Control
and Cybernetics
VOL. 1 (1972) No. 1/2

Multinomial distribution and its application to inspection of product quality

by

JAN BUTKIEWICZ, SZYMON FIRKOWICZ

Institute o Applied Cybernetics
Polish Academy of Sciences,
Warszawa

There are given the basic properties of the multinomial distribution and its connection with a multinomial Poisson distribution and Dirichlet distribution. There is considered an application of this distribution for estimation and inspection of product quality as well as to statistical inferring on the basis of many random samples by using multidimensional alternative classification of the product items being tested.

1. Basic properties of the multinomial distribution

The multinomial distribution is a multivariate generalization of the binomial distribution.

Let us consider k attributes of an inspected product and use an alternative classification of items into defective and nondefective relatively to each attribute. Hence, we have 2^k classes of the product: one class of nondefective items of the product which we will index with 0 and 2^k-1 classes of defective items of product (having defects) which we will index with $1, 2, \ldots, 2^k-1$. Taking into account the possibility of the emptiness of some classes of defective items (when joint outcomes of some particular defects may be imposible) we shall let m denote the number of not empty classes of defective items of the product where $1 \leqslant m \leqslant 2^k-1$. If p_j denotes a fraction of items contained in the j-class, so $0<p_j<1$ $(j=0, 1, \ldots, m)$

$$\sum_{j=1}^{m} p_j=1 \tag{1}$$

and a result of an inspection of a simple random sample of size n may be described by m-dimensional random vector $Z=(Z_1, Z_2, \ldots, Z_m)$ and $Z_0=n-\sum_{j=1}^{m} Z_j$, having a $(m+1)$-nomial distribution with the probability function

$$p(z; n, p)=P\left(\bigcap_{j=1}^{m}(Z_j=z_j);\ n, p\right)=n!\prod_{j=0}^{m}\frac{p_j^{z_j}}{z_j!} \tag{2}$$

where $p=(p_1, p_2, ..., p_m)$ is a vector of parameters and $p_0=1-\sum_{j=1}^{m} p_j$, $z=(z_1, ..., z_m)$ is an observed realization of Z, $z_0=n-\sum_{j=1}^{m} z_j$ is an observed realization of Z_0; Z_j denotes a random number of items in the j-class.

When n is fixed the random vector Z may have $\binom{n+m}{m}$ realizations, which is easy to prove using the induction method.

The characteristic function of random vector $Z=(Z_1, ..., Z_m)$, which has the $(m+1)$-nomial distribution whose probability function is defined by the equation (2), is specified at any point $t=(t_1, ..., t_m)$ by the formula

$$\varphi(t)=\left(p_0+\sum_{j=1}^{m} p_j \exp it_j\right)^n. \tag{3}$$

The expectations, variances and covariances of components of the vector Z are given by the following formulas

$$\mathrm{E}(Z_j)=np_j, \qquad j=1, 2, ..., m, \tag{4}$$

$$\mathrm{Var}(Z_j)=np_j(1-p_j), \quad j=1, 2, ..., m, \tag{5}$$

$$\mathrm{Cov}(Z_i, Z_j)=-np_i p_j, \qquad i\neq j;\ i, j\in\{1, 2, ..., m\}. \tag{6}$$

The binomial distribution is a particular case of the multinomial distribution $(m=1)$. If the random vector $Z=(Z_1, Z_2, ..., Z_m)$ has the multinomial distribution, then its components $Z_1, Z_2, ..., Z_m$ have marginal binomial distributions.

On assuming $np_j=\Lambda_j$, where Λ_j are constant positive numbers $(j=1, 2, ..., m)$, it may be proved that by $n\to\infty$ a limit of the probability function of $(m+1)$-nomial distribution (2) is a probability function of m-dimensional Poisson distribution

$$P\left(\bigcap_{j=1}^{m}(Z_j=z_j);\ \Lambda\right)=\prod_{j=1}^{m}\frac{\Lambda_j^{z_j}}{z_j!}\exp(-\Lambda_j) \tag{7}$$

where $\Lambda=(\Lambda_1, ..., \Lambda_m)$, $z_j\in\{0, 1, ...\}$.

Hence, in order to approximate the multinomial distribution in the case of large values of n and small numbers $p_1, ..., p_m$ we can take the m-dimensional Poisson distribution (7) with $\Lambda_j=np_j$. To approximate the binomial distribution by the Poisson distribution it is required that the inequalities $n>20$, $p<0.2$ be satisfied, similarly to approximate the $(m+1)$-nomial distribution by the m-dimensional Poisson distribution (7) it is required that the inequalities $n>20$, $\max(p_1, ..., p_m)<0.2$ be satisfied.

It should be stressed that the multinomial distribution is reproductive (additive) with respect to n. Distributions of its components and distributions of sums containing any number of these components are multinomial distributions; so they are also reproductive with respect to n.

The multinomial distribution is connected with a Dirichlet distribution being a multidimensional generalization of.a beta distribution.

The Dirichlet distribution [1] of a random vector $X=(X_1, ..., X_m)$ whose realizations $x=(x_1, ..., x_m)$ are in a simplex

$$S_m=\left\{x: x_j>0,\ j=1, 2, ..., m; \sum_{j=1}^{m} x_j<1\right\} \tag{8}$$

is described by the probability density function

$$f_D(x; \nu)=\begin{cases} \Gamma\left(\sum_{j=0}^{m} \nu_j\right) \prod_{j=0}^{m} \frac{x_j^{\nu_j-1}}{\Gamma(\nu_j)} & \text{for} \quad x \in S_m \\ 0 \quad \text{otherwise} \end{cases} \tag{9}$$

where $\nu=(\nu_0, \nu_1, ..., \nu_m)$ is a vector of positive parameters and $x_0=1-\sum_{j=1}^{m} x_j$.

The particular case of the Dirichlet distribution $(m=1)$ is the beta distribution. If the random vector $X=(X_1, ..., X_m)$ has the Dirichlet distribution, then its components $X_1, X_2, ..., X_m$ have marginal beta distributions [1]. The binomial distribution of random variable Z_j is connected with the beta distribution of random variable X_j [2] by the relation

$$P(Z_j \leqslant z_j;\ n, p_j)=P(X_j \geqslant p_j;\ z_j+1, n-z_j)=1-J_{p_j}(z_j+1, n-z_j) \tag{10}$$
$$j=1, 2, ..., m)$$

where $J_{p_j}(z_j+1, n-z_j)$ denotes the distribution function of the beta distribution at the point p_j with parameters z_j+1, $n-z_j$. Further the multinomial distribution of a random vector $Z=(Z_1, Z_2, ..., Z_m)$ is connected with the Dirichlet distribution of a random vector $X=(X_1, X_2, ..., X_m)$ by the relations [3]:

$$\begin{aligned} P\left(\bigcap_{j=1}^{m} (Z_j \leqslant z_j); n, p\right) &= \\ =P\left(\bigcap_{j=1}^{m} (X_j \geqslant p_j);\ \nu_j=z_j+1, j=1, ..., m;\ \nu_0=n+1-m-\sum_{j=1}^{m} z_j\right) &= \\ =1-F_D\left(p; \nu_j=z_j+1,\ j=1, ..., m;\ \nu_0=n+1-m-\sum_{j=1}^{m} z_j\right) & \end{aligned} \tag{11}$$

$$\begin{aligned} P\left(\bigcap_{j=1}^{m} (Z_j \leqslant z_j); n, p\right) &= \\ =P\left(\bigcap_{j=1}^{m} (X_j \geqslant p_j);\ \nu_j=z_j+1, j=1, ..., m;\ \nu_0=n+1-m-\sum_{j=1}^{m} z_j\right) &= \\ =\int_{S_{m,p}} f_D\left(x;\ \nu_j=z_j+1,\ j=1, ..., m;\ \nu_0=n+1-m-\sum_{j=1}^{m} z_j\right) \mathrm{d}x & \end{aligned} \tag{12}$$

where $F_D(p; \nu)$ denotes the distribution function of the Dirichlet distribution and S_m contains the simplex

$$S_{m,p}=\{x: x \in S_m,\ x_j>p_j,\ j=1, ..., m\}. \tag{13}$$

2. Estimation of product quality by multinomial alternative classification

Using the multidimensional alternative classification to estimate a product quality it is convenient to apply the following quality index

$$Q=\sum_{j=0}^{m} q_j p_j = q_0 - \sum_{j=1}^{m} (q_0 - q_j) p_j \tag{14}$$

where m denotes the number of not empty classes, $q_0, q_1, \ldots, q_m$ denote partial quality indices of particular product classes provided that the following relations are satisfied

$$0 < q_j < \infty \quad (j=1, \ldots, m)$$
$$\max(q_1, q_2, \ldots, q_m) < q_0 < \infty \tag{15}$$

and p_j $(j=0, 1, \ldots, m)$ have previous meanings.

We can give some interpretations: a partial quality index q_j may denote a market value of the product in the j-classe, a quality index Q (14) may denote the probability that the product is having the ability to function correctly, so Q may denote a measure of product reliability.

The estimator of quality index Q is given by the statistic

$$\hat{Q} = q_0 - \frac{1}{n} \sum_{j=1}^{m} (q_0 - q_j) Z_j \tag{16}$$

and its realization

$$Q^* = q_0 - \frac{1}{n} \sum_{j=1}^{m} (q_0 - q_j) z_j \tag{17}$$

represents an estimate of an unknown value of Q in the population which is described by the random sample. Being a linear function of random variables $Z_1, Z_2, \ldots, Z_m$, the statistic $\hat{Q}$ (16) has an asymptotically normal distribution with the mean

$$E(\hat{Q}) = Q \tag{18}$$

and the variance

$$\mathrm{Var}(\hat{Q}) = \frac{1}{n}\left[\sum_{j=1}^{m} (q_0 - q_j)^2 p_j (1 - p_j) - \sum_{i \neq j}^{m} (q_0 - q_i)(q_0 - q_j) p_i p_j\right.$$

which follow from the equations (16) and (4)—(6). To estimate this variance we put $\frac{z_1}{n}, \ldots, \frac{z_m}{n}$ instead of unknown probabilities $p_1, \ldots, p_m$ in the relation (18) and we have:

$$\mathrm{Var}^*(\hat{Q}) = \frac{1}{n}\left[\sum_{j=1}^{m} (q_0 - q_j)^2 \frac{z_j}{n}\left(1 - \frac{z_j}{n}\right) - \sum_{i \neq j}^{m} (q_0 - q_i)(q_0 - q_j) \frac{z_i}{n} \cdot \frac{z_j}{n}\right]. \tag{19}$$

Then we obtain confidence limits of confidence interval for Q with confidence coefficient β by approximate formulas

$$\bar{Q} \approx Q^* + y_{1-\alpha}\sqrt{\mathrm{Var}^*(\hat{Q})}$$
$$\underline{Q} \approx Q^* - y_{1-\alpha}\sqrt{\mathrm{Var}^*(\hat{Q})} \tag{20}$$

where

$$\alpha = \begin{cases} 1-\beta \text{ for an one-sided interval,} \\ \dfrac{1-\beta}{2} \text{ for a two-sided interval,} \end{cases}$$

$y_{1-\alpha}$, satisfying the equation $P(Y \geqslant y_{1-\alpha}) = \alpha$, denotes the $(1-\alpha)$-th quantile of the normal distribution $N(0, 1)$.

Taking into account the realization $z = (z_1, z_2, \ldots, z_m)$ of the random vector $Z = (Z_1, Z_2, \ldots, Z_m)$ in the sample of size n, and using the relations (10) and (14), it is easy to prove that the expectation of quality index Q of the inspected product is given by the formula

$$E(Q|\, z, n) = q_0 - \frac{1}{n+1} \sum_{j=1}^{m} (q_0 - q_j)(z_j + 1). \tag{21}$$

3. Inspection of product quality by multidimentional alternative classification

Suppose partial quality indices are given $q_0, q_1, \ldots, q_m$ and it is required that for the products the inequality $Q \geqslant Q_0$ holds. Using the remarks of the section 2 we see that the requirement $Q \geqslant Q_0$ is not satisfied with the confidence coefficient α provided that the following inequality

$$Q^* + y_{1-\alpha}\sqrt{\mathrm{Var}^*(Q)} \leqslant Q_0 \tag{22}$$

holds, otherwise there is no reason for taking this decision.

However, a postulated requirement for a product quality may be connected with probabilities $p_1, p_2, \ldots, p_m$ and not with quality index Q. Let us suppose that it is required to satisfy the inequalities

$$p_1 \leqslant p_{10},\ p_2 \leqslant p_{20},\ \ldots,\ p_m \leqslant p_{m0} \tag{23}$$

where there are given $p_{j0} > 0$ and

$$0 < \sum_{j=1}^{m} p_{j0} < 1. \tag{23a}$$

Certainly the satisfying of the above requirement implies the satisfying of inequalities

$$p_0 \geqslant 1 - \sum_{j=1}^{m} p_{j0} \tag{24}$$

$$Q \geqslant q_0 - \sum_{j=1}^{m} (q_0 - q_j) p_{j0}. \tag{25}$$

So the fact that the requirement (23) is satisfied implies the fact that the requirement $Q \geqslant Q_0$ is satisfied, where Q_0 denotes the right-handed side of the inequality (25). It should be stressed that the requirement (23) is stronger than $Q \geqslant Q_0$, because the inequality $Q \geqslant Q_0$ may be also held in the case when any p_j satisfies the inequality $p_j > p_{j0}$. The requirement (23) is not satisfied if there exists such p_j that $p_j > p_{j0}$. Hence, if $z = (z_1, ..., z_m)$ is the observed realization of the random vector $Z = (Z_1, ..., Z_m)$ in the sample of size n, we approve the decision that the requirement (23) is not satisfied provided that the probability (11), in which we put $p_j = p_{j0}$, $j = 1, ..., m$, does not exceed α. This is equivalent to the statement that the distribution function of Dirichlet distribution with parameters $\nu_0 = n + 1 - m - \sum_{j=1}^{m} z_j$, $\nu_j = z_j + 1$, $j = 1, ..., m$, exceeds the value $1 - \alpha$, that is $F_D(p_{10}, ..., p_{m0};\ \nu_j = z_j + 1,\ j = 1, ..., m;\ \nu_0 = n + 1 - m - \sum_{j=1}^{m} z_j) \geqslant 1 - \alpha$.

Otherwise the result of the inspection by sampling does not contradict the fact that the product quality satisfies the imposed requirement (23). It may be proved, by using the relations (10) and (14), that if inspected products satisfy the requirement (23) and it is known from sampling the realization $z = (z_1, ..., z_m)$ of the random vector $Z = (Z_1, ..., Z_m)$ in the sample of size n, then the expectation of product index Q is given by the relation

$$E(Q \mid z, n;\ p_j \leqslant p_{j0},\ j = 1, ..., m) = \\ = q_0 - \frac{1}{n+1} \sum_{j=1}^{m} (q_0 - q_j)(z_j + 1) J_{p_{j0}}(z_j + 2, n - z_j) \tag{27}$$

where $J_{p_{j0}}(z_j + 2, n - z_j)$ denotes the distribution function of the beta distribution with parameters $z_j + 2$, $n - z_j$ at the point p_{j0}.

4. Inferring from many samples

Suppose $r \geqslant 2$ random samples, of sizes n_i, $i = 1, ..., r$, of the same product were tested and the realizations $z_i = (z_{1i}, ..., z_{mi})$ of the radom vector $Z = (Z_1, ..., Z_m)$ were observed.

If populations represented by these samples do not differ, then unknown probabilities of defective classes p_j $(j = 1, ..., m)$ may be evaluated by the joint sample using the relation

$$p_j = \frac{\sum_{i=1}^{r} z_{ji}}{\sum_{i=1}^{r} n_i} \qquad (j = 1, ..., m). \tag{28}$$

Furthermore to evaluate a quality index Q we use the relations of section 2 and 3 taking into account

$$n=\sum_{i=1}^{r} n_i, \quad z_j=\sum_{i=1}^{r} z_{ji} \tag{29}$$

To verify the hypothesis that the tested samples are derived from the same population it is possible to apply the following inference. If a random vector $Z=(Z_1, Z_2, ..., Z_m)$ has the multinomial distribution (2), then the statistic

$$U=\sum_{j=0}^{m} \frac{(Z_j-np_j)^2}{np_j} \tag{30}$$

where $Z_0=n-\sum_{j=1}^{m} Z_j$, $p_0=1-\sum_{j=1}^{m} p_j$, has asymptotically χ^2 distribution with m degrees of freedom [1].

The χ^2_m distribution is reproductive with respect to m, so putting instead of an unknown probability p_j its evaluation p_j^* we lose one degree of freedom. Hence, taking into account α as the level of significance of the test, we can describe the region of rejection of the hypothesis that samples are derived from the same population using the inequality

$$\sum_{i=1}^{r}\sum_{j=0}^{m} \frac{(z_{ji}-n_i p_j)^2}{n_i p_j} \geqslant \chi^2_{m(r-1),\,1-\alpha} \tag{31}$$

where $\chi^2_{m(r-1),\,1-\alpha}$ denotes the $(1-\alpha)$-th quantile of the statistique $\chi^2_{m(r-1)}$ being described by the equality

$$P\,(\chi^2_{m(r-1)} \geqslant \chi^2_{m(r-1),\,1-\alpha})=\alpha.$$

References

1. Wilks S. S., Mathematical statistics. New York, 1963.
2. Firkowicz S., Statystyczne badanie wyrobów. Warszawa 1970.
3. Butkiewicz J., Wykorzystanie wielomianowego rozkładu do k-wymiarowej alternatywnej klasyfikacji. *Arch. Autom. i Telemech.* **16**, 3 (1971).

Rozkład wielomianowy i jego zastosowanie w badaniach wyrobów

Podano podstawowe właściwości rozkładu wielomianowego oraz jego powiązanie z wielowymiarowym rozkładem Poissona i rozkładem Dirichleta. Omówiono wykorzystanie tego rozkładu do oceny i kontroli jakości wyrobów oraz do wnioskowania statystycznego z wielu próbek losowych przy stosowaniu wielowymiarowej klasyfikacji alternatywnej badanych wyrobów.

Полиномиальное распределение и его применение в испытаниях изделий

Приведены основные свойства полиномиального распределения и его связь с распределениями многомерным Пуассона и Дирихле. Оговорено использование этого распределения для оценки и контроля качества изделий, а также для статистического анализа результатов испытаний нескольких выборок.