

Chapter 6

Inference for categorical data¹

(Author Name)

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

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

Inference for a single proportion

Practice

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

- A) All 1000 get the drug.
- B) 500 get the drug, 500 don't

Practice

Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug?

- A) All 1000 get the drug.
- B) **500 get the drug, 500 don't**

Results from the GSS

The GSS asks the same question, below is the distribution of responses from the 2010 survey:

| | |
|----------------------------|-----|
| All 1000 get the drug | 99 |
| 500 get the drug 500 don't | 571 |
| <hr/> | |
| Total | 670 |

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

- ▶ **Parameter of interest:** Proportion of **all** Americans who have good intuition about experimental design.

p (a population proportion)

Parameter and point estimate

We would like to estimate the proportion of all Americans who have good intuition about experimental design, i.e. would answer "500 get the drug, 500 don't"? What are the parameter of interest and the point estimate?

- ▶ **Parameter of interest:** Proportion of **all** Americans who have good intuition about experimental design.

p (a population proportion)

- ▶ **Point estimate:** Proportion of **sampled** Americans who have good intuition about experimental design.

\hat{p} (a sample proportion)

Inference on a proportion

What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug, 500 don't"?

Inference on a proportion

What percent of all Americans have good intuition about experimental design, i.e. would answer "500 get the drug, 500 don't"?

- ▶ We can answer this research question using a confidence interval, which we know is always of the form

$$\text{point estimate} \pm \text{ME}$$

- ▶ And we also know that $\text{ME} = \text{critical value} \times \text{standard error}$ of the point estimate

Standard error of a sample proportion

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p , and standard error equal to

$$\sqrt{\frac{p(1-p)}{n}}.$$

$$\hat{p} \sim N \left(\text{mean} = p, SE = \sqrt{\frac{p(1-p)}{n}} \right)$$

But of course this is true only under certain conditions...

Any guesses?

Sample proportions are also nearly normally distributed

Central limit theorem for proportions

Sample proportions will be nearly normally distributed with mean equal to the population proportion, p , and standard error equal to $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N \left(\text{mean} = p, SE = \sqrt{\frac{p(1-p)}{n}} \right)$$

But of course this is true only under certain conditions...

Any guesses?

Independent observations, at least 10 successes and 10 failures

Note: If p is unknown (most cases), we use \hat{p} in the calculation of the standard error.

Back to experimental design...

The GSS found that 571 out of 670 (87%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Back to experimental design...

The GSS found that 571 out of 670 (87%) Americans answered the question on experimental design correctly. Estimate (using a 95% confidence interval) the proportion of all Americans who have good intuition about experimental design?

Given: $n = 670, \hat{p} = 0.85$. First, check conditions.

- ▶ **Independence:** The sample is random, and $670 < 10\%$ of all Americans, therefore we can assume that one respondent's response is independent of another.
- ▶ **Success-Failure:** 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

Practice

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

A) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

B) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

C) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

D) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Practice

We are given that $n = 670$, $\hat{p} = 0.85$, we also just learned that the standard error of the sample proportion is $SE = \sqrt{\frac{p(1-p)}{n}}$. Which of the below is the correct calculation of the 95% confidence interval?

A) $0.85 \pm 1.96 \times \sqrt{\frac{0.85 \times 0.15}{670}} \rightarrow (0.82, 0.88)$

B) $0.85 \pm 1.65 \times \sqrt{\frac{0.85 \times 0.15}{670}}$

C) $0.85 \pm 1.96 \times \frac{0.85 \times 0.15}{\sqrt{670}}$

D) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

Choosing a sample size

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%?

Choosing a sample size

How many people should you sample in order to cut the margin of error of a 95% confidence interval down to 1%?

$$ME = z^* \times SE$$

$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \text{Use } \hat{p} \text{ from previous study}$$

$$0.01^2 \geq 1.96^2 \times \frac{0.85 \times 0.15}{n}$$

$$n \geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2}$$

$$n \geq 4898.04 \rightarrow n \text{ should be at least 4,899}$$

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Why?

What if there isn't a previous study?

... use $\hat{p} = 0.5$

Why?

- ▶ If you don't know any better, 50-50 is a good guess.
- ▶ $\hat{p} = 0.5$ gives the most conservative estimate — highest possible sample size.

CI vs. HT for proportions

- ▶ Success-Failure conditions:

- ▶ CI: At least 10 **observed** successes and failures.
- ▶ HT: At least 10 **expected** successes and failures, calculated using the null value.

- ▶ Standard error:

- ▶ CI: Calculate using observed sample proportion:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- ▶ HT: Calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

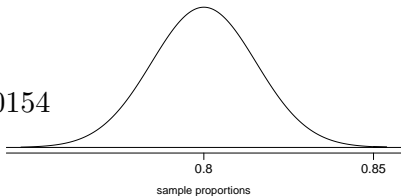
The GSS found that 571 out of 670 (85%) Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

$$H_0 : p = 0.80 \quad H_A : p > 0.80$$

$$SE = \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154$$

$$Z = \frac{0.85 - 0.80}{0.0154} = 3.25$$

$$p\text{-value} = 1 - 0.9994 = 0.0006$$



Since the p-value is low, we reject H_0 . The data provide convincing evidence that more than 80% of Americans have a good intuition on experimental design.

Practice

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

- A) Yes
- B) No
- C) Cannot tell

Practice

11% of 1,001 Americans responding to a 2006 Gallup survey stated that they have objections to celebrating Halloween on religious grounds. At 95% confidence level, the margin of error for this survey is $\pm 3\%$. A news piece on this study's findings states: "More than 10% of all Americans have objections on religious grounds to celebrating Halloween." At 95% confidence level, is this news piece's statement justified?

- A) Yes
- B) **No**
- C) Cannot tell

Recap - Inference for one proportion

- ▶ Popular parameter: p . point estimate: \hat{p}
- ▶ Conditions:
 - ▶ Independence
 - ▶ Random sample and 10% condition
 - ▶ At least 10 successes and failures
 - ▶ If not \rightarrow randomization
- ▶ Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
 - ▶ for CI: use \hat{p}
 - ▶ for HT: use p_0

Difference of two proportions

Melting ice cap

Scientists predict that global warming may have big effects on the polar regions within the next 100 years. One of the possible effects is that the northern ice cap may completely melt. Would this bother you a great deal, some, a little, or not at all if it actually happened?

- A) A great deal
- B) Some
- C) A little
- D) Not at all

Results from the GSS

The GSS asks the same question, below are the distributions of responses from the 2020 GSS as well as from a group of introductory statistics students at Duke University:

| | GSS | Duke |
|--------------|-----|------|
| A great deal | 454 | 69 |
| Some | 124 | 30 |
| A little | 52 | 4 |
| Not at all | 50 | 2 |
| Total | 680 | 105 |

Parameter and point estimate

- ▶ **Parameter of interest:** Difference between the proportions of **all** Duke students and **all** Americans who would be bothered a great deal by the northern ice cap completely melting.

$$p_{Duke} - p_{US}$$

- ▶ **Point estimate:** Difference between the proportions of **sampled** Duke students and **sampled** Americans who would be bothered a great deal by the northern ice cap completely melting.

$$\hat{p}_{Duke} - \hat{p}_{US}$$

Inference for comparing proportions

- ▶ The details are the same as before...
- ▶ CI: *point estimate \pm margin of error*.
- ▶ HT: Use $Z = \frac{\text{point estimate} - \text{null value}}{SE}$ to find appropriate p-value.
- ▶ We just need the appropriate standard error of the point estimate ($SE_{\hat{p}_{Duke} - \hat{p}_{US}}$). which is the only new concept.

Standard error of the difference between two sample proportions

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Conditions for CI for difference of proportions

- ▶ **Independence within groups:**

- ▶ The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.
- ▶ $105 < 10\%$ of all Duke students and $680 < 10\%$ of all Americans.

We can assume that the attitudes of Duke students in the sample are independent of each other, and attitudes of US residents in the sample are independent of each other as well.

- ▶ **Independence between groups:** The sampled Duke students and the US residents are independent of each other.
- ▶ **Success-Failure:** At least 10 observed successes and 10 observed failures in the two groups.

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|------|-----|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$(\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} & (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\ &= (0.657 - 0.668) \end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} & (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\ &= (0.657 - 0.668) \pm 1.96 \end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} & (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\ = & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\&= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\&= -0.011 \pm\end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\&= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\&= -0.011 \pm 1.96 \times 0.0497\end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\&= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\&= -0.011 \pm 1.96 \times 0.0497\end{aligned}$$

Practice

Construct 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by melting of the northern cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^* \times \sqrt{\frac{\hat{p}_{Duke}(1 - \hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1 - \hat{p}_{US})}{n_{US}}} \\&= (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\&= -0.011 \pm 1.96 \times 0.0497\end{aligned}$$

Practice

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

- A) $H_0 : p_{Duke} = p_{US}$
 $H_A : p_{Duke} \neq p_{US}$
- B) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
 $H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$
- C) $H_0 : p_{Duke} - p_{US} = 0$
 $H_A : p_{Duke} - p_{US} \neq 0$
- D) $H_0 : p_{Duke} = p_{US}$
 $H_A : p_{Duke} < p_{US}$

Practice

Which of the following is the correct set of hypotheses for testing if the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do?

- A) $H_0 : p_{Duke} = p_{US}$
 $H_A : p_{Duke} \neq p_{US}$
- B) $H_0 : \hat{p}_{Duke} = \hat{p}_{US}$
 $H_A : \hat{p}_{Duke} \neq \hat{p}_{US}$
- C) $H_0 : p_{Duke} - p_{US} = 0$
 $H_A : p_{Duke} - p_{US} \neq 0$
- D) $H_0 : p_{Duke} = p_{US}$
 $H_A : p_{Duke} < p_{US}$

Both A) and C) are correct.

Flashback to working with one proportion

- ▶ When constructing a confidence interval for a population proportion, we check if the **observed** number of successes and failures are at least 10.

$$n\hat{p} \geq 10 \quad n(1 - \hat{p}) \geq 10$$

- ▶ When conducting a hypothesis test for a population proportion, we check if the **expected** number of successes and failure are at least 10.

$$np_0 \geq 10 \quad n(1 - p_0) \geq 10$$

Pooled estimate of a proportion

- ▶ In the case of comparing two proportions where $h_0 : p_1 = p_2$, there isn't a given null value we can use to calculate the **expected** number of successes and failures in each sample.
- ▶ Therefore, we need to first find a common (**pooled**) proportion for the two groups, and use that in our analysis.
- ▶ This simply means finding the proportion of total successes among the total number of observations.

Pooled estimate of a proportion

$$\hat{p} = \frac{\# \text{ of successes}_1 + \# \text{ of successes}_2}{n_1 + n_2}$$

Practice

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

Practice

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\hat{p} = \frac{\# \text{ of successes}_1 + \# \text{ of successes}_2}{n_1 + n_2}$$

Practice

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}\hat{p} &= \frac{\# \text{ of successes}_1 + \# \text{ of successes}_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680}\end{aligned}$$

Practice

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}\hat{p} &= \frac{\# \text{ of successes}_1 + \# \text{ of successes}_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680} = \frac{523}{785}\end{aligned}$$

Practice

Calculate the estimated pooled proportion of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap. Which sample proportion (\hat{p}_{Duke} or \hat{p}_{US}) the pooled estimate is closer to? Why?

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned}\hat{p} &= \frac{\# \text{ of successes}_1 + \# \text{ of successes}_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666\end{aligned}$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$Z = \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}}$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\ &= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \end{aligned}$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\ &= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} \end{aligned}$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\ &= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \end{aligned}$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\ &= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \end{aligned}$$

$$p\text{-value} = 2 \times P(Z < -0.22)$$

Practice

Do these data suggest that the proportion of all Duke students who would be bothered a great deal by the melting of the northern ice cap differs from the proportion of all Americans who do? Calculate the test statistic, the p-value, and interpret your conclusion in context of the data.

| Data | Duke | US |
|------------------|-------|-------|
| A great deal | 69 | 454 |
| Not a great deal | 36 | 226 |
| Total | 105 | 680 |
| \hat{p} | 0.657 | 0.668 |

$$\begin{aligned} Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\ &= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \end{aligned}$$

$$p\text{-value} = 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82$$

Recap - comparing two proportions

- ▶ Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$

Recap - comparing two proportions

- ▶ Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- ▶ Conditions:

Recap - comparing two proportions

- ▶ Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- ▶ Conditions:
 - ▶ independence within groups
 - random sample and 10% condition met for both groups
 - ▶ independence between groups
 - ▶ at least 10 successes and failures in each group
 - if not → randomization (Section 6.4)

Recap - comparing two proportions

- ▶ Population parameter: $(p_1 - p_2)$, point estimate: $(\hat{p}_1 - \hat{p}_2)$
- ▶ Conditions:
 - ▶ independence within groups
 - random sample and 10% condition met for both groups
 - ▶ independence between groups
 - ▶ at least 10 successes and failures in each group
 - if not \rightarrow randomization (Section 6.4)
- ▶ $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
 - ▶ for CI: use \hat{p}_1 and \hat{p}_2
 - ▶ for HT:
 - ▶ when $H_0 : p_1 = p_2$: use $\hat{p}_{pool} = \frac{\# suc_1 + \# suc_2}{n_1 + n_2}$
 - ▶ when $H_0 : p_1 - p_2 = (\text{some value other than } 0)$: use \hat{p}_1 and \hat{p}_2
 - this is pretty rare

Reference - Standard error calculations

| | one sample | two samples |
|------------|--------------------------------|---|
| mean | $SE = \frac{s}{\sqrt{n}}$ | $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| proportion | $SE = \sqrt{\frac{p(1-p)}{n}}$ | $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ |

Reference - Standard error calculations

| | one sample | two samples |
|------------|--------------------------------|---|
| mean | $SE = \frac{s}{\sqrt{n}}$ | $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| proportion | $SE = \sqrt{\frac{p(1-p)}{n}}$ | $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ |

- When working with means, it's very rare that σ is known, so we usually use s .

Reference - Standard error calculations

| | one sample | two samples |
|------------|--------------------------------|---|
| mean | $SE = \frac{s}{\sqrt{n}}$ | $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| proportion | $SE = \sqrt{\frac{p(1-p)}{n}}$ | $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ |

- ▶ When working with means, it's very rare that σ is known, so we usually use s .
- ▶ When working with proportions,
 - ▶ if doing a hypothesis test, p comes from the null hypothesis
 - ▶ if constructing a confidence interval, use \hat{p} instead

Chi-square test of GOF

Weldon's dice

- ▶ Walter Frank Raphael Weldon (1860 - 1906), was an English evolutionary biologist and a founder of biometry. He was the joint founding editor of *Biometrika*, with Francis Galton and Karl Pearson.
- ▶ In 1894, he rolled 12 dice 26,306 times, and recorded the number of 5s or 6s (which he considered to be a success).
- ▶ It was observed that 5s or 6s occurred more often than expected, and Pearson hypothesized that this was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to 7, the face with 6 pips is lighter than its opposing face, which has

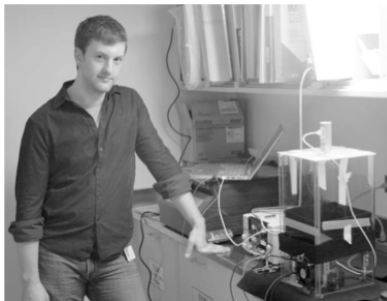


Labby's dice

- ▶ In 2009, Zacariah Labby (U of Chicago), repeated Weldon's experiment using a homemade dice-throwing, pip counting machine.

<http://www.youtube.com/watch?v=95EErdouO2w>

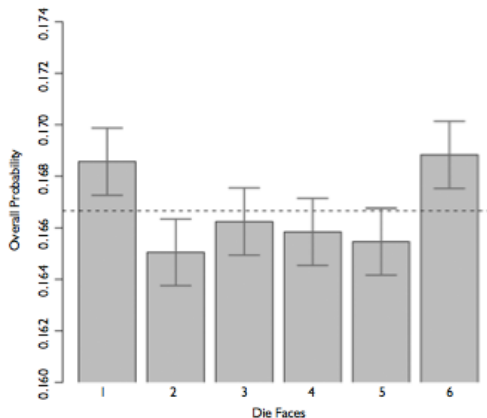
- ▶ The rolling-imaging process took about 20 seconds per roll.



- ▶ Each day there were ~ 150 images to process manually.
- ▶ At this rate Weldon's experiment was repeated in a little more than six full days.

Labby's dice

- ▶ Labby did not actually observe the same phenomenon that Weldon observed (higher frequency of 5s and 6s).
- ▶ Automation allowed Labby to collect more data than Weldon did in 1894, instead of recording “successes” and “failures”, Labby recorded the individual number of pips on each die.



Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, ... , 6s would he expect to have observed?

A) $\frac{1}{6}$

B) $\frac{12}{6}$

C) $\frac{26,306}{6}$

D) $\frac{12 \times 26,306}{6}$

Expected counts

Labby rolled 12 dice 26,306 times. If each side is equally likely to come up, how many 1s, 2s, ... , 6s would he expect to have observed?

A) $\frac{1}{6}$

B) $\frac{12}{6}$

C) $\frac{26,306}{6}$

D) $\frac{12 \times 26,306}{6} = 52,612$

Summarizing Labby's results

| Outcome | Observed | Expected |
|---------|----------|----------|
| 1 | 53,222 | 52,612 |
| 2 | 52,118 | 52,612 |
| 3 | 52,465 | 52,612 |
| 4 | 52,338 | 52,612 |
| 5 | 52,244 | 52,612 |
| 6 | 53,285 | 52,612 |
| Total | 315,672 | 315,672 |

Summarizing Labby's results

| Outcome | Observed | Expected |
|---------|----------|----------|
| 1 | 53,222 | 52,612 |
| 2 | 52,118 | 52,612 |
| 3 | 52,465 | 52,612 |
| 4 | 52,338 | 52,612 |
| 5 | 52,244 | 52,612 |
| 6 | 53,285 | 52,612 |
| Total | 315,672 | 315,672 |

Why are the expected counts the same for all outcomes but the observed counts are different? At a first glance, does there appear to be an inconsistency between the observed and expected counts?

Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

H_0 : There is no inconsistency between the observed and the expected counts. **The observed counts follow the same distribution as the expected counts.**

Setting the hypotheses

Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

H_0 : There is no inconsistency between the observed and the expected counts. **The observed counts follow the same distribution as the expected counts.**

H_A : There is an inconsistency between the observed and the expected counts. **The observed counts do not follow the same distribution as the expected counts.** There is a bias in which side comes up on the roll of a die.

Evaluating the hypotheses

- ▶ To evaluate these hypotheses, we quantify how different the observed count are from the expected counts.

Evaluating the hypotheses

- ▶ To evaluate these hypotheses, we quantify how different the observed count are from the expected counts.
- ▶ Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.

Evaluating the hypotheses

- ▶ To evaluate these hypotheses, we quantify how different the observed count are from the expected counts.
- ▶ Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.
- ▶ This is called a **goodness of fit** test since we're evaluating how well the observed data fit the expected distribution.

Anatomy of a test statistic

- ▶ The general form of a test statistic is

$$\frac{\textit{point estimate} - \textit{null value}}{\textit{SE of point estimate}}$$

Anatomy of a test statistic

- ▶ The general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{SE \text{ of point estimate}}$$

- ▶ This construction is based on
 1. Identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
 2. standardizing that difference using the standard error of the point estimate.

Anatomy of a test statistic

- ▶ The general form of a test statistic is

$$\frac{\text{point estimate} - \text{null value}}{SE \text{ of point estimate}}$$

- ▶ This construction is based on
 1. Identifying the difference between a point estimate and an expected value if the null hypothesis was true, and
 2. standardizing that difference using the standard error of the point estimate.

These two ideas will help in the construction of an appropriate test statistic for count data.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the **chi-square χ^2 statistic**.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the **chi-square χ^2 statistic**.

χ^2 statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O-E)^2}{E} \text{ where } k = \text{total number of cells}$$

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52,465-52,612)^2}{52,612} = 0.41$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52,465-52,612)^2}{52,612} = 0.41$ |
| 4 | 52,338 | 52,612 | $\frac{(52,338-52,612)^2}{52,612} = 1.43$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52,465-52,612)^2}{52,612} = 0.41$ |
| 4 | 52,338 | 52,612 | $\frac{(52,338-52,612)^2}{52,612} = 1.43$ |
| 5 | 52,244 | 52,612 | $\frac{(52,244-52,612)^2}{52,612} = 2.57$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52,465-52,612)^2}{52,612} = 0.41$ |
| 4 | 52,338 | 52,612 | $\frac{(52,338-52,612)^2}{52,612} = 1.43$ |
| 5 | 52,244 | 52,612 | $\frac{(52,244-52,612)^2}{52,612} = 2.57$ |
| 6 | 53,285 | 52,612 | $\frac{(53,285-52,612)^2}{52,612} = 8.61$ |

Calculating the chi-square statistic

| Outcome | Observed | Expected | $\frac{(O-E)^2}{E}$ |
|---------|----------|----------|---|
| 1 | 53,222 | 52,612 | $\frac{(53,222-52,612)^2}{52,612} = 7.07$ |
| 2 | 52,118 | 52,612 | $\frac{(52,118-52,612)^2}{52,612} = 4.64$ |
| 3 | 52,465 | 52,612 | $\frac{(52,465-52,612)^2}{52,612} = 0.41$ |
| 4 | 52,338 | 52,612 | $\frac{(52,338-52,612)^2}{52,612} = 1.43$ |
| 5 | 52,244 | 52,612 | $\frac{(52,244-52,612)^2}{52,612} = 2.57$ |
| 6 | 53,285 | 52,612 | $\frac{(53,285-52,612)^2}{52,612} = 8.61$ |
| Total | 315,672 | 315,672 | 24.73 |

Why square?

Squaring the difference between the observed and the expected outcome does two things:

Why square?

Squaring the difference between the observed and the expected outcome does two things:

- ▶ Any standardized difference that is squared will now be positive.

Why square?

Squaring the difference between the observed and the expected outcome does two things:

- ▶ Any standardized difference that is squared will now be positive.
- ▶ Differences that already looked unusual will become much larger after being squared.

Why square?

Squaring the difference between the observed and the expected outcome does two things:

- ▶ Any standardized difference that is squared will now be positive.
- ▶ Differences that already looked unusual will become much larger after being squared.

When have we seen this before?

The chi-square distribution

- ▶ In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.

The chi-square distribution

- ▶ In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.
- ▶ The chi-square distribution has one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

The chi-square distribution

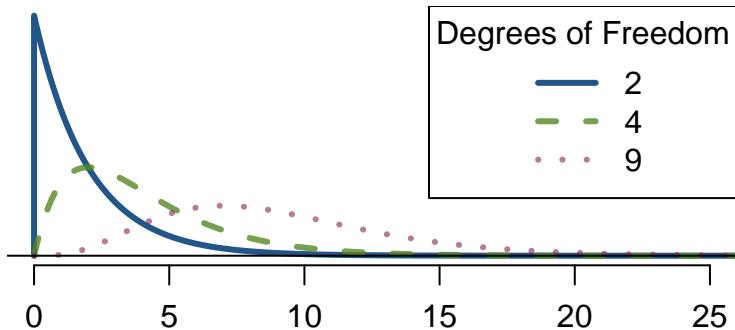
- ▶ In order to determine if the χ^2 statistic we calculated is considered unusually high or not we need to first describe its distribution.
- ▶ The chi-square distribution has one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

Remember: So far we've seen three other continuous distributions:

- ▶ Normal distribution: unimodal and symmetric with two parameters: mean and standard deviation.
- ▶ T distribution: unimodal and symmetric with one parameter: degrees of freedom.
- ▶ F distribution: unimodal and right skewed with two parameters: degrees of freedom or numerator (between group variance) and denominator (within group variance)

Practice

Which of the following is false?

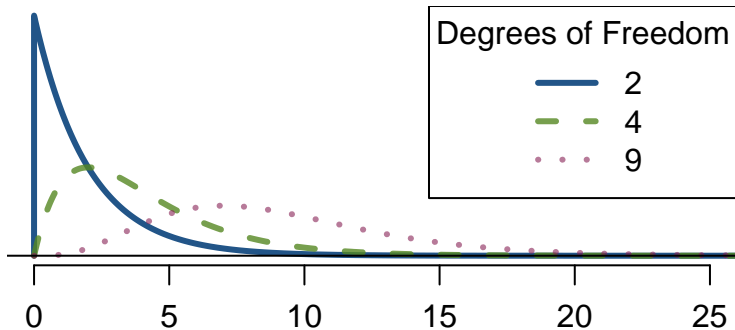


As the df increases,

- A) The center of the χ^2 distribution increases as well.
- B) The variability of the χ^2 distribution increases as well.
- C) The shape of the χ^2 distribution becomes more skewed (less like a normal).

Practice

Which of the following is false?



As the df increases,

- A) The center of the χ^2 distribution increases as well.
- B) The variability of the χ^2 distribution increases as well.
- C) **The shape of the χ^2 distribution becomes more skewed (less like a normal).**

Finding areas under the chi-square curve

- ▶ p-value = tail area under the chi-square distribution (as usual).

Finding areas under the chi-square curve

- ▶ p-value = tail area under the chi-square distribution (as usual).
- ▶ For this we can use technology, or chi-square probability table.

Finding aread under the chi-square curve

Estimate the shaded area (above the cutoff value of 10) under the χ^2 curve with $df = 6$.

Finding area under the chi-square curve

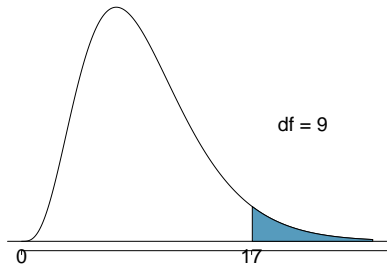
Estimate the shaded area (above the cutoff value of 10) under the χ^2 curve with $df = 6$.

```
pchisq(q = 10, df = 6, lower.tail = FALSE)
```

```
## [1] 0.124652
```

Finding area under the chi-square curve

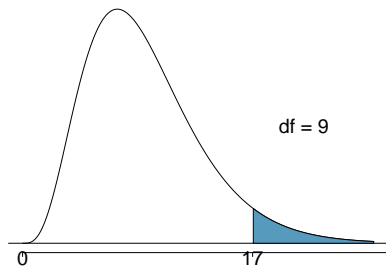
Estimate the shaded area (above the cutoff value of 10) under the χ^2 curve with $df = 6$.



- A) 0.05
- B) 0.02
- C) between 0.02 and 0.05
- D) between 0.05 and 0.1
- E) between 0.01 and 0.02

Finding area under the chi-square curve

Estimate the shaded area (above the cutoff value of 10) under the χ^2 curve with $df = 6$.



A) 0.05

B) 0.02

C) **between 0.02 and 0.05**

D) between 0.05 and 0.1

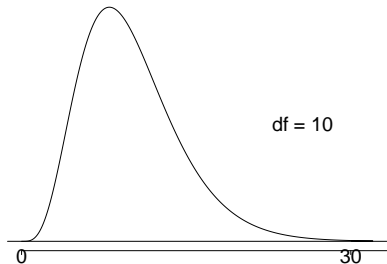
E) between 0.01 and 0.02

```
pchisq(q = 17, df = 9, lower.tail = FALSE)
```

```
## [1] 0.04871598
```

Finding area under the chi-square curve

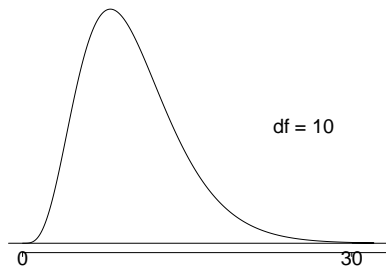
Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



- A) greater than 0.3
- B) between 0.005 and 0.001
- C) less than 0.001
- D) greater than 0.001
- E) cannot tell using this table

Finding area under the chi-square curve

Estimate the shaded area (above 30) under the χ^2 curve with $df = 10$.



- A) greater than 0.3
- B) between 0.005 and 0.001
- C) **less than 0.001**
- D) greater than 0.001
- E) cannot tell using this table

```
pchisq(q = 30, df = 10, lower.tail = FALSE)
```

```
## [1] 0.0008566412
```

Back to Labby's dice

- ▶ The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?

Back to Labby's dice

- ▶ The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- ▶ The hypotheses were:
 - H_0 : There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.
 - H_A : There is an inconsistency between the observed and the expected counts. The observed counts **do not** follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.

Back to Labby's dice

- ▶ The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- ▶ The hypotheses were:
 - H_0 : There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.
 - H_A : There is an inconsistency between the observed and the expected counts. The observed counts **do not** follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.
- ▶ We had calculated a test statistic of $\chi^2 = 24.67$.

Back to Labby's dice

- ▶ The research question was: Do these data provide convincing evidence of an inconsistency between the observed and expected counts?
- ▶ The hypotheses were:
 - H_0 : There is no inconsistency between the observed and the expected counts. The observed counts follow the same distribution as the expected counts.
 - H_A : There is an inconsistency between the observed and the expected counts. The observed counts **do not** follow the same distribution as the expected counts. There is a bias in which side comes up on the roll of a die.
- ▶ We had calculated a test statistic of $\chi^2 = 24.67$.
- ▶ All we need is the df and we can calculate the tail area (the p-value) and make a decision on the hypotheses.

Degrees of freedom for a goodness of fit test

- ▶ When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

$$\mathbf{df = k - 1}$$

Degrees of freedom for a goodness of fit test

- ▶ When conducting a goodness of fit test to evaluate how well the observed data follow an expected distribution, the degrees of freedom are calculated as the number of cells (k) minus 1.

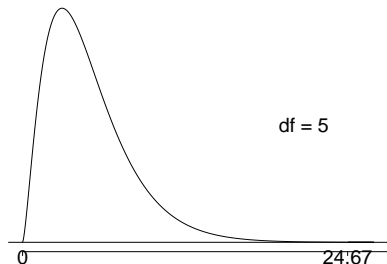
$$df = k - 1$$

- ▶ For dice outcomes, $k = 6$, therefore

$$df = 6 - 1 = 5$$

Finding a p-value for a chi-square test

The **p-value** for a chi-square test is defined as the **tail area above the calculated test statistic**.



p-value = $P(\chi_{df=5}^2 > 24.67)$ is less than 0.001

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. At 5% significance level, what is the conclusion of the hypothesis test?

- A) Reject H_0 , the data provide convincing evidence that the dice are fair.
- B) Reject H_0 , the data provide convincing evidence that the dice are biased.
- C) Fail to reject H_0 , the data provide convincing evidence that the dice are fair.
- D) Fail to reject H_0 , the data provide convincing evidence that the dice are biased.

Conclusion of the hypothesis test

We calculated a p-value less than 0.001. At 5% significance level, what is the conclusion of the hypothesis test?

- A) Reject H_0 , the data provide convincing evidence that the dice are fair.
- B) **Reject H_0 , the data provide convincing evidence that the dice are biased.**
- C) Fail to reject H_0 , the data provide convincing evidence that the dice are fair.
- D) Fail to reject H_0 , the data provide convincing evidence that the dice are biased.

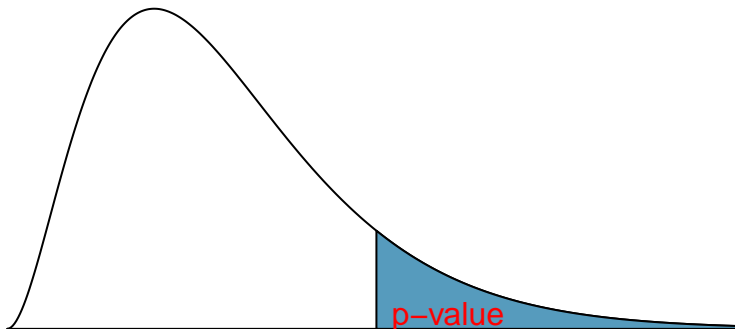
Turns out...

- ▶ The 1-6 axis is consistently shorter than the other two (2-5 and 3-4), thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces.
- ▶ Pearson's claim that 5s and 6s appear more often due to the carved-out pips is not supported by these data.
- ▶ Dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balance.



Recap: p-value for a chi-square test

- ▶ The p-value for a chi-square test is defined as the tail area **above** the calculated test statistic.
- ▶ This is because the test statistic is always positive, and a higher test statistic means a stronger deviation from the null hypothesis.



Conditions for the chi-square test

1. **Independence:** Each case that contributes a count to the table must be independent of all the other cases in the table.

Conditions for the chi-square test

1. **Independence:** Each case that contributes a count to the table must be independent of all the other cases in the table.
2. **Sample size:** Each particular scenario (i.e. cell) must have at least 5 **expected** cases.

Conditions for the chi-square test

1. **Independence:** Each case that contributes a count to the table must be independent of all the other cases in the table.
2. **Sample size:** Each particular scenario (i.e. cell) must have at least 5 **expected** cases.
3. **df > 1:** Degrees of freedom must be greater than 1.

Conditions for the chi-square test

1. **Independence:** Each case that contributes a count to the table must be independent of all the other cases in the table.
2. **Sample size:** Each particular scenario (i.e. cell) must have at least 5 **expected** cases.
3. **df > 1:** Degrees of freedom must be greater than 1.

Failing to check conditions may unintentionally affect the test's error rates.

2009 Iran Election

There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.

| Candidate | Observed # of voters in poll | Reported % of votes in election |
|----------------------|---------------------------------|------------------------------------|
| (1) Ahmedinajad | 338 | 63.29% |
| (2) Mousavi | 136 | 34.10% |
| (3) Minor candidates | 30 | 2.61% |
| Total | 504 | 100% |

2009 Iran Election

There was lots of talk of election fraud in the 2009 Iran election. We'll compare the data from a poll conducted before the election (observed data) to the reported votes in the election to see if the two follow the same distribution.

| Candidate | Observed # of voters in poll | Reported % of votes in election |
|----------------------|---------------------------------|------------------------------------|
| (1) Ahmedinajad | 338 | 63.29% |
| (2) Mousavi | 136 | 34.10% |
| (3) Minor candidates | 30 | 2.61% |
| Total | 504 | 100% |

↓
observed

↓
expected
distribution

Hypotheses

What are the hypotheses for testing if the distributions of reported and polled votes are different?

Hypotheses

What are the hypotheses for testing if the distributions of reported and polled votes are different?

H_0 : The observed counts from the poll follow the same distribution as the reported votes.

H_A : The observed counts from the poll do not follow the same distribution as the reported votes.

Calculation of the test statistic

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|----------------------|------------------------------|---------------------------------|-----------------------------|
| (1) Ahmaddinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

Calculation of the test statistic

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|----------------------|------------------------------|---------------------------------|-----------------------------|
| (1) Ahmaddinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

Calculation of the test statistic

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|----------------------|------------------------------|---------------------------------|-----------------------------|
| (1) Ahmaddinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

Calculation of the test statistic

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|----------------------|------------------------------|---------------------------------|-----------------------------|
| (1) Ahmaddinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_3 - E_3)^2}{E_3} = \frac{(30 - 13)^2}{13} = 22.23$$

Calculation of the test statistic

| Candidate | Observed # of voters in poll | Reported % of votes in election | Expected # of votes in poll |
|----------------------|------------------------------|---------------------------------|-----------------------------|
| (1) Ahmedinajad | 338 | 63.29% | $504 \times 0.6329 = 319$ |
| (2) Mousavi | 136 | 34.10% | $504 \times 0.3410 = 172$ |
| (3) Minor candidates | 30 | 2.61% | $504 \times 0.0261 = 13$ |
| Total | 504 | 100% | 504 |

$$\frac{(O_1 - E_1)^2}{E_1} = \frac{(338 - 319)^2}{319} = 1.13$$

$$\frac{(O_2 - E_2)^2}{E_2} = \frac{(136 - 172)^2}{172} = 7.53$$

$$\frac{(O_3 - E_3)^2}{E_3} = \frac{(30 - 13)^2}{13} = 22.23$$

$$\chi^2_{df=3-1=2} = 30.89$$

Conclusion

Based on these calculations what is the conclusion of the hypothesis test?

- A) p-value is low, H_0 is rejected. The observed counts from the poll do not follow the same distribution as the reported votes.
- B) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.
- C) p-value is low, H_0 is rejected. The observed counts from the poll follow the same distribution as the reported votes.
- D) p-value is low H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

Conclusion

Based on these calculations what is the conclusion of the hypothesis test?

- A) **p-value is low, H_0 is rejected. The observed counts from the poll do not follow the same distribution as the reported votes.**
- B) p-value is high, H_0 is not rejected. The observed counts from the poll follow the same distribution as the reported votes.
- C) p-value is low, H_0 is rejected. The observed counts from the poll follow the same distribution as the reported votes.
- D) p-value is low H_0 is not rejected. The observed counts from the poll do not follow the same distribution as the reported votes.

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 |

| | 4th | 5th | 6th |
|---------|-----|-----|-----|
| Grades | | | |
| Popular | | | |
| Sports | | | |

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

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

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 χ_{df}^2 curve, above the calculated test statistic.

Expected counts in two-way tables

Expected counts in two-way tables

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

Expected counts in two-way tables

Expected counts in two-way tables

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

| | 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

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

| | 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 |

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

Expected counts in two-way tables

Expected counts in two-way tables

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

| | 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 |

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

$$E_{\text{row } 1, \text{col } 2} = \frac{119 \times 141}{478} = 35$$

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

→ 52

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

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

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

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 |

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

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 |

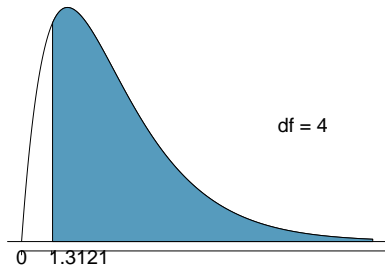
$$\chi^2 = \sum \frac{(63 - 61)^2}{61} + \frac{(31 - 35)^2}{35} + \frac{(32 - 34)^2}{34} = 1.3121$$

$$df = (R - 1) \times (C - 1) = (3 - 1) \times (3 - 1) = 2 \times 2 = 4$$

Calculating the p-value

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

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

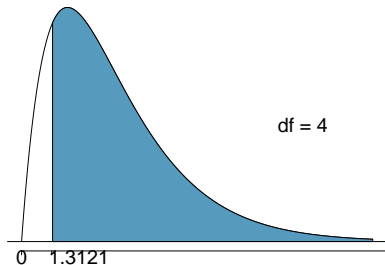


- A) More than 0.3
- B) Between 0.3 and 0.2
- C) Between 0.2 and 0.1
- D) Between 0.1 and 0.05
- E) Less than 0.001

Calculating the p-value

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

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



- 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.