Assignment 1 (three pages)

Statistics 32950-24620 (Spring 2023)

Due 9 am, Tuesday, March 28th.

Requirements

- Your answers should be typed or clearly written, started with your name, Assignment 1, STAT 32950 or 24620; saved as `LastnameFirstnamePset1.pdf`, and uploaded to Gradescope under either 329Pset1 or 246Pset1.
  Make sure to **submit to the correct course number** you registered, and tag the pages for each question.
- When you use R (or others) to solve problems such as Question 2 in this assignment, select only relevant parts of the output, edit, then insert in your writing.
- You may discuss approaches with others. However the assignment should be devised and written by yourself.

Problem assignments

(Corresponding to Johnson and Wichern's chapters 1, 2, 3, and related background for chapters 4 and 8)

1. (*Formulation of joint, marginal and conditional distributions of discrete random variables*)

   This Easter Sunday you are playing a game with a not-so-good friend Mr Trick, so you don't mind not seeing him again after the game. Take three identical looking eggs and label them as follows. The first egg is marked with the number 1 in blue, and the number 2 in red. The second egg is marked 3 in blue and 4 in red. The third ball is marked 5 in blue and 0 in red. Let B denote the blue number and R the red number on an egg. (B and R are variables with random outcomes.)

   Put the three eggs in a basket and cover with some grass. You pick a color. Mr Trick takes the other color. The rules are one of the following stated in (b), (c), and (d). The large number wins.

   (a) Let R denote the outcome number in red, B the number in blue. List the joint probabilities of (R, B) in a 3×3 table, then add the margins (row/column probabilities).

   (b) Rule-I: A single egg is taken at random from the basket, determining both the blue (B) and the red (R) numbers. Which color should you choose under Rule-I?
   Support your choice by computing your winning probability, using joint probabilities of (R, B).

   (c) Rule-II: You take an egg at random and read the number of your color. Then you put the egg back into the basket and cover all eggs. Then Mr Trick chooses an egg at random and reads the number of his color. Which color should you choose under Rule-II?
   Compute your winning probability, using marginal probabilities of R and B.

   (d) Rule-III: Like Rule-II, but your egg is not put back, so Mr Trick only has two remaining eggs (properly covered) to choose from. Which color should you choose under Rule-III?
   Compute your winning probability, using conditional probabilities of $B|R$ or $R|B$, by applying the law of total probability $\mathbb{P}(Y = y) = \sum_x \mathbb{P}(Y = y|X = x)\,\mathbb{P}(X = x)$.

2. (*Basic description of multivariate data*)

   Download/input the dataset https://www.stat.uchicago.edu/~meiwang/courses/s23-mva/ladyrun23.dat(clickable).

   The data are on national track records for women, updated from Table 1-9 in J&W. Measurements for 100m, 200m, and 400m are in seconds, longer distance records are in minutes. Variable names are not included.

   The following R command can be used to input the data (after saving the data in your working directory):

```
ladyrun = read.table("ladyrun23.dat")
colnames(ladyrun)=c("Country","100m","200m","400m","800m","1500m","3000m","Marathon")
```

Compute the following (rounded to 2 decimal places) for the dataset.

(a) Sample means.
Is there any variable for which the mean is not meaningful (same judgement for the following questions)?

(b) Sample covariance matrix and correlation matrix. Just the R command, no need to print the output.

(c) Sample correlation matrix using Kendall's τ. Just the R command, no need to print the output.

(d) Sample correlation matrix using Spearman's ρ. Just the R command, no need to print the output.

(e) All three types of correlation matrix (Pearson, Kendall, Spearman) on the logarithm of the data. Again, just the R command, no need to print the output.
Are the results the same as in (b), (c), and (d)? Why?

(f) Obtain the eigenvalues (show only 2 decimal places) and the eigenvectors (command only, no need of output) for the sample correlation matrix $\boldsymbol{R}$ (for all numerical valued variables).
What is the sum of all eigenvalues? Compare it to the dimensions of the variables.

(Useful R commands: `mean, cov, cor, eigen`)

3. (*Conditional expectation and conditional variance, discrete case*)

The following table lists the join probabilities of random variables X and Y.

	Y=1	Y=2	Y=3	Y=4
X=1	c	c	0	0
X=2	c	c	c	0
X=3	c	c	c	c

(a) Find the value of c. Derive the marginal probability mass functions $f_X(x) = \mathbb{P}(X = x)$ and $f_Y(y)$.

(b) Find the conditional expectation $g(x) = \mathbb{E}(Y \mid X = x)$ for $x = 1, 2, 3$.

(c) Find the conditional variance $Var(Y \mid X = x)$ for $x = 1, 2, 3$.

(d) Evaluate $\mathbb{E}\big[\mathbb{E}(Y \mid X)\big] = \mathbb{E}[g(X)]$. Verify that it equals $\mathbb{E}(Y) = \sum_y y f_Y(y)$ using results in (b).

(e) Evaluate $Var\big[\mathbb{E}(Y \mid X)\big]$, then derive the variance of Y by using $Var(Y) = Var\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[Var(Y|X)\big]$.

4. (*Derivations*)

(a) (*Spectral decomposition*) Let $A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ for $\rho \in (0, 1)$.

i. Derive the eigenvalues (λ_i's) of A (by hand, show work).
ii. Derive unit-length eigenvectors (v_i's) of A and show that they are orthogonal (by hand, show work).
iii. Write out the spectral decomposition (*a.k.s* eigen-decomposition) $A = V\Lambda V^T$, where the columns of V are orthonormal eigenvectors, and Λ is the diagonal matrix of eigenvalues of A.
iv. Use the spectral decomposition to write A^{-1} in terms of (matrix operations of) V and Λ.
v. Use the spectral decomposition to find $R = A^{1/2}$ (in terms of operations of V and Λ) such that $A = R^2$.

(b) (*Positive semi-definiteness of covariance matrix*) Show that the covariance matrix $\Sigma = Cov(\boldsymbol{X})$ of a random vector $\boldsymbol{X} = [X_1, , ..., X_p]^T \in \mathbb{R}^p$ must have all eigenvalues nonnegative.

(c) (*Integral form for Kendall's τ*) For continuous independent bivariate random vectors (X_1, Y_1) and (X_2, Y_2) with the same distribution function, Kendall's τ can be defined as

$$\tau = \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - \mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Show that

$$\tau = 2\,\mathbb{P}[(X_1 - X_2)(Y_1 - Y_2) > 0] - 1 \;=\; 4 \iint_{\mathbb{R}^2} F(x, y) dF(x, y) - 1$$

where the cumulative distribution function $F(x, y) = \mathbb{P}(X_i < x, Y_i < y)$ is continuous.

(d) (*Expectation of random matrix*) Let $\boldsymbol{C} = A\boldsymbol{X}B$, where $\boldsymbol{X}$ is a $p \times p$ random matrix, A, B are scalar (non-random) matrices of dimensions $k \times p$ and $p \times r$ respectively.

i. What are the dimensions of matrix $\boldsymbol{C}$?

ii. Write down c_{ij}, the (i, j)th entry of $\boldsymbol{C}$, in terms of elements of A, B and $\boldsymbol{X}$. (Note: The expression has to be general, not for specific values.)

iii. Show that $\mathbb{E}(\boldsymbol{C}) = A\,\mathbb{E}(\boldsymbol{X})B$.

5. (*Properties related to principal component analysis*)

The trivariate random vector $X = (X_1, X_2, X_3) = [X_1\ X_2\ X_3]'$ has covariance matrix

$$Cov(X) = \Sigma = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 4 & 0 \\ 1 & 0 & 3 \end{bmatrix}$$

(a) What are the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of Σ (ordered as $\lambda_1 \geq \lambda_2 \geq \lambda_3$)? (Keep two decimal places.)

(b) Write (principal component) random variable Y_i as $Y_i = \boldsymbol{a}_i' X = a_{i1}X_1 + a_{i2}X_2 + a_{i3}X_3$, where $\boldsymbol{a}_i$ is an eigenvector of Σ with eigenvalue λ_i such that $\|\boldsymbol{a}_i\|^2 = a_{i1}^2 + a_{i2}^2 + a_{i3}^2 = 1$, for $i = 1, 2, 3$.

(c) Derive the value of $Var(Y_1)$, the variance of Y_1, using the variance properties

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j \leq n} Cov(X_i, X_j), \qquad Var(cX_i) = c^2\, Var(X_i) \quad \text{for constant } c.$$

Then compare the value of $Var(Y_1)$ with the λ_i's.

The following R commands can be used to input the given covariance matrix and obtain eigenvalues and eigenvectors:

```
> sigma=matrix(c(2,-1,1,-1,4,0,1,0,3),3,3)
> eigen(sigma)
```

6. (*Joint, marginal, conditional density and expectation, continuous case*)

The joint density of random variables (X, Y) is

$$f_{XY}(x, y) = \begin{cases} \dfrac{c}{(1 + x + y)^3} & \text{if } \ 0 \leq x, y, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Derive the value of c.

(b) Derive the marginal density of X.

(c) Derive the conditional density $f_{Y|X}(y \mid x)$ for $x > 0$.

(d) (**Required for 32950 students only**. Optional for 24620)

i. What is the value of $\mathbb{E}(Y)$? Show your derivations.

ii. Derive the conditional expectation $g(x) = \mathbb{E}(Y \mid X = x)$ for $x > 0$, with detailed integration steps without quoting integral formulas.