Statistics 110: Introduction to Probability | Section 7 Solutions, Fall 2018

Multivariate Distributions, Covariance & Correlation

TF: Justin Zhu (justinzhu@college.harvard.edu)

Practice Problems

Example 1. Housing Day.
Suppose Harvard College is conducting its housing lottery. For simplicity's sake, we'll say that there are 1200 Freshmen that will be randomly assigned to 12 houses. Let $X_1, X_2, \ldots, X_{12}$ count how many students are place in Dunster (X_1), all the way to Pfoho (X_{12}) (organized by best house to worst).

(a) Are X_1 and X_2 independent?

(b) What is the joint distribution of $X_1, X_2, \ldots, X_{12}$?

(c) What is the marginal distribution of X_1, the number of students who are placed into Dunster House, and the joint distribution of X_1 and $1200 - X_1$?

(d) What is the conditional distribution of X_1 given $X_{10} + X_{11} + X_{12} = 450$?

Solution

(a) No they are not. Since the number of Freshmen is constrained to 1200, knowing that a lot of people got into one house decreases the number of people that could be in the remaining houses.

(b) By the story of the Multinomial distribution,

$$(X_1, X_2, \ldots, X_{12}) \sim \text{Mult}_{12}(1200, (1/12, \ldots, 1/12))$$

(c) In this case, we can group together bins that are not in Dunster House together. We have

$$X_1 \sim \text{Bin}(1200, 1/12)$$

$$(X_1, 1200 - X_1) \sim \text{Mult}_2(1200, (1/12, 11/12))$$

(d)

$$X_1 | X_{10} + X_{11} + X_{12} = 450 \sim \text{Bin}(750, 1/9)$$

Example 2. Subsets of MVNs.
Let (X_1, X_2) be BVN. Marginally, suppose that X_1 and X_2 are $\mathcal{N}(0, 1)$, and $\text{Corr}(X_1, X_2) = \rho$.

(a) Find the distribution of $X_1 - 3X_2$.

(b) Find c such that $X_1 - 3X_2 \perp\!\!\!\perp X_1 + cX_2$.

Solution

(a) Because $X_1 - 3X_2$ is a linear combination of a BVN, we know that $X_1 - 3X_2$ is Normally distributed. We also know that

$$E(X_1 - 3X_2) = 0$$
$$\text{Var}(X_1 - 3X_2) = \text{Var}(X_1) + \text{Var}(3X_2) - 2\text{Cov}(X_1, 3X_2) = 10 - 6\rho$$

So it follows that $X_1 - 3X_2 \sim \mathcal{N}(0, 10 - 6\rho)$

(b) We know that $(X_1 - 3X_2, X_1 + cX_2)$ is BVN, and within MVN, uncorrelated implies independent. So we just need to find c in order for $\text{Cov}(X_1 - 3X_2, X_1 + cX_2) = 0$.

$$\text{Cov}(X_1 - 3X_2, X_1 + cX_2) = \text{Var}(X_1) - 3\text{Cov}(X_1, X_2) + c\text{Cov}(X_1, X_2) - 3c\text{Var}(X_2) = 1 - 3\rho + c\rho - 3c$$

Setting the above equal to 0 and solving for c, we get that $c = \frac{1-3\rho}{3-\rho}$.

Example 3. Jellybeans.
I have a jar of 30 jellybeans: 10 red, 8 green, 12 blue. I draw a sample of 12 jellybeans without replacement. Let X be the number of red jellybeans in the sample, Y the number of green jellybeans. Find $\text{Cov}(X, Y)$.

Solution
Let $X = I_1 + \ldots + I_{12}$, and $Y = J_1 + \ldots + J_{12}$, where

$$I_i = \begin{cases} 1 & \text{if } i\text{th jellybean in sample is red} \\ 0 & \text{otherwise} \end{cases}$$

$$J_i = \begin{cases} 1 & \text{if } i\text{th jellybean in sample is green} \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned}
\text{Cov}(I_1, J_1) &= E(I_1 J_1) - E(I_1)E(J_1) \\
&= 0 - \left(\frac{10}{30}\right)\left(\frac{8}{30}\right) \\
\text{Cov}(I_1, J_2) &= E(I_1 J_2) - E(I_1)E(J_2) \\
&= \left(\frac{10}{30}\right)\left(\frac{8}{29}\right) - \left(\frac{10}{30}\right)\left(\frac{8}{30}\right) \\
\text{Cov}(X, Y) &= \sum_{i=1}^{12} \text{Cov}(I_i, J_i) + 2\sum_{i<j} \text{Cov}(I_i, J_j) \\
&= \sum_{i=1}^{12} \text{Cov}(I_1, J_1) + 2\binom{12}{2}\text{Cov}(I_1, J_2) \\
&= 12 \cdot \text{Cov}(I_1, J_1) + 12 \cdot 11 \cdot \text{Cov}(I_1, J_2) \\
&= -\frac{96}{145}
\end{aligned}$$

It's good to do a little sanity check at the end: it makes sense that the covariance is negative. If the sample contains a lot of red jellybeans, the sample probably has fewer green jellybeans.
Another way to solve this is to create an indicator for each red jellybean and each green jellybean in the jar, where the indicator equals 1 if the jellybean is in the sample and 0 otherwise.

Example 4. Stat Courses.
Let X be the number of statistics majors in a certain college in the class of 2030, viewed as an r.v. Each statistics major chooses between two tracks: a general track in statistical principles, and a track in quant finance. Suppose that each statistics major chooses randomly which of these two tracks to follow, independently, with probability p of choosing the general track. Let Y be the number of statistics majors who choose the general track, and Z be the number of statistics majors who choose the quantitative finance track.

(a) Suppose that $X \sim \text{Pois}(\lambda)$. Find the correlation between X and Y.

(b) Let n be the size of the Class of 2030, where n is a known constant. For this part and the next, instead of assuming that X is Poisson, assume that each of the n students chooses to be a statistics major with

probability r, independently. Find the joint distributions of Y, Z, and the number of non-statistics majors, and their marginal distributions.

(c) Continuing as in the previous part, find the correlation between X and Y.

Solution

(a) By the chicken-egg story, we know that Y and Z are independent Poisson random variables, with rate parameters λp and λq, respectively. We must first find the covariance between X and Y.

$$Cov(X, Y) = Cov(Y + Z, Y) = Var(Y) + Cov(Y, Z) = \lambda p$$

We now plug this into the equation for correlation:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\lambda p}{\sqrt{\lambda\lambda p}} = \sqrt{p}$$

(b) Under this new model, we have that $X \sim \text{Bin}(n, r)$. By the multiplication rule, we have that the probability of a student becoming a general Statistician is rp, a Goldman-Sachs Statistician is rq, and a non-Statstician (lame) is $1 - r$. Therefore, we can apply the story of the Multinomial here:

$$(Y, Z, n - X) \sim Mult_3(n, (rp, rq, 1 - r))$$

(c) We use the fact that covariance of the marginal distributions in a multinomial is given by $-np_ip_j$.

$$Cov(X, Y) = Cov(Y + Z, Y) = Var(Y) + Cov(Z, Y) = nrp(1 - rp) - n(rp)(rq) = npr(1 - r)$$