Unit 3: Inference for Categorical and Numerical Data

1. Inference for a Single Proportion

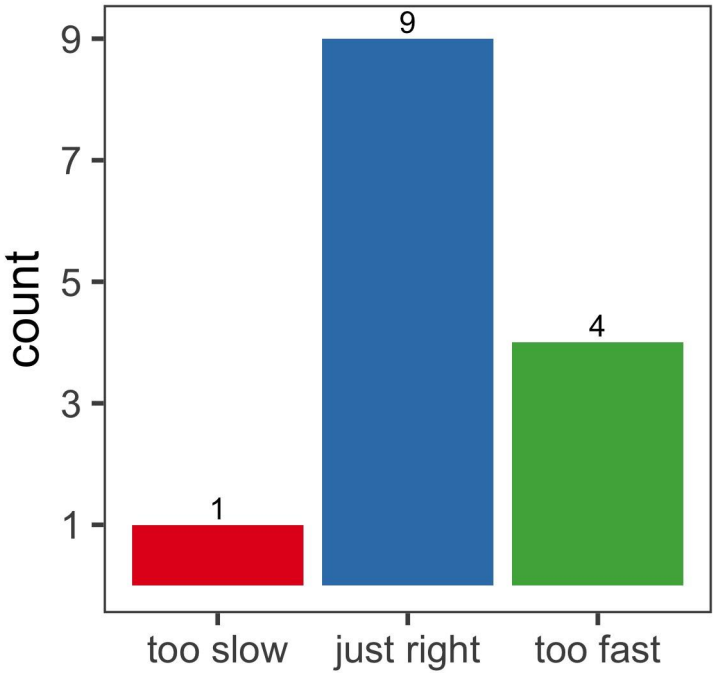
2/17/2020

Results of the Eberly Early Course Survey

14 total responses (21 total students)

My impression:

We're probably moving a little too fast.
(this seems to come more from labs)



The Good

Lecture

- Organization, availability of slides, variety
  ability to ask questions

Examples in lecture

- Simulations, R

Quizzes

- Frequent feedback, going over the expected answers

The Bad

More office hours, closer to deadlines

- Can do!

More guidance in labs

- Walk through what the code does
- Check in about labs in class
- Leave exercise solutions up for longer
- Talk about on your own questions
- Hard to learn R and concepts at the same time

Recap from last time

1. Statistical inference methods based on the CLT depend on the same conditions as the CLT

2. We can use confidence intervals to estimate population parameters

3. Critical values depend on the confidence interval

Key ideas

1. We can use the CLT to make inferences about proportions

2. Confidence intervals can be used to make inferences about a population proportion

3. Confidence intervals can be used to do hypothesis tests

Practice 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 people get the drug
(b) 500 people get the drug, 500 don't

Practice 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 people get the drug

(b) 500 people get the drug, 500 don't

Results from the NSF SEEI2012

National Science Board
SCIENCE & ENGINEERING INDICATORS 2016

The National Science Foundation asked this question as part of a survey on general scientific literacy in 2010. Here are the results:

All 1000 get the drug	99
500 get the drug 500 don't	571
Total	670

Estimating the population parameter

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 for a proportion

What proportion of Americans would answer “500 get the drug 500 don't?”

Let’s try to answer this question with a confidence interval.
We know this should have the form:
point estimate ± margin of error

point estimate ± Standard Error(SE) x Critical value (Z)*

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

Sample proportions are also nearly Normally Distributed

The Central Limit theorem for proportions says that the sample proportion will be nearly normal with mean equal to the population mean p and standard error $\sqrt{\frac{p(1-p)}{n}}$

$$\hat{p} \sim Normal\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Only holds under the assumptions of the Central Limit Theorem:

- Independent samples
- N large enough (~10 success, ~10 failures)

Let’s estimate the population proportion

The SEEI found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Let’s 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: We’re told the sample is random
Success-failure: 571 people answered correctly (successes) and 99 answered incorrectly (failures), both are greater than 10.

Practice Question 2

We are given that $n = 670$ and $\hat{p} = 0.85$.
We just learned that the standard error of the sample proportion is $\sqrt{\frac{p(1-p)}{n}}$

Which of the following is the correct calculation of the 95% confidence interval?

(a) $.85 \pm 1.96 \times \sqrt{\frac{.85 \times .15}{670}}$

(b) $.85 \pm 1.65 \times \sqrt{\frac{.85 \times .15}{670}}$

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

(d) $.85 \pm 1.96 \times \frac{.85 \times .15}{\sqrt{670}}$

Practice Question 2

We are given that $n = 670$ and $\hat{p} = 0.85$.
We just learned that the standard error of the sample proportion is $\sqrt{\frac{p(1-p)}{n}}$

Which of the following is the correct calculation of the 95% confidence interval?

(a) $.85 \pm 1.96 \times \sqrt{\frac{.85 \times .15}{670}}$

(b) $.85 \pm 1.65 \times \sqrt{\frac{.85 \times .15}{670}}$

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

(d) $.85 \pm 1.96 \times \frac{.85 \times .15}{\sqrt{670}}$

Choosing a sample size

Our 95% CI with n = 670 is (.82, .88).
What if we want our 95% to be more precise?

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

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

use $\hat{p}$ from the previous study

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

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

$$n \geq 4898.04$$

But what if there is no previous study?

Use $\hat{p} = 0.5$. But why?

$\hat{p} = 0.5$ gives the most conservative estimate.

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

Is largest when $p = .5$.

Hypothesis testing using CIs with Proportions

"Do more people get the question right than we would expect by chance?"

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: $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
- HT: calculate using the null value: $SE_{p_0} = \sqrt{\frac{p_0(1-p_0)}{n}}$

Hypothesis testing for the SEEI data

“Do more people get the question right than we would expect by chance?”

“Do more people get the question right than we expected?

Practice Question 3

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) Can’t tell

Practice Question 3

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) Can’t tell

Key ideas

1. We can use the CLT to make inferences about proportions

2. Confidence intervals can be used to make inferences about a population proportion

3. Confidence intervals can be used to do hypothesis tests