Wednesday, November 8 **Fall 2023**

Section 8: Multinomial and Multivariate Normal

Srihari Ganesh

Based on section note formatting template by Rachel Li and Ginnie Ma '23

1 Summary

1.1 Multinomial

We first generalize the notion of Bernoulli trials to many categories; this vocabulary for "categorical trials" is not standard/necessary for the class, just introduced for to help define the Multinomial.

Definition 1 (Categorical trials)**.** Consider **categorical trials**, where the outcome of a trial falls into one of k categories (e.g., the roll of a die has 6 categories, the flip of a coin has 2, etc.). Let $\mathbf{p} \in \mathbb{R}^k$ be a probability vector (where each entry is in $[0, 1]$ and the entries add up to p), where p_i is the probability that the outcome falls into the i^{th} category.

Story 2 (Multinomial)**.** Suppose we run n independent and identically distributed (i.i.d.) categorical trials with k categories and probability vector $\mathbf{p}$. Let $\mathbf{X}$ (a k-dimensional random vector) count the number of trials that fell into each category. Then $\mathbf{X}$ is distributed **Multinomial**: $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$.

Result 3 (Marginal)**.** For $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, $X_j \sim \text{Bin}(n, p_j)$.

Result 4 (Conditioning)**.** For $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$,

$$(X_2, \dots, X_n) | X_1 = x_1 \sim \text{Mult}_{k-1}(n - x_1, \left(\frac{p_2}{1 - p_1}, \dots, \frac{p_n}{1 - p_1}\right)).$$

Result 5 (Lumping)**.** Suppose $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$. Then we can group (**lump**) categories in any way to get a new Multinomial random variable by adding up the associated probabilities. For example, if $(X_1, X_2, X_3, X_4, X_5) \sim \text{Mult}_5\left(n, (p_1, p_2, p_3, p_4, p_5)\right)$, then some valid examples are

$$(X_1 + X_4, X_2, X_3 + X_5) \sim \text{Mult}_3\left(n, (p_1 + p_4, p_2, p_3 + p_5)\right),$$
$$(X_1 + X_2, X_3, X_4, X_5) \sim \text{Mult}_4\left(n, (p_1 + p_2, p_3, p_4, p_5)\right).$$

Result 6 (Covariance)**.** For $\mathbf{X} \sim \text{Mult}_k(n, \mathbf{p})$, $\text{Cov}(X_i, X_j) = -np_ip_j$.

Result 7 (Chicken-Egg extension, Joe might've called this Fish-Egg)**.** Suppose $N \sim \text{Pois}(\lambda)$ and $\mathbf{X}|N = n \sim \text{Mult}_k(n, \mathbf{p})$ where $k, \mathbf{p}$ don't depend on n. Then for $j = 1, 2, \dots, k$,

$$X_j \sim \text{Pois}(\lambda p_j).$$

1.2 Multivariate Normal

Definition 8 (Multivariate Normal (MVN))**.** Suppose $\mathbf{X}$ is a k-dimensional random vector. Then $\mathbf{X}$ follows **Multivariate Normal (MVN)** distribution if for any constants $t_1, \dots, t_k \in \mathbb{R}$,

$$t_1X_1 + \cdots + t_kX_k$$

is Normal (where 0 is consider to follow a *degenerate* Normal distribution).

Definition 9 (Bivariate Normal (BVN))**.** **X** follows a **Bivariate Normal (BVN)** distribution if it is a 2-dimensional MVN.

Result 10 (Uncorrelated MVN implies independence)**.** Suppose (X, Y) is bivariate normal with $\mathrm{Cov}(X, Y) = 0$ (i.e., X and Y are uncorrelated). Then X and Y are independent.
More generally, if **X** and **Y** (potentially vectors) are components of the same MVN and X_i, Y_j are uncorrelated for any i, j, then **X** and **Y** are independent.

☣ **11.** Please note the specific conditions under which Result 10 holds. It is always true that independent random variables are uncorrelated, but the converse is rarely a general truth. For example, two uncorrelated Normal random variables are not necessarily independent; we could only make that statement if we knew they were components of the same MVN.

Result 12 (Independence of sum and difference)**.** Suppose $X \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ are independent. Then $X + Y$ and $X - Y$ are also independent.

Result 13 (Concatenation)**.** Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots Y_m)$ are both Multivariate Normal with $\mathbf{X}, \mathbf{Y}$ independent of each other. Then $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ is also Multivariate Normal.

Result 14 (Subvector)**.** Suppose (X, Y, Z) is Multivariate Normal. Then (X, Y) is also Multivariate Normal. In general, any subvector of a Multivariate Normal still follows a Multivariate Normal distribution.

2 Practice Problems

1. Suppose

$$(X_1, X_2, X_3, X_4, X_5) \sim \text{Mult}_5(n, (p_1, p_2, p_3, p_4, p_5)).$$

   (a) What is the distribution of $X_1 + X_3 | X_2 + X_4 = x$?

   Solution
   By Multinomial lumping (Result 5)

$$(X_1 + X_3, X_2 + X_4, X_5) \sim \text{Mult}_3(n, (p_1 + p_3, p_2 + p_4, p_5)).$$

   By Multinomial conditioning (Result 4),

$$(X_1 + X_3, X_5) | X_2 + X_4 = x \sim \text{Mult}_2(n - x, \left(\frac{p_1 + p_3}{1 - p_2 - p_4}, \frac{p_5}{1 - p_2 - p_4}\right))$$

   By the Multinomial marginal (Result 3),

$$X_1 + X_3 | X_2 + X_4 = x \sim \text{Bin}\left(n - x, \frac{p_1 + p_3}{1 - p_2 - p_4}\right)$$

   (b) What about the distribution of $X_1 + X_3 | X_2 + X_4 + X_5 = x$? *Hint: should require little-to-no math.*

   Solution
   By definition, $X_1 + X_2 + X_3 + X_4 + X_5 = n$. So $P(X_1 + X_3 = n - x | X_2 + X_4 + X_5 = x) = 1$.

2. *(MVN operations)*

   (a) Suppose (X, Y, Z) is distributed Multivariate Normal. Show that $(X + Y, X + Z)$ is MVN.

   Solution

$$t_1(X + Y) + t_2(X + Z) = (t_1 + t_2)X + t_1 Y + t_2 Z$$

   which is Normal since (X, Y, Z) is MVN.

(b) Suppose $(X - Y, X, Z)$ is MVN. Show that (X, Y) is Bivariate Normal.

> **Solution**
>
> $$t_1 X + t_2 Y = -t_2(X - Y) + (t_1 + t_2)X + 0Z,$$
>
> which is Normal since $(X - Y, X, Z)$ is MVN.

(c) *(Example 7.5.2 in Blitzstein & Hwang)* Suppose $X \sim \mathcal{N}(0, 1)$ and S is a random sign, i.e.,

$$P(S = s) = \begin{cases} 1/2 & s \in \{-1, 1\}, \\ 0 & \text{else.} \end{cases}$$

Take for granted that $SX \sim \mathcal{N}(0, 1)$. Why is (X, SX) not Bivariate Normal? *Hint: show that $X + SX$ is not a solely continuous random variable: i.e., $P(X + SX = x) > 0$ for some $x \in \mathbb{R}$.*

> **Solution**
> See that
>
> $$P(X + SX = 0) = P(S = -1) = 1/2,$$
>
> so $X + SX$ is not continuous, and thus is not Normal (since the Normal is continuous). (X, SX) cannot be Bivariate Normal since $X + SX$ is a linear combination of its elements that is not Normal.

3. *(Example 7.5.10, Blitzstein & Huang)* Suppose $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $\rho \in (0,1)$. Let $Z = aX + bY$. Find the values of a and b (real valued constants) that give

$$\begin{aligned} Z &\sim \mathcal{N}(0,1), \\ \text{Cov}(X,Z) &= \rho. \end{aligned}$$

Hint: calculate the variance of Z and the covariance between Z and X in terms of a and b. Set them equal to the desired values.

Solution
By the independence of X and Y,

$$\text{Var}(Z) = a^2\text{Var}(X) + b^2\text{Var}(Y) = a^2 + b^2.$$

So we need $a^2 + b^2 = 1$. Then the covariance is

$$\text{Cov}(X,Z) = \text{Cov}(X, aX + bY) = a\text{Var}(X) + b\text{Cov}(X,Y) = a.$$

So $a = \rho$. That makes $b = \sqrt{1-\rho^2}$.