Lecture 19: Inference for Categorical Data

- Statements on hypothesis testing and p-value
- Inference for a single proportion
- Difference of two proportions

Introduction

– Today we will start introducing inference in the setting of categorical data. We use these methods to answer questions like the following:

– What proportions of the American public approves of the job the Supreme Court is doing?

– The Pew Research Center conducted a poll about support for the 2010 health care law, and they used two forms of the survey question. Each respondent was randomly given one of the two questions. What is the difference in the support for respondents under the two question orderings?

– The methods we learned will continue to be useful and the core ideas remain the same.

Statements about Hypothesis Testing

Decide whether the following statements are true or false. Explain your reasoning.
a) A p-value of .08 is more evidence against the null hypothesis than a p-value of .04.
False. A small p-value means the value of the statistic we observed in the sample is unlikely to have occurred when the null hypothesis is true. Hence, a .04 p-value means it is even more unlikely the observed statistic would have occurred when the null hypothesis is true than a .08 p-value. The smaller the p-value, the stronger the evidence against the null hypothesis.

b) The statement, "the p-value is .003", is equivalent to the statement, "there is a 0.3% probability that the null hypothesis is true".
False. The null hypothesis is either true, or it is not true. Hence, the probability that it is true equals either zero or one. The p-value is not interpreted as a probability that the null hypothesis is true. It is the probability of observing a value of the test statistic that is as or more extreme than what was observed in the sample, assuming the null hypothesis is true.

c) Even though you rejected the null hypothesis, it may still be true.
True. Just by chance it is possible to get a sample that produces a small p-value, even though the null hypothesis is true. This is called a Type I error. A Type II error is when the null hypothesis is not rejected when it is in fact false.

d) A researcher who tried to learn statistics without taking a formal course does a hypothesis test and gets a p-value of .014. He says, "there is a 98.6% chance that the alternative hypothesis is true, so the null hypothesis is false." What, if anything, is wrong with his statement?
False. The researcher is claiming that (1 - p-value) is the probability that the alternative hypothesis is true. The p-value is not a probability of an alternative (or null) hypothesis being true or false. See the answer to part b.

e) You perform a hypothesis test using a sample size of four units, and you do not reject the null hypothesis. Your research colleague says this statistical test provides conclusive evidence that the null hypothesis is true. Do you agree or disagree with his conclusion? Explain your reasoning in three or less sentences.

With four units, the null hypothesis is unlikely to be rejected because the variability in the sample mean will be large, i.e. the standard error will be large. Hence, all we can say is that there is not enough data to determine whether or not the null hypothesis is correct. In fact, whenever you fail to reject a null hypothesis, essentially you are saying that the evidence in the data does not contradict the null hypothesis. You never can conclude from a hypothesis test that the null hypothesis is true.

f) If you get a p-value of 0.13, it means that when the null hypothesis is true, a value of the test statistic as or more extreme than what was observed occurs in about 13% of all samples.

True. This is a re-expression of the definition of p-values. That is, saying there is a 13% chance of observing results as or more extreme than what was observed is equivalent to saying that you'd observe results as or more extreme than what was observed in 13% of (random) samples.

g) You are the head of the Food and Drug Administration (F.D.A.), in charge of deciding whether new drugs are effective and should be allowed to be sold to people. A pharmaceutical company trying to win approval for a new drug they manufacture claims that their drug is better than the standard drug at curing a certain disease. The company bases this claim on a study in which they gave their drug to 1000 volunteers with the disease. They compared these volunteers to a group of 1000 hospital patients who were treated with the standard drug and whose information is obtained from existing hospital records. The company found a "statistically significant" difference between the percentage of volunteers who were cured and the percentage of the comparison group who were cured. That is, they did a statistical hypothesis test and rejected the null hypothesis that the percentages are equal. As director of the F.D.A., should you permit the new drug to be sold? Explain your reasoning in three or less sentences.

You should not allow the drug to be manufactured based on this evidence. The study was not a randomized study, which means there may be differences in the background charcteristics of the people who got the new drug and the people who got the old drug. Hypothesis tests cannot fix poorly designed studies.

h) If you get a p-value of 0.13, it means there is a 13% chance that the sample average equals the population average.
False. In fact, it's almost guaranteed that the sample average won't exactly equal the population average, because the process of taking a random sample guarantees variability in the sample average. The p-value does not give the probability that the sample average equals the population average. See part b for the precise definition of p-values.

i) If you get a p-value of 0.13, it means there is a 13% chance that the sample average does not equal the population average.
False. See the answer to part h.

j) If you get a p-value of 0.13, it means there is an 87% chance that the sample average equals the population average.
False. Computing (1 - p-value) does not give the probability that the sample average equals the population average. See the answer to part h.

k) If you get a p-value of 0.13, it means there is an 87% chance that the sample average does not equal the population average.
False. See the answer to part h.

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

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

$$p \ \left(\text{a population proportion}\right)$$

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

$$\hat{p} \ \left(\text{a sample proportion}\right)$$

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

$$point\ estimate \pm ME$$

- And we also know that $ME = critical\ value \times standard\ error$ of the point estimate.

$$SE_{\hat{p}} = ?$$

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 mean, p, and standard error equal to $\sqrt{\frac{p\,(1-p)}{n}}$.

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

- 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 (85%) of 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.

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

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

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%.

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

$$\begin{aligned}
0.01 &\geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Use estimate for } \hat{p} \textit{ from previous study} \\
0.01^2 &\geq 1.96^2 \times \frac{0.85 \times 0.15}{n} \\
n &\geq \frac{1.96^2 \times 0.85 \times 0.15}{0.01^2} \\
n &\geq 4898.04 \rightarrow n \textit{ should be at least 4,899}
\end{aligned}$$

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

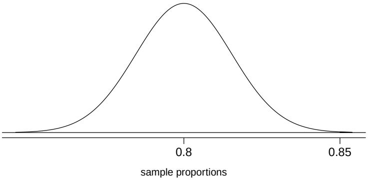
CI vs. HT for proportions

- Success-failure condition:
  - 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: $SE = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

The GSS found that 571 out of 670 (85%) of 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?

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

$$\begin{aligned} SE &= \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154 \\ Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \\ p - value &= 1 - 0.9994 = 0.0006 \end{aligned}$$



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.

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 ±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

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 ±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

Examples

From decades of experience, the statistics department knows that 30% of students fall asleep in class. One of the professors wants to prove that his teaching is more lively. Suppose he collects data and out of 120 students, only 35 fall asleep. Is this evidence that his teaching is better?

First, we must find the null and alternative hypotheses. We put what he wants to show in the alternative.

$$\mathbf{H}_0\textbf{:}\ p \geq 0.3 \qquad \textbf{vs.} \qquad \mathbf{H}_A : p < 0.3$$

Next, we must find his test statistic, the one-number summary of all the information in the sample regarding the null hypothesis.

$$ts = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{35}{120} - .3}{\sqrt{\frac{.3*.7}{120}}} = -.1992.$$

Examples

Finally, we need to find the significance probability, or P-value. To determine the significance probability, from the standard normal table, the P-value is:

$$\mathbf{P}[z < ts] = \mathbf{P}[z < -.1992] = .4211.$$

This result is not unlikely. Just by chance, 42% of the time he would get a result like this if his teaching were no better than anyone else's.

If $\alpha = 0.1$, we fail to reject the null hypothesis and conclude that his teaching were no better than anyone else's.

To determine the significance probability, we are trying to decide whether our sample statistic is improbably different from the null, improbably larger than the null, or improbably smaller than the null, respectively in the three different kinds of hypothesis testing setup.

Summary - inference for one proportion

- Population 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

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 2010 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 $Z = \frac{point\ estimate - null\ value}{SE}$ 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

1. *Independence within groups:*
   - The US group is sampled randomly and we're assuming that the Duke group represents a random sample as well.
   - $n_{Duke} < 10\%$ 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.
2. *Independence between groups:* The sampled Duke students and the US residents are independent of each other.
3. *Success-failure:*

   At least 10 observed successes and 10 observed failures in the two groups.

Construct a 95% confidence interval for the difference between the proportions of Duke students and Americans who would be bothered a great deal by the melting of the northern ice cap $(p_{Duke} - p_{US})$.

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

$$
\begin{aligned}
& (\hat{p}_{Duke} - \hat{p}_{US}) \pm z^{\star} \times \sqrt{\frac{\hat{p}_{Duke}(1-\hat{p}_{Duke})}{n_{Duke}} + \frac{\hat{p}_{US}(1-\hat{p}_{US})}{n_{US}}} \\
= \quad & (0.657 - 0.668) \pm 1.96 \times \sqrt{\frac{0.657 \times 0.343}{105} + \frac{0.668 \times 0.332}{680}} \\
= \quad & -0.011 \pm 1.96 \times 0.0497 \\
= \quad & -0.011 \pm 0.097 \\
= \quad & (-0.108, 0.086)
\end{aligned}
$$

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) $H_0 : p_{Duke} = p_{US}$
$H_A : p_{Duke} < p_{US}$

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.

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

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

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

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 calculated 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

$$\hat{p} = \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2}$$

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

$$\begin{aligned}\hat{p} &= \frac{\#\, of\ successes_1 + \#\, of\ successes_2}{n_1 + n_2} \\ &= \frac{69 + 454}{105 + 680} = \frac{523}{785} = 0.666\end{aligned}$$

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

$$
\begin{aligned}
Z &= \frac{(\hat{p}_{Duke} - \hat{p}_{US})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_{Duke}} + \frac{\hat{p}(1-\hat{p})}{n_{US}}}} \\
&= \frac{(0.657 - 0.668)}{\sqrt{\frac{0.666 \times 0.334}{105} + \frac{0.666 \times 0.334}{680}}} = \frac{-0.011}{0.0495} = -0.22 \\
p - value &= 2 \times P(Z < -0.22) = 2 \times 0.41 = 0.82
\end{aligned}
$$

Summary - 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 $H_0 : p_1 - p_2 =$ *(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	$SE = \sqrt{\frac{p(1-p)}{n}}$	$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- 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

Recap

Today we covered

1. Some common misinterpretations of confidence intervals
2. Inference for a single proportion
3. Inference of the difference of proportions

Suggested reading:

- OpenIntro3: Sec. 6.1, 6.2