Boletín de Estadística e Investigación Operativa
Vol. 26, No. 3, Octubre 2010, pp. 189-191

Estadística

A simple proof of Fisher's Theorem and of the distribution of the sample variance statistic

Luis María Sánchez-Reyes and Luis González Abril

Departamento de Economía Aplicada I
Universidad de Sevilla
✉ luiss-rf@us.es, luisgon@us.es

Abstract

In this paper a very simple and short proofs of Fisher's theorem and of the distribution of the sample variance statistic in a normal population are given.

Keywords: Normal population, Chi-squared population.
AMS Subject classifications: 62F99.

1. Introduction

Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random vector such that $E[X_i] = \mu$ and $Cov[X_i, X_j] = \sigma^2\delta_{ij}$ with $\sigma^2 > 0$ and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. For $n \geq 2$, the mean of order k $(1 \leq k \leq n)$, defined as $\overline{X}_k = \frac{1}{k}\sum_{i=1}^{k} X_i$, is considered, which verifies:

I. $E\left[\overline{X}_k\right] = \mu$, since $E\left[\overline{X}_k\right] = E\left[\frac{1}{k}\sum_{i=1}^{k} X_i\right] = \frac{1}{k}\sum_{i=1}^{k} E\left[X_i\right] = \mu$.

II. $Cov\left[\overline{X}_\ell, \overline{X}_k\right] = \frac{\sigma^2}{k}$ for any $1 \leq \ell \leq k \leq n$ because

- If $\ell = k$ then $Cov\left[\overline{X}_\ell, \overline{X}_k\right] = Var[\overline{X}_k] = \frac{1}{k^2}\sum_{i=1}^{k} Var[X_i] = \frac{\sigma^2}{k}$.
- If $\ell < k$ then $Cov\left[\overline{X}_\ell, \overline{X}_k\right] = \frac{1}{k}Cov\left[\overline{X}_\ell, X_1 + \ldots + X_\ell + \ldots + X_k\right] = \frac{1}{k}Cov\left[\overline{X}_\ell, X_1 + \ldots + X_\ell\right] = \frac{1}{k}Cov\left[\overline{X}_\ell, \ell\overline{X}_\ell\right] = \frac{\ell}{k}Var\left[\overline{X}_\ell\right] = \frac{\sigma^2}{k}$.

Let $\mathbf{Y} = (Y_2, \cdots, Y_n)$ be a random vector where

$$Y_k = \frac{\sqrt{k(k-1)}}{\sigma}\left(\overline{X}_k - \overline{X}_{k-1}\right) \tag{1.1}$$

for any $2 \leq k \leq n$. The most relevant properties of $\mathbf{Y}$ are:

III. $E[Y_k] = 0$ since $E\left[\overline{X}_k - \overline{X}_{k-1}\right] = \mu - \mu = 0$ from (I).

IV. $Cov[Y_\ell, Y_k] = \delta_{k\ell}$, since from (II),

- If $\ell = k$ then $Var\left[\overline{X}_k - \overline{X}_{k-1}\right] = Var[\overline{X}_k] - 2\,Cov[\overline{X}_k, \overline{X}_{k-1}] + Var[\overline{X}_{k-1}] = \sigma^2\left(\frac{1}{k} - \frac{2}{k} + \frac{1}{k-1}\right) = \frac{\sigma^2}{k(k-1)}$, which implies that $Var[Y_k] = 1$
- For $2 \le \ell < k \le n$, then $Cov\left[\overline{X}_\ell - \overline{X}_{\ell-1}, \overline{X}_k - \overline{X}_{k-1}\right] =$ $Cov\left[\overline{X}_\ell, \overline{X}_k\right] - Cov\left[\overline{X}_\ell, \overline{X}_{k-1}\right] - Cov\left[\overline{X}_{\ell-1}, \overline{X}_k\right] + Cov\left[\overline{X}_{\ell-1}, \overline{X}_{k-1}\right]$ $= \sigma^2\left(\frac{1}{k} - \frac{1}{k-1} - \frac{1}{k} + \frac{1}{k-1}\right) = 0$, which implies that $Cov[Y_\ell, Y_k] = 0$.

V. $Cov[Y_\ell, \overline{X}_k] = 0$ for $2 \le \ell \le k \le n$ as:
$Cov\left[\overline{X}_\ell - \overline{X}_{\ell-1}, \overline{X}_k\right] = Cov\left[\overline{X}_\ell, \overline{X}_k\right] - Cov\left[\overline{X}_{\ell-1}, \overline{X}_k\right] = \frac{\sigma^2}{k} - \frac{\sigma^2}{k} = 0.$

VI. The statement $\sigma^2 \sum_{k=2}^{n} Y_k^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ holds for $n \ge 2$. This equality is proven by induction on n.
For $n = 2$: $\overline{X}_2 = \frac{1}{2}(X_1 + X_2)$ and as $\overline{X}_1 = X_1$, it follows that:
$\left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2 = 2\left(\frac{X_2 - X_1}{2}\right)^2 = 2\left(\frac{X_2 + X_1 - 2X_1}{2}\right)^2 = 2\left(\overline{X}_2 - \overline{X}_1\right)^2 = \sigma^2 Y_2^2$
Assume the equality holds for n.
$\sum_{i=1}^{n+1} (X_i - \overline{X}_{n+1})^2 = \sum_{i=1}^{n} (X_i - \overline{X}_{n+1})^2 + (X_{n+1} - \overline{X}_{n+1})^2 =$
$= \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + n(\overline{X}_n - \overline{X}_{n+1})^2 + (X_{n+1} - \overline{X}_{n+1})^2$.
But $X_{n+1} = (n+1)\overline{X}_{n+1} - n\overline{X}_n$ and therefore $(X_{n+1} - \overline{X}_{n+1})^2 = n^2(\overline{X}_{n+1} - \overline{X}_n)^2$. Hence $\sum_{i=1}^{n+1} (X_i - \overline{X}_{n+1})^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + (n^2 + n)(\overline{X}_{n+1} - \overline{X}_n)^2 =$ (applying the induction hypothesis) $= \sigma^2 \sum_{k=2}^{n} Y_k^2 + \sigma^2 Y_{n+1}^2 = \sigma^2 \sum_{k=2}^{n+1} Y_k^2$, and the equality holds for $n + 1$.

From these results, it follows that:

Proposition 1.1. *If the population model X is normal, then $nS^2/\sigma^2 \sim \chi^2_{n-1}$, where S^2 is the sample variance statistic of $\mathbf{X}$ and χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom.*
Proof. $\mathbf{Y} = (Y_2, \cdots, Y_n)$ with Y_k defined in (1.1) is normal since it is obtained from $\mathbf{X}$ by a linear transformation. Furthermore, from (VI): $\frac{nS^2}{\sigma^2} = \sum_{k=2}^{n} Y_k^2 \sim \chi^2_{n-1}$ because the Y_k random variables are normal, their mean is zero (III), their variance is one (IV), and they are independent since they are uncorrelated (IV).

Theorem 1.1 (Theorem of Fisher). *The statistics $\overline{X}$ and S^2 are independent if the population model is normal.*
Proof. The $(Y_2, \cdots, Y_n, \overline{X}_n)$ vector is normal since it is obtained from $\mathbf{X}$ by a linear transformation and it follows from (II), (IV) and (V) that the variance-covariance matrix of this vector is diagonal which determinant equals to σ^2/n. Hence, its joint density function is

$$f_{(\mathbf{Y}, \overline{X}_n)}(y_2, \cdots, y_n, \overline{x}_n) = \frac{\sqrt{n}}{(2\pi)^{n/2}\sigma} e^{-\frac{1}{2}\left(\sum_{i=2}^{n} y_i^2 + \frac{n}{\sigma^2}(\overline{x}_n - \mu)^2\right)}$$

which can be written as

$$\frac{1}{(2\pi)^{(n-1)/2}} e^{-\frac{1}{2}\left(\sum_{i=2}^{n} y_i^2\right)} \frac{\sqrt{n}}{(2\pi)^{1/2}\sigma} e^{-\frac{1}{2}\left(\frac{n}{\sigma^2}(\overline{x}_n-\mu)^2\right)} = f_{\mathbf{Y}}(y_2,\cdots,y_n)\cdot f_{\overline{X}_n}(\overline{x}_n)$$

where the first function is the joint density function of the vector $(Y_2,\cdots,Y_n)$ and the second function is the marginal density function of the variable $\overline{X}_n$. Therefore, $\mathbf{Y}$ and $\overline{X}_n$ are independent and so is $\overline{X}_n$ of any transformation of the $\mathbf{Y}$ vector. Thus, as $S^2 = \frac{\sigma^2}{n}\sum_{k=2}^{n} Y_k^2$, the independence between the mean and the sample variance is proved.

About the authors

Luis María Sánchez Reyes is a Lecturer in Department of Applied Economics I in the University of Seville. He obtained the M.Sc. degree in Mathematics from the University of Seville. His research interests include statistical methodology and classification methods.

Luis González Abril is a Lecturer in Department of Applied Economics I in the University of Seville. He obtained the M.Sc. degree in Mathematics and the Ph.D. degree in Economics from the University of Seville. His research interests include machine learning, statistical methodology and classification methods.