K O N I N K L I J K N E D E R L A N D S
M E T E O R O L O G I S C H I N S T I T U U T

Wetenschappelijk Rapport W.R. 60-7 (III-258)

Dr. C. Levert

Some statistical aspects on accuracy requirements arising from a non-routine use of climatological data

De Bilt, 1960

Kon. Ned. Meteor. Inst.
De Bilt

KONINKLIJK NEDERLANDS
METEOROLOGISCH INSTITUUT

Wetenschappelijk Rapport
W.R. 60-7 (III-258)
U.D.C.: 551.501 : 519.2 :
551.508 ~~519.2~~

Dr. C. Levert

Some statistical aspects on accuracy requirements arising from a non-routine use of climatological data

Summary

It is obvious that climatological purposes and accuracy of measurements and methods must be interdependent. This accuracy plays a role in climatological problems of significance, minimum values, densities of network, relationships, associations and representativity.

In this short and not exhausting study the author gives the statistical reasoning with respect to only four typical, partly numerical, examples. The desirability is stressed strongly that these be analogized as much as possible by the reader himself.

The paper ends with tentative definitions of some well known concepts as accuracy, precision, reliability, random and persistent errors, reproducibility and repeatability.

0. Introduction

The interdependence between climatological purposes, especially with a non-routine use of data in scientific research, and accuracy of measurements or methods may be read in at least two ways:

a) given the accuracies of measurement (method), one may ask which purposes of a climatological nature can be fulfilled and which not

b) given a set of different types of climatological purposes, one may ask which accuracies of measurement (method) will be required to achieve such purposes

A careful analysis of the problem seems to lead to the conclusion that the main effects of the inaccuracy of measurements (methods) refer to the statistical parameters standard deviation (σ), correlation coefficient (ρ) and regressioncoefficient (α).

Requirements of accuracy are met, we think, in five classes of climatological problems:

1. significance (as in example 1.1)
2. minimum values (as in example 1.2)
3. densities of network (spatial and time correlation coefficients)
4. relationships or associations (as in example 1.3)
5. representativity (as in example 1.4)

1. Four typical examples, which illustrate the statistical reasoning

These examples should be analogized as much as possible.

1.1. Significant differences

Suppose somebody makes at one and the same moment at each of two different points A and B only one temperature measurement (x and y). For some reason he likes to know whether the true temperature α at A is higher than the true temperature β at B for at least $1\,^{o}C$ (the actual question may look unrealistic, but only the principle is pointed out). The question arises: what must be the minimum value of the difference $d = x - y$, and what is the influence of the errors of measurement, for which a conclusion such as desired above is justified except for, say, 5 per cent of being false?

Suppose the true temperatures α and β are defined sufficiently well, then α may be measured as $x = \alpha + e$ and β as $y = \beta + f$. The e and f are errors of measurement, which are distributed, say, normally (systematic errors being absent) around zero.

The standard deviations of these normal distributions are σ_e and σ_f; suppose $\sigma_e = \sigma_f$, say σ. Even if $\alpha = \beta$ there is a probability 0.05 that only due to errors of observation $|d| = |x - y| > 2\sigma\sqrt{2}$. Here 1.96 ($\approx 2$) is the value which is exceeded with a probability 0.025 in the so-called standardized normal distribution.

The interpretation is: when measuring $|x - y| < 2\sigma\sqrt{2}$ (σ being known), then the conclusion "$\alpha = \beta$" is right in 95 out of 100 cases.

Suppose, however, that it is known a priori that $\alpha > \beta$, but the difference $\delta = \alpha - \beta$ is unknown. When measuring the same values α and β again (suppose that would be possible), the difference $d = x - y$

would follow a normal distribution around a mean value $\delta = \alpha - \beta$, with a standard deviation $\sigma\sqrt{2}$. There is a chance 0.05 that $d > \delta + 2.58 \cdot \sigma\sqrt{2}$. Now 2.58 is the value exceeded with a probability of 0.05 in the standard normal distribution.

Interpretation: when measuring $d > 1 + 2.5 \cdot \sigma\sqrt{2}$ (again σ being known), then the conclusion $\delta > 1$ is wrong in only 5 out of 100 cases.

For instance: with $\sigma = 0.1^\circ$, then the requirement is $d > 1.4^\circ C$
and with $\sigma = 1^\circ$, then the requirement is $d > 4.6^\circ C$

Hence it is necessary that $x - y > 1^\circ$ in order to be allowed to conclude that $\alpha - \beta > 1^\circ$ (with a 0.05 probability that the conclusion is false). This condition, however, is not sufficient, because $x - y$ should also exceed this 1°-level sufficiently much and this "sufficiently much" depends on the errors of measurement. The greater the measuring error the higher the level which should be surpassed by d.

N.B. In this case the best part of the measuring error consists of the reading error, but generally the reading error is only one of several components of the total measuring error. Then a careful investigation into the error components will be needed. Moreover the way in which these error components are distributed should be studied. It is unlikely that all distributions will be exactly normal. For instance, the reading error usually satisfies a rectangular distribution.

This example can be analogized easily in many ways. We think, for instance, of the establishment of two automatic weather stations at two different points in certain inhospitable regions. Suppose these stations can work during no more than four years. Nevertheless we want to know (on the basis of data received by radiograph) whether, for instance, the mean values of daily minimum temperatures differ more than $10^\circ C$. What is the upper limit of the standard deviation of measuring errors (now consisting of many components), provided this question is to be answered in the affirmative with a probability of, say, 0.05 of being wrong?

1.2. Minimum values

Suppose in some country a large lake will be reclaimed, causing probably a little change of the climate in the neighbouring region. Suppose the absolute minimum temperatures in each of several winters have already been measured. Now the question may arise: what is the smallest number of years (winters) after the completion of the reclaim on the basis of which the conclusion would be justified that there is a large probability that the change in the mean value of this element exceeds a given amount? It is also possible to read the question in the opposite direction. When the number of such measurements is given (by, for instance, financial restrictions), which changes in the element examined can be proved sufficiently sure and which not?

Since the problem refers to several parameters it can be formulated in several ways.

We prefer here a purely symbolic statistical formulation. Let be given a normal population (I), characterized fully by the mean value μ_I and standard deviation σ_I. This is the so-called null hypothesis H_o. Next this population is affected in some way or another. Suppose, if there is any change, this change will be a "shifting" (only the mean value alters), but it is questionable whether this can be measured in a statistically significant way; we write: $\mu_{II} = \mu_I + \gamma$ (?); $\sigma_{II} = \sigma_I$. This is called the alternative hypothesis H_a.

Suppose a random sample of m elements $x_1, x_2, \ldots x_m$ is drawn from population I and also a random sample of n elements $y_1, y_2 \ldots y_n$ from II. Generally $\bar{x} \neq \bar{y}$ and $s_I \neq s_{II}$; $\bar{x}$ and s_I are resp. mean value and standard deviation of the x-sample; $\bar{y}$ and s_{II} the same in the y-sample. Even if $\gamma = 0$ the result $c \equiv \bar{y} - \bar{x} > 0$ is possible ($c = 0$ would be very improbable), only due to sample effects. Therefore a measurement $\bar{y} > \bar{x}$ does not necessary imply $\mu_{II} > \mu_I$.

Let be:

α = probability of an "error of the first kind" (e.g. 0.05), that is the probability of rejecting H_o whereas H_o is correct

β = probability of an "error of the second kind" (e.g. 0.10), that is the probability of non-rejecting H_o whereas H_a is true

τ_α (or $\tau_{1-\beta}$) = value of τ, which is exceeded with a probability α (or $1 - \beta$); this τ follows the standard normal distribution. For instance $\alpha = 0.05 \therefore \tau_\alpha = 1.645$; $\beta = 0.90 \therefore \tau_{1-\beta} = 1.282$

$\tau \equiv \tau_\alpha + \tau_{1-\beta}$; e.g. $\alpha = 0.05$; $\beta = 0.90 \therefore \tau = 2.93$

Then the following relation can be derived:

$$1 = \frac{\sigma \cdot \tau}{\gamma} \sqrt{\frac{m + n}{m \cdot n}}$$

where $\sigma^2 = \sigma_w^2 + \sigma_e^2$, with

σ_w = standard deviation of the true values of the meteorological element in question (for instance the standard deviation of the true absolute minimum temperatures of $n \to \infty$ winters). These deviations may be termed "errors of nature". Each winter nature may be supposed to make a shot at the average value of the true absolute winterminima, but she always misses by a variable amount.

σ_e = standard deviation of the measuring errors, defined as follows.

Suppose the true value of the absolute minimum temperature in winter i is ξ_i, but the measurement gives $x_i = \xi_i + e_i$ ($e_i \gtrless 0$); $i = 1, 2, \ldots m$. The deviations e may be called "errors of measurement". Then suppose $\mathcal{E}e = 0$.[1] The standard deviation of the symmetric e-distribution is called σ_e. This e is considered as independent of ξ; moreover, systematic errors are absent.

The relation mentioned above, may be read as follows:

1) for given m, n, α, β, σ only γ values $> \gamma_o$ are measurable. The larger the m, the lower the level γ_o.
2) for given m, σ, α, β the level γ_o is higher (lower), the smaller (larger) the n.
3) If γ_o is prescribed and m, n, α, β are given, then γ's $> \gamma_o$ can be "measured" in a sense as described above, if σ is smaller than a specified upper limit σ_o.

$$\gamma_o = \sigma \cdot \tau \sqrt{\frac{m + n}{m \cdot n}} \quad ;$$

[1] $\mathcal{E}$ stands for $\lim \frac{1}{n} \sum e$ for $n \to \infty$

Here the accuracy enters into the computation, because, as said already, $\sigma = \sqrt{\sigma_w^2 + \sigma_e^2}$. There are two cases:

a) If $\sigma_o < \sigma$, then for given m, n, α, β (and whatever these values may be) a shifting of the universe can never be proved.

b) If $\sigma_o > \sigma$, then, for given σ, the σ_e should not exceed a well defined value: $\sqrt{\sigma_o^2 - \sigma_w^2}$. One reads: the larger the σ_w (that is the larger the natural variation of the phenomenon), the lower the σ_e (that is: the more difficult the measurement, since the "construction" of instruments or methods with a smaller σ_o is much more difficult).

N.B. Here the change of the population was supposed to be a "shifting". There may be circumstances in which it is obvious that only the standard deviation changes. Then a different statistical reasoning should be made and again the question of measuring errors enters into the computation.

1.3. Correlation

Suppose the measurements of two climatological quantities ξ and η are subject to errors d and e and that these errors are random, hence $\mathcal{E}d = \mathcal{E}e = 0$. The n pairs of values ξ_i, η_i are measured as $x_i = \xi_i + d_i$; $y_i = \eta_i + e_i$; $i = 1, 2 \ldots n$.

E.g. the rainfall totals in many months of April at two neighbouring stations S_1 and S_2 are measured; or else ξ = total sunshine duration in the growing season and η = the yield of some crop.

In both examples we want to know the universe correlation coefficient ρ between ξ and η (the ρ may measure the association between the populations). The sample of n pairs x_i, y_i gives the value r between x and y. If d and e are independent of each other and of ξ and η, then it can be shown that

$$\rho(x, y) \equiv \lim_{n\to\infty} r = f \cdot \rho(\xi, \eta), \text{ with } f = \frac{\sigma(\xi) \cdot \sigma(\eta)}{\sigma(x) \cdot \sigma(y)}$$

where $\sigma^2(x) = \sigma^2(\xi) + \sigma^2(d)$; $\sigma^2(y) = \sigma^2(\eta) + \sigma^2(e)$; hence $f \leqslant 1$ if $\sigma(d) \geqslant 0$, $\sigma(e) \geqslant 0$.

The correlation coefficient between the ξ, η values turns out

to be larger than the computed one, r, not only in the sample but also if the sample size increases infinitely. In other words: the existence of measuring errors diminishes the correlation coefficient. As soon as $\sigma(\xi)$, $\sigma(\eta)$, $\sigma(d)$, $\sigma(e)$ are known, the $r(\xi, \eta)$ is readily computed, when $r(x, y)$ is measured. If now ξ and η follow a binormal population, special formulae or nomographs enable us to find the so-called 95 per cent confidence (reliability) interval for $\rho(\xi, \eta)$. Let the lower and upper limits or boundaries of this interval be ρ_l and ρ_u. These are functions of n and ρ. If it is required that the relative width $(\rho_u - \rho_l) : \frac{1}{2}(\rho_u + \rho_l)$ of the confidence interval should not exceed a prescribed value, it is possible to compute the minimum number n of pairs of measurements for which this is the case. But it is also possible to give n and, provided that $\sigma(\xi)$ and $\sigma(\eta)$ are known, to compute those upper limits of $\sigma(d)$ and $\sigma(e)$ for which the relative broadth of the 0.95 reliability region of ρ remains beneath a prescribed value.

Without working out fully mathematically this aspect, we easily feel that the measuring accuracy may be important in such statistical computations.

Remark:

Sometimes ξ and η are not directly measured, but are derived by formulae from the original measurements (e.g. evaporation; humidity computed on the basis of wet and drybulb temperatures, etc.). Then the errors d and e will not always be independent of each other and of ξ and η. In such cases correlations may enter into the formulae for $\sigma(x)$ and $\sigma(y)$. The relations become less simple mathematically. However, notwithstanding such a mathematical-statistical complication, the principle remains the same.

1.4. Representativity

There exist many definitions of the concept "representativity". One of these is the following one: with regard to an unknown, but well defined, true mean value μ, a sample mean $\bar{x}$ may be called "representative", if there is a prescribed large probability P that the absolute difference $|\bar{x} - \mu|$ between the true and measured

mean value does not surpass a prescribed amount $\Delta > 0$. Again the accuracy of measurements plays a role.

Mathematically: let the variable x follow a normal distribution with unknown mean value μ and unknown variance σ^2. Let be drawn a random sample of n elements. The sample mean value $\bar{x}$ and variance s^2 are estimates of resp. μ and σ^2. Let t be the value, which is exceeded with a probability P in the so-called Student- or t-distribution with n - 1 degrees of freedom (for $n \to \infty$ this distribution changes into the standard normal one). It can be shown that

$$t/\sqrt{n} \leqslant \Delta/s \quad \text{or} \quad n \geqslant \left(\frac{t \cdot s}{\Delta}\right)^2 ; \; t \text{ is a function of } n.$$

Usually this relation is used to compute the minimum number of measurements for which the mean value has a prescribed "degree of representativity" (especially in normal period - considerations). Then, for the sake of simplicity, s is supposed equal to σ ; moreover, the standard normal distribution is used instead of the Student one and, for given Δ, the unknown n can be solved.

Here I propose a different use, explained numerically as follows (the example may look unrealistic; however, only the principle is stressed).

Let be x = total duration of sunshine in summer, at station A, as measured, for instance, with the Campbell Stokes sunshine recorder. The true summer mean value $\mu = \frac{1}{n} \Sigma_1^n \xi_i$ is unknown. In summer i the true value (suppose this value is defined sufficiently well) may be ξ_i, but $x_i = \xi_i + e_i$ is measured; e_i is called the random error of measurement and therefore satisfies a distribution with mean zero; let the standard deviation be σ_e (systematic errors are absent). Let the true value ξ follow a normal distribution, characterized by μ and σ_ξ. If e_i and ξ_i are independent, then $\sigma_x^2 = \sigma_\xi^2 + \sigma_e^2$ in the population and $s_x^2 = s_\xi^2 + s_e^2$ in the sample.

Definition: $\bar{x}$ is called representative for μ if $\frac{t}{\sqrt{n}} \leqslant \frac{\Delta}{s}$ or $s \leqslant \frac{\Delta}{t}\sqrt{n}$.

For instance: with Δ = 10 h; n = 5; P = 0.05 . (t = 2.78), then $s_x \leqslant 8.04$ h, but what about σ_x?

Now s_x is an estimate of σ_x. The value s_x is subject to sample effects, in particular if n is small. There is a probability $P^1 = 0.95$ that the unknown σ^2 is situated in the following confidence interval

$$\frac{n}{\chi_1^2} s^2 < \sigma^2 < \frac{n}{\chi_2^2} s^2; \chi_1^2 \text{ and } \chi_2^2 \text{ are functions of n.}$$

Here $\chi_1^2(\chi_2^2)$ is the value of χ^2 which is exceeded with a probability 0.025 (0.975) in the so-called χ^2-distribution with n - 1 degrees of freedom. For instance: with n = 5, then $\chi_1^2 = 11.0$; $\chi_2^2 = 0.46$ and $0.67\ s < \sigma < 3.28\ s$. But also $s_x \leqslant 8.04$ h. Hence $\sigma^2 = \sigma_\xi^2 + \sigma_e^2 < 700$ h. The σ_ξ characterizes the natural variation of the element, here the true total sunshine duration per summer. This inequality learns: if $\sigma_\xi > \sqrt{700} = 26.4$ h then the inequality can never be satisfied, even not for errorless measurements. If however $\sigma_\xi < 26.4$ h, then the measuring errors should possess a standard deviation smaller than $\sqrt{700 - \sigma_\xi^2}$ (σ_ξ should be known). Substituting, for instance, $\sigma_\xi = 26$ h, then the result is $\sigma_e < 4$ h; with, for instance, $\mu = 200$ h, then $\sigma_e/\mu < 2$ %. This very small percentual accuracy cannot be reached generally with the Campbell Stokes autograph. Suppose we double the number of measurements; n becomes 10. Then, again for $\Delta = 10$, $P = 0.05$ and $P^1 = 0.95$, the result is $\sigma^2 < 960$. Now, for $\sigma_\xi = 26$, it is necessary that $\sigma_e < 16.7$; hence $\sigma_e/\mu < 8.4$ % and now, probably, this requirement can be satisfied, provided that the measurements are carried out as carefully as possible. Here the inaccuracies chiefly refer to difficulties of analyzing the cards of the Campbell Stokes sunshine recorder, depending on quality of paper, quality of burning glass, sharpness of analyzing instruction and so on.

2. Appendix

Tentative definitions of the concepts accuracy, precision and reproducibility.

Accuracy Let be given a well defined, unknown, value ξ. Let be made n independent measurements of ξ: $x_1, x_2 \ldots x_n$.

The error (deviation of the truth) e_i of the ith measurement is defined by $e_i = x_i - \xi$ $(i = 1, 2 \ldots n)$.
Let be $\bar{x} = \frac{1}{n} \Sigma_1^n x_i$; $\bar{e} = \frac{1}{n} \Sigma_1^n e_i$; $\mathcal{E}x \equiv \lim_{n\to\infty} . \bar{x} = \mu$;
$\mathcal{E}e \equiv \lim_{n\to\infty} . \bar{e} = \delta \lessgtr 0$, and hence $\mu = \xi + \delta$ and $\bar{e} = \bar{x} - \xi$
This δ is called the systematic error (persistence error or bias), which persists during a series of the same or similar measurements and which is therefore not eliminated by any process of averaging.
Let be $s_x^2 = \frac{1}{n-1} \Sigma_1^n (x_i - \bar{x})^2$; $s_e^2 = \frac{1}{n-1} \Sigma_1^n (e_i - \bar{e})^2$
and $a^2 \equiv \frac{1}{n} \Sigma_1^n e_i^2$; further $\sigma_x^2 = \lim_{n\to\infty} . s_x^2$; $\sigma_e^2 = \lim_{n\to\infty} . s_e^2$
and $\alpha^2 = \lim_{n\to\infty} a^2$; then $s_x = s_e$; $\sigma_x = \sigma_e$ and $\alpha = \sqrt{\sigma_x^2 + \delta^2}$.

Precision

This σ_x measures the precision, whereas δ or α measures the accuracy; $\alpha \geqslant \sigma_x$ if $|\delta| \geqslant 0$.
The error e is called random if $\mathcal{E}e = 0$, that is if $\delta = 0$. Then $\mathcal{E}x = \xi$; such an error is individually unpredictable, but its average tends to zero in the long run.
Measurements may be highly precise (then σ_x is very small), but at the same time extremely inaccurate or unreliable (large α because of large δ).
The accuracy is sometimes called the reliability.

Reproducibility or repeatability

When, although the true value remains the same, the total series of measurements shows separate groups the mean value of which differ significantly or (and) the standard deviations of which differ significantly (σ_x or, and, δ are not constant in time), then the measurements (method) is called not reproducible. A measure of reproducibility may be based on the variation in time of δ or (and) σ_x.
Measurements (or a method) may be highly reproducible and extremely precise and at the same time very little accurate.

Literature: "The design and analysis of industrial experiments" O.L. Davies; London 1954.