Two Categorical Variables
Sections 25.1, 25.2, 25.3, 25.4, 25.8
Lecture 45

Robb T. Koether

Hampden-Sydney College

Mon, Apr 11, 2016

Outline

Outline

Two Categorical Variables

- Given two categorical variables, such as sex and political affiliation, we may wonder whether they are related.
- If they are not related, then we say that they are independent.
- If sex and political affiliation are independent, then we should see the same split between Republican, Democrat, and Independent among men as we see among women.
- That is, the proportions should be equal.
- Likewise, we should see the same male/female split whether we are looking at Republicans, Democrats, or Independents.

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	
Male	108	92	200	
Female	112	218	270	

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108	92	200	400
Female	112	218	270	600

- We have the row totals.

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108	92	200	400
Female	112	218	270	600
Total	220	310	470	

- We have the row totals.
- We have the column totals.

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108	92	200	400
Female	112	218	270	600
Total	220	310	470	1000

- We have the row totals.
- We have the column totals.
- We have the grand total.

Outline

The Expected Counts

- We need to compare the observed counts to the expected counts.
- Consider the first column, the Republicans.
  - There were 220 Republicans.
  - Overall, the sample was 40% male and 60% female.
  - Assuming independence, we would expect 40% of the Republicans to be male and 60% to be female.

The Expected Counts

- So the expected count of Republican males is

$$E = 40\% \text{ of } 220$$

The Expected Counts

- So the expected count of Republican males is

$$E = 40\% \text{ of } 220$$
$$= \left(\frac{400}{1000}\right) \times 220$$

.

The Expected Counts

- So the expected count of Republican males is

$$E = 40\% \text{ of } 220$$
$$= \left(\frac{400}{1000}\right) \times 220$$
$$= \frac{400 \times 220}{1000}$$

.

- So the expected count of Republican males is

$$\begin{aligned} E &= 40\% \text{ of } 220 \\ &= \left(\frac{400}{1000}\right) \times 220 \\ &= \frac{400 \times 220}{1000} \\ &= \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}. \end{aligned}$$

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108	92	200	400
Female	112	218	270	600
Total	220	310	470	1000

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92	200	400
Female	112	218	270	600
Total	220	310	470	1000

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92	200	400
Female	112 132	218	270	600
Total	220	310	470	1000

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200	400
Female	112 132	218	270	600
Total	220	310	470	1000

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200	400
Female	112 132	218 186	270	600
Total	220	310	470	1000

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200 188	400
Female	112 132	218 186	270	600
Total	220	310	470	1000

Two-Way Tables

- Suppose we survey 1000 individuals and note their sex and the party affiliation (Rep, Dem, Ind).
- We may display the results in a two-way table.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200 188	400
Female	112 132	218 186	270 282	600
Total	220	310	470	1000

Outline

The Chi-Square Statistic

- To measure how close the observed counts (O) are to the expected counts (E), we compute the fraction

$$\frac{(O-E)^2}{E}$$

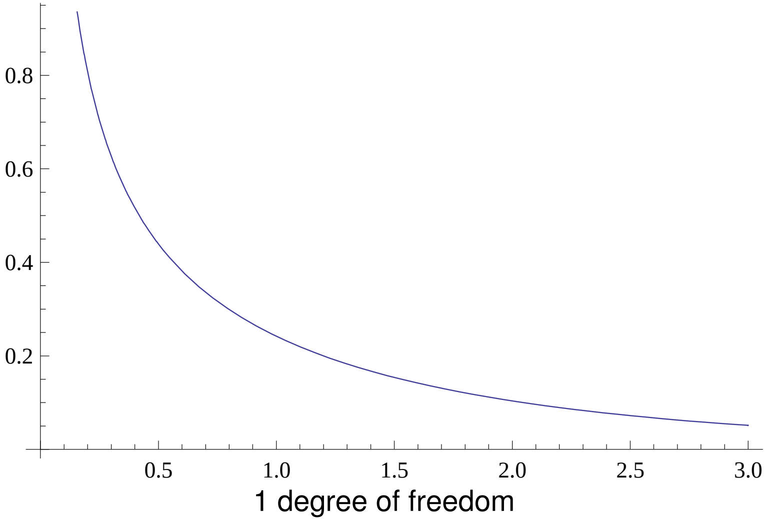
  for each cell in the table.
- The chi-square statistic χ^2 is the sum of these fractions:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}.$$

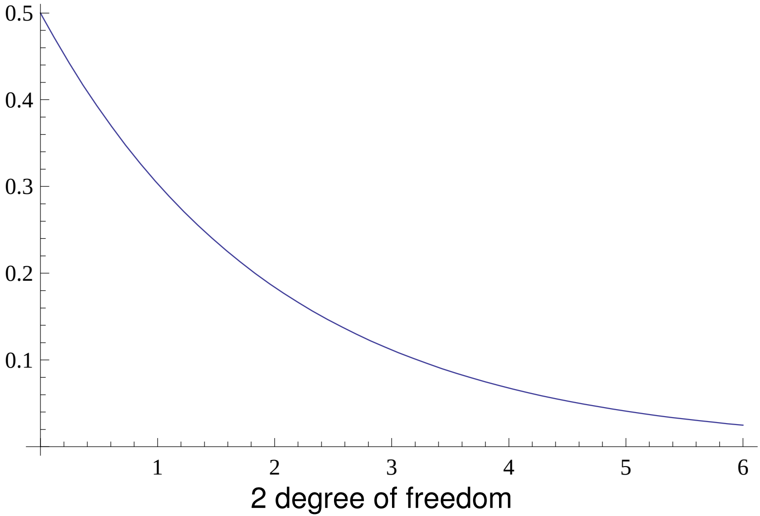
The Chi-Square Distribution

- The distribution of the χ^2 statistic is not symmetric.
- Rather, it is skewed right.
- It also has a different shape for each table size.
- Thus, we must specify the number of degrees of freedom.

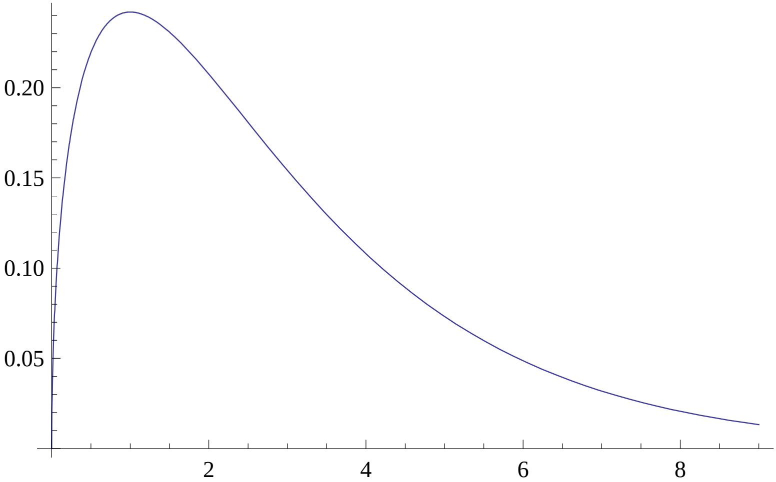
The Chi-Square Distribution



The Chi-Square Distribution

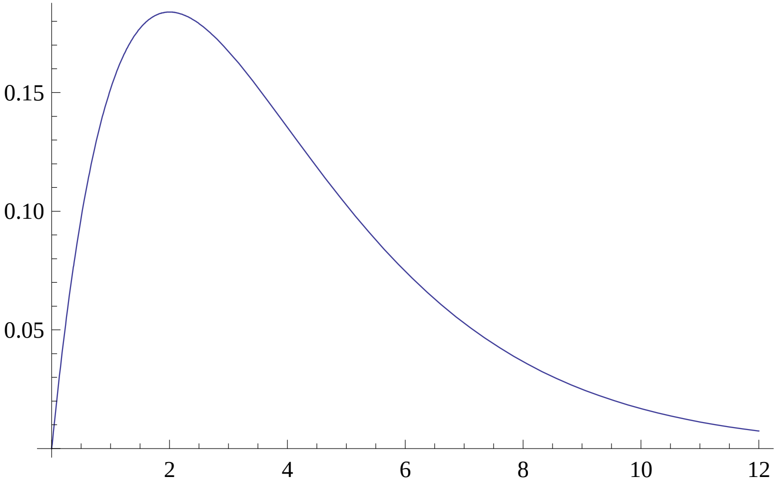


The Chi-Square Distribution



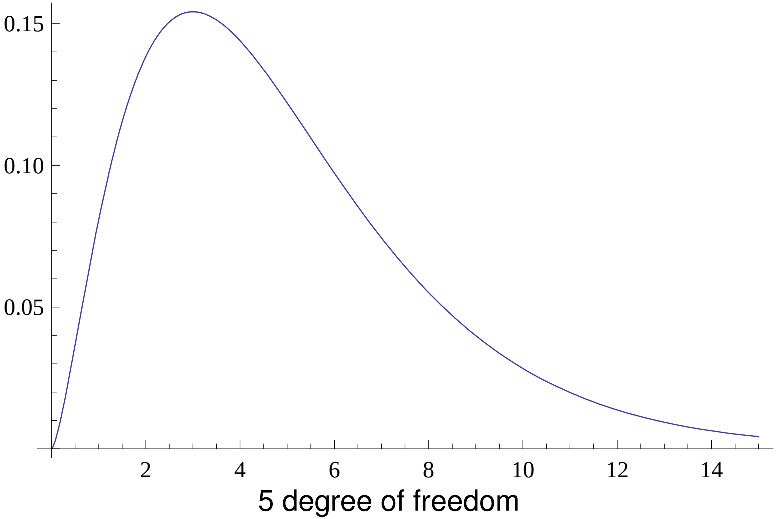
3 degree of freedom

The Chi-Square Distribution

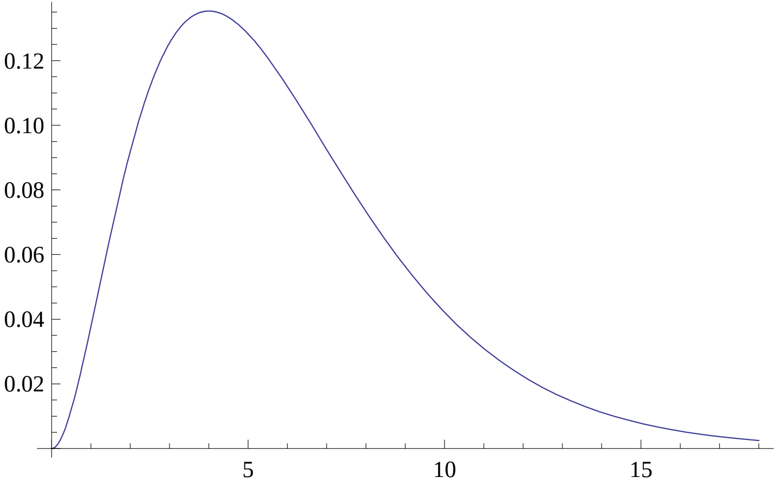


4 degree of freedom

The Chi-Square Distribution

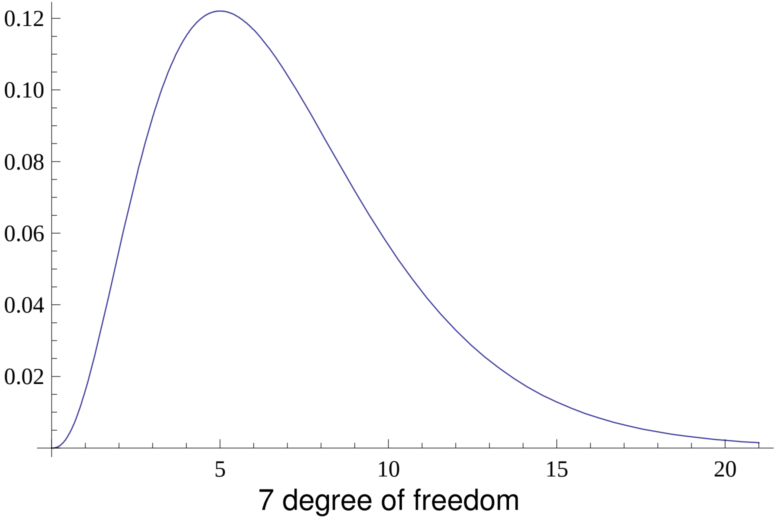


The Chi-Square Distribution

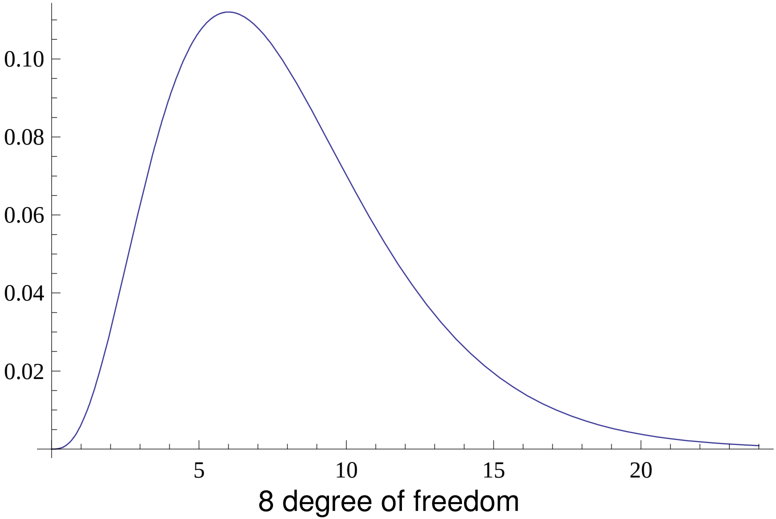


6 degree of freedom

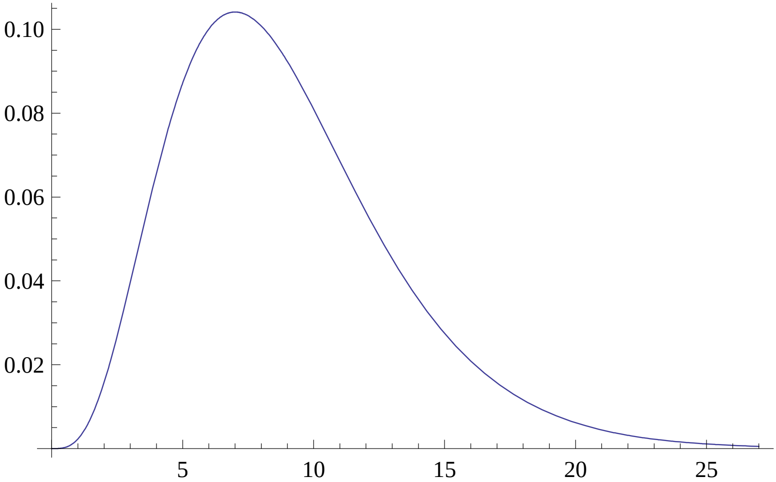
The Chi-Square Distribution



The Chi-Square Distribution

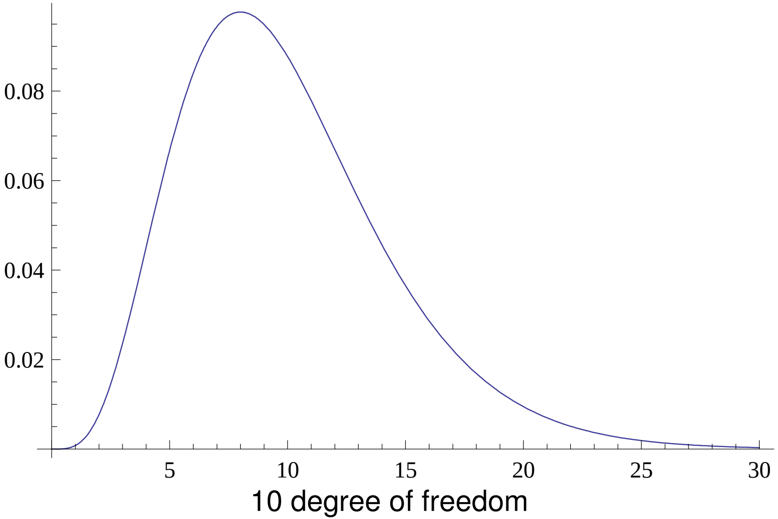


The Chi-Square Distribution

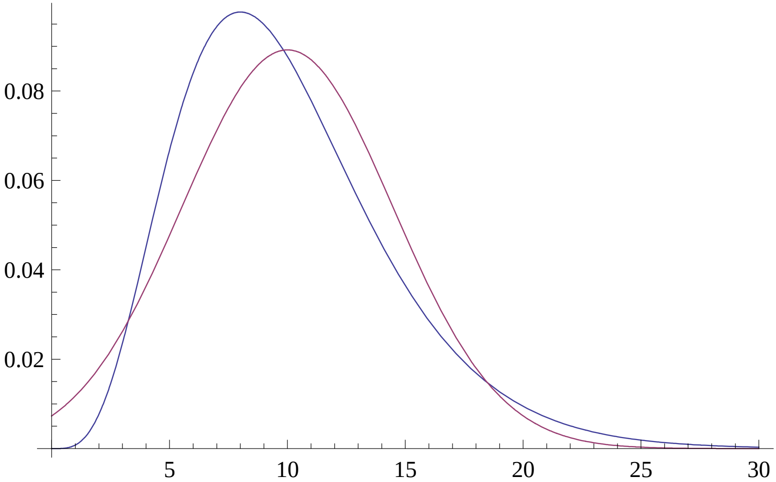


9 degree of freedom

The Chi-Square Distribution

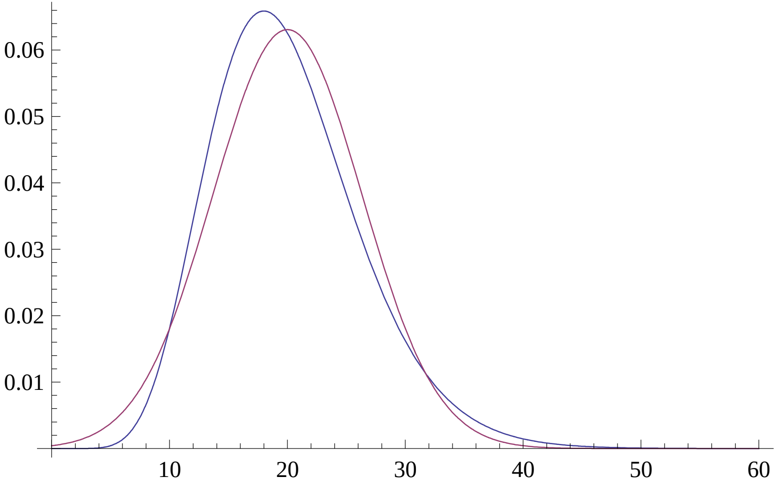


The Chi-Square Distribution



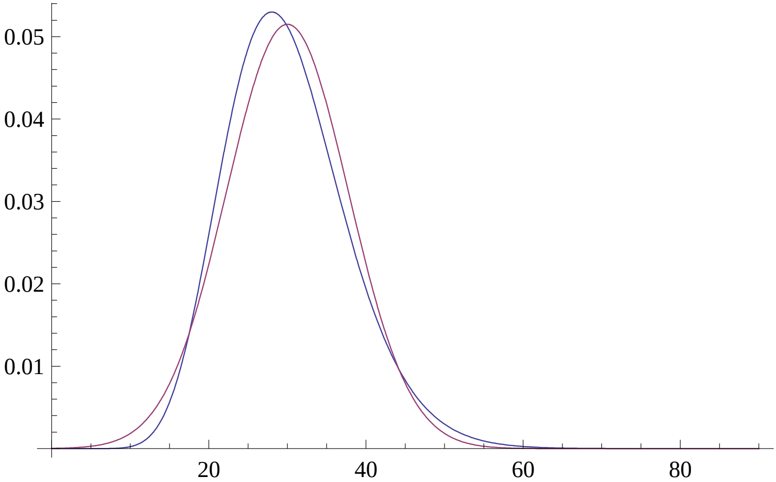
10 degree of freedom vs. $N(10, \sqrt{20})$

The Chi-Square Distribution



20 degree of freedom vs. $N(20, \sqrt{40})$

The Chi-Square Distribution



30 degree of freedom vs. $N(30, \sqrt{60})$

The Chi-Square Distribution

- Characteristics of the χ^2 distributions:
  - The mean of χ^2 equals the degrees of freedom *df*.
  - The standard deviation of χ^2 equals $\sqrt{2df}$.
  - The shape is skewed right, but as *df* increases, the shape approaches the normal distribution $N(df, \sqrt{2df})$.

Degrees of Freedom

- How many degrees of freedom are there in a two-way table? And why are they called “degrees of freedom?”
- Suppose we know the row and column totals, but not the counts.

	Rep	Dem	Ind	Total
Male				400
Female				600
Total	220	310	470	1000

- How many count values can we fill in before the remaining counts are “forced?”

Degrees of Freedom

- In a two-way table, the number of degrees of freedom is

$$\text{df} = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1).$$

Example (Computing χ^2)

- Calculate χ^2 for the following table.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200 188	400
Female	112 132	218 186	270 282	600
Total	220	310	470	1000

Computing the χ^2 Statistic

Example (Computing χ^2)

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

Computing the χ^2 Statistic

Example (Computing χ^2)

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}$$

$$= \frac{(108 - 88)^2}{88} + \frac{(92 - 124)^2}{124} + \frac{(200 - 188)^2}{188} + \frac{(112 - 132)^2}{132} + \frac{(218 - 186)^2}{186} + \frac{(270 - 282)^2}{282}$$

Computing the χ^2 Statistic

Example (Computing χ^2)

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

$$= \frac{(108-88)^2}{88} + \frac{(92-124)^2}{124} + \frac{(200-188)^2}{188} + \frac{(112-132)^2}{132} + \frac{(218-186)^2}{186} + \frac{(270-282)^2}{282}$$

$$= 4.545 + 8.258 + 0.766 + 3.030 + 5.505 + 0.511$$

Computing the χ^2 Statistic

Example (Computing χ^2)

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

$$= \frac{(108-88)^2}{88} + \frac{(92-124)^2}{124} + \frac{(200-188)^2}{188} + \frac{(112-132)^2}{132} + \frac{(218-186)^2}{186} + \frac{(270-282)^2}{282}$$

$$= 4.545 + 8.258 + 0.766 + 3.030 + 5.505 + 0.511$$

$$= 22.615.$$

Outline

The *chi*^{2} Test

- Our procedure will follow the same 6 steps as always.
  1. State the hypotheses.
  2. Give the value of α.
  3. Write the formula for the test statistic.
  4. Calculate the value of the test statistic.
  5. Calculate the *p*-value.
  6. Draw a conclusion.

The *chi*^{2} Test

- The null hypothesis says that there is no difference in the distributions among the rows or among the columns.
- That is, the two variables are independent.

$$H_0: \text{The variables are independent}$$

- The alternative hypothesis says the opposite.

$$H_a: \text{The variables are not independent}$$

The *chi*^{2} Test

- The test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}.$$

- The degrees of freedom is

$$df = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1).$$

- To find the *p*-value of χ^2, use the χ^2`cdf` function on the TI-83.

Outline

Computing the χ^2 Statistic

Example (Computing χ^2)

- Test whether a person's sex and a person's political affiliation are independent.

	Rep	Dem	Ind	Total
Male	108 88	92 124	200 188	400
Female	112 132	218 186	270 282	600
Total	220	310	470	1000

Computing the χ^2 Statistic

Example (Computing χ^2)

(1)

$$H_0: \text{The variables are independent}$$
$$H_a: \text{The variables are not independent}$$

(2) Let $\alpha = 0.05$.

Computing the χ^2 Statistic

Example (Computing χ^2)

(1) The test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}.$$

(2) We calculate $\chi^2 = 22.615$.

(3) The p-value is

$$\begin{aligned} p\text{-value} &= \chi^2\texttt{cdf(22.615,E99,2)} \\ &= 1.228 \times 10^{-5}. \end{aligned}$$

(4) Reject H_0 and conclude that sex and political affiliation are not independent.

Outline

Assignment

Assignment

- Read Sections 25.1, 25.2, 25.3, 25.4, 25.8.
- Apply Your Knowledge: 1, 2, 3, 5, 6.
- Check Your Skills: 19, 20, 21, 24, 25.
- Exercises 30, 31, 32, 34, 35.