Probability & Statistics for EECS: Homework 10 # Solution

Problem 1

Show the proof of general LOTP (four cases).

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

Solution

1. X discrete, Y discrete

   By conditional probability,

$$P(X=x)=\sum_y P(X=x,Y=y)=\sum_y P(X=x|Y=y)P(Y=y)$$

2. X discrete, Y continuous

   By Bayes' Law of X discrete, Y continuous: $f_Y(y|X=x)=\frac{P(X=x|Y=y)f_Y(y)}{P(X=x)}$:

$$\begin{aligned}\int_{-\infty}^{\infty} P(X=x|Y=y)f_Y(y)dy &= \int_{-\infty}^{\infty} f_Y(y|X=x)P(X=x)dy \\ &= P(X=x)\int_{-\infty}^{\infty} f_Y(y|X=x)dy \\ &= P(X=x)\end{aligned}$$

3. X continuous, Y discrete

   By Bayes' Law of X continuous, Y discrete: $f_X(x|Y=y)=\frac{P(Y=y|X=x)f_X(x)}{P(Y=y)}$:

$$\begin{aligned}\sum_y f_X(x|Y=y)P(Y=y) &= \sum_y P(Y=y|X=x)f_X(x) \\ &= f_X(x)\sum_y P(Y=y|X=x) \\ &= f_X(x)\end{aligned}$$

   Alternatively,

$$\begin{aligned}f_X(x) &= \lim_{\varepsilon\to 0}\frac{P(X\in(x-\varepsilon,x+\varepsilon))}{2\varepsilon} \\ &= \lim_{\varepsilon\to 0}\sum_y \frac{P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)P(Y=y)}{2\varepsilon} \\ &= \sum_y \lim_{\varepsilon\to 0}\frac{P(X\in(x-\varepsilon,x+\varepsilon)|Y=y)P(Y=y)}{2\varepsilon} \\ &= \sum_y f_X(x|Y=y)P(Y=y)\end{aligned}$$

Problem 1 continued on next page...

4. X continuous, Y continuous

   By conditional probability,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy$$

Problem 2

The bus company from Blissville decides to start service in Blotchville, sensing a promising business opportunity. Meanwhile, Fred has moved back to Blotchville. Now when Fred arrives at the bus stop, either of two independent bus lines may come by (both of which take him home). The Blissville company's bus arrival times are exactly 15 minutes apart, whereas the time from one Blotchville company bus to the next is Expo($\frac{1}{15}$). Fred arrives at a uniformly random time on a certain day.

(a) What is the probability that the Blotchville company bus arrives first?

(b) What is the CDF of Fred's waiting time for a bus?

Solution

(a) Let $U \sim \text{Unif}(0, 15)$ be the arrival time of the next Blissville company bus, and $X \sim \text{Expo}\left(\frac{1}{15}\right)$ be the arrival time of the next Blotchville company bus (the latter is $X \sim \text{Expo}\left(\frac{1}{15}\right)$ by the memoryless property). Then

$$\begin{aligned} P(X < U) &= \int_0^{15} P(X < U \mid U = u) \frac{1}{15} du \\ &= \frac{1}{15} \int_0^{15} P(X < u \mid U = u) du \\ &= \frac{1}{15} \int_0^{15} \left(1 - e^{-u/15}\right) du \\ &= \frac{1}{e} \end{aligned}$$

(b) Let $T = \min(X, U)$ be the waiting time. Then

$$P(T > t) = P(X > t, U > t) = P(X > t)P(U > t)$$

So the CDF of T is

$$P(T \le t) = 1 - P(X > t)P(U > t) = 1 - e^{-\frac{t}{15}}(1 - \frac{t}{15})$$

for $0 < t < 15$ (and 0 for $t \le 0$, and 1 for $t \ge 15$).

Problem 3

A chicken lays a Pois(λ) number N of eggs. Each egg hatches a chick with probability p, independently. Let X be the number which hatch, and Y be the number which do NOT hatch.

(a) Find the joint PMF of N, X, Y. Are they independent?

(b) Find the joint PMF of N, X. Are they independent?

(c) Find the joint PMF of X, Y. Are they independent?

(d) Find the correlation between N and X. Your final answer should work out to a simple function of p and the λ should cancel out.

Solution

(a) For non-negative integer x, y, n, if $x + y \neq n$, $P(X = x, Y = y, N = n) = 0$.

If $x + y = n$, then we have:

$$\begin{aligned} P(N = n, X = x, Y = y) &= P(N = n)P(X = x, Y = y|N = n) \\ &= P(N = n)P(X = x|N = n) \\ &= \frac{\lambda^n}{n!}e^{-\lambda}\binom{n}{x}p^x(1-p)^{n-x} \end{aligned}$$

Since $N = X+Y$, so N, X, Y are not independent. (We could also see this by observing that for $x, y, n > 0$ such that $x+y \neq n$ we have that $P(X = x, Y = y, N = n) = 0$, while $P(X = x)P(Y = y)P(N = n) > 0$.)

(b) Since $N = X + Y$, for $n \geq i \geq 0$,

$$P(N = n, X = x) = P(N = n, X = x, Y = n - x) = \frac{\lambda^n}{n!}e^{-\lambda}\binom{n}{x}p^x(1-p)^{n-x}$$

Otherwise, $P(N = n, X = x) = 0$. By (c) we know that $X \sim Pois(\lambda p)$ and $N \sim Pois(\lambda)$, so $P(X = x)P(N = n) \neq P(X = x, N = n)$. Therefore, X and N are not independent. (We could also see this by observing that for $i > n$we have that $P(X = x, N = n) = 0$, while $P(X = x)P(N = n) > 0$.)

(c) Since $N = X + Y$, so:

$$\begin{aligned} P(X = x, Y = y) &= P(X = x, N = x + y) \\ &= \frac{\lambda^{x+y}}{(x+y)!}e^{-\lambda}\binom{x+y}{x}p^x(1-p)^y \\ &= \frac{(\lambda p)^x}{x!}e^{-\lambda p}\frac{(\lambda(1-p))^y}{y!}e^{-\lambda(1-p)} \\ &= P(X = x)P(Y = y). \quad x, y \geq 0. \end{aligned}$$

Therefore, X, and Y are independent. $X \sim Pois(\lambda p), Y \sim Pois(\lambda(1-p))$.

(d) By the chicken-egg story, X is independent of Y, with $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda q)$, for $q = 1 - p$. So

$$\text{Cov}(N, X) = \text{Cov}(X + Y, X) = \text{Cov}(X, X) + \text{Cov}(Y, X) = \text{Var}(X) = \lambda p,$$

giving

$$\text{Corr}(N, X) = \frac{\lambda p}{SD(N)SD(X)} = \frac{\lambda p}{\sqrt{\lambda \lambda p}} = \sqrt{p}$$

Problem 4

A scientist makes two measurements, considered to be independent standard Normal random variables. Find the correlation between the larger and smaller of the values.

Solution

Hint: Note that $\max(x,y)+\min(x,y)=x+y$ and $\max(x,y)-\min(x,y)=|x-y|$.
Solution: Let X and Y be i.i.d $\mathcal{N}(0,1)$ and $M=\max(X,Y), L=\min(X,Y)$. By the hint,

$$\begin{aligned} E(M)+E(L) = E(M+L) = E(X+Y) = E(X)+E(Y) = 0, \\ E(M)-E(L) = E(M-L) = E|X-Y| = \frac{2}{\sqrt{\pi}}, \end{aligned}$$

where the last equality was shown in Example 7.2.3 . So $E(M)=1/\sqrt{\pi}$, and

$$\operatorname{Cov}(M,L) = E(ML)-E(M)E(L) = E(XY)+(EM)^2 = (EM)^2 = \frac{1}{\pi},$$

since $ML=XY$ has mean $E(XY)=E(X)E(Y)=0$. To obtain the correlation, we also need $\operatorname{Var}(M)$ and $\operatorname{Var}(L)$. By symmetry of the Normal, $(-X,-Y)$ has the same distribution as (X,Y), so $\operatorname{Var}(M)=\operatorname{Var}(L)$; call this v. Then

$$\begin{aligned} E(X-Y)^2 = \operatorname{Var}(X-Y) = 2, \text{ and also} \\ E(X-Y)^2 = E(M-L)^2 = EM^2+EL^2-2E(ML) = 2v+\frac{2}{\pi}. \end{aligned}$$

So $v = 1-\frac{1}{\pi}$ (alternatively, we can get this by taking the variance of both sides of $\max(X,Y)+\min(X,Y) = X+Y$). Thus

$$\operatorname{Corr}(M,L) = \frac{\operatorname{Cov}(M,L)}{\sqrt{\operatorname{Var}(M)\operatorname{Var}(L)}} = \frac{1/\pi}{1-1/\pi} = \frac{1}{\pi-1}.$$

Problem 5

This problem explores a visual interpretation of covariance. Data are collected for $n \geq 2$ individuals, where for each individual two variables are measured (e.g., height and weight). Assume independence across individuals (e.g., person l's variables gives no information about the other people), but not within individuals (e.g., a person's height and weight may be correlated).

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be the n data points. The data are considered here as fixed, known numbers-they are the observed values after performing an experiment. Imagine plotting all the points (x_i, y_i) in the plane, and drawing the rectangle determined by each pair of points. For example, the points $(1,3)$ and $(4,6)$ determine the rectangle with vertices $(1,3), (1,6), (4,6), (4,3)$.

The signed area contributed by (x_i, y_i) and (x_j, y_j) is the area of the rectangle they determine if the slope of the line between them is positive, and is the negative of the area of the rectangle they determine if the slope of the line between them is negative. (Define the signed area to be 0 if $x_i = x_j$ or $y_i = y_j$, since then the rectangle is degenerate.) So the signed area is positive if a higher x value goes with a higher y value for the pair of points, and negative otherwise. Assume that the x_i are all distinct and the y_i are all distinct.

(a) The sample covariance of the data is defined to be

$$r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

are the sample means. (There are differing conventions about whether to divide by $n-1$ or n in the definition of sample covariance, but that need not concern us for this problem.)

Let (X, Y) be one of the (x_i, y_i) pairs, chosen uniformly at random. Determine precisely how $\operatorname{Cov}(X, Y)$ is related to the sample covariance.

(b) Let (X, Y) be as in (a), and $(\tilde{X}, \tilde{Y})$ be an independent draw from the same distribution. That is, (X, Y) and $(\bar{X}, \tilde{Y})$ are randomly chosen from the n points, independently (so it is possible for the same point to be chosen twice).

Express the total signed area of the rectangles as a constant times $E((X - \bar{X})(Y - \tilde{Y}))$. Then show that the sample covariance of the data is a constant times the total signed area of the rectangles.

Hint: Consider $E((X - \tilde{X})(Y - \tilde{Y}))$ in two ways: as the average signed area of the random rectangle formed by (X, Y) and $(\bar{X}, \bar{Y})$, and using properties of expectation to relate it to $\operatorname{Cov}(X, Y)$. For the former, consider the n^2 possibilities for which point (X, Y) is and which point $(\tilde{X}, \bar{Y})$; note that n such choices result in degenerate rectangles.

(c) Based on the interpretation from (b), give intuitive explanations of why for any r.v.s W_1, W_2, W_3 and constants a_1, a_2, covariance has the following properties:

(i) $\operatorname{Cov}(W_1, W_2) = \operatorname{Cov}(W_2, W_1)$;

(ii) $\operatorname{Cov}(a_1 W_1, a_2 W_2) = a_1 a_2 \operatorname{Cov}(W_1, W_2)$;

(iii) $\operatorname{Cov}(W_1 + a_1, W_2 + a_2) = \operatorname{Cov}(W_1, W_2)$;

(iv) $\operatorname{Cov}(W_1, W_2 + W_3) = \operatorname{Cov}(W_1, W_2) + \operatorname{Cov}(W_1, W_3)$.

Problem 5 continued on next page...

Solution

(a) After doing the experiment of choosing randomly among the n points, (X, Y) crystallizes into some (x_i, y_i), at which time $(X-\bar{x})(Y-\bar{y})$ crystallizes to $(x_i-\bar{x})(y_i-\bar{y})$. So

$$\operatorname{Cov}(X, Y)=E((X-\bar{x})(Y-\bar{y}))=\frac{1}{n} \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})=r.$$

(b) The total signed area is

$$A=\sum_{i<j}(x_i-x_j)(y_i-y_j)$$

Now compare this with the average signed area of the random rectangle formed by (X, Y) and $(\tilde{X}, \tilde{Y})$, which by linearity is

$$E((X-\tilde{X})(Y-\tilde{Y}))=E(XY)+E(\tilde{X}\tilde{Y})-E(X\tilde{Y})-E(\tilde{X}Y)$$

This simplifies to

$$2E(XY)-2E(X)E(Y)=2\operatorname{Cov}(X, Y)$$

since $E(\tilde{X}\tilde{Y})=E(XY)$ (because XY and $\tilde{X}\tilde{Y}$ have the same distribution), $E(X\tilde{Y})=E(X)E(\tilde{Y})=E(X)E(Y)$ (because X and $\tilde{Y}$ are independent, and hence uncorrelated), and $E(\tilde{X}Y)=E(X)E(Y)$. On the other hand, $E((X-\tilde{X})(Y-\tilde{Y}))$ is the arithmetic mean of n^2 values (consider all pairs (i, j) where i is the index for which point (X, Y) is and j is the index for which point $(\tilde{X}, \tilde{Y})$ is), consisting of n 0's (for the degenerate rectangles formed when $(\tilde{X}, \tilde{Y})=(X, Y)$) and the $\binom{n}{2}$ signed rectangle areas, listed twice each. So

$$E((X-\tilde{X})(Y-\tilde{Y}))=\frac{n \cdot 0+2 \sum_{i<j}(x_i-x_j)(y_i-y_j)}{n^2}=\frac{2}{n^2} A$$

Thus,

$$\operatorname{Cov}(X, Y)=\frac{A}{n^2}$$

(c) Fundamental properties of covariance follow readily from the above interpretation:
(i) Reversing which axis is which has no effect on the areas of the rectangles.
(ii) Stretching (or shrinking) a rectangle along one axis changes the area by the same factor. For example, if we double all the widths and triple all the lengths of the rectangles, then the areas all increase by a factor of 6.
(iii) Shifting a rectangle horizontally or vertically has no effect on its area.
(iv) A rectangle with dimensions a by $b+c$ can be split into two rectangles, one a by b and the other a by c, and the area of the original rectangle equals the sum of the areas of the smaller rectangles. Here we are dealing with signed areas, but in this interpretation a positive-area rectangle splits into two positive-area rectangles and a negative-area rectangle splits into two negative-area rectangles.