Journal of Natural Sciences and Mathematics of UT, Vol. 4, No. 7-8, 2019

UDC: 519.224/.226
Conference paper

A NOTE ON AN ALTERNATIVE FORMULA FOR CALCULATING S^2 AND ITS USES

Denajda Çibuku[1], Entela Kostrista[2]

[1*]*Business Management Department, Faculty of Economy, LOGOS University College*

[2] *Scientific Research Center, Faculty of Applied Sciences, LOGOS University College*

[*]*Corresponding Author: e-mail: denajda.cibuku@kulogos.edu.al*

Abstract

This article is referred to the article of Stigler. That gives the proof of the joint distribution between the sample mean and the sample variance with the assumption that the population is normally. The machine formula of S^2 has been used during the proof by Stigler, the bivariate distribution is being needed and also the Induction Theory.

Starting in this context and as part of statistical education, it has been brought a different viewpoint of doing the proofs of some important results and derivations of the sample mean distribution. It has been changed the scheme of proof replaying the machine formula with an alternative formula of S^2, which is a less used definition of the sample variance.

Compared with Stigler this method of proof gives less calculation. It is of interest to know what the covariance of the sample mean and the sample variance is without the assumption of normality. One method of proof is being done by Zhang (2007). He did not use the well-known formula of the sample variance, but it has been used the alternative formula to give the same result in an easer way. It, also, has been used the fact that the third moment of the sample population exists. The computations were straightforward and did not require advanced mathematical notions.

Is been showed an example of Poisson distributed to illustrate the fact that the covariance of the joint distribution of the sample variance and the sample mean were zero, but they were not independent. It has been constructed a table of joint probability and from this has been displayed the result.

Starting from this example and regarding challenges in teaching Statistics we pretend in the future to apply the distribution of an average of Poisson random variables in advance Statistics.

Keywords: *Sample Variance, Sample Mean, Independence, Poisson.*

1. Introduction

In introduction courses in mathematical statistics, the proof that the sample mean $\overline{X}$ and sample variance S^2 are independent when one is sampling from normal populations is commonly deferred until substantial mathematical machinery has been developed. The proof uses properties of moment-generating functions, algebra [1]. As shown in [3] all that is needed to complete the proof are some facts about bivariate normal distribution: two linear combinations of a pair of independent normally distributed random variables are themselves bivariate normal, and hence if they are uncorrelated, are independent.

We know the machine formula:

$$S^2 = \frac{1}{(n-1)} \sum_{j=1}^{n} (X_j - \overline{X})^2.$$

Both numerical and theoretical computations can be simplified from the use of one of the two equivalent forms of the machine formula:

$$\sum_{j=1}^{n}(X_j - \overline{X})^2 = \sum_{j=1}^{n} X_j - n\overline{X}^2 = \sum_{j=1}^{n} X_j - \frac{(\sum_{j=1}^{n} X_j)^2}{n}.$$

Thus, whether one is dealing with a sample of 100 or 1000 observations the only computationally demanding part of the calculation of S^2 is determining the sum of the sample values and their squares. The machine formula not only simplifies the computation of S^2, but it also permits a simple derivation of the following important result $E(S^2) = \sigma^2$.

2. An alternative formula of S^2 and its uses

This alternative formula S^2 is less useful for computation purposes than the machine formula, but it is of considerable theoretical interest. It is defined as:

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2 \qquad \text{Equation 1}$$

The double summation in Equation 1 can be rewritten in the form:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \overline{X} + \overline{X} - X_j)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - \overline{X})^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} (\overline{X} - X_j)^2 + 2n \sum_{j=1}^{n} (X_j - \overline{X})^2 \qquad \text{Equation 2}$$

Recall that $\sum_{j=1}^{n}(X_j - \overline{X})^2 = 0$, divide Equation 2 by 2n(n-1) and get the formula in Equation 1.

Let $\theta_j = E(X_i - \mu)^2$ for j=1, 2, 3, 4 denote the first four central moments of the r.v X. Using a little algebra we need some preliminary (calculation) results to find out a formula for Var S^2 and Cov($S^2, \overline{X}$). The following results are well-known:

$$E(X_i - X_j)^2 = 2\theta_2$$

$$E[(X_i - X_j)^4] = 2\theta_4 + 6{\theta_2}^2 \quad (i \neq j)$$

$$E[(X_i - X_j)^2 (X_i - X_k)^2] = \theta_4 - {\theta_2}^2 \quad (i \neq j \neq k).$$

These moments are obtained by adding and subtracting μ to $X_i - X_j$ inside the expectations. Let us consider now alternative derivations of some important theorems based on formula in Equation 1 for the sample variance.

Theorem 2.1

The variance of S^2 is $\frac{1}{n}\left(\theta_4 - \frac{n-3}{n-1}\,{\theta_2}^2\right)$.

Proof: Appealing successively to the fact that

$$\text{Var}(S^2) = \text{Cov}(S^2, S^2)$$

$$= \text{Cov}\left[\frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2, \frac{1}{2n(n-1)}\sum_{k=1}^{n}\sum_{l=1}^{n}(X_k - X_l)^2\right]$$

$$= \frac{1}{4n^2(n-1)^2}\text{Cov}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2, \sum_{k=1}^{n}\sum_{l=1}^{n}(X_k - X_l)^2\right)$$

$$= \frac{1}{4n^2(n-1)^2}\sum_{i,j}\sum_{k,l}\text{Cov}\left((X_i - X_j)^2, (X_k - X_l)^2\right).$$

It thus remains only to evaluate all terms $\text{Cov}((X_i - X_j)^2, (X_k - X_l)^2)$ in the summation recognizing that these depend solely on the relations between the i, j, k, l and then counting how many of each type occur. Specifically, we need only to consider the cases when $i \neq j, k \neq l$ and then the issue becomes whether the (i, j) and (k, l) pairs have 0, 1 or 2 elements in common.

If i, j, k and l are all distinct i.e. the pairs do not have any element in common then $(X_i - X_j)^2$ and $(X_k - X_l)^2$ are independent. It means that there are $n(n-1)(n-2)(n-3)$ terms such that $\text{Cov}((X_i - X_j)^2, (X_k - X_l)^2) = 0$.

Following the same reasoning, we observe that there are $4n(n-1)(n-2)$ terms such that $\text{Cov}((X_i - X_j)^2, (X_i - X_l)^2) = \theta_4 - {\theta_2}^2$ representing pairs that have only one element in common.

If the two pairs are the same i=k, j=l, that is for terms $2n(n-1)$, we have that:

$$\text{Cov}((X_i - X_j)^2, (X_i - X_j)^2) = \text{Var}((X_i - X_j)^2)$$

$$= 2\theta_4 + 2{\theta_2}^2.$$

Putting this altogether give the result.

Using the alternative formula in Equation 1 we can compute:

$$\text{Cov}(\bar{X}, S^2) = \text{Cov}\left(\bar{X}, \frac{1}{2n(n-1)}\sum_{j=1}^{n}\sum_{k=1}^{n}(X_j - X_k)^2\right)$$

$$= \frac{1}{2n^2(n-1)}\text{Cov}\left(\sum_{i=1}^{n}X_i, \sum_{j=1}^{n}\sum_{k=1}^{n}(X_j - X_k)^2\right)$$

$$= \frac{1}{2n^2(n-1)}\sum_{i,j,k=1}^{n}\text{Cov}\left(X_i, (X_j - X_k)^2\right).$$

The only terms that contribute to the summation are those covariances where $j \neq k$ but i= j or i= k (otherwise gives 0). Here we have 2n(n-1) such terms:

$\text{Cov}(X_i, (X_j - X_k)^2) = E((X_i - \mu)^3) = \theta_3$ we have 2n(n-1) such terms.

Substitution then gives:

$$\text{Cov}(\bar{X}, S^2) = \theta_3/n$$

Observe that $\overline{X}$ and S^2 are uncorrelated if and only if $\theta_3 = 0$. This does not mean that $\overline{X}$ and S^2 are independent. Uncorrelation implies independence only when the sample $X_1, X_2, \ldots, X_n$ comes from a normal distribution [4].

It is shown in 2007 by Zhang [6] that using the alternative formula gets the result easer less calculations.

An example

A direct application of the formula above is that; if the population distribution is symmetric about its mean, also suppose that the third moment exists, then the covariance of the sample mean is zero. According to this result is possible the construction of numerous examples of "zero covariance without independence".

I have applied a no simple (trivial) example where the population distribution is Poisson with parameter λ and the sample size is n=2.

Let X_1 and X_2 be a sample of two independent observations drawn from a population having Poisson distributed. In the table below it is shown the distribution of the discrete random vector $(S^2, \overline{X})$. This table can extend to infinity but we can take just one case for example when,

Table 1. The distribution of the discrete random vector $(S^2, \overline{X})$

$\overline{X}$ / S^2	0	1/2	1	1 1/2
0	$e^{-2\lambda}$	0	$e^{-2\lambda}$ λ^2	0
1/2	0	$e^{-2\lambda}$ 2λ	0	$e^{-2\lambda}$ λ^2
2	0	0	$e^{-3\lambda}$ $\lambda^2/$ 2!	0

$$P(S^2 = \frac{1}{2}, \overline{X} = \frac{1}{2})$$

Let see:

$$P(S^2 = 0 | \overline{X} = 1) = e^{-2\lambda}\,\lambda^2 \neq P(X_1 = 0, X_2 = 0) + \cdots + \cdots = P(S^2 = 0)$$

Now, we convinced that we have zero covariance without independence".

The goal now is to prove that $\overline{X}$ and S^2 are independent using the alternative formula in Equation 1. Let $X_1, X_2, \ldots, X_n$ be independent, identically normally distributed random variables, i.e $X_i \sim N(\mu, \sigma^2)$. Let:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S_n{}^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2.$$

Recall that the chi-squared distribution $\chi^2(k)$ can be defined as the distribution of $U_1{}^2 + U_2{}^2 + \cdots U_k{}^2$ where $U_i \sim N(0,1)$ for $i = 1,2,\ldots,n$ are independent identically distributed standard normals.

Theorem 2.2

The following propositions are true if X_i are indipendent and identically distributed:

a) $\overline{X}_n$ has a $N(\mu, \frac{\sigma^2}{n})$ distribution

b) $\overline{X}_n$ and $S_n{}^2$ are independent.

Proof: Proceed by induction. First consider the case n = 2, that is $\overline{X}_2 = (X_1 + X_2)/2$ and using the alternative formula observe that $S_2{}^2 = (X_1 - X_2)^2/2$ and $(X_1 - X_2)/\sqrt{2}$ is N(0, 1). Point a) in the theorem is an immediate consequence of the assumed knowledge of normal distributions. Point b) follows from the definition of $\chi^2(1)$. Since $\text{cov}(X_1 - X_2, X_1 + X_2) = \text{cov}(X_1, X_1) - \text{cov}(X_2, X_2) = 0$, $X_1 + X_2$ and $X_1 - X_2$ are independent and b) follows.

Now assume the theorem holds for a sample of size n. We prove it holds for a sample of size $n + 1$. First establish the two relationships:

$$\overline{X}_{n+1} = (n\overline{X}_n + X_{n+1})/(n+1) \qquad \text{Equation 3}$$

$$S_{n+1}{}^2 = \left(\frac{n-1}{n+1}\right) S_n{}^2 + \frac{1}{n+1}\sum_{i=1}^{n+1}(X_{n+1} - X_i)^2. \qquad \text{Equation 4}$$

Now $\overline{X}_n$ and X_{n+1} are independent, and their distributions are $N(\mu, \frac{\sigma^2}{n})$ (by the induction hypothesis) and $N(\mu, \sigma^2)$, respectively. Hence $\overline{X}_{n+1}$ is a linear combination of two independent normal random variables, and a) follows by simply computing $E\,\overline{X}_{n+1}$ and $\text{Var}(\overline{X}_{n+1})$.

X_{n+1} is independent of $S_n{}^2$ and $\overline{X}_n$ is also independent of $S_n{}^2$ by the induction hypotheses. This shows that $\overline{X}_{n+1}$ is independent of $S_n{}^2$ and b) follows by noting that:

$$\text{Cov}(n\overline{X}_n + X_{n+1}, \sum_{i=1}^{n+1}(X_{n+1} - X_i)^2) = 0$$

The relationships in Equation 3 and 4 are themselves exercises in summation notation. The relation in Equation 4 is direct, as it is based on the useful consequence $\overline{X}_{n+1} - \overline{X}_n = (X_{n+1} - \overline{X}_n)/(n+1)$.

Formula in Equation 4 follows by expanding $S_{n+1}{}^2$ using the alternative formula:

$$\sum_{i=1}^{n+1}\sum_{j=1}^{n+1}(X_i - X_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 + \frac{1}{n(n+1)}\sum_{j=1}^{n}(X_{n+1} - X_j)^2.$$

This proof involves bivariate distribution using the alternative formula of S_n^2.

Uncorrelated variables are more complex than the independence and orthogonality. Using a geometric portrayal [3] we convince about the relation among independence and uncorrelated variables.

3. Conclusion

The results obtained from the use of the alternative formula for the sample variance coincided with those already existing in the literature but the derivation followed is different**.** The computations were straightforward and did not require advanced mathematical notions.

References

[1]. Casella, G. and Berger, R. L., (2001). Statistical Inference. (2nd ed.). Duxbury Press.

[2]. Mukhopadhyay, N., (2000). Probability and Statistical Inference, 162, 131.

[3]. Rogners, J. L., and Nicerwander, W. A., and Toothaker, L. (1984). Linearly Independent, Orthogonal, and Uncorrelated Vaariables. The American Statistician, 38, No.2, 133.

[4]. Shanmugam, R., (2007). Correlation between Sample Mean and Sample Variance. Journal of Modern Applied Statistical Methods, 7, 408-415.

[5]. Stigler, S. (1984). Kruskal's Proof of the Joint Distribution of $\overline{X}$ and S^2**.** The American Statistician, 38, No.2, 134-135.

[6]. Zhang, L. (2007). Sample Mean and Sample Variance: Their Covariance and Their Independence. The American Statistician, 63, No.2, 159-160.