

ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

HYPOTHESIS TESTING FOR A PROPORTION

PART 1

UC San Diego

COMPUTER SCIENCE & ENGINEERING
HALICIOĞLU DATA SCIENCE INSTITUTE

Hypothesis Testing

- The Hypothesis testing framework allows us to formally evaluate claims about a population, such as whether a survey provides strong evidence that a candidate has the support of a majority of the voting population.
- Hypothesis testing usually involves a
 - **Null Hypothesis** (denoted H_0), representing a “skeptical” perspective, and an
 - **Alternative Hypothesis** (denoted H_1 or H_A), representing an alternative claim being considered.
- Rough Analogy: In the US court system, a defendant is presumed innocent (Null Hypothesis) until proven guilty (Alternative Hypothesis).

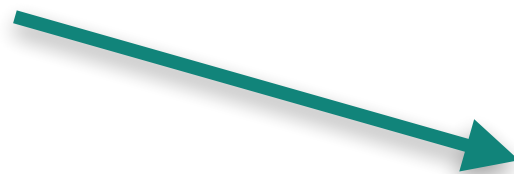
Hypothesis Testing

- **Example:** We would like to determine if vitamin C can cure the common cold. In this case, our null and alternative hypotheses **could be**
 - H_0 : People who take vitamin C are **no less likely** to become ill
 - H_A : People who take vitamin C are **are less likely** to become ill
- **Example:** We would like to determine if children of professional athletes are more likely to become professional athletes.
 - What should our H_0, H_A be?

Hypothesis Testing

- The null and alternative hypotheses are contradictory, by design.
- Our goal is to decide whether we have enough evidence to reject the null hypothesis or not.
- Our evidence comes from sample data.
- After examining the evidence, we have two options:
 - "**reject H_0** " if the sample information favors the alternative hypothesis
 - "**do not reject H_0** " or "**decline to reject H_0** " if the sample information is not sufficient to reject H_0 .

Testing Hypothesis Using Confidence Intervals

- **Example:** A multiple choice question has 4 options, and the correct answer is A.
- **Hypotheses:**
 - H_0 : Adults are as accurate as random guessing, i.e., the proportion of adults who get the question right is $p = 0.25$
 - H_A : $p \neq 0.25$.

Null value, often denoted p_0
- **Remark:** Even though it seems the null-hypothesis is very precise and easy to reject, the hypothesis testing framework requires that there be strong evidence before we reject the null hypothesis

Testing Hypothesis Using Confidence Intervals

- **Example (continued):** We randomly sample 100 adults and ask them the multiple choice question. We find that 21 % of them answered correctly.
- This is a sample proportion! $\hat{p} = 0.21$
- We'll use confidence intervals to decide if the difference between 0.21 and 0.25 is due to chance, or if the data provides strong evidence that the population proportion is different from 0.25.
- **CI calculation at the 95% level:** Samples are independent, success/failure condition holds (check this!)

$$\hat{p} = 0.21, \quad SE = \sqrt{\frac{0.21(1 - 0.21)}{100}} = 0.0407$$
$$I = (\hat{p} - 1.96 \cdot SE_{\hat{p}}, \quad \hat{p} + 1.96 \cdot SE_{\hat{p}}) = (0.1302, \quad 0.2898)$$

Testing Hypothesis Using Confidence Intervals

- **Example (continued):** $I = (0.1302, 0.2898)$
- So we are 95% confident that the proportion of all adults to correctly answer this multiple choice question is between 13.02% and 28.98%.
- Notice that $p_0 = 0.25$ falls within the confidence interval.
- So the data does not provide enough evidence to reject the null hypothesis (that the performance of adults on this multiple choice question is better than random guessing).
- **We do not reject H_0**
- **Note/Reminder:** Not rejecting H_0 does not mean we are saying it is correct.

Testing Hypothesis Using Confidence Intervals

- **Example (Modified):** We randomly sample 100 adults and ask them a multiple choice question (recall there were 4 options). We find that 37 % of them answered correctly.
- **CI calculation at the 95% level:** Samples are independent, success/failure condition holds (check this!)

$$\hat{p} = 0.37, \quad SE = \sqrt{\frac{0.37(1 - 0.37)}{100}} = 0.0483$$

$$I = (\hat{p} - 1.96 \cdot SE_{\hat{p}}, \quad \hat{p} + 1.96 \cdot SE_{\hat{p}}) = (0.2754, \quad 0.4646)$$

- *This time p_0 does not fall within the 95% confidence interval. In fact the interval is greater than p_0 .*
- *We can reject the null-hypothesis, and conclude (at the 95% confidence level) that adults do better than random guessing on this question.*

Testing Hypothesis Using Confidence Intervals

Remarks and reminders:

- Different confidence level (e.g. 99.9%, 99%, 95%) could well give different results
 \implies it is important to state the confidence level
- For a fixed confidence level, a larger sample size (n) gives a smaller confidence interval.
- Failing to reject the null hypothesis does not mean that we are saying it is correct.