Sta 111 - Summer II 2017
Probability and Statistical Inference

13. Inference for a single proportion

Lu Wang

Duke University, Department of Statistical Science

July 23, 2017

Outline

1. Confidence intervals for a proportion
    1. Central limit theorem for proportions

2. Hypothesis testing for a proportion

3. Only use CLT based methods if the sample size is large enough for a nearly normal sampling distribution

We would like to estimate the proportion of all Americans who have good intuition about experimental design. A 2010 General Social Survey (GSS) asks the following question.

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

Parameter and point estimate

Below is the distribution of responses from the GSS results:

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

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.

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

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

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

- We can do inference on the proportion using a confidence interval, which we know is always of the form

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

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

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

Standard error of a sample proportion

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

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

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?

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 a 95% confidence interval for 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 this condition is met.

2. *Success-failure (S-F)*: 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) $571 \pm 1.96 \times \sqrt{\frac{571 \times 99}{670}}$

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

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

$$0.01 \geq 1.96 \times \sqrt{\frac{0.85 \times 0.15}{n}} \rightarrow \textit{Plug in p or } \hat{p}$$

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

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

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

CI vs. HT for proportions

Remember, when doing a HT always start with assuming H_0 is true!

- Success-failure condition:
  - CI: At least 10 *observed* successes and failures
  - HT: At least 10 *expected* successes and failures, calculated using the null value: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

- Standard error:
  - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

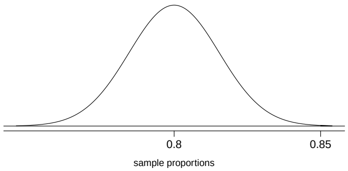
The GSS found that 571 out of 670 (85%) of Americans answered the question on experimental design correctly. Do these data provide convincing evidence that more than 80% of Americans have a good intuition about experimental design?

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

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

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

$$p - value = 1 - 0.9994 = 0.0006$$



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

- If the S-F condition is met, can do theoretical inference: Z test, Z interval

- If the S-F condition is not met, must use simulation based methods: randomization test, bootstrap interval
  - *HT*: Randomization test - simulate under the assumption that H_0 is true, then find the p-value as proportion of simulations where the simulated $\hat{p}$ is at least as extreme as the one observed.

  - *CI*: Bootstrap interval - resample with replacement from the original sample, and construct interval using percentiles.

Recap - 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 / bootstrap
- Standard error: $SE = \sqrt{\frac{p(1-p)}{n}}$
  - for CI: use $\hat{p}$
  - for HT: use p_0