SCOPUA Journal of

Applied Statistical Research

Vol. 1, Issue 2, August 2025
DOI: https://doi.org/10.64060/JASR.v1.i2.3

On the Mathematical Expectation of the Sample Variance in Simple Sampling Technique

Yasser Al Zaim [1*],Abdulrahman AlAita [2]

[1]Department of Statistics, Faculty of Science, Damascus University, Syria

[2]Department of Agricultural Economy, Faculty of Agricultural Engineering, Damascus University, Syria

* Corresponding Email: y.alzaim@damascusuniversity.edu.sy

Received: 02 April 2025 / Revised: 10 July 2025 / Accepted: 13 July 2025 / Published online: 16 August 2025

ABSTRACT

Drawing random samples is the core of modern life jobs. In manufacturing, it is important to inspect deficiencies by only sampling items from a production line, to meet quality worldwide standards and to maintain sufficient statistical quality control. Furthermore, in today's survey research, the theory of sampling technique is foundational to ensure that all inquired and essential information is gathered. In effect, one may name thousands of practical applications that rely on taking samples, like climatic studies, industry, ecology, and so on. In effect, many studies were designed and proposed in searching for an effective sampling technique. It is, in fact, both an art and a robust science. So many strategies and considerations were plotted to determine the proper sample size and the proper sampling technique, like simple sampling and stratified sampling. This paper is a brief study focusing on the behaviour of the mathematical expectation of the sample variance in sampling without replacement and in sampling with replacement. Formally, we show that when sampling is with replacement, there exists a crucial difference between the two situations, namely, distinct samples and indistinct samples. Namely, by a series of simulation studies and a famous historical example, it will be shown that there is a faulty fact concerning the unbiasedness of sample variance when drawing indistinct samples with replacement.

Keywords*:* Finite Population; Simple Random Sampling; Sample Variance; Unbiased Estimator

1. Introduction and preliminaries

It is a well-known fact today that studying the entire statistical population is not an effective approach to handling challenging problems. As modern studies are forced to interact with massive big data, more and more studies deal with how one can choose a good random sample. In fact, many statistical packages emerged to assist researchers in determining the sample size, like OpenEpi, GPower, Epi-Calc2000, Minitab, PASS, and PS Power, and many free online websites. The use of sampling theory in data mining and other related data sciences is crucial. Pujar et al. (2020) evaluated an ANOVA-based study to monitor water quality by taking random samples from the Krishna River. The statistical steganography methods have also been developed to hide information and to embed secret data. These methods are all built based on sampling techniques, which involve randomly selecting data points from an image or an audio file, and then modifying the sampled points by ensuring that all modifications are made with minimising statistical anomalies. Indeed, good sampling techniques ensure distributing embedded data in a way that makes it harder to be detected. We may refer the interested reader to Majeed et al. (2021); Sumathi et al. (2013). On the other hand, by sampling in high-dimensional data space, Ji et al. (2022) analysed the performance of many multivariate time series classifiers by applying deep learning approaches in order to see the role of data augmentation on time series-oriented flare prediction methods. In another high-dimensional study based on simple sampling, Al Zaim and Faridrohani (2022) defined the famous Anderson-Darling goodness-of-fit test to check the Gaussianity of three-dimensional random fields. Again, Al Zaim and Faridrohani (2021) proposed a Bayesian approach to detect a signal in a Gaussian scale space random field. Further, Alnajdi and Al Zaim (2025) applied a variety of approaches to forecast the gold price based on a simple random sample drawn from 8/8/2022 to 8/6/2024, namely, they performed a wide comparative study to evaluate the performance of statistical, machine learning, and deep learning models in predicting the gold price. In total quality management (TQM), Eissa et al. (2025) employed simple random sampling to evaluate the TQM awareness and its application in improving healthcare basic services at Al-Mouwasat University Hospital in Damascus, Syria. In a similar recent study, AlAita et al. (2025) adopted simple random sampling to run a field study at Isfahan University of Technology, Iran, to analyse a breeding program of safflower plant.

As stated above, the suitable way to deal with big and complex data that emerges in different fields of study, is to draw samples from the underlying studied population(s). And since drawn samples are random, one should study the behaviour of sample quantities, like the sample variance s^2.

Many references deal with studying the sample variance s^2in two cases, drawing with replacement and drawing without replacement (Cochran, 1977). But, in the case of drawing with replacement, investigating the behaviour of $\mathbb{E}[s^2]$ When the chosen samples are distinct or indistinct are omitted from many references, and this is what we will talk about in this paper. Formally speaking, this paper aims to correct a false hidden conclusion considering computing $\mathbb{E}[s^2]$ for indistinct samples are drawn with replacement from a finite population. In other words, both novelty and main contribution of this work arise from spotting a deficiency in a fundamental statistical theorem, and by doing so, we may have successfully turned other researchers' attention to this specific problem when drawing indistinct samples with replacement. We believe that recognising that error is new and a sign of significant improvement, which would open eyes to seek urgent proof and correction to that fundamental statistical theorem.

Actually, in the modern history of statistics, there are many examples of research correcting old false claims. For example, Westfall (2014) corrected a false claim considering the kurtosis coefficient, and he has shown that the kurtosis tells us little about the peak or the centre of any probability distribution. He also concluded that kurtosis should never be related or defined in terms of peakedness, and officially ended the relationship of kurtosis with peakedness.

2. The unbiased estimator of population variance

Let $x_1, x_2, \cdots, x_n$denote a simple sample characteristic that was drawn from a finite population with elements $y_1, y_2, \cdots, y_N$, with $n \leqslant N$. Denote by $|\Omega|$ to the number of simple samples of size$n$drawn from N . That is, (Kirk, 2011, p. 1329)

$$|\Omega| = \begin{cases} \binom{N}{n}; drawingindistinctsampleswithoutreplacement \\ NP_n; drawingdistinctsampleswithoutreplacement \\ \binom{N+n-1}{n}; drawingindistinctsampleswithreplacement \\ N^n; drawingdistinctsampleswithreplacement \end{cases} \qquad (2.1)$$

Where $\binom{v}{k}$ refers to the combinations of v elements, $k$at a time, and vP_k refers to the permutations of v elements, $k$at a time. Now, if we denote by $\bar{x}$ to the sample mean, then its variance is (Cochran, 1977, p. 23)

$$\mathbb{V}(\bar{x}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \qquad (2.2)$$

When the sampling is without replacement, and it is

$$\mathbb{V}(\bar{x}) = \frac{\sigma^2}{n} \qquad (2.3)$$

When sampling with replacement (Cochran, 1977, p. 30).

It is well known that$\bar{x}$, the sample mean, is an unbiased estimator of $\bar{y}$, whatever the sampling is. As a result, we have that

$$\mathbb{E}[\sum_{i=1}^{n} x_i] = \frac{n}{N}\sum_{i=1}^{N} y_i \qquad (2.4)$$

The variance of y_i's is given as follows

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N} \qquad (2.5)$$

And it is called the complete variance. We take also

$$S^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1} \qquad (2.6)$$

as the variance of y_i's. Similarly, the variance of x_i's is similarly defined as

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (2.7)$$

Theorem 2.1. For a simple random sample and the drawing is without replacement, s^2, the sample variance, is an unbiased estimator of S^2.

Proof. We follow (Cochran, 1977, p. 26).

$$\mathbb{E}[s^2] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(x_i \pm \bar{y} - \bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(x_i \pm \bar{y} - \bar{x})^2\right]$$

$$= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}[(x_i - \bar{y})^2 - 2(x_i - \bar{y})(\bar{x} - \bar{y}) + (\bar{x} - \bar{y})^2]$$

$$= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}(x_i - \bar{y})^2 - \frac{n}{n-1}\mathbb{E}[\bar{x} - \bar{y}]^2$$

$$= \frac{1}{n-1}\mathbb{E}\sum_{i=1}^{n}(x_i - \bar{y})^2 - \frac{n}{n-1}\mathbb{V}(\bar{x})$$

The last statement using (2.2) and (2.4) becomes

$$\mathbb{E}[s^2] = \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{N-n}{N-1}\frac{\sigma^2}{n} = \frac{(n-1)N}{(n-1)(N-1)}\sigma^2$$

$$= \frac{N}{N-1}\sigma^2 = S^2$$

Theorem 2.1 is still true, whether chosen samples are distinct or indistinct. But, at the end of the next theorem, we tell another story.

Theorem 2.2. For a simple random sample and the drawing is with replacement, s^2, the sample variance, is an unbiased estimator of σ^2.

Proof. The proof begins by repeating the same steps as in Theorem 2.1 and noticing that (2.3), we find that

$$\mathbb{E}[s^2] = \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} = \sigma^2$$

Both Theorems 2.1 and 2.2 are well-known to all statisticians, and we did not provide any new ideas yet. But, what we would like to pay attention to is the fact that in proving Theorem 2.2, knowing whether drawn samples are distinct or indistinct did not play any role. This remark plays the central role in the following section and reveals a hidden mistake in Theorem 2.2.

3. Simulation Study

In this section, we present different scenarios to show that when sampling is without replacement, $\mathbb{E}[s^2] = S^2$, whether samples are distinct or not. On the other hand, when sampling is with replacement and the samples are distinct, it is true that $\mathbb{E}[s^2] = \sigma^2$. But, when sampling is with replacement and the samples are indistinct, it is false to assume that $\mathbb{E}[s^2] = \sigma^2$.

All *R language* codes are available by contacting the first author. In this section, assume that a rural population is distributed in five villages as shown in Table 3.1.

Table 3.1 *Residents distribution in five villages*

village	a	b	c	d	e
residents	700	500	550	625	600

And we are interested in drawing simple samples of size $n = 3$in each of the following four scenarios.

3.1. First Scenario

Suppose that the object in this scenario is to draw, without replacement, indistinct simple samples of size $n = 3$. Obviously, $|\Omega| = 10$ according to equation (2.1) . Table 3.2 contains the possible simple samples ω_k along with their means $\bar{x}_k$ and variances s_k^2, where$k = 1, 2, \cdots, 10$.

Table 3.2 *Possible chosen samples when sampling is without replacement*

k	1	2	3	4	5	6	7	8	9	10
ω_k	abc	abd	abe	bcd	bce	cde	acd	ace	ade	bde
$\bar{x}_k$	583.33	608.33	600	625	616.67	641.67	558.33	550	575	591.67
s_k^2	10833.33	10208.33	10000	5625	5833.33	2708.33	3958.33	2500	4375	1458.33

Results in Table 3.2 show that

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2 + s_2^2 + \cdots + s_{10}^2}{10} = 5750$$

$$S^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N-1} = \frac{(700-595)^2 + (500-595)^2 + \cdots + (600-595)^2}{4} = 5750$$

and this is consistent with Theorem 2.1.

3.2. Second Scenario

We turn to the second case, namely, the target is to draw, without replacement, distinct simple samples of size$n$. By equation (2.1), these are $|\Omega| = 60$ samples. Table 3.3 shows all drawn simple samples, and in this case, we still have

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2+s_2^2+\cdots+s_{60}^2}{60} = \frac{345000}{60} = 5750$$

$$S^2 = 5750$$

This coincides with Theorem 2.1.

Table 3.3 *Possible chosen distinct samples without replacement*

k	ω_k	k	ω_k	k	ω_k	k	ω_k	k	ω_k	k	ω_k
1	*abc*	11	*abd*	21	*abe*	31	*acb*	41	*acd*	51	*ace*
2	*adb*	12	*adc*	22	*ade*	32	*aeb*	42	*aec*	52	*aed*
3	*bac*	13	*bad*	23	*bae*	33	*bca*	43	*bcd*	53	*bce*
4	*bda*	14	*bdc*	24	*bde*	34	*bea*	44	*bec*	54	*bed*
5	*cab*	15	*cad*	25	*cae*	35	*cba*	45	*cbd*	55	*cbe*
6	*cda*	16	*cdb*	26	*cde*	36	*cea*	46	*ceb*	56	*ced*
7	*dab*	17	*dac*	27	*dae*	37	*dba*	47	*dbc*	57	*dbe*
8	*dca*	18	*dcb*	28	*dce*	38	*dea*	48	*deb*	58	*dec*
9	*eab*	19	*eac*	29	*ead*	39	*eba*	49	*ebc*	59	*ebd*
10	*eca*	20	*ecb*	30	*ecd*	40	*eda*	50	*edb*	60	*edc*

3.3. Third Scenario

Now, assume that indistinct simple samples of size $n = 3$ are to be drawn with replacement. As a result, by equation (2.1) , $|\Omega| = 35$. The results listed in Table 3.4 show that

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2 + s_2^2 + \cdots + s_{35}^2}{35} = \frac{134166.7}{35} = 3833.33$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N} = \frac{N-1}{N} S^2 = 4600$$

This seems to be a contradiction with Theorem 2.2, and counterproductive to the purpose of fostering statistical literacy.

Table 3.4 *Possible chosen indistinct samples and sampling is with replacement. The $\bar{x}_k$ and s_k^2 are reported in parentheses*

k	$\omega_k(\bar{x}_k, s_k^2)$	k	$\omega_k(\bar{x}_k, s_k^2)$	k	$\omega_k(\bar{x}_k, s_k^2)$	k	$\omega_k(\bar{x}_k, s_k^2)$
1	$bbb(500, 0)$	11	$bed(575, 4375)$	21	$ced(591.67, 1458.33)$	31	$eaa(666.67, 3333.33)$
2	$bbc(516.67, 833.33)$	12	$bea(600, 10000)$	22	$cea(616.67, 5833.33)$	32	$ddd(625, 0)$
3	$bbe(533.33, 3333.33)$	13	$bdd(583.33, 5208.33)$	23	$cdd(600, 1875)$	33	$dda(650, 1875)$
4	$bbd(541.67, 5208.33)$	14	$bda(608.33, 10208.33)$	24	$cda(625, 5625)$	34	$daa(675, 1875)$
5	$bba(566.67, 13333.33)$	15	$baa(633.33, 13333.33)$	25	$caa(650, 7500)$	35	$aaa(700, 0)$
6	$bcc(533.33, 833.33)$	16	$ccc(550, 0)$	26	$eee(600, 0)$		
7	$bce(550, 82500)$	17	$cce(566.67, 833.33)$	27	$eed(608.33, 208.33)$		
8	$bcd(558.33, 3958.33)$	18	$ccd(575, 1875)$	28	$eea(633.33, 3333.33)$		
9	$bca(583.33, 10833.33)$	19	$cca(600, 7500)$	29	$edd(616.67, 208.33)$		
10	$bee(566.67, 3333.33)$	20	$cee(583.33, 833.33)$	30	$eda(641.67, 2708.33)$		

3.4. Fourth Scenario

In this scenario, we consider drawing with replacement all distinct simple samples of size $n = 3$, hence, by equation (2.1), $|\Omega| = 125$. Table 3.5 reflects all possible simple samples, and as a result, we have that

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2 + s_2^2 + \cdots + s_{125}^2}{125} = \frac{575000}{125} = 4600$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N} = \frac{N-1}{N} S^2 = 4600$$

This coincides with Theorem 2.2.

Table 3.5 *Possible chosen distinct samples with replacement*

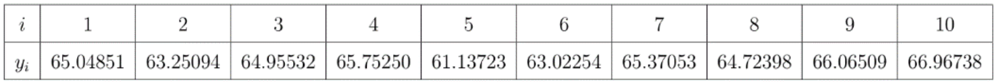
k	ω_k	k	ω_k	k	ω_k	k	ω_k	k	ω_k	k	ω_k	k	ω_k
1	aaa	21	aea	41	bda	61	cca	81	dba	101	eaa	121	eea
2	aab	22	aeb	42	bdb	62	ccb	82	dbb	102	eab	122	eeb
3	aac	23	aec	43	bdc	63	ccc	83	dbc	103	eac	123	eec
4	aad	24	aed	44	bdd	64	ccd	84	dbd	104	ead	124	eed
5	aae	25	aee	45	bde	65	cce	85	dbe	105	eae	125	eee
6	aba	26	baa	46	bea	66	cda	86	dca	106	eba		
7	abb	27	bab	47	beb	67	cdb	87	dcb	107	ebb		
8	abc	28	bac	48	bec	68	cdc	88	dcc	108	ebc		
9	abd	29	bad	49	bed	69	cdd	89	dcd	109	ebd		
10	abe	30	bae	50	bee	70	cde	90	dce	110	ebe		
11	aca	31	bba	51	caa	71	cea	91	dda	111	eca		
12	acb	32	bbb	52	cab	72	ceb	92	ddb	112	ecb		
13	acc	33	bbc	53	cac	73	cec	93	ddc	113	ecc		
14	acd	34	bbd	54	cad	74	ced	94	ddd	114	ecd		
15	ace	35	bbe	55	cae	75	cee	95	dde	115	ece		
16	ada	36	bca	56	cba	76	daa	96	dea	116	eda		
17	adb	37	bcb	57	cbb	77	dab	97	deb	117	edb		
18	adc	38	bcc	58	cbc	78	dac	98	dec	118	edc		
19	add	39	bcd	59	cbd	79	dad	99	ded	119	edd		
20	ade	40	bce	60	cbe	80	dae	100	dee	120	ede		

4. Historic Example

Pearson and Lee (1903) collected the data of $N = 1078$ fathers' heights and their fully grown sons in inches. But, since the computer was not there, the original data was rounded to the nearest integer. So, Freedman et al. (2007) added random noise as a continuity correction to the data. In this section, we use the corrected data which is available from the authors in this link https://myweb.uiowa.edu/pbreheny/data/pearson.html.

We will focus on the heights of fathers. But, since N is large, we must reduce the computational burden, and so, we considered the first $N = 10$ as the population size (see Table 4.1). As for n , we take into account drawing simple samples of size $n = 5$.

Table 4.1 *The first* 10 *fathers' heights in inches*



i	1	2	3	4	5	6	7	8	9	10
y_i	65.04851	63.25094	64.95532	65.75250	61.13723	63.02254	65.37053	64.72398	66.06509	66.96738

First, when sampling indistinct samples with replacement, we have $|\Omega| = 2002$ samples.

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2 + s_2^2 + \cdots + s_{2002}^2}{2002} = \frac{4787.73}{2002} = 2.391473$$

$$\sigma^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{N} = \frac{N-1}{N} S^2 = 2.630621$$

Again, these results reflect the fact that Theorem 2.2 is not verified.

Second, let us turn to the other case, namely, sampling distinct samples with replacement. Here, we have that $|\Omega| = 100000$ samples, and hence,

$$\mathbb{E}[s^2] = \sum_{k=1}^{|\Omega|} \frac{s_k^2}{|\Omega|} = \frac{s_1^2 + s_2^2 + \cdots + s_{100000}^2}{100000} = \frac{263062.1}{100000} = 2.630621$$

Which equals to $\sigma^2 = 2.630621$. This final result rhymes with Theorem 2.2.

5. Conclusion

This paper has clarified the effect of drawing a simple sample methodology on computing the mathematical expectation of the sample variance s^2. We have seen, through a series of simulation studies and a real historic example, that s^2, is an unbiased estimator of the population variance in three cases, namely, when sampling is without replacement for distinct and indistinct samples, and when sampling is with replacement for distinct samples.

Declaration

Conflict of Study: *The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper*

Funding Statement: *This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.*
Availability of Data and Material: *Data will be made available by the corresponding author on reasonable request.*
Consent to Publish: All authors have agreed to publish this manuscript in the SCOPUA Journal of Applied Statistical Research (JASR) ISSN(e): 3104-4794.
Ethical Approval: *Not Applicable*
Consent to Participate: *Not Applicable*

References

1. AlAita, A., Talebi, H. and Al Zaim Y. (2025), "Estimating and Testing Augmented Randomized Complete Block Designs: The Neutrosophic Approach", *Neutrosophic Sets and Systems*, **82**, 639-654.
2. Al Zaim, Y. and Faridrohani, M. R. (2021), Bayesian random projection-based signal detection for gaussian scale space random fields. *AStA Advances in Statistical Analysis*, **105**:503–532.
3. Al Zaim, Y. and Faridrohani, M. R. (2022), Random projection-based anderson-darling test for random fields, *Journal of Iranian Statistical Society*, **20(2)**:1–28.
4. Alnajdi, R. and Al Zaim, Y. (2025), Time-series machine learning approaches in analyzing and forecasting gold price, *Damascus university journal (in press)*.
5. Cochran, W. G. (1977), *Sampling Techniques, third edition*, John Wiley & Sons.
6. Eissa, A., Al-Tarrab, F., Kasem, E., Al Zaim, Y., Hmidoush, A., Salloum, J. and Ataya, J. (2025), "A comparative study of total quality management in healthcare from provider and patient perspectives at Al-Mouwasat University Hospital", *Scientific Reports*, **15**:23746. https://doi.org/10.1038/s41598-025-08512-2
7. Freedman, D., Pisani, R. and Purves, R. (2007), *Statistics, 4th edition*, Norton.
8. Ji, A., Wen, J., Angryk, R. and Aydin, B. (2022), Solar flare forecasting with deep learning-based time series classifiers, *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2907–2913.
9. Kirk, R. (2011), *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, Berlin, Heidelberg.
10. Majeed, M. A., Sulaiman, R., Shukur, Z. and Hasan, M. K. (2021) A review on text steganography techniques, *Mathematics*, **9(21)**:2829.
11. Pearson K. and Lee, A. (1903), On the laws of inheritance in man. *Biometrika*, **2**:357-462.
12. Pujar, P. M., Kenchannavar, H., Kulkarni, R. and Kulkarni, U. (2020), Real-time water quality monitoring through internet of things and anova-based analysis: a case study on river krishna. *Applied Water Science*, **10(22)**.
13. Sumathi, C. P., Santanam, T. and U. G. Umamaheswari (2013), A study of various steganographic techniques used for information hiding, *International Journal of Computer Science & Engineering Survey*, **4(6)**.
14. Westfall, P. H. (2014), Kurtosis as peakedness, 1905-2014. r.i.p., *The American Statistician*, **68(3)**:191–195.