Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 259-293
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Measures of Variability for Qualitative Variables Using the R Software

José Moral de la Rubia

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

How to cite this paper: Moral de la Rubia, J. (2024) Measures of Variability for Qualitative Variables Using the R Software. *Open Journal of Statistics*, **14**, 259-293.
https://doi.org/10.4236/ojs.2024.143013

Received: May 1, 2024
Accepted: June 10, 2024
Published: June 13, 2024

Abstract

Although there are many measures of variability for qualitative variables, they are little used in social research, nor are they included in statistical software. The aim of this article is to present six measures of variation for qualitative variables of simple calculation, as well as to facilitate their use by means of the R software. The measures considered are, on the one hand, Freeman's variation ratio, Moral's universal variation ratio, Kvalseth's standard deviation from the mode, and Wilcox's variation ratio which are most affected by proximity to a constant random variable, where the measures of variability for qualitative variables reach their minimum value of 0. On the other hand, the Gibbs-Poston index of qualitative variation and Shannon's relative entropy are included, which are more affected by the proximity to a uniform distribution, where the measures of variability for qualitative variables reach their maximum value of 1. Point and interval estimation are addressed. Bootstrap by the percentile and bias-corrected and accelerated percentile methods are used to obtain confidence intervals. Two calculation situations are presented: with a sample mode and with two or more modes. The standard deviation from the mode among the six considered measures, and the universal variation ratio among the three variation ratios, are particularly recommended for use.

Keywords

Variation Ratio, Relative Entropy, Index of Qualitative Variation, Standard Deviation from Mode, Bootstrap Confidence Interval

1. Introduction

Many measures of variation have been defined for qualitative variables [1] [2] [3]; however, they are little known and utilized, even though qualitative variables

are frequent and highly important in social and health research [4] [5]. In fact, many basic or applied statistics manuals do not even mention them, and commercial or open-access statistical software does not include them. It is important to note that their development is relatively recent, beginning in the mid-20th century with the publication of the seminal article by the American mathematician Claude Elwood Shannon (1916-2001) on the mathematical theory of communication [6], and with the diversity index proposed by Simpson [7]. These measures proliferated between the 1960s, and 1980s [2], and continue to be developed today [3] [8] [9].

In this article, six measures are considered: Freeman's [10] Variation Ratio (*FVR*), Wilcox's [1] Variation Ratio (*WVR*), Moral's [3] Universal Variation Ratio (*UVR*), Kvalseth's [11] Standard Deviation from Mode (*SDM*), Gibbs-Poston's [12] Index of Qualitative Variation (*IQV*), and Shannon's [13] Relative Entropy (*RelE*). The selected measures are simple to calculate, practical to use, clear to interpret, and have been applied in social research [3] [14] [15].

A characteristic shared by all these measures of variability is to have a range from 0 to 1. In this range, 0 corresponds to the distribution of a constant random variable, in which one value concentrates all the probability at the population level or all the frequency at the sample level. The value of 1 corresponds to a uniform distribution, in which all its values have the same probability at the population level or the same frequency at the sample level.

The aim of this article is twofold. On the one hand, the six measures of variation are presented in a simple way for better understanding among social researchers, since they are not included in commonly used statistical software, whether commercial or freely available. On the other hand, an additional objective is to facilitate their use through the R software, which is freely available. Two calculation situations are considered: qualitative data random samples with a single mode and with two or more modes.

2. Freeman's Variation Ratio

The variation ratio was developed by the American psychology Larry C. Freeman for unimodal distributions and can be denoted by *FVR* [10]. It starts from a formula of variation around the mode, and its expression is simplified to the complement of the frequency of the single mode. Refer to Formula (1), where n_i represents absolute frequency of each value (i = 1, 2, …, k), k is the number of qualitative categories, n_{mo} denotes absolute frequency of the single mode (*mo*), n signifies sample size, f_i stands for relative frequency, and f_{mo} represents relative frequency of the single mode.

$$FVR = \frac{\sum_{i=1}^{k}\left(n_i - \frac{n_{mo}}{k}\right)}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}\left(n_i - \frac{n_{mo}}{k}\right)}{n} = \sum_{i=1}^{k}\left(\frac{n_i}{n} - \frac{n_{mo}}{k \times n}\right)$$
$$= \sum_{i=1}^{k}\left(f_i - \frac{f_{mo}}{k}\right) = \sum_{i=1}^{k} f_i - \sum_{i=1}^{k}\frac{f_{mo}}{k} = 1 - \frac{k \times f_{mo}}{k} = 1 - f_{mo} \quad (1)$$

It only applies to a unimodal distribution. In the case of a constant random variable, where a single value *a* concentrates all the frequency ($f_a = 1$), the variation ratio is 0, since the value *a* is the mode and its relative frequency is one ($FVR = 1 - f_{mo} = 1 - 1 = 0$). This represents the minimum variability condition. In the case of a uniform distribution, it can be argued that its value is 1 because this distribution has no mode, and consequently, the relative frequency of its mode is 0 ($FVR = 1 - f_{mo} = 1 - 0 = 1$). This represents the situation of maximum variability. Its advantage is the simplicity of calculation and its disadvantage is the scarce information on the distribution used for its computation.

3. Wilcox's Variation Ratio

The American political scientist Allen R. Wilcox published a standardized variation ratio based on the mode, which can be denoted by *WVR* [1]. Refer to Formula (2).

$$\begin{aligned} WVR &= 1 - \frac{\sum_{i=1}^{k}\left(n_{mo} - n_i\right)}{n\left(k-1\right)} = 1 - \frac{\sum_{i=1}^{k} n_{mo} - \sum_{i=1}^{k} n_i}{n\left(k-1\right)} = 1 - \frac{kn_{mo} - n}{n\left(k-1\right)} \\ &= 1 - \frac{\frac{1}{n}}{\frac{1}{n}} \times \frac{kn_{mo} - n}{n\left(k-1\right)} = 1 - \frac{k\frac{n_{mo}}{n} - \frac{n}{n}}{k-1} = 1 - \frac{kf_{mo} - 1}{k-1} \end{aligned} \tag{2}$$

It is straightforward to establish the relationship between *WVR* and *FVR*, as shown in Formula (3).

$$WVR = 1 - \frac{kf_{mo} - 1}{k-1} = \frac{k - 1 - kf_{mo} + 1}{k-1} = \frac{k}{k-1}\left(1 - f_{mo}\right) = \frac{k}{k-1} FVR \tag{3}$$

Based on this equality, it can be stated that *WVR* is greater than or equal to *FVR*. This measure of variability necessarily requires that the distribution be unimodal. When there are two or more modes, it cannot be calculated. In the case of a uniform distribution, the modal frequency cannot be given a value of 0, as was argued with *FVR*, since the *WVR* statistic falls outside the range of 0 to 1 stipulated for standardized indices of variation of qualitative variables: $WVR = k/(k-1) \times FVR = k/(k-1) \times 1 = k/(k-1) > 1$. Consequently, it has more limitations than *FVR*; however, it includes additional information on the number of qualitative categories of the variable.

4. Moral's Universal Variation Ratio

The Spanish-Mexican psychologist José Moral proposed a modification of Freeman's formula [3]. This new proposal allows its application in cases of multiple modes and considers the number of categories (*k*), as does the *WRV*. The author called this modified statistic the Universal Variation Ratio (*UVR*), because it can be applied to any type of qualitative variable distribution. Refer to Formula (4), where *k* represents number of qualitative categories of the variable, *c* stands for number of values with maximum (absolute or relative) frequency and this value can vary from 1 to *k*, $f_{\max}$ denotes maximum relative frequency

corresponding to the mode (f_{mo}), except in a uniform distribution ($c = k$), in which it is considered that there is no mode and the value of all frequencies is $1/k$.

$$UVR = \frac{1-\frac{1}{c}\times f_{\max}}{1-\max\left(\frac{1}{c}\times f_{\max}\right)} = \frac{1-\frac{f_{\max}}{c}}{1-\frac{1}{k^2}} = \frac{1-\frac{f_{\max}}{c}}{\frac{k^2-1}{k^2}} = \frac{k^2}{k^2-1}\times\left(1-\frac{f_{\max}}{c}\right) \quad (4)$$

Starting from Freeman's formula, $1 - f_{mo}$, the proposed formula weights the modal relative frequency by the inverse of the number of modes ($1/c$) and divides the expression by its maximum value. This maximum is reached with the uniform distribution, when $c = k$, $f_{\max} = 1/k$ y $1 - 1/c \times f_{\max} = 1 - 1/k \times 1/k = 1 - 1/k^2$.

When a value monopolizes the entire frequency (constant random variable), where the modal frequency is unique and has a unit value ($c = 1$ and $f_{\max} = 1$), the Universal Variation Ratio (*UVR*) reaches its minimum value of 0 (Formula (5)).

$$UVR = \frac{k^2}{k^2-1}\times\left(1-\frac{f_{\max}}{c}\right) = \frac{k^2}{k^2-1}\times\left(1-\frac{1}{1}\right) = \frac{k^2}{k^2-1}\times 0 = 0 \quad (5)$$

If there is no mode (uniform distribution), all categories have the same frequency and this frequency is the maximum ($1/k$), the value of c is k and the Universal Variation Ratio (*UVR*) reaches its maximum value of 1 (Formula (6)).

$$UVR = \frac{1-f_{\max}/c}{\left(k^2-1\right)/k^2} = \frac{1-(1/k)/k}{\left(k^2-1\right)/k^2} = \frac{1-1/k^2}{1-1/k^2} = 1 \quad (6)$$

As c approaches k, where k is the number of categories (uniform distribution), the result of the proposed modification approaches 1, as shown in Formula (7).

$$UVR = \lim_{c\to k}\frac{1-\frac{1}{c}\times f_{\max}}{1-1/k^2} = \frac{1-\frac{1}{k}\times\frac{1}{k}}{1-1/k^2} = \frac{1-1/k^2}{1-1/k^2} = 1 \quad (7)$$

This is because, to the extent that the sample of qualitative variable A presents more categories with maximum frequency (c), the subtractive effect of the maximum frequency decreases in the modified variation ratio ($1 - f_{\max}/c$), and consequently, the value of this measure of variability increases (*UVR*). The lower the number of categories (k), the greater the increase in the Universal Variation Ratio (*UVR*), since the variability is more evenly distributed, moving the distribution of variable A away from that of a constant random variable (minimum value) and closer to that of a uniform distribution (maximum value).

In the case of a mode ($c = 1$), which is the situation in which *FVR* and Moral's *UVR* are comparable, *UVR* yields a value greater than or equal to *FVR*, as does *WRV*. Refer to Formula (8).

$$UVR = \frac{1-f_{Max}/c}{\left(k^2-1\right)/k^2} = \frac{1-f_{Mod}/1}{\left(k^2-1\right)/k^2} = \frac{k^2\left(1-f_{Mod}\right)}{k^2-1} = \frac{k^2}{k^2-1}FVR \geq FVR \quad (8)$$

When the number of qualitative categories (k) is very small, there is a greater difference between *UVR* and *FVR*. With two categories, the difference or increase is one third: $UVR\text{-}FVR = k^2/(k^2 - 1) = 0.333$. With three categories, the difference or increase is one-eighth: $UVR\text{-}FVR = k^2/(k^2 - 1) = 0.125$. However, as the number of categories increases, the difference becomes smaller. With four categories, the difference or increment is 0.067, and with five, it is 0.042. When the number of categories tends to infinity, *UVR* converges to *FVR*, as shown in Formula (9), where $c = 1$.

$$UVR = \lim_{k\to\infty} \frac{k^2}{k^2-1}\left(1-\frac{f_{\max}}{c}\right) = \lim_{k\to\infty} \frac{k^2}{k^2-1}\left(1-f_{mo}\right) = 1-f_{mo} = FVR \tag{9}$$

In contrast to *FVR*, the new measure of variation proposed by Moral [3] is less than or equal to Wilcox's *WVR* [1], when there is a single qualitative category with maximum frequency ($c = 1$). Consequently, *FVR* always takes a value less than or equal to *UVR*, and *UVR* always takes a value less than or equal to *WVR*. Refer to Formula (10).

$$\begin{aligned} &UVR = \frac{k^2}{k^2-1} \times FVR \rightarrow FVR = \frac{k^2-1}{k^2} \times UVR = \left(1-\frac{1}{k^2}\right) \times UVR \\ &WVR = \frac{k}{k-1} \times FVR \rightarrow FVR = \frac{k-1}{k} \times WVR = \left(1-\frac{1}{k}\right) \times WVR \\ &FVR = \left(1-\frac{1}{k^2}\right) \times UVR = \left(1-\frac{1}{k}\right) \times WVR \\ &FVR \le UVR \le WVR \end{aligned} \tag{10}$$

5. Kvalseth's Standard Deviation from Mode

The Norwegian-born American engineer Tarald O. Kvalseth proposed a new standardized index named Standard Deviation from Mode (*SDM*) [11]. Its advantage is that it utilizes all frequencies. Although it can be calculated with multiple modes (two or more of the k values with modal frequency), or even with a uniform distribution (the k values with maximum frequency), the author proposed the measure of variation for a unimodal distribution (Formula (11)). This restriction allows him to determine an asymptotic standard error, define an asymptotic confidence interval, and specify the use of inferential statistics to compare two or more *SDM* statistics.

$$\widehat{SDM} = 1-\sqrt{\frac{\sum_{i=1}^{k}\left(f_{mo}-f_i\right)^2}{k-1}} \tag{11}$$

Algebraically, it can be demonstrated that the value of Kvalseth's *SDM* estimator is less than or equal to *WVR*. Refer to Formula (12).

$$\begin{aligned} &\widehat{SDM} \le WVR \\ &1-\sqrt{\frac{\sum_{i=1}^{k}\left(f_{mo}-f_i\right)^2}{k-1}} \le 1-\frac{k\times f_{mo}-1}{k-1} \end{aligned} \tag{12}$$

As the sample size increases ($n \rightarrow \infty$), the sampling distribution of the *SDM*

statistic converges to a normal distribution with mean *SDM* and variance σ_{DEM}^2. Refer to Formula (13).

$$\widehat{SDM} \xrightarrow{d} N\left(SDM, \sigma_{SDM}^2\right)$$
$$\sigma_{SDM}^2 = \frac{p_{Mo} \times \left(1 - kp_{Mo}\right)^2 + \sum_{i=1}^{k} p_i \times \left(p_{Mo} - p_i\right)^2}{n \times \left(k-1\right)^2 \times \left(1 - SDM\right)^2} - \frac{\left(1 - SDM\right)^2}{n} \tag{13}$$

The square root of the above expression gives the standard deviation of the sampling distribution of *SDM* or standard error of *SDM*, as shown in Formula (14).

$$\sigma_{SDM} = \sqrt{\frac{p_{Mo} \times \left(1 - kp_{Mo}\right)^2 + \sum_{i=1}^{k} p_i \times \left(p_{Mo} - p_i\right)^2}{n \times \left(k-1\right)^2 \times \left(1 - SDM\right)^2} - \frac{\left(1 - SDM\right)^2}{n}} \tag{14}$$

Substituting the probability with its estimator $\hat{p}_i = n_i / n = f_i$ and using the sample mode, we obtain the estimator of the standard deviation of the sampling distribution of *SDM* or standard error of *SDM* (Formula (15))

$$\hat{\sigma}_{SDM} = \sqrt{\frac{f_{Mo}\left(1 - kf_{mo}\right)^2 + \sum_{i=1}^{k} f_i \left(f_{mo} - f_i\right)^2}{n\left(k-1\right)^2 \left(1 - \widehat{SDM}\right)^2} - \frac{\left(1 - \widehat{SDM}\right)^2}{n}} \tag{15}$$

When the sample size is reasonably large ($n \geq 30$), the standard error allows for interval estimates of *SDM*. See Formula (16), where $z_{1-\alpha/2}$ represents quantile of order $1 - \alpha/2$ in a standard normal distribution and $1 - \alpha$ = level of confidence.

$$P\left(\widehat{SDM} - z_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{SDM} \leq SDM \leq \widehat{SDM} + z_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{SDM}\right) = 1 - \alpha \tag{16}$$

6. Gibbs-Poston Index of Qualitative Variation

The American sociologists Jack P. Gibbs and Dudley L. Poston [12] took up the M_1 index (Formula (17)), which had been defined as a diversity index by the English statistician Edward Hugh Simpson [7], who derived it from the Italian statistician and sociologist Corrado Gini [16].

$$M_1 = 1 - \sum_{i=1}^{k} f_{x_i}^2 \tag{17}$$

This index can be interpreted as the complement of the probability that a random pair of samples belongs to the same category; in other words, it estimates the probability that the pair does not belong to the same category. This index has also been referred to as the differentiation index, livelihood differentiation index, and geographical differentiation index, depending on the context in which it has been used [17].

Gibbs and Poston [12] proposed a second index (M_2), which is the standardized version of the previous one, called the Index of Qualitative Variation (*IQV*). Refer to Formula (18). It is the most widely used measure of qualitative

variation, especially in the social sciences [11].

$$M_2 = IQV = \frac{k\left(1-\sum_{i=1}^{k} f_{x_i}^2\right)}{k-1} = \frac{k \times M_1}{k-1} = \frac{k}{k-1} \times M_1 \tag{18}$$

In the case of a constant random variable or random sample in which the n data points correspond to a single value a, the value of IQV is 0 (Formula (19)).

$$IQV = \frac{k\left(1-\sum_{i=1}^{k} f_i^2\right)}{k-1} = \frac{2 \times (1-1)}{2-1} = \frac{0}{1} = 0 \tag{19}$$

In the case of a uniform distribution, the value of IQV is 1 (Formula (20)).

$$IQV = \frac{k\left(1-\sum_{i=1}^{k} f_{x_i}^2\right)}{k-1} = \frac{k(1-1/k)}{k-1} = \frac{k(k-1)/k}{k-1} = \frac{k-1}{k-1} = 1 \tag{20}$$

This measure of variation applies to any type of distribution. It uses all the information of the distribution, but usually gives high values, especially as the frequency is more distributed (proximity to a uniform distribution).

7. Shannon's Relative Entropy

Entropy is the measure of disorder in a system of elements [18]. In probability theory, entropy is at its maximum when all elements are equiprobable and the presence of some elements does not allow predicting the appearance of others. The concept originates from thermodynamics and was applied to information theory by the American mathematician and electrical engineer Claude Elwood Shannon [13]. From there, it transitioned to statistics as a property to characterize both discrete and continuous distributions and to measure variability in qualitative variables (classification systems). At the population level, it is denoted by the capital Greek letter eta (H), and at the sample level, by the capital Latin letter E. Entropy is the mathematical expectation of Shannon's information or logarithm of the probability mass function [13]. Refer to Formula (21).

$$H(X) = E(I) = E\left[-\log_b\left(P(X=x)\right)\right] = E\left[-\log_b\left(f_X(x)\right)\right] \tag{21}$$

When the information (I_X) is in base 2, we speak of bits or binary units of information. When the information is in decimal base, we speak of dits or decimal units of information. When the information is in the natural base, we speak of nats or natural units of information. This last option is the most used.

For an empirical or sample frequency distribution of a qualitative variable A (with k values), the entropy is calculated as shown in Formula (22), where $f_n(a_i)$ represents relative frequency of the value a_i of the qualitative variable A, $n(a_i)$ denotes absolute frequency of the value a_i of the qualitative variable A, and n stands for sample size.

$$E = -\sum_{i=1}^{k} f_n(a_i)\ln\left(f_n(a_i)\right) = -\sum_{i=1}^{k} \frac{n_{a_i}}{n}\ln\left(\frac{n_{a_i}}{n}\right) \tag{22}$$

Any null frequency must be omitted in the entropy calculation. Proceeding in this manner, the statistic H_A can take values in the interval [0, ln(k)], where k is

the number of categories, including the categories omitted due to null frequency.

Knowing the maximum value of entropy in a discrete distribution, which corresponds to the discrete uniform distribution $U\{a, b\}$, the Shannon's normalized or relative entropy can be calculated. This maximum value is the natural logarithm of the cardinality or number of values (k) in the bounded interval from a (minimum) to b (maximum) of the support of a discrete uniform distribution: $\ln(k)$, where $k = \#\{a, b\}$. The relative entropy is denoted by *RelH* at the population level and *RelE* at the sample level, and it is the average entropy or information divided by its maximum value [13]. For an empirical or sample frequency distribution, the relative entropy is calculated using Formula (23). In the case of a constant random variable, the number of categories (k) is given a value of 2.

$$RelE = \frac{E}{\max(E)} = \frac{-\sum_{i=1}^{k} f_n(a_i)\ln(f_n(a_i))}{\ln(k)} = \frac{-\sum_{i=1}^{k} \frac{n_{x_i}}{n}\ln\left(\frac{n_{x_i}}{n}\right)}{\ln(k)} \quad (23)$$

Like the Gibbs-Poston measure [12], it applies to any type of distribution. It utilizes all the information of the distribution, but usually gives high values, particularly when the frequency is more evenly distributed (approaching a uniform distribution).

8. Pattern of Behavior of the Six Defined Variation Measures

Moral [3] studied the behavior of these six measures in relation to different distributions: the distribution of a constant random variable and eight discrete distributions with five nominal categories: the first with a distribution close to that of a constant random variable; the second with a distribution close to symmetry around a single mode; the third with strict symmetry around a single mode; the fourth close to a uniform distribution, although with a single mode; the fifth with a bimodal distribution; the sixth with a trimodal distribution; the seventh with a quadrimodal distribution; and the eighth with a uniform distribution.

After analyzing this data, Moral [3] notes that *FVR*, *UVR*, *WVR*, and *SDM* are more sensitive to the proximity to a distribution in which a category concentrates all the probability than *IQV* and *RelE*, as their value is closer to the expected value for the distribution of a constant random variable, which is 0. On the contrary, the last two indices are more sensitive to the proximity to a uniform distribution, albeit with a poorly defined mode, *i.e.*, a unimodal distribution with very similar frequencies, so their value is closer to the expected value for a uniform distribution, which is 1. When the distribution is uniform, implying that all categories have exactly the same probability or frequency, the value of all indices is 1, except for Wilcox's variation ratio, which cannot be calculated. It should be noted that the universal variation ratio closely resembles Freeman's variation ratio when the distribution is unimodal. The more defined the single mode is, the closer the latter two indices are to the value of 0, indicating the presence of a constant random variable.

In the case of more than one mode, Freeman's and Wilcox's variation ratios

cannot be calculated. With the presence of more than one mode, the increase in the index value is experienced more strongly in *IQV* and *RelE*. The universal variation ratio shows the most moderate increase with two modes, while the standard deviation from the mode is the most moderate with more than two modes.

9. Confidence Intervals for Qualitative Measures of Variability

There are essentially five methods to define the confidence interval of a statistic:

1) Asymptotic normal method: Also known as the Wald-type confidence interval [19]. This method relies on the central limit theorem. It is based on the convergence in distribution of the sampling distribution of the standardized statistic to a standard normal distribution. It necessitates a large random sample and that the sampling distribution has finite mean and variance. The delta method is typically employed to determine such mean and variance [20]. If the sampling distribution is normal, a large but random sample is not necessary.

2) Student's t-distribution method: It is a robust variant of the asymptotic normal method, utilized when the finite population standard deviation is unknown and the sample size is small [21].

3) Resampling methods: This approach involves obtaining the sampling distribution of the statistic by generating numerous samples from the original random sample and calculating the statistic in each of them. The point estimator of the parameter is the mean of this generated distribution, and the standard error is the sample standard deviation of the generated distribution. The first technique developed was the permutation of the data from the original sample, followed by the jackknife technique, in which one element is removed, and n samples of $n - 1$ data are generated. Finally, the technique of sampling with replacement improved the jackknife procedure. Confidence intervals, such as Wald-type or t-type, or using the percentiles of the simulated sampling distribution, are defined [22].

4) Bayesian method: In Bayesian statistics, confidence intervals are replaced by credible intervals. These intervals represent the range within which a parameter value falls with a certain probability, given the observed data and prior knowledge [23].

5) Exact method: The confidence interval is constructed using the probability distribution of the sample statistic. Unlike approximate methods, exact confidence intervals provide precise coverage probabilities for the true parameter of interest, regardless of the sample size or distributional assumptions. They have primarily been developed for statistics that follow a discrete distribution, such as binomial, multinomial, hypergeometric, or Poisson distribution [24].

Each method has its strengths and weaknesses. The choice of method depends on the sample size and the population distribution or sampling distribution of the statistic. For example, when the finite population standard deviation is un-

known and the sample size is small, the t-distribution method is preferred over the Wald-type method. Bootstrap methods are useful when the underlying distribution is unknown or non-normal. Bayesian methods are powerful when prior knowledge about the parameter of interest is available. When the probability distribution of the statistic is known, the exact confidence interval is preferred, especially with small samples.

Since the sampling distributions of the six measures of variation considered are unknown, repetitive sampling with replacement (bootstrap) can be used to obtain these distributions. Through nonparametric methods: percentile (PERC) and Corrected-Bias and Accelerated (BCa) percentile, confidence intervals can be defined [25]. In cases of bias and skewness, the latter method is preferred over the former [26]. When the bias-corrected percentile has an extreme value, the confidence interval cannot be calculated by the BCa method. In this case, the bias-corrected percentile can be given a value of 0, which yields a result equivalent to the percentile method, as suggested by Efron [27]. The calculation algorithms to obtain the confidence interval for both methods are shown below [28].

1) A random sample, denoted by *x*, of size *n* from the variable *X* is utilized as the starting point: $x=\{x_1,x_2,\cdots,x_n\}\subseteq X$. The variable *X* can be qualitative, ordinal, or quantitative. In this work, our focus is on qualitative variables.

2) The measure of variability for qualitative variable is calculated: $\hat{\theta}=t(x)$, which is the estimate of the parameter θ in the original sample of size *n*.

3) *B* independent samples of size *n* are drawn with replacement from the original random sample x of size *n*. It is recommended that there be at least 1000 draws ($B \geq 1000$), that the sample size be at least 30 ($n \geq 30$), and that the random sample be representative of the population from which it was collected.

4) In each of the B samples, the measure of variability $\theta_i^*\left(i=1,2,\cdots,B\right)$ is calculated, which generates the bootstrap sampling distribution of the statistic or estimator. This distribution can be represented by means of a histogram, using the Freedman-Diaconis rule to determine the uniform width of class intervals [29]. This method searches for an optimal interval width without making assumptions about the distribution of the statistic.

5) The bootstrap estimate of the parameter θ is the arithmetic mean of the bootstrap sampling distribution of the statistic or estimator and is denoted by $\hat{\theta}_{bootstrap}$ (Formula (24)). Bootstrap standard error is the sample or bias-corrected standard deviation of the bootstrap sampling distribution of the statistic or estimator and is denoted by $se_{boostrap}$ (Formula (25)). Bootstrap bias is the difference between the bootstrap estimate and the estimate in the original sample ($\hat{\theta}$) and is denoted by $bias_{boostrap}$ (Formula (26)).

$$\hat{\theta}_{bootstrap}=\bar{\hat{\theta}}^*=\frac{\sum_{i=1}^{B}\hat{\theta}_i^*}{B} \tag{24}$$

$$se_{bootstrap}=\sqrt{\frac{\sum_{i=1}^{B}\left(\hat{\theta}_i^*-\hat{\theta}_{bootstrap}\right)^2}{B-1}} \tag{25}$$

$$bias_{boostrap} = \hat{\theta}_{bootstrap} - \hat{\theta} \tag{26}$$

6) The $\alpha/2$ order quantile of the bootstrap sampling distribution of the statistic or estimator defines the lower limit, and the $1 - \alpha/2$ order quantile constitutes the upper limit of the confidence interval at $(1 - \alpha) \times 100$ by the percentile method (Formula (27)).

$$P\left(q_{\alpha/2}\left(\left\{ \hat{\theta}_i^* \right\}_{i=1}^{B} \right) \le \theta \le q_{1-\alpha/2}\left(\left\{ \hat{\theta}_i^* \right\}_{i=1}^{B} \right) \right) = 1 - \alpha \tag{27}$$

The quantile can be obtained using Rule 8 of the R software, as recommended by Hyndman and Fan [30] for calculating sample quantiles (Formula (28)). This rule expresses the order of the quantile (p) as the median of the order statistic i of a standard continuous uniform distribution U [0, 1], which serves as the non-informative prior when estimating a probability in Bayesian inference and is the distribution followed by randomly chosen probability values. If many random samples of size n are drawn from a standard continuous uniform distribution and the data of each sample are sorted in ascending order, the statistic of order i follows the beta distribution of shape parameters: $\alpha = i$ and $\beta = n + 1 - i$. Since α and β are not less than 1 and when they are equivalent are greater than 1, its median is approximately: $(\alpha - 1/3)/(\alpha + \beta - 2/3)$.

$$\begin{aligned} p &= \frac{\alpha - 1/3}{\alpha + \beta - 2/3} = \frac{i - 1/3}{i + n + 1 - i - 2/3} = \frac{i - 1/3}{n + 1/3} \\ i &= \frac{1}{3} + p\left(n + \frac{1}{3} \right) = \lfloor i \rfloor + \left(i - \lfloor i \rfloor \right) \\ x_{(p)} &= x_{(i)} + \left(i - \lfloor i \rfloor \right)\left(x_{(i+1)} - x_{(i)} \right) \end{aligned} \tag{28}$$

7) To obtain the confidence interval using the second method, we start by calculating the bias-corrected percentile, as shown in Formula (29), where I is the indicator function (0 when the condition is not met and 1 when it is met), and Φ^{-1} represents the probit function or quantile function of the standard normal distribution.

$$z_0^* = \Phi^{-1}\left(\frac{\sum_{i=1}^{B} I\left(\hat{\theta}_i^* \le \hat{\theta} \right)}{B} \right) \tag{29}$$

For extreme values, when the argument of the probit function in the Formula (29) approaches 0 or 1, z_0^* becomes undefined. In this case, z_0^* can be assigned a value of 0, so that the bootstrap Bias-Corrected and accelerated (BCa) confidence interval corresponds to the percentile bootstrap confidence interval [27].

8) The asymmetry correction factor (acceleration) is calculated using Formula (30) based on the jackknife method. Following this method, n samples are generated from the random sample x of size n, removing one sample datum in each sample, and the statistic or estimator $\hat{\theta}_{(-i)}$ is calculated in each of the n samples.

$$a = \frac{\sum_{i=1}^{n}\left(\sum_{i=1}^{n}\hat{\theta}_{(-i)}/n - \hat{\theta}_{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\sum_{i=1}^{n}\hat{\theta}_{(-i)}/n - \hat{\theta}_{(-i)}\right)^2\right]^{3/2}} \tag{30}$$

9) The orders of the bias-corrected and accelerated quantiles corresponding to the lower and upper limits of the confidence interval at $(1 - \alpha) \times 100$ are determined. Refer to Formula (31), where Φ is probit function, and $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are quantiles of a standard normal distribution.

$$p_{LL} = \Phi\left(z_0^* + \frac{z_0^* + z_{\alpha/2}}{1 - a\left(z_0^* + z_{\alpha/2}\right)}\right) \text{ and } p_{UL} = \Phi\left(z_0^* + \frac{z_0^* + z_{1-\alpha/2}}{1 - a\left(z_0^* + z_{1-\alpha/2}\right)}\right) \tag{31}$$

10) The p_{LL} order quantile of the bootstrap sampling distribution of the statistic or estimator defines the lower limit, and the p_{UL} order quantile defines the upper limit of the confidence interval at $(1 - \alpha) \times 100$ by the bias-corrected and accelerated percentile method, as shown in Formula (32). These quantiles can be computed using rule 8 of the R software [30].

$$P\left(q_{p_{LL}}\left(\left\{\hat{\theta}_i^*\right\}_{i=1}^{B}\right) \le \theta \le q_{p_{UL}}\left(\left\{\hat{\theta}_i^*\right\}_{i=1}^{B}\right)\right) = 1 - \alpha \tag{32}$$

10. Computation of Measures of Variability for Qualitative Variables with the R Software

One of the reasons for the underutilization of these measures of variability is their unavailability in statistical software. Below are two scripts for the R software, which can be adjusted to accommodate sample data other than those provided. The elements requiring change for such adjustment are highlighted in blue.

The first script pertains to a random sample with one mode, whereas the second script is designed for a random sample with two modes. In the first sample, the script calculates the six measures of variation and the asymptotic confidence interval for Kvalseth's *SDM*, given the unique mode.

For the two-mode sample, commands for *FVR* and *WRV* are excluded, along with the standard error and asymptotic confidence interval for the *SDM*, as they cannot be computed. In the first script, bootstrap confidence intervals are computed using the percentile method. In the second script, these confidence intervals are calculated using the bias-corrected and accelerated percentile method. All results are rounded to four decimal places. The first script provides low-definition graphics output generated by the R program, while the second script displays the commands used to save these graphics as high-definition image files (*.jpeg).

10.1. Example 1

From the qualitative variable A = "marital status in a population of adult men aged 20 to 60 years living in a northern Mexican border city", a random sample of 72 participants was drawn. It is required to represent its distribution by means

of a frequency table and a bar chart. Additionally, calculating the mode of the 72 sample data as a measure of central tendency, along with Freeman's Variation Ratio (*FVR*), Wilcox's Variation Ratio (*WVR*), Moral's Universal Variation Ratio (*UVR*), Kvalseth's Standard Deviation from Mode (*SDM*) with its asymptotic standard error, Gibbs-Poston Index of Qualitative Variation (*IQV*), and Shannon's Relative Entropy (*RelE*) as measures of variation is desired.

Furthermore, obtaining the 95% confidence interval for the six measures of variation is necessary. For this purpose, it is advisable to utilize bootstrap using the percentile method.

As additional information, including the bootstrap point estimate, standard error, and bias is necessary, along with representing the bootstrap distribution using a histogram with the density curve superimposed, following the Freedman-Diaconis rule to determine the width of class intervals [29].

---R script for Example 1---

```
# Data definition (common for all six measures of variation)
A<- c("Single", "Married", "Living together", "Separated", "Divorced", "Widowed")
a<- c(2, 1, 3, 3, 6, 5, 3, 2, 1, 1, 3, 2, 5, 6, 1, 1, 2, 3, 4, 1, 2, 1, 5, 2, 3, 2, 2, 3, 2, 2, 1, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 4, 2, 2, 4, 1, 3, 2, 2, 1, 5, 1, 1, 3, 2, 1, 1, 2, 2, 2, 1, 4, 2, 1, 4, 2, 1, 3, 2, 1)

# Frequency table and bar plot (common for all six measures)
frequency_table<- table(factor(a, levels = 1:6, labels = A))
Marital_status<- names(frequency_table)
abs_freq<- as.vector(frequency_table)
N <- sum(abs_freq)
rel_fre<- abs_freq/N
rel_freq = round(rel_fre, 4)
pct <- rel_freq * 100
complete_table<- data.frame(Marital_status, abs_freq, rel_freq, pct)
print_table<- complete_table[, c("Marital_status", "abs_freq", "rel_freq", "pct")]
cat("Frequency distribution of marital status among men\n")
print(print_table)
barplot(rel_freq, names.arg = A, xlab = "MS", ylab = "Relative frequency", main = "Bar chart: Marital status", las = 2, col = "lightgray")

# Mode calculation (common for all six measures)
modes <- Marital_status[abs_freq == max(abs_freq)]
mode_frequency<- max(rel_fre)
n <- length(a)
k <- length(Marital_status)
c <- length(modes)
cat("Sample size: n =", n, "\n")
```

```
cat("Number of nominal categories: k =", k, "\n")
cat("Modal categories: mo =", modes, "\n")
cat("Number of modal values: c =", c, "\n")
cat("Relative frequency of the mode: fmo =", round(mode_frequency, 4), "\n")

# Calculation of FVR and its 95% PERC bootstrap confidence interval
FVR <- 1 - mode_frequency
cat("Freeman's (1965) Variation Ratio: FVR =", round(FVR, 4), "\n")
set.seed(123)
B <- 1000
boot_FVR<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(a, replace = TRUE)
boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = A))
boot_mode<- max(boot_freq_table)/sum(boot_freq_table)
boot_FVR[i] <- 1 - boot_mode
}
BE_FVR <- mean(boot_FVR)
bias_FVR<- mean(boot_FVR) - FVR
se_FVR<- sd(boot_FVR)
cat("Bootstrap estimation for FVR:", round(BE_FVR, 4), "\n")
cat("Bootstrap bias for FVR:", round(bias_FVR, 4), "\n")
cat("Bootstrap standard error for FVR:", round(se_FVR, 4), "\n")
PERC_CI_FVR <- quantile(boot_FVR, c(0.025, 0.975), type = 8)
cat("The 95% PERC bootstrap confidence interval for FVR: [", round
(PERC_CI_FVR[1], 4), ",", round(PERC_CI_FVR[2], 4), "]\n")

# Freedman-Diaconis histogram with overlaid density curve for FVR
hist(boot_FVR, breaks = "fd", col = "lightgrey", border = "black", freq =
FALSE, main = "Freedman-Diaconis histogram with overlaid density curve",
xlab = "Bootstrap Freeman Variation Ratio", ylab = "Density", xlim = c(0, 1))
lines(density(boot_FVR), col="black", lwd=4)

# Calculation of WVR and its 95% PERC bootstrap confidence interval
WVR <- k/(k - 1) * (1 - mode_frequency)
cat("Wilcox's (1973) Variation ratio: WVR =", round(WVR, 4), "\n")
set.seed(123)
boot_WVR<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(a, replace = TRUE)
boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = A))
boot_mode<- max(boot_freq_table)/sum(boot_freq_table)
boot_WVR[i] <- k/(k - 1) * (1 - boot_mode)
```

```
}
BE_WVR <- mean(boot_WVR)
bias_WVR<- mean(boot_WVR) - WVR
se_WVR<- sd(boot_WVR)
cat("Bootstrap estimation for WVR:", round(BE_WVR, 4), "\n")
cat("Bootstrap bias for WVR:", round(bias_WVR, 4), "\n")
cat("Bootstrap standard error for WVR:", round(se_WVR, 4), "\n")
percentile_conf_interval_WVR<- quantile(boot_WVR, c(0.025, 0.975), type = 8)
cat("The 95% PERC bootstrap confidence interval for WVR: [", round (percentile_conf_interval_WVR[1], 4), ",", round(percentile_conf_interval_WVR[2], 4), "]\n")

# Freedman-Diaconis histogram with overlaid density curve for WVR
hist(boot_WVR, breaks = "fd", col = "lightgrey", border = "black", freq = FALSE,
main = "Freedman-Diaconis histogram with overlaid density curve",
xlab = "Bootstrap Wilcox Variation Ratio", ylab = "Density", xlim = c(0, 1))
lines(density(boot_WVR), col = "black", lwd = 4)

# Calculation of UVR and its 95% PERC bootstrap confidence interval
UVR <- k^2 /( k^2-1) * (1 - mode_frequency/c)
cat("Moral's (2022) Universal Variation Ratio: UVR =", round(UVR, 4), "\n")
set.seed(123)
boot_UVR<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(a, replace = TRUE)
boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = A))
boot_mode<- max(boot_freq_table)/sum(boot_freq_table)
boot_mode_num<- sum(boot_freq_table == max(boot_freq_table))
boot_UVR[i] <- k^2/(k^2 - 1) * (1 - boot_mode/boot_mode_num)
}
BE_UVR <- mean(boot_UVR)
bias_UVR<- mean(boot_UVR) - UVR
se_UVR<- sd(boot_UVR)
cat("Bootstrap estimation for UVR:", round(BE_UVR, 4), "\n")
cat("Bootstrap bias for UVR:", round(bias_UVR, 4), "\n")
cat("Bootstrap standard error for UVR:", round(se_UVR, 4), "\n")
PERC_CI_UVR <- quantile(boot_UVR, c(0.025, 0.975), type = 8)
cat("The 95% PERC bootstrap confidence interval for UVR: [", round (PERC_CI_UVR[1], 4), ",", round(PERC_CI_UVR[2], 4), "]\n")

# Freedman-Diaconis histogram with overlaid density curve for UVR
hist(boot_UVR, breaks = "fd", col = "lightgrey", border = "black", freq =
```

```
FALSE, main = "Freedman-Diaconis histogram with overlaid density curve", xlab = "Bootstrap Moral Universal Variation Ratio", ylab = "Density", xlim = c(0, 1))
  lines(density(boot_UVR), col="black", lwd=4)

  # Calculation of Tvalseth's Standard Deviation from Mode (SDM)
  f_mo<- mode_frequency
  SDM <- 1 - sqrt(sum((f_mo - rel_fre)^2)/(k - 1))
  cat("Tvalseth's (1988) Standard Deviation from Mode: SDM =", round(SDM, 4), "\n")

  # Calculating the asymptotic standard error and 95% confidence interval for SDM
  ase <- sqrt((f_mo * (1 - k * f_mo)^2 + sum(rel_fre * (f_mo - rel_fre)^2))/(N * (k - 1)^2 * (1 - SDM)^2) - (1 - SDM)^2/N)
  cat("Asymptotic Standard Error of SDM: ASE(SDM) =", round(ase, 4), "\n")
  alpha <- 0.05
  z_crit<- qnorm(1 - alpha/2)
  ACI<- c(SDM - z_crit * ase, SDM + z_crit * ase)
  cat("The 95% asymptotic confidence interval for SDM: [", round(ACI[1], 4), ",", round(ACI[2], 4), "]\n")

  # Calculation of the 95% PERC bootstrap confidence interval for SDM
  set.seed(123)
  boot_SDM<- numeric(B)
  for (i in 1:B) {
  boot_sample<- sample(a, replace = TRUE)
  boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = A))
  boot_mode_freq<- max(boot_freq_table)/sum(boot_freq_table)
  boot_SDM[i] <- 1 - sqrt(sum((boot_mode_freq - rel_fre)^2)/(k - 1))
  }
  BE_SDM <- mean(boot_SDM)
  bias_SDM<- mean(boot_SDM) - SDM
  se_SDM<- sd(boot_SDM)
  cat("Bootstrap estimation for SDM:", round(BE_SDM, 4), "\n")
  cat("Bootstrap bias for SDM:", round(bias_SDM, 4), "\n")
  cat("Bootstrap standard error for SDM:", round(se_SDM, 4), "\n")
  PERC_CI_SDM <- quantile(boot_SDM, c(0.025, 0.975), type= 8)
  cat("The 95% PERC bootstrap confidence interval for SDM: [", round(PERC_CI_SDM[1], 4), ",", round(PERC_CI_SDM[2], 4), "]\n")
  # Freedman-Diaconis histogram with overlaid density curve for SDM
  hist(boot_SDM, breaks = "fd", col = "lightgrey", border = "black", freq = FALSE, main = "Freedman-Diaconis histogram with overlaid density curve", xlab = "Bootstrap Tvalseth Standard Deviation from Mode", ylab = "Density",
```

```
xlim = c(0, 1))
  lines(density(boot_SDM), col="black", lwd=4)

  # Calculation of IQV and its 95% PERC bootstrap confidence interval
  IQV <- k/(k - 1) * (1 - sum(rel_fre^2))
  cat("Gibbs-Poston (1975) Index of Qualitative Variation: IQV =", round(IQV,
4), "\n")
  boot_IQV<- numeric(B)
  set.seed(123)
  for (i in 1:B) {
  boot_sample<- sample(a, replace = TRUE)
  boot_frequency_table<- table(factor(boot_sample, levels = 1:6, labels = A))
  boot_rel_fre<- as.vector(boot_frequency_table)/sum(boot_frequency_table)
  boot_IQV[i] <- k/(k - 1) * (1 - sum(boot_rel_fre^2))
  }
  BE_IQV <- mean(boot_IQV)
  bias_IQV<- mean(boot_IQV) - IQV
  se_IQV<- sd(boot_IQV)
  cat("Bootstrap estimation for IQV:", round(BE_IQV, 4), "\n")
  cat("Bootstrap bias for IQV:", round(bias_IQV, 4), "\n")
  cat("Bootstrap standard error for IQV:", round(se_IQV, 4), "\n")
  PERC_CI_IQV <- quantile(boot_IQV, c(0.025, 0.975), type = 8)
  cat("The 95% PERC bootstrap confidence interval for IQV: [", round
(PERC_CI_IQV[1], 4), ",", round(PERC_CI_IQV[2], 4), "]\n")

  # Freedman-Diaconis histogram with overlaid density curve for IQV
  hist(boot_IQV, breaks = "fd", col = "lightgrey", border = "black", freq =
FALSE, main = "Freedman-Diaconis histogram with overlaid density curve",
xlab = "Bootstrap Gibbs-Poston Index of Qualitative Variation", ylab = "Densi-
ty", xlim = c(0, 1))
  lines(density(boot_IQV), col="black", lwd=4)

  # Calculation of RelE and its 95% PERC bootstrap confidence interval
  entropy <- -sum(rel_fre * log(rel_fre + (rel_fre == 0)))
  RelE <- entropy/log(k)
  cat("Shannon's (1948) Entropy: E =", round(entropy, 4), "\n")
  cat("Shannon's (1948) Relative Entropy: RelE =", round(RelE, 4), "\n")
  set.seed(123)
  boot_RelE<- numeric(B)
  for (i in 1:B) {
  boot_sample<- sample(a, replace = TRUE)
  boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = A))
  boot_rel_fre<- as.vector(boot_freq_table)/sum(boot_freq_table)
  boot_entropy<- -sum(boot_rel_fre* log(boot_rel_fre + (boot_rel_fre== 0)))
```

```
boot_RelE[i] <- boot_entropy/log(k)
}
BE_RelE<- mean(boot_RelE)
bias_RelE<- mean(boot_RelE) - RelE
se_RelE<- sd(boot_RelE)
cat("Bootstrap estimation for RelE:", round(BE_RelE, 4), "\n")
cat("Bootstrap bias for RelE:", round(bias_RelE, 4), "\n")
cat("Bootstrap standard error for RelE:", round(se_RelE, 4), "\n")
PERC_CI_RelE<- quantile(boot_RelE, c(0.025, 0.975), type = 8)
cat("The 95% PERC bootstrap confidence interval for RelE: [", round
(PERC_CI_RelE[1], 4), ",", round(PERC_CI_RelE[2], 4), "]\n")

# Freedman-Diaconis histogram with overlaid density curve for RelE
hist(boot_RelE, breaks = "fd", col = "lightgrey", border = "black", freq =
FALSE, main = "Freedman-Diaconis histogram with overlaid density curve",
xlab = "Bootstrap Shannon Relative Entropy", ylab = "Density", xlim = c(0, 1))
lines(density(boot_RelE), col="black", lwd=4)
```

This script can be executed online as the R software is accessible at https://rdrr.io/snippets/ with over 19000 pre-installed packages available for free. Table 1 along with the requested statistics and confidence intervals are displayed as output. The graphs are excluded because they are low-definition. It should be noted that the ggplot2 library of the R software can be utilized to save these graphics as high-definition image files (*.jpeg), as shown the second script.

Sample size: $n = 72$
Number of nominal categories: $k = 6$
Modal categories: *mo* = Married
Number of modal values: $c = 1$
Relative frequency of the mode: *fmo* = 0.3889
Freeman's (1965) Variation Ratio: *FVR* = 0.6111
Bootstrap estimation for *FVR*: 0.597
Bootstrap bias for *FVR*: −0.0142
Bootstrap standard error for *FVR*: 0.0459
The 95% PERC bootstrap confidence interval for *FVR*: [0.5, 0.6806]

Table 1. Frequency distribution of marital status among men.

Marital_status	abs_freq	rel_freq	pct
Single	23	0.3194	31.94%
Married	28	0.3889	38.89%
Free union	10	0.1389	13.89%
Separated	5	0.0694	6.94%
Divorced	4	0.0556	5.56%
Widowed	2	0.0278	2.78%

Wilcox's (1973) Variation ratio: $WVR = 0.7333$
Bootstrap estimation for WVR: 0.7163
Bootstrap bias for WVR: −0.017
Bootstrap standard error for WVR: 0.055
The 95% PERC bootstrap confidence interval for WRV: [0.6, 0.8167]
Moral's (2022) Universal Variation Ratio: $UVR = 0.6286$
Bootstrap estimation for UVR: 0.6229
Bootstrap bias for UVR: −0.0057
Bootstrap standard error for UVR: 0.0682
The 95% PERC bootstrap confidence interval for UVR: [0.5143, 0.85]
Tvalseth's (1988) Standard Deviation from Mode: $SDM = 0.7133$
Asymptotic Standard Error of SDM: $ASE(SDM) = 0.061$
The 95% asymptotic confidence interval for SDM: [0.5939, 0.8328]
Bootstrap estimation for SDM: 0.6990
Bootstrap bias for SDM: −0.0143
Bootstrap standard error for SDM: 0.0434
The 95% PERC bootstrap confidence interval for SDM: [0.6047, 0.7743]
Gibbs-Poston (1975) Index of Qualitative Variation: $IQV = 0.8625$
Bootstrap estimation for IQV: 0.8505
Bootstrap bias for IQV: −0.012
Bootstrap standard error for IQV: 0.0359
The 95% PERC bootstrap confidence interval for IQV: [0.775, 0.9126]
Shannon's (1948) Entropy: $E = 1.4514$
Shannon's (1948) Relative Entropy: $RelE = 0.81$
Bootstrap estimation for $RelE$: 0.7896
Bootstrap bias for $RelE$: −0.0204
Bootstrap standard error for $RelE$: 0.0486
The 95% PERC bootstrap confidence interval for $RelE$: [0.6894, 0.8771]

10.2. Example 2

From the qualitative variable C = marital status of the population of adult women aged 18 to 60 years living in a border city in northern Mexico, a random sample was drawn. It is necessary to represent these data with a frequency table and a bar chart, calculate the mode as a measure of central tendency, as well as Moral's Universal Variation Ratio (UVR), Kvalseth's Standard Deviation from Mode (SDM), Gibbs-Poston Index of Qualitative Variation (IQV), and Shannon's Relative Entropy ($RelE$) as measures of variation. Additionally, obtaining the 95% confidence interval for the four qualitative measures of variation is desired. For this purpose, it is advisable to utilize bootstrap using the bias-corrected and accelerated percentile method. Furthermore, information on the bootstrap point estimate, standard error, and bias is needed, along with displaying the bootstrap distribution using a histogram, determining the width of class intervals by the Freedman-Diaconis rule [29].

```
---R script for Example 2---
# Data definition
C<- c("Single", "Married", "Free union", "Separated", "Divorced", "Widowed")
c<- c(2, 2, 1, 2, 2, 2, 2, 1, 2, 3, 1, 2, 5, 1, 2, 2, 1, 3, 2, 1, 2, 1, 5, 2, 2, 2, 2, 1, 2, 2, 4, 5, 3, 3, 3, 2, 1, 1, 5, 1, 1, 1, 2, 2, 3, 1, 2, 3, 4, 1, 4, 4, 1, 1, 3, 3, 1, 1, 1, 1, 6, 1, 4, 1, 1, 2, 2, 6, 2, 1, 1, 1, 1, 2, 3, 2, 2)

# Frequency table (common for all four measures of variation)
frequency_table<- table(factor(c, levels = 1:6, labels = C))
Marital_status<- names(frequency_table)
abs_freq<- as.vector(frequency_table)
N <- sum(abs_freq)
rel_fre<- abs_freq/N
rel_freq = round(rel_fre, 4)
pct <- rel_freq * 100
complete_table<- data.frame(Marital_status, abs_freq, rel_freq, pct)
print_table<- complete_table[, c("Marital_status", "abs_freq", "rel_freq", "pct")]
cat("Frequency distribution of marital status among women\n")
print(print_table)

# Bar plot using ggplot2 and saved as a JPEG file (common for all four measures)
library(ggplot2)
df<- data.frame(Marital_Status = factor(C, levels = C), Relative_Frequency = rel_fre)
plot <- ggplot(df, aes(x = Marital_Status, y = Relative_Frequency)) +
geom_bar(stat = "identity", fill = "lightgray", color = "black") +
labs(x = "Marital status", y = "Relative frequency") +
theme(axis.text.x.bottom = element_text(angle = 10, hjust = 0.5, size = 7), axis.text.y = element_text(size = 7), axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), panel.background = element_rect(fill = "white"), panel.grid.major = element_blank(), panel.grid.minor = element_blank(), axis.line = element_line(color = "black"))
jpeg("bar_chart.jpeg", width = 800, height = 600, units = "px", res = 300)
print(plot)
dev.off()
plot

# Mode calculation (common for all four measures)
modes <- Marital_status[abs_freq == max(abs_freq)]
mode_frequency<- max(rel_fre)
n <- length(c)
k <- length(Marital_status)
```

```
nmo<- length(modes)
cat("Sample size: n =", n, "\n")
cat("Number of nominal categories: k =", k, "\n")
cat("Modal categories: mo =", modes, "\n")
cat("Number of modal values: c =", nmo, "\n")
cat("Relative frequency of the mode: fmo =", round(mode_frequency, 4), "\n")

# Calculation of UVR and its bootstrap sampling distribution
UVR <- k^2 /( k^2-1) * (1 - mode_frequency/nmo)
cat("Moral's (2022) Universal Variation Ratio: UVR =", round(UVR, 4), "\n")
set.seed(123)
B <- 1000
boot_UVR<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(c, replace = TRUE)
boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = C))
boot_mode_freq<- max(boot_freq_table)/sum(boot_freq_table)
boot_mode_num<- sum(boot_freq_table == max(boot_freq_table))
boot_UVR[i] <- k^2/(k^2 - 1) * (1 - boot_mode_freq/boot_mode_num)
}
BE_UVR <- mean(boot_UVR)
bias_UVR<- mean(boot_UVR) - UVR
se_UVR<- sd(boot_UVR)
cat("Bootstrap estimation for UVR:", round(BE_UVR, 4), "\n")
cat("Bootstrap bias for UVR:", round(bias_UVR, 4), "\n")
cat("Bootstrap standard error for UVR:", round(se_UVR, 4), "\n")

# Histogram with overlaid density curve for UVR (save as JPEG file)
hist_data1 <- data.frame(boot_UVR)
q25 <- quantile(hist_data1$boot_UVR, 0.25, type = 8)
q75 <- quantile(hist_data1$boot_UVR, 0.75, type = 8)
iqr<- q75 - q25
FD <- 2 * iqr/(length(hist_data1$boot_UVR)^(1/3))
hist_plot1 <- ggplot(hist_data1, aes(x = boot_UVR)) +
geom_histogram(binwidth = FD, fill = "lightgrey", color = "black", aes(y = ..density..)) + geom_density(color = "black", size = 1.5) + labs(x = "Bootstrap Moral's UVR", y = "Density") + theme(panel.background = element_rect(fill = "white"), axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text (size = 8), axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line = element_line(color = "black"))
jpeg("histogram_UVR.jpeg", width = 800, height = 600, units = "px", res = 300)
print(hist_plot1)
```

```
dev.off()
hist_plot1

# Bias-corrected percentile for UVR
z_0_UVR<- qnorm(sum(boot_UVR<= UVR)/B)
if (is.infinite(z_0_UVR)) {z_0_UVR<- 0} else {z_0_UVR<- z_0_UVR}
cat("Bias-corrected percentile for UVR:", round(z_0_UVR, 4), "\n")

# Skewness correction factor (acceleration) using jackknife estimation for UVR
jackknife_UVR<- numeric(n)
for (i in 1:n) {
jackknife_sample<- c[-i]
jackknife_freq_table<- table(factor(jackknife_sample, levels = 1:6, labels = C))
jackknife_mode_freq<- max(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_mode_num<- sum(jackknife_freq_table == max(jackknife_freq_table))
jackknife_UVR[i] <- k^2/(k^2 - 1) * (1 - jackknife_mode_freq/jackknife_mode_num)
}
jackknife_UVR_mean<- sum(jackknife_UVR)/n
a_UVR<- sum((jackknife_UVR_mean - jackknife_UVR)^3)/(6 * sum ((jackknife_UVR_mean - jackknife_UVR)^2)^(3/2))
cat("Skewness correction factor (acceleration):", round(a_UVR, 4), "\n")

# BCa bootstrap confidence interval for UVR
z_LL<- qnorm(0.025)
z_UL<- qnorm(0.975)
LL_BCa_UVR<- pnorm(z_0_UVR + (z_0_UVR + z_LL)/(1 - a_UVR* (z_0_UVR + z_LL)))
UL_BCa_UVR<-pnorm(z_0_UVR + (z_0_UVR + z_UL)/(1 - a_UVR* (z_0_UVR +z_UL)))
BCa_confidence_interval_UVR<- quantile(boot_UVR, probs = c(LL_BCa_UVR, UL_BCa_UVR), type = 8)
cat("The 95% BCabootstrapconfidenceintervalfor UVR: [", round(BCa_confidence_interval_UVR[1], 4), ",", round(BCa_confidence_interval_UVR[2], 4), "]\n")

# Calculation of SDM and its bootstrap sampling distribution
f_mo<- mode_frequency
SDM <- 1 - sqrt(sum((f_mo - rel_fre)^2)/(k - 1))
cat("Tvalseth's (1988) Standard Deviation from Mode: SDM =", round(SDM, 4), "\n")
set.seed(123)
```

```
boot_SDM<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(c, replace = TRUE)
boot_frequency_table<- table(factor(boot_sample, levels = 1:6, labels = C))
boot_mode_frequency<-
max(boot_frequency_table)/sum(boot_frequency_table)
boot_rel_fre<- as.vector(boot_frequency_table)/sum(boot_frequency_table)
boot_SDM[i] <- 1 - sqrt(sum((boot_mode_frequency-boot_rel_fre)^2)/(k -
1))
}
BE_SDM <- mean(boot_SDM)
bias_SDM<- mean(boot_SDM) - SDM
se_SDM<- sd(boot_SDM)
cat("Bootstrap estimation for SDM:", round(BE_SDM, 4), "\n")
cat("Bootstrap bias for SDM:", round(bias_SDM, 4), "\n")
cat("Bootstrap standard error for SDM:", round(se_SDM, 4), "\n")

# Histogram with overlaid density curve for SDM (save as JPEG file)
hist_data2 <- data.frame(boot_SDM)
q25 <- quantile(hist_data2$boot_SDM, 0.25, type = 8)
q75 <- quantile(hist_data2$boot_SDM, 0.75, type = 8)
iqr<- q75 - q25
FD <- 2 * iqr/(length(hist_data2$boot_SDM)^(1/3))
hist_plot2 <- ggplot(hist_data2, aes(x = boot_SDM)) + geom_histogram
(binwidth = FD, fill = "lightgrey", color = "black", aes(y = ..density..)) +
geom_density(color = "black", size = 1.5) + labs(x = "Bootstrap Tvalseth's SDM",
y = "Density") + theme(panel.background = element_rect(fill = "white"),
axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text(size = 8),
axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line
= element_line(color = "black"))
jpeg("histogram_SDM.jpeg", width = 800, height = 600, units = "px", res =
300)
print(hist_plot2)
dev.off()
hist_plot2

# Bias-corrected percentile for SDM
z_0_SDM <- qnorm(sum(boot_SDM<= SDM)/B)
if (is.infinite(z_0_SDM)) {z_0_SDM <- 0} else {z_0_SDM <- z_0_SDM}
cat("Bias-corrected percentile for SDM:", round(z_0_SDM, 4), "\n")

# Skewness correction factor (acceleration) using jackknife estimation for
SDM
jackknife_SDM<- numeric(n)
```

```
for (i in 1:n) {
jackknife_sample<- c[-i]
jackknife_freq_table<- table(factor(jackknife_sample, levels = 1:6, labels = C))
jackknife_mode_freq<- max(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_rel_fre<- as.vector(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_SDM[i] <- 1 - sqrt(sum((jackknife_mode_freq - jack-
knife_rel_fre)^2)/(k - 1))
}
jackknife_SDM_mean<- sum(jackknife_SDM)/n
a_SDM<- sum((jackknife_SDM_mean - jackknife_SDM)^3)/(6 *
sum((jackknife_SDM_mean - jackknife_SDM)^2)^(3/2))
cat("Skewness correction factor (acceleration):", round(a_SDM, 4), "\n")

# BCa bootstrap confidence interval for SDM
LL_BCa_SDM<- pnorm(z_0_SDM + (z_0_SDM + z_LL)/(1 - a_SDM* (z_0_SDM
+ z_LL)))
UL_BCa_SDM<-pnorm(z_0_SDM + (z_0_SDM + z_UL)/(1- a_SDM* (z_0_SDM
+ z_UL)))
BCa_confidence_interval_SDM<- quantile(boot_SDM, probs = c(LL_BCa_
SDM, UL_BCa_SDM), type = 8)
cat("The 95% BCa bootstrap confidence interval for SDM: [", round
(BCa_confidence_interval_SDM[1], 4), ",", round(BCa_confidence_ inter-
val_SDM[2], 4), "]\n")

# Calculation of IQV and its bootstrap sampling distribution
IQV <- k/(k - 1) * (1 - sum(rel_fre^2))
cat("Gibbs-Poston (1975) Index of Qualitative Variation: IQV =", round(IQV,
4), "\n")
set.seed(123)
boot_IQV<- numeric(B)
for (i in 1:B) {
boot_sample<- sample(c, replace = TRUE)
boot_frequency_table<- table(factor(boot_sample, levels = 1:6, labels = C))
boot_rel_fre<- as.vector(boot_frequency_table)/sum(boot_frequency_table)
boot_IQV[i] <- k/(k - 1) * (1 - sum(boot_rel_fre^2))
}
BE_IQV <- mean(boot_IQV)
bias_IQV<- mean(boot_IQV) - IQV
se_IQV<- sd(boot_IQV)
cat("Bootstrap estimation for IQV:", round(BE_IQV, 4), "\n")
cat("Bootstrap bias for IQV:", round(bias_IQV, 4), "\n")
cat("Bootstrap standard error for IQV:", round(se_IQV, 4), "\n")
# Histogram with overlaid density curve for IQV (save as JPEG file)
hist_data3 <- data.frame(boot_IQV)
```

```
q25 <- quantile(hist_data3$boot_IQV, 0.25, type = 8)
q75 <- quantile(hist_data3$boot_IQV, 0.75, type = 8)
iqr<- q75 - q25
FD <- 2 * iqr/(length(hist_data3$boot_IQV)^(1/3))
hist_plot3 <- ggplot(hist_data3, aes(x = boot_IQV)) + geom_histogram (binwidth = FD, fill = “lightgrey”, color = “black”, aes(y = ..density..)) + geom_density(color = “black”, size = 1.5) + labs(x = “Bootstrap Gibbs-Poston IQV”, y = “Density”) + theme(panel.background = element_rect(fill = “white”), axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text(size = 8), axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line = element_line(color = “black”))
jpeg(“histogram_IQV.jpeg”, width = 800, height = 600, units = “px”, res = 300)
print(hist_plot3)
dev.off()
hist_plot3

# Bias-corrected percentile for IQV
z_0_IQV <- qnorm(sum(boot_IQV<= IQV)/B)
if (is.infinite(z_0_IQV)) {z_0_IQV <- 0} else {z_0_IQV <- z_0_IQV}
cat(“Bias-corrected percentile for IQV:”, round(z_0_IQV, 4), “\n”)

# Skewness correction factor (acceleration) using jackknife estimation for IQV
jackknife_IQV<- numeric(n)
for (i in 1:n) {
jackknife_sample<- c[-i]
jackknife_freq_table<- table(factor(jackknife_sample, levels = 1:6, labels = C))
jackknife_mode_freq<- max(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_rel_fre<- as.vector(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_IQV[i] <- k/(k - 1) * (1 - sum(jackknife_rel_fre^2))
}
jackknife_IQV_mean<- sum(jackknife_IQV)/n
a_IQV<- sum((jackknife_IQV_mean - jackknife_IQV)^3)/(6 * sum((jackknife_IQV_mean - jackknife_IQV)^2)^(3/2))
cat(“Skewness correction factor (acceleration):”, round(a_IQV, 4), “\n”)

# BCa bootstrap confidence interval for IQV
LL_BCa_IQV<- pnorm(z_0_IQV + (z_0_IQV + z_LL)/(1 - a_IQV* (z_0_IQV + z_LL)))
UL_BCa_IQV<- pnorm(z_0_IQV + (z_0_IQV + z_UL)/(1 - a_IQV* (z_0_IQV + z_UL)))
BCa_confidence_interval_IQV<- quantile(boot_IQV, probs = c(LL_BCa_IQV, UL_BCa_IQV), type = 8)
cat(“The 95% BCa bootstrap confidence interval for IQV: [”, round (BCa_confidence_interval_IQV[1], 4), “,”, round(BCa_confidence_ inter-
```

```
val_IQV[2], 4), "]\n")

  # Calculation of RelE and its bootstrap sampling distribution
  entropy <- -sum(rel_fre * log(rel_fre + (rel_fre == 0)))
  RelE <- entropy/log(k)
  cat("Shannon's (1948) Entropy: E =", round(entropy, 4), "\n")
  cat("Shannon's (1948) Relative Entropy: RelE =", round(RelE, 4), "\n")
  set.seed(123)
  boot_RelE<- numeric(B)
  for (i in 1:B) {
  boot_sample<- sample(c, replace = TRUE)
  boot_freq_table<- table(factor(boot_sample, levels = 1:6, labels = C))
  boot_rel_fre<- as.vector(boot_freq_table)/sum(boot_freq_table)
  boot_entropy<- -sum(boot_rel_fre * log(boot_rel_fre + (boot_rel_fre == 0)))
  boot_RelE[i] <- boot_entropy/log(k)
  }
  BE_RelE<- mean(boot_RelE)
  bias_RelE<- mean(boot_RelE) - RelE
  se_RelE<- sd(boot_RelE)
  cat("Bootstrap estimation for RelE:", round(BE_RelE, 4), "\n")
  cat("Bootstrap bias for RelE:", round(bias_RelE, 4), "\n")
  cat("Bootstrap standard error for RelE:", round(se_RelE, 4), "\n")

  # Histogram with overlaid density curve for RelE (save as JPEG file)
  hist_data4 <- data.frame(boot_RelE)
  q25 <- quantile(hist_data4$boot_RelE, 0.25, type = 8)
  q75 <- quantile(hist_data4$boot_RelE, 0.75, type = 8)
  iqr<- q75 - q25
  FD <- 2 * iqr/(length(hist_data4$boot_RelE)^(1/3))
  hist_plot4 <- ggplot(hist_data4, aes(x = boot_RelE)) + geom_histogram
(binwidth = FD, fill = "lightgrey", color = "black", aes(y = ..density..)) +
geom_density(color = "black", size = 1.5) + labs(x = "Bootstrap Shannon's RelE",
y = "Density") + theme(panel.background = element_rect(fill = "white"),
axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text(size = 8),
axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line
= element_line(color = "black"))
  jpeg("histogram_RelE.jpeg", width = 800, height = 600, units = "px", res =
300)
  print(hist_plot4)
  dev.off()
  hist_plot4

  # Bias-corrected percentile for RelE
  z_0_RelE <- qnorm(sum(boot_RelE<=RelE)/B)
```

```
if (is.infinite(z_0_RelE)) {z_0_RelE <- 0} else {z_0_RelE <- z_0_RelE}
cat("Bias-correctedpercentile for RelE:", round(z_0_RelE, 4), "\n")

# Skewness correction factor (acceleration) using jackknife estimation for RelE
jackknife_RelE<- numeric(n)
for (i in 1:n) {
jackknife_sample<- c[-i]
jackknife_freq_table<- table(factor(jackknife_sample, levels = 1:6, labels = C))
jackknife_mode_freq<- max(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_rel_fre<- as.vector(jackknife_freq_table)/sum(jackknife_freq_table)
jackknife_entropy<- -sum(jackknife_rel_fre * log(jackknife_rel_fre + (jackknife_rel_fre == 0)))
jackknife_RelE[i] <- jackknife_entropy/log(k)
}
jackknife_RelE_mean<- sum(jackknife_RelE)/n
a_RelE<- sum((jackknife_RelE_mean - jackknife_RelE)^3)/(6 * sum((jackknife_RelE_mean - jackknife_RelE)^2)^(3/2))
cat("Skewness correction factor (acceleration):", round(a_RelE, 4), "\n")

# BCa bootstrap confidence interval for RelE
LL_BCa_RelE<- pnorm(z_0_RelE + (z_0_RelE + z_LL)/(1 - a_RelE* (z_0_RelE + z_LL)))
UL_BCa_RelE<- pnorm(z_0_RelE + (z_0_RelE + z_UL)/(1 - a_RelE* (z_0_RelE + z_UL)))
BCa_confidence_interval_RelE<- quantile(boot_RelE, probs = c(LL_BCa_RelE, UL_BCa_RelE), type = 8)
cat("The 95% BCa bootstrap confidence interval for RelE: [", round (BCa_confidence_interval_RelE[1], 4), ",", round(BCa_confidence_interval_ RelE[2], 4), "]\n")
```

When the script is executed with the R software installed on the personal computer, Table 2, along with the requested statistics and confidence intervals, is displayed as output alongside Figures 1-5. If the script is run online, the table, statistics, and low-definition graphs appear, but the *.jpeg files are not created.

Table 2. Sample frequency distribution of marital status in women.

Marital_status	abs_freq	rel_freq	pct
Single	28	0.3636	36.36%
Married	28	0.3636	36.36%
Free union	10	0.1299	12.99%
Separated	5	0.0649	6.49%
Divorced	4	0.0519	5.19%
Widowed	2	0.0260	2.60%

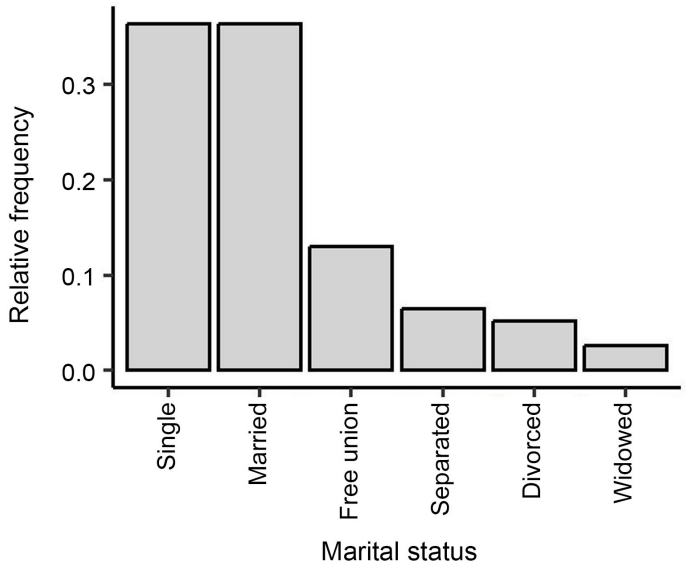

Figure 1. Bar chart of marital status in women.

Sample size: $n = 77$
Number of nominal categories: $k = 6$
Modal categories: mo = Single Married
Number of modal values: $c = 2$
Relative frequency of the mode: $fmo = 0.3636$
Moral's (2022) Universal Variation Ratio: $UVR = 0.8416$
Bootstrap estimation for UVR: 0.6257
Bootstrap bias for UVR: −0.2158
Bootstrap standard error for UVR: 0.064
Bias-corrected percentile for UVR: 2.1201
Skewness correction factor (acceleration): −0.0194
The 95% BCa bootstrap confidence interval for UVR: [0.8482, 0.875]

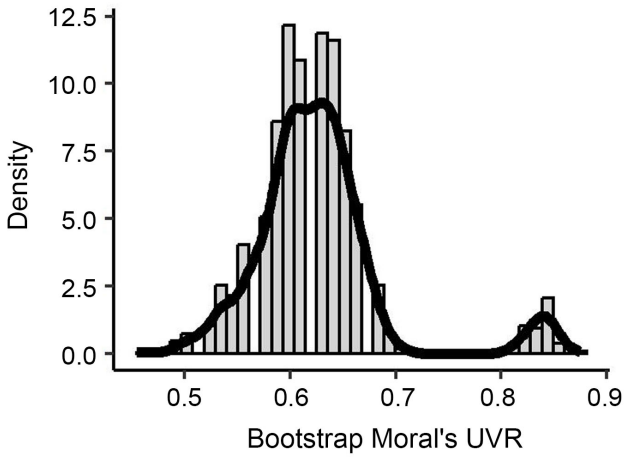

Figure 2. Freedman-Diaconis histogram and density curve of the bootstrap UVR distribution.

Tvalseth's (1988) Standard Deviation from Mode: $SDM = 0.7335$
Bootstrap estimation for *SDM*: 0.6963
Bootstrap bias for *SDM*: −0.0372
Bootstrap standard error for *SDM*: 0.0438
Bias-corrected percentile for *SDM*: 0.8134
Skewness correction factor (acceleration): 0.0211
The 95% BCa bootstrap confidence interval for *SDM*: [0.6853, 0.8162]

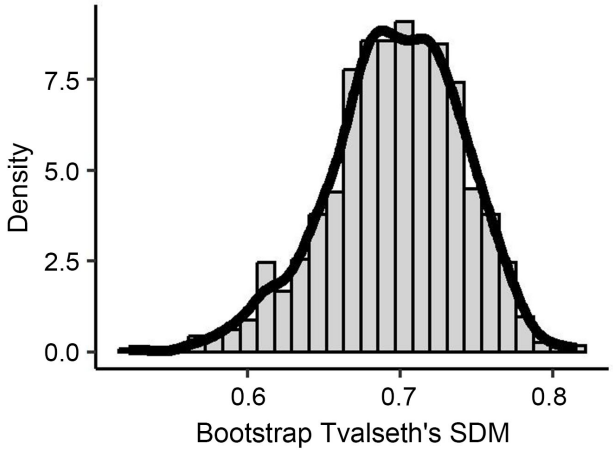

Figure 3. Freedman-Diaconis histogram and density curve of the bootstrap SDM distribution.

Gibbs-Poston (1975) Index of Qualitative Variation: $IQV = 0.8533$
Bootstrap estimation for *IQV*: 0.8435
Bootstrap bias for *IQV*: −0.0098
Bootstrap standard error for *IQV*: 0.0337
Bias-corrected percentile for *IQV*: 0.2378
Skewness correction factor (acceleration): 0.0212
The 95% BCa bootstrap confidence interval for *IQV*: [0.7942, 0.9229]

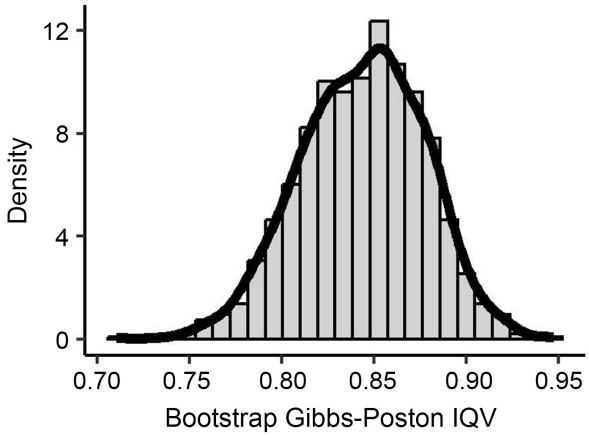

Figure 4. Freedman-Diaconis histogram and density curve of the bootstrap IQV distribution.

Shannon's (1948) Entropy: $E = 1.4268$
Shannon's (1948) Relative Entropy: $RelE = 0.7963$
Bootstrap estimation for *RelE*: 0.779
Bootstrap bias for *RelE*: −0.0173
Bootstrap standard error for *RelE*: 0.047
Bias-corrected percentile for *RelE*: 0.3319
Skewness correction factor (acceleration): 0.0303
The 95% BCa bootstrap confidence interval for *RelE*: [0.7207, 0.903]

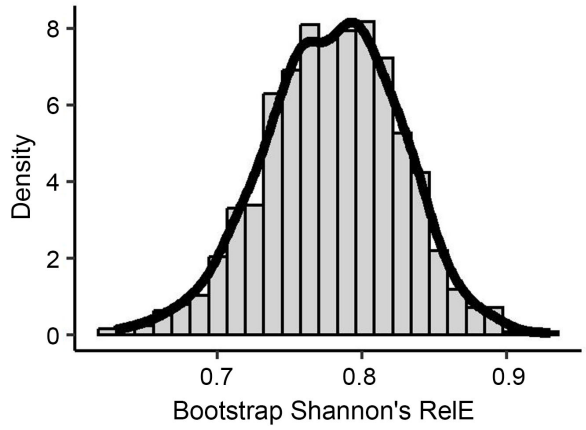

Figure 5. Freedman-Diaconis histogram and density curve of the bootstrap RelE distribution.

11. Conclusions

The present article focuses on nominal category qualitative variables but can also be applied to ordered category variables with a discrete set of categories. For these variables, Kvalseth's *SDM* is highly recommended as it utilizes more information than the variation ratios and is less affected by the ceiling effect induced by the proximity to the upper bounding distribution (uniformitarian distribution) than *IQV* and *RelE* [3]. Although Kvalseth [11] initially restricted the use of *SDM* to unimodal distributions, this measure is applicable to multimodal distributions and even to a uniform distribution, just like Moral's *UVR*, which is the best option among the variation ratios that are very easy to calculate. It should be noted that the most commonly used measure of variability is the Gibbs-Poston *IQV* [12] [31], and another important measure is Shannon's relative entropy [32]. Nevertheless, both measures are significantly influenced by proximity to the uniform distribution [3].

By utilizing bootstrap, it is feasible to derive the confidence interval and even the standard error. When the distribution is unknown, the preferable option is a non-parametric bootstrap method. Among these, the percentile method and the bias-corrected and accelerated percentile method stand out [26]. Both methods are perfectly valid when the bias and acceleration are minimal. However, when one of these two indices becomes moderate ($|bias| \geq 0.1$ or $|a| \geq 0.025$), the bi-

as-corrected and accelerated percentile method outperforms [33]. The primary issue with the latter method arises when the bias-corrected percentile (z_0) exhibits an extreme value. In such instances, assigning a value of 0 to z_0 yields a result equivalent to the percentile method, as proposed by Efron [27].

In each of the two scripts presented, its bootstrap sample is the same for all calculated measurements by using one seed and staying constant when generating the six or four sampling distributions. The selection of the seed (123) is arbitrary and could be any number. However, it is customary to employ fixed numbers or straightforward patterns to facilitate code reproducibility [28]. At the same time, the assessment of asymmetry (acceleration) is conducted using the jackknife method, ensuring complete stability and consistency in the confidence intervals across various measures of variation. It should be noted that these scripts should be applied with random, representative, and sufficiently large samples, with a minimum of 30 data points for the estimation to be deemed valid [26] [27].

The application of these six measures of variability is feasible with discrete and continuous variables, particularly when their distribution exhibits a defined peak. Nonetheless, it is discouraged due to the underutilization of the information inherent in the data compared to absolute or relative measures based on the average or median of differential scores concerning the arithmetic mean or median [34]. With discrete quantitative variables, the mode is still estimated using the maximum frequency value method, akin to qualitative variables [35]. However, for continuous quantitative variables, it necessitates employing a density estimation method to identify the value with the highest density [36].

An additional issue with samples of continuous variables is the necessity to discretize the distribution into k class intervals, which further exacerbates the loss of information and disregards the quantitative nature of the sample data [37]. To determine the number of class intervals, a viable option is Freedman-Diaconis rule [29]. Among all measures of variation for quantitative variables, Kvalseth's *SDM* and estimation of the mode via the maximum density value would be the optimal choice for continuous quantitative variables with unimodal distribution [36]. It's worth noting that, from the density function of a continuous distribution, Shannon's information can be directly calculated using integrals. Additionally, with certain continuous functions, one can ascertain the maximum entropy and obtain the relative entropy through the quotient between the entropy and its maximum [38].

It is recommended to utilize measures of variability, such as Kvalseth's *SDM* [11] and Moral's *UVR* [3], alongside well-known measures like the Gibbs-Poston *IQV* [12] and Shannon's relative entropy [13]. Additionally, measures of shape, such as Moral's skewness and peakedness, [39] should be used in conjunction with frequency tables, bar charts, and the measure of central tendency, the mode, when describing qualitative variables. Statistical tests to check whether one or more point data represent outliers are well known with quantitative variables [40] [41] [42]. However, there are new methods for detecting outliers with qualitative

variables, such as k-mode based cluster analysis [43].

Persisting on the development of descriptive measures for these variables, which are highly relevant in the social and health sciences, is important, as this area remains scarcely addressed in mathematical statistics [4] [8]. What might be the directions of future research in qualitative variation? Apart from confidence intervals, inferential tests for comparing measures of variation are an important future direction. Kvalseth's work with standard deviation from the mode [11] and a later-introduced measure of the odds measure of qualitative variation [44] are situated in this line. A related approach to qualitative variation measures that has been developed for some time in social research is segregation indices [45], among which the Hutchens index [46] [47] can be highlighted. Interval estimation and inferential comparisons with these indices may also be interesting lines of development.

Acknowledgements

The author expresses gratitude to 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] Wilcox, A.R. (1973) Indices of Qualitative Variation and Political Measurement. *The Western Political Quarterly*, **26**, 325-343. https://doi.org/10.1177/106591297302600209

[2] Agresti, A. and Agresti, B.F. (1978) Statistical Analysis of Qualitative Variation. *Sociological Methodology*, **9**, 204-237. https://doi.org/10.2307/270810

[3] Moralde la Rubia, J. (2022) Una Medida de Variación para Datos Cualitativos con Cualquier Tipo de Distribución [A Meausre of Variation for Qualitattive Data with Any Type of Distribution]. *Psychologia*, **16**, 63-76. https://doi.org/10.21500/19002386.5642

[4] Levitt, H.M. (2021) Qualitative Generalization, Not to the Population but to the Phenomenon: Reconceptualizing Variation in Qualitative Research. *Qualitative Psychology*, **8**, 95-110. https://doi.org/10.1037/qup0000184

[5] Maxwell, J.A. (2021) Why Qualitative Methods Are Necessary for Generalization. *Qualitative Psychology*, **8**, 111-118. https://doi.org/10.1037/qup0000173

[6] Golan, A. and Harte, J. (2022) Information Theory: A Foundation for Complexity Science. *Proceedings of the National Academy of Sciences of the United States of America*, **119**, e2119089119. https://doi.org/10.1073/pnas.2119089119

[7] Simpson, E.H. (1949) Measurement of Diversity. *Nature*, **163**, 688. https://doi.org/10.1038/163688a0

[8] Li, Y., Garg, H. and Deng, Y. (2020) A New Uncertainty Measure of Discrete Z-Numbers. *International Journal of Fuzzy Systems*, **22**, 760-776. https://doi.org/10.1007/s40815-020-00819-8

[9] Weiss, C.H. (2019) On the Sample Coefficient of Nominal Variation. In: Steland, A., Rafajłowicz, E. and Okhrin, O., Eds., *Stochastic Models, Statistics and Their Applications*, Springer, Cham, 239-250.

[10] Freeman, L.C. (1965) Elementary Applied Statistics for Students in Behavioral Sciences. John Wiley and Sons, New York.

[11] Kvalseth, T.O. (1988) Measuring Variation for Nominal Data. *Bulletin of the Psychonomic Society*, **26**, 433-436. https://doi.org/10.3758/BF03334906

[12] Gibbs, J.P., and Poston Jr., D.L. (1975) The Division of Labor: Conceptualization and Related Measures. *Social Forces*, **53**, 468-476. https://doi.org/10.2307/2576589

[13] Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[14] Deacon, D. and Stanyer, J. (2021) Media Diversity and the Analysis of Qualitative Variation. *Communication and the Public*, **6**, 19-32. https://doi.org/10.1177/20570473211006481

[15] Evren, A., Tuna, E., Ustaoglu, E. and Sahin, B. (2021) Some Dominance Indices to Determine Market Concentration. *Journal of Applied Statistics*, **48**, 2755-2775. https://doi.org/10.1080/02664763.2021.1963421

[16] Gini, C. (1912) Variabilitá e Mulabilitá: Contributo allo Studio delle Distribuzioni edelle Relazioni Statistiche. Tipografia di Paolo Cuppini, Bologna.

[17] Keren, S., Svoboda, M., Janda, P. and Nagel, T.A. (2020) Relationships between Structural Indices and Conventional Stand Attributes in an Old-Growth Forest in Southeast Europe. *Forests*, **11**, Article 4. https://doi.org/10.3390/f11010004

[18] Sharp, K. (2019). Entropy and the Tao of Counting: A Brief Introduction to Statistical Mechanics and the Second Law of Thermodynamics. Springer, Cham. https://doi.org/10.1007/978-3-030-35457-2

[19] Wald, A. (1939) Contributions of the Theory of Statistical Estimation and Testing Hypotheses. *Annals of Mathematical Statistics*, **10**, 299-326. https://doi.org/10.1214/aoms/1177732144

[20] Zepeda-Tello, R., Schomaker, M., Maringe, C., Smith, M.J., Belot, A., Rachet, B., Schnitzer, M.E. and Luque-Fernandez, M.A. (2022) Delta Method in Epidemiology: An Applied and Reproducible Tutorial. arXiv: 2206.15310.

[21] Janczyk, M. and Pfister, R. (2023) Confidence Intervals. In: Janczyk, M. and Pfister, R., Eds., *Understanding Inferential Statistics: From A for Significance Test to Z for Confidence Interval*, Springer, Heidelberg, 69-80. https://doi.org/10.1007/978-3-662-66786-6_6

[22] James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J. (2023) Resampling Methods. In: James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., Eds., *An Introduction to Statistical Learning*, Springer, Cham, 201-228. https://doi.org/10.1007/978-3-031-38747-0_5

[23] Van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. and Yau, C. (2021) Bayesian Statistics and Modelling. *Natural Reviews Methods Primers*, **1**, Article No. 1. https://doi.org/10.1038/s43586-020-00001-2

[24] Blaker, H. (2000) Confidence Curves and Improved Exact Confidence Intervals for Discrete Distributions. *Canadian Journal of Statistics*, **28**, 783-798. https://doi.org/10.2307/3315916

[25] Zelikman, E., Wu, Y., Mu, J. and Goodman, N. (2022) Star: Bootstrapping Reasoning with Reasoning. *Advances in Neural Information Processing Systems*, **35**,

15476-15488.

[26] Rousselet, G.A., Pernet, C.R. and Wilcox, R.R. (2021) The Percentile Bootstrap: A Primer with Step-by-Step Instructions in R. *Advances in Methods and Practices in Psychological Science*, **4**. https://doi.org/10.1177/2515245920911881

[27] Efron, B. (1987) Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, **82**, 171-185. https://doi.org/10.1080/01621459.1987.10478410

[28] Canty, A. (2022) Package 'boot'. https://cran.r-project.org/web/packages/boot/boot.pdf

[29] Freedman, D. and Diaconis, P. (1981) On the Histogram as a Density Estimator: L_2 Theory. *Probability Theory and Related Fields*, **57**, 453-476. https://doi.org/10.1007/BF01025868

[30] Hyndman, R.J. and Fan, Y. (1996) Sample Quantiles in Statistical Packages. *American Statistician*, **50**, 361-365. https://doi.org/10.1080/00031305.1996.10473566

[31] Guajardo, S.A. (2024) Assessing Organizational Diversity with the Index of Qualitative Variation. Cambridge Scholars Publishing, Cambridge.

[32] Feutrill, A. and Roughan, M. (2021) A Review of Shannon and Differential Entropy Rate Estimation. *Entropy*, **23**, Article 1046. https://doi.org/10.3390/e23081046

[33] Efron, B. and Narasimhan, B. (2020) The Automatic Construction of Bootstrap Confidence Intervals. *Journal of Computational and Graphical Statistics*, **29**, 608-619. https://doi.org/10.1080/10618600.2020.1714633

[34] Ramachandran, K.M. and Tsokos, C.P. (2020) Mathematical Statistics with Applications in R. Academic Press, San Diego.

[35] Coolidge, F.L. (2020) Statistics: A Gentle Introduction. 4th Edition, Sage Publications, Thousand Oaks. https://doi.org/10.4135/9781071939000

[36] Poncet, P. (2022) Package 'Modeest'. Mode Estimation. https://cran.r-project.org/web/packages/modeest/modeest.pdf

[37] Banić, N. and Elezović, N. (2021) TVOR: Finding Discrete Total Variation Outliers among Histograms. *IEEE Access*, **9**, 1807-1832. https://doi.org/10.1109/ACCESS.2020.3047342

[38] Nielsen, F. and Nock, R. (2017) MaxEnt Upper Bounds for the Differential Entropy of Univariate Continuous Distributions. *IEEE Signal Processing Letters*, **24**, 402-406. https://doi.org/10.1109/LSP.2017.2666792

[39] Moral de la Rubia, J. (2023) Shape Measures for the Distribution of a Qualitative Variable. *Open Journal of Statistics*, **13**, 619-634. https://doi.org/10.4236/ojs.2023.134030

[40] Grubbs, F.E. (1950) Sample Criteria for Testing Outlying Observations. *Annals of Mathematical Statistics*, **21**, 27-58. https://doi.org/10.1214/aoms/1177729885

[41] Dixon, W.J. (1951) Ratios Involving Extreme Values. *Annals of Mathematical Statistics*, **22**, 68-78. https://doi.org/10.1214/aoms/1177729693

[42] Rosner, B. (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, **25**, 165-172. https://doi.org/10.1080/00401706.1983.10487848

[43] Nowak-Brzezińska, A. and Łazarz, W. (2021) Qualitative Data Clustering to Detect Outliers. *Entropy*, **23**, Article 869. https://doi.org/10.3390/e23070869

[44] Kvålseth, T.O. (1991) Statistical Inference for the Odds Measure of Qualitative Variation. *Perceptual and Motor Skills*, **72**, 115-118.

https://doi.org/10.2466/pms.1991.72.1.115

[45] Fossett, M. (2017) New Methods for Measuring and Analyzing Segregation. Springer, Cham. https://doi.org/10.1007/978-3-319-41304-4

[46] Hutchens, R.M. (2004) One Measure of Segregation. *International Economic Review*, **45**, 555-578. https://doi.org/10.1111/j.1468-2354.2004.00136.x

[47] Chakravarty, S.R. and Silber, J. (2007) A Generalized Index of Employment Segregation. *Mathematical Social Sciences*, **53**, 185-195. https://doi.org/10.1016/j.mathsocsci.2006.11.003