Homework_12

April 22, 2025

Probability for Data Science

UC Berkeley, Spring 2025

Michael Xiao and Ani Adhikari

1 Homework 12 (Due Monday, April 21st at 5 PM)

1.0.1 How to Do Your Homework

The point of homework is for you to try your hand at using what you've learned in class. The steps to follow:

- Go to lecture and sections, and also go over the relevant text sections before starting on the homework. This will remind you what was covered in class, and the text will typically contain examples not covered in lecture. The weekly Study Guide will list what you should read.
- Work on some of the practice problems before starting on the homework.
- Attempt the homework problems by yourself with the text, section work, and practice materials all at hand. Sometimes the week's lab will help as well. The two steps above will help this step go faster and be more fruitful.
- At this point, seek help if you need it. Don't ask how to do the problem — ask how to get started, or for a nudge to get you past where you are stuck. Always say what you have already tried. That helps us help you more effectively.
- For a good measure of your understanding, keep track of the fraction of the homework you can do by yourself or with minimal help. It's a better measure than your homework score, and only you can measure it.

1.0.2 Rules for Homework

- Every answer should contain a calculation or reasoning. For example, a calculation such as $(1/3)(0.8) + (2/3)(0.7)$ or `sum([(1/3)*0.8, (2/3)*0.7])`is fine without further explanation or simplification. If we want you to simplify, we'll ask you to. But just $\binom{5}{2}$ by itself is not fine; write "we want any 2 out of the 5 frogs and they can appear in any order" or whatever reasoning you used. Reasoning can be brief and abbreviated, e.g. "product rule" or "not mutually exclusive."
- You may consult others (see "How to Do Your Homework" above) but you must write up your own answers using your own words, notation, and sequence of steps.

- We'll be using Gradescope. You must submit the homework according to the instructions at the end of homework set.

1.1 We will not grade assignments which do not have pages correctly selected for each question.

1.2 1. Prediction with Sums

Let $X_1, X_2, \ldots, X_n$ be i.i.d. with expectation μ and variance σ^2. Let $S = \sum_{i=1}^{n} X_i$.

a) Find the least squares predictor of S based on X_1, and find the mean squared error (MSE) of the predictor.

b) Find the least squares predictor of X_1 based on S, and find the MSE of the predictor. Is the predictor a linear function of S? If so, it must also be the best among all linear predictors based on S, which is commonly known as the regression predictor.

[Consider whether your predictor in (b) would be different if X_1 were replaced by X_2, or by X_3, or by X_i for any fixed i. Then use symmetry and the additivity of conditional expectation.]

1.3 2. A Covariance Identity

Let X and Y be jointly distributed random variables, and as in Section 22.1 let $b(X) = E(Y \mid X)$. Show that $Cov(X, Y) = Cov(X, b(X))$.

1.4 3. Sections ## -

A class of 60 students has three sections. Summary statistics for scores on Quiz 7: - Section 1: 25 students, mean 25, SD 3 - Section 2: 20 students, mean 23, SD 2 - Section 3: 15 students, mean 27, SD 4

Let S be the Quiz 7 score of a student picked at random from the class. Use the code cell below to calculate numerical answers for parts a and b.

a) Find $E(S)$.

b) Find $SD(S)$.

[1]:
```
# Calculation for a

exp_s = (25*25)/60 + (20*23)/60 + (15*27)/60
exp_s
```

[1]:
```
24.833333333333332
```

[2]:
```
# Calculation for b

e_var = (25*9)/60 + (20*4)/60 + (15*16)/60
var_e = (25*(25-exp_s)**2)/60 + (20*(23-exp_s)**2)/60 + (15*(27-exp_s)**2)/60

var_s = e_var + var_e
sd_s = var_s**(1/2)
sd_s
```

[2]:
```
3.3747427885527643
```

1.5 4. Accident Claims

Poisson processes are standard models used by actuaries and industrial engineers.

At a street intersection, accidents happen according to a Poisson process at a rate of λ per month. Each accident leads to an insurance claim of a random dollar amount. Assume that the amounts of the claims are i.i.d. with mean μ and SD σ, independent of the process of accidents.

Find the expectation and variance of the total amount claimed for accidents at that intersection in a year.

1.6 5. Two-Colored Die

A die that has two green faces and four blue faces is rolled repeatedly. Let R be the number of rolls till both colors have appeared.

In each part below, find the numerical value.

a) Find $E(R)$.

b) Find $SD(R)$.

1.7 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

1.7.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using applications such as CamScanner. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image in CamScanner or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.
- If you used LaTeX to do the written portions, you do not need to do any scanning; you can just download the whole notebook as a PDF via LaTeX.

1.7.2 Code Portion

- Save your notebook using `File > Save and Checkpoint`.
- Generate a PDF file using `File > Download As > PDF via LaTeX`. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please post a follow-up on the general Homework 12 Ed thread.

1.7.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Homework 12 on Gradescope.
- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

If you are having difficulties scanning, uploading, or submitting your work, please read the Ed Thread on this topic and post a follow-up on the general Homework 12 Ed thread.

[ ]:

1a. $X_1, X_2, \ldots, X_n \sim iid$

$E[X_i] = \mu$

$Var(X_i) = \sigma^2$

$S = \sum_{i=1}^{n} X_i$

Least Squares Predictor : $E[S|X_1]$

$$E[S|X_1] = E[X_1 + X_2 + \ldots + X_n | X_1]$$
$$= E[X_1|X_1] + E[X_2|X_1] + \ldots + E[X_n|X_1]$$
$$= X_1 + \mu + \ldots + \mu$$
$$= X_1 + (n-1)\mu$$

$$MSE(g(x)) = E[(x - g(x))^2]$$
$$MSE(E[S|X_1]) = E[(S - E[S|X_1])^2]$$
$$= E[(S - (X_1 + (n-1)\mu))^2]$$
$$= E[(S - X_1 - (n-1)\mu)^2]$$
$$= E[(X_1 + X_2 + \ldots + X_n - X_1 - (n-1)\mu)^2]$$
$$= E[(X_2 + \ldots + X_n - (n-1)\mu)^2]$$
$$= E[(X_2 + \ldots + X_n) - (E[X_2 + \ldots + X_n])^2]$$
$$= Var(X_2 + \ldots + X_n)$$
$$= (n-1)\sigma^2$$

1b. $E[X_1|S] = E[X_2|S] = \ldots = E[X_n|S]$

$$E[X_1 + X_2 + \ldots + X_n | S] = E[S|S]$$
$$= S$$

$$E[X_1|S] + E[X_2|S] + \ldots + E[X_n|S] = S$$
$$n\,E[X_1|S] = S$$
$$E[X_1|S] = \frac{S}{n}$$
$$= \bar{X}$$

$$MSE(\bar{x}) = E[(x_1 - \bar{x})^2]$$

$$E[x_1 - \bar{x}] = E[\mu - \mu]$$
$$= E[0]$$
$$= 0$$

$$= E[((x_1 - \bar{x}) - \underbrace{E(x_1 - \bar{x})}_{=0})^2]$$
$$= Var(x_1 - \bar{x})$$
$$= Var(x_1) + Var(\bar{x}) - 2Cov(x_1, \bar{x})$$
$$= \sigma^2 + \frac{1}{n^2} \cdot n\sigma^2 - 2Cov(x_1, \bar{x})$$

$$Cov(x_1, \bar{x}) = Cov(x_1, \frac{1}{n}(x_1 + \dots + x_n))$$
$$= \frac{1}{n} Cov(x_1, (x_1 + \dots + x_n))$$
$$= \frac{1}{n} (Cov(x_1, x_1) + \underbrace{\dots + Cov(x_1, x_n)}_{=0})$$

$\therefore$ i.i.d

$$= \frac{1}{n} Var(x_1)$$
$$= \frac{1}{n} \sigma^2$$

$$= \sigma^2 + \frac{\sigma^2}{n} - \frac{2\sigma^2}{n}$$
$$= \frac{n\sigma^2 + \sigma^2 - 2\sigma^2}{n}$$
$$= \sigma^2 \left(\frac{n-1}{n}\right)$$

2. show $Cov(X, Y) = Cov(X, b(X))$ ✓

$Dw = Y - b(X)$

$Cov(X \mid Dw) = 0$ textbook 22.1.3 "orthogonality principle"

$Cov(X \mid Y - b(X)) = 0$

$Cov(X, Y) - Cov(X, b(X)) = 0$

$Cov(X, Y) = Cov(X, b(X))$ ✓

3a. $S \sim$ quiz 7 score of a student picked at random from the class

$$E[S] = P_1 E[S_1] + P_2 E[S_2] + P_3 E[S_3]$$

$$E[S] = \frac{25(25) + 20(23) + 15(27)}{25 + 20 + 15}$$

$$= \frac{625 + 460 + 405}{60}$$

$$= \frac{1490}{60}$$

$$\approx 24.8333$$

3b. S_1 : 25 students

$\mu = 25$

$\sigma = 3^2 = 9$

S_2 : 20 students

$\mu = 23$

$\sigma = 2^2 = 4$

S_3 : 15 students

$\mu = 27$

$\sigma = 4^2 = 16$

$$SD(S) = \sqrt{Var(S)}$$

$$Var(S) = E[Var(Y|X)] + Var(E[Y|X])$$

$$= E[Var(S|S_i)] + Var(E[S|S_i])$$

$$E[Var(S|S_i)] = \frac{25(9) + 20(4) + 15(16)}{60}$$

$$= \frac{225 + 80 + 240}{60}$$

$$= \frac{545}{60}$$

$$\approx 9.0833$$

$$\text{Var}(E[S|S_i]) = \frac{25(25 - 24.8333)^2 + 20(23 - 24.8333)^2 + 15(27 - 24.8333)^2}{60}$$

$$= \frac{25(0.0278) + 20(3.3610) + 15(4.6946)}{60}$$

$$= \frac{0.695 + 67.22 + 70.419}{60}$$

$$= \frac{138.334}{60}$$

$$\approx 2.3056$$

$$\therefore \text{Var}(S) = 9.0833 + 2.3056$$

$$= 11.3889$$

$$\therefore SD(S) = \sqrt{11.3889}$$

$$\approx 3.375$$

4. $Y \sim$ # accidents in a year

$\sim$ Poisson $(\lambda) \longrightarrow$ per month

$\sim$ Poisson $(12\lambda) \longrightarrow$ 12 months

$X_i \sim$ total amount claimed at i^{th} accident

$E[X_i] = \mu$

$SD(X_i) = \sigma$

$T \sim$ total amount claimed for all accidents in a year

$T = X_1 + X_2 + \dots + X_Y$

$= \sum_{i=1}^{Y} X_i$

$E[T] = E\left[E\left[\sum_{i=1}^{Y} X_i \mid Y\right]\right]$

$= E\left[\sum_{i=1}^{Y} E[X_i]\right]$

$= E[Y\mu]$

$= \mu E[Y]$

$= 12\lambda\mu$

$Var(T) = E[Var(T \mid Y)] + Var(E[T \mid Y])$

$Var(T \mid Y) = Var\left(\sum_{i=1}^{Y} X_i \mid Y\right)$

$= \sum_{i=1}^{Y} Var(X_i)$

$= Y\sigma^2$

$E[Var(T \mid Y)] = E[Y\sigma^2]$

$= \sigma^2 E[Y]$

$= 12\lambda\sigma^2$

$E[T \mid Y] = Y\mu$

$$\begin{aligned} \mathrm{Var}(E[T|Y]) &= \mathrm{Var}(Y\mu) \\ &= \mu^2 \mathrm{Var}(Y) \\ &= 12\lambda\mu^2 \end{aligned}$$

$$\begin{aligned} \therefore \mathrm{Var}(T) &= 12\lambda\sigma^2 + 12\lambda\mu^2 \\ &= 12\lambda(\sigma^2 + \mu^2) \end{aligned}$$

5a. $R \sim$ # rolls till both colors have appeared

$P(\text{green face}) = \frac{2}{6}$

$P(\text{blue face}) = \frac{4}{6}$

$R = 1 + X$

↳ additional # rolls to get both colors

$E[R] = 1 + P(\text{green first})\, E[X \mid \text{green first}] + P(\text{blue first})\, E[X \mid \text{blue first}]$

$E[X \mid \text{green}] = \frac{1}{4/6}$ geometric

$= \frac{3}{2}$

$E[X \mid \text{blue}] = \frac{1}{2/6}$

$= 3$

$\therefore E[R] = 1 + \frac{1}{3} \cdot \frac{3}{2} + \frac{2}{3} \cdot 3$

$= 1 + \frac{1}{2} + 2$

$= 3.5$

5b. $\text{Var}(R) = \text{Var}(1 + X)$

$= \text{Var}(X)$

$= E[\text{Var}(X \mid \text{first roll})] + \text{Var}(E[X \mid \text{first roll}])$

$E[X \mid \text{green}] = 3/2$

$\text{Var}(X \mid \text{green}) = \frac{1 - 2/3}{(2/3)^2}$

$= \frac{1}{3} \cdot \frac{9}{4}$

$= \frac{3}{4}$

$$E[X \mid \text{blue}] = 3$$

$$\text{Var}(X \mid \text{blue}) = \frac{1 - 1/3}{(1/3)^2}$$

$$= \frac{2}{3} \cdot 9$$

$$= 6$$

$$E[\text{Var}(X \mid \text{first roll})] = \frac{1}{3} \cdot \frac{3}{4} + \frac{2}{3} \cdot 6$$

$$= \frac{1}{4} + 4$$

$$= \frac{17}{4}$$

$$E[X \mid \text{first roll}] = \frac{1}{3} \cdot \frac{3}{2} + \frac{2}{3} \cdot 3$$

$$= \frac{1}{2} + 2$$

$$= \frac{5}{2}$$

$$\text{Var}(E[X \mid \text{first roll}]) = \frac{1}{3}\left(\frac{3}{2} - \frac{5}{2}\right)^2 + \frac{2}{3}\left(3 - \frac{5}{2}\right)^2$$

$$= \frac{1}{3}(1) + \frac{2}{3}\left(\frac{1}{4}\right)$$

$$= \frac{1}{3} + \frac{1}{6}$$

$$= \frac{3}{6}$$

$$= \frac{1}{2}$$

$$\therefore \text{Var}(R) = \frac{17}{4} + \frac{1}{2}$$

$$= \frac{19}{4}$$

$$SD(R) = \sqrt{19/4}$$

$$\approx 2.179$$