Lecture 23: Tests for Independence in Two-Way Tables

Chapter 6.4

Previously... Chi-Square Tests

Chi-square χ^2 tests allow us compare observed counts to expected counts.

Previously... Chi-Square Tests

Chi-square χ^2 tests allow us compare observed counts to expected counts.

We compute a χ^2 statistic, and then compare it to the χ^2 distribution to compute *p*-values.

$$\chi^2 = \frac{(\mathsf{obs}_1 - \mathsf{exp}_1)^2}{\mathsf{exp}_1} + \ldots + \frac{(\mathsf{obs}_k - \mathsf{exp}_k)^2}{\mathsf{exp}_k}$$

Previously... Assumptions for Chi-Square Test

1. Independence: Each case is independent of the each other
2. Sample size/distribution: We need at least 5 cases in each scenario i.e. each cell in the table
3. Degrees of freedom: We need at least $df = 2$, i.e. $k \geq 3$

Today's Example

Google is always tinkering with its search ranking algorithm.

Today's Example

Google is always tinkering with its search ranking algorithm. Say we want to compare the following 3 algorithms:

Today's Example

Google is always tinkering with its search ranking algorithm. Say we want to compare the following 3 algorithms:

1. The current version

Today's Example

Google is always tinkering with its search ranking algorithm. Say we want to compare the following 3 algorithms:

1. The current version
2. test 1: test algorithm that boosts the search rank of sites with pictures of cats

Today's Example

Google is always tinkering with its search ranking algorithm. Say we want to compare the following 3 algorithms:

1. The current version
2. test 1: test algorithm that boosts the search rank of sites with pictures of cats
3. test 2: test algorithm that boosts the search rank of sites with pictures of dogs

Today's Example

Among the many metrics for user satisfaction with the results for a particular `query` (search term) is the `new search` variable:

Today's Example

Among the many metrics for user satisfaction with the results for a particular `query` (search term) is the `new search` variable:

- No new search: User clicked on a result. User is likely satisfied with result.

Today's Example

Among the many metrics for user satisfaction with the results for a particular query (search term) is the `new search` variable:

- No new search: User clicked on a result. User is likely satisfied with result.
- New search: User did not click on a result and tried a new related search. User is likely unsatisfied with result.

Today's Example

So we have two categorical variables:

- Categorial variable `algorithm` (current, test 1, and test 2)
- Binary Categorial variable `new search` (yes/no)

Today's Example

So we have two categorical variables:

- Categorial variable `algorithm` (current, test 1, and test 2)
- Binary Categorial variable `new search` (yes/no)

Are they independent? i.e. for different algorithms, do we have different levels of new search?

Today's Example

Let's select queries to evaluate each algorithm and organize them in a contingency table:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search				
New search				
Total	5000	2500	2500	10000

Today's Example

Say we observed the following results:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	4000	2000	2000	8000
New search	1000	500	500	2000
Total	5000	2500	2500	10000

Today's Example

Say we observed the following results:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	4000	2000	2000	8000
New search	1000	500	500	2000
Total	5000	2500	2500	10000

We observe that for all 3 algorithms, there is no new search $\frac{4}{5}$ of the time.

Today's Example

Say we observed the following results:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	4000	2000	2000	8000
New search	1000	500	500	2000
Total	5000	2500	2500	10000

We observe that for all 3 algorithms, there is no new search $\frac{4}{5}$ of the time.

In this case, `algorithm` and `new search` are independent: regardless of which `algorithm` used, the proportion of new searches stays the same.

Today's Example

Now say instead we observed the following results:

	algorithm			
new search	Current	Test 1	Test 2	Total
No new search	4000	2500	1500	8000
New search	1000	0	1000	2000
Total	5000	2500	2500	10000

Today's Example

Now say instead we observed the following results:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	4000	2500	1500	8000
New search	1000	0	1000	2000
Total	5000	2500	2500	10000

In this case, `algorithm` and `new search` are not independent: depending on which `algorithm` used, the proportion of new searches is different.

Hypothesis Test

How now test at the $\alpha = 0.05$ significance level:

$$\begin{aligned} H_0: &\quad \text{The algorithms each perform equally well} \\ \text{vs } H_A: &\quad \text{The algorithms do not perform equally well} \end{aligned}$$

i.e. are the categorial variables `algorithm` and `new search` independent?

Hypothesis Test

How now test at the $\alpha = 0.05$ significance level:

$$\begin{aligned} H_0 &: \quad \text{The algorithms each perform equally well} \\ \text{vs } H_A &: \quad \text{The algorithms do not perform equally well} \end{aligned}$$

i.e. are the categorial variables `algorithm` and `new search` independent?

We can do this via χ^2 tests: comparing observed vs expected counts.

Different Names

The following all refer to the same test:

Different Names

The following all refer to the same test:

- χ^2 test for two-way tables

Different Names

The following all refer to the same test:

- χ^2 test for two-way tables
- χ^2 test for independence of two categorical variables

Different Names

The following all refer to the same test:

- χ^2 test for two-way tables
- χ^2 test for independence of two categorical variables
- χ^2 test for homogeneity: are the algorithms homogeneous in their performance?

Different Names

The following all refer to the same test:

- χ^2 test for two-way tables
- χ^2 test for independence of two categorical variables
- χ^2 test for homogeneity: are the algorithms homogeneous in their performance?
- χ^2 test for contingency tables

Example from Textbook

Let's make the values match the example from the textbook:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Example from Textbook

Before we start, let's make each column reflect a proportion and not a count.

Example from Textbook

Before we start, let's make each column reflect a proportion and not a count.

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	0.7022	0.6996	0.7272	0.7078
New search	0.2978	0.3004	0.2728	0.2922
Total	1	1	1	1

Example from Textbook

Before we start, let's make each column reflect a proportion and not a count.

	algorithm			
new search	Current	Test 1	Test 2	Total
No new search	0.7022	0.6996	0.7272	0.7078
New search	0.2978	0.3004	0.2728	0.2922
Total	1	1	1	1

If all algorithms performed the same, we'd expect

- 0.7078 for all 3 values in the top row
- 0.2922 for all 3 values in the bottom row

Example from Textbook

Before we start, let's make each column reflect a proportion and not a count.

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	0.7022	0.6996	0.7272	0.7078
New search	0.2978	0.3004	0.2728	0.2922
Total	1	1	1	1

If all algorithms performed the same, we'd expect

- 0.7078 for all 3 values in the top row
- 0.2922 for all 3 values in the bottom row

The question is: what is the degree of this deviation?

χ^2 Statistic

To build the χ^2 statistic, we need a notion of what's observed vs what's expected.

χ^2 Statistic

To build the χ^2 statistic, we need a notion of what's observed vs what's expected.

As just shown, the overall proportions were 0.7078 and 0.2978.

χ^2 Statistic

To build the χ^2 statistic, we need a notion of what's observed vs what's expected.

As just shown, the overall proportions were 0.7078 and 0.2978.

i.e. under the null hypothesis, we expect these proportions to hold for all algorithms.

What's Expected

We expect:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search				$7078 = 0.7078 \times 10000$
New search				$2922 = 0.2922 \times 10000$
Total	5000	2500	2500	10000

What's Expected

We expect:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search			$1769.5 = 0.7078 \times 2500$	7078
New search			$730.5 = 0.2922 \times 2500$	2922
Total	5000	2500	2500	10000

What's Expected

We expect:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search		$1769.5 = 0.7078 \times 2500$	1769.5	7078
New search		$730.5 = 0.2922 \times 2500$	730.5	2922
Total	5000	2500	2500	10000

What's Expected

We expect:

new search	algorithm Current	Test 1	Test 2	Total
No new search	$3539 = 0.7078 \times 5000$	1769.5	1769.5	7078
New search	$1461 = 0.2922 \times 5000$	730.5	730.5	2922
Total	5000	2500	2500	10000

Observed vs. Expected

We expect:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	3539	1769.5	1769.5	7078
New search	1461	730.5	730.5	2922
Total	5000	2500	2500	10000

Observed vs. Expected

We expect:

`new search`	algorithm Current	Test 1	Test 2	Total
No new search	3539	1769.5	1769.5	7078
New search	1461	730.5	730.5	2922
Total	5000	2500	2500	10000

We observed:

`new search`	algorithm Current	Test 1	Test 2	Total
No new search	3511	1749	1818	7078
New search	1489	751	682	2922
Total	5000	2500	2500	10000

Computing Expected Counts in a Two-Way Table

In effect, we compute the expected count for the i^{th} row and j^{th} column via:

$$\text{Expected Count for Row } i \text{, Col } j = \frac{(\text{Row } i \text{ Total}) \times (\text{Column } j \text{ Total})}{\text{Table Total}}$$

Chi-Square Statistic

We compute χ^2 test statistic: for all $i = 1, \ldots, 6$ cells

$$\frac{(\text{observed count}_i - \text{expected count}_i)^2}{\text{expected count}_i}$$

Chi-Square Statistic

We compute χ^2 test statistic: for all $i = 1, \ldots, 6$ cells

$$\frac{(\text{observed count}_i - \text{expected count}_i)^2}{\text{expected count}_i}$$

$$\begin{aligned}
\text{Row 1, Col 1} &= \frac{(3511 - 3539)^2}{3539} = 0.222 \\
\vdots \quad & \quad \vdots \\
\text{Row 2, Col 3} &= \frac{(682 - 730.5)^2}{730.5} = 3.220
\end{aligned}$$

Chi-Square Statistic

We compute χ^2 test statistic: for all $i = 1, \ldots, 6$ cells

$$\frac{(\text{observed count}_i - \text{expected count}_i)^2}{\text{expected count}_i}$$

$$\begin{aligned}
\text{Row 1, Col 1} \quad &= \quad \frac{(3511 - 3539)^2}{3539} = 0.222 \\
\vdots \quad & \quad \vdots \\
\text{Row 2, Col 3} \quad &= \quad \frac{(682 - 730.5)^2}{730.5} = 3.220
\end{aligned}$$

So

$$\begin{aligned}
\chi^2 \quad &= \quad 0.222 + 0.237 + \ldots + 3.220 \\
&= \quad 6.120
\end{aligned}$$

Chi-Square Distribution

We compare this to a χ^2 distribution to get the p-value. What are the degrees of freedom?

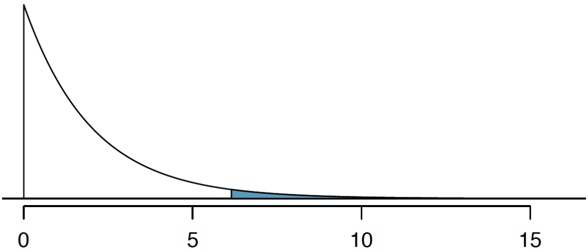
Chi-Square Distribution

We compare this to a χ^2 distribution to get the p-value. What are the degrees of freedom?

$$
\begin{aligned}
df &= (\#\text{ of rows - 1}) \times (\#\text{ of columns - 1}) \\
&= (R-1) \times (C-1) \\
&= (2-1) \times (3-1) = 2 \text{ in our case}
\end{aligned}
$$

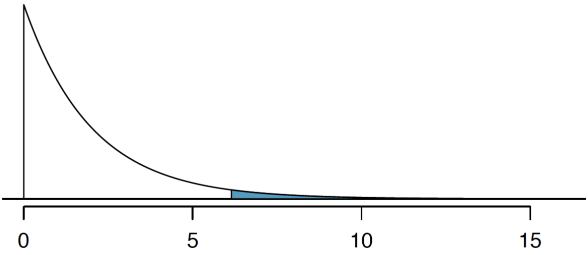
Chi-Square Distribution

Looking up 6.120 in the χ^2 table on page 412 on the *df* $= 2$ row, it would be between 0.05 and 0.02. Since our $\alpha = 0.05$, we reject the null hypothesis and accept the alternative that the algorithms do not perform equally well.



Chi-Square Distribution

Looking up 6.120 in the χ^2 table on page 412 on the $df = 2$ row, it would be between 0.05 and 0.02. Since our $\alpha = 0.05$, we reject the null hypothesis and accept the alternative that the algorithms do not perform equally well.



We can also compute the p-value exactly in R by typing `pchisq(6.120, df=2, lower.tail=FALSE)`. It was 0.047.

Conditions/Assumptions

Nearly identical to conditions/assumptions for χ^2 tests for goodness-of-fit:

Conditions/Assumptions

Nearly identical to conditions/assumptions for χ^2 tests for goodness-of-fit:

1. Independence: Each case is independent of the other

Conditions/Assumptions

Nearly identical to conditions/assumptions for χ^2 tests for goodness-of-fit:

1. Independence: Each case is independent of the other
2. Sample size/distribution: We need at least 5 cases in each scenario i.e. each cell in the table

Conditions/Assumptions

Nearly identical to conditions/assumptions for χ^2 tests for goodness-of-fit:

1. Independence: Each case is independent of the other
2. Sample size/distribution: We need at least 5 cases in each scenario i.e. each cell in the table
3. Degrees of freedom: (Different than before) We need $df = (R-1) \times (C-1) \geq 2$.

Why Are They Called Degrees of Freedom?

In the case of χ^2 tests, the degrees of freedom is the number of values needed before you specify all values in the cells of the table.

Why Are They Called Degrees of Freedom? Rows

Each row has 2 degrees of freedom because once we've specified 2 values, all values in the row are specified.

Why Are They Called Degrees of Freedom? Rows

Each row has 2 degrees of freedom because once we've specified 2 values, all values in the row are specified.

Say in the 1st row, we specify two (arbitrarily chosen) values:

	algorithm			
new search	Current	Test 1	Test 2	Total
No new search	X	Y		7078
New search				2922
Total	5000	2500	2500	10000

Why Are They Called Degrees of Freedom? Rows

Each row has 2 degrees of freedom because once we've specified 2 values, all values in the row are specified.

Say in the 1st row, we specify two (arbitrarily chosen) values:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	X	Y		7078
New search				2922
Total	5000	2500	2500	10000

then the missing value is $7078 - X - Y$.
i.e. the wiggle room we have is $C - 1$ two cells

Why Are They Called Degrees of Freedom? Columns

Each column has 1 degree of freedom because once we've specified 1 value, all values in the column are specified.

Say in the 1st column, we specify one (arbitrarily chosen) value:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	X			7078
New search				2922
Total	5000	2500	2500	10000

Why Are They Called Degrees of Freedom? Columns

Each column has 1 degree of freedom because once we've specified 1 value, all values in the column are specified.

Say in the 1st column, we specify one (arbitrarily chosen) value:

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	X			7078
New search				2922
Total	5000	2500	2500	10000

then the missing value is $5000 - X$.
i.e. the wiggle room we have is $R - 1$ one cell

Why Are They Called Degrees of Freedom? Columns

So the overall *df* is the row degree of freedom times the column degree of freedom $(C - 1) \times (R - 1)$, in our case $df = 2$.

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	X	Y		7078
New search				2922
Total	5000	2500	2500	10000

Why Are They Called Degrees of Freedom? Columns

So the overall *df* is the row degree of freedom times the column degree of freedom $(C-1) \times (R-1)$, in our case $df = 2$.

	algorithm			
`new search`	Current	Test 1	Test 2	Total
No new search	X	Y		7078
New search				2922
Total	5000	2500	2500	10000

i.e. if we know these two values, we can fill the rest of the table.

Next Lecture

We kick off Chapter 7: Introduction to Linear Regression.