W203 Statistics for Data Science

Aditya Mengani

aditya.mengani@berkeley.edu

Monday, 4pm

Summer 2020

Unit 4 Homework: Joint Probability Distributions

May 25, 2020

Problem 1 :

Unladen Swallows

In the async lecture, we built a model consisting of two random variables: Let W represent the wingspan of a swallow, and V represents the velocity.

We assume W has a normal distribution with mean 10 and standard deviation 4.
We assume that $V=0.5 \cdot W+U$V=0.5□W+U, where UU is a random variable (which we might call error). We assume that UU has a standard normal distribution and is independent of WW.

Using properties of variance and covariance, derive each element of the variance-covariance matrix for W and V. (3 points)

Given $V = 0.5W + U$ $\mu_W = 10$ $\sigma_W^2 = 16$ (since $\sigma_W = 4$)

It is also given that U is a random variable having standard normal distribution, independent of W

The variance of 'V' can be calculated as below:-

$Var[V] = Var[0.5W + U]$ (since $V = 0.5W + U$)

- According to rules of variance, $var(ax)$ can be written as $a^2 . var(x)$
- Similarly $var[x+y] = var(x) + var(y)$ if x and y are independent

So, we know W and U are independent, hence their variance can be written as below:

$\boxed{Var[V] = (0.5)^2 \, var(W) + var(U)}$ —— (a)

we know $var(W) = \sigma_W^2 = 16$

- From the rules of standard normal distribution, variance of a standard normal distribution r.v. is always 1

thus substituting $var(W) = 16$ and $var(U) = 1$ in (a) above

$$\boxed{Var(V) = \frac{1}{4} \times 16 + 1 = 4 + 1 = 5}$$

Thus $var(V) = 5$ and $Var(W) = 16$

Now, we can calculate the covariance of V and W as below:

$$\boxed{Cov(X,Y) = E(X,Y) - \mu_x \mu_y} \quad \text{--- (c)}$$

using this proposition, $cov(V,W)$ can be written as:

$$cov(V,W) = E(VW) - \mu_V \cdot \mu_W$$

we know $\mu_W = 10$. To calculate $E(V)$ we use the below approach:

$$E(V) = E(0.5W + U)$$

$$= 0.5E(W) + E(U)$$

(using the additive rule the expectation can be expressed as a summation)

$$E(V) = 0.5 \times 10 + E(U)$$

Since, 'U' is a standard normal distribution, its expectation is always 0 $\Rightarrow E(U) = 0$

Thus, $\boxed{E(V) = 5 + 0 = 5}$ --- (b)

To calculate $E(V \cdot W)$ we use the below approach:

$$E(VW) = E((0.5W + U) \cdot W)$$

$$= E(0.5W^2 + UW)$$

$$\boxed{E(VW) = E(W^2) \times 0.5 + E(UW)} \quad \text{--- (d)}$$

$E(UW) = 0$, since U and W are independent,
$E(UW) = E(U) \cdot E(W)$, we know $E(U) = 0$
thus $\boxed{E(UW) = 0}$ --- (e)

substituting (e) into (d),

$$\boxed{E(VW) = 0.5 \cdot E(W^2)} \quad - (f)$$

From rules of variance

$$var(W) = E(W^2) - [E(W)]^2$$

Thus $16 = E(W^2) - 100$

$$\boxed{E(W^2) = 116} \quad \text{—— } (g)$$

Thus $E(VW) = \frac{116}{2} = 58$ (substituting (g) in (f) above)

~~Note~~ earlier, $Cov(V,W) = E(VW) - \mu_V \cdot \mu_W$

$$= 58 - 10 \times 5$$

$$= 58 - 50$$

$$= 8$$

Thus $cov(V,W) = 8$

The covariance - variance matrix of V and W can be written as follows :-

$$\begin{bmatrix} 16 & 8 \\ 8 & 5 \end{bmatrix}$$

Problem 2 :

Relating Min and Max

Continuous random variables XX and YY have a joint distribution with probability density function,

$$f(x,y)=\{2,0,0<y<x<1 otherwise.$$ f(x,y)={2,0<y<x<10,otherwise.

You may wonder where you would find such a distribution. In fact, if A_1A1 and A_2A2 are independent random variables uniformly distributed on $[0,1]$[0,1], and you define $X=max(A_1,A_2)$X=max(A1,A2), $Y=min(A_1,A_2)$Y=min(A1,A2), then XX and YY will have exactly the joint distribution defined above.

a. Draw a graph of the region for which XX and YY have positive probability density. (3 points)

b. Derive the marginal probability density function of XX, $f_X(x)$fX(x). Make sure you write down a complete expression. (3 points)

c. Derive the unconditional expectation of XX. (3 points)

d. Derive the conditional probability density function of YY, conditional on XX, $f_{Y|X}(y|x)$fY|X(y|x) (3 points)

e. Derive the conditional expectation of YY, conditional on XX, $E(Y|X)$E(Y|X). (3 points)

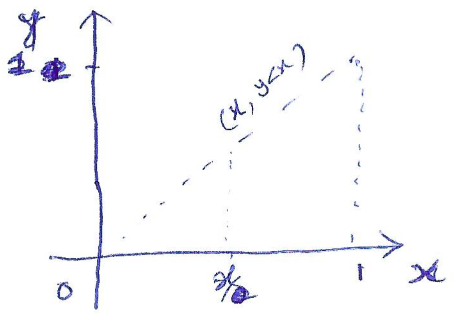
f. Derive $E(XY)$E(XY). Hint: if you take an expectation conditional on XX, XX is just a constant inside the expectation. This means that $E(XY|X)=XE(Y|X)$E(XY|X)=XE(Y|X). (3 points)

g. Using the previous parts, derive $cov(X,Y)$cov(X,Y) (3 points)

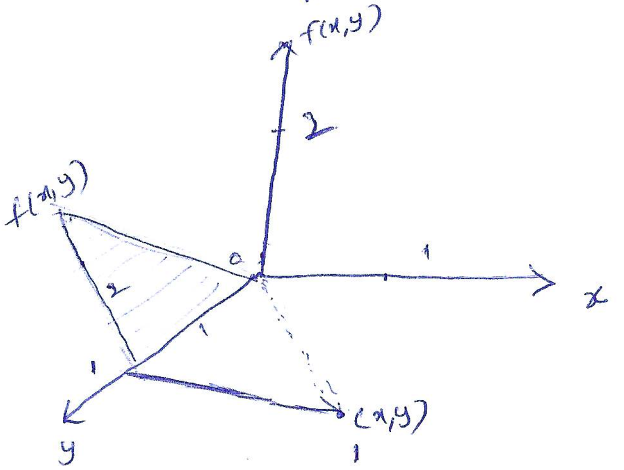
Given $f(x,y) = \begin{cases} 2 & 0<y<x<1 \\ 0 & \text{otherwise} \end{cases}$

(a). Graph for the probability density of X and Y :-

In 2-dimensional representation the graph of X and Y can be represented as follows :-



This can be represented in 3-d as below:



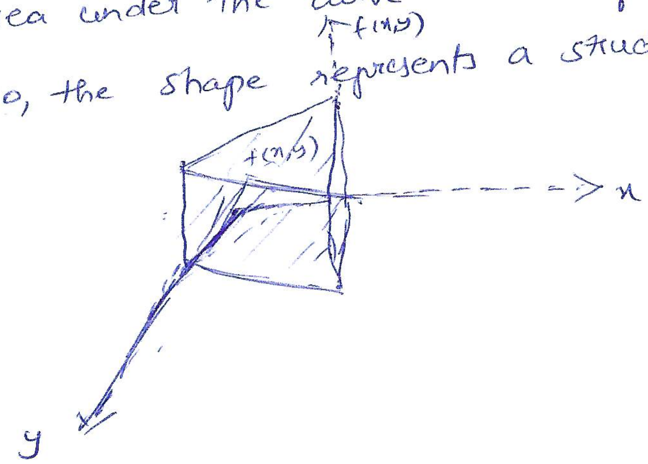
Area of triangle $= \frac{1}{2} \times 2 \times 1 = 1$

Assume the triangle is projected on the y axis is oneside of it's shape.

$f(x,y)$ is a surface at height 2 ($f(x,y)$) above the point (x,y) in 3-dimensional coordinate system

Then $P((x,y) \in A)$ is the volume underneath this surface and above the 2-d region A (x,y), analogous to the area under the curve in case of a single random variable.

So, the shape represents a structure like below:



Hence $P(X,Y \in A) = P(a \leq x \leq b, c \leq Y \leq d)$

$= P(0 \leq x < 1, 0 < y < x) = \int_0^1 \int_0^x f(x,y)\, dy\, dx$

(b). Derive Marginal probability density function of X $f_X(x)$

The definition of marginal probability density function of X ~~and~~ is denoted by $f_X(x)$ is given by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \quad \text{for } -\infty < x < \infty$$

Substituting $f(x,y) = 2$ and changing the limits to positive density, $(0 < y < x)$

$$f_X(x) = \int_0^x 2 \cdot dy$$

$$= 2 \cdot \int_0^x dy = 2y \Big|_0^x$$

Thus $\boxed{f_X(x) = 2x}$

(c). Derive unconditional expectation of X :-

The unconditional Expectation of X can be written as :-

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

Substituting limits with positive probability and $f(x) = f_X(x) = 2x$ from 2(b) solution :-

$$E(X) = \int_0^1 x \cdot 2x \cdot dx \quad \text{since } 0 < x < 1$$

$$= 2 \cdot \int_0^1 x^2\,dx = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}$$

Thus $\boxed{E(X) = \frac{2}{3}}$

(d). Derive conditional probability density function of Y, conditional of X, $f_{X/Y}(y/x)$

$$\boxed{f_{(Y/X)}(y/x) = \frac{f(x,y)}{f_X(x)}}$$

from problem (b) $f_X(x) = 2x$

from problem definition $f(x,y) = 2$

Thus, $f_{(Y/X)}(y/x) = \frac{2}{2x} = \frac{1}{x}$

Hence $\boxed{f_{(Y/x)}(Y/x) = \frac{1}{x}}$

(e). Derive conditional expectation of Y, conditional on X

$$\boxed{E(Y|X) = \int_{-\infty}^{\infty} y \cdot f(Y|X)\, dy}$$

According to definition of Expectation $E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$

we know from rules of conditional probability

$$E(Y/X) = \int_{-\infty}^{\infty} y \cdot f_{(Y/X)}(Y/X)\, dy$$

$$= \int_{0}^{x} y \cdot \frac{f(X,Y)}{f(X)}\, dy = \int_{0}^{x} y \cdot \frac{2}{2x}\, dy$$

$$= \frac{1}{x}\left[\frac{y^2}{2}\right]_0^x = \frac{x^2}{2x} = \frac{x}{2}$$

Thus $\boxed{E(Y/X) = \frac{x}{2}}$

2

(f). Derive $E(XY)$ where X is constant.

$E(XY)$ can be written as below; according to Law of iterated expectations; $E(XY) = E(E(XY/X))$ $\longrightarrow$ $\boxed{\begin{array}{l}\text{Rule:-}\\ E(Y) = E(E(Y/X))\end{array}}$

$\boxed{E(XY/X) = X \cdot E(Y/X)}$ $\longrightarrow$ (given)

thus, $E(XY) = E(E(XY/X)) = E(X \cdot E(Y/X))$

$$= E\left(X \cdot \frac{X}{2}\right) = E\left(\frac{X^2}{2}\right)$$

$$\boxed{\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx \\ E(h(X)) &= \int_{-\infty}^{\infty} h(x) \cdot f_x(x)\,dx \end{aligned}}$$ ← Expectation definitions

So, $E\left(\frac{X^2}{2}\right) = \int_{-\infty}^{\infty} \frac{x^2}{2} \cdot 2x \cdot dx$ $\left(\begin{array}{l}\text{since } f_X(x) = 2x \\ h(x) = x^2/2\end{array}\right)$

$$= \int_0^1 x^3 dx = \left.\frac{x^4}{4}\right|_0^1$$

$$= 1/4$$

Thus, $\boxed{E(XY) = 1/4}$ —— (f)

(g) Derive $\text{Cov}(X,Y)$:

$\boxed{\text{Cov}(X,Y) = E(XY) - E(X)E(Y)}$ → definition

$E(XY) = 1/4$ from problem (f)

$E(X) = 2/3$ from problem (c)

Marginal pdf $f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$ for $0 < x < 1$

$$= \int_0^1 2 \cdot dx \quad \text{for } 0<x<1 \quad \left(\text{since } f(x,y) = 2 \ \cancel{\text{and}}\right)$$

$$= \left.2x\right|_0^1 = 2$$

$\boxed{f_Y(y) = 2}$

$$E(Y) = \int_0^x y \cdot f_Y(y)\, dy \qquad \text{for } 0<y<x$$

Thus, $\boxed{f_Y(Y) = 2}$ which was calculated earlier can be substituted above

$$E(Y) = \int_0^x y \cdot 2\, dy$$

$$= 2 \cdot \frac{y^2}{2} \Big|_0^x = x$$

Thus $\boxed{E(Y) = x}$

we know $E(XY) = 1/4$ and $E(X) = 2/3$ $\quad E(Y) = x$

Thus $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$

$$= \frac{1}{4} - \frac{2}{3} \cdot x = \frac{3-8x}{12}$$

If ~~x→0~~, $x \to 0$, $\text{cov}(x,y) = \frac{1}{4}$

If $x \to 1$, $\text{cov}(x,y) = \frac{-7}{12}$

Thus for $0<x<1$,

$\text{Cov}(X,Y) = \frac{1}{4} < \text{cov}(x,y) < -\frac{7}{12}$

Problem 3:

Great Time to Watch Async

Suppose your waiting time in minutes for the Caltrain in the morning is uniformly distributed on [0, 5], whereas waiting time in the evening is uniformly distributed on [0, 10]. Each waiting time is independent of all other waiting times.

a. If you take the Caltrain each morning and each evening for 5 days in a row, what is your total expected waiting time? (3 points)

b. What is the variance of your total waiting time? (3 points)

c. What is the expected value of the difference between the total evening waiting time and the total morning waiting time over all 5 days? (3 points)

d. What is the variance of the difference between the total evening waiting time and the total morning waiting time over all 5 days? (3 points)

(a). Given the distributions are uniform

Let M_1, M_2, M_3, M_4, M_5 be the random variables that define wait times for each morning ride

Let E_1, E_2, E_3, E_4, E_5 be the random variables that define wait times for each evening ride

Since the morning wait times are uniformly distributed, the pdf is defined as $\frac{1}{5-0} = \frac{1}{5}$ for each of the variables $(M_1, M_2 \ldots M_5)$

Similarly pdf for all evening wait time variables is same and is defined as $= \frac{1}{10-0} = \frac{1}{10}$ for each variable $(E_1, E_2 \ldots E_5)$

The Expectation for each of these random variables in morning is $E(M_1) = E(M_2) = E(M_3) = E(M_4) = E(M_5) = \int_0^5 \frac{1}{5} \cdot M_i \, dM_i$

$= \frac{5+0}{2} = 2.5$

Similarly expectation for each of the evening rv is same and is $E(E_1) = E(E_2) = E(E_3) = E(E_4) = E(E_5) = \frac{10+0}{2} = 5$

(we can derive this from, $E(X) = \frac{a+b}{2}$ where X is a uniform distribution between a and b, $a \le X \le b$)

Thus,

$$E(M_1) = E(M_2) = E(M_3) = E(M_4) = E(M_5) = 2.5$$

$$E(E_1) = E(E_2) = E(E_3) = E(E_4) = E(E_5) = 5$$

From the rules of multivariate expressions,

Let T be a random variable defining the total wait time of travel for 5 days in a row, both morning and evening

Thus $T = M_1 + M_2 + M_3 + M_4 + M_5 + E_1 + E_2 + E_3 + E_4 + E_5$

$$E(T) = E(M_1 + \cdots M_5 + E_1 \cdots E_5)$$

According to Linearity property $E(M_1 + \cdots M_n) = E(M_1) + E(M_2) - E(M_n$

$$= E(M_1) + E(M_2) + \cdots + E(M_5) + E(E_1) + E(E_2) + \cdots E(E_5)$$

$$= 2.5(5) + 5(5)$$

$$= 37.5$$

$$E(T) = 37.5$$

(b). Variance of total waiting time

$$Var(T) = Var(M_1 + M_2 + M_3 \cdots M_5 + E_1 + E_2 \cdots E_5)$$

Since each of the defined variables are indepent there won't be any co-variance between them

Thus $Var(T) = Var(M_1) + Var(M_2) + \cdots Var(M_5) + Var(E_1) + Var(E_2) + \cdots Var(E_5)$

$$Var(M_1) = Var(M_2) = Var(M_3) = Var(M_4) = Var(M_5) = \frac{(b-a)^2}{12} = \frac{25}{12}$$

$$Var(E_1) = Var(E_2) = Var(E_3) = Var(E_4) = Var(E_5) = \frac{(10-0)^2}{12} = \frac{100}{12}$$

(since $Var(X) = \frac{(b-a)^2}{12}$ where X is uniform)

Thus $Var(T) = \frac{25}{12} \times 5 + \frac{100}{12} \times 5$

$$= \frac{5^3}{12} + \frac{5^2 \times \cancel{4}}{12}$$

$$= 10.41 + 41.66 = 52.07$$

(c). Let $Z = T_E - T_M$

where $T_E = E_1 + E_2 \dots E_5$ (sum (or) total evening wait time)

$T_M = M_1 + M_2 \dots M_5$ (total morning wait time)

$E(Z) = E(T_E - T_M) = E(T_E) - E(T_M)$

→ according to linearity property

$E(T_E) = 5 \times \cancel{25}\, 5 = \cancel{125}\ 25$ (these values were calculated in problems (a) and (b))

$E(T_M) = 5 \times 2.5 = 12.5$

$E(Z) = E(T_E - T_M) = E(T_E) - E(T_M) = 25 - 12.5 = 12.5$

$$\boxed{E(Z) = 12.5}$$

Thus expected value of difference between total evening ~~and~~ wait time and total morning wait time is

$E(Z) = 12.5$

(d). If we assume Z ~~T_E~~ is the rv,

where $Z = T_E - T_M$

$$Var(T_E - T_M) = Var(T_E) - var(T_M)$$

(from rules of linearity $Var(a+b) = var(a) + var(b)$, when a and b are independent and there won't be any covariance between them).

thus,

$$var(z) = var(T_E - T_M) = 5 \cdot \frac{(10-0)^2}{12} - 5 \cdot \frac{(5-0)^2}{12}$$

$$= 5\left[\frac{100}{12} - \frac{25}{12}\right] = 5\left[\frac{75}{12}\right] = 31.25$$

Thus variance of difference between the total evening waiting time and total morning wait time over all 5 days is

$$Var(z) = 31.25$$

Problem 4:

Maximizing Correlation

Show that if $Y=aX+b$Y=aX+b where XX and YY are random variables and $a\neq 0$a≠0, $corr(X,Y)=-1$corr(X,Y)=−1 or $+1$+1. (3 points)

$$\boxed{corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y}} \rightarrow \text{definition} \quad \text{— (a)}$$

$$E(XY) = E(X(aX+b)) = a \cdot E(X^2) + b\,E(X)$$
$$E(Y) = E(aX+b) = a \cdot E(X) + b$$

— according to summation properties of expectation

— (b)

$$Cov(X,Y) = E(XY) - E(X) \cdot E(Y)$$
$$= a \cdot E(X^2) + b \cdot E(X) - E(X) \cdot (a \cdot E(X) + b)$$
$$= a \cdot E(X^2) + b \cdot E(X) - a[E(X)]^2 - b \cdot E(X)$$
$$= a[E(X^2) - [E(X)]^2] = a \cdot \sigma_X^2 \quad \left(\text{since } Var(X) = E(X^2) - [E(X)]^2\right)$$

— (c)

$$Var(Y) = var(aX+b) = \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X$$

$\left(\text{since } \sigma^2_{aX+b} = a^2 \cdot \sigma^2_X \text{ where } b \text{ is a constant}\right)$

$$\Rightarrow \sigma_Y = \sigma_{aX+b} = |a| \cdot \sigma_X \quad \text{— (d)}$$

So, substituting (d), (c) in (a)

$$corr(X,Y) = \frac{a \cdot \sigma_X^2}{|a| \cdot \sigma_X \cdot \sigma_X} = \frac{a}{|a|} \cdot \frac{\sigma_X^2}{\sigma_X^2} = \pm 1 \text{ (based on 'a')}$$

$corr(X,Y) = \pm 1$ (where the value depends on value of 'a')

If a is -ve, $corr(X,Y) = -1$

If a is +ve, $corr(X,Y) = +1$

Thus $\boxed{corr(X,Y) = \pm 1}$