

Covariance Identity for Multinomial Trials

BY ANDREW. L. RUKHIN ¹

*Department of Mathematics and Statistics, University of Maryland,
Baltimore County Campus, 1000 Hilltop Circle, Baltimore, Maryland
21250, U. S. A.; rukhin@math.umbc.edu*

ABSTRACT

A formula for the covariance of duration times between successive occurrences of two different outcomes in multinomial trials is derived. Its relationship to the Banach match-box problem and restricted randomization designs is mentioned.

Keywords: Allocation rules, Banach match-box problem, Combinatorial extreme-value theory; Incomplete beta-function; Multinomial trials; Probability generating function, Restricted randomization design.

¹This research was supported by NSA grant #MDA904-00-1-0033.

1 The main result

Consider the classical setting of independent multinomial trials with possible outcomes $1, \dots, K$, whose probabilities are p_1, \dots, p_K . Denote by $X_j^{(k)}$ the number of the trial resulting in the j -th occurrence of the outcome k . For example, $X_1^{(k)}$ the trial at which the outcome k , $k = 1, \dots, K$, occurs for the first time, and $X_2^{(k)} > X_1^{(k)}$, is the trial corresponding to the second occurrence of this outcome. Clearly, $X_j^{(k)}$ has the negative binomial distribution on positive integers $1, 2, \dots$ with parameters j and p_k . It is well known that

$$X_r^{(k)} = \sum_{j=0}^{r-1} U_j^{(k)},$$

where $U_j^{(k)}$ are independent random variables with the geometric, parameter p_k , distribution. The positive random variable $U_i^{(k)} = X_{i+1}^{(k)} - X_i^{(k)}$, $i = 1, 2, \dots$ has the meaning of the duration time between two successive occurrences (i and $i + 1$) of the outcome k ; $U_0^{(k)} = X_1^{(k)}$ is the number of the first trial resulting in the outcome k .

Unless K is large, these classical random variables are highly dependent. However, the author was unable to find in the literature the expression for their covariances, and the goal of this note is to establish the following beautiful formulas.

PROPOSITION For $k \neq \ell$, $i, j = 1, 2, \dots$,

$$\text{Cov} \left(X_i^{(k)}, X_j^{(\ell)} \right) = - \sum_{0 \leq m < i, 0 \leq n < j} \binom{m+n}{m} \frac{p_k^m p_\ell^n}{(p_k + p_\ell)^{m+n+1}}. \quad (1)$$

Also

$$\text{Cov}(U_i^{(k)}, U_j^{(\ell)}) = - \binom{i+j}{i} \frac{p_k^i p_\ell^j}{(p_k + p_\ell)^{i+j+1}}. \quad (2)$$

The most striking form of (2) happens when $K = 2$ and $p_1 + p_2 = 1$. According to (2), $\text{Cov}(U_i^{(1)}, U_j^{(2)})$ is merely negative binomial probability of i successes in $i + j$ trials,

$$\text{Cov}(U_i^{(1)}, U_j^{(2)}) = - \binom{i+j}{i} p_1^i (1 - p_1)^j.$$

An alternative expression for (1) (which can be used to prove Proposition 1) is

$$\begin{aligned} & \text{Cov} \left(X_i^{(k)}, X_j^{(\ell)} \right) \\ &= -\frac{1}{p_k p_\ell} \left[i p_\ell I_{\frac{p_k}{p_k+p_\ell}}(i, j) + j p_k I_{\frac{p_\ell}{p_k+p_\ell}}(j, i) \right] + \frac{p_k^{i-1} p_\ell^{j-1}}{(p_k + p_\ell)^{i+j-1} B(i, j)}. \end{aligned} \quad (3)$$

Here $I_p(i, j)$ denotes the incomplete beta-function.

According to (2), the correlation coefficient of duration times between successive occurrences of outcomes k and ℓ has the form

$$\text{corr}(U_i^{(k)}, U_j^{(\ell)}) = -\binom{i+j}{i} \frac{p_k^{i+1} p_\ell^{j+1}}{(p_k + p_\ell)^{i+j+1} \sqrt{(1-p_k)(1-p_\ell)}}.$$

When $p_k \equiv 1/K$, the formula (3) shows that

$$\text{Cov} \left(X_m^{(k)}, X_m^{(\ell)} \right) = -mK \left[1 - \binom{2m}{m} \frac{1}{2^{2m}} \right],$$

so that as $m \rightarrow \infty$

$$\text{corr} \left(X_m^{(k)}, X_m^{(\ell)} \right) \sim -\frac{1}{K-1}. \quad (4)$$

2 Relationship to Classical Probability and An Application

The interest in the covariance of duration periods between successive occurrences of two different outcomes in multinomial trials arises in the study of the restricted randomization designs for clinical trials involving several treatments. A survey of these designs and their comparative merits is given in Rosenberger and Lachin (2002). According to a classical truncated multinomial design, the allocation process starts with the uniform distribution, until a treatment receives the prescribed number m of patients, after which this uniform distribution switches to the remaining treatments, and so on. If K denotes the number of treatments, then for independent K -nomial trials with uniform probabilities,

$$\tau = \min_{k=1, \dots, K} [X_m^{(k)}],$$

is the random moment at which the first treatment receives the given number m of patients.

Clearly, the distribution of τ is related to that of the maximal coordinate of a multinomial vector as

$$P(\tau > x) = P(\text{after } x \text{ trials frequencies of all outcomes} < m).$$

Levin (1981) had suggested an efficient algorithm for the numerical evaluation of these probabilities. The generating function for this distribution can be found in Riordan (1958), ex 6. Ch 5. It is also used by David and Barton (1962) in Ch 6 to derive the combinatorial extreme value distribution of the largest cell frequency. Our proof of Proposition is also based on the probability generating function in Lemmas 2 and 3.

When $K = 2$, the stopping rule τ has the probability distribution

$$P(\tau = m + x) = \binom{m + x - 1}{x} \frac{1}{2^{m+x-1}}, \quad x = 0, 1, \dots, m - 1,$$

derived first by Blackwell and Hodges (1957). This is merely the probability of $m - 1$ successes in $m + x - 1$ symmetric Bernoulli trials.

The random variable τ is closely related to the classical Banach match-box problem as described in Feller, 1968, p 166, or in Blom, Holst, and Sandell (1994), pp 9–11 and p 125. Let B denote the random variable equal to the number of matches remaining in one match box when another box is emptied. Then $\tau = 2m - B$, given that the original number of matches in each box is m . An extension of this distribution for unequal probabilities of choosing the boxes is given by Uppuluri and Blot (1970).

According to (4) for $K = 2$, the correlation coefficient between $X_m^{(1)}$ and $X_m^{(2)}$, tends to -1 as $m \rightarrow \infty$. As the limiting joint distribution of $(X_m^{(1)}, X_m^{(2)})$ is normal, it must be concentrated on the line $z_1 = -z_2$. Thus, $(\tau - 2m)/(2m)^{1/2} = (\min(X_m^{(1)}, X_m^{(2)}) - 2m)/(2m)^{1/2}$ is seen to be asymptotically distributed as $\min(Z, -Z) = -|Z|$ with Z denoting the standard normal random variable.

This result was originally proven by Blackwell and Hodges (1957). Holst (1989) studied the match-box problem in the case when the demand for the matches is described by a Poisson process. He has shown that even in this general case, for $m \rightarrow \infty$ the distribution of $(2m - \tau)/(2m)^{1/2}$ converges weakly to that of $|Z|$. Note that in the continuous time case the analogs of

the variables $X_j^{(k)}$ are independent as they correspond to independent Poisson processes describing demands for matches in each pocket.

3 Proofs

LEMMA 1. For $k \neq \ell$ and $i \leq x < y$,

$$p_{xy} = P\left(X_i^{(k)} = x, X_j^{(\ell)} = y\right) = \sum \binom{x-1}{i-1 \quad \nu_2 \quad \nu_3} \binom{y-x-1}{\mu_1 \quad \mu_2 \quad \mu_3} p_k^{i+\mu_1} p_\ell^j (1-p_k-p_\ell)^{\nu_3+\mu_3}, \quad (5)$$

where the summation is taken with regard to indices $\nu_2, \nu_3, \mu_1, \mu_2, \mu_3$ such that $\nu_2 + \mu_2 = j - 1$, $\nu_2 + \nu_3 = x - i$, and $\mu_1 + \mu_2 + \mu_3 = y - x - 1$.

Proof Introduce new trinomial trials where the first outcome corresponds to the outcome k in the original trials, the second outcome is the outcome ℓ in the original trials, and the remaining outcome is the union of all other outcomes in the original trials.

The event defining the probability p_{xy} can be described in the following way. In the trinomial trials, the trial x results in the first outcome, and the trial y results in the second outcome. In the first $x - 1$ trinomial trials, the first outcome occurs exactly $i - 1$ times, the second outcome occurs, say, ν_2 times, so that the last, third outcome happens $\nu_3 = x - i - \nu_2$ times. Between trials x and y , these outcomes occur μ_1, μ_2 and μ_3 times respectively, while the total number of the second outcome is j .

Our event is the disjoint union of events corresponding to these indices. The multinomial probabilities of these events give (5). ■

A similar formula for the probabilities p_{xy} holds when $j \leq y < x$.

LEMMA 2. With p_{xy} denoting the probabilities in (5), one has for any positive z and v ,

$$\sum_{x < y} \frac{p_{xy} z^{x-i} v^{y-x-1}}{(x-1)!(y-x-1)!} = \frac{p_k^i p_\ell^j (z+v)^{j-1}}{(i-1)!(j-1)!} \exp\{-zp_k + (z+v)(1-p_\ell)\}. \quad (6)$$

Proof According to Lemma 1, the sum in the left-hand side of (6) can be written as

$$\sum \frac{z^{\nu_2+\nu_3} v^{\mu_1+\mu_2+\mu_3}}{(i-1)!\nu_2!\nu_3!\mu_1!\mu_2!\mu_3!} p_k^{i+\mu_1} p_\ell^j (1-p_k-p_\ell)^{\nu_3+\mu_3},$$

where the summation is taken over $\nu_2 + \mu_2 = j - 1$.

By summing first over ν_2 and μ_2 , and then over ν_3, μ_1, μ_3 , one obtains (6). ■

LEMMA 3. For any positive u and s ,

$$\sum_{x < y} p_{xy} u^x s^y = \frac{p_k^i p_\ell^j}{(i-1)!(j-1)!}$$

$$\int \int_{0 < z < w < \infty} \exp \left\{ -z \left[\frac{1}{us} - \frac{1}{s} + p_k \right] - w \left[\frac{1}{s} - 1 + p_\ell \right] \right\} z^{i-1} w^{j-1} dz dw. \quad (7)$$

Proof Multiplying (6) by $z^{i-1} \exp\{-z/t - v/s\}$ and integrating over positive z, v , one obtains after transformation of variables, $w = z + v$,

$$\sum_{x < y} p_{xy} t^x s^{y-x} = \frac{p_k^i p_\ell^j}{(i-1)!(j-1)!}$$

$$\int \int_{0 < z < w < \infty} \exp \left\{ -z(1/t - 1/s + p_k) - w(1/s - 1 + p_\ell) \right\} z^{i-1} w^{j-1} dz dw.$$

Putting $t = us$, gives (7). ■

Proof of Proposition Differentiate (7) with respect to u and s , and put $u = s = 1$ to get

$$\sum_{x < y} x y p_{xy}$$

$$= \frac{p_k^i p_\ell^j}{(i-1)!(j-1)!} \int \int_{0 < z < w < \infty} \exp \left\{ -z p_k - w p_\ell \right\} z^i w^{j-1} (w-1) dz dw.$$

By combining this formula with the one corresponding to $y < x$, one obtains

$$EX_i^{(k)} X_j^{(\ell)} = \frac{p_k^i p_\ell^j}{(i-1)!(j-1)!}$$

$$\times \left[\int \int_{0 < z < w < \infty} z^i w^{j-1} (w-1) \exp \left\{ -z p_k - w p_\ell \right\} dz dw \right.$$

$$\left. + \int \int_{0 < z < w < \infty} z^j w^{i-1} (w-1) \exp \left\{ -z p_\ell - w p_k \right\} dz dw \right].$$

Repeated application of the well known formula

$$\frac{a^j}{(j-1)!} \int_z^\infty e^{-aw} w^{j-1} dw = e^{-az} \sum_{k=0}^{j-1} \frac{(az)^k}{k!}, \quad a > 0,$$

leads to the identity

$$EX_i^{(k)} X_j^{(\ell)} = \frac{ij}{p_k p_\ell} - \sum_{0 \leq m < i, 0 \leq n < j} \binom{m+n}{m} \frac{p_k^{m+1} p_\ell^{n+1}}{(p_k + p_\ell)^{m+n+1}},$$

which is equivalent to (1). Formula (2) easily follows from the definition of $U_i^{(k)}$. ■

REFERENCES

- BLACKWELL, D. AND HODGES, J. L., JR. (1957). Design for the control of selection bias. *Annals of Mathematical Statistics* **28** 449-460.
- BLOM, G., HOLST, L., AND SANDELL, D. (1994). *Problems and Snapshots from the World of Probability*. Springer, New York.
- DAVID, F. N., AND BARTON D. E. (1962). *Combinatorial Chance*. New York, Hafner Publishing Co.
- FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol I*. Wiley, New York.
- HOLST, L. (1989). A note on Banach match box problem. *Statistics & Probability Letters*, **8** 441-443.
- LEVIN, B. (1981). A representation for multinomial distribution function. *Annals of Statistics* **9** 1123-1126.
- RIORDAN, J. (1958). *An Introduction to Combinatorial Analysis*. Wiley, New York.
- ROSENBERGER, W. F. AND J. M. LACHIN (2002). *Randomization in Clinical Trials: Theory and Practice*. Wiley, New York.
- UPPULURI, R. R. AND BLOT, J. (1970). A probability distribution arising in a riff-shuffle. In PATIL, G. P., ED. *Random Counts in Scientific Work*. Pennsylvania State University Press, University Park.