On the dispersion of data in nonsymmetric distributions

L. BARONE
Department of Mathematics. University of Lecce, Via Arnesano,
73 100 – Lecce, I - Italy
Email: barone@ingle01.unile.it

G. Z. VOULGARIDIS
Okeanidon 19, 117 45 - Ag. Sostis, Athens,Greece
Email: g_voulgaridis@yahoo.com

A.H. JOARDER
Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
Email: anwarj@kfupm.edu.sa

The use of the left and right variance is proposed and an index of asymmetry based on them is introduced. Several examples demonstrate its usefulness.

1. Introduction

The question of evaluating more accurately the dispersion of a random variable emerges in all non-symmetric probability distributions. When the population distribution is non-symmetric, the mean and variance (or standard deviation) do not provide a precise idea of the shape and symmetry of the distribution. It is argued that the average, the proposed left variance (or left standard deviation) and right variance (or right standard deviation) describe them more accurately.

Let X be a real random variable with F as its cumulative distribution function so that $F(x) = P(X < x)$, for every real number x. Then the mean μ mean (average) is defined by

$$\mu = E(X) = \int_{-\infty}^{+\infty} x \; dF .$$

The median of X, denoted by $\tilde{\mu}$ is generally defined by the inequalities:

$$F(\tilde{\mu}) = P(X < \tilde{\mu}) \leq \tfrac{1}{2} \leq P(X \leq \tilde{\mu}) = F(\tilde{\mu}+)$$

where $F(\tilde{\mu}+)$ is the limit from the right of $F(x)$ in $\tilde{\mu}$. If F is continuous, its median is unique and equal to the value $\tilde{\mu}$ for which $F(\tilde{\mu}) = \tfrac{1}{2}$.

In the case of a discrete random variable X, mode (μ_0) is defined as the value (or values) for which $f(x)$ attains its maximum value. When X is an absolutely continuous random variable with probability density $f(x)$, mode is defined as the point (or points) where $f(x)$ presents the maximum. If there is a unique mode of X, then X is called unimodal.

A random variable X is said to have a symmetric distribution about its average μ, if and only if, for every real number x:

$$F(x+\mu) = P(X-\mu < x) = P(-(X-\mu) < x) = P(X > \mu - x) = 1 - F((\mu - x)+)$$

If X is absolutely continuous, with probability density function $f(x)$, then its distribution is symmetric if and only if $f(x+\mu) = f(\mu - x)$. Usually, there is no relation between a distribution being symmetric and the equality $\mu = \tilde{\mu} = \mu_0$. This is demonstrated in the following examples:

Example 1.1 Consider the following probability mass function **:**

$$P(X=1) = P(X=3) = P(X=5) = 1/9,\ P(X=2) = P(X=4) = 3/9$$

This distribution is certainly symmetric about $\mu = 3$, even though it is bimodal. The mean (or the expected value) of the distribution is $\mu = \sum xp(x) = 3$ while mode occurs at 2 and 4. The median is given by $\tilde{\mu} = 3$ since $P(X<3) = 4/9 < 0.5$ and $P(X \le 3) = 5/9 \ge 0.5$.

Example 1.2 Consider the following probability mass function **:**

$$P(X=8) = P(X=12) = P(X=25) = P(X=35) = 1/7,\ P(X=20) = 3/7$$

This distribution is certainly non-symmetric, even though $\mu = \tilde{\mu} = \mu_0$. The mean (or the expected value) of the distribution is $\mu = \sum xp(x) = 20$, the mode is $\mu_0 = 20$ and the median is given by $\tilde{\mu} = 20$ since $P(X<20) = 2/7 < 0.5$ and $P(X \le 20) = 5/7 \ge 0.5$.

Example 1.3 Consider the following uniform probability mass function **:**

$$P(X = x) = 1/4,\ x = 1, 2, 3, 4$$

The median is given by $\tilde{\mu} = 2$ and 3 since $P(X<2) = 1/4 < 0.5$, $P(X \le 2) = 2/4 \ge 0.5$, and $P(X<3) = 2/4 \le 0.5$, $P(X \le 3) = 3/4 \ge 0.5$. The mean (or the expected value) of the distribution is $\mu = \sum xP(X = x) = 2.5$. This distribution is symmetric, even though $\tilde{\mu} \ne \mu$.

If the distribution is unimodal and the median is defined by the central element of the closed interval, then the distribution is symmetric (with respect to μ), if and only if, it results that $\mu = \tilde{\mu} = \mu_0$. If the distribution is not symmetric, the asymmetry is called right if $\mu_0 \le \tilde{\mu} \le \mu$, and the asymmetry is called left if $\mu \le \tilde{\mu} \le \mu_0$. Even if $\mu = \tilde{\mu} = \mu_0$, the distribution is not always symmetric (see Example 1.2). But it is not adequate to know that a distribution is non-symmetric. In order to compare different non-symmetric distributions, one needs a measurement of the asymmetry of each distribution. Various measurements of asymmetry have been proposed; the most remarkable ones are described below:

(i) Pearson's Index of "skewness"

It is given by

$$\beta_1 = (\mu - \mu_0)/\sigma, \tag{1}$$

where $\sigma^2 = E(X - \mu)^2$ is the second central moment of a random variable X. This index is independent of the measurement units and indeed it is ideal for comparing asymmetry between different distributions. Since mode is difficult to determine for sample, the following measure has been more popular.

(ii) Hotelling and Solomons measure of Skewness

Hotelling and Solomons (1932) proved that $|\mu - \tilde{\mu}| < \sigma$, and over time the following measure of skewness

$$\beta_2 = \frac{\mu - \tilde{\mu}}{\sigma/3} \tag{2}$$

which varies in $[-3,3]$ became popular. In sample $\mu, \tilde{\mu}$ and σ are replaced by their counterparts $\bar{x}, \tilde{x}$ and s, and we denote the measure by b_2.

(iii) Fisher's Index

It is given by

$$\beta_3 = \mu_3/\sigma^3, \tag{3}$$

where $\mu_3 = E(X - \mu)^3$ is the third order central moment. Pearson's Index given by β_1 historically precedes Fisher's index β_3. In sample μ_3 and σ are replaced by their sample counterparts $m_3 = \sum(x - \bar{x})^3/n$ and s where $(n-1)s^2 = \sum(x - \bar{x})^2$. Let $z = (x - \bar{x})/s$ be the *z*-score of an observation x. Then the resulting quantity

$$b_3 = \frac{1}{n-1}\sum\left(\frac{x - \bar{x}}{s}\right)^3 = \frac{1}{n-1}\sum z^3 \tag{4}$$

appears to be the third order standardised moment, and is the Fisher's index of asymmetry in a sample.

It is reminded that all moments of odd order ($k = 3, 5, 7, \ldots$) are proper to evaluate the asymmetry. In fact, they are zero when the distribution is symmetric, positive when the distribution presents asymmetry on the right and negative when the distribution presents asymmetry on the left. Note that all the measures of skewness are unit free to allow them to compare samples with different units. If $0 < |b_3| < 1/2$, the asymmetry is insignificant, if $1/2 < |b_3| < 1$, the asymmetry is moderate and if $|b_3| > 1$, the asymmetry is strong .

In this paper the use of left variance and the right variance is proposed, and consequently an index of asymmetry is proposed. The idea is illustrated by some examples

2. The left and right variance

In this study, we propose the use of the left and the right variance, σ_l^2 and σ_r^2 respectively as follows:

If $\mu = E(X)$ and $\sigma^2 = V(X)$, the "left variance σ_l^2 of *X*" is defined by

$$\sigma_l^2 = E\left((X-\mu)^2 I(X \le \mu)\right), \tag{5}$$

and "right variance σ_r^2 of *X*" by

$$\sigma_r^2 = E\left((X-\mu)^2 I(X \ge \mu)\right). \tag{6}$$

where *I* is the indicator function.

In addition σ_l^2/σ^2 will be called standard left variance and σ_r^2/σ^2 will be called standard right variance. The β index of dispersion is introduced by

$$\beta = \frac{\sigma_r^2}{\sigma^2} - \frac{\sigma_l^2}{\sigma^2} = \frac{\sigma_r^2 - \sigma_l^2}{\sigma^2} \tag{7}$$

If $\beta = 0$, then f is symmetric around the average μ. If $-1 < \beta \le 0$, then f is non-symmetric, and the distribution is more dispersed on the left than on the right of the average μ. If $0 \le \beta < 1$, then f is non-symmetric, and the distribution is more dispersed on the right than on the left of the average μ. It should be remarked here that even if the distribution is non-symmetric, its data can be equally dispersed about the average.

The proposed index of skewness for the sample is given by

$$b = \frac{s_r^2}{s^2} - \frac{s_l^2}{s^2} = \frac{s_r^2 - s_l^2}{s^2} = \frac{1}{n_2 - 1}\sum_{z \ge 0} z^2 - \frac{1}{n_1 - 1}\sum_{z \le 0} z^2 \tag{8}$$

where

$$s_l^2 = \frac{1}{n_1 - 1}\sum_{x \le \bar{x}} (x - \bar{x})^2 \quad \text{and} \quad s_r^2 = \frac{1}{n_2 - 1}\sum_{x \ge \bar{x}} (x - \bar{x})^2 . \tag{9}$$

If $x_i \ge \bar{x}$ then $z \ge 0$ so that (i) $z^2 \le z^3$ when $z \ge 1$, and (ii) $z^2 > z^3$ otherwise. Therefore, a comparison between b_3 and b appears to be difficult. Let us describe some example in which $b < b_3$ and one in which $b > b_3$.

Example 2.1 Consider the sample: 8, 12, 20, 20, 20, 25, 35. The z-scores are approximately given by

$$-1.373485669, -0.915657112, 0, 0, 0, 0.572285695, 1.716857086$$

so that $b \approx 0.818777292 - 0.681222707 \approx 0.1376$ and $b_3 \approx 0.3149$.

Example 2.2 Consider the data x: 0.08 0.12 0.2 0.2 0.2 0.25 0.35. Since the data are obtained by dividing the data in Example 2.1, b and b_3 remains the same as in Example 2.1.

Example 2.3 For the data $x: -9, -3, -3, 0, 2, 4, 9$, the z-scores are approximately given by

$$-1.558845727,\ -0.519615242,\ -0.519615242,\ 0,\ 0.346410161,\ 0.692820323, 1.558845727$$

so that $b = 1.01 - 0.99 = 0.02$ and $b_3 \approx 0.3149$.

Example 2.3 For the data $x: -1.85, -0.62, 0, 0, 0, 0.99, 1.48$ we the z-scores are approximately given by

$$-1.715538793, -0.574937325, 0, 0, 0, 0.918045084, 1.078378412$$

so that $b \approx 0.818406569 - 0.501426693 \approx 0.3170$ and $b_3 \approx 0.3149$.

Proposition 2.1 Let n_1 be the number of observations not exceeding the sample mean and n_2 be the number of observations that are at least as large as the sample mean. Then the following inequality hold

$$s^2 \le s_l^2 + s_r^2 \le \frac{n}{\nu} s^2 \text{ where } \nu = \min\{n_1, n_2\}$$

where $(n-1)s^2 = \sum\left(x - \bar{x}\right)^2$, and s_l^2 and s_r^2 are defined in (9).

Proof. Since

$$\begin{aligned}(n-1)s^2 &= \sum\left(x-\bar{x}\right)^2 \\ &= \sum_{x \le \bar{x}}\left(x-\bar{x}\right)^2 + \sum_{x \ge \bar{x}}\left(x-\bar{x}\right)^2 \\ &= (n_1 - 1)s_l^2 + (n_2 - 1)s_r^2\end{aligned}$$

and $0 \le \frac{n_1 - 1}{n-1} \le 1$ and $0 \le \frac{n_2 - 1}{n-1} \le 1$, it follows that

$$s^2 = \frac{n_1 - 1}{n-1} s_l^2 + \frac{n_2 - 1}{n-1} s_r^2 \le s_l^2 + s_r^2 .$$

Again since

$\frac{1}{n_1}\sum_{x\le\bar{x}}\left(x-\bar{x}\right)^2 \le \frac{1}{v}\sum_{x\le\bar{x}}\left(x-\bar{x}\right)^2$ and $\frac{1}{n_2}\sum_{x\ge\bar{x}}\left(x-\bar{x}\right)^2 \le \frac{1}{v}\sum_{x\ge\bar{x}}\left(x-\bar{x}\right)^2$

it follows that

$$s_l^2+s_r^2 \le \frac{1}{v}\sum(x_i-\bar{x})^2 \le \frac{n}{v}\frac{1}{n-1}\sum(x_i-\bar{x})^2 = \frac{n}{v}s^2 \ .$$

Hence the proof.

The asymmetry that many distributions present, has been a subject of study by many authors and various index of asymmetry have been proposed. The use of left and right variance permits us to describe in a more accurate way the dispersion of data in these distributions. In addition, the index of dispersion, proposed in the present paper, provides us not only with the sense of asymmetry of a distribution, but also with whether this asymmetry is significant or not. In comparison to the Fisher’s index, this is more immediate as it needs fewer computations. Considering the fact that the computation of Fisher’s index needs more multiplication, which could amplify the propagation of errors, the proposed index could be considered more accurate.

It seems it is a difficult job to study the random sampling distribution of the proposed index of asymmetry. Further, the study of its moments, properties and percentile points remains open to be studied.

Acknowledgements

The third author is grateful to the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia**.**

References

Hotelling, H., and Solomons, L.M. (1932). The limits of a measure of skewness. *The Annals of Mathematical Statistics*, 3, 141-142.