STAT 512: Statistical Inference **Autumn 2020**

Lecture 7: Multinomial distribution

Instructor: Yen-Chi Chen

The multinomial distribution is a common distribution for characterizing categorical variables. Suppose a random variable Z has k categories, we can code each category as an integer, leading to $Z \in \{1, 2, \cdots, k\}$. Suppose that $P(Z = k) = p_k$. The parameter $\{p_1, \cdots, p_k\}$ describes the entire distribution of k (with the constraint that $\sum_j p_j = 1$). Suppose we generate $Z_1, \cdots, Z_n$ IID from the above distributions and let

$$X_j = \sum_{i=1}^{n} I(Z_i = j) = \# \text{ of observations in the category } j.$$

Then the random vector $X = (X_1, \cdots, X_k)$ is said to be from a multinomial distribution with parameter $(n, p_1, \cdots, p_k)$. We often write

$$X \sim M_k(n; p_1, \cdots, p_k)$$

to denote a multinomial distribution.

Example (pet lovers). The following is a hypothetical dataset about how many students prefer a particular animal as a pet. Each row (except the 'total') can be viewed as a random vector from a multinomial distribution. For instance, the first row $(18, 20, 6, 4, 2)$ can be viewed as a random draw from a multinomial distribution $M_5(n = 50; p_1, \cdots, p_5)$. The second and the third row can be viewed as other random draws from the same distribution.

	cat	dog	rabbit	hamster	fish	total
Class 1	18	20	6	4	2	50
Class 2	15	15	10	5	5	50
Class 3	17	18	8	4	3	50

7.1 Properties of multinomial distribution

The PMF of a multinomial distribution has a simple closed-form. If $X \sim M_k(n; p_1, \cdots, p_k)$, then

$$p(X = x) = p(X_1 = x_1, \cdots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

The *multinomial coefficient* $\frac{n!}{x_1!x_2!\cdots x_k!} = \binom{n}{x_1, \cdots, x_n}$ is the number of possible ways to put n balls into k boxes.

The famous *multinomial expansion* is

$$(a_1 + a_2 + \cdots + a_k)^n = \sum_{x_i \geq 0, \sum_i x_i = n} \frac{n!}{x_1! x_2! \cdots x_k!} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}.$$

This implies that $\sum_{x_i \geq 0, \sum_i x_i = n} p(X = x) = 1$.

By the construction of a multinomial $M_k(n; p_1, \cdots, p_k)$, one can easily see that if $X \sim M_k(n; p_1, \cdots, p_k)$, then

$$X = \sum_{i=1}^{n} Y_i,$$

where $Y_1, \cdots, Y_n \in \{0, 1\}^k$ are IID multinomial random variables from $M_k(1; p_1, \cdots, p_k)$.

Thus, the moment generating function of X is

$$M_X(s) = \mathbb{E}[e^{s^T X}] = \mathbb{E}[e^{s^T Y_1}]^n = \left(\sum_{j=1}^{k} p_j e^{s_j}\right)^n$$

The multinomial distribution has a nice additive property. Suppose $X \sim M_k(n; p_1, \cdots, p_k)$ and $V \sim M_k(m; p_1, \cdots, p_k)$ and they are independent. It is easy to see that

$$X + V \sim M_k(n + m; p_1, \cdots, p_k).$$

Suppose we focus on one particular category j, then you can easily show that

$$X_j \sim \mathsf{Bin}(n, p_j).$$

Note that $X_1, \cdots, X_k$ are not independent due to the constraint that $X_1 + X_2 + \cdots + X_k = n$. Also, for any X_i and X_j, you can easily show that

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

An intuitive way to think of this is that the number $X_i + X_j$ is the number of observations in either category i or categeory j. So we are essentially pulling two categories together.

7.2 Conditional distribution of multinomials

The multinomial distribution has many interesting properties when conditioned on some other quantities. Here we illustrate the idea using a four category multinomial distribution but the idea can be generalized to other more sophisticated scenarios.

Let $X = (X_1, X_2, X_3, X_4) \sim M_4(n; p_1, p_2, p_3, p_4)$. Suppose we combine the last two categories into a new category. Let $W = (W_1, W_2, W_3)$ be the resulting random vector. By construction, $W_3 = X_3 + X_4$ and $W_1 = X_1, W_2 = X_2$. Also, it is easy to see that

$$W \sim M_3(n, q_1, q_2, q_3), \quad q_1 = p_1, q_2 = p_2, q_3 = p_3 + p_4.$$

So pulling two or more categories together will result in a new multinomial distribution.

Let $Y = (Y_1, Y_2)$ such that $Y_1 = X_1 + X_2$ and $Y_2 = X_3 + X_4$. We know that $Y \sim M_2(n; p_1 + p_2, p_3 + p_4)$. What will the conditional distribution of $X|Y$ be?

$$\begin{aligned}
p(X = x|Y = y) &= \frac{p(x_1, x_2, x_3, x_4)}{p(y_1, y_2)} I(y_1 = x_1 + x_2, y_2 = x_3 + x_4) \\
&= \frac{\frac{n!}{x_1! x_2! x_3! x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}}{\frac{n!}{y_1! y_2!} (p_1 + p_2)^{y_1} (p_3 + p_4)^{y_2}} I(y_1 = x_1 + x_2, y_2 = x_3 + x_4) \\
&= \frac{(x_1 + x_2)!}{x_1! x_2!} \left(\frac{p_1}{p_1 + p_2}\right)^{x_1} \left(\frac{p_2}{p_1 + p_2}\right)^{x_2} \times \frac{(x_3 + x_4)!}{x_3! x_4!} \left(\frac{p_3}{p_3 + p_4}\right)^{x_3} \left(\frac{p_4}{p_3 + p_4}\right)^{x_4} \\
&= p(x_1, x_2|y_1) p(x_3, x_4|y_2)
\end{aligned}$$

so we conclude that (1)

$$(X_1, X_2) \perp (X_3, X_4)|Y,$$

i.e., they are conditionally independent, and (2)

$$X_1, X_2|X_1{+}X_2 \sim M_2\left(X_1+X_2; \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right), \quad X_3, X_4|X_3{+}X_4 \sim M_2\left(X_3+X_4; \frac{p_3}{p_3+p_4}, \frac{p_4}{p_3+p_4}\right).$$

Because $X_1 + X_2 = n - X_3 - X_4$, the above result also implies that

$$X_1, X_2|X_3, X_4 \stackrel{d}{=} X_1, X_2|n - X_3 - X_4 \sim M_2\left(n - X_3 - X_4; \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right),$$

where $X \stackrel{d}{=} Y$ means that the two random variables have the same distribution. Thus, one can see that (X_1, X_2) and (X_3, X_4) are negatively correlated.

General case. Suppose that we can partition $X_1, \cdots, X_k$ into r blocks

$$\underbrace{(X_1, \cdots, X_{k_1})}_{B_1}, \underbrace{(X_{k_1+1}, \cdots, X_{k_2})}_{B_2}, \cdots, \underbrace{(X_{k_{r-1}+1}, \cdots X_{k_r})}_{B_r}.$$

Then we have $B_1, \cdots, B_r$ are conditionally independent given $S_1, \cdots, S_r$, where $S_1 = \sum_{i=1}^{k_1} X_i = \sum_j B_{i,j}$ and $S_r = \sum_{i=k_{r-1}}^{k_r} X_i = \sum_j B_{r,j}$ are the block-specific sum.

Also,

$$B_j|S_j \sim M_{k_j - k_{j-1}}\left(S_j; \frac{p_{k_{j-1}+1}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}, \cdots, \frac{p_{k_j}}{\sum_{\ell=k_{j-1}+1}^{k_j} p_\ell}\right).$$

Now we turn to a special case, consider $X \sim M_k(n; p_1, \cdots, p_k)$. We focus on only two variables X_i and X_j $(i \neq j)$. What will the conditional distribution of $X_i|X_j$ be?

Using the above formula, we choose $r = 2$ and the first block contains everything except X_j and the second block only contains X_j. This implies that $S_1 = n - S_2 = n - X_j$. Thus,

$$(X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k)|X_j \stackrel{d}{=} (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k)|n{-}X_j \sim M_{k-1}\left(n - X_j; \frac{p_1}{1-p_j}, \cdots, \frac{p_k}{1-p_j}\right).$$

So the marginal distribution

$$X_i|X_j \sim \mathsf{Bin}\left(n - X_j, \frac{p_i}{1-p_j}\right).$$

As a result, we see that X_i and X_j are negatively correlated. Also,

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathbb{E}[\underbrace{\mathsf{Cov}(X_i, X_j|X_j)}_{=0}] + \mathsf{Cov}(\mathbb{E}[X_i|X_j], \underbrace{\mathbb{E}[X_j|X_j]}_{=X_j}) \\
&= \mathsf{Cov}(\mathbb{E}[X_i|X_j], X_j) \\
&= \mathsf{Cov}\left((n - X_j)\frac{p_i}{1-p_j}, X_j\right) \\
&= -\frac{p_i}{1-p_j}\mathsf{Var}(X_j) \\
&= -np_ip_j.
\end{aligned}$$

7.3 Estimating the parameter of multinomials

In reality, we observe a random vector X from a multinomial distribution. We often know the total number of individuals n but the parameters $p_1, \cdots, p_k$ are often unknown that have to be estimated. Here we will explain how to use the MLE to estimate the parameter.

In a multinomial distribution, the parameter space is $\Theta = \{(p_1, \cdots, p_k) : 0 \leq p_j, \sum_{j=1}^k p_j = 1\}$. We observe the random vector $X = (X_1, \cdots, X_k) \sim M_k(n; p_1, \cdots, p_k)$. In this case, the likelihood function is

$$L_n(p_1, \cdots, p_k|X) = \frac{n!}{X_1! \cdots X_k!} p_1^{X_1} \cdots p_k^{X_k}$$

and the log-likelihood function is

$$\ell_n(p_1, \cdots, p_k|X) = \sum_{j=1}^k X_j \log p_j + C_n,$$

where C_n a constant is independent of p. Note that naively computing the score function and set it to be 0 will not grant us a solution (think about why) because we do not use the constraint of the parameter space – the parameters are summed to 1. To use this constraint in our analysis, we consider adding the Lagrange multipliers and optimize it:

$$F(p, \lambda) = \sum_{j=1}^k X_j \log p_j + \lambda \left(1 - \sum_{j=1}^k p_j\right).$$

Differentiating this function with respect to $p_1, \cdots, p_k$, and λ and set it to be 0 gives

$$\frac{\partial F}{\partial p_j} = \frac{X_j}{p_j} - \lambda \Rightarrow X_j = \hat{\lambda} \cdot \hat{p}_{MLE,j}$$

and $1 - \sum_{j=1}^k \hat{p}_{MLE,j} = 0$. Thus, $n = \sum_{j=1}^k X_j = \hat{\lambda} \sum_{j=1}^k p_j = \hat{\lambda}$ so $\hat{p}_{MLE,j} = \frac{X_j}{n}$, which is just the proportion of category j.

7.4 Dirichlet distribution

The Dirichlet distribution is a distribution of continuous random variables relevant to the Multinomial distribution. Sampling from a Dirichlet distribution leads to a random vector with length k and each element of this vector is non-negative and summation of elements is 1, meaning that it generates a random probability vector.

The Dirichlet distribution is a multivariate distribution over the simplex $\sum_{i=1}^k x_i = 1$ and $x_i \geq 0$. Its probability density function is

$$p(x_1, \cdots, x_k; \alpha_1, \cdots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i - 1},$$

where $B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha)}{\Gamma(\sum_{i=1}^k \alpha_i)}$ with $\Gamma(a)$ being the Gamma function and $\alpha = (\alpha_1, \cdots, \alpha_K)$ are the parameters of this distribution.

You can view it as a generalization of the Beta distribution. For $Z = (Z_1, \cdots, Z_k) \sim \mathsf{Dirch}(\alpha_1, \cdots, \alpha_k)$, $\mathbb{E}(Z_i) = \frac{\alpha_i}{\sum_{j=1}^k \alpha_j}$ and the mode of Z_i is $\frac{\alpha_i - 1}{\sum_{j=1}^k \alpha_j - k}$ so each parameter α_i determines the relative importance of category (state) i. Because it is a distribution putting probability over K categories, Dirchlet distribution is very popular in social sciences and linguistics analysis.

The Dirchlet distribution is often used as a prior distribution for the multinomial parameter $p_1, \cdots, p_k$ in Bayesian inference. The fact that it generates a probability vector makes it an excellent candidate for this job.

Let $p = (p_1, \cdots, p_k)$. Assume that

$$X|p = (X_1, \cdots, X_k)|p \sim M_k(n; p_1, \cdots, p_k)$$

and we place a prior

$$p \sim \mathsf{Dirch}(\alpha_1, \cdots, \alpha_k).$$

The two distributional assumptions imply that the posterior distribution of p will be

$$\begin{aligned}\pi(p|X) &\propto \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \times \frac{1}{B(\alpha)} p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \\ &\propto p_1^{x_1+\alpha_1-1} \cdots p_k^{x_k+\alpha_k-1} \\ &\sim \mathsf{Dirch}(x_1 + \alpha_1, \cdots, x_k + \alpha_k).\end{aligned}$$

If we use the posterior mean as our estimate, then

$$\hat{p}_{\pi,i} = \frac{x_i + \alpha_i}{\sum_{j=1}^{k} x_j + \alpha_j},$$

which is the MLE when we observe the counts $x' = (x'_1, \cdots, x'_k)$ such that $x'_j = x_j + \alpha_j$ (but note that α_j does not have to be an integer). So the prior parameter α_j can be viewed as a pseudo count of the category j before collecting the data.