

Exact theoretical distributions around the replicate results of a germination test

Jean-Louis Laffont¹, Bonnie Hong², Bo-Jein Kuo³ and Kirk M. Remund⁴

Research Paper

Cite this article: Laffont J-L, Hong B, Kuo B-J, Remund KM (2019). Exact theoretical distributions around the replicate results of a germination test. *Seed Science Research* **29**, 64–72. <https://doi.org/10.1017/S0960258519000011>

Received: 11 July 2018
Revised: 21 December 2018
Accepted: 13 January 2019

Keywords:

germination; over-dispersion; partitions; range; tolerances; under-dispersion; variance ratio

Author for correspondence:

Jean-Louis Laffont,
Email: jean-louis.laffont@pioneer.com

¹Pioneer Génétique SARL, 1131 Chemin de l'Enseigne, 31840 Aussonne, France; ²Corteva Agriscience, Agriculture Division of DowDuPont, 7300 NW 62nd Avenue, Johnston, IA 50131, USA; ³Biostatistics Division, Department of Agronomy, National Chung Hsing University, 250 Kuo Kuang Road, Taichung, 40227, Taiwan, ROC and ⁴Bayer Crop Science, 800 North Lindbergh Blvd, St Louis, Missouri 63167, USA

Abstract

Many seed quality tests are conducted by first randomly assigning seeds into replicates of a given size. The replicate results are then used to check whether or not any problems occur in the realization of the test. The two main tools developed for this verification are the ratio of the observed variance of the replicate results to a theoretical variance and the tolerance for the range of the results. In this paper, we derive the theoretical distribution and its related properties of the sequence of numbers of seeds with a given quality attribute present in the replicates. From these theoretical results, we revisit the two quality checking tools widely used for the germination test. We show a precaution to be taken when relying on the variance ratio to check for under- or over-dispersion of the replicate results. This has led to the development of tables providing credible intervals of the variance ratio. The International Seed Testing Association tolerance tables for the range of the results are also compared with tolerances computed from the exact theoretical distribution of the range, leading us to recommend a revision of these tables.

Introduction

Many seed quality tests use k replicates of m seeds. For example, four replicates of 100 seeds are recommended by the International Seed Testing Association (ISTA) for the germination test (ISTA, 2018). The objective of these replicates is to ensure that no particular anomaly such as a seed analyst mistake occurs in the test. This is accomplished by comparing the observed variation among the replicates with the variation due only to the random distribution of the seeds into the replicates. Miles (1963) developed tolerance tables for different tests (purity, germination and other seed count tests), setting limits for the range of the replicate results above which it is considered that the observed variation is not only due to the random allocation of seeds to replicates. These tolerances are based on strong assumptions regarding the nature of the distribution of the number of seeds with the quality attribute in the different replicates. Therefore, for the germination test, Miles considered the binomial distribution in one replicate and the studentized range distribution for the range of the replicate results in the absence of any test issues (examples of test issues include failing to appropriately randomize, abnormal growth chamber conditions, and human errors that may occur during seed germination process or results evaluation process).

Let us assume that of 400 seeds submitted to the germination test, 360 seeds germinated. In the random assignment of these germinated seeds into the four replicates, 100 germinated seeds could have occurred in three replicates and 60 in the remaining one. Alternatively, 90 germinated seeds could have occurred in each of the four replicates. These are just two of the many possible patterns of assigning 360 germinated seeds into 4×100 seed replicates. What is the probability to get a given pattern based solely on the random assignment of the 360 germinated seeds into four 100 seed replicates? The formulation of this probability will lead to the definition of the distribution of the number of seeds with the quality attribute (in this case, ability to germinate under the test conditions) in the different replicates and then, the derivation of the distributions of the statistics related to the replicate results such as the range or the ratio of the observed variance of the replicates to the binomial theoretical variance.

Using the urn model and number theory in this paper, we develop the exact distributions and we discuss the implications for the germination test.

Theoretical results

Consider an urn with n white balls and $N - n$ black balls and let W be the random variable 'number of white balls in a random sample of m balls'. $P(W = w)$ is then given by the

hypergeometric probability mass function (pmf):

$$P(W = w) = \frac{\binom{n}{w} \binom{N-n}{m-w}}{\binom{N}{m}}$$

Now, let Y be the random variable ‘tuple¹ with 1st element the number of white balls in a random sample of m balls out of N balls, 2nd element the number of white balls in a random sample of m balls out of the remaining balls (i.e. $N - m$), ..., i^{th} element the number of white balls in a random sample of m balls out of the remaining balls, ...’. The derivation of $P(Y = (n_1, n_2, \dots, n_i, \dots, n_k)_{\text{ordered}})$ is as follows:

$$\begin{aligned} P(Y = (n_1, n_2, \dots, n_i, \dots, n_k)_{\text{ordered}}) &= \frac{\binom{n}{n_1} \binom{N-n}{m-n_1}}{\binom{N}{m}} \\ &\times \frac{\binom{n-n_1}{n_2} \binom{N-m-n+n_1}{m-n_2}}{\binom{N-m}{m}} \\ &\times \dots \times \frac{\binom{n-\sum_{j=1}^{i-1} n_j}{n_i} \binom{N-(i-1)m-n+\sum_{j=1}^{i-1} n_j}{m-n_i}}{\binom{N-(i-1)m}{m}} \\ &\times \dots \times 1 = \frac{\binom{m}{n_1} \binom{m}{n_2} \dots \binom{m}{n_i} \dots \binom{m}{n_k}}{\binom{N}{n}} \end{aligned}$$

As Y can be viewed as the joint distribution of the random variables Y_i ($i = 1, 2, \dots, k$) ‘number of white balls in box i ’ and noting that the pmf of Y is the pmf of a multivariate hypergeometric distribution with parameters $N, (m, m, \dots, m)$ and n , where $N = km$ and $\sum_{i=1}^k n_i = n$, we have (Bishop *et al.*, 1975):

$$E[Y_i] = \frac{nm}{N}, \text{Var}[Y_i] = \frac{nm}{N} \left(1 - \frac{m}{N}\right) \frac{N-n}{N-1}$$

and

$$\text{Cov}(Y_i, Y_{i'}) = -\frac{nm^2}{N^2} \frac{N-n}{N-1} \quad (i \neq i').$$

Suppose now that the tuple has repeated elements (i.e. some elements have the same number of white balls):

$$\left(\underbrace{n_1, \dots, n_1}_{k_1}, \underbrace{n_2, \dots, n_2}_{k_2}, \dots, \underbrace{n_j, \dots, n_j}_{k_j} \right), \quad 1 \leq j \leq k, \quad \sum_{i=1}^j k_i = k.$$

¹A tuple is an ordered set of elements. When there are two elements, the tuple is called an ordered pair and when there are three elements, it is called a triplet.

Then the number of tuples with the same set of elements in different orders is $\frac{k!}{k_1!k_2!\dots k_j!}$.

Finally, the probability of the unordered set $(n_1, n_2, \dots, n_i, \dots, n_k)_{\text{unordered}}$ is given by:

$$\begin{aligned} P(X = (n_1, n_2, \dots, n_i, \dots, n_k)_{\text{unordered}}) &= \frac{k!}{k_1!k_2!\dots k_j!} \frac{\binom{m}{n_1} \binom{m}{n_2} \dots \binom{m}{n_i} \dots \binom{m}{n_k}}{\binom{N}{n}} \end{aligned}$$

The unordered set $(n_1, n_2, \dots, n_i, \dots, n_k)_{\text{unordered}}$ corresponds to a partition² of n into at most k parts of maximum size m . Closed-form formulas have been developed for easily calculating the number of partitions $p(n, k)$ of n into at most k parts for some small values of k [Andrews (2003) provides formulas of $p(n, k)$ which are easy to compute for $k \leq 9$]. For the number, $u(n, k, m)$, of partitions of n into k parts with maximum size m , some results are provided in Appendix A to facilitate its computation in some situations. When no closed-form formula is available for a particular triplet (n, k, m) , one way to get $u(n, k, m)$ is to enumerate all the partitions of n into at most k parts, to suppress those with $n_i > m$, and to count the remaining ones. Figure 1 provides $u(n, k, m)$ for $m = 100, k = 2, 3, 4$, and for $n = 1, 2, \dots, km$. We can see the symmetry of the plots around $km/2$ for each k and that the number of all the possible values of X increases slowly for $k = 2$ reaching a maximum of 51, whereas it increases rapidly for $k = 4$ reaching a maximum of 29,920. For $m = 50$ and $k = 8$, the maximum for $n = 200$ is very large and is equal to 16,909,449.

The first property related to the probability of the unordered set $(n_1, n_2, \dots, n_i, \dots, n_k)_{\text{unordered}}$ is:

$$\begin{aligned} P(X = (n_1, n_2, \dots, n_i, \dots, n_k)_{\text{unordered}}) &= P(X = (m-n_1, m-n_2, \dots, m-n_i, \dots, m-n_k)_{\text{unordered}}) \end{aligned}$$

The proof of this property is obvious from the combinatorial identity $\binom{n}{k} = \binom{n}{n-k}$.

The second property is the link with the hypergeometric distribution when $k = 2$. We have:

If $n_1 = n_2$, $P(X = (n_1, n_2)_{\text{unordered}}) = p_{n_1}$, else $P(X = (n_1, n_2)_{\text{unordered}}) = 2p_{n_1}$, where p_{n_1} is from the hypergeometric pmf: $p_{n_1} = \frac{\binom{n}{n_1} \binom{N-n}{m-n_1}}{\binom{N}{m}}$. The proof of this property is obvious from the equality $\frac{\binom{m}{n_1} \binom{m}{n-n_1}}{\binom{N}{n}} = \frac{\binom{n}{n_1} \binom{N-n}{m-n_1}}{\binom{N}{m}}$.

We now define the random variable $S_{n_1, n_2, \dots, n_k}^2 = \frac{1}{k-1} \sum_{i=1}^k (Y_i - \frac{n}{k})^2$. Multiple unordered sets can lead to the same realization of $S_{n_1, n_2, \dots, n_k}^2$. For example, for $n = 12, k = 4$ and $m = 10$, the observed variance of the elements of the unordered sets $(6,4,1,1), (6,3,3,0)$ and $(5,5,2,0)$ is the same and is equal to 6. The distribution of $S_{n_1, n_2, \dots, n_k}^2$ can therefore be assessed through

²A partition is a way of writing an integer as a sum of positive integers where the order of the addends is not significant, possibly subject to one or more additional constraints. By convention, partitions are normally written from largest to smallest addends.’ (Weisstein, 2018). For example, $8 = 4 + 2 + 1 + 1$.

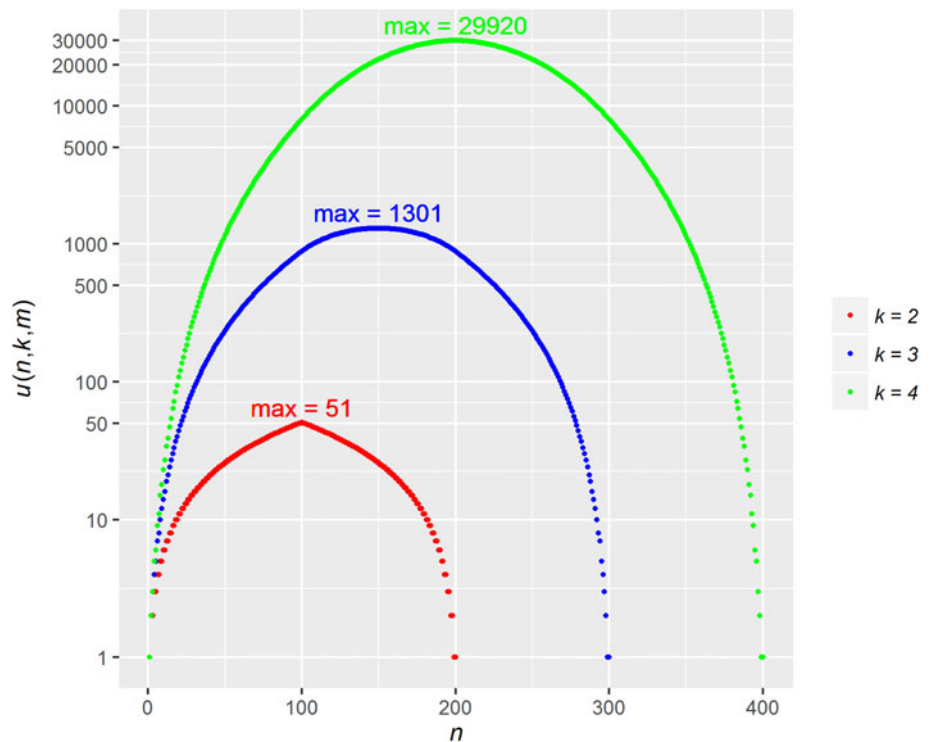


Fig. 1. Graph of $u(n, k, m)$ (\log_{10} scale) vs n for $k = 2, 3, 4$ and $m = 100$ (see text for details). The maximum values of $u(n, k, m)$ are displayed at the peaks of the three curves.

the distribution of X by summing the probabilities of the unordered sets with the same observed variance. **Figure 2** provides example pmf and cumulative distribution function (cdf) of $S_{n_1, n_2, \dots, n_k}^2$ given $n = 12, k = 4$ and $m = 10$.

The third property is:

$$E[S_{n_1, n_2, \dots, n_k}^2] = \frac{nm(N - n)}{N(N - 1)}.$$

The proof of this property is as follows:³

$$\begin{aligned} E[S_{n_1, n_2, \dots, n_k}^2] &= E\left[\frac{1}{k-1} \sum_{i=1}^k \left(Y_i - \frac{n}{k}\right)^2\right] = \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i'=i+1}^k E[(Y_i - Y_{i'})^2] \\ &= \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i'=i+1}^k \text{Var}[Y_i - Y_{i'}] = \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i'=i+1}^k [\text{Var}[Y_i] + \text{Var}[Y_{i'}] - 2\text{Cov}(Y_i, Y_{i'})] \\ &= \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i'=i+1}^k \left[\frac{nm}{N} \left(1 - \frac{m}{N}\right) \frac{N-n}{N-1} + \frac{nm}{N} \left(1 - \frac{m}{N}\right) \frac{N-n}{N-1} + 2 \frac{nm^2}{N^2} \frac{N-n}{N-1} \right] \\ &= \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{i'=i+1}^k \left[2 \frac{nm(N-n)}{N(N-1)} \right] = \frac{nm(N-n)}{N(N-1)}. \end{aligned}$$

Application to the germination test

Germination tests are performed in laboratories to predict the emergence of seedlings in field conditions from seeds sampled from the same lot. ISTA has developed rules for testing germination for a wide range of plant species (ISTA, 2018). The test is usually based on 400 seeds tested in replicates of 100 seeds. Fewer than

400 seeds can be tested (e.g. two replicates of 100 seeds), but not less than 100 in replicates of 25 or 50 seeds.

In the ISTA rules, replicate results are used to assess the reliability of the germination test: the range of the germination results in the replicates is compared with limits developed by Miles (1963). These replicate results are also used to form variance ratios for assessing over-dispersion or under-dispersion (Deplewski *et al.*, 2016). In light of the theoretical results developed in the previous section, new insights regarding these two tools are apparent.

Distribution of the variance ratio

Let k be the number of replicates, m the number of seeds per replicate, N the total number of seeds used in the test ($N = km$) and n_i the number of germinating seeds in replicate i ($\sum_{i=1}^k n_i = n$). The variance ratio f is defined as the ratio of the observed variance between the replicates to the theoretical binomial variance:

$$f = \frac{s^2}{s_B^2}$$

³ $\text{Var}[Y_i - Y_{i'}] = E[(Y_i - Y_{i'}) - E[Y_i - Y_{i'}]]^2$. We have: $E[Y_i - Y_{i'}] = E[Y_i] - E[Y_{i'}] = 0$. Therefore: $\text{Var}[Y_i - Y_{i'}] = E[(Y_i - Y_{i'})^2]$

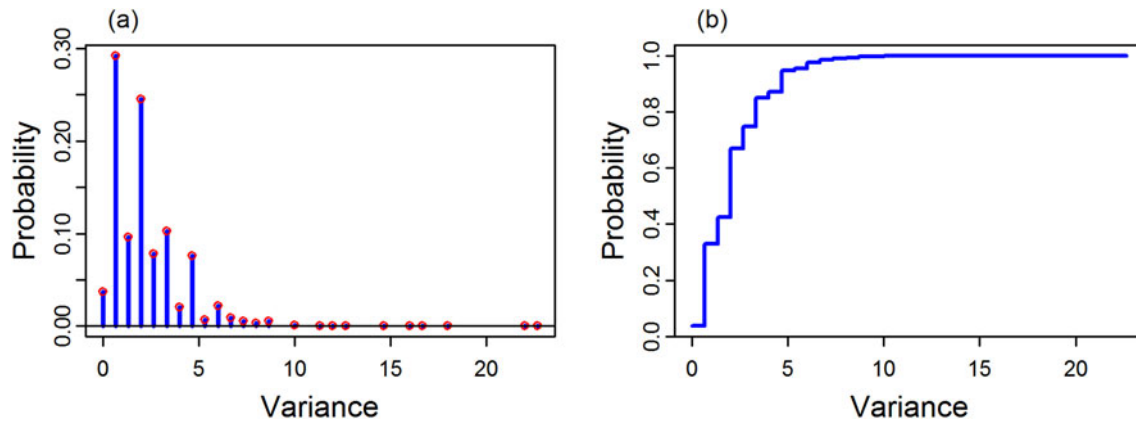


Fig. 2. Graphs of (a) the probability mass function of $S^2_{n_1, n_2, \dots, n_k}$ and (b) the cumulative distribution function of $S^2_{n_1, n_2, \dots, n_k}$ for $n=12, k=4$ and $m=10$.

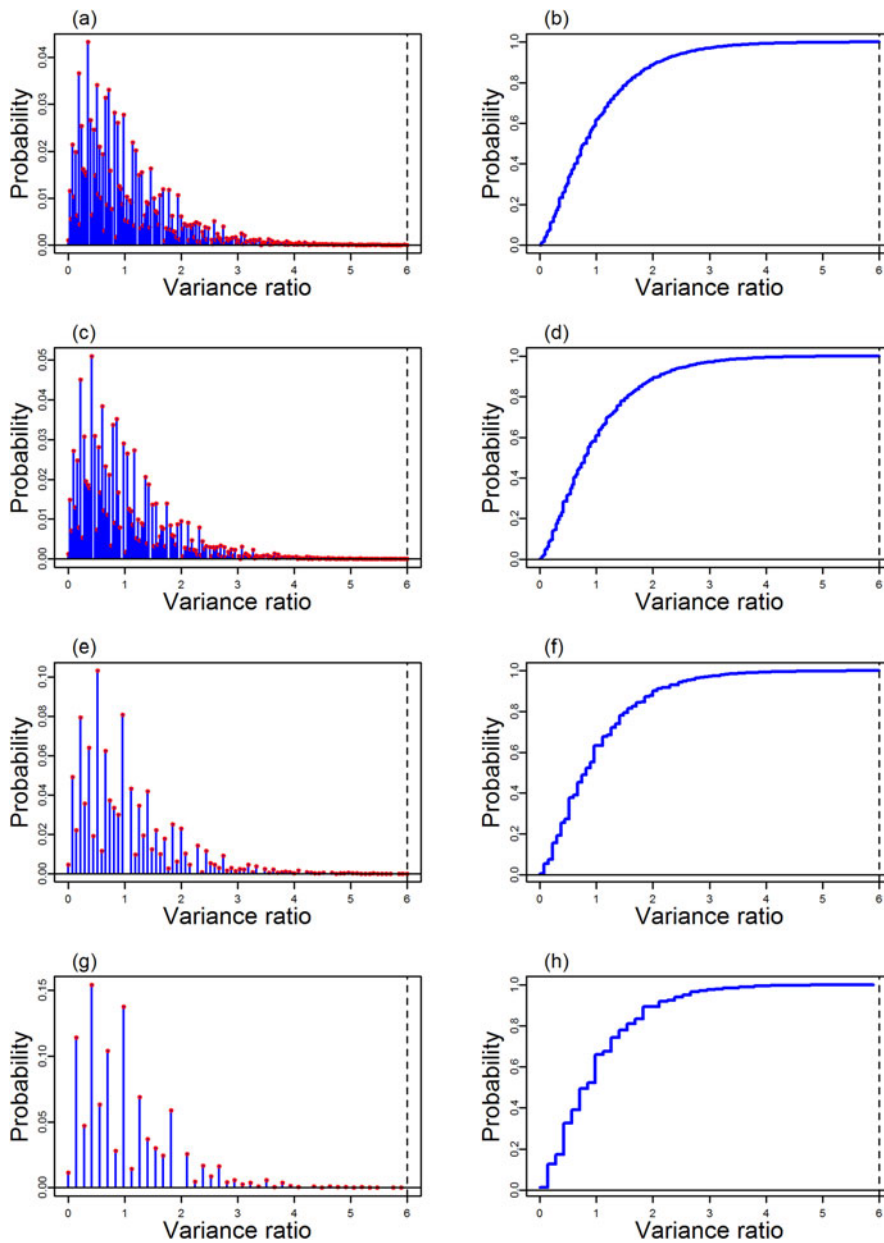


Fig. 3. Graphs of the probability mass function and the cumulative distribution function of F for germination tests performed on four replicates of 100 seeds and for \bar{p} equal to (a, b) 50%, (c, d) 70%, (e, f) 90% and (g, h) 95%. The variance ratios are truncated at 6 for all graphs, with maximums (a, b) 133.33, (c, d) 107.94, (e, f) 44.44 and (g, h) 21.05.

Table 1. Mode, median, 95% credible interval (CI) and 90% CI of the variance ratio for germination tests performed on four replicates of 100 seeds

Germination percentage		Variance ratio			
51–99%	1–50%	Mode	Median	95% CI	90% CI
99	1	0.67	0.48	0–1.98	0–1.89
98	2	0.34	0.74	0–2.72	0.01–2.27
97	3	0.69	0.66	0–2.85	0.03–2.43
96	4	0.52	0.73	0.01–2.97	0.04–2.33
95	5	0.42	0.73	0.02–2.93	0.05–2.53
94	6	0.83	0.76	0.02–3.02	0.05–2.47
93	7	0.72	0.71	0.02–2.98	0.06–2.54
92	8	0.63	0.76	0.03–3.02	0.06–2.47
91	9	0.57	0.75	0.03–2.99	0.07–2.58
90	10	0.52	0.77	0.03–3.10	0.07–2.55
89	11	0.48	0.78	0.03–3.03	0.08–2.51
88	12	0.44	0.77	0.03–3.06	0.09–2.59
87	13	0.41	0.75	0.04–3.07	0.10–2.56
86	14	0.39	0.72	0.04–3.04	0.11–2.53
85	15	0.37	0.74	0.04–3.06	0.11–2.55
84	16	0.35	0.79	0.04–3.04	0.11–2.58
83	17	0.33	0.78	0.04–3.12	0.10–2.57
82	18	0.59	0.78	0.04–3.07	0.10–2.57
81	19	0.56	0.78	0.04–3.12	0.10–2.58
80	20	0.54	0.77	0.05–3.09	0.10–2.55
79	21	0.52	0.76	0.05–3.06	0.10–2.59
78	22	0.51	0.76	0.05–3.12	0.10–2.59
77	23	0.49	0.77	0.06–3.10	0.10–2.57
76	24	0.48	0.76	0.06–3.09	0.10–2.59
75	25	0.46	0.78	0.06–3.08	0.10–2.58
74	26	0.45	0.78	0.06–3.09	0.10–2.58
73	27	0.44	0.78	0.07–3.07	0.10–2.57
72	28	0.43	0.80	0.07–3.07	0.10–2.57
71	29	0.42	0.79	0.07–3.12	0.09–2.61
70	30	0.41	0.78	0.07–3.10	0.09–2.60
69	31	0.41	0.78	0.06–3.09	0.10–2.58
68	32	0.40	0.78	0.06–3.10	0.10–2.59
67	33	0.39	0.79	0.06–3.09	0.10–2.59
66	34	0.39	0.78	0.06–3.07	0.10–2.58
65	35	0.38	0.78	0.06–3.09	0.10–2.59
64	36	0.38	0.77	0.06–3.10	0.10–2.59
63	37	0.37	0.77	0.06–3.09	0.10–2.59
62	38	0.37	0.76	0.06–3.09	0.10–2.57
61	39	0.36	0.76	0.06–3.10	0.10–2.58
60	40	0.36	0.76	0.06–3.12	0.10–2.57
59	41	0.36	0.76	0.06–3.12	0.10–2.58

(Continued)

Table 1. (Continued.)

Germination percentage		Variance ratio			
51–99%	1–50%	Mode	Median	95% CI	90% CI
58	42	0.36	0.76	0.06–3.12	0.10–2.63
57	43	0.35	0.76	0.06–3.11	0.11–2.62
56	44	0.35	0.76	0.06–3.10	0.11–2.61
55	45	0.35	0.76	0.06–3.09	0.11–2.60
54	46	0.35	0.77	0.06–3.10	0.11–2.60
53	47	0.35	0.77	0.06–3.11	0.11–2.59
51-52	48-50	0.35	0.77	0.06–3.10	0.11–2.59

Table 2. Mode, median, 95% credible interval (CI) and 90% CI of the variance ratio for germination tests performed on four replicates of 50 seeds

Germination percentage		Variance ratio			
50–99%	1–50%	Mode	Median	95% CI	90% CI
99	1	0.67	0.67	0–1.88	0–1.75
98	2	0.68	0.48	0–2.00	0–1.91
97	3	1.15	0.71	0–2.42	0–2.14
96	4	0.35	0.74	0–2.66	0.01–2.30
95	5	0.70	0.61	0–2.76	0–2.32
94	6	0.71	0.67	0–2.90	0.03–2.42
93	7	0.51	0.64	0–2.88	0–2.47
92	8	0.54	0.73	0.01–3.00	0.04–2.34
91	9	0.41	0.71	0–2.93	0–2.35
90	10	0.44	0.72	0.02–2.93	0.05–2.52
89	11	0.34	0.71	0–2.97	0–2.48
88	12	0.38	0.78	0.02–3.07	0.05–2.47
87	13	0.29	0.71	0–3.01	0.06–2.52
86	14	0.78	0.73	0.02–2.97	0.05–2.53
85	15	0.26	0.72	0–3.05	0.07–2.53
84	16	0.69	0.70	0.02–3.02	0.06–2.58
83	17	0.24	0.74	0–3.00	0.08–2.49
82	18	0.63	0.76	0.03–3.04	0.06–2.48
81	19	0.22	0.75	0–2.98	0.08–2.55
80	20	0.58	0.75	0.03–3.04	0.06–2.54
79	21	0.20	0.76	0–3.06	0.09–2.52
78	22	0.54	0.76	0.03–3.05	0.07–2.56
77	23	0.64	0.74	0–3.05	0.09–2.55
76	24	0.51	0.77	0.03–3.09	0.07–2.57
75	25	0.60	0.73	0.04–3.04	0.10–2.58
74	26	0.49	0.78	0.03–3.07	0.07–2.54
73	27	0.58	0.72	0.04–3.02	0.10–2.57
72	28	0.46	0.79	0.03–3.03	0.08–2.55

(Continued)

Table 2. (Continued.)

Germination percentage		Variance ratio			
50–99%	1–50%	Mode	Median	95% CI	90% CI
71	29	0.55	0.76	0.04–3.11	0.10–2.55
70	30	0.44	0.78	0.03–3.07	0.09–2.60
69	31	0.53	0.79	0.04–3.10	0.10–2.56
68	32	0.43	0.76	0.03–3.08	0.09–2.60
67	33	0.51	0.79	0.04–3.04	0.10–2.56
66	34	0.42	0.75	0.04–3.09	0.10–2.56
65	35	0.50	0.79	0.05–3.05	0.10–2.55
64	36	0.41	0.74	0.04–3.10	0.10–2.57
63	37	0.49	0.78	0.05–3.05	0.10–2.54
62	38	0.40	0.73	0.04–3.08	0.10–2.54
61	39	0.48	0.77	0.05–3.05	0.09–2.55
60	40	0.39	0.72	0.04–3.05	0.11–2.55
59	41	0.47	0.77	0.05–3.07	0.09–2.56
58	42	0.38	0.73	0.04–3.07	0.11–2.56
57	43	0.46	0.76	0.05–3.08	0.09–2.61
56	44	0.38	0.73	0.04–3.08	0.11–2.59
55	45	0.46	0.76	0.05–3.10	0.09–2.63
54	46	0.38	0.74	0.04–3.08	0.11–2.59
53	47	0.45	0.75	0.05–3.09	0.09–2.63
52	48	0.37	0.74	0.04–3.08	0.11–2.58
51	49	0.45	0.75	0.05–3.09	0.09–2.63
50	50	0.37	0.74	0.04–3.08	0.11–2.58

where $s^2 = [1/(k - 1)] \sum_{i=1}^k (p_i - \bar{p})^2$, $s_B^2 = \bar{p}(1 - \bar{p})/m$, $p_i = n_i/m$ and $\bar{p} = (\sum_{i=1}^k p_i)/k = n/N$.

Considering the variance from a binomial distribution with parameter ‘number of trials’ equal to m in the denominator of f is justified by the fact that the distribution of the number of germinating seeds in a random sub-sample from a representative sample from a seed lot is also binomial with probability parameter equal to the proportion of germinating seeds in the lot (see theorem in Appendix B). We also note that f is the dispersion factor of a simple binomial generalized linear model (McCullagh and Nelder, 1989): $n_i \sim \text{Binomial}(m, \pi)$, $\text{logit}(\pi) = \mu$.

We are interested in the distribution of the discrete random variable F associated with f . It is easily derived from the distribution of $S_{n_1, n_2, \dots, n_k}^2$, as s_B^2 is a constant for a given number of n germinating seeds out of N and $s^2 = \frac{1}{m^2(k - 1)} \sum_{i=1}^k (n_i - n/k)^2$.

Figure 3 provides the pmf and cdf for germination tests performed using four replicates of 100 seeds and for 50, 70, 90 and 95% germinating seeds. These distributions are highly skewed to the right and have multiple peaks. Extreme values can be very large but with a very low probability.

From property 3, the expectation of the variance ratio is:

$$E[F] = \frac{mN^2}{n(N - n)m^2} E[S_{n_1, n_2, \dots, n_k}^2] = \frac{N}{N - 1} .$$

For a germination test involving 400 seeds, the expectation is therefore equal to 1.002506. Given that the distribution of F is highly skewed, a more appropriate measure of central tendency than the mean is the mode or the median, with the latter being preferable because of the presence of numerous peaks in the pmf. For germination percentages of tests involving 400 seeds, Table 1 provides these measures (which are identical for complementary percentages due to property 1) along with 95 and 90% credible intervals⁴ (CI) computed using linear interpolation as F is discrete. We can see in Table 1 that the central tendency for the variance ratio is much lower than the mean: the maximum of the medians for all germination percentages is not greater than 0.80, which provides the theoretical explanation for the mean of the variance ratios being 0.836 as reported by Deplewski *et al.* (2016) by averaging variance ratios observed in 51,581 germination tests. Finding a f value around 0.8 should not therefore be an indication of a particular issue with the test, and the CIs provided in Table 1 should help in deciding whether there is really an issue related to under- or over-dispersion. Table 2 is the counterpart of Table 1 for a germination test involving four replicates of 50 seeds.

Distribution of the range

The distribution of the range of replicate germination results is derived from the distribution of X , similarly to the distribution of the variance ratio. The pmf and the cdf of the range for germination tests performed on four replicates of 100 seeds and for \bar{p} equal to 50, 70, 90 and 95% are provided in Fig. 4. These distributions are skewed to the right and are more regular than the distributions of F .

The ISTA tables used to ensure reliability of a germination test from the range of the germination results in the replicates (table 5B, parts 1 to 3 in ISTA, 2018) were developed by Miles (1963) as follows: calculation of $s_{Miles} = q_{1-\alpha; k; \infty} \sqrt{p(1 - p)/m}$ where $p = \bar{p} - 0.005$ and $q_{1-\alpha; k; \infty}$ is the $(1 - \alpha)$ quantile of a studentized range distribution for k groups and infinite degrees of freedom; the tolerated range is then taken as the next larger whole number than s_{Miles} when the fraction in s_{Miles} is greater than or equal to 0.8, the next smaller whole number otherwise. In ISTA tables, α is equal to 0.025.

This calculation of the tolerated range assumes that the replicate results are normally distributed with a binomial variance. We can now use the theoretical cdf of the range to compute the $(1 - \alpha)$ quantile of the range distribution. The tolerances from ISTA table 5B and the 0.975 quantiles of the range are visualized in Fig. 5. We can see that the ISTA tolerances are conservative (red points are below blue points for a given germination average, with very few exceptions for tests with four replicates of 100 seeds), especially when the number of seeds involved in the test is low. For tests performed with four replicates of 100 seeds, the approximation due to the normality and the binomial assumptions is quite good: the maximum difference is equal to 2.05 (for a germination percentage equal to 99% or 1%), the other differences being below 1.15. For tests performed with a lower num-

⁴A credible interval is a range of values within which an unobserved parameter value falls with a given probability. In practice, credible intervals are used in the same way as confidence intervals. In concept, they are different because bounds of credible intervals are fixed and the parameter of interest is random.

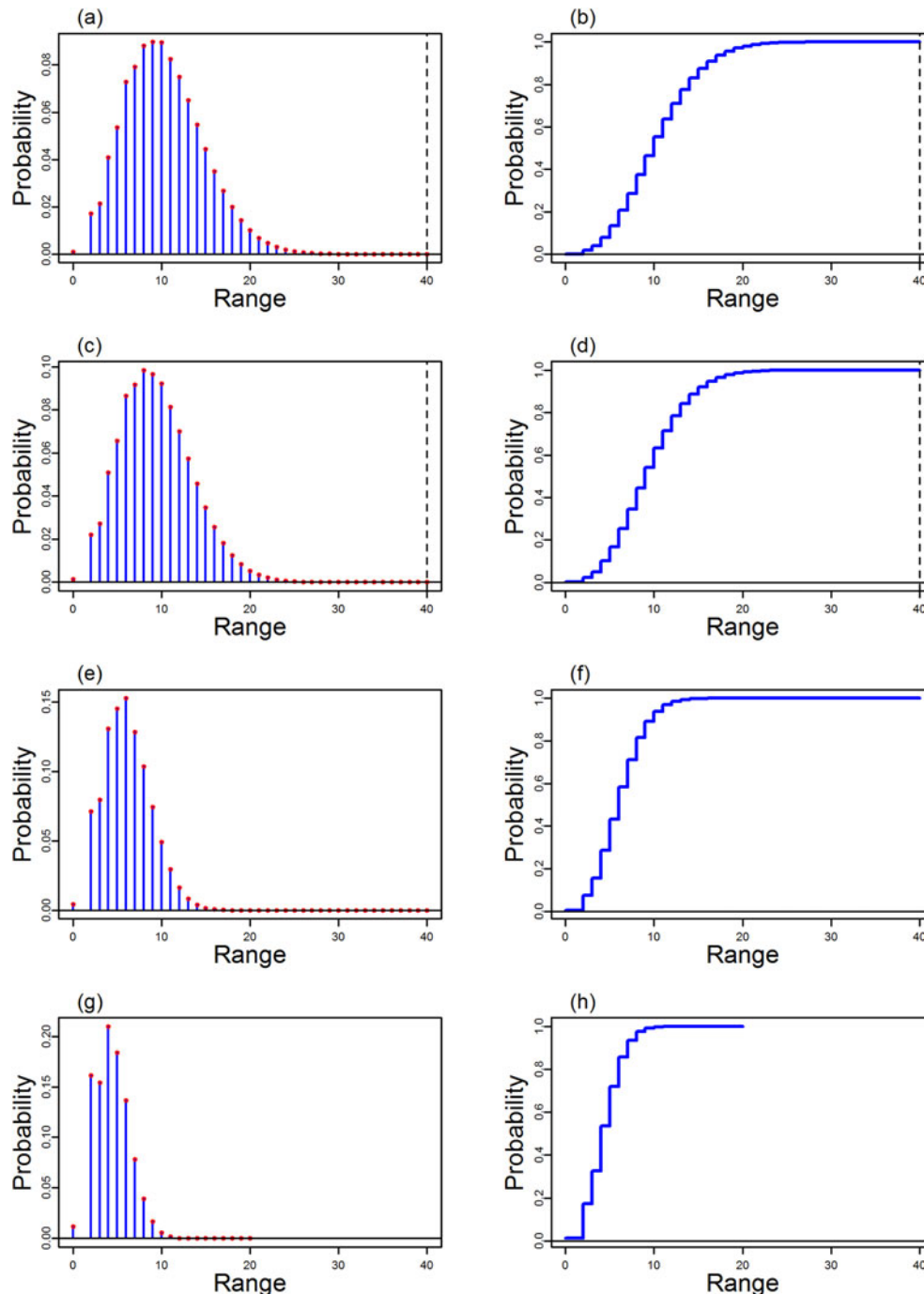


Fig. 4. Graphs of the probability mass function and the cumulative distribution function of the range for germination tests performed on four replicates of 100 seeds and for \bar{p} equal to (a, b) 50%, (c, d) 70%, (e, f) 90% and (g, h) 95%. (a, b, c, d) Range truncated at 40, maximum would be 100.

ber of replicates and a lower number of seeds per replicate, the approximation is inadequate.

Conclusion

The theoretical results we have obtained using number theory and, more specifically, the theory of partitions, are of importance in practical terms. We have focused on applications around the germination test. The methodology can also be applied to other seed tests as long as the results are binary, for example, purity tests.

We have thus developed the theoretical distribution of the variance ratio when the only source of variation is the random assignment of the seeds in the replicates. This enabled us to prove that the variance ratio, in the absence of any analytical problems, is likely to be well below unity. We have also been able to construct credibility intervals for the variance ratio which could be used in the area of the validation of new germination test methods using collaborative studies.

Another theoretical distribution we have derived is the distribution of the range of the germination results in the replicates,

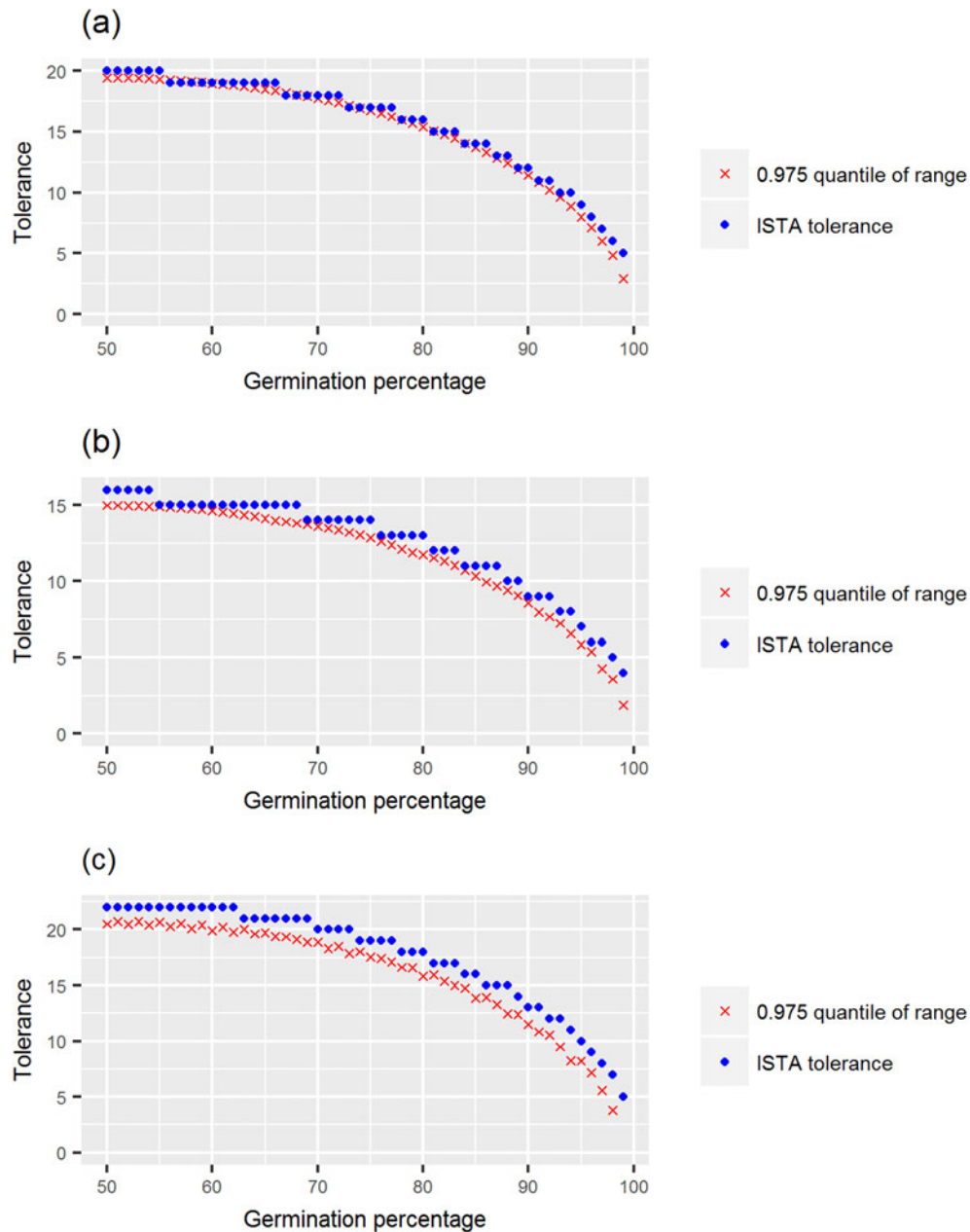


Fig. 5. Tolerances from ISTA (2018) table 5B (blue points) and 0.975 quantiles of the range (red points) for germination tests performed on (a) four replicates of 100 seeds, (b) two replicates of 100 seeds and (c) two replicates of 50 seeds.

here again in the absence of any analytical problems. This has allowed us to compare tolerances derived from a given quantile of this theoretical distribution with the tolerances derived by Miles (1963) using a normal and binomial approximation. Miles' tolerances which are used in ISTA table 5B are shown to be conservative (i.e. the risk of falsely rejecting valid germination tests is below the nominal risk of 2.5%). The tolerances in table 5B part 1 for tests performed on four replicates of 100 seeds are very close to the exact tolerances. However, the differences between ISTA tolerances and exact tolerances are in the order of 1% germination tolerance difference for the tests performed on two replicates of 100 seeds and in the order of 2% germination tolerance difference for the tests performed on two replicates of 50 seeds. We therefore recommend an

adjustment of the ISTA table 5B based on the exact theoretical distribution of the range.

This work, which confirms some practices (i.e. use of ISTA table 5B part 1 to evaluate the germination range of four replicates of 100 seeds) and dispels some myths (i.e. why dispersion factor is often less than 1) could be used in other areas, for example in group testing (Dorfman, 1943) for finite populations.

The computations in this paper have been performed using R (R Core Team, 2012) and in particular using the R packages *partitions* (Hankin, 2006) and *ggplot2* (Wickham, 2009).

Acknowledgements. The authors would like to thank the referees and the editor for detailed comments and corrections leading to a significant improvement of the paper.

References

- Andrews G.E. (2003) Partitions: at the interface of q -series and modular forms. *The Ramanujan Journal* 7, 385–400.
- Bishop Y.M.M., Fienberg S.E. and Holland P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. Reprinted (2007), New York, Springer.
- Deplewski P., Kruse M. and Piepho H.-P. (2016) Underdispersion of replicate results in germination tests is species and laboratory specific. *Seed Science and Technology* 44, 1–17.
- Dorfman R. (1943) The detection of defective members of large populations. *Annals of Mathematical Statistics* 14, 436–440.
- Hankin R.K.S. (2006) Additive integer partitions in R. *Journal of Statistical Software, Code Snippets* 16, 1–3.
- ISTA (2018) International Rules for Seed Testing, Chapter 5: The Germination Test. International Seed Testing Association, Bassersdorf, Switzerland.
- McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models*. London: Chapman and Hall.
- Miles S.R. (1963) *Handbook of Tolerances and Measures of Precision for Seed Testing*. Proceedings of the International Seed Testing Association 28/3.
- R Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>
- Weisstein E.W. (2018) Partition. *MathWorld – A Wolfram Web Resource*. Available at: <http://mathworld.wolfram.com/Partition.html> (accessed 16 May 2018).
- Wickham H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Appendix A

Let $p(n, k)$ be the number of partitions of n into at most k parts and let $u(n, k, m)$ be the number of partitions of n into k parts with maximum size m .

Proposition: $u(n, k, m) = u(km - n, k, m)$.

Proof: Let $n = n_1 + n_2 + \dots + n_k$ be a partition of n into k parts with maximum size m . Then, $km - n = (m - n_k) + (m - n_{k-1}) + \dots + (m - n_1)$ is a partition of $km - n$ into k parts with maximum size m . We have a one-to-one correspondence of the partitions of n and the partitions of $km - n$ so $u(n, k, m) = u(km - n, k, m)$.

Formulas for $u(n, k, m)$ when $n \leq m$: If $n \leq m$, $u(n, k, m) = p(n, k)$. Then we can use closed-form formulas available for $p(n, k)$ for some k 's. For example (Andrews, 2003): $p(n, 2) = \lfloor (n+2)/2 \rfloor$, $p(n, 3) = \lfloor (n+3)^2/12 \rfloor$, $p(n, 4) = \lfloor (n+5)(n^2 + n + 22 + 18 \lfloor n/2 \rfloor) / 144 \rfloor$ where $\lfloor \cdot \rfloor$ represents the floor function and $\lceil \cdot \rceil$ the nearest integer function.

Formula for $u(n, 3, m)$ when $m < n < 2m$:

$$u(n, 3, m) = \left\lceil \frac{(n+3)^2}{12} \right\rceil - \left\lceil \frac{(n-m)(n-m+2)}{4} \right\rceil$$

where $\lceil \cdot \rceil$ represents the ceiling function.

Proof: We consider the partitions of n into almost three parts for which $n > m$. The number S of such partitions is equal to the sum of the number of partitions of $i = 0, 1, 2, \dots, (n - m - 1)$ into almost two parts: $S = \sum_{i=0}^{n-m-1} \lfloor (i+2)/2 \rfloor$. If $(n - m - 1)$ is odd, the sequence numbers in S is $(1, 1, 2, 2, \dots, (n - m)/2, (n - m)/2)$. Then $S = 2 \lfloor (n - m)/2 \rfloor \lfloor (n - m)/2 + 1 \rfloor / 2 = (n - m)(n - m + 2)/4 = \lfloor (n - m)(n - m + 2)/4 \rfloor$. If $(n - m - 1)$ is even, the sequence numbers in S is $(1, 1, 2, 2, \dots, (n - m - 1)/2, (n - m -$

$1)/2, (n - m + 1)/2)$. Then $S = 2 \lfloor (n - m - 1)/2 \rfloor \lfloor (n - m - 1)/2 + 1 \rfloor / 2 + (n - m + 1)/2 = (n - m)(n - m + 2)/4 + 1/4$. As $(n - m)$ and $(n - m + 2)$ are two consecutive odd integers, $\text{frac}[(n - m)(n - m + 2)/4] = 3/4$ and as $\text{frac}(x) = x - \lfloor x \rfloor + 1$ ($x > 0$), then $S = \lfloor (n - m)(n - m + 2)/4 \rfloor$. Finally, $\mu(n, 3, m) = p(n, 3) - S = \lfloor (n + 3)^2/12 \rfloor - \lfloor (n - m)(n - m + 2)/4 \rfloor$.

Appendix B

Theorem: If the distribution of the number of successes in a primary sample of size N is Binomial(N, π), then the distribution of the number of successes in a subsample (from the primary sample) of size n ($n < N$) is also Binomial(n, π).

Proof: Consider a very large number of white and black balls with a proportion π of white balls. Let Y be the random variable 'number of white balls in a random sample of N balls'. $P(Y = k)$ is then given by the binomial probability:

$$P(Y = k) = \binom{N}{k} \pi^k (1 - \pi)^{N-k} = f(k).$$

Now let X_i be the random variable 'number of white balls in a random subsample of n balls from the N balls previously sampled'. The conditional probability $P(X_i = k_i | Y = k)$ is then given by the hypergeometric probability:

$$P(X_i = k_i | Y = k) = \frac{\binom{k}{k_i} \binom{N-k}{n-k_i}}{\binom{N}{n}} = g(k_i | k).$$

We now derive the marginal distribution of X_i , $g(k_i)$:

$$\begin{aligned} g(k_i) &= \sum_{k=0}^N g(k_i | k) f(k) \\ &= \sum_{k=0}^N \left[\frac{\binom{k}{k_i} \binom{N-k}{n-k_i}}{\binom{N}{n}} \binom{N}{k} \pi^k (1 - \pi)^{N-k} \right] \\ &= \sum_{k=0}^N \left[\binom{n}{k_i} \binom{N-n}{k-k_i} \pi^k (1 - \pi)^{N-k} \right] \\ &= \binom{n}{k_i} \sum_{k=0}^N \left[\binom{N-n}{k-k_i} \pi^k (1 - \pi)^{N-k} \right]. \end{aligned}$$

Noting that $\binom{N-n}{k-k_i} = 0$ for $k < k_i$ or for $k > N - n + k_i$, we have:

$$\begin{aligned} g(k_i) &= \binom{n}{k_i} \sum_{k=k_i}^{N-n+k_i} \left[\binom{N-n}{k-k_i} \pi^k (1 - \pi)^{N-k} \right] \\ &= \binom{n}{k_i} \pi^{k_i} (1 - \pi)^{n-k_i} \sum_{k=k_i}^{N-n+k_i} \left[\binom{N-n}{k-k_i} \pi^{k-k_i} (1 - \pi)^{N-k-k_i} \right]. \end{aligned}$$

Now, substituting $k - k_i$ by j in the above sum and using the binomial theorem:

$$\begin{aligned} g(k_i) &= \binom{n}{k_i} \pi^{k_i} (1 - \pi)^{n-k_i} \sum_{j=0}^{N-n} \left[\binom{N-n}{j} \pi^j (1 - \pi)^{N-n-j} \right] \\ &= \binom{n}{k_i} \pi^{k_i} (1 - \pi)^{n-k_i}. \end{aligned}$$

The distribution of X_i is therefore a binomial distribution with parameters n and π which proves the theorem.