

PBIO 504

Chi-Square Tests

Contingency Tables for Analysis of Categorical Data

- When working with categorical variables (such as gender) or continuous data grouped into categories (such as BMI categorized as underweight, normal, overweight, obese), we first arrange the data in tables.
- **Contingency tables** tabulate categorical data describing the joint distribution of a row variable with a column variable (e.g. gender by race)

2x2 Table

	Caucasian	African-American	<i>row totals</i>
Female	A	B	$A+B$
Male	C	D	$C+D$
<i>column totals</i>	$A+C$	$B+D$	$A+B+C+D$ ($=N$)

If there are 2 rows and 2 columns, we call it a 2 by 2 table; for 2 rows and 3 columns it is a 2 by 3 table, etc...

Example: Head Injury and Helmets

Do bicycle safety helmets prevent injury?	Wearing a helmet	Not wearing a helmet	<i>totals</i>
Head injury	17	218	235
No head injury	130	428	558
<i>totals</i>	147	646	793

Hypothesis Testing in the Example

- To examine the effectiveness of wearing a helmet to protect against a head injury while bicycle riding, we need to decide on a null hypothesis and its alternative.
- H_0 : *the proportion of riders suffering head injury among those who wore helmets is identical to the proportion of riders with a head injury among those who did not wear a helmet.*
- H_A : *the proportions of riders suffering head injuries are different for these two groups.*
- The chi-square test is useful for testing these hypotheses, since it will evaluate the difference between the observed and the expected distributions of counts.

How Do We Figure Out the Expected Distribution?

- We will assume that, under the null hypothesis, the proportions of head injuries will be the same for users and nonusers of helmets, so we can just treat the study as one single population.
- We will calculate the overall rate of injury, and then apply that rate to the users and nonusers.
- See the next slide

Back to the Example: Expected Counts

Do bicycle safety helmets prevent injury?	Wearing a helmet	Not wearing a helmet	<i>totals</i>
Head injury	17	218	235
No head injury	130	428	558
<i>totals</i>	147	646	793

- Sample prevalence of injury is $235/793 = 29.6\%$
- So, of the **147 people wearing helmets**, the expected injury count = $147 * 0.296 = 44$ persons, and the rest $(147 - 44) = 103$ are expected to be injury free.
- In the **group who did not wear helmets**, the expected counts are $646 * 0.296 = 191$ injured and $(646 - 191) = 455$ not injured.

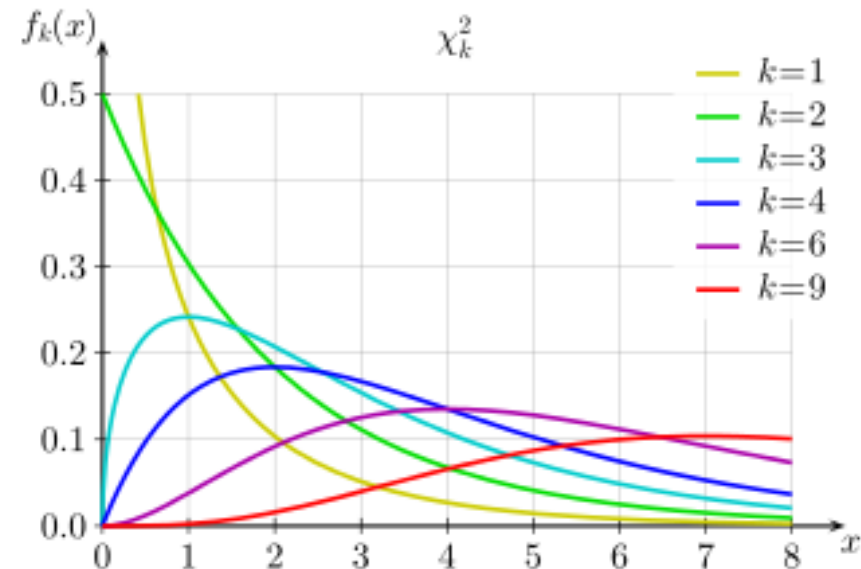
Observed (O) and Expected (E) Counts

Do bicycle safety helmets prevent injury?	Wearing a helmet		Not wearing a helmet		<i>Totals*</i>
	O	E	O	E	
Head injury	17	44	218	191	235
No head injury	130	103	428	455	558
<i>Totals*</i>	147		646		793

* Note that the totals are the same, whether we add the observed counts or the expected counts, since we derived all of them from the same total population of 793.

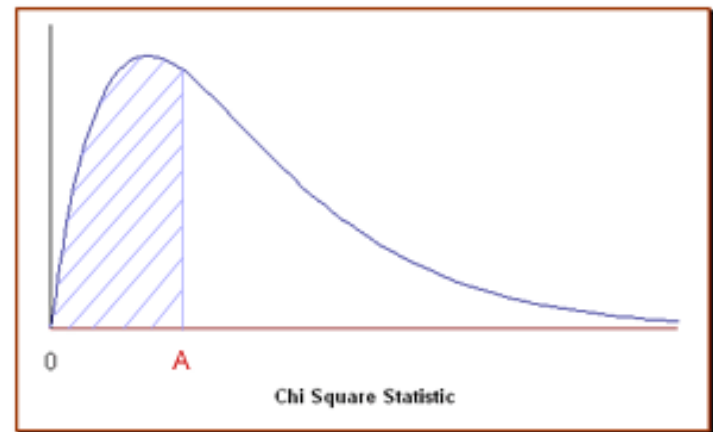
The Chi-Square Distribution

- The statistic $X^2 = [(n - 1) * s^2] / \sigma^2$ follows a chi-square distribution.
- The mean of the distribution is equal to the number of degrees of freedom: $\mu = \nu = n - 1$.
- The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2 * \nu$
- The chi-square curve approaches a normal distribution as the degrees of freedom increase



The Chi-Square Test

- Chi-Square is the sum of the squared differences between the observed and expected counts, divided by the expected counts.
- It ranges from zero to infinity.
- $X^2 = \sum (O-E)^2 / E$



If its p-value < 0.05 then reject the null hypothesis.

Solving the Head Injury Example

$$\begin{aligned}X^2 &= |17-44|^2/44 + |130-103|^2/103 + \\&\quad |218-191|^2/191 + |428-455|^2/455 \\&= 16.5+7.1+3.8+1.6 = 29 \\X^2 &= 29\end{aligned}$$

$$df=(r-1)*(c-1)=(2-1)*(2-1)=1, \text{ p-value} = 0.0000001$$

What can we conclude about wearing helmets and head injury prevention?

- Conclude the alternative that wearing helmet prevents injuries

STATA – summary, tables, tests -----

tabi - Two-way tables

Main Advanced

User-supplied cell frequencies: (space separated with "\" for new rows)

17 218\130 428

Test statistics

- ☒ Pearson's chi-squared
- ☐ Fisher's exact test
- ☐ Goodman and Kruskal's gamma
- ☐ Likelihood-ratio chi-squared
- ☐ Kendall's tau-b
- ☐ Cramer's V

Cell contents

- ☐ Pearson's chi-squared
- ☐ Within-column relative frequencies
- ☐ Within-row relative frequencies
- ☐ Likelihood-ratio chi-squared
- ☐ Relative frequencies
- ☒ Expected frequencies
- ☐ Suppress frequencies

☐ Treat missing values like other values

☐ Do not wrap wide tables

☐ Show cell contents key

☐ Suppress value labels

☐ Suppress enumeration log

OK Cancel Submit

```
. tabi 17 218\130 428, chi2
```

row	col		Total
	1	2	
1	17	218	235
2	130	428	558
Total	147	646	793

```
Pearson chi2(1) = 28.2555 Pr = 0.000
```

Head Injury and Wearing a Helmet

Do bicycle safety helmets prevent injury?	Wearing a helmet	Not wearing a helmet	<i>totals</i>
Head injury	17	218	<i>235</i>
No head injury	130	428	<i>558</i>
<i>totals</i>	<i>147</i>	<i>646</i>	<i>793</i>

Notice that the rate of head injury is $17/147=11.5\%$ in the bicycle riders who wore helmets, compared to $218/646=33.7\%$ in the group who did not use helmets: quite a difference! The chi-square test result confirms our impression that there is an association.

Assumptions about the Test

- The chi-square test assumes independent observations and a large sample size.
- In practice, it has been observed that the large sample size assumption is satisfied when all **expected cell counts** are larger than 5.
- If this assumption is violated, then we can use the **Fisher's Exact Test**, which is a type of permutation test that uses a different method ("*computationally arduous*" according to your text book!) to calculate the p-value. STATA and other software tools routinely provide Fisher's exact test.

Example of Fisher's Exact Test

Association between using a tax service and making a mistake on the tax return	Made a mistake	No mistakes	<i>totals</i>
Used a service	2	496	<i>498</i>
Did their own taxes	3	441	<i>444</i>
<i>totals</i>	<i>5</i>	<i>937</i>	<i>942</i>

Note that actually one of the column total is 5, which indicates that the expected counts for that column are going to be less than 5 since they have to add up to 5.

- The **Fisher's exact test** p-value is $p=0.671$
- What would you conclude about the null hypothesis? Fail to reject.

Chi-Square Test in STATA

***** head injury example from lecture;

The frequency table looks as follows:

17	218
130	428

'tabi' tells
STATA to
make a table

The chi2
option tells
STATA to do
the chi square
test

The '\'
tells
STATA to
create a new
row

tabi 17 218\130 428, chi2

STATA Output

row	col		Total
	1	2	
1	17	218	235
2	130	428	558
Total	147	646	793

Pearson $\chi^2(1) = 28.2555$ Pr = 0.000

We are interested in the chi square test itself, which is the last row of results.

Fisher's Exact Test in STATA

2	496
3	441

The usual chi-square test will not give a reliable result, since it assumes a large sample size (namely that all expected cell counts are larger than 5).


`tabi 2 496\3 441, exact`

The exact option tells STATA to do the Fisher's Exact test

STATA Output

row	col		Total
	1	2	
1	2	496	498
2	3	441	444
Total	5	937	942

Fisher's exact = 0.671
1-sided Fisher's exact = 0.446



This is the
one we want