August 21, 2018

Some Facts About the Multinomial Distribution

Dennis D. Cox
Department of Statistics
Rice University

August 21, 2018

1 Some Definitions and Notations

A *multi-index* $\underline{m} = (m_1, \ldots, m_k)$ is a vector of non-negative integers. We use $|\underline{m}| = \sum_i m_i$ to denote the sum of the components (which is also the ℓ_1 length of the vector). We define a *monomial* as an expression

$$\underline{a}^{\underline{m}} = \prod_{i=1}^{k} a_i^{m_i}.$$

where $\underline{a} = (a_1, \ldots, a_k)$ is a vector of real numbers and $\underline{m}$ is a multi-index of the same dimension as $\underline{a}$. Since each m_i is a positive integer, $a_i^{m_i}$ is always defined (which wouldn't be the case if $m_i = 1/2$ or $m_i = -1$). One special case we need to stipulate is that

$$0^0 = 1.$$

Thus, the function $x \mapsto x^0$ is continuous, but the function $x \mapsto 0^x$, for $x \geq 0$ is not continuous at $x = 0$.

Next, we introduce the *multinomial coefficient*,

$$\left(\begin{array}{c} n \\ \underline{m} \end{array} \right) = \left\{ \begin{array}{cl} \frac{n!}{\prod_{i=1}^{k} m_i!} & \text{if } |\underline{m}| = n; \\ 0 & \text{otherwise.} \end{array} \right.$$

Here, n is a positive integer and $\underline{m} = (m_1, \ldots, m_k)$ is a multi-index. The usual binomial coefficient is a special case of the multi-nomial with $k = 2$, i.e.

$$\binom{n}{(m_1, m_2)} = \binom{n}{m_1},$$

provided $m_1 + m_2 = n$. One might think the notation above is confusing, but just remember that when the bottom expression in the notation is a vector, then the multi-nomial notation is in use, and when it is a scalar, then it is the binomial coefficient notation.

The multinomial coefficients have two important and useful mathematical properties, one algebraic and one combinatorial. We first consider the algebraic application.

Theorem 1 (Multinomial Formula)

$$\left(\sum_{i=1}^{k} a_i\right)^n = \sum_{\underline{m}} \binom{n}{\underline{m}} \underline{a}^{\underline{m}},$$

where the sum is over all multi-indices $\underline{m}$ of dimension k. (Of course, all terms that don't satisfy $|\underline{m}| = n$ are 0.)

Proof: We proceed by induction on k, the number of terms. We assume the reader is familiar with the binomial formula, which is the result when $k = 2$. Now assume it is true for some $k \geq 2$ and show that it follows for $k + 1$. Suppose $\underline{a} = (a_1, \ldots, a_k, a_{k+1})$. For $1 \leq j \leq k$, let

$$b_j = \begin{cases} a_j & \text{if } j < k, \\ a_k + a_{k+1} & \text{if } j = k. \end{cases}$$

Then, by the induction hypothesis that the formula is true for k terms, we get

$$\begin{aligned} \left(\sum_{j=1}^{k} b_j\right)^n &= \sum_{\underline{m}} \binom{n}{\underline{m}} \underline{b}^{\underline{m}} \\ &= \sum_{\underline{m}:|\underline{m}|=n} \frac{n!}{\prod_{i=1}^{k} m_i!} \prod_{i=1}^{k} b_i^{m_i} \\ &= \sum_{\underline{m}:|\underline{m}|=n} \frac{n!}{\prod_{i=1}^{k} m_i!} \left(\prod_{i=1}^{k-1} a_i^{m_i}\right) (a_k + a_{k+1})^{m_k} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{\underline{m}:|\underline{m}|=n} \frac{n!}{\prod_{i=1}^{k} m_i!} \left(\prod_{i=1}^{k-1} a_i^{m_i}\right) \sum_{i=0}^{m_k} \frac{m_k!}{i!(m_k-i)!} a_k^i a_{k+1}^{m_k-i} \\
&= \sum_{\underline{m}:|\underline{m}|=n} \sum_{\ell=0}^{m_k} \frac{n!}{\prod_{i=1}^{k} m_i!} \frac{m_k!}{\ell!(m_k-\ell)!} \left(\prod_{i=1}^{k-1} a_i^{m_i}\right) a_k^{\ell} a_{k+1}^{m_k-\ell}.
\end{aligned}
$$

(Question: Why did we need to change the notation for the index of summation in "$\sum_{\ell=0}^{m_k}$" to ℓ from i in the line before that?) The binomial formula was used in the fourth equation of the last calculation. In the last expression, replace the multi-index $\underline{m}$ with $\underline{\alpha} = (m_1, \ldots, m_{k-1}, \ell, m_k - \ell)$. Note that this is a general $k+1$ dimensional multi-index with $|\underline{\alpha}| = n$. Also,

$$
\begin{aligned}
\underline{a}^{\underline{\alpha}} &= \left(\prod_{i=1}^{k-1} a_i^{m_i}\right) a_k^{\ell} a_{k+1}^{m_k-\ell}, \\
\left(\begin{array}{c} n \\ \underline{\alpha} \end{array}\right) &= \frac{n!}{\prod_{i=1}^{k} m_i!} \frac{m_k!}{\ell!(m_k-\ell)!}, \\
\sum_{\underline{m}:|\underline{m}|=n} \sum_{\ell=0}^{m_k} &= \sum_{\underline{\alpha}:|\underline{\alpha}|=n} .
\end{aligned}
$$

Putting these into the previous calculations and using the definition of $\underline{b}$ we have

$$
\begin{aligned}
\left(\sum_{j=1}^{k} b_j\right)^n &= \left(\sum_{j=1}^{k+1} a_j\right)^n \\
&= \sum_{\underline{\alpha}:|\underline{\alpha}|=n} \left(\begin{array}{c} n \\ \underline{\alpha} \end{array}\right) \underline{a}^{\underline{\alpha}},
\end{aligned}
$$

which is the result needed to prove the theorem.

The other property we need is a combinatorial result. If A is a set, then $|A|$ is the number of elements of A, which is either a nonnegative integer or ∞. Given an integer $n > 0$ and a multi-index $\underline{m}$ such that $|\underline{m}| = n$, define

$$
\begin{aligned}
\mathcal{C}(n, \underline{m}) \quad = \quad & \{h : \ h \text{ is a function mapping } \{1, \ldots, n\} \longrightarrow \{1, \ldots, k\} \\
& \quad \text{such that } \forall i \in \{1, \ldots, k\}, \sum_{j=1}^{n} I_{\{i\}}(h(j)) = m_i\}.
\end{aligned}
$$

Note that $\sum_{j=1}^{n} I_{\{i\}}(h(j))$ is the number of values $(h(1), \ldots, h(n))$ for which $h(j) = i$. Expressed differently, $\mathcal{C}(n, \underline{m})$ is all possible ways we can assign n

things to k classes with m_1 things in the first class, m_2 things in the second class, …, and m_k things in the k-th class.

Theorem 2

$$|\mathcal{C}(n, \underline{m})| \quad = \quad \left(\begin{array}{c} n \\ \underline{m} \end{array} \right).$$

Proof: We prove this by induction on k. The case $k = 2$ is the combinatorial property of the binomial coefficient. Note that a mapping $h : \{1, \ldots, n\} \longrightarrow \{1, 2\}$ can be regarded as assigning "success" (i.e., $h(x) = 1$) or "failure" ($h(x) = 2$) to each of the integers $1, \ldots, n$, and the number of ways this can be done such that the number of successes is $\sum_{i=1}^n I_{\{1\}}(h(i)) = m_1$ and the number of failures is $\sum_{i=1}^n I_{\{2\}}(h(i)) = m_2$ is

$$\left(\begin{array}{c} n \\ m_1 \end{array} \right) = \left(\begin{array}{c} n \\ (m_1, n - m_1) \end{array} \right).$$

We assume this combinatorial property of the binomial coefficient is already known.

So assume the result is true for some $k \geq 2$ and we will show it is true for $k + 1$. For any map $h : \{1, \ldots, n\} \longrightarrow \{1, \ldots, k+1\}$ in $\mathcal{C}(n, \underline{m})$, define $h' : \{1, \ldots, n\} \longrightarrow \{1, \ldots, k\}$ by

$$h'(i) \ = \ \min\{h(i), k\}.$$

Thus, h' agrees with h except when $h(i) = k + 1$ and then $h'(i) = k$. Note that h' is in $\mathcal{C}(n, \underline{m}')$ where $\underline{m}'$ is a k-dimensional multi-index with $m'_j = m_j$ for $j < k$ and $m'_k = m_k + m_{k+1}$. By the induction hypothesis, the number of such h' maps is

$$\left(\begin{array}{c} n \\ \underline{m}' \end{array} \right).$$

Now, for each such h' map, there are

$$\left(\begin{array}{c} m'_k \\ m_k \end{array} \right)$$

h maps in $\mathcal{C}(n, \underline{m})$ that give rise to the same h' map. Thus, the total number of h maps in $\mathcal{C}(n, \underline{m})$ is

$$\left(\begin{array}{c} n \\ \underline{m}' \end{array} \right) \left(\begin{array}{c} m'_k \\ m_k \end{array} \right) \quad = \quad \frac{n!}{\left(\prod_{i=1}^{k-1} m_i!\right) m'_k!} \frac{m'_k!}{m_k! m_{k+1}!}$$

$$\begin{aligned} &= \frac{n!}{\left(\prod_{i=1}^{k-1} m_i!\right) m_k! m_{k+1}!} \\ &= \left(\begin{array}{c} n \\ \underline{m} \end{array}\right), \end{aligned}$$

which completes the proof.

One can use this theorem to prove the multinomial formula of the first theorem by observing that when we expand out all the terms in the n-fold product of the summation, the number of terms equal to any particular monomial is given by the multinomial coefficient.

2 Derivation of the Multinomial Distribution

Suppose $\tilde{Y}$, $\tilde{Y}_1$, $\tilde{Y}_2$, …, $\tilde{Y}_n$ are i.i.d. random variables taking values in $\{1, 2, \ldots, c\}$ where $c > 1$ is a positive integer. We think of the Y's as categorical random variables with the numerical value being a code for the particular categories. For instance, if the categories are one of the political affiliations "republican", "democrat", or "independent", then we may code them as 1,2,3, repectively. Denote the probability of observing a particular class j by

$$\pi_j \;=\; P[\tilde{Y} = j].$$

Of course, the vector $\pi = (\pi_1, \ldots, \pi_c)$ has nonnegative entries that sum to 1. We will refer to such a vector as a *probability vector*.

Define the general indicator function

$$I_A(y) \;=\; \begin{cases} 1 & \text{if } y \in A, \\ 0 & \text{otherwise,} \end{cases}$$

where A is a set of possible y values. Now consider the random c-dimensional vectors

$$\underline{Y}_i \;=\; \left(I_{\{1\}}(\tilde{Y}_i), I_{\{2\}}(\tilde{Y}_i), \ldots, I_{\{c\}}(\tilde{Y}_i)\right).$$

This is a vector of 0's and 1's with a single 1. Note that $\underline{Y}_i$ and the $\tilde{Y}_i$ are basically equivalent. We can write down the probability mass function for the random data matrix as

$$P[\underline{Y}_1 = \underline{y}_1, \ldots, \underline{Y}_n = \underline{y}_n] \quad = \quad \prod_{i=1}^{n} P[\underline{Y}_i = \underline{y}_i]$$

$$\begin{aligned} &= \prod_{i=1}^{n} \prod_{j=1}^{c} \pi_j^{\underline{y}_{ij}} \\ &= \prod_{j=1}^{c} \pi_j^{\sum_i \underline{y}_{ij}} \end{aligned}$$

If we think of this as a likelihood function for the parameter vector π, then it depends on the observed data $\underline{y}_1$, …, $\underline{y}_n$ only through their sum. Hence, the sum

$$T \;=\; \sum_{i=1}^{n} \underline{Y}_i$$

is a sufficient statistic. (Note: we will begin writing vectors without the under lines. It should be clear from context which quantities are vectors and which are not.) Note that the j'th component of T is the number of Y_i where the category j occurred. Clearly we can reconstruct the data set from T, except for the particular order that the categories occurred, and since we are assuming i.i.d. observations, there is no information about π in the ordering.

Now we claim that the probability mass function (p.m.f.) of the multinomial is given by

$$P[T = t] \;=\; \begin{pmatrix} n \\ t \end{pmatrix} \pi^t .$$

Let $p(t)$ denote the r.h.s. of this equation. We first show it defines a valid p.m.f., i.e., that it is nonnegative and sums to 1. Since the components of π are nonnegative and the multinomial coefficient is nonnegative, it follows that $p(t) \geq 0$. To show it sums to 1, we use the fact that the components of π sum to 1 and apply the multinomial formula:

$$\begin{aligned} 1 &= \left(\sum_{j=1}^{c} \pi_j \right)^n \\ &= \sum_t \begin{pmatrix} n \\ t \end{pmatrix} \pi^t , \end{aligned}$$

where the last summation is over all c-dimensional multi-indices t. We call the probability distribution with the p.m.f. $p(t)$ above the $\mathbf{Mult}(n, \pi)$ distribution. It assigns positive probability only to the c-dimensional multi-indices t such that $|t| = n$.

Now, the moment generating function (m.g.f.) for the $\mathbf{Mult}(n, \pi)$ distribution is

$$\begin{aligned}
\psi(u) &= \sum_t e^{u^T t} p(t) \\
&= \sum_t \begin{pmatrix} n \\ t \end{pmatrix} \prod_{i=1}^{c} e^{u_i t_i} \pi_i^{t_i} \\
&= \left(\sum_{i=1}^{c} e^{u_i} \pi_i \right)^n ,
\end{aligned}$$

where the last line follows from the multinomial formula. Now, it is easy to check that for $n = 1$ categorical observation, the $\mathbf{Mult}(1, \pi)$ is the correct distribution. Recall that the m.g.f. of a sum of independent random vectors is the product of the corresponding m.g.f.'s of the summands, so we see that the $\mathbf{Mult}(n, \pi)$ is the distribution for the sum of n i.i.d. $\mathbf{Mult}(1, \pi)$ random vectors, which is how we defined T above. This verifies that we have the right p.m.f.

It is clear that each component T_j of a $\mathbf{Mult}(n, \pi)$ random vector has binomial $\mathbf{Bin}(n, \pi_j)$ distribution. This follows by considering an observation in class j as "success" and in any other class not equal to j as "failure." Thus, from the (presumably) known facts about the binomial, the component mean and variance is

$$\mathrm{E}[T_j] = n\pi_j, \quad \mathrm{Var}[T_j] = n\pi_j(1 - \pi_j).$$

We are also interested in the covariance between components of a $\mathbf{Mult}(n, \pi)$ vector. This can be computed using the m.g.f. Assume $j \neq k$, then

$$\begin{aligned}
\mathrm{E}[T_j T_k] &= \left. \frac{\partial^2}{\partial u_j \partial u_k} \psi(u) \right|_{u=0} \\
&= \left. \frac{\partial^2}{\partial u_j \partial u_k} \left(\sum_{i=1}^{c} e^{u_i} \pi_i \right)^n \right|_{u=0} \\
&= \left. \frac{\partial}{\partial u_j} n e^{u_k} \pi_k \left(\sum_{i=1}^{c} e^{u_i} \pi_i \right)^{n-1} \right|_{u=0} \\
&= \left. n(n-1) e^{u_j} e^{u_k} \pi_j \pi_k \left(\sum_{i=1}^{c} e^{u_i} \pi_i \right)^{n-2} \right|_{u=0} \\
&= n(n-1)\pi_j \pi_k .
\end{aligned}$$

Note that $j \neq k$ was used to obtain the fourth equality from the third in the previous calculation. Thus, we have for $j \neq k$ that

$$\mathrm{Cov}(T_j, T_k) \;=\; E[T_jT_k]-E[T_j]E[T_k] \;=\; n(n-1)\pi_j\pi_k-(n\pi_j)(n\pi_k) \;=\; -n\pi_j\pi_k.$$

The fact that the covariance is negative should not come as a surprise. If T_j happens to be larger than its mean, then we would expect other component of T to be smaller since the constraint $\sum_k T_k = n$ must hold.

3 A Connection Between the Multinomial and Poisson Distributions

Now we look at an important property of the multinomial when we randomize the sample size n according to a Poisson distribution. The $\mathbf{Pois}(\mu)$ distribution (with parameter $\mu \geq 0$) is defined on the nonnegative integers and has p.m.f.

$$p(x) \;=\; e^{-\mu}\mu^x/x!, \quad x = 0, 1, 2, \ldots.$$

We assume the reader is familiar with this distribution. Now consider a two-stage random experiment: generate N according to a $\mathbf{Pois}(\mu)$ distribution with $\mu > 0$, and given a value of $N = n$, generate a random k-dimensional multi-index T according to

$$T \sim \begin{cases} \mathbf{Mult}(n, \pi) & \text{if } n > 0, \\ \delta_0 & \text{if } n = 0. \end{cases}$$

Here, δ_0 is a unit point mass (degenerate probability) distribution that puts probability 1 at the point 0. Also, π is a given k-dimensional probability vector. If $n = 0$, then of course always T (the vector of numbers of observations in each category value) is also 0, and the factorials and powers in the formula for the multinomial p.m.f. will all be 1 at $t = 0$ and 0 if $t \neq 0$. Now, we compute the unconditional distribution of T. Given a multi-index t, if we observe $T = t$, then there is only one value of N that could occur, namely $N = |t|$. So

$$\begin{aligned} P[T=t] \;&=\; \sum_{n=0}^{\infty} P[T=t|N=n]P[N=n] \\ &=\; \sum_{n=0}^{\infty} \binom{n}{t} \pi^t e^{-\mu}\mu^n/n! \end{aligned}$$

$$
\begin{aligned}
&= \frac{|t|!}{\prod_{i=1}^{k} t_i!} \left(\prod_{i=1}^{c} \pi_i^{t_i} \right) e^{-\mu} \mu^{|t|} / |t|! \\
&= \prod_{i=1}^{c} e^{-\pi_i \mu} (\pi_i \mu)^{t_i} / t_i!.
\end{aligned}
$$

But this is precisely the p.m.f. for independent $\mathbf{Pois}(\pi_i \mu)$. Of course, our sample size n is typically not random, much less Poisson distributed, but if we imagine it is, then we can use this last result. In particular, it provides a new likelihood for the multinomial data, although it introduces an additional parameter, namely μ, which would be a nuisance parameter for making inference about the multinomial parameter π.