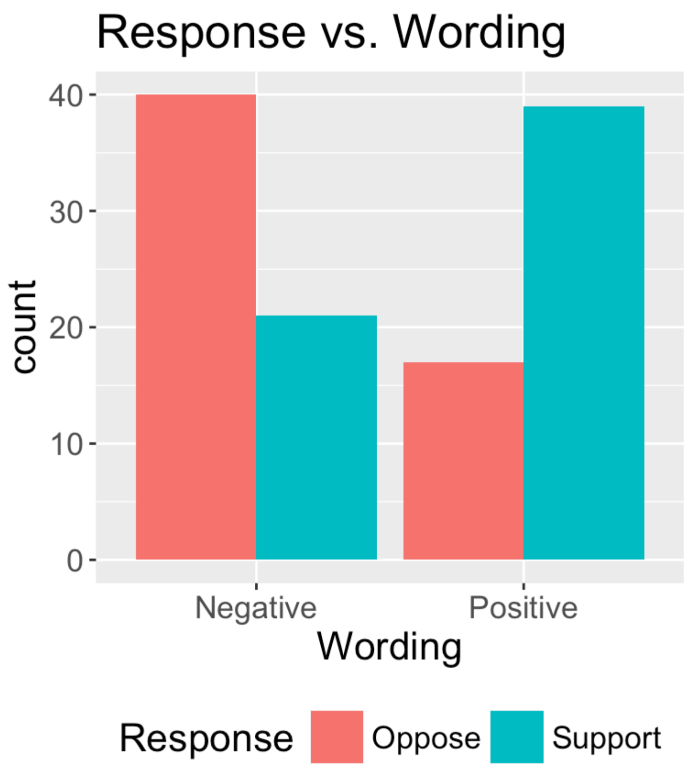
Review

- Inference for a population porportion
  - Large sample C.I. for a population proportion
    $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - Large sample significance test for a population proportion
    $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{approx.}{\sim} N(0, 1)$
- Inference for a difference in two proportions
  - Large sample C.I. for a difference in proportions
    $\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
  - Large sample significance test for a difference in proportions
    $z = \frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \overset{approx.}{\sim} N(0, 1)$ where $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$

Outline

- Chi-square test
  - Null and alternative hypotheses
  - Test statistic
  - Chi-square distribution
  - P-value and conclusion
- Using chi-square test
  - Example 1: *Response* vs. *Wording* (2×2 table)
  - Example 2: 376 polls of 2016 presidential election (2×3 table)
- Chi-square goodness-of-fit test
  - Example 1: STAT 11 2019 *ClassYear*
  - Example 2: Breast cancer missing data

Example 1 - *Response* vs. *Wording*



	Negative	Positive	Total
Oppose	40	17	**57**
Support	21	39	**60**
Total	**61**	**56**	**117**

- Does wording affect students' responses?
- Denote p_1 and p_2 as the proportion of support for the negative and positive group.
- $n_1 = 61, n_2 = 56$
- $\hat{p}_1 = \frac{21}{61} = 0.344, \hat{p}_2 = \frac{39}{56} = 0.696$

Example 1 - *Response* vs. *Wording*

$$n_1 = 61, n_2 = 56, \hat{p}_1 = \frac{21}{61} = 0.344, \hat{p}_2 = \frac{39}{56} = 0.696$$

- **95% C.I.** $\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
  $= (0.344 - 0.696) \pm 1.96 \times \sqrt{\frac{0.344 \times 0.656}{61} + \frac{0.696 \times 0.304}{56}} = -0.352 \pm 0.169$
- We are 95% confident that the interval $[-0.521, -0.183]$ will contain the true difference in population proportions.
- **0.05 Test** $H_0 : p_1 = p_2$ vs. $H_a : p_1 \neq p_2$; $\hat{p} = \frac{21+39}{61+56} = 0.513$
  $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.344 - 0.696 - 0}{\sqrt{0.513 \times 0.487 \times (1/61 + 1/56)}} = -3.81$
  $P = 2P(Z \geq |-3.81|) = 0.00014 < 0.05$
- We reject H_0 at level 0.05. There is a significant difference in proportion of support between the two wording groups. Wording significantly affects students' responses.

Another approach to the problem

	Negative	Positive	Total
Oppose	n_{11}	n_{12}	r_1
Support	n_{21}	n_{22}	r_2
Total	c_1	c_2	n

- **Does *Wording* affect *Response*?** If *Wording* does NOT affect *Response*, what would happen?
- $p_1 = p_2 = p$, proportion of support is the same for both groups and overall
$$p = \frac{n_{21}+n_{22}}{c_1+c_2} = \frac{r_2}{n}$$
- We expect number of supporters in the negative and positive group will be $c_1 p = \frac{r_2 c_1}{n}$ and $c_2 p = \frac{r_2 c_2}{n}$, i.e. numbers of students in the negative and positive group multiplying the proportion of support.

Another approach to the problem

Observed counts vs. Expected counts

	Negative	Positive	Total
Oppose	n_{11} vs. $\frac{r_1 c_1}{n}$	n_{12} vs. $\frac{r_1 c_2}{n}$	r_1
Support	n_{21} vs. $\frac{r_2 c_1}{n}$	n_{22} vs. $\frac{r_2 c_2}{n}$	r_2
Total	c_1	c_2	n

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{n}$$

Expected counts are computed by assuming **no association** between *Response* and *Wording*.

Another approach to the problem

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{n}$$

Observed

	Negative	Positive	Total
Oppose	40	17	**57**
Support	21	39	**60**
Total	**61**	**56**	**117**

Expected (no association)

	Negative	Positive	Total
Oppose	$29.7 = \frac{57\times61}{117}$	$27.3 = \frac{57\times56}{117}$	**57**
Support	$31.3 = \frac{60\times61}{117}$	$28.7 = \frac{60\times56}{117}$	**60**
Total	**61**	**56**	**117**

The difference between the observed counts and the expected counts measures the strength of the association between the two categorical variables.

Chi-square Test - Hypotheses

H_0 : There is no association between the two categorical variables.

H_a : The two categorical variables are associated.

- Under H_0,

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{n}$$

- If H_0 is true, the observed counts should be close to the expected counts.
- If H_a is true, the observed counts should be very different from the expected counts.
- The difference is measured by $O - E$, where O represents the observed counts and E for the expected counts.

Chi-square Test - Test statistic

For a $r \times c$ two-way table (r is numbers of rows; c is number of columns), 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. The formula for the statistic is

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

where "O" represents the observed cell count, "E" represents the expected count for the same cell, and the sum is over all $r \times c$ cells in the table.

- The chi-square statistic follows a **chi-square distribution** with $(r-1)(c-1)$ degrees of freedom.

$$X^2 \sim \chi^2_{(r-1)(c-1)}$$

Chi-square Test - Test statistic

Observed

	Negative	Positive	Total
Oppose	40	17	**57**
Support	21	39	**60**
Total	**61**	**56**	**117**

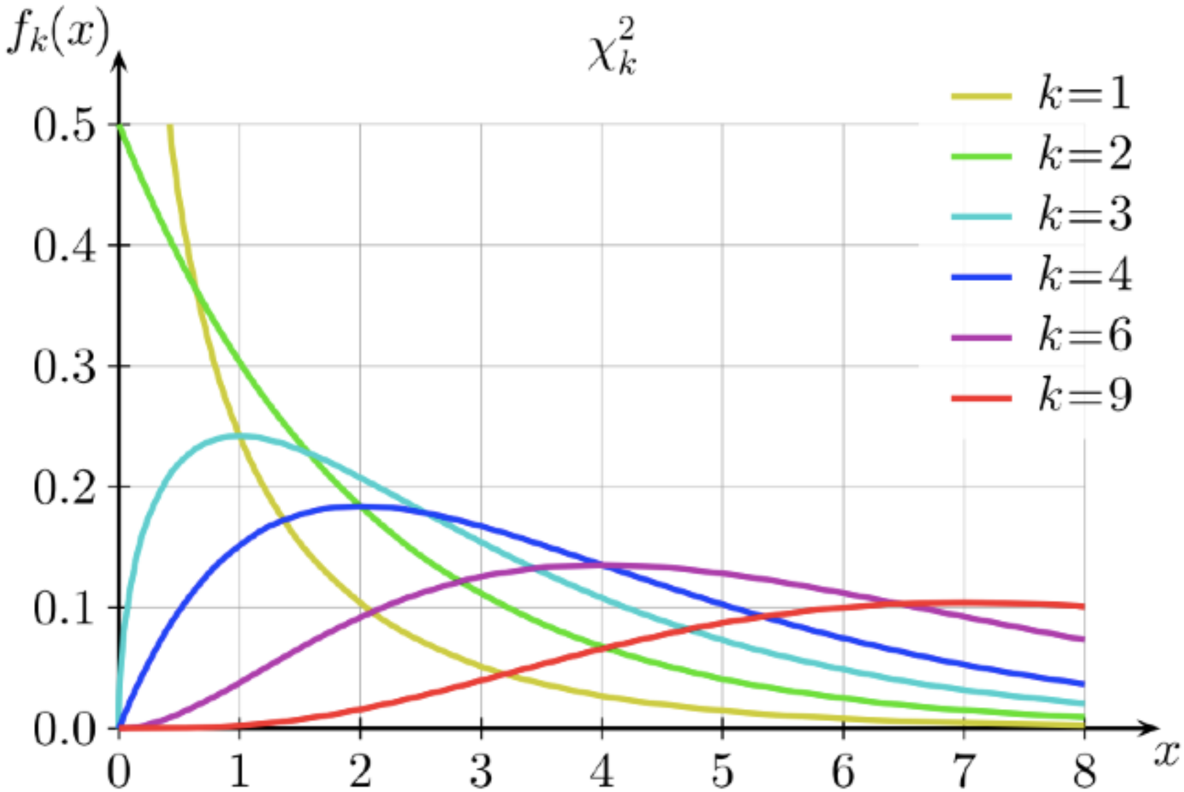
Expected (no association)

	Negative	Positive	Total
Oppose	29.7	27.3	**57**
Support	31.3	28.7	**60**
Total	**61**	**56**	**117**

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(40-29.7)^2}{29.7} + \frac{(17-27.3)^2}{27.3} + \frac{(21-31.3)^2}{31.3} + \frac{(39-28.7)^2}{28.7} = 14.5$$

where $X^2 \sim \chi_1^2$ since $(r-1)(c-1) = (2-1)(2-1) = 1$

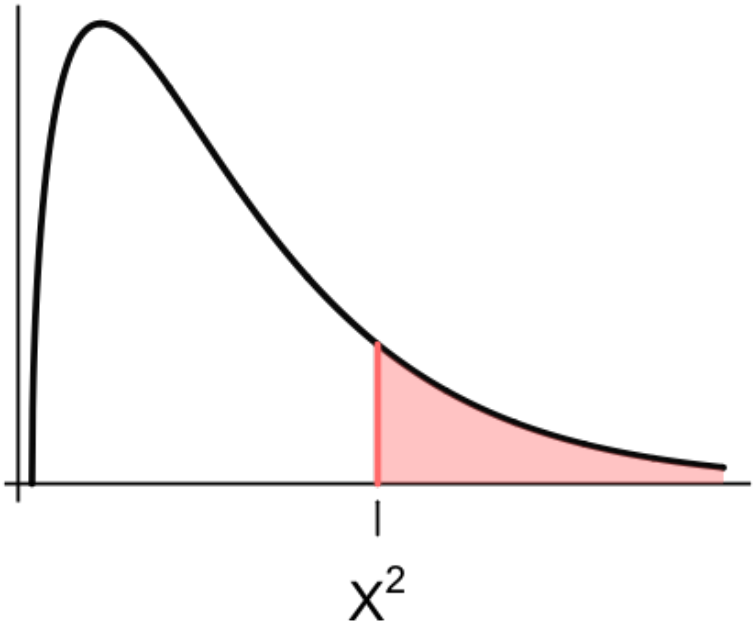
Chi-square Test - Chi-square distribution



$X^2 \sim \chi^2_k = \chi^2_{(r-1)(c-1)}$,

- A 2×2 table has $k = (2-1)(2-1) = 1$.
- A 2×3 table has $k = (2-1)(3-1) = 2$.
- A 4×5 table has $k = (4-1)(5-1) = 12$.

Chi-square Test - *P*-value



- $X^2 = \sum \frac{(O-E)^2}{E}$ is always positive.
- The larger X^2 value, the stronger evidence against H_0
  - Large X^2: large difference between the observed and expected counts - reject H_0.
  - Small X^2: the observed are close to the expected - cannot reject H_0.
- P-value $= P(\chi^2 \geq X^2)$

```
1-pchisq(x2, df = )
```

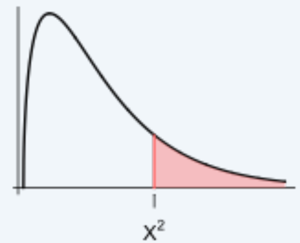
- Chi-square test is a one-sided test.

The chi-square test for two-way tables

H_0: there is no association between the row and column variables in a two-way table.
H_a: these variables are related.

If H_0 is true, the chi-square statistic $X^2 = \sum \frac{(O-E)^2}{E}$ has approximately a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. The P-value for the chi-square test is

$$P(\chi^2 \geq X^2)$$



- For 2×2 tables, we require all four expected cell counts to be 5 or more.
- For tables larger than 2×2, we will use this approximation whenever the average of the expected counts is 5 or more and the smallest expected count is 1 or more.

Example 1 - *Response* vs. *Wording*

Observed

	Negative	Positive	Total
Oppose	40	17	**57**
Support	21	39	**60**
Total	**61**	**56**	**117**

Expected (no association)

	Negative	Positive	Total
Oppose	29.7	27.3	**57**
Support	31.3	28.7	**60**
Total	**61**	**56**	**117**

H_0 : there is no association between *Response* and *Wording*.

H_a : there is an association between the two variables.

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(40-29.7)^2}{29.7} + \frac{(17-27.3)^2}{27.3} + \frac{(21-31.3)^2}{31.3} + \frac{(39-28.7)^2}{28.7} = 14.5 \sim \chi_1^2$$

$P = P(\chi^2 \geq 14.5) = 0.00014 < 0.05$(`1-pchisq(14.5, df = 1)`). We reject H_0 at level 0.05 and conclude that the wording of questions is significantly associated with the response.

Notes

- The z test for the difference in two proportions of support has $z = -3.81$ and $P = 0.00014$
- The chi-square test for the association between wording and response has $X^2 = 14.5$ and $P = 0.00014$
- In fact, $X^2 = 14.5 = z^2 = (-3.81)^2$.
- For a 2×2 table, $k = 1$, $X^2 \sim \chi^2$ is the square of $Z \sim N(0, 1)$.
- The chi-square test of a 2×2 table is equivalent to a two-sided z test for the difference in two proportions.
- But chi-square test does not tell us the direction of the association. z test does.

Example 1 - *Response* vs. *Wording*

Chi-square test in R

```r
table.RW <- table(Survey$Response, Survey$Wording)
table.RW
```

```
##
##            Negative Positive
##   Oppose         40       17
##   Support        21       39
```

```r
chisq.test(table.RW, correct = F) # Without Yates' continuity correction
```

```
##
##  Pearson's Chi-squared test
##
## data:  table.RW
## X-squared = 14.493, df = 1, p-value = 0.0001406
```

- The `chisq.test()` applies the Yates' continuity correction on 2×2 tables by default. For larger tables, no correction is applied

Example 1 - *Response* vs. *Wording* by *Gender*

```
F.RW<-table(Survey$Response[Survey$Gender=="F"],Survey$Wording[Survey$Gender=="F"])
M.RW<-table(Survey$Response[Survey$Gender=="M"],Survey$Wording[Survey$Gender=="M"])
F.RW # Female students only
```

```
##
##           Negative Positive
##   Oppose        25        7
##   Support       10       21
```

```
M.RW # Male students only
```

```
##
##           Negative Positive
##   Oppose        14       10
##   Support       11       18
```

Example 1 - *Response* vs. *Wording* by *Gender*

```r
chisq.test(F.RW, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  F.RW
## X-squared = 13.416, df = 1, p-value = 0.0002495
```

```r
chisq.test(M.RW, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  M.RW
## X-squared = 2.1935, df = 1, p-value = 0.1386
```

- *Wording* significantly affects the response of female students in STAT011 2016 but not male students at level 0.05.
- Note: this does not mean that *Wording* has NO effect on male students but that the effect is not statistically significant.

Example 2 - 376 polls

```r
table.polls
```

```
##                    Method
## Result              Online Live Phone I.V.R.
##   Clinton led          188        126     27
##   Trump led or even      8         14     13
```

```r
chisq.test(table.polls)
```

```
## Warning in chisq.test(table.polls): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  table.polls
## X-squared = 31.906, df = 2, p-value = 1.179e-07
```

- Method is significantly associated with the results of election polls at level 0.05.
- R shows this warning message when any of the expected counts is less than 5.

Example 2 - 376 polls

Check the expected counts of a two-way table.

```r
table.polls
```

```
##                          Method
## Result                    Online Live Phone I.V.R.
##   Clinton led                188        126     27
##   Trump led or even            8         14     13
```

```r
chisq.test(table.polls)$expected
```

```
##                          Method
## Result                        Online Live Phone    I.V.R.
##   Clinton led             177.75532  126.96809 36.276596
##   Trump led or even        18.24468   13.03191  3.723404
```

- The last cell has expected count less than 5.

Chi-square goodness-of-fit test

- For a categorical variable with more than two categories, we may be interested in testing whether the proportions of the several categories are the same as some specific proportions.
- For example, in STAT 11 2016, are the numbers of freshmen, sophomores, junior and seniors the same?

Class Year	Fr	So	Jr	Sr
Counts	31	36	13	14
Proportions	0.33	0.38	0.14	0.15

Here we want to test whether the observed proportions are the same as 0.25, 0.25, 0.25 and 0.25 for the four categories.

Chi-square goodness-of-fit test

Data for n observations of a categorical variable with k possible outcomes are summarized as observed counts, $n_1, n_2, \cdots, n_k$, in k cells. The null hypothesis specifies proportions $p_1, p_2, \cdots, p_k$ for the possible outcomes. The alternative hypothesis says that the true proportions of the possible outcomes are not the proportions specified in the null hypothesis.

For each cell, multiply the total number of observations n by the specified proportions to determine the expected counts: Expected count $= np_i$. The chi-square statistic measures how much the observed cell counts differ from the expected cell counts. The formula for the statistic is

$$X^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{k-1}$$

The degrees of freedom are $k-1$, and P-values are computed from the chi-square distribution. Use this procedure when the expected counts are all 5 or more.

Chi-square goodness-of-fit test

1. Specify the null hypothesized proportions $p_1, p_2, \cdots, p_k$.
2. Calculate the expected counts under H_0

$$\text{Expected count} = np_i$$

   where $i = 1, 2, \cdots, k$.

3. Compare the observed counts to the expected counts and compute the chi-square test statistic

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

4. Calculate the P-value for $X^2 \sim \chi^2_{k-1}$ and state the conclusion.

Chi-square goodness-of-fit test

Class Year	Fr	So	Jr	Sr
Counts	31	36	13	14
Proportions	0.33	0.38	0.14	0.15

1. $H_0 : p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p4 = 0.25$
   H_a : the true proportions are not the proportions specified in H_0
2. Expected counts are $E_1 = E_2 = E_3 = E_4 = 94 \times 0.25 = 23.5$
3. The chi-square statistic is

$$X^2 = \frac{(31-23.5)^2}{23.5} + \frac{(36-23.5)^2}{23.5} + \frac{(13-23.5)^2}{23.5} + \frac{(14-23.5)^2}{23.5} = 17.57$$

4. $P = P(\chi^2 \geq 17.57) = 0.0005 < 0.05$`1-pchisq(17.57, df = 3)`. We reject H_0 at level 0.05. The true proportions are not 0.25 for the four class years.

Example - Breast density

The distribution of breast density values in the female population is

Breast density value	0	0~0.25	0.25~0.50	0.50~0.75	0.75~1
Proportions in population	0.139	0.365	0.285	0.181	0.030

In a large study for breast cancer, about half of the subjects have missing data in breast density. A new statistical method was developed to estimate the missing breast density values for the 1616 subjects. The new method resulted in

Breast density value	0	0~0.25	0.25~0.50	0.50~0.75	0.75~1
Counts based on the new method	208	625	478	259	46

Does the method work well in estimating the missing values?

Example - Breast density

Breast density value	0	0~0.25	0.25~0.50	0.50~0.75	0.75~1
Observed	208	625	478	259	46
Expected	224.6	589.8	460.6	292.5	48.5

```r
observed <- c(208, 625, 478, 259, 46)
p0 <- c(0.139, 0.365, 0.285, 0.181, 0.030)
chisq.test(x=observed, p=p0)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 7.9493, df = 4, p-value = 0.09345
```

Example - Breast density

$X^2 = 7.9, df = 4$ and $P = 0.09 > 0.05$.

- We cannot reject H_0 at level 0.05. The true proportions from the new method are not significantly different from the known population proportions.
- The new method fits the population distribution quite well.
- This is why the test is called goodness-of-fit test. Sometimes we do not want to reject H_0 in goodness-of-fit tests.

Summary

- Chi-square test
  - Null and alternative hypotheses
  - Test statistic $X^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{(r-1)(c-1)}$ `chisq.test(table, correct=F)`
  - Chi-square distribution
  - P-value and conclusion
- Using chi-square test
  - Example 1: *Response* vs. *Wording* (2×2 table)
  - Example 2: 376 polls of 2016 presidential election (2×3 table)
- Chi-square goodness-of-fit test $X^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2_{k-1}$ `chisq.test(x = observed counts, p = null proportions)`
  - Example 1: STAT 11 2019 *ClassYear*
  - Example 2: Breast cancer missing data