Welcome to 142: Introduction to Probability and Statistics in Biology and Public Health

Guess the date

In this year, UC Berkeley established a statistics department (split from mathematics) and hired David Blackwell - the first African American to receive tenure at UC Berkeley, and the first African American elected to the National Academy of Science (10 years later)

Quote from Dr. Blackwell

Basically, I'm not interested in doing research and I never have been. . . .I'm interested in understanding, which is quite a different thing. And often to understand something you have to work it out yourself because no one else has done it.

- quoted in a 2007 New York Times article

Today's Goals

Welcome and orientation to the class

My goals for our time together

Course logistics

Talk about the framework we use in the class (PPDAC)

Who am I?

Our Teaching team

Lead GSI: Juliet Del Core

Section GSIs:

- Benny Sun
- Kaiwen Hou
- Kirsten Landsiedel
- Shouxun Xu
- Yichen Xu

Tech GSI: Sean Yu

What is this class?

What is this class?

Figure 2: What do you think of when you think about statistics?

What does ChatGPT say?

- "why should you learn probability and statistics"
- Probability and statistics are important because they provide tools for understanding uncertainty and making informed decisions based on data. They help us evaluate evidence, assess risk, and distinguish meaningful patterns from random noise—skills that are essential in science, medicine, economics, technology, and everyday life.

My goals for you

Foundational concepts in probability and biostatistics

How to answer questions with data:

- your ability to critically assess statistical information presented to you in scientific and non-scientific fora
- your sense of how to approach answering real world questions with data
- develop your statistical intuition around variability and chance
- develop your toolkit for visualization, summarizing and testing simple relationships
- your ability to concisely and accurately describe statistical methods and results

What is this class?

In this class we are going to think about

- DATA - How we gather, display and summarize information
- Probability - the role of chance
- Statistics - the science of drawing statistical conclusions from data using a knowledge of probability

Three parts

- Part I: learning to explore and summarize univariate and bivariate distributions.
- Part II: classical problems in probability and the some commonly used probability distributions and the central limit theorem
- Part III: statistical inference, the process of estimating statistics from samples to make inference about populations

This is not a math class

Statistics is often classified as a branch of math, but I'd argue that it is more important to focus on the connections that statistics has with science (how we can learn about the world through data)

Though it is true that statistics uses math (and sometimes fairly advanced math!), not much math is needed to learn introductory statistics

In this class we will try, as much as possible, to emphasize concepts and help you develop your statistical intuition

This is not a programming class

Statistics is often viewed as "just computer programming," but this is an incorrect and dangerous characterization: computer programming is simply a tool for conducting statistical analysis

The use of computer programming in statistics is—and should be—quite different than approaches to non-statistical programming

We are using r programming in this course because it is an extremely useful skill, facilitates computation, and is desired in the job market

This is a relevant class

I hope to convince everyone here that statistics is relevant to everyone

As is more and more apparent, public health statistics have relevance to important policy decisions

You also make many decisions during your day that are influenced by statistics

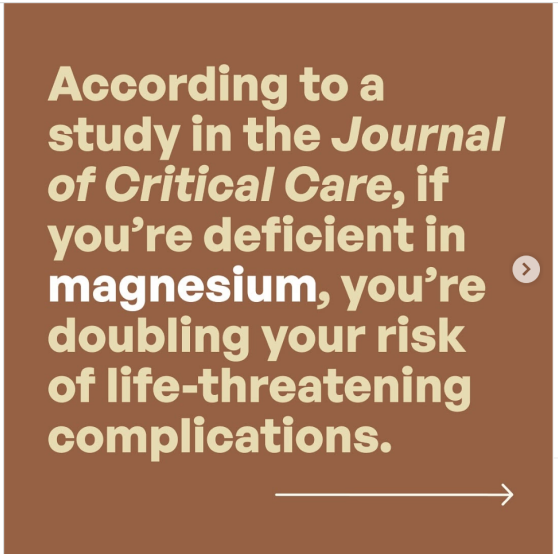
Statistics is not just relevant for public health, but also for other professions, including: education, journalism and law

As we'll try to illustrate via the recurring "statistics is everywhere" segments, statistics is useful for understanding the news and the world around us

Statistics is Everywhere

Social media and magnesium

Instagram

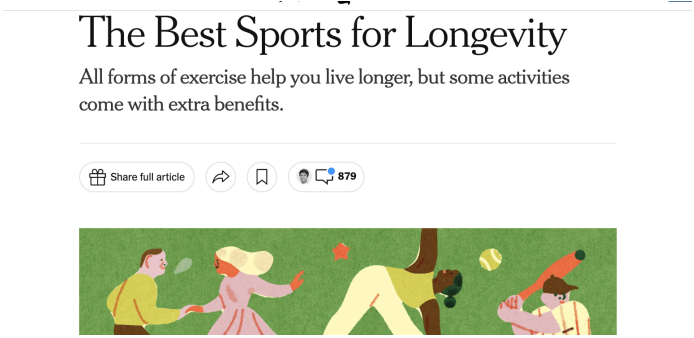


Magnesium

Jessica Fairley, Neil John Glassford, Ling Zhang, Rinaldo Bellomo, Magnesium status and magnesium therapy in critically ill patients: A systematic review, Journal of Critical Care, Volume 30, Issue 6, 2015, Pages 1349-1358

"In a systematic review of the literature, we found that all studies were of low quality, grossly heterogeneous, and contradictory. We identify that hypomagnesemia is associated with increased mortality. However, we found no robust evidence of any effect of the administration of magnesium in this population. Our observations imply that our knowledge of magnesium assessment and use in general ICU patients is poor and that higher-quality studies are sorely needed in these patients"

NYT “best sports for longevity”



- New York Times article Jan 15th 2026

NYT "best sports for longevity"

"One study from Denmark found that tennis players lived almost 10 years longer than their sedentary peers — and longer than soccer players, swimmers and the other recreational athletes included in the analysis. Other research from Britain and the United States followed people for about a decade and found that playing racket sports was linked to a lower risk of death during the follow-up period than any other sport or form of exercise studied."

Articles

Various Leisure-Time Physical Activities Associated With Widely Divergent Life Expectancies: The Copenhagen City Heart Study Schnohr, Peter et al. Mayo Clinic Proceedings, Volume 93, Issue 12, 1775 - 1785

Oja P, Kelly P, Pedisic Z, et al. British Journal of Sports Medicine 2017;51:812–817.

Watts EL, Matthews CE, Freeman JR, et al. Association of Leisure Time Physical Activity Types and Risks of All-Cause, Cardiovascular, and Cancer Mortality Among Older Adults. JAMA Netw Open. 2022;5(8):e2228510.

Consequences of poor communication

Logistics

Lecture/Section/Office Hours

Rationale for structure

Use Ed for substantive questions - gsi email for dsp/administrative issues

When in doubt - check the website and the ed announcements

There is lab and a quiz this week!

Also - please complete the survey

Data project teams - these will be assigned based on your lab section and survey responses

Ongoing evolution of the course

From Derivation to hands on programming

Co-Development of course with Dr. Riddell

How to get help with code

- Ask questions during labs/discussion sections, office hours, or on Ed discussion forum. Use the appropriate thread!
- Develop your online search skills. For example if you have a `ggplot2` question, begin your google search with "r ggplot" and then describe your issues, e.g., "r ggplot how do I make separate lines by a second variable".
- The most common links that will appear are:
  - https://stackoverflow.com: Crowd-sourced answers that have been up-voted. The top answer is often the best one.
  - https://ggplot2.tidyverse.org/: The official ggplot2 webpage is very helpful.
  - https://community.rstudio.com/: The RStudio community page.
  - https://rpubs.com/: Web pages made by R users that often contain helpful tutorials.

A word about AI

AI has become much more common in the last few years.

We recommend a number of resources for help with R code, chatGPT or other AI sources may be helpful as well - with some important caveats.

Frequently asked questions so far

Do I have to attend lecture/section?

Do I need the textbook?

Do I need to know programming?

Will I get off the waitlist?

Concern about grading.

Frequently asked questions so far

Figure 4: Will I get an A?

PPDAC - the approach we will use to answering questions with statistics

Problem

A clear statement of what we are trying to achieve.

Three main problem types

- Descriptive: learning about some particular attribute of a population
- Causative/Etiologic: do changes in an explanatory variable cause changes in a response variable?
- Predictive: how can we best predict the value of the response variable for an individual?

Problem type?

- Insurance company: What is the probability (how likely is it) that a 25 year old unmarried male driver has a car accident?
- Health department: How many cases of influenza have we seen this season compared to last season?
- Health care system: If we treat patients with diabetes using medication X, will thier insulin regulation be better or worse than medication y?

Plan

The procedures we use to carry out the study.

- Census or sample from the target population?
  - How was the sampling conducted?
  - Was the sample random?
- Is the study prospective or retrospective?
- Is the study observational or experimental?

Data

The data which is collected according to the Plan.

- How many observations to we have?
- How relaible are the measures?

Analysis

The data is summarized and analysed to answer the questions posed by the Problem.

We use our knowledge about probabilities to assess the role of chance in our findings.

Conclusion

Conclusions are drawn about what has been learned about answering the Problem.

PPDAC Example 1: A smoking behaviour study

PPDAC Example

Problem: Suppose we wish to study the smoking behavior of California residents aged 14-20 years.

In particular, we are interested in the *prevalence* of current smoking by gender.

What type of problem is this?

PPDAC Example

Plan: We need to first choose a time period, because we know that smoking behavior has changed immensely over time. It is unfeasible to gather these data for all residents in California who are 14-20 years old.

Instead we conduct a *random sample* of size n persons. We collect their: age, gender, and smoking status.

Note that we need to decide how large n should be, and how to obtain the random sample. The latter question is, in particular, very important if we want to ensure that our sample is representative of the population of interest. Time and money also constrain how the sample will be collected.

PPDAC Example

Data: Suppose that a random sample of 200 persons aged 14-20 was selected, yielding these data:

Gender	Number of smokers	Number of non-smokers	Total
Teen girls and women	32	66	98
Teen boys and men	27	75	102
Total	59	141	200

PPDAC Example

Analysis: The proportion of women in the sample who smoke is $32/98 = 33\%$. The proportion of men in the sample who smoke is $27/102 = 26\%$.

We would also like some idea as to how close this estimate is likely to be from the actual proportion in the population.

If we selected a second random sample of the same size, we would likely estimate different proportions for men and women. We will learn how to estimate the precision of these estimates.

PPDAC Example

Conclusion: 33% of girls and women aged 14-20 and 26% of boys and men of the same age group are current smokers in California in 2018 (plus a measure of uncertainty).

Example 2: Life expectancy for non-Hispanic black and white men and women in California between 1969-2013

Introduction

Life expectancy is one of the core measures used in public health to comment on the well-being of groups of people. Differences in life expectancy by race/ethnicity, for individuals living in the same region can reflect underlying inequalities in policies, access to care, food environments, structural and systemic racism, among other potential causes.

Research objective (Problem)

The purpose of this short report is to visualize life expectancy among black and white men and women in California between 1969 and 2013.

We are interested in whether there are differences by group and whether these differences have changed over time.

What type of problem is this?

Plan

Death certificates in the United States include race/ethnicity, age at death, and date of death and capture all deaths of US residents. These data are aggregated by the CDC's National Cancer Institute into the SEER*Stat software. Previously, Riddell et al.[1], analyzed these data to compute estimated trends in life expectancy for non-Hispanic black and white men and women, for 40 US states between 1969 and 2013. States without enough data were excluded from these analyses.

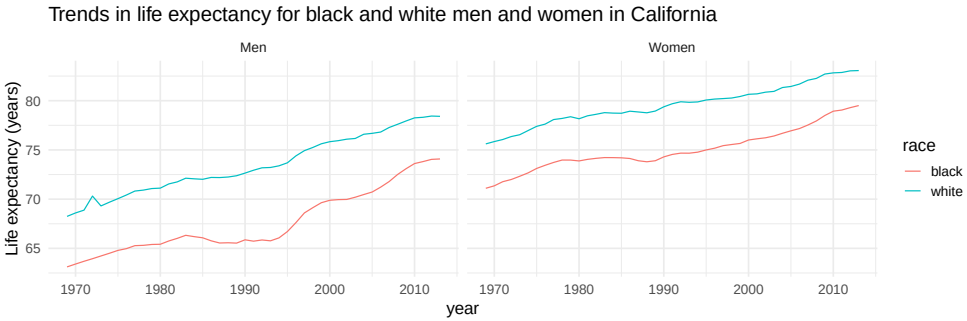
To carry out this short report, we will use data from Riddell et al. to visualize trends in life expectancy as part of an exploratory data analysis. In particular, we will plot time trends for black and white men and women in California.

Data

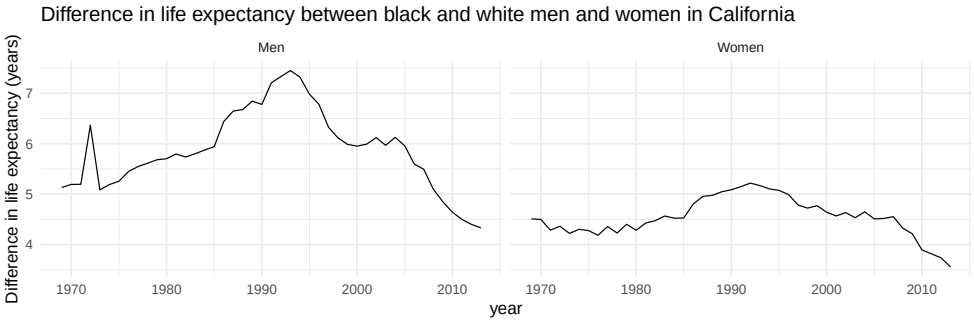
Here are the first few rows of these data for California:

state	stabbrs	year	sex	Census_Region	Census_Division	LE	race
California	CA	1969	Female	West	Pacific	75.61137	white
California	CA	1969	Male	West	Pacific	68.24766	white
California	CA	1970	Female	West	Pacific	75.84916	white
California	CA	1970	Male	West	Pacific	68.59865	white
California	CA	1971	Female	West	Pacific	76.05663	white

Analysis


Trends in life expectancy for black and white men and women in California
Men
Women
Life expectancy (years)
80
75
70
65
1970
1980
1990
2000
2010
year
race
black
white

Analysis


Difference in life expectancy between black and white men and women in California
Men
Women
Difference in life expectancy (years)
7
6
5
4
1970
1980
1990
2000
2010
year

Conclusion

The difference in life expectancy in 1969 between non-Hispanic blacks and whites was 5.1 years for men and 4.5 for women in California.

By 2013, the difference was 4.3 years for men and 3.6 for women in California.

References

The PPDAC method is described based on course notes from STAT 231 from the University of Waterloo (Ontario, Canada). Spring 2006 Course Packet.

1. Riddell CA, Morrison KT, Harper S, Kaufman JS. Trends in the contribution of major causes of death to the black-white life expectancy gap by US state. Health & Place. 2018. 52:85-100. doi: 10.1016/j.healthplace.2018.04.003.

a pre-emptive appology

(credit to xkcd.com for the comic)