Seed Science Research

cambridge.org/ssr

On testing for seed sample heterogeneity with the exact probability distribution of the germination count range

Anderson Rodrigo da Silva

Statistics and Geoprocessing Lab., Instituto Federal Goiano, Rod. Geraldo S. Nascimento, km 2.5, Urutaí CEP 75790-000, GO, Brazil

Technical Update

Cite this article: da Silva AR (2020). On testing for seed sample heterogeneity with the exact probability distribution of the germination count range. *Seed Science Research* **30**, 59–63. https://doi.org/10.1017/S0960258520000112

Received: 30 December 2019
Accepted: 7 April 2020
First published online: 30 April 2020

Key words:
germination test; hypergeometric distribution; normal seedlings; *R*-value; *S*-value

Author for correspondence:
Anderson Rodrigo da Silva,
E-mail: anderson.silva@ifgoiano.edu.br

Abstract

Seed lot heterogeneity is often evaluated through the range between germination percentages of four seed samples, considering normal and binomial approximations for calculating the tolerated range (S). In this paper, an exact test for the germination count range (R) is derived based on the hypergeometric and the binomial probability model for germination count. Through Monte Carlo simulations, the empirical distribution of R is built to evaluate the quantiles of the exact distributions. Moreover, a power analysis is performed by simulation. Sample size and germination rate effects are evaluated. In lots with a high germination rate, the proposed test based on the hypergeometric model is about 20% more powerful than the test based on the S-value. A table containing the critical values is presented and recommended to be used in *off*-range heterogeneity testing.

Introduction

A seed lot is characterized by a set of variables such as the number of pure seeds, normal and abnormal seedlings, the number of dead and dormant seeds and the number of seeds damaged by insects. In seed analysis, the standard procedure for germination testing is to use four samples (replicates) of 100 seeds each, as recommended by the International Seed Testing Association (ISTA, 2017). In order to assure the germination test reliability, a seed lot is expected to have an acceptable level of heterogeneity, which is evaluated through the *in-range* heterogeneity test with the H-value and the *off-range* heterogeneity test with the R-value.

According to Piepho et al. (2018), it is important to measure and quantify that variation between seed samples because, if the four replicates results would vary significantly more than expected, this would indicate that something went wrong with the germination test, for example that the seeds in one sample died but not in the others, and the test would have to be repeated.

The test for the *off-range* heterogeneity between seed samples consists of evaluating the maximum difference between germination percentages and to compare it with a tolerated value (S), calculated considering the theoretical variance of the binomial distribution and a critical quantile of the studentized range (q), as proposed by Miles (1963). In a formal way, consider $p_1, p_2, \ldots, p_m$ as realizations of the germination percentage of m independent samples containing n seeds each. Then, compare $R = \max(p_i) - \min(p_i)$ to $S = q\sqrt{n^{-1}\bar{p}(1-\bar{p})}$, where $\bar{p} = m^{-1}\sum_{i=1}^{m} p_i$. When $R \geq S$, the samples are considered heterogeneous and further sampling should be done. Note that this approach requires assuming that all the p_i are independent and identically distributed as Normal variables with mean np and variance $np(1-p)$, at least approximately.

In a germination test, seed samples of similar size (n) are drawn from the seed lot without replacement, and the number of normal seedlings is computed. In this case, the theoretical probability distribution is not binomial(n, p), but hypergeometric(N, K, n), where K is the number of normal seedlings of the seed lot containing N seeds. This result was previously identified (Piepho et al., 2018; Laffont et al., 2019). When searching for genetically modified events in seed lots, a similar test procedure is adopted. According to Herman and Robbins (2013), for large seed lots, a binomial distribution is typically assumed, but for seed lots for which the tested sample is a substantial proportion of the overall seed lot, a hypergeometric distribution is typically assumed.

In this paper, a test based on the exact probability distribution of the germination count range is presented and evaluated by Monte Carlo simulation.

Materials and methods

The exact test

Consider a seed lot of size N from which K seeds form normal seedlings. In a germination test, m samples of size n each are drawn from that seed lot without replacement, generating the

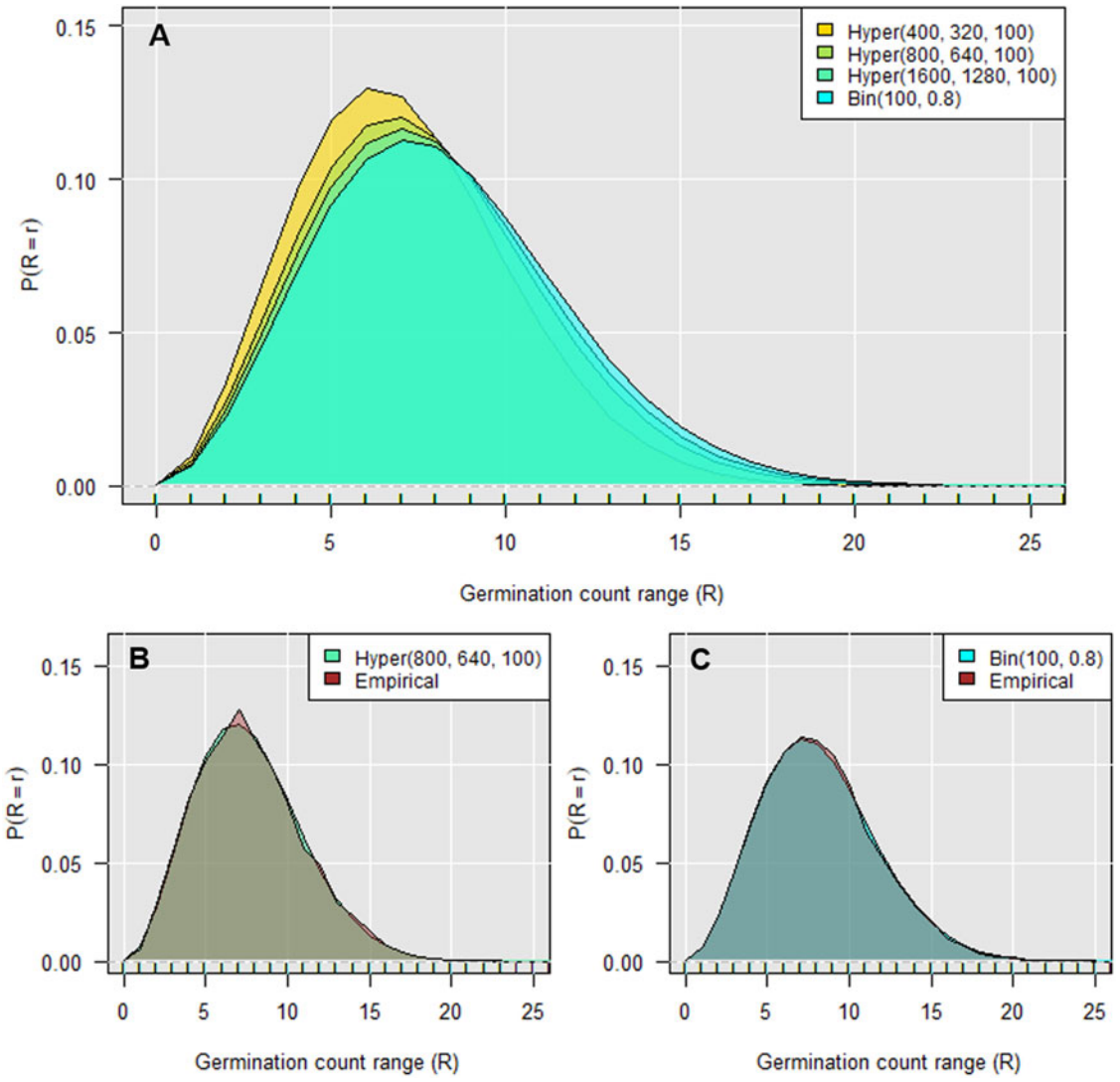

Fig. 1. (a) Exact probability distributions of the germination count range (R) built under hypergeometric(N, K, n) and binomial(n, p) models for germination count, based on four samples. Mark ticks on the x-axis indicate the discrete values of R. (b) Empirical probability distribution of the germination count range considering the hypergeometric distribution for germination count. (c) Empirical probability distribution of the germination count range considering the binomial distribution for germination count.

random variables $X_1, X_2, \ldots, X_m$ that represent the number of normal seedlings (germination count). Let us assume that all the X_i are independent and identically distributed according to the hypergeometric model with parameters N, K and n. Now take the order statistics $X_{(1)} = \min(X_1, X_2, \ldots, X_m)$ and $X_{(m)} = \max(X_1, X_2, \ldots, X_m)$ as random variables with distribution functions $F_{X_{(1)}}$ and $F_{X_{(m)}}$, respectively. Let us define the variable $R = X_{(m)} - X_{(1)}$ as the range of germination count for the m samples being evaluated. Under the null hypothesis that $X_{(1)}$ and $X_{(m)}$ share the same distribution parameters (N, K, n), the exact probability distribution function of R can be derived (Arnold et al., 2008), as follows:

$$\begin{aligned}&\mathbb{P}_R(R=0|N,K,n)\\&=\sum_{x=0}^{n}\mathbb{P}_{X_{(m)}X_{(1)}}\left(X_{(m)}=x,X_{(1)}=x\right)\\&=\sum_{x=0}^{n}[\mathbb{P}_X(X=x)]^m\end{aligned}\tag{1}$$

and

$$\begin{aligned}&\mathbb{P}_R(R=r|N,K,n)\\&=\sum_{x=0}^{n}\mathbb{P}_{X_{(m)}X_{(1)}}\left(X_{(m)}=x+r,X_{(1)}=x\right)\\&=\sum_{x=0}^{n}\left.\begin{matrix}[\mathbb{P}_X(X\le x+r)-\mathbb{P}_X(X\le x-1)]^m\\-[\mathbb{P}_X(X\le x+r)-\mathbb{P}_X(X\le x)]^m\\-[\mathbb{P}_X(X\le x+r-1)-\mathbb{P}_X(X\le x-1)]^m\\+[\mathbb{P}_X(X\le x+r-1)-\mathbb{P}_X(X\le x)]^m\end{matrix}\right\}I_R\quad(R=1,2,\ldots,n)\end{aligned}\tag{2}$$

where $I_R(\cdot)$ is an indicator function and

$$\mathbb{P}_X(X=x|N,K,n)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}\tag{3}$$

https://doi.org/10.1017/S0960258520000112 Published online by Cambridge University Press

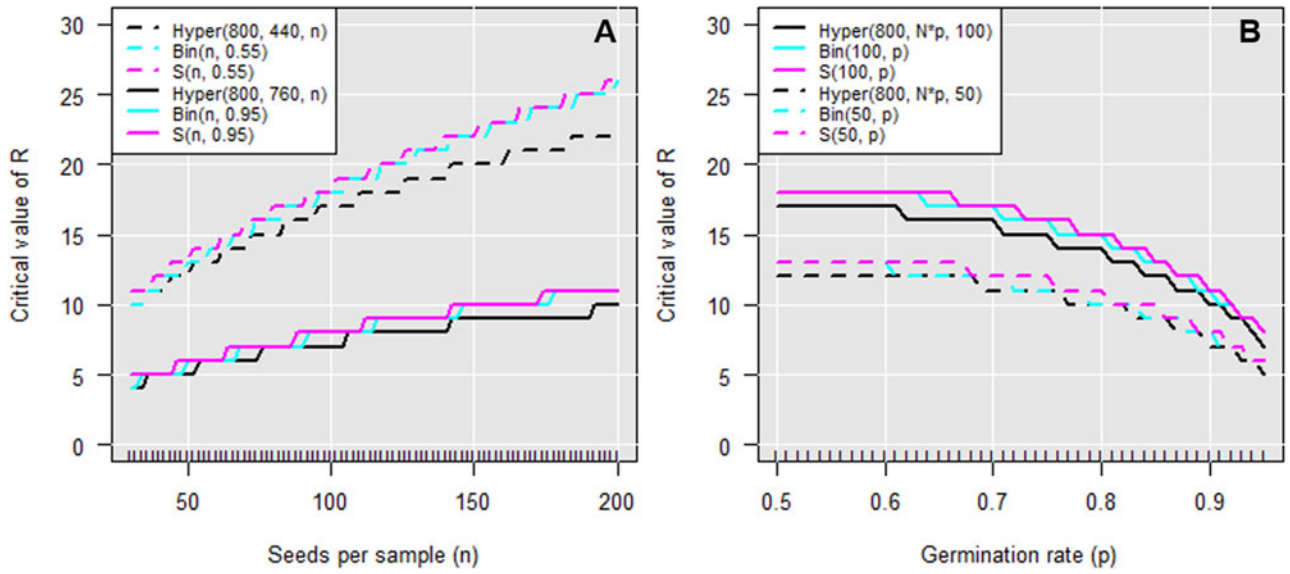

Fig. 2. Critical values (5% significance) of the germination count range between four seed samples calculated using the exact probability distributions based on the hypergeometric(N, K, n) and binomial(n, p) models, and the S-values (ISTA, 2017). Variations according to (a) the sample size (n) and (b) the germination rate (p).

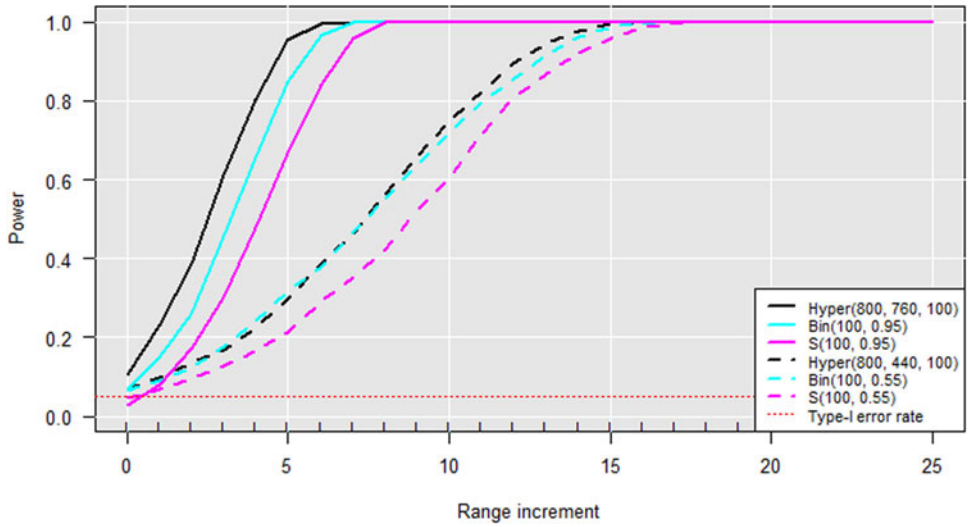

Fig. 3. Power analysis of the tests for the germination count range based on the exact probability distributions derived from the hypergeometric and binomial models, and assuming the Normal distribution for calculating the S-value.

Expectation and variance of the range are given in Supplementary Appendix A, while Supplementary Appendix B gives the codes in R for the probability mass function and the cumulative distribution.

The exact test consists of calculating the one-sided P-value for the realization $r = x_{(m)} - x_{(1)}$ as $\mathbb{P}_R(R > r) = 1 - \sum_{i=0}^{r} \mathbb{P}_R(R = i)$. In this sense, if the P-value does not exceed the nominal level of significance α, the seed samples are considered *off-range* heterogeneous.

Evaluation by simulation and computing

Once there is a relation between the hypergeometric and the binomial models, the distribution of R is built considering the binomial probability mass function for the random variable X, here defined as the number of normal seedlings observed in a seed sample. In fact, when $N \gg n$, it can be shown that

$$\mathbb{P}_X(X = x|N,\ K, n) \cong \binom{n}{x} p^x (1-p)^{n-x} \qquad (4)$$

where $p \cong K/N$ is the probability of success (normal seedling).

The quantiles of R obtained with both exact distributions were compared with the sample quantiles from empirical distribution functions $\hat{F}_R$ built through Monte Carlo simulation processes, one for each base distribution. Ten thousand series of size $m = 4$ (seed samples) were generated for hypergeometric (800, 640, 100) and binomial (100, 0.8) counts, from which 10,000 estimates of R were obtained in order to calculate the empirical probability mass.

https://doi.org/10.1017/S0960258520000112 Published online by Cambridge University Press

Table 1. Critical values[a] of the germination count range between four samples of n seeds each drawn without replacement from the seed lot of size (N) with the germination rate (K/N) varying from 0.50 to 0.95

		Seeds per sample (n)					
N	K/N	30	40	50	60	100	200
200	0.50	9	10	11	–	–	–
	0.55	9	10	11	–	–	–
	0.60	9	10	11	–	–	–
	0.65	9	10	11	–	–	–
	0.70	8	9	10	–	–	–
	0.75	8	9	10	–	–	–
	0.80	7	8	9	–	–	–
	0.85	6	7	8	–	–	–
	0.90	5	6	7	–	–	–
	0.95	4	4	5	–	–	–
400	0.50	10	11	12	13	16	–
	0.55	9	11	12	13	16	–
	0.60	9	11	12	13	15	–
	0.65	9	10	11	12	15	–
	0.70	9	10	11	12	14	–
	0.75	8	9	10	11	14	–
	0.80	8	9	10	10	13	–
	0.85	7	8	9	9	11	–
	0.90	6	6	7	8	9	–
	0.95	4	5	5	6	7	–
800	0.50	10	11	12	13	17	22
	0.55	10	11	12	13	17	22
	0.60	10	11	12	13	17	22
	0.65	9	11	12	13	16	21
	0.70	9	10	11	12	16	20
	0.75	8	10	11	12	15	19
	0.80	8	9	10	11	14	18
	0.85	7	8	9	10	12	16
	0.90	6	7	7	8	10	13
	0.95	4	5	5	6	7	10
2000	0.50	10	11	13	14	18	24
	0.55	10	11	13	14	18	24
	0.60	10	11	12	14	17	24
	0.65	9	11	12	13	17	23
	0.70	9	10	12	13	16	22
	0.75	8	10	11	12	15	21
	0.80	8	9	10	11	14	19
	0.85	7	8	9	10	13	17
	0.90	6	7	8	8	11	15
	0.95	4	5	5	6	8	11
>4000	0.50	10	11	13	14	18	25
	0.55	10	11	13	14	18	25
	0.60	10	11	12	14	18	25
	0.65	9	11	12	13	17	24
	0.70	9	10	12	13	16	23
	0.75	9	10	11	12	16	22
	0.80	8	9	10	11	14	20
	0.85	7	8	9	10	13	18
	0.90	6	7	8	8	11	15
	0.95	4	5	6	6	8	11

[a]Based on the exact distribution of the germination count range and 5% significance.

The effects of sample size (n) and germination rate (p) on the sensitivity of the critical values of R were evaluated by comparing the 0.95 quantiles of the exact distributions with the S-value calculated according to Miles (1963), with 5% nominal significance.

The power of the tests was also calculated by simulating 10,000 values of germination count range according to the base distribution models, that is, hypergeometric and binomial. The range of germination percentages was simulated considering the Normal distribution, with which the S-values were calculated at 5% significance. Formally, consider the respective cumulative distribution functions, $F_{R(\text{Hyp})}$ and $F_{R(\text{Bin})}$ under the following parametrization: $m = 4$, $N = 800$, $n = 100$, $P = 0.55$ and 0.95. And take r_i as the i-th ($i = 1, 2, \ldots, 10{,}000$) simulated value of range under the null hypothesis (lot homogeneity). An increment parameter δ varying from 0 to 25 was added to the simulated range values in order to evaluate the null hypothesis rejection rate (test's power), that is,

$$\text{Power}(\delta) = \frac{1}{10,\,000}\sum_i I[\mathbb{P}_R(R > r_i + \delta) \leq \alpha] \qquad (5)$$

The statistical procedures, simulations and general computing were performed with the software R version 3.4.3 (www.r-project.org). The codes are available with the author.

Results and discussion

The exact probability distributions of the germination count range are presented in Fig. 1, considering the hypergeometric and the binomial models as the base distribution for the germination count of four samples of size $n = 100$ drawn from a seed lot of different sizes ($N = 400$, 800, 1600), with a fixed germination rate ($p = K/N = 0.8$). In Fig. 1a, the approximation to the exact distribution obtained with the binomial model can be verified as lot size increases. When the seed sample size gets close to the lot size, the theoretical distribution gets more skewed to the right, as observed by Laffont et al. (2019). Figure 1b,c shows the Monte Carlo distributions overlapping the theoretical distributions of R.

https://doi.org/10.1017/S0960258520000112 Published online by Cambridge University Press

Tolerated values (S) of germination percentages between four samples were calculated and rounded up (transformed) to germination count values in order to compare them with the 0.95 quantiles of the exact distributions. Figure 2a shows the effect of sample size (n) on the estimates of S and critical values of R in seed lots with germination rates of 0.55 and 0.95, respectively. The critical values obtained using the hypergeometric distribution were the most sensitive on detecting sample heterogeneity, especially for $n > 50$. The S-values were similar to the critical values obtained with the binomial model, as expected, since the first statistics assumes the binomial variance. For $n = 100$, the hypergeometric-based estimates are one seed lower. For $n = 200$, they are four seeds lower. From 50 to 200 seeds per sample, the critical values increase, in average, twice. This is also the average effect of the germination rate (from 55 to 95% germination) on the critical values for a given sample size.

The effect of the lot germination rate (p) on S and the 0.95 quantiles of the exact distributions are shown in Fig. 2b. The same behaviour was observed by Laffont et al. (2019), who calculated 0.975 quantiles, which stand for two-sided P-values. However, when testing for *off*-range heterogeneity through the germination range, only the right side of the distribution is of interest. That is why the critical values presented here have the whole nominal significance (0.05) to the right side. The authors also observed that the S-values are more conservative than the exact quantiles. In average, the difference between the hypergeometric-based values is one seed lower. Piepho et al. (2018) observed that using the hypergeometric model can lead to significantly improved results in heterogeneity testing, especially in all applications where the sample size is low and the percentage value is very high or very low.

In terms of the power of the tests, the germination rate has a considerable effect (Fig. 3). All of them are more powerful when the seed lot has high physiological potential. For example, in average, the germination range between four samples of size $n = 100$ drawn from a lot of size $N = 800$ with 95% germination is equal to four seeds. To detect a significant (P-value < 0.05) range increased by five seeds (range = 9 seeds) with the hypergeometric-based test, the power is equal to 0.96, which is greater than the power of the binomial-based (0.84) or the S-value (0.66). Increases of seven seeds in range promote power above 0.98 for all tests. However, in a seed lot of 55% germination, the power would be much lower, around 0.3, 0.3 and 0.2, respectively. In the case of the low germination rate, using the S-value is not recommended, as it presented approximately 10% less power than the exact tests. In lots with the high germination rate, the proposed test based on the hypergeometric model is about 20% more powerful than the test based on the S-value. In fact, the proposed test is generally more powerful than the other two.

Finally, the critical values with 5% significance calculated using the hypergeometric-based model for several combinations of lot size, germination rate and sample size are given in Table 1, which is recommended to be used in *off*-range heterogeneity testing. Note that variations in germination rate and sample size affect significantly the critical values.

Supplementary material. To view supplementary material for this article, please visit: https://doi.org/10.1017/S0960258520000112.

Financial support. This work was financially supported by the Instituto Federal Goiano (www.ifgoiano.edu.br) and by the Brazilian National Council for Scientific and Technological Development – CNPq [grant number: 307334/2018-0].

Conflicts of interest. None declare.

References

Arnold BC, Balakrishnan N and Nagaraja HN (2008) *A first course in order statistics*. Philadelphia, SIAM.

Herman RA and Robbins KR (2013) Use of hypergeometric distribution for estimating adventitious presence of GM traits in small seed lots may be misleading. *Seed Science Research* **23**, 211–212.

ISTA (2017) *International rules for seed testing*. Bassersdorf, Switzerland, International Seed Testing Association.

Laffont J-L, Hong B, Kuo B-J and Remund KM (2019) Exact theoretical distributions around the replicate results of a germination test. *Seed Science Research* **29**, 64–72.

Miles SR (1963) Handbook of tolerances and measures of precision for seed testing. *Proceedings of the International Seed Testing Association* **28**, 681–685.

Piepho H-P, Kruse M and Deplewski PM (2018) Expected variance between seed germination test replicate results. *Seed Science and Technology* **46**, 197–209.