A robust measure of skewness using cumulative statistic calculation

Mario Schlemmer

Abstract

An important aspect of the shape of a distribution is the level of asymmetry. Strong asymmetries play a role in many ecosystems and are found in the size and reproductive success of individuals. But the standard third moment coefficient of skewness has the drawback that it is very sensitive to outliers, which can lead to incorrect interpretations. A new metric is introduced that is based on calculating the cumulative statistics of the Lorenz curve framework, but it evaluates the asymmetry of the underlying distribution. The standard requirements for skewness measures which the proposed measure satisfies are briefly described and it is compared to the moment-based measure using the lognormal distribution with and without outliers. The results demonstrate that the proposed measure behaves similarly for ‘normal’ distributions, but is robust (not overly sensitive) if there are outliers.

Keywords: asymmetry; Lorenz curve; robustness; size distribution

Introduction

If both sides of a univariate continuous distribution mirror one another the distribution is symmetric. If the left tail is longer it is called left-skewed, and right-skewed if the right tail is stretched out more. But if there are values that lie outside the overall pattern of the distribution the classical moment coefficient of skewness $b_1$ lacks robustness. Consider the example dataset that contains the length/width ratio of 88 butter clams collected at Puget Sound (Langkamp & Hull, 2022). The dataset is summarized by the boxplot in Figure 1, without the outlier the sample is normally distributed and symmetric. With the outlier $b_1 = 1.49$, a value that indicates a distribution that is highly skewed to the right, but if the outlier is removed $b_1 = -0.04$. Similarly, $b_1$ would indicate a heavily left-skewed distribution if there would be a very low value in the dataset instead of the high value.

Several statistics that are not affected by outliers have been suggested, among them the well-known Paerson mode skewness and the closely related medcouple (Brys et al.,
2004). The main drawback of these resistant statistics that use the median and quantiles is that they are also not sensitive to extreme tail behavior which can be vital to capturing information about the shape of the distribution. A measure that strikes a balance between tail sensitivity and robustness is based on the Lorenz curve, for which the values are sorted by size and the cumulative proportion of observations(e.g.butter clams) is plotted on the x-axis against the cumulative proportion of measured ratios(y-axis).

This approach provides a graphical representation of inequality among species and individuals that can be used for size, biomass, or fecundity. In economics, it is also used to represent the distribution of assets or the inequality in water and energy consumption. The Lorenz curve coincides with one diagonal of the unit square if there is no dispersion in the dataset($p = q$), this diagonal is sometimes called the 45-degree line. If the Lorenz curve is symmetric both sides mirror each other along the other diagonal of the unit square, known as the axis of symmetry(Damgaard, & Weiner, 2000). But generally, a symmetric Lorenz curve does not imply that the underlying distribution of values is symmetric. It has been pointed out that the most useful content of the Lorenz curve is its distance, $p - q$ from the 45-degree line(Clementi et al., 2019).

**Cumulative Skew**

We assume that the sample of a univariate dataset consists of $n$ independent observations $X_n = x_1, x_2, x_3, ..., x_n$, say the size of individuals, the cumulative proportion of individuals is denoted as $p_i$, the cumulative size as $q_i$. The vertical distances $D_i = d_1, d_2, d_3, ..., d_{n-1}$ between the Lorenz curve and the 45-degree line are calculated by subtracting each $q_i$ from the corresponding $p_i$. If the distribution of $X_n$ is symmetric, then distances at opposite sides of the curve match one another, such a case is presented in Figure 2 which also shows that if the number of distances is uneven,
then the distance in the middle has no counterpart. To obtain a measure of skewness positive and negative weights denoted $w_i$ are attached to the distances that lie above and below the median of $p_i$, respectively. No weight is attached to the distance that corresponds to the median. To standardize the sum of weighted distances it is divided by the sum of unweighed distances.

![Figure 2: Distances between the Lorenz curve and the 45-degree line for a symmetric distribution.](image)

If $r_i$ denotes the ranks of values in the cumulative distribution function, then cumulative skew($CS$) is given by:

$$CS = \frac{\sum(D_i \times w_i)}{\sum D_i}, \text{ where}$$

$$w_i = (2 \times r_i - n) \times 3/n$$

The weights are bounded by $[-3, 3]$ and $CS$ by $[-1, 1]$. For finite $n$ the skewness measure is bounded by $[-1 + 2/n, 1 - 2/n]$.

The proposed measure $CS$ meets the four requirements for an appropriate skewness measure that have been described in the literature (Groeneveld & Meeden, 1984; Brys et al., 2004). These properties are:

**Property 1:** $CS$ is scale and location invariant, i.e.

$$F = a \times F + b, \text{ for constants } a > 0 \text{ and } b$$
**Property 2:** Inverting an asymmetric distribution changes the sign of the skewness.

\[ F(x \times -1) = -F(x) \]

These two properties imply Property 3:

**Property 3:** For a symmetric distribution \( F = 0 \).

**Property 4:** If \( F <_c G \), then \( CS(F) \leq CS(G) \)

\( CS \) respects the \( c \)-ordering as defined by van Zwet(1964). A class of distributions that satisfy this \( c \)-ordering are Tukey’s g-distributions which are calculated from values that are drawn from a normal distribution(Brys et al., 2004; Hoaglin et al, 2011). I calculated g-distributions \( G_g \) with different parameters \( g(G_{0.1},G_{0.2},...,G_{1.5}) \) and found \( G_{g1} \) to \( c \)-precede \( G_{g2} \) for any \( g_1 < g_2 \). In Figure 3 I plotted the skewness for g-distributions with different values for parameter \( g \). Skewness is increasing faster if there is more variability in the underlying normal distribution, but both lines show a monotonic increase of measured skewness.

![Figure 3: Monotone increasing values of CS as a function of g.](image)

**Robustness**

Let us first consider symmetric distributions, property 3 ensures that the probability of negative and positive values is the same if the values are drawn from a distribution that has no bias. I analyzed the behavior for 100,000 samples of size \( n = 100 \) from a normal distribution and the mean skewness was \(-0.000925691\) with a standard error of 0.000135877. For the heavy-tailed Cauchy distribution the mean was \(0.0029298121\) with a standard error of \(0.001287708\).
The proposed skewness measure $CS$ is more robust against outliers than the conventional measure $b_1$. For the complete butter clam dataset presented in Figure 1 $CS = 0.06$, and when the obvious outlier is removed $CS = -0.02$. To compare the robustness of $CS$ with that of the conventional measure $b_1$ the average skewness for 10,000 random samples of size $n = 200$ was calculated for lognormal distributions with different parameter $\sigma$, in addition to two conditions that contaminated lognormal distributions with weak and strong outliers. The lognormal distribution is widely used in ecology to model non-Gaussian dynamics like population abundance and reproductive success of individuals. It is the probability distribution of a random variable whose logarithm is normally distributed. A crucial property of the lognormal distribution is that it is close to a normal distribution for small values of the shape parameter $\sigma$ and gets more right-skewed as $\sigma$ increases.

| condition         | $b_1(ave)$ | $CS(ave)$ |
|-------------------|------------|-----------|
| 1. $\sigma = 0.2$| 0.586      | 0.104     |
| 2. $\sigma = 0.5$| 1.598      | 0.241     |
| 3. $\sigma = 1.0$| 3.724      | 0.448     |
| 4. $\sigma = 2.0$| 7.635      | 0.741     |
| 5. $\sigma = 0.5$ outliers(high) | 4.589      | 0.385     |
| 6. $\sigma = 1.0$ outliers(low)   | $-1.275$   | 0.153     |

Table 1. Evaluation of skewness.

In Table 1 we see that both $CS$ and $b_1$ behave similarly in the first four conditions. As expected, distributions get more and more right-skewed with larger $\sigma$. Condition 5 simulated the effect of 1 to 5 large values in the distribution (0.5% to 2.5% of the sample). Here the value of $b_1$ is much larger than in condition 2 which has the same $\sigma$ but no outliers. The effect of the contamination on $CS$ is more moderate. In condition 6 the outliers were 1 to 5 low values. Here the effect of outliers on the average value of $b_1$ is so strong that it now indicates that the distributions are left-skewed, although the vast majority of values were sampled from the heavily right-skewed lognormal distribution ($\sigma = 1$). The new measure is again more robust if there are values that lie outside the overall pattern of the distribution. The strongest differences between the two measures for this sample size were observed if there is only a single strong outlier. This robustness against outliers together with the fulfillment of the four requirements described above and the intuitive appeal through its connection with the Lorenz framework suggests that the new measure $CS$ is a viable alternative to existing measures of skewness.
References

Brys, G., Hubert, M., & Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics, 13*(4), 996-1017.

Clementi, F., Gallegati, M., Gianmoena, L., Landini, S., & Stiglitz, J. E. (2019). Mis-measurement of inequality: a critical reflection and new insights. *Journal of Economic Interaction and Coordination, 14*(4), 891-921.

Damgaard, C., & Weiner, J. (2000). *Describing inequality in plant size or fecundity. Ecology, 81*(4), 1139-1142.

Groeneveld, R. A., & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician), 33*(4), 391-399.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (Eds.). (2011). *Exploring data tables, trends, and shapes.* John Wiley & Sons.

Langkamp, G. & Hull, J. (2022). QELP Data Set 002. [online] Seattlecentral.edu. Available at: https://seattlecentral.edu/qelp/sets/002/002.html [Accessed 26 September 2022]

Zwet, W.R. van (1964). *Convex Transformations of Random Variables,* Mathematisch Centrum: Amsterdam.