Three Essays on the Contributions of Wassily Hoeffding to Statistics and Probability

The Impact of Wassily Hoeffding's Research on Nonparametrics
by Pranab Kumar Sen

The Impact of Wassily Hoeffding's Work on Sequential Analysis
by Gordon Simons

Wassily Hoeffding's Work in the Sixties
by Kebus Oosterhoff & Willem van Zwet

February 1994

Memor Series #2510

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina
Three Essays on the Contributions of Wassily Hoeffding to Statistics and Probability

The Impact of Wassily Hoeffding's Research on Nonparametrics
by Pranab Kumar Sen

The Impact of Wassily Hoeffding's Work on Sequential Analysis
by Gordon Simons

Wassily Hoeffding's Work in the Sixties
by Kobus Oosterhoff & Willem van Zwet

The Impact of Wassily Hoeffding’s Research on Nonparametrics

Pranab K. Sen

1. Introduction

Wassily Hoeffding earned his Ph.D. degree in mathematics from the Berlin University in 1940 for a dissertation in correlation theory which dealt with some aspects of bivariate probability distributions that are invariant under monotone transformations of the marginals. This dissertation was primarily devoted to some (descriptive) studies of certain measures of rank correlations. With the impending Second World War, for Wassily, living in Berlin in the early forties was not that comfortable. Nevertheless, he managed to advance his basic research on nonparametric correlation theory. It was only after his eventual migration to the United States (in the Fall of 1946) that he started to appreciate the full depth of probability theory and statistics (during his sojourn at the Columbia University, New York), and most of his pioneering work emerged during his longtime residence at Chapel Hill (1947–1991). He felt that “...probability and statistics were very poorly represented in Berlin at that time (1936–45) ...”. Notwithstanding this, his early work on correlation theory was not just a landmark in nonparametrics; it also endowed him with a career-long zeal and affection for the pursuit of the most fundamental research in mathematical statistics, probability theory, numerical analysis and a variety of other related areas. In this respect, nonparametrics was indisputedly the “jewel in the crown” of Wassily’s creativity and ingenuity in research. Wassily Hoeffding indeed played a seminal role in stimulating basic research in a broad domain of mathematical statistics and probability theory, and his “collected work” in this volume reflects the genuine depth and immense breadth of his research contributions. In this article, I shall mainly confine myself to describing the profound impact of his work in the general area of nonparametrics, with occasional detours to some other related areas.

1University of North Carolina at Chapel Hill
Prior to World War II, the developments in nonparametric statistical inference were somewhat piecemeal. In fact, the terminology "nonparametrics" (or nonparametrics methods) was coined later on. Further, although there were some developments on measures of association for ranked and/or ordinal data, and some goodness-of-fit tests, their theoretical foundations were yet to be laid down properly. There were also some randomization/permutation tests which were not "nonparametric" in a conventional sense. There were only a few simple tests of significance for some simple statistical hypotheses where, under the null hypotheses, the test statistics have distributions independent of the underlying distributions of sample observations. This is why they were referred to as distribution-free tests and the related methodology as distribution-free methods. These distribution-free methods are the precursors of the modern nonparametrics, and Wassily Hoeffding played a very significant role in this evolution of nonparametrics.

The role of invariance, sufficiency, completeness and ancillarity in statistical inference (both estimation and testing theory) was not properly explored (at that time), and asymptotic considerations were not particularly appealing to statisticians. Not much was known about the power properties of the so-called distributions-free tests (even for large samples), and very little about genuine nonparametric estimation theory in a rational framework. The ideas of S(α)-structure tests (now identified as tests of Neyman structure), permutational central limit theorems and local optimality of nonparametric tests, percolated a few years later on; also yet to occur was the evolution of the important concept of "asymptotic efficiency", sparked by the formulation of the Pitman alternatives and the Pitman measure of ARE (asymptotic relative efficiency). The role of large deviation probabilities in nonparametrics was explored at an even later stage. In this context, Wassily Hoeffding's research led to many original and outstanding contributions to all these vital areas of nonparametrics, and it constitutes a major part of Wassily's collected work.

We shall describe his major contributions and their profound impacts under the following broad sectors:


3. Dependent Central Limit Theorem: Introduction to m-Dependence.

4. Permutational Central Limit Theorem: Ramifications and Refinements.

5. Optimal Nonparametric Tests: Concept of LMPR.


9. Large Deviational Theorems: A Landmark in Statistical Inference.


In research developments in most areas of statistics and probability theory, mathematical abstractions have a tendency to invade statistical ingenuity and concepts to a great extent. In this respect, the contributions made by Wassily Hoeffding are characterized by their elegance and lack of unnecessary abstraction. This facilitated adaptation of his work by researchers to many other areas of statistics and probability theory. It is no surprise, therefore, that some of his contributions are among those most cited in the literature.

2. Measures of Association and Dependence

The problem of correlation (and more generally, association and dependence) arising in multivariate situations can be regarded as that of determining those properties of multivariate distributions which depend on the relationships of the variables to each other (and not just on the marginal distributions). Such multivariate distributions may relate to quantitative characters (possibly discrete in nature) or to qualitative characters (such as arise in common contingency tables). They may also relate to a mixed model admitting both quantitative and qualitative variables.

The conventional measure of association based on the product moment correlation coefficient has its genesis in the family of multivariate normal distributions whose association patterns are completely characterized by the associated covariance matrix. Such a family does not include the discrete random variables nor the categorical models in a broad setup. Although, for a bivariate normal distribution, lack of correlation implies independence of the two variates, in a general nonnormal setup (where the linearity of regression may not be taken for granted), zero correlation and independence need not be isomorphic. Moreover, in a multinormal framework, because of linearity of regressions, invariance under linear transformations (viz. change of scale of measurements) holds for the correlation coefficient, but the rationality of this particular invariance property remains unclear for a general multivariate model. This prompted Wassily to consider a general class of transformations from the original variables \((X, Y, \ldots)\) to the new variables
\((X^*, Y^*, \ldots)\) defined by

\[
X^* = L_1(X), \quad Y^* = L_2(Y), \quad \ldots, \tag{2.1}
\]

where the functions \((L_i)\) are all one-to-one (i.e. single valued and uniquely invertible) and they are strictly monotone (so that the ordering is preserved). He referred to these as changes of scale and he formulated all the properties of a multivariate distribution which pertain to the topic of correlation into two broad classes, depending on whether or not they are invariant to such arbitrary changes of scale.

Suppose the joint distribution function (d.f.) of \((X, Y)\) is denoted by \(F(x, y), (x, y) \in \mathbb{R}^2\), and let \(F_1(x) = F(x, \infty)\) and \(F_2(y) = F(\infty, y)\) be the two marginal d.f.'s (of \(X\) and \(Y\) respectively). Then, we have (assuming the existence of the integrals),

\[
EX^* = \int_{\mathbb{R}} L_1(x) \, dF_1(x), \quad EY^* = \int_{\mathbb{R}} L_2(y) \, dF_2(y) \tag{2.2}
\]

and

\[
\text{Cov}(X^*, Y^*) = E(X^*Y^*) - E(X^*)E(Y^*)
= \iint_{\mathbb{R}^2} L_1(x)L_2(x) \, d[F(x, y) - F_1(x)F_2(y)], \tag{2.3}
\]

and, by partial integration, we have

\[
\text{Cov}(X^*, Y^*) = \iint_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)] \, dL_1(x)dL_2(y). \tag{2.4}
\]

In the literature, (2.4) is known as the Höffding Identity. Under independence of \(X\) and \(Y\), \(F(x, y) = F_1(x)F_2(y)\) almost everywhere \((x, y)\), so that (2.4) is equal to zero. If we define the positive (or negative) dependence (association) of \(X, Y\) by \(F(x, y) - F_1(x)F_2(y)\) being nonnegative (or nonpositive) everywhere, it readily follows that for all one-to-one \(L_1\) and \(L_2\), the nature of association remains the same. In this context, it is not even necessary that \(L_1\) and \(L_2\) are uniquely invertible; it suffices to assume that they are monotone nondecreasing (or nonincreasing). On the other hand, the nature of \(L_1\) and \(L_2\) can greatly influence the degree of dependence in (2.4). For example, consider the probability integral transformations \(L_i(t) = F_i(t), t \in \mathbb{R}, \ i = 1, 2\). Then (2.4) reduces to the Spearman grade covariance function

\[
\iint_{\mathbb{R}^2} [F(x, y) - F_1(x)F_2(y)] \, dF_1(x)dF_2(y), \tag{2.5}
\]
which is the average magnitude of the discrepancy function $\Delta(x, y) = F(x, y) - F_1(x)F_2(y)$ with respect to the two marginal d.f.'s. The original (Pearsonian) product-moment covariance is the counterpart with respect to the Lebesgue measure on $\mathbb{R}^2$. Similarly, if we let $L_i$ have unit mass at the point $a_i$ (i.e. $L_i(t) = 0$ for $t < a_i$ and 1 for $t \geq a_i$), $i = 1, 2$, then (2.4) reduces to $\Delta(a_1, a_2)$ which is known as a Quadrant Measure of Association.

It is clear how the basic Hoeffding Identity in (2.4) paved the way for the formulation and study of a class of measures of association and dependence, and a very systematic account of this work is due to Hoeffding [19, 20, 21]. Lehmann [40] contains some further development of these concepts. The interesting feature is that (2.5), (2.4) and all other related functions are essentially functionals of the d.f. $F$ (defined on $\mathbb{R}^2$). This may be what led Wassily to formulate the most general theory of estimable parameters or regular functionals, and that is perhaps one of the milestones in the theory of statistics. We discuss that in the next section.


In classical parametric estimation theory, the form of the underlying distribution is tacitly assumed to be given, and the unknown algebraic constants appearing in this functional form are regarded as statistical parameters. Choice of a suitable loss function and an estimation criterion having some optimal (or at least, desirable) properties constitute the major statistical task in this theory.

In the nonparametric case, the situation is quite different. To start with, the functional form of the underlying probability law may not be known precisely, so that statistical parameters are to be defined and interpreted in a different way. Secondly, since the unknown distribution is allowed to be a member of a general class, any optimality or desirability properties to be ascribed have to be with respect to this class of distributions.

Such considerations generally lead to estimators which are symmetric functions of the sample observations and are unbiased. Although justifications for the use of symmetric and unbiased estimators were partially made earlier by Halmos [17] and others, their treatment lacked the complete generality of the subject matter as treated in the basic paper [22] of Hoeffding. Statistical parameters are expressed as functionals of the underlying distribution functions, and are therefore termed regular functionals or estimable parameters. The estimators of such regular functionals may not generally be linear in the sample observations, and as a result, their distribution theory may require much more delicate treatment. Hoeffding's $U$-statistics are
unbiased and symmetric functions of the sample observations which enjoy some optimal properties among the class of estimators which are unbiased in a nonparametric sense.

A year earlier, von Mises considered a class of differentiable statistical functions, now referred to as $V$-statistics. These $U$- and $V$-statistics are indeed very close to each other (although the $V$-statistics are generally not unbiased). Whereas the asymptotic distribution theory of $V$-statistics rests on some intricate differentiability properties of statistical functionals and the sample or empirical distribution function (which were not properly known or established at that time), Hoeffding's treatment of $U$-statistics rests on an elegant projection result that goes far beyond the domain of $U$-statistics and avoids these differentiability prerequisites to a greater extent. (This remark should not, however, undermine the impact of von Mises' [61] work which led to the foundation of asymptotic theory of statistical functionals: see Fernholz [9] for a more detailed account.) It was no surprise that, almost forty years later when Norman Johnson and Sam Kotz were planning to bring out a volume "Breakthrough in Statistics: 1890–1989", Hoeffding's $U$-statistics paper was included in this set. I can't check the temptation of summarizing some of the comments I wrote down ([58]) as an introduction to this article in the aforesaid volume.

To appreciate fully the novelty of the Hoeffding [22] approach and its impact on mathematical statistics and probability theory, let us focus on the very definition of $U$-statistics as given in this paper [22]. For a kernel of degree 2 or more, although the $U$-statistic is an unbiased and symmetric estimator, it does not have independent summands, and that creates some complications in the study of its distributional and moment properties. The moment problem is less complicated than the other one, and Hoeffding employed some clever combinatorial arguments to provide exact expressions for the variance-covariance in a general setup.

The more difficult problem was to incorporate the standard central limit theorem for the study of the asymptotic normality of the standardized version of $U$-statistics. In the process of taking appropriate conditional expectations (in order to derive a neat expression for the variance of a $U$-statistic), he managed to obtain a projection of a $U$-statistic as a sum of independent random variables plus a remainder term of stochastically smaller order of magnitude, and this opened up a new and unified approach to studying the asymptotic normality and other properties of nonlinear estimators.

This was by far the most significant contribution of this outstanding paper. The very nonlinear nature of $U$-statistics (as well as other nonparametric ones) may not invariably meet the regularity conditions pertaining to their asymptotic normality. Indeed, in the context of goodness-of-fit problems, under the null hypotheses, the limiting distributions may not be typically
asymptotically normal. In a subsequent paper [24], Hoeffding managed to exhibit this very neatly with another U-statistic which is not asymptotically normal. This feature has also been explained thoroughly in his basic paper on U-statistics where, depending on the nature of the conditional expectations of the kernel given a subset of its arguments, he was able to classify the estimable parameter (say, \( \theta(F) \)) as stationary of order \( d \), if this \( d \)th order conditional expectation is nonstochastic, where \( d \) may range between 0 and the degree of the kernel. He then showed that it is only when the parameter \( \theta(F) \) is stationary of order zero that asymptotic normality may typically hold under very mild regularity conditions. When \( d = 1 \), one usually has a representation (in law) of the standardized form \( n[U_n - \theta(F)] \) as \( \sum_{j \geq 0} \lambda_j(Z_j^2 - 1) \), where \( Z_j \) are independent standard normal variables and the \( \lambda_j \) are the eigenvalues corresponding to (complete) orthonormal functions relating to the second-order kernel of the U-statistic \( U_n \).

Actually, this basic paper of Hoeffding contains all the ingredients for a far deeper result that incorporates the orthogonal partition of U-statistics and yields the penultimate step for the asymptotic behavior of U-statistics depending on the order of stationarity. In response to a query (Sen [54]), Hoeffding [31] showed that for a symmetric kernel of degree \( m (\geq 1) \), one has for every \( n \geq m \),

\[
U_n = \theta(F) + \binom{m}{1} U_{n,1} + \cdots + \binom{m}{m} U_{n,m},
\]

where the \( U_{n,h} \) are themselves U-statistics (of order \( h \)), such that they are pairwise uncorrelated, \( EU_n = 0 \) and \( E(U_{n,h}^2) = O(N^{-h}) \), for \( h = 1, \ldots, m \).

In fact, for each \( h (= 1, \ldots, m) \), \( U_{n,h} \) is stationary of order \( h - 1 \), and the normalized forms \( n^{h/2}U_{n,h} \) have asymptotically nondegenerate distributions which are independent for different \( h \). The representation in (3.1) is referred to in the literature as the Hoeffding (or \( H^- \)) decomposition of U-statistics. This decomposition has been made even more popular by van Zwet [60] who, amongst other things, extended the result to a more general class of symmetric statistics; a more detailed account of this is given in the accompanying review article of Oosterhoff and van Zwet in this volume. In equation (3.1), \( mU_{n,1} \) is termed the Hoeffding (or \( H^- \)) projection of the U-statistic, and is clearly a linear statistic which is readily amenable to asymptotic analysis under the usual regularity conditions for independent summands.

The \( H^- \)-projection and \( H^- \)-decomposition have been landmarks in the area of nonparametrics, and it is not surprising to see that in this field, in the 1950s and 1960s, the literature was flooded with estimators and test statistics that were expressible as U-statistics or their natural generalizations for which similar projection and decomposition results work out neatly. Estimable parameters for more than one distribution function led to the development of so-called generalized U-statistics; for a discussion of the
$H$-decomposition of such statistics, see Chapter 3 of Sen [56]. The treatment of Hoeffding [24] was by no means limited to identically distributed random variables: it covered the case of vector-valued $U$-statistics as well.

Another important observation: referred to (3.1), for each $h = 1, \ldots, m$,\break \{$(\cdot)^h) U_{n,h}, \ n > h$\} is a (zero-mean) forward martingale sequence, and such a martingale property holds even when the underlying random elements are not necessarily identically distributed. In the case of independent and identically distributed random variables (and vectors), of course, one can also claim that for each $h = 1, \ldots, m$, \{\(U_{n,h}, \ n \geq h\)\} is a reversed martingale sequence. This was observed for the first time by Berk [3], although I have the feeling that Hoeffding was aware of it a while earlier. Back in the late 1940s and early 1950s, martingales and reversed martingales were not so popular with mathematical statisticians and perhaps he did not want to put his results in such abstract coatings. However, nearly, a quarter-century later on, they proved to be very useful tools for the study of some deeper results for $U$-statistics covering both weak and strong invariance principles for (generalized) $U$-statistics, and some of these are reported in Chapter 3 of Sen [56]. Two recent monographs on $U$-statistics by Korolouk and Borovskikh [36] and Lee [38] are useful references for exploring the depths of (the theory and applications of) $U$-statistics and at every stage, one can feel the impact of the basic contributions of Wassily Hoeffding in this fertile area of research. There are more than four hundred published articles dealing with Hoeffding's $U$-statistics in some way or other.

4. Whither Independence in Central Limit Theorems?

Sparked by the $H$-projection of a possibly nonlinear statistic into a linear one, it was natural to ask to what extent independence of the summands would be necessary for adaptation of the classical central limit theorems for linear statistics. In time-series analysis and other fields of applications too, random elements at different points in time may have some serial dependence which may become less and less significant as the distance between the two points becomes large. Another notable application is that of sampling from a finite population (without replacement), where the sampled units are not stochastically independent although they exhibit some symmetric dependence. The Wald–Wolfowitz [62] permutational central limit theorem made it possible to use asymptotic normality results in the latter context, while much remained to be done with respect to the former situation. In both cases, Hoeffding contributed richly.

Hoeffding and Robbins [23] were both at the University of North Carolina at Chapel Hill, and together they worked out neatly a somewhat different
decomposition for an m-dependent sequence to establish its asymptotic normality property. In a certain sense, the ideas of a mixing property germinated from this fundamental paper, relaxing the need for independence for asymptotic normality results. The current literature is flooded with the notions of phi-mixing, star-mixing, regular-mixing and strong-mixing sequences which are attracted to asymptotically normal laws, and yet the simplicity and usefulness of the Hoeffding–Robbins paper remain intact. Also, most of the extensions of the asymptotic normality for U-statistics for weakly dependent sequences of random elements are based on this fundamental work in one way or another.

Simple random sampling without replacement and distribution theory of the usual nonparametric test statistics under suitable hypotheses of invariance share a common feature: permutational invariance under suitable groups of transformations which map the sample space onto itself. Such permutational invariance properties yield some sort of symmetric dependence among the observations which may not be strictly independent in a statistical sense. Asymptotic normality results pertaining to such symmetric dependent cases are generally referred to as permutational central limit theorems (PCLT). In nonparametrics, all permutation (or randomization) tests are based on such permutation principles (explicitly or implicitly), and in dealing with their large-sample properties, it becomes necessary to appeal to PCLT’s under appropriate regularity conditions. While Wald and Wolfowitz [62] incorporated a “method of moments” type approach to derive the PCLT under asymptotic convergence of all finite order permutational moments, they clearly posed this problem as an open one, and soon a series of research papers addressed this problem in greater generality. Noether [42] was able to relax the Wald–Wolfowitz conditions to a certain extent, albeit in the same setup of a linear statistic. Hoeffding [26] not only extended the basic result to a larger class of (bi-linear) statistics, but also incorporated Noether-type condition instead of the classical Wald–Wolfowitz type ones. Specifically, he considered a statistic of the form

$$B_N = \sum_{i=1}^{N} b_N(i, R_{Ni}),$$

where \((R_{N1}, \ldots, R_{NN})\) takes on each permutation of \((1, \ldots, N)\) with the common probability \((N!)^{-1}\), and for every \(N \geq 1\), \(B_N = \{b_N(i, j), 1 \leq i, j \leq N\}\) is a suitable set of real numbers. Let us define

$$b_N(i, \cdot) = N^{-1} \sum_{j=1}^{N} b_N(i, j); \quad b_N(\cdot, j) = N^{-1} \sum_{i=1}^{N} b_N(i, j);$$

$$\bar{b}_N = N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} b_N(i, j)$$
and
\[ d_N(i, j) = b_N(i, j) - b_N(i, \cdot) - b_N(\cdot, j) + \bar{B}_N, \text{ for } i, j = 1, \ldots, N. \]

Then it can be easily shown that, under the permutation structure on the \( R_{Ni} \), the mean and variance of \( B_N \) are respectively \( N\bar{b}_N \) and \((N - 1)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j)\). Hoeffding [26] assumed that as \( N \) increases,
\[
\frac{\max_{1 \leq i, j \leq N} d_N^2(i, j)}{N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j)} = o(N),
\]

and this enabled him to deriving the desired PCLT for \( B_n \). He also showed that the above condition also implies the Wald–Wolfowitz type condition that
\[
\frac{N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} |d_N(i, j)|^r}{\{N^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j)\}^{r/2}} = o(N^{r/2-1}),
\]

for \( r = 3, 4, \ldots \); this Hoeffding condition was further relaxed by Motoo [41] to a Lindeberg–type condition: for every \( \varepsilon > 0 \),
\[
\frac{\sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j)I(d_N^2(i, j) > \varepsilon \sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j))}{\sum_{i=1}^{N} \sum_{j=1}^{N} d_N^2(i, j)} \rightarrow 0,
\]
as \( n \to \infty \). The ultimate answer to the underlying regularity conditions pertaining to the PCLT was provided by Hájek [13] who showed that under Noether type conditions on the individual sequences \( a_N(i), \ i = 1, \ldots, N; \ c_N(j), \ j = 1, \ldots, N \), when \( b_N(i, j) = a_N(i)c_N(j) \) for every \( i, j = 1, \ldots, N \), the Hoeffding–Motoo condition on the \( d_N(i, j) \) is a necessary and sufficient one for the PCLT to hold.

In nonparametrics, these PCLTs occupy a very special place. As we shall note later on, locally most powerful rank tests are typically based on statistics of the type \( B_N \), hence their asymptotic normality (under suitable hypotheses of invariance) can be derived by using such PCLTs. In this context, the \( b_N(\cdot, \cdot) \) satisfy certain additional conditions under which some martingale characterizations hold. As such, suitable martingale central limit theorems (and their functional variations) can be incorporated in providing alternative methods for deriving such PCLTs. This approach has been studied in detail in Chapters 3–8 of Sen [56]. Moreover, such martingale approaches have paved the way for multivariate permutational central limit theorems for nonparametric statistics [viz. Sen [57] ], and it is not at all surprising that Hoeffding–Motoo type conditions underlie these developments as well. Typically, in such a multivariate setup, we have a \( q \times p \) matrix–statistic
\[
\sum_{i=1}^{N} c_{Ni}[a_{N1}(R_{Ni}^{(1)}), \ldots, a_{NP}(R_{Ni}^{(p)})] = L_N, \text{ say,}
\]
where the \( c_{N_i} \) are \( q \)-vectors of known regression constants and

\[
R_N = \begin{pmatrix}
R_{N1}^{(1)} & \cdots & R_{NN}^{(1)} \\
\vdots & \ddots & \vdots \\
R_{N1}^{(p)} & \cdots & R_{NN}^{(p)}
\end{pmatrix}
\]

The score functions \( a_{N_i}(i) \), \( i = 1, \ldots, N, \ j = 1, \ldots, p \), are defined suitably. The permutation group is generated by the column-permutations of \( R_N \) and these \( N! \) possible column permutations are conditionally equally likely (given the collection). If one wants to use the classical Cramér–Wold theorem in establishing the desired permutational limit law for \( L_N = (L_{Nrs}) \), say, then it is necessary to consider an arbitrary matrix \( A = (\lambda_{rs}) \) of real constants, and formulate a linear combination

\[
L_N^0 = \sum_{r=1}^{q} \sum_{s=1}^{p} \lambda_{rs} L_{Nrs},
\]

on which the classical PCLTs may be used. However, in this setup, \( L_N^0 \) may not conform to a linear statistic (especially when \( q \geq 2 \)), and the Hoeffding formulation appears to be much more useful. Often, the scores \( a_{N_i}(i) \) are so defined that, under the usual hypothesis of invariance, \( L_N \) can be represented as a martingale. Again, in such cases, the Hoeffding projections referred to earlier may be used to provide a considerably simpler proof of the asymptotic normality results (under the permutational setup). There are other situations in which a linear statistic may not crop up, but Hoeffding's bilinear statistic may turn out to be more applicable, and in that sense his PCLT remains as a landmark in this area of research.

5. Optimal Nonparametric Tests: Concept of LMPR

As has been mentioned before, classically, nonparametric tests used to be referred to as distribution-free tests because, under suitable hypotheses of invariance, the distribution of such a test statistic is generally independent of the underlying distribution. However, this simple prescription may not be generally true, as may easily be verified with a bivariate or multivariate hypothesis of invariance or with a composite hypothesis problem. Moreover, the distribution-freeness under the null hypothesis provides some convenience in terms of the characterization of the size of the tests although it is usually necessary resort to a randomization test to achieve an exact significance level. The power of a test depends not only on the chosen level of significance but also on the specific alternative hypothesis one may have in mind. In this respect, the power is highly dependent on the underlying
distribution (unless one considers a Lehmann [39] type of alternative or the so-called proportional hazards model).

To illustrate this point, let us briefly consider the simple regression model which contains the classical two-sample model as a special case. Suppose that $X_1, \ldots, X_N$ are independent random variables with d.f.'s $F_1(x), \ldots, F_N(x)$ respectively, all defined on $\mathbb{R}$. Consider the model

$$F_j(x) = F(x, \alpha - \beta c_j), \quad x \in \mathbb{R}, \; j = 1, \ldots, N,$$

where the $c_j$ are known (regression) constants (not all equal), $\alpha$ and $\beta$ are the intercept and regression parameters (unknown), and the d.f. $F$ is continuous but of unknown form. Consider the null hypothesis of randomness (or homogeneity of the d.f.'s $F_j$) which may also be stated as $H_0 : \beta = 0$. Under $H_0$, all the $F_j$ are the same, so that the $X_i$ are all i.i.d.r.v.'s. This enables one to construct a large class of tests based on the ranks $(R_{NM})$ of the $X_i$ (among themselves) and the given $c_i$. Such a test will be distribution-free under $H_0$ and robust in a global sense. Therefore, it is natural to ask: which test should one prescribe within this class of distribution-free tests?

In this regression setup, it can be assumed that the $c_i$ are ordered (i.e. $c_1 \leq \cdots \leq c_N$). Suppose we are interested in the alternatives that $\beta > 0$. The distribution of a rank statistic (say, $T_N$) under such an alternative (say, $H_1$) is no longer distribution-free and it will depend not only on the specific $\beta$ but also on the d.f. $F$.

As we shall note later on, the developments in the forties related to using permutation tests and vaguely justifying them on their optimality properties under a specific parametric setup. Hoeffding [25] came up with the novel idea of characterizing optimality of such a test in a local sense. Suppose that against $H_0 : \beta = 0$, we consider an indexed set of alternatives $H_{1\Delta} : \beta - \Delta (\geq 0)$, where $\Delta$ belongs to the interval $(0, \eta)$, for some $\eta > 0$. A test is then termed locally most powerful (LMP) for $H_0$ against the class $H(\eta) = \{H_{1\Delta} : \beta - \Delta \in (0, \eta)\}$, if it is uniformly most powerful against $H(\eta)$ when $\eta$ tends to 0. If this class of tests is restricted to rank-based ones, we shall term such an optimal test as a locally most powerful rank test (LMPR). Suppose now that the d.f. $F$ admits an absolutely continuous density $f(\cdot)$ with first derivative $f'(\cdot)$, such that $\int_\mathbb{R} |f'(x)| \, dx < \infty$. Let

$$\psi_F(u) = -f'(F^{-1}(u))/f(F^{-1}(u)), \quad u \in (0, 1).$$

Suppose that $U_{N;1} < \ldots < U_{N;N}$ denote the order statistics of a sample of size $N$ from the uniform$(0,1)$ distribution, and let $a_N^\psi(k) = E\psi_F(U_{N;k})$, for $k = 1, \ldots, N$. If a linear rank statistic is based on the score function $\psi_F(\cdot)$ and the scores $a_N^\psi(k)$, it will be LMPR for the alternatives mentioned above. In particular, if $F$ is normal, the $a_N^\psi(k)$ are known as the normal scores, so that $T_N$ reduces to the so-called Fisher–Yates–Hoeffding–Terry test statistic.
These findings are by no means restricted to a normal $F$ and they encompass a more general class of densities (including the ones with finite Fisher information). This work of Hoeffding [25] opened up a broad avenue of research on locally optimal rank and other robust tests in various models. Terry [59], under the able guidance of Wassily, formulated some rank order tests which are most powerful against specific parametric alternatives. An excellent treatment of LMPR tests with general parametric alternatives is contained in Hájek and Šidák ([16], Ch.III). This last reference is also an excellent source of further research work done in this context to tie in the LMPR tests with asymptotically optimal rank tests (for contiguous alternatives) and asymptotically optimal adaptive procedures. The score function $\psi_f(u), u \in (0, 1)$, depends on the unknown density $f(\cdot)$, and, hence, there is a natural question: can this density $f(\cdot)$ and/or some other related functional be estimated from the sample data and incorporated in the formulation of suitable rank tests (or estimates) which would remain asymptotically optimal for all $f(\cdot)$ belonging to a class? This is the genesis of adaptive procedures, and the very formulation of LMPR and its amalgamation with asymptotically optimal tests for local alternatives played a fundamental role in the subsequent developments. Hoeffding’s [25] work is another landmark in nonparametrics.

6. Asymptotics of Permutation Tests

R. A. Fisher [10] initiated the idea of permutation tests, and E. J. G. Pitman made some major contributions in 1937–38. Schéffe [53] unified the theory in a more comprehensive manner, although the real breakthrough came in 1952, and this was due to Hoeffding [27]. These permutation tests are the precursors of general nonparametric tests, and they are often referred to as randomization tests. In nonparametrics, it is often the case that a hypothesis of invariance generates a finite group of transformations which maps the sample space onto itself, and under the null hypothesis, the joint distribution of the sample observations remains invariant with respect to this group of transformations. As such, suitable orbits for the sample point can be defined with reference to such groups of transformations, and a (conditional) test based on the uniformity of the conditional distribution of the sample point on such orbits (having a discrete structure) works out well under comparatively less stringent regularity assumptions.

In the forties, researchers were tempted to use standard parametric test statistics in this conditional setup, and the main emphasis was on computation of the (conditional) critical level using the allied permutation laws, and on justifying such optimal parametric tests in a comparatively more general setup. Given the undeveloped status of optimal nonparametric tests (at that time), such permutation tests had a predominantly parametric struc-
ture, and there remained some basic questions about the structure of such permutation tests and their optimality properties, if any. Although the permutational (conditional) distribution of a test statistic provides access to computing the exact (conditional) critical levels, it suffers from two basic drawbacks. Firstly, these critical levels are themselves random variables, and computationally they become messy when the sample size becomes large (this was the main reason for the adoption of the PCLTs). Secondly, these permutation tests are basically conditional tests, but studies of their power properties generally demand the knowledge of the unconditional distributions of such test statistics when the hypothesis of invariance may not hold. While the PCLTs take care of the convergence of the conditional distributions under the null hypothesis, they provide very little help in dealing with the non-null hypothesis situations. The exact enumeration of the unconditional non-null distribution of nonparametric (even, permutation) test statistics is generally a laborious task (even for small sample sizes), and this becomes prohibitively laborious as the sample size increases. This is the main reason why, in nonparametrics, limit theorems have found a special place, and for the permutation tests, Hoeffding [27] provided the basic results in a general form. This fundamental work centers around the use of some "optimal" parametric test statistics in a permutational setup to render them permutationally (conditionally) distribution-free, and then proceeds to establish the following two important asymptotic results:

1. Whenever the actual null hypothesis d.f. of a test statistic (say, $T_n$) and the permutation d.f. of $T_n$ both converge to a common d.f. (say $G$) which satisfies certain (mild) continuity conditions, then both the permutational critical level and the unconditional critical level converge to a common limit, and the permutation test is asymptotically (in probability) size-equivalent to the unconditional test.

2. The asymptotic power of the permutation test based on $T_n$ is in agreement with the asymptotic power of the unconditional test based on $T_n$. In this context, it may be noted that for any fixed alternative, by virtue of the usual consistency property of tests, both the asymptotic powers would be equal to one, so that the problem becomes non-trivial only for local alternatives.

Judged from these aspects, there were certain mathematical intricacies which Wassily had to deal with in a rigorous manner. Firstly, the permutational d.f. of $T_n$ is itself a random function (although of bounded variation), so that he needed some convergence properties of random d.f.'s, and some of these results are of independent interest. In fact, in connection with the multivariate rank tests, such permutational laws are, in general, conditional ones, and Hoeffding's lemmas have been very useful in guaranteeing the required convergence results (in probability). Secondly, in dealing with a sequence of local alternatives, one is confronted with a triangular scheme
of random elements, so that the proofs of the desired convergence results needed some extra care, which he exhibited in his characteristic manner. Thirdly, it was before the days when the completeness of order statistics or of other sample functions which may not conform to the notion of minimal sufficiency was properly established, so the direct proofs Hoeffding provided cast light on their basic nature in a much less abstract fashion. We may even note that the notion of "contiguity of probability measures" introduced by LeCam [37] but popularized mostly through the pioneering efforts of Hájek [14] was not yet developed, so that Hoeffding's treatment also retains its original flavor in its simplicity, elegance and applicability aspects.

7. Order Statistics, Bernoulli Schemes and Allied Distributional Problems in Nonparametrics

We noted earlier that, in the context of LMPR tests, the optimal scores are given by $a_{Y}(k) = E(-f'(X_{N:k})/f(X_{N:k})), k = 1, \ldots, N$, where the $X_{N:k}$ are the ordered r.v.'s in a sample of size $N$ drawn from the distribution $F$ whose density function is given by $f(\cdot)$. In the study of the asymptotic theory of such LMPR tests and in other contexts too, one encounters a problem which may be posed as follows.

Let $c_{N} = EX_{N:k}, i = 1, \ldots, N$ (so that $c_{N1} \leq \cdots \leq c_{NN}$), and assume that $E_{F}X$ (and hence the $c_{N1}$) exists. Define a d.f. $G_{N}$ which puts probability mass $N^{-1}$ at each of the points $c_{Ni}, i = 1, \ldots, N$ (and zero elsewhere). Then, if for some $r \geq 1$, $E|X|^r$ exists, the d.f. $G_{N}$ converges to $F$ as $N \to \infty$, and $\int |x|^r dG_{N}(x) \to \int |x|^r dF(x), as N \to \infty$.

In this setup, the result relates to the weak convergence as well as the moment convergence of the particular sequence $\{G_{N}\}$. In fact, Hoeffding [28] proved a more general result. Suppose that $E_{F}X$ exists, and let $g(x)$ be a real-valued continuous function, such that $|g(x)| \leq h(x)$, where $h(\cdot)$ is convex and $\int h(x) dF(x) < \infty$. Then

$$\lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} g(c_{Ni}) = \lim_{N \to \infty} \int g(x) dG_{N}(x) = \int g(x) dF(x).$$

He also listed some other extensions of this basic result. In passing, we may remark that the $c_{Ni}$ are not r.v.'s, hence $G_{N}(\cdot)$, as defined above, is a discrete d.f. but is not a random d.f. (as was the case with the permutation distributions treated in the earlier section). The discrete nature of $G_{N}$, combined with the fact that the number of mass points increases with $N$, poses some problems with the conventional treatment for the weak convergence results (viz. the Helly-Bray lemma) and related moment convergence.
results. Hoeffding [28] did not hesitate to use explicitly the algebraic structures of the $c_{Nt}$ (in terms of $F(\cdot)$), thereby providing a direct proof with a minimal coating of abstraction.

Let me comment on the significance of this result from the statistical point of view. If we define the score function $\psi_f(u), u \in (0, 1),$ as in Section 5, then allied to the scores $a_N^0(k)$ are the van der Waerden scores $a_N(k) = \psi_f(k/(N+1)), k = 1, \ldots, N.$ The affinity of the $a_N^0(k)$ and $a_N(k)$ can of course be studied under appropriate smoothness conditions on the score generating function $\psi_f(\cdot).$ The first mean value theorem in integral calculus leads us to conclude that whenever, for $r > 0$, $\int_0^1 |\psi_f(u)|^r \, du$ exists, $N^{-1} \sum_{i=1}^N |a_N(i)|^r$ converges to the former integral. The basic result of Hoeffding [28] extends this convergence result to the case of the $a_N^0(k)$ as well. In fact, it follows under the same condition on the score generating function, that $N^{-1} \sum_{i=1}^N |a_N^0(i) - a_N(i)|^r$ converges to zero as $N \to \infty.$ Moreover, for the $a_N^0(k),$ when $r = 1$ the mean is exactly equal to $\int_0^1 \psi_f(u) \, du,$ whereas the exactness may be vitiated for the $a_N(k).$ This convergence result, in turn, permits us to use interchangeably the two sets of score functions. Since the $a_N(k)$ are computationally simpler, we may take advantage of this when $N$ is not so small. This fundamental result has some important use in centering of linear rank statistics, and we shall refer to that later on.

Combinatorial methods have special appeal as ways of solving a variety of problems in mathematical statistics, probability theory and stochastic process as well. This is particularly overwhelming in nonparametrics, and it is no exception that the genius in Hoeffding found an original way of incorporating combinatorics in a simple yet very interesting problem in Poisson sampling which opened up a wide avenue of research in subsequent years. Hoeffding [30] considered the following simple model. Let $S_n = X_1 + \cdots + X_n$ where the $X_i$ are independent Bernoulli r.v.'s, such that $P(X_i = 1) = 1 - P(X_i = 0) = p_i, i = 1, \ldots, n,$ and the $p_i$ need not be the same. Let $F_n = n^{-1} \sum_{i=1}^n p_i.$ It is easy to check that $E S_n = np$ and $\text{Var}(S_n) = np(1-p) - \sum_{i=1}^n (p_i - p)^2 \leq np(1-p),$ where the inequality sign is attained only when all $p_i$ are equal to $p.$ Starting with this simple observation, Wessely considered an arbitrary convex $g(\cdot)$ for which

$$g(k + 2) - 2g(k + 1) + g(k) > 0,$$

for every $k = 0, 1, \ldots, n - 2.

and showed that

$$E[g(S_n)] \leq \sum_{k=0}^n g(k) \binom{n}{k} p^k (1-p)^{n-k},$$

(7.1)

where the equality sign holds if and only if $p_1 = \cdots = p_n = p.$ Two other results obtained by him in this context have had a lot of statistical impact
in recent years. Firstly, if \( c \) is any nonnegative integer, then under the above scheme, for every \( n \geq 1 \),
\[
0 \leq P(S_n \leq c) \leq \sum_{k=0}^c \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{if } 0 \leq c \leq np - 1;
\]
\[
\sum_{k=0}^c \binom{n}{k} p^k (1 - p)^{n-k} \leq P(S_n \leq c) \leq 1, \quad \text{if } np \leq c \leq n,
\]
and the bounds are attained only if \( p_1 = \cdots = p_n = p \). Secondly, as a consequence, if \( b \) and \( c \) are two nonnegative integers such that \( 0 \leq b \leq np \leq c \leq n \), then for every \( n \geq 1 \),
\[
\sum_{k=0}^c \binom{n}{k} p^k (1 - p)^{n-k} \leq P(b \leq S_n \leq c) \leq 1,
\]
where the lower bound is attained only if \( p_1 = \cdots = p_n = p \) (unless \( b = 0 \) and \( c = n \)).

It may be noted that (7.1) (which is also affectionately termed the Hoeffding theorem on the Poisson–binomial distribution) has led to a series of developments in diverse setups. Samuels [52] made some further interesting remarks on (7.1) while Anderson and Samuels [1] provided further insights to this inequality with more delicate conditions on the \( p_i \). Rinott [50] and Gleiser [11] derived parallel results using majorization and Schur functions, while a very elaborate study of (7.1) for more general form of distributions (and some characterizations) is due to Bickel and van Zwet [4].

The inequality (7.1) has been extended in a yet another direction, initiated by Hoeffding [32] himself. Consider simple random sampling without replacement (SRSWOR) and denote the r.v.'s by \( X_1, \ldots, X_n \); as well as the i.i.d. case under SRSWR (with replacement) with the corresponding r.v.'s denoted by \( X_1^0, \ldots, X_n^0 \). Also, let \( S_n = X_1 + \cdots + X_n \) and \( S_n^0 = X_1^0 + \cdots + X_n^0 \). Then \( E_S(S_n) \leq E_S(S_n^0) \). Rosen [51] extended this inequality to the case in which SRSWOR has been replaced by any symmetric sampling plan. Karlin [35] contains an extensive investigation of various systematic and rejective sampling plans for which the Hoeffding inequality holds. Combinatorics play a basic role in this context, and so does the concept of total positivity (due to Karlin himself). In view of our primary emphasis on the nonparametric aspects, we shall not pursue this further.

Next, we note that (7.3) is a basic inequality geared towards the effect of heterogeneity of the underlying d.f.'s on the probability laws for symmetric sample functions. In this sense, for the binomial case, the coverage probability is a minimum in the homogeneous case (i.e. for \( p_1 = \cdots = p_n \)), so that if the \( X_i \) do not have the same Bernoulli law, there is increased concentration around the expected value. In the recent past, considerable attention has been paid to robustness aspects when the sample observations may not have a common distribution (the heteroscedastic model in the linear parametric case is akin to this too). Hoeffding's result is probably
the first step in this direction. Speaking of nonparametric rank tests, this assumption of identically distributed r.v.'s has been dispensed with to a greater extent. There is a closely related problem in order statistics where an analogue of the Hoeffding inequality has been derived under parallel regularity conditions [Sen [55]]. Let \( X_1, \ldots, X_n \) be independent r.v.'s with continuous d.f.'s \( F_1, \ldots, F_n \), all defined on \( \mathbb{R} \). Let \( X_{n:1} < \cdots < X_{n:n} \) be the corresponding order statistics. Also, let \( F_n = (F_1, \ldots, F_n) \) and \( \bar{F}_n = n^{-1} \sum_{i=1}^{n} F_i \) be the average d.f. Assume that \( \xi_{n,r} \) is a unique solution of \( \bar{F}_n(\xi_{n,r}) = r/n \) for \( r = 1, \ldots, n-1 \) and \( \xi_{n,0} = -\infty, \xi_{n,n} = \infty \). Also, let \( P_r(x; F_n) = P\{X_{n,r} \leq x; F_n\} \) and \( P^*_r(x; F_n) = P\{X_{n,r} \leq x; F_1 = \cdots = F_n = \bar{F}_n\} \). Further, let \( P_r(\eta_{n,r}; F_n) = P^*_r(\eta_{n,r}; \bar{F}_n) = 1/2 \) for \( r = 1, \ldots, n \). Then, for \( 2 \leq r \leq n-1 \), and for all \( x \leq \xi_{n,r-1} \leq \xi_{n,r} \leq y \), \( P_r(y; F_n) - P_r(x; F_n) \geq P^*_r(y; \bar{F}_n) - P^*_r(x; \bar{F}_n) \), where the equality sign holds only if \( F_1 = \cdots = F_n \). For the two extreme order statistics, \( P_1(x; F_n) \geq P^*_1(x; \bar{F}_n) \) and \( P_n(x; F_n) \leq P^*_n(x; \bar{F}_n) \), for all \( x \in \mathbb{R} \), with strict inequalities unless \( F_1 = \cdots = F_n = \bar{F}_n \) at \( x \). Therefore, the d.f.'s of the individual order statistics are more concentrated around the corresponding quantiles of the d.f. \( \bar{F}_n \) in the heterogeneous case than in the homogeneous case. Further, it follows from the above that for all \( r : 2 \leq r \leq n-1 \), \( \xi_{n,r-1} \leq \eta_{n,r} \), \( \eta_{n,r} \leq \xi_{n,r} \), so that \( |\eta_{n,r} - \xi_{n,r}| \leq |\xi_{n,r} - \xi_{n,r-1}| \). Like (7.3), these inequalities are not of asymptotic form, and their analogues in the asymptotic cases are easy to conceive [see Chapter 3 of Puri and Sen [46]]. The Hoeffding inequality in (7.3), as extended here to sample order statistics in a much more general setup, has been instrumental in the development of some reliability inequalities; see e.g. Pledger and Proschan [45]. In an asymptotic setup, it is not difficult to verify that distributional inequalities for the empirical distribution processes and partial sum processes also satisfy this Hoeffding-type behavior. In the context of rank tests for the multi-sample/regression models as well as the independence models, it has been observed that the i.d. nature of the r.v.'s can be dispensed with to a certain extent, and Hoeffding-type inequalities lie at the root of all such developments; see Puri and Sen [46] for some details.

8. Asymptotic Efficiency in Nonparametrics

In the context of nonparametric testing problems, the limiting power of a test provides the essential information on the asymptotic performance properties of the test. Thus, a comparison of the asymptotic power functions of competing tests reveals their asymptotic relative efficiency (A.R.E) properties. As has been noted earlier, if the competing tests are all consistent then their asymptotic powers for a fixed alternative are all equal to one, so there is little information one can gather from this consistency property alone. Pitman [44] introduced the concept of a sequence of local
alternatives (to a basic hypothesis of invariance), which may generally be of
parametric nature, and showed that under such alternatives the asymptotic
powers are different from one, so that they may be combined to produce a
meaningful measure of the A.R.E. Basically, this amounts to choosing two
different sequences, say \( \{N_{n1}\} \) and \( \{N_{n2}\} \), of sample sizes depending on
the index \( n \) (which is tied to the formulation of the Pitman–Type alterna-
tives), such that two sequences of test statistics, say \( \{T_{N_{n1}}^{(1)}\} \) and \( \{T_{N_{n2}}^{(2)}\} \),
both have the same asymptotic power functions (and sizes too) with re-
spect to the sequence of alternatives under consideration, and in that case,
whenever \( \lim_{n \to \infty} \{N_{n1}/N_{n2}\} \) exists, it can be interpreted as a convenient
measure of the A.R.E. of the test based on \( T^{(2)} \) with respect to that on
\( T^{(1)} \).

Whenever the competing statistics are attracted by suitable nondegen-
erate limit laws (under the null hypothesis as well as under Pitman–type
alternatives), the parameters associated with the limit laws may provide
this measure in a convenient way. Along the same lines as Pitman [44],
Noether [43] generalized this definition of A.R.E., which covers both the
asymptotic normal and chi-squared distributions. A somewhat different no-
tion of efficiency had been investigated by Hoeffding and Rosenblatt [29].
They formulated a general analogue of the Pitman A.R.E. for a broader
family of tests and covering a broader class of alternatives (including com-
posite ones). However, their work was by no means confined to classical
nonparametric tests for the classical problems in statistical inference, and
their setup included this latter one as a special case. On the other hand,
like the earlier developments, their measure of A.R.E. related to so-called
local alternatives. In this direction, the general A.R.E. results presented
in Chapter VII of Hájek and Šidák [16] incorporate the notion of contiguous
alternatives, for which a greater amount of unification of the diverse
measures is possible. However, even for such contiguous alternatives, for
two statistics having non-conformable limit distributions (viz. Wilcoxon
test vs. Kolmogorov–Smirnov test in the two–sample problem), the con-
ventional measures of A.R.E. may not work out well, and some further
modifications may therefore be necessary. Another of Hoeffding’s students,
Dana Quade [49], worked out the case with the Kolmogorov–Smirnov test.
Bahadur [2] considered an alternative measure of A.R.E. (known as the
Bahadur efficiency) for possibly non–local alternatives, and in a certain
sense for local alternatives they become isomorphic (under additional con-
ditions).
9. Large Deviations and Probability Inequalities

Wassily Hoeffding made some notable contributions to probability inequalities for sample sums (averages) as well as for suitable (viz. convex) functions of them when the basic assumption of independence and/or identity of distributions of the underlying r.v.'s may not hold. Moreover, he incorporated such inequalities (with some other ones) in the development of some general large deviation limit theorems. In a sense, his [33] Annals of Mathematical Statistics paper is a landmark in this direction (he has a companion article in the Fifth Berkeley Symposium), and David Herr [18], another of his students, carried out the study for multivariate normal distributions and published a fine paper in the Annals of Mathematical Statistics in 1967. Wassily delivered a series of (three) lectures (designated as the Wald Lectures) on the asymptotic optimality of likelihood ratio and related tests at the 1967 annual meeting of the Institute of Mathematical Statistics (in Washington D.C). I had the privilege of listening to him, and I still regard this as being among his most thoughtfull sets of presentations. Since the details of this aspect (along with the other developments during the sixties) are covered in the accompanying review article by Oosterhoff and van Zwet (in this volume), I omit further deliberations. However, without mention of this aspect, my review of Wassily's contributions in nonparametrics would have been rather incomplete. I would also like to point out in this context that the previous work Wassily had done on efficiency and A.R.E. of statistical tests and the deep interaction he had with Professor Raghu Raj Bahadur (with whom he had a friendship from the very beginning when both came to Chapel Hill in the late forties) were instrumental in his deep interest and creative research in this novel area. Actually, Wassily made an excellent contribution towards the asymptotic optimality of statistical tests wherein the concept of Bahadur efficiency (along with other non-local measures of the A.R.E) of statistical tests laid down the foundation and the large deviation probabilities provided elegant results in this direction.

10. Significance of Wassily's Work on Nonparametrics during the Seventies and Eighties

In nonparametrics, developments on limit theorems constitute one of the most important achievements. In this sense, the real breakthrough occurred with the asymptotic theory of $U$-statistics in 1948. This article by itself cleared the way for many subsequent developments. The second phase of developments began with the notable work of Chernoff and Savage [5] who established the asymptotic normality of a general class of linear rank statistics when the null hypothesis (of randomness) may not hold. In this respect
too, it would not be improper to refer to the work of Dwass [6, 7, 8] who, under the able guidance of Hoeffding, established similar results (although for a somewhat restricted class of rank statistics) incorporating suitable approximations based on (generalized) $U$–statistics where Hoeffding's 1948 article remained as fundamental. Nevertheless, the Chernoff and Savage [5] regularity conditions were also not the most general ones, and there was ample room for further theoretical developments. Hájek [14] considered an alternative approach where the regularity assumptions concerning the score functions appeared to be the least stringent, although the developments were confined to contiguous alternatives and demanded finite Fisher information. Hájek and Šidák [16] explained the relative merits and demerits of these two alternative approaches in a very convincing manner.

During the mid–sixties, other people started looking into this problem from different angles. Govindarajulu, LeCam and Raghavachari [12] attempted to relax the regularity conditions in Chernoff and Savage [5] through some weak convergence results on sample (empirical) distribution processes. In this respect, the Pyke and Shorack [48] approach gained momentum because of its intrinsic sophistication based on almost sure representations, and their regularity conditions appeared to be the most general ones. However, their developments were mostly confined to the single or multi–sample situations, whereas Hájek [14] related to a comparatively more general (simple) regression model. It was quite natural for Hájek [15] to investigate the asymptotics for a general linear rank statistic, for possibly non–contiguous alternatives.

Wassily became aware of this fundamental work of Hájek at a very early stage, and he offered various constructive comments which made the presentation even better. Some exchange of ideas of these two intellectual giants in this context can be seen in the comments Wassily made on the Hájek review article which appeared in the 1970 proceedings of the nonparametric statistics conference held at Bloomington, Indiana. (These comments appear in this present volume.) The main achievement of Hájek [15] is a powerful variance inequality for linear rank statistics in a completely general setup which, when combined with another elegant polynomial approximation for the score function (under additional regularity conditions), yields the desired asymptotic normality result under quite general regularity conditions.

However, in this respect, Hájek [15] had to restrict attention to the actual expectation of the linear rank statistic as a centering constant. On the other hand, in the developments in Chernoff and Savage [5] and Hájek [14], this centering constant came out in an alternative simpler form. Although Hájek [15] addressed the issue of replacing this centering constant in a somewhat more restricted setup, there remained the basic issue: can the expectation of the linear rank statistic be replaced by a more natural parameter in this asymptotic normality result? In Section 7, we have made some more general comments on the genuine interest Wassily had in com-
binatorics which led him to formulate the inequalities in (7.1), (7.2) and (7.3) in a very elegant manner. This interest prompted him to investigate the Hájek [15] centering of a linear rank statistic problem in a completely different manner. His interest in the Bernstein polynomial approximation to absolutely continuous, square integrable and monotone (or a difference of two monotone) functions led him to establish possible replacement of centering constants under the general setup [34], where he [34] needed only to replace the square–integrability of the score function in Hájek [15] by a slightly more stringent one:

$$\int_0^1 \{t(1-t)\}^{-1/2} d\phi_k(t) < \infty, \quad \text{for } k = 1, 2, \quad (10.1)$$

where the actual score generating function $[\phi(t)]$ is expressed as the difference of two monotone score functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$. In the literature, this is referred to as the Hoeffding [34] condition, and is satisfied by all the usual score functions used in practice. Puri and Sen [47] contains a detailed treatment of this basic work of Hoeffding in various problems in nonparametrics.

Wassily Hoeffding taught the courses on nonparametric statistics, statistical decision theory, and sequential analysis at the University of North Carolina at Chapel Hill for more than 25 years. Although he never cared to gather his excellent classmates into monographs or texts, those of us who had the opportunity to attend some of these courses have gleaned a basic appreciation of his vast research interest.

During the seventies, he started teaching an advanced course on “Large Sample Theory” which included both parametric and nonparametric inference procedures having some asymptotic optimality properties. Practically nothing was available in an advanced textbook at that time, and we were all hoping that after a few years of offering this course, he might turn his notes into a monograph. His health became one of the impasses and his sense of perfection was also the other factor. He would not release any article for publication unless he was himself fully satisfied with the developments contained in it. This might have been the reason why his classical [31] University of North Carolina Mimeo Series Report (on Strong Law of Large Numbers for $U$–statistics) never met the light of publication in a natural journal. It might also have been the reason why he waited for nearly three years to publish the Bernstein polynomial approximation paper.

In the seventies and early eighties, he looked into the problem of some range preserving (unbiased) estimation problems which extended the original $U$–statistics paper in a somewhat decision theoretic setup, thus adding to his notable contributions to decision theory. This material itself would have constituted an outstanding monograph.

Once I asked him (long time ago): why he never cared to write a book or
advanced monograph in one of these three areas? The answer was a long pause and a characteristic Wassily-smile; I have the feeling that another nonparametric text published in the late fifties dealt with many of the fine papers Wassily had written in this area (and in a way not too different from the original sources), so that there was no point in his rewriting his own work in a more lucid way to convince the audience of the role of unification in writing a monograph. We are, of course, fortunate to have these constructive criticisms and encouraging comments on various endeavors. His writing is lucid and rigorous, yet not so abstract that there may not be any profound need to further unify his work into a monograph/textbook form. This is the main reason why we felt this present collection of work would fill a genuine need to reveal the vast and fundamental work Wassily Hoeffding had made in various areas in Statistics and Probability Theory, to appreciate his mastery, and to incorporating this as a useful source of reference to some authentic work on which modern nonparametric methods (as well as the other areas reviewed) stand.

Wassily, a person of Danish origin, born in Russia, educated in Berlin (Germany) and professionally recognized in USA, was international in a true sense. He was indeed able to combine the mathematical insights of eastern European tradition with the statistical perceptions of the west, and throughout his professional career, he maintained a true academic spirit and contributed generously through his outstanding research creativity. Nonparametrics has been the "jewel in the crown" of his research accomplishments, and Wassily's actual writings convey this picture in a more visible way than this little introduction I have put together.

11. References


972–994.


98, 39–82.


zations of theorems of Chernoff and Savage on asymptotic normality
*Statist. Probab.* 1, 609–638.


Math. Statist.* 33, 1124–1147.


Prague.

*Statist.* 17, 34–44.

[18] Herr, D.G. (1967). Asymptotically optimal tests for multivariate nor-

*Schriften des Mathematischen Instituts und des Instituts für An-

kontinuierliche Verteilungen. *Archiv für mathematische Wirtschaften
und Sozialforschung.* 7, 49–70.


The Impact of Wassily Hoeffding's Work on Sequential Analysis

Gordon Simons

1. Prefatory remarks.

I was tremendously privileged to have been one of Wassily Hoeffding's colleagues. While his demeanor and utter clarity of thought could be intimidating at times, it truthfully can be stated that he was always gracious and generous in his assessments of his colleagues, reserving for himself standards that only someone of his exceptional intellectual stature could hope to achieve.

Wassily was, by temperament, quiet and not given to verbosity, neither in speech nor in his writings. He found it easier to express his thoughts on paper, something he did extremely well. I was reminded of this recently as I read an interview of Boris Gnedenko (Statistical Science, Volume 7, Number 2, May 1992). When he was asked by Nozer Singpurwalla whether, for his many writings, he uses a word processor or computer, Gnedenko replied, "No, no, I use a typewriter." And when asked about the number of drafts, Gnedenko stated "One draft. It is necessary to think first and only then to write. At this stage I am almost finished." The same could be said of Hoeffding's writings. I can well recall sitting in Wassily's office as he wrote for my benefit, in beautifully composed prose, with pen and ink, the details of some statistical thoughts, first thinking and then writing, never once needing to cross out some part of a disorganized thought, never correcting a misused or misplaced word, and never needing to insert so much as a missing word. While it is a loss to the profession that he never wrote a textbook or research monograph, his clear and elegant writing style can be found throughout all of his research papers.

---

1 University of North Carolina at Chapel Hill
2 This was partially alleviated through the efforts of David Kikuchi, a former student who organized some of Hoeffding's lecture notes. Of course, graduate students in statistics at UNC benefited immensely under Hoeffding's instruction.
Some great thinkers are best appreciated through the writings of those who follow in their wake. Not so for Wassily Hoeffding. While he is being honored here for the excellence of his published research, let it also be noted here that the Hoeffding legacy includes his quintessential presentation of this research. This seductive but often subtle constituent of his legacy is available only to those who are willing to spend time with the original documents republished here.

SO READ, ... ENJOY, ... AND BE ENRICHED!

2. Hoeffding's work in sequential analysis.

Of his approximately 40 research papers, only four address topics in sequential analysis. All four are concerned with the expected sample size required by sequential procedures. Three obtain general lower bounds for the expected sample size of sequential tests. One of these extends the discussion from "expected sample size" to "average risk", and hence from a testing context to a more general decision-theoretic setting. The fourth paper addresses the task of efficiently "producing an unbiased coin" (of generating a dichotomous sequential statistic with mean one half) by repeatedly tossing a biased coin of unknown bias.

2.1 LOWER BOUNDS FOR EXPECTED SAMPLE SIZE.

Following Hoeffding's formulation (1960a,b), let $X_1, X_2, \ldots$ be a sequence of independent random variables having a common probability density $f$ with respect to a $\sigma$-finite measure $\mu$. One of two decisions, $d_1$ and $d_2$, is to be made. Let $f_1$ and $f_2$ be two probability densities such that $d_2(d_1)$ is considered as wrong if $f = f_1(f_2)$. Consider the class $C$ of sequential tests (decision rules) for making decision $d_1$ or $d_2$ such that the probability of a wrong decision is at most a (small) specified constant $\alpha_i$ when $f = f_i(i = 1, 2)$. Hoeffding produced various lower bounds for $E_0N$, the expected sample size when $f$ is $f_0$, where $f_0$ is in general different from $f_1$ and $f_2$. The statistical need for these lower bounds is described in the next several paragraphs.

Suppose $f$ depends on a real parameter $\theta$ and $f_i$ corresponds to the value $\theta_i(i = 0, 1, 2)$ with $\theta_1 < \theta_2$. Suppose further that decision $d_1$ or $d_2$ is preferred, respectively, if $\theta \leq \theta_1$ or $\theta \geq \theta_2$, and that neither is strongly preferred if $\theta_1 < \theta < \theta_2$. In this context, it is reasonable to require the probability of a wrong decision never to exceed $\alpha_1(\alpha_2)$ when $\theta \leq \theta_1(\theta \geq \theta_2)$. As Hoeffding observed, any sequential test meeting these requirements automatically meets the requirements described in the previous paragraph (i.e., belongs to $C$).
The converse is also true in many important cases. For instance, suppose the iid observations $X_1, X_2, \ldots$ are normally distributed with unknown mean $\theta$ and known variance $\sigma^2$, hereafter to be referred to as the "canonical example". Then the sequential probability ratio test (SPRT) for testing between the simple hypotheses $\theta = \theta_1 (f = f_1)$ and $\theta = \theta_2 > \theta_1 (f = f_2)$, with error probabilities $\alpha_1$ and $\alpha_2$, has an increasing power function in $\theta$. Consequently, the probability of decision $d_2 (d_1)$ never exceeds $\alpha_1 (\alpha_2)$ when $\theta \leq \theta_1 (\theta \geq \theta_2)$.

According to Abraham Wald and Jacob Wolfowitz (1948), an SPRT is optimal in the sense that $E_0 N$ is minimised when $\theta_0 = \theta_1$, and also when $\theta_0 = \theta_2$, among tests in the class $C$. (Somewhat stronger statements are possible. See Simons (1976).) The real difficulty is posed by $\theta_0$ in the interval $(\theta_1, \theta_2)$, where there is no strong preference for either decision, $d_1$ or $d_2$, but where $E_0 N$, for the SPRT, can be quite large.

Jack Kiefer and Lionel Weiss (1957) showed that the class of sequential tests which minimize $E_0 N$ for various specified pairs $\alpha_1$ and $\alpha_2$ is essentially the class of Bayes solutions to the problem of minimizing the linear combination

$$
\xi_0 E_0 N + \xi_1 P(d_2 | f = f_1) + \xi_2 P(d_1 | f = f_2)
$$

for various specified prior probabilities $\xi_i, i = 0, 1, 2$, adding to unity. Moreover, these Bayes solutions can be found within the class of "generalized SPRT's" (GSPRT's) when the densities $f_i, i = 0, 1, 2$, are members of a Koopman–Darmois family, and, hence, for the canonical example.

The stopping time for a GSPRT assumes the form

$$
N = \inf \{ n \geq 1 : \lambda_n \notin I_n \}
$$

where $\lambda_n$ is the likelihood ratio for data $X_1, \ldots, X_n$ and the pair of densities $f = f_i, i = 1, 2$, and where $I_n$ is a positive interval, usually an open interval $(a_n, b_n), n \geq 1$. (Degenerate intervals are describable by setting $b_n = a_n$, thus permitting truncated stopping times.)

While Kiefer and Weiss provide some helpful theoretical information about the intervals $(a_n, b_n)$, it is inadequate; the practitioner is still left with a very nasty numerical problem to solve. Future development of the subject depended on finding good approximations: GSPRT's, possessing easy descriptions, which come close to minimizing $E_0 N$.

In order to assess potential good approximations, it is very helpful to know, fairly accurately, how small $E_0 N$ can be made for all sequential tests in the class $C$. General lower bounds for $E_0 N$ provide one means of addressing this question.

Hoeffding (1960a) described three different lower bounds for $E_0 N$: under the reasonable assumption $\alpha_1 + \alpha_2 < 1$,

$$
E_0 N \geq \frac{1 - \alpha_1 - \alpha_2}{1 - \int \min(f_0, f_1, f_2) \, d\mu};
$$  \hspace{1cm} (2.1)
\[ E_0 N \geq \sup_{0 < c < 1} \frac{-\log(\alpha_1, (1 - \alpha_2)^{1-c} + (1 - \alpha_1)^c \alpha_2^{1-c})}{c \int f_0(\log(f_0/f_1))d\mu + (1 - c) \int f_0(\log(f_0/f_2))d\mu}; \quad (2.2) \]

and

\[ E_0 N \geq \frac{\{[(\tau/4)^2 - \zeta \log(\alpha_1 + \alpha_2)]^{\frac{1}{2}} - \tau/4\}^2}{\zeta^2} \quad (2.3) \]

where

\[ \zeta = \max(\zeta_1, \zeta_2) \]  
\[ \zeta_i = \int f_0 \log(f_0/f_i)d\mu \]  
\( i = 1, 2 \),

and

\[ \tau^2 = \int \{\log(f_2/f_1) - \zeta_1 + \zeta_2\}^2 f_0d\mu. \]

Apparently, Hoeffding never noticed that (2.3) is equivalent to

\[ E_0 N \geq \min_{i=1,2} \frac{\{[(\tau/4)^2 - \zeta_i \log(\alpha_1 + \alpha_2)]^{\frac{1}{2}} - \tau/4\}^2}{\zeta^2}, \quad (2.4) \]

because the right side of (2.3) is decreasing in the variable \( \zeta \), and \( \zeta = \max(\zeta_1, \zeta_2) \).

While equality in (2.1) is obtainable, Hoeffding asserts that it is inferior to (2.2) and (2.3) in "most of the more common cases," and, by means of an asymptotic argument, he shows it can be of the wrong order of magnitude.

For \( f_0 = f_1 \) and \( c \to 1 \), (2.2) reduces to Wald's well known lower bound for \( E_1 N \):

\[ E_1 N \geq \frac{\alpha_1 \log(\alpha_1/(1 - \alpha_2)) + (1 - \alpha_1) \log((1 - \alpha_1)/\alpha_2)}{\int f_1 \log(f_1/f_2)d\mu}, \quad (2.5) \]

and a similar inequality holds for \( E_2 N \). Thus (2.2) can be expected to be quite good when \( f_0 \) is near \( f_1 \) or \( f_2 \) (i.e., when \( \theta_0 \) is close to either \( \theta_1 \) or \( \theta_2 \)).

Hoeffding's lower bound (2.3) has generated the most interest among statisticians despite its inferiority to (2.2) when \( f_0 \) is near \( f_1 \) or \( f_2 \). For the canonical example,

\[ f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(x - \theta_i)^2}{2\sigma^2}\right\}, \quad i = 0, 1, 2, \]

\[ \zeta_i = \frac{(\theta_i - \theta_0)^2}{2\sigma^2} \]  
\( i = 1, 2 \), and \( \tau = \frac{|\theta_2 - \theta_1|}{\sigma} \) (independent of \( \theta_0 \)).

Thus,

\[ E_0 N \geq \min_{i=1,2} \frac{\sigma^2((\theta_2 - \theta_1)^2 - 8(\theta_i - \theta_0)^2 \log(\alpha_1 + \alpha_2))^{\frac{1}{2}} - |\theta_2 - \theta_1|)^2}{4(\theta_i - \theta_0)^4}; \quad (2.6) \]
and, for fixed $\theta_1$ and $\theta_2$, this is maximized when $\theta_0 = (\theta_1 + \theta_2)/2$, at which (2.6) becomes

$$E_0N \geq \frac{4\sigma^2([1 - 2 \log(\alpha_1 + \alpha_2)]^{\frac{1}{2}} - 1)^2}{(\theta_2 - \theta_1)^2}$$

(2.7)

The lower bound (2.3) (more specifically, (2.7)) was used by T. W. Anderson (1960) to show that his modification to the SPRT effectively reduces $E_0N$. Of course, in the process of showing the worth of his modification, Anderson was at the same time providing convincing evidence of the worth of Hoeffding's lower bound (2.3).

Anderson's real interest was in minimizing the maximum expected sample size over the full range of $\theta$, not simply at a single point $\theta_0$ in the interval $(\theta_1, \theta_2)$. As one might guess, the two problems are equivalent when the testing problem is symmetrical: $\theta_2 = -\theta_1$ and $\alpha_2 = \alpha_1$. What matters then is the expected sample size at $\theta_0 = (\theta_1 + \theta_2)/2 = 0$. For this, Anderson found it convenient to use a symmetric triangular stopping boundary. In discussing this application, David Siegmund (1985) makes the following observations: "An examination of the preceding argument (a derivation of (2.6)) shows that in the symmetric triangular case only the Schwartz inequality (used in Hoeffding's derivation of (2.3)) fails to be an equality. Since the Schwartz inequality is in general rather crude, it is surprising that the lower bound is so close to the correct value" (parenthetical remarks added).

While Anderson's calculations of power and expected sample size are merely approximations based upon probabilistic formulas for Brownian motion, his approximations are quite good, and his basic conclusions are still completely sound. Similar calculations (based on solutions of the heat equation satisfying appropriate boundary conditions) had already been worked out by Hoeffding's Ph.D. student, Thomas Donnelly (1957). Given the complexity of their formulas, it is not too surprising that Anderson failed to reconcile his own and Donnelly's formulas.

Gary Lorden (1967 and 1976) was well aware of the importance of Hoeffding's lower bound (2.3), and he used it when justifying his "2-SPRT" (a specific, theoretically appealing version of the modified SPRT's described by Anderson).

T.L. Lai, Herbert Robbins and David Siegmund (1983) were the first, apparently, to use (2.3) in a purely analytical context, to derive an asymptotic conclusion. T.L. Lai (1988, see Lemma 11) used it again for a similar purpose.

W. J. Hall (1980) introduced the class of "sequential minimum probability ratio tests", commenting in the summary of his paper: "The method of test construction is essentially implicit in a paper of Wassily Hoeffding's (1960),
in which he developed a lower bound on the ASN of sequential tests at an intermediate hypothesis. A subset of them have been independently introduced by Lorden (1976) — his 2-SPRT’s.” Hall is referring to Hoeffding’s lower bound (2.3).

Simons (1967) derived various generalizations of (2.3) and successfully used them for specific numerical applications.

Thus, to conclude: It is abundantly clear that Hoeffding’s lower bound (2.3) has contributed significantly to the development of sequential analysis.

In addition to the lower bounds (2.1)–(2.3), Hoeffding discovered a useful but, in fact, trivial generalization of Wald’s lower bound (2.5) (probably in the middle sixties), which he allowed Robert Bechhofer, Jack Kiefer and Milton Sobel (1966, pages 33–35 and 151–156) to use, and to publish: Let \( f = f_i \) for some \( i = 1, ..., S \) \( (S \geq 2) \), and allow decision rules \( d_1, ..., d_R \). Finally, let \( \alpha_{rs} \) be the probability of decision \( d_r \) given \( f = f_s \), \( \sum_{r=1}^{R} \alpha_{rs} = 1 \) for each \( s = 1, ..., S \). Then

\[
E_i N \geq \max_{s \neq i, 1 \leq s \leq S} \left( \frac{\sum_{r=1}^{R} \alpha_{ri} \log (\alpha_{ri}/\alpha_{rs})}{\int f_i \log(f_i/f_s) d\mu} \right), \quad 1 \leq i \leq S. \tag{2.8}
\]

Simons (1967) rediscovered (2.8) and used it to obtain other lower bounds for expected sample sizes in a variety of contexts. The most interesting example of these, for the present discussion, is a lower bound closely related to (2.2) (in the same context of three densities \( f_0, f_1, \) and \( f_2 \), two decisions \( d_1 \) and \( d_2 \), and with controls of error, \( \alpha_1 \) and \( \alpha_2 \), when \( f = f_1 \) and \( f_2 \), respectively):

\[
E_0 N \geq \inf_{0 < \alpha < 1} \max_{i = 1, 2} \left( \frac{\alpha \log \frac{\alpha}{1-\alpha} + (1-\alpha) \log \frac{1-\alpha}{\alpha}}{\int f_0 \log(f_0/f_i) d\mu} \right). \tag{2.9}
\]

Surprisingly, the right sides of (2.2) and (2.9) are equal! Since the maximum in (2.9) is a convex function of \( \alpha \), the right side of (2.9) is the easier version to evaluate numerically. Generalizations of (2.2) and (2.9), which remain mathematically equivalent, appear in Simons (1967).

Finally, it should be mentioned that Hoeffding (1960a) described, in Section 8, “a sequence of increasingly better lower bounds for \( E_0 N \), arising indirectly from “a sequence of increasingly better lower bounds for the average risk of a general sequential procedure” (described in Section 7), “similar to bounds obtained by Blackwell and Girshick”, “obtained as a consequence of results of Wald and Wolfowitz (1950) which are also contained in Wald’s book (1950)” (parenthetical remarks in quotes modified for present needs). Thus, Hoeffding’s approach in this part of his paper is fundamentally decision-theoretic.
The only lower bound for $E_0N$ explicitly worked out by Hoeffding in Section 8 is equivalent to (2.1), something he and others might have found discouraging. Whatever the case, this part of Hoeffding's work has been ignored, and perhaps unwisely neglected.

2.2 UNBIASED COIN TOSSING WITH A BIASED COIN

Wassily Hoeffding drew great pleasure from what he viewed as his most enjoyable research project. His interest in the subject of unbiased coin tossing began when, as President of the Institute of Mathematical Statistics, he was obliged to make an appointment choice between two possible candidates. Perhaps because he didn't really want to burden either candidate with the job, he entertained the idea of flipping a coin. But how could one avoid a possible bias in the coin? A casual conversation with his colleague Norman Johnson directed him to von Neumann's (1951) simple but ingenious solution: Flip the biased coin twice at a time until a "head-tail" or "tail-head" occurs. Then either can be viewed as a "head" occurring with probability one half. No knowledge of the original bias is required.

The von Neumann procedure requires $p^{-1}q^{-1}$ coin tosses on average to come to a decision, where $p (0 < p < 1)$ is the probability the biased coin generates a head and $q = 1 - p$. Are there more efficient procedures (requiring fewer coin tosses on average)? This intriguing question was Hoeffding's starting point.

A procedure must do two things:

(a) decide the number of coin tosses, i.e., describe a stopping rule;

(b) based solely on the data generated in step (a), define a Bernoulli random variable with mean one half (a statistic, not depending on $p$).

Obviously, not every stopping rule makes step (b) possible. It is easy to see that at least one head and one tail are necessary. But this condition is not sufficient, something Hoeffding was finally able to demonstrate — by means of a very clever argument. Had this condition been sufficient, Hoeffding's question would have led to a simple conclusion, and probably the publication of a very short note.

Hoeffding and Simons (1970) focused attention on Markovian stopping rules, stopping rules define by a prescribed set $S$ of positive integer pairs $(i, j)$ which dictate stopping as soon as the current numbers of tails and heads reach $i$ and $j$, respectively. By plotting the current number of heads against the current number of tails one generates a "sample path" beginning at the origin $(0, 0)$ and terminating at a "stopping point" $(i, j)$ in $S$. Every point $(i, j)$ in $S$ is required to be the terminus of at least one realizable sample path. The von Neumann procedure is based upon a Markovian stopping rule.
Before I became involved with the problem, Hoeffding had already developed the notion of an “even procedure”, a procedure for which step (b) above is obvious. A stopping set \( S \) produces an even procedure if there exists an even number of (equally likely) distinct sample paths terminating at \((i, j)\) for each \((i, j)\) in \( S \). For each stopping point, half the sample paths can be viewed as generating a “head”, the other half a “tail”, thereby fulfilling step (b).

While the von Neumann procedure is even, it is not the most efficient of the even procedures. A best (most efficient) even procedure does exist. It is defined by the stopping set

\[
S = \left\{ (i, j) : \text{the binomial coefficient} \binom{i + j}{i} \text{ is even} \right\}. \tag{2.10}
\]

Since each stopping point in this set is approachable by precisely two distinct sample paths, no more efficient even procedure is possible. (Indeed, the stopping set in (2.10) contains the stopping set of every other even procedure.)

The procedure defined by (2.10) requires

\[
2 \prod_{r=1}^{\infty} \left( 1 + p^{2r} + q^{2r} \right) \tag{2.11}
\]

coin tosses on average to come to a decision. This product is minimized when \( p = \frac{1}{2} \), where it assumes the value 3.40 approximately. This represents a 15% improvement over the von Neumann procedure, which requires 4 tosses on average when \( p = \frac{1}{2} \).

A nagging question remained: Is the best even procedure the best possible (most efficient) procedure? Since we expected a positive answer, we were surprised to discover an explicit improvement (called \( Q_3 \)) on the best even procedure, which stops as soon or sooner than the best even procedure, yielding about a 9% improvement when \( p = \frac{1}{2} \).

The required number of coin tosses by \( Q_3 \), on average, can be reduced further, selectively for \( p \) sufficiently close to zero, but with some diminution of efficiency when \( p \) is large. While the existence of a single uniformly better procedure than \( Q_3 \) (more efficient for every value of \( p \)) seems fairly unlikely, Quentin Stout and Bette Warren (1984) showed that such an improvement is possible with the use of a suitable non-Markovian stopping time.

One of the fun aspects of this problem for us was finding the formula shown in (2.11) for the expected number of tosses required by the best even procedure, and a related formula for \( Q_3 \). Recursive arguments yield the desired results.

An intriguing unsolved question remains: Can the class of Markovian stopping rules that permit step (b) (above) be suitably characterized? Such a
characterization could be expected to shed additional light on the efficiency question.

It should be mentioned that Paul Samuelson (1968) showed that a variant of the von Neumann procedure could be used when the input stream of Bernoulli random variables arises from a Markov chain with unknown parameters. His procedure, and a more efficient one suggested by John Pratt, produce output streams of independent Bernoulli random variables with common mean one half. More recent work of this sort has been described by Manuel Blum (1986).

The Hoeffding–Simons paper spawned activity in a variety of new directions:

(i) Meyer Dwass (1972) works with an input stream of iid discrete random variables having an unknown distribution and focuses attention on the production of a discrete random variable with \( r \geq 2 \) equiprobable values, emphasizing "even procedures" when \( r = 2 \), and a close analogue to the even procedures when \( r \) is a prime number.

(ii) Similar to Dwass' work, Jacques Bernard and Gérard Letac (1973), in addition, discuss properties modulo \( r \) of the multinomial coefficients when \( r \) is the power of a prime \( P \), thereby generalizing and extending results for \( r = 2 \) appearing in Hoeffding and Simons (1970). A typical result: When the binomial coefficient \( \binom{n}{i} \) is divisible by 2 for some pair \( (i,j) \), then the same is true of the pair \( (iP + a, jP + b) \) for \( |a| < P, |b| < P \).

(iii) Peter Elias (1972), James Lechner (1972) and (according to Elias) J. Gill (unpublished) described techniques for producing a stream of iid Bernoulli variables with mean one half from a stream of iid Bernoulli variables with unknown mean \( p \) with the ratio of output variables to input variables approaching the Shannon information-theoretic upper bound \[-p \log_2 p - q \log_2 q\]. Thus, it is possible to obtain a long term efficiency close to 1 when \( p \) is close one half. For comparison, the repetitious use of a Markov stopping time can yield a long term efficiency of slightly less than one third when \( p \) is close to one half. The additional efficiency is accomplished by taking advantage of the fact that, when many output variables are desired, one can use the input variables more flexibly (essentially, to produce blocks of output variables). In addition, Elias discusses similar issues for the context, envisioned by Paul Samuelson (1968), of a Markov chain input stream, and shows that comparable efficiencies are achievable.

(iv) Quentin Stout and Bette Warren (1984) described some improvements that are possible with tree-structured (not necessarily Markovian) stopping times.

(v) Paul Camion (1974) described a link between the present topic and recurrent events, drawing heavily on "some algebraic methods" appearing in lecture notes by M. P. Schützenberger at Toulouse University in 1965, which was partly published in (1961).
3. References


Wassily Hoeffding's Work in the Sixties

Kobus Oosterhoff and Willem R. van Zwet

1. Introduction

The nineteen sixties were a very significant period for Wassily Hoeffding's research. Never a prolific writer but rather a careful polisher, he published eight papers during this decade. More importantly, he developed a number of significant new ideas. Here we shall discuss four papers, three of which were published in 1963, 1964 and 1967. These three are landmark papers dealing with probability inequalities, optimal tests for the multinomial distribution and large deviations. The fourth paper was never published and only appeared as a technical report in 1961. It introduced what is now called Hoeffding's decomposition, which may well make this unpublished paper one of his major contributions.

We shall discuss these papers in chronological order, thus starting with the unpublished technical report.

2. Hoeffding's decomposition

The technical report entitled "The strong law of large numbers for U-statistics" appeared in July 1961 as Institute of Statistics Mimeograph Series No. 302 of the University of North Carolina at Chapel Hill. The title page announces somewhat threateningly that in order to obtain the report, "Department of Defense contractors must be established for the ASTIA (Document Center) services, or have their "need-to-know" certified by the cognizant military agency of their project or contract". Of course such restrictions were customary at the time, but this particular report really became one of the best kept secrets in the recent history of statistics. It was never published and few people were aware of its existence. Though it

---

1 Free University of Amsterdam, University of Leiden & University of North Carolina at Chapel Hill
was quoted from time to time, copies were almost unavailable. One of the present authors managed to find a copy in Wassily Hoeffding's office when visiting Chapel Hill in 1989.

The problem that Hoeffding set out to solve in this paper is as follows. Let \( X_1, X_2, \ldots \) be i.i.d. random variables, \( h \) a symmetric function of \( r \) variables with \( E|h(X_1, \ldots, X_r)| < \infty \), and \( U_n \) a U-statistic defined for \( n \geq r \) by

\[
U_n = \frac{1}{\binom{n}{r}} \sum_{1 \leq i_1 < i_2 < \cdots < i_r \leq n} h(X_{i_1}, \ldots, X_{i_r}). \tag{2.1}
\]

Prove that \( U_n \) converges to \( Eh(X_1, \ldots, X_r) \) almost surely as \( n \to \infty \). This was known to be true when \( h(X_1, \ldots, X_r) \) has a finite moment of order higher than \( 2 - (1/r) \) (cf. Sen (1960)).

Hoeffding established this result by showing that \( U_n \) can be written as a linear combination of \( r \) random variables with a martingale property, and putting martingale technique to work. Apparently unaware of this technical report, the same result was proved in Berk (1966) as an immediate consequence of the reverse martingale property of U-statistics. In Sen (1977) the result is extended to so-called generalized U-statistics under a slightly stronger moment condition. It seems plausible that Hoeffding became aware of Berk's proof at an early stage and decided not to publish his own, more involved proof. However, the main interest of the paper is not the result but the method of proof. Though it may not be optimal for the particular problem at hand, it contains an idea that has proved extremely fruitful in many other situations.

Let \( T = t(X_1, X_2, \ldots, X_n) \) be a function of the independent random variables \( X_1, \ldots, X_n \), with \( ET^2 < \infty \), and let \( \Omega = \{1, 2, \ldots, n\} \). Let \(|A|\) denote the cardinality (i.e. the number of elements) of a set \( A \), and define, for subsets \( A \) and \( D \) of \( \Omega \),

\[
E(T|A) = E(T \mid X_i : i \in A), \tag{2.2}
\]

\[
T_D = \sum_{|A|=0}^{|D|} (-1)^{|D|-|A|} E(T \mid A). \tag{2.3}
\]

Thus \( E(T|A) \) denotes the conditional expectation of \( T \) given those \( X_i \) with index \( i \in A \). In particular \( E(T|\emptyset) = ET \) and \( E(T|\Omega) = T \). Next, \( T_D \) is defined by (2.3) as an alternating sum of \( E(T|A) \) over all subsets \( A \subset D \) including the empty set as well as \( D \) itself. For \( D = \emptyset \), we find \( T_\emptyset = E(T|\emptyset) = ET \).

Equation (2.3) expresses \( T_D \) in terms of the conditional expectations
\( E(T|A) \). It is easy to see that there is also an inverse relation

\[
E(T|A) = \sum_{D \subseteq A} T_D. \tag{2.4}
\]

This is known to algebraists as Möbius' inversion formula. For \( A = \Omega \) it yields

\[
T = \sum_{D \subseteq n} T_D \tag{2.5}
\]

which is nowadays called Hoeffding's decomposition of the random variable \( T \). One easily verifies that the components \( T_D \) of \( T \) have the very special property

\[
E(T_D|D') = 0 \text{ unless } D \subset D', \tag{2.6}
\]

and hence in particular

\[
ET_D = 0 \text{ if } D \neq \emptyset, \tag{2.7}
\]

\[
ET_D T_{D'} = 0 \text{ if } D \neq D'. \tag{2.8}
\]

Thus Hoeffding's decomposition represents \( T \) as the sum of its expectation \( T_k \) and uncorrelated, mean zero random variables \( T_D \) for \( D \neq \emptyset \). By (2.4) and (2.5) this implies that

\[
\sigma^2(E(T|A)) = \sum_{\substack{D \subseteq A \\mid |D| = 1}} ET_D^2, \tag{2.9}
\]

\[
\sigma^2(T) = \sum_{\substack{D \subseteq n \\mid |D| = 1}} ET_D^2. \tag{2.10}
\]

The decomposition simplifies considerably in the special case where \( X_1, \ldots, X_n \) are not only independent but also identically distributed, and \( t \) is a symmetric function of its \( n \) arguments. For \( D = \{i_1, \ldots, i_k\} \), we find that

\[
T_D = \psi_k(X_{i_1}, \ldots, X_{i_k}), \tag{2.11}
\]

where \( \psi_k \) is symmetric in its \( k \) arguments and depends on \( D \) only through its cardinality \( |D| = k \). Hoeffding's decomposition now assumes the form

\[
T = \sum_{k=0}^{n} \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \psi_k(X_{i_1}, \ldots, X_{i_k})
\]
\[
= ET + \sum_{i=1}^{n} \psi_1(X_i) + \sum_{1 \leq i < j \leq n} \psi_2(X_i, X_j) \\
+ \sum_{1 \leq i < j < k \leq n} \psi_3(X_i, X_j, X_k) + \cdots + \psi_n(X_1, \ldots, X_n),
\]

and the variance decompositions (2.9) and (2.10) become

\[
\sigma^2(E(T|A)) = \sum_{k=1}^{[A]} \binom{[A]}{k} E \psi_k^2(X_1, \ldots, X_k), \quad (2.13)
\]

\[
\sigma^2(T) = \sum_{k=1}^{n} \binom{n}{k} E \psi_k^2(X_1, \ldots, X_k). \quad (2.14)
\]

For further properties of Hoeffding's decomposition the reader is referred to van Zwet (1984).

In his report, Hoeffding wrote down the decomposition on the right in (2.12) for the case where \( T \) is a \( U \)-statistic with a symmetric kernel function \( h \) as in (2.1). On the one hand this is a special case, but on the other it contains the result for the general case of an arbitrary symmetric function \( T = t(X_1, \ldots, X_n) \), since this is a \( U \)-statistic of order \( r = n \). The martingale property used by Hoeffding is an immediate consequence of the decomposition.

It turns out, however, that Hoeffding's decomposition has many more interesting applications. In the first place, it is an inexhaustible source of variance inequalities. The recipe is simple. To show that \( \sigma^2(T) \) is larger (or smaller) than a given bound, one tries to express this bound as a linear combination of the \( ET_D^2 \) for \( 1 \leq |D| \leq n \) and show that the coefficients of the \( ET_D^2 \) in this expression are smaller (larger) than 1. Perhaps the best known example of this technique is the Efron–Stein inequality which states that the jackknife estimate of \( \sigma^2(T) \) is biased upwards (cf. Efron and Stein (1981)). For other examples see Karlin and Rinott (1982) and van Zwet (1984).

Perhaps even more importantly, Hoeffding's decomposition is a major tool in asymptotic statistics. Because of the orthogonality (2.8) of the \( T_D \),

\[
\hat{T} = \sum_{|D| \leq 1} T_D = ET + \sum_{i=1}^{n} T_{\{i\}} = ET + \sum_{i=1}^{n}[E(T|X_i) - ET]
\]
is the \( L^2 \)-projection of \( T \) on the space of sums of functions of single variables \( X_i \). This means that among all random variables \( L = \sum f_i(X_i), \hat{T} \) minimizes \( E(T - L)^2 \). Also, because of (2.8) and (2.10),
\[ E(T - \hat{T})^2 = \sigma^2(T) - \sigma^2(\hat{T}). \] (2.15)

Now consider a sequence of statistics \( T_n = t_n(X_1, \ldots, X_n) \) with \( \sigma^2(T_n) = 1 \) and projections \( \hat{T}_n \). It follows from (2.15) that if

\[ \sigma^2(\hat{T}_n) \to 1, \] (2.16)

then \( T_n \) and \( \hat{T}_n \) are asymptotically equivalent. Typically, \( \hat{T}_n \) will be asymptotically normal and \( T_n \) will have the same limit distribution.

This type of proof of asymptotic normality is commonly known as Hájek's projection method. It has the advantage over older methods based on Taylor expansion of \( T_n \) that no smoothness assumptions on the functions \( t_n \) are needed. Of course we have no guarantee that the higher order terms in the Hoeffding decomposition will be asymptotically negligible so that (2.16) will hold, but this is all we have to check. Also we have the assurance that if \( E(T_n - L_n)^2 \to 0 \) for some \( L_n = \sum f_{i,n}(X_i) \), then this will be true for \( L_n = \hat{T}_n \).

If one wishes to obtain more precise information about the distribution of \( T_n \) than a limit distribution can provide, one clearly has to take higher order terms in Hoeffding's decomposition into account. In van Zwet (1984) it is shown that Hoeffding's decomposition produces a Berry–Esseen bound for the distribution of \( T_n \) under (conditional) moment conditions only. More generally, it appears that approximation of a statistic by \( U \)-statistics through Hoeffding's decomposition is a powerful tool to obtain Edgeworth expansions.

3. Probability inequalities

Of Wassily Hoeffding's papers, the one that is referred to most often to the present day is probably "Probability inequalities for sums of bounded random variables", \textit{J. Am. Statist. Assoc.} 58 (1963), 13–30. The first group of bounds in the paper concern the mean \( \bar{X} = (X_1 + \ldots + X_n)/n \) of \( n \) independent random variables \( X_1, \ldots, X_n \), each of which has expected value 0 and is bounded above and below, thus \( a_i \leq X_i \leq b_i \) for \( i = 1, \ldots, n \). We write \( \mu \) and \( \sigma^2/n \) for the expectation and the variance of \( \bar{X} \). Hoeffding shows that

\[ P(|\bar{X} - \mu| \geq t) \leq 2 \exp \left\{ \frac{-2(nt^2)}{\sum_{i=1}^{n} (b_i - a_i)^2} \right\} \] (3.1)

This remains true if the \( (X_i - EX_i) \) are martingale summands. The bound (3.1) is known as Hoeffding's inequality.
Hoeffding also gives sharper but somewhat more complicated bounds, as well as bounds involving the variance \( \sigma^2 \). These bounds are then carefully compared. There are also inequalities for the general dependent case and the case of \( m \)-dependence. The inequalities hold for arbitrary sample sizes and, as usual in Hoeffding’s work, the bounds contain no unknown constants. Everything has been explicitly calculated and the proofs are all elementary.

Finally, there is the well-known result that if \( S \) and \( T \) denote the sample sums of random samples of the same size drawn with and without replacement from the same population, then for any convex function \( f \),

\[
Ef(S) \geq Ef(T).
\]

This confirms our intuitive feeling that in some appropriate sense, sampling with replacement produces more spread than sampling without replacement. The result was generalized by Rosén (1967) to the case where \( S \) denotes the sample sum for any symmetric sampling plan, i.e. a plan that is invariant under permutation of the population elements. Karlin (1974) generalized Rosén’s result to a more general class of functions of the sample values and gave some further extensions as well as a conjecture which generated considerable research activity. See van Zwet (1982), Krafft and Schaefer (1984), Schaefer (1987), Bhandari (1987) and Klaassen (1990).

Let us return to Hoeffding’s inequality (3.1) which is easily the most popular inequality in the paper. In recent years it has played an important role in the theory of empirical processes indexed by sets and functions. In this context, it is usually applied in combination with a counting argument as well as a symmetrization procedure (cf. Pollard (1984), section II.3). We shall demonstrate the power of Hoeffding’s inequality coupled with symmetrization in a much simpler case. The problem we consider is that of bounding the tail of Student’s distribution for symmetrically distributed underlying random variables.

Let \( X_1, \ldots, X_n \) be i.i.d. random variables which are distributed symmetrically about 0. Suppose that \( P(X_i = 0) = 0 \) and define Student’s \( T \) as

\[
T = n^{\frac{1}{2}} \frac{\hat{X}}{\hat{S}},
\]

where

\[
\hat{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \hat{S}^2 = \frac{\sum_{i=1}^{n} (X_i - \hat{X})^2}{(n-1)}.
\]

For \( t > 0 \), it is easy to see that \( |T| \geq t \) if and only if
\[ |\bar{X}| \geq t \left( \frac{\sum_{i=1}^{n} X_i^2}{n(n - 1 + t^2)} \right)^\frac{1}{2}. \]

Conditionally on \(|X| = (|X_1|, \ldots, |X_n|)|, the random variables \(X_1, \ldots, X_n\) are independent with expectation 0, and \(X_i\) is bounded in absolute value by \(|X_i|\) for \(i = 1, \ldots, n\). Hence Hoeffding's inequality (3.1) applies to this conditional situation with \(b_i = -a_i = |X_i|\) and we obtain

\[
P(|T| \geq t \mid |X|) = P\left( |\bar{X}| \geq t \left( \frac{\sum_{i=1}^{n} X_i^2}{n(n - 1 + t^2)} \right)^\frac{1}{2} \mid |X| \right)
\leq 2 \exp \left\{ -\frac{nt^2}{2(n - 1 + t^2)} \right\}.
\]

Since the right-hand side is independent of \(|X|\), we find that

\[
P(|T| \geq t) \leq 2 \exp \left\{ -\frac{nt^2}{2(n - 1 + t^2)} \right\} \tag{3.3}
\]

which is the desired result. Note that there is no moment assumption on the \(X_i\); it is only assumed that their distribution is symmetric about 0.

4. Tests for the multinomial distribution

In the paper "Asymptotically optimal tests for multinomial distributions", \(Ann. Math. Statist. 36 (1965), 369–408,\) Hoeffding moved in an entirely new direction. The paper is devoted to substantiating the following proposition within a multinomial setting: "If a given test of size \(\alpha_N\) is sufficiently different from a likelihood ratio (LR) test, then there is a LR test of size \(\leq \alpha_N\) which is considerably more powerful than the given test at most points \(p\) of the set of alternatives when the sample size \(N\) is large enough, provided \(\alpha_N \to 0\) at a suitable rate."

Roughly speaking, the convergence of \(\alpha_N\) to zero must be more rapid than any (negative) power of \(N\); as a result, the ratio of the error probabilities at \(p\) of the LR test with respect to a competing test also vanishes at a rate faster than any power of \(N\). The regularity conditions on the given test are very weak, but there are some other complications; the set of alternatives \(p\) for which the result holds is not very transparent and some elaboration is necessary of what is meant by "sufficiently different tests". In the applications discussed by Hoeffding this can be made explicit.

This type of result is in striking contrast to the more classical approach where the sizes of the tests are kept fixed as \(N \to \infty\) and local alternatives
are primarily considered. With such a local approach the LR test is asymptotically optimal only in a very weak sense, quite different from the strong optimality for non-local alternatives implied by Hoeffding's theorem. As an example Hoeffding compares the chi-squared test and the LR test for a simple hypothesis in some detail. According to local theory both tests are asymptotically equivalent, but Hoeffding shows that for "most" fixed alternatives the LR test is asymptotically superior.

Not only is the statistical approach to testing problems novel, the tools employed by Hoeffding to obtain his results are also untraditional. Local theory is typically based on the central limit theorem and its ramifications. The non-local approach, however, is based on large deviation theorems of the following sort:

\[ P_N(A_N \mid p) = \exp\{-N I(A_N, p) + O(\log N)\} , p \in \Omega , \quad (4.1) \]

where \( A_N \) is a (regular) subset of the simplex

\[ \Omega = \left\{ z \in \mathbb{R}^k : z_1 \geq 0, \ldots, z_k \geq 0, \sum_i z_i = 1 \right\} \]

and \( I(A_N, p) \) is the Kullback-Leibler information number

\[ I(A_N, p) = \inf \left\{ \sum_i z_i \log \left( \frac{z_i}{p_i} \right) : z \in A_N \right\} . \]

In applications, \( A_N \) is the acceptance region or the rejection region of a test. Large deviation theorems of type (4.1) were first obtained by Sanov (1957). There is earlier work of Chernoff (1952) relating a large deviation measure of efficiency to moment generating functions, but this approach has not found recognition until much later. Compared with the central limit theorem, large deviation results like (4.1) are very crude and require few regularity conditions. Of course they can be supplemented by further refinements, but such niceties play no role in the applications dealt with in this paper.

It is not surprising that this pioneering paper attracted much attention and inspired new research in several directions. Oosterhoff and van Zwet (1972) proved in the same multinomial setup that the shortcoming of the LR test of a simple hypothesis vanishes uniformly as \( N \to \infty \) if \( \alpha_N \to 0 \), implying that for the exceptional alternatives appearing in Hoeffding's theorems the gain of a competing test over the LR test is asymptotically very small (but note that shortcomings are concerned with differences, not with ratios of error probabilities). This result was extended to \( k \)-parameter exponential families and composite hypotheses by Kallenberg (1978). A quite general
result in the same spirit was proved by Brown (1971): in general families of distributions satisfying rather stringent regularity conditions it holds that

$$\alpha^LR_N \leq \alpha_N, \quad \limsup N^{-1} \log \alpha_N < 1$$

(4.2)

$$\liminf_N \{N^{-1} \log \beta_N(\theta) - N^{-1} \log \beta^LR_N(\theta)\} \geq 0 \quad \text{for all } \theta \in \Theta_1,$$

(4.3)

where $\alpha_N, \alpha^LR_N, \beta_N$ and $\beta^LR_N$ denote the error probabilities of an arbitrary test and an appropriate LR test, respectively, and $\Theta_1$ denotes the set of alternatives. Hoeffding’s result is stronger in the sense that the above inequality (4.3) is strict for “most” alternatives in multinomial families. An alternative form of Brown’s result is: for fixed $\theta \in \Theta_1$ there exists an appropriate LR test such that

$$\beta^LR_N(\theta) \leq \beta_N(\theta), \quad \limsup N^{-1} \log \beta_N(\theta) < 1,$$

(4.4)

$$\liminf_N \{N^{-1} \log \alpha_N - N^{-1} \log \alpha^LR_N\} \geq 0.$$  

(4.5)

The non-local approach to the performance of tests has been systematically developed by Bahadur in a series of papers starting in the early sixties. The central concept in Bahadur’s approach is the “exact slope” of a test statistic $T_N$ under an alternative $\theta$, i.e. the strong limit

$$c(\theta) = -2 \lim_N N^{-1} \log L_N$$

(4.6)

where $L_N$ denotes the level (or $P$-value) attained by the test statistic $T_N$. Roughly speaking, the exact slope is the exponential rate of convergence of $\alpha_N$ to zero if at a fixed alternative $\theta$ the power is also kept fixed. See Bahadur (1960), (1967), (1971) and in particular his nice monograph (1971). The relation with the work of Hoeffding and Brown is obvious from (4.5). The non-local optimality of LR tests discussed by Hoeffding and Brown is reflected in the property of maximal slope of the LR statistic, which Bahadur proves under conditions reminiscent of those in Brown (1971). This answers the question raised by Chapman in his invited discussion of Hoeffding’s (1965) paper about the relation with the work on slopes by Bahadur (Chapman’s remarks are obscured by a computational error due to the confusion with “approximate slopes” in the early days).

There is vast literature in this area. We mention only a few papers. Tusnády (1977) further explores the relation between exact slopes and large deviations. Sievers (1969) studies a series of examples employing the moment
generating function approach to large deviations. Groeneboom and Oosterhoff (1977) review much of the earlier work. Intermediate approaches where one considers alternatives converging to the null hypothesis at a slow rate are studied in Kallenberg (1981), (1983) and Kallenberg and Ledwina (1987). Connections between local and non-local measures of optimality are very intriguing; a first paper on this subject is Wieand (1976). Another question concerns the the speed of convergence in large deviation limit theorems; there is some evidence that in many instances the convergence is slow. This explains a reluctance to attach too much weight to non-local efficiency measures, cf. Groeneboom and Oosterhoff (1981).

5. Large deviations

The paper "On probabilities of large deviations", Proc. Fifth Berkeley Symp. Math. Statist. Prob. I (1967), 203–219, is the last of the four major papers we discuss here. It is of a different nature than the paper discussed in the previous section. It focuses on more precise estimates of large deviations of empirical distributions of N i.i.d. random vectors. After a survey of known results it concentrates on refinements of Sanov-type large deviation theorems for functions of multinomial vectors. Let $X^{(N)}$ have a $k$-dimensional multinomial distribution and let $f$ be a real-valued function. It is shown for several functions $f$ that

$$P_N(f(X^{(N)}) \geq c|p) \approx CN^{-s} \exp\{-NI(A(c),p)\}$$  \hspace{1cm} (5.1)

for a broad range of values $c$, where $A(c) = \{x : f(x) \geq c\}$ and $s$ depends on the function $f$. If $f$ is a linear combination of the components of $x$ and $f(p) = 0$, then $s = 1/2$; if $f(x) = I(x,p)$, then $s = (k-3)/2$; if $f(x) = \sum_i(x_i - p_i)^2/p_i$, then $s = (k-3)/2$ if $cN^{2/3}$ is bounded above and $s = 1/2$ if $c$ is bounded away from 0.

These results were partially extended to general $d$-parameter exponential families by Efron and Truax (1968) who proved that for non-lattice exponential families and sample sums $X^{(N)}$

$$P_N(I(X^{(N)},\theta) \geq c|\theta) = CN^{-(d-1)/2} \exp\{-Nc\(1 + o(1)\)$$  \hspace{1cm} (5.2)

uniformly in $c$ for $\varepsilon \leq c \leq \max_x I(x,\theta) - \varepsilon$. In the 1-lattice case the last factor must be replaced by $\exp(\mathcal{O}(1))$. They also discuss the accuracy of approximations to the power of the LR test based on such refined large deviation results. Hoadley (1967) further pursued the work of Sanov and Hoeffding on large deviations of functions of empirical distributions to study nonparametric tests. In a more general setting, this line of research was continued by Groeneboom, Oosterhoff and Ruymgaart (1979).
The important statistical applications of large deviation theory, first recognized by Chernoff, Hoeffding and Bahadur, have been a constant source of inspiration for further research in this field. The impressive growth of literature on large deviations bears evidence of this fact.

6. References


