Some Notes on Variation

AUTHOR:Lloyd S. Nelson

TITLE:Some Notes on Variation

SOURCE:Journal of Quality Technology 31 no4 459-62 O 1999

Key Words: Bias, Coefficient of Variation, Range, Standard Deviation, Variance.

WE shall consider some of the measures of variation, how they are calculated, and certain of their properties. Specifically, we shall discuss the following sample measures: variance, standard deviation, coefficient of variation, and range. We will assume that these statistics are based on random samples of independent observations from normal populations. Remember, if the sample is not drawn by using a bona fide procedure for random selection, information derived from it is questionable and hence untrustworthy.

VARIANCE

The extent to which individual observations deviate from the mean (used throughout to refer to the arithmetic mean) is a measure of the sample variation. What is needed is a procedure for combining these deviations to give a useful statistic. A first thought might be to take the mean of these deviations. But this won't work because the result will always be zero. The sum of the positive deviations always exactly balances the sum of the negative deviations.

Now, since the signs appear to be the cause of the trouble, let's get rid of them. Both of two ways of doing this have been formalized. The simplest is to drop the signs; that is, take the absolute values of the deviations. The mean of these is called the "mean deviation about the mean." The "mean deviation about the median" has also been tried. These statistics are generally not used because of the mathematical intractability of dealing with combinations of them.

Another way of getting rid of the signs is by squaring the deviations from the mean. The population variance (named by R. A. Fisher in 1918) is commonly estimated by dividing the sum of these squares by the associated degrees of freedom. For a simple random sample of n observations the divisor is n - 1. Using a divisor of n - 1 instead of n produces an unbiased estimate of the population variance. This is to say, as the sample size increases, the sample estimate of the variance approaches ever closer to the population variance. For an empirical demonstration of this fact, see Nelson (1974). If an average of a number of variances is required, the individual variances are weighted by their degrees of freedom.

There are several approaches to calculating a variance estimate. The first approach is

s[sup2] = [Graphic Character Omitted](X[subi] - X)[sup2]) / (n - 1) (1)

This is the definition formula. The second approach,

s[sup2] = ([Graphic Character Omitted]X[sup2[sub[subi] - ([Graphic Character Omitted] X[subi])[sup2]) / n) / (n - 1), (2)

is derived from Equation (1) and was devised to take advantage of a mechanical calculator that cumulated the values of the observations and their squares simultaneously. It is easy to see (once having seen it!) that Equation (2) is algebraically equivalent to Equation (1). Equation (2) can be dangerous to use because it can yield inaccurate results if the two terms in the numerator are of nearly equal size.

A third procedure deals only with the differences between all pairs of the data:

s[sup2] = ([Graphic Character Omitted][Graphic Character Omitted] (X[subi] - X[subj])[sup2]) / 2n(n - 1). (3)

It is interesting to note that the mean never comes into play in Equation (3). This formulation demonstrates that the variance (and, consequently, the standard deviation) is in no way "tied" to the mean despite what Equation (1) might lead you to believe.

A fourth procedure was mentioned in Nelson (1978) by which a variance estimate can be updated as new observations arrive one at a time. This could prove useful if there were restrictions on computer memory size.

A useful technique to preserve accuracy is to condition the data before carrying out the calculation. The data can be centered by substracting the mean value from each observation. If the mean is not known, the first observation could be used. It might also prove useful to scale the values by multiplying them by some factor. An extreme example would be the set: 1.0000002, 1.0000005, 1.0000003, 1.0000006, and 1.0000008. By subtracting the first value and then multiplying by 10[sup7], we convert the sample to 2, 5, 3, 6, and 8, for which round-off in the squaring process would not be a problem for calculators.

Of course, having coded the data prior to calculating the required statistics, it is necessary to uncode the results. (We do not "decode" the data because this is not an encryption.) Uncoding is accomplished by applying the inverse of the coding operations in reverse order. In the preceding example, the new mean is 4.8 and the new standard deviation is 2.3875. The mean of the original numbers is 4.8(10[sup-7]) + 1 = 1.00000048. The standard deviation of the original numbers is 2.3875(10[sup-7]) = 0.00000023875. It is unaffected by addition or subtraction in the coding.

Uncertainty in the estimate of the variance of a normal population is given by the following confidence expression:

vs[sup2] / X[sup2[sub[sub1-alpha/2] < sigma[sup2] < vs[sup2] / X[sup2[sub[subalpha/2], (4)

where X[sup2[sub[sub1-alpha/2] and X[sup2[sub[subalpha/2] are quantiles of a X[sup2] distribution, v equals degrees of freedom, and alpha is usually in the range of 0.10 to 0.01. Suppose, for example, that v = 14 and alpha = 0.05, then the left-hand divisor in Equation (4) is 26.12 and the right-hand divisor is 5.63. This gives a two-sided 95 percent confidence interval.

The dimension of the variance is the square of the dimension of the observations. If, for example, the observations are given in square inches, the dimension of the variance is inches raised to the fourth power. This puts it outside of our three-dimensional physical world, making it difficult to think about and impossible to visualize.

STANDARD DEVIATION

The standard deviation (named by Karl Pearson in 1893) is the positive square root of the variance. Contrary to expectation, it does not inherit the property of being unbiased. Do not try to "improve" your analyses by correcting for this bias. Tables of critical values for significance tests for statistics involving this standard deviation (e.g., Student's t) take into account this bias. How much bias is present can be judged from the value of the factor c4 (see any quality control book).

The standard deviation is a measure relating to how observations cluster around the mean. In a normal distribution, about 68 percent of the values are within +/-sigma, about 90 percent are within +/-1.28sigma, and so forth. Whereas the mean is a location parameter, the standard deviation is a scale parameter. The standard deviation has the same dimension as the observations. This makes it easy to contemplate and useful to combine with the mean (which has the same dimension) to yield confidence and tolerance intervals.

Uncertainty in the estimate of the standard deviation of a normal population is expressed by taking the square root of each term in Equation (4). A table given by Nelson (1997) simplifies this calculation for lower and upper one-sided confidence limits with confidences 0.90, 0.95, and 0.99 and for two-sided limits with confidences 0.80, 0.90, and 0.98.

An important question is: What size sample is required to estimate a standard deviation of a normal population to within some percentage of its true value? Unfortunately, even modest precision is obtained only with fairly large samples. For example, suppose it is required to estimate the standard deviation to within +/-10%; that is, s lies in the range 0.9sigma to 1.1sigma, with a confidence of 0.95. From a nomograph given in Nelson (1976), it can be seen that a sample of about 200 is needed.

The mean can be analogized to the center of gravity of a set of equally weighted disks whose distances along a bar are equivalent to the individual values. The mean is at the point where the bar balances. In a similar manner, the standard deviation is the distance along a weightless bar for which a single disk equal in weight to the total of the individual disks would give the same radius of gyration about the mean point. The term "standard error" refers to the standard deviation of a statistic. For example, s/[square root]n is the standard error of the distribution of the means of n observations.

COEFFICIENT OF VARIATION

The coefficient of variation (CV) is frequently expressed as a percentage of the mean; that is,

CV = s / X (5)

multiplied by 100. It is important that the observations be made using a scale that has a true zero (a so-called "interval" scale). For example, weight and length have true zeros; degrees Celsius does not. When comparing two or more coefficients of variation that do not have true zeros, they should have nearly the same zero points.

A common use of the coefficient of variation is to compare the relative variation of observations having quite different means--for instance, weights of elephants and mice. This statistic is a dimensionless number. Consequently, it makes possible the comparison of measurement variation for observations that have different dimensions. For example, suppose that a group of people have height and weight measurements as follows:

```
           Height      Weight
X     69 inches    145 pounds
s        2.6 inches   21 pounds.
```

Although we cannot compare inches and pounds, we can compare relative variability: for height, %CV = 100(2.6/69) = 3.8% and for weight, %CV = 100(21/245) = 14.5%. Other examples of coefficients of variation for people are oral temperature, 0.5%; pulse rate, 15%; and intelligent quotient, 18%. In my experience, I have found %CV = 5-10% for carefully controlled research and development work, 10-30% for factory operations, and over 30% to be not unusual for some biological studies.

An approximate significance test of the difference between two %CV's can be carried out with the proviso that the %CV's are less than about 10% and the two samples are at least 50. For economy of symbolism, let c[sub1] and c[sub2] stand for the two CV's. Then,

c[sub1] = s[sub1] / X[sub1], c[sub2] = s[sub2] / X[sub2],

sigma[subc[sub1]] = c[sub1] / [square root]2n[sub1], sigma[subc[sub2]] = c[sub2] / [square root]2n[sub2],

and

s[subc[sub1] - c[sub2]] = [square root](c[sup2[sub[sub1] / 2n[sub1]) + (c[sup2[sub[sub2] / 2n[sup2].) (6)

The difference between the coefficients of variation divided by its standard error (Equation (6)) is approximately normally distributed.

EXAMPLE

Suppose

```
n[sub1] = 50        n[sub2] = 60
X[sub1] = 27.66  X[sub2] = 14.81
s[sub1] = 2.601     s[sub2] = 1.022
c[sub1[sup = 0.094     c[sub2] = 0.069
```

Then,

sigma[subc[sub1] - c[sub2]] = [square root]((2.601/27.66)[sup2] / 2(50)) + ((1.022/14.81)[sup2] / 2(60)) = 0.0113

and

Z = (0.094 - 0.069)/0.0113 = 2.21,

where Z is the standardized difference between the coefficients of variation and is assumed to be approximately normally distributed. Reference to a table of the normal distribution shows a one-sided significance level of 0.014. The two-sided significance would be double this.

RANGE

The range (R) is the difference between the largest value (Max) and the smallest value (Min) in a sample:

R = Max - Min. (7)

Its use in control charting was promoted by Egon Pearson, who argued that it was easier to calculate and, for the small samples required, had adequate efficiency. Shewhart preferred the standard deviation because the efficiency of the range falls off rapidly as the sample size increases. Practice has favored Pearson's point of view. Nowadays, if the calculations are carried out by a computer, there is no reason not to use the standard deviation.

The range can be used in place of the standard deviation in any kind of statistical analysis provided that it is divided by a factor called d*[sub2] (which is different from the quality control factor d[sub2]). This operation produces a standard deviation estimate similar to what would be obtained from the positive square root of Equation (1). A table of d*[sub2] factors together with their accompanying degrees of freedom is given in Nelson (1975). Although it is obvious that the range can be spuriously increased by outliers, it should be noted that the standard deviation is also inflated by such values.

ADDED MATERIAL

REFERENCES

NELSON, L. S. (1974). "Showing the Properties of Statistics by Enumeration Experiments". Journal of Quality Technology 6, pp. 210-211.

NELSON, L. S. (1975). "Use of the Range to Estimate Variability". Journal of Quality Technology 7, pp. 46-48.

NELSON, L. S. (1976). "Nomograph of Sample Size for Estimating Standard Deviation". Journal of Quality Technology 8, pp. 179-180.

NELSON, L. S. (1978). "Combining Statistics from Two Groups and Some Updating Calculations". Journal of Quality Technology 10, pp. 180-181.

NELSON, L. S. (1997). "Factors for Confidence Limits on Standard Deviations". Journal of Quality Technology 29, pp. 485-487.

Source: Journal of Quality Technology, Oct99, Vol. 31 Issue 4, p459, 4p
Item: 2511701