# Chapter 6

## Inference for categorical data[1]

Department of Mathematics & Statistics
North Carolina A&T State University

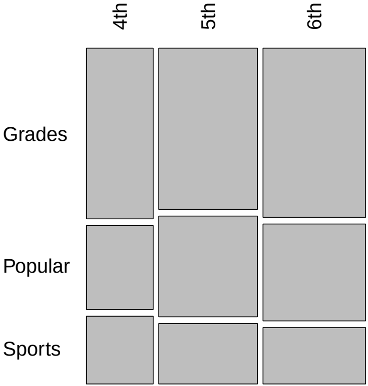
[1] These notes use content from OpenIntro Statistics Slides by Mine Cetinkaya-Rundel.

# Chi-square test of independence

# Popular kids

In the dataset `popular`, students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below. Do these data provide evidence to suggest that goals vary by grade?

| | Grades | Popular | Sports |
|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 |
| $5^{th}$ | 88 | 55 | 33 |
| $6^{th}$ | 96 | 55 | 32 |



# Chi-square test of independence

- The hypotheses are:

  $H_0$: Grade and goals are independent. Goals do not vary by grade.

  $H_A$: Grade and goals are dependent. Goals vary by grade.

# Chi-square test of independence

- The hypotheses are:

  $H_0$: Grade and goals are independent. Goals do not vary by grade.
  $H_A$: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

where $k$ is the number of cells, $R$ is the number of rows, and $C$ is the number of columns.

---

Note: We calculate $df$ differently for one-way and two-way tables.

# Chi-square test of independence

- The hypotheses are:

  $H_0$: Grade and goals are independent. Goals do not vary by grade.
  $H_A$: Grade and goals are dependent. Goals vary by grade.

- The test statistic is calculated as

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E} \quad \text{where} \quad df = (R-1) \times (C-1),$$

  where $k$ is the number of cells, $R$ is the number of rows, and $C$ is the number of columns.

  ---

  Note: We calculate $df$ differently for one-way and two-way tables.

- The p-value is the area under the $\chi^2_{df}$ curve, above the calculated test statistic.

# Expected counts in two-way tables

**Expected counts in two-way tables**

$$Expected\ Count = \frac{(row\ total) \times (column\ total)}{table\ total}$$

# Expected counts in two-way tables

**Expected counts in two-way tables**

$$Expected\ Count = \frac{(row\ total) \times (column\ total)}{table\ total}$$

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 | 119 |
| $5^{th}$ | 88 | 55 | 33 | 176 |
| $6^{th}$ | 96 | 55 | 32 | 183 |
| Total | 247 | 141 | 90 | 478 |

# Expected counts in two-way tables

**Expected counts in two-way tables**

$$Expected\ Count = \frac{(row\ total) \times (column\ total)}{table\ total}$$

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 | 119 |
| $5^{th}$ | 88 | 55 | 33 | 176 |
| $6^{th}$ | 96 | 55 | 32 | 183 |
| Total | 247 | 141 | 90 | 478 |

$$E_{row\ 1, col\ 1} = \frac{119 \times 247}{478} = 61$$

# Expected counts in two-way tables

**Expected counts in two-way tables**

$Expected\ Count = \frac{(row\ total)\times(column\ total)}{table\ total}$

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 | 119 |
| $5^{th}$ | 88 | 55 | 33 | 176 |
| $6^{th}$ | 96 | 55 | 32 | 183 |
| Total | 247 | 141 | 90 | 478 |

$E_{row\ 1,col\ 1} = \frac{119\times247}{478} = 61$

$E_{row\ 1,col\ 2} = \frac{119\times141}{478} = 35$

# Expected counts in two-way tables

What is the expected count for the highlighted cell?

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 | 119 |
| $5^{th}$ | 88 | 55 | 33 | 176 |
| $6^{th}$ | 96 | 55 | 32 | 183 |
| Total | 247 | 141 | 90 | 478 |

A) $\frac{176 \times 141}{478}$

B) $\frac{119 \times 141}{478}$

C) $\frac{176 \times 247}{478}$

D) $\frac{176 \times 478}{478}$

# Expected counts in two-way tables

What is the expected count for the highlighted cell?

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 | 31 | 25 | 119 |
| $5^{th}$ | 88 | 55 | 33 | 176 |
| $6^{th}$ | 96 | 55 | 32 | 183 |
| Total | 247 | 141 | 90 | 478 |

A) $\frac{176\times141}{478}$ $\rightarrow$52

B) $\frac{119\times141}{478}$

C) $\frac{176\times247}{478}$

D) $\frac{176\times478}{478}$

more than expected # of 5th graders have a goal of being popular

# Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed count.

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 61 | 31 35 | 25 23 | 119 |
| $5^{th}$ | 88 91 | 55 52 | 33 33 | 176 |
| $6^{th}$ | 96 95 | 55 54 | 32 34 | 183 |
| Total | 247 | 141 | 90 | 478 |

# Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed count.

| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 61 | 31 35 | 25 23 | 119 |
| $5^{th}$ | 88 91 | 55 52 | 33 33 | 176 |
| $6^{th}$ | 96 95 | 55 54 | 32 34 | 183 |
| Total | 247 | 141 | 90 | 478 |

$$\chi^2 \quad = \quad \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \quad + \frac{(32-34)^2}{34} = 1.3121$$

# Calculating the test statistic in two-way tables

Expected counts are shown in blue next to the observed count.

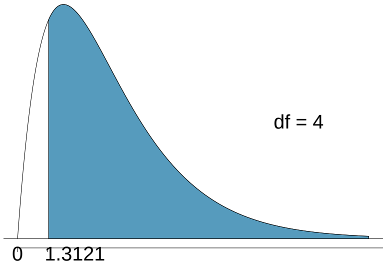
| | Grades | Popular | Sports | Total |
|---|---|---|---|---|
| $4^{th}$ | 63 61 | 31 35 | 25 23 | 119 |
| $5^{th}$ | 88 91 | 55 52 | 33 33 | 176 |
| $6^{th}$ | 96 95 | 55 54 | 32 34 | 183 |
| Total | 247 | 141 | 90 | 478 |

$$\begin{aligned} \chi^2 &= \sum \frac{(63-61)^2}{61} + \frac{(31-35)^2}{35} + \quad + \frac{(32-34)^2}{34} = 1.3121 \\ df &= (R-1)\times(C-1) = (3-1)\times(3-1) = 2\times 2 = 4 \end{aligned}$$

# Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$$\chi^2 = 1.3121 \qquad df = 4$$




A) More than 0.3

B) Between 0.3 and 0.2

C) Between 0.2 and 0.1

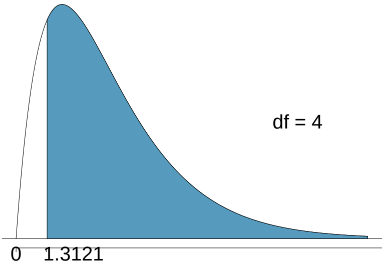
D) Between 0.1 and 0.05

E) Less than 0.001

# Calculating the p-value

Which of the following is the correct p-value for this hypothesis test?

$$\chi^2 = 1.3121 \qquad df = 4$$




A) **More than 0.3**

B) Between 0.3 and 0.2

C) Between 0.2 and 0.1

D) Between 0.1 and 0.05

E) Less than 0.001

# Conclusion

Do these data provide evidence to suggest that goals vary by grade?

$H_0$: Grade and goals are independent. Goals do not vary by grade.

$H_A$: Grade and goals are dependent. Goals vary by grade.

# Conclusion

Do these data provide evidence to suggest that goals vary by grade?

$H_0$: Grade and goals are independent. Goals do not vary by grade.
$H_A$: Grade and goals are dependent. Goals vary by grade.

Since p-value is high, we fail to reject $H_0$. The data do not provide convincing evidence that grade and goals are dependent. It doesn't appear that goals vary by grade.