STAT 252

Week 1: Review

2024-04-01

Day Two

Purpose of Statistics: Make inferences about a population from a sample

---

Statistical Question: A question that can be answered by collecting data that varies

For example...

How old am I?

(not statistical)

How old are Cal Poly students on average?

(statistical)

How tall is Jamear?

(not statistical)

How tall is the average 12-year-old in the US?

(statistical)

Let's Practice

Statistical or Not?

1. **How many hours per week do students spend studying for exams?**
2. **How many siblings do you have?**
3. **What was the temperature at 12pm today?**

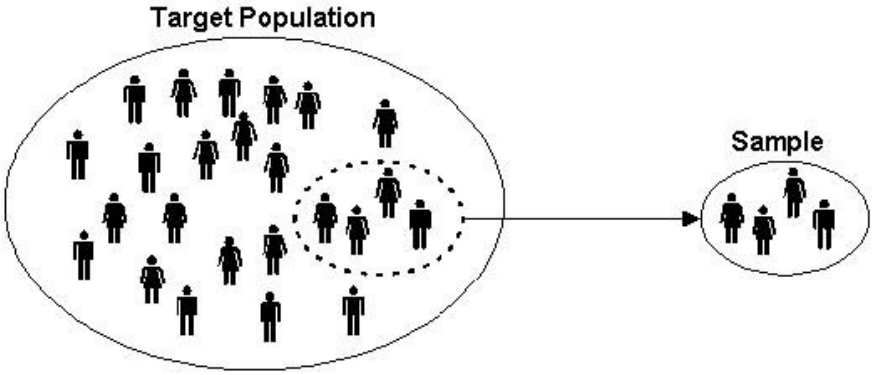
Classifying Units of Study

Population: The entire group of study from which a sample is drawn

Sample: Part of the population from which data is gathered

Observational Unit: A single person, place, or thing from the sample

Sample Size: Total number of observational units in a sample (denoted n)



For example...

Population, Sample, Observational Unit: Cal Poly Students

Population: All students at Cal Poly

Sample: Students currently enrolled in a STAT 252 course

Observational Unit: A Cal Poly student

Population, Sample, Observational Unit: Sodas in college dining halls

Population: All sodas served in college dining halls across the US

Sample: Sodas served at Cal Poly's dining halls

Observational Unit: A soda

Let's Practice

State the population, sample, and observational unit.

Statistical Question: According to current Cal Poly students, what dining hall venue is the best?

Population:

Sample:

Observational Unit:

Statistical Question: What proportion of Cal Poly students hand-write versus type their notes?

Population:

Sample:

Observational Unit:

Statistical Question: How tall are buildings in SLO on average?

Population:

Sample:

Observational Unit:

Variables

Variable: A characteristic that can be measured and assume different values

Categorical: Variable that takes on three or more category designations

For example...

1. **What is the most common hair color at Cal Poly?**

- Variable of Interest: Hair Color
- Variable Type: Categorical

2. **What is the least liked ice cream flavor among Cal Poly students?**

- Variable of Interest: Ice Cream Flavor
- Variable Type: Categorical

Categorical Binary: Variable that takes on two category designations

For example...

1. **What proportion of US citizens are married?**

- Variable of Interest: Marital Status
- Variable Type: Categorical Binary

2. **What percentage of Cal Poly students have a job?**

- Variable of Interest: Employment Status
- Variable Type: Categorical Binary

Quantitative: Variable that takes on a continuous range of numerical values

For example...

1. **What is the temperature on average in a Cal Poly dorm?**

- Variable of Interest: Temperature
- Variable Type: Quantitative

2. **How heavy is a student's backpack on average?**

- Variable of Interest: Weight
- Variable Type: Quantitative

Let's Practice

State the variable of interest and its type (Categorical, Categorical Binary, Quantitative)

1. **How many students at Cal Poly participate in sports?**

- Variable of Interest:
- Variable Type:

2. **What mode of transportation (bus, Uber, personal vehicle) is most commonly used by Cal Poly students to go downtown?**

- Variable of Interest:
- Variable Type:

3. **What is the average monthly revenue for a small business in SLO?**

- Variable of Interest:
- Variable Type:

Variables (cont.)

Explanatory Variable: Variable that explains variation in the response variable

Response Variable: Variable of interest that the outcome of a study measures

For example...

1. Does going to office hours affect a student's performance on a test?

- Explanatory Variable: Whether a student went to office hours
- Response Variable: Score on test

2. Does eating before or after a workout allow you to squat more?

- Explanatory Variable: Whether you ate before or after your workout
- Response Variable: Squat weight

Let's Practice

State the explanatory and response variable for each scenario.

1. How does the type of fertilizer used affect plant growth?

- Explanatory Variable:
- Response Variable:

2. Does the type of exercise regimen (aerobic or strength training) impact weight loss?

- Explanatory Variable
- Response Variable:

Day Three

Purpose of Statistics: Make inferences about a population from a sample

---

Parameter v. Statistics

Proportion vs. Mean

Parameter vs. Statistic

Symbols

Types of Study Designs

Observational Study

Experimental Design

Importance of Randomness

Review of Statistical Inference

Two Types (goals of each)

Async Work

Video 1: One-Population Hypothesis Test (Mean, Proportion)

On top of basic instruction / questions, I will include space for your R output, visualization, and analysis

Video 2: Confidence Interval (Mean, Proportion)

On top of basic instruction / questions, I will include space for your R output, visualization, and analysis

Video 3: Types of Errors (I, II), Types of Distributions, and Confidence Interval Manipulations (widening, sample size, etc)

Note: The content below could either be deleted other than the glossary once I put everything together or kept as a review of the week for students

Statistical Techniques

1. One-Population Hypothesis Test

- If TS < -3.5 OR TS > 3.5, we reject H_0
- If p-value < 0.1, we reject H_0 (calculator needed to find p-value)
- Type I Error: Rejecting a null hypothesis that is true
- Type II Error: Failing to reject a null hypothesis that is false

One-Population Mean Hypothesis Test

Assumptions: Data is random; observations are independent

Hypotheses for Two-Tailed Test:

- Null Hypothesis:
  - Words: The population mean of [context] is equal to μ_o
  - Symbols: H_0: $\mu = \mu_o$
- Alternative Hypothesis:

  - Words: The population mean of [context] is not equal to μ_o
  - Symbols: H_1: $\mu \neq \mu_o$

Test Statistic: $TS = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

One-Population Proportion Hypothesis Test

Assumptions: Data is random; $np \geq 10$ & $n(1 - p) \geq 10$

Hypotheses for Two-Tailed Test:

- Null Hypothesis:
  - Words: The population proportion of [context] is equal to p_o
  - Symbols: H_0: $p = p_o$
- Alternative Hypothesis:
  - Words: The population proportion of [context] is not equal to p_o
  - Symbols: H_1: $p \neq p_o$

Test Statistic: $TS = \frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$

2. One-Population Confidence Intervals

- Critical Value: Number of standard errors from the parameter needed to achieve a certain level of confidence (CV)
- Margin of Error: ME = CV * SE
- Upper Bound: UB = PE + ME
- Lower Bound: LB = PE - ME

One-Population Mean Confidence Interval

Assumptions: Data is random; observations are independent

Point Estimate: $\bar{x}$

Standard Error: $\frac{s}{\sqrt{n}}$

One-Population Proportion Confidence Interval

Assumptions: Data is random; $np \geq 10$ & $n(1 - p) \geq 10$

Point Estimate: $\hat{p}$

Standard Error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Other Key Concepts

1. Types of Distributions

- If $n < 30$, s is given, or nothing is stated about the distribution, use:
  - t-distribution
- Otherwise, use:
  - z-distribution

2. Study Designs

- Types:
  - Experiment
  - Observational Study
- Randomness for Experiments:
  - Random Selection
  - Random Assignment (typically incorporates a control group)

Glossary

- Alternative Hypothesis: Statement that proposes a statistically significant relationship or difference exists in the data
- Categorical Variable: Variable that takes on three or more category designations
- Categorical Binary Variable: Variable that takes on two category designations
- Confidence Interval: Statistical inference that estimates the value of a parameter with a range of plausible values
- Control Group: Participants who do not receive the experimental treatment
- Experimental Study: A study in which the researcher actively manipulates participants and splits them into a control group and treatment group(s)

- Explanatory Variable: Variable that explains variation in the response variable
- Hypothesis Test: Statistical inference that assesses the plausibility of a particular claim about the parameter
- Lower Bound: Lower limit of plausible values a parameter within a confidence interval can take on
- Mean: A measure of center for quantitative variables representing the average of a data set that takes on parameter μ and statistic $\bar{x}$
- Null Hypothesis: Statement that proposes no statistically significant relationship or difference exists in the data
- Observational Study: A study in which the researcher collects data through observations without any manipulation of participants
- Observational Unit: A single person, place, or thing within data
- Parameter: Numbers that summarize data for an entire population
- Point Estimate: A sample statistic collected for statistical inference
- Population: The entire group of study from which a sample is drawn
- Proportion: The ratio of the frequency of a specific level of a categorical (binary) variable to the total count of the categorical (binary) variable that takes on parameter p and statistic $\widehat{p}$
- p-value: Probability of observing sample data given the null hypothesis is true
- Quantitative Variable: Variable that takes on a continuous range of numerical values
- Randomness: Lack of predictability and patterns in events
- Random Assignment: Process by which the treatment is given to the observational unit by chance to reduce confounding variables
- Random Selection: Process by which participants are chosen for a study by chance to reduce bias
- Response Variable: Variable of interest that the outcome of a study measures
- Sample: Part of the population from which data is gathered
- Sample Size: The total number of observational units in a sample
- Standard Deviation: A measure of variability for quantitative variables representing the average distance between each value of a quantitative variable and its mean that takes on parameter σ and statistic s
- Standard Error: A measure of variability for statistical inferences representing the average distance between each sample statistic from its population parameter

- Statistic: Numbers that summarize data from a sample
- Statistical Inference: Statistical technique that draws a conclusion about a population parameter based on a sample statistic
- Statistical Question: A question that can be answered by collecting data that varies
- Test Statistic: A measure within hypothesis testing representing the number of standard errors a sample statistic is away from the population parameter
- Upper Bound: Upper limit of plausible values a parameter within a confidence interval can take on
- Variable: A characteristic that can be measured and assume different values