PH142: Introduction to Probability and Statistics in Biology and Public Health

Course Syllabus

Table of Contents

Course Information

Instructor	Mi-Suk Kang Dufour (she/her/hers)
Email:	mi-suk@berkeley.edu
Instructor Availability:	https://mi-suk.youcanbook.me/
GSIs:	Lead GSI: Joan Shim (she/her/hers) GSI & Section(s): Hemakshat Sharma (he/him/his) 101 & 102 Kaiyuan (Jacky) Li (he/them/them) 104 & 109 Julia Piccirillo-Stosser (she/her) 106 & 107 Junzhe Shao (he/him/his) 103 & 112 Paula Marquez (she/her) 105 Qiuran Rita Lyu (she/her) 108 & 111 Sylvia Song (she/her) 110
Email:	142gsi@berkeley.edu
Course Canvas/bCourses link:	https://bcourses.berkeley.edu/courses/1532521
Course Units:	4
Contributing Instructors:	Corinne Riddell, Maureen Lahiff

Course Description

This course is an introduction to statistics and data science, primarily for MPH and undergraduate public health majors, and others interested in public health topics. The course can be divided into three parts. In Part I, we will focus on learning to use R to explore and summarize univariate and bivariate distributions. Specifically, we will use the dplyr and ggplot2 packages to manipulate and visualize data sets in R. Part II of the course introduces classical problems in probability and the Normal, binomial, and Poisson distributions. The most important topic we will cover in Part II is the Central Limit Theorem. In Part III, we introduce statistical inference, the process of estimating statistics from samples to make inference about populations. During all parts of the course we will use real and simulated data sets to gain experience conducting biostatistical analyses using R. We will follow the PPDAC model, which stands for "Problem, Plan, Data, Analysis, and Conclusion".

Prerequisites

High school algebra

Course Learning Objectives

After successfully completing Part I of the course, you will be able to:
- Extract relevant statistical information from published articles in the scientific and popular press
- Describe distributions of variables visually and calculate summary statistics for measures of centrality and spread
- Determine the appropriate graphic to plot distributions and provide R code to manipulate and visualize data frames
- Identify basic sampling strategies and study designs used in Public Health
- Describe core concepts of ethics in Public Health
- Perform basic data manipulation in R
- Interpret output from a simple linear regression model

After successfully completing Part II of the course, you will be able to:
- Compute probabilities using the general rules
- Identify and describe binomial and Poisson random variables
- Compute probabilities using basic properties of the Normal distribution
- Express epidemiologic measures as probabilities
- Describe the central limit theorem
- Write R code to compute probabilities for the Normal, binomial, and Poisson distributions

After successfully completing Part III of the course, you will be able to:
- Estimate means, proportions, and differences between means and proportions, compute their confidence intervals and perform statistical tests
- State the assumptions and importance of the assumptions for statistical tests
- Perform a simple chi-squared test
- Perform a matched t-test
- Describe and check the assumptions for simple linear regression. Interpret the confidence interval and statistical test of regression intercept and slope coefficients
- Describe ANOVA, including the null and alternative hypotheses, and interpret output
- Describe when bootstrapping can be used
- Describe a permutation test
- Demonstrate knowledge that has been used throughout the term, in terms of data visualization and data manipulation
- Write R code snippets to perform hypothesis tests and calculate p-values

Instructor Information

Dr. Kang Dufour is an epidemiologist and biostatistician at University of California Berkeley Division of Biostatistics. Her work focuses on implementation research and evaluation of public health programming for infectious diseases including HIV, CMV, STDs, and Malaria.

Course Format

PH142 is a 4 unit course:

Synchronous video lectures will be held M, W, F from 8:10 to 9:00 am. Lectures will be recorded and made available as soon as possible following each session. Although lecture attendance is not required it is strongly recommended. Please keep in mind that there may be some delay between live lecture and the posting of the lecture videos. Lectures will primarily cover material in the lecture slides and handouts. All lecture materials are made available. Code used during lecture will be shared in markdown format through the datahub, slides will be shared in PDF format. Lectures will introduce concepts, highlight examples of these concepts in the popular and scientific press, and walk through some examples and calculations.

Lab/discussion sections will be held each week. These sections will be held in person for 2 hours each. Lab sessions are not recorded.

Course Schedule (subject to change - see course website for most up to date schedule)

Week	Topics	Assignments	Book chapter(s)	Due Dates
1	Introduction to Datahub and PPDAC, Working with Data in R and R studio	Lab 1, Homework 1	1	Quiz 1: January 20 Lab 1: January 23
2	Visualizing Data in R and	Lab 2,	2,3	Quiz 2: January 27

	Rstudio (ggplot2), Describing data with numbers, Exploring relationships between two variables	Homework 2		Lab 2: January 30
3	Introduction to Regression, Two-way tables, Samples and observational studies	Lab 3, homework 3	4,5,6,7	Quiz 3: February 3 Lab 3: February 6
4	Designing experiments, Part 1 summary, Introduction to probability	No lab – midterm review during lab session	8,9	None
5	Probability continued	Homework 4, lab 4, Midterm I	10	**Midterm 1: February 12** Lab 4: February 20
6	The normal distribution, Binomial distribution	Lab 5, Homework 5, data project part I	10,11,12	Data Project Part I: February 27 Lab 5: February 27 Quiz 4: February 24
7	Poisson distribution, Sampling and the central limit theorem	Lab 6, Homework 6	12,13	Quiz 5: March 2 Lab 6: March 5
8	Confidence intervals, hypothesis testing, power and sample size	Lab 7, homework 7	13,14,15	Quiz 6: March 9 Lab 7: March 12
9	z testing Recap of part II	Midterm II	15,16	Midterm 2 March 15

10	Inference about a population mean, tests for two means, ANOVA	Data project part II Lab 8	17	Data Project Part II: March 22 Quiz 7: March 23
11	Spring Break – no instruction			
12	ANOVA recap–post–hoc testing, Non parametrics, Inference for regression	Lab 9, Homework 9	17,18,24	Quiz 8: April 6 Lab 8: April 2 Lab 9: April 9
13	Inference for proportions,Comparing two proportions, Bootstrapping confidence intervals	Lab 10, Homework 10	23,19	Quiz 9: April 13 Lab 10: April 16
14	goodness of fit, chi–squared, Permutations testing	Lab 11, homework 11	20,21,22	Quiz 10: April 20 Lab 11: April 23
15	Summary of course concepts, structured review		none	
16	Reading week			
17	Finals week	Final exam, Data project part III		Data Project Part III: May 6

Course Grading

Grading is based on the following:

- **Participation**. Throughout the term you will be asked to provide feedback, practice with/test course systems and participate in course discussions. Participation assignments will be announced on Ed. These will be marked for completion only. You may miss one participation activity without penalty.

- **Homework Assignments** will be distributed as R markdown files on datahub on Wednesdays (no assignments will be released the weeks of midterm exams). **Homework will not be submitted for marks** and you are encouraged to

work on it in groups if that is how you learn best. **Completing the homework is your best preparation for the midterms and final examination.** All solutions will be posted on datahub each Tuesday after the homework was made available.

- **Weekly quizzes** will be available from 9 am on Fridays until noon on the following Saturday. Quizzes will be relatively short and meant to encourage you to keep on top of weekly content. Once opened, you will have 1 hour to complete the quiz. Quizzes will cover material from Wednesday and Friday of the previous week and Monday of the current week. **Your lowest quiz score will be dropped.**

- **Lab exercises** are intended to practice the concepts from lecture in a practical programming environment. You can complete and submit these during the lab section, or on your own beforehand. Students find it much more helpful to complete this in lab rather than independently, but we understand students learn differently, so feel free to do what works best for you. Lab exercises are **graded on correct completion**, so you must complete the lab fully, passing all tests, in order to receive credit for the assignment. Since we provide all tests for correctness of your code, **grading is all-or-nothing**. Labs will be released on Wednesdays and due the following Tuesdays. You may miss one lab assignment without penalty. Lab attendance is not required.

- **Midterm 1,2 and Final Exam.** There are two midterms offered on February 12 and March 15. The midterms will be administered in class that day and are approximately 45 mins long. **The final exam is on Monday, May 6th . If you have a conflict with any of the exam dates, please email the instructor immediately so that we can discuss possible accommodations. Accommodations cannot be made for individuals enrolled in another class at the same time as this one, so please take this class in another semester if doubly enrolled.** Appropriate accommodations for the midterm will be made for those with disabilities (please refer to the "Disabilities" section, below). Please note that only in extremely rare circumstances such as illness (with a doctor's note) will the in-class midterm be given to individual students after the scheduled examination date. Exams will cover the material presented in lecture, discussion, and lab sections, including R coding syntax, unless otherwise noted.

- **Exam policies.** Exams will be administered in person and on paper. You may bring one page (front and back) of notes with you to use on the day of the exam. Notes may be hand printed or computer printed with a minimum of 10pt font. You should also bring with you a simple scientific calculator. While you take the exam, you are prohibited from discussing the test with anyone other than the PH142 instructional team. Evidence of cheating may result in a 0 on

the test or further disciplinary action. We will strive to return graded examinations within one week of the exam date.

- **Data skills demonstration group project.** The purpose of the group project will be to use public health or biological data that you find or have access to and use it to demonstrate the statistical concepts that you've learned throughout the course. You have two options for the data project: 1)you may self-select into a group of up to 5 students or 2) we can assign you to a group if that is preferable for you. You will be asked about your preferences at the beginning of the term. You will have 2 deliverables spread out across the term, covering each section of the course and 1 deliverable due at the end which will cover part 3 and will also synthesize parts 1 and 2 into a cohesive report. At least one member of each group will be responsible for checking in with their assigned GSI before submitting each part of the project, either during section or office hours, as part of your grade.

- **Course extra credit**. Throughout the course we provide examples of statistical information in the scientific and popular press. You may choose to provide an analysis and discussion of a recently published scientific article which is also covered in the popular press, or a critique and explanation of a study that you found personally compelling. You may earn up to 2 points on your course grade through extra credit.   Details will be provided in the first weeks of the term.

Final grades will be assigned according to the following percentages:

Weekly Quizzes	15%
Participation	5%
Lab completion	10%
Midterm 1	15%
Midterm 2	15%
Group project (in 3 parts)	20%
Final Exam	20%

S/U (satisfactory/unsatisfactory) grading is permitted for graduate students in this course. There are no differences in the course requirements or the grading for students who choose this option. "S" will appear on transcripts for grades of "B-" or above. P/NP (pass, no pass) grading is permitted for undergraduate students. A P will be assigned to those with a grade of "C-" or higher.

Course Materials

Courses website

To access the course website, go to https://ph142-ucb.github.io/sp24/.

Here you will find links to required and optional readings, the syllabus, assignment descriptions and additional course resources. The course website will contain the most up to date schedule and assignment information.

Required Materials

We will be using **R**, a statistical programming language, and **RStudio**, an integrated development environment on **Datahub**, a cloud computing environment created at Berkeley. Use of R, RStudio, and Datahub is required for homework assignments and lab exercises and requires an internet connection and web browser. You will learn how to use R, RStudio, and Datahub during the first week of classes. You can access datahub from the links on the course website.

Optional Materials

The course textbook is "The practice of statistics in the life sciences" by Brigitte Baldi and David S. Moore. You can find the **textbook online here** (if not, it's in the queue for digitization at the university library and will be made available to you when it's ready). The 4th edition is the latest one, but previous editions are fine. You can also purchase or rent the book here. We rely on it more during Part II and III of the course than we do in Part I. It is possible to complete the course using course materials and we note that the instructors differ in some places from the opinions presented in the text. Where the textbook and course materials differ the course materials will take precedence.

Other resources

In addition, here are some free online resources available as supplementary material. We link to these specific resources in the lecture slides when applicable:

- Learning statistics with R: https://learningstatisticswithr.com/lsr-0.6.pdf
- OpenIntro Statistics: https://drive.google.com/file/d/0B-DHaDEbiOGkc1RycUtIcUtIelE/view
- A ModernDive into R and the Tidyverse: https://moderndive.com/9-hypothesis-testing.html#ht-infer
- Statistical Thinking for the 21st Century: https://statsthinking21.github.io/statsthinking21-core-site/ci-effect-size-power.html#statistical-power
- R for Data Science: https://r4ds.had.co.nz/data-visualisation.html

Announcements

Course announcements will be sent out through a once-weekly email blast from Ed.

Questions and comments

Questions during lecture and lab are strongly encouraged. If something is unclear to you, it is probably unclear to many others in the room. There may be times, however, when the instructor or the GSI decides that a particular question or discussion is not helpful to the entire class or will take too long to address satisfactorily. In these cases, we may defer the question to be answered outside of class time. If important questions are answered outside of class, the answers will be shared with the entire class on Ed.

We will use Ed for class discussion. The system is highly catered to getting you help fast and efficiently from classmates, the GSIs, and the instructor. In general, you can expect that the GSIs will respond to posted questions within 24 hours. Please do not email course content questions directly to the instructor or GSIs. The instructor and GSIs will not respond to questions about course content by email.

GSIs will not respond to Ed questions during holidays and breaks and you should expect longer turnaround times on the weekend (48 hours compared to 24 hours). GSIs will respond to Ed questions up until 24 hours before exams. However, students may continue to post and answer each other's questions during breaks and in the last 24 hours before an exam.

For questions and concerns that are not related to course content, please email ph142@berkeley.edu. GSIs will respond to the course email account within 2 business days. Email responses may be slower over weekends and breaks.

Office Hours

Each GSI will hold office hours at least once per week. GSI office hours are an opportunity for you to access more individualized support. We will also ask that your group attend office hours to discuss your data project at least once for each part of the project. Individual office hours with the professor are offered by appointment. The professor will also hold drop-in group office hours every second week that will be open for questions and discussion of course material. Scheduling of office hours will be determined according to student preferences following the beginning of term needs assessment.

Tutoring

The Dream Office offers peer-to-peer group tutoring sessions for Residential PH142 Graduate students. They do not offer 1:1 tutoring nor partake in grading and decision-making regarding course assignments. For more information about tutoring hours and offerings, see their most up-to-date semester schedule or contact biotutor_sphdiversity@berkeley.edu.

DREAM Office Graduate Student Tutors:

> Hillary Nguyen

> Khanh Thai

Course Email

We strive to reduce email as much as possible. All questions about course material should be asked on Ed. This allows us to reduce email and also allows other students to benefit from the questions and answers. We will not answer any questions about course material via email. Email the GSI account to: make DSP accommodations for tests or homework or request an assignment extension (see "Late Assignments" below). Email both the instructor and GSI account for personal concerns or disruptions that affect your performance in the course or during an emergency that will result or has resulted in a missed test.

Policies

Due Dates

If you have a conflict with <u>any</u> of the midterm/exam dates, please email the instructor by February 1st so that we can discuss possible accommodations.

Grace Period

All assignments, unless stated otherwise, are **due on the specified day at 10:00 pm.** Due to the nature of electronic submission, we understand that some students may experience technical difficulties with submission close to the deadline. Therefore, we are offering a **grace period of two hours, until 11:59 pm,** to account for these submission issues. If you are having issues during this time, please email your submission to the GSI email at (ph142@berkeley.edu) before 10:30 pm. **We will not be accepting requests regarding submission errors after 11:59 pm on the due date.** The grace period applies by default, you do not need to notify us to use it.

Regrades

Regrades will be allowed on quizzes and midterm exams, and must be submitted **within three days** after the grades are released using Gradescope. Note that if you request reconsideration of a graded question, instructors may reconsider grades on the entire assignment. Due to the short turnaround time for final grade submissions we generally cannot accommodate re-grade requests for the final exam.

Late Submissions

Assignments submitted 24 hours after the due date will be penalized by 50%. Extensions can be made for DSP students but should be requested ideally before the due date by emailing the GSI email account. Anyone else requesting an exemption should email the GSI account explaining their situation. If an emergency event prevents submitting an assignment by the deadline, please contact the GSI email account as soon as reasonably possible, including documentation with your request for extension.

Technology

Zoom will be used to conduct lectures. Zoom links will be shared on the calendar embedded on the course website. **You *must* use a Zoom account affiliated with your Berkeley email to access the invite!** Lecture will be recorded each week and one lab section will be recorded each week and made available as soon as we receive the processed videos (generally 2-3 days). Students will be muted and their videos turned off by default since there are so many of you! I encourage you to ask questions using the chat or by using the raise your hand feature. Feel free to turn on your camera when you ask a question orally. Please note that questions asked are part of the lecture/lab recording.

Anti-racist and inclusive learning environment

As we, at Berkeley Public Health, strive to create an anti-racist learning environment, we commit to teaching this course, to the best of my/our ability, with an antiracist, racial justice, and equity-minded lens. We invite you to take this journey with us by being fully present. We are interested in your perspectives and in the value and knowledge you bring to help make this an enriching classroom environment.
We view this syllabus as a dynamic document oriented toward learning and not just coverage of material; thus, we may add or modify topics covered, assignments, and resources (e.g., required readings/videos) slightly based on the needs and interests of students in the course. We welcome feedback and input at any time and invite careful reflection of any modifications that may help improve the course in the future.

As your teaching team, we agree that:

- We will do our best to include course content that include examples relevant to BIPOC communities (e.g., readings; examples; data, etc.)
- Students are the experts of their own experiences. Your world lens is welcomed; and as students, you are invited to lift up information and/or data that is relevant to the course material. Everyone is a teacher and everyone is a student.
- We cannot speak on behalf of all groups, or fully understand the issues, concerns and history of all BIPOC. However, we are willing to listen and learn, admit mistakes and engage in ongoing cultural humility practices.
- We welcome feedback and input at any time during the course without fear of reprisal; in our mid-semester evaluation there will be specific language about antiracism teaching practices.

SPH Course Policies

Descriptions of and relevant campus links to SPH school wide course policies on Disability Support Services, Accommodation of Religions Creed, Course Evaluations, Academic Integrity can be found at:
https://berkeley.box.com/s/knh3rbk9ikgvmca4ymy93msgj9bkebq5

Disabilities: The Disabled Students Program (DSP)

The mission of the Disabled Students' Program (DSP) is to ensure that all students with disabilities have equal access to educational opportunities at UC Berkeley. The DSP offers a wide range of services, accommodations, and auxiliary services for students with disabilities. These services are individually designed and based on the specific needs of each student as identified by DSP's Specialists.

The mission of the Disabled Students' Program (DSP) is to ensure that all students with disabilities have equal access to educational opportunities at UC Berkeley. The DSP offers a wide range of services, accommodations, and auxiliary services for students with disabilities. These services are individually designed and based on the specific needs of each student as identified by DSP's Specialists.

We will accommodate disabled students' needs according to DSP documentation; please notify the DSP if you require such accommodation (DSP will then contact the instructor). **Note that this may take several weeks, so please initiate this process ASAP so that any accommodations can be implemented in time for the first midterm exam**. Steps to the application process: https://dsp.berkeley.edu/students/new-students.

If you require DSP accommodations for a test, please email the GSI email account at ph142@berkeley.edu with your request and write "DSP accommodation" in the subject heading as soon as you know accommodations are required. If your accommodation allows for extension on take-home assignments, we ask that you discuss your request no later than 24 hours after the assignment is posted.

Mental Health

If you are experiencing stress, anxiety, or other forms of distress during the semester, we hope to be a resource for you–**please don't hesitate to reach out to a GSI or the Professor for support.** You are not alone.

There are also many resources available to you. All registered Berkeley students are eligible to use Counseling Psychological Services. **You do not have to purchase the Student Health Insurance Plan to use these services.** The first five counseling sessions are free for registered Berkeley students. Counselors can provide support in

academic success, life management, career and life planning, and personal growth and development.

UC Berkeley, Counseling and Psychological Services

- Please call (510) 642-9494 or stop by the office on the 3rd floor of the Tang Center to make an appointment with a counselor.
- **Drop-in counseling for emergencies:** Monday - Friday, 10:00AM-5:00PM
- **After hours counseling:** In the case of emergencies at night or on weekends, call (855) 817-5667 for free assistance and referrals. Request to speak with a counselor.
- **For emergency support:** Call UCPD 911 or (510) 642-3333

24 Hour Crisis Hotlines

- **Alameda County Crisis Line:** Call 1-800-309-2131 *(offers confidentiality, TDD services for deaf and hearing impaired callers and translation in 140 languages)*
- **National Crisis Help Line:** Call 1-800-273-TALK
- **Crisis Text Line:** Text HOME to 741741
- **National HopeLine Network:** Call 1-800-SUICIDE

We also ask that you look out for your fellow peers. If you see any of the signs below that may indicate your classmate may need assistance, please use the resources above or reach out to any of the GSIs or Professors.

- Withdrawing from other people
- Changes in weight or eating patterns
- Changes in sleeping patterns
- Fatigue or lack of energy
- Increased anxiety or irritability
- Feeling worthless or hopeless

Other Campus Resources:

- **Let's Talk:** *Informal Drop-In Counseling*
- **Self-Help Resources**
- **Be Well at Cal**
- **Center for Support and Intervention:** *Violence Prevention, Harm Reduction and Case Management Services*

Academic Honesty

Learning is hard work–we encourage everyone to work together and support one another. However, while group work is encouraged, with the exception of the group project, **students must submit their own code and answers** for grading. Students can not work together on the quizzes, midterm, or final examinations. **Tests that show evidence of academic misconduct will be immediately flagged and reported to the Center for Student Conduct for review.** This can result in a grade of 0 on an assignment or a harder penalty depending on the degree of the offense.

Each term, a few students in this class are reviewed by the Center for Student Conduct as we take cheating very seriously.

Berkeley's code of conduct is **here**. See Section V and Appendix II for information about how UC Berkeley defines academic misconduct. In particular, the sections on cheating and plagiarism are most relevant for this class.

If you are not clear about the expectations for writing a test or examination, be sure to seek clarification from the instructors or your GSI beforehand.

Harassment Policy

We are all responsible for creating an environment that is welcoming, civil, safe, and tolerant. UC Berkeley does not tolerate harassment of PH142 students, GSIs, or instructors.
- Instructors and GSIs will act to stop acts of harassment in the classroom.
- Students experiencing harassment can contact the office for the prevention of harassment and discrimination. To file a report, you can email ask_ophd@berkeley.edu or call them at (510) 643-7984. For more information, see: https://ophd.berkeley.edu/.
- Please note that Instructors and GSIs are Responsible Employees and must report incidents of sexual violence and harassment to the Office for Prevention of Harassment and Discrimination. Please see this website for confidential reporting resources: http://survivorsupport.berkeley.edu/Confidential-Resources-Anonymous-Reporting-and-Privacy.