# CAAP Statistics
# Final Review

Aug 9, 2022

# Topics Overview

- Introduction to data: Observational vs Experimental
- Summarizing data: Numerical vs Categorical
- Probability: Law of Large Numbers
- Distribution: Normal, Bernoulli, Binomial, Poisson
- Foundations for Inference: Central Limit Theorem
- Inference for Numerical Data: When to use t-distribution?
- Inference for Categorical Data
- Introduction to Linear Regression
  - Residuals = observed - fitted
  - Types of Outliers
  - Inference for the slope - $H_0 : \beta_1 = 0$

# Introduction to Data

- What is Sampling? Sample vs. Population
- Observational Data
  - Random Sampling(Simple random sampling, cluster sampling, stratified sampling…)
- Experimental Data
  - Random Assignment
  - Treatment vs. Control
- Causation vs Correlation

## Summarizing Data

- Graphical Summary
  - Scatterplot
  - Histogram
  - Boxplot
  - Barplot
- Numerical Summary
  - Mean, variance
  - Q1, Q3, IQR
  - Median

# Probability

- Definition of Probability
  - Frequentist vs Bayesian
  - Law of Large Numbers
- Probability Distribution
  - Independent vs Disjoint vs Complement
  - Product Rule: $P(A \cap B) = P(A) \times P(B)$
  - Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Sampling with, without replacement
- Random Variables
  - Expectation, Variance

# Distributions

- Normal Distribution
  - Continuous
  - Unimodal, symmetric and bell-shaped
- Bernoulli Distribution
  - Discrete
  - Binary outcome: 1(success), 0(failure)
- Binomial Distribution
  - Discrete
  - Sum of independent Bernoulli trials
- Poisson distribution
  - Discrete
  - Number of rare events in a unit amount of time

# Foundations for Statistical Inference

- Central Limit Theorem
  - Sample mean/proportion follows normal distribution as the sample size increases
  - Law of Large Numbers: sample mean/proportion approaches population mean/proportion as the sample size increases
- Null hypothesis vs. Alternative hypothesis
- P-value: probability of having more extreme value than the observed value under null distribution
- Type I error vs. Type II error

# Review of Hypothesis Testing

1. Set the Hypotheses

   $$H_0 : \mu = \mu_0$$

   $H_A : \mu \neq \mu_0$ or $H_A : \mu < \mu_0$ or $H_A : \mu > \mu_0$

2. Check assumptions/conditions
   - Independence
   - Normality
     - The distribution is normally distributed
     - Sample size is large enough that we can apply CLT

3. Calculate a test-statistic and a p-value

   $z = \dfrac{\bar{x} - \mu_0}{SE}$ where $SE = \sigma / \sqrt{n}$

4. Make a decision
   - If p-value < $\alpha$, reject $H_0$
   - If p-value > $\alpha$, do not reject $H_0$

# Inference for Numerical Data

- Normality condition
  - **If the population distribution is normal**, CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size.
  - **If the sample size is large enough**, by CLT, the sampling distributions will be nearly normal.
  - What if the sample size is **small? t-distribution**
- One sample mean with t-test
- Paired t-test: When to use?
- Differences in two means

# Inference for Categorical Data

- Proportion is a special case of mean— the only difference is the formula for the standard error
- Test for one proportion
- Differences in two proportions
    - Think of Malaria Vaccine example

# Introduction to Linear Regression

- Line Fitting, Residuals and Correlation
  - $\hat{y} = b_0 + b_1 x$
  - Residuals = Observed(y) - Fitted($\hat{y}$)
  - Assumptions: Linearity, Normality, Constant variance
- Fitting a line by Least Squares Regression
  - Idea: want to minimize the sum of squared residuals
- Types of outliers in Linear Regression
  - Outliers, High leverage point
  - Influential point
- Inference for Linear Regression
  - Hypothesis testing for the slope: $H_0 : \beta_1 = 0$

# Thank you!