On the 'independence of trials-assumption' in geometric distribution

Mohammad Fraiwan Al-Saleh [a]

[a] Department of Mathematics and Physics, Qatar University, Doha, Qatar

PLEASE SCROLL DOWN FOR ARTICLE

Acknowledgement

The author wishes to thank J.R. Gosselin for his interest in this work.

References

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.
[2] H. Bergmann, *Eine explizite darstellung der bernoullischen zahlen*, Math. Nachr. 34 (1967), pp. 377–378.
[3] K.-W. Chen, *A summation on Bernoulli numbers*, J. Number Theory 111 (2005), pp. 372–391.
[4] K. Dilcher, *Sums of products of Bernoulli numbers*, J. Number Theory 60 (1996), pp. 23–41.
[5] E.R. Hansen, *Tables of Series and Products*, Prentice-Hall, Englewood Cliffs, NJ, 1975.
[6] K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, Springer-Verlag, New York, 1982.
[7] N.E. Nörlund, *Vorlesungen über Differenzenrechnung, in Grundlehren der mathematischen Wissenschaften*, Vol. XIII, Springer, Berlin, 1924.
[8] J. Riordan, *Combinatorial Identities*, Wiley and Sons, New York, 1968.
[9] L. Saalschütz, *Neue formeln für die Bernoullischen zahlen*, J. Reine Angew. Math. 126 (1903), pp. 99–101.
[10] N.J.A. Sloane, A091885 formatted as a simple table, in *The On-Line Encyclopedia of Number Sequences*, Available at www.research.att.com/~njas/sequences (Last updated: 3 July 2007).
[11] N. Gauthier and P.S. Bruckman, *Sums of the even integral powers of the cosecant and secant*, Fibonacci Quart. 44 (2006), pp. 264–273.
[12] K. Knopp, *Theory and Application of Infinite Series*, Dover, New York, 1990.
[13] P.G. Todorov, *On the theory of the Bernoulli polynomials and numbers*, J. Math. Anal. Appl. 104 (1984), pp. 309–350.

On the 'independence of trials-assumption' in geometric distribution

Mohammad Fraiwan Al-Saleh[a]*

Department of Mathematics and Physics, Qatar University, Doha, Qatar

In this note, it is shown through an example that the assumption of the independence of Bernoulli trials in the geometric experiment may unexpectedly not be satisfied. The example can serve as a suitable and useful classroom activity for students in introductory probability courses.

Keywords: Bernoulli experiment; geometric; independence of trials; sampling with replacement

*Email: malsaleh@qu.edu.qa; m-saleh@yu.edu.jo

1. Introduction

A Bernoulli experiment (BE), is a random experiment for which the sample space consists only of two outcomes, success (s) and failure (f). Let $p = \Pr(s)$; $q = 1 - p = \Pr(f)$. Many of the well-known discrete distributions are usually defined based on the repetition of this experiment. Assume that a BE is repeated n times, so that the n trials are independent and p kept fixed for all the trials. If X_1 be the total number of successes, then X_1 is a binomial random variable, denoted by $X_1 \sim B(n, p)$, with probability density function (pdf), $f_1(x) = \binom{n}{x} p^x (1 - p)^{n-x}$, for $x = 0, 1, \ldots, n$ and $0 < p < 1$. If instead, a BE is repeated with the same conditions above (independence of trials and fixed p), until 1 success is obtained and if we let X_2 be the required number of trials, then X_2 is a geometric random variable, denoted by $X_2 \sim G(p)$, with pdf, $f_2(x) = pq^{x-1}$, for $x = 1, 2, \ldots$. Similarly, if X_3 denotes the required number of trials to obtain r successes, then X_3 is a negative binomial random variable, denoted by $X_3 \sim NB(r, p)$, with pdf, $f_3(x) = \binom{x-1}{r-1} p^r q^{x-r}$, for $x = r, r + 1, \ldots$.

It is well known that X_3 has the same distribution as $\sum_{i=1}^{r} Y_i$, where $Y_1, \ldots, Y_r$ are independent and identically distributed (iid) $G(p)$. Thus,

$$X_3 \stackrel{d}{=} \sum_{i=1}^{r} Y_i.$$

As a consequence of this relation, we have

$$E(X_3) = rE(Y_1) = \frac{r}{p} \quad \text{and} \quad \text{Var}(X_3) = \frac{rq}{p}.$$

The two assumptions, the independence of trials and fixed p, are essential to guarantee that the trials are identical (see [1] and [2]). In other words, the same BE is being repeated.

In classroom teaching, teachers usually simplify the above two assumptions to students by saying that the assumptions are satisfied, if we randomly sample with replacement from a population, that consists of items of two kinds with $100p\%$ of one kind (s) and $(1 - p)$ 100% of the other kind (f). Note that in the above-mentioned population, the assumption of fixed p is satisfied whether sampling is with or without replacement, i.e. the $\Pr(s)$ is the same for all trials. However, if sampling is without replacement then, the conditional probability of s at any trial depends on the results of the previous trials, i.e. the trials are dependent. If the population is infinite, then sampling with or without replacement does not make any difference.

The purpose of this article is to provide a case were sampling with replacement may not be enough for the 'independence of trials' assumption. Hence, we may conclude mistakenly that the distribution is binomial, geometric or negative binomial, etc. The example shows that the satisfaction of the independence of trials assumption may depend on the value of the common probability of success of each trial, p.

2. The example

Assume that we are sampling randomly from a population of items of k different types and the proportions of these types are, respectively, $p_1, p_2, \ldots, p_k$, with $\sum_{i=1}^{k} p_k = 1$. Items are selected, at random, one at a time independently (with replacement) and the type of each selected item is noted. Let X be the minimum number of items need to be selected to obtain one item at least of each type. The support of X is the set $\{k, k + 1, \ldots\}$.

Let Y_1 be the total number of items that need to be drawn to obtain any type. Clearly $Y_1 = 1$ with probability 1; Y_1 can also be regarded as a geometric random variable with $p = 1$, i.e. $Y_1 \sim G(1)$. Let Y_2 be the total number of items that need to be drawn to obtain a type different than the first drawn type. In general, for $i = 2, \ldots, k$, let Y_i be the total number of additional items that need to be drawn, after obtaining the $(i-1)$ types, to obtain a type different than the $(i-1)$ previously drawn types. Clearly

$$X \overset{d}{=} \sum_{i=1}^{k} Y_i. \tag{1}$$

We pose the following question: Are the Y_i's geometrically distributed? A first reaction to this question is 'YES'. Each trial is either a success or a failure, where a trial is a success if its outcome is a different type than the previously obtained types.

For simplicity, consider first the simple case of $k = 2$. Then X represents the minimum number of items that need to be taken to obtain at least one of each of the two types, t_1 and t_2, say. Thus, X can take the values $2, 3, 4 \ldots$

$$\Pr(X = x) = \Pr(\overset{x-1 \quad \text{times}}{\overleftarrow{t_1 t_1 \ldots t_1 t_2}}) + \Pr(\overset{x-1 \quad \text{times}}{\overleftarrow{t_2 t_2 \ldots t_2 t_1}})$$

$$= p_1 p_2^{x-1} + p_2 p_1^{x-1}.$$

Thus, the pdf of X is

$$f(x) = \begin{cases} p_1(1 - p_1)^{x-1} + (1 - p_1)p_1^{x-1} & \text{if} \quad x = 2, 3, \ldots \\ 0 & \text{if} \quad \text{otherwise} \end{cases}, \tag{2}$$

which is a convex combination of two truncated geometric's densities. Hence, it can be easily verified that

$$E(X) = \frac{1}{p_1(1 - p_1)} - 1 = \frac{1}{p_1 p_2} - 1. \tag{3}$$

Since, $Y_1 = 1$ with probability 1. Y_2 is the number of trials, excluding the first one, till obtaining a success. The success here is the event of obtaining a type different from the type obtained in the first trial. Thus, the probability of success here is,

$$p = \Pr(s) = \Pr(t_1 t_2 \quad \text{or} \quad t_2 t_1) = 2p_1 p_2. \tag{4}$$

Now, $Y_2 = X - 1$. Thus, the pdf of Y_2, $g(y_2)$, is

$$g(y_2) = \Pr(Y_2 = y_2) = \Pr(X = y_2 + 1).$$

Therefore,

$$g(y_2) = \begin{cases} p_1(1 - p_1)^{y_2} + (1 - p_1)p_1^{y_2} & \text{if} \quad y_2 = 1, 2, \ldots \\ 0 & \text{if} \quad \text{otherwise} \end{cases}, \tag{5}$$

$$E(Y_2) = \frac{1}{p_1(1 - p_2)} - 2. \tag{6}$$

If Y_2 were geometric then, the pdf should have been, for $y_2 = 1, 2, \ldots$

$$g^*(y_2) = p(1 - p)^{y_2 - 1} = 2p_1(1 - p_1)(1 - 2p_1(1 - p_1))^{y_2 - 1}. \tag{7}$$

Thus, for $0 < p_1 < 1$, $g^*(y_2) \neq g(y_2)$ unless $p_1 = 0.5$. Hence, Y_2 has a geometric distribution if and only if $p_1 = 0.5$.

One question that may immediately arise: What went wrong? The probability of success at any trial is fixed at $p = 2p_1p_2$. This means that the problem is with the second assumption: the trials are not independent. To see this, let A_1 and A_2 denotes, respectively, the result of the first and second trial.

$$\Pr(A_1 = s) = \Pr(t_1 t_2 \quad \text{or} \quad t_2 t_1) = 2p_1(1 - p_1);$$

$$\begin{aligned}\Pr(A_2 = s) &= \Pr(t_1 t_1 t_2, t_1 t_2 t_2, t_2 t_2 t_1 \quad \text{or} \quad t_2 t_1 t_1) \\ &= 2p_1^2(1 - p_1) + 2p_1(1 - p_1)^2 \\ &= 2p_1(1 - p_1).\end{aligned}$$

However, the conditional probability that $A_2 = s$ given that $A_1 = s$ is

$$\Pr(A_2 = s | A_1 = s) = \frac{\Pr(t_1 t_2 t_2 \text{ or } t_2 t_1 t_1)}{\Pr(t_1 t_2 \text{ or } t_2 t_1)} = \frac{1}{2}.$$

Hence, $\Pr(A_2 = s | A_1 = s) \neq \Pr(A_2 = s)$, unless $p_1 = (0.5)$, which says that the trials are not independent unless the two types have equal proportions. This is actually very surprising, because we are sampling from the population with replacement. In fact, since the maximum value of $p_1(1 - p_1)$ is 0.25 assumed at $p_1 = (0.5)$, $\Pr(A_2 = s | A_1 = s) > \Pr(A_2 = s)$ for all $p_1 \neq (0.5)$.

The result of this example can be easily generalized for $k > 2$. Thus, $Y_1, Y_2, \ldots, Y_k$ are geometrically distributed if and only if $p_1 = p_2 = \cdots = p_k = (1/k)$.

For example, if a balanced die with six faces is rolled until each face is obtained at least once, then the minimum required number of trials is $X = \sum_{i=1}^{6} Y_i$, where Y_i are independent with $Y_i \sim G((7 - i)/6)$. Thus, the mean and variance of X can be easily obtained. Also, if a balanced coin is tossed until each face is obtained at least once, then again, $X = Y_1 + Y_2$, where $Y_1 \sim G(1)$ and $Y_2 \sim G(1/2)$. These two examples are familiar examples in any introductory course of probability. On the other hand, if individuals are sampled from the population repeatedly until obtaining at least one individual for each of the four blood types, then $X = \sum_{i=1}^{4} Y_i$, but the Y_i are not geometrically distributed. The reason here is that p_i's, which are the proportions of the four blood types are not equal. In other words, the trials to obtain the second blood type after obtaining one of them are not independent, as are the trials for obtaining the third and the fourth blood types.

3. Concluding remarks

It is trivial in many classroom examples to check the conditions on Bernoulli trials so that the distribution is Binomial, geometric, etc. However, examples of the above types require extra care when checking the assumption of the independence of trials.

Acknowledgements

References

[1] S. Ghahramani, *Fundamentals of Probability*, Prentice-Hall Inc., New Jersey, U.S.A, 2000.
[2] R. Hogg and E. Tanis, *Probability and Statistical Inference*, MacMillan Publishing Co., Inc., New York, USA, 1983.

Unification and infinite series

J.V. Leyendekkers[a] and A.G. Shannon[bc]*

[a]The University of Sydney, Faculty of Science, Australia; [b]Master Warrane College, The University of New South Wales, Kensington, NSW, Australia; [c]Provost, RafflesKvB Institute Pty Ltd, North Sydney, NSW, Australia

Some infinite series are analysed on the basis of the hypergeometric function and integer structure and modular rings. The resulting generalized functions are compared with differentiation of the 'mother' series.

Keywords: unification; infinite series; hypergeometric

1. Introduction

Infinite series have intrigued mathematicians for centuries, as has the concept of unification, whereby seemingly unrelated functions can be shown to be particular cases of some general function. A case in point occurred early in the nineteenth century when many known functions were shown to be particular cases of the hypergeometric function [1]. Another unifying approach is to analyse systems using integer structure (IS) and modular rings [2]. Here, we illustrate how some series may be thus interpreted, both for their mathematical interest and pedagogical value, so far as they link a variety of ideas.

2. Hypergeometric functions

Many functions can be represented by infinite series. For example, for $-1 < x \leq 1$,

$$\ln(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots$$
$$= xF(1, 1; 2; -x), \tag{2.1}$$

is a hypergeometric function, defined in general by

$$F(a, b; c; x) = 1 + \frac{ab}{c}x + \frac{a(a+1)b(b+1)}{2c(c+1)}x^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{3!c(c+1)(c+2)}x^3 + \cdots$$
$$= \sum_{n=0}^{\infty} \frac{a^{\overline{n}}b^{\overline{n}}}{c^{\overline{n}}}\frac{x^n}{n!}, \tag{2.2}$$

*Corresponding author. Email: tony@warrane.unsw.edu.au