Midterm Review Session problems

October 6, 2025

Exercise 1 (Filling a contingency table from conditionals)**:** You are told that (X, Y) is a pair of random variables where $X \in \{0, 1\}$ and $Y \in \{0, 1, 2\}$. The joint distribution is not given, but you are told the following information:

$$P(Y = 0) = 1/2, \quad P(Y = 1) = 1/4, \quad P(Y = 2) = 1/4$$

and the conditional probabilities of X given Y are

$$P(X = 1 \mid Y = 0) = 1/2, \quad P(X = 1 \mid Y = 1) = 1/4, \quad P(X = 1 \mid Y = 2) = 1/4$$

Fill in the contingency table for the joint distribution $P(X = x, Y = y)$:

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$			
$X = 1$			

Exercise 2 (Covariance of two Bernoulli random variables)**:** Let (X, Y) be a pair of Bernoulli random variables with joint distribution given by

	$Y = 0$	$Y = 1$
$X = 0$	0.3	0.2
$X = 1$	0.1	0.4

Compute $\mathrm{cov}(X, Y)$.

Exercise 3 (Sample distribution)**:** Consider the problem of estimating the mean of a normal random variable $X \sim \text{Normal}(\mu, \sigma^2)$ with known variance σ^2. You have N iid samples $X_1, \ldots, X_N$ of X where μ is unknown. Write down an estimator whose sample distribution is

$$\hat{\mu}_{☺} \sim \text{Normal}\left(\mu a + c/N, \frac{a^2\sigma^2}{N}\right), \quad a, c \neq 0$$

Exercise 4: In class we learned that given N samples $X_1, \ldots, X_N$ from a Normal distribution with unknown mean

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

is a biased estimator of the variance. Explain (without writing out any code) how you could test that this is indeed biased in Python. Specifically, describe what data you would generate and a plot you could make.

Exercise 5 (Translating code to math)**:** What will the following code print?

```
> import numpy as np
> n = 1000
> x = np.random.binomial(3,1/2,size=n)
> y = np.random.normal(1+ x,1,n)
> print(np.mean(y))
```

Exercise 6 (Translating math to code)**:** Consider the model

$$X_1 \sim \text{Bernoulli}(1/2)$$
$$X_2|X_1 \sim \text{Bernoulli}(X_1/2 + 1/4)$$
$$X_3|X_2 \sim \text{Normal}(X_1 - X_2, X_1 + 1)$$

Write a python code to approximate $P(X_3 < 2|X_2 = 0)$.

Exercise 7 (Recovering regression parameters)**:** Let

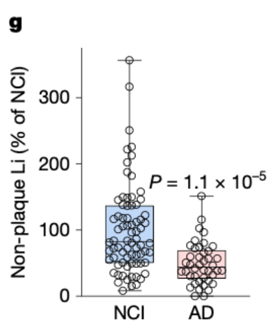
$$Y \mid X \sim \text{Normal}(\beta_0 + \beta_1 X,\ \sigma_\varepsilon^2), \tag{1}$$

Given the marginal mean, expectation and correlation coefficient

$$\mu_X = 2, \qquad \sigma_X^2 = 1, \qquad \mu_Y = 5, \qquad \sigma_Y^2 = 4, \qquad \rho = 1/2$$

FInd β_1, β_0 and σ_ε^2.

Exercise 8: The following is a plot from the paper *Lithium deficiency and the onset of Alzheimer's disease* published in Nature. It shows Lithium levels in two groups (NCI = no cognitive impairment, AD = Alzheimer's disease). Why might a linear regression model with the group as the predictor X and Lithium level as the response Y be inappropriate for this data?



Solutions

Exercise 1: Filling a contingency table from conditionals

Using

$$P(X=1, Y=y) = P(X=1 \mid Y=y)\, P(Y=y), \qquad P(X=0, Y=y) = (1 - P(X=1 \mid Y=y))\, P(Y=y). \tag{2}$$

we get

	$Y=0$	$Y=1$	$Y=2$
$X=0$	1/4	3/16	3/16
$X=1$	1/4	1/16	1/16

which sum to 1 as a check.

Exercise 2: Covariance of two Bernoulli random variables

Compute the needed expectations:

$$\mathbb{E}[X] = P(X=1) = 0.1 + 0.4 = 0.5, \tag{3}$$

$$\mathbb{E}[Y] = P(Y=1) = 0.2 + 0.4 = 0.6, \tag{4}$$

$$\mathbb{E}[XY] = 1 \cdot 1 \cdot P(X=1, Y=1) = 0.4. \tag{5}$$

Hence,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.4 - (0.5)(0.6) = 0.1. \tag{6}$$

Exercise 3: Sample distribution

Let $\overline{Y}_N = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Then

$$\overline{Y}_N \sim \mathrm{Normal}\left(\mu,\ \frac{\sigma^2}{N}\right). \tag{7}$$

Define the estimator

$$\widehat{\mu}_{☺} \ = \ a\overline{Y}_N + \frac{c}{N}. \tag{8}$$

By linear transformation of a Normal,

$$\widehat{\mu}_{☺} \sim \mathrm{Normal}\left(a\mu + \frac{c}{N},\ \frac{a^2\sigma^2}{N}\right), \tag{9}$$

which matches the required form.

Exercise 4: Testing bias of $\hat{\sigma}_1^2$ in Python (no code required)

To empirically demonstrate bias of

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2, \tag{10}$$

you could:

1. Fix a ground-truth Normal model, e.g. $X_i \sim \mathrm{Normal}(\mu_0, \sigma_0^2)$ with known μ_0, σ_0 and a chosen N.
2. Repeat for many trials (e.g. 50,000): simulate a size-$N = 10$ sample, compute $\hat{\sigma}_1^2$ for each trial.
3. Plot the *sampling distribution* (histogram) of $\hat{\sigma}_1^2$ and overlay a vertical line at σ_0^2. The sample mean

   $$\overline{\hat{\sigma}_1^2} \ = \ \frac{1}{T}\sum_{t=1}^{T}\hat{\sigma}_{1,t}^2 \tag{11}$$

   will lie systematically below σ_0^2, visualizing the negative bias.

Exercise 5: Translating code to math

The model is a linear regression model, using the formula $\mu_Y = \beta_0 + \beta_1\mu_X$ we get

$$\mathbb{E}[y] = 1 + \tfrac{3}{2} = \tfrac{5}{2} \tag{12}$$

Therefore the code will print a number close to 2.5.

Exercise 6: Translating math to code

```
import numpy as np
N = 100000
x1 = np.random.binomial(1, 1/2, size=N)
x2 = np.zeros(N)
x3 = np.zeros(N)
for i in range(N):
    x2[i] = np.random.binomial(1, x1[i]/2 + 1/4)
    x3[i] = np.random.normal(x1[i] - x2[i], np.sqrt(x1[i] + 1))

print(np.mean(x3[x2==0]<2))

# or
print(len(x3[(x2==0) & (x3<2)])/len(x3[x2==0]))
```

Exercise 7: Recovering regression parameters

Under the model $Y \mid X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$, we have

$$\text{Cov}(X, Y) = \beta_1 \, \text{Var}(X), \tag{13}$$

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \beta_1 \frac{\sigma_X}{\sigma_Y}. \tag{14}$$

Hence

$$\beta_1 = \rho \frac{\sigma_Y}{\sigma_X} = \frac{1}{2} \cdot \frac{2}{1} = 1. \tag{15}$$

Then

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] = 5 - 1 \cdot 2 = 3, \tag{16}$$

and using $\text{Var}(Y) = \beta_1^2 \text{Var}(X) + \sigma_\varepsilon^2$,

$$\sigma_\varepsilon^2 = \sigma_Y^2 - \beta_1^2 \sigma_X^2 = 4 - 1^2 \cdot 1 = 3. \tag{17}$$

Exercise 9: Regression appropriateness for the lithium plot

The variance $\text{var}(Y|X)$ appears to differ between groups.