STANDARDISATION OF SYMBOLS IN MATHEMATICAL STATISTICS AND BIOMETRICS

Proposals by Prof.Dr. D. van Dantzig.

Introductory note:

This document contains the draft proposals considered by the Netherlands Committee for the Standardisation of Statistical Nomenclature and Symbols (Standardisation Committe No.73 of the "Hoofdcommissie voor de Normalisatie in Nederland", the Netherlands Standardisation Institute).

They have been prepared by consideration and amendment of proposals submitted in the first instance by Prof.Dr. D. van Dantzig, the chairman of the committee.

It is recommended that these proposals be adopted in all cases except in those where
(a) a strongly established tradition exists (e.g. the well known use of χ^2) and
(b) no confusion is to be expected.

\- . -

§ 1. In order to obtain a practically useful and logically consistent notation for mathematical statistics and biometry, it is desirable to distinguish between the following types of quantities (1):

a) Satistical quantities, viz. quantities which are either given (or prescribed or otherwise predetermined) or observed, or computed from given and observed quantities by means of a determined computation-scheme, not depending upon unknown parameters or functions.

Notation: Latin letters, preferably printed in italics.
Examples: number of observations n; observed values x_1, ..., x_n; sample-mean m, sample-variance s^2.

b) Stochastic quantities; which are subject to some specified (though often partially or completely unknown) distribution function. They can be considered as variables on a probabilityfield (according to A. Kolmogoroff's definition), which serves as a mathematical model for the totality of all possible results of the experiments. It is necessary that the simultaneous distribution-function of all stochastic quantities together is specified; this is e.g. the case if the separate distribution-functions af all

(1) The terminological differences, necessary if instead of quantities, qualitative marks are considered, will not be mentioned explicitly. Also the question, whether the term "statistical" applied to predetermined quantities is an adequate one, is left out of consideration at present.

stochastic quantities under consideration are specified and these quantities are known to be independent.

Notation: underlined Latin letters (or Latin letters provided with some other distinguishing mark).

Examples: $\underline{x}_1, \ldots, \underline{x}_n$; $\underline{m}$, $\underline{s}^2$, etc. To the statistical quantities (belonging to a given system of observations) correspond (on the base of definite mathematical model) stochastic quantities (variable over the set of all systems of observations).

Remark 1) In most French and several other publications stochastic quantities ("variables aléatoires") are denoted by Latin capitals. The use of the same symbol without and with underlining for a statistic quantity and the corresponding stochastic quantity, however, has a certain advantage in avoiding duplication of formulae (cf § 2).

Remark 2) The distinction sometimes made between "estimates" (computed from a given sample) and "estimators" (variable over the set of all samples) of an unknown parameter is expressed in our system by underlining the latter ones, e.g. s^2 (estimate) and $\underline{s}^2$ (estimator).

Remark 3) When there is no confusion likely to arise the underlining of stochastic quantities may be omitted.

c) Model-parameters, in particular unknown ones determining the statistical model.

Notation: preferably Greek small letters.

Examples: $\mu, \sigma^2, \beta_1, \beta_2, \mu_k, \gamma_k, \varkappa_k$, etc.

Remark 1) The letters p and q = 1 - p for the probabilities of an alternative, accepted by tradition, should here be considered as "honorary Greek" letters.

Remark 2) Known parameters should strictly speaking be denoted by Latin letters (cf. a). If e.g. the mean and the standard deviation are given, they ought to be denoted, however, by letters different from m and s, which denote different quantities (cf. 3).

§ 2. It is desirable to distinguish between:

a) The statistical mean of a statistical or a stochastic quantity and

b) its stochastical mean or (mathematical) expectation.

Notation: It is proposed to designate the operators of forming the statistical mean by a curved letter $\mathcal{M}$ (or sometimes by a bar above the operand), and the expectation by a curved letter $\mathcal{E}$.
It is further proposed to designate the (cumulative) distribution-function preferably by a capital, e.g. $F(x) = \mathcal{P}[\underline{x} \leqq x]$ (2), and (if it exists) the distribution-density (or probability-density) by a small letter, e.g. $f(x)$ (2).

Remark: The use of a curved $\mathcal{P}$ for "the probability of" would make the ordinary P free for other purposes. So do abbreviations like Pr or Prob.

Examples: Ordinary moments:

$$\mu_k = \mathcal{E}\,\underline{x}^k = \int x^k dF(x) = \int x^k f(x)\,dx$$

$$\underline{m}_k = \mathcal{M}\,\underline{x}^k = \frac{1}{n}\sum_i^n \underline{x}_i^{\,k}$$

Remark 1) As long as only statistical quantities are considered, the underlining is dropped:

$$m_k = \mathcal{M}x^k = \frac{1}{n}\sum_i^n x_i^{\,k}$$

The operator $\mathcal{E}$ can be applied to stochastic quantities only (3):

$$\mathcal{E}\,\underline{m}_k = \frac{1}{n}\sum_i^n \mathcal{E}\,\underline{x}_i^{\,k} = \mu_k$$

if all $\underline{x}_i$ have the same distribution.

Remark 2) The suffix 1 for the ordinary moment of order 1 may be omitted:

$$\mu_1 = \mu\,; \quad m_1 = m$$

§ 3. It is proposed to denote quantities, formed in the same way by consistent use of the operators $\mathcal{E}$ and $\mathcal{M}$ respectively by corresponding Greek and Latin letters, e.g.

$$\sigma^2 = \mathcal{E}(\underline{x} - \mathcal{E}\,\underline{x})^2$$

$$\underline{s}^2 = \mathcal{M}(\underline{x} - \mathcal{M}\,\underline{x})^2 = \frac{1}{n}\sum_i^n (\underline{x}_i - \underline{m})^2$$

Further it is proposed to denote statistical quantities, used as "best" estimates in some sense for model-parameters, expressed by Greek letters, by other symbols or by the corresponding Latin letters provided with some distinguishing mark, e.g. a prime.

---

(2) The letters F and f may also be considered here as "honorary Greek" ones.

(3) $\mathcal{E}m_k$ should preferably not be used. Strictly speaking $\mathcal{E}m_k$ (not underlined) $= m_k$, as m_k is a constant.

Example: $\breve{V} = s'^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \sum (x_i - m)^2$; $\mathcal{E}\, \underline{s}'^2 = \sigma^2$

k'_p Fisher's k-statistics; $\mathcal{E}\, \underline{k}'_p = \varkappa_p$

Remark: The custom of using corresponding Greek and Latin letters to denote model-parameters and their unbiased estimates, is not advocated. It is applied by several authors when denoting $\frac{1}{n-1} \sum (x_i - m)^2$ by s^2, and also when denoting by k_p Fisher's statistics, denoted here by k'_p. It is, however, apparently never followed consistently. This would appear to require the use of $\breve{m}_k$ (4) (m_k according to the customary notation) for the quantity $\frac{n^{!k}}{n^k} \mathcal{M}(x - m)^k$ (5) instead of $\mathcal{M}(x - m)^k$ itself. This, however, is not done usually.

We did not succeed in finding any general principle determining uniquely the use of corresponding latin and Greek letters which justifies the use of s^2 for V and which might be followed consistently. We assume that the use of the letter V (= variance) for $\frac{n}{n-1} s^2$ might remove the objections which some biometrists may have against the use of s'^2.

§ 4. Absolute moments. It is proposed to denote absolute moments of (not necessarily integral) order k, by:

$$|\mu|_k = \mathcal{E}\, |\underline{x}|^k, \quad |m|_k = \mathcal{M}\, |x|^k$$

Remarks: The suffix k must remain outside the vertical lines, as $|\mu_k| = |\mathcal{E}\, \underline{x}^k|$. The suffix 1 for the absolute moment of order 1 can not be omitted.

§ 5. Factorial moments. Instead of the multitude of notations like $x^{[k]}$, $x^{(k)}$, $x^{\lfloor k}$, etc. for factorial powers, the notation $x^{!k}$ is proposed, which suggests the relation with factorials and the analogy with powers.

$$x^{!k} = x(x - 1)(x - 2) \ldots (x - k+1).$$

Accordingly the notations $\mu_{!k}$ and $m_{!k}$ for the factorial moments are proposed:

$$\mu_{!k} = \mathcal{E}\, \underline{x}^{!k} = \mathcal{E}\, \underline{x}(\underline{x} - 1) \ldots (\underline{x} - k+1)$$

$$\underline{m}_{!k} = \mathcal{M}\, \underline{x}^{!k} = \frac{1}{n} \sum \underline{x}_i (\underline{x}_i - 1) \ldots (\underline{x}_i - k+1).$$

---

(4) cf. §6.

(5) cf. §5 for the meaning of the symbol $n^{!k}$.

§ 6. Reduced moments. It is proposed to use the terms <u>reduced</u> value of a stochastic or statistical variable for its difference from the mean, <u>standardised</u> value for the quotient by the standarddeviation, and <u>normalised</u> value for the standardised reduced value. It is proposed to introduce a special notation for the <u>reduced</u> values only, the other ones being used far less frequently. Moreover it is necessary to distinguish reduction with regard to the stochastic ("true") and to the statistic ("sample") mean.

<u>Notation</u>: It is proposed to denote reduction with regard to the <u>statistic</u> mean m by an arc $\breve{}$ and to the <u>stochastic</u> mean by a sinusoidal line $\sim$ above the letter:

$$\breve{\underline{x}} = \underline{x} - \underline{m} \qquad \tilde{\underline{x}} = \underline{x} - \mu$$

$$\breve{\underline{m}}_k = \mathcal{M}(\underline{x} - \underline{m})^k \qquad \tilde{\mu}_k = \mathcal{E}(x - \mu)^k$$

Moreover:

$$\breve{\mu}_k = \mathcal{E}(\underline{x} - \underline{m})^k \qquad \tilde{\underline{m}}_k = \mathcal{M}(\underline{x} - \mu)^k$$

The quantity $\tilde{m}_k$, which is not statistic, is preferably to be avoided.

<u>Remark</u>: The current notation, viz μ'_k for the non-reduced and μ_k for the reduced moments, has the disadvantages (first) of not admitting the distinction between the two kinds of reduction and (second) of not corresponding with a notation for reduced variates. If it were to be maintained, it would be desirable (first) to restrict it to one of the two types of reduction only, e.g. to reduction with regard to the stochastic mean; (second) to denote any <u>non-reduced variate</u> by a letter <u>with a prime</u>, e.g. x', and its reduced value by the same letter without the prime.

§ 7. Cumulants. It is proposed to maintain the current notation $\varkappa_k$ for the cumulant (Thiele's semi-invariant) of the order k with regard to the stochastic mean, and the notation k_k for the quantity formed in the same way with regard to the statistic mean. Hence:

$$\mathcal{E} \exp \underline{x}t = \exp \tilde{\sum} \varkappa_k t^k/k!$$

$$\mathcal{M} \exp \underline{x}t = \exp \breve{\sum} \underline{k}_k t^k/k!$$

Further it is proposed to use the notations σ^2 and s^2 for the second cumulants:

$$\sigma^2 = \mathcal{E}(\underline{x} - \mu)^2 = \mathcal{E}\,\tilde{\underline{x}}^2 \qquad \underline{s}^2 = \mathcal{M}(\underline{x} - \underline{m})^2 = \mathcal{M}\,\breve{\underline{x}}^2$$

§ 8. Invariants. It is proposed to maintain the notation γ_{k-2} for the invariant of order k, and g_{k-2} for the statistic quantity formed in the same way:

$$\gamma_{k-2} = \varkappa_k / \sigma^k \; ; \; g_{k-2} = k_k / s^k$$

Remark: The use of K. Pearson's invariants $\beta_1 = \gamma_1^2$ and $\beta_2 = \gamma_2 + 3$ remains free, but is not advocated. In any case it is incorrect to write a minus-sign before the value of β_1, if it is meant that γ_1 is negative. E.g. $\beta_1 = -0{,}3794$ ought to be written as $\gamma_1 = -\sqrt{0{,}3794}$. The use of $\sqrt{\beta_1}$ instead of γ_1 is contrary to the customary use of the $\sqrt{\ }$-sign as denoting the positive root.

§ 9. Order-statistics and quantiles. If $x_1, \ldots, x_n$ are n statistical (e.g. observed) quantities, the smallest and largest are denoted by $x_{(1)}$ and $x_{(n)}$ respectively, the m^{th} according to increasing order by $x_{(m)}$. As an alternative notation also the symbols $q_{(m/(n+1))}$ for these quantities are proposed. If n+1 is a multiple of r, the h^{th} r-ile is therefore $x_{(h(n+1)/r)} = q_{(h/r)}$. In particular the median is $x_{((n+1)/2)} = q_{(0,5)}$ when n+1 is even, the 1^{st} and 3^{rd} quantile are $x_{((n+1)/4)} = q_{(0.25)}$ and $x_{(3(n+1)/4)} = q_{(0,75)}$ when n+1 is a multiple of 4, the percentiles $x_{((n+1)/100)} = q_{(0,01)}$, etc., if n+1 is a multiple of 100, etc.
It is proposed to use the symbol $\varphi_{(c)}$ for the corresponding model-parameters, if they are unique:

$$F(\varphi_{(c)}) = c$$

In particular the distribution median (when it is unique) is denoted by $\varphi_{(0,5)}$, etc.

Remark: The older Greek letter φ (koppa), corresponds with our q (first letter of quantile). Instead of φ, also the letter ξ might be used.

§ 10. Multi-dimensional distributions and other topics. We hope to come back upon these subjects on another occasion.
