STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

Publicatie

77

Een voorstel tot normalisatie van de symbolen in de 'statistica' en de 'biometrica'

D. van Dantzig

MC

1950

2

Overdruk uit Statistica, Jaargang 4, no. 1-2, 1950

Een voorstel tot normalisatie van de symbolen in de ‚statistica' en de ‚biometrica'

door „Normalisatie Commissie 73"

Proposals for the standardization of symbols in mathematical statistics and biometrics. (Introduction in Dutch, main text in English).

Inleiding.

Ongeveer anderhalf jaar geleden werd op aandringen van de „Vereniging voor Statistiek" door de „Hoofdcommissie voor de Normalisatie in Nederland" (HCNN) ingesteld de „Normalisatie Commissie 73" voor de normalisatie van de nomenclatuur en de symboliek der wiskundige statistiek.

Door de voorzitter van deze laatste commissie, Prof. Dr D. van Dantzig, werden in de loop van het vorig jaar voorstellen uitgewerkt tot normalisatie van de symbolen. Deze voorstellen zijn, na in een vergadering van Commissie 73 te zijn besproken en geamendeerd, voorlopig als discussiebasis aanvaard. Zij zijn nadien ook ter tafel gebracht op het van 30 Augustus tot 2 September 1949 te Genève gehouden Congres van de „Biometric Society".

In deze voorstellen is getracht, zoveel mogelijk aansluitend aan hetgeen thans algemeen gebruikelijk is, te komen tot een consequente symboliek, waarin het essentiële onderscheid tussen verschillende typen van statistische grootheden duidelijk tot uiting komt. Met toestemming van de HCNN en van Commissie 73 drukken we deze voorstellen hieronder af. Zij zullen de lezers ongetwijfeld enige stof tot nadenken bieden.

De Heer J. Sittig*), secretaris van Commissie 73, zal eventuele aanvullingen of kritische opmerkingen over deze voorstellen gaarne in ontvangst nemen, aangezien het werk van de commissie hierbij zal zijn gebaat.

Ten slotte zij er met nadruk op gewezen dat het hier slechts *voorstellen* geldt, en dat, zoals de „insiders" in het normalisatiebedrijf maar al te goed weten, nog veel water door de Rijn zal moeten vloeien, alvorens deze voorstellen tot een definitief Nederlands normblad zullen zijn uitgegroeid; om van de internationale normalisatie maar liever niet te spreken.

§ 1. *Fundamental distinction of different types of quantities*

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]):

*) Adviesbureau voor Toegepaste Statistiek, Wassenaarseweg 2, 's-Gravenhage.

[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.

a) *Statistic(al) quantities* [2]), 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: Roman letters, preferably printed in italics.
Examples: number of observations n; observed values $x_1 \ldots, x_n$; sample-mean m, sample-variance s^2.

b) *Stochastic quantities* [3]), which are subject to some specified (though often partially or completely unknown) distribution function. They can be considered as variables on a probability field (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 of all stochastic quantities under consideration are specified and these quantities are known to be independent.

Notation: underlined Roman letters (or Roman letters provided with some other distinguishing mark) [4]).
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 basis of a 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 Roman capitals [5]). The use of the *same* symbol without and with underlining for a statistic quantity and the corresponding stochastic quantity has, however, a certain advantage in avoiding duplication of formulae (cf § 2).

[2]) This term seems preferable to the shorter 'statistics' which also denotes the whole field of study.

[3]) The denomination 'eventual quantities' should also have some advantages, and the terminology adopted here must be considered as only provisional.

[4]) The underlining of symbols may sometimes lead to difficulties in printing. In such cases printing in bold type may be used instead, though this has the disadvantage of being also used for vectors, matrices, etc. Other distinguishing marks (dots, asterisks, etc.) seem to be less advisable.

[5]) It should be observed, however, that this notation 'spoils' a complete alphabet; besides, a consistent application will be difficult, Roman capitals being hardly avoidable for several quantities other than statistic ones.

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 no confusion is likely to arise, the underlining of stochastic quantities may be omitted.

c) *Model-parameters*, in particular unknown ones, determining the statistical model.
Notation: preferably small Greek letters.
Examples: μ, σ^2, β_1, β_2, μ_k, γ_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 Roman 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. *Statistical mean and stochastical mean*

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. $F(x) = \mathcal{P}\,[\underline{x} \leqq x]$ [6]), and (if it exists) the *distribution-density* (or probability-density) by a small letter, e.g. $f(x)$ [6]).

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:

$$\mu_k = \mathcal{E}\,\underline{x}^k = \int x^k \,\mathrm{d}\,F(x) = \int x^k f(x)\,\mathrm{d}x;$$

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

[6]) The symbols F and f may also be considered here as „honorary Greek" letters.

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

$$m_k = \mathfrak{M}\, x^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

The operator $\mathfrak{E}$ can be applied to stochastic quantities only [7]):

$$\mathfrak{E}\, \underline{m}_k = \frac{1}{n} \sum_{i=1}^{n} \mathfrak{E}\, x_i^k = \mu_k,$$

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

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

$$\mu_1 = \mu;\; m_1 = m.$$

§ 3. *Use of the operators* $\mathfrak{E}$ *and* $\mathfrak{M}$.

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

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

$$\underline{s}^2 = \mathfrak{M}\, (\underline{x} - \mathfrak{M}\, \underline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (\underline{x}_i - \underline{m})^2.$$

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

Example:

$$V = s'^2 = \frac{n}{n-1} s^2 = \frac{1}{n-1} \Sigma\, (x_i - m)^2;\; \mathfrak{E}\, \underline{s}'^2 = \sigma^2;$$

k'_p = F i s h e r's k-statistics; $\mathfrak{E}\, \underline{k}'_p = \varkappa_p$.

Remark: The custom of using corresponding Greek and Roman letters to denote model-parameters and their unbiased estimates, is not recommended. It is applied by several authors when denoting $\frac{1}{n-1} \Sigma (x_i - m)^2$ by s^2, and also when denoting by k_p F i s h e r's statistics, denoted here by k'_p. It is, however,

[7]) $\mathfrak{E} m_k$ should preferably not be used. Strictly speaking $\mathfrak{E} m_k$ (not underlined) $= m_k$, as m_k is a constant.

apparently never followed consistently. This would appear to require the use of $\breve{m}_k$ [8]) (m_k according to the customary notation) for the quantity $\frac{n^k}{n^{!k}} \mathfrak{M} (x - m)^k$ [9]) instead of $\mathfrak{M} (x - m)^k$ itself. This, however, is not done usually. We did not succeed in finding any general principle determining uniquely the use of corresponding Roman 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:

$$|\mu|_k = \mathfrak{E} |\underline{x}|^k, \; |m|_k = \mathfrak{M} |x|^k.$$

Remarks: The suffix k must remain *outside* the vertical lines, as $|\mu_k| = |\mathfrak{E} \underline{x}^k|$. 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*):

$$\mu_{!k} = \mathfrak{E} \underline{x}^{!k} = \mathfrak{E} \underline{x} (\underline{x} - 1) \ldots (\underline{x} - k + 1) \text{ [10]};$$

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

§ 6. *Reduced moments.*

It is proposed to use the terms *reduced* value of a stochastic or statistical variable for its difference from the mean, *standardized* value for the quotient by the standard deviation, and *normalized* value for the standardized reduced value. It is proposed to introduce a special notation for the *reduced* value 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.

8) cf. § 6.

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

10) The notations $|\mu_k|$, $\mu_{!k}$, $\tilde{\mu}_k$ are proposed because they save letters.

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

$$\breve{\underline{x}} = \underline{x} - \underline{m} \quad ; \; \tilde{\underline{x}} = \underline{x} - \mu;$$
$$\breve{\underline{m}}_k = \mathfrak{M}\,(\underline{x} - \underline{m})^k \; ; \; \tilde{\mu}_k = \mathcal{E}\,(\underline{x} - \mu)^k.$$

Moreover:

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

The quantity $\breve{\mu}_k$, which is not statistic, is preferably to be avoided.
Remark: The current notation, viz μ'_k for the non-reduced and μ_k for the reduced moments, has the disadvantages (1) of not admitting the distinction between the two kinds of reduction and (2) of not corresponding with a notation for reduced variates. If it were to be maintained, it would be desirable (*a*) to restrict it to one of the two types of reduction only, e.g. to reduction with regard to the stochastic mean; and (*b*) to denote any *non-reduced variate* by a letter *with a prime*, 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 (or $\underline{k}_k$) for the quantity formed in the same way with regard to the statistic mean. Hence.:

$$\mathcal{E} \exp \underline{x}t = \exp \sum_1^\infty \varkappa_k\, t^k/k!;$$

$$\mathfrak{M} \exp \underline{x}t = \exp \sum_1^\infty \underline{k}_k\, t^k/k!.$$

Further it is proposed to use the notations σ^2 and s^2 (or $\underline{s}^2$) for the second cumulants:

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

§ 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 recommended. 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.