AD-A070 023 CALIFORNIA UNIV BERKELEY STATISTICAL LAB F/G 12/1
SOME MEMORABLE INCIDENTS IN PROBABILISTIC/STATISTICAL STUDIES,(U)
1979 J NEYMAN N00014-75-C-0159
UNCLASSIFIED CU-SL-79-02-ONR NL

END
DATE FILMED
7-79
DDC

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
Statistical Laboratory
University of California
Berkeley, California 94720

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

(6) 3 REPORT TITLE
Some memorable incidents in probabilistic/statistical studies,

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Scientific

(10) 5 AUTHOR(S) *(First name, middle initial, last name)*
Jerzy Neyman

(12) 35p.

(11) 6 REPORT DATE
1979 (15)

7a. TOTAL NO. OF PAGES
32

7b. NO. OF REFS
28

8a. CONTRACT OR GRANT NO.
ONR N000 14-75-C-0159

(14) CU-SL-ONR 79-02-GNR

b. PROJECT NO.

c. DAAG29-76-G-0167

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10 DISTRIBUTION STATEMENT
This document has been approved for public release; its distribution is unlimited.

11 SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research
Washington, D.C. 20014

13 ABSTRACT

This paper has been prepared for delivery as an opening address at a Symposium at Chapel Hill, N.C., intended to honor Professor Wassily Hoeffding. The contents of the paper are summarized in the titles of its five chapters marked with Roman numerals II to VI, as follows: II. The Cramér-Hoeffding research incident (= the importance the theory of large deviations initiated by Cramér for the asymptotic theory of statistical tests). III. Two different strategies in mathematical statistics. IV. The Yule-Pólya research incident: ~~(i)~~ mechanism of a natural phenomenon, and ~~(ii)~~ non-identifiability. V. Some modern recurrences of the Yule-Pólya problem. VI. Effort at an "optimal" competitor to the K.P.'s chi square test. Possibly, the most important unresolved problem is that of the "residual" non-identifiability of the serial sacrifice experimental design, discussed in Chapter V.

DD FORM 1 NOV 65 1473

Unclassified

Security Classification

333 400

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Cramér-Hoeffding						
Mathematical Statistics						
Yule-Pólya						
Mechanism of Natural Phenomenon						
Non-identifiability						
Optimal competitor						
K.P.'s Chi Square Test						

SOME MEMORABLE INCIDENTS IN PROBABILISTIC/STATISTICAL STUDIES

Jerzy Neyman

Statistical Laboratory

University of California, Berkeley, CA 94720

CONTENTS

I. INTRODUCTION

1. Congratulations to Professor Hoeffding. I am very grateful to Professor Chakravarti for his invitation to open the discussion at this Symposium intended to honor Professor Wassily Hoeffding. We met long ago and from the very beginning, it was a pleasure to find a marked similarity in our research interests. After the joint work with Egon S. Pearson [1933] concerned with power functions and later, after the development of the theory of confidance intervals [1937a], my research efforts focused on the deduction of variously defined "optimal" statistical methodologies [1959] that could be easily used in studies of natural phenomena. Against this, here is the title of Professor Hoeffding's paper: "Optimal nonparametric tests" [1951] he delivered at the Second Berkeley Symposium on Statistics and Probability held during the summer of 1950, more than a quarter of a century ago. Since that time our intellectual contacts continued, but our personal encounters were "like Victoria Regina: seldom, seldom in bloom."

Incidentally, the problem of the optimal non-parametric tests of composite statistical hypotheses is still "on the books."

II. THE CRAMÉR-HOEFFDING RESEARCH INCIDENT

2. The Harald Cramér Ground Breaking Paper of 1938.

The mathematical tool most frequently used in the development of statistical methods is the Central Limit Theorem on probabilities, roughly as follows. Let $\{X_n\}$ be a sequence of random variables each having two moments, $EX_n=0$ and $EX^2=\sigma^2<\infty$ and let

$$S_n = \sum_{i=1}^{n} X_i. \tag{1}$$

Then, under certain conditions,

$$\lim_{n\to\infty} P\{S_n \leq t\sigma\sqrt{n}\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2}\, du = \Phi(t) \tag{2}$$

for any preassigned real number t. This theorem preoccupied mathematicians for a couple of centuries now [Loève, 1960]. The successive proofs given differ in the generality of the "certain conditions" just mentioned. This long duration of efforts to prove the validity of formula (2) resulted in the establishment of a "routine of thought." Whenever some particular problems of mathematical statistics involved the consideration of sums of random variables like (1), with the value of n considered "large," it became customary to presume that formula (2) gives a satisfactory approximation of the true distribution of S_n. The word "customary" is not adequate. The breaking of a "routine of thought" stimulates opposition.

Among other things, the classical central limit theorem was used to compare the effectiveness of statistical tests. Here, the term Pitman asymptotic efficiency comes to my mind.

As described by Yu. V. Linnik [1961], the honor of breaking this firmly established routine of thought belongs to Harald Cramér. In 1938, just before the beginning of World War II, there appeared Cramér's paper [1938] offering the first solution to a novel question that Cramér dared to ask. Briefly, it is as follows.

With reference to formula (1) assume that all the variables of the sequence $\{X_n\}$ are mutually independent and identically distributed. Consider the probability

$$F_n(t_n) = P\{S_n \leq t_n\sigma\sqrt{n}\}, \qquad (3)$$

where t_n grows to infinity as n is increased. Cramér's ground breaking question was about the asymptotic behavior of the ratio

$$\frac{1-F_n(t_n)}{1-\Phi(t_n)}, \qquad (4)$$

depending on properties of the variables X_n and on the rate of increase of t_n. This paper generated a new chapter of probability theory, labeled "theory of large deviations" [Linnik, 1961]. Briefly and roughly, the important question was whether $1-\Phi(t_n)$ can be considered as a satisfactory approximation of the probability that the sum S_n will exceed a limit proportional to $t_n\sqrt{n}$.

3. Professor Hoeffding's Initiative to Use the Novel Probabilistic Tool. While it is obvious that Cramér's limit theorem on large deviations must be a better tool for studying the asymptotic properties of statistical tests than is the classical central limit theorem, the disasters and the length of World War II were not conducive to the development of conceptual mathematical subdisciplines. In consequence, the relevance of the Cramér ground breaking work remained unnoticed for almost two decades. Here, a paper by Professor Hoeffding [1965] played a special role.

The title of this paper is: "Asymptotically optimal tests for multinomial distribution." Professor Hoeffding begins by formulating his own definition of asymptotic optimality and then states: "To attack these problems, the theory of probabilities of large deviations is needed." This is followed by proofs that, under specified conditions, certain familiar tests (the likelihood ratio and the chi square tests) are asymptotically optimal in the sense of the new, call it, Hoeffding definition of optimality.

Professor Hoeffding's paper was presented at a meeting of the IMS and the discussion that followed is recorded in the Annals. It appeared that, even though Cramér's theorem on large deviations was familiar to several statisticians, including H. Chernoff, R.A. Wijsman and D.G. Chapman, Professor Hoeffding must be credited with the first serious effort to

see what the novel probabilistic tool can contribute to the theory of asymptotic tests.

Incidentally, published in 1965, fourteen years ago, Hoeffding's paper continues to affect the thinking of this day. The following quote is from a paper published in the last issue of the Zeitschrift für Wahrscheinlichkeitstheorie und varwandete Gebiete [Berk and Jones, 1979]: "The first [lemma] is actually a special case of a theorem of Hoeffding (1965), Theorem 2.1."

My hearty compliments to Professor Hoeffding!

4. Reasons for Preferring the Theory of Large Deviations as a Tool for Studying Asymptotic Tests. The word "preferring" in the title of the present section emphasizes its subjective characters. It has to do with the meaning I attach to the terms "errors of the first and second kinds" possible to commit in testing a statistical hypothesis.

As described in [1977a] in the course of an empirical study one is frequently faced with a two-decision problem. Depending upon the outcome of the statistical test used, one has to decide to go, say, either "right" or "left," and either decision can be erroneous. Depending upon personal attitudes, one of the two errors will be judged more important to avoid than the other. My definition is: the error that is more important to avoid is called the error of the "first kind." In consequence, when selecting a test to be used in a particular empirical study, my first concern is to make sure that the probability of com-

mitting an error of the first kind does not exceed a preassigned level α, now called "level of significance." Depending upon the subjective feeling of importance, the chosen level of significance may be $\alpha=0.10$, or $\alpha=0.05$ or $\alpha=0.01$, etc.

When the problem of the desired level of significance is solved and if it can be ensured by any test of some determined class, the time comes to think of the less important error, the error of the "second kind," which means to determine the most powerful test within the class considered.

This is the background of my preference for the theory of large deviations as a tool in the theory of asymptotic tests as compared with the classical central limit theorem.

In an empirical study involving a two-decision problem, one is faced with some real life situation, with some hypothesis which can be true or false and with the degree of its falsehood measured by a parameter ζ, the value of which is unknown. The only thing that is under our control, at least to some extent, is the number n of observations that can be used to test the hypothesis that $\zeta=0$. The all important question is whether this particular number n is large enough to achieve the chosen level of significance α. The answer depends on how close the ratio (4) is to unity, which is the subject of Cramér's theory of large deviations, including its modern descendants. The use of this theory does not violate the real life situation of the problem, with ζ having some unknown fixed value.

Now consider the asymptotic test possibilities offered by the classical central limit theorem on probabilities. As is well-known, both the Pitman asymptotic efficiency theory and the theories of asymptotic tests developed by Cramér [1928] and by myself [1959] depend on visualizing that the real life problem, say, the problem of testing considered today, is a member of a hypothetical sequence with the fixed unknown ζ replaced by ζ_n, such that the product $\zeta_n\sqrt{n}$ is bounded away from zero and infinity, preferably tending to some known limit. This is something very different from and much less inspiring than the question of how close to unity is the value of (4).

III. TWO DIFFERENT STRATEGIES IN MATHEMATICAL STATISTICS

5. A Curious Detail of the History of Statistical Tests.

The Cramér-Hoeffding research incident described in sections 2 and 3 illustrates a curious detail of the history of statistical tests, particularly of the early history. The customary strategy is composed of two consecutive steps. (i) A statistician concerned with some empirical domain proposes a testing procedure suggested by his intuition. Then, (ii) an effort is made to investigate the properties of this procedure, occasionally leading to the conclusion that it is in some sense "optimal." Examples of this sequence (i)-(ii) are countless.

The first test procedure, still in very frequent use, is the chi square test introduced by Karl Pearson in 1900. It was one of the subjects studied in the Hoeffding paper just discussed. The other test discussed in the same Hoeffding paper is the likelihood ratio test. As stated by Professor Hoeffding, the likelihood ratio criterion was suggested by E.S.P. and myself in 1928. However, this suggestion was made on intuitive grounds. The criterion suggested did not result from a search for a procedure satisfying a defined concept of optimality. The intuitive background of the likelihood ratio test was simply as follows: if among the contemplated admissible hypotheses there are some that ascribe to the facts observed probabilities much larger than that ascribed by the hypothesis tested, then it appears "reasonable" to reject that hypothesis.

As another example, I wish to mention a test criterion competitive to the chi square, first suggested by Harald Cramér [1928] and somewhat later also advanced by Richard von Mises [1931].

The alternative philosophy, or strategy, is just the opposite to the sequence (i) and (ii). When one has to deal with an empirical domain of study and one feels in need of a statistical procedure, it seems natural to visualize the properties that this procedure should have to deserve the description "optimal." Naturally, such concept of optimality can depend

upon the domain of empirical study and it must depend on the subjective preferences of its author. However, once the optimality is defined, the mathematical problem occurs: to find the "optimal," if such exists. On occasion one finds that the initially defined optimal procedures do not exist. Too bad! Then one has to look for a "compromise optimality," etc. One example is the concept of "unbiased most powerful tests" [Neyman and Pearson, 1936]. Here, the word unbiased marks the compromise optimality. In the case considered, the "uniformly" most powerful test does not exist.

IV. THE YULE-PÓLYA RESEARCH INCIDENT: (i) MECHANISM OF A NATURAL PHENOMENON, AND (ii) NON-IDENTIFIABILITY

6. My Contacts with George Udny Yule. During my four year long activities at the Department of Statistics, University College, London (1934-1938), I had the privilege of meeting quite a few outstanding scholars. This included G. U. Yule for whom I developed great respect and warm feelings.

The studies of Yule that attracted my particular attention were preformed jointly with M. Greenwood [1920]. Subsequently, a related paper was published by E. M. Newbold [1928].

My preferred way of describing these studies is as follows: They are concerned with the chance mechanism operating in real life, the mechanism that determines the distribution of an ob-

servable random variable X. If this mechanism is understood, it could be used to solve an important practical problem.

The particular random variable X of the Greenwood-Yule-Newbold studies was the number of accidents per unit of time, per bus driver in London. The important practical problem considered was the means to diminish the frequency of accidents involving the buses. Exactly similar problems are important in the present epoch, even though the actual domain of study can be very different. One example is the question: how can one diminish the frequency of deaths from cancer?

The problem of accidents was studied in our Stat. Lab. in the early 1950's. Here Professor Grace E. Bates played an important role [Bates and Neyman, 1952a, 1952b].

The first of these papers is dedicated to the memory of George Udny Yule and is preceded by a one page biographical sketch. It includes the following passage: "In 1931 Yule felt that he was too old to hold the position of Reader at Cambridge University and retired. At the same time he felt young enough to learn to fly. Accordingly, he went through the intricacies of training, got a pilot's license and bought a plane. Unfortunately, a heart attack cut short both the flying and, to a considerable degree, his scholarly work."

It happened that my personal contacts with Yule were very limited. They occurred during the period when he was recovering from his heart attack. However, these contacts affected my thinking. In particular, they contributed to the formulation

of my paper of 1937[b].

The attempts to decrease the frequency of accidents taking into account the "human factors," mentioned in the title of Miss Newbold's report, are connected with the concept now called "accident proneness." There is little doubt that particular individuals do differ in their proneness to accidents of some specified categories. However, the details of this variability are not clear and here empirical studies are important. During our studies in the early 1950's our thinking was affected by two contrasting hypothetical mechanisms. One of them is the Greenwood-Yule-Newbold (GYN, for short) hypothetical mechanism, the properties of which can be summarized as the "mixture - no contagion - no time effect" mechanism. The other hypothetical mechanism, implied by studies of George Pólya[1930], was just the contrary: "identity of individuals, contagion and time effect."

To be more specific: the GYN mechanism presupposed that the number of accident incurred by a particular individual per unit of time, such as a year, is a Poisson variable with a fixed expectation λ, representing this individual's personal accident proneness, which remains unchanged throughout his active life (= "no time effect"). Another basic assumption is that the value of λ varies from one individual to the next (= "mixture"). More particularly, the assumption was adopted that the variation of λ within a relevent population, such as the population of

actual or potential bus drivers in London, can be adequately represented by a gamma distribution.

Starting with these basic assumptions it was easy to deduce that the number of accidents per year incurred by individual bus drivers must have a negative binomial distribution. Actually, using the data on accidents involving bus drivers it was found that this distribution could be well fitted by a negative binomial so that the GYN mechanism (or shall we call it "model?") appeared to have been "confirmed."

Everything appeared nice and smooth until the Pólya "model" was examined. As described above, this model denied the existence of a "mixture." The basic assumption was that all individuals forming the population of actual or potential employees in a particular industry were "born equal." However, it was assumed that the number of accidents in a time interval [t, t+h), where h is a small positive number, depends upon the number of accidents incurred before time t (= "contagion"). Also, there was the assumption that, as the duration of employment increases, the experience gained may diminish the individual's accident proneness (= "time effect").

Using these specific assumptions suggested by the famous Pólya paper of 1930, it was easy to calculate the distribution of the number of accidents per year in a population comparable to that of the London bus drivers. Because of the contrast between the two hypothetical mechanisms, the GYN and the

Pólya mechanisms. the expectation was that the two distributions would be very different. If this happened, then the empirical data, such as the data resulting from Miss Newbold's study of the London bus drivers could be used to resolve questions like that in the title of our study [1952b]: "true of false contagion?"

When the easy calculations of the relevent probability generating function were performed, Dr. Bates and I experienced a little shock: with reference to a single observational period, such as a year, the Pólya "no mixture - contagion - time effect" model implied that the distribution of the number of accidents per driver must be a negative binomial, coinciding with that implied by the Greenwood-Yule-Newbold model! This finding brought to our minds several ideas that appear important to this day. One is the concept of non-identifiability. The other related idea is that the problem of validation of a hypothetical mechanisms of a natural phenomenon deserves a serious effort. One hopeful possibility is that the non-identifiability of some two (or more) hypothetical mechanisms, the non-identifiability with respect to the distribution of a specific single random variable X, may disappear just as soon as one supplements X by some other appropriately selected variables, say X_1, X_2,...X_s.

The second of our joint papers considers a number of not too difficult empirical studies capable of providing a definitive answer to the all important question about the reality

of "contagion" in accidents. E.g., The identifiability can be achieved by counting accidents of each driver not just in one particular year (say X_1 of them), but also those incurred during the following year, say X_2 of them, etc. See Grace E. Bates [1955].

This section concludes my description of the Yule-Pólya problem as it came to my attention with reference to industrial accidents: what is the governing chance mechanism? Without much risk of exaggeration one may assert that this type of problem is encountered in every serious study of a complex natural phenomenon. In cosmology: what is the chance mechanism governing the dispersal of clusters of galaxies? How can one verify any relevent hypothesis? In public health: what is the mechanism behind the observed geographic variability in the incidence of cancer? Through what experiments and with what statistical methodology can one gain reliable information? In weather modification experiments: what are the processes in the atmosphere that follow "cloud seeding?" What statistical methodology is likely to provide the desired information through the analysis of the many completed experiments?

Here, a remark on terminology seems in order. It seems to me that the common use of the term "model" deserves a modification or restriction. My preference would be to restrict the use of this term to sets of (customarily) qualitative assumptions advanced to explain a natural phenomenon. One example is the GYN model suggested to explain the notorious

driver to driver variability in the number of accidents per year, the "mixture - no contagion - no time effect" model. The same applies to the Pólya "no mixture - contagion - time effect" model. This use of the term "model" appears quite different from the designation of a mathematical formula that fits the observations. One frequently encountered example is the phrase "linear model," etc.

Discussions of the Yule-Pólya dilemma relating to the problem of public health will be found in the next chapter.

V. SOME PRESENT DAY RECURRENCES OF THE YULE-PÓLYA DILEMMA

7. Public Health Policy and Basic Research. The importance and the difficulty of the present day public health problems overshadow those of industrial accidents symbolized by the names of Yule and Pólya. However, the broadly understood research problems remain similar.

One of the typical contemporary public health problems is concerned with the hazards from electricity producing plants [1977b], briefly as follows. A locality L, marked by a rapidly growing population, is in need of a new electricity producing plant. This may be either a nuclear facility or a fossil fuel burning unit and the choice is up to some decision making authorities. Among other things, the choice must be made taking into account some public health questions. Whatever type of

plant is constructed, it will contribute to the local pollution in its own way. The important questions are: how many more cancer cases, heart attacks, etc. are to be expected in this locality L as a result of the predictable extra pollution from the normal operation of the novel electric generator? How can one answer this question reliably?

The reliability of the answer depends upon the understanding of two different mechanisms. One mechanism is concerned with the happenings in experimental animals, mice, dogs, etc., subjected to a specified change in the environmental pollution. The other important mechanism is that of the dependence of the effects of the first mechanism on the identity of the species concerned, whether mouse, or rat, or dog, or man. Obviously, the complexity of the problem is tremendous. It splits itself into a number of subproblems. In the next section, we shall consider one of these subproblems. It involves the ubiquitous phenomenon of non-identifiability.

8. Typical "Survival Experiment" and the Methodology of "Potential Survival Times." The customary source of information on the happenings in the experimental animals, say mice, exposed to some "agents" studied is a "survival experiment." There are two substantial groups of mice, one labeled "experimental" and the other "controls." The experimental mice are exposed to the agents studied and the controls are not. When a mouse of either group dies, its body is subjected to a path-

ological study and an effort is made to determine the cause of its death. With a degree of oversimplification, it is postulated that there is a somewhat limited number of possible causes of death, say K of them. The problem studied is that of the difference in death rates from the different causes among the experimental and the control mice. This is only a rough description of the problem. One of the difficulties that became obvious on closer examination is due to the omnipresent phenomenon of "competing risks." One illustrative example is as follows.

All of us alive today are exposed to a variety of risks of death, including street traffic and cancer. If I am run over and killed by a car tonight, it would be impossible for me to die later from cancer and, in due course, this would affect the published death rates from cancer. In consequence, the numerical results of a survival experiment with mice do not characterize "net rates" of deaths from the various causes of death studied but only the "crude rates." These crude rates corresponding to the different causes (or "risks") studied characterize not only the intensities of particular risks, but they also reflect the combined property of all of them that is due to competition. Now, let us visualize the results of a completed survival experiment after all the mice, say of the experimental group, have died.

Table 1 illustrates the obtainable results.

Table 1

Illustration of the results of a survival experiment.

Cause of Death	Survival Times of Particular Mice
C_1	$t_{11} \leq t_{12} \leq t_{13} \cdots \leq t_{1n_1}$
C_2	$t_{21} \leq t_{22} \leq t_{23} \cdots \leq t_{2n_2}$
...	.....
...	.....
C_K	$t_{K1} \leq t_{K2} \leq \ldots\ldots \leq t_{Kn_K}$

The first column of Table 1 enumerates all the K causes of death. The wide second column gives the corresponding consecutive survival times of mice that died from the particular causes. Thus, for example, the symbol t_{11} stands for the time of the first recorded death from cause C_1. Similarly, the last symbol in the same line, namely t_{1n_1} represents the time of death of the last mouse that died from the same ause C_1, etc. Here, then, the subscripts n_1, n_2, ..., n_K denote the numbers of mice that died from causes C_1, C_2, ... C_K, respectively. Naturally, these numbers n_1, n_2, ..., n_K will not be all equal and their variability will reflect both the severity of particular causes and their competition. The

reader will have no difficulty in visualizing an exactly similar table compiled for the control mice. These two tables would then be ready for the evaluation of the effects of the agents studied on the survival experience of the mice.

Having in one's mind the problem of a new electric generator in locality L, one might think of the question: how many more deaths from cancer (perhaps cause C_1) should one expect among mice if the "agents" studied included irradiation? What about the methodology of evaluating the experiment that could answer reliably a question of this kind?

One of the methodologies used is that, based on the concept of "potential survival times." For an experimental animal exposed to K possible risks (or causes) of death, the term i-th potential survival time designates a random variable Y_i supposed to represent the age at death of this animal in the hypothetical condition in which C_i is the only possible cause of death. The probability that Y_i will exceed a preassigned value t is called the "net survival probability."

Unfortunately, while a survival experiment can be conducted to investigate a great variety of different "agents," the resulting "causes" of death are not under control of the experimentor. Thus, no direct empirical counterpart of the net survival probability can be available. All that the results of a survival experiment illustrated in Table 1 can provide is the empirical

counterparts of the so-called "crude survival probabilities." For the i-th cause the crude probability of surviving up to time t, say $Q_i(t)$ is the probability that $Y_i = \min(Y_1, Y_2, \ldots Y_K)$ and that $Y_i > t$. Here, then, the question arises whether a statistical methodology could be developed to use the crude survival data as in Table 1, perhaps somehow supplemented, in order to estimate the net survival probabilities.

As interestingly described by David [1974], the competing risk phenomenon occurs not only in problems of public health but also in problems of technological reliability. Here, the most attractive presumption supplementing the data of a survival experiment is the assumption that the potential survival times Y_i are mutually independent. However, the hypothesis of independence cannot be tested using the data of a survival experiment and the publications of Tsiatis [1975] and of Peterson [1976] document the presence of non-identifiability. The crude survival probabilities are consistent with an infinity of systems of widely different net survival probabilities. The conclusion is that the survival experiment of the type described is too simplistic to provide all the valuable information for studies of problems of health.

9. Survival Experiments with Serial Sacrifice. The "serial sacrifice" methodology [Upton, 1969] represents a very important advance in the health related experimentation. Rather than focus

on the diagnosed "causes" of death of the experimental animals, the serial sacrifice experimentation deals with what I like to call "elementary pathological states," say S_1, S_2, ... S_K. For example S_1 may stand for thymic lymphoma (a cancer), S_2 for reticulum cell sercoma, another cancer, etc. At selected times, say t_1, t_2, ... samples of mice alive at these times are killed (= "sacrifice") and their bodies are subjected to a pathological analysis. The result of such analysis for a particular mouse may be that, at the time of its sacrifice, it was affected by, say, three elementary pathological states, S_4, S_5, S_6, and no others.

The above methodology provides empirical counterparts to the following type of questions: how frequently the mice alive at the preassigned times t_1, t_2, ... are affected by this or that combination of pathological states? Combined with similar data for mice that died on their own (not through "sacrifice") the amount of information from a serial sacrifice experiment is very much richer than from the "typical" survival experiment illustrated in Table 1. Also, there is an important difference in the nature of the information.

Here, I wish to call the reader's attention to the analogy between the serial sacrifice vs. "typical" survival experiment situation, on the one hand, and the multiple periods of counting accidents vs. just one such period, on the other. As discussed in

in Section 6, the non-identifiability of two contrasting mechanisms of accident proneness was due to the insufficiency of observational data: numbers of accidents incurred during a single year. The counts of accidents incurred by each driver the following year made the non-identifiability disappear. It is this analogy that is symbolized by reference to the "Yule-Pólya dilemma" in the title of the present Chapter V.

I learned about the serial sacrifice design during a visit to the Oak Ridge National Laboratory and, particularly, through conversations with Dr. John B. Storer. At the time Dr. Storer was in charge of the continuing experiment set up by Upton. Later, we had the pleasure of Dr. Storer's visit to Berkeley. Also, we received from him a substantial sample of data from the experiment in question. In these data, the total number of elementary pathological states was eight. The further difference with the "typical" survival experiment was that there were no "causes" of death indicated.

While all human determinations are subject to error, the determination of particular pathological states is comparable to chemical analyses and represents an effort at objectivity. On the other hand, the diagnosis of a "cause" of death is a conclusion likely to be affected by subjective attitudes of the pathologists.

10. Another Shock of Non-Identifiability. As mentioned in Section 6, the finding of non-identifiability affecting the study

of accident proneness caused Dr. Bates and myself to experience a shock. Here, I have to admit a somewhat explosive feeling of enthusiasm I felt when contemplating the experimental results obtainable through serial sacrifice experiment. I rather felt that these results, without any additional observations, provide data for the study of a stochastic process representing the natural succession of life and death events: birth at time zero, followed by first illness at age t_1, then by recovery at time t_2, etc. etc., and finally death at some observable time. Because the domain of stochastic processes is now well developed, I expected that a statistical methodology could be discovered to use the serial sacrifice data in order to estimate the mechanism of treatment effects in mice contemplated, perhaps, or a realization of a finite states Markov chain, with all the transition probabilities possible to estimate. Due to the work of Clifford [1977], I experienced a shock. Even with some over-simplifying assumptions (denying the possibility of "recovery," etc.) a discrete time Markov chain model proved to be unidentifiable with respect to the data of a serial sacrifice experiment! The details are described in the analysis of Storer's data performed with Clifford's active participation [Berlin et al, 1979].

While the serial sacrifice data provide answers to the questions "how frequently mice sacrificed at age t are affected by a stated combination of pathological states," the missing

information relates to mice alive at age t and having at that age a stated pathological combination. During the subsequent unit of time, say during the next 100 days, the health state of these mice can change in many different ways: recover from some illnesses, contract some others, etc. With the present design of serial sacrifice experiments there is no information on the frequency of such transitions. The tantalizing question is whether some not too difficult modification of the methodology could provide information to fill in the now existing gaps. The way of discovering such effective modifications requires a reasonably close cooperation between an intensely interested statistician and an equally intensely interested experimenting biologist. The questions to resolve are of the following type: could the analysis of urine of a mouse provide enough information on its health state? Could the analysis of a blood sample be sufficient? However, can this sample of blood be taken without altering the contemporary transition probabilities of the mouse, i.e., without hurting the mouse? Who knows? However, unless one tries, one can hardly hope to succeed.

VI. EFFORT AT AN "OPTIMAL" COMPETITOR TO K.P.'S χ^2 TEST FOR GOODNESS OF FIT

10. Introductory Remarks. This chapter is to illustrate my preferred strategy of studying or of developing statistical

tests: begin by defining the optimal performance of the test, and then try to deduce the desired criterion. As indicated in the title of the chapter, the example chosen for illustration is the Karl Pearson's test "for goodness of fit" symbolized by χ^2.

As is well known, the χ^2 test is now being used for a variety of purposes, such as contingency tables, etc. In these circumstances, I wish to emphasize the limited scope of the following discussion: it is concerned with the problem of "goodness of fit" as comtemplated in olden days by K.P. My actual effort to formulate the problem and to solve it was published in 1937[b]. It is limited to the case of a "simple hypothesis," this is, to the case in which the problem is to decide whether a completely specified probability density, say $p_\chi(x)$ fits the empirical distribution of an observable random variable X. Another limitation consists in the assumption that the number N of observed values of X is "large." The problem of extending the methodology to the case of composite parametric hypotheses has been treated by Javitz [1975].

11. Criticism of the K.P.'s χ^2 Test for Goodness of Fit. An effort at an "optimal" competitor of an existing test intended for use in some specified conditions must begin by the unavoidably subjective criticism of the original test. The well known procedure of the χ^2 test for goodness of fit begins by dividing the range of variation of the observable X into a certain number, say s, of "cells,"

with boundaries

$$a_0<a_1<a_2 \ldots <a_s, \tag{5}$$

where a_0 may mean $-\infty$ and a_s may be $+\infty$. Next, the probability density $p_\chi(x)$ is used to compute the expected number of independent observations, say n_i, falling into the i-th cell for i=1, 2, ... s. Let m_i denote the actual number out of the total N observations that fall into the same i-th cell. Then, K.P.'s test criterion for goodness of fit is given by

$$\chi^2 = \sum_{i=1}^{s} \frac{(m_i-n_i)^2}{n_i} . \tag{6}$$

The fit is considered "bad" if the calculated χ^2 exceeds the tabled limit corresponding to the chosen level of significance. Otherwise, the fit is considered "good."

My own subjective criticism of the test includes the fact that the value of the criterion (6) does not depend on the order of positive and negative differences (m_i-n_i). The extreme example is represented by the following possibilities. In one case, the signs of the consecutive differences m_i-n_i and $m_{i+1}-n_{i+1}$ are not the same. In the other case one can observe a substantial number of consecutive differences m_i-n_i that are all negative while all the others are positive. While these two possibilities are consistent with the same value of the criterion (6), my intuitive feeling is that in the second case the "goodness of fit" is subject to a rather strong doubt, irrespective of the actual computed value of (6), even if it happens to be small.

12. "Smooth Test" for Goodness of Fit. The first step in the deduction of the "smooth test" intended as an "optimal" competitor to χ^2, consisted in standardizing the analytical developments. Rather than consider the great variety of distributions $p_X(x)$ that may come under consideration, I proposed to replace the observable X by its function Y defined by the relation

$$y = \int_{-\infty}^{x} p_X(x)dx \qquad (7)$$

when y and x designate particular values of the two random variables. As it is easy to check, the range of variation of Y is from zero to unity, with its probability density

$$p_Y(y) = 1, \qquad (8)$$

this, irrespective of the distribution of X.

As contemplated by Karl Pearson, the background of the problem of goodness of fit admits the possibility that the specified density of $p_X(x)$ may not correspond to reality. However, there are no general indications as to what the alternatives might be. In my attempt to deduce an optimal competitor to the chi square test, I contemplated the set of alternatives vaguely described as "smooth."

In terms of the variable Y, with its range of variation limited to the interval (0, 1) where its density is equal to unity, the contemplated "smooth" alternatives are those with densities the logarithms of which are polynomials of orders 1, 2, ... K.

The theory published in 1937 develops an asymptotic version of optimal unbiased type C tests of orders K=1, 2, ... with K denoting the order of polynomial used. The study of asymptotic power of these tests indicates that, generally, adequate results could be obtained with K not exceeding 4. The tests so deduced are not open to the criticism of the original test for goodness of fit indicated above.

In recent times quite a few non-parametric tests for goodness of fit have been considered with emphasis on their robustness. It would be interesting to use the Monte Carlo methodology to compare the performance of these tests with that of the smooth test of a limited order $K \leq 4$.

ACKNOWLEDGEMENTS

This paper was prepared using the facilities of the Statistical Laboratory with partial support from the Office of Naval Research (ONR N00014 75 C 0159), the Department of the Army (Grant DA AG 29 76 G 0167), and the National Institute of Environmental Health Sciences (2 R01 ES01299-16). The opinions expressed are those of the author.

REFERENCES

Grace E. Bates (1955), "Joint distribution of time intervals for the occurrence of successive accidents in a generalized Pólya scheme," Ann. Math. Stat., Vol. 21, pp. 705-720.

Grace E. Bates and J. Neyman (1952a), "Contribution to the theory of accident proneness, I. An optimistic model of correlation between light and severe accidents," Univ. of Calif. Publ. in Stat., Vol. I, pp. 215-254.

Grace E. Bates and J. Neyman (1952b), "Contribution to the theory of accident proneness, II. True or false contagion," Univ. of Calif. Publ. in Stat., Vol. I, pp. 255-276.

Robert H. Berk and Douglas M. Jones (1979), "Goodness-of-Fit Test Statistics that Dominate the Kologorov Statistics," Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, Vol. 47, pp. 47-59.

Bengt Berlin, Joel Brodsky and Peter Clifford (1979), "Testing Disease Dependence in Survival Experiments with Serial Sacrifice," JASA, Vol. 74, No. 1.

Harald Cramér (1928), "On the composition of elementary errors," Skandinavisk Aktuarietidskrift, Vol. 11, pp. 13-74 and 141-180.

Harald Cramér (1938), "Sur un nouveau théorème-limite de la théorie des probabilitiés," Actualités Sci. Ind. No. 736, pp. 5-23.

H. A. David (1974), "Parametric approaches to the theory of competing risks," Reliability and Biometry, Statistical Analysis of Lifelength" (F. Proschan and R.J. Serfling, eds.), SIAM, Philadelphia, pp. 275-290.

M. Greenwood and G. U. Yule, (1920), "An inquiry into the nature of frequency distributions...with particular reference to... repeated accidents," J. Roy. Stat. Soc., Vol. 83, pp. 255-279.

Wassily Hoeffding (1951), "'Optimum' Nonparametric Tests," Proc. Second Berkeley Symp. Math. Stat. and Prob., Univ. of Calif. Press, Berkeley, CA, pp. 33-92.

Wassily Hoeffding (1965), "Asymptotically optimal tests of multinomial distribution," Annals of Math. Stat., Vol. 36, pp. 369-401.

Harold S. Javitz (1975), "Generalized smooth tests of goodness of fit, independence, and equality of distributions," unpublished doctoral dissertation, Univ. of Calif., Berkeley.

Yu. V. Linnik (1961), "On the probability of large deviations for the sums of independent variables," Proc. Fourth Berkeley Symp. Math. Stat. and Prob., Vol. II, Univ. of Calif. Press, Berkeley, CA, pp. 289-306.

Michel Loève (1960), Probability Theory, Van Nostrand, p. 268.

Richard von Mises (1931), Wahrscheinlichkeitsrechnung, Leipzig u. Wien, pp. 316-335.

E. M. Newbold (1928), "A contribution to the study of human factors in the causation of accidents," Industr. Health Res. Board Report No. 34, London, H.M. Stationary Office.

J. Neyman (1937a), "Outline of a theory of statistical estimation based on the classical theory of probability," Philos. Trans. Roy. Soc. of London, Ser. A., Vol. 236, pp. 333-380.

J. Neyman (1937b), "'Smooth test for goodness of fit," Skandinavisk Aktuarietidsckrit, Vol. 20, pp. 149-199.

J. Neyman (1959), "Optimal asymptotic tests of composite statistical hypotheses," Probability and Statistics (The Harald Cramér Volume), (U. Grenander, ed.), Almquist and Wiksells, Uppsala, Sweden, pp. 213-234.

J. Neyman (1977a), "Frequentist probability and frequentist statistics," Synthese, Vol. 36, pp. 97-131.

J. Neyman (1977b), "Public health hazards from electricity-producing plants," Science, Vol. 195, pp. 754-758.

J. Neyman and E.S. Pearson (1933), "On the problem of the most efficient tests of statistical hypotheses," Philos. Trans. Roy. Soc. of London, Ser.A., Vol. 231, pp. 289-337.

J. Neyman and E.S. Pearson (1936), "Contributions to the theory of testing statistical hypotheses (1) Unbiased critical regions of Type A and Type A_1," Stat. Res. Memoirs, Vol. 1, pp. 1-37.

Karl Pearson (1900), "On the criterion that a given system of deviations from the probabile in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Phil Mag. and J. of Sci., Vol. 50, pp. 157-175.

A.V. Peterson (1976), "Bounds for a joint distribution function with fixed sub-distribution functions: application to competing risks," Proc. Natl. Acad. Sci., USA, Vol. 73, pp. 11-13.

G. Pólya (1930), "Sur quelques points de la théorie des probabilités," Ann. de l'Institut Henri Poincaré, Vol. 1, pp. 117-161.

A. Tsiatis (1975), "A nonidentifiability aspect of the problem of competing risks," Proc. Natl. Acad. of Sci., Vol. 72, No. 1, pp. 20-22.

A.C. Upton et al (1969), Radiation Induced Cancer, International Atomic Energy Agency, Vienna, p. 425.