Chapter 5. Multiple Random Variables

5.8: The Multinomial Distribution

Slides (Google Drive) Alex Tsun Video (YouTube)

As you've seen, the Binomial distribution is extremely commonly used, and probably the most important discrete distribution. The Normal distribution is certainly the most important continuous distribution. In this section, we'll see how to generalize the Binomial, and in the next, the Normal.

Why do we need to generalize the Binomial distribution? Sometimes, we don't just have two outcomes (success and failure), but we have $r > 2$ outcomes. In this case, we need to maintain counts of how many times each of the r outcomes appeared. A single random variable is no longer sufficient; we need a vector of counts!

Actually, the example problems at the end could have been solved in Chapter 1. We will just formalize this situation so that we can use it later!

5.8.1 Random Vectors (RVTRs) and Covariance Matrices

We will first introduce the concept of a random vector, which is just a collection of random variables stacked on top of each other.

Definition 5.8.1: Random Vectors

Let $X_1, ..., X_n$ be arbitrary random variables, and stack them into a vector like such:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

We call $\mathbf{X}$ an n-dimensional **random vector (rvtr)**.
We define the expectation of a random vector just as we would hope, coordinate-wise:

$$\mathbb{E}\left[\mathbf{X}\right] = \begin{bmatrix} \mathbb{E}\left[X_1\right] \\ \vdots \\ \mathbb{E}\left[X_n\right] \end{bmatrix}$$

What about the variance? We cannot just say or compute a single scalar $\mathsf{Var}\left(\mathbf{X}\right)$ because what does that mean for a random vector? Actually, we need to define an $n \times n$ covariance matrix, which stores all pairwise covariances. It is often denoted in one of three ways: $\Sigma = \mathsf{Var}\left(\mathbf{X}\right) = \mathsf{Cov}(\mathbf{X})$.

Definition 5.8.2: Covariance Matrices

The **covariance matrix** of a random vector $\mathbf{X} \in \mathbb{R}^n$ with $\mathbb{E}\left[\mathbf{X}\right] = \boldsymbol{\mu}$ is the matrix denoted $\Sigma =$

$\mathsf{Var}\left(\mathbf{X}\right) = \mathsf{Cov}(\mathbf{X})$ whose entries $\Sigma_{ij} = \mathsf{Cov}\left(X_i, X_j\right)$. The formula for this is:

$$\Sigma = \mathsf{Var}\left(\mathbf{X}\right) = \mathsf{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathbf{T}}\right] = \boxed{\mathbb{E}\left[\mathbf{X}\mathbf{X}^{\mathbf{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathbf{T}}}$$

$$= \begin{bmatrix} \mathsf{Cov}\left(X_1, X_1\right) & \mathsf{Cov}\left(X_1, X_2\right) & \dots & \mathsf{Cov}\left(X_1, X_n\right) \\ \mathsf{Cov}\left(X_2, X_1\right) & \mathsf{Cov}\left(X_2, X_2\right) & \dots & \mathsf{Cov}\left(X_2, X_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}\left(X_n, X_1\right) & \mathsf{Cov}\left(X_n, X_2\right) & \dots & \mathsf{Cov}\left(X_n, X_n\right) \end{bmatrix}$$

$$= \begin{bmatrix} \mathsf{Var}\left(X_1\right) & \mathsf{Cov}\left(X_1, X_2\right) & \dots & \mathsf{Cov}\left(X_1, X_n\right) \\ \mathsf{Cov}\left(X_2, X_1\right) & \mathsf{Var}\left(X_2\right) & \dots & \mathsf{Cov}\left(X_2, X_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}\left(X_n, X_1\right) & \mathsf{Cov}\left(X_n, X_2\right) & \dots & \mathsf{Var}\left(X_n\right) \end{bmatrix}$$

Notice that the covariance matrix is **symmetric** ($\Sigma_{ij} = \Sigma_{ji}$), and contains variances along the diagonal.

Note: If you know a bit of linear algebra, you might like to know that covariance matrices are always symmetric **positive semi-definite**.

We will not be doing any linear algebra in this class - think of it as just a place to store all the pairwise covariances. Now let us look at an example of a covariance matrix.

Example(s)

If $X_1, X_2, ..., X_n$ are iid with mean μ and variance σ^2, then find the mean vector and covariance matrix of the random vector $\mathbf{X} = (X_1, \dots, X_n)$.

Solution The mean vector is:

$$\mathbb{E}\left[\mathbf{X}\right] = \begin{bmatrix} \mathbb{E}\left[X_1\right] \\ \vdots \\ \mathbb{E}\left[X_n\right] \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \mu \mathbf{1_n}$$

where $\mathbf{1_n}$ denotes the n-dimensional vector of all 1's. The covariance matrix is (since the diagonal is just the individual variances σ^2 and the off-diagonals ($i \neq j$) are all $\mathsf{Cov}\left(X_i, X_j\right) = 0$ due to independence)

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

where I_n denotes the $n \times n$ identity matrix. □

Theorem 5.8.1: Properties of Expectation and Variance Hold for RVTRs

An important theorem is that properties of expectation and variance still hold for RVTRs.

Let $\mathbf{X}$ be an n-dimensional RVTR, $A \in \mathbb{R}^{n\times n}$ be a constant matrix, $\mathbf{b} \in \mathbb{R}^n$ be a constant vector. Then:

$$\mathbb{E}\left[A\mathbf{X}+\mathbf{b}\right] = A\mathbb{E}\left[\mathbf{X}\right]+\mathbf{b}$$
$$\mathsf{Var}\left(A\mathbf{X}+\mathbf{b}\right) = A\mathsf{Var}\left(\mathbf{X}\right)A^T$$

Since we aren't expecting any linear algebra background, we won't prove this.

5.8.2 The Multinomial Distribution

Suppose we have scenario where there are $r = 3$ outcomes, with probabilities p_1, p_2, p_3 respectively, such that $p_1 + p_2 + p_3 = 1$. Suppose we have $n = 7$ independent trials, and let $Y = (Y_1, Y_2, Y_3)$ be the rvtr of counts of each outcome. Suppose we define each X_i as a one-hot vector (exactly one 1, and the rest 0) as below, so that $Y = \sum_{i=1}^{n} X_i$ (this is exactly like how adding indicators/Bernoulli's gives us a Binomial):

	X_1	X_2	X_3	X_4	X_5	X_6	X_7		Y
OUTCOME 1	1	0	0	0	0	0	1		2
OUTCOME 2	0	0	0	1	0	0	0	Σ	1
OUTCOME 3	0	1	1	0	1	1	0		4

Now, what is the probability of this outcome (two of outcome 1, one of outcome 2, and four of outcome 3) - that is, $(Y_1 = 2, Y_2 = 1, Y_3 = 4)$? We get the following:

$$\begin{aligned} p_{Y_1,Y_2,Y_3}(2,1,4) &= \frac{7!}{2!1!4!}\cdot p_1^2 \cdot p_2^1 \cdot p_3^4 && \text{[recall from counting]} \\ &= \binom{7}{2,1,4}\cdot p_1^2 \cdot p_2^1 \cdot p_3^4 \end{aligned}$$

This describes the joint distribution of the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$, and its PMF should remind of you of the binomial PMF. We just count the number of ways $\binom{7}{2,1,4}$ to get these counts (multinomial coefficient), and make sure we get each outcome that many times $p_1^2 p_2^1 p_3^4$.

Now let us define the Multinomial Distribution more generally.

Definition 5.8.3: The Multinomial Distribution

Suppose there are r outcomes, with probabilities $\mathbf{p} = (p_1, p_2, ..., p_r)$ respectively, such that $\sum_{i=1}^{r} p_i = 1$. Suppose we have n independent trials, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$ be the rvtr of counts of each outcome. Then, we say:

$$\mathbf{Y} \sim \text{Mult}_r(n, \mathbf{p})$$

The joint PMF of $\mathbf{Y}$ is:

$$p_{Y_1,\ldots,Y_r}(k_1,\ldots k_r) = \binom{n}{k_1,\ldots,k_r}\prod_{i=1}^{r} p_i^{k_i}, \quad k_1,\ldots k_r \geq 0 \text{ and } \sum_{i=1}^{r} k_i = n$$

Notice that each Y_i is marginally $\text{Bin}(n, p_i)$. Hence, $\mathbb{E}[Y_i] = np_i$ and $\mathsf{Var}(Y_i) = np_i(1-p_i)$. Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\mathbf{p} = \begin{bmatrix} np_1 \\ \vdots \\ np_r \end{bmatrix} \qquad \mathsf{Var}(Y_i) = np_i(1-p_i) \qquad \mathsf{Cov}(Y_i, Y_j) = -np_ip_j \quad (\text{for } i \neq j)$$

Notice the covariance is negative, which makes sense because as the number of occurrences of Y_i increases, the number of occurrences of Y_j should decrease since they can not occur simultaneously.

Proof of Multinomial Covariance. Recall that marginally, X_i and X_j are binomial random variables; let's decompose them into their Bernoull trials. We'll use different dummy indices as we're dealing with covariances.

Let X_{ik} for $k = 1, \ldots, n$ be indicator/Bernoulli rvs of whether the k^{th} trial resulted in outcome i, so that $X_i = \sum_{k=1}^{n} X_{ik}$

Similarly, let $X_{j\ell}$ for $\ell = 1, \ldots, n$ be indicators of whether the ℓ^{th} trial resulted in outcome j, so that $X_k = \sum_{\ell=1}^{n} X_{j\ell}$.

Before we begin, we should argue that $\mathsf{Cov}(X_{ik}, X_{j\ell}) = 0$ when $k \neq \ell$ since k and ℓ are different trials and are independent.

Furthermore, $\mathbb{E}[X_{ik}X_{jk}] = 0$ since it's not possible that both outcome i and j occur at trial k.

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathsf{Cov}\left(\sum_{k=1}^{n} X_{ik}, \sum_{\ell=1}^{n} X_{j\ell}\right) && \text{[indicators]} \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathsf{Cov}(X_{ik}, X_{j\ell}) && \text{[covariance works like FOIL]} \\
&= \sum_{k=1}^{n} \mathsf{Cov}(X_{ik}, X_{jk}) && \text{[independent trials, cross terms are 0]} \\
&= \sum_{k=1}^{n} \mathbb{E}[X_{ik}X_{jk}] - \mathbb{E}[X_{ik}]\,\mathbb{E}[X_{jk}] && \text{[def of covariance]} \\
&= \sum_{k=1}^{n} -p_ip_j && \text{[first expectation is 0]} \\
&= -np_ip_j
\end{aligned}$$

Note that in the third line we dropped one of the sums because the indicators across different trials k, ℓ are independent (zero covariance). Hence, we just need to sum when $k = \ell$. □

There is an example of the Multinomial distribution at the end of the section!

5.8.3 The Multivariate Hypergeometric (MVHG) Distribution

Suppose there are $r = 3$ political parties (Green, Democratic, Republican). The senate consists of $N = 100$ senators: $K_1 = 45$ Green party members, $K_2 = 20$ Democrats, and $K_3 = 35$ Republicans.

We want to choose a committee of $n = 10$ senators.

Let $Y = (Y_1, Y_2, Y_3)$ be the number of each party's members in the committee (G, D, R in that order). What is the probability we get 1 Green party member, 6 Democrats, and 3 Republicans? It turns out is just the following:

$$p_{Y_1,Y_2,Y_3}(1, 6, 3) = \frac{\binom{45}{1}\binom{20}{6}\binom{35}{3}}{\binom{100}{10}}$$

This is very similar to the univariate Hypergeometric distribution! For the denominator, there are $\binom{100}{10}$ ways to choose 10 senators. For the numerator, we need 1 from the 45 Green party members, 6 from the 20 Democrats, and 3 from the 35 Republicans.

Once again, let us define the MVHG Distribution more generally.

Definition 5.8.4: The Multivariate Hypergeometric Distribution

Suppose there are r different colors of balls in a bag, having $\mathbf{K} = (K_1, ..., K_r)$ balls of each color, $1 \le i \le r$. Let $N = \sum_{i=1}^r K_i$ be the total number of balls in the bag, and suppose we draw n without replacement. Let $\mathbf{Y} = (Y_1, ..., Y_r)$ be the rvtr such that Y_i is the number of balls of color i we drew. We write that:

$$\mathbf{Y} \sim \text{MVHG}_r(N, \mathbf{K}, n)$$

The joint PMF of Y is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \frac{\prod_{i=1}^r \binom{K_i}{k_i}}{\binom{N}{n}}, \quad 0 \le k_i \le K_i \text{ for all } 1 \le i \le r \text{ and } \sum_{i=1}^r k_r = n$$

Notice that each Y_i is marginally $\text{HypGeo}(N, K_i, n)$, so $\mathbb{E}[Y_i] = n\frac{K_i}{N}$ and $\mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N - K_i}{N} \cdot \frac{N - n}{N - 1}$. Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\frac{\mathbf{K}}{N} = \begin{bmatrix} n\frac{K_1}{N} \\ \vdots \\ n\frac{K_r}{N} \end{bmatrix} \qquad \mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N - K_i}{N} \cdot \frac{N - n}{N - 1} \qquad \mathsf{Cov}(Y_i, Y_j) = -n\frac{K_i}{N}\frac{K_j}{N} \cdot \frac{N - n}{N - 1}$$

Proof of Hypergeometric Variance. We'll prove the variance of a univariate Hypergeometric finally (the variance of Y_i), but leave the covariance matrix to you (can approach it similarly to the multinomial covariance matrix).

Let $X \sim \text{HypGeo}(N, K, n)$ (univariate hypergeometric). For $i = 1, \ldots, n$, let X_i be the indicator of whether or not we got a success on trial i (not independent indicators). Then, $\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = \frac{K}{N}$ for every trial i, so $\mathbb{E}[X] = n\frac{K}{N}$ by linearity of expectation.

First, we have that since $X_i \sim \text{Ber}\left(\frac{K}{N}\right)$:

$$\mathsf{Var}\left(X_i\right) = p(1-p) = \frac{K}{N}\left(1 - \frac{K}{N}\right)$$

Second, for $i \neq j$, $\mathbb{E}\left[X_i X_j\right] = \mathbb{P}\left(X_i X_j = 1\right) = \mathbb{P}\left(X_i = 1\right)\mathbb{P}\left(X_j = 1 \mid X_i = 1\right) = \frac{K}{N} \cdot \frac{K-1}{N-1}$, so

$$\mathsf{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] = \frac{K}{N} \cdot \frac{K-1}{N-1} - \frac{K^2}{N^2}$$

Finally,

$$\begin{aligned}
\mathsf{Var}\left(X\right) &= \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) && \text{[def of } X\text{]} \\
&= \mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) && \text{[covariance with self is variance]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{Cov}\left(X_i, X_j\right) && \text{[bilinearity of covariance]} \\
&= \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right) && \text{[split diagonal]} \\
&= n\frac{K}{N}\left(1 - \frac{K}{N}\right) + 2\binom{n}{2}\left(\frac{K}{N} \cdot \frac{K-1}{N-1} - \frac{K^2}{N^2}\right) && \text{[plug in]} \\
&= n\frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1} && \text{[algebra]}
\end{aligned}$$

□

5.8.4 Exercises

These won't be very interesting since this could've been done in chapter 1 and 2!

1. Suppose you are fishing in a pond with 3 red fish, 4 green fish, and 5 blue fish.

   (a) You use a net to scoop up 6 of them. What is the probability you scooped up 2 of each?

   (b) You "catch and release" until you caught 6 fish (catch 1, throw it back, catch another, throw it back, etc.). What is the probability you caught 2 of each?

   Solution:

   (a) Let (X_1, X_2, X_3) be how many red, green, and blue fish I caught respectively. Then, $X \sim \text{MVHG}_3(N = 12, \mathbf{K} = (3, 4, 5), n = 6)$, and

   $$\mathbb{P}\left(X_1 = 2, X_2 = 2, X_3 = 2\right) = \frac{\binom{3}{2}\binom{4}{2}\binom{5}{2}}{\binom{12}{6}}$$

(b) Let (X_1, X_2, X_3) be how many red, green, and blue fish I caught respectively. Then, $X \sim \text{Mult}_3(n = 6, \mathbf{p} = (3/12, 4/12, 5/12))$, and

$$\mathbb{P}\left(X_1 = 2, X_2 = 2, X_3 = 2\right) = \binom{6}{2, 2, 2} \left(\frac{3}{12}\right)^2 \left(\frac{4}{12}\right)^2 \left(\frac{5}{12}\right)^2$$