Ratio of Vector Lengths as an Indicator of Sample Representativeness

Hee-Choon Shin
National Center for Health Statistics, 3311 Toledo Rd., Hyattsville, MD 20782

Abstract

The main objective of sampling is to obtain a representative sample for an unbiased and efficient estimate within a budget constraint. The current paper is to develop a new measure of representativeness of a sample. A population characteristics or measures of N population elements could be interpreted as a vector on N-dimensional space. Similarly, a sample characteristics or measures of n sample elements could be interpreted as a vector on n-dimensional subspace, imbedded in N-dimensional space. The length of a population vector is defined as the square root of the sum of squares of N components. The length of a sample vector is the square root of the sum of squares of n components. A weighted length of a sample vector could be obtained by weighting the n components with sampling weights. We can measure the sample representativeness as a ratio of the weighted length of a sample vector to the length of the population vector.

1. Introduction

Desirability of representativeness of a sample is accepted by professional statisticians and lay observers, even though the meaning or definition of “representativeness” varies, as discussed by many noted statisticians (Kruskal & Mosteller, 1979a, 1979b, 1979c, 1980; Snedecor, 1939; Stephan & McCarthy, 1958). A representative sample is a sample which is representative of a population. The method of selection may be random or purposive (Cramer, 1946, p. 331; Neyman, 1934). As Kendall and Buckland (1960) defined, we use the term “representative” to describe samples which “turn out to be so, however it chosen.” To the current author, no measure of sample representativeness has yet been proposed. In this paper, we propose an index of representativeness of a sample. Our measure is not equivalent to Yates’ balancing, which requires the average size of the selected units to be equal to the average size of the units of the population (Yates, 1971, Pp. 39-41).

2. Data: Two Perspectives

Consider an $N \times P$ rectangular data matrix with P variables on N units. Usually N is much larger than P, *i.e.*, $N \gg P$. There are two ways to present the data: One is the row picture, and the other is the column picture. The row picture would show N points on the P-dimensional space, familiar to us from when we began to study algebra using Cartesian diagrams. The column picture would present P vectors on N dimensional space. As an example, consider the following simple 2×1 matrix **A** which shows two measures of a single variable/characteristic:

$$\mathbf{A} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The matrix **A** contains information on two measures/observations of a single variable. Fig. 1(a) shows 2 points on a number line, and Fig. 1(b) shows a vector on a 2-dimensional space. With a million observations, for example, there would be one million points on a number line in Fig. 1(a). In Fig. 1(b), there would be still a single vector but on a million-dimensional space.

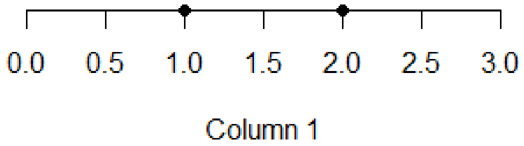


Fig. 1(a). Row Picture

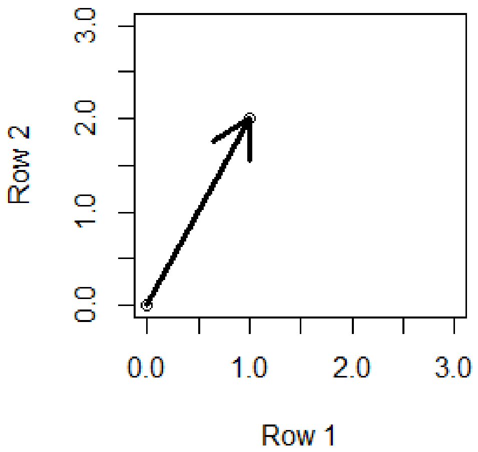


Fig. 1(b). Column Picture

Let **Y** be a $N \times 1$ population vector of Y_i $(i = 1, 2 \ldots N)$,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

Again, the vector **Y** could be interpreted as N points on *one*-dimensional space, or a vector on N-dimensional space.

Let **S** be a $N \times 1$ matrix containing sample indicators,

$$\mathbf{S} = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{bmatrix},$$

where $S_i = 1$ if $Y_i \in \mathbf{y}$; and $S_i = 0$ if $Y_i \notin \mathbf{y}$. The **y** stands for $n \times 1$ matrix containing a sample set of size n. Let $\mathbf{D} = diag(\mathbf{S})$. Then,

$$\mathbf{Y}_s = \mathbf{DY}.$$

The *i*th element of $\mathbf{Y}_s$ is either Y_i or 0 depending on the specific outcome of sampling algorithm for $i = 1, 2 \dots N$. The **y** would be obtained by omitting "zero" entries[1] from $\mathbf{Y}_s$,

$$\mathbf{Y}_s \Rightarrow \mathbf{y} = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \\ \vdots \\ Y_{(n)} \end{bmatrix}.$$

Note that we use parenthesis for subscripts to refer to sample elements as opposed to population elements. Sampling is to determine a particular **S,** based on specified methods.

3. Vector Length in Sample Subspace

Now, **y** is in a n-dimensional subspace imbedded in the *N*-dimensional space. The length of **Y** , ||**Y**||, is defined to be

$$||\mathbf{Y}|| = \sqrt{Y_1^2 + Y_2^2 + \cdots + Y_N^2} \tag{1}$$

Similarly, the length of **y** , ||**y**||, is defined to be

$$||\mathbf{y}|| = \sqrt{Y_{(1)}^2 + Y_{(2)}^2 + \cdots + Y_{(n)}^2}. \tag{2}$$

The length of a vector in the sample subspace is less than, or equal to the length of the vector in the population space, i.e.,

$$||\mathrm{y}|| \leq ||\mathrm{Y}||. \tag{3}$$

The equality occurs when all the population elements are included in the sample, i.e., $n = N$.

4. Weighted Length

Let $||\mathrm{y}||_w$ be an weighted length of a vector in sample subspace, i.e.,

$$||\mathrm{y}||_w = \sqrt{w_{(1)}Y_{(1)}^2 + w_{(2)}Y_{(2)}^2 \dots w_{(n)}Y_{(n)}^2}, \tag{4}$$

where $w_{(i)}$ is an inverse of inclusion probability, $p_{(i)}$ of the (i)-th unit. The weighted length, $||\mathrm{y}||_w$ of a sample vector could be larger than the length of a population vector, ||Y||, and therefore, the inequality (||y|| ≤ ||Y||) is no longer true. When the sample is selected with equal probability, i.e., $p_{(i)} = n/N$ for all (i),

[1] We omit the "zero" entries due to "0" sample indicator. We should not omit the entries when the value of Y_i is 0. For practical purposes, we could assume that all Y_i are positive.

$$||\mathrm{y}||_w = \sqrt{\frac{N}{n}}||\mathrm{y}||. \tag{5}$$

5. An Index of Sample Representativeness

There would be $N!/\{n!\,(N-n)!\}$ ways to select a size n sample from a population of N elements. Now, consider a particular sample of size n and a ratio (λ) of its weighted length to the length of population values,

$$\lambda = \frac{||\mathbf{y}||_w}{||\mathbf{Y}||}. \tag{6}$$

When population vector $\mathbf{Y}$ is a normalized vector of length 1, λ is simply a weighted length of sample vector $\mathbf{y}$. The λ would be 0 when every sample $Y_{(i)}$ is 0, i.e., $0 \leq \lambda < \infty$. The value of λ of a representative sample would be near 1.

6. Example 1: An Illustrative Example

Let us consider selecting 2 elements without replacement from population set, $\mathbf{Y}^T = [1,2,3,4]$, where $\mathbf{Y}^T$ stands for the transpose of $\mathbf{Y}$. The population mean is 2.5000 and its length is 5.4772. Table 1 shows weighted length and λ of each sample. The λ of Sample C is 1.0646 which is closest to 1. We argue that Sample C is more representative even though sample means of Samples C and D are the same and equal to the population mean.

Table 1. Weighted Lengths and λ's

Sample Identifier	Sample Elements	Sample Mean of $\mathbf{y}$	Weighted Length ($\|\|\mathbf{y}\|\|_w$)	λ
A	(1,2)	1.5000	3.1623	0.5774
B	(1,3)	2.0000	4.4721	0.8165
C	(1,4)	2.5000	5.8310	1.0646
D	(2,3)	2.5000	5.0990	0.9309
E	(2,4)	3.0000	6.3246	1.1547
F	(3,4)	3.5000	7.0711	1.2910

7. Utilizing Auxiliary Variable

Up to this point, we discussed sample representativeness with the variable of interest, **Y**. If we had complete information on **Y,** in fact, sampling is unnecessary. Without any prior information on **Y**, we have no choice but to randomly select a sample to make inference about the variable of interest. In practice, however, we have some auxiliary variables at our disposal such as geographic identifier or previous Census counts. For a list frame for human population, there might be a whole array of auxiliary variables including demographics.

Now, let **X** be the population vector of an auxiliary variable and **x** be a sample vector. Now, λ can be approximated with

$$\lambda \approx \frac{||\mathbf{x}||_w}{||\mathbf{X}||}. \tag{7}$$

The λ is simply a weighted length of sample vector $\mathbf{x}$ when population vector $\mathbf{X}$ is a normalized vector of length 1.

8. Simulation with Fisher's *Iris* data

Fisher (1936) observed four variables/characteristics (sepal length, sepal width, petal length, and petal width) of 50 plants in each type of three irises (*iris setosa*, *iris versicolor*, and *iris virginica*). Overall, there were 150 units and there were no missing values. For our simulation, the 150 units serve as the universe or population. For our simulation, petal width is the variable of interest, $\mathbf{Y}$, and the other three are auxiliary variables, $\mathbf{X}_p{}'$s. Table 2 shows descriptive statistics and lengths of the four variables. Also shown in Table 2 are correlation coefficients and cosines[2] of the three vectors with petal width. Again, in actual application, the correlation coefficients and cosines needs to be estimated since $\mathbf{Y}$ is not available before data collection. The large correlation coefficient (.9629) and *cosine* (.9836) indicates a close relationship between petal width and petal length.

Table 2. Vector Lengths in cm and Cosines with Petal Width: Fisher's *Iris* Data (1936)

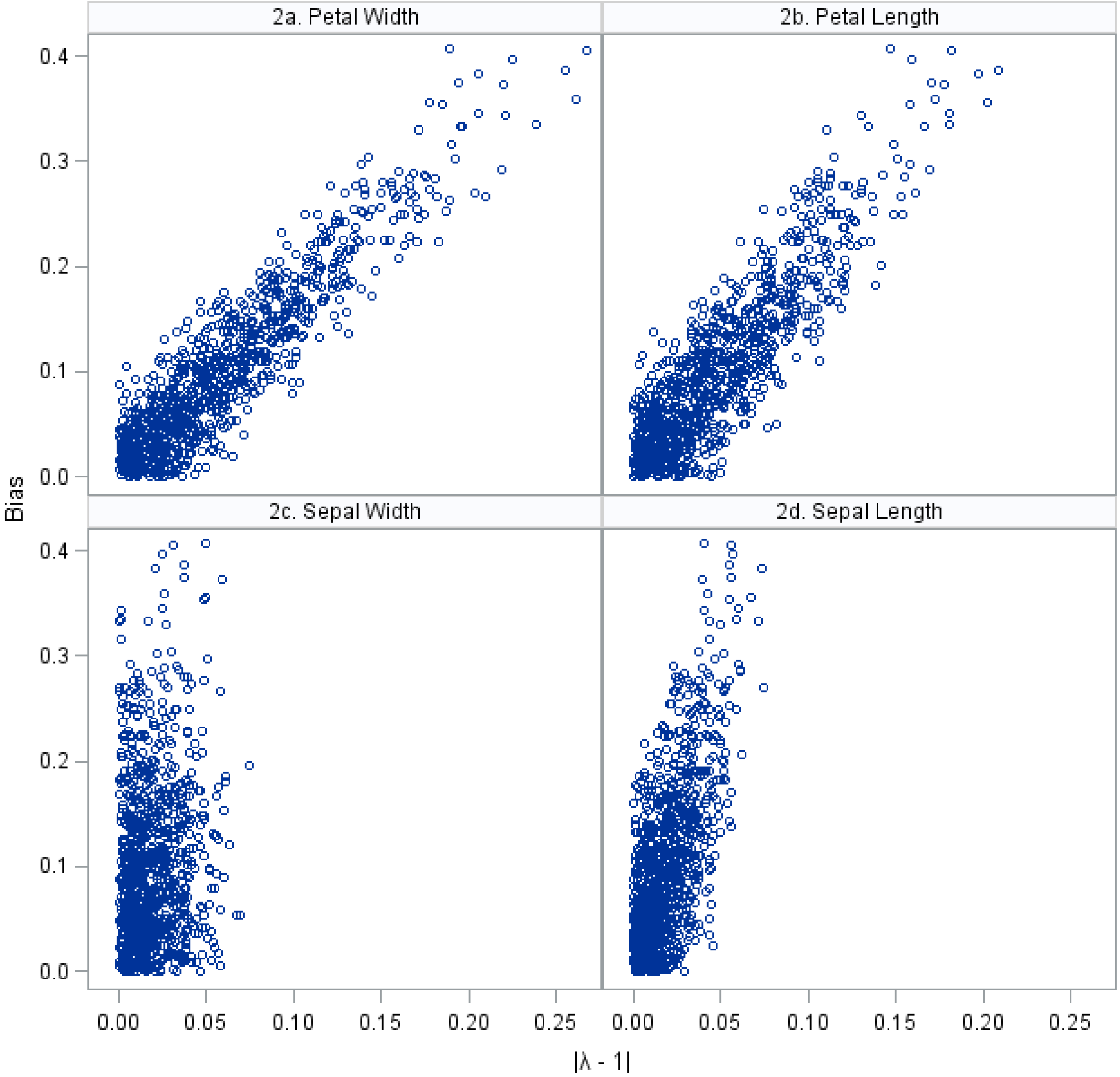
Measures	Variables/Characteristics			
	Sepal Length	Sepal Width	Petal Length	Petal Width
N	150	150	150	150
Mean	5.8433	3.0573	3.7580	1.1993
Standard Deviation	0.8281	0.4359	1.7653	0.7622
Vector Length	72.2762	37.8206	50.8204	17.3876
Correlation Coefficient (r) with Petal Width	0.8179	-0.3661	0.9629	1.0000
Cosine with Petal Width	0.8977	0.8088	0.9836	1.0000

Let us consider selecting 30 units from the 150 units without replacement to estimate the mean of petal width. There would be $150!/(30!\,120!) \approx 3.2199 \times 10^{31}$ ways to select a sample of size 30 from 150 units. At each draw, we calculated index of sample representativeness using Equation (7), and evaluated the size of bias by comparing sample mean of petal width to the population mean (1.1993) of petal width. We generated 1,000 samples for our simulation. As discussed, $\lambda = 1$ indicates that the sample exactly represents the population. We presented the indices using $|\lambda - 1|$. Each panel of Figure 2 shows relationships between the absolute bias and the index based on particular auxiliary variable. Fig. 2a is based on petal width, $\mathbf{Y}$. In actual sampling, $\mathbf{Y}$ would not be available. The vertical axis was so scaled that all the points may be spread around the 45 degree line when $\mathbf{Y}$ was used for simulation. As we see, all

[2] $\cos\theta = \frac{\mathbf{Y}^T\boldsymbol{X}}{||\mathbf{Y}||||X||}$, where θ is the angle between $\mathbf{Y}$ and $\mathbf{X}$.

the points are spread along the 45 degree line, and in general shows a positive relationship between the size of $|\lambda - 1|$ and bias. Fig. 2b is based on petal length, which is highly correlated with petal width ($r = .9629; cosine = .9836$). All the points in Fig. 2b are spread a bit off from the 45 degree line but in general shows a positive relationship between the size of $|\lambda - 1|$ and bias, and indicates that petal length is a good auxiliary variable. Fig. 2c indicates that the index of sample representativeness using sepal width is less effective in discriminating a representative sample from a non-representative sample, and sepal width is not a good auxiliary variable.

Figure 2. Index of Representativess and Bias by Auxiliary Variable Used


2a. Petal Width

2b. Petal Length

2c. Sepal Width

2d. Sepal Length

9. Concluding Remarks

We proposed an index of sample representativeness using vector length and showed that our index was effective in choosing a representative sample with a simulation using Fisher's *Iris* data. Our index could be an approximate quality measure of a sample. We also found that auxiliary variable needed to be moderately correlated with the variable of interest to be useful. Choosing a good auxiliary variable is difficult since the correlation between the auxiliary variable and the item of interest is unavailable since the item of interest is unknown before data collection. Developing an index using multiple auxiliary variables would be desirable. What would be the effect of using multiple auxiliary variables on the quality of the index? Also we should note that most of large-scale surveys are multipurpose (Kish, 1988). That is, there are many items of interest (**Y**'s) in a single survey. The best auxiliary variable for a particular item may not be the best for the other items. Further research is needed to develop an index for multipurpose surveys.

10. Disclaimer and Acknowledgements

The findings and conclusions stated in the manuscripts are solely those of the author. They do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention. We thank Alan Dorfman for helpful comments and suggestions.

References

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics, 7*, 179-188.

Kendall, M. G., & Buckland, W. R. (1960). *A Dictionary of Statistical Terms* (2nd ed.). New York: Hafner.

Kish, L. (1988). Multipurpose Sample Designs. *Survey Methodology, 14*(1), 19-32.

Kruskal, W., & Mosteller, F. (1979a). Representative Sampling, I: Non-Scientific Literature. *International Statistical Review, 47*(1), 13-24.

Kruskal, W., & Mosteller, F. (1979b). Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review, 47*(2), 111-127.

Kruskal, W., & Mosteller, F. (1979c). Representative Sampling, III: The Current Statistical Literature. *International Statistical Review, 47*(3), 245-265.

Kruskal, W., & Mosteller, F. (1980). Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939. *International Statistical Review, 48*(2), 169-195.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society, 97*(4), 558-625.

Snedecor, G. W. (1939). Design of Sampling Experiments in the Social Sciences. *Journal of Farm Economics, 21*(4), 846-855.

Stephan, F. F., & McCarthy, P. J. (1958). *Sampling Opinions: An Analysis of Survey Procedure.* New York: Wiley.

Yates, F. (1971). *Sampling Methods for Censuses and Surveys* (3rd ed.). London: Griffin.