JOINT DISTRIBUTIONS OF TIME INTERVALS FOR THE OCCURRENCE OF SUCCESSIVE ACCIDENTS IN A GENERALIZED POLYA SCHEME[1]

By Grace E. Bates

Mount Holyoke College and University of California

0. Summary. The main body of this paper has two rather distinct parts, the first part (Section 2) containing the derivation of an explicit representation for the joint distribution of accident times in quite a general situation, and the remainder of the paper dealing with a very special case which leads to a discussion of testing the hypothesis that a distribution is uniform over (0, 1) against alternatives which are exponential truncated to (0, 1). The second part may be read profitably with little reference to the probabilistic arguments of Section 2.

1. Introduction. In [1] a comparison was made between two models often used in studies of accident proneness. The first model, due to Greenwood, Yule, and Newbold [2], [3], postulates variability among the individuals of a population with respect to accident proneness, assumes that previous accidents do not change the probabilities of future accidents and that experience gained in the particular occupation giving rise to the risk of these accidents does not modify these probabilities. The combined term "mixture-no contagion-no time effect" model was used to symbolize this first scheme. The second model, due to Polya [4], postulates identity of the individuals with respect to accident proneness, possible presence of contagion, and possible effect of experience gained since entering the particular occupation.

Using in [1] the scheme of Polya in a slightly more generalized form, it was shown that the multivariate distribution of the number of accidents incurred in several successive periods of observation, as soon as two or more periods of observation were used, was distinguishable from the corresponding distribution implied by the mixture-no contagion-no time effect scheme, barring an exceptional particular case.

The last section of [1] considered the same problem of distinguishing between the two models when the random variables used were the time intervals between successive accidents incurred by an individual in one period of observation. In formulating this scheme it was found possible to liberalize a little the scheme of Polya by not insisting that the contagion be a linear type. This approach was applied only to the case of individuals who, during the period of observation, sustain exactly one accident.

Received December 22, 1952, revised May 6, 1955.

[1] This work was done with the partial support of the Office of Naval Research. Reproduction in whole or in part permitted for any U. S. Government purpose.

The present paper applies the outline given in Section 10 of [1] to the general case in which one makes use of data relating to several groups, say G_i, $i = 1, 2, \cdots, k$, of individuals, each group composed of individuals who sustain the same number, n_i, of accidents in the period of observation.

Section 2 of the present paper establishes, for an individual who sustains n successive accidents in the period of observation, the joint distribution of the time intervals measured from the beginning of the period to the occurrence of each accident. Some interesting special cases of this distribution are enumerated and the remainder of the paper is concerned with a comparison of one of these special cases with the case of no mixture-no contagion-no time effect. The special case selected is one implying that there is no mixture-no time effect and that contagion, when present, is of the linear type. One advantage of dealing with the case of linear contagion is that it is not necessary in this case to assume that all the individuals under consideration have sustained the same number of accidents prior to the period of observation.

In section 3, preparatory to constructing tests of the hypothesis of no-mixture-no contagion-no time effect versus alternatives of no mixture-linear contagion-no time effect, the distribution of the mean of the random variables τ_i, representing the length of the time interval from the start of a unit period of observation to the occurrence of the i-th accident, is determined under the hypothesis tested and under the alternative hypotheses. The distribution of the mean of the τ_i's under the hypothesis tested is the well-known distribution of the mean of n completely independent random variables, each uniformly distributed on (0, 1).

Section 4 considers the construction of these tests. It is found that there are uniformly most powerful tests of the hypothesis of no mixture-no time effect-no contagion against each of the classes of one-sided alternatives (termed "positive linear contagion" and "negative linear contagion", respectively) and a uniformly most powerful unbiased test of the original hypothesis versus the set of alternatives, no mixture-no time effect-linear contagion. Given accident data, including times of occurrence of the accidents relating to several groups of individuals, each group containing individuals who sustain the same number of accidents in the period of observation, the statistic required for the tests is the grand mean of all the time intervals for all the individuals.

In the last section we treat the problem of computing the power of the uniformly most powerful unbiased test. The exact power function is obtained explicitly from the distributions of Section 2. However, since actual computation of this power is very tedious, an approximation to the power function is desirable. This approximation is effected in two stages: first, the critical region boundaries for a specified test level are approximated by using the normal approximation to the distribution of the mean of the time intervals under the hypothesis of no mixture-no time effect-no contagion—i.e., to Laplace's distribution of the mean of n completely independent random variables uniform in (0, 1). Then the Central Limit Theorem is applied to the distribution of the mean in the set of alternative hypotheses to find the approximate power of this test.

2. Joint distribution of time intervals for n successive accidents sustained by an individual in one period of observation. As in section 10 of [1] we consider an individual I who, from the moment $t = 0$ on, is exposed to the risk of nonfatal accidents of a particular kind. For this individual we shall consider probabilities $P_{m,n}(T_1, T_2)$ defined as follows, for $0 \leqq T_1 \leqq T_2$: $P_{m,n}(T_1, T_2)$ is the conditional probability that, during the time interval (T_1, T_2), the individual I will incur exactly n accidents, given that at time T_1 or before he had sustained exactly m accidents. We impose on the probabilities $P_{m,n}(T_1, T_2)$ the three postulates given below. The totality of these three postulates we shall describe as the Polya contagious scheme.

POSTULATE 1. *If $T_2 \to T_1$, then all the probabilities $P_{m,n}(T_1, T_2)$ converge to limits $P_{m,n}(T_1, T_1)$. More specifically,*

$$P_{m,0}(T_1, T_1) = 1 \quad \textit{for every } m, \tag{1}$$

and consequently,

$$P_{m,n}(T_1, T_1) = 0 \quad \textit{for every } m, \textit{ and for } \quad n \geqq 1. \tag{2}$$

POSTULATE 2. *The probabilities $P_{m,n}(T_1, T_2)$ depend on the number m of accidents sustained up to and including the moment T_1 and also on the value of T_1, but not on the moments at which these previous accidents occurred.*

POSTULATE 3. *At least at $T_2 = T_1$, the probabilities $P_{m,n}(T_1, T_2)$ are differentiable with respect to T_2, and specifically,*

$$\frac{\partial}{\partial T_2} P_{m,n}(T_1, T_2)\Big|_{T_2=T_1} = \begin{cases} \dfrac{-\lambda_m}{1 + \nu T_1} & \text{if } n = 0 \\ \dfrac{\lambda_m}{1 + \nu T_1} & \text{if } n = 1 \\ 0 & \text{if } n > 1 \end{cases} \tag{3}$$

where $\nu \geqq 0$ and $\lambda_0, \lambda_1, \cdots, \lambda_m, \cdots$ are arbitrary positive numbers, with possibly $\lambda_{m+k} = 0$, $k \geqq M$, for some positive integer M.

In applying the probabilistic scheme formulated above, one may consider the probability space as the accident histories of a large population of individuals. Then, at the outset, the conditions above require that the λ_m's and ν's be the same for all individuals. It will be pointed out later, however, that the tests of the hypothesis of no mixture-no time effect-no contagion in the special set of alternatives of no mixture-no time effect-linear contagion derived in this paper, imply only that ν be zero for each individual and that each individual have the same constant increment $\psi = \lambda_{m+k+1} - \lambda_{m+k}$ in the sequence of his λ_m's.

Following the usual procedure ([5], Chapter 17) one obtains the differential equations

$$\frac{\partial}{\partial T_2} P_{m,0}(T_1, T_2) = \frac{-\lambda_m}{1 + \nu T_2} P_{m,0}(T_1, T_2), \tag{4}$$

$$\frac{\partial}{\partial T_2} P_{m,n}(T_1, T_2) = \frac{-\lambda_{m+n}}{1 + \nu T_2} P_{m,n}(T_1, T_2) + \frac{\lambda_{m+n-1}}{1 + \nu T_2} P_{m,n}(T_1, T_2) \tag{5}$$

for $n \geqq 1$.

Under initial conditions given by Postulate 1, the solution of (4) is

$$P_{m,0}(T_1, T_2) = \left(\frac{1 + \nu T_1}{1 + \nu T_2}\right)^{\lambda_m/\nu}, \tag{6}$$

and the solution (Cf., [6] pp. 406–407 where the general solution of a similar set of differential equations is given) of (5) when all the λ's are unequal and $\nu \neq 0$ is

$$P_{m,n}(T_1, T_2) = (-1)^n \lambda_m \lambda_{m+1} \cdots \lambda_{m+n-1} \sum_{k=0}^{n} \left(\frac{1 + \nu T_1}{1 + \nu T_2}\right)^{\lambda_{m+k}/\nu} D_{k,n}^{-1} \tag{7}$$

for $n \geqq 1$, with

$$D_{k,n} = \prod_{\substack{j=0 \\ j \neq k}}^{n} (\lambda_{m+k} - \lambda_{m+j}). \tag{8}$$

Since the set of equations (4) and (5) may be solved recursively for the $P_{m,n}(T_1, T_2)$, the solutions (6) and (7) are unique. From the familiar formulas for solving linear differential equations of this type it is easily seen that the solutions, $P_{m,n}(T_1, T_2)$ are non-negative. Furthermore, it is easily verified that

$$\sum_{n=0}^{\infty} P_{m,n}(T_1 T_2) \leqq 1. \tag{9}$$

If the system of equations (4), (5), is finite (i.e., $\lambda_{m+k} = 0$ for $k \geqq M$) then $P_{m,n}(T_1, T_2) = 0$ for $n > M$ and the equality holds in (9) so that the solutions (6) and (7) for $0 \leqq n \leqq M$, form a proper probability distribution.

In the case of an infinite set of equations (4), (5), it may happen that (9) is a true inequality. This type of situation has been discussed by Feller ([5], pp. 369–371) as implying nonzero probability of the occurrence of an infinite number of events in the finite period. Feller derives a necessary and sufficient condition for the equality in (9) to hold. Using equations (4) and (5), Feller's argument goes through almost verbatim to yield the same result—namely, a necessary and sufficient condition for the equality in (9) to hold is that the series $\sum_{n=0}^{\infty} \lambda_{m+n}^{-1}$ diverges.

In the application of the distribution of $P_{m,n}$'s made in the remainder of this paper, it will be seen that we have either a finite set of equations (4) and (5), or, in the infinite case, the divergence condition on the λ_{m+n}'s is fulfilled.

Forms obtained from solutions (6) and (7) by a passage to the limit as $\nu \to 0$ and/or as some or all of the λ's become equal, may be shown by direct verification also to satisfy (4) and (5), a fact which will be used later in this paper.

It is clear that $P_{m,0}(T_1, T_2)$ is a decreasing function of λ_m. Furthermore, if all the λ's have the same value, the model implies the absence of contagion in the accidents. As in [1], if the λ's form an increasing sequence $\lambda_0 < \lambda_1 < \cdots < \lambda_m < \cdots$, we use the term "regular positive contagion", meaning that the more accidents the individual had in the past, the more intense his risk of accidents in the future. If the λ's form a monotonic decreasing sequence, we speak of "regu-

lar negative contagion"—past accidents "teach" the individual how better to avoid accidents in the future. We use the term "irregular contagion" for a non-monotone sequence of λ's. The constant ν in the probabilities $P_{m,n}(T_1, T_2)$ is termed "time effect" in the model and may be attributed to the effect of previous experience gained by the individual in the occupation which gives rise to the risks of these particular accidents. From (3) it may be seen that with increasing T_1 and $\lambda_m > 0$ the rate of increase of the probabilities of further accidents in (T_1, T_2) is slowed down by the presence of $\nu > 0$. When $\nu = 0$ there is an absence of time effect.

We now assume that an individual I is observed for a unit of time from T_1 to $T_1 + 1$. The first problem considered is that of finding the joint distribution of random variables τ_i defined as follows: τ_i is the time from T_1 to the occurrence of the i-th accident for the individual I. We are concerned with the joint distribution of $\tau_1 < \tau_2 < \cdots < \tau_n$ conditional on the occurrence of n accidents in the period of observation and m accidents previous to this period.

In order to solve the problem under consideration, we now compute the probability of the following two events, each of which is to be understood as conditioned by the occurrence of exactly m accidents at or before T_1 :

(i) I incurs exactly n accidents in $(T_1, T_1 + 1)$.

(ii) I incurs n accidents in $(T_1, T_1 + 1)$ and the random variables τ_i satisfy conditions $\tau_i \leqq t_i$, $i = 1, 2, \cdots, n$ with $0 < t_1 < t_2 < \cdots < t_n < 1$.

The probability of (i) is $P_{m,n}(T_1, T_1 + 1)$, and from (7) we have

$$P_{m,n}(T_1, T_1 + 1) = (-1)^n \lambda_m \lambda_{m+1} \cdots \lambda_{m+n-1} \sum_{k=0}^{n} \left(\frac{a}{a + \nu}\right)^{\lambda_{m+k}/\nu} D_{k,n}^{-1} \tag{10}$$

where, for convenience,

$$a = 1 + \nu T_1 . \tag{11}$$

The probability of (ii) is

$$\sum_{\{J_k\}} \prod_{k=0}^{n} P_{m+J_k, J_{k+1}-J_k}(T_1 + t_k, T_1 + t_{k+1}) \tag{12}$$

with $t_0 = 0$ and $t_{n+1} = 1$, and where the sum is over all sequences $\{J_k\}$ such that $J_0 = 0$, $J_n = J_{n+1} = n$, with $\{J_k\}$ a non-decreasing sequence of integers, each $J_k \geqq k$ for $k = 0, 1, \cdots, n$.

The joint density function of the τ_i's conditioned by the occurrence of m previous accidents and n accidents in $(T_1, T_1 + 1)$ will be the n-th partial derivative with respect to $t_1, t_2, \cdots, t_n$ of the expression in (12), divided by $P_{m,n}$ $(T_1, T_1 + 1)$ given in (10). Of the terms in the sum in (12), the only term which does not contribute zero to this density is

$$\prod_{k=0}^{n-1} P_{m+k,1}(T_1 + t_k, T_1 + t_{k+1}) P_{m+n,0}(T_1 + t_n, T_1 + 1), \tag{13}$$

i.e., the term involving the sequence of J_k's with $J_k = k$, $0 \leqq k \leqq n$. This may be seen as follows:

Consider the term in the sum in (12) for any one of the possible sequences of J_k's for which $J_{n-1} = n$. Then the first difference with respect to t_n of this term is identically zero, since, for $0 < \epsilon < 1 - t_n$ we have for this difference:

$$(14)\quad \prod_{k=0}^{n-2} P_{m+J_k, J_{k+1}-J_k}(T_1 + t_k\,,\ T_1 + t_{k+1}) \cdot \{P_{m+n,0}(T_1 + t_{n-1}\,,\ T_1 + t_n + \epsilon)P_{m+n,0}(T_1 + t_n + \epsilon,\ T_1 + 1) - P_{m+n,0}(T_1 + t_{n-1}\,,\ T_1 + t_n)P_{m+n,0}(T_1 + t_n\,,\ T_1 + 1)\}$$

where the expression in the braces is certainly zero. Hence only terms of the sum in (12) for which the sequence of J_k's has $J_{n-1} = n - 1$ will contribute to the density function.

Next, considering sequences of J_k's with $J_{n-1} = n - 1$, we take the class of those for which $J_{n-2} = n - 1$. But for each such J_k the corresponding term in the sum in (12) will have second difference with respect to t_n, t_{n-1} identically zero and hence in order for a sequence of J_k's to contribute non-zero density, we must have also $J_{n-2} = n - 2$. Continuing inductively, we find that unless the particular sequence of J_k's is chosen for which $J_k = k$, $0 \leqq k \leqq n$, the n-th partial derivative in (12) with respect to t_n, t_{n-1}, $\cdots$, t_1 is 0.

Differentiating (13) with respect to t_n, t_{n-1}, $\cdots$, t_1 and dividing by (10), we have the density function

$$(15)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1\,,\ t_2\,,\ \cdots\,,\ t_n \mid \psi_m\,,\ \psi_{m+1}\,,\ \cdots\,,\ \psi_{m+n-1}\,,\ \nu) = \frac{\prod_{k=1}^{n} (1 + \nu t_k/a)^{(\psi_{m+k-1})/\nu - 1}}{a^n \sum_{k=0}^{n} (-1)^k (1 + \nu/a)^{(\psi_{m+k} + \cdots + \psi_{m+n-1})/\nu} R_{k,n}^{-1}}, \qquad 0 < t_1 < \cdots < t_n < 1,$$

with

$$(16)\quad R_{k,n} = \{(\psi_m + \psi_{m+1} + \cdots + \psi_{m+k-1})(\psi_{m+1} + \cdots + \psi_{m+k-1}) \cdots (\psi_{m+k-2} + \psi_{m+k-1})\psi_{m+k-1}\psi_{m+k}(\psi_{m+k} + \psi_{m+k+1}) \cdots (\psi_{m+k} + \psi_{m+k+1} + \cdots + \psi_{m+n-1})\},$$

and

$$(17)\quad \psi_{m+k} = \lambda_{m+k+1} - \lambda_{m+k}\,, \qquad k = 0,\ 1,\ \cdots,\ n - 1.$$

The following special cases are of interest and were obtained by considering the limiting forms of the density function (15). The equivalence of these results to those obtained by a passage to the limit before differentiation may be verified directly.

Case (i) No mixture-possible time effect-no contagion:

$$(18)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_i = \psi = 0; \nu) = n!\left[\frac{\nu}{a \log_e (1 + \nu/a)}\right]^n \prod_{k=1}^{n} (1 + \nu t_k/a)^{-1}.$$

Case (ii) No mixture-no time effect-possible contagion:

$$(19)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_m, \cdots, \psi_{m+n-1}; \nu = 0) = \frac{\exp\left(\sum_{k=1}^{n} \psi_{m+k-1} t_k\right)}{\sum_{k=0}^{n} (-1)^k \exp(\psi_{m+k} + \cdots + \psi_{m+n-1}) R_{k,n}^{-1}}.$$

Case (iii) No mixture-possible time-effect-linear contagion:

$$(20)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_m = \psi_{m+1} = \cdots = \psi_{m+n-1} = \psi; \nu) = \frac{n!\psi^n}{a^n[(1 + (\nu/a))^{\psi/\nu} - 1]^n} \prod_{k=1}^{n} (1 + (\nu t_k/a))^{\psi/\nu - 1}.$$

Case (iv) No mixture-no time effect-linear contagion:

$$(21)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_i = \psi; \nu = 0) = n!\left(\frac{\psi}{e^\psi - 1}\right)^n \exp \psi \sum_{k=1}^{n} t_k$$

Case (v) No mixture-no time-effect-no contagion:

$$(22)\quad p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_i = \psi = 0; \nu = 0) = n'.$$

We not that the joint density function (15) takes the form of that in (22) also when all the ψ_i are equal to ψ and $\psi = \gamma$, whether or not $\psi = 0$. In this case then, the presence or absence of contagion is unidentifiable.

For the remainder of the present paper we shall be concerned with a comparison of the models implied by (21) and (22), so that contagion is absent if and only if $\psi = 0$. We first note some of the implications arising from the model implied by (21).

It is clear from the model that for $\psi_i = \psi$, the contagion is of the regular positive type for $\psi > 0$ and is regular negative contagion for $\psi < 0$. Furthermore, it is obvious that this condition on the ψ_i implies that the contagion is linear; specifically,

$$(23)\quad \lambda_{m+k} = \lambda_m + k\psi.$$

To see the effect of this linearity more clearly, we return to (6) and see that for $\nu = 0$,

$$(24)\quad P_{m,0}(T_1, T_2) = e^{-\lambda_m(T_2 - T_1)}$$

so that, in the unit time interval, $(T_1, T_1 + 1)$, we have

$$e^{-\psi} = \frac{P_{m+k+1,0}(T_1, T_1, + 1)}{P_{m+k,0}(T_1, T_1, + 1)}, \qquad k = 0, 1, \cdots, n - 1. \tag{25}$$

That is, $e^{-\psi}$ is the factor by which the probability of avoiding accidents in a unit time interval, for an individual I who had previously sustained $m + k$ accidents, must be multiplied in order to find the probability of his avoiding accidents in this time interval had he sustained $m + k + 1$ previous accidents. Thus, for example, a value of ψ of about -0.7 means that an increase of 1 in the number of previous accidents would double his chances of avoiding further accidents in this period, while a ψ of 0.7 means that a similar increase in the number of previous accidents would halve his chances of avoiding further accidents in the period. The actual number of previous accidents sustained by the individual, of course, still determines his probability of avoiding accidents in the period of observation, but the condition that all the ψ_i be equal implies that the relative increase or decrease of this probability depends only on the increase in the number of previous accidents. One important consequence of the condition that all the ψ_i's be equal (that is, that the contagion be linear) is that in testing the hypothesis of no contagion versus contagion of this type it is not necessary to assume that all the individuals under observation have incurred the same number of accidents previously or that λ_m is the same for each individual.

We now examine the probability distribution given by the $P_{m,n}(T_1, T_2)$ in the case under consideration of no mixture-no time effect-linear contagion. Returning to the differential equations (4) and (5), with $\nu = 0$ and $\lambda_{m+k} = \lambda_m + k\psi$, we see that in the case of positive linear contagion ($\psi > 0$) the system of equations may be infinite. In this case, however, the series of λ's clearly diverges so that equality in (9) holds. In the case of negative linear contagion ($\psi < 0$) it is evident that the condition that the λ_{m+k}'s be non-negative places a restriction on n, namely

$$n \leqq -\lambda_m/\psi, \qquad \psi < 0. \tag{26}$$

Thus the system of differential equations in (4), (5) must be finite when $\psi < 0$ and the $P_{m,n}$'s form a proper probability distribution.

In the case of negative linear contagion, if one knew λ_m (or could conjecture an upper bound for λ_m), it would be possible to reject at the outset certain alternative values of ψ—those such that $n > -\lambda_m/\psi$. Thus the particular model of linear contagion may be criticized in that it places this added restriction on the degree of negative contagion permissible in the model. In the tests of the next sections it is assumed that one has at hand only the accident data in the period of observation, with no knowledge of the number of previous accidents or of the λ_m's. Subject to the limitations of the model, we are then testing the hypothesis of no contagion versus linear contagion.

3. Joint density functions for the means of the n random variables τ_i, having density functions (21) **and** (22). In Section 4 we construct tests of the hypothesis of no contagion in the class of admissible hypotheses each implying no mixture-no time effect-linear contagion. The statistic needed for these tests is found to be the mean of the time intervals, and it is for this reason that the present section is needed.

We rewrite formulas (22) and (21) here for convenience:

$$p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_i = \psi = \nu = 0) = n! \tag{27}$$

$$p_{\tau_1,\tau_2,\cdots,\tau_n}(t_1, t_2, \cdots, t_n \mid \psi_i = \psi; \nu = 0) \tag{28}$$
$$= n!\left(\frac{\psi}{e^{\psi} - 1}\right)^n \exp \psi \sum_{k=1}^{n} t_k, \quad 0 < t_1 < t_2 < \cdots < t_n < 1.$$

From the sampling theory of order statistics (cf. [7] p. 90) we note that the unordered τ_i in (27) are distributed as a random sample from the uniform distribution on (0, 1), which we shall denote by $p(t \mid \psi = 0)$, and that the unordered τ_i in (28) are distributed as a random sample from the distribution with density on (0, 1) given by

$$p(t \mid \psi) = \frac{\psi e^{\psi t}}{e^{\psi} - 1}. \tag{29}$$

The density in (29) is equal to that of the exponential function, $f(t) = -\psi e^{\psi t}$, $\psi < 0$, truncated to (0, 1).

The distribution of the mean of n independent random variables, each with uniform distribution on (0, 1), is well known. Laplace [8] derived this distribution in his Mémoire on the mean inclination of the orbits of comets. (For a more accessible reference, see [9], Vol. 1, p. 244). Writing $(\bar{\tau} \mid \psi = 0)$ for the mean of the τ_i corresponding to (27) and $\varphi_{\bar{\tau}}(u \mid \psi = 0)$ for the characteristic function, we have

$$\varphi_{\bar{\tau}}(u \mid \psi = 0) = \left(\frac{e^{iu/n} - 1}{iu/n}\right)^n \tag{30}$$

$$E(\bar{\tau} \mid \psi = 0) = \tfrac{1}{2}; \qquad \sigma^2(\bar{\tau} \mid \psi = 0) = 1/(12n), \tag{31}$$

$$p_{\bar{\tau}}(t \mid \psi = 0) = \frac{n^n}{(n-1)!} \sum_{j \leqq nt} (-1)^j \binom{n}{j} (t - j/n)^{n-1}, \quad 0 \leqq t \leqq 1. \tag{32}$$

As n increases, the distribution function of the standardized variable

$$z = \sqrt{12n}\,(\bar{\tau} - \tfrac{1}{2}), \tag{33}$$

when $\psi = 0$, rapidly approaches (cf. [10] p. 245) that of the standardized normal variable.

If we now write $(\bar{\tau} \mid \psi)$ for the mean of the τ_i corresponding to (28), with $\varphi_{\bar{\tau}}(u \mid \psi)$ for the characteristic function, we have

$$\varphi_{\bar{\tau}}(u \mid \psi) = \left(\frac{\psi}{e^{\psi} - 1}\right)^n \left(\frac{e^{(iu/n)+\psi} - 1}{(iu/n) + \psi}\right)^n \tag{34}$$

$$E(\bar{\tau} \mid \psi) = \frac{\psi e^{\psi} - e^{\psi} + 1}{\psi(e^{\psi} - 1)}; \qquad \sigma^2(\bar{\tau} \mid \psi) = \frac{1}{n} \frac{(e^{\psi} - 1)^2 - \psi^2 e^{\psi}}{[\psi(e^{\psi} - 1)]^2}, \tag{35}$$

$$\begin{aligned} p_{\bar{\tau}}(t \mid \psi) &= \left(\frac{\psi}{e^{\psi} - 1}\right)^n \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iut} \left(\frac{e^{(iu/n)+\psi} - 1}{(iu/n) + \psi}\right) du \\ &= \left(\frac{\psi e^{\psi t}}{e^{\psi} - 1}\right)^n \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ivt} \left(\frac{e^{iv/n} - 1}{iv/n}\right)^n dv \qquad (36) \\ &= \left(\frac{\psi e^{\psi t}}{e^{\psi} - 1}\right)^n p_{\bar{\tau}}(t \mid \psi = 0). \end{aligned}$$

4. Tests of the hypothesis of no contagion in the set of admissible hypotheses each implying no mixture-no time effect-linear contagion. We suppose that we have observations on several groups of individuals, the r-th group being determined by the integer r, $r = 1, 2, 3, \cdots$, of accidents sustained by each individual of that group in a unit time interval $(T_1, T_1 + 1)$. Suppose further that we have N_r individuals in the r-th group. In this set of observations, if for some integer j, there are no individuals who have sustained exactly j accidents, then N_j is zero for this integer j. We define random variables τ_{irs} in the following way:

τ_{irs} is the time from T_1 to the i-th accident for the s-th individual in the r-th group; $i = 1, \cdots, r$; $r = 1, 2, 3, \cdots$; $s = 1, 2, \cdots, N_r$.

Let

$$N = \sum_r r N_r . \tag{37}$$

That is, let N be the total number of accidents.

We now make the following assumption: From individual to individual, among the $\sum N_r$ individuals, the accident times are independent.

Under this assumption, the unordered random variables τ_{irs} act like N independent observations from a parent population that is either uniform on (0, 1) if no contagion is present, or has density that of (29) if contagion is present. Denoting by $\{\tau_j\}$ the unordered set of τ_{irs}'s, we have

$$p_{\{\tau_j\}}(\{t_j\} \mid \psi = 0) = 1 \tag{38}$$

and

$$p_{\{\tau_j\}}(\{t_j\} \mid \psi) = \left(\frac{\psi}{e^{\psi} - 1}\right)^N \exp \psi \sum_{j=1}^{N} t_j = \left(\frac{\psi e^{\bar{t}\psi}}{e^{\psi} - 1}\right)^N, \tag{39}$$

where $\bar{t} = (1/N) \sum_{j=1}^{N} t_j$.

We note that $\bar{\tau}$ is a sufficient statistic for the distribution of τ_j's, where

$$\bar{\tau} = (1/N) \sum_{j=1}^{N} \tau_j, \tag{40}$$

that is, the grand mean of all the time intervals for all the individuals.

Clearly, then, under the present conditions of no mixture-no time effect-linear contagion, we have a uniformly most powerful test of the hypothesis $\psi = 0$ against either (but not both) of the one-sided alternatives $\psi < 0$ or $\psi > 0$, with critical regions of type $\bar{\tau} < t_0$ or $\bar{\tau} > t_0$, respectively, where t_0 is determined by the level of significance α.

Furthermore, letting

$$\varphi = \frac{\partial}{\partial \psi} \log_e p_{\{\tau_j\}}(\{t_j\} \mid \psi) = N\left(\frac{1}{\psi} + \bar{t} - \frac{e^{\psi}}{e^{\psi} - 1}\right) \tag{41}$$

$$\varphi' = \frac{\partial \varphi}{\partial \psi} = N\left(\frac{-1}{\psi^2} + \frac{e^{\psi}}{(e^{\psi} - 1)^2}\right) \tag{42}$$

we have

$$\varphi' = A + B\varphi, \tag{43}$$

with

$$A = N\left(\frac{-1}{\psi^2} + \frac{e^{\psi}}{(e^{\psi} - 1)^2}\right) \text{ and } B = 0. \tag{44}$$

That is, [11], we have a uniformly most powerful unbiased (U.M.P.U.) test of the hypothesis $\psi = 0$ versus alternatives $\psi \neq 0$, having critical region

$$\bar{\tau} < \tfrac{1}{2} + c_1 \quad \text{and} \quad \bar{\tau} > \tfrac{1}{2} + c_2 \tag{45}$$

where

$$\int_{c_1+\frac{1}{2}}^{c_2+\frac{1}{2}} p_{\bar{\tau}}(t \mid \psi = 0)\, dt = 1 - \alpha \tag{46}$$

and, since the test is U.M.P.U.,

$$\int_{c_1+\frac{1}{2}}^{c_2+\frac{1}{2}} (t - \tfrac{1}{2}) p_{\bar{\tau}}(t \mid \psi = 0)\, dt = 0. \tag{47}$$

From (32), we see that

$$p_{\bar{\tau}}(t \mid \psi = 0) = p_{\bar{\tau}}(1 - t \mid \psi = 0) \tag{48}$$

so that condition (47) requires that the test be one with "equal tails".

The writer is grateful to the referee for pointing out that results in a recent paper by Lehmann [12] applied to the distribution of time intervals here considered, enable one to conclude further that the above uniformly most powerful unbiased test is also a uniformly most powerful of all most stringent tests (as defined by Wald [13]).

Note that the statistic needed for these tests is the grand mean of all the time intervals between T_1 and the occurrence of each accident for each of the individuals observed. Since N is the total number of accidents, it is clear that an N of, say 12, may be obtained by observing 12 individuals, each of whom incurs 1 accident in the period of observation; or 6 individuals, each of whom incurs 2 accidents, etc. Consider, for example, one of the sets of data used in Part I of the paper [1]. This set of data was taken from the publication [14] of Farmers and Chambers and was concerned with the accident records of 166 London bus drivers during five successive years of service. The total number of accidents sustained by these 166 individuals in the five-year period was 1297. Thus, if the times of occurrence of these accidents were available, tests of the hypothesis of no contagion under the present conditions would have an N of 1297. It is true, however, in this example, that the underlying assumption of the independence of accident times among the individuals seems unrealistic.

Note that the manner of construction of the tests of this section require that N be fixed, since we are using the distribution of the r's, given the total number of accidents N. Since N is itself a random variable, the question may be raised as to whether we are not losing some of the information in defining our test (choosing the critical region) conditioned by the value of N. Furthermore, since the N_r, the number of individuals having exactly r accidents, may still vary with fixed N, subject only to the condition that $\sum_r rN_r = N$, one may wonder about possible loss of information in making tests independent of the particular set of values of the observed random variables N_r. We shall show that in the class of tests satisfying the requirement that the critical regions defining the tests be similar with respect to the parameters involved, the tests of this section have the property, roughly speaking, of using all of the information provided by the data. This fact is a consequence of the special form of the frequency function, under the null hypothesis $\psi = 0$, of the number of accidents sustained by an individual.

Let X be the random variable which equals the number of accidents sustained by an individual in a unit period of observation. Returning to Section 2 of this paper and letting $\lambda_m = \lambda_{m+1} = \cdots = \lambda_{m+n} = \lambda$, (or by considering the limiting form of solutions (6) and (7) with $T_2 = T_1 + 1$) we see that the probability of an individual sustaining exactly n accidents in a unit time interval when $\psi = 0$, is given by

$$p_X(n) = \frac{e^{-\lambda}\lambda^n}{n!}, \qquad n = 0, 1, 2, \cdots. \tag{49}$$

Now consider the accident data for which the tests of this section were devised. We have a set of N accident times in a unit period of observation, with $N = \sum_r rN_r$, involving $\sum_r N_r = K$, say, individuals. For convenience we define variables M_i as follows:

M_i = the number of accidents sustained by the i-th individual in the unit time interval, $i = 1, 2, \cdots, K$. Then $N = \sum_{i=1}^{K} M_i$ and each M_i when $\psi = 0$ has a Poisson distribution with unknown mean λ_i, where λ_i is the parameter per-

taining to the Poisson frequency function for the i-th individual's number of accidents.

Hence, under the assumptions of this section, N is a sum of K independent Poisson variates and is again a Poisson variate. Then the distribution of N is complete (cf. [15]). Furthermore, N is also sufficient for the distribution of $\bar{\tau}$ when $\psi = 0$, since the unordered τ_i's are known to act, conditionally on N, like independent random variables uniformly distributed on (0, 1) so that the distribution of τ is then independent of the parameters $\lambda_1, \lambda_2, \cdots, \lambda_K$. It follows from the Lehmann-Scheffé results, [15], that the only similar regions are those of Neyman structure with respect to N, i.e., roughly speaking, those for which the conditional probability of a point falling therein is equal to α, whatever the value of N.

Furthermore, considering the joint distribution of the M_i, $i = 1, 2, \cdots, K$, we note that it is the product distribution of K independent Poisson variates, the i-th variate having unknown mean λ_i. The set $(M_1, M_2, \cdots, M_K)$ being sufficient for the distribution of τ's when $\psi = 0$, we need only show the (bounded) completeness of the joint distribution of the M_i to apply the Lehmann-Scheffé results. Given that

$$(50) \qquad \sum_{m_1, m_2, \cdots, m_K} f(m_1, m_2, \cdots, m_K) \left(\exp - \sum_{i=1}^{K} \lambda_i\right) \prod_{i=1}^{K} \frac{\lambda_i^{m_i}}{(m_i)!} = 0$$

for every choice of $\lambda_1, \lambda_2, \cdots, \lambda_K$ (and hence in particular, for the case in which all the λ_i's are different), this implies that $f(m_1, m_2, \cdots, m_K) \equiv 0$, since $f(m_1, m_2, \cdots, m_K)$ is the coefficient of $\lambda_1^{m_1}\lambda_2^{m_2} \cdots \lambda_K^{m_K}$ in the expansion of 0 in powers of $\lambda_1, \lambda_2, \cdots, \lambda_K$. That is, the joint distribution of the M_i is complete. By the Lehmann-Scheffé results, we need then only work conditionally on the M_i's. But since N_r = the number of individuals sustaining r accidents, the N_r's are then fixed, which shows that the critical regions selected in our tests are independent of the particular set of N_r's used subject to the condition $\sum rN_r = N$.

The writer is indebted to the referee for pointing out the above analysis of the implications of the tests of this section.

Finally, it should be noted that the tests of this section apply to accident data in which each individual has the same length of time of exposure to accidents, which (clearly without loss of generality) we took to be a unit period of observation. It may be worthwhile to mention the fact that these tests may be generalized to use accident data involving periods of exposure to accidents which vary with the individual.

Given, say, K individuals, the i-th individual incurring n_i accidents in an exposure period of length L_i, and given the times τ_{ij} of occurrence of the j-th accident for the i-th individual (the times measured from the start of the individual's period of exposure), we may consider normalized random variables σ_{ij} defined by $\sigma_{ij} = \tau_{ij}/L_i$, $i = 1, \cdots, K; j = 1, \cdots, n_i$. Then if N is the total number of accidents, it may easily be shown that $\sum\sum\sigma_{ij}/N$ has exactly the same distribution under the null hypothesis $\psi = 0$ as $\sum\sum\tau_{ij}/N$ in the body of

the paper. Hence a test of the hypothesis $\psi = 0$ in this case having desired size may be obtained just as described in this paper. However, $\sum\sum\sigma_{ij}/N$ is now not sufficient and the test will not have the optimum properties of the tests of this paper. If the L_i's do not differ too much, it is probable that using the $\sum\sum\sigma_{ij}/N$ will result in not too great a loss of power.

Actually a sufficient test statistic for the case of varying times of exposure to accidents is the weighted mean $\sum\sum L_i\sigma_{ij}/N$ of the σ_{ij} (or, the grand mean of all the actual accident times, τ_{ij}) and the distribution of this statistic under the null hypothesis is that of the mean of K independent sets of random variables, the i-th set consisting of n_i independent identical variables uniform on $(0, L_i)$. This distribution (in the case of $n_i = 1, i = 1, \cdots, K$) has been studied recently by Olds [16].

Further discussion of the more general situation of varying times of exposure will be left for a later paper. In the following section we return again to the case of one period of observation for the accident times.

5. Power function for the UMPU test of the hypothesis of no contagion in the set of admissible hypotheses implying no mixture-no time effect-linear contagion. Since we shall be interested in detecting the presence of contagion of either kind (positive or negative) we shall consider in this section the problem of computing the power function for the UMPU test of Section 4.

Using (32) and (36), the exact power function $P(\psi)$ for a given test size α, is given by

$$P(\psi) = 1 - \int_{\frac{1}{2}-c_N}^{\frac{1}{2}+c_N} p_{\bar{\tau}}(t \mid \psi)\, dt \tag{51}$$

with

$$\alpha/2 = \int_0^{\frac{1}{2}-c_N} p_{\bar{\tau}}(t \mid \psi = 0)\, dt, \tag{52}$$

where the notation c_N is used to emphasize the dependence of this value on N.

It is obvious from the form of the density functions (32) and (36), that the computation of the exact power is a tedious procedure, even for N relatively small. Indeed, just the determination of the critical region boundaries by the numerical solution of the polynomial equation obtained from (32) is time-consuming. Since the variable $\bar{\tau}$ with density function (32) is asymptotically normal $(\frac{1}{2}, 1/\sqrt{12N})$, a first step in obtaining the approximate power consists in approximating for a given level of significance α, the critical region boundaries.

$$\bar{\tau} = \tfrac{1}{2} \pm c_N \qquad c_N \doteq c/\sqrt{12N} \tag{53}$$

where

$$\int_{-c}^{c} (1/\sqrt{2\pi})e^{-x^2/2}\, dx = 1 - \alpha. \tag{54}$$

Table I below gives a comparison of this approximation to c_N with the true values of c_N, for $N = 3, 6, 9, 12$ and $\alpha = .05$.

TABLE I

Comparison of c_N and $c\sqrt{12N}$ for $N = 3, 6, 9, 12$ and $\alpha = .05$.

N	c_N	$c/\sqrt{12N}$
3	.32289	.32667
6	.22931	.23099
9	.18769	.18860
12	.16275	.16333

Secondly, by the Central Limit Theorem, the variable $\bar{\tau}$ with density function (36) is asymptotically normal (μ, σ) with

$$\mu = \frac{\psi e^{\psi} - e^{\psi} + 1}{\psi(e^{\psi} - 1)}; \qquad \sigma = \frac{1}{\sqrt{N}} \frac{[(e^{\psi} - 1)^2 - \psi^2 e^{\psi}]^{\frac{1}{2}}}{\psi(e^{\psi} - 1)}. \tag{55}$$

We obtain, then, as an approximation, say $P_1(\psi)$, to the power function,

$$P_1(\psi) = 1 - \int_{w_1}^{w_2} 1/\sqrt{2\pi}\, e^{-x^2/2}\, dx \tag{56}$$

with

$$w_1 = \frac{\frac{1}{2} - \mu - c/\sqrt{12N}}{\sigma}, \qquad w_2 = \frac{\frac{1}{2} - \mu + c/\sqrt{12N}}{\sigma}. \tag{57}$$

Using the above approximation, Table II shows values of N needed for the power (as approximated) to exceed .90 for various values of ψ in the case of the UMPU test at the 5 per cent level of significance. The particular values of ψ used in this table were selected because of the significance of the factor $e^{-\psi}$ discussed in section 2. Thus, the detection of a ψ such that $e^{-\psi} = 2$ (or $\frac{1}{2}$) would seem to imply a rather high order of contagion in that the occurrence of each additional accident tends to double (or halve) the previous probability of avoiding accidents in a unit period.

TABLE II

Values of N required for power to be at least .90 for 5 per cent UMPU *test, when $e^{-\psi}$ has specified values of table.*

$e-\psi$	1.2	1.4	1.5	1.6	1.8	2.0
N	3825	1120	770	575	370	265

In view of the fact that N here is the total number of accidents, rather than the number of individuals, it would seem that the power of this test is fairly good.

6. Acknowledgments. The writer wishes to acknowledge her appreciation of the assistance provided by the Statistical Laboratory at Berkeley, and especially the many suggestions of the Director, Professor Neyman, for the preparation of this paper.

REFERENCES

[1] GRACE E. BATES AND JERZY NEYMAN, "Contributions to the theory of accident proneness," *Univ. of Calif. Publ. Statistics*, Vol. 1 (1952), pp. 215–275.

[2] M. GREENWOOD AND G. U. YULE, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents," *J. Roy. Stat. Soc.*, Vol. 83 (1920), pp. 255–279.

[3] E. M. NEWBOLD, "A contribution to the study of the human factor in the causation of accidents," *Industrial Health Research Board, Report No. 34*, H. M. Stationery Office, London, 1926.

[4] G. POLYA, "Sur quelque points de la théorie des probabilités," *Ann. Inst. H. Poincare*, Vol. 1 (1930), pp. 117–161.

[5] WILLIAM FELLER, *Probability Theory and its Applications*, Vol. 1, John Wiley and Sons, New York, 1950.

[6] WILLIAM FELLER, "On the theory of stochastic processes, with particular reference to applications," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 404–431.

[7] S. S. WILKS, *Mathematical Statistics*, Princeton University Press, 1943.

[8] P. S. DE LAPLACE, *Oeuvres Complètes*, Tome 8, Paris, pp. 279–321.

[9] M. G. KENDALL, *The Advanced Theory of Statistics*, Vols. 1 and 2, Charles Griffin and Co., London, 1948.

[10] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[11] J. NEYMAN AND E. S. PEARSON, "Contribution to the theory of testing statistical hypotheses, Part II," *Stat. Res. Memoirs*, Vol. 2 (1938) pp. 25–53 (p. 33).

[12] E. L. LEHMANN, "On families of admissible tests," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 97–104.

[13] A. WALD, "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Trans. Amer. Math. Soc.*, Vol. 54 (1943), pp. 426–482.

[14] E. FARMER AND E. G. CHAMBERS, "A study of accident proneness among motor drivers," *Industrial Health Research Board, Report No. 84, H. M. Stationery Office, London*, 1939.

[15] E. L. LEHMANN AND HENRY SCHEFFÉ, "Completeness, similar regions, and unbiased estimation, Part 1," *Sankhyā*, Vol. 10 (1950), pp. 305–340.

[16] EDWIN G. OLDS, "A note on the convolution of uniform distributions," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 282–285.