Sample Mean and Sample Variance: Their Covariance and Their (In)Dependence

Lingyun ZHANG

It is of interest to know what the covariance of sample mean and sample variance is without the assumption of normality. In this article we study such a problem. We show a simple derivation of the formula for computing covariance of sample mean and sample variance, and point out a way of constructing examples of "zero covariance without independence." A small example is included to help teachers explain to students.

KEY WORDS: Bernoulli distribution; Non-normal sample; Zero covariance without independence.

1. INTRODUCTION

We define sample mean $\bar{X} = \sum_{i=1}^{n} X_i/n$ and sample variance $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, where $\{X_1, X_2, \ldots, X_n\}$ comprises a random sample from some a population. It is well known that $\bar{X}$ and S^2 are independent if the population is normally distributed. Now, naturally we can ask a question: Are $\bar{X}$ and S^2 independent without the assumption of normality?

The answer to the above question is "No" according to the following theorem found in Lukacs (1942).

Theorem: 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 $\bar{X}$ and S^2 are mutually independent.

Remark: That the normality is a necessary condition for the independence between $\bar{X}$ and S^2 was first proved by Geary (1936) using a mathematical tool provided by R. A. Fisher, but the proof in Lukacs (1942) is easier to understand.

This theorem is beautiful. The theorem shows that the independence between $\bar{X}$ and S^2 is unique—it holds only for normally distributed populations (provided that the second moment of the population distribution exists). If the population is normally distributed, the covariance of $\bar{X}$ and S^2, denoted by $\text{cov}(\bar{X}, S^2)$, is 0 because of their independence; but generally what is $\text{cov}(\bar{X}, S^2)$? This article is to answer such a question.

The rest of the article is organized as follows. In Section 2, we show a simple derivation of the formula for computing covariance of $\bar{X}$ and S^2, followed by a small example in Section 3.

2. COVARIANCE OF $\bar{X}$ AND S^2

Proposition: If the third moment of the population distribution exists, then

$$\text{cov}(\bar{X}, S^2) = \frac{\mu_3}{n}, \tag{1}$$

where μ_3 is the third central moment of the population distribution and n is the sample size.

Formula (1) was published by Dodge and Rousson (1999), and to comment on its elegance we quote a sentence from the article: "... statistical theory provides beautiful formulas when they involve the first three moments (with a special prize for the insufficiently known formula $\text{cov}(\bar{X}, S^2) = \mu_3/n$)..." No proof of the formula was given by Dodge and Rousson (1999); the following is our derivation of (1).

Derivation of (1): Let μ and σ^2 denote the population mean and variance, respectively.

$$\begin{aligned}\text{cov}(\bar{X}, S^2) &= E(\bar{X}S^2) - E(\bar{X})E(S^2)\\ &= E\Big((\bar{X} - \mu + \mu)S^2\Big) - \mu\sigma^2\\ &= E\Big((\bar{X} - \mu)S^2\Big).\end{aligned}$$

Let $Y_i = X_i - \mu$, for $i = 1, 2, \ldots, n$; and denote the sample mean and variance of the Y_i by $\bar{Y}$ and S_Y^2, respectively. Then

$$\begin{aligned}\text{cov}(\bar{X}, S^2) &= E(\bar{Y}S_Y^2)\\ &= \frac{1}{n(n-1)}E\left(\sum_{i=1}^{n} Y_i\left[\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2\right]\right)\\ &= \frac{1}{n(n-1)}\left[E\left(\sum_{i=1}^{n} Y_i \sum_{j=1}^{n} Y_j^2\right)\right.\end{aligned}$$

Lingyun Zhang is Lecturer, IIST, Massey University, Private Box 756, Wellington, New Zealand (E-mail: *l.y.zhang@massey.ac.nz*). The author thanks the editor, Professor Peter Westfall, an associate editor, and a referee for valuable comments and suggestions which have helped to improve the quality of this article. He also thanks his colleague Dr. Mark Bebbington for proofreading an earlier version of this article and useful suggestions.

 DOI: 10.1198/000313007X188379

$$-\frac{1}{n}E\left(\sum_{i=1}^{n}Y_i\left(\sum_{j=1}^{n}Y_j\right)^2\right)\Bigg]$$

$$\equiv \frac{1}{n(n-1)}(I_1 - I_2), \qquad (2)$$

where

$$\begin{aligned} I_1 &= E\left(\sum_{i=1}^{n}Y_i\sum_{j=1}^{n}Y_j^2\right) \\ &= E\left(\sum_{i=1}^{n}Y_i^3\right) \\ &= n\mu_3, \end{aligned}$$

and

$$\begin{aligned} I_2 &= \frac{1}{n}E\left(\sum_{i=1}^{n}Y_i\left(\sum_{j=1}^{n}Y_j\right)^2\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n}Y_i\left(\sum_{j=1}^{n}Y_j^2 + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n}Y_jY_k\right)\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n}Y_i^3\right) \\ &= \mu_3. \end{aligned}$$

Substituting I_1 and I_2 into Equation (2) completes the proof.

A direct application of formula (1) is that, if the population distribution is symmetric about its mean (also suppose that its third moment exists), then the covariance of the sample mean and variance is 0. According to this result and the theorem in Section 1, we can construct numerous examples of "zero covariance without independence."

3. AN EXAMPLE

With a wish to help teachers explain to students, we apply (1) to a simple case, where the population distribution is Bernoulli and sample size $n = 2$.

Let X_1 and X_2 be a sample of two independent observations drawn from a population having a Bernoulli distribution with parameter p $(0 < p < 1)$. The sample mean and sample variance now can be written down as

Table 1. The joint probability distribution of $\bar{X}$ and S^2.

$(\bar{X}, S^2)$	$(0, 0)$	$(1/2, 1/2)$	$(1, 0)$
Probability	$(1-p)^2$	$2p(1-p)$	p^2

$$\bar{X} = \frac{X_1 + X_2}{2}, \quad \text{and} \quad S^2 = \frac{(X_1 - X_2)^2}{2}.$$

By the two equations and because of the population having a Bernoulli distribution, we can easily obtain the joint probability distribution of $\bar{X}$ and S^2, which is summarized in Table 1.

The third central moment of X_1 is equal to $p(1-p)(1-2p)$ (see Johnson, Kotz, and Kemp 1992, p. 107, or derive it directly.) According to (1), we have

$$\operatorname{cov}(\bar{X}, S^2) = \frac{p(1-p)(1-2p)}{2}. \qquad (3)$$

We see from (3) that:

- $\operatorname{cov}(\bar{X}, S^2) > 0$, if $p < \frac{1}{2}$;
- $\operatorname{cov}(\bar{X}, S^2) = 0$, if $p = \frac{1}{2}$;
- $\operatorname{cov}(\bar{X}, S^2) < 0$, if $p > \frac{1}{2}$.

A by-product of the above discussion (no need of the aid of the theorem in Section 1) is an example of "zero covariance without independence." To produce such an example we simply let $p = 1/2$, in which case $\operatorname{cov}(\bar{X}, S^2) = 0$. However, $\bar{X}$ and S^2 are not independent because

$$\Pr(S^2 = 0 | \bar{X} = 1) = 1 \neq \frac{1}{2} = \Pr(S^2 = 0). \quad \text{(by using Table 1).}$$

[Received May 2006. Revised January 2007.]

REFERENCES

Dodge, Y., and Rousson, V. (1999), "The Complications of the Fourth Central Moment," *The American Statistician*, 53, 267–269.

Geary, R. C. (1936), "The Distribution of 'Student's' Ratio for Non-Normal Samples," *Supplement to the Journal of the Royal Statistical Society*, 3, 178–184.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992), *Univariate Discrete Distributions* (2nd ed.) New York: Wiley.

Lukacs, E. (1942), "A Characterization of the Normal Distribution," *The Annals of Mathematical Statistics*, 13, 91–93.