Chi Square Test

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2011

What is this Presentation?

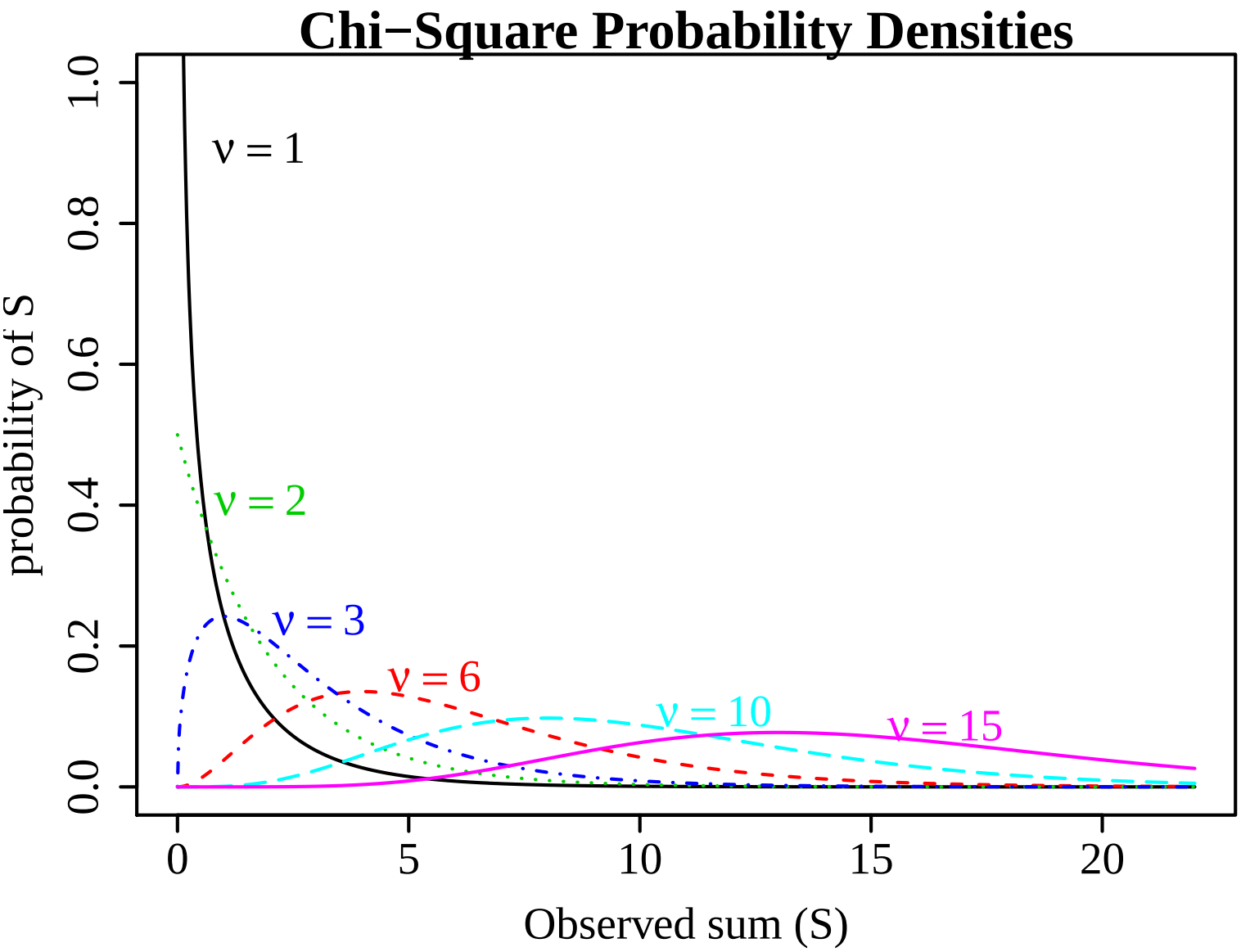
- Chi Square Distribution
- Find Things to Compare With That

The χ^2 Distribution

- Describe accumulated deviations
- Example: Add up ν (Greek letter "nu") Standardized Normal Observations

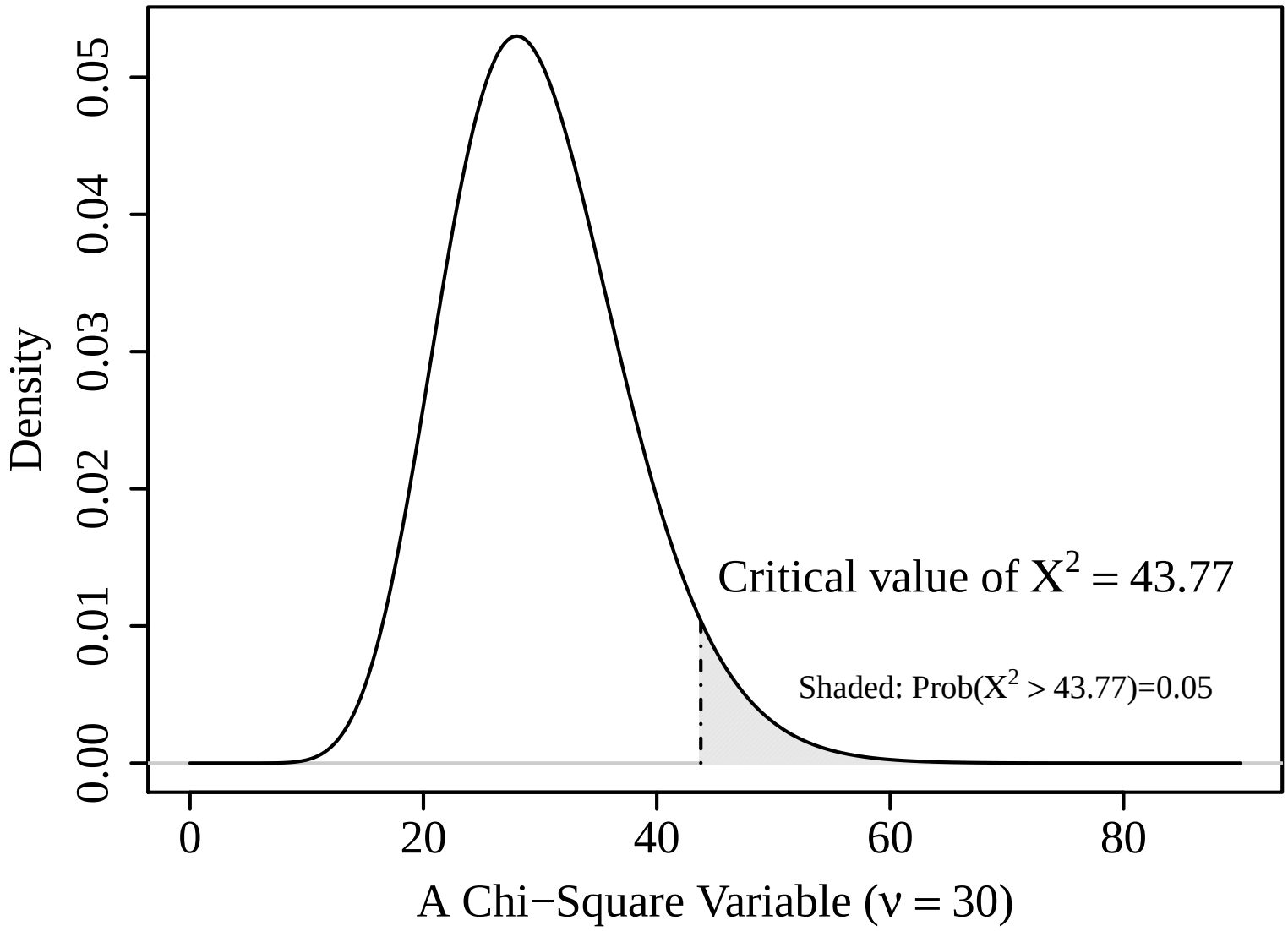
$$S_\nu = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$$

- For each ν, we expect a different probability distribution (obvious!)

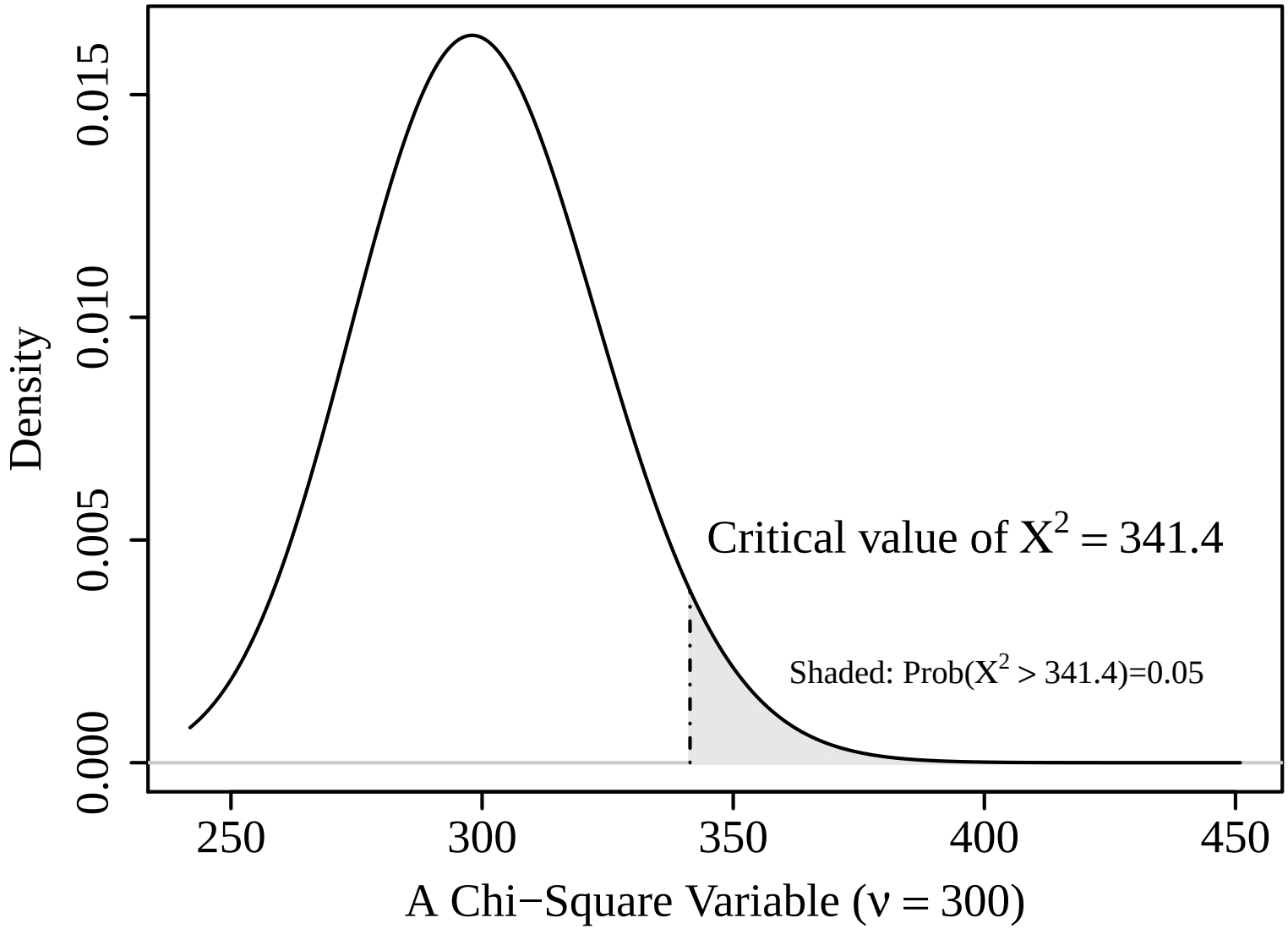

Chi−Square Probability Densities
probability of S
0.0
0.2
0.4
0.6
0.8
1.0
ν = 1
ν = 2
ν = 3
ν = 6
ν = 10
ν = 15
0
5
10
15
20
Observed sum (S)

Leads to a "natural" one tailed test

- Many models have a "predicted" number and "deviations" in the form of a difference b/t observed and predicted
- The sum of squared deviations can be "scaled" onto the χ^2 distribution


Density
0.00
0.01
0.02
0.03
0.04
0.05
0
20
40
60
80
A Chi−Square Variable (ν = 30)
Critical value of $X^2 = 43.77$
Shaded: Prob($X^2 > 43.77$)=0.05

Check a Larger ν



Moments

- Expected Value of S: ν
- Variance of S: 2ν
- Standard Deviation of S: $\sqrt{2\nu}$

ν has a Nickname

- ν is sometimes known as "degrees of freedom."
- Conceptually, it is the number of unrestricted squares added together.

How To Use For Hypothesis Tests

- "Wrestle" some observed problem into a format that resembles a sum of squares
- Get a statistician to figure out a transformation that can be used to compare it with a χ^2.

- Consider a table that is prepared in accordance with the Iron Law of Crosstabs

		Hair			
Juggling		brown	red	yellow	Totals
	great				
	OK				
	none				
Totals					

You Make a Prediction For Each Cell

		Hair			
Juggling		brown	red	yellow	Totals
	great				
	OK				
	none				
Totals					

Now Some Statistical Hocus Pocus

- Consider any Cell, row i and column j
- Here's my first guess. Square the mistakes:

$$(Obs_{ij} - Pred_{ij})^2 \tag{1}$$

- Add across all cells

$$\sum_{cells}(Obs_{ij} - Pred_{ij})^2 \tag{2}$$

- That's not quite right yet, because
  - it is not a Normal variable
  - it is not Standardized either

But You Can See Where This is Heading, Can't you?

- Suppose $\sum_{cells}(Obs_{ij} - Pred_{ij})^2$ is very small. Then your predictions are pretty good. Don't reject them.
- Suppose $\sum_{cells}(Obs_{ij} - Pred_{ij})^2$ is HUGE. Your predictions were bad. Reject the model on which you based them.
- We just need a way to "scale" this sum of squares.

They Call it Pearson's χ^2 Test

- Karl Pearson claimed that if we sum this quantity across cells

$$\frac{(Obs_{ij} - Pred_{ij})^2}{Pred_{ij}} \tag{3}$$

  we can compare the result against a χ^2 distribution.
- The Pearson Chi Square Statistic (suppose we call that X^2)

$$X^2 = \sum_{i=1}^{\#rows} \sum_{j=1}^{\#columns} \frac{(Obs_{ij} - Pred_{ij})^2}{Pred_{ij}} \tag{4}$$

- Pedantic. The number X^2 is not the same as the distribution χ^2. X^2 is "some number" Karl Pearson figured out that has a distribution similar to χ^2 *under some conditions*.

Start Approximating

- Pearson showed that $X \sim \chi^2_\nu$, but only
  - "asymptotically" as the N of cases used to calculate $Pred_{j1}$ tends to infinity.
  - and only if the "true probability" that a case will land in row i, column j is greater than 0
- What's the correct value of ν?
  - A little bit tricky. Depends on how you calculate $Pred_{ij}$

Why so Vague about ν

- (Here's my best guess)
- The ν depends on the amount of information you used from the data to make your predictions
- If you just put in your personal "wild guesses," then use $\nu = r \cdot c$
- If you make a silly prediction that each cell should have an equal proportion of all observed cases, then you are only using the N of cases to make your prediction, so $\nu = r \cdot c - 1$
- If you do the more-or-less standard "identically distributed columns" prediction, then you use $\nu = (r - 1)(c - 1)$

Standard Story about Identical Column Proportions

- The "null hypothesis" is that all columns are samples from identical random processes.
- Multinomial random variable assigns outcomes to row 1 , 2, 3 with probability (p_1, p_2, p_3).
- Note: If only 2 rows, then we have a Binomial distribution (coin flips).
- We don't say what the p_i might be, only that they the same for each column.

Standard Story (cont.)

- Use the observed proportions as estimates of $(p_1, p_2 p_3)$. Suppose we call them $\hat{p}_1$, $\hat{p}_2$, $\hat{p}_3$.
- If each column were drawn from that "multinomial distribution,"
  - we should predict each column's cells as a reflection of those probabilities.
  - For a column with T_j cases, we predict the cell counts are

$$(T_j \hat{p}_1, T_j \hat{p}_2, T_j \hat{p}_3), \textit{so call those predictions } (Pred_{j1}, Pred_{j2}, Pred_{j3}) \tag{5}$$

- By estimating $\hat{p}_i$ from the data, we "use up one degree of freedom" for each row.
- By using T_j from each column, we "use up a degree of freedom" for each column.
- Hence, the correct reference value for the X statistic is χ_ν where $\nu = (r-1)(c-1)$

Crosstabs I

```
options(width = 55)
library(gmodels)
CrossTable(infert$education, infert$induced,
      expected = TRUE, format = "SPSS", chisq = T,
      fisher = T, mcnemar = T)
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|            Total Percent |
|-------------------------|

Total Observations in Table:  248

                | infert$induced
infert$education |         0 |         1 |         2 | Row Total |
-----------------|-----------|-----------|-----------|-----------|
          0-5yrs |         4 |         2 |         6 |        12 |
                 |     6.919 |     3.290 |     1.790 |           |
                 |     1.232 |     0.506 |     9.898 |           |
                 |    33.333% |    16.667% |    50.000% |     4.839% |
                 |     2.797% |     2.941% |    16.216% |           |
                 |     1.613% |     0.806% |     2.419% |           |
-----------------|-----------|-----------|-----------|-----------|
         6-11yrs |        78 |        27 |        15 |       120 |
```

Crosstabs II

	69.194	32.903	17.903	
	1.121	1.059	0.471	
	65.000%	22.500%	12.500%	48.387%
	54.545%	39.706%	40.541%	
	31.452%	10.887%	6.048%	
12+ yrs	61	39	16	116
	66.887	31.806	17.306	
	0.518	1.627	0.099	
	52.586%	33.621%	13.793%	46.774%
	42.657%	57.353%	43.243%	
	24.597%	15.726%	6.452%	
Column Total	143	68	37	248
	57.661%	27.419%	14.919%	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 16.53059 d.f. = 4 p = 0.002383898

McNemar's Chi-squared test

Crosstabs III

```
Chi^2 =  128.0159     d.f. =  3     p =  1.447511e-27


Fisher's Exact Test for Count Data
---------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.007819568


      Minimum expected frequency: 1.790323
Cells with Expected Frequency < 5: 2 of 9 (22.22222%)
```