A few properties of sample variance

Eric Benhamou

A few properties of sample variance

Eric Benhamou [*, †, ‡]

Abstract

A basic result is that the sample variance for i.i.d. observations is an unbiased estimator of the variance of the underlying distribution (see for instance Casella and Berger (2002)). Another result is that the sample variance 's variance is minimum compared to any other unbiased estimators (see Halmos (1946)). But what happens if the observations are neither independent nor identically distributed. What can we say? Can we in particular compute explicitly the first two moments of the sample mean and hence generalize formulae provided in Tukey (1957a), Tukey (1957b) for the first two moments of the sample variance? We also know that the sample mean and variance are independent if they are computed on an i.i.d. normal distribution. This is one of the underlying assumption to derive the Student distribution Student alias W. S. Gosset (1908). But does this result hold for any other underlying distribution? Can we still have independent sample mean and variance if the distribution is not normal? This paper precisely answers these questions and extends previous work of Cho, Cho, and Eltinge (2004). We are able to derive a general formula for the first two moments and variance of the sample variance under no specific assumptions. We also provide a faster proof of a seminal result of Lukacs (1942) by using the log characteristic function of the unbiased sample variance estimator.

AMS 1991 subject classification: 62E10, 62E15

Keywords: sample variance, variance of sample variance, independence between sample mean and variance

[*]A.I. SQUARE CONNECT, 35 Boulevard d'Inkermann 92200 Neuilly sur Seine, France
[†]LAMSADE, Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny,75016 Paris, France
[‡]E-mail: eric.benhamou@aisquareconnect.com, eric.benhamou@dauphine.eu

1. Introduction

Let $X_1, \ldots, X_n$ be a random sample and define the sample variance statistic as:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad X_n = (X_1, \ldots, X_n)^T \tag{1}$$

where $\bar{X}_n$ is the empirical mean, s_n^2 the empirical Bessel corrected empirical variance also called sample variance, and X_n the vector of the full history of this random sample.

We are interested in the first two moments of the sample variance as well as its relationship with the sample mean. A basic result is that the sample variance for i.i.d. observations is an unbiased estimator of the variance of the underlying distribution. But what happens if the observations are neither independent nor identically distributed. What can we say?

Can we in particular compute explicitly the first two moments of the sample variance without any particular assumptions on the sample? Can we generalize standard formula for the first two moments of the sample variance as provided in Tukey (1957a), Tukey (1957b). We also know that the sample mean and variance are independent if they are computed from an i.i.d. normal distribution. But what about any other underlying distribution? Can we still have independent sample mean and variance if the distribution is not normal for an i.i.d. sample? These are the motivations of this paper. This paper extends classical statistical results found in Cho et al. (2004) but also Tukey (1950), Tukey (1956), Tukey (1957a), Tukey (1957b). It is organized as follows. First we derive the fist two moments for the sample variance. We then examine the condition for the sample mean and variance to be independent. We show that it is only in the specific case of an underlying normal distribution that they are independent. We conclude on possible extensions.

2. Moment properties

2.1. symmetrical form of the sample variance

A first property that will be useful in the rest of th paper is the writing of the sample variance as a "U-statistic" (or symmetric) form as given by the following lemma

Lemma 2.1. *The sample variance can be defined as the average of the kernel* $h(x_1, x_2) = (x_1 - x_2)^2/2$ *over all* $n(n-1)$ *pairs of observations* (X_i, X_j) *for* $i \neq j$*:*

$$s_n^2 = \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{(X_i - X_j)^2}{2} = \frac{1}{n(n-1)}\sum_{i\neq j}\frac{(X_i - X_j)^2}{2} \tag{2}$$

Proof. See proof A □

This symmetric form for the sample variance helps us computing the various moments of the sample variance. Denoting by $\mu_k = \mathbb{E}\left[X^k\right]$ the various moment of the variable X and assuming that $(X_i)_{i=1,\dots,n}$ are n observations of the variable X (not necessarily independent), we can start computing the sample variance moments.

2.2. First moment of sample variance

Lemma 2.2. *The expectation of the sample variance is given by:*

$$\mathbb{E}\left[s_n^2\right] = \frac{\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]}{n} - \frac{\sum_{i\neq j} \mathbb{E}\left[X_i X_j\right]}{n(n-1)} \tag{3}$$

Hence if $(X_i)_{i=1,\dots,n}$ is independent and identically distributed, we get that s_n^2 is an unbiased estimator of the variance :

$$\mathbb{E}\left[s_n^2\right] = \mu_2 - \mu_1^2 = Var[X] \tag{4}$$

where $\mu_2 = \mathbb{E}\left[X^2\right]$ and $\mu_1 = \mathbb{E}\left[X\right]$.

Proof. See proof B.1 □

This lemma calls various remarks. First of all, the fact that for iid sample, the sample variance is unbiased is very well know (see for instance Casella and Berger (2002)). Secondly, the cross term $\mathbb{E}\left[X_i X_j\right]$ implies that this estimator will not be unbiased for correlated sample as the expectation can rewite as

$$\mathbb{E}\left[s_n^2\right] = \frac{\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]}{n} - \frac{\sum_{i\neq j} \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]}{n(n-1)} - \frac{\sum_{i\neq j} \left(\mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]\right)}{n(n-1)} \tag{5}$$

Hence for a non independent sample the term $\mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]$ does not cancel, while the first and second terms can be interpreted as estimator of the second and first moment of the sample. More generally, the interest of this general lemma is its application to non independent and non identically distributed samples.

2.3. application to AR(1)

Let us apply our result to a non independent sample. For instance, assume that the sample $(X_i)_{i=1,\dots,n}$ is generated by an auto regressive process of order 1 (AR(1)). We impose that the process is stationary with mean 0, variance $\frac{\sigma^2}{1-\rho^2}$ where σ is the variance of the underlying noise and ρ is the first order correlation. In this specific case, our general formula

provides the expectation of the sample variance. We find that the sample variance is biased and given by

$$\mathbb{E}\left[s_n^2\right] = \frac{\sigma^2}{1-\rho^2}\left(1 - \frac{2\rho}{(1-\rho)(n-1)} + \frac{2\rho(1-\rho^n)}{n(n-1)(1-\rho^2)}\right) \tag{6}$$

Proof. See proof B.2 □

2.4. Second moment of sample variance

Lemma 2.3. *The second moment of the sample variance is given by:*

$$\mathbb{E}\left[s_n^4\right] = \frac{\mathbb{E}[\hat{\mu}_4]}{n} - \frac{4\mathbb{E}[\hat{\mu}_1\hat{\mu}_3]}{n} + \frac{(n^2-2n+3)\mathbb{E}[\hat{\mu}_2^2]}{n(n-1)} - \frac{2(n-2)(n-3)\mathbb{E}[\hat{\mu}_1^2\hat{\mu}_2]}{n(n-1)} + \frac{(n-2)(n-3)\mathbb{E}[\hat{\mu}_1^4]}{n(n-1)} \tag{7}$$

where we have adopted the following notation:

$$\begin{aligned} \hat{\mu}_4 &= \frac{\sum_{i=1}^n X_i^4}{n} & \hat{\mu}_3\hat{\mu}_1 &= \frac{\sum_{i\neq j} X_i^3 X_j}{n(n-1)} & \hat{\mu}_2^2 &= \frac{\sum_{i\neq j} X_i^2 X_j^2}{n(n-1)} \\ \hat{\mu}_1^2\hat{\mu}_2 &= \frac{\sum_{i\neq j\neq k} X_i^2 X_j X_k}{n(n-1)(n-2)} & \hat{\mu}_1^4 &= \frac{\sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l}{n(n-1)(n-2)(n-3)} \end{aligned} \tag{8}$$

Proof. See proof B.3 □

Like previously, the expression for the second moment of the sample mean is very general and an extension of previous results. Its interest is precisely to apply without any restriction on the underlying observation. This generalizes in particular Cho et al. (2004), but also Tukey (1950), Tukey (1956), Tukey (1957a), and Tukey (1957b).

2.5. Variance of sample variance

Lemma 2.4. *The variance of the sample variance is given by:*

$$\begin{aligned} Var\left[s_n^2\right] &= \frac{\mathbb{E}[\hat{\mu}_4]}{n} - \frac{4\mathbb{E}[\hat{\mu}_1\hat{\mu}_3]}{n} + \frac{(n^2-2n+3)\mathbb{E}[\hat{\mu}_2^2]}{n(n-1)} - (\mathbb{E}[\tilde{\mu}_2])^2 \\ &\quad - \frac{2(n-2)(n-3)\mathbb{E}[\hat{\mu}_1^2\hat{\mu}_2]}{n(n-1)} + 2\mathbb{E}[\tilde{\mu}_1^2]\mathbb{E}[\tilde{\mu}_2] \\ &\quad + \frac{(n-2)(n-3)\mathbb{E}[\hat{\mu}_1^4]}{n(n-1)} - (\mathbb{E}[\tilde{\mu}_1^2])^2 \end{aligned} \tag{9}$$

where $\tilde{\mu}_2 = \dfrac{\sum_i X_i^2}{n}$ *and* $\tilde{\mu}_1^2 = \dfrac{\sum_{i \neq j} X_i X_j}{n(n-1)}$. *If the observations* $(X_i)_{i=1,\dots,n}$ *are independent and identically distributed and if we denote by* μ_i^c *the central moments of the population, this simplifies into:*

$$Var\left[s_n^2\right] = \frac{\mu_4^c}{n} - \frac{(n-3)(\mu_2^c)^2}{n(n-1)} \tag{10}$$

If the observations are from an i.i.d. normal distribution, this results in the traditional result

$$Var\left[s_n^2\right] = \frac{2\sigma^4}{n-1} \tag{11}$$

Proof. See proof B.4 □

Like previous results, equation (9) is the most general one and encompasses cases where observations are not necessarily independent nor identically distributed. To our knowledge, these results are new and give as a byproduct all standard results about the first, second and variance of the sample mean that can be found in textbook like Casella and Berger (2002).

3. Relationship between sample mean and variance

We finally tackle the question of the condition for the sample mean and variance to be independent. This is a strong result that for instance enables us to derive the Student distribution as in the normal case of iid variables, the sample mean and variance are clearly independent. We are interested in the opposite. What is the condition to impose on our distribution for iid variable to make our sample mean and variance independent? We shall prove that it is only in the case of normal distribution that these two estimators are independent as the following proposal states

Proposition 1. *The sample mean and variance are independent if and only if the underlying (parent) distribution is normal.*

Proof. This result was first proved by Geary (1936) and later by Lukacs (1942). We provide in C a proof that uses modern notations. It is an adaptation of the proof in Lukacs (1942) but with a simpler approach as we work with the log characteristic function and the unbiased sample variance. This makes the resulting differential equation trivial to solve as this is just a constant second order derivative constraint. □

This result implies consequently that it will not easy to derive the underlying distribution of the t-statistic for a non normal distribution. Indeed the t-statistic is defined as the ratio of the sample mean over the sample variance. If the sample mean and sample variance are

not independent, the computation of the underlying distribution does not decouple. This makes the problem of the computation of the underlying distribution an integration problem that has no closed form. This kills in particular our hope to derive other distribution that generalizes the case of the Student distribution to non normal underlying assumptions.

4. Conclusion

In this paper, we have derived the most general formula for the fist, second moment and variance of the sample variance. Our formula does not assume that the underlying sample is independent neither identically distributed. We also show that for an i.i.d. sample, the independence between the sample mean and variance is characteristic of the normal distribution. Possible extensions are to computer higher moments for the sample variance.

Appendix A. Symmetry for the sample variance

Let us first prove equation

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right) \tag{12}$$

with $\bar{X}_n$ defined by equation (1). Expanding the left hand side (LHS) leads to

$$\begin{aligned} LHS &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2+\bar{X}_n^2-2X_i\bar{X}_n) && (13)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2+n\bar{X}_n^2-2\bar{X}_n\sum_{i=1}^{n}X_i\right) && (14)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right) && (15)\end{aligned}$$

We want to prove equation (2):

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2=\frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{(X_i-X_j)^2}{2}=\frac{1}{n(n-1)}\sum_{i\neq j}\frac{(X_i-X_j)^2}{2}, \tag{16}$$

Note that the forms where we sum over all pairs of i,j and where we sum over all pairs that are different $i\neq j$ are equal (middle and right hand side) as the missing terms between the two sides are equal to zero. Some routine algebraic reduction on the middle hand side (MHS) gives:

$$\begin{aligned} MHS &= \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\frac{X_i^2+X_j^2-2X_iX_j}{2} && (17)\\ &= \frac{1}{n(n-1)}\frac{2n\sum_{i=1}^{n}X_i^2}{2}-\frac{1}{n(n-1)}\sum_{i,j=1}^{n}X_iX_j && (18)\\ &= \frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\frac{1}{n(n-1)}(n\bar{X}_n)^2 && (19)\\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}_n^2\right) && (20)\end{aligned}$$

We can easily conclude using equation (12) □

Appendix B. Moment of sample variance

B.1. First moment of sample variance

The result of lemma 2.2 is immediate expanding lemma 2.1 equality:

$$s_n^2 = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(X_i - X_j)^2}{2} \tag{21}$$

$$= \frac{1}{n(n-1)} \left((n-1) \sum_{i=1}^{n} X_i^2 - \sum_{i \neq j} X_i X_j \right), \tag{22}$$

and taking the expectation. The case of i.i.d. variables is also trivial as independence implies

$$\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j]$$

The identically distributed assumption implies $\mathbb{E}[X_i^2] = \mu_2$ and $\mathbb{E}[X_i] = \mu_1$. Finally, we have

$$\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] = n\mu_2, \quad \text{and} \quad \sum_{i \neq j} \mathbb{E}[X_i] \mathbb{E}[X_j] = n(n-1)\mu_1^2$$

□.

B.2. Application to AR(1)

Lemma 2.2 can be rewritten as

$$\mathbb{E}\left[s_n^2\right] = \frac{\sum_{i=1}^{n} \mathbb{E}[X_i^2]}{n} - \frac{\sum_{i \neq j} \mathbb{E}[X_i] \mathbb{E}[X_j]}{n(n-1)} - \frac{1}{n(n-1)} \sum_{i \neq j} \left(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] \right) \tag{23}$$

In the case of an AR(1) process, we have

$$\mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] = \frac{\sigma^2}{1-\rho^2} \rho^{|i-j|} \tag{24}$$

Hence, the term due to non independent is computed as follows:

$$\frac{1}{n(n-1)} \sum_{i \neq j} \left(\mathbb{E}[X_i X_j] - \mathbb{E}[X_i] \mathbb{E}[X_j] \right) = \frac{\sigma^2}{1-\rho^2} \frac{\sum_{i \neq j} \rho^{|i-j|}}{n(n-1)} \tag{25}$$

To conclude, one can use that

$$\sum_{i=1}^{n}(n-i)\rho^i = \rho\frac{n(1-\rho)-(1-\rho^n)}{(1-\rho)^2} \tag{26}$$

□.

B.3. Second moment of sample variance

Let us do some routine algebraic computation. We have

$$\begin{aligned} s_n^4 &= \frac{1}{n^2(n-1)^2}\left((n-1)\sum_{i=1}^{n}X_i^2 - \sum_{i\neq j}X_iX_j\right)^2 \qquad (27)\\ &= \frac{1}{n^2(n-1)^2}\left((n-1)^2(\sum_{i=1}^{n}X_i^2)^2 + (\sum_{i\neq j}X_iX_j)^2 - 2(n-1)(\sum_{i=1}^{n}X_i^2)(\sum_{k\neq l}X_kX_l)\right) \end{aligned}$$

Let us expand. The first expansion $(\sum_{i=1}^{n}X_i^2)^2$ is easy and immediate:

$$(\sum_{i=1}^{n}X_i^2)^2 = \sum_{i=1}^{n}X_i^4 + \sum_{i\neq j}X_i^2X_j^2 \tag{28}$$

In the expansion of $(\sum_{i\neq j}X_iX_j)^2$ we have that the squared terms are with same indexes $((i\neq j)=(k\neq l))$. The cross terms are $(i\neq j),(k\neq l)$ with the constraint that they are different $((i,j)\neq(k,l))$. There are three possibilities for these cross terms. These cross terms can either be only two real indexes $(i\neq j)$ and $(j\neq i)$ or vice versa leading to two times the squared terms, or we have that (i,j,k,l) are in fact only three numbers and this can happen 4 times as it is either j,j,k or l that coincides with the other indexes or there all different, and this can happen only once. Hence, we have:

$$(\sum_{i\neq j}X_iX_j)^2 = 3\sum_{i\neq j}X_i^2X_j^2 + 4\sum_{i\neq j\neq k}X_i^2X_jX_k + \sum_{i\neq j\neq k\neq l}X_iX_jX_kX_l \tag{29}$$

To expand $(\sum_{i=1}^{n}X_i^2)(\sum_{k\neq l}X_kX_l)$, we can notice that either there is no intersection of indexes between i,j and k, or i coincides with either k or l. And this can happen 4 times. So the expansion is given by

$$(\sum_{i=1}^{n}X_i^2)(\sum_{k\neq l}X_kX_l) = \sum_{i\neq j\neq k}X_i^2X_jX_k + 4\sum_{i\neq j}X_i^3X_j \tag{30}$$

Regrouping terms leads to

$$
\begin{aligned}
s_n^4 \;=\; & \frac{\frac{\sum_{i=1}^n X_i^4}{n}}{n} - \frac{\frac{4\sum_{i\neq j} X_i^3 X_j}{n(n-1)}}{n} + \frac{\frac{(3+(n-1)^2)\sum_{i\neq j} X_i^2 X_j^2}{n(n-1)}}{n(n-1)} \\
& - \frac{\frac{(2(n-1)(n-2)-4(n-2))\sum_{i\neq j\neq k} X_i^2 X_j X_k}{n(n-1)(n-2)}}{n(n-1)} + \frac{\frac{(n-2)(n-3)\sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l}{n(n-1)(n-2)(n-3)}}{n(n-1)}
\end{aligned} \tag{31}
$$

We can conclude by using the notation given in 8 and taking the expectation □

B.4. Variance of sample variance

The variance of the sample variance writes

$$
\mathrm{Var}[s_n^2] = \mathbb{E}[s_n^4] - (\mathbb{E}[s_n^2])^2. \tag{32}
$$

Using lemma 2.2, the square of the expectation of s_n^2 writes:

$$
\begin{aligned}
(\mathbb{E}\left[s_n^2\right])^2 &= \left(\mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}\right] - \mathbb{E}\left[\frac{\sum_{i\neq j} X_i X_j}{n(n-1)}\right]\right)^2 && (33) \\
&= \left(\mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}\right]\right)^2 - 2\mathbb{E}\left[\frac{\sum_{i\neq j} X_i X_j}{n(n-1)}\right]\mathbb{E}\left[\frac{\sum_{i=1}^n X_i^2}{n}\right] \\
&\quad + \left(\mathbb{E}\left[\frac{\sum_{i\neq j} X_i X_j}{n(n-1)}\right]\right)^2 && (34)
\end{aligned}
$$

Combining this result with 2.3, leads to equation (9). To use central moment, it is equivalent to impose that observations have null odd moments. Adding the condition of i.i.d observations implies that the following equalities:

$$
\hat{\mu}_4 = \mu_4^c \qquad \hat{\mu}_3\hat{\mu}_1 = 0 \qquad \hat{\mu}_2^2 = (\mu_2^c)^2 \tag{35}
$$

$$
\tilde{\mu}_2 = \mu_2^c \qquad \tilde{\mu}_1^2 = 0 \qquad \hat{\mu}_1^2\hat{\mu}_2 = 0 \qquad \hat{\mu}_1^4 = 0 \tag{36}
$$

Hence, we get

$$
\mathrm{Var}\left[s_n^2\right] = \frac{\mu_4^c}{n} - \frac{(n-3)(\mu_2^c)^2}{n(n-1)} \tag{37}
$$

If the observations are from an i.i.d. normal distribution with zero mean and a variance σ^2, we have $\mu_4^c = 3\sigma^4$ and $\mu_2^c = \sigma^2$ which leads to the result □.

Appendix C. Proof of the condition for sample mean and variance to be independent

The assumption of i.i.d. sample for $(x_1, \ldots, c_n)$ implies that the joint distribution of $(x_1, \ldots, x_n)$ denoted by $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ is equal to $\prod_{i=1}^{n} f_X(x_i)$, which we will write $\prod_{i=1}^{n} f(x_i)$ dropping the $._X$ to make notation lighter.

The log of the characteristic function of the joint variable $(\bar{X}_n, s_n^2)$ writes

$$\ln(\phi_{(\bar{X}_n,s_n^2)}(t_1, t_2)) = \ln\left(\iiint e^{it_1\bar{x}_n + it_2 s_n^2} \prod_{i=1}^{n} f(x_i)dx_i\right). \tag{38}$$

Similarly, the log of the characteristic function for the sample mean $\bar{X}_n$ writes

$$\ln(\phi_{\bar{X}_n}(t_1)) = \ln\left(\iiint e^{it_1\bar{x}_n} \prod_{i=1}^{n} f(x_i)dx_i\right), \tag{39}$$

and similarly for the sample variance

$$\ln(\phi_{s_n^2}(t_2)) = \ln\left(\iiint e^{it_2 s_n^2} \prod_{i=1}^{n} f(x_i)dx_i\right). \tag{40}$$

The assumption of independence between sample mean $\bar{X}_n$ and variance s_n^2 is equivalent to the fact that the characteristic function of the couple decouples, or that the log characteristic functions sum up.

$$\ln(\phi_{(\bar{X}_n,s_n^2)}(t_1, t_2)) = \ln(\phi_{\bar{X}_n}(t_1)) + \ln(\phi_{s_n^2}(t_2)). \tag{41}$$

Differentiating condition 41 with respect to t_2 in $t_2 = 0$ leads to

$$\frac{1}{\phi_{(\bar{X}_n,s_n^2)}(t_1, t2)} \left.\frac{\partial\phi_{(\bar{X}_n,s_n^2)}(t_1, t_2)}{\partial t_2}\right|_{t_2=0} = \frac{1}{\phi_{s_n^2}(t2)} \left.\frac{\partial\phi_{s_n^2}(t_2)}{\partial t_2}\right|_{t_2=0}. \tag{42}$$

Noticing that $\phi_{s_n^2}(0) = 1$ and $\phi_{(\bar{X}_n,s_n^2)}(t_1, 0) = \phi_{\bar{X}_n}(t_1)$, the condition 41 writes

$$\frac{1}{\phi_{\bar{X}_n}(t_1)} \left.\frac{\phi_{(\bar{X}_n,s_n^2)}(t_1, t_2)}{\partial t_2}\right|_{t_2=0} = \left.\frac{\partial\phi_{s_n^2}(t_2)}{\partial t_2}\right|_{t_2=0}. \tag{43}$$

Using the fact that $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$, it is easy to see that

$$\phi_{\bar{X}_n}(t_1) = \prod_{i=1}^{n} \int e^{it_1 x_i/n} f(x_i) dx_i = [\phi_X(t_1/n)]^n \tag{44}$$

For the sample variance, we can use the "U-statistic" (or symmetric) form as shown in lemma 2.1, to see that

$$s_n^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \frac{\sum_{i \neq j} X_i X_j}{n(n-1)} \tag{45}$$

Hence, the derivative of the characteristic function of the couple $(\bar{X}_n, s_n^2)$ writes

$$\left.\frac{\partial \phi_{(\bar{X}_n, s_n^2)}(t_1, t_2)}{\partial t_2}\right|_{t_2=0} = \iiint i s_n^2 \prod_{i=1}^{n} e^{it_1 x_i/n} f(x_i) dx_i \tag{46}$$

$$= i \iiint \left(\frac{\sum_{i=1}^n x_i^2}{n} - \frac{\sum_{i \neq j} x_i x_j}{n(n-1)} \right) \prod_{i=1}^{n} e^{it_1 x_i/n} f(x_i) dx_i \tag{47}$$

$$= i[\phi_X(\frac{t_1}{n})]^{n-2} \left(\phi_X(\frac{t_1}{n}) \int x^2 e^{\frac{it_1 x}{n}} f(x) dx - (\int x^2 e^{\frac{it_1 x}{n}} f(x) dx)^2 \right) \tag{48}$$

In the latter equation, if we set $t_1 = 0$, we get in particular that

$$\left.\frac{\partial \phi_{s_n^2}(t_1, t_2)}{\partial t_2}\right|_{t_2=0} = \left.\frac{\partial \phi_{(\bar{X}_n, s_n^2)}(0, t_2)}{\partial t_2}\right|_{t_2=0} = i\sigma^2 \tag{49}$$

Hence, condition (43) writes

$$\frac{\phi_X(\frac{t_1}{n}) \int x^2 e^{\frac{it_1 x}{n}} f(x) dx - (\int x^2 e^{\frac{it_1 x}{n}} f(x) dx)^2}{[\phi_X(\frac{t_1}{n})]^2} = \sigma^2 \tag{50}$$

We also have that the derivative of the characteristic function $\phi_X(\frac{t_1}{n})$ with respect to $u = t_1/n$ gives

$$\frac{\partial \phi_X(u)}{\partial u} = \int ix e^{ixu} f(x) dx \tag{51}$$

To simplify notation, we drop the index in ϕ_X and writes this function ϕ. Using equation (51), condition (50) writes

$$\frac{-\phi(u) \frac{\partial^2 \phi(u)}{\partial u^2} + \left(\frac{\partial \phi(u)}{\partial u} \right)^2}{\phi(u)^2} = \sigma^2 \tag{52}$$

The log of the characteristic function of $\phi(u) = \mathbb{E}[e^{iuX}]$, denoted by $\Psi(u) = \ln \phi(u)$, first

and second derivatives with respect to u are given by:

$$\frac{\partial \Psi(u)}{\partial u} = \frac{\partial \ln \phi(u)}{\partial u} = \frac{1}{\phi(u)} \frac{\partial \phi(u)}{\partial u} \tag{53}$$

$$\frac{\partial^2 \Psi(u)}{\partial^2 u} = \frac{\partial}{\partial u} \frac{\partial \Psi(u)}{\partial u} = \frac{1}{\phi(u)} \frac{\partial^2 \phi(u)}{\partial u^2} - \frac{1}{\phi(u)^2} \left(\frac{\partial \phi(u)}{\partial u} \right)^2 \tag{54}$$

Hence, condition (52) writes

$$\frac{\partial^2 \Psi(u)}{\partial^2 u} = -\sigma^2 \tag{55}$$

Using the boundary conditions $\Psi(0) = 0$ and $\Psi^{'}(0) = i\mathbb{E}[X] = i\mu$, it is easy to integrate condition 55 which is a constant second order derivative to get

$$\Psi(u) = i\mu u - \frac{\sigma^2 u^2}{2} \tag{56}$$

Condition 56 states that a necessary and sufficient condition for the sample mean and variance to be independent is that the log characteristic function of X is a quadratic form. But a quadratic form for the log characteristic function of X is a characterization of a normal distribution, which concludes the proof. □

References

Casella, G., Berger, R., 2002. Statistical inference.

Cho, E., Cho, M. J., Eltinge, J., 2004. The variance of sample variance from a finite population. Proceedings of Joint American Statistical Association and International Statistical Institute Conference, Toronto .

Geary, R. C., 1936. Distribution of student's ratio for nonnormal samples. Royal Statistical Society Journal, Supplement 3.

Halmos, P. R., 1946. The theory of unbiased estimation. Ann. Math. Stat. pp. 34–43.

Lukacs, E., 1942. A characterization of the normal distribution. Ann. Math. Statist. 13, 91–93.

Student alias W. S. Gosset, 1908. The probable error of a mean. Biometrika pp. 1–25.

Tukey, J. W., 1950. Some sampling simplified? Journal of the American Statistical Association pp. 501–519.

Tukey, J. W., 1956. Variances of variance components i. The Annals of Mathematical Statistics pp. 722–736.

Tukey, J. W., 1957a. Variances of variance components ii. The Annals of Mathematical Statistics pp. 43–56.

Tukey, J. W., 1957b. Variances of variance components iii. The Annals of Mathematical Statistics pp. 378–384.