Advanced Statistics for Social Sciences 1

Statistical Thinking and Data Analysis with R

Instructor: Rachit Dubey

Email: rdubey@ucla.edu

Office Hours: [Thursdays 4-4:30 pm]

Course Overview

What we will cover

- **Statistical thinking:** Core concepts, inference, hypothesis testing, regression
- **R programming:** Data manipulation, visualization, statistical analysis
- **Practical applications:** Real-world datasets and research scenarios

Course Components

- **Lectures:** Theory and concepts, with (hopefully) fun examples
- **Labs:** Hands-on R programming and data analysis
- **Readings:** Textbook chapters and supplementary materials

Required Materials:

- **Textbook:** *Statistical Thinking for the 21st Century*, Russell Poldrack
- R and RStudio (**https://cran.r-project.org/**)

Assessments & Grading

My goals

- To help you think critically about the scientific process + develop a solid math and programming foundation (if you need it)

- ***This class isn't meant to stress you****!*
    - **If you come from a computational background:** I want to teach you how to test your models/algorithms
    - **If you come from a social science background:** I want you to become comfortable with math and develop a solid computational background

- **Note:** <u>I am NOT an R-expert. The R-component of the class is meant to give you background but this is NOT a programming class</u>

Assessments

- Attendance (40%), please don't be late to the class and please let me know beforehand if you can't make the class

- Final Presentation (60%)
    - **Option A (Research focused):** Present your ongoing research project, emphasizing the statistical methods used and why you used them

    - **Option B (Course synthesis):** Select 2-3 statistical concepts from the course and present: (1) what they are, (2) when to use them, (3) a worked example with real or simulated data

About me

Rachit Dubey

Assistant Professor, Department of Communication (I just joined 2 months ago!)

Background

- **Computational Cognitive Scientist** by training; **Focus areas:** Climate Change, AI & Society, Public Policy
- Undergrad in Computer Science (NTU, Singapore)
- Masters in Education (UC Berkeley)
- PhD in Computer Science (Princeton)
- Postdoc at MIT Sloan School of Management

- **Office Hours:** Thursdays 4-4:30 PM (PST) and by appointment (**rdubey@ucla.edu**)
- **Response Time:** Within 1-2 days on weekdays

Introduction to Statistical Thinking

What is statistical thinking?

> *“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”* - H.G. Wells

Statistical thinking is a way of understanding a complex world by describing it in **relatively simple terms** that nonetheless capture essential aspects of its structure.

Key Characteristics:

- Simplifies complexity while preserving essential information
- Quantifies uncertainty in our knowledge
- Draws from mathematics, statistics, computer science, and psychology

Why do we need it?

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.” - H.G. Wells

One very simple reason: Human intuition (especially about statistics) can be quite wrong!

- Availability bias: Most Americans believe violent crime is increasing, but statistical analysis shows it has **steadily decreased since the 1990s**.)
- Anchoring bias: We rely too heavily on the first piece of information we encounter (e.g., asking 'Is the population of Turkey more or less than 15 million?' vs. 'more or less than 75 million?' produces drastically different estimates)

For computer scientists: How do we reliably know that our algorithm or model innovation is better than baselines

For cognitive scientists: How do we know human behavior is influenced by certain factors?

For social scientists: How do we know a policy or an intervention actually cause outcomes (vs. correlation)?

Why do we need it?

> *“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.”* - H.G. Wells

It also makes us smarter citizens..

- Does the covid vaccine reduce or increase risk of hospitalization?
- Does watching opera improve lifespan?
- Does eating ice cream increase the risk of drowning?

What can statistics do for us?

1. Describe

The world is complex and we need to describe it in simplified, understandable ways.

2. Decide

We need to make decisions based on data, usually in the face of uncertainty.

3. Predict

We wish to make predictions about new situations based on knowledge of previous situations.

Example: A/B Testing

- **Describe:** Website A has 5.2% click-through rate, Website B has 5.8%
- **Decide:** Is B actually better, or just random variation? (Need hypothesis testing)
- **Predict:** If we switch to B, what click-through rate can we expect next month?

The big ideas of statistics

1. Learning from data

Statistics provides us with a way to describe how new data can be best used to update our beliefs

- We start with **prior knowledge** (what we already know)
- New data helps us update and refine our understanding
- Strength of prior knowledge matters (e.g., restaurant with 3 reviews vs. 300 reviews)

The big ideas of statistics

2. Aggregation

Statistics is the “science of throwing away data”

- We condense large amounts of data into meaningful summaries
- Allows us to see patterns that would be hard to parse in raw data

Example: A/B Testing

- **Raw data:** 50,000 visitors to Website A, each person clicked an ad (1) or didn’t (0). That’s 50,000 individual data points..!
- **Aggregated:** “5.2% click-through rate”
- We “threw away” individual variation and kept just one number that tells the story

Important Caveat

Aggregation can go too far. A summary can sometimes provide a misleading picture of the underlying data. Always consider what information is lost when we aggregate data (more on this later).

The big ideas of statistics

3. Uncertainty

The world is inherently uncertain, and statistics helps us quantify this uncertainty.

Example: Cigarette smoking and lung cancer

A 68-year old man who smoked 2 packs/day for 50 years has a **15% risk** of lung cancer (1 in 7).

This means:

- Risk is much higher compared to non-smokers
- But many lifelong smokers may never get lung cancer!
- The relationship is **probabilistic**, NOT **deterministic**

Important:
Statistics provides **evidence**, not **proof**. We can never “prove” a hypothesis in the logical sense.

The big ideas of statistics

4. Sampling

Aggregation implies that we can make useful insights by collapsing across data – but how much data do we need?

The idea of *sampling* says we can summarize entire populations based on samples, if obtained properly

Example: Election polling

- Polling 1000 registered voters to predict how millions of people will vote.
- How the sample is obtained is **critical** for generalization
- Polling only works if sample is representative, not just calling landlines in a single city!
- Larger samples are better, but with **diminishing returns.**

The big ideas of statistics

5. Causality and statistics

Well-designed studies, along with statistics, can help us demonstrate causation

Observational Studies

- Can show relationships and patterns
- Cannot conclusively demonstrate causation (recall the opera and ice cream examples!)
- Other factors may explain the relationship

Randomized Controlled Trials (RCTs)

- Well-designed RCTs are the gold standard for demonstrating causation
- Random assignment to treatment/control groups
- Especially meaningful when they control for confounding variables

Working with data

What is data?

Data are observations or measurements we record about the world

- Examples: survey responses, sensor readings, reaction times, brain scans, etc.

Person	Age	Favorite Food	Hours Sleep
Alice	24	Pizza	7.5
Bob	52	Tacos	6.3
..			
Eugene	40	Ice cream	7.0

- With this data, we can summarizing the “world” e.g., what is the mean age of members in this town?
- We can find relationships using this data e.g., does hours of sleep relate to age?
- We can predictions using this data e.g., given age = 25, can we predict hours of sleep?

Types of data

Qualitative Data

Describes qualities rather than quantities (e.g., favorite food, colors, categories)

Quantitative Data

Numerical measurements that can be analyzed mathematically

Three types of numbers:

- **Binary:** Yes/No, True/False, 0/1 (e.g., "Have you experienced migraines?")
- **Integers:** Whole numbers (e.g., Likert scales: 1-7, counts of events)
- **Real numbers:** Measurements with decimal precision (e.g., weight, temperature, reaction time)

Discrete vs. continuous measurements

Discrete

- Takes one of a finite set of values
- No “in-between” values make sense
- Examples:
    - Number of friends (can't have 33.7 friends)
    - Dog breeds
    - Number of children

Continuous

- Can fall anywhere in a range
- Defined in terms of real numbers
- Examples:
    - Weight (theoretically infinite precision)
    - Temperature
    - Reaction time

Descriptive Statistics

Descriptive statistics

The first step with any dataset: **summarize it in compact, understandable ways**

Measures of Central Tendency

Where is the "center" or "average" of our data?

- **Mean:** The arithmetic average
- **Median:** The middle value
- **Mode:** The most frequent value

The Mean

The mean is just average: add all values and divide by how many you have.

Example: Winning margins from 5 games: 56, 31, 56, 8, 32

Mean = (56+31+56+8+32)/5 = 36.6 points

Properties of the Mean:

- Uses **all** information in the data
- The "center of gravity" of the distribution
- Very sensitive to extreme values (outliers)

The Mean

When we have lots of data, writing everything out becomes impractical. We use notation as shorthand.

Building blocks:

- **N** = number of observations (in our example, *N* = 5)
- **X** = our data variable (winning margin)
- x_i = the ith observation (x_1 = 56, x_2 = 31, and so forth)
- $\sum$ = “add them all”
- μ = the mean of **X**

Formula:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Read as: "The mean equals the sum of all X values, divided by N"

The Median

The middle value when data is sorted in order

Calculation:

- Sort data from smallest to largest
- If odd number of observations: take the middle one
- If even number of values: take the average of the two middle values

Properties:

- The “middle observation” — half are smaller, half are larger
- Robust to extreme value

Example: Winning margins from 5 games: 56, 31, 56, 8, 32

Sort: 8, 31, 32, 56, 56. **Median = 32 points**

Mean vs. median

Scenario: People sitting at a table

- Bob (income **$50,000**)
- Kate (income **$60,000**)
- Jane (income **$65,000**)

Mean: **$58,333** | Median: **$60,000**

Then Bill Gates sits down (income $100,000,000):

Mean: **$25,043,750** 😱

Median: **$62,500**

When to use which?

- Means are useful when your problem is concerned with **expected** values.
- Medians are useful when your problem is concerned with a "typical" sample.

Real-world consequences

2010 Australian Housing Debate

Commonwealth Bank’s claim: Housing price-to-income ratio = 5.6 (comparable to major cities.. SF, NYC= 7.0)
“Housing is affordable!”

Independent analysts: Housing price-to-income ratio = 9.0
“Housing is severely unaffordable!”

The Bank’s calculation:

- Ratio = median house price/average (mean) income.
- Mean income > median income (inflated by wealthy outliers)
- Diving by larger number makes ratio appear smaller (“more affordable”)

The correct calculation:

- Ratio = median house price/median income (apples to apples)

Note: The bank had financial incentives to show housing as affordable (they profit from mortgages). Always consider who benefits from a particular statistical choice!

The Mode

The values that occurs ***most frequently*** in the dataset.

When to use the mode:

- **Nominal or categorical data:** Mean and median are meaningless here (e.g., everyone's favorite food.. pizza, tacos, pasta etc)

Example: NBA finals

Which team has had the most appearances in the NBA Finals?

Answer: LA Lakers (32 times)

This is the only sensible measure for this categorical data

Takeaways

Statistics helps us:

- **Describe** complex phenomena simply
- **Decide** under uncertainty
- **Predict** future outcomes

Core principles:

- Learn from data systematically
- Aggregate information meaningfully
- Quantify uncertainty
- Sample populations appropriately
- Understand correlation ≠ causation

Descriptive statistics:

- **Mean:** Center of gravity, uses all data, sensitive to outliers
- **Median:** Middle value, robust to outliers
- **Mode:** Most frequent, useful for categorical data