Statistics 110: Introduction to Probability Section 7, Fall 2018

Multivariate Distributions, Covariance & Correlation

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

Multinomial (Multivariate Discrete)

Review - Binomial is a simple case of multinomial.

Let us say that the vector $\vec{\mathbf{X}} = (X_1, X_2, X_3, \ldots, X_k) \sim \text{Mult}_k(n, \vec{p})$ where $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Story - We have n items, and then can fall into any one of the k buckets independently with the probabilities $\vec{p} = (p_1, p_2, \ldots, p_k)$.

Example - Let us assume that every year, 100 students in the Harry Potter Universe are randomly and independently sorted into one of four houses with equal probability. The number of people in each of the houses is distributed $\text{Mult}_4(100, \vec{p})$, where $\vec{p} = (.25, .25, .25, .25)$. Note that $X_1 + X_2 + \cdots + X_4 = 100$, and they are dependent.

Multinomial Coefficient The number of permutations of n objects where you have $n_1, n_2, n_3 \ldots, n_k$ of each of the different variants is the **multinomial coefficient**.

$$\binom{n}{n_1 n_2 \ldots n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}$$

Joint PMF - For $n = n_1 + n_2 + \cdots + n_k$

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = x_k) = \binom{n}{n_1 n_2 \ldots n_k} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Lumping - If you lump together multiple categories in a multinomial, then it is still multinomial. A multinomial with two dimensions (success, failure) is a binomial distribution.

Marginal PMF and Lumping

$$X_i \sim \text{Bin}(n, p_i)$$

$$X_i + X_j \sim \text{Bin}(n, p_i + p_j)$$

$$X_1, X_2, X_3 \sim \text{Mult}_3(n, (p_1, p_2, p_3)) \Longrightarrow X_1, X_2 + X_3 \sim \text{Mult}_2(n, (p_1, p_2 + p_3))$$

$$X_1, X_2, \ldots, X_{k-1} | X_k = n_k \sim \text{Mult}_{k-1}\left(n - n_k, \left(\frac{p_1}{1 - p_k}, \frac{p_2}{1 - p_k}, \ldots, \frac{p_{k-1}}{1 - p_k}\right)\right)$$

Multivariate Uniform (Multivariate Continuous)

If $\mathbf{X} = (X_1, X_2, \ldots X_n)$ where $X_i \sim \text{Unif}(a_i, b_i)$, then X is distributed multivariate uniform and can be thought of pictorally as a box in n dimensions, where any point within the box is equally likely to be selected.
The PDF of any point inside the box (the support of $\mathbf{X}$) is equal to 1 over the volume of the box, so for all $\mathbf{x} \in \mathbb{R}^n$ in the box, we have:

$$f_{\mathbf{X}}(x) = \frac{1}{(b_1 - a_1)(b_2 - a_2) \ldots (b_n - a_n)}$$

An analogous result is true for the discrete case. The PMF at each point is simply 1 divided by the total number of points.

Multivariate Normal (MVN)

A vector $\vec{X} = (X_1, X_2, X_3, \ldots, X_k)$ is declared Multivariate Normal if any linear combination is normally distributed (e.g. $t_1X_1 + t_2X_2 + \cdots + t_kX_k$ is Normal for any constants $t_1, t_2, \ldots, t_k$). The parameters of the Multivariate normal are the mean vector $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$ and the covariance matrix Σ where the $(i,j)^{\text{th}}$ entry is $\text{Cov}(X_i, X_j)$. For any MVN distribution: 1) Any sub-vector is also MVN. 2) If any two elements of a multivariate normal distribution are uncorrelated, then they are independent. Note that this does not apply to most random variables.

Covariance and Correlation

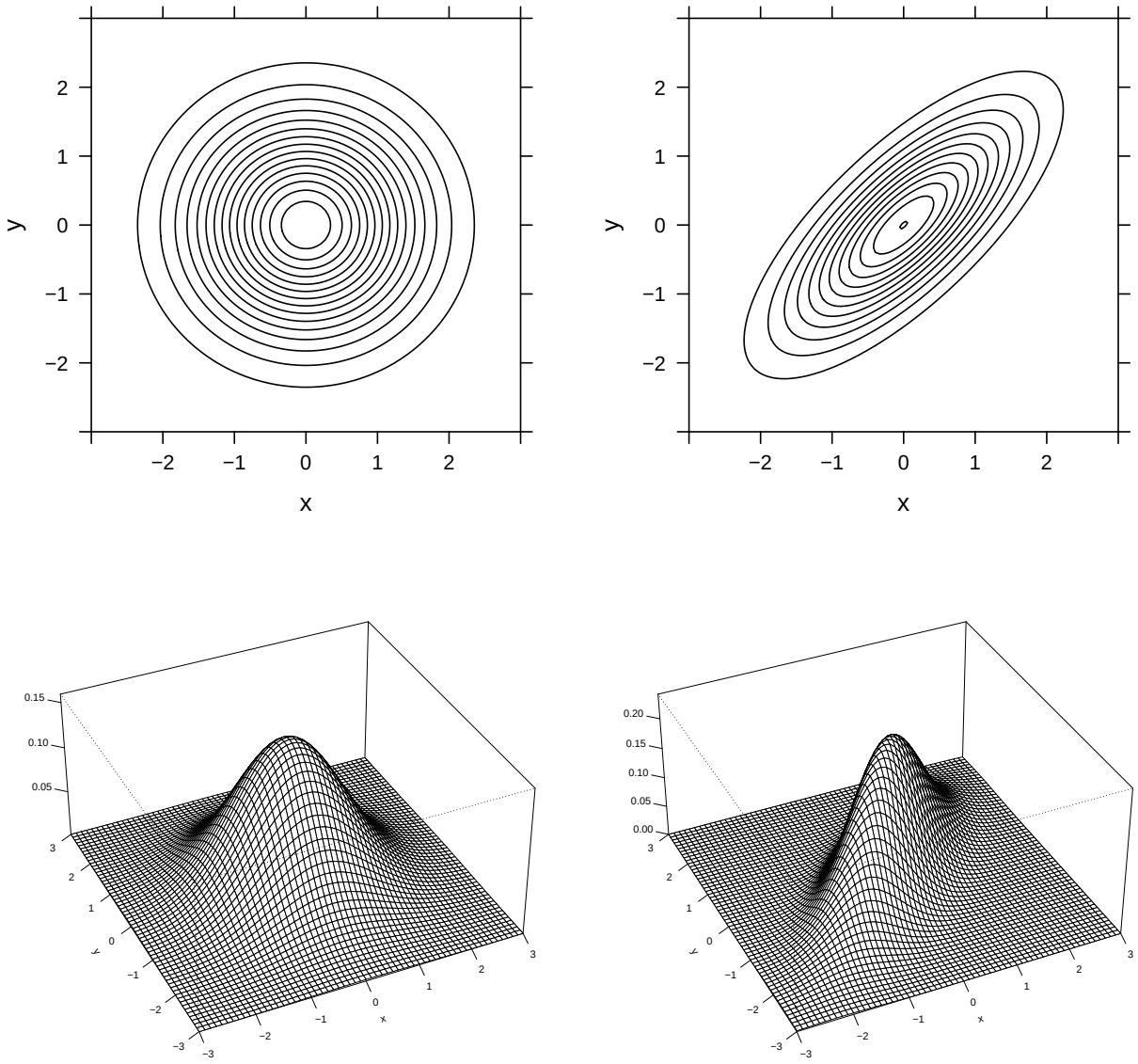

Figure 1: (left) We have two uncorrelated distributions that are marginally identical. (right) We have two positively correlated distributions that are marginally identical. If we know that one of them is high relative to the mean, then we know that the the other one is likely to be high relative to the mean too. Plots courtesy of Jessy Hwang.

Covariance and Correlation (cont'd)

Covariance is the two-random-variable equivalent of Variance, defined by the following:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Note that

$$\text{Cov}(X, X) = E(XX) - E(X)E(X) = E(X^2) - [E(X)]^2 = \text{Var}(X)$$

Correlation is a rescaled variant of Covariance that is always between -1 and 1.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Covariance and Indepedence - If two random variables are independent, then they are uncorrelated. The inverse is not necessarily true.

$$X \perp\!\!\!\perp Y \longrightarrow \text{Cov}(X, Y) = 0$$

Covariance and Variance - Note that

$$\begin{aligned} \text{Cov}(X, X) &= \text{Var}(X) \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \\ \text{Var}(X_1 + X_2 + \cdots + X_n) &= \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j} \text{Cov}(X_i, X_j) \end{aligned}$$

In particular, if X and Y are independent then they have covariance 0 thus

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

In particular, If $X_1, X_2, \ldots, X_n$ are i.i.d. and all of them have the same covariance relationship, then

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = n\text{Var}(X_1) + 2\binom{n}{2}\text{Cov}(X_1, X_2)$$

Covariance and Linearity - For random variables W, X, Y, Z and constants b, c:

$$\begin{aligned} \text{Cov}(X + b, Y + c) &= \text{Cov}(X, Y) \\ \text{Cov}(2X, 3Y) &= 6\text{Cov}(X, Y) \\ \text{Cov}(W + X, Y + Z) &= \text{Cov}(W, Y) + \text{Cov}(W, Z) + \text{Cov}(X, Y) + \text{Cov}(X, Z) \end{aligned}$$

Covariance and Invariance - Correlation, Covariance, and Variance are addition-invariant, which means that adding a constant to the term(s) does not change the value. Let b and c be constants.

$$\begin{aligned} \text{Var}(X + c) &= \text{Var}(X) \\ \text{Cov}(X + b, Y + c) &= \text{Cov}(X, Y) \\ \text{Corr}(X + b, Y + c) &= \text{Corr}(X, Y) \end{aligned}$$

In addition to addition-invariance, Correlation is *scale-invariant*, which means that multiplying the terms by any constant does not affect the value. Covariance and Variance are not scale-invariant.

$$\text{Corr}(2X, 3Y) = \frac{\text{Cov}(2X, 3Y)}{\sqrt{\text{Var}(2X)\text{Var}(3Y)}} = \frac{6\text{Cov}(X, Y)}{\sqrt{36\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \text{Corr}(X, Y)$$

Intuitive Explanation of Covariance - See
`https://www.quora.com/Probability/What-is-an-intuitive-explanation-of-covariance`

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 $X_{10} + X_{11} + X_{12} = 450$?

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 $X_1 - 3X_2 \perp\!\!\!\perp X_1 + cX_2$.

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 $\text{Cov}(X, Y)$.

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.