Unit 7: Topics in inference

2. χ^2 test

Stat 140 - 02

Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*

The data in StudentSurvey includes two categorical variables:

- Award = Academy, Nobel, or Olympic
- HigherSAT = Math or Verbal

Do you think there is a relationship between the award preference and which SAT is higher? If so, in what way?

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math	21	68	116	205
Verbal	10	79	61	150
Total	31	147	177	355

Data are summarized with a 2×3 table for a sample of size n=355.

H_0 : Award preference is not associated with which SAT is higher

H_a : Award preference is associated with which SAT is higher

If H_0 is true $\Longrightarrow$ The award distribution is expected to be the same in each row.

$$\text{Expected Count} = \frac{\text{row total} \times \text{column total}}{n}$$

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math				205
Verbal				150
Total	31	147	177	355

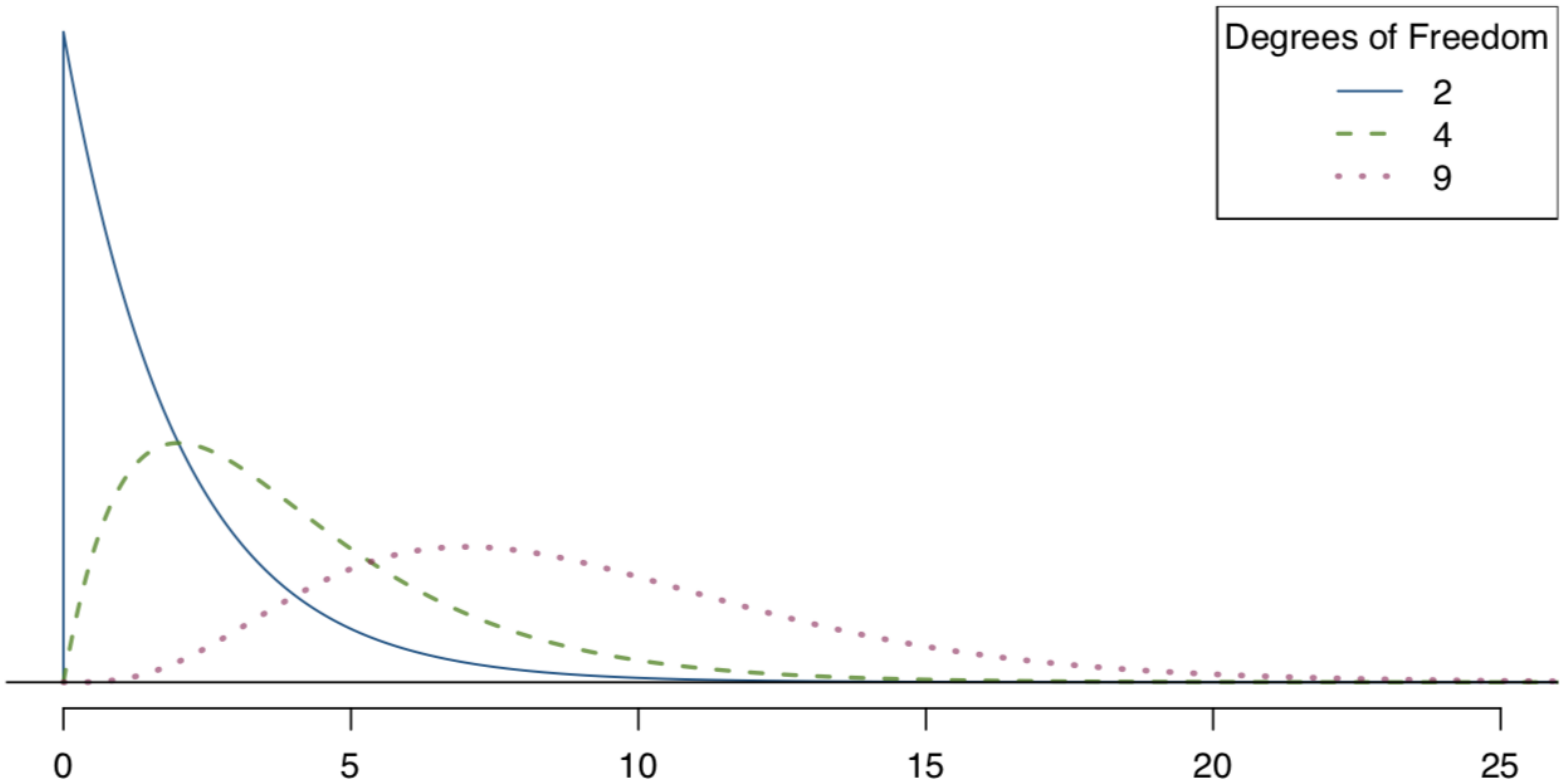
Note: The expected counts maintain row and column totals, but redistribute the counts as if there were no association.

HigherSAT	**Academy**	**Nobel**	**Olympic**	Total
Math	21 (17.9)	68 (84.9)	116 (102.2)	205
Verbal	10 (13.1)	79 (62.1)	61 (74.8)	150
Total	31	147	177	355

HigherSAT	**Academy**	**Nobel**	**Olympic**
Math			
Verbal			

$$\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

The χ^2 distribution has just one parameter, degrees of freedom (df), which influences the shape, center, and spread of the distribution.

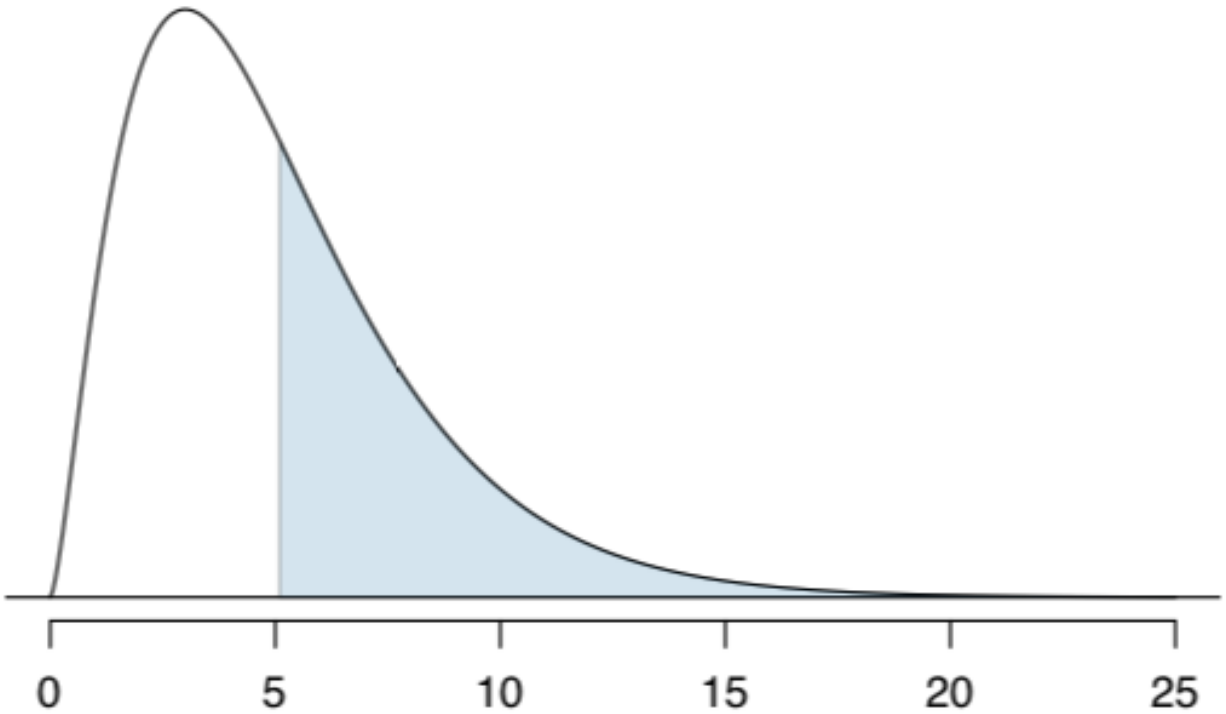


- Independence: In addition to what we previously discussed for independence, each case that contributes a count to the table must be independent of all the other cases in the table.
- Sample size / distribution: Each cell must have at least 5 expected cases.

Under the null hypothesis, the χ^2-statistic follows a χ^2 distribution with the following df

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

p-value $=$ tail area under the chi-square distribution (as usual)
We can compute this using the pchisq() function in R.



1. H_0 : The two variables are not associated
   H_a : The two variables are associated

2. Calculate the expected counts for each cell:

$$\text{Expected Count} = \frac{\text{row total} \times \text{column total}}{n}$$

3. Calculate the χ^2 statistic: $\chi^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$

4. Compute the p-value as the area in the tail above the χ^2 statistic using either a randomization distribution, or a χ^2 distribution with df = (r – 1) (c – 1) if all expected counts > 5

5. Interpret the p-value in context.