Statistics for Data Science

Unit 5 Homework: Joint Distributions

Solutions

1. Unladen SwallowsIn 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 $V = 0.5 \cdot W + U$, where U is a random variable (which we might call error). We assume that U has a standard normal distribution and is independent of W.

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

Given:

- W has a normal distribution with $\mu_W = 10$ and $\sigma_W = 4$.
- $V = 0.5\ W + U$, where U is a random variable, so $cov(W, V) = (W, 0.5W+U)$.
- U has a standard normal distribution so $\mu_U = 0$ and $\sigma_U = 1$.
- U is independent of W (i.e. $cov(W, V) = 0$).

To derive each element of the variance-covariance matrix for W and V, we need to find (a) variance of W, (b) variance of V, and covariance of W and V.

(a) Find variance of W.
Since $\sigma_W = 4$, $var(W) = \sigma_W^2 = 16$.

(b) Find variance of V.
By definition $var(V) = var(0.5W+U)$.

Recall $var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$ and $var(aX) = a^2 var(x)$.

So $var(V) = var(0.5W+U)$

$= var(0.5W)+var(U)+2cov(0.5W,U)$.

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

Since U and W are independent, $cov(W,U) = 0$.

$= (0.5)^2 var(W)+1+2cov(0.5W,U)$

$= 0.25 \cdot 16+1$

$= 5$

(c) Find cov (W, V).

Let $X = W$ and $Y = 0.5W+U$.

Recall $cov(X,Y) = E(XY) - E(X)E(Y)$.

So cov(W,V) = cov(W, 0.5W+U)

<u>Approach # 1:</u>
cov(W, 0.5W + U) = cov(W, 0.5W) + cov(W, U) = 0.5 var(W) + 0.

<u>Approach # 2:</u>
cov(W, 0.5W + U) = cov(W, 0.5W) + cov(W, U) = 0.5 var(W) + 0.

$$= E(0.5W^2 + WU) - E(W)E(0.5W + U)$$
$$= 0.5E(W^2) + E(W)E(U) - E(W)[0.5E(W) + E(U)]$$
$$= 0.5E(W^2) \cancel{+ E(W)E(U)} - 0.5E(W)^2 \cancel{- E(W)E(U)}$$
$$= 0.5[E(W^2) - E(W)^2]$$
$$= 0.5(Var(W)) \qquad \textit{(by definition, } var(W) = E(W^2) - E(W)^2\textit{)}$$
$$= 0.5 \cdot 16$$
$$= 8$$

Using (a), (b), and (c), we get

$$\text{var(W,V)} = \begin{bmatrix} var(W) & cov(W,V) \\ cov(V,W) & var(V) \end{bmatrix} = \begin{bmatrix} 16 & 8 \\ 8 & 5 \end{bmatrix}$$

2. Broken Rulers

You have a ruler of length 1 and you choose a place to break it using a uniform probability distribution. Let random variable, X, represent the length of the left piece of the ruler. X is distributed uniformly in [0,1]. You take the left piece of the ruler and once again choose a place to break it using a uniform probability distribution. Let random variable Y be the length of the left piece from the second break.

(a) Find the conditional expectation of Y given X, E(Y|X).

(b) Find the unconditional expectation of Y. One way to do this is to apply the law of iterated expectations, which states that E(Y) = E(E(Y|X)). The inner expectation is the conditional expectation computed above, which is a function of X. The outer expectation finds the expected value of this function.

(c) Compute E(XY). Hint: if you take an expectation conditional on a value of X, X is just a constant inside the expectation. This means that E(XY|X) = X•E(Y|X)

(d) Using the previous results, compute cov(X, Y).

Let X and Y be random variables, such that

X = the length of the left piece of the ruler, where X is distributed uniformly in [0,1], and

Y = the length of the left piece from the second break.

(a) Find E(Y|X).

Since Y has uniform distribution, E(Y) = X/2. Hence, E(Y |X) = X/2.

Note: The hardest this about this question is probably just understanding that the problem is telling you what the conditional distribution of Y is, conditional on X. Notice that the second break is chosen using a uniform distribution. That distribution must be over the interval from 0 up to X, the length of the left piece from the first break. This reminds us that the first break has already occurred, effectively fixing X, which may now be treated as a constant. This effectively means that we've conditioned on X. So the problem tells us that $f_{Y|X}$ is uniform on [0, 1].

Since the expectation of a uniform distribution is the midpoint of its interval, E(Y|X) = X/2.

(b) Find E(Y).

By the law of iterated expectations, E(Y) = E(E(Y |X)).

From (a), we know E(Y|X) = X/2, so E(Y) = E(E(Y|X)) = $E\left(\frac{X}{2}\right)$

Since X is uniformly distributed, E(Y) = $E\left(\frac{X}{2}\right) = \frac{1}{2} \cdot E(X)$ = ½•½ = ¼

Note: Again, the expectation of a uniform random variable is the midpoint of its interval.

(c) Find (XY).

Using the hint, we can rewrite the law of iterated expectations:

E(XY) = E(E(XY|X)) = E(X•E(Y|X)) = $E\left(\frac{X^2}{2}\right) = \frac{1}{2} \cdot E(X^2)$.

By definition of uniform distribution, $E(X^2) = \int_0^1 \frac{1}{1-0} x^2 dx = \frac{1}{3}x^3 \Big|_0^1 = \frac{1}{3}$

Thus, E(XY) = $\frac{1}{2} E(X^2) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$

Note: From the hint, we have E(E(XY|X)) = E(X•E(Y|X)). At this point, we can pull the X out from inside the inner expectation, since we've conditioned on X, so we're just treating X as a constant. (To see this another way, notice that the inner expectation is integrated over Y, so X can be pulled out in front of the integral). Hence, we get E(XY) = E(E(XY|X)) = E(X•E(Y|X)) = $\mathrm{E}\left[\mathrm{X}\left(\frac{\mathrm{X}}{2}\right)\right] = \frac{1}{2} \cdot \mathrm{E}(\mathrm{X}^2)$.

(d) Find (XY).

Using our answers from (a), (b), and (c), we get:

cov(X, Y) = E(XY) – E(X)E(Y) = $\frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{24}$

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?

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

(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?

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

(a) Let X_i and Y_i be random variables, such that for i = 1, …, 5,
X_i = number of minutes spent waiting for Caltrain in the morning on the ith day, and
Y_i = number of minutes spent waiting for Caltrain in the evening on the ith day.

We know that X_i is uniformly distributed on [0,5], Y_i is uniformly distributed on [0,10], and each waiting time is independent of all other waiting times (i.e. X_i is independent of Y_i and X_i is independent of X_j for any i, j = 1,…,5 s.t. i ≠ j).

Since all the X_i's and Yi's have identical uniform probability distributions, the total expected waiting time for 5 days in a row is

$$\sum_{i=1}^{5} E(Xi + Yi) = \sum_{i=1}^{5} [E(Xi) + E(Yi)]$$

$$= 5 \cdot [E(Xi) + E(Yi)]$$

$$= 5 \cdot \left[\int_0^5 x_i f(x_i) dx + \int_0^5 y_i f(y_i) dy\right]$$

$$= 5 \cdot \left[\frac{1}{5}\frac{x^2}{2}\Big|_0^5 + \frac{1}{10}\frac{y^2}{2}\Big|_0^{10}\right]$$

$$= 5 \cdot \left[\frac{1}{5}\frac{25}{2} + \frac{1}{10}\frac{100}{2}\right]$$

$$= 5 \cdot \left[\frac{5}{2} + \frac{10}{2}\right]$$

$$= 37.5 \; minutes$$

(b) $\sum_{i=1}^{5} Var(Xi + Yi) = \sum_{i=1}^{5} [Var(Xi) + Var(Yi) + 2cov(Xi, Yi)]$

$= \sum_{i=1}^{5} [Var(Xi) + Var(Yi)]$ (Note: X and Y are independent)

$= 5 \cdot \left[\frac{5^2}{12} + \frac{10^2}{12}\right]$

$= \frac{625}{12}$

≈ 52.08

(c) $\sum_{i=1}^{5} E(Xi - Yi) = \sum_{i=1}^{5} [E(Yi) - E(Xi)]$

From (a) we know $E(Xi) = \frac{5}{2}$ and $E(Yi) = \frac{10}{2}$.

Hence, $\sum_{i=1}^{5} E(Xi - Yi) = \sum_{i=1}^{5} \left[\frac{10}{2} - \frac{5}{2}\right] = 12.5 \; minutes$

(d) $\sum_{i=1}^{5} Var(Yi - Xi) = \sum_{i=1}^{5} [Var(Xi) + Var(Yi) - 2cov(Xi, Yi)]$

$= \sum_{i=1}^{5}[Var(Xi) + Var(Yi)]$ (Note: X and Y are independent)

$= 5 \cdot \left[\frac{5^2}{12} + \frac{10^2}{12}\right]$

$= \frac{625}{12}$

≈ 52.08

Q4

$Corr(X,Y) = Corr(X, aX+b) = \frac{Cov(X,aX+b)}{\sqrt{Var(X)Var(aX+b)}}$

The numerator of this fraction can be simplified as follows.

$Cov(X, aX+b) = Cov(X, aX) + Cov(X, b) = aCov(X, X) + 0 = aVar(X)$

There are other ways we could have done that. Some students may find it easier to first write down the covariance in terms of expectation. That leads to the following simplification.

$Cov(X, aX+b) = E[X(aX+b)] - E[X]E[aX+b] = aE[X^2] + bE[X] - aE[X]^2 - bE[X]$

$= a\big(E[X^2] - E[X]^2\big) = a\ Var(X)$

We know that $var(aX+b) = a^2 var(X)$ so the denominator can be written as

$\sqrt{Var(X)Var(aX+b)} = \sqrt{a^2 Var(X)Var(X)} = |a|\ Var(X)$

So, we have

$$Corr(X, aX+b) = \frac{a\ Var(x)}{|a|\ Var(x)} = \frac{a}{|a|}$$

Which means:

$$Corr(X, aX+b) = \begin{cases} -1 & a < 0 \\ \text{not defined} & a = 0 \\ +1 & a > 0 \end{cases}$$

Q5

i- $f_{N|M}(n|m) = 1/(m+1) \quad \forall 0 \le n \le m$

ii- $f_N(n) = \sum_{m=n} f_{N|M}(n|m) f_M(m)$

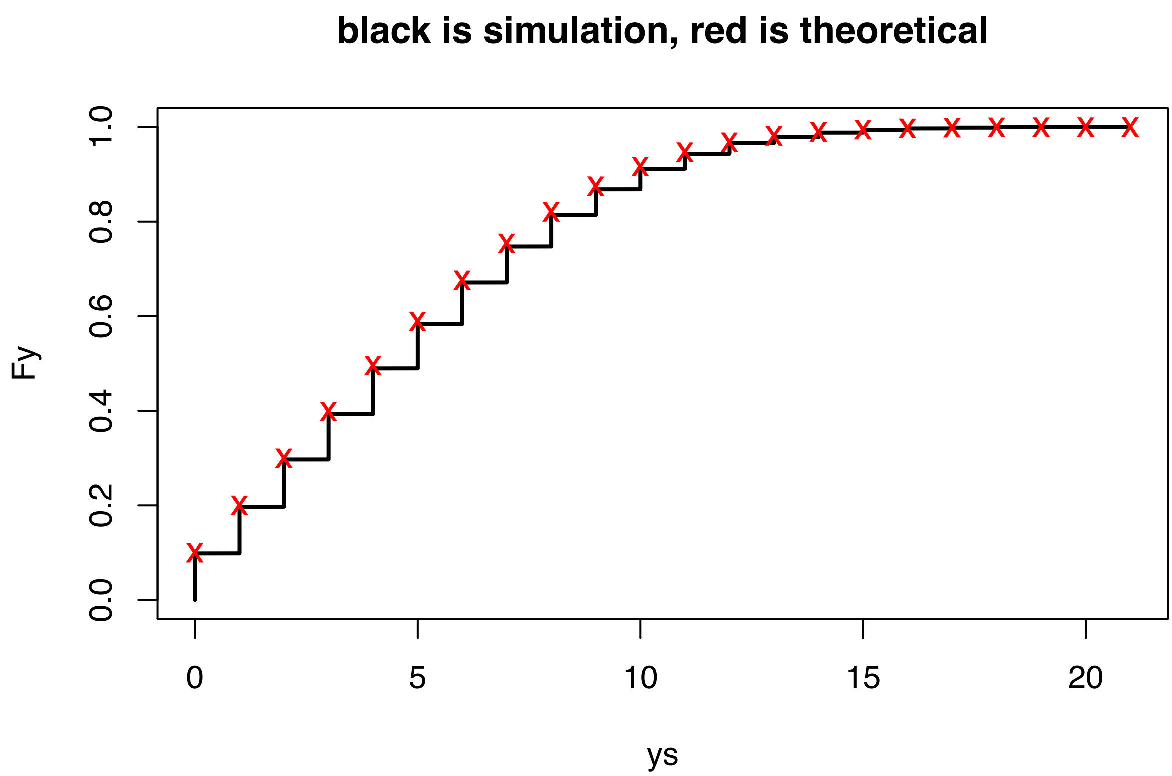
$= \sum_{m=n} \frac{1}{\alpha} f_M(m+1)$

$= \frac{1}{\alpha}(1 - F_M(n))$

Simulation

```
alpha = 10
N = 1e4
x = rpois(N, lambda =alpha)
y = sapply(x+1, sample, size= 1)-1
ys = sort(y)
Fy = (1:N)/N  # simulation
basey = 0:max(y) #basis for theoretical
th.p = 1/alpha * (1-ppois(basey,lambda = alpha) ) # density
th.c = cumsum(th.p)  # cumulative (theoretical)

plot(Fy~ys, ty='l',lwd=2, main='black is simulation, red is theoretical')
points(basey,th.c,col='red',pch='x',cex=1.2)
```



black is simulation, red is theoretical
Fy
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
ys