Lecture 23: Tests for Independence in Two-Way Tables

Chapter 6.4

Today's Example

Example from Chapter 6.4 page 297: Google is always tinkering with its search ranking algorithm. Say we want to compare the following 3 algorithms:

- 1. the current version
- 2. new test algorithm 1
- 3. new test algorithm 2

Today's Example

User satisfaction is measured with the new search variable:

- no new search: User clicked on a result. Suggests user is satisfied with result.
- ► new search: User did not click on a result and tried a new related search. Suggests user is dissatisfied with result.

Today's Example

Today's Example

Say we observe the following contingency table:

	a.	Lgorithr	n	
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

For all 3 algorithms, there is a new search $\frac{1}{5}$ of the time.

The two levels of new search are independent of algorithm: regardless of which algorithm used, the proportion of new searches stays the same.

E / 2

Today's Example

Now say instead we observed the following results:

	a.	lgorithm	n	
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, they are dependent: depending on which algorithm used, the proportion of new search is different.

Hypothesis Test		
		7/23
Different Names		

Example from Textbook

Let's make the values match the example from the textbook on page 299:

	a.	lgorithm	n	
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.

	a	lgorithm	n	
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

Are we observing what we expect? i.e. What is the degree of this deviation?

What's Expected

We expect:

	a.	Lgorithr	n	
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

11 / 23

What's Expected

We expect:

		a	lgorithm	
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

13 / 23

What's Expected

We expect:

	algorit	hm		
new search	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

Expected Counts:

	a	lgorithm	n	
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

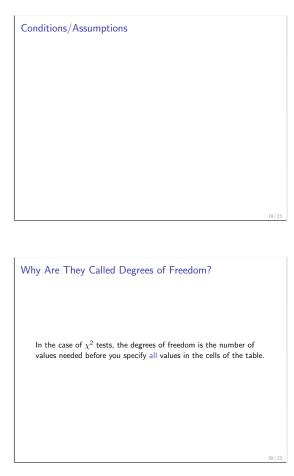
Observed Counts:

	a.	Lgorithm	n	
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

15 / 23

Chi-Square Statistic

Chi-Square Distribution	
	17 / 23
	17 / 23
	17 / 23
Chi-Square Distribution	17/23
Chi-Square Distribution	17 / 23
Chi-Square Distribution	17/23
Chi-Square Distribution	17/23
Chi-Square Distribution	17/23



Why Are They Called Degrees of Freedom? Rows

Each row has df=2 because if we specify 2 values, all values in the row are specified.

Example:

	algorithm			
new search	Current	Test 1	Test 2	Total
No new search	Х	Υ		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 $\mathcal{C}-1$ two cells

21 / 22

Why Are They Called Degrees of Freedom? Columns

Each column has df=1 because if we specify 1 value, all values in the column are specified.

Example:

	a.			
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 $(C-1) \times (R-1)$, in our case df = 2.

	algorithm			
new search	Current	Test 1	Test 2	Total
No new search	Х	Υ		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.

23 / 23