The Libyan Journal of Science (An International Journal): Volume 19A, 2016

On the Measures of Dispersion for Qualitative Data

Bahlul O. Shalabi

Department of Statistics, Faculty of Science, Tripoli University – Tripoli – Libya
E-mail: b.shalabi@uot.edu.ly

Abstract

The index of qualitative variation (IQV) is defined as a measure of dispersion for qualitative (non-numerical) data. There are many indices of qualitative variation, but they are rarely mentioned in introductory statistics textbooks. Wilcox (1967) presented seven indices that can be used to measure qualitative variation. The main purpose of this paper is to mention the concept of dispersion for a qualitative data, submit seven formulas for its measurement, which is appropriate for an elementary course in statistics, and make comparison between the seven indices of qualitative variation (IQV) that are presented by Wilcox (1967). A Matlab computer program has been written to calculate such indices.

Keywords: Dispersion; Index of qualitative variation (IQV), Numerical data; Non-numerical data; Quantitative data; Qualitative data.

المستخلص

يعرف مؤشر الاختلاف النوعي (IQV) بأنه مقياس لتشتت البيانات النوعية (غير العددية). توجد هناك العديد من مؤشرات الاختلاف النوعي، ولكن نادرا ما يتم ذكرها في كتب الإحصاء التمهيدية. الغرض الرئيسي من هذه الورقة هو الإشارة إلى مفهوم تشتت البيانات النوعية، وتقديم العديد من الصيغ لقياسه، التي هي مناسبة لمقرر مبادئ الإحصاء، كذلك إجراء مقارنة عددية بين مؤشرات الاختلاف النوعي السبعة المقترحة من قبل Wilcox في سنة 1967 حيث تم كتابة برنامج كمبيوتر بلغة ماتلاب لحساب هذه المؤشرات.

Accepted for publication: 14/7/2016

Introduction

The measures of dispersion for numerical data (interval or ratio level data) such as: the range, the interquartile range, the semi-interquartile range, the absolute mode deviation, the absolute mean deviation, the absolute mean difference, the standard deviation, and the variance are the most commonly measures of variation presented and discussed in introductory statistics textbooks.

Most introductory statistics textbooks do not mention the concept of dispersion for qualitative (non-numerical) data (nominal or ordinal level data), such as political party, religion, marital status, ethnicity, gender and race. Many students and teachers think that there are no measures of dispersion for qualitative data.

The measure of dispersion for qualitative data is called an index of qualitative variation (IQV). There are many indices of qualitative variation, but they are rarely used. Wilcox (1967) presented seven indices that can be used to measure qualitative variation. The main objective of this paper is to present the seven indices of qualitative variation and make numerical comparison between them.

Notations

In order to present some of indices of qualitative variation for a given set of a grouped qualitative data, let us introduce the following notations which shall be used throughout this paper:

K : denotes the number of categories;
f_i : denotes the frequency of the ith category;
f_m : denotes the frequency of the modal category;
f_L : denotes the lowest frequency;
N : denotes the size of the qualitative data, where $N = \sum_{i=1}^{K} f_i$;
$p_i = f_i / N$: denotes the proportion of cases in the ith category.

Seven indices of qualitative variation (IQV)

Wilcox (1967, 1973) gives seven indices of qualitative variation each index of them has the following properties:

- The index is between 0 and 1.
- The index is 0 if and only if all the cases belong to one category.

- The index is 1 if and only if all the cases are distributed equally over all categories
- When a qualitative data has a frequency distribution closer to a uniform, the value of the index becomes closer to 1 and indicates that there is a larger variation, but when the differences in frequencies across categories is larger, then the value of the index becomes closer to 0 and indicates that there is a smaller variation.

The seven indices of the qualitative variation that Wilcox developed are:

IQV # 1: *MODVR*
The first index, *MODVR*, is an analog of the standard deviation. This index is an index of deviation from the mode. A computational formula for *MODVR* is

$$MODVR = 1 - \frac{K f_m - N}{N(k-1)}. \qquad (1)$$

The low values of this index (and the others which follows) will stand for low variation and high values for high variation.

IQV # 2: *RANVR*
The second index, *RANVR*, is based on the range. A computational formula for *RANVR* is

$$RANVR = 1 - \frac{f_m - f_L}{f_m}. \qquad (2)$$

IQV # 3: *AVDEV*
The third index, *AVDEV*, is an analog of the mean absolute deviation. A computational formula for *AVDEV* is

$$AVDEV = 1 - \frac{\sum_{i=1}^{K} \left| f_i - N/K \right|}{2N(K-1)/K}. \qquad (3)$$

IQV # 4: *MNDIF*

The forth index, *MNDIF*, is an analog of the absolute mean difference. A computational formula for *MNDIF* is

$$MNDIF = 1 - \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\left|f_i - f_j\right|}{N(K-1)}. \qquad (4)$$

IQV # 5: *VARNC*

The fifth index, *VARNC*, is an analog of the variance. A computational formula for *VARNC* is

$$VARNC = 1 - \frac{\sum_{i=1}^{K}\left(f_i - N/K\right)^2}{N^2(K-1)/K}. \qquad (5)$$

IQV # 6: *STDEV*

The sixth index, *STDEV*, is an analog of the standard deviation. A computational formula for *STDEV* is

$$STDEV = 1 - \sqrt{\frac{\sum_{i=1}^{K}\left(f_i - N/K\right)^2}{\left(N - N/K\right)^2 + \left(K-1\right)\left(N/K\right)^2}}. \qquad (6)$$

IQV # 7: *HREL*

The seventh and the last index of qualitative variation is *HREL*. A formula for such index is

$$HREL = \frac{-\sum_{i=1}^{K} p_i \log_2 p_i}{\log_2 K}.; \qquad (7)$$

where $p_i = f_i / N$ denotes the proportion of cases in the ith category, and $\log_2$ is the logarithm of the base 2.

Example 1:
Suppose we have three groups of data with 1000 cases on a qualitative variable with two possible outcomes – category A or category B.
Group 1: 750 cases in category A; 250 cases in category B
Group 2: 500 cases in category A; 500 cases in category B
Group 3: 50 cases in category A; 950 cases in category B

Fig. 1 below shows the bar chart for each group of these three different groups of data.

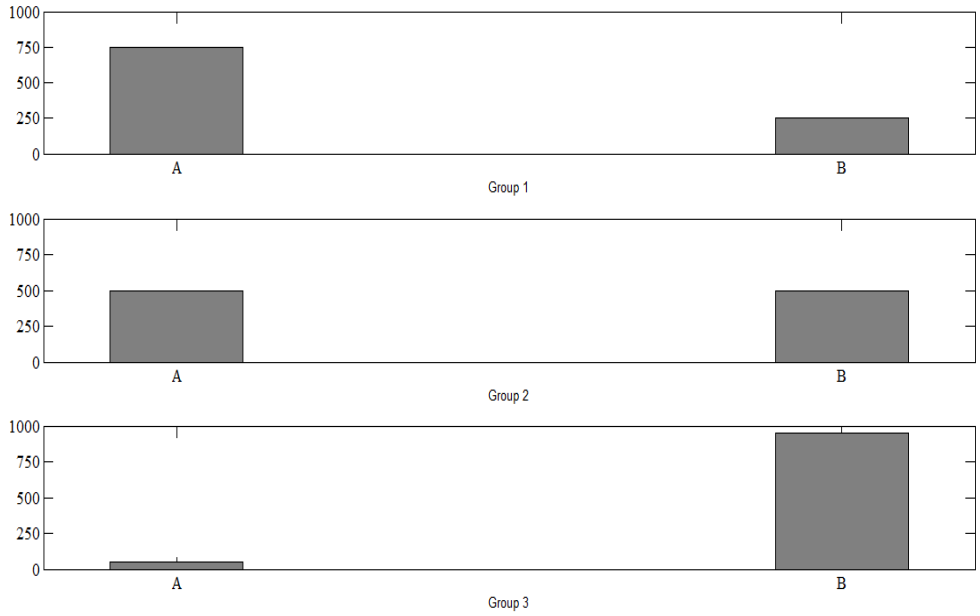

Figure 1. Bar chart representation for the three groups of qualitative data

Table 1., below and Fig. 2 show the calculated values of the seven indices of qualitative variation for each group of data using the Matlab function iqv (see the appendix).

Table 1. The values of the seven indices of qualitative variation for the three groups of data.

IQV	*MODVR*	*RANVR*	*AVDEV*	*MNDIF*	*VARNC*	*STDEV*	*HREL*
Group 1	**0.5**	**0.33333**	**0.5**	**0.5**	**0.75**	**0.5**	**0.81128**
Group 2	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Group 3	**0.1**	**0.052632**	**0.1**	**0.1**	**0.19**	**0.1**	**0.2864**

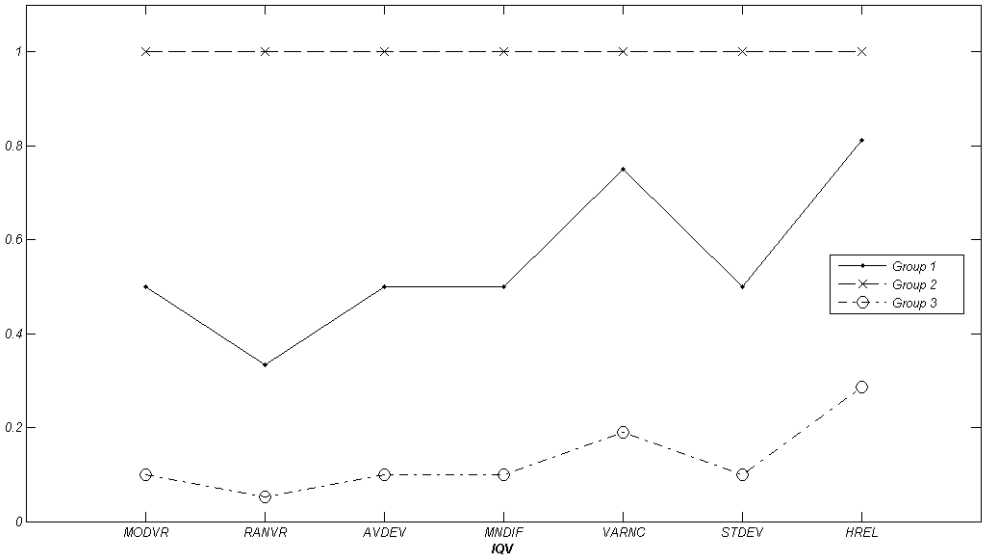

Figure 2. Values of the seven indices of qualitative variation for the three groups of data.

According to the entries in table 1, and Fig. 1 and Fig. 2 above, we note that the data in group 1 has less variation than the data in group 2 because 750 of the cases are the same in group 1, while only 500 of the cases in group 2 are the same. Consequently, there is more variation in group 2 than in group 1. Since 950 of the cases in group 3 are the same then, among the three groups, group 3 has the least variation. In this example we observe that the value of the seven indices of qualitative variation for groups 2 and 3 of data are nearly the same except the seventh index of qualitative variation, *HREL*, have value larger than the others and the second index of qualitative variation, *RANVR*, have value smaller than the others.

Example 2:
Table 2 below summarizes data on marital status for 3530 participants.

Table 2. Frequency Distribution Table for Marital Status

Marital Status	**Frequency**
Single	203
Married	2580
Widowed	334
Divorced	367
Separated	46
Total	**3530**

Fig. 3 below shows the bar chart of marital status.

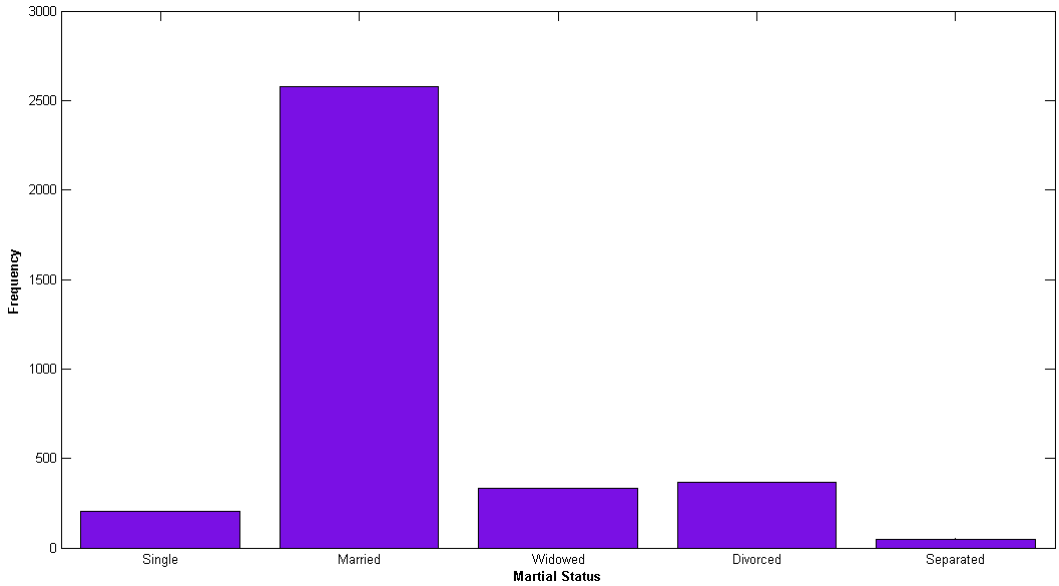

Figure 3. Bar chart of martial status.

Table 3. below and Fig. 4 show the calculated values of the seven indices of qualitative variation for frequency distribution table for marital status using the Matlab function iqv (see the appendix).

Table 3. The values of the seven indices of qualitative variation for Marital Status**.**

	MODVR	*RANVR*	*AVDEV*	*MNDIF*	*VARNC*	*STDEV*	*HREL*
IQV	0.3364	0.0178	0.3364	0.2589	0.5532	0.3316	0.5644

In this example we observe that the value of the seven indices of qualitative variation for Marital Status are nearly the same except the seventh index of qualitative variation, *HREL*, have value larger than the others and the second index of qualitative variation, *RANVR*, have value smaller than the others.

Although all the seven indices produced similar results one point should be highlighted: In the two examples given above, we have observed that all these indices give nearly similar results except the *HERL* and the *RANVR* as shown in Fig. 2 and Fig. 4.

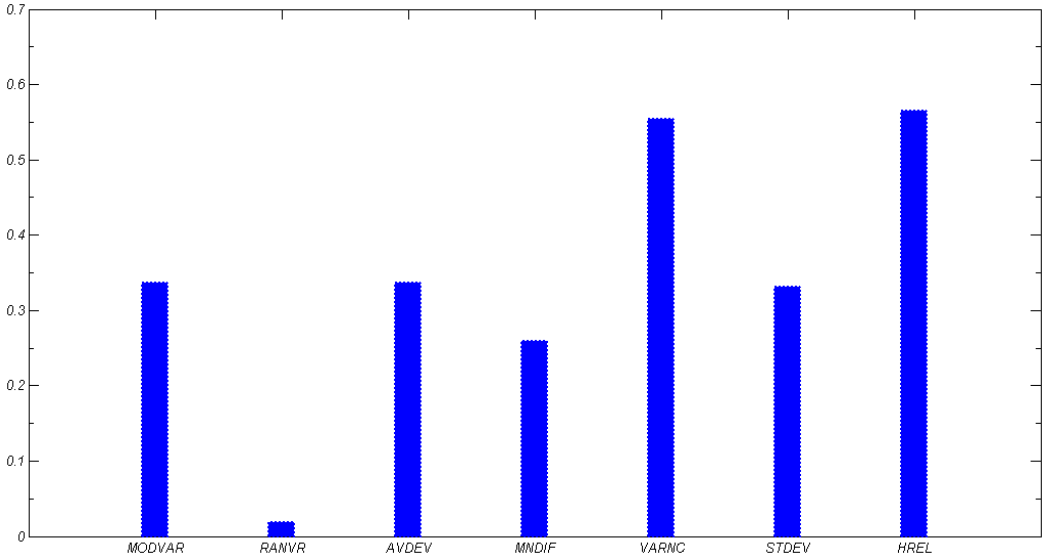

Figure 4. The values of the seven indices of qualitative variation for **Marital Status**..

2. Conclusion

In this paper I have presented seven indices of qualitative variation, these indices were developed by Wilcox A. R. (1967, 1973). The main purpose of this paper was to mention the concept of dispersion for a qualitative data, and present several formulas for its measurement, which is appropriate for an elementary course in statistics. In the two examples given in this paper, we have observed that all the seven indices give nearly similar results except the *HERL* and the *RANVR*. Further research work can be done on the mathematical aspects of these seven indices of qualitative variation.

References

Wilcox A. R. (1967). Indices of Qualitative Variation, Oak Ridge National Laboratory; ORNL-TM-1919

Wilcox A. R. (1973). Indices of qualitative variation and political measurement, Western Political Quarterly, **26** (2), 325–343.

Appendix

This Matlab function computes the seven indices of the qualitative variation due to Wilcox (1967, 1973).

```
function IQV1_7=iqv(x,f);
k = length(x); d = [];
for s = 1:k,       d = [d, x(s)*ones(1,f(s))]; end
 n = length(d); T=zeros(n, n); r = 0; c = 0;
for ii = d,
   r = r + 1; a = d(r);
   for jj = d
      c = c + 1; b = d(c);
          if  a ~= b,  T(r, c) = 1;  end
    end
   c = 0;
end
u = sum(T(:))/(n*n); fm = max(f); N = sum(f); fL = min(f);
% IQV#1:  IQV based on the mode, MODVR
MODVR = 1- (k*fm - N)/(N*(k - 1));
% IQV#2: IQV based on the range, RANVR
RANVR = 1- (fm - fL)/fm
 % IQV#3: IQV based on the average, AVDEV
AVDEV = 1- sum(abs(f - N/k))/((2*N/k)*(k - 1));
% IQV#4: IQV based on the mean difference, NDIF
a=[];
for i=1:k -1
             for j = i+1:k
                    a=[a; abs(f(i) - f(j))];
            end
end
D = sum(a); MNDIF=1- D/(N*(k- 1));
 % IQV#5: IQV based on the variance, VARNC
VARNC = 1-sum((f - N/k).^2)/(N^2*(k - 1)/k);
 % IQV#6: IQV based on the standard deviation, STDEV
          STDEV=1- sqrt(sum((f - N/k).^2)/((N - N/k)^2+(k -1)*(N/k)^2));
% IQV#7: HREL
P = f/N;  HREL=-1*sum(P.*log2(P))/log2(k)
IQV1_7 = [MODVR; RANVR; AVDEV; MNDIF; VARNC; STDEV; HREL];
```