A STATISTICAL PROBLEM FROM AGRICULTURAL PRACTICE [1])

C. LEVERT
Royal Neth. Meteorological Institute, de Bilt, Netherlands

SUMMARY

Somebody wants to start the cultivation of a certain crop in some country only provided the variability of the total rainfall amounts in the growth periods is sufficiently small.

He wants to know whether the given measurements of a few years give such an indication but also how many more years of measurements could give him a sufficiently small inaccuracy interval of the variability.

The author deals with the main points of the solution of this statistically and practically interesting problem, theoretically and with the help of a nomogram.

A numerical example illustrates the use of the nomogram.

1 INTRODUCTION

Suppose, somebody wants to start the cultivation of a certain crop in some country only provided the climate is "sufficiently reliable". Let be understood by this concept that, for instance, the rains will fall in such a manner that, although the total amounts of precipitation in the "growthperiods" will never be exactly equal year for year, their percentual variability is sufficiently small. If the investigator disposes of the climatic data of some years, he wants to know, firstly, whether these few data give an indication that the rainfall climate will be sufficiently unvariable in the given sense and, if the answer is in the affirmative, in which way it is possible to estimate the minimum number of years of measurements, which will be necessary to assure him that the defined reliability of the rainfall climate will possess a prescribed accuracy.

This problem turned out to be statistically very interesting. Since a solution seems to be also of practical importance, the author was induced to deal here with its main points.

2 STATISTICAL FORMULATION

First of all the problem should be reformulated in a more strictly statistical way. Given a random sample of n elements x_1, x_2, - - - x_n drawn from a normal universe with unknown mean value μ and unknown standard deviation σ (and consequently unknown variation coefficient $\gamma = \sigma : \mu$), the questions are firstly, how to evaluate the reliability interval for γ on the base of the sample values $\overline{x}$, s and c = $s/\overline{x}$ (estimates of μ, σ and γ) and, secondly, how to calculate the minimum number of measurements for which the upper limit of this reliability interval does not surpass a given value. Now, statistical theory (1) proves that, when random samples of fixed size n are drawn, the value c follows a well defined probability distribution such that
a)   c is approximately normally distributed,
b)   the mean value of c is γ,

c) for not too small values of n (say > 5) the standard devation σ_c of c is approximately

$$\sigma_c = \gamma \sqrt{\frac{1 + \dfrac{2\,n - 3}{n}\,\gamma^2}{2\,(n-1)}} \tag{1}$$

The larger the n, the better the approximations. Then there is a probability of 0.95 that, for given γ and given n, the value of c can be found within the following interval

$$\gamma - 2\,\sigma_c \leqslant c \leqslant \gamma + 2\,\sigma_c \tag{2}$$

If a probability of 0.90 or 0.99 is required, the factor 2 (precise: 1.96) in expression (2) must be replaced by 1.64 or 2.58 respectively.

3 Nomogram

Practical applirations of this theory are facilitated by the construction of a nomogram in cartesian coordinates with c as ordinate and γ as abscis, both in the same linear scale. For each of a large set of value of n (the larger the set, the better), e.g. 6, 10, 15, 20, 30, 40, 50, 75, 100, both curves c =
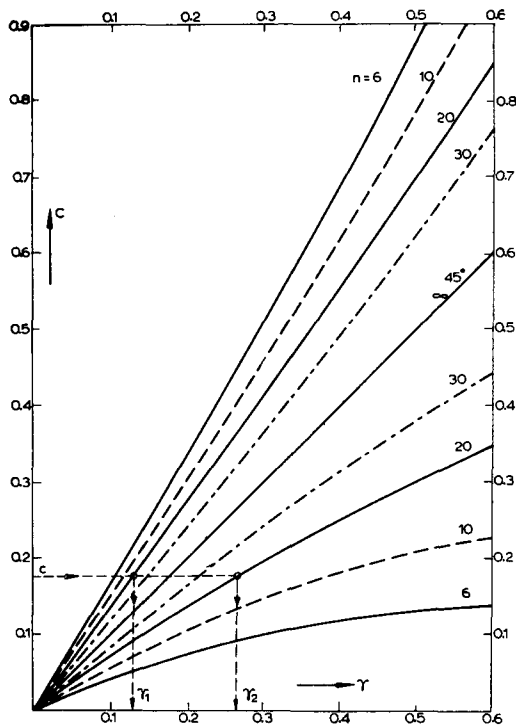

Fig. 1   Let be given a normal population, with mean value μ and standard deviation σ; γ is defined as σ / μ.
Let be drawn from this populations a random sample of n elements, furnishing a mean value $\bar{x}$, a standarddeviation s and $c = s/\bar{x}$. The values $\bar{x}$, s, c are estimates of μ, σ, γ.

$\gamma - 2\sigma_c$ and $c = \gamma + 2\sigma_c$ are drawn. In the nomogram represented in fig. 1, for the sake of clearnes, only four couples of curves are drawn: n = 6, 10, 20 and 30.

These curves are situated symmetrically in a vertical direction with respect to the 45° line (n = ∞); the larger the n, the nearer the curves from one and the same pair are to the hypotenuse. For a given value of γ and given sample size n this nomogram enables us to read off the reliability (uncertainty, inaccuracy) interval of c. Of course one can also compute the limits of this uncertainty range with the help of (1) and (2).

But it is more important to use the same nomogram in the opposite sense, because generally γ is unknown and c is calculated on the base of n independent measurements x_1, x_2, ... x_n. To this purpose we draw a horizontal line, the ordinate of which equals the calculated c and cut this line with the curves of the couple with parameter n. We read off the abscissi γ_1 and γ_2 of the intersection points on the upper and lower curve and conclude: the unknown γ must lie between the limits γ_1 and γ_2, in other words the reliability interval for γ is γ_1 up to γ_2. There is only a probability of $1 - 0.95 = 0.05$ that this conclusion is false. These value γ_1 and γ_2 can also be computed algebraically in stead of estimated graphically, but then two fourth degree equations should be solved:

$$c = \gamma + 2\gamma \sqrt{\frac{1 + \dfrac{2\,n-3}{n}\gamma^2}{2\,(n-1)}} \tag{3a}$$

$$c = \gamma - 2\gamma \sqrt{\frac{1 + \dfrac{2\,n-3}{n}\gamma^2}{2\,(n-1)}} \tag{3b}$$

N.B. If a smaller probability that the conclusion with regard to the reliability interval of γ may be false is preferred, say, for instance, 0.01, the nomogram should be drawn for the curves $c = \gamma - 2.58\,\sigma_c$ and $c = \gamma + 2.58\,\sigma_c$. Consequently this interval becomes broader, because γ_1 diminishes and γ_2 increases.

4 NUMERICAL EXAMPLE

A numerical example may illustrate the use of the nomogram. Suppose an investigator prefers a "sufficiently reliable" climate, i.e. with $\gamma \leqslant \gamma_0$; the critical level γ_0 may be, for instance, 0.15 (the choice of this level is not a statistical question). Suppose be calculates, based on 20 x − values, c = s : $\overline{x}$ = 0.175. Then the nomogram gives (see figure) a 0.95 reliability interval for γ, $0.13 \leqslant \gamma \leqslant 0.26$, in general $\gamma_1 \leqslant \gamma \leqslant \gamma_2$, with $\gamma_2 - c > c - \gamma_1$. Consequently the critical level $\gamma_0 = 0.15$ lies between the limits γ_1 and γ_2. If one wants to find a reliability interval smaller than the calculated (or graphically estimated) one, it is necessary to continue the measurements, but such decision should only be made in case γ_0 is situated within the reliability range. If, for instance, the investigator requires to find an upper limit

not larger than 0.22, it is with help of the nomogram (finding, by interpolation, the parameter of the curve for which the intersection point with the horizontal line $c = 0.175$ gives a $\gamma_2 = 0.22$) that the sample size is proved to be at least 30. This means that the measurements should be continued for at least $30-20 = 10$ years. At the end of this period the c value should be recalculated and a new decision should be made on the ground of the new γ_1 and γ_2-values.

5   CONCLUSION

This numerical example has shown that the nomogram can give either an indication that it is not senseful to continue the measurements or an indication that it is advisable to make further measurements during at least a certain number of years.

Needless to say that some knowledge of the climate in question is required (n, the number of years of previous observations, should be at least 5, see (1)). No statistical method or reasoning can lead to conclusions without any basic data being available.

REFERENCES
1   HALD, A. : Statistical theory with engineering application. 1952.