PHP2510

RECAP WEEKS 1-4

STUDENT FEEDBACK

Thank you for identifying your challenge areas. I appreciate your candor. Below outlines the common feedback and my responses. All feedback is paraphrased and expressed without judgment.

CALC REQUIREMENTS

FEEDBACK, PARAPHRASED

- I have a hard time using integrals and summation
- I don't know what is expected

NOTES / ACTION

- The level of calculus required is specified by the pre-req self-assessment (see announcements in Canvas)
- For any question that is difficult in the assessment, you need to self-study (this website is particularly useful) or use TA office hours. I won't be devoting lecture time to the pre-reqs.
- I will never make you do any integral/derivative that is harder than ax^n on an exam, but I do expect you to be comfortable with that (for any value of a, n). When building models (now and in PHP2511), you need to be comfortable with exp, log, summations, the choose function, etc.
- After this week, the amount of calculus lessens; but formal probability notation does not go away.

LEARNING R & SIMULATIONS

FEEDBACK, PARAPHRASED

- I still have a hard time getting started
- I don't know how to use R for a solution
- I want more practice, especially doing simulations

NOTES / ACTION

- I view R as a tool for reinforcing the content of the class (e.g. simulations show a result that may be hard to otherwise understand) and for applying the content to datasets (e.g. see homework #1). If you are struggling on the steps that come *before* that, you need to self-study.
    - Watch this video: R programming for ABSOLUTE beginners
    - Read this blog: https://www.datacamp.com/tutorial/r-studio-tutorial
- I recommend reviewing these cheat sheets (run all the code locally for each relevant topic; devote X min/day): https://posit.cloud/learn/recipes
- I will work more R coding into the existing lecture materials and practice problems (e.g. see below)
- When developing labs, I will add a few extra problems for self-study. For example, a lab may have 5 questions, but I will instruct the TAs to run the lab as if there are only 3 (but post a solutions guide for all 5). You can then try to do 4-5 yourself (after completing 1-3 with the TAs).
- Developing coding skills is time-consuming, and likely takes place outside of a classroom. It's important you think first what you need your code to do (consider using pseudocode for each step), then slowly add functionality and ensure any code (if borrowed from an AI agent or online resources) actually does what you intend. Expect this skill to develop over time, not immediately. Don't use code you don't understand - it's too risky.
- Exams will not require direct coding, but may require reading code, specifying pseudo-code, and/or interpreting code output. Homework will require coding (but elegance is not graded).

CONDITIONAL AND JOINT PROBABILITIES

FEEDBACK, PARAPHRASED

- This is still a general area of confusion

NOTES / ACTION

- Try to solve the questions below
- I will go over the answers, and present some lecture content, on Thursday (10/2)

EXPECTATIONS AND VARIANCE

FEEDBACK, PARAPHRASED

- This is still a general area of confusion

NOTES / ACTION

- I will devote more lecture time to this (today, Tuesday 9/30)
- We will do the practice problems below.
- Whatever we don't get to, please do on your own. I'll post solutions on Ed Discussion

MISC / NON-CONTENT BASED

FEEDBACK, PARAPHRASED

1. I want more practice problems
2. I don't know what to expect on the exam and how to study for it
3. How does this play into real life (specifically random variable distributions)?

NOTES / ACTION

1. I do expect you to self-check your comprehension of the readings by doing the problems that the book provides (and escalating when you are stuck). I can pick out specific problems that I think are instructive and assign these as ungraded homework, but you do need to study material beyond what I do in lecture. Please ask on Ed Discussion and I will respond.
2. The exam will be closed-book and have a collection of question types (multiple choice, long-form response, answer derivations). No memorizing of random variable formulae is required. The exams will pull from both the readings and lecture materials. Checking your comprehension of provided code (aligned with the content) is fair game.
3. I acknowledge that direct discussion of public health examples is useful. I will try to pull those in more. We will work with public health datasets in homeworks.

CONDITIONAL & JOINT PROBABILITIES

Q1
A town holds an annual health fair offering free blood pressure screenings. A year later, a researcher is given the following joint probability table based on a sample of the town's residents:

	Attended Fair	**Did Not Attend**
Controlled BP	0.24	0.56
Uncontrolled BP	0.06	0.14

Are attending the health fair and having controlled blood pressure associated?

Q2
A company tracks which of its employees received a flu vaccine ("Vaccinated" or "Unvaccinated") and whether they had more than 5 sick days ("High Absence") during flu season. Assume:

- 60% of employees were "Vaccinated."
- The probability of an employee having "High Absence" is 20%.
- Being "Vaccinated" and having "High Absence" are independent events.

1. What is the probability of an employee being "Vaccinated" and having "High Absence"?
2. What is the probability of an employee being "Unvaccinated" and not having "High Absence"?
3. Why is the core assumption of this problem unusual/unlikely?

Q3
A rural area has two primary water sources: Well A and Well B.

- 70% of residents use Well A, and 30% use Well B.
- Well A has a 5% probability of being contaminated.
- Well B has a 20% probability of being contaminated.
- A person who drinks contaminated water has a 50% chance of getting sick; assume drinking uncontaminated water leads to a 0% chance of getting sick.

1. What is the overall probability that a randomly selected resident will get sick from drinking the contaminated water?
2. Use R simulations to confirm.

Q4
An epidemiologist investigates a potential link between daily diet soda consumption and Type 2 Diabetes. They conduct a case-control study, selecting 100 patients with diabetes (cases) and 200 patients without diabetes (controls). They then ask about past daily soda consumption.

- In the diabetes group, 60 people were daily soda drinkers.
- In the control group, 80 people were daily soda drinkers.

1. What does this suggest about the association between diet soda and diabetes in this study? Identify two different ways you can answer this mathematically.
2. What is different about this 2x2 from the others that we have considered?

Note: in the future, we will formalize association tests for Q1+Q4; some deviation from perfect independence is likely in a sample even when the two events are truly independent in the population

EXPECTATIONS AND VARIANCES

Q1
A hospital administrator needs to plan the budget for the upcoming flu season. Based on past data, patients who get the flu have three possible outcomes:

- Mild Case: Requires $100 in resources.
- Severe Case: Requires $1,500 in resources.
- Hospitalization: Requires $8,000 in resources.

The probability of a mild case is 65%, a severe case is 30%, and hospitalization is 5%.

1. What is the expected resource cost per flu patient?
2. What is the standard deviation in resource cost per flu patient?
3. Calculate theoretically then use R simulations to confirm

Q2
A mobile health unit screens for a genetic marker that is present in 2% of the population. They screen people one by one until they find the first positive case, then move to a new neighborhood. Let X be the number of people screened to find the first positive case.

1. What is the expected number of people the unit will screen in a neighborhood? What is the variance? *Hint: use a well-known random variable*
2. Confirm the results with R's built-in version of the random variable
3. CHALLENGE: Run the following R code and understand how it's an effective simulation that mimics (2) but doesn't use the built-in random variable (which hides what is really happening)

```
tries <- c()
for (i in 1:1000){
  find <- F
  tries[i] <- 0
  while(find == F){
    tries[i] <- tries[i] + 1
    check <- rbinom(1,1,.02)
    if(check == 1){
      find <- T
    }
  }
}
```

Q3
A city is considering two different public health programs to reduce the number of weekly traffic accidents.

- Program A (Education Campaign): Would result in an expected 10 accidents/week with a standard deviation of 4.
- Program B (New Traffic Lights): Would result in an expected 11 accidents/week with a standard deviation of 1.

Question:

1. Which program would you advocate for and why?
2. Run a simulation in R, with a Normality assumption, to get a feel for what these standard deviations mean for a given year. How many weeks is (A) strictly worse than (B)? By more than 2?

Q4 [from tidyverse handout, additional questions section]
You develop a patient-reported outcomes survey that asks each subject to score their daily level of pain on a scale of 1, 2, 3, 4, 5. After a pilot data collection, you detect that respondents are half as likely to answer (1) and (5) as they are (2), (3), (4), which they score in equal likelihoods.

1. Write down the distribution f(x).
2. What is the expectation of a patient's pain?

3. What is the variance?
4. Can you explain the variance calculation (via a step-by-step transformation or visualization, perhaps)?

Q5 [from class]
Consider the dice-throwing exercise we did that created a 10x10 joint distribution table (slide 16 here).

- Calculate, by hand, the expected value of D1
- Calculate, by hand, the expected value of D2
- Review the R code posted on Ed Discussion and confirm you understand what it is doing
- Use R simulations to calculate the variance of: D1, D2, D1+D2