On the Independence of Sample Mean and Variance

[1]Anwar H. Joarder and [2]M. Hafidz Omar
[1]School of Business, University of Liberal Arts Bangladesh, Dhanmondi, Dhaka 1209, Bangladesh,
Email: anwar.joarder@ulab.edu.bd
[2]Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals Dhahran 1261,
Saudi Arabia, Email: omarmh@kfupm.edu.sa

Abstract

We present two new proofs of the independence of sample mean and variance for two independently, identically and normally distributed random variables.

Key Words: Sample mean, sample variance, independence, moment generating function

Manuscript received on 31 July 2016 and accepted for publication on 10 September 2016.

1 Introduction

The independence of sample mean and variance of independently, identically and normally distributed variables is essential in the basic definition of Student t-statistic [1] and also in the development of many statistical methods. For a sample of size n, it is usually proved using the independence of $\bar{X}$ and $(X_1-\bar{X},\ X_2-\bar{X},\ \ldots,X_n-\bar{X}),$ (see e.g. Theorem 1, p.340, [2]), but this requires background on independence of functions of random variables (Theorem 2, p.121, [2]) that may not be easily accessible to beginning undergraduate students.

There are several proofs of the independence of sample mean and variance, the popular of which seems to be the one due to[3]. It uses moment generating function of chi-square variable conditional on sample mean and as such not straightforward. Therefore, proofs accessible to undergraduate students have been an issue of discussion. See for example, [4] and also American Statistician, 1992, Volume 46, No. 1, pp. 72-75.

In this note, we present two new proofs. The first one is a direct proof of independence of sample mean and variance but for a sample of size 2 with a view to shedding light on the topic and to inspire students and instructors. Though we use moment generating function but unlike Shuster's proof we avoid the use of conditional distributions. The second one seems to be simplest of all other proofs though it requires the evaluation of a double integral and a sense of origin invariance of sample variance.

2 The Importance of the Result

The derivation of t-distribution by [1] required the independence of sample mean and variance. Though it was not clearly mentioned in the paper, [5] figured out that the independence of sample mean and variance was implicitly assumed. The independence of mean and variance was proved by [3] using a mathematical tool provided by Fisher. This happened one year before Student passed away. [7] points that Helmert proved that sample mean and variance are independent. Because of this historical fact, [8] recommended calling the joint distribution of the two random variables "Helmert's Distribution". Lukacs (1942) presented an easier proof of the result:

If the variance (or second moment) of a population distribution exists, then a necessary and sufficient condition for the normality of the population distribution is that sample mean and variance are mutually independent. Since then many authors including [8], [9], [10], [11], [3], [4] and [12] came up with different proofs and characterizations.

The independence of mean and variance extends to a broader class of distributions when the iid restriction of the sample is relaxed. For example, it is well known that if $(X_1, X_2, ..., X_n)$ has an *n*-variate normal distribution with an exchangeable covariance pattern, that is, $V(X_j) = \sigma^2$,

$Cov(X_j, X_k) = \rho\sigma^2, \ -\frac{1}{n-1} < \rho < 1,$ < 1, for $j = 1, 2, ..., n$ and $k = 1, 2, ..., n$

(c.f. [10], pp. 196–197). [13], in an expository article, also demonstrated independence of mean and variance based on non-iid samples from populations with specific mixture structures. It has been proved, see for example, [14] that *t*-statistic based on elliptically symmetric distributions has Student *t*-distribution. The *t*-statistic based on many skew normal distributions has also Student *t*-distribution. See, for example, [15].

Also *t*-statistics based on a joint distribution proposed by [16] follow Student *t*-distribution. [17] came up with vertical density representation that includes a brod class of distributions that guarantees that *t*-statisic has a Student *t*-distribution.

The above proves that normailty is not a necessity as has been thought of over the decades.

However, even with all these developments, we feel there is no direct and simple proof of the independence of mean and variance.

3 Main Results

Let X_1 and X_2 have an arbitrary 2-dimensional joint distribution. We define the sample mean $\bar{x}$ and variance s^2 by $n\bar{x} = x_1 + x_2$ and $2s_x^2 = (x_1 - x_2)^2$ respectively. Since $\bar{X} \sim N(\mu, \ \sigma^2/n),$ the moment generating function $M_{\bar{X}}(t_1) = E[\exp(t_1\bar{X})]$ of $\bar{X}$ is known to be

$M_{\bar{X}}(t_1) = \exp\left(t\mu + \frac{t_1^2}{2} \times \frac{\sigma^2}{n}\right)$, which simplifies to $M_{\bar{X}}(t_1) = \exp\left(t_1\mu + \frac{t_1^2\sigma^2}{4}\right)$ for a sample of size 2.

The moment generating function $M_{S^2}(t_2) = E[\exp(t_2 S^2)]$ of S^2 with degrees of freedom $m = 1$ is given by

$$M_{S^2}(t_2) = E\left[\exp\left(\frac{mS^2}{\sigma^2}\frac{t_2\sigma^2}{m}\right)\right] = M_U\left(\frac{t_2\sigma^2}{m}\right),$$

where, U has a χ^2 with $m = 1$ degrees of freedom. Then for $n = 2,$ we have,

$$M_{S^2}(t_2) = \frac{1}{\sqrt{1 - 2t_2\sigma^2}}, \text{ where } -\infty < t_2 < \frac{m}{2\sigma^2}.$$

For the history of the distribution of sample variance or of chi-square based on normal distribution, see [6].

Theorem 3.1 Let the random variables X_1 and X_2 be independently, identically and normally distributed with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Then the joint moment generating function of the sample mean $\bar{X}$ and variance S_X^2 satisfies $M_{\bar{X}, S_X^2}(t_1, t_2) = M_{\bar{X}}(t_1) M_{S_X^2}(t_2)$ and hence $\bar{X}$ and S_X^2 are independent.

First Proof. For two observations $2s_x^2 = (x_1 - x_2)^2$. The joint moment generation function of $\bar{X} = (X_1 + X_2)/2$ and s_x^2 is given by $M_{\bar{X}, s_X^2}(t_1, t_2) = E(e^{t_1\bar{X} + t_2 s_X^2})$. Letting $X = \mu + \sigma Z$, or, $X_j = \mu + \sigma Z_j$, $j = 1, 2,$ we have $M_{\bar{X}, s_X^2}(t_1, t_2) = E\exp[t_1(\mu + \sigma\bar{Z}) + t_2(\sigma^2 s_Z^2],$ where, $X_j = \mu + \sigma Z_j$, $j = 1, 2.$ Obviously $s_Z^2 = \frac{1}{2}[(\mu + \sigma Z_1) - (\mu + \sigma Z_2)]^2 = \frac{1}{2}(Z_1 - Z_2)^2$, and $f_{Z_j}(z_j)$ is the density function of standard normal variable Z_j, $j = 1, 2.$ Then we have,

$$M_{\bar{X}, S^2}(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left[t_1(\mu + \sigma\bar{z})] + t_2\sigma^2 s_Z^2\right] f_{Z_1}(z_1) f_{Z_2}(z_2) dz_1 dz_2,$$

which can be written as,

$$M_{\bar{X},S^2}(t_1,t_2)=e^{t_1\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[\frac{1}{2}t_1\sigma(z_1+z_2)+\frac{1}{2}t_2\sigma^2(z_1-z_2)^2\right]f_{Z_1}(z_1)f_{Z_2}(z_2)dz_1dz_2. \qquad (3.1)$$

Since the exponent of the exponential function in the above integrand is

$$q=\frac{1}{2}t_1\sigma(z_1+z_2)+\frac{1}{2}t_2\sigma^2(z_1-z_2)^2-\frac{1}{2}(z_1^2+z_2^2),$$

(3.1) can be written as

$$M_{\bar{X},S^2}(t_1,t_2)=e^{t_1\mu}\exp\left(\frac{t_1^2\sigma^2}{4(1-t_2\sigma^2)}\right)\int_{-\infty}^{\infty}f_W(w)M_Y(-t_2\sigma^2w)dw, \qquad (3.2)$$

where, $W\sim N\left(\frac{t_1\sigma}{2(1-t_2\sigma^2)},\ \frac{1}{1-t_2\sigma^2}\right)$ and $Y\sim N\left(\frac{t_1\sigma}{2(1-t_2\sigma^2)},\ \frac{1}{1-t_2\sigma^2}\right)$.

Having written out $M_Y(-t_2\sigma^2w)$ and used in (3.2), we have

$$M_{\bar{X},S^2}(t_1,t_2)=\exp\left(t_1\mu+\frac{t_1^2\sigma^2}{4}\right)\frac{1}{\sqrt{1-2t_2\sigma^2}}I(t_2,\sigma),$$

where, $I(t_2,\sigma)=\int_{-\infty}^{\infty}\frac{\sqrt{1-2t_2\sigma^2}}{\sqrt{2\pi(1-t_2\sigma^2)}}\exp\left[-\frac{1-2t_2\sigma^2}{2(1-t_2\sigma^2)}\left(y-\frac{t_1\sigma}{2}\right)^2\right]dy.$

The above integral is 1 as the integrand is the density function of $Y\sim N\left(\frac{t_1\sigma}{2},\ \frac{1-t_2\sigma^2}{1-2t_2\sigma^2}\right)$.

Thus we have,

$$M_{\bar{X},S^2}(t_1,t_2)=\exp\left(t_1\mu+\frac{t_1^2\sigma^2}{4}\right)\frac{1}{\sqrt{1-2t_2\sigma^2}},$$

which is the product of $M_{\bar{X}}(t_1)=\exp\left(t_1\mu+\frac{t_1^2\sigma^2}{4}\right)$ and $M_{S^2}(t_2)=\frac{1}{\sqrt{1-2t_2\sigma^2}}$.

That is, $M_{\bar{X},S^2}(t_1,t_2)=M_{\bar{X}}(t_1)M_{S^2}(t_2)$. By uniqueness property of moment generating function, it implies the independence of sample mean and variance.

The proof for n observations seems to be straightforward but it is quite involved.

Second Proof. Without any loss of generality, we may assume that $\mu=0$ and $\sigma=1.$ Then the joint density function of the sample is given by $f(x_1,x_2)=\frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1^2+x_2^2)\right)$, so that the joint moment generating function of $\bar{X}$ and s_X^2 is given by

$$M_{\bar{X},s_X^2}(t_1,t_2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(t_1\bar{x}+t_2s_x^2)\times\frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1^2+x_2^2)\right)dx_1dx_2,$$

which can be written as,

$$M_{\bar{X},s_X^2}(t_1,t_2)=\exp\left(\frac{t_1^2}{4}\right)\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(t_2s_x^2)\exp\left[-\frac{1}{2}\left(x_1-\frac{t_1}{2}\right)^2-\frac{1}{2}\left(x_2-\frac{t_1}{2}\right)^2\right]dx_1dx_2.$$

Then, with the transformation $u_i = x_i - t_1/2,\ (i=1,2),$ we have $s_x^2 = s_u^2,\ dx_1dx_2 = du_1du_2$, and hence

$$M_{\bar{X},s_X^2}(t_1,t_2) = \exp(t_1^2/4)I(t_2;u), \tag{3.3}$$

$$\text{where } I(t_2;u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}\exp\left(t_2 s_u^2\right)\exp\left(-\frac{1}{2}(u_1^2+u_2^2)\right)du_1du_2.$$

The above integral is exactly the same as,

$$I(t_2;x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}\exp\left[\frac{1}{2}t_2(x_1-x_2)^2\right]\exp\left(-\frac{1}{2}(x_1^2+x_2^2)\right)dx_1dx_2 = M_{S_X^2}(t_2).$$

Since $M_{\bar{X}}(t_1) = \exp(t_1^2/4),$ it follows from (3.3), that $M_{\bar{X},s_X^2}(t_1,t_2) = M_{\bar{X}}(t_1)M_{s_X^2}(t_2).$ By uniqueness property of the moment generating function, this proves that the sample mean and variance are independent. The proof for n observations seems to be straightforward.

4. Conclusion

The independence of sample mean and variance has been proved in Section 3 for a sample of size 2. Since the proofs are simple and direct, it will make students and instructors confident about the fundamental theorems of statistics. An open problem is to generalize the proofs of this paper to any sample size. Another challenging problem would be to identify the family of distributions that enjoy the property of independence of sample mean and variance. This will prove the robustness of *t*-test under broader distributional assumptions.

Acknowledgement

The authors gratefully acknowledge the research facilities provided by King Fahd University of Petroleum and Minerals, Saudi Arabia through the Project IN131015.

References

[1] Student. The probable error of a mean. *Biometrika,* 6: 1-25, 1908.

[2] Rohatgi, V.K. and Saleh, A.K.M.E. *Probability and Statistics*. Wiley, 2001.

[3] Shuster, J. A simple method of teaching the independence of $\bar{X}$ and S^2. *The American Statistician*, 27(1), 29-30, 1973.

[4] Zehna, P.W. On Proving that $\bar{X}$ and S^2 are independent. *The American Statistician*, 45(2), 121-122, 1991

[5] Fisher, R.A. Student. *Annals of Eugenics*, 9, 1-9, 19390

[6] Geary, R.C. The distribution of Student's ratio for non-normal samples. *Supplement to the Royal Statistical Society*, 3, 178-184, 1936.

[7] Johnson, N.L.; Kotz, S. and Balakrishnan, N. Continuous Multivariate Distributions, v-2. New York: John Wiley and Sons, 1994.

[8] Kruskal, W.H. Helmert's distribution. *American Mathematical Monthly*, 53, 435-438, 1946.

[9]Zinger, A.A. On independent samples from a normal population. Uspekhi Matem. Nauk, 6:5(45), 172-175, 1951.

[10] Kagan, A.M.; Linnik, Yu. V. and Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*. John Wiley and Sons: New York, 1973.

[11] Rao, C. R. *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley, 1973

[12]Stigler, Stephen M. Kruskal's proof of the joint distribution of $\bar{X}$ and S^2. *The American Statistician*, 46, 72-73, 1992.

[13] Mukhopadhyay, N., "Dependence or Independence of the Sample Mean and Variance in Non-iid or Nonnormal Cases and the Role of Some Tests of Independence," in *Recent Advances in Applied Probability, IWAP 2002 116 Teacher's Corner* Downloaded by [32.209.230.42] at 18:03 09 September 2014 *Caracas Proceedings*, eds. R. Baeza-yates, J. Glaz, J. Husler, and J. L. Palacios, New York: Springer-Verlag, pp. 397–426, 2005.

[14] Fang, K.T; Kotz, S. and Ng, K.W. *Symmetric Multivariate and Related Distributions.* Chapman and Hall, 1990.

[15] Gupta, A.K.; Omar, M. Hafidz and Joarder, A.H. On a Generalized Mixture of Standard Normal and Skew Normal Probability Distributions. *Pan American Mathematics Journal,* 23(3), 1-14, 2013.

[16] Gupta, Rameshwar D. and Richards, Donald St. P. Multivraite Liouville Distributions –V. 377-396. Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz. Eds: Norman L. Johnson and N. Balakrishnan (1997). John Wiley and Sons, New York, 1997.

[17] Yang, Zhenhai; Fang, Kai -Tai and Kotz, Samuel. On the Student's *t*-distribution and the *t*-statistic. *Journal of Multivariate Analysis*, 98, 1293-1304, 2007.

Anwar H. Joarder obtained PhD and MS degrees in Statistics from the University of Western Ontario in 1992 and 1988, respectively. Currently he is a professor at University of Liberal Arts Bangladesh (ULAB). Before joining ULAB, he worked at many institutions including the University of Western Ontario, Monash University and University of Sydney. He serves on the editorial board of about 10 international research journals. He has authored or co-authored about 80 methodical papers published in internationally reputed research journals. His research area includes robustness of statistical methods, correlated statistical analysis, multivariate parametric estimation, survey sampling and pedagogical statistics. He has been a lifelong Fellow of Royal Statistical Society, England, since 2005.

M. Hafidz Omar obtained PhD degree in Educational Measurement and Statistics in 1995 from the University of Iowa, USA. He obtained MS degree in Statistics in 1992 and BS in Actuarial Science in 1990 from the same university. He worked in companies and taught statistics at several institutions in the USA including the University of Iowa, the University of Kansas, and Riverside Publishing Company. Currently he is an Associate Professor of Statistics and Actuarial Science at King Fahd University of Petroleum and Minerals. He has refereed many research papers from reputable international journals. He has authored or co-authored more than 50 papers published in internationally reputed research journals. His research area includes applied multivariate statistics, correlated statistical analysis, probability modeling, mathematical statistics, educational Statistics, actuarial statistics, and statistics education.