STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

Publicatie

91

Statistical methods based on few assumptions

D. van Dantzig and J. Hemelrijk

1954

28 SESSION DE L'INSTITUT INTERNATIONAL DE STATISTIQUE
SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE
SESSIONE DELL'ISTITUTO INTERNAZIONALE DI STATISTICA

ROME, 6-12 SEPTEMBRE 1953

D. van DANTZIG and J. HEMELRIJK

STATISTICAL METHODS BASED ON FEW ASSUMPTIONS

ROME
1954

REPRINTED FROM
BULLETIN OF THE INTERNATIONAL STATISTICAL INSTITUTE
TOME XXXIV - 2nd PART

STABILIMENTO TIPOGRAFICO FAUSTO FAILLI — ROMA

STATISTICAL METHODS BASED ON FEW ASSUMPTIONS

by D. van Dantzig and J. Hemelrijk

Mathematical Centre, Amsterdam

1. — Broadly considered four stages can be distinguished in the development of mathematical statistics [1]. The first stage, beginning e.g. with John Graunt (1662) is characterized by the treatment of statistical quantities as if they were constant, as long as no obvious changes in the situation had occurred. E.g. the ratio of the yearly number of deaths and of living was estimated by Graunt as 1:32, in his summary as 1:30. Hence Sir William Petty " computed " the population of Paris, Rome, Amsterdam etc., simply by multiplying the known yearly number of deaths by 30. Graunt apparently knew, that his numbers were mean values, and had some awareness of the phenomenon of statistical variability, but not of its dependence on the numbers of observations. E. g. Graunt believed that he could draw conclusions about country-life being healthier than town-life from the fact that the ratio of the greatest to the smallest death rate during a number of years was at most 2:1 in London, but 5:1 in a (small) country parish. Using modern terminology we can say that in the first stage the distribution of a demographic quantity was characterized by *one* number, e.g. a mean ratio or a ratio of means, etc.

The second stage is characterized by the growing awareness of the phenomenon of variability. Its main historical sources were the efforts to find laws for the errors made in astronomical observations. It culminated in Laplace's discovery in 1778 that the normal law of errors results from a large number of independent elementary errors, whatever their individual " laws " (assumed to be identical) may be. Its somewhat more elementary treatment by Poisson and, in particular, its imbedding in the formalism of least squares by Gauss (1809) rapidly worked as the Mephistophelian drink : " Mit diesem Trank im Leibe siehst eine Helena in jedem Weibe " : many statisticians soon believed to find the normal distribution almost always and everywhere. In particular Adolphe Quételet (1796-1874), under direct influence of Laplace, and in his track Francis Galton (1822-1911) did much to spread knowledge of the normal distribution, which in that stage doubtless was necessary for their most important contribution, which was the introduction of methods and results, hitherto mainly used in astronomy and geodesy, into the social and biological sciences. Mathematically speaking the second stage is characterized by the description of empirical (univariate) distributions by means of *two* parameters, e.g. mean and standard deviation, and, more generally, if multivariates are considered, by means of the moments of first and second order. Hence the method of least squares, the simple and multiple correlation- and regression-analysis, the old theory of risk by Hattendorff, and in a way also the analysis of variance and covariance can be considered to belong to this stage.

---

[1] Cf. D. van Dantzig (1951).

The third stage is characterized by the discovery by Edgeworth, Kapteyn and in particular Karl Pearson of the fact that under closer examination really occurring distributions rarely prove to be normal, and that their description requires more constants, e.g. the moments of third and fourth order. This led to the system of Pearson curves, to the Gram-Charlier developments, and, more generally, to the theory of curve-fitting. Pearson's "goodness of fit" criterion χ^2 proved to be a useful tool for the judgment of the degree of fitting reached. The third stage, like the second one, had been based on the more or less explicit belief that statistical phenomena were governed by laws of general validity albeit that they showed somewhat greater complexity than just the normal law Notwithstanding the brilliant results obtained, in particular by Karl Pearson, it ended more or less in disappointment. The parameter values (and sometimes even the types of the curves) obtained by adjustment showed hardly any constancy or regularity. Moreover some other laws, partly dating from an older period, as Gompertz-Makeham's law of mortality (1825-1860), P.F. Verhulst's law of growth (1845) rediscovered by R. Pearl and L.J. Reed (1920), and known under the queer name of "logistic curve"), Pareto's law of income distribution (1896), J. C. Kapteyn's logarithmically normal law (1903) for the distributions of the dimensions of biological individuals etc. proved to fit rather badly in many cases, and to resist decisive improvement by introducing a greater number of parameters. Also the Gram-Charlier and similar developments were found to be of rather limited usefulness.

The growing uneasiness about the possiblity of mastering distributions depending on four or more parameters led the way to the fourth stage together with a renewed critical attitude towards the foundations of probability theory in the twenties (John Maynard Keynes, Richard Von Mises, Ronald A. Fisher) and thirties (Hans Reichenbach, A. Kolmogoroff, Jerzy Neyman, e.a.). Asfar as practical statistical methods are concerned this increased desire for logical rigour showed itself a.o. in the gradual replacement of asymptotic relations, which refer to indefinitely increasing numbers of observations, by exact relations, valid for restricted sample-sizes (cf. "Student's" and R.A. Fisher's "theory of small samples") in R.A. Fisher's refutation of the Bayes-Laplace theory of inverse probability and its replacement by his "maximum likelihood methods", and in Jerzy Neyman's revision, partly in common with Egon S. Pearson, of the principles of testing hypotheses. It is as an outcome of this desire for logical rigour that we see today's greatly increased interest in the class of methods, designated by various terms, as "non-parametric", "distribution-free" and "rank-invariant" methods in mathematical statistics, which we consider as being more or less characteristic for the fourth stage of this science. Hence, grosso modo we could characterize the four stages up to now by the use (in univariate distributions) of one parameter, two parameters, many parameters, and no parameters respectively[1] .

---

[1] By subdividing the history of mathematical statistics into this four stages, we, of course, do not wont to stress this point of view as the only reasonable one. Other subdivisions, or greater stress laid upon other aspects of modern statistics may be equally justified.

2. — The general form of statistical inference is as follows. Some observations, which we represent in their totality by the letter z have been made; some more observations, which we represent in their totality by the letter w will be made; it is requested to make some prediction on w, based on some assumption concerning the simultaneous probability distribution of z and w. The variables involved are thus considered as random variables or *variates*. This will be denoted by *underlining* their symbols. The same symbols, not underlined, may then be used to denote values, which these variates may assume.

One of the simplest cases occurring in practice is the one where z, called the "*evidence*", consists of n numbers $\underline{z}_1, \ldots, \underline{z}_n$ whereas w consists of N numbers $\underline{w}_1, \ldots, \underline{w}_N$, the $\underline{z}_i$ and $\underline{w}_j$ being stochastically independent [1] and all having the same probability distribution [2]. The problem is, to determine a region S in the space R_N of all possible w, such that the probability that w actually will be contained in S is at least equal to a given number $1 - \alpha$. We denote this condition by saying that w is contained in S *spr* α, where "*spr* α" is an abbreviation for the expression "*salva probabilitate* α", meaning "except for a probability *at most* equal to α". In other cases more restrictive, and often more complicated assumptions than the ones mentioned above have to be made. The number α will be called the "*unreliability threshold*". In the theory of testing hypotheses it is often called the "*level of significance*", the exact probability β ($\leqq \alpha$) of the excepted cases being called "*size*". In general β might be called the (true) "*unreliability*". In the theory of confidence regions $1 - \alpha$ is called the "*confidence coefficient*".

Empirically the assumption that z and w have a common probability distribution means that the way in which z and w actually have been or will be obtained, can, with an accuracy sufficient for practical purposes, be assimilated with a "*random choice*" of an element underlying this probability distribution. This random choice is a procedure which can only be described in empirical terms, e.g. as drawing a lot from a lottery under definite empirical conditions which we shall not try to describe here. In order to ensure that this "*probability model*" can be used, it is not strictly necessary that the combined observation of z and w itself is a repeatable phenomenon, but it suffices that either some natural cause is at work, ascertaining the requested similarity with the model, or otherwise some "*randomization procedure*" is applied. The replacement of these empirical conditions by the mathematical model of probability theory is called the "*switching on*" of the latter, whereas the "*switching off*" is performed by applying the law of large numbers together with the d'Alembert-Borel principle of neglecting sufficiently small probabilities. In particular this is done by neglecting the probability that among a large number of predictions, all made with the same unreliability threshold α, an appreciably larger fraction than α will prove to be failures, provided the "switching on" conditions always are satisfied and the predictions

---

[1] "Mutually completely independent" according to J. NEYMAN'S (1950) terminology.

[2] More generally n and/or N may also be variates (as in stochastic processes, e.g. sequential analysis). For the present we leave this generalization out of consideration.

are stochastically independent. It is irrelevant, whether the $\underline{z}$ and $\underline{w}$ always are of the same kind or nor.

If, in the special case mentioned above, n is so large that the deviation from the law of large numbers can be neglected, we have essentially to do with a prediction based on a *known* probability distribution. If this is the case for N instead of n, the prediction can be considered as (viz. is in the limit for $N \to \infty$ *spr* 0 equivalent with) a statement on the probability distribution.

A statement about a probability distribution is usually called an (in general) "*composite hypothesis*" and, if it determines the probability distribution uniquely, a "*simple hypothesis*". Whereas often the term "composite" is omitted, we prefer to drop the term "simple", i.e. to use the term "*hypothesis*" only for simple ones and to call a "composite hypothesis" a "*set (or region) of hypotheses*". In particular the set of all possible hypotheses with respect to a given situation is called the "*hypothesis space*". A set of hypotheses, stated, on the basis of a given evidence z, to contain *spr* α the unknown underlying probability distribution, is called a (safe) "*confidence set*" (or confidence region) *spr* α for this probability distribution. The adjective "safe" is added because the true unreliability β may be smaller than α.

It is of importance to remark that *some* condition, e.g. of stochastical independence and constancy of probability distributions is unavoidable. Such an assumption can be tested as a hypothesis, but only by means of other assumptions of a similar nature. Without *any* such assumption nothing at all can be done. For instance the two hypotheses

a) $\underline{z}_1, \ldots, \underline{z}_n$ are "*univalued*", i.e. each takes one unknown value *spr* 0 (creed of complete determinacy)

b) $\underline{z}_1, \ldots, \underline{z}_n$ are independent random variables, e.g. all normally distributed with positive standard deviations and unknown means (creed of complete indeterminacy)[1] are always irrefutable.

The fundamental difference between the "fourth stage methods" and the previous ones is contained in the greater liberality with which such assumptions previously were admitted. At present one prefers to admit assumptions only, which, with no more than a relatively slight degree of idealization, can be considered as being guaranteed by the empirical "switching on conditions". The set of hypotheses, singled out in this way by these conditions, is called the "*class of admissible hypotheses*", denoted by Ω. Each confidence region has to be contained in this class.

There are cases where the experimental conditions "guarantee" (with an accuracy and a certainty sufficient for practical purposes) that the class of admissible hypotheses has a finite number of dimensions only, so that every admissible hypothesis can be determined by specifying the values of a finite number of "parameters", constant for every admissible hypothesis, variable over the whole class In such cases the second or third stage methods, "*parametric methods*", can be applied.

---

[1] If we take the standard deviation to be σ, case *b*) corresponds with any $\sigma \neq 0$, whereas case a) is the special case $\sigma = 0$.

Abundant, however, are the cases where this is not so and this has led to the development of statistical methods based on as few assumptions as seemed practicable. These methods are usually called " non parametric ", " distributionfree ", etc., without precise definitions of these terms being given. Indeed, such definitions prove to be difficult to give. An interesting analysis of these terms is given in a forthcoming publication by M.G. Kendall and R.M. Sundrum, which the authors had the kindness to show to us in manuscript.

Without trying to give definitions for these notions certain types of conditions will be indicated, which are usually admitted in this type of work and which often happen to be " guaranteed " by the empirical conditions (always with the abovementioned proviso).

a) *Conditions of stochastic independence.* The class Ω of admissible hypotheses is often restricted to distributions, where all z_i (or groups of them) are independent variates. More generally this may hold for some known *functions* of the z_i instead of for the z_i themselves.

b) *Conditions of identity of distribution functions.* In particular it often happens that it is known that some of the z_i have the same distribution function (which itself is unknown), that some other ones also have the same distribution function (which may be the same as the first one or not), etc. For brevity we shall call two variates " *isomorous* " if they have the same distribution functions, so that conditions *b*) may be referred to as " *conditions of isomory* ".

c) *Conditions of continuity.* — Rather often the condition can be imposed that all z_i have continuous probability distributions.

Although further conditions are to be mentioned later, it may be remarked at this point, that a number of statistical methods, and in particular many statistical tests, have been developed, based on conditions of these three types only. Some of these will be used in later sections of this paper to illustrate the progress made in this direction during the last few years and we will return to this set of conditions in the next section.

Further it may be noted that a set of conditions of type *a*) *b*) and *c*) always is *rank invariant*, i.e. invariant under simultaneous transformations of all z_i into variables $z_i' = \varphi(z_i)$, where $\varphi(z)$ is a monotonous increasing continuous function[1]. The reverse, however, is not true; conditions like $F_1(z) < F_2(z)$, $\int F_1(z)\, dF_2(z) < \frac{1}{2}$, etc., where F_i denotes the distribution function of z_i, are rank invariant, but are not covered by the above conditions. We therefore generalize *a*), *b*) and *c*) together by introducing:

d) *Conditions of a rank invariant character.* The notion of rank invariance seems to be a rather fundamental one in this context, but it is not sufficient to

---

[1] For reasons of simplicity we here leave the possibility of discontinuities or of intervals of constancy out of consideration.

characterize the statistical methods which form the subject of this paper. It is e.g. noteworthy that Pitman's tests, one of which is treated in a later section, are based on rank invariant conditions, but are themselves not rank invariant, i.e. their result is not invariant under rank invariant transformations of the observations.

Apart from the conditions mentioned already, others are sometimes used. Without laying any claim to completeness, we give some more types of these.

e) *Algebraic relations between distribution functions.* Some algebraic relations between the distribution functions F_i of $\underline{z}_i$ $(i = 1, \ldots, n)$, e.g.

$$F_2(z) = p\,F_1(z) + (1-p)\,F_1^2(z) \qquad (0 \leqq p \leqq 1)$$

may be considered. Also conditions of identity or algebraic equalities may hold for the distribution functions of some *known* functions of the $\underline{z}_i$ instead of for the $\underline{z}_i$ themselves (e.g. for $\underline{z}_2 + a$ and $\underline{z}_1$, a being a known or unknown constant, etc.).

f) *Conditions of boundedness or symmetry.* Examples :

$$F_i(a) = 1 - F_i(b) = 0 \qquad (a < b),$$

$$F_i(-z) = 1 - F_i(z)$$

for all i or for some of them, etc.

3. — The field of all methods of such character as has been indicated in the foregoing section, is so large already, that it would be impracticable to give a complete survey of what has been done, as may be illustrated by the 72 pages of titles of papers on these subjects in the bibliography compiled by I.R. Savage (1952). On the other hand some excellent surveys of parts of the field have been given already, e.g. by H.Scheffé (1943), J. Wolfowitz (1949), P.A.P. Moran, J.W. Whitfield and H.E. Daniels (1950) and W.H. Kruskal and W.A. Wallis (1952). Therefore no attempt at completeness in any sense has been made in this paper : a rather special complex of distributionfree tests has been chosen more or less arbitrarily for its illustrative qualities. In particular a number of tests for the following hypothesis will be treated.

H_0 : *a*) the variates $\underline{z}_1, \ldots, \underline{z}_n$ are independent,

b) they are isomorous (i.e. all of them have the same distribution function),

c) their distribution function is continuous.

The class Ω of admissible hypotheses is determined by requiring *a*) and *c*) to hold, and *b*) to hold for some subsets into which the set of variates $(\underline{z}_1, \ldots, \underline{z}_n)$ can be divided. Often special subclasses of Ω are considered, in particular with regard to the power function.

The hypothesis H_0 tested implies the hypothesis H_{00} : *the simultaneous distribution function of* $\underline{z}_1, \ldots, \underline{z}_n$ *is invariant under the group G of all permutations of these variates,*

whereas Ω corresponds with the class of hypotheses ω : *the simultaneous distribution function of* $z_1, \ldots, z_n$ *is invariant under the group* K *of all such permutations of these variates which leave each of the subsets of isomorous variates invariant.* Evidently K is a subgroup of G.

Now H_0 may be tested by testing H_{00} and this is done according to the following principle (due to R.A. Fisher (1935)). If $z_1, \ldots, z_n$ are any observations of the variates $z_1, \ldots, z_n$, which may be supposed to be all different (as this is true *spr* 0), then H_{00} implies the probability of the inequalities $z_{i_1} < \ldots < z_{i_n}$ for any permutation $i_1, \ldots, i_n$ of the suffixes $1, \ldots, n$ to be the same for all permutations of these suffixes, hence $= 1/g$, $g = n!$ being the number of permutations in G. Choosing some set M, consisting of m permutations, as a critical region, its " size " is m/g, hence $\leqq \alpha$ if $m \leqq \alpha g$. The critical set M in G for testing H_{00} corresponds with a critical region in the complete sample space for testing H_0, consisting of all $\{z_1, \ldots, z_n\}$, such that the permutation $\{i_1, \ldots, i_n\}$ is contained in M if and only if $z_{i_1} < \ldots < z_{i_n}$. Then the probability that $(z_1, \ldots, z_n)$ is contained in this region is m/g, as the *conditional* probability that this is so, given $z_1, \ldots, z_n$, has the same value for all samples $\{z_1, \ldots, z_n\}$, the probability of equal values among the z_i being zero.

The condition of continuity, although convenient, is not necessary for applying this principle. If it is dropped, sets of observed values $\{z_1, \ldots, z_n\}$ which are not all unequal have to be considered too. The above principle, however, may also be formulated as follows. Let H_{00} be true and let the set of variables $\{z_1, \ldots z_n\}$ assume the values $z_1, \ldots, z_n$ in any order. We define the random permutation $i_1, \ldots, i_n$ of the numbers $1, \ldots, n$ by z_i taking the value z_{i_j}. Then all $g = n!$ possible permutations have equal probabilities. This formulation implies the one given above, but now equal values among the z_i are permitted. To give a popular picture of the principle : H_{00} implies that the values, taken by the variables $z_1, \ldots, z_n$ might be written down on n lottery tickets and then, by successively drawing these tickets at random, be assigned to the variables z_1, z_2, etc. without changing the simultaneous probability distribution of $z_1, \ldots, z_n$.

4. — Thus critical regions for testing H_{00} may easily be formed, their sizes may be computed exactly or approximately ; it also often is feasible to take into account certain subclasses of Ω as classes for which the test is meant to be especially powerful and consistency for this class may then often be proved ; but the computation of the powerfunction of these tests is very complicated and usually not much is known as yet about this very important function even for restricted classes of alternatives, except sometimes for large samples. Accordingly most head way has been made in the directions mentioned first and only during the last few years the problem of the power function has yielded to the efforts of a number of prominent statisticians, among whom W. Hoeffding has obtained the most important results.

Following the general line of the historical development a description of the tests without bothering much about their power functions will be given first, some remarks about the power functions in special cases being given afterwards.

5. — Our starting point is the *method of rank correlation*, based on the rank correlation coefficient t, which was first considered by R. Greiner (1907) and F. Esscher (1924) and which was rediscovered by M. G. Kendall (1938), who gave the theory its present form.

Consider a set $(u_1, v_1), \ldots, (u_n, v_n)$ of n pairs of arbitrary real numbers, among which at least two of the u_i (and two of the v_j) are different from each other. Arranging the numbers $u_1, \ldots, u_n$ according to increasing magnitude, assigning an arbitrary order to equal numbers, we obtain a ranking, which may contain groups of equal numbers, called ties. In this ranking each number u_i has a rank; to all numbers of a tie the arithmetical mean of the ranks of these numbers is assigned, equal numbers u_i thus having the same rank. The same procedure is applied to $v_1, \ldots, v_n$. Denoting the ranks obtained in this way by $s_1, \ldots, s_n$ and $r_1, \ldots, r_n$ respectively, we have a set $(s_1, r_1), \ldots, (s_n, r_n)$ of pairs of ranks.

From these pairs of ranks Kendall computes a quantity S by scoring

$$
\begin{aligned}
-1, \quad & \text{if} \quad (s_h - s_k)(r_h - r_k) < 0 \\
0, \quad & \text{if} \quad (s_h - s_k)(r_h - r_k) = 0 \\
+1, \quad & \text{if} \quad (s_h - s_k)(r_h - r_k) > 0,
\end{aligned}
$$

and by adding the scores for all pairs (h, k) with $h < k$. The definitions may also be given precisely in the same way with the numbers (u_i, v_i) themselves instead of their ranks and in words it may be given as follows. For every pair (h, k) $(h, k = 1, \ldots, n;\ h < k)$ $+1$ is scored if the order of magnitude of u_h and u_k is the same as that of v_h and v_k, -1 if these two pairs have opposite order and 0 if none of these two cases is fulfilled, i.e. if $u_h = u_k$ or $v_h = v_k$ or both. It is clear that the value of S only depends on the pairs of ranks $(s_1, r_1), \ldots, (s_n, r_n)$ but not on the arrangement of these pairs.

Given the set of numbers $u_1, \ldots, u_n, v_1, \ldots, v_n$ (or the set of ranks $s_1, \ldots s_n, r_1, \ldots, r_n$) there are $g = n!$ ways of forming sets of n pairs (u, v) (or (s, r)). Supposing these g sets of pairs to have equal probabilities $1/g$, the probability distribution of S may be derived. As will be seen later this supposition reduces in a number of special cases to the hypothesis H_{00} of section 3 and it will therefore also be denoted by H_{00}, although it is of a more general form.

When *no ties* are present the probability distribution of S under H_{00} may be computed directly by means of a recursion formula (cf. M.G. Kendall (1938) and (1948)); tables up to $n = 10$ are given there, a more extensive table up to $n = 40$ has been given by L. Kaarsemaker and A. Van Wijngaarden (1952)). Furthermore in this case S is asymptotically normally distributed with mean 0 and variance $n\,(n\text{-}1)\,(2n + 5)\,18$ (cfr. M.G. Kendall (1938), G.B. Dantzig (1939)).

Tests for H_{00}, developed along the lines indicated in section 3, may be used to test several hypotheses implying H_{00}. E.g. if u and v are two independently distributed random variables and $(u_1, v_1), \ldots, (u_n, v_n)$ are n independent pairs of observations of these variables, then H_{00} is satisfied and the statistic S may be used as a test statistic for this independence, large values of $|S|$ being critical values.

Moreover W. Hoeffding (1948 b) proved, that for $n \geqq 5$ not only H_{00} follows from the independence of u and v but on the other hand, if u and v have continuous joint and marginal probability distributions, then H_{00} also implies independence of u and v.

If only one of the two rows of numbers $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ is a row of observations of one or more random variables, the other row may e.g. be used to order these observations. Taking $u_i = i$ and $v_i = x_i$ $(i = 1, \ldots, n)$, where x_i denotes an observation of a random variable x_i, H.B. Mann (1945) uses S [1] to test the hypothesis H_0, that the random variables $x_1, \ldots, x_n$ are independently distributed according to the same continuous probability distribution. H_0 implies H_{00} and thus the distribution of S under H_0 is known. Defining ε_{ij} by the relation

$$P\,[x_i < x_j] = \frac{1}{2} + \varepsilon_{ij},$$

Mann proves the onesided test with large values of S critical to be consistent for alternatives satisfying

$$\lim_{n \to \infty} n^{-\frac{3}{2}} \sum_{i<j} \varepsilon_{ij} = +\infty.$$

The test may then be used as a test against (upward, or, with small values of S critical, downward) trend. This result also throws some light on the kind of alternatives for which the abovementioned test of independence is consistent. Mann gives a condition for unbiasedness of the onesided tests and discusses a class of alternatives for which the test is most powerful among all tests based on ranks. These conditions, being rather involved and not easily expressible in simple properties of the distributions of the x_i, will not be discussed here.

When there are *ties in one ranking* only (e.g. when equal values occur among the u_i but not among the v_i) not so much is known about the distribution of S under H_{00}. G.P. Sillitto (1947) tabulated the exact distribution for $n = 3, \ldots, 10$ with pairs and triplets of equal values allowed in one ranking and T.J. Terpstra (1952 *a*) proved the asymptotic normality of S under mildly restrictive conditions. Terpstra uses his result to construct a test against trend for groups of observations from a number of random variables $x_1, \ldots, x_h$ with continuous distribution functions, the hypothesis tested being, that these distribution functions are identical. Given n_j $(j = 1, \ldots, h)$ independent observations of x_j, he takes $u_1 = u_2 = \ldots = u_{n_1} = 1$, $u_{n_1+1} = \ldots = u_{n_1+n_2} = 2$, etc. and substitutes for $v_1, \ldots, v_{n_1}$ the n_1 observations of x_1, for $v_{n_1+1}, \ldots, v_{n_1+n_2}$ the n_2 observations of x_2, etc. Then S (or a linear function of S) may be used as a test statistic for the abovementioned hypothesis (which again implies H_{00}), large and small va-

---

[1] His statistic, denoted by T, is in fact a linear function of S.

lues of S being critical for an upward and a downward trend respectively in the arrangement of variables $x_1, \ldots, x_h$. This test is consistent for a iarge class of alternatives similar to the alternatives considered by Mann [1]. On the other hand Terpstra's result may also be used to generalize Mann's test against trend, with one observation of each of the random variables $x_1, \ldots, x_n$, for the case that the distribution functions of these variables are not continuous. The ties then occur in the v_i and not in the u_i.

Another test, which may be derived from Kendall's S, is the well-known test of Wilcoxon (1945) for the problem of two samples. As a matter of fact Terpstra's test against trend is a generalization of this test and reduces to Wilcoxon's test when $h = 2$, i.e. when there are two groups of observations. This test, which has been developed independently by a number of authors (cf. W. M. Kruskal and W.A. Wallis (1952) for historical details) is a test for the hypothesis, that two independent samples $x_1, \ldots, x_n$, and $y_1, \ldots, y_{n_2}$ have been taken from the same continuous distribution. Putting $u_1 = \ldots = u_{n_1} = 1$, $u_{n_1+1} = \ldots = u_{n_1+n_2} = 2$, $v_i = x_i$ $(i = 1, \ldots, n_1)$ and $v_{n_1+j} = y_j$ $(j = 1, \ldots, n_2)$, Kendall's S becomes a linear function of the test statistic used by Wilcoxon, which is the sum of the ranks of the x_i, i.e. $\sum_{i=1}^{n_1} r_i$. The hypothesis, that the two samples are taken from the same population implies H_{00} and thus S may be used to test this hypothesis. The power of Wilcoxon's test will be discussed in a later section of this paper.

For the case, when *ties* occur *in both rankings*, the mean (which is equal to 0) and the variance under H_{00} are known (M.G. Kendall (1947)), but no general theorem about the limiting distribution of S for large n seems to have been given as yet. For some special cases asymptotic normality has been proved and it seems likely that this property holds under very general conditions concerning the ties. It is e.g. likely, that in the case of Terpstra's test against trend applied to variables with discontinuous distributions S is asymptotically normal, but this has not been proved as yet. It has been proved, however, for Wilcoxon's test, when ties are present by W.H. Kruskal (1952) (cf. also J. Hemelrijk (1952)). In all these cases the test statistic S is a linear function of the ranks of the observations when arranged according to size.

Another special case, noted by Kendall (1948) p. 35, is the 2 x 2 table. When n objects, possessing or not-possessing a quality A and a quality B, are inspected and when u_i is taken to be 1 when the i^{th} object possesses the quality A and 2 otherwise, v_i taking the same values according to the presence or absence of quality B, then S^2 is proportional to the usual χ^2 of a 2 x 2 table, the marginal totals being fixed. This shows S to be normally distributed in the limit for the extreme case, when both ranking consist of a dichotomy. In a similar way S may be brought into relation with the general contingency table and with tests which may be derived from 2 x 2- and contingency tables, like the median test of Westenberg (1948) and G.W. Brown and A.M. Mood (1948) and generalizations of this test (cf. e.g. A.M. Mood (1950), G.W. Brown, and A.M.Mood (1951), J. Hemelrijk (1950 *b*), N. Blomquist (1951)).

---

[1] TERPSTRA's condition $\lambda^{-1} = o((l\,n)^{1/2})$ should be read $\lambda^{-1} = o(l^2\,n^{1/2})$.

6. — The trend tests mentioned in section 5 and Wilcoxon's test are based on linear functions of the ranks of the n observations when arranged according to size. Generalizations have been made by means of quadratic functions of the ranks. Wilcoxon's test has been generalized to a test for k samples independently by W. H. Kruskal (1952) (cf. also W. H. Kruskal and W. A. Wallis (1952), by P. J. Rijkoort (1952) and by T. J. Terpstra (1952 *b*) in two different ways. Let n_i $(i = 1 , \ldots , k ; \Sigma\, n_i = n)$ independent observations of x_i be given and let R_i be the sum of the ranks of this sample of x_i in the ranking of all n observations together, then Kruskal uses as test statistic [1]

$$\sum_{i=1}^{k} \frac{R_i^2}{n_i} ,$$

and Rijkoort uses

$$\sum_{i=1}^{k} R_i^2 .$$

Terpstra's first test coincides with that of Kruskal ; his second test is more elaborate. Apart from the sums of ranks R_i he introduces the quantities $R_{h,j}$ $(h , j = 1 , \ldots , k ; h < j)$, defined as the sum of the ranks of the h^{th} sample, computed from the pooled h^{th} and j^{th} sample, arranged according to size. His test statistic is then [1]

$$\sum_{h<j} \frac{R_{h,j}^2}{n_h\, n_j} - \frac{1}{n+1} \sum_{i=1}^{k} R_i^2 .$$

In all three cases the hypothesis H_0 tested is again, that the variables $x_1 , \ldots , x_k$ are isomorous (i.e. have the same distribution). As before H_0 implies H_{00} and the exact distribution of the test statistic as well as approximations and the limiting distribution for large (or many) samples may be derived from H_{00}. No comparison of the power functions of the three tests has been made as yet. For $k = 2$ they reduce to Wilcoxon's test.

Another general method, which strictly speaking, is not a generalization of Kendall's rank correlation method but of C. Spearman's (1904), is M. Friedmann's (1937) method of m rankings. In this case there are m rankings of equal length, the hypothesis tested being that for each of these all permutations of the ranks have equal probabilities. The variance of the column totals of the ranks is used as a test statistic for this hypothesis, large values being critical. The original theory is only applicable to a rectangular scheme of plots with exactly one observation in each plot. Lately this rather severe restriction has been partly removed by J. Durbin (1951), who generalized the method to incomplete block designs, and practically completely by A. Benard and Ph. Van Elteren (1953), who generalized it to arbitrary numbers of observations in the plots (empty plots being permitted also), subject to weak restrictions. Their method contains several

---

[1] Strictly speaking the test statistics in the original papers are linear functions of the statistics given here ; this does not change the test.

others as special cases, in particular the method of m rankings itself, Durbin's generalization, the k sample test of Kruskal and Terpstra, Wilcoxon's two sample test and the sign test. In this method also a quadratic function of the ranks in the m rankings is used as a test statistic, but its general form is too involved to be given here.

7. — In the foregoing sections the power function of the tests has not been mentioned. For most of the tests considered not much is known about the power as yet, but in some cases important progress has been made. A number of general theorems of great interest have been given by W. Hoeffding (1948 *a*) (1951) (1952), and these have been applied to several problems by other authors, e.g. by E. L. Lehmann (1951) (1953) and M. E. Terry (1952). The theorems of Hoeffding, which are too technical by nature to be given here, refer to the asymptotic distribution of statistics based on ranks under several hypotheses.

To illustrate the results in this direction, the *problem of two samples* will be discussed a little further and five tests for this problem will be compared.

Let $\underline{x}_1, \ldots, \underline{x}_{n_1}$ and $\underline{y}_1, \ldots, \underline{y}_{n_2}$ be $n = n_1 + n_2$ independently distributed random variables, one observation of each of these variables being available. The hypothesis tested is, that the probability distributions of $\underline{x}_1, \ldots, \underline{y}_{n_2}$ are identical, the alternative hypotheses being, that this is true for $\underline{x}_1, \ldots, \underline{x}_{n_1}$ and for $\underline{y}_1, \ldots, \underline{y}_{n_2}$ separately, but that the probability distributions of these two groups of variables differ in some specified way. Let us denote in general the cumulative distribution functions of the x_i by F and of the y_i by G and omit the indices i and j when this is convenient; then we have

$$H_0 : F \equiv G$$

for the hypothesis tested.

The following tests will be considered.

I. Student's test (W. S. Gosset (1908), R. A. Fisher (1926)), using the statistic

$$t_S = \frac{\bar{\underline{x}} - \bar{\underline{y}}}{\sqrt{\underline{S}_1^2 + \underline{S}_2^2}} \sqrt{\frac{n_1 n_2 (n-2)}{n}} \qquad (n = n_1 + n_2),$$

with

$$\bar{\underline{x}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \underline{x}_i , \quad \bar{\underline{y}} = \frac{1}{n_2} \sum_{j=1}^{n_2} \underline{y}_j , \quad \underline{S}_1^2 = \sum_{i=1}^{n_1} (\underline{x}_i - \bar{\underline{x}})^2 , \quad \underline{S}_2^2 = \sum_{j=1}^{n_2} (\underline{y}_j - \bar{\underline{y}})^2 .$$

II. Pitman's test (E. J. G. Pitman (1937)), based on the statistic

$$\underline{t}_P = \sum_{i=1}^{n_1} \underline{x}_i .$$

III. Wilcoxon's test (F. Wilcoxon (1945)) using

$$\underline{t}_W = \sum_{i=1}^{n_1} \underline{r}_i ,$$

where r_i is the rank of x_i ($i = 1, \ldots, n_1$), when all x_i and y_j together are arranged according to size.

IV. Terry's test (M. E. Terry (1952)). The test statistic of this test is

$$t_T = \sum_{i=1}^{n_1} E\, Z_{n,r_i},$$

where r_i is again the rank of x_i and $E\, Z_{n,r}$ [1] is the mathematical expectation of the r^{th} order statistic of a random sample of size n ($= n_1 + n_2$) from a standard normal distribution.

V. Van der Waerden's test (B. L. Van der Waerden (1952), (1953)), based on

$$t_X = \sum_{i=1}^{n_1} \Psi\left(\frac{r_i}{n+1}\right),$$

where $\Psi(q)$ denotes the q-quantile of the standard normal distribution [1].

The first of these tests is not distributionfree. However, several of the other tests have been constructed with this test in mind and their power functions have been investigated especially in comparison with the power of Student's test, which is uniformly most powerful if applied onesided for onesided alternatives implying that x and y are normally distributed with equal variances but different means.

The test statistic of the other four tests are seen to be closely related. If we denote the two samples taken together by $z_1, \ldots, z_n$, a general expression for these statistics is

$$t^* = \sum_{i=1}^{n_1} \varphi_i\,, \quad \varphi_i = \varphi\,(z_i\,, r_i)\,,$$

---

[1] Let $\psi\,(q)$ denote the q-quantile of the standard normal distribution, i. e. let.

$$\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\psi(q)} \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}\, x = q\,,$$

then

$$E\, Z_{n,r} = \frac{1}{\mathrm{B}\,(r\,,\, n - r + 1)} \int_0^1 F^{r-1}\, \psi\,(F)\,(1-F)^{n-r}\, \mathrm{d}\, F\,,$$

so that:

$$t_T = \int_0^1 \frac{dF}{F}\, \psi\,(F)\,(1-F)^n \sum_{i=1}^{n_1} \frac{1}{\mathrm{B}\,(r_i\,, n - r_i + 1)} \left(\frac{F}{1-F}\right)^{r_i}.$$

where $\varphi(z, r)$ stands for $z, r, E\underline{Z}_{n,r}$ and $\psi\left(\frac{r}{n+1}\right)$ respectively, $\underline{r}_i$ being the rank of z_i if $\underline{z}_1, \ldots, \underline{z}_n$ are arranged according to size and where the summation takes place over those i for which the corresponding $\underline{z}_i$ constitute the sample $\underline{x}_1, \ldots \underline{x}_{n_1}$. There is no special reason, except perhaps simplicity, for confining φ to one of these four functions. In general any monotonous function of z and or r, would give a useful test statistic for the problem considered.

The hypothesis H_0 tested implies again H_{00}, which may here be expressed by considering the sample $\underline{x}_1, \ldots, \underline{x}_{n_1}$ to be generated by random sampling without replacement from the values $z_1, \ldots, z_n$, found in the experiment, i.e. the values of both samples pooled. The sampling moments of $\underline{t}^*$ may thus be derived by means of the well known formulae for sampling without replacement from a finite population (this method has in fact been used by the authors of the tests in one form or another).

With

$$\mu = \frac{1}{n} \sum_{h=1}^{n} \varphi(z_h, r_h)$$

and

$$\sigma^2 = \frac{1}{n} \sum_{h=1}^{n} (\varphi(z_h, r_h) - \mu)^2,$$

this leads to

$$E(t^* \mid H_{00}; z_1, \ldots, z_n) = n_1 \mu$$

and

$$\sigma^2(\underline{t}^* \mid H_{00}; z_1, \ldots, z_n) = \frac{n_1 n_2}{n-1} \sigma^2.$$

These formulae being valid irrespective of the values of the z_i, they may also be used when ties are present in the observations, provided some convention has been adopted for determining the value of φ for these tied observations. The simplest way of allocating values of φ to tied individuals is to average in each tie the values of φ which the members of this tie would have had if they had been unequal but otherwise in the same position with respect to all z_i not belonging to this tie. This method has been used by many authors, especially when ranks are concerned; cf. e.g. M. G. Kendall (1948) and for historical references W. H. Kruskal and W. A. Wallis (1952) footnote 1, page 11. It was also proposed, in a letter to the authors, by Van der Waerden for his test and it seems to the authors to constitute an improvement (i.e. it is deemed likely that the power function is improved) on the randomization procedures, which are sometimes proposed as a mean of dealing with ties (cf. e.g. M. E. Terry (1952)). It may be remarked that the variance of t^* always decreases when untied values are replaced by their average and that the difference is usually small when there are no large ties. This means that the formulae for the variance without taking

the presence of ties into account can safely be used as a first approximation, correcting for ties only if there are large ties. The mean $E(\underline{t}^* \mid H_{00}; z_1, \ldots, z_n)$ does not depend on the presence or absence of ties. [1]

We thus find for the four tests considered the following expressions for the mean and variance of the test statistic under H_{00}.

	mean	variance	
$\underline{t}_P$	$\frac{n_1}{n} \sum_{h=1}^{n} z_h$	$\frac{n_1 n_2}{n(n-1)} \sum_{h=1}^{n} \left(z_h - \frac{1}{n} \sum_{k=1}^{n} z_k \right)^2$	(a)
$\underline{t}_W$	$\frac{1}{2} n_1 (n+1)$	$\frac{1}{12} n_1 n_2 (n+1) - \frac{1}{12} \frac{n_1 n_2}{n(n-1)} \sum_t (t-1)\, t\, (t+1)$	(b)
$\underline{t}_T$	0	$\frac{n_1 n_2}{n(n-1)} \sum_{h=1}^{n} (E\underline{Z}_{n,h})^2$	(c)
$\underline{t}_X$	0	$\frac{n_1 n_2}{n(n-1)} \sum_{h=1}^{n} \left\{ \psi\left(\frac{h}{n+1} \right) \right\}^2$	(c)

(*a*) Equal values of z_h being permitted.

(*b*) t denotes the number of individuals of a tie of $z_1, \ldots, z_n$, the Σ - sign denoting summation over all ties (cf. J. Hemelrijk (1952) and W.H. Kruskal and W. A. Wallis (1952)).

(*c*) In these expressions average values of $E\underline{Z}_{n,h}$ and of $\psi\left(\frac{h}{n+1}\right)$ being substituted in the case of ties ; the mean is not affected by this.

The asymptotic distribution of $\underline{t}^*$, under H_{00}, is a normal distribution as has been proved for $\underline{t}_P$ by A. Wald and J. Wolfowitz (1944), for $\underline{t}_W$ by H. B. Mann and D. R. Whitney (1947) and for $\underline{t}_T$ and $\underline{t}_X$ by the authors of these tests. The conditions for this asymptotic behaviour are slightly different for the different tests, but this does not seem to be essential. For $\underline{t}_P$ and $\underline{t}_T$ also other approximations of the distribution of the test statistic have been given, which are more accurate for smaller samples. The asymptotic normality when ties are present has only been proved as yet for $\underline{t}_W$ by W. H. Kruskal (1952).

Pitman's test II is clearly not rank invariant. It differs from the rank invariant tests III-V in that it is based on the *conditional* distribution of $\underline{t}_P$ for given values of the $\underline{z}_i$, whereas the tests III-V depend on the ranks only, and are therefore, if ties have probability zero, unconditional. In the case, however, when ties are present, the latter also become conditional.

---

[1] Instead of arranging the values of φ itself, one might also average the arguments of φ, when φ is defined for these average arguments. Then both the mean and the variance of t^* depend on the thies, whereas the method proposed above does, not change the mean under H_{00}.

The difference between the tests III-V is illustrated in figure 1, where the curve C represents the standard normal distribution function.

For an observation with rank r the random variable $\underline{Z}_{n,r}$ has a probability distribution, which, when represented by a mass-distribution on the z-axis and then projected on C, has a centre of gravity G (cf. fig. 1) with ordinate $\frac{r}{n+1}$ and, for $r > \frac{n+1}{2}$ with an abscissa $E\,\underline{Z}_{n,r}$, which, because of the concavity of the right hand half of C is somewhat larger than the abscissa $\psi\left(\frac{r}{n+1}\right)$ of the intersection G' of C with the horizontal line through G. The three quantities

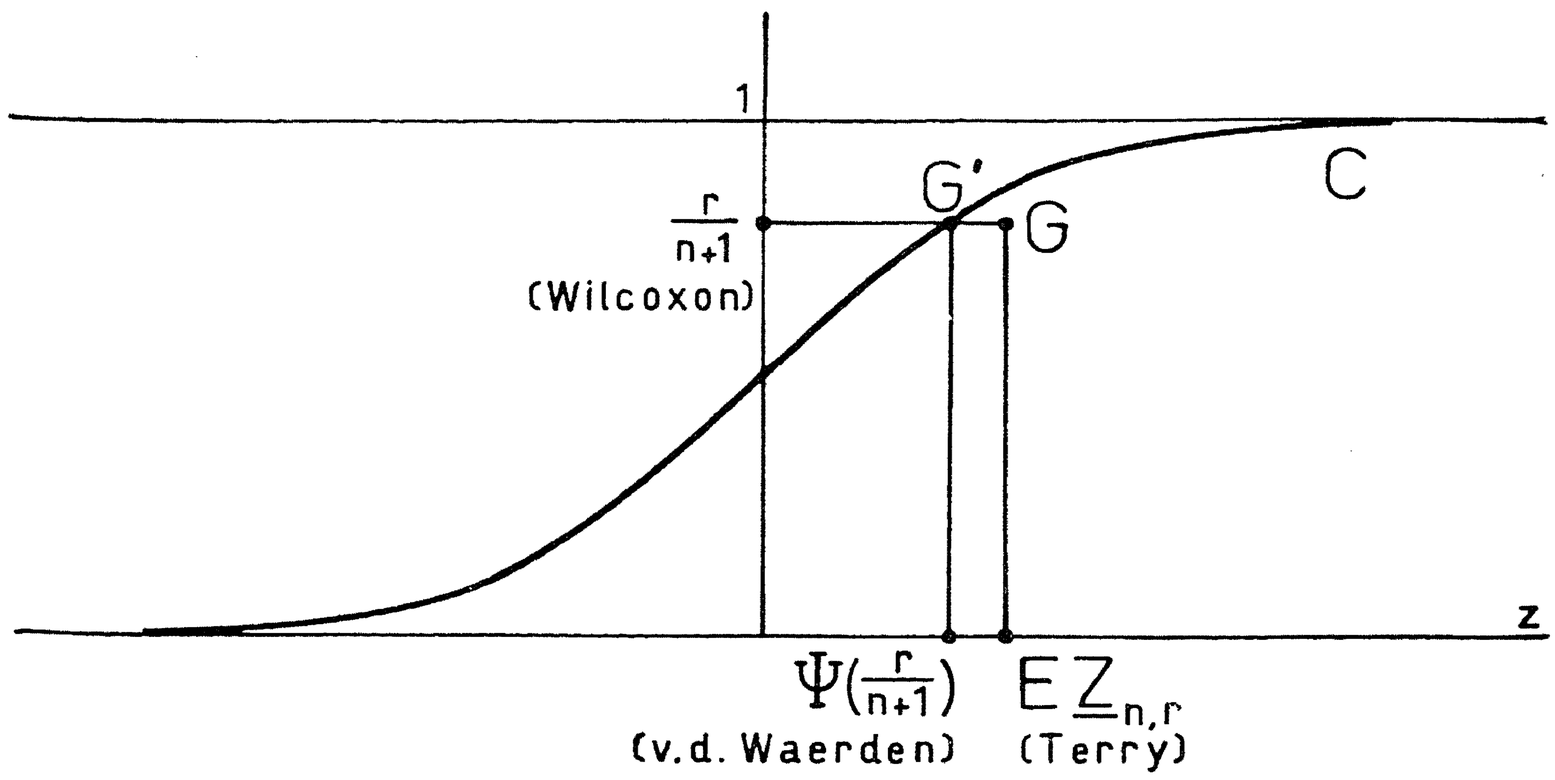

Fig. 1. — Tests III, IV and V.

$\varphi(z, r)$ used by Wilcoxon, Terry and Van der Waerden are thus clearly indicated in this diagram.

For small samples, when one wishes to use the exact probability distribution, Pitman's test has the disadvantage, caused by the use of the observations themselves, that no general expression for the probability distribution under H_{00}, not involving the values $z_1, \ldots, z_n$, can be given. This means, that the distribution has to be worked out for every case separately, which is rather an elaborate procedure. For the other tests the exact distribution under H_{00} may be calculated once and for all for small samples, as has been done by the authors of the tests and others (for references about the distribution of t_H cf. W. H. Kruskal and W. A. Wallis (1952)).

When choosing between these tests for applications the main point of interest is their power under alternative hypotheses. Unfortunately not much is known about this for small samples, but a number of results have been presented for large samples.

Pitman's test statistic t_P is, given the values $z_1, \ldots, z_n$, equivalent with Student's statistic t_S computed from the same observations, as may be seen as follows. Introducing

$$w = \frac{\frac{n}{n_1 n_2}(t_P - n_1 \bar{z})^2}{\sum_{h=1}^{n}(z_h - \bar{z})^2} \qquad \left(\bar{z} = \frac{1}{n}\sum_{h=1}^{n} z_h\right),$$

it follows that a critical region based on large values of w coincides with a symmetrical bilateral critical region for t_P, $n_1 \bar{z} = \frac{n_1}{n}\sum_{h=1}^{n} z_h$ being the expected value of t_P under H_{00}. However, computation shows that

$$\frac{w}{1-w} = \frac{t_S^2}{n-2},$$

hence w is a monotonous function of t_S^2 and Pitman's test may be described as using Student's t_S as test statistic, but deriving its distribution under H_0 by means of the equality of the probabilities of all permutations of the observations. The *difference* between Pitman's and Student's test can be seen as follows. Taking, according to Pitman, for every set of z_i a critical region with size β, the sum of these regions constitutes also a critical region with size β, under *any* common distribution function of the z_i. In particular this is the case if this distribution is normal. Then Student's critical region with size β differs from Pitman's, as the former consists (in the onesided case) of values $t_S \geqq t_S(\beta)$, where $t_S(\beta)$ is some constant depending on n_1, n_2, and β only, whereas the latter consists of values $t_S \geqq t_S(\beta, z)$, where the right hand member depends on $z \equiv \{z_1, \ldots, z_n\}$ and on n_1, n_2, and β. Considering Student's critical region for fixed values of $z_1, \ldots, z_n$, this conditional region will thus for some sets $\{z_1, \ldots, z_n\}$ be larger and for other ones be smaller in size than Pitman's conditional critical region.

A. Wald and J. Wolfowitz (1944), however, proved the asymptotic normality of Pitman's t_P under H_{00} and J. Wolfowitz (1949) states that the test is "asymptotically the same" as Student's test as a consequence of this property. The meaning of this expression is not made clear, but it probably is, that the abovementioned difference between the critical regions decreases for increasing n, or, more precisely, that if x and y have the same normal distribution, the probability of rejecting H_0 with one of the tests and not with the other one tends to zero for $n \to \infty$ (with n_1/n_2 and n_2/n_1 both bounded). For alternative hypotheses for which both tests are consistent, this property is obvious, the probability of rejecting H_0 tending to 1 for both tests. The further consequences of the "asymptotic identity" for the power of the tests under alternative hypotheses seem not yet to be completely clarified. As far as the authors are aware nothing is known about the power function of Pitman's test for small samples.

The power of Wilcoxon's test has been investigated for very small samples from normal distributions with equal variances by H. R. Van der Vaart (1950).

He found (for $n \leqq 5$) that the slope of the power function of the onesided test of Wilcoxon at the point $H_0 : F \equiv G$ differs only slightly from the corresponding slope of the power function of Student's onesided test (the ratio having values between 1 and 0,94). The same was proved to hold (again under normal alternatives with equal variances) for the difference between the second derivatives of the power functions of the twosided tests in the point H_0. For large samples the ratio of the second derivatives approaches the value $3/\pi$ ($\approx 0{,}955$)[1]. A similar result (as yet unpublished) was obtained by G. E. Noether and E. J. G. Pitman for both the onesided and twosided tests. They proved that the relative *asymptotic* efficiency, defined as the ratio of the numbers of observations necessary to give the two tests locally the same powerfunction in the neighbourhood of the point $F \equiv G$ (cf. G. E. Noether (1950)), tended also to $3/\pi$ for $n \to \infty$ with n_1/n_2 and n_2/n_1 bounded. On the other hand they found Wilcoxon's test to be far more efficientt han Student's test when F and G are not normal but skew distributions of a special type E.g. the ratio mentioned proves to be 3/2 for χ^2-distributions with 4 degrees of freedom (Student's test is not applicable without changing the probability-distri bution of t_S for small samples in this case, but asymptotically it remains valid). Another important result is due to E. L. Lehmann (1953), who proved that the test is most powerful among all rank tests in the point $H_0 : F \equiv G$ with regard to alternatives of the form

$$G \equiv p\,F + q\,F^2\,, \qquad 0 \leqq p \leqq 1\,,\ p + q = 1\,,$$

i.e. that the first derivative of the power function with regard to p in the point $p = 1$ is maximized by this test ; F denotes any continuous distribution function.

As for other asymptotic results, D. van Dantzig (1951*b*) and E. L. Lehmann (1951) proved the consistency of the test for alternatives with

$$P\,[x > y] \neq \frac{1}{2}\,.$$

The onesided test was proved by Lehmann (1951) to be unbiased against the alternatives that

$$F\,(u) > G\,(u) \qquad \text{for all } u$$

(or $<$ instead of $>$ for the other onesided test). Van der Vaart proved (as yet unpublished) that the twosided test is biased for a large class of skew distributions as alternatives, when $n_1 \neq n_2$.

Terry designed his test to be asymptotically a locally most powerful rank order test at the point $H_0 : F \equiv G$ for normal alternatives, i.e. the slope of the power function of the onesided test at this point is asymptotically a maximum for all tests based on ranks. Terry also investigated the power of his test under normal alternatives experimentally for $n_1 = n_2 = 4$ and found the difference with the power of Student's test to be considerable.

Van der Waerden proved his test to be asymptotically the same as Student's test for the case, that $\underline{y}$ is distributed normally and $n_2 \to \infty$, n_1 being bounded ;

---

[1] In a letter to the authors VAN DER VAART has announced that the corresponding ratio for the onesided tests has the limiting value $\sqrt{3/\pi}$.

the x_i only need to be distributed independently according to continuous distributions, which need not be identical.

The tests of Terry and Van der Waerden are closely related (cf. fig. 1) ; in fact their critical regions differ only very slightly. For $n_1 = n_2 = 5$ e.g. computation shows the exact onesided critical regions to coincide up to a level of significance 0,08 and above that level only incidental differences of little importance are present. An important difference between these tests and Wilcoxon's test is, that more different levels of significance are obtained by the former ones : in many cases permutations of the observations giving the same value of t_W yield different values for t_T and t_X . A consequence of the structural difference between the three test statistics is, that in t_T and t_X much greater weight is laid upon the extreme observations than in t_W . We have e.g. for $n = 20$ for the largest z the following weights.

Wilcoxon's test : $20 \Big/ \sum_{h=1}^{20} h = 0.065,$

Van der Waerden's test : $\psi\left(\frac{20}{21}\right) \Big/ \sum_{h=1}^{20} \psi\left(\frac{h}{21}\right) = 0.117,$

Terry's test : $E\,Z_{20,20} \Big/ \sum_{h=1}^{20} E\,Z_{20,h} = 0.122.$

The weight of the extreme observations is largest for Terry's test, but the difference between the weights for this test and Van der Waerden's is small compared with the difference with the weights for Wilcoxon's test.

For Pitman's test no fixed weights are attached to the observations, the z_i being used themselves.

It is not quite clear, what the consequences of these different weights on the power function will be. Van der Waerden compared the power of t_W and t_X for a number of numerical examples with small n_1 and n_2 , including normal alternatives and some distributions satisfying the relation $G(u) = F(u + d)$ for all u. He found t_X to have more power than t_W for these cases. On the other hand it follows from Lehmann's result, that for alternatives of the form $G \equiv p\,F + q\,F^2$ Wilcoxon's test has the largest power. A further investigation seems desirable. The larger number of different values assumed by t_X and t_T in comparison with t_W certainly is a point in favour of these two tests, but this could also be obtained by substituting much simple functions for $\varphi(z, r)$ in the formula for t^* and there is no special need to use the normal distribution function. On the other hand the smaller weights of the extreme observations for t_W, has the important advantage, that the influence of outlying observations, which may (but need not) be caused by mistakes of some kind and which may have a bad effect on the reliability of a statistical analysis — both when they are used and when they are eliminated, the elimination often being of an arbitrary character — is much smaller for this test than for the other two tests. Furthermore Wilcoxon's test is the easiest of the three as far as computations are concerned.

8. — It is clear that results like those mentioned in section 7 only form the beginning of a statistical theory for situations, where no assumptions about the

form of the underlying probability distribution are warranted. Especially the lack of knowledge of the power functions for small samples is irritating, because it really is the small sample theory which we are after, so that asymptotic results usually are not of great importance. Although there is nothing against the use of limiting distributions when the degree of approximation is known, or may be estimated by comparison with exact distributions — for these cases the limiting distributions are in fact very useful —, one should not forget that, asymptotically speaking nearly all statistical methods are distributionfree owing to the central limit theorem.

The possibility of making as many distributionfree tests as one wishes is indicated by the method outlined in section 7 for the problem of two samples. Pitman, Terry and Van der Waerden used functions for $\varphi(z, r)$ which connect their tests with the normal distribution. The same may be done analogously with other distributions and their principle of " normalizing ranks " may also be applied to the *generalizations* of Wilcoxon's test which have been described in earlier sections of this paper. On the other hand the influence of the tails of the underlying distributions may also easily be diminished still further than is done by the use of ranks, by choosing smaller weights for the small and large values among $z_1, \ldots, z_n$, e.g. by taking $\varphi(z, r) = \left(r - \frac{n+1}{2}\right)^{1/2}$. It seems, however, rather useless to go further in this direction of developing new tests, where tests are available already, without first developing methods to evaluate the powerfunctions for small numbers of observations, with respect to different classes of alternative hypotheses.

On the other hand it is important that one should not be compelled to make unwarranted suppositions (like normality and equality of variances, if these are not known to be fulfilled) only because of the lack of methods which do not need these suppositions. Some headway has been made in this direction especially by the recent development of trend tests and k sample tests mentioned in section 5 and 6 and by the generalization of the method of m rankings by Benard and Van Elteren. Also there are many other developments which have not been mentioned in this paper. However, large fields of statistical methods like those commanded by the classical theory of regression and analysis of variance and covariance have not yet been conquered completely by methods which do not depend on normality and homoscedacity. Several attempts in this direction have been made and incidental results have been obtained. Cf. e.g. G. W. Brown and A. M. Mood (1951), J. Hemelrijk (1950 *a*), H. Theil (1950), J. E. Walsh (1952). The method of Mood and Brown goes farthest in the direction of an analogon of the analysis of variance with more than one classification, but it is based on the median and is probably not very powerful. The method of m rankings and its generalizations seem more promising in this respect. The development of a general analogue of the classical theory in this field would be very important.

It would also be very important, if a unifying theory of ranking methods, distributionfree methods etc. were built up. The papers of Hoeffding are a starting point for this, but even the conceptional background of the methods in question is not yet clear, as may be seen from section 2 of this paper and from the forthcoming paper of M. G. Kendall and R. M. Sundrum.

As a final remark we draw attention to the fact, that the foregoing methods all deal with the testing of hypotheses and that this only is a first step in the

theory of statistical analysis. More important is the determination of confidence regions and the obtainment of predictions on future observations. To find confidence regions by means of tests of hypotheses it is desirable to have a method which enables us to test every hypothesis of the hypothesis space separately or at least to test different groups of hypotheses. Then the set of all hypotheses or groups of hypotheses which are not rejected, with a given level of significance α, on a given evidence Z, form a confidence set for the true hypothesis with confidence coefficient $1 - \alpha$. In many cases, however, only a restricted group of hypotheses — often of the character of "null-hypotheses" like H_{00} — can be tested and for the larger part of the hypothesis space no appropriate tests are available. This is a weak point of the theory. Nevertheless something can be done in the direction of determining a confidence set, even if for only one critical region R some knowledge about the power function $\beta(H)$ is available. A situation often occurring is: $\beta(H) = P[Z \in R \mid H]$, i.e. the probability that the evidence Z lies in R, if H is true, is exactly computable and $\leqq \alpha$ for some especially simple hypothesis H_0; for hypotheses near H_0, an upper estimate $\overline{\beta}(H)$ of $\beta^*(H)$ is known, and for hypotheses greatly differing from H_0, a lower estimate $\beta^*(H)$ of $\beta(H)$ can be computed. If, in that case, a $Z \in R$ is found, not only H_0, but also all H for which $\overline{\beta}^*(H) \leqq \alpha$ may be rejected, and the set of all remaining H is a safe confidence set *spr* α. If, on the other hand a Z *outside* R is found, all H with $\beta^*(H) \geqq 1 - \alpha$ can be rejected. For, if such a hypothesis were true, $P[Z \in R]$ would be $= 1 - \beta^*(H) \leqq 1 - \beta(H) \leqq \alpha$. Hence, for all $Z \in R$ all H with $\beta^*(H) < 1 - \alpha$ form a confidence set. In general, of course, both confidence sets will be too large, i.e. with more mathematical trouble it would be possible to obtain smaller sets in which we could already have confidence *spr* α. For the case of a single unknown parameter with a power function of the ordinary type we illustrate the situation in fig. 2[1]. For H near H_0 $\overline{\beta}(H) \geqq \beta(H)$, $\overline{\beta}(H_0) \leqq \alpha$; for H far from

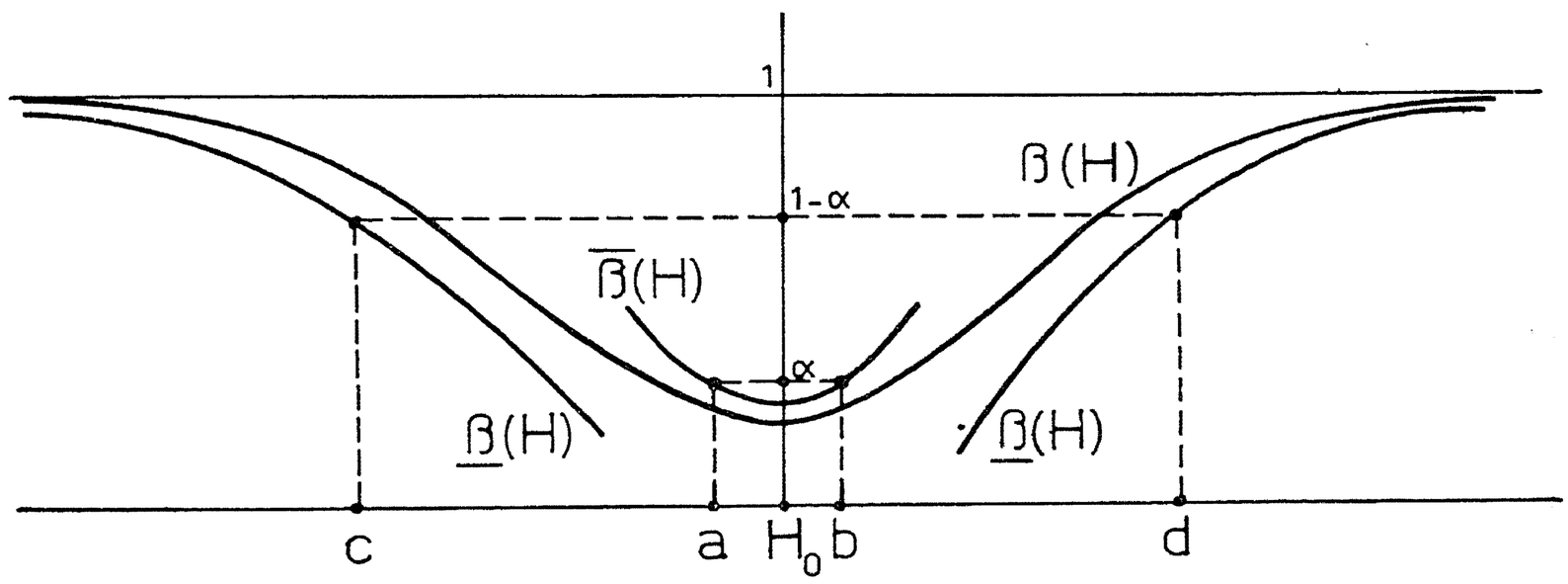

Fig. 2. Safe confidence sets.[2]

H_0 $\beta^*(H) \leqq \beta(H)$. All H *outside* the interval (a, b) form a safe confidence set if $Z \in R$ is found; all H in (c, d) form a safe confidence set if a Z outside R is found.

---

[1] For an analogous discussion, based on the critical region belonging to WILCOXON's test, cf. D. VAN DANTZIG (1951).

[2] In fig. 2 $\overline{\beta}$ and β stand for β^* and β^* respectively.

We see that from this point of view H_0 is not of particular importance, except for the fact that it may be helpful in computing $\beta^*(H)$ and $\beta(H)$. On the other hand the method described here is very primitive and only two different confidence sets are possible as the result of the experiment. Nevertheless we must insist that only by determining (safe) confidence sets within the whole class Ω of all empirically guaranteed hypotheses we can keep the unreliability threshold of the complete statistical procedure under control. This clearly indicates the fact, that the methods of statistical analysis based on few assumptions (i.e. with large classes Ω of admissible hypotheses) have only just been started on their way of development and that much remains to be done to give them more scope and power.

Appendix

Let $z_1, \ldots, z_n$ be a set of n real numbers. If all z_i are different, the rank of any one of them, say z_i, after arrangement according to increasing order is

$$r_i = \sum_{j=1}^{n} \iota(z_i - z_j), \tag{1}$$

where $\iota(z)$ denotes the "unit function"

$$\iota(z) \underline{\underline{\text{def}}} \begin{cases} 1 & \text{if} \quad z \geqq 0 \\ 0 & \text{if} \quad z < 0. \end{cases} \tag{2}$$

The mean rank $= \frac{1}{2}(n+1)$; the "reduced rank", i.e. the difference of the rank from its mean, which we shall denote by $\breve{r}_i$, is

$$\breve{r}_i = \frac{1}{2} \sum_{j=1}^{n} \operatorname{sgn}(z_i - z_j), \tag{3}$$

where the "signum function" $\operatorname{sgn} z$ is defined by

$$\operatorname{sgn} z \underline{\underline{\text{def}}} \begin{cases} 1 & \text{if} \quad z > 0 \\ 0 & \text{if} \quad z = 0 \\ -1 & \text{if} \quad z < 0. \end{cases} \tag{4}$$

If among the z_i equal values occur, a "tie" is defined as a set of *all* z_i equal to one of them. If the ranks of tied values are defined according to the mean-rank-method proposed by M.G. Kendall, the relation (3) remains valid, whether ies occur [1] or not.

For abbreviation we write

$$z_{ij} \underline{\underline{\text{def}}} sgn(z_i - z_j), \tag{5}$$

[1] It is customary to say that "no ties" are present if all ties have size 1, i. e. f the number of ties is n.

so that for arbitrary i and j

$$z_{ij} = -z_{ji}\,, \quad z_{ii} = 0. \tag{6}$$

Moreover, if $I \equiv \{1, \ldots, n\}$ is the set of all suffixes, and if S and T are arbitrary subsets of I, we put

$$\begin{cases} z_{i,T} \overset{\text{def}}{=\!=} \sum_{j \in T} z_{ij}\,, \\ z_{S,j} \overset{\text{def}}{=\!=} \sum_{i \in S} z_{ij} \end{cases} \tag{7}$$ [1]

and

$$z_{S,T} \overset{\text{def}}{=\!=} \sum_{i \in S} \sum_{j \in T} z_{ij}\,. \tag{8}$$

We have then

$$\begin{cases} z_{i,T} = -z_{T,i} \\ z_{S,T} = -z_{T,S} \\ z_{S,S} = 0\,. \end{cases} \tag{9}$$

By means of (3), (5), (7) we find

$$\breve{r}_i = \frac{1}{2} z_{i,I} = \frac{1}{2} z_{i,I'(i)} \tag{10}$$

by (10), if $I'(i)$ denotes the complement of i in I, i.e. the set of all $j \in I$ which are $\neq i$. In the same way, if $i \in T$, $z_{i,T}$ is twice the reduced rank of z_i in the ranking [2] of the elements of T, and, if $i \in T$, $z_{i,T}$ is the difference between the number of elements in T which are $< z_i$ and the number of elements in T which are $> z_i$.

If the sets S and T are disjoint (= have no elements in common), the quantity $z_{S,T}$, is twice the reduced value of the test statistic U (according to Mann and Whitney's notation) of Wilcoxon's test :

$$U_{S,T} = \frac{1}{2}(z_{S,T}\,; +n_1\,n_2)\,, \quad \breve{U}_{S,T} = \frac{1}{2} Z_{S,T}\,, \tag{11}$$

if $n_1 = |S|$ and $n_2 = |T|$ are the sizes of S and T respectively [3]. We remark that $z_{S,T}$ is an *additive* setfunction with respect to both its arguments [4].

---

[1] We remind that the symbol $j \in T$ means "j belongs to T" (j is an element of T), so that the sum in (7) is to be extended over all j belonging to T. ($\in$ is G. Peanó's « esti-symbol »).

[2] We omit further the condition "according to non decreasing order, ties being accounted for in the customary manner".

[3] Generally we denote the size (here : = number of elements) of a set S by $|S|$.

[4] For some purposes it is easier to use, instead of the *sums* $z_{S,T}$, the corresponding *means* $z^{S,T} = z_{S,T}/n_1\,n_2$, which lie always between — 1 and + 1. Then the additivity is not replaced by a property of similar simplicity. Whereas, if S and T are disjoint, $z_{S,S+T} = z_{S,T}$, we have $z^{S,S+T} = \frac{n_1}{n_1 + n_2} z^{S,T}$.

Now, let I be the sum (union) of k disjoint subsets ("samples") S_λ ($\lambda = 1, \ldots, k$) of sizes $n_\lambda = |S_\lambda|$, so that $n = \Sigma n_\lambda$. Then, the quantity T used in Terpstra's test against trend is

$$T = \sum_{\lambda > \mu} U_{S_\lambda, S_\mu} = \frac{1}{4} \left\{ \sum_{\lambda,\mu} \operatorname{sgn} (\lambda - \mu)\, z_{S_\lambda, S_\mu} + n^2 - \Sigma\, n_\lambda^2 \right\} . \tag{12}$$

The quantity

$$H = \frac{12}{k\,(k+1)} \sum_\lambda \frac{\breve{u}_\lambda^2}{n_\lambda} , \tag{13}$$

used in the k samples test, introduced independently by Terpstra and Kruskal, as well as Rijkoort's

$$\chi_R^2 = \frac{12\,(k-1)}{(n+1)\,(n^2 - \Sigma\, n_\lambda^2)} \sum_\lambda \breve{u}_\lambda^2 , \tag{14}$$

used asymptotically for the same purpose, both are quadratic forms in the quantities

$$\breve{u}_\lambda = \sum_{i \in S_\lambda} \breve{r}_i = \frac{1}{2} Z_{S_\lambda, I} . \tag{15}$$

By the additivity of $z_{R_\lambda, S}$ as a function of S, together with (9), I can be replaced by the complement of S_λ in I, making obvious that (according to Terpstra's original definition) u_λ is Wilcoxon's between R_λ and its complement.

Terpstra's second test statistic (cf. section 6) is the quadratic form

$$6 \sum_{\lambda,\mu} \frac{\breve{U}^2_{S_\lambda, S_\mu}}{n_\lambda\, n_\mu} - n\, H .$$

Finally, if I is the sum of $n\,m$ disjoint subsets ("cells") $C_{\mu\nu}$, where $\mu = 1, \ldots, m$; $\nu = i, \ldots, n$, arranged in m "rows"

$$R_\mu = \sum_\nu C_{\mu\nu}$$

and n "columns" (belonging to different "observators")

$$o_\nu = \sum_\mu C_{\mu\nu} ,$$

then the quantity used in Bernard and Van Elteren's generalized m-ranking test is a quadratic form in the quantities $\breve{v}_\nu$ each of which is the sum over rows of

the ranks occurring in the ν^{th} column, each row being ranked separately. Hence

$$\breve{v}_\nu = \frac{1}{2} \sum_\mu z_{C_{\mu\nu}} , R_\mu . \tag{16}$$

Comparing this with (15), we see that $\breve{v}_\nu$ is the sum over rows of the quantities (15), obtained by considering each row as being built up out of the cells as samples.

Generally speaking we may ask to test the hypothesis that the z_i $(i \in I)$ are independent variates, the common distribution of which is invariant under a given group G of permutations of I, within the set of admissible hypotheses stating that this invariance is required under a subgroup H of G only. We shall, however, not go into those, as yet incomplete, results in this direction, which have been obtained.

Instead, we want to make a remark on the possibility of generalizing the theory for the case where the z_i are multivariates. Let us assume that each z_i is a vector in an Euclidean f-dimensional space, its components being $z_{i,1}, \ldots, z_{i,f}$ or, generally, $z_{i,\alpha}$ $(\alpha = 1, \ldots, f)$. Then it seems a natural generalization, to consider statistics which are functions of the quantities

$$Z_{i_0}, \ldots, i_f = \operatorname{sgn} \det \begin{pmatrix} z_{i_0,1} , \ldots , z_{i_0,f} , 1 \\ \cdot \quad\quad\quad \cdot \\ \cdot \quad\quad\quad \cdot \\ \cdot \quad\quad\quad \cdot \\ z_{i_f,1} , \ldots , z_{i_f,f} \; 1 \end{pmatrix} \tag{17}$$

only, as for $f = 1$ these reduce to

$$z_{i_0}, i_1 = \operatorname{sgn} \det \begin{pmatrix} z_{i_0} , 1 \\ \\ z_i \; , 1 \end{pmatrix} = \operatorname{sgn} (z_{i_0} - z_{i_1}) , \tag{18}$$

in accordance with (5). To which transformations $z \to z^1 = \varphi(z)$ may the n vectors be subjected simultaneously without altering the quantities (17)? For $f = 1$ we know that (18) (with i and j instead of i_0 and i_1) remains invariant if $z'_i = \varphi(z_i)$, $z'_j = \varphi(z_j)$, where $\varphi(z)$ is any strictly increasing continuous function of z. For $f = 2$ the invariance of (17) requires that

$$\operatorname{sgn} \begin{vmatrix} z_{i,1} & z_{i,2} & 1 \\ z_{j,1} & z_{j,2} & 1 \\ z_{k,1} & z_{k,2} & 1 \end{vmatrix} = \operatorname{sgu} \begin{vmatrix} z'_{i,1} & z'_{i,2} & 1 \\ z'_{j,1} & z'_{j,2} & 1 \\ z'_{k,1} & z'_{k,2} & 1 \end{vmatrix} , \tag{19}$$

where $(z'_{i,1} , z'_{i,2}) = \varphi(z_{i,1} , z_{i,2})$. In particular both members of (19) must vanish simultaneously. Hence φ must transform straight lines into straight lines, i.e. the transformation must be *affine* and, moreover, orientation preserving. The

same holds true for any $f \geqq 2$. For $f = 1$ the condition of simultaneous vanishing of (18) and its transform requires only that the transformation is bi-univoque.

The group of all affine transformations, however, depends on a finite number of constants only, whereas for $f = 1$ we had the group of all orientation preserving *topological* transformations, depending on an arbitrary function Here we have an analogy with the group of conformal transformations, which in two dimensions depends on an arbitrary function, but in any larger number of dimensions (where it must transform spheres into spheres) on a finite number of constants only.

The generalization (18) has two other disadvantages. Firstly, the quantities which (apart from a factor $\{(f+1)!\}^{-1}$) generalize the Wilcoxonion, viz.

$$z_{S_0,\ldots,S_f} = \sum_{i_0 \in S_0} \cdots \sum_{i_f \in S_f} z_{i_0,\ldots,i_f} \tag{20}$$

do not lead in a natural way to a two samples test, but to an $(f+1)$ samples test only. Secondly, even for $f = 2$ and an n which is not extremely small, the actual computation of (20) becomes very cumbersome as it requires the determination of the orientations of *all* triangles which can be formed out of the n points $(z_{i,1}, \ldots, z_{i,f})$.

At first sight one might think that the natural generalization of rank-invariant statistics, say for $f = 2$, were statistics invariant under arbitrary orientation — preserving topological transformations of the plane into itself. This, however, can not be the case, as any point cloud can by such a transformation be transformed into *any* other point cloud having the same number of (different) points. Some further restriction of the transformation group is therefore unavoidable. One can, of course, admit rank-invariant transformations of each of the coordinates separately, but one might admit other simple transformations, e.g. rotations also. As yet the problem of finding the more dimensional generalization of rank invariant statistics remains open.

Introduction

(No references are given for the historical notes of the first section)

A. Benard and Ph. van Elteren (1953), *A generalization of the method of m rankings*, "Proc. Kon. Ned. Ak. van Wet *A 56*; Indagationes Mathematicae *15*, 58-869", In press.

N. Blomquist (1951), *Some tests based on dichotomization*, "Ann. Math. Stat." 22, 362-371.

G. W. Brown and A. M. Mood (1948), *Homogenity of several samples*, "The American Statistician", *2*, 22.

G. W. Brown and A. M. Mood (1951), *On median tests for linear hypotheses*, Proc. Second Berkeley Symp., Univ. of Calif. Press, Berkeley and Los Angeles, 159-166.

D. van Dantzig (1951*a*), *De natuur als tegenopeler*, « Statistica » 5, 149-159.

D. van Dantzig (1951*b*), *On the consistency and the power of Wilcoxon's two sample test*, "Ned. Akad. v. Wet., Proc." A *54*, 'Indagationes Mathematicae" *13*, 1-8.

G. B. Dantzig (1939), *On a class of distributions that approach the normal distribution function*, "Ann. Math. Stat." *10*, 247-253.

J. DURBIN (1951), *Incomplete blocks in ranking experiments*, "British Journal of Psychology", *4*, 85-90.

F. ESSCHER (1924), *On a method of determining correlation from the ranks of variates*, "Skand. Akt. 7", 201-219.

R. A. FISCHER (1926), *Applications of Student's distribution*, "Metron" *5*, 90-104.

R. A. FISHER (1935), *The design of experiments*, Oliver and Boyd, London and Edinburgh.

M. FRIEDMAN (1937), *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, "J. Am. Stat. Ass." *32*, 675-699.

W. S. GOSSET (1908), *On the probable error of a mean*, "Biometrika" *6*, 1.

R. GREINER (1909), *Ueber das Fehlersystem der Kollektivmasslehre*, "Z. für Math und Phys." *57*, 121, 225, 337.

J. HEMELRIJK (1950*a*), *A family of parameterfree tests for symmetry with respect to a given point*, I, II, "Proc. Kon. Ned. Ak. v. Wet." *53*, 945-955 and 1186-1198, "Indagationes Mathematicae" *12*, 340-350 and 419-431.

J. HEMELRIJK (1950*b*), *Symmetrietoetsen en andere toepassingen van de theorie van Neyman en Pearson*, Thesis, Amsterdam.

J. HEMELRIJK (1952), *Note on Wilcoxon's two sample test when ties are present*, "Ann. Math. Stat." *23*, 133-135.

W. HOEFFDING (1948*a*), *A class of statistics with asymptotically normal distribution*, "Ann. Math. Stat." *19*, 293-325.

W. HOEFFDING (1948*b*), *A non-parametric test of independence*, "Ann. Math. Stat." *19*, 546-557.

W. HOEFFDING (1951), "*Optimum*" *non-parametric tests*, Proc. Second Berkeley Symp. on Math. Stat. and Prob., Univ. of Calif. Press, Berkeley and Los Angeles, 83-92.

W. HOEFFDING (1952), *The large sample power of tests based on permutations of observations*, "Ann. Math. Stat." *23*, 169-191.

L. KAARSEMAKER and A. VAN WIJNGAARDEN (1952), *Tables for use in rank correlation*, Report R 73 Computation Department Mathematical Centre, Amsterdam.

M. G. KENDALL (1938), *A new measure of rank correlation*, "Biometrika" *30*, 81-93,

M. G. KENDALL (1947), *The variance of τ when both rankings contain ties*, "Biometrika" *34*, 297-298.

M. G. KENDALL and R. M. SUNDRUM, *Distributionfree methods and order properties*, In press.

W. H. KRUSKAL (1952), *A non-parametric test for the several sample problem*, "Ann., Math. Stat. *23*, 525-539.

W. H. KRUSKAL and W. A. WALLIS (1952), *Use of ranks in one-criterion analysis of variance*, "Journ. Am. Stat. Ass." *47*, 583-621.

E. L. LEHMANN (1951), *Consistency and unbiasedness of certain non-parametric tests*, "Ann. Math. Stat." *22*, 165-179.

E. L. LEHMANN (1953), *The power of rank tests*, "Ann. Math. Stat." *24*, 23-43.

H. B. MANN (1945), *Non-parametric tests against trend*, "Econometrica" *13*, 245-259

H. B. MANN and D. R. WHITNEY (1947), *On a test of whether one of two random variables is stochastically larger than the other*, "Ann. Math. Stat." *18*, 50-60.

A. M. MOOD (1950), *Introduction to the theory of statistics*, Mc Graw-Hill, New York, Toronto, London.

P. A. P. MORAN, J. W. WHITFIELD and H. E. DANIELS (1950), *Symposium on ranking methods*, "J. R. Stat. Soc." B *12*, 153-191.

J. NEYMAN (1950), *First course in probability and statistics*, Holt, New York.

G. E. NOETHER (1950), *Asymptotic properties of the wald-Wolfowitz test of randomness*, "Ann. Math. Stat." *21*, 231-246.

E. J. G. PITMAN (1937), *Significance tests which may be applied to samples from any populations*, I. "J. Roy. Stat. Soc. Suppl.," 4, 119-129.

P. J. RIIKOORT (1952), *A generalization of Wilcoxon's test*, "Proc. Kon. Ned. Ak v. Wet." *A 55*, Indagationes Mathematicae 14, 394-404.

I. R. SAVAGE (1952), *Bibliography of non-parametric statistics and related topics*, "Nat. Bur. Stand. Rep.", nr. 1828.

H. SCHEFFÉ (1943), *Statistical inference in the non-parametric case*, "Ann. Math. Stat." *14*, 305-332.

G. P. SILLITTO (1947), *The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties*, "Biometrika" *34*, 36-40.

C. SPEARMAN (1904), *The proof and measurement of association between two things*, "Am. Journ. Psych." *15*, 88.

T. J. TERPSTRA (1952*a*), *The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking*, "Proc. Kon. Ned. Ak. v. Wet." *A 55*, "Indagationes Mathematicae" *14*, 327-333.

T. J. TERPSTRA (1952*b*), *A non-parametric k sample test and its connection with the H-test*, Report *S* 92 (*VP* 2) of the Statistical Department of the Mathematical Centre, Amsterdam.

M. E. TERRY (1952), *Some rank order tests, which are most powerful against specific parametric alternatives*, "Ann. Math. Stat." *23*, 346-366.

H. THEIL (1950), *A rank invariant method of linear and polynomial regression analysis*, I, II and III, "Proc. Kon. Ned. Ak. v. Wet." *53*, 386-392, 521-525 and 1397-1412, "Indagationes Mathematicae" *12*, 85-91, 172-177 and 467-482.

H. R. VAN DER VAART (1950), *Some remarks on the power function of Wilcoxon's test for the problem of two samples*, I and II, "Proc. Kon. Ned. Ak. v. Wet." *53*, 494-506 and 507-520, "Indagationes Mathematicae" *12*, 146-158 and 159-172.

B. L. VAN DER WAERDEN (1952), *Order tests for the two-sample problem and their power*, "Proc. Kon. Ned. Ak. v. Wet. *A 55*, "Indagationes Mathematicae" *14*, 453-458. Corrigenda ibid. (1953) p. 80.

B. L. VAN DER WAERDEN (1953), *Ein neuer Test für das Problem der zwei Stichproben*, "Mathem. Annalen", *126*, 93-107.

A. WALD and J. WOLFOWITZ (1944), *Statistical tests based on permutations of the observations*, "Ann. Math. Stat." *15*, 358-372.

J. E. WALSH (1952), *Some non-parametric tests for Student's hypothesis in experimental designs*, "J. Am. Stat. Ass." *47*, 401-415.

J. WESTENBERG (1948), *Significance test for median and interquartile range in samples from continuous populations of any form*, "Proc. Kon. Ned. Ak. v. Wet." *51*, 252-261.

F. WILCOXON (1945), *Individual comparisons by ranking methods*, "Biometrics" *1* 80-82.

J. WOLFOWITZ (1949), *Non-parametric statistical inference*, Proc. of the Berkeley Symp. on Math. Stat. and Prob., Univ. of Calif. Press, Berkeley and Los Angeles, 93-113.

Résumé

Cette communication est un exposé de quelques uns des derniers résultats et montre le développement dans le domaine des méthodes statistiques ayant pour base quelques hypothèses, méthodes généralement désignées par « distribution libre », « non-paramétrique », « ordre inchangé ».

L'on indique spécialement un ensemble de méthodes étroitement liées à la méthode de la corrélation, en y comprenant aussi des « tests for trend » des tests d'échantillon k et des généralisations de la méthode m « rankings ». Deux tests d'échantillons (Students, Pitman, Wilcoxon, Terry, Van der Waerden) sont examinés plus largement ; sont discutés aussi les théorèmes concernant leurs fonctions, théorèmes dûs à plusieurs auteurs. L'importance des enquêtes concernant la portée des échantillons moindres est mise en évidence aussi bien en ce qui concerne l'opportunité d'une généralisation des méthodes employées, par analogie à l'analyse des variations comportant un plus grand nombre de classifications, non basées sur la normalité des distributions premières.

A l'appendice se trouve un résumé des groupes des méthodes exposées.