Journal of Statistics Education

Variability for Categorical Variables

Gary D. Kader
Appalachian State University

Mike Perry
Appalachian State University

Journal of Statistics Education Volume 15, Number 2 (2007),
http://www.amstat.org/publications/jse/v15n2/kader.html

Keywords: Variability, Categorical Variable, Unalikeability

Abstract

Introductory statistics textbooks rarely discuss the concept of variability for a categorical variable and thus, in this case, do not provide a measure of variability. The impression is thus given that there is no measurement of variability for a categorical variable.

A measure of variability depends on the concept of variability. Research has shown that "unalikeability" is a more natural concept than "variation about the mean" for many students. A "coefficient of unalikeablity" can be used to measure this type of variability.

Variability in categorical data is different from variability in quantitative data. This paper develops the coefficient of unalikeability as a measure of categorical variability.

1. Introduction

Introductory statistics textbooks give considerable attention, as they should, to the distribution of quantitative variables and measures of their variability. Discussions of categorical variables, however, typically do not. The treatment of categorical data analysis usually moves immediately to the more interesting questions formulated in terms of contingency tables, with the focus of the analysis on variability among counts in the

table. There is usually no discussion of the concept of variability for a categorical variable, and thus no mention of a measure of variability that plays the role that standard deviation plays in the quantitative case. The impression is thus given that there is no concept of variability for a categorical variable, or, if there is one, there is no known way of measuring it. This impression is incorrect. There is a concept of variability for a categorical variable, and there are ways of measuring it. We suspect that a significant percentage of the teachers of introductory statistics are unaware of these ideas, and readily admit that we were not until we investigated the ideas presented in this paper.

1.1 Objectives

The purposes of this paper are several fold, including:

1. Describe a concept of variability for a categorical variable, and provide a method for its measurement. This is done at an elementary level which requires no probability or statistics background and thus is appropriate for an introductory course.

2. Show how these ideas evolved from research results on students' concepts of variability for quantitative variables.

3. Although our development is done independently of previous ideas, we point out that the underlying ideas have been around for at least ninety years. The early uses were for specialized applications or in statistically sophisticated settings and thus not presented in a fashion appropriate for a student's first exposure to variability.

1.2 Students' Concepts of Variability

Intuitive concepts of variation might differ among our students; we may be talking about one concept of variation in our classes while our students are thinking about another! In a study by Loosen, Lioen, and Lacante (1985), students were shown two sets of blocks, referred to here as set I and set II (see Figure 1). In the original study the blocks in set I were painted red and were 10, 20, 30, 40, 50, 60 cm high. The blocks of set II were painted yellow and were 10, 10, 10 and 60, 60, 60 cm high. Note that for quantitative data the height of each block in this physical representation indicates the magnitude of the corresponding value.

The students were instructed as follows: “These are two sets of blocks: a set of red blocks and a set of yellow ones. In which set do the blocks have the greater variation among themselves?” Fifty percent selected set I for the greater variation, 36% selected II and 14% said there was no difference.

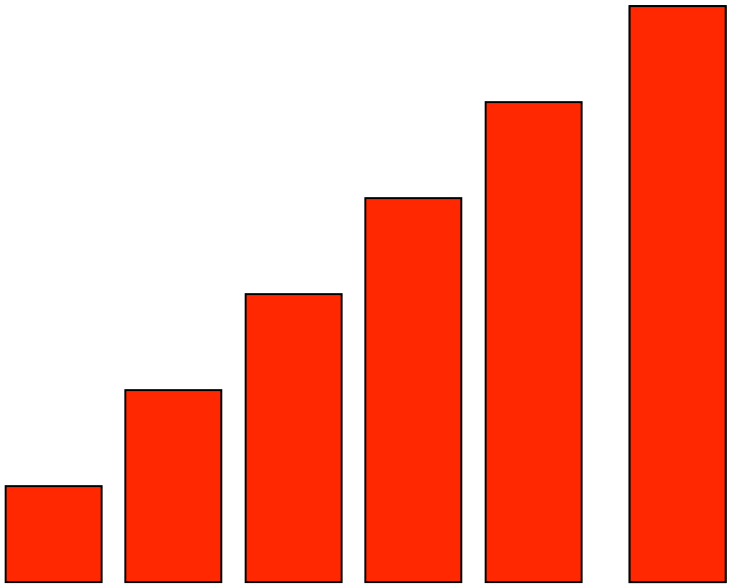
The 50% who selected set I are making their judgment on the observation that no two blocks have the same length. These students are basing their choice on an intuitive concept of variability - *unalikeability* – the lack of bars of the same size or the lack of clusters of bars of the same size. These learners do not think of variation as “how much the values differ from the mean.” Their perception has to do with “how often the

observations differ from one another." The authors point out that this can be an important part of a classroom lesson. The teacher can show the students that the standard deviation would indicate that set II has the greater variation because its standard deviation is larger than that of set I, and that the standard deviation is not measuring the concept of variation for students who selected set I.

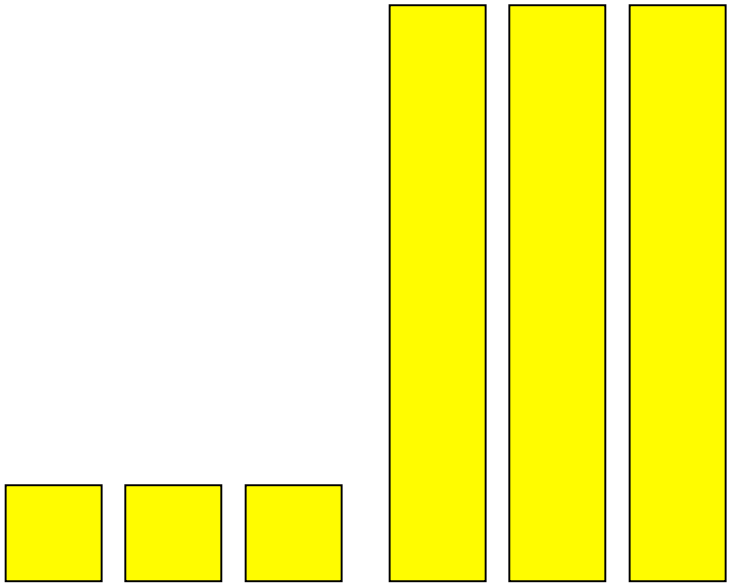
1.3 The Coefficient of Unalikeability

Unalikeability is defined to mean *how often observations differ from one another.*

The concept of *unalikeability* focuses on how often observations differ, not how much. The incidence of differences for the six blocks of set I and set II are indicated in Table 1. Each table gives all possible pairings of the sizes of the bars, and table entries are either 0 or 1 to indicate whether the block sizes are equal or different, respectively. Note that all pairs are indicated twice -- once in each half of the table. Comparisons of a block with itself are not of interest and are indicated with an asterisk.



Set I



Set II

Figure 1. Physical Representation for Two Sets of Numerical Data

Table 1. Incidence of Differences

Set I

	10	20	30	40	50	60
10	*	1	1	1	1	1
20	1	*	1	1	1	1
30	1	1	*	1	1	1
40	1	1	1	*	1	1
50	1	1	1	1	*	1
60	1	1	1	1	1	*

Set II

	10	10	10	60	60	60
10	*	0	0	1	1	1
10	0	*	0	1	1	1
10	0	0	*	1	1	1
60	1	1	1	*	0	0
60	1	1	1	0	*	0
60	1	1	1	0	0	*

If the 1's in a table are added up, we obtain the number of differences that occur when all possible comparisons are made, one observation with another. If we divide by 36-6=30, the number of comparisons, then we get the proportion of differences that occur.

For set I, where all of the data differ from one another, this proportion is 30/30 =1. For set II, the proportion is 18/30 = 0.60. Note that since all pairs appear twice, only half of the entries need to be counted. In the case of set II, there would be 15 comparisons, and the proportion would be 9/15 = 0.60. Also note that if all of the data are equal in value, this proportion is 0.

This provides a *coefficient of unalikeability* on a scale from 0 to 1. The higher the value, the more unalike the data are. If $x_1, x_2, \ldots, x_n$ are n observations on a quantitative variable, x, Perry and Kader (2005) give a general definition for the coefficient of unalikeability as:

$$u = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n}$$

where

$$c(x_i, x_j) = \begin{cases} 1, & x_i \neq x_j \\ 0, & x_i = x_j \end{cases}$$

This coefficient was suggested by the idea of a "within data" variance. Gordon (1986) reminds us that standard deviation and variance can be defined independently of the mean by taking the average of the squares of the differences between each pair of values:

$$W_1 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2}{n^2 - n}$$

The coefficient of unalikeability mimics this idea by replacing the squares of distances with the 0 - 1 indicator of differences. Gordon points out that

$$W_1 = 2S^2, \text{ where } S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

2. ANOTHER LOOK AT UNALIKEABILITY

We were recently examining some of the ideas underlying the coefficient of unalikeability and in doing so took a look at the coefficient from the perspective of categorical variables. Although the *length* of the bars in Figure 1 is a quantitative variable, the students who think of variability as unalikeability are forming categories. A category consists of all bars of the same length; once the categories are formed, the actual lengths are ignored.

Note that in the case of a categorical variable, x, each observation is classified into one of m distinct categories. In this case, the definition for quantity $c(x_i, x_j)$ becomes:

$$c(x_i, x_j) = \begin{cases} 1, & \text{if } x_i \text{ and } x_j \text{ are in different categories} \\ 0, & \text{if } x_i \text{ and } x_j \text{ are in the same category} \end{cases}$$

2.1 Visualizing Variability in Categorical Data

Variability in categorical data is somewhat different than variability in numerical data. Let's begin by examining three groups of data with ten responses on a variable with two possible outcomes – Category A or Category B.

Group 1: Seven responses in Category A; three responses in Category B
Group 2: Five responses in Category A; five responses in Category B
Group 3: One response in Category A; nine responses in Category B

Figure 2 provides a physical representation for these three different situations. Note that, unlike numerical data, the bar height in this representation for categorical data does not indicate the magnitude of a response; it indicates only whether the response was in Category A or Category B.

Which group of data has the most variability? the least variability? For categorical data, the notion "how far apart?" does not make sense; however, the notion of unalikeability does make sense. Within a particular group two responses differ if they are in different categories and are the same if they are in the same category. That is, the two responses are either unalike (different categories) or alike (same category). Consequently, variability in categorical data is equivalent to unalikeability in numerical data.

Comparing Groups 1 and 2, the data in Group 1 are more alike since seven of the values are the same (i.e., 7 are in Category A), while only five of the values in Group 2 are the same (i.e., 5 are in either Category A or B). Consequently, the data in Group 2 are more unalike. That is, there is more variability in Group 2 than in Group 1. Since nine of the values in Group 3 are the same (i.e., 9 are in Category B) then, among the three groups, Group 3 has the least variability.

2.2 Quantifying Variability with Two Categories

For reasons that will evolve in the following discussion, we propose an alternative definition for the coefficient of unalikeability as:

$$u_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c(x_i, x_j)}{n^2}, \text{ where } c(x_i, x_i) = 0, \; i = 1, 2, \ldots, n$$

When u is defined with the divisor $n^2 - n$, the coefficient has value 1 with all distinct measurements. Using n^2 as the divisor instead produces a value close to 1 for large n since:

$$\frac{1}{n^2 - n} = \frac{n^2}{n^2 - n}\left(\frac{1}{n^2}\right) = \frac{n}{n-1}\left(\frac{1}{n^2}\right)$$

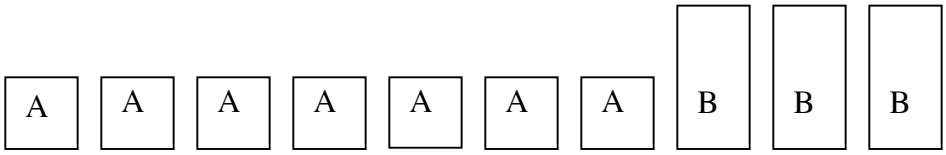
This second coefficient is analogous to the other "within data" variance proposed by Gordon (1986):

$$W_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} (x_i - x_j)^2}{n^2}$$

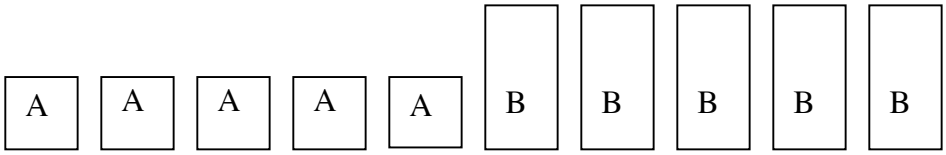
Gordon points out that

$$W_2 = 2V, \text{ where } V = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

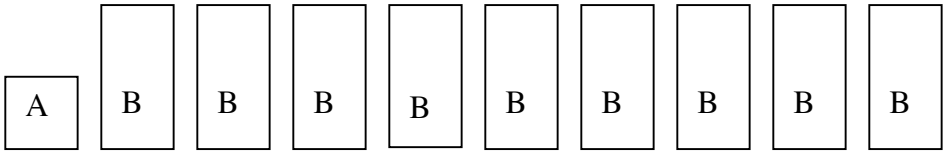
The incidence of differences for the ten responses of Groups 1, 2, and 3 are shown in Table 2. Each table gives all possible pairings of responses, and table entries are either 1 or 0 to indicate whether the responses are unalike or alike, respectively. The corresponding values for u_2 are indicated in Table 3.



Group 1



Group 2



Group 3

Figure 2. Physical Representations for Three Groups of Categorical Data

Table 2. Incidence of Differences for Three Groups of Categorical Data

Group 1

	A	A	A	A	A	A	A	B	B	B
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
A	0	0	0	0	0	0	0	1	1	1
B	1	1	1	1	1	1	1	0	0	0
B	1	1	1	1	1	1	1	0	0	0
B	1	1	1	1	1	1	1	0	0	0

Group 2

	A	A	A	A	A	B	B	B	B	B
A	0	0	0	0	0	1	1	1	1	1
A	0	0	0	0	0	1	1	1	1	1
A	0	0	0	0	0	1	1	1	1	1
A	0	0	0	0	0	1	1	1	1	1
A	0	0	0	0	0	1	1	1	1	1
B	1	1	1	1	1	0	0	0	0	0
B	1	1	1	1	1	0	0	0	0	0
B	1	1	1	1	1	0	0	0	0	0
B	1	1	1	1	1	0	0	0	0	0
B	1	1	1	1	1	0	0	0	0	0

Group 3

	A	B	B	B	B	B	B	B	B	B
A	0	1	1	1	1	1	1	1	1	1
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	0	0	0	0	0

Table 3. Value of u_2 for Three Groups of Categorical Data

Group	u_2
1	42/100 = .42
2	50/100 = .50
3	18/100 = .18

The values for u_2 indicate that the data in Group 3 are most alike and the data in Group 2 are most unalike. That is, Group 3 has the least variation and Group 2 has the most variation.

A second look at the table of incidences for Group 1 (Table 2) reveals that the 1's occur in the array in blocks.

The sum of the 1's can be determined by:

$$7 \cdot 3 + 3 \cdot 7 = 2 \cdot 7 \cdot 3$$

Thus $u_2 = 2 \cdot \frac{7 \cdot 3}{10^2} = 2 \cdot \frac{7}{10} \cdot \frac{3}{10}$

Note that here u_2 has the form:

$$2p_1p_2 \qquad (1)$$

where p_1, p_2 are the proportion of responses in categories A, B respectively.

The sum of the 1's can also be determined by:

$$7 \cdot (10-7) + 3 \cdot (10-3)$$

Thus $u_2 = \frac{7}{10} \cdot \frac{(10-7)}{10} + \frac{3}{10} \cdot \frac{(10-3)}{10} = \frac{7}{10} \cdot \left(1 - \frac{7}{10}\right) + \frac{3}{10} \cdot \left(1 - \frac{3}{10}\right)$

Note that here u_2 has the form $p_1(1-p_1) + p_2(1-p_2)$.

The sum of the 1's can also be determined by:

$$10 \cdot 10 - (7 \cdot 7 - 3 \cdot 3)$$

Thus $u_2 = \frac{10 \cdot 10 - (7 \cdot 7 + 3 \cdot 3)}{10^2} = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$

Note that here u_2 has the form $1 - {p_1}^2 - {p_2}^2$.

In each case we get .42, the proportion of possible pairings which are unalike. Note that the three formulas for finding u_2 work for the Groups 2 and 3 as well.

2.3 Connections to a Bernoulli Variable

The widely used Bernoulli variable codes responses for a two-outcome categorical variable as 1 (Category A) or 0 (Category B). With

p_1 = the proportion of 1's or the proportion of responses in Category A, and
$p_2 = 1 - p_1$ = the proportion of 0's or the proportion of responses in Category B.

It is well known that the mean of a Bernoulli variable is p_1 and the variance, V, is $p_1 p_2$. So, like the second form of Gordon's within variance, W_2, the coefficient of unalikeability as described in Equation (1) can be expressed as:

$$u_2 = 2V$$

2.4 Quantifying Variability with Three Categories

Consider the following data on ten responses for a variable with three possible outcomes – Category A, Category B or Category C:

Group 4: Two responses in Category A; three responses in Category B; and five responses in Category C

The table of incidences for Group 4 (Table 4) reveals that the 1's again occur in the array in blocks.

Table 4. Incidence of Differences for Three Outcome Categorical Variable

Group 4

	A	A	B	B	B	C	C	C	C	C
A	0	0	1	1	1	1	1	1	1	1
A	0	0	1	1	1	1	1	1	1	1
B	1	1	0	0	0	1	1	1	1	1
B	1	1	0	0	0	1	1	1	1	1
B	1	1	0	0	0	1	1	1	1	1
C	1	1	1	1	1	0	0	0	0	0
C	1	1	1	1	1	0	0	0	0	0
C	1	1	1	1	1	0	0	0	0	0
C	1	1	1	1	1	0	0	0	0	0
C	1	1	1	1	1	0	0	0	0	0

The sum of the 1's in Table 4 can be determined by:

$$(2 \cdot 3 + 2 \cdot 5) + (3 \cdot 2 + 3 \cdot 5) + (5 \cdot 2 + 5 \cdot 3) = 2(2 \cdot 3 + 2 \cdot 5 + 3 \cdot 5)$$

Thus

$$u_2 = 2 \cdot \left(\frac{2}{10} \cdot \frac{3}{10} + \frac{2}{10} \cdot \frac{5}{10} + \frac{3}{10} \cdot \frac{5}{10} \right)$$

Note that here u_2 has the form $2(p_1p_2 + p_1p_3 + p_2p_3)$, where p_1, p_2, p_3 are the proportion of responses in categories A, B, and C, respectively.

The sum of the 1's can also be determined by:

$$2 \cdot (10-2) + 3 \cdot (10-3) + 5 \cdot (10-5)$$

Thus

$$u_2 = \frac{2}{10} \cdot \frac{(10-2)}{10} + \frac{3}{10} \cdot \frac{(10-3)}{10} + \frac{5}{10} \cdot \frac{(10-5)}{10} = \frac{2}{10} \cdot \left(1 - \frac{2}{10}\right) + \frac{3}{10} \cdot \left(1 - \frac{3}{10}\right) + \frac{5}{10} \cdot \left(1 - \frac{5}{10}\right)$$

Note that here u_2 has the form $p_1(1-p_1) + p_2(1-p_2) + p_3(1-p_3)$.

The sum of the 1's can also be determined by:

$$10 \cdot 10 - (2 \cdot 2 + 3 \cdot 3 + 5 \cdot 5)$$

Thus $$u_2 = \frac{10 \cdot 10 - (2 \cdot 2 + 3 \cdot 3 + 5 \cdot 5)}{10^2} = 1 - \left(\frac{2}{10}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{5}{10}\right)^2$$

Note that here u_2 has the form $1 - {p_1}^2 - {p_2}^2 - {p_3}^2$.

In each case we get .62, the proportion of possible pairings which are unalike.

2.5 The Coefficent of Unalikeability for a Categorical Variable

For the case of a finite number of observations (n), a finite number of categories (m) and a finite number of objects, k_i, within Category i, as previously illustrated the pattern of blocks will allow expression of the coefficient of unalikeablity as:

$$u_2 = 2\sum_{i<j} p_i p_j \text{ or,} \tag{2}$$

$$u_2 = \sum_i p_i(1-p_i) \text{ or,} \tag{3}$$

$$u_2 = 1 - \sum_i p_i^2 \tag{4}$$

where $p_i = \frac{k_i}{n}$.

The interpretation of u_2 is that *it represents the proportion of possible comparisons (pairings) which are unalike*. Note that u_2 includes comparisons of each response with itself.

3. Measuring Population Diversity

"Unalikeability" is essentially equivalent to the several concepts of "population diversity" that have been developed in a variety of disciplines such as sociology, economics, linguistics and ecology. In conjunction with these concepts measures of diversity have been developed, and the simplest of these are essentially equivalent to "the coefficient of unalikeability." Because they evolved from applications within disciplines, they appear in the literature under several different names and are described with different notation. These kinds of measures are rarely mentioned in introductory statistics textbooks of a general nature, although they are sometimes presented in discipline based textbooks such as introductory "statistics for the social sciences" or "statistics for the biological sciences."

3.1 Diversity Within a Population

Lieberson (1969) describes diversity from the perspective of sociology as "the position of a population along a continuum ranging from homogeneity to heterogeneity with respect to one or more qualitative variables." Through the following example, he illustrates a general method for describing the magnitude of diversity within social aggregates by pairing all possible units in the population and determining the proportion of pairs of responses that are in different categories.

> "Suppose an investigator wishes to measure the degree of religious diversity within a specified aggregate, e.g., a city. A very simple operational solution is to describe the city in terms of the probability that randomly paired members of the population will hold different religious affiliations."

Lieberson points out that his index is essentially identical to each of the following measures:

Gini's index of mutability (1912)
Simpson's measure of diversity (1949)
Bachi's index of linguistic homogeneity (1956)
Greenberg's Monolingual Non-Weighted Method for measuring linguistic diversity (1956)
The index of qualitative variation described by Mueller and Schuessler (1961)
Gibbs and Martin's measurement of industry diversification (1962)

In order to convey the operational interpretations of these measures, discussions of Simpson's and Greenberg's ideas are presented in the following section.

3.2 Operational Interpretations of Two Diversity Measures.

Pielou (1969) discusses Simpson's measure of diversity within the context of mathematical ecology as follows.

> "Suppose two individuals are drawn at random and without replacement from an S-species collection containing N individuals, of which N_j belong to the j^{th} species (j=1,2,...,S; $\sum_j N_j = N$). If the probability is great that both individuals will belong to the same species, we can say that the diversity of the collection is low. This probability is
>
> $$\sum_j \frac{N_j(N_j-1)}{N(N-1)}$$
>
> and so we may use
>
> $$D = 1 - \sum_j \frac{N_j(N_j-1)}{N(N-1)}$$
>
> as a measure of the collection's diversity."

Greenberg (1956) describes the Monolingual Nonweighted Methods for measuring linguistic diversity as follows.

> "If from a given area we choose two members of the population at random, the probability that these two individuals speak the same language can be considered a measure of its linguistic diversity. If everyone speaks the same language, the probability that two such individuals speak the same language is obviously 1, or certainty. If each individual speaks a different language, the probability is zero. Since we are measuring diversity rather than uniformity, this measure may be subtracted from 1, so that our index will vary from 0, indicating the least diversity, to 1, indicating the greatest.
>
> $$A = 1 - \sum_i (i^2)$$
>
> where i is the proportion of speakers for a particular language."

Note that in both discussions, the measure of diversity is described in terms of the likelihood of two responses being either the same or being different, and the measure of diversity is expressed in a form similar to Equation (4).

4. Textbooks

Agresti (1990) develops the idea of variability in categorical data though contingency tables in which data on two categorical variables (an explanatory variable and a response variable) are summarized in a contingency table. He presents the general idea of measuring variability for a single nominal response variable *Y* in the following way:

$$V(Y) = \sum \pi_{+j}(1-\pi_{+j}) = 1 - \sum \pi_{+j}^2$$

where π_{+j} is the probability a response is in Category *j*. Note that Agresti's first expression for *V(Y)* is equivalent to u_2 as described in Equation (3), and his second expression is equivalent to u_2 as described in Equation (4).

Agresti points out that this quantity "is the probability that two independent observations from the marginal distribution of *Y* fall in different categories." He also notes that in the case of *m* distinct categories for *Y*, *V(Y)* is maximized when $\pi_{+j} = 1/m$ for all *j* and the maximum value is *(m-1)/m*. It is minimized when all responses are in the same category, in which case is 0. Of course, Agresti's book is not an introductory textbook and his presentation of the notion of variability for a categorical variable is not at an elementary level. Also, the presentation of this idea seems to have been deleted from the latest edition of his book.

Although some may exist, we have not seen a general introductory level statistics text that includes a discussion on measuring variability in qualitative data. However, some introductory statistics books designed for the social sciences do include such a discussion. For example, the book Social Statistics for a Diverse Society (Leon-Guerrero and Frankfort-Nachmias, 2000), presents the *index of qualitative variation (IQV)* as a measure of variability for nominal variables. The IQV is described as a measure of variability for qualitative variables "based on the ratio of the total number of differences in the distribution to the maximum number of possible differences within the same distribution." This definition is equivalent to the coefficient of unalikeability. Their presentation does not develop the underlying ideas but is formula driven and moves immediately from the definition to how to calculate the IQV from a frequency table.

5. Summary

A measure of variability depends on the concept of variability. In the case of quantitative measurements, the standard deviation is measuring variation about the mean. Our students, however, may be thinking about other concepts such as unalikeability. We should emphasize these distinctions in our teaching.

Variability in categorical data is based on unalikeability (diversity), which is quite different from variability in quantitative data. Thus the coefficient of unalikeability is a

natural measure of variability that has a well-defined interpretation. The concept and its measurement are appropriate for an introductory statistics course.

The evolution of ideas is often ignored in the teaching of statistics. It is important, in our opinion, to show students how definitions and formulas evolve. The coefficient of unalikeability is a fairly straightforward illustration of how measures of statistical concepts can be invented. We have found this sort of development effective for other concepts. For example, developing the mean absolute deviation as a prelude to the standard deviation. The idea of a ratio based on counts as a correlation coefficient can be introduced before the full development of Pearson's correlation coefficient (Holmes 2001).

The distinction between "unalikeability" and "variation about the mean" is based on the difference between "how often" and "how much." Throughout statistical analysis we see this type of distinction, especially the difference between measures based on distance and those which are not based on distance. Most introductory presentations of statistics emphasize the differences between measures based on distance and measures based on order for quantitative data. We believe the development of statistical thinking should include a discussion on measuring variability in categorical data as well.

References

Agresti, Alan (1990). *Categorical Data Analysis*. John Wiley and Sons, Inc. 24-25.

Bachi, R. (1956). "A statistical analysis of the revival of Hebew in Israel." in Roberto Bachi (ed.), Scripta Hierosolymitana, Vol. III, Jerusalem: Magnus press. 179-247

Gibbs, J. P. and Martin, W. T. (1962). "Urbanization, technology and division of labor: International patterns." *American Sociological Review* 27: 667-677.

Gini, C. W. (1912). "Variability and Mutability, contribution to the study of statistical distributions and relations." *Studi Economico-Giuricici della R. Universita de Cagliari*.

Gordon, T. (1986). "Is the standard deviation tied to the mean?" *Teaching Statistics*, 8(2), 40-2. (Reprinted in Green, D.R. (ed.) (1994).

Greenberg, J. H. (1956). "The measurement of linguistic diversity." *Language* 32, 109-115.

Holmes, P. (2001). "Correlation: From Picture to Formula." *Teaching Statistics* 23(3), 67-70.

Leon-Guerrero, Anna and Frankfort-Nachmias, Chava (2000). *Social Statistics for a Diverse Society*. 2nd edition, Pine Forge Press: Thousand Oaks, California. 153-162.

Lieberson, S. (1969). “Measuring Population Diversity.” *American Sociological Review*, 34(6), 850-862.

Loosen, F., Lioen, M. and Lacante, M. (1985). “The standard deviation: some drawbacks to an intuitive approach.” *Teaching Statistics*, **7**(1), 2-5.

Mueller, J. H. and Schuessler, K. F. (1961). *Statistical Reasoning in Sociology*. Boston: Houghton Mifflin.

Perry, M. and Kader, G. (2005). “Variation as Unalikeability.” *Teaching Statistics*, **27** (2), 58-60.

Pielou, E. C. (1969). *An Introduction to Mathematical Ecology*. John Wiley and Sons, Inc. 223.

Simpson, E. H. (1949). "Measurement of diversity.″ *Nature*, 163, 688.

Gary D. Kader
Department of Mathematical Sciences
Appalachian State University
Boone, NC 28608
U.S.A.
gdk@math.appstate.edu

Mike Perry
Department of Mathematical Sciences
Appalachian State University
Boone, NC 28608
U.S.A.
perrylm@appstate.edu
