Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 620-635
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Shape Measures for the Distribution of a Qualitative Variable

José Moral de la Rubia

School of Psychology, Universidad Autónoma de Nuevo León, Nuevo León, Monterrey, México
Email: jose.morald@uanl.edu.mx

How to cite this paper: Moral de la Rubia, J. (2023) Shape Measures for the Distribution of a Qualitative Variable. *Open Journal of Statistics*, **13**, 620-635.
https://doi.org/10.4236/ojs.2023.134030

Received: July 28, 2023
Accepted: August 20, 2023
Published: August 23, 2023

Abstract

There are several shape measures for quantitative variables, some of which can also be applied to ordinal variables. In quantitative variables, symmetry, peakedness, and kurtosis are essential properties to evaluate the deviation from assumptions, particularly normality. They aid in selecting the most appropriate method for estimating parameters and testing hypotheses. Initially, these properties serve a descriptive role in qualitative variables. Once defined, they can be considered to check for non-compliance with assumptions and to propose modifications for testing procedures. The objective of this article is to present three measures of the shape of the distribution of a qualitative variable. The concepts of qualitative asymmetry and peakedness are defined. The measurement of the first concept involves calculating the average frequency difference between qualitative categories matched by frequency homogeneity or proximity. For the second concept, the peak-to-shoulder difference and the qualitative percentile kurtosis are taken into consideration. This last measurement is a less effective option than the peak-to-shoulder difference to measure peakedness. A simulated example of the application of these three measures is given and the paper closes with some conclusions and suggestions.

Keywords

Symmetry, Peakedness, Descriptive Measures, Nominal Measurement Scale, Qualitative Variables

1. Introduction

The measurement of the shape of a distribution has a descriptive interest and an application to assess the effect of non-compliance with assumptions and develop corrections, as well as to create measures of some aspects of empirical reality [1].

Many measures of shape have been developed with quantitative variables that take special relevance with respect to the normality assumption [2], even applied to ordinal variables [3]. For the moment, it has been considered that the shape of the distribution is not important for qualitative variables and it is not possible to propose a measurement approach [4]. However, recently, there has been a proposal for the definition and measurement of skewness and peakedness in distributions of qualitative variables in scientific literature in Spanish [5] [6].

This article presents this new conceptual and measurement approach, combines both concepts, and gives exemplified application for wider dissemination. It begins with the definition of qualitative skewness and its measurement by the average frequency difference between qualitative categories matched for homogeneity or frequency proximity. Interpretative data for this measure are shown. The definition of qualitative peakedness and its measurement by peak-to-shoulder difference and qualitative percentile kurtosis is continued. Data on the behavior of these two statistics are given. A simulated example in the field of psychopathological epidemiology on the calculation and interpretation of these three measures is presented. Finally, conclusions are drawn and suggestions for their use are given.

2. Skewness Measure for Qualitative Variables

2.1. Concept of Skewness in the Distribution of a Qualitative Variable

The concept of skewness implies an axis that allows the distribution to be divided into two parts. If one part is a reflection of the other, the distribution is considered to be symmetrical and the measure of skewness should yield a value of 0. On the other hand, if they are different, we speak of asymmetry, and the value of the measure should be different from 0. The more disparate the two sides divided by the axis of symmetry, the further away from 0 the value of the measure of skewness should be.

In continuous random distributions, a measure of central tendency, such as the arithmetic mean, median, mode or mid-rank, is taken as the axis of symmetry. In the case of qualitative variables, the only option would be the mode, that is, the nominal category with the highest frequency in the sample. However, the mode is not always unique. There may be two modal categories (bimodal distribution), three or more modal categories (multimodal distribution) or none (uniform distribution). Consequently, if the mode is adopted as the axis of symmetry, the concept could only be applied to unimodal distributions.

If the categories of a variable A on a nominal measurement scale are represented by numbers, they lack any algebraic property. The categories can perfectly well be identified by letters, words or non-numerical symbols to highlight the fact that they represent the classification options within an inclusive (every element of the population can be classified) and exhaustive (in a single category) system and not a measurement in the strict sense (objective determination of how many times the measured characteristic of the object is the unit of measurement agreed upon by experts). The only quantification that qualitative variables admit is the counting of the number of times that each of their categories appears in the sample or population, that is, the frequency or probability of each category. Thus, from its relative frequency or probability, a possibility of transformation opens up. The qualitative categories of the variable can be transformed into ordered categories ($X^{\downarrow}$). In this way, an ordinal frequency metric can be created. See Equation (1).

The qualitative variable A has k attributes or nominal categories:

$$A=\left\{a_i\right\}_{i=1}^{k}=\left\{a_1,a_2,\cdots,a_k\right\}$$

We create the ordinal variable $X^{\downarrow}$ which is a function of the qualitative variable A and has k ordinal categories. It is obtained by sorting its k attributes in decreasing order by their probability or frequency:

$$X^{\downarrow}=f\left(A\right)=\left\{i^{\downarrow}\right\}_{i=1}^{k}=\left\{1^{\downarrow},2^{\downarrow},\cdots,k^{\downarrow}\right\} \tag{1}$$

Thus, the values of $\mathrm{X}^{\downarrow}$ have decreasing probabilities:

$$f_{X^{\downarrow}}\left(1^{\downarrow}\right)\geq f_{X^{\downarrow}}\left(2^{\downarrow}\right)\geq\cdots\geq f_{X^{\downarrow}}\left(k^{\downarrow}\right)$$

First of all, skewness is a property of the shape of a distribution. When elaborating the bar chart to study skewness in qualitative variables, we do not proceed to arrange the sequence of categories (ordered by frequency) in ascending order, as in the cumulative distribution function, since a staircase shape typical of an increasing monotonic function would appear, but rather we try to create a more or less triangular or trapezoid shape, giving rise to the variable X^{Δ}.

If the number of k categories is odd, X^{Δ} is generated by placing the category with the highest frequency (modal category) or one of the maximum frequency categories, chosen at random, in the center (peak). After matching the remaining categories by frequency homogeneity or proximity, these pairs are arranged in descending order on either side of the mode. The category with the highest frequency of each pair is placed on the left and the category with the lowest frequency of the pair is placed on the right. The pair with the highest frequencies will be the one closest to the central category and the pair with the lowest frequencies will be the one furthest away from the central category. See Equation (2).

The ordinal variable X^{Δ}, which has a triangular or trapezoidal shape in a bar chart, is created from the ordinal variable $X^{\downarrow}$, which has a descending ladder shape, by relocating its k categories. In the case where k is odd:

$$X^{\Delta}=f\left(X^{\downarrow}\right)=\left\{i^{\Delta}\right\}_{i=1}^{k}=\left\{1^{\Delta},2^{\Delta},\cdots,k^{\Delta}\right\}$$

If $k\neq\dot{2}$,

$$\begin{aligned}&f_{X^{\Delta}}\left(1^{\Delta}\right)=f_{X^{\downarrow}}\left(\left(k-1\right)^{\downarrow}\right)\geq f_{X^{\Delta}}\left(2^{\Delta}\right)=f_{X^{\downarrow}}\left(\left(k-3\right)^{\downarrow}\right)\geq\cdots\geq f_{X^{\Delta}}\left(\left(\frac{k+1}{2}\right)^{\Delta}\right)\\&=f_{X^{\downarrow}}\left(1^{\downarrow}\right)\leq\cdots\leq f_{X^{\Delta}}\left(\left(k-1\right)^{\Delta}\right)=f_{X^{\uparrow}}\left(\left(k-2\right)^{\downarrow}\right)\leq f_{X^{\Delta}}\left(k^{\Delta}\right)=f_{X^{\uparrow}}\left(k^{\downarrow}\right)\end{aligned} \tag{2}$$

If the number of categories is even, the pair with the highest frequencies are placed in the center and proceed in the same way. If there are only two categories, the highest is placed in first order and the lowest in second order. See Equation (3).

In case k is even: If $k = \dot{2}$,

$$\begin{aligned} & f_{X^{\Delta}}\left(1^{\Delta}\right) = f_{X^{\downarrow}}\left(\left(k-1\right)^{\downarrow}\right) \geq f_{X^{\Delta}}\left(2^{\Delta}\right) = f_{X^{\downarrow}}\left(\left(k-3\right)^{\downarrow}\right) \geq \cdots \geq f_{X^{\Delta}}\left(\left(\frac{k}{2}\right)^{\Delta}\right) \\ & = f_{X^{\downarrow}}\left(1^{\downarrow}\right) \leq f_{X^{\Delta}}\left(\left(\frac{k}{2}+1\right)^{\Delta}\right) = f_{X^{\downarrow}}\left(2^{\downarrow}\right) \leq \cdots \leq f_{X^{\Delta}}\left(\left(k-1\right)^{\Delta}\right) = f_{X^{\downarrow}}\left(\left(k-2\right)^{\downarrow}\right) \\ & \leq f_{X^{\Delta}}\left(k^{\Delta}\right) = f_{X^{\downarrow}}\left(k^{\downarrow}\right) \end{aligned} \quad (3)$$

A distribution can be considered symmetrical if the two parts on either side of the maximum frequency category located in the center (odd number of categories) or of the imaginary line perpendicular to the abscissa axis between the two central categories of highest frequency (even number of categories) are equal. Conversely, there is asymmetry if they are dissimilar. This concept is called *qualitative asymmetry*.

2.2. Measurement of Skewness in the Distribution of a Qualitative Variable

Qualitative skewness can be measured by means of the average frequency difference between qualitative categories matched by frequency homogeneity or proximity. This measure does not require the existence of a single mode, even if it applies to a uniform distribution, whose shape in the bar chart is not trapezoidal, but rectangular.

In the following, this measurement approach is expressed in algebraic terms. Let A be a qualitative variable with a number of k nominal categories and each with relative frequency $f_A(a_i)$. The frequencies are ordered in descending order, that is, from the highest to the lowest to create $X^{\downarrow}$. From $X^{\downarrow}$, X^{Δ} is obtained.

If k is odd, the maximum frequency category that was located in the center of the bar chart, $f_{X^{\Delta}}\left[\left(\left(k+1\right)/2\right)^{\Delta}\right] = f_{X^{\downarrow}}\left(1^{\downarrow}\right)$, is excluded from frequency matching. The remaining frequencies are paired: $f_{X^{\downarrow}}\left(2^{\downarrow}\right)$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(3^{\downarrow}\right)$, $f_{X^{\downarrow}}\left(4^{\downarrow}\right)$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(5^{\downarrow}\right)$, $\cdots$, $f_{X^{\downarrow}}\left[\left(k-1\right)^{\downarrow}\right]$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(k^{\downarrow}\right)$. The $(k - 1)/2$ pairs of similar or closer frequencies are subtracted, the differences are summed and divided by the number of differences summed: $(k - 1)/2$. See Equation (4). In this way, the Average Frequency Difference between qualitative categories matched by frequency homogeneity or proximity is obtained, denoted by *AFD*. The formula for calculating *AFD* from X^{Δ} is also shown. See Equation (5).

$$f_{X^{\downarrow}}\left(1^{\downarrow}\right) \geq f_{X^{\downarrow}}\left(2^{\downarrow}\right) \geq \cdots \geq f_{X^{\downarrow}}\left(k^{\downarrow}\right)$$

$$\text{If} \quad k \neq \dot{2}, \quad AFD = \frac{\sum_{i=2}^{k-1}\left(f_{X^{\downarrow}}\left(i^{\downarrow}\right) - f_{X^{\downarrow}}\left(\left(i+1\right)^{\downarrow}\right)\right)}{\left(k-1\right)/2} \tag{4}$$

$$\text{If} \quad k \neq \dot{2}, \quad AFD = \frac{\sum_{i=1}^{(k-1)/2}\left(f_{X^{\Delta}}\left(i^{\Delta}\right) - f_{X^{\Delta}}\left(\left(k+1-i\right)^{\Delta}\right)\right)}{\left(k-1\right)/2} \tag{5}$$

If *k* is even, the frequencies are paired: $f_{X^{\downarrow}}\left(1^{\downarrow}\right)$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(2^{\downarrow}\right)$, $f_{X^{\downarrow}}\left(3^{\downarrow}\right)$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(4^{\downarrow}\right)$, $\cdots$, $f_{X^{\downarrow}}\left(\left(k-1\right)^{\downarrow}\right)$ is equal to or immediately greater than $f_{X^{\downarrow}}\left(k^{\downarrow}\right)$. See Equation (6). The *k*/2 pairs of similar or nearest similar frequencies are subtracted, the differences are summed, and divided by the number of summed differences (*k*/2), yielding *AFD*. The formula for calculating *AFD* from X^{Δ} is also shown. See Equation (7).

$$\text{If} \quad k = \dot{2}, \quad AFD = \frac{\sum_{i=1}^{k-1}\left(f_{X^{\downarrow}}\left(i^{\downarrow}\right) - f_{X^{\downarrow}}\left(\left(i+1\right)^{\downarrow}\right)\right)}{k/2} \tag{6}$$

$$\text{If} \quad k = \dot{2}, \quad AFD = \frac{\sum_{i=1}^{k/2}\left(f_{X^{\Delta}}\left(i^{\Delta}\right) - f_{X^{\Delta}}\left(\left(n+1-i\right)^{\Delta}\right)\right)}{\left(k-1\right)/2} \tag{7}$$

The *AFD* statistic is bounded from 0 to 1. A value of 0 indicates symmetry, which may correspond to a triangular (unimodal distribution), trapezoid (bi- or multimodal distribution) or rectangular (uniform distribution) profile. A value of 1 represents the maximum asymmetry and is reached with the distribution of a constant discrete random variable in which one value concentrates all the probability or frequency (Bernoulli distribution of parameter $p = 1$).

2.3. Interpretative Rules for Average Frequency Difference

Moral [5], based on the binomial distribution with parameter $p = 0.5$, obtained cut-off points for *AFD* suggestive of asymmetry (Table 1). This distribution was used because its domain (0 to *n*) allows us to establish a parallelism with the number of categories of a nominal variable (1 to $k = n + 1$). Besides, this distribution allows to define the distributions of the summands in the numerator of the *AFD* statistic as binomial proportions and to use the approximation to the normal distribution, which facilitates the simulation of data by the Monte Carlo method [7]. Finally, the probability of success of one-half guarantees perfect symmetry and the null value in *AFD* at the population level.

The 95th percentile can be used as a cut-off point to establish whether there is asymmetry. Its median is 0.09, which would constitute the most generalized cut-off point. However, the value of the cut-off point is higher the smaller the sample size, since there is a very high inverse linear relationship between the value of the 95th percentile and the sample size. In turn, the highest values of the 95th percentile appear with small and large nominal category numbers and the

Table 1. Descriptives and norms for *AFD* in the presence of a symmetric binomial distribution ($p = 0.5$) of two to eleven categories for sample sizes of 20, 40, 100, 200, 500, and 1000.

Statistics	Sample size											
	20	40	100	200	500	10^3	20	40	100	200	500	1000
	2 categories						3 categories					
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	0.627	0.550	0.319	0.196	0.127	0.090	0.545	0.388	0.242	0.183	0.110	0.087
*P*50	0.107	0.075	0.048	0.033	0.021	0.015	0.093	0.065	0.041	0.030	0.018	0.013
*P*75	0.183	0.129	0.082	0.057	0.036	0.026	0.158	0.111	0.070	0.051	0.031	0.023
*P*80	0.204	0.143	0.091	0.064	0.040	0.029	0.176	0.124	0.078	0.056	0.035	0.026
*P*90	0.262	0.184	0.117	0.081	0.051	0.037	0.228	0.159	0.100	0.072	0.044	0.033
*P*95	0.312	0.220	0.139	0.096	0.060	0.044	0.270	0.188	0.119	0.085	0.053	0.039
	4 categories						5 categories					
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	0.422	0.269	0.178	0.118	0.081	0.054	0.334	0.200	0.145	0.110	0.071	0.044
*P*50	0.093	0.068	0.043	0.030	0.019	0.013	0.075	0.053	0.035	0.025	0.016	0.012
*P*75	0.134	0.097	0.061	0.043	0.028	0.019	0.110	0.077	0.051	0.036	0.023	0.017
*P*80	0.145	0.105	0.066	0.046	0.030	0.021	0.120	0.084	0.055	0.039	0.025	0.018
*P*90	0.176	0.127	0.080	0.056	0.037	0.025	0.146	0.102	0.067	0.048	0.030	0.022
*P*95	0.203	0.146	0.091	0.065	0.042	0.029	0.170	0.119	0.077	0.056	0.034	0.025
	6 categories						7 categories					
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	0.431	0.299	0.186	0.145	0.079	0.058	0.375	0.258	0.170	0.122	0.079	0.050
*P*50	0.124	0.083	0.053	0.038	0.024	0.017	0.086	0.074	0.045	0.032	0.020	0.014
*P*75	0.167	0.113	0.073	0.052	0.033	0.023	0.123	0.101	0.062	0.044	0.028	0.019
*P*80	0.179	0.121	0.078	0.055	0.035	0.025	0.133	0.108	0.067	0.047	0.030	0.021
*P*90	0.211	0.144	0.092	0.065	0.041	0.029	0.161	0.127	0.079	0.056	0.035	0.024
*P*95	0.237	0.162	0.104	0.074	0.047	0.033	0.185	0.144	0.089	0.064	0.040	0.028
	8 categories						9 categories					
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	0.446	0.342	0.207	0.136	0.093	0.068	0.368	0.264	0.182	0.171	0.076	0.064
*P*50	0.124	0.089	0.062	0.043	0.028	0.020	0.110	0.074	0.048	0.052	0.023	0.016
*P*75	0.167	0.118	0.082	0.057	0.037	0.026	0.148	0.100	0.065	0.067	0.031	0.022
*P*80	0.179	0.127	0.087	0.060	0.039	0.027	0.159	0.107	0.069	0.071	0.033	0.023
*P*90	0.212	0.150	0.102	0.070	0.046	0.032	0.187	0.127	0.081	0.082	0.038	0.027
*P*95	0.238	0.170	0.114	0.080	0.051	0.036	0.210	0.144	0.092	0.092	0.043	0.031

Continued

	10 categories						11 categories					
Min	0	0	0	0	0	0	0	0	0	0	0	0
Max	0.956	0.651	0.384	0.290	0.176	0.128	0.748	0.533	0.360	0.247	0.168	0.108
*P*50	0.262	0.207	0.130	0.091	0.060	0.042	0.221	0.159	0.111	0.080	0.049	0.036
*P*75	0.352	0.269	0.170	0.119	0.077	0.055	0.296	0.214	0.146	0.105	0.064	0.047
*P*80	0.375	0.285	0.180	0.126	0.082	0.058	0.317	0.230	0.156	0.112	0.068	0.050
*P*90	0.439	0.330	0.211	0.146	0.094	0.067	0.372	0.269	0.182	0.130	0.080	0.058
*P*95	0.496	0.369	0.236	0.163	0.106	0.075	0.418	0.304	0.203	0.145	0.089	0.065

Note. Number of simulations = 20,000. Descriptive statistics: *Min* = minimum value, *Max* = maximum value, *P*50 = median or 50th percentile, *P*75 = upper quartile or 75th percentile, *P*80 = 80th percentile, *P*90 = 90th percentile, and *P*95 = 95th percentile.

lowest values of the 95th percentile with the central categories of X^{Δ}, reaching the minimum with five categories (2 to 11), since there is a non-linear relationship between the 95th percentile and the number of categories. Although there is a tendency for the 95th percentile value to be higher when the number of categories is even than when it is odd, this difference is not significant. From this pattern, it can be deduced that the minimum cut-off point appears with five nominal categories and a sample size of 1000 (*P*95 = 0.03) and the maximum with 10 nominal categories and a sample size of 20 (*P*95 = 0.47), as can be seen in Table 1.

The proposed measurement not only conformed to the expected behavior with the binomial distribution, but also showed a high correlation with the average inter-judge skewness, r_S = 0.87, 95% CI: [0.74, 1]. To obtain the average inter-judge skewness, five expert judges visually assessed on a scale of five ordered categories the skewness of 60 bar diagrams of binomial distributions B (n, p). The number of nominal categories ranged from 2 to 11 (n = 1 to 10). To achieve varying degrees of skewness, six different values were given to the probabilities of success: p = 0.01 (maximum skewness), 0.1, 0.2, 0.3, 0.4 and 0.5 (symmetry). The ordering criterion was whether the arrangement of the bars on either side of the central bar (excluded) in the case of an odd number of categories or of the imaginary line between the two central bars (included) in the case of an even number of categories can be considered: 1 = totally symmetrical arrangement, 2 = very slightly asymmetrical, 3 = slightly asymmetrical, 4 = quite asymmetrical, and 5 = very asymmetrical. Each graph is scored on a continuum from 1 to 5, which corresponded to the arithmetic mean of the five judges [5].

3. Peakedness Measure for Qualitative Variables

3.1. The Concept of Peakedness in the Distribution of a Qualitative Variable

Like skewness, peakedness is a property of the shape of a distribution. When a discrete distribution is represented by a bar chart, peakedness is defined as the vertical distance (frequency) between the peak and the shoulders.

The categories of a variable on a nominal measurement scale are the options of a classification system; however, they admit quantification based on their frequency in the sample or population. As previously stated, from their relative frequency or probability, a possibility of transformation opens up. The qualitative categories of the variable can be transformed into ordered categories, that is, an ordinal frequency metric can be created.

When elaborating the bar chart to visualize the peakedness, the sequence of categories (ordered by frequency) is not arranged in ascending order, but in a triangular or trapezoidal shape as previously described when presenting the qualitative asymmetry. Once the bar chart has been prepared, the k categories of the qualitative variable A are numbered from 1 to k, thus creating the ordinal variable X^{Δ}.

The peak is the value or values on the abscissa axis to which the maximum value on the ordinate axis corresponds, and it is located in the central category or categories within the bar plot. To measure peakedness, the simple frequency of the peak, *i.e.* the maximum simple frequency, is taken.

The concept of shoulder (left or right) implies an axis of symmetry and a distance from this axis. The way the diagram is constructed places the mode in the center as the axis of symmetry. A distance on each side of the axis of symmetry of about 25% of the distribution is established, so that both shoulders cover about the central 50%, as suggested by Horn [8]. As the categories have been enumerated in the diagram from 1 to k, creating the variable X^{Δ}, the percentiles of this variable can be calculated. The 25th percentile is the category that has a cumulative frequency of at least 0.25 in the plot or distribution of X^{Δ} and the 75th percentile is the category that has a cumulative frequency of at least 0.75 in the plot or distribution of X^{Δ}. However, if those percentiles correspond to the modal value or values, then the immediate adjacent non-modal category, if any, is taken. Once the two bounding categories of the shoulders are located, their simple relative frequencies are taken to measure the peakedness. Since these frequencies may be disparate, the arithmetic mean of the two frequencies is calculated. The left shoulder bounding category is denoted by h_1 and the right shoulder bounding category is denoted by h_2. The notation is taken from Tukey [9] and refers to his concept of hinges. The concepts of peak, shoulders, hinges and tails are illustrated for easier understanding in **Figure 1**.

3.2. Measurement of Peakedness in the Distribution of a Qualitative Variable

The measure of qualitative peakedness is the difference between the frequency of the peak (f_{peak}) and the average of the frequencies of the two bounding categories of the shoulders ($\left[f_{h_1} + f_{h_2} \right] / 2$) in the triangular (one mode), trapezoidal (two or more modes) or rectangular (uniform distribution) arrangement. When there are only two nominal categories, the difference is calculated between the frequencies of the two attributes, thus coinciding with the measure of qualitative

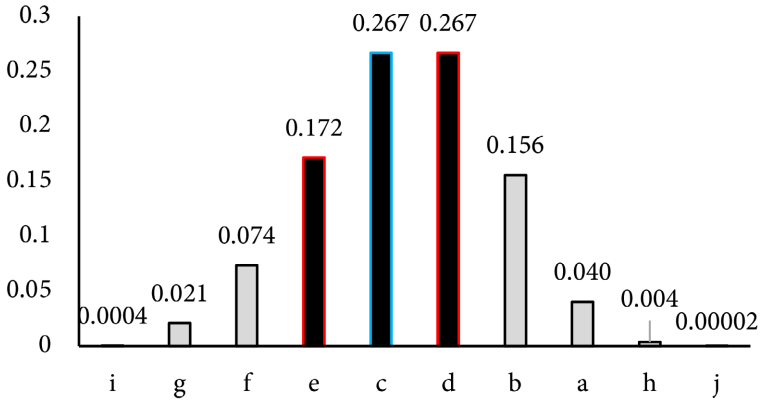

Figure 1. The shoulder region (P25 to P75) is filled in black, and the tails (left < P25 and right > P75) are in grey. The hinges are edged in red, and the peak is edged in blue. The bar-chart represents a binomial destruction: B ($n = 9$, $p = 0.3$).

skewness [5]. This statistic is called the peak-to-shoulder distance or difference and is denoted by *PSD*. See Equation (8).

$$PSD = f_{\text{peak}} - \frac{f_{h_1} + f_{h_2}}{2} \tag{8}$$

It should be noted that the definition adopted for percentile P_X ($p \times 100$) of the distribution of X^{Δ} is the minimum value of the ordered categories $(1, 2, \cdots, k)$ that accumulates at least a probability or frequency equal to the order p of the percentile. This is the simplest definition and corresponds to the inverse of the cumulative distribution function [10]. See equation (9).

$$X^{\Delta} = f\left(X^{\downarrow}\right) = \left\{i^{\Delta}\right\}_{i=1}^{k} = \left\{1^{\Delta}, 2^{\Delta}, \cdots, k^{\Delta}\right\}$$

$$P_X\left(p \times 100\right) = inf.\left(i^{\Delta} \in X^{\Delta} = \left\{1^{\Delta}, 2^{\Delta}, \cdots, k^{\Delta}\right\} \mid P\left(X^{\Delta} \le i^{\Delta}\right) \ge p\right) \tag{9}$$

The *PSD* statistic is bounded from 0 to 1. A value of 0 indicates a null peakedness, and appears with a uniform distribution. A value of 1 represents the maximum peakedness and is reached with the distribution of a constant discrete random variable in which one value concentrates all the probability or frequency. To interpret *PSD*, the interval [0, 0.25) can be considered to indicate low, [0.25, 0.5) medium, [0.5, 0.75) high, and [0.75, 1] very high peakedness.

A second measure of peakedness has been defined for the distribution of a qualitative variable. It is calculated from the ordinal variable X^{Δ} generated from the frequencies of the qualitative variable *A*. This second measure is the qualitative percentile kurtosis (*QPC*) or quotient between the semi-interquartile range (R_{SIQ}) and the percentile range (R_P) of X^{Δ}. See Equation (10).

This second measure of shape is taken from Truman Lee Kelley [11], which is a ratio between the semi-distance of the shoulders and the distance between the extreme tails. It was originally proposed to measure kurtosis, which is a complex concept that includes both aspects of thickening or thinning of the tails and pointing or flattening of the peak Therefore, *QPC* is a statistic that has been adapted from the context of quantitative and ordinal variables to apply to qualitative variables.

QPC can take values in the interval [0, 0.5]. As R_{SIQ} and R_P are closer, *QPC* tends to 0.5, showing shortened tails or platykurtosis. On the contrary, as the percentile range is larger than the semi-interquartile range, *QPC* approaches 0, showing elongated tails or leptokurtosis [11]. For the *QPC* values to correspond to the common interpretative logic of kurtosis measures [1], its value in the sample can be subtracted from 0.5 (maximum value). Thus, as the value gets closer to 0, there is a shortening of the tails, and as it gets closer to 0.5, there is a lengthening of the tails. It could be interpreted that values from 0 to 0.16 show platicurtosis (shortened tails), from 0.17 to 0.33 mesocurtosis (medium tails) and from 0.34 to 0.5 leptokurtosis (elongated tails).

$$\begin{aligned} QPC &= \frac{1}{2} - \frac{R_{SIQ}\left(X^{\Delta}\right)}{R_P\left(X^{\Delta}\right)} = \frac{1}{2} - \frac{\dfrac{P_{75}\left(X^{\Delta}\right) - P_{25}\left(X^{\Delta}\right)}{2}}{P_{90}\left(X^{\Delta}\right) - P_{10}\left(X^{\Delta}\right)} \\ &= \frac{1}{2} - \frac{P_{75}\left(X^{\Delta}\right) - P_{25}\left(X^{\Delta}\right)}{2 \times \left(P_{90}\left(X^{\Delta}\right) - P_{10}\left(X^{\Delta}\right)\right)} \end{aligned} \tag{10}$$

A problem with *QPC* is its indeterminacy (0/0) when the distances between quartiles and extreme deciles are zero. In this situation a value of 0.5 is given to *QPC*. This indeterminacy appears in distributions in which the modal category subsumes the first and last deciles and, therefore, the first and last quartiles. This is an extreme situation of shortened tails and prominence of the peak; hence the value of the statistic should be 0.5.

3.3. Behavior and Validity of *PSD* and *QPC*

What is the expected behavior of the *PSD* statistic? The more prominent the mode, the closer *PSD* should be to 1, and the less prominent, the closer to 0. The number of qualitative categories necessarily affects, since a greater number of categories causes the frequency to be more distributed among the categories and, consequently, the mode to have less prominence.

Moral [6] studied the behavior of this measure using the binomial distribution to generate a great diversity of profiles. This distribution was chosen because the parameter *p* (probability of success) allows manipulation of the prominence of the mode. The closer *p* is to 0, the more prominent the mode is, and the closer *p* is to 0.5, the less prominent the mode is. On the other hand, the parameter *n* (number of independent trials with constant probability of success) makes it possible to manipulate the number of attributes of the qualitative variable.

In the study, a total of 120 distributions were generated with six *p* values and twenty n values. As expected, the *PSD* statistic correlated with the *p* parameter of the binomial distribution with 46.8% of the shared variance and with the number of *k* categories with 27.6% of the shared variance. When comparing these two dependent Spearman's rank correlation coefficients with one variable in common using the two-tailed Hotelling's [12] t-test, the correlation of *PSD* showed a significantly higher association with *p* than with *k* ($t_{[117]} = -2.40$, $p = 0.009$).

The expectation was not for a very high correlation with either of the two parameters (*n*, *p*) of the binomial distributions generating the 120 random samples. This is because the frequency distribution among the categories becomes asymmetrical when *p* takes a value other than 0.5, resulting in bounding categories with disparate heights. This complexity increases as the number of categories increases.

The expectation that the parity or non-parity of the number of categories lacks relevance for *PSD* was also confirmed. The comparison of central tendency using the Mann-Whitney U test was not significant, and the effect size was trivial: Mann-Whithey *U* statistic = 1670, $z = 0.68$ with correction for ties and continuity, two-tailed $p = 0.497$, Rosenthal's *r* as effect size measure = 0.06. This was because the bounding categories of the shoulders are not subsumed in the categories with maximum (peak) frequency, which, if present, would determine the difference between an odd or even number of categories.

The validity of *PSD* was tested by its correlation with the average level of peakedness, which was assessed visually by five expert judges in 66 bar-charts. The correlation between *PSD* and the average inter-judge peakedness was significant and positive ($r_S = 0.87$, 95% *CI* [0.78, 0.93], $t_{[64]} = 14.18$, two-tailed $p < 0.001$) with a very high strength of association ($r_S = 0.871$) and a sharing variance of 75.9% ($r_S^2 \times 100$).

The correlation of *QPC* with the *p* parameter was positive and significant with a medium strength of association (sharing a variance of 17.8%). However, *QPC* was independent of the number of categories (*n*) and of having an even or odd number of categories. The correlation between *PSD* and *QPC* was significant ($r_S = 0.38$, 95% *CI* [0.53, 0.21], $t_{[64]} = 4.45$, $p < 0.001$) with a medium strength of association and a shared variance of 14.4% [6].

If *QPC* is considered as a criterion of validity for *PSD*, one might wonder why the correlation between *PSD* and *QPC* was not high. The reason is that *QPC* is an indirect measure of peakedness, as shown by its medium correlation with the average inter-judge peakedness, with only one-sixth of shared variance, while the correlation between *PSD* and average inter-judge peakedness was very high, with three-quarters of shared variance. Furthermore, the correlation between *QPC* and the parameter *p* was medium, with less than one-fifth of shared variance, whereas the correlation between *PSD* and *p* was high, with a shared variance close to one-half. Consequently, *PSD* is a better measure of peakedness than *QPC*. The latter statistic probably quantifies the movement of the probability mass from the shoulders to the tails, which relates to kurtosis, a more complex concept.

4. Example of Calculation of Qualitative Asymmetry and Peakedness

Be a random (dummy) sample of 200 persons (100 men and 100 women) attending urban mental health consultation. Personality disorders were assessed using the Structured Clinical Interview for DSM-IV Axis II Personality Disorders [13]. Diagnostic interviews were conducted by five psychiatrists and five clinical psychologists. Patients were classified into 10 specific DSM-V categories (schizoid, schizotypal, and paranoid; antisocial, narcissistic, histrionic, and borderline; avoidant, dependent, and obsessive-compulsive) and two additional ones: unspecified and no disorder [14]. Calculate the central tendency (mode), variability (universal variation ratio from Moral [15]), qualitative skewness (average frequency difference), and qualitative peakedness (peak-to-shoulder difference and qualitative percentile kurtosis) of the distribution of this polychotomous variable with the sample data in **Table 2**.

Mode as a measure of central tendency.

$$\begin{aligned} Mode &= \left\{ a_i \mid \max\left[\left\{ f_A\left(a_i\right)\right\}_{i=1}^{12}\right] = f_A\left(a_1 = \text{no disorder}\right) = 0.2 \right\} \\ &= a_1 = \text{no disorder} \end{aligned}$$

Universal Variation Ratio (UVR) as a measure of variability.

$$\begin{aligned} UVR &= \frac{k^2}{k^2-1}\times\left(1-\frac{f_{\max}}{c}\right) = \frac{k^2}{k^2-1}\times\left(1-\frac{\max\left[\left\{ f_A\left(a_i\right)\right\}_{i=1}^{12}\right]}{c}\right) \\ &= \frac{12^2}{12^2-1}\left(1-\frac{0.2}{1}\right) = 0.806 \end{aligned}$$

Table 2. Frequency distribution of ten specific and two additional categories of personality disorders.

A	n_x	f_x	$X^{\downarrow}$	$f_{X^{\downarrow}}$	X^{Δ}	$f_{X^{\Delta}}$	$f_{dX^{\Delta}}$	$F_{X^{\Delta}}$
1 = no disorder	40	0.2	1 = no disorder	0.2	1 = schizotypal	0.03	0.02	0.03
2 = schizoid	2	0.01	2 = dependent	0.15	2 = paranoid	0.05	0.02	0.08
3 = schizotypal	6	0.03	3 = borderline	0.13	3 = narcissistic	0.06	0.01	0.14
4 = paranoid	10	0.05	4 = histrionic	0.11	4 = avoidant	0.1	0.02	0.24
5 = antisocial	10	0.05	5 = avoidant	0.1	5 = borderline	0.13	0.02	0.37
6 = narcissistic	12	0.06	6 = obses-comp	0.08	6 = no disorder	0.2	0.05	0.57
7 = histrionic	22	0.11	7 = narcissistic	0.06	7 = dependent	0.15		0.72
8 = borderline	26	0.13	8 = paranoid	0.05	8 = histrionic	0.11		0.83
9 = avoidant	20	0.1	9 = antisocial	0.05	9 = obses-comp	0.08		0.91
10 = dependent	30	0.15	10 = schizotypal	0.03	10 = antisocial	0.05		0.96
11 = obses-comp	16	0.08	11 = nonspecific	0.03	11 = nonspecific	0.03		0.99
12 = not specific	6	0.03	12 = schizoid	0.01	12 = schizoid	0.01		1
Σ	200	1		1		1	0.14	

Note. A = personality disorders, n_X = simple absolute frequencies of A, f_X = simple relative frequencies of A, $X^{\downarrow}$ = personality disorders ordered descending by sample frequencies, $f_{X^{\downarrow}}$ = simple relative frequencies of $X^{\downarrow}$, X^{Δ} = personality disorders arranged trapezoidally by sample frequencies, $f_{X^{\Delta}}$ = simple relative frequencies of X^{Δ}, $f_{dX^{\Delta}}$ = frequency differences between qualitative categories matched by frequency homogeneity or proximity, $F_{X^{\Delta}}$ = cumulative relative frequencies of X^{Δ}, and Σ = sum by column.

k = number of attributes or categories of the variable in nominal measurement scale.

$f_{\max}$ = highest relative frequency in the sample.

c = number of attributes with the highest relative frequency.

Average Frequency Difference (*AFD*) as a measure of skewness. The number of categories is even. It is first calculated from $X^{\downarrow}$.

$$AFD = \frac{\sum_{i=1}^{k-1}\left(f_{X^{\downarrow}}(i) - f_{X^{\downarrow}}(i+1)\right)}{k/2}$$

$$= \frac{2}{k}\Big[\left(f_{X^{\downarrow}}(1) - f_{X^{\downarrow}}(2)\right) + \left(f_{X^{\downarrow}}(3) - f_{X^{\downarrow}}(4)\right) + \left(f_{X^{\downarrow}}(5) - f_{X^{\downarrow}}(6)\right)$$

$$+ \left(f_{X^{\downarrow}}(7) - f_{X^{\downarrow}}(8)\right) + \left(f_{X^{\downarrow}}(9) - f_{X^{\downarrow}}(10)\right) + \left(f_{X^{\downarrow}}(11) - f_{X^{\downarrow}}(12)\right)\Big]$$

$$ADF = \big[(0.2-0.15)+(0.13-0.11)+(0.1-0.08)+(0.06-0.05)$$

$$+(0.05-0.03)+(0.03-0.01)\big]/6$$

$$= 0.14/6 = 0.023$$

The calculation is repeated from X^{Δ} yielding the same result which is very close to 0, showing symmetry.

The calculation is repeated starting from X^{Δ}, obtaining a result very close to zero, which shows symmetry.

$$ADF = \frac{\sum_{i=1}^{k/2} df_{X^{\Delta}}(i)}{k/2} = \frac{\sum_{i=1}^{k/2}\left(f_{X^{\Delta}}(x_{n+1-i}) - f_{X^{\Delta}}(x_i)\right)}{k/2}$$

$$= \frac{2}{k}\Big[\left(f_{X^{\Delta}}(x_{12}) - f_{X^{\Delta}}(x_1)\right) + \left(f_{X^{\Delta}}(x_{11}) - f_{X^{\Delta}}(x_2)\right) + \left(f_{X^{\Delta}}(x_{10}) - f_{X^{\Delta}}(x_3)\right)$$

$$+ \left(f_{X^{\Delta}}(x_9) - f_{X^{\Delta}}(x_4)\right) + \left(f_{X^{\Delta}}(x_8) - f_{X^{\Delta}}(x_5)\right) + \left(f_{X^{\Delta}}(x_7) - f_{X^{\Delta}}(x_6)\right)\Big]$$

$$ADF = \big[(0.2-0.15)+(0.13-0.11)+(0.1-0.08)+(0.06-0.05)$$

$$+(0.05-0.03)+(0.03-0.01)\big]/6$$

$$= 0.14/6 = 0.023$$

Peak-to-Shoulder Difference (*PSD*) as a measure of peakedness.

$$PSD = f_{\text{peak}} - \frac{f_{h_1} + f_{h_2}}{2} = f_{\text{no disorder}} - \frac{f_{\text{borderline}} + f_{\text{histrionic}}}{2}$$

$$= 0.2 - \frac{0.13 + 0.11}{2} = 0.08$$

Qualitative percentile kurtosis (*QPC*) as a measure of peakedness.

$$P_{75}\left(X^{\Delta}\right) = 8 = \text{histrionic};\ F_{X^{\Delta}}(8) = 0.83 > 0.75$$

$$P_{25}\left(X^{\Delta}\right) = 5 = \text{borderline};\ F_{X^{\Delta}}(5) = 0.37 > 0.25$$

$$P_{90}\left(X^{\Delta}\right) = 9 = \text{obsessive} - \text{compulsive};\ F_{X^{\Delta}}(9) = 0.91 > 0.90$$

$$P_{10}\left(X^{\Delta}\right) = 3 = \text{narcissistic};\ F_{X^{\Delta}}(3) = 0.14 > 0.10$$

$$QPC = 0.5 - \frac{R_{SIQ}\left(X^{\Delta}\right)}{R_{P}\left(X^{\Delta}\right)} = 0.5 - \frac{8-5}{2[9-3]} = 0.5 - \frac{3}{2\times 6} = 0.5 - \frac{3}{12} = 0.5 - \frac{1}{4} = 0.25$$

The peak-to-shoulder difference (*PSD*) shows low peakedness (0 to 0.24), and the qualitative percentile kurtosis (*QPC*) reveals medium tails or mesocurtosis (0.17 to 0.33). The bar plot of X^{Δ} shows a clearly symmetrical and flattened profile (**Figure 2**), evidencing that *PSD* is a better measure for pointing than *QPC*. The latter seems appears to measure the displacement of the probability mass from the shoulders to the tails, *i.e.*, kurtosis [16].

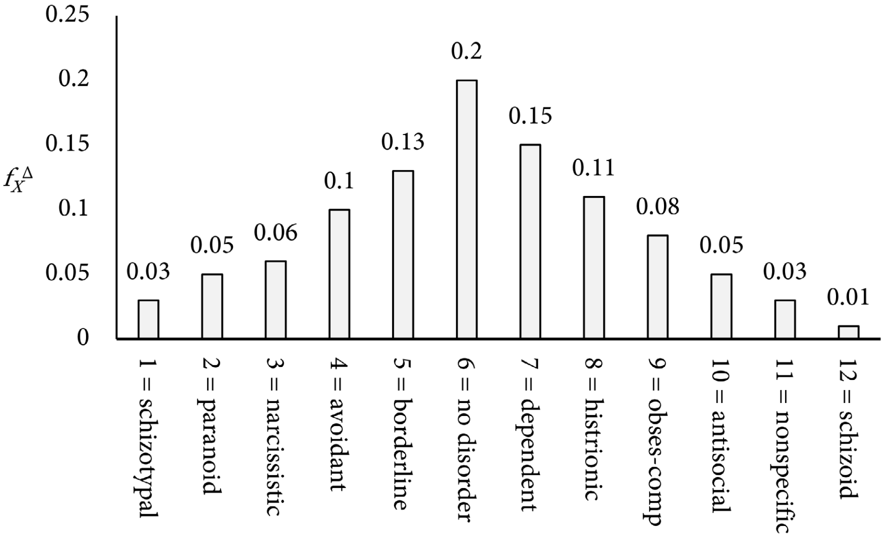

Figure 2. Bar chart of X^{Δ}.

5. Conclusions

A definition of the concept of skewness for the distribution of a qualitative variable is possible by creating an axis of symmetry with the modal frequency, pairing the qualitative categories by frequency proximity and distributing them on both sides, on the left the higher of the pair and on the right the lower of the pair. The average of the differences between the frequencies in this arrangement allows to define a statistic ranging from 0 to 1, where 0 indicates symmetry and 1 the maximum asymmetry. Moreover, this arrangement allows a reliable visual assessment of skewness. This statistic is valid by showing a behavior adjusted to the definition of qualitative skewness. It approaches 0 the more symmetrical or similar are the frequencies or heights of the bars equidistant to the symmetry axis and approaches 1 the more disparate they are. In turn, it is precise as indicated by its very high correlation with the average of the expert judges' skewness ratings. Even cut-off points (95th percentile), obtained by Monte Carlo simulation, are available from a binomial distribution with parameter p = 0.5. These points are suggestive of asymmetry depending on the number of nominal categories (2 to 11) and the sample size (20 to 1000).

It is also possible to define a concept of peakedness for qualitative variables from the qualitative asymmetry approach and a measure can be established through the difference or distance of the peak frequency to the average of the frequencies of the shoulders in the bar chart (triangular, trapezoidal or rectangular profile). It is simple to calculate and valid, as demonstrated by its behavior from binomial distribution models and its correlation with two criteria: the average inter-judge evaluation of the peakedness and the qualitative percentile kurtosis. Peak-to-shoulder distance is a better measure of peakedness than qualitative percentile kurtosis. This latter probably measures the movement of the probability mass from peak to tails (kurtosis) and not the vertical distance from peak to shoulders (peakedness).

It is suggested to use the average frequency difference between qualitative categories matched for frequency homogeneity or proximity to measure skewness and the peak-to-shoulder difference to measure peakedness when describing the distribution of a qualitative variable. As their use becomes more widespread, it may find utility and create new measures for inclusion in the study of noncompliance with the assumptions underlying the testing tests and estimation techniques developed for nominal variables.

Acknowledgements

The author thanks the reviewers and editor for their helpful comments.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Bono, R., Arnau, J., Alarcón, R. and Blanca, M.J. (2020) Bias, Precision, and Accuracy of Skewness and Kurtosis Estimators for Frequently Used Continuous Distributions. *Symmetry*, **12**, Article No. 19. https://doi.org/10.3390/sym12010019

[2] Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A. (2019) Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, **22**, 67-72. https://doi.org/10.4103/aca.ACA_157_18

[3] Kishore, K. and Kapoor, R. (2020) Statistics Corner: Reporting Descriptive Statistics. *Journal of Postgraduate Medicine Education and Research*, **54**, 66-68. https://doi.org/10.5005/jp-journals-10028-1364

[4] Sarka, D. (2021) Descriptive Statistics. In: *Advanced Analytics with Transact-SQL*, Apress, Berkeley, 3-29. https://doi.org/10.1007/978-1-4842-7173-5_1

[5] Moral de la Rubia, J. (2022) Una medida de asimetría unidimensional para variables cualitativas [A Measure of One-Dimensional Asymmetry for Qualitative Variables]. *Revista de Psicología*, **40**, 519-551. https://doi.org/10.18800/psico.202201.017

[6] Moral de la Rubia, J. (2022) Medición del apuntamiento en variables en escala nominal [Measurement of Peakedness in Nominal Scale Variables]. *Revista de Psicología*, **41**, 421-459. https://doi.org/10.18800/psico.202301.016

[7] Lumivero (2021) The Leading Data Analysis and Statistical Solution for Microsoft Excel. https://www.xlstat.com/en

[8] Horn, P.S. (1983) A Measure for Peakedness. *The American Statistician*, **37**, 55-56. https://doi.org/10.1080/00031305.1983.10483090

[9] Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA.

[10] Han, P., Kong, L., Zhao, J. and Zhou, X. (2019) A General Framework for Quantile Estimation with Incomplete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **81**, 305-333. https://doi.org/10.1111/rssb.12309

[11] Kelley, T.L. (1923) Statistical Methods. McMillan Company, New York.

[12] Hotelling, H. (1940) The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters. *Annals of Mathematical Statistics*, **11**, 271-283. https://doi.org/10.1214/aoms/1177731867

[13] First, M.B., Williams, J.B.W., Benjamin, L.S. and Spitzer, R.L. (2016) American Psychiatry Association Structured Clinical Interview for DSM-5-PD. Personality Disorders. American Psychiatric Publishing, Washington DC.

[14] Weekers, L.C., Hutsebaut, J., Zimmermann, J. and Kamphuis, J.H. (2022). Changes in the Classification of Personality Disorders: Comparing the DSM-5 Section II Personality Disorder Model to the Alternative Model for Personality Disorders Using Structured Clinical Interviews. *Personality Disorders: Theory, Research, and Treatment*, **13**, 527-535. https://doi.org/10.1037/per0000512

[15] Moral de la Rubia, J. (2022) Una medida de variación para datos cualitativos con cualquier tipo de distribución [A Measure of Variation for Qualitative Data with Any Type of Distribution]. *Psychologia: Avances de la Disciplina*, **16**, 63-76. https://doi.org/10.21500/19002386.5642

[16] Westfall, P.H. (2014) Kurtosis as Peakedness, 1905-2014. R.I.P. *The American Statistician*, **68**, 191-195. https://doi.org/10.1080/00031305.2014.917055