

Degrees of freedom

The corollary of the requirement of functional independence in probability theory is the concept of degrees of freedom in mathematical statistics. The difference between μ , the unknown true population mean, and \bar{x} , the central value of a set of independently measured values, determined in samples selected from the population, lies at the divide between probability theory and sampling theory on one side, and mathematical statistics and sampling practice on the other.

Degrees of freedom are pervasive in mathematical statistics and sampling practice. In geostatistics, however, randomly distributed sets of independently measured values give degrees of freedom but ordered sets are not always similarly blessed. Koch & Link suggest, “The name *degrees of freedom* is somewhat troublesome because its usage is partly descriptive”, and compare it with the phase rule in chemistry and mineralogy. A simple rule to determine whether two or more values give degrees of freedom is to find out whether these values are calculated or measured. Calculated values are functionally dependent and deprived of degrees of freedom whereas measured values are not.

Randomized; equal weights

The divisors in the formulas for the variance of the randomly distributed or randomized set and the variance terms of temporally or *in situ* ordered sets are, in fact, the degrees of freedom for the set. The concept of degrees of freedom for a randomly distributed or randomized set evolved logically from the formula for the sample mean. For a set of n independently measured values, the deviation between each measured value in the set and its arithmetic mean is,

$$\Delta x_1 = x_1 - \bar{x}, \dots, \Delta x_i = x_i - \bar{x}, \dots, \Delta x_n = x_n - \bar{x}$$

so that the sum of n deviations is,

$$\Sigma[\Delta x_i] = [x_1 \dots + \dots x_i \dots + \dots x_n] - n \cdot \bar{x}$$

By definition, the arithmetic mean of a set of n measured values is,

$$\bar{x} = [x_1 \dots + \dots x_i \dots + \dots x_n] \div n$$

which implies that,

$$\Sigma[\Delta x_i] = [x_1 \dots + \dots x_i \dots + \dots x_n] - n \cdot \bar{x} = 0$$

If $n-1$ deviations are given, the missing one is determined simply because the sum of n deviations equals zero. By implication, a set of n measured values has $n-1$ independent deviations and one (**1**) dependent deviation. Hence, a randomly distributed or randomized set of n independently measured values with equal weights has **$n-1$** degrees of freedom.

Ordered sets; equal weights

The deviations between n measured values in a temporally of *in situ* ordered set are,

$$\Delta x_2 = x_1 - x_2, \Delta x_3 = x_2 - x_3, \dots, \Delta x_i = x_{i-1} - x_i, \dots, \Delta x_n = x_{n-1} - x_n$$

so that the sum of all deviations is,

$$\Sigma[\Delta x_i] = [x_1 - x_2] + [x_2 - x_3] \dots [x_{i-1} - x_i] + \dots + [x_{n-2} - x_{n-1}] + [x_{n-1} - x_n]$$

which implies that,

$$\Sigma[\Delta x_i] = x_1 - x_2 + x_2 - x_3 + \dots x_{i-1} - x_i + \dots x_{n-2} - x_{n-1} + x_{n-1} - x_n = x_1 - x_n$$

and that,

$$\Sigma[\Delta x_i] = x_1 - x_2 + x_2 - x_3 + \dots x_{i-1} - x_i + \dots x_{n-2} - x_{n-1} + x_{n-1} - x_n - x_1 + x_n = 0$$

Given that all of the measured values in the ordered set but the first and the last are used twice to compute the first variance term, it follows that an ordered of n measured values with equal weights has $df_o = 2n - 2 = 2 \cdot [n - 1]$ degrees of freedom.

The fact that the j^{th} variance term of a temporally or *in situ* ordered set with equal weights has $df_o = 2 \cdot [n - j]$ degrees of freedom can be proved by induction. Intuitively, it is logical because only the first variance term is based on the complete set whereas the second term does not include x_{n-1} , the third term does not include x_{n-2} and x_{n-1} , and so on.

Degrees of freedom are positive integers for sets of measured values with equal weights such as equidistant boreholes but become positive irrationals for sets of measured values with variable weights such as core samples of variable length and density.

References

Statistical analysis of geological data

Koch, G S Jr, and Link R F, Volume I, Wiley and Sons, 1970

Applied statistics for engineers

Volk, W, Krieger Publishing, 1980

Abuse of statistics

Merks, J W, CIM Bulletin, Vol 86, No 966, Jan 1993

Sampling in mineral processing

Merks, J W, SME Symposium, Vancouver, BC, Oct 20-24, 2002