HOMEWORK 8

PSTAT 120A: Summer 2022

Due: 11:59pm Saturday, July 23
Instructor: Ethan P. Marzban

Instructions:

- Please submit your work to Gradescope by no later than the due date posted above.
- Be sure to show your work; correct answers with no supporting work will not be awarded full points.
- 2 randomly selected questions/parts will be graded, but you must still turn in your work for all problems in order to be eligible to earn full credit.

. . . .

1. **The Multinomial Distribution.** Recall that the Binomial distribution arises in the context of tracking the number of successes across n independent Bernoulli(p) trials. Definitionally, then, we require a binary division; namely a well-defined notion of "success" and "failure." Oftentimes, in Statistical Modeling, this is too stringent of a restriction.

   Suppose our n independent trials each result in one of r outcomes; as a simple case, when $r = 3$, we might say that our outcomes are "success," "failure," and "neutral." Additionally, suppose that each trial results in outcome i with probability p_i, for $i = 1, \cdots, r$. Let X_i denote the number of outcomes of type i we see (again, for $i = 1, \cdots, r$); then the random vector $(X_1, \cdots, X_r)$ is said to follow the **Multinomial Distribution** with parameters n (total number of trials), r (number of possible outcomes on each trial), and $p_1, \cdots, p_r$ (the probability of each outcome). We denote this:

$$(X_1, \cdots, X_r) \sim \text{Multi}(n, r, p_1, \cdots, p_n)$$

   Over the next few parts, we will investigate the Multinomial distribution in greater detail.

   PART I: Deriving the Joint P.M.F.

   (a) Suppose that PSTAT 120A has 100 students and 4 Discussion Sections (we can call them Sections 1 through 4). Further suppose that section 1 must contain 15 students, section 2 must contain 35, Section 3 must contain 20, and Section 4 must contain 30. In how many ways can we divide the students among these 4 sections?

   Solution: We can consider this akin to one of our poker questions:

   - From the 100, pick 15 to be in Section 1: $\binom{100}{15}$
   - From the remaining $100 - 15 = 85$, pick 35 to be in Section 2: $\binom{85}{35}$
   - From the remaining $85 - 35 = 50$, pick 20 to be in Section 3: $\binom{50}{20}$
   - From the remaining $50 - 20 = 30$, pick 30 to be in Section 4: $\binom{30}{30}$

   Hence, our final answer is

$$\binom{100}{15}\binom{84}{35}\binom{50}{20}\binom{30}{30}$$

(b) If n and r are positive integers, and $k_1, \cdots, k_r$ are nonnegative integers that sum to n (i.e. $k_1 + \cdots, k_r = n$), then the number of ways of assigning lables $1, 2, \cdots, r$ to n items so that, for each $i = 1, 2, \cdots, r$ exactly k_i items receive label i, is the **multinomial coefficient**

$$\binom{n}{k_1, k_2, \cdots, k_r} = \frac{n!}{(k_1!) \times (k_2!) \times \cdots \times (k_r)!}$$

Rewrite your answer to part (a) using a multinomial coefficient.

Solution: Let's rewrite our answer to part (a) a bit:

$$\begin{aligned}\binom{100}{15}\binom{85}{35}\binom{50}{20}\binom{30}{30} &= \frac{100!}{15! \cdot \cancel{85!}} \times \frac{\cancel{85!}}{35! \cdot \cancel{50!}} \times \frac{\cancel{50!}}{20! \cdot \cancel{30!}} \times \frac{\cancel{30!}}{30! \cdot 0!} \\ &= \frac{100!}{15! \cdot 35! \cdot 20! \cdot 30!} =: \binom{100}{15, 35, 20, 30}\end{aligned}$$

(c) Now, let's return to the Multinomial distribution. Find $p_{X_1, \cdots, X_r}(k_1, \cdots, k_r)$, the joint p.m.f. of $(X_1, \cdots, X_r)$. You may find it useful to revisit the methodology we used when deriving the p.m.f. of the Binomial distribution.

Solution: Iltimately, we wish to compute the probability of the event $\{X_1 = k_1, \cdots, X_r = k_r\}$. One possible configuration of outcomes that is included in this even is:

$$\underbrace{(\text{Type } 1) \ \cdots \ (\text{Type } 1)}_{k_1 \text{ times}} \ \underbrace{(\text{Type } 2) \ \cdots \ (\text{Type } 2)}_{k_2 \text{ times}} \ \cdots \ \underbrace{(\text{Type } r) \ \cdots \ (\text{Type } r)}_{k_r \text{ times}}$$

The probability of this particular outcome is

$$p_1^{k_1} \times p_2^{k_2} \times \cdots \times p_r^{k_r}$$

However, this is not the only outcome contained in the event $\{X_1 = k_1, \cdots, X_r = k_r\}$. We must multiply by the total number of ways to distribute the k_1 Type 1's, k_2 Type 2's, etc. across the n trials. The number of ways can be seen to be, by way of the work we did on parts (a) and (b) above,

$$\binom{n}{k_1, \cdots, k_r}$$

meaning

$$p_{X_1, \cdots, X_r}(k_1, \cdots, k_r) = \binom{n}{k_1, \cdots, k_r} \cdot p_1^{k_1} \times p_2^{k_2} \times \cdots \times p_r^{k_r}$$

(d) Speaking of the Binomial Distribution, show that the $\text{Multi}(n, 2, p_1, p_2)$ distribution is equivalent to the Binomial distribution.

Solution: Substituting above, we see

$$p_{X_1, X_2}(k_1, k_2) = \binom{n}{k_1, k_2} p_1^{k_1} p_2^{k_2}$$

Now, this might not immediately seem like the Binomial distribution. However, recall that $k_1 + k_2 = n$ by construction; therefore, $k_2 = n - k_1$. Additionally, $p_1 + p_2 = 1$ (also by construction), so $p_2 = 1 - p_1$. Therefore,

$$\begin{aligned} p_{X_1,X_2}(k_1,k_2) &= \binom{n}{k_1,k_2} p_1^{k_1} p_2^{k_2} \\ &= \binom{n}{k_1, n-k_1} p_1^{k_1}(1-p)^{n-k_1} \\ &= \binom{n}{k_1} p_1^{k_1}(1-p)^{n-k_1} \end{aligned}$$

which is perhaps more recognizable as a Binomial p.m.f..

PART II: Using the Joint P.M.F. In all parts that follow, continue to take $(X_1, \cdots, X_r) \sim \text{Multi}(n, r, p_1, \cdots, p_r)$

(e) What is the marginal distribution of X_1? (No summations needed; just make an argument about exactly *what* X_1 measures.)

Solution: X_1 measures the number of occurrences of type 1. Therefore, we can reclassify our outcomes as "type 1" and "not type 1," which reveals that

$$X_1 \sim \text{Bin}(n, p_1)$$

(f) Give an expression for $\text{Cov}(X_i, X_j)$, for $i, j = 1, \cdots, r$. **Hint:** There are two possible ways to solve this part.

(1) Consider the indicator defined by

$$\mathbb{1}_{k,i} = \begin{cases} 1 & \text{if trial } k \text{ gives outcome } i \\ 0 & \text{if trial } I \text{ gives an outcome other than } i \end{cases}$$

and express X_i as a suitable sum of these indicators.

(2) Alternatively, you can recognize the distribution of $(X_1 + X_2)$, compute its variance, and then use previously-derived results about variances of sums of random variables to obtain an equation involving $\text{Cov}(X_i, X_j)$ that you can solve for.

Solution: I'll illustrate using method 2 first. By a similar logic as in part (e), $(X_i + X_j) \sim \text{Bin}(n, p_i + p_j)$. Therefore,

$$\text{Var}(X_1 + X_2) = n(p_i + p_j)(1 - p_i - p_j)$$

We also have that, in general, $\text{Var}(X_i + X_j) = \text{Var}(X_i) + \text{Var}(X_j) + 2\text{Cov}(X_i, X_j)$. By part (e), $X_i \sim \text{Bin}(n, p_1)$ and $X_j \sim \text{Bin}(n, p_2)$ meaning

$$\begin{aligned} \text{Var}(X_1 + X_2) &= \text{Var}(X_i) + \text{Var}(X_j) + 2\text{Cov}(X_i, X_j) \\ &= np_i(1-p_i) + np_j(1-p_j) + 2\text{Cov}(X_i, X_j) \end{aligned}$$

Therefore, putting everything together, we find

$$n(p_i + p_j)(1 - p_i - p_j) = np_i(1-p_i) + np_j(1-p_j) + 2\text{Cov}(X_i, X_j)$$

which allows us to solve for $\mathrm{Cov}(X_i, X_j)$:

$$\begin{aligned} 2\mathrm{Cov}(X_i, X_j) &= n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j) \\ &= \cancel{np_i} - \cancel{np_i^2} - np_ip_j + \cancel{np_j} - np_ip_j - \cancel{np_j^2} - \cancel{np_i} + \cancel{np_i^2} - np_j + \cancel{np_j^2} \\ &= -2np_ip_j \end{aligned}$$

and hence

$$\mathrm{Cov}(X_j, X_j) = np_ip_j$$

This is, of course, true only if $i \neq j$; if $i = j$ then $\mathrm{Cov}(X_i, X_j) = \mathrm{Var}(X_i) = np_i(1 - p_i)$ meaning

$$\boxed{\mathrm{Cov}(X_i, X_j) = \begin{cases} np_i(1 - p_i) & \text{if } i = j \\ -np_ip_j & \text{if } i \neq j \end{cases}}$$

2. **Question 2 has been removed.**

3. Let X and Y be two continuous random variables with: $\mathbb{E}[X] = 6$, $\mathrm{Var}(X) = 4$, $\mathbb{E}[Y] = 6$, $\mathrm{Var}(Y) = 3$, and $\mathrm{Cov}(X, Y) = -1$. Use Chebyshev's Inequality to provide a bound for $\mathbb{P}(9 \leq X + Y \leq 15)$; be sure to specify whether this bound is an *upper* or *lower* bound.

Solution: Let $Z := X + Y$. Since X and Y are both nonnegative, their sum will also be nonnegative and so Z will be nonnegative. Additionally,

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 6 + 6 = 12 \\ \mathrm{Var}(Z) &= \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) = 4 + 3 - 2(1) = 5 \end{aligned}$$

Now, we write

$$\begin{aligned} \mathbb{P}(9 \leq X + Y \leq 15) &= \mathbb{P}(9 - 12 \leq Z - 12 \leq 15 - 12) \\ &= \mathbb{P}(-3 \leq Z - 12 \leq 3) = \mathbb{P}(|Z - 12| \leq 3) \geq 1 - \frac{\mathrm{Var}(Z)}{3^2} = 1 - \frac{5}{9} = \boxed{\frac{4}{9}} \end{aligned}$$

We can see that this is a lower bound.