Typical values and Measures of variability

Helgi Thorsson
Reykjavik University
Notes to Methodology, T-701-Rem4

September 9, 2007

1 Introduction

Some of the most elementary topics of Statistics deal with defining and computing typical values for a data set, as well as measures of variability among the data. These are used in descriptive statistics to summarize a data set as well as to describe parameters of statistical models.

2 Typical values

The *mean* (or *average*, depending on context) is what everybody gets if everything is put together and distributed evenly. Sometimes this has a concrete meaning, but often the mean should be looked upon as a normalised sum. The average class of (say) 25.2 pupils is nowhere to be found. The close relationship to the sum of all the values is evident from the definition of the average of the mumbers $x_1, x_2, \quad , x_i, \quad , x_n$:

$$x = \frac{x_1 + x_2 + \quad + x_i + \quad + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The *median* is the value in the middle when the dataset is ordered by size. The most common notation for the median is x. For an even number of values, the median is not well defined, any value in the middle interval is a median. A classic solution is to use the mean of the two adjacent values. If the arithmetic mean is in the middle interval, it may also be used as a median.

The mean and the median are related through the concept of *truncated mean.* A certain number or a certain proportion of the largest and smallest values are dropped from the data before the mean is computed. The 50% truncated mean is the median.

The *mode* or the most common value is often accessible and informative, for example together with a graphical representation of the data.

A *weighted average* is a weighted sum of the numbers where the weights are not neccessarily all equal, but nevertheless sum to one. The ususal definition of the mean may be written as a weighted sum where each value has the weight $\frac{1}{n}$:

$$x = \sum_{i=1}^{n} \frac{1}{n} x_i$$

One use of weighted average is for data where there are only few different values, but each one occurs often. The usual formula for the average is modified to sum over the distinct values of x, rather than all measurements, but the weight of each value is its proportion within the data set.

Another use of weighted means is to compute the mean of a *discrete probability distribution*. The weights then are the probabilites of obtaining each value andsummation is over all possible values of the distribution, finite or infintie. In this case, we talk about the *mean of the distribution* and represent it by , rather than x.

Another extension of the concept of the mean is functional means. Functional averages modify the influence of extreme values (the smallest or the largest or both) and may correspond to specific models ofthe data. With an invertible function, f, we compute

$$f^{\;1}\left(\sum_{i=1}^{n} \frac{1}{n} f\left(x_i\right)\right)$$

Common cases here are

The harmonic mean obtained by $f(x) = \frac{1}{x}$

the geometric mean obtained by $f(x) = log(x)$

Though not invertible in a unique way, the function $f(x) = x^2$ plays a special role in measuring variabilty among data. As does $f(x) = |x|$ to a lesser extent. In either case, the x would not be the original data, but their deviation from a chosen value such as the mean.

Each value computed from a sample is called a sample statistic.

3 Is there a most typical (or best) value?

Which is the correct or the best typical value for a set of data? All of the typical values have their merits, though their relevance may vary from one situation to another. The mean and the median of a symmetric dataset are the same value, but for asymmetric data, they can be quite different from one another. A good description of the data would make use of both values and a brief description of the form of the asymmetry.

In farily symmetric data, the mean is the best typical value in the sense that it varies less from one sample to another than the median. In very asymmetric data this may be the other way round, for example in data with occational single values large enough to make up more than half of the sum.

3.1 Excercise

Imagine we have a set of 9 individuals and the following readings of a variable, x, for each individual: $x = \{11, 12, 13, 14, 15, 16, 17, 18, 19\}$. On another set of 9 we have the readings $y = \{11, 12, 13, 14, 15, 16, 17, 5000, 10000\}$. Use R (or any other suitable software) to repeatedly sample 5 values (without replacement) from the vector, compute each sample average and give an overview of the various values. Study the sample medians in the same way.

4 Measures of variablility

The mean and the median each optimizes a measure of variability of the data.

For a given data set $\{x_i\}$ the function of a

$$s^2(a) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i \quad a)^2$$

is minimized by $a \;=\; x$ as seen by the usual method of differentiating the function. The minimum is called the *sample variance* of x. Its square root is the *standard deviation*.

$$var(x) = s^2(x) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i \quad x)^2$$

There is also a *population variance* where the sum is divided by $n \quad 1$ instead of n.

For values of a different from $\quad x$, the function $s^2(a)$ may be written in a form showing that the penalty for deviating from x is the square of the off-center:

$$s^2(a) \;=\; var(x) + (a \quad x)^2$$

This formula is derived by expanding and manipulating

$$s^2(a) \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i \quad a)^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}((x_i \quad x) \quad (x \quad a))^2$$

The formula has two uses:

1. By setting a = 0, we get the usual formula for simply computing the variance
2. The formula can be taken as the beginning of the very improtant theory of least squares, where it shows the relationship between the means of different groups.

The variance and related values of the *Sum of Squares* family are very important in classical Statistics, in particular in relation to the *Normal Distribution* where they are fundamental to computing probabilities of obtaining by chance at least such and such deviation from the mean.

Another measure of variability is the *Mean Absolute Deviation (MAD)* where absolute value is used instead of squaring the individual deviations:

$$MAD\,(a) \;=\; \frac{1}{n}\sum_{i=1}^{n}|x_i - a|$$

This function is minimized by $a = \tilde{x}$. To see that, rearrange the sum and join the terms for the largest x_i and the smallest, then the second largest and the second smallest and so on. We get

$$MAD\,(\tilde{x}) \;=\; \frac{1}{n}\sum_{i=1}^{[n/2]}(x_{n+1-i} - x_i)$$

The positive terms in this sum give the sum of all values above the median, the negative term give the sum of all values below the median, showing that the minimum is proportional to the difference between the average of the upper half and the average of the lower half of the numbers.

The first term is the *renge* of the x-values, the difference between the largest and the smallest.