Vol.10.Issue.1.2022 (Jan-Mar)
©KY PUBLICATIONS

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

http://www.bomsr.com
Email:editorbomsr@gmail.com

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

RESEARCH ARTICLE

SAMPLE VARIANCE - IS IT RALLY AN UNBAISED ESTIMATE OF THE POPULATION VARIANCE?

Dr.RAMNATH TAKIAR
Scientist G – (Retired), National Centre for Disease Informatics and Research (NCDIR),
Indian Council of Medical Research (1978-2013)
Bangalore – 562110, Karnataka, India.
Email: ramnathtakiar@gmail.com, ramnath_takiar@yahoo.co.in

DOI:10.33329/bomsr.10.1.21

Dr.RAMNATH TAKIAR

ABSTRACT

The present paper explores empirically the property of unbiasedness of sample variance. For the study purposes, three finite populations termedas P1, P2 and P3 of size 7 were considered. For each population, all the possible distinct samples of size 2, 3, 4, 5 and 6 were generated. Thus, accordingly, the number of samples generated were 21, 35, 35, 21 and 7, respectively. In total for each population, 119 samples were generated. For each population and for each sample size, three statistics were calculated namely mean, V(n-1), V(n) where V(n-1) and V(n) is the sum of squared deviations, divided by (n-1) and n, respectively. Against the popular claim, V(n-1) is not foundto be an unbiased estimate of the Population Variance denoted as V(N). Further, it was found to be consistently overestimating the V(N) by 17% approximately. In case of V(n), it was found to be underestimating the population variance by 19%. Based on the study results, it is suggested that the preference of using V(n-1) over V(n) should be assessed critically and may be dropped from further use. Consequence to this, and in view of my previous study, the use of t-test for small samples should be viewed critically, again and may be the use of Z-test in place of t-test to be encouraged, hereafter.

Keywords: Sample variance, unbiasedness, t-test, Z-test

Introduction

The mean is the most popular average which is commonly used to characterize a set of data. It is a statistical constant derived from the data and gives us an idea about the central part of the distribution. Besides mean, another measure which is commonly used to characterize the data is the variance or standard deviation. The mean along with the standard deviation can be used to compare different series of data. In most of the applications, we deal with the sample mean and the standard deviation. While mean is considered as an unbiased estimate, the sample variance is not considered as an unbaised estimate of the population variance. Hence, the use of sample standard deviation become questionable. To overcome this difficulty in using the sample standard deviation,acorrection was suggested and it was advocated that in calculation of the sample variance, the sum of squared deviations should be divided by (n-1) instead of n. It is claimed that by doing so, the estimate of the sample variance becomes an unbiased estimate of the population variance (Gupta 2012, Gupta and Kapoor 2001, Snedecor & Cochran 1968). The present paper explores the validity of the above calim. The objectives of the present study therefore is:

- To explore empirically whether the Sample Variance is really an unbaised estimate of the Population Variance or not?

Material and Methods

Let us define first the terms which I am going to use in the paper.

Mean

The arithmetic mean of a set of n observations is their sum divided by the number of observations. In other words, if x_1, x_2, x_3, ...x_n are the observations of a variable then the arithmetic mean is given by the formula:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + + + x_n}{n} = \frac{\sum x}{n}$$

Where ∑ is the notation for representing the sum of the observations.

Variance

It is the average of the squared deviations of the observations from their arithmetic mean. The formula for a population variance is given as follows:

$$Variance = \sigma^2 = V(N) = \frac{1}{N} \sum(x - \mu)^2$$

Following the above definition, we define V(N-1) as follows:

$$V(N-1) = \frac{1}{N-1} \sum(x - \mu)^2$$

Where μ is the population mean and N is the population size. For the sample, two types of formulae can be defined for the calculation of variance, namely

Sample Variance = V(n) = $\frac{1}{n} \sum (x_i - \overline{x})^2$

Sample Variance = V(n-1) = $\frac{1}{(n-1)} \sum (x_i - \overline{x})^2$

The sample variance, calculated by the formula of V(n) which is a natural choice, is shown to be a biased estimate while the sample variance calculated by V(n-1) is shown to be an unbiased estimate.

Unbiased Estimate

An estimate is said to be an unbaised estimate if its expected value equals to the parameter value. In notation,

$$E(\bar{x}) = \mu\,; \quad E[V(n)] = \sigma^2$$

If the Expected value of the selected statistic is not equal to the parameter value, then that statisticis considered as a biased estimate.

Selection of Populations

For the study purposes, it was thought logical to consider finite populations. A finite population allow you to generate all the possible samples and thus a decision can be taken about the estimates of the population parameters, based on all the possible samples of different size.Thethree finite populations considered are shown in Table 1.

Table 1: Description of Three Finite Populations Selected for the Study Purposes

Sl. No.	Population		
	P1	P2	P3
1	8	48	81.1
2	42	55	88.8
3	33	27	105.8
4	76	55	81.3
5	46	70	70.4
6	39	78	88.1
7	58	54	80.4
Mean	43.1	55.3	85.1
Median	42.0	55.0	81.3
V(N-1)	444.14	264.57	119.97
V(N)	380.69	226.78	102.83

In theory, V(n) is considered as a biased estimate while V(n-1) is shown to be an unbiased estimate of the population variance. Hence, in practice, V(n-1) is preferred over V(n). It isalsopossible to show that the sample variance V(n-1) is really an unbiased estimate of the population variance by selecting a population and then drawing all the possible samples of a given size. If the mean ofallsuch sample variances defined by V(n-1) of different sample sizes, corresponds to the populationvariance, then we can say that the V(n-1) is an unbiased estimate of the population variance.

For each population, all the possible distinct samples of size 2,3,4,5 and 6 are considered. In case of a Population with the elements (A,B,C), AB, AC and BC are considered as the distinct samples. The number of all possible samples for a population size of 7 are shown in the Table 2.

The distribution of means and sample variances V(n) and V(n-1) are obtained for each of the generated samples. For each sample size, an attempt is made to check the property of unbiasedness by obtaining the mean in addition to the Variance of the sampling distribution of all the above three selected statistics . It is expected that to be an unbaised estimate, the mean of the selected statistic should coincide with the population parameter.

Table 2: Number of Distinct Samples according to Sample Size When Population size is 7

Variable	Sample size					
	2	3	4	5	6	Total
Number of Samples	21	35	35	21	7	119

Results

The all-possible samples of size 2, for the Population of P1, are generated and then for each sample, Mean, V(n-1) and V(n) are calculated and shown in Table 3.

Table 3: Mean, V(n-1) and V(n) for the Samples of size 2 From the Population P1

Sl. No	Samples	Mean	V(n-1)	V(n)
1	8,42	25.0	578.0	289.0
2	8,33	20.5	312.5	156.3
3	8,76	42.0	2312.0	1156.0
4	8,46	27.0	722.0	361.0
5	8,39	23.5	480.5	240.3
6	8,58	33.0	1250.0	625.0
7	42,33	37.5	40.5	20.3
8	42,76	59.0	578.0	289.0
9	42,46	44.0	8.0	4.0
10	42,39	40.5	4.5	2.3
11	42,58	50.0	128.0	64.0
12	33,76	54.5	924.5	462.3
13	33,46	39.5	84.5	42.3
14	33,39	36.0	18.0	9.0
15	33,58	45.5	312.5	156.3
16	76,46	61.0	450.0	225.0
17	76,39	57.5	684.5	342.3
18	76,58	67.0	162.0	81.0
19	46,39	42.5	24.5	12.3
20	46,58	52.0	72.0	36.0
21	39,58	48.5	180.5	90.3
Mean		43.1	444.1	222.1
V(n)		158.62	285566.19	71388.23
Parameter		43.1	380.7	380.7

The all-possible samples of size 3, 4, 5 and 6 for the Population P1, are generated and shown in Table 4 to Table 7 mainly for the readers to understand and verify themselves as to how all possible samples, for different sample sizes, can be generated.

Table 4: All Possible Samples of size 3 from the Population P1

8, 42, 33	8, 76, 46	42, 33, 58	33, 76, 58
8, 42, 76	8, 76, 39	42, 76, 46	33, 46, 39
8, 42, 46	8, 76, 58	42, 76, 39	33, 46, 58
8, 42, 39	8, 46, 39	42, 76, 58	33, 39, 58
8, 42, 58	8, 46, 58	42, 46, 39	76, 46, 39
8, 33, 76	8, 39, 58	42, 46, 58	76, 46, 58
8, 33, 46	42, 33, 76	42, 39, 58	76, 39, 58
8, 33, 39	42, 33, 46	33, 76, 46	46, 39, 58
8, 33, 58	42, 33, 39	33, 76, 39	

Table 5: All Possible Samples of size 4 from the Population P1

8, 42, 33, 76	8, 42, 39, 58	8, 76, 39, 58	42, 76, 46, 58
8, 42, 33, 46	8, 33, 76, 46	8, 46, 39, 58	42, 76, 39, 58
8, 42, 33, 39	8, 33, 76, 39	42, 33, 76, 46	42, 46, 39, 58
8, 42, 33, 58	8, 33, 76, 58	42, 33, 76, 39	33, 76, 46, 39
8, 42, 76, 46	8, 33, 46, 39	42, 33, 76, 58	33, 76, 46, 58
8, 42, 76, 39	8, 33, 46, 58	42, 33, 46, 39	33, 76, 39, 58
8, 42, 76, 58	8, 33, 39, 58	42, 33, 46, 58	33, 46, 39, 58
8, 42, 46, 39	8, 76, 46, 39	42, 33, 39, 58	76, 46, 39, 58
8, 42, 46, 58	8, 76, 46, 58	42, 76, 46, 39	

Table 6: All Possible Samples of size 5 from the Population P1

8, 42, 33, 76, 46	8, 42, 76, 46, 39	8, 33, 76, 39, 58	42, 33, 46, 39, 58
8, 42, 33, 76, 39	8, 42, 76, 46, 58	8, 33, 46, 39, 58	42, 76, 46, 39, 58
8, 42, 33, 76, 58	8, 42, 76, 39, 58	8, 76, 46, 39, 58	33, 76, 46, 39, 58
8, 42, 33, 46, 39	8, 42, 46, 39, 58	42, 33, 76, 46, 39	
8, 42, 33, 46, 58	8, 33, 76, 46, 39	42, 33, 76, 46, 58	
8, 42, 33, 39, 58	8, 33, 76, 46, 58	42, 33, 76, 39, 58	

Table 7: All Possible Samples of size 6 from the Population P1

8, 42, 33, 76, 46, 39

8, 42, 33, 76, 46, 58

8, 42, 33, 76, 39, 58

8, 42, 33, 46, 39, 58

8, 42, 76, 46, 39, 58

8, 33, 76, 46, 39, 58

42, 33, 76, 46, 39, 58

For each sample size of the Population P1, the mean of the Sample Mean, V(n-1) and V(n) are calculated, separately, and are shown in Table 8.

Table 8: Mean of the Sample Means, V(n-1) and V(n) for the Population P1

Sample Size	No. of Samples	Mean	V(n-1)	V(n)
2	21	43.1	444.14	222.07
3	35	43.1	444.14	296.1
4	35	43.1	444.14	333.11
5	21	43.1	444.14	355.31
6	7	43.1	444.14	370.12
Pooled	119	43.1	444.14	308.72
Parameter value		**43.1**	**380.69**	**380.69**

From the data provided in the Table 8, it is clear that the Sample mean is an unbiasedestimate. However, the Sample variances namely V(n-1), V(n) can be considered as the biased estimatesforall the sample sizes as they differ from that observed for the population. It is interesting to note that V(n-1) remained constant for all the sample sizes and is an unbiased estimate of the modified Population Variance V(N-1) and not of V(N) as popularly, believed. When V(n-1) iscomparedwithV(N), it is observed that it overestimates, consistently, on an average by 17%. In case of V(n), it tends to underestimate the Population Variance V(N) by 19%.

For the Population P1, for each sample size, the Variance of the Sample Means, V(n-1) and V(n) are calculated, separately, and shown in Table 9.

Table 9: Variance of Sample Means, V(n-1) and V(n) for the Population P1

Sample Size	No. of Samples	Mean	Variance V(n-1)	Variance V(n)	$\frac{\mathbf{V(n-1)}}{\mathbf{V(n)}}$
2	21	158.6	285566.19	71388.23	4.00
3	35	84.6	124623.17	55388.08	2.25
4	35	47.6	64910.76	36512.30	1.78
5	21	25.4	33235.13	21270.48	1.56
6	7	10.6	13501.99	9376.38	1.44
Pooled	**119**	**72.0**	**112798.56**	**46084.81**	**2.45**

It is clear that for all the sample sizes, the variation in V(n-1) estimates as compared to V(n) is always higher ranging from 1.44 times to 4.00 times. On an average, the estimate is 2.45 times higher.

For each sample size of the Population P2, the mean of the Sample Means, V(n-1) and V(n) are calculated, separately, and shown in the Table 10.

Table 10: Mean of Sample Means, V(n-1) and V(n) for the Population P2

Sample Size	No. of Samples	Mean	V(n-1)	V(n)
2	21	55.3	264.57	132.29
3	35	55.3	264.57	176.38
4	35	55.3	264.57	198.43
5	21	55.3	264.57	211.66
6	7	55.3	264.57	220.48
Pooled	119	55.3	264.57	183.9
Parameter value		**55.3**	**226.78**	**226.78**

From the data provided in the Table 10, it is clear that for all the sample sizes, the Sample mean is an unbiased estimate while, the variances V(n-1) and V(n) cannot be considered as the unbiased estimates. As before, it is noted that V(n-1) remained constant for all the sample sizes and is an unbiased estimate of the modified Population Variance V(N-1) and not V(N), as expected.When V(n-1) is compared with V(N), it is observed that it overestimates, consistently, on an averageby17%. The V(n) is observed to be underestimating the V(N) by approximately 19%.

For the Population P2, for each sample size, the Variance of the estimates of Sample Means, V(n-1) and V(n) are calculated, separately, and shown in the Table 11.

Table 11: Variance of Sample Means, V(n-1) and V(n) for the Population P2

Sample Size	No. of Samples	Mean	Variance V(n-1)	Variance V(n)	$\frac{V(n-1)}{V(n)}$
2	21	94.5	99667.82	24916.95	4.00
3	35	50.4	43112.53	19161.12	2.25
4	35	28.4	22367.62	12581.78	1.78
5	21	15.1	11427.21	7313.42	1.56
6	7	6.3	4635.79	3219.3	1.44
Pooled	**119**	**42.9**	**39136.57**	**15976.73**	**2.45**

It is clear from the above table that for all the sample sizes, the variation in V(n-1) estimates as compared to V(n), is always higher, ranging from 1.44 times to 4.00 times. On an average, the estimate is 2.45 times higher.

For the Population P3, for each sample size, the Mean of the Sample Means, V(n-1) and V(n) are calculated, separately, and shown in Table 12. From the data provided in Table 12, it is clear that for all the sample sizes, the Sample mean is an unbiased estimate while V(n-1) and V(n) are the biased estimates.

Table 12: Means of Sample Means, V(n-1) and V(n) for the Population P3

Sample Size	No. of Samples	Mean	V(n-1)	V(n)
2	21	85.1	119.97	59.98
3	35	85.1	119.97	79.98
4	35	85.1	119.97	89.97
5	21	85.1	119.97	95.97
6	7	85.1	119.97	99.97
Pooled	119	85.1	119.97	83.39
Parameter value		**85.1**	**102.83**	**102.83**

It is noted that V(n-1) remained constant for all the sample sizes as noted before and is an unbiased estimate of the modified Population Variance V(N-1) and not that of V(N). When V(n-1)was compared with V(N), it is observed that it overestimates, consistently, on an average by 17%.TheV(n) is observed to be underestimating the V(N) by approximately 19%.

For each sample size of the Population P3, the Variance of the Sample Means, V(n-1) andV(n) are calculated, separately, and shown in Table 13.

Table 13: Variance of Sample Means, V(n-1) and V(n) for the Population P3

Sample Size	No. of Samples	Mean	Variance V(n-1)	Variance V(n)	$\frac{\mathbf{V(n-1)}}{\mathbf{V(n)}}$
2	21	42.8	23352.68	5838.17	4
3	35	22.9	10771.19	4787.2	2.25
4	35	12.9	5743.13	3230.51	1.78
5	21	6.9	2978.82	1906.45	1.56
6	7	2.9	1220.13	847.31	1.44
Pooled	**119**	**19.4**	**9575.66**	**3931.66**	**2.44**

As before, It is clear that for all the sample sizes, the variation in V(n-1) estimates ascompared to V(n), is always higher, ranging from 1.44 times to 4.00 times. On an average, the estimate is 2.44 times higher.

Discussion

It is well documented that the sample mean is an unbiased estimate of the population mean. In the present study, also, for 15 different sampling distribution of mean, it is confirmed that mean is an unbaised estimate of the population mean. However, against the popular claim that V(n-1) is an unbaised estimate of the population variance, in the present study, it is observed that it is not an unbaised estimate of the population variance. This is surprising and unbelievable to find that V(n-1) is not an unbaised estimate. This raised the doubt that so called theoretical proof which is offered to prove that it is an unbaised estimate against the natural choice of V(n), has to be checked again, critically. This raise the question further whether there exists an unbiased estimate for thepopulation variance at all? Till date, it was believed that V(n-1) is an unbiased estimate. However, it provedtobe wrong in the present study. It is also noted that V(n-1) consistently overestimate while V(n) underestimate the population variance. However, when you compare its variance with that of V(n), it exhibits more than double the variation seen in V(n) making it undesirable. In view of this, the preference of V(n-1) over V(n) is not justifiable. We can as well continue to use V(n) in place ofV(n-1).

The sample variance V(n-1) has been preferably used over V(n) in the pretext that it is an unbiased estimate. Now, that it is shown in the present study that it is not an unbiased estimate of the population variance, its use should be stopped. This also raised the doubt in the use oft-testwhich essentially uses V(n-1) in place of V(n). In my previous study (Takiar 2021), I have shown that t-test as believed is not that superior to Z-test and now that it is shown that V(n-1) is not an unbaised estimate, the continuous use of t-test has to be reviewed seriously and can be stopped. The result ofthecurrent study strengthens the point raised in my previous study that in place of t-test, Z-test can be used,here after.

Conclusion

The study results have clearly shown that the sample variance V(n-1) like V(n) is a biased estimate of the population variance. Further, the variation in V(n-1) estimates was observed to be more than 2.5 times as compared to that seen in the case of V(n). This suggests that there is no reason left for the preferential use of V(n-1) over V(n). This also raises doubt in the use of t-test for small samples as it uses V(n-1) in place of V(n). It is therefore suggested that for small samples, the Z-test can be used in place of t-test.

References

[1]. Gupta S C (2012): Fundamentals of Statistics, Himalaya Publishing House, 7th Edition, Page 16.2.

[2]. Gupta S C and Kapoor V K (2001): Fundamentals of Mathematical Statistics, Sultan Chand & Sons, 10h Edition, Page 14.1.

[3]. Snedecor G W and Cochran W G 1968: Statistical Methods, Oxford and IBH Publishing Co., Indian Edition, Page 45.

[4]. Takiar R (2021): The Validity of $\boldsymbol{t}$-test and $\boldsymbol{Z}$-test for Small One sample and Small Two Sample test, *Bulletin of Mathematics and Statistics Research*, Vol. 9(4), Page 42-57.

Biography

Dr. Ramnath Takiar

I am a Post graduate in Statistics from Osmania University, Hyderabad. I did my Ph.D. from Jai Narain Vyas University of Jodhpur, Jodhpur, while in service, as an external candidate. I worked as a research scientist (Statistician) for Indian Council of Medical Research from 1978 to 2013 and retired from the service as Scientist G (Director Grade Scientist). I am quite experienced in large scale data handling, data analysis and report writing. I have 65 research publications in national and International Journals related to various fields like Nutrition, Occupational Health, Fertility and Cancer epidemiology. During the tenure of my service, I attended three International conferences namely in Goiana (Brazil -2006), Sydney (Australia-2008) and Yokohoma (Japan-2010) and presented a paper in each. I also attended the Summer School related to Cancer Epidemiology (Modul I and Module II) conducted by International Agency for Research in Cancer (IARC), Lyon, France from 19th to 30th June 2007. After my retirement, I joined my son at Ulaanbaatar, Mongolia. I worked in Ulaanbaatar as a Professor and Consultant from 2013-2018 and was responsible for teaching and guiding Ph.D. students. I also taugth Mathematics to undergraduates and Econometrics to MBA students. During my service there, I also acted as the Executive Editor for the in-house Journal "International Journal of Management". I am still active in research and have published 5 research papers in year 2021.