STAT 113
Tests for One and Two Multi-category Variables

Colin Reimer Dawson

Oberlin College

December 1, 2017

Outline

Testing for Association Between Categorical Variables

Handout

Driving While Black

	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	510	109	240	16	7	882
Not Searched	1826	2081	1008	486	104	5505
Total	2336	2190	1248	502	111	6387

H_0 : No association between driver race and vehicle search

H_1 : Some association between driver race and vehicle search

A Randomization Scheme

- We can simulate data with no association by randomly pairing the two variables (pair 6387 driver race values, with counts matching the data, with 6387 search/no search values, with counts matching the data)
- Count how often each combination occurs in the randomization sample to build a random 2-way contingency table.

Deriving Expected Counts

- We can use the same Chi-square (χ^2) statistic as we used for the Goodness of Fit test:

$$\chi^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{\left(\text{Observed Count}_{r,c} - \text{Expected Count}_{r,c}\right)^2}{\text{Expected Count}_{r,c}}$$

- The overall search rate is 13.8%. If driver race and search rate are not associated, we expect 13.8% of the cases in each column to be in the "Search" row.

Driving While Black: Expected Counts

	H/L	White	Black	Asian	Other	Total Prop.
Searched	$0.138 \times 0.366 \times 6387$					0.138
Not Searched	$0.862 \times 0.366 \times 6387$					0.862
Total Prop.	0.366	0.343	0.195	0.079	0.017	1.000

Driving While Black: Expected Counts

	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	322.37	302.22	172.22	69.28	15.32	882
Not Searched	2013.63	1887.78	1075.78	432.72	95.68	5505
Total	2336	2190	1248	502	111	6387

Driving While Black: Observed Counts

	Hisp./Lat.	White	Black	Asian	Other	Total
Searched	510	109	240	16	7	882
Not Searched	1826	2081	1008	486	104	5505
Total	2336	2190	1248	502	111	6387

$$\begin{aligned}\chi^2 =& (510 - 322.37)^2/322.37 + (109 - 302.22)^2/302.22 + \ldots \\ & \cdots + (104 - 95.68)^2/95.68 \\ =& 353.51\end{aligned}$$

What distribution?

- If we are doing a randomization test, we are all set to go: randomly re-pair the Search and Driver Race variables for each randomization sample; compute the χ^2 stats; and see what proportion are at least 353.51.
- The appropriate theoretical distribution is the χ^2. But how many df?

How many degrees of freedom are there?

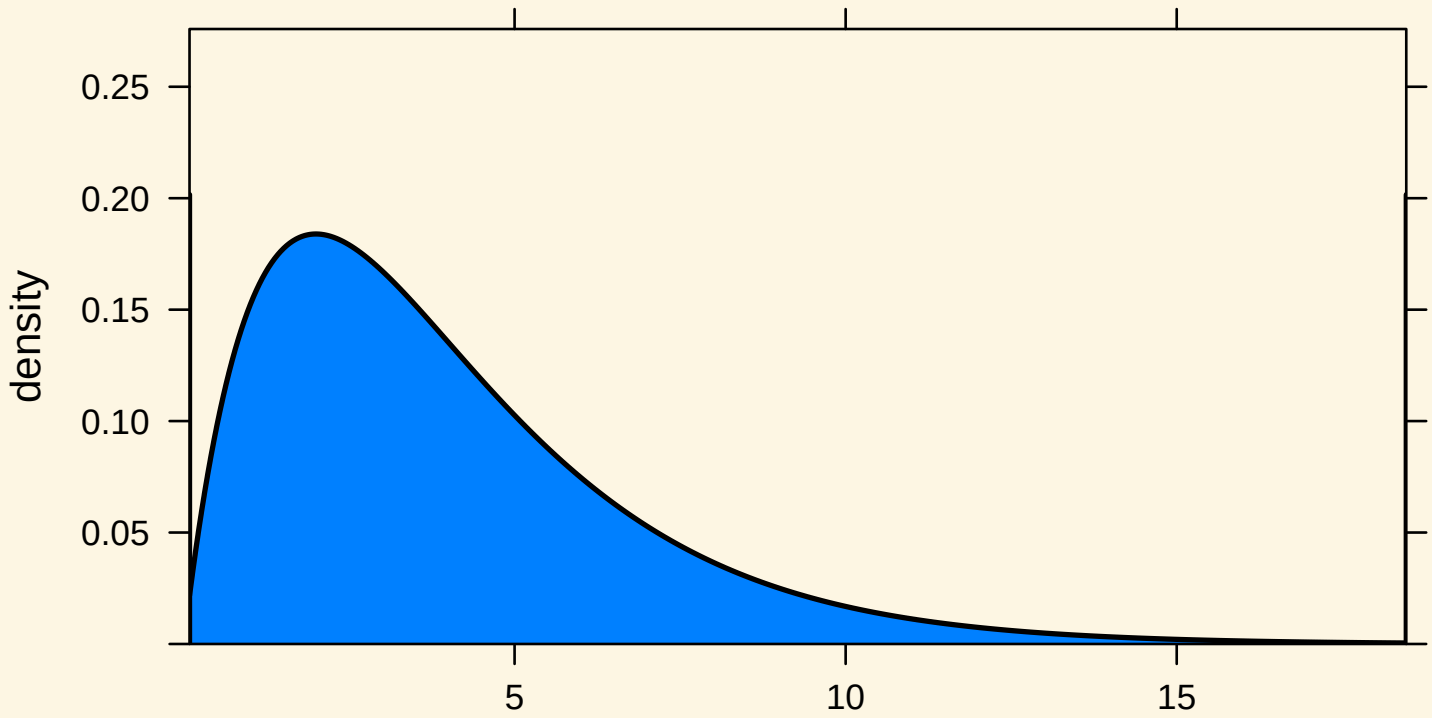
- We are treating the row and column totals as *fixed*. Given that, how many pieces of information are there in one sample?
- The last row and column must consist of "whatever is left" to reach the row and column totals. So, there are only (R-1)(C-1) "free" values.

Conditions Needed to use χ^2 Distribution

Need $n \geq 5 \cdot R \cdot C$; i.e., an average of five observations per cell.

Finding the P-value

```
xpchisq(353.51, df = (2-1)*(5-1), lower.tail = FALSE)
```




```
     [1] 3.062673e-75

## It's literally "off the chart"!
```

The "Pre-packaged" Method

```
observed.counts <-
    rbind(  ## Stands for "row bind" --- creates a two-way table
        Searched = c(HL = 510, W = 109, B = 240, A = 16, O = 7),
        NotSearched = c(HL = 1826, W = 2081, B = 1008, A = 486, O = 104))
## Also possible to read raw data from a file and use tally() to get counts
observed.counts  ## Look at the data to make sure it looks right
```

```
                HL    W    B   A   O
Searched       510  109  240  16   7
NotSearched   1826 2081 1008 486 104
```

```
## Since expected proportions come from the data, R will compute them for you
xchisq.test(observed.counts)


     Pearson's Chi-squared test

    data:  x
    X-squared = 353.51, df = 4, p-value < 2.2e-16

      510.00    109.00    240.00     16.00      7.00
    ( 322.59) ( 302.42) ( 172.34) (  69.32) (  15.33)
    [108.88] [123.71] [ 26.56] [ 41.02] [  4.53]
    < 10.43> <-11.12> <  5.15> < -6.40> < -2.13>

     1826.00   2081.00   1008.00    486.00    104.00
    (2013.41) (1887.58) (1075.66) ( 432.68) (  95.67)
    [ 17.45] [ 19.82] [  4.26] [  6.57] [  0.72]
    < -4.18> <  4.45> < -2.06> <  2.56> <  0.85>

    key:
     observed
     (expected)
     [contribution to X-squared]
     <Pearson residual>
```

```
## Can use a randomization test instead
xchisq.test(observed.counts, simulate.p = TRUE, B = 10000)
```

```
 Pearson's Chi-squared test with simulated p-value (based on 10000
 replicates)

data:  x
X-squared = 353.51, df = NA, p-value = 9.999e-05

  510.00    109.00    240.00     16.00      7.00
( 322.59) ( 302.42) ( 172.34) (  69.32) (  15.33)
[108.88] [123.71] [ 26.56] [ 41.02] [  4.53]
< 10.43> <-11.12> <  5.15> < -6.40> < -2.13>

 1826.00   2081.00   1008.00    486.00    104.00
(2013.41) (1887.58) (1075.66) ( 432.68) (  95.67)
[ 17.45] [ 19.82] [  4.26] [  6.57] [  0.72]
< -4.18> <  4.45> < -2.06> <  2.56> <  0.85>

key:
 observed
 (expected)
 [contribution to X-squared]
 <Pearson residual>
```

Summary: Chi-Square Test of Goodness of Fit

- One categorical variable; testing whether the set of sample proportions reflect particular population proportions
- Randomization procedure: Sample n values with replacement based on null proportions
- Test statistic:

$$\chi^2 = \sum_{c=1}^{C} \frac{(\text{Observed count}_c - \text{Expected count}_c)^2}{\text{Expected count}_c}$$

- Theoretical distribution: χ^2 with $C - 1$ degrees of freedom
- Condition for analytic approximation: All *expected* counts ≥ 5

Summary: Chi-Square Test of Association

- Association: Two categorical variables; testing whether there is an association between the two (do the conditional proportions differ across rows/columns?)
- Randomization procedure: Randomly pair values, as in testing correlation/regression slope for quantitative variables
- Test statistic:

$$\chi^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$

- Theoretical distribution: χ^2 with $(R-1)(C-1)$ degrees of freedom
- Condition for analytic approximation: All *expected* counts ≥ 5