# Chapter 6: Inference for categorical data

Yingbo Li

Southern Methodist University

STAT 2331

# Outline

1. Single population proportion

2. Difference of two proportions

## Question

1. 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 General Social Survey (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 a 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 a good intuition about experimental design.

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

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

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

# Inference on a proportion

What percent of all Americans have a 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

- Then, according to the CLT:

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

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

Any guesses?

Assumptions/conditions:

1. *Independence*:
   - ★ *Random sample*
   - ★ *10% condition*: If sampling without replacement, $n <$ 10% of the population.
2. *Normality*: 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 a good intuition about experimental design?

Given: $n = 670, \hat{p} = 0.85$. First check assumptions & 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. *Normality*: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

Question

2. 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%.

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

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

# 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 $n$

## Question

3. Which of the following are the correct set of hypotheses for testing if more than 80% of Americans have a good intuition about experimental design?

(a) $H_0 : \mu = 0.80$
$H_A : \mu > 0.80$

(b) $H_0 : p = 0.85$
$H_A : p > 0.85$

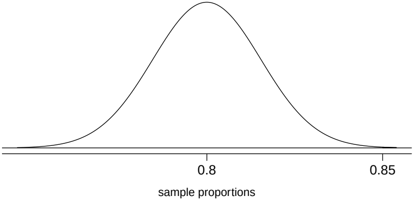
(c) $H_0 : p = 0.80$
$H_A : p > 0.80$

(d) $H_0 : \hat{p} = 0.80$
$H_A : \hat{p} > 0.80$

# Hypothesis testing for a proportion

$$\hat{p} \sim N\left(mean = 0.80, SE = \sqrt{\frac{0.80 \times 0.20}{670}} = 0.0154\right)$$

*Note:* *The SE is different, because now we are conducting a hypothesis test assuming $H_0$ is true, and $H_0$ says $p = 0.80$.*



$$\begin{aligned} Z &= \frac{0.85 - 0.80}{0.0154} = 3.25 \\ \text{p-value} &= 1 - 0.9994 = 0.0006 \end{aligned}$$

Since 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.

# Recap: inference for a single population proportion

1. Confidence interval
   - Parameter of interest: $p$
   - Point estimate: $\hat{p}$,
   - $SE = \sqrt{\frac{\hat{p}\ (1-\hat{p})}{n}}$
   - Assumptions:
     (1) independence
     (2) $n\hat{p} > 10$ and $n(1-\hat{p}) > 10$
2. Hypothesis testing
   - Compare $p$ with $p_0$
   - $SE = \sqrt{\frac{p_0\ (1-p_0)}{n}}$
   - Assumptions:
     (1) independence
     (2) $np_0 > 10$ and $n(1-p_0) > 10$

# Melting ice cap

## Question

4. 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 is the distribution of responses from the 2010 survey:

| | |
|---|---|
| A great deal | 454 |
| Some | 124 |
| A little | 52 |
| Not at all | 50 |
| Total | 680 |

Question

5. Which of the following is the correct set of hypotheses for testing if the proportion of college 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) $H_0 : p_{coll} = p_{US}$
$H_A : p_{coll} \neq p_{US}$

(b) $H_0 : \hat{p}_{coll} = \hat{p}_{US}$
$H_A : \hat{p}_{coll} \neq \hat{p}_{US}$

(c) $H_0 : p_{coll} - p_{US} = 0$
$H_A : p_{coll} - p_{US} \neq 0$

(d) $H_0 : p_{coll} = p_{US}$
$H_A : p_{coll} < p_{US}$

*Both (a) and (c) are correct.*

## Parameter and point estimate

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

$$p_{coll} - p_{US}$$

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

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

# Exploratory analysis

| | College students | US |
|---:|:---:|:---:|
| # of successes | 58 | 454 |
| n | 85 | 680 |
| $\hat{p}$ | 0.682 | 0.668 |

## Checking assumptions & conditions

1. *Independence within groups:*
   - The US group is sampled randomly and we're assuming that the college group represents a random sample as well.
   - $85 < 10\%$ of all college students and $680 < 10\%$ of all Americans.

   We can assume that the attitudes of college 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 college students and the US residents are independent of each other.
3. *Normality:*

   We need at least 10 *expected* successes and 10 *expected* failures in the two groups.

# 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.

$$np \geq 10 \qquad\qquad n(1-p) \geq 10$$

In the above formula $p$ comes from the null hypothesis.

# Finding expected number of successes when comparing two population proportions

- Similar to the one sample case, when constructing a confidence interval for the difference between two population proportions, we check if the *observed* number of successes in each group and failures are at least 10.

$$
\begin{aligned}
&Group1 : n_1\hat{p}_1 \geq 10 &\qquad& n_1(1-\hat{p}_1) \geq 10\\
&Group2 : n_2\hat{p}_2 \geq 10 &\qquad& n_2(1-\hat{p}_2) \geq 10
\end{aligned}
$$

- But in this case the null hypothesis simply states the population proportions are equal to each other, and doesn't set them equal to a given value.

$$H_0 : p_1 = p_2$$

  Then, we need to first find a common proportion for the two groups, and use that in our analysis.

# Pooled estimate of a proportion

- Since $H_0$ implies that both samples come from the same population, we pool the two samples to calculate a *pooled* estimate of the sample proportion.
- This simply means finding the proportion of total successes among the total number of observations.

### Pooled estimate of a proportion

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

# Pooled estimate of a proportion - in context

| | College students | US |
|---|---|---|
| # of successes | 58 | 454 |
| n | 85 | 680 |
| $\hat{p}$ | 0.682 | 0.668 |

$$\begin{aligned}\hat{p} &= \frac{\#\ of\ successes_1 + \#\ of\ successes_2}{n_1 + n_2} \\ &= \frac{58 + 454}{85 + 680} \\ &= \frac{512}{765} = 0.669\end{aligned}$$

Why is the pooled estimate closer to $\hat{p}_{US}$ than $\hat{p}_{coll}$?

# Number of expected successes and failures

College students

$$\begin{aligned} 85 \times 0.669 &= 56.865 \\ 85 \times (1 - 0.669) &= 28.135 \end{aligned}$$

US

$$\begin{aligned} 680 \times 0.669 &= 454.92 \\ 680 \times (1 - 0.669) &= 225.08 \end{aligned}$$

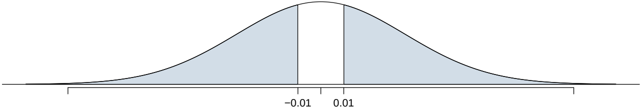
There are at least 10 expected successes and 10 expected failures in both groups. So we can proceed with the test.

# The hypothesis test

- Hypotheses: $H_0 : p_{coll} = p_{US}$; $H_A : p_{coll} \neq p_{US}$
- Assumptions and conditions are satisfied
- Test statistic:

$$
\begin{aligned}
Z &= \frac{(0.682 - 0.668) - 0}{\sqrt{\frac{0.669 \times (1-0.669)}{85} + \frac{0.669 \times (1-0.669)}{680}}} \\
&= \frac{0.014}{0.054} \\
&= 0.26
\end{aligned}
$$

- p-value: $2 \times P(Z > 0.26) = 2 \times (1 - 0.6026) = 0.7948$



What's the conclusion?

# The 95% confidence interval

$$
\begin{aligned}
\textit{point estimate} \quad &\pm \quad ME \\
(0.682 - 0.668) \quad &\pm \quad 1.96 \times \sqrt{\frac{\color{red}{\mathit{0.682}} \times (1 - 0.682)}{85} + \frac{\color{red}{\mathit{0.668}} \times (1 - 0.668)}{680}} \\
0.014 \quad &\pm \quad 1.96 \times 0.0536 \\
0.014 \quad &\pm \quad 0.105 \\
(-0.091 \quad &, \quad 0.119)
\end{aligned}
$$

Question

6. Which of the below is the correct interpretation of this 95% confidence interval?

$$p_{coll} - p_{US} = (-0.091, 0.119)$$

We are 95% confident that college students who would be bothered a great deal about the melting of the northern ice cap are

(a) 9.1% to 11.9% lower

(b) 9.1% to 11.9% higher

(c) *9.1% lower to 11.9% higher*

(d) 9.1% higher to 11.9% lower

than those in the US population.

# Recap - comparing two proportions

- Population parameter: $p_1 - p_2$, point estimate: $\hat{p}_1 - \hat{p}_2$
- Assumptions and conditions:
  - independence within groups (random sample and 10% condition met for both groups)
  - independence between groups
  - at least 10 successes and failures
- $SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
- *Only* when conducting a hypothesis test where $H_0 : p_1 = p_2$
  - Pooled proportion: $\hat{p} = \frac{\#\ suc_1 + \# suc_2}{n_1 + n_2}$
  - Use the pooled proportion for calculating expected number of successes and failures, and in the calculation of the standard error

# Reference - standard error calculations

| | one sample | two samples | distribution |
|---|---|---|---|
| mean | $SE = \frac{s}{\sqrt{n}}$ | $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ | T, or<br>normal $(n \geq 30)$ |
| proportion | $SE = \sqrt{\frac{p(1-p)}{n}}$ | $SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ | Normal |

- When working with means, it's very rare that $\sigma$ 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