Review: Statistical Inference

Overview

ClassData

Study units, types of variables

- The **observational units or study units** in a statistical study are the objects described by a set of data (people, animals, things).
- The characteristics recorded about each individual (or case or experimental unit or observational unit) are called **Variables**.
- *Quantitative (numerical) variable* are measurements that are recorded on a naturally occurring numerical scale.
  - Variables that can assume a countable number (finite or infinite) of values are called **discrete**.
  - Variables that can assume values corresponding to any of the points contained in one or more intervals (i.e., values that are infinite and uncountable) are called **continuous**.
- *Qualitative (categorical) variable* are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories. A categorical variable with only two categories is called **binary**.

Tidy data: observational units on rows and variables are in the columns.

Variables:Predictor vs Response

- The variable whose effect you want to study is the **explanatory (predictor) variable**.
- The variable that measures the outcome of interest, that you suspect might be affected by the other, is the **response variable**.

Types of Data

- Variables that are measured at regular intervals over time are called a time series. Typical measuring points are months, quarters, or years. Time series connotes a single study unit observed at regular intervals over a very long period of time. A prototypical example would be the annual GDP growth of a country over decades or even more than a hundred years.
- Longitudinal data is acquired from some measurements over a larger number of study units. A prototypical example might be a drug trial, where there are hundreds of patients measured at baseline (before treatment), and monthly for the next 3 months.
- When several variables are all measured at the same time point, the data is called cross-sectional data. For example, data on sales revenue, number of customers, and expenses for last month at each Starbucks (more than 20,000 locations as of 2012) at one point in time would be cross-sectional data.

Data Sources

- Observation study: units are observed in natural setting and variables of interest are recorded. In an observational study, researchers observe individuals and measure variables of interest but do not attempt to influence responses.
  1. The explanatory variable is not imposed by the researchers.
  2. The goal is to describe the situation and perhaps discover association between variables.
- Designed experiment: An experiment is a study in which the experimenter actively imposes the explanatory variable group on the subjects (observational units).
  1. The explanatory variable group is called a treatment.
  2. In an experiment the researcher can legitimately draw a *cause-and-effect* conclusion between the explanatory and response variables.

More on Variables: Confounding Variables

A **confounding variable** is one whose potential effects on a response variable cannot be distinguished from those of the explanatory variable

- A confounding variable is related to both the explanatory and response variable.
- Because of a potential for confounding variables, one cannot legitimately draw **cause-and-effect** conclusions from observational studies.

CLASS ACTIVITY

The 2008-09 Oklahoma City Thunder, an NBA team in its second year after moving from Seatle, found that their win-loss record was actually worse for home games with a sell-out crowd (3 wins and 15 losses) than for home games without have a sell-out crowd (12 wins 11 losses). *Source of data: April 10th 2009 issue of Sports Illustrated.*

- What are the observational units?
- What are the response and explanatory variables?
- Is this an experiment or observational study?
- Is it reasonable to conclude that sell-out crowd caused the team to play worse?
- Suggest a confounding variable that plausibly explains the observed relationship.

Elements of Statistical Inference

Much of statistics involves making inferences from a sample to a population.

- **Population**: the entire group of observational units (people or objects) about which information is desired
  - A **parameter** is a number that describes a population.
- **Sample**: a (typically small) part of the population from which data are gathered
  - A statistic is a number that describes a sample.
  - Sample size: number of observational units in the sample.

Motivation

As a new graduate student at UVA

1. You are interested in estimating the proportion of all UVA graduate students who owns a Mac computer. This will help inform you on what type of computer to buy for your graduate study.
2. The mean number of hours that all UVA graduate students sleep at night. This might help inform you on the number of hours to sleep as a graduate student.

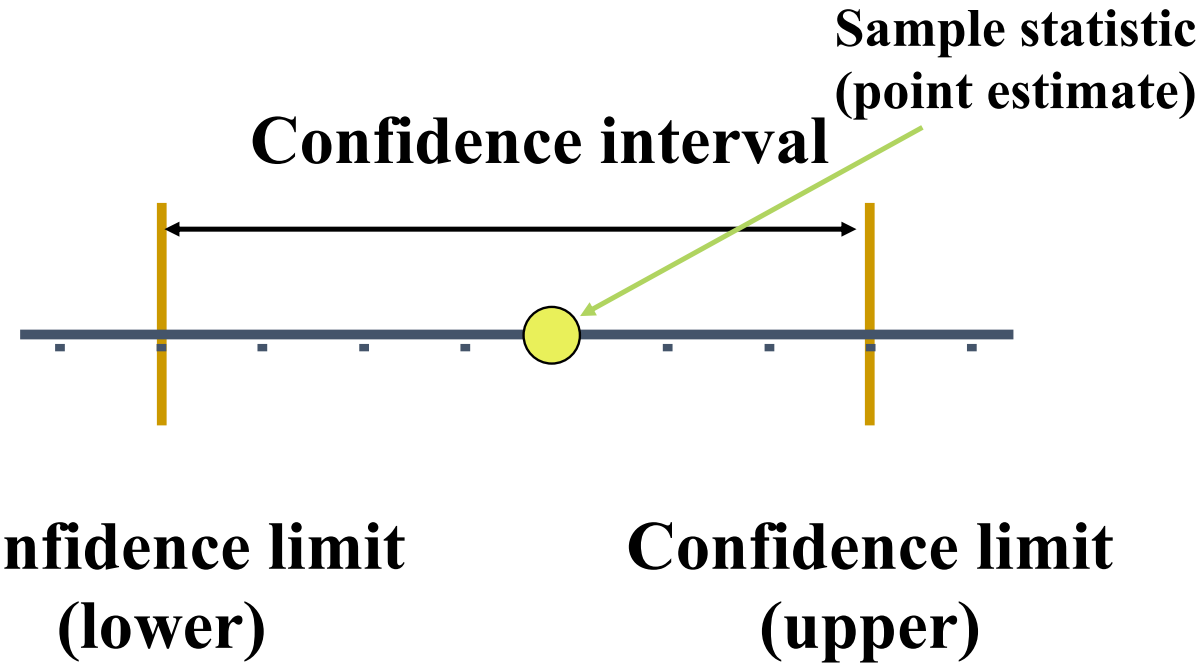
Identifying and Estimating the Target Parameter

The unknown population parameter (e.g., mean or proportion) that we are interested in estimating is called the target parameter

Parameter	Statistic	Key words	Data Type
μ	$\bar{x}$	Mean, Average	Numerical
p	$\hat{p}$	Proportion, Percentage	Categorical

Estimation Methods

- A **point estimate** is a single number that is our best guess of the parameter
- An **interval estimate** is an interval of numbers within which the parameter value is believed to fall



Confidence Interval for Proportions

Confidence Interval for a Population Proportion, p

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

where

- $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, known as the **standard error**, is the estimated standard deviation of the sampling distribution of the population proportion
- z is the z-score

Assumptions

- Independence Assumption.
- Sample Size Assumption: the sample size, n, must be large enough, that is, $n\hat{p} \geq 10$ and $n\hat{q} \geq 10$.
- Randomization Condition: the samples are randomly selected from a large population.

Example

Find the 95% confidence interval to estimate the proportion of UVA graduate students who owns a Mac computer.

Confidence Interval for Means

Small Sample Confidence Interval for a Population Mean, μ

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

where

- t is the t-score based on $(n-1)$ degrees of freedom

Assumptions

- Independence Assumption: the sampled values are independent of each other
- Data obtained by randomization
- The population has a distribution that is approximately normal

CLASS ACTIVITY: Confidence Intervals

Find a 95% confidence interval to estimate the mean number of hours that all UVA graduate students sleep at night.

Hypothesis Testing

Motivation

- Your best friend (whom you know to be really bad at basketball) tells you he's been practicing shooting free throws for the past year. He's been getting some instructions from the best coach in the game and his free throw percentage is 85%. You take him to a basketball court and he makes 40 out of 100 free throws. Do you believe your friend's claim?
- A final project group member tells you that all UVa students sleep for 6 hours on average. You use students from this class as a random sample. Do you believe your classmate?

Identifying the Target Parameter

The unknown population parameter (e.g., mean or proportion) that we are interested in estimating is called the target parameter

Key words	Data Type
Mean, Average	Numerical
Proportion, Percentage	Categorical

What is a hypothesis

- A hypothesis is a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.
- The purpose of hypothesis testing is to determine whether there is enough statistical evidence in favor of a certain belief (or hypothesis) about a parameter.

Elements of Hypothesis Test

- The null hypothesis, denoted H_0, usually represents the "status quo" or some claim about the population parameter that the researcher wants to test.
- The alternative (research) hypothesis, denoted H_A, usually represents the values of a population parameter for which the researcher wants to gather evidence to support. The alternative hypothesis contains the values of the parameter that we consider plausible if we reject the null hypothesis.
- The p-value (or the observed significance level) is the probability of obtaining a test statistic more extreme than actual sample value, given that the null hypothesis is true.
- Write the conclusion (in context) of your hypothesis test.

Logic of Jury Trials [Logic of Hypothesis Testing]

- Hypotheses: H_0 : Defendant is innocent, H_A : Defendant is not innocent.
- Prosecutor gathers and presents evidence. [For us, this means collect data - assumptions must hold].
- $*$Judge the evidence - by jury or judge. [For us, this means compute the test statistics in R].
- Make a decision "beyond reasonable doubt". [For us, this means make a decision based on the p-value].
- Judge concludes on the sentence (or no sentence). [For us, this means write the conclusion of your hypothesis test].

Type I and Type II Errors

Conclusion	H_0 is True	H_0 is not true
Do not reject H_0	Good decision	Type II Error
Reject H_0	Type I Error	Good decision

- Significance level $= \alpha =$ the likelihood of making a Type I error.

CLASS ACTIVITY: Make a similar table for jury trials.
Note: "beyond reasonable doubt" means keep α small.

Testing Hypothesis about Proportions

Testing Hypotheses about Proportions

1. Check Assumptions: Random sample? Independent? $np_0 \geq 10$ and $nq_0 \geq 10$
2. State Hypotheses: $H_0 : p = p_0$; $H_A : p \neq p_0$ (or $H_A : p > p_0$ or $H_A : p < p_0$)
3. Compute Test statistics: $Z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$
4. Compute p-value. Make a decision based on a pre-specified α. [Note: if p-value$< \alpha$ reject H_0, else do not reject.]
5. Conclude

Example

Your best friend (whom you know to be really bad at basketball) tells you he's been practicing shooting free throws for the past year. He's been getting some instructions from the best coach in the game and his free throw percentage is 85%. You take him to a basketball court and he makes 40 out of 100 free throws. Is your friend still a bad basketball player? That is, does your friend shoot fewer than 85% of free throws? Test at $\alpha = 0.01$.

Testing Hypothesis about Means

Testing Hypotheses about Means

1. Check Assumptions: Random sample? Independent? $n \geq 30$ or normal population?
2. State Hypotheses: $H_0 : \mu = \mu_0$; $H_A : \mu \neq \mu_0$ (or $H_A : \mu > \mu_0$ or $H_A : \mu < \mu_0$)
3. Compute Test statistics: $Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
4. Compute p-value. Make a decision based on a pre-specified α. [Note: if p-value$< \alpha$ reject H_0, else do not reject.]
5. Conclude

Example

A final project group member tells you that all UVa students sleep for 6 hours on average. Use students from this class as a random sample. Do you believe your classmate - test your conjecture that all UVa students sleep more than 6 hours on average? Test at $\alpha = 0.05$.