

Normalizing Data

Hervé Abdi

1 Overview

We often want to compare scores or sets of scores obtained on different scales. For example, how do we compare a score of 85 in a cooking contest with a score of 100 on an I.Q. test? In order to do so, we need to “eliminate” the unit of measurement, this operation is called to *normalize* the data. There are two main types of normalization. The first type of normalization originates from linear algebra and treats the data as a *vector* in a multidimensional space. In this context, to normalize the data is to transform the data vector into a new vector whose *norm* (i.e., length) is equal to one. The second type of normalization originates from statistics, and eliminates the unit of measurement by transforming the data into new scores with a mean of 0 and a standard deviation of 1. These transformed scores are known as Z-scores.

Hervé Abdi
The University of Texas at Dallas
Address correspondence to:
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083-0688, USA
E-mail: herve@utdallas.edu <http://www.utd.edu/~herve>

2 Normalization to a norm of one

2.1 The norm of a vector

In linear algebra, the *norm* of a vector measures its *length* which is equal to the Euclidean distance of the endpoint of this vector to the origin of the vector space. This quantity is computed (from the Pythagorean theorem) as the square root of the sum of the squared elements of the vector. For example, consider the following data vector denoted \mathbf{y} :

$$\mathbf{y} = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix}. \quad (1)$$

The norm of vector \mathbf{y} is denoted $\|\mathbf{y}\|$ and is computed as

$$\|\mathbf{y}\| = \sqrt{35^2 + 36^2 + 46^2 + 68^2 + 70^2} = \sqrt{14,161} = 119. \quad (2)$$

2.2 Normalizing with the norm

In order to normalize \mathbf{y} , we divide each element by $\|\mathbf{y}\| = 119$. The normalized vector, denoted $\tilde{\mathbf{y}}$, is equal to

$$\tilde{\mathbf{y}} = \begin{bmatrix} \frac{35}{119} \\ \frac{36}{119} \\ \frac{46}{119} \\ \frac{68}{119} \\ \frac{70}{119} \end{bmatrix} = \begin{bmatrix} 0.2941 \\ 0.3025 \\ 0.3866 \\ 0.5714 \\ 0.5882 \end{bmatrix}. \quad (3)$$

The norm of vector $\tilde{\mathbf{y}}$ is now equal to one:

$$\|\tilde{\mathbf{y}}\| = \sqrt{0.2941^2 + 0.3025^2 + 0.3866^2 + 0.5714^2 + 0.5882^2} = \sqrt{1} = 1. \quad (4)$$

3 Normalization using centering and standard deviation: Z-scores

3.1 The standard deviation of a set of scores

Recall that the standard deviation of a set of scores expresses the dispersion of the scores around their mean. A set of N scores, each denoted Y_n , whose mean is equal to M , has a standard deviation denoted \hat{S} which is computed as

$$\hat{S} = \sqrt{\frac{\sum (Y_n - M)^2}{N - 1}} \quad (5)$$

For example, the scores from vector \mathbf{y} (see Equation 1) have a mean of 51 and a standard deviation of

$$\begin{aligned} \hat{S} &= \sqrt{\frac{(35 - 51)^2 + (36 - 51)^2 + (46 - 51)^2 + (68 - 51)^2 + (70 - 51)^2}{5 - 1}} \\ &= \frac{1}{2} \sqrt{(-16)^2 + (-15)^2 + (-5)^2 + 17^2 + 19^2} \\ &= 17. \end{aligned} \quad (6)$$

3.2 Z-scores: Normalizing with the standard deviation

In order to normalize a set of scores using the standard deviation, we divide each score by the standard deviation of this set of scores. In this context, we almost always subtract the mean of the scores from each score prior to dividing by the standard deviation. This normalization is known as Z-scores. Formally, a set of N scores each denoted Y_n and whose mean is equal to M and whose standard deviation is equal to \hat{S} is transformed in Z-scores as

$$Z_n = \frac{Y_n - M}{\hat{S}}. \quad (7)$$

With elementary algebraic manipulations, it can be shown that a set of Z-score has a mean equal of zero and a standard deviation of one. Therefore, Z-scores constitute an unit free measure which can be used to compare observations measured with different units.

3.3 An example

For example, the scores from vector \mathbf{y} (see Equation 1) have a mean of 51 and a standard deviation of 17. These scores can be transformed into the vector \mathbf{z} of Z – scores as:

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}. \quad (8)$$

The mean of vector \mathbf{z} is now equal to zero and its standard deviation is equal to one.

Related entries

Mean, normal distribution, standard deviation, standardization, standardized score, variance, Z-scores.

Further readings

Abdi, H., Edelman, B., Valentin, D., & Dowling, W.J. (2009). *Experimental Design and Analysis for Psychology*. Oxford: Oxford University Press.