THE STATISTICAL WORK OF DAVID VAN DANTZIG (1900–1959)

By J. Hemelrijk

Technological University of Delft; Mathematical Centre, Amsterdam

Reprinted from The Annals of Mathematical Statistics
Vol. 31, No. 2, June, 1960
Printed in U.S.A.

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 31, No. 2, June, 1960
Printed in U.S.A.

THE STATISTICAL WORK OF DAVID VAN DANTZIG (1900–1959)

BY J. HEMELRIJK

Technological University of Delft; Mathematical Centre, Amsterdam

1. Introduction. The death of David van Dantzig on July 22nd 1959 bereaved Holland of its foremost mathematical statistician and put an untimely end to his work. His importance for the development of research in and application of probability and statistics in Holland has been large; we lose not only an outstanding mathematician but also a pioneer. A short outline of his life and of his statistical work is given in the following sections. One topic is discussed in some detail: an application of his theory of "collective marks" to rank correlation. The theory of collective marks is perhaps his most interesting and most promising contribution to the theory of probability and statistics. The example mentioned has as yet only been published in mimeographed form [34]; it is a nice demonstration of the power of the method used and it may show the way to further applications.

A bibliography of probabilistic and statistical papers is given; a bibliography of all van Dantzig's writings can be found at the end of [40] and a more extensive discussion of van Dantzig's statistical work in [41].

2. His life. van Dantzig started his mathematical career as a pure mathematician: his main subjects before the war were topology, differential geometry, philosophy (in particular significs–"significa"–and the foundations of mathematics and physics) and mathematical physics. From 1932 onward he lectured at the Technical University at Delft, first as a Lecturer and later (1938) as Professor.

During the war he was discharged by the Germans and studied probability and statistics. Holland was at that moment decidedly an underdeveloped country as far as these subjects were concerned and he put himself to the task of changing this. He emerged from the war as an outstanding statistician and was at once given a central place as such at the municipal University of Amsterdam as Professor in the Theory of Collective Phenomena. Together with Prof. J. G. Van der Corput and Prof. J. F. Koksma he founded the "Mathematical Centre" in Amsterdam, an institution, subsidised by government and industry, which unites all branches of pure and applied mathematics in one organisation. As head of the Department of Mathematical Statistics he was able to stimulate research and consultation to such a degree that general recognition nationally and internationally soon followed. He was appointed Fellow of the Institute of Mathematical Statistics, the American Statistical Association and the Royal Statistical Society. He was a member of the International Statistical Institute and, in Holland, of the Royal Academy of Sciences (*Koninklijke Akademie van Vetenschappen*) and, of course, an outstanding member of the Dutch Statistical Association (*Vereniging voor Statistiek*). He was Visiting Professor at the Uni-

Received February 5, 1960.

versity of California (Berkeley) in 1951 and worked for some time at the National Bureau of Standards in Washington, D. C.

His impact on statistics in Holland was formidable. He instilled his pupils with his sharp and critical way of thinking, which is manifest in his mathematical work as well as in his views on foundations and consultation. As an objectivist and frequentist he was apt to be sharp in his criticism of other views, [31], [33], and in consultation he adhered strictly to high standards of exactness and precision.

During the last years of his life he supervised a group of mathematicians in the departments of Mathematical Statistics and of Applied Mathematics at the Mathematical Centre, who worked on the problem of the high water levels of the North Sea. This work was initiated by the Dutch government's "Delta Commission" following the disastrous flood of 1953. Partial publications on this subject are [25], [29], [35] and [36]. A final version of a report of several hundreds of pages was lying on his desk on the day of his death. He had been putting final touches to the manuscript, which will be published as a book, one volume of the report of the Delta Commission to the government.

3. His theory of collective marks. The theory of collective marks has been extensively described and applied in [4], [5], [22], [27], [34] and [37]. Here only the fundamental ideas will be given and one example of their application. For simplicity, we only consider the case of a discrete random variable assuming the values

$$x_1, x_2, \cdots \text{ with probabilities } p_1, p_2, \cdots \ (\textstyle\sum p_i = 1). \tag{1}$$

Generalisation to other cases is straightforward, although sometimes complicated.

The *collective mark* of the probability distribution is defined by

$$C \underset{=\!=}{\text{def}} \sum_i p_i A_i, \text{[1]} \tag{2}$$

where the A_i $(i = 1, 2, \cdots)$ and C are abstract symbols, for which all kinds of substitutions can be considered.

If E is an arbitrary possible event and $P(\bar{E} \mid x_i)$ is the conditional probability, given x_i, that E will *not* occur, then substitution of $P(\bar{E} \mid x_i)$ for A_i leads to

$$\sum_i p_i P(\bar{E} \mid x_i) = P(\bar{E}); \tag{3}$$

i.e. this substitution transforms C into $P(\bar{E})$.[2] This simple property is fundamental for the application of the theory.

Another interesting substitution is A^i for A_i, where the As. are numbers of some kind, giving

$$C(A) \underset{=\!=}{\text{def}} \sum_i p_i A^i. \tag{4}$$

[1] $\underset{=\!=}{\text{def}}$ indicates that the left hand member is defined by the right hand one.

[2] The use of $\bar{E}$ instead of E in this formula turns out later to be convenient for application, so it is introduced from the start.

The p_i are uniquely determined if $C(A)$ is known. Taking A as a complex number this is the *generating function* of the p_i; with $A_i = \exp \tau x_i$ (2) yields the *moment generating function* and with $A_j = \exp [itx_j]$ (real t) the *characteristic function* is obtained. Thus the collective mark is a generalisation of these functions.

Van Dantzig was not the only one considering this kind of general functional, but he gave a new turn to the theory by introducing an imaginary event E (which he called "a catastrophe; e.g. the ignition of a red light") which was associated with a realization of the random variable by means of a random experiment ("lottery") in the following way: if x_i occurred then $\bar{E}$ (non-E) occurs with a prescribed probability

$$P(\bar{E} \mid x_i). \tag{5}$$

According to (3), the unconditional probability $P(\bar{E})$ that no catastrophe occurs is then identical with the collective mark if (5) is substituted for A_i. However, the probabilities (5) can be chosen arbitrarily, giving ample latitude for manipulation. An adroit choice of (5) may make it easy to compute $P(\bar{E})$ directly and then C, so that e.g., the characteristic function follows at once.

Thus the collective mark can be given an interpretation as a probability and the way to short-cut derivations of important results is opened. This method has been applied successfully by van Dantzig and some of his pupils, and it certainly deserves further consideration.

4. Application to rank correlation. [34] To illustrate this method consider m independent random samples $S_1, \cdots, S_m$ from the same continuous population

$$x_{\lambda,i} \qquad (\lambda = 1, \cdots, m; \quad i = 1, \cdots, n_\lambda; \quad \textstyle\sum n_\lambda = n). \tag{6}$$

The statistic

$$T \underline{\underline{\mathrm{def}}} \text{ the number of pairs } (x_{\lambda,i}, x_{\mu,j}) \text{ with } \lambda < \mu \text{ and } x_{\lambda,i} > x_{\mu,j} \tag{7}$$

$$(\lambda, \mu = 1, \cdots, m; \quad i = 1, \cdots, n_\lambda; \quad j = 1, \cdots, n_\mu)$$

is fundamental in rank correlation, and its distribution is known. T can assume integer values from 0 to $N = \sum_{\lambda<\mu} n_\lambda n_\mu$. The method of collective marks allows of a straightforward derivation of the collective mark and the characteristic function of T.

Let h marbles, numbered $1, \cdots, h$, roll along the x-axis starting from $+\infty$ and coming to rest according to independent random drawings from the continuous population considered. The probability that two marbles come to rest at the same point is then zero, and all permutations of the numbers $1, \cdots, h$ as read from the marbles along the x-axis have the same probability, $(h!)^{-1}$. Also, for any given subgroup of marbles all permutations are equally probable and independent of the permutations in other subgroups having no marbles in common with the former. For any two or more subgroups without common elements all possible situations with respect to each other are equally probable.

All this remains valid if the marbles are rolled one at a time in the order of their numbers, no two being in motion at the same time. Now each time a rolling marble on its way passes an "earlier" marble which has come to rest already, let a ticket be drawn at random and with replacement from a lottery L with probability $1 - A$ of a catastrophe. Consider the conduct of the kth marble. All permutations of the first k marbles have equal probability and therefore the probability that the kth marble passes $0, 1, \cdots, k - 1$ marbles is k^{-1} for each of these possibilities, independent of the number of passings realised by the first $k - 1$ marbles. The marginal probability that the kth marble does not cause a catastrophe is therefore

$$k^{-1}(1 + A + A^2 + \cdots + A^{k-1}) = (1 - A^k)/[k(1 - A)]. \tag{8}$$

For all h marbles together we thus find, because of the independence mentioned already, that the probability of no catastrophe is

$$P_h(\bar{E}) = (1 - A)^{-h} \prod_{k=1}^{h} (1 - A^k)/k = \{h!(1 - A)^h\}^{-1} \prod_{k=1}^{h} (1 - A^k). \tag{9}$$

According to Section 3 this is also the collective mark of the total number of passings during the whole process of rolling h marbles,

$$P_h(\bar{E}) = C_h(A). \tag{10}$$

Rolling n marbles consecutively for the m samples $S_1, \cdots, S_m$, starting with n_1 marbles for sample S_1 and ending with n_m for S_m, this also holds for all passings. In order to link this up with the statistic T we consider two kinds of passings,

1. "Same sample passings" of marbles passing earlier marbles of the same sample;
2. "Different sample passings" of marbles of a sample S_μ passing marbles S_λ with $\mu > \lambda$.

T is the total number of different sample passings.

For the same sample passings of S_λ we have, according to (9), in an obvious notation,

$$P_{\lambda,\text{ same}}(\bar{E}) = C_{n_\lambda}(A) \qquad (\lambda = 1, \cdots, m), \tag{11}$$

and, again using the independence of passings within different samples, we find for all same sample passings:

$$P_{\text{same}}(\bar{E}) = \prod_{\lambda=1}^{m} C_{n_\lambda}(A). \tag{12}$$

The total number of passings is equal to the sum of the numbers of both kinds and these two numbers are again stochastically independent. Thus we find

$$P_{\text{total}}(\bar{E}) = P_{\text{same}}(\bar{E}) \cdot P_{\text{different}}(\bar{E}), \tag{13}$$

where the last factor may also be written $P_T(\bar{E})$. Therefore

$$C_n(A) = \prod_{\lambda=1}^{n} C_{n_\lambda}(A) \cdot C_T(A), \tag{14}$$

where

$$C_T(A) = \sum_{T=0}^{N} P_T \cdot A^T = P_{\text{different}}\,(\bar{E}) \qquad (N = \sum_{\lambda<\mu} n_\lambda n_\mu) \tag{15}$$

is the collective mark of T.

From (9), (14), and (15) we obtain

$$\begin{aligned} C_T(A) &= \sum_{T=0}^{N} P_T \cdot A^T = C_n(A) \Big/ \prod_{\lambda=1}^{m} C_{n_\lambda}(A) \\ &= [(n_1! \cdots n_m!)/n!] \left[\left(\prod_{k=1}^{n} (1 - A^k) \right) \Big/ \left(\prod_{\lambda=1}^{m} \prod_{j=1}^{n_\lambda} (1 - A^j) \right) \right]. \end{aligned} \tag{16}$$

Substituting $A = e^{it}$ the characteristic function is obtained. For $n_\lambda = 1$ ($\lambda = 1, \cdots, m$), i.e. for samples of one element each, $n = m$ and $2T - N$ is equal to M. G. Kendall's ranking statistic S; the characteristic function can then be reduced to

$$\mathcal{E} e^{itS} = \prod_{k=1}^{m} = (\sin kt)/(k \sin t), \tag{17}$$

a formula derived earlier by Kendall by means of recurrence relations. For $n_\lambda > 1$ the method pertains to a test for trend in the several-sample problem.

The rest of van Dantzig's statistical and probabilistic work is here presented by title only in the bibliography.

5. Bibliography of van Dantzig's statistical and probabilistic papers.

[1] "Mathematical and empirical foundations of probability theory" (in Dutch), *Ned. Tijdschr. voor Nat.*, Vol. 8 (1941), pp. 70–93.

[2] "Correspondence with B. de Finetti" (Punti di Vista), *Statistica* Vol. 1 (1941), pp. 229–242 (Milano).

[3] *Probability theory*, (in Dutch; mimeographed course), Mathematisch Centrum, Amsterdam, 1947.

[4] *Mathematical Statistics*, (in Dutch, mimeographed course), Mathematisch Centrum, Amsterdam, 1947–48.

[5] "Sur la méthode des fonctions génératrices," Le Calcul des Probabilités, *Coll. Int. du Centre Nat. de la Rech. Sc.*, Vol. 13 (1949), pp. 29–46.

[6] "Blaise Pascal and the meaning of mathematical thinking for the study of human society" (in Dutch; inaugural address), *Euclides*, Vol. 25 (1948), pp. 203–232.

[7] "Sur l'analyse logique des relations entre le Calcul des Probabilités et ses applications," Congrès Intern. de Phil. des Sciences, 1–8 July 1949, Vol. 4 Calcul des Probabilités; *Actualités Scientif. et Industr. Nr. 1146*.

[8] "Proposals for the standardization of symbols in mathematical statistics and biometrics," *Statistica (Neerlandica)*, Vol. 4 (1950), pp. 80–86.

[9] "Laplace, probabiliste et statisticien et ses précurseurs," *Archives Intern. d'Histoire des Sciences*, Vol. 8 (1955), pp. 27–37.

[10] "Une nouvelle généralisation de l'inégalité de Bienaymé," *Ann. de l'Inst. Henri Poincaré*, Vol. 12 (1951), pp. 31–43.

[11] "Some historical relations between mathematical and descriptive statistics" (in Dutch), *Statistica*, Vol. 4 (1950), pp. 233–248.

[12] "On the consistency and the power of Wilcoxon's two-sample test," *Proc. Kon. Ned. Ak. v. Wet.* 54A; *Indag. Math.*, Vol. 13 (1951), pp. 1–8.

[13] "Les problèmes que pose l'application du Calcul des Probabilités," *Collection de Logique mathématique, Ser. B*, Vol. 1 (1952), pp. 53–65.

[14] "Time-discrete stochastic processes in arbitrary sets, with applications to processes with absorbing regions and to the problem of loops in Markov chains," mimeographed, Mathematisch Centrum, Amsterdam (1952).

[15] "Carnap's foundation of probability theory," *Synthese*, Vol. 8 (1953), pp. 459–470.

[16] "Mathematical consultation for medical, biological and other research" (in Dutch), *Proc. Kon. Ned. Ak. v. Wet.*, Vol. 61, *Indag. Math.*, Vol. 14 (1952), pp. 13–18.

[17] "Nature as opponent" (in Dutch), *Statistica*, Vol. 5 (1951), pp. 149–159.

[18] "Utilité d'une distribution de probabilités, ou Distribution des probabilités d'une utilité," *Coll. d'Econométrie*, Paris, Centre National de la Recherche Scientifique, 1953.

[19] "Another form of the weak law of large numbers," *Nieuw Archief v. Wiskunde (3)*, Vol. 1 (1953), pp. 129–145.

[20] "Prediction and prophesy" (in Dutch), *Statistica*, Vol. 6 (1952), pp. 195–204.

[21] "Statistical methods based on few assumptions" (with J. Hemelrijk), *Bull. of the I.S.I.*, Vol. 34 (1954), pp. 3–31.

[22] "On arbitrary hereditary time-discrete stochastic processes, considered as stationary Markov chains, and the corresponding general form of Wald's fundamental identity" (with C. Scheffer), *Proc. Kon. Ned. Ak. v. Wet.*, Vol. A 57; *Indag. Math.*, Vol. 16 (1954), pp. 377–388.

[23] "The responsibilities of the statistician" (in Dutch), *Statistica*, Vol. 7 (1954), pp. 199–208.

[24] "Mathematical consultation in practice" (in Dutch), *Euclides*, Vol. 30 (1954), pp. 53–67.

[25] "Mathematical problems raised by the flood disaster 1953," *Proc. of the Int. Math. Congress* (Amsterdam 1954), Vol. I, pp. 218–239.

[26] "Sur les ensembles de confiance généraux et les méthodes dites non paramétriques," *Coll. sur l'Analyse Statistique*, Bruxelles 1954, pp. 73–91 (édition du Centre Belge de Recherches Mathématiques).

[27] "Chaînes de Markov dans les ensembles abstraits et applications aux processus avec régions absorbantes et au problème des boucles," *Ann. de l'Inst. Henri Poincaré XIV, fasc. III* (1955), pp. 145–199.

[28] "Ten years of mathematical statistics" (in Dutch), *Statistica*, Vol. 9 (1955), pp. 233–242.

[29] "Econometric decision problems for flood prevention," *Econometrica*, Vol. 24 (1956), pp. 276–287.

[30] "*A course in Markov chains*" (in Dutch; edited by G. Zoutendijk and J. Hirsch), mimeographed, Mathematisch Centrum, Amsterdam, 1956.

[31] "Statistical Priesthood (Savage on personal probabilities)," *Statistica Neerlandica*, Vol. 11 (1957), pp. 1–16.

[32] "From "Rekeningh in Spelen van Geluck" to decision theory" (in Dutch), Yearbook II of the University of Amsterdam (1956–57), pp. 39–50.

[33] "Statistical Priesthood II (Sir Ronald on Scientific Inference)", *Statistica Neerlandica*, Vol. 11 (1957), pp. 185–200.

[34] "Les fonctions génératrices liées à quelques tests nonparamétriques," mimeographed, Mathematisch Centrum, Amsterdam 1957.

[35] *Extrapolation of the frequency-line of high water levels at Hoek van Holland by means of selected storms*, (in Dutch; with J. Hemelrijk), Report for the Delta Commission, 1959.

[36] *The econometric decision problem concerning the prevention of floods in Holland*, (in Dutch), Report for the Delta Commission, 1959.

[37] "Itérations markoviennes dans les ensembles abstraits" (with G. Zoutendijk), *J. de Math. pure et appl.*, Vol. 38 (1959), pp. 183-200.

[38] "Sur quelques questions de la théorie mathématique du choix pondéré," *Coll. "Sur la Décision"*, Paris, mai 1959, to appear.

[39] "Prediction and Prophesy", to appear.

Papers about van Dantzig's Life and Work:

[40] J. Hemelrijk, "In memoriam Prof. Dr. D. van Dantzig," *Statistica Neerlandica*, Vol. 13 (1959), pp. 415–432 (in Dutch).

[41] J. Hemelrijk, David van Dantzig's statistical work, *Synthese*, Vol. 11 (1959), pp. 335–351.