Course Syllabus

PH 241 Intermediate Biostatistics for Public Health (Spring 2024)

Course Team

Instructor	Prof. Alejandro Schuler (he/his) alejandro.schuler@berkeley.edu (mailto:alejandro.schuler@berkeley.edu)
TAs	Shirley Sui (Fully Online course only) shouchu@berkeley.edu (mailto:shouchu@berkeley.edu) Xin Zhou (Residential course only) xinzhou@berkeley.edu (mailto:xinzhou@berkeley.edu) Kevin Benac (Residential course only) benac@berkeley.edu (mailto:benac@berkeley.edu) Andy Kim (Residential course only) a_kim@berkeley.edu (mailto:a_kim@berkeley.edu)

Audience

This class is for anyone who wants to be able to think critically about data analysis. The only requirements are pre-calculus and proficiency in R programming with the tidyverse to the level of exploratory data analysis. If you're feeling shaky on these topics here are some helpful resources:

Subject	Resources
Pre-Calculus	- Khan Academy Pre-Calculus (https://www.khanacademy.org/math/precalculus) - Berkeley Math 1: Foundations of Lower Div Math (https://sites.google.com/berkeley.edu/pre-calculusessentials2022/pre-calculus-essentials-program?authuser=0) (see book and notes) focus on understanding functions, graphs, and vectors
R Programming	- teachR (https://teachingr.com/) : a quick series of youtube videos that focus on what we'll need for this course - R for Data Science (https://r4ds.hadley.nz/) : best reference for learning R and the tidyverse. Focus on ch 1-4, 25, 26

Apart from this, undergraduate calculus and linear algebra are certainly helpful but not required.

Content

This course will provide you with the ability to think critically about how to answer complex public health questions with data. The fundamental problem in doing this is that data are random: if you do the exact same study twice, you'll get different datasets and different answers. How can we define the truth we're after?

Part I Learning Goals:

1. translate between mathematical notation, diagrams, english prose, and code
2. demonstrate fundamental ideas in probability and statistics (e.g. conditioning, independence, expectation, consistency) using multivariate simulation
3. distinguish between estimands and estimates

In the first part of the course we'll develop unambiguous language that we can use to talk about the sort of randomness we observe in the real world. This language takes shape as the fundamentals of probability and statistics: random variables, estimands, estimators, sampling distributions, etc. Once we have those frameworks we can work within them to clearly define what it is we're after when we analyze data and transparently quantify our uncertainty in the result.

Week	Topic	References
1	Course Intro	- functional programming in R (https://github.com/alejandroschuler/r4ds-courses/blob/summer-2023/lectures/5-functional-programming.md)
2	Probability	- Stat20 3.1 (https://stat20staff.netlify.app/3-probability/01-chance-intro/notes) , 3.2 (https://stat20staff.netlify.app/3-probability/02-cond-prob-indep/notes)

Week	Topic	References
		- Pitman 1 - Fox 1.1, 1.3 - Wasserman 1 - Aronow 1.1.(1-3)
3	Distributions	- Stat20 **3.3** (https://stat20staff.netlify.app/3-probability/03-random-variables/notes) - Pitman 3.1, 3.4, 4.1, 6.1, 6.3 - Wasserman 2 - Aronow 1.(2-4)
4	Independence	- Stat20 **3.2** (https://stat20staff.netlify.app/3-probability/02-cond-prob-indep/notes#independence) - Pitman 1.4 - Wasserman 1.5, 2.7, 2.9, 17 - Aronow 1.1.4, 1.3.4
5	Expectation	- Stat20 **3.4** (https://stat20staff.netlify.app/3-probability/04-ev-se/notes) - Pitman 3.2, 6.2, 6.4 - Aronow 2
6	Estimators	- Stat20 **4.1** (https://stat20staff.netlify.app/4-generalization/01-sampling-distributions/notes) - Wasserman 6.1-2, 6.3.1, 9 - Aronow 3.1, 3.2.(1-2), 3.3
7	Inference	- Stat20 **4.2** (https://stat20staff.netlify.app/4-generalization/02-confidence-intervals/notes) , **4.3** (https://stat20staff.netlify.app/4-generalization/04-hypothesis-tests/notes) , **4.4** (https://stat20staff.netlify.app/4-generalization/05-hypothesis-tests-2/notes) - Wasserman 6.3.2, 7, 10 - Aronow 3.2, 3.4

Part II Learning Goals:

1. state the assumptions under which a generalized linear model coefficient is unbiased for a marginal causal effect
2. state the assumptions under which maximum likelihood and sandwich variance estimators for generalized linear model coefficients are consistent
3. use simulation to evaluate consistency and coverage of marginal effect estimators under different data-generating processes
4. apply regression models to infer marginal causal effects from randomized interventional data using plug-in estimation

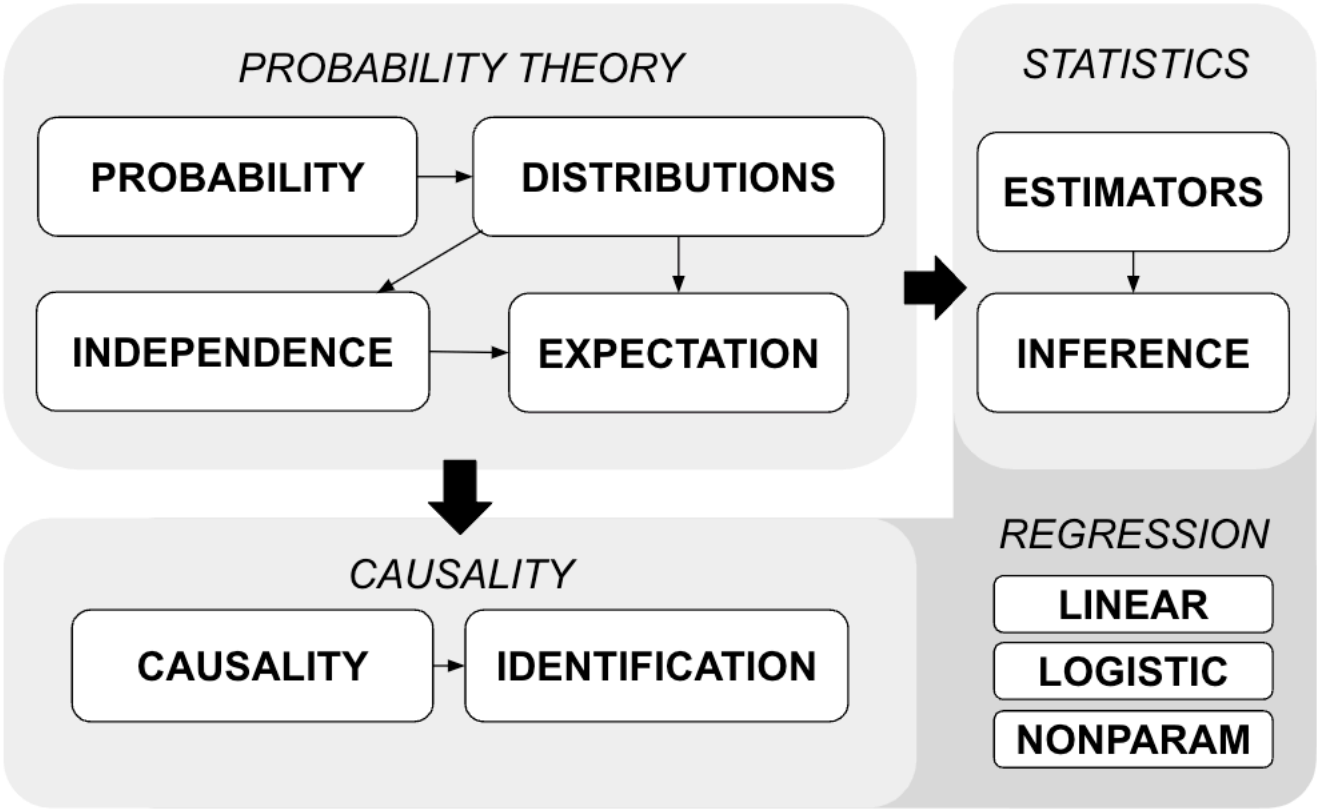
In the second part of the course we'll apply our statistical thinking to one of the most basic problem settings for epidemiology: figuring out whether one treatment works better than another. We'll dig into linear regression and see that despite its rampant use and popularity it is not always a viable approach. We'll see that it depends on whether or not the treatment is experimentally randomized, what method is used to quantify uncertainty, and exactly what we mean by "treatment effect". Understanding these nuances will require us to carefully apply the concepts we learned in the first part of the course.

My hope is that by the end of the course you will be reasonably competent in the analysis of standard observational data. You won't know how to get everywhere and solve every breakdown, but you'll have the fundamental tools to figure it out from the ground up. You'll also be a veteran of the linear regression circuit, especially as it pertains to the estimation of causal effects.

Putting it all together, you can refer to the concept map for the course to see how the foundational knowledge we'll develop in part I sets the stage for the regression models we discuss in part II. Probability theory serves as a prerequisite for statistics and causality, at the intersection of which sit regression analyses.

Week	Topic	References
8	Causality	- Stat20 **5.1** (https://stat20staff.netlify.app/5-causation/01-defining-causality/notes) - Morgan 1, 2, 8, 10 - Wasserman 16 - Aronow 7.1.(1-2)
9	Identification	- Stat20 **5.2** (https://stat20staff.netlify.app/5-causation/02-experiments/notes) - Morgan 3 - Aronow 7.1.(4-5), 7.1.7

Week	Topic	References
10	Linear Regression: Mechanics	- Stat20 6.1 (https://stat20staff.netlify.app/2-summarizing-data/06-summarizing-associations/notes) , 6.2 (https://stat20staff.netlify.app/2-summarizing-data/07-multiple-linear-regression/notes) , 6.3 (https://stat20staff.netlify.app/6-prediction/01-method-of-least-squares/notes) , 6.4 (https://stat20staff.netlify.app/6-prediction/02-improving-predictions/notes) - James 2.1, 2.2.1, 3.1.1, 3.2.1, 3.6.(1-2) - Fox 2, 5 - Wasserman 13.1, 13.5 - Aronow 4.1
11	Linear Regression: Inference	- James 3.1.(2-3), 3.3.(1-2), 3.6.(4-6) - Fox 6, 7, 19 - Aronow 4.2 - Shalizi 10 (https://www.stat.cmu.edu/~cshalizi/TALR/TALR.pdf)
12	Linear Regression: Causality	- Fox 13 - Morgan 5.2, 5.4.2 - Aronow 7.2.(1-2)
13	Logistic Regression	- Stat20 6.6 (https://stat20staff.netlify.app/6-prediction/05-logistic-regression/notes) - James 4.(1-3) - Fox 14 - Wasserman 13.7
14	Nonparametric Regression	- Stat20 6.5 (https://stat20staff.netlify.app/6-prediction/03-overfitting/notes) - James 1-2, 3.5, 5.1, 8
15	Observational Inference	- Morgan 4, 5.3 - Aronow 7.1.6, 7.2



References

In the calendar above I've provided weekly references. Theses are not required readings, they are meant to **supplement** your understanding as sources for you to consult if you get stuck, want to see things phrased a different way, or are curious to learn more. They cover more than what I will ask you to understand in this course but I've tried to pick out the specific sections that are most relevant to each week.

The sources I refer to in the reading lists are:

Book	Reference
Stat20	Bray, A., Charmichael, I., Gifford, S., Sanchez, J. Stoyanov, S. **UCB Stat 20 Lecture Notes.** (https://www.stat20.org/notes)
Pitman	Pitman, Jim. Probability. Springer Science & Business Media, 2012. **Available online from the Berkeley Library** (https://search.library.berkeley.edu/discovery/fulldisplay?docid=cdi_askewsholts_vlebooks_9781461243748&context=PC&vid=01UCS_BER:UCB&lang=en&search_scope=DN_and_CI&adaptor=Primo%20Cer
James	James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning **https://www.statlearning.com/** (https://www.statlearning.com/) New York: springer, 2013.
Fox	Fox, John. Applied regression analysis and generalized linear models. Sage Publications, 2015.

	Available online from the Berkeley Library (https://search.library.berkeley.edu/discovery/fulldisplay?docid=cdi_askewsholts_vlebooks_9781483386478&context=PC&vid=01UCS_BER:UCB&lang=en&search_scope=DN_and_CI&adaptor=Primo%20Cer
Morgan	Morgan, Stephen L., and Christopher Winship. Counterfactuals and causal inference. (https://edisciplinas.usp.br/pluginfile.php/3984640/mod_resource/content/2/%5BStephen_L._Morgan%2 Cambridge University Press, 2015.
Wasserman	Wasserman, Larry. All of statistics: a concise course in statistical inference. Vol. 26. New York: Springer, 2004. Available online from the Berkeley Library (https://search.library.berkeley.edu/discovery/fulldisplay?docid=alma991039625019706532&context=L&vid=01UCS_BER:UCB&lang=en&search_scope=DN_and_CI&adaptor=Local%20Search%20Engine&tab
Aronow	Aronow, Peter M., and Benjamin T. Miller. Foundations of agnostic statistics. Cambridge University Press, 2019. Available online from the Berkeley Library (https://search.library.berkeley.edu/discovery/fulldisplay?docid=cdi_askewsholts_vlebooks_9781316836309&context=PC&vid=01UCS_BER:UCB&lang=en&search_scope=DN_and_CI&adaptor=Primo%20Cer

Shalizi	Shalizi, C.R. The truth about linear regression **PDF (https://www.stat.cmu.edu/~cshalizi/TALR/TALR.pdf)**

The books are available in hard copy from many places but first check online by searching “[book title] [author] pdf” because many of the authors put preprints of their work up for free. They may also be available from the library in online or hardcopy format. For the rest, if you’re unsure about purchasing, I’ve heard that some students use library genesis to have a quick look at what’s in each book before committing.

In each week I have ordered the sources roughly from basic to advanced. For any given topic the more advanced sources may be beyond you — don’t worry about that, just get what you can out of them and feel free to bring any questions to office hours or the forum. I recommend looking at the equivalent section in the easier references first and then building your way up, using google or chatGPT to smooth out any bumps you encounter. Be aware that the notational conventions and terminology may differ from book to book.

Course Components

There are three sources of points:

Participation

Each lab attendance and lecture reflection count for a single point of participation credit.

Lecture Reflections

The lecture reflections are simple surveys that need to be completed each week. They ask you to summarize the lecture material, comment on what you didn’t think was clear, and describe anything that you found particularly interesting or relevant. They should take fewer than 10 minutes to complete.

Lectures can be found under the "Topics" section of each weekly page on bcourses. Lecture reflections are due every week on Monday by 11:59 PM and can be completed within the **bcourses "Discussion" tab (https://bcourses.berkeley.edu/courses/1531053/discussion_topics)** .

Labs

During labs the TAs will review the previous week’s assignments, going through all the questions and clarifying any confusions. If time permits they may answer questions about the current week’s lectures or give time to work on the current assignment. For the first week the TAs will lead a quick refresher course in R programming since there will be no assignment to review. During weeks 9 and 15 an hour will be set aside from lab to administer the make-up quizzes (see below).

Labs are led live during whatever section you signed up for.

> **FOR THE FULLY-ONLINE VERSION (241W)**: we will still host synchronous labs (on zoom, to be scheduled first week of class) but your attendance is optional to accommodate the fully-online setting. These sessions will be recorded so you have the opportunity to watch them later, but we **highly encourage you to attend live** so that you can ask questions and interact with your classmates. We will schedule the live lab to accommodate as many students as possible.

Regular Assignments

All (sub)questions on each assignment are worth the same amount. An assignment is “passed” with a score of 80% or more. If you “fail” a regular assignment, you will have the opportunity to recoup that point of assignment credit by passing the relevant portion of the make-up quiz (see below). The reason these are pass/fail grading is because the regular assignments test the "non-negotiable" core learning goals. However, you have the opportunity to make up the credit later (see below).

You may collaborate with your classmates on regular assignments, but you must write the names of anyone you collaborate with on your assignment sheet.

The problems in the regular assignments closely follow what is discussed in lecture. Course staff will be generous in discussing these problems during office hours.

Regular assignments can be found **on the assignments section on bcourses (https://bcourses.berkeley.edu/courses/1531053/assignments)** . They are due every week on Monday by 11:59pm PST via gradescope and must be submitted as pdf documents.

Make-Up Quizzes

The idea of the make-up quizzes is to give you another opportunity to demonstrate your knowledge of the basic course material. They are not a traditional “midterm” and “final” because they are not required if you are not interested in recouping credit from failed regular assignments. Moreover they can only be used to recoup regular assignment credit, not for challenge assignment credit (more on that below).

Quizzes 1 and 2 correspond to the first and second parts of the course respectively. There will be 2-3 questions per week of material and they will be similar to what is in the regular assignments. If you get all the questions correct pertaining to a given week, this will count as having passed the regular assignment for that week.

No additional credit will be given if an assignment was already passed: the quizzes are just a backup source of credit for the regular assignments and **can only help you**. However, we do encourage all students to take them even if they do not need the points because these quizzes also serve as tests *for the course staff*. If the scores are poor then *we* are doing a bad job and we want to know!

Make-up quizzes will be administered during regular lab time on weeks 9 and 15. They will be paper-and-pencil and all questions will be multiple-choice. You are allowed to bring up to 5 pages of notes that you can reference that can of course include corrected homework assignments, lecture slides, notes, etc. No computers, phones, etc. are allowed or needed.

Make-Up Final

Students **who have less than a B grade** after week 15 will have the opportunity to take a final, comprehensive make-up test that will be scheduled for finals week. Made-up credit from this test can only raise your grade to a B and no further. This test will be a little harder than the make-up quizzes so that you are motivated to do well earlier in the course and not take your chances on one final exam. The point of the make-ups really is to give you multiple opportunities and maximum flexibility to demonstrate your learning.

> 🌐 **FOR THE FULLY-ONLINE VERSION (241W):** quizzes will be done online during lab time using test-taking software. If you are unable to attend the regular lab time and wish to take the quiz we will make an alternative arrangement. **All 241W students MUST show up to least one of the two quizzes** in order to pass the course. This is to satisfy a Berkeley requirement that fully online courses have at least one proctored assignment.

Challenge Assignments

All (sub)questions on each assignment are worth the same amount. Challenge assignments are graded on a continuous basis: each question is worth the same amount and the final score for the assignment is a number between 0 and 1.

Challenge assignments must be completed individually.

The challenge assignments are meant to be more difficult and the problems will require you to extend what you learned in lecture to new situations or to combine two or more ideas. The exact ideas are not always core to the course. The purpose of the challenge assignments is to give you a bit of practice doing really hard things and finding information on your own when you are at a loss. As such, it doesn't make sense to grade these pass/fail like the regular assignments and that's why we use a continuous measure of effort.

Course staff will discuss these problems more indirectly in office hours, providing supporting information and resolving student misunderstandings but being careful not to excessively lead.

Challenge assignments are completely optional in the sense that if you are happy with a B+ there is no need for you to complete them at all (see grading scheme below). I encourage everyone to attempt the challenge questions each week but I’ve set the grading up so that you can get a good grade without doing them in case they are taking too much time away from your other courses and activities.

Challenge assignments can be found **on the assignments section on bcourses (https://bcourses.berkeley.edu/courses/1531053/assignments)** . They are due every week on Monday by 11:59pm PST via gradescope and must be submitted as pdf documents.

Grading

This course uses **specification grading** (https://www.insidehighered.com/views/2016/01/19/new-ways-grade-more-effectively-essay) . The point of specification grading is to clearly define what it takes to achieve each grade. That makes the grades meaningful and transparent and it completely also frees students from doing work they do not want to do if they are not interested in receiving a certain grade.

A final score will be computed for each course component (participation, regular assignments, challenge assignements). For example, if a student passes 12/14 of the challenge assignments, their challenge assignment score will be 86%.

Final grades are assigned on the basis of the scheme in the following table. To earn the indicated grade, the student must attain **at least** the number of points indicated **in each column** of the table below. For example, if a student gets a total of 3.5 challenge assignment points (e.g. 50% on 7 challenge assignments) and has 24/30 participation points but doesn't do any regular assignments (0/14), they would be limited to a C. To achieve a given grade, you need at least the listed score in participation AND regular AND challenge assignments. The total points possible for each category are shown in the header row. **Note that these were modified in light of the mid-course survey.**

PH 241 (Residential) Grade Table

	Participation (30)	Regular (14)	Challenge (14)
A+	24	13	**8**
A	24	**13**	**5**
A-	24	**12**	**2**
B+	24	**11**	0
B	24	**10**	0
B-	24	**8**	0
C	**24**	**5**	0
D	**14**	0	0

PH 241W (Fully Online) Grade Table

	Participation (15)	Regular (14)	Challenge (14)
A+	12	13	**8**
A	12	**13**	**5**
A-	12	**12**	**2**
B+	12	**11**	0
B	12	**10**	0
B-	12	**8**	0
C	**12**	**5**	0
D	**7**	0	0

FOR THE FULLY ONLINE VERSION (241W) lab attendance is optional so participation credit is out of 15 lecture reflections only instead of out of 15 lecture reflections + 15 labs. **All 241W students MUST show up to at least one of the two quizzes** in order to pass the course.

Note that an A+ confers no additional GPA benefit.

Late Policy

Please do not email the course staff asking for one-time participation excusals or late submissions for any assignments. The requirements above are designed to be lenient and to structurally accommodate standard absences and exceptions. We also have the make-up quizzes and final which are opportunities for you to demonstrate your knowledge and recoup any missing points from the regular assignments.

We don't want to impair anyone's learning, but at the end of the day it's a matter of respecting the TAs' time. It's also an issue for the other students in the class because if not everyone has submitted on time then we cannot start giving out solutions promptly, etc. It also doesn't matter for you: I have been on many admissions and hiring committees and never have I seen the difference between e.g. a B and a B+ grade in any course make any difference at all.

However if you have a **DSP exception** for extra time on assignments or anything else please let us know and we can of course accommodate you. Also if you have more **serious long-term issues** that require an extended absence or recurring delays (e.g. hospital stay) then please do get in touch and we will

arrange an alternative schedule that is best for your learning. In these situations you can even take an incomplete grade and have up to a year to finish the course.

Schedule

The regular schedule for the course will look like this every week:

Week	Monday Night	Monday - Friday
k	submit week $k-1$ **reflection** submit week $k-1$ **assignments**	watch week k **lectures** work on week k **assignments** review week $k-1$ **assignments during lab**

Office Hours, Forum, and Tutoring

We highly encourage everyone to take advantage of office hours and to ask questions on the forum. I am usually highly responsive.

Office hours are optional and details can be found on **the Office Hours page** (https://bcourses.berkeley.edu/courses/1531053/pages/office-hours). Come say hi!

The course discussion board is **on Ed Discussion** (https://bcourses.berkeley.edu/courses/1531053/external_tools/77220?display=borderless).

The Dream Office offers peer-to-peer group tutoring sessions for residential PH241 students. We do not offer 1:1 tutoring nor partake in grading and decision-making regarding course assignments. For more information about tutoring hours and offerings, please contact **biotutor_sphdiversity@berkeley.edu** (mailto:biotutor_sphdiversity@berkeley.edu).

Pedagogical Approach

- Our abilities are learned and can be changed with practice. We can even shift our beliefs and preferences.
- We are all inherently valuable beyond measure regardless of our histories and abilities. We all bring unique and meaningful perspectives to every community.
- We each embody a unique constellation of privileges and experiences of oppression, albeit not necessarily of the same scope or scale. We must work to recognize our privileges, dismantle oppression, and create an equitable society.
- We are accountable for our choices.
- Math is important and necessary. More on that:

To be a competent data analyst you have to be able to think rigorously and clearly within an abstract "mathematical" framework. It's just not possible to teach a "baby version" of statistics: imagine giving out pilot licenses without explaining what all the buttons in the cockpit do! That's why this class covers the usual fundamentals that you'd find in an introductory class taught in a statistics department.

But learning those abstract concepts is hard because, well, they're abstract! The typical approach I've seen is to describe these things using mathematical notation and proof, show a few figures, give a metaphor or two (if you're lucky) and pray that the students understand. That tends to works alright for people who are very used to mathematical abstractions and pretty terribly for everyone else (myself included).

My approach is to use lots of simulations and examples to make these concepts *tangible*. A paragraph of text and an equation is not tangible. But when you have a bit of R code you can break it down and understand it line-by-line: you can run it by yourself. You can even change it and design your own experiments! That's *tangible*. I use these hands-on experiments to teach the concepts, but I also use them to evaluate understanding. I'm not going to ask you to solve many equations: being able to perform algebraic manipulations isn't a good measure of how well you understand a concept. Instead, I'll more often ask you to put together a simulation to demonstrate an idea or to draw two pictures and point out the differences. I firmly believe that is a more productive way to learn. So, yes, this course has "math" but it's not about the symbols and numbers, it's about the ideas.

Some of you received messages when you were young that you aren't a "math person" and those messages likely vary based on gender, race and ethnicity, and where you went to school. That probably creates some feelings of **math anxiety** (https://en.wikipedia.org/wiki/Mathematical_anxiety), which is basically endemic, even in academia. But it's actually completely normal to struggle with math or abstraction: humans kind of **suck at thinking clearly and slowly** (https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow) because we usually don't have to in order to survive. I myself struggle with math *daily*. The good news is that, with practice, (a) you get more forgiving of yourself, (b) you start to find things more interesting and fun, and (c) as a consequence of both of these things, you get much better at math. People who are "good" at math don't have bigger brains, they've just learned how to be patient with themselves, have fun, and keep trying.

Lastly, I want to acknowledge that the course is really hard and for many of you it may be difficult to feel any sense of accomplishment. That's completely normal. It's unfortunate, but **studies repeatedly show** (https://twitter.com/emollick/status/1756396139623096695) that paradoxically you are learning way

more when you feel like you're "not getting it" and much less when you have a lot of easy victories that you can notch on your belt. There are **whole books about it** (https://www.hup.harvard.edu/books/9780674729018). And, of course, **the more you know, the dumber and more inadequate you feel** (https://en.wikipedia.org/wiki/Dunning%E2%80%93Kruger_effect). Such is the human condition (I am suffering from it as well) so please just know that you are doing great and that feeling stupid means you're getting smarter.

Additional Information

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** (https://berkeley.box.com/s/knh3rbk9ikgvmca4ymy93msgj9bkebq5)