БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ ACADÉMIE BULGARE DES SCIENCES
ИЗВЕСТИЯ НА МАТЕМАТИЧЕСКИЯ ИНСТИТУТ
BULLETIN DE L'INSTITUT DE MATHÉMATIQUES
Том (Tome) XV

THE CONCEPT OF ACCIDENT PRONENESS

Violet R. Cane

1 INTRODUCTION

Accident data is often well fitted by a negative binomial distribution. It is well known that two explanations, one in terms of accident proneness, the other involving contagion, can be given for the occurrence of such a distribution. It is, however, sometimes said that we can decide between these explanations if complete information about the accidents is available—by this I mean that the time of every accident for each person in the sample is known. In this paper it is shown that one cannot, even with complete information, decide between the two explanations; that no experiment which involves increasing the risk of accidents for some subjects and not for others will help us to decide; and, moreover, that there are not just two explanations but an infinite number.

2. BASIC MODELS

Statisticians are naturally tempted to describe the occurrence of accidents in terms of a Poisson process. Call this Model 1 it is assumed that accidents occur at rate λu, where λ refers to some property of the person at risk, u refers to the danger of the situation in which accidents occur. The distribution of the number of accidents in a given time T then has p. g. f. $E(z^n) = \exp\{\lambda u T(z-1)\}$, a Poisson distribution.

Factory accidents did not conform with this model. Greenwood and Yule [4] therefore proposed Model 2: it is assumed that Model 1 holds for any given individual but that individuals may have different λ values; since λ is proportional to the expected number of accidents in unit time it can be thought of as measuring accident proneness. If the distribution of λ in the population at risk can be described by the probability density

$$c^k \lambda^{k-1} \exp(-\lambda c)\, \Gamma(k)$$

the p.g.f. for the distribution of accidents in time T is $E(E(z^n/\lambda))$ which is $c^k(c-uT(z-1))^{-k}$. It is convenient to absorb c into u and so replace c by 1.

McKendrick [7] was, it seems, the first to point out that the same distribution could be obtained from a contagion model, a particular form of

which is Model 3: it is assumed that a person who has had n accidents in time $(0, t)$ has a probability $\frac{k+n}{1+ut}\,udt$ (independent of the times of the preceding accidents) of having another in $(t, t+dt)$. All members of the population have the same chance, $kudt$ of an accident in $(0, dt)$.

By replacing u by $u(t)$, ut by $U(t) \int_0^t u(t)dt$, in each model we can allow for temporal fluctuations in the danger of the accident situation. In effect, we change the time scale, transforming t to $U(t)$ τ, say.

Irwin [6], Ashton [2] and Arbous and Kerrich [1] give summaries of previous work on accident data and list a large number of references.

3. EQUIVALENCE OF MODELS 2 AND 3

Suppose we know that a given person has n accidents at times t_i, $0<t_1< \quad <t_n<T$. The probability of such a result is

for Model 2: $\prod_{i=1}^{n} \exp\{-\lambda u(t_i-t_{i-1})\}\lambda u dt_i \exp\{-\lambda u(T-t_n)\}$

for Model 3: $\prod_{i=1}^{n}\left(\frac{1+ut_{i-1}}{1+ut_i}\right)^{k\ i\ 1}{}^{k+i-1}_{1+ut_i}\ udt_i\left(\frac{1+ut_n}{1+uT}\right)^{k\ n}$

(t_0 is to be taken as 0).

These expressions may be rewritten as

$$(n!\ dt_1 \qquad dt_n T^{-n})(\lambda u T)^n \exp(\quad \lambda u T)\ n!$$

and

$$(n!\ dt_1 \ldots\ dt_n T^{-n})\ (uT)^n(1+uT)^{-n} \quad \binom{k\ n\ 1}{n},$$

where the first bracket gives the probability that accidents occur at the specified times given that there are n accidents in all. Thus the distribution of $t_1, \ldots, t_n$, conditional on n accidents in time T, is the same in each case — it is, in fact, the distribution of the order statistic for a sample of n from a uniform distribution on $(0, T)$. Consequently, we can only hope to distinguish between the two models by reference to the total number of accidents sustained by each person in the sample. If, however, the observed distribution of the total number of accidents is consistent with a negative binomial distribution, we may equally well explain this result in terms of Model 2 or in terms of Model 3.

If therefore we have the detailed accident records of individuals covering a period of time T we can use these to check that the distribution of accidents at any time t $(t \leqq T)$ has a p. g. f. of the form $(1-utz+ut)^{-k}$. If this is not true then neither model fits the data, if it is true then both models fit and the observed distribution of the total number of accidents must be used to estimate u and k.

4. DIFFICULTIES OF EXPERIMENTAL COMPARISON

Since in Model 2 the accident proneness of an individual is constant, whereas in Model 3 if a person has one accident he is more likely to have another, we might think of trying to compare the two models experimentally by choosing two groups of people at random, one group to act as control and the other to be exposed to greater risk of accident. Suppose that, for the control group, $u(t)$ u throughout the period $(0, T)$ and that for the experimental group $u(t)$ u on (T_1, T), $u(t)$ αu on $(0, T_1)(\alpha>1, 0<T_1<T)$. On Model 2 the distribution of accidents during (T_1, T) is the same for each group and has p. g. f.

$$[1 \quad u(T \quad T_1)(1 \quad z)]^{k}$$

The same is true on Model 3; for consider the distribution of accidents in (T_1, T) given that r have occurred in $(0, T_1)$. The p. g. f. is $[1+u(T-T_1)(1+uT)^{-1}(1 \quad z)]^{k \quad r}$ and the unconditional p. g. f. for the number of accidents in (T_1, T) is obtained by taking the expectation of this expression with respect to r. Since

$$E(z^{k \quad r}) \quad z^{r}[1 \quad uT_1(1-z)]^{k}$$

we find on substitution that the required p. g. f. is $[1 \quad u(T-T_1)(1-z)]^{-k}$, as for Model 2.

5. TRUE OR FALSE CONTAGION

It is clear that, provided we have data which fit a negative binomial distribution, we cannot distinguish between true contagion (Model 3) and false contagion or accident proneness (Model 2) by any statistical treatment. There may be other reasons for choosing one model rather than the other; such reasons must be described in terms of prior information or prior belief. Bates and Neyman [3], for example, considered the two models for the case $u(t) \quad \alpha^{-1}(1 \quad \alpha t)^{\alpha^{-1}-1}$ and asserted, in effect, that the value 1 for α had a high prior probability if Model 2 held and a low prior probability if Model 3 held. This would lead one to accept Model 2 if α were close to 1, and to reject it otherwise.

6. OTHER MODELS

There is no mathematical difference between the two models, the difference lies in their meaning. Model 2 means that people differ (in respect of accident proneness) in ways which remain fixed; they do not modify their behaviour through experience. We may call this an interpretation in terms of heredity or nature.

Model 3 means that people are alike to begin with, they differ subsequently because they have changed through experience. We may call this an interpretation in terms of environment or nature.

Neither interpretation is easy to accept. An interpretation involving both nature and nurture would seem much more reasonable. We can readily provide a suitable model by using the following result (Gurland [5]):

If we have a p. g. f. of the form

$$(1) \qquad E(z^n/k_j)=\left(\frac{P_1}{1-Q_1z}\right)^{k_j}, \quad P_1+Q_1 \quad 1$$

and if k_j itself has a distribution given by

$$(2) \qquad E(w^{k_j}) \quad \left(\frac{P_2w}{1-Q_2w}\right)^{k}, \quad P_2+Q_2 \quad 1$$

then the mixed distribution has p. g. f.

$$(3) \qquad E[E(z^n\ k_j)]-E\left(\frac{P_1}{1-Q_1z}\right)^{k_j} \quad \left(\frac{P_2P_1}{Q_1+P_2P_1 \quad Q_1z}\right)^{k}$$

We can consider therefore Model 4 take (1) as $(1-bT(z \quad 1))^{-k_j}$ and regard it as arising from Model 3, so that it gives the distribution of the number of accidents for a person whose initial rate is k_jb, and interpret (2) as a description of the variation of k_j in the population. Then (3) becomes

$$[1-bTP_2\ ^1(z-1)]^{-k}.$$

Note: a. We could also allow k to vary in a distribution given by

$$E(s^k) \quad \left(\frac{P_3s}{1-Q_3s}\right)^{l}, \quad P_3\ ^1-Q_3 \quad 1$$

but then

$$E[E(w^{k_j}/k)]=\left(\frac{P_3P_2w}{1-(1-P_2P_3)w}\right)^{l}$$

is also of type (2).

b. The p. d. for any individual's accident record can be written as in section 3, so that the distribution of the times at which accidents occur, conditional on the total number of accidents, is the same for each model.

c. The result obtained in section 4 is a particular case of the result (3).

7. HOW MANY VERSIONS OF MODEL 4?

Suppose that we have accident data relating to the time range $0- \ t \ T$ which is well fitted by the distribution

$$(1-ut(z-1))^{-k};$$

u and k are estimated from the observations and are to be taken as given constants. We can interpret the data in terms of Model 4 if $u=b/P_2$, i. e. we can choose any P_2, $0<P_2\leq 1$, and then take $b=P_2u$.

If $P_2=1$, then $b=u$ and we have Model 3, i. e. a trivial mixture in which $k_j=k$ for all j. At the other extreme, if P_2 is small, we have $[1-$

$bt(z-1)]^{-k_j} = [1-P_2ut(z-1)]^{-k_j}$ $\exp[P_2k_jut(z-1)]$ approximately (since $k_j \geq k>0$) and $E(\exp i\theta k_jP_2)$ $(1-i\theta)^{-k}$ approximately. Thus as $P_2 \to 0$, the distribution for each individual tends to a Poisson distribution with mean λut (for time t), where $\lambda = P_2k_j$, and the distribution of λ tends to a Gamma distribution with index k, i. e. we have Model 2. Thus Models 2 and 3 are the extreme versions of an infinite set of models of type 4.

The square of the coefficient of vatiation for k_j in model 4, is Q_2/k; for fixed k we might describe the proportion of nature to nurture as given by $Q_2 : P_2$.

8. INTERPRETATION

The reason for the multiplicity of equivalent models lies in the form of the instantaneous risk for the contagion model. The probability of an accident in $(t, t+dt)$ given that n have occurred in $(0, t)$ may be written $\frac{k+n}{u^{-1}+t}\,dt$; if we regard k as the number of accidents occurring in a time u^{-1} preceding the beginning of observation, this probability depends only on the average number of accidents in the past. For Model 4 the corresponding probability is $\frac{k_j+n}{(P_2u)^{-1}+t}\,dt$, as if k accidents occurred in a time of length $(P_2u)^{-1}$; moreover, since equation (3) can be written $w^k[1-Q_2P_2^{-1}(w-1)]^k$, it is as though the last (k_j-k) accidents had occurred in a time interval of length $Q_2(P_2u)^{-1}$. Thus each model effectively assumes that at some fixed time every individual in the population at risk had an equal chance of sustaining an accident; the models differ in the fixed time chosen — for the pure contagion model it is taken to be the beginning of the observation period, for the mixed Poisson model it is taken to be a long time in the past; the ratio $Q_2 P_2$ specifies how far in the past we set this point of equality of experience.

9. DISCUSSION

If all that we require of a mathematical model is that it should produce equations which represent well the data so far observed and which may therefore be useful in predicting future results, then the fact that a number of different concepts may lead to the same equations is of no concern. All the accident models described would lead us to predict that the more accidents a person has had in the past, the more he is likely to have in the future. Unfortunately, mathematical models are not used only in this way. Often a model is based on some concepts which seem plausible and if the model works the concepts are held to be justified; it is then felt that some insight into the mechanism by which the data was produced has been obtained and this insight is applied to much more general situations. As an example of this, consider the problem of population growth.

Suppose we observe a number of colonies each initially of size k and find that the colony sizes at later times fit the distribution whose p.g. f. is

$$z^k(1-e^t(z-1))^{-k}.$$

We can interpret the data as the outcome of a pure birth process, i. e. as if each individual can reproduce with a constant birth rate (of unity). For such a process the probability of a birth in a given colony in $(t, t+dt)$, given that n births have occurred in this colony in $(0, t)$, is $(k+n)dt$. This is an example of Model 3 with $u(t)$ $\exp t$. (Waugh [8] has exploited this representation of the pure birth process to prove limit theorems.) Consequently there are other interpretations, for example: only founder members reproduce, the birth rate for any one of them is λe^t at time t (where λ is a random variate with p. d. exp (λ)). Another possibility (more appropriate for a bee-hive) is: only one of the founder members reproduces, the birth rate depends linearly on the population size. The explanation of the data as a pure birth process of the usual sort has a pleasing symmetry and also has some justification for demographic work with human and animal populations where we know the reproductive mechanism. It does not follow that we should accept this explanation of population size in other cases.

In the case of accident data variation in the observed results can be attributed partly to variation in the population, partly to variation inherent in the process and it has been shown that the partition can be effected in infinitely many ways; this is because differences between individuals can only be displayed by behaviour which is exhibited in the course of time and therefore can be "explained" by the process. This sort of difficulty can arise in a number of different fields, from the study of the development of mental illness to the study of the movement of allegedly identical elementary particles. It is clearly necessary to try to formulate classes of explanations which may apply to a given type of phenomena rather than attempt to defend one explanation at all costs or wrangle unprofitably over two.

REFERENCES

1. Arbous, A. G., J. E. Kerrich. — Biometrics, **7**, 1951, 340—342.
2. Ashton, W. The theory of road traffic flow. London, 1966.
3. Bates, G. E., J. Neyman. — Univ. Calif. Publ. Statist., 1, 1952, 215—275.
4 Greenwood, M., Udny Yule. — J. R. Statist. Soc., 83, 1920, 255—275.
5. Gurland, J. — Biometrika, **44**, 1957, 265—268.
6. Irwin, J. O. — J. R. Statist. Soc., A, **127**, 1964, 438—451.
7. McKendrick, A. G. — Proc. Edin. Math. Soc., **44**, 1926, 98—130.
8. Waugh, W. A. O'N. — J. R. Statist. Soc. B, **32**, 1970, 418—431.

Received 2. X. 1972

ВЪРХУ ПОНЯТИЕТО ЗА СКЛОННОСТ КЪМ ЗЛОПОЛУКА

Вайолет Кейн

(Резюме)

Обикновено данните за злополуките се описват добре с помощта на отрицателно биномно разпределение. Известно е, че за това могат да се дадат две обяснения: едното е свързано със „склонност към злополука", а другото — със „заразност от злополука". Понякога обаче казват, че можем да решим кое от двете обяснения е задоволително, ако имаме пълна информация за съществуващите злополуки, като под това се разбира, че времето на всяка една злополука за всеки един човек от извадката е известно. В тази работа се показва, че не можем, дори и ако разполагаме с пълна информация, да решим кое от двете е в сила и че нито един експеримент, при който за едни субекти рискът расте, а за други намалява, няма да ни помогне да вземем решение; нещо повече, съществуват не само две, а безкраен брой обяснения за злополуките.

О ПОНЯТИИ ПОДВЕРЖЕННОСТИ ПРОИСШЕСТВИЯМ

Вайолет Кейн

(Резюме)

Обычно данные о происшествиях хорошо описываются отрицательным биномиальным распределением. Известно, что этому можно предпослать два объяснения — первое связано с „подверженностью происшествиям", второе — с „заражением происшествием" Иногда считается, что можно выбирать из двух объяснений более удачное, если имеется полная информация о существующих происшествиях. Под этим подразумевается, что время для каждого происшествия с каждым человеком из выборки известно. В статье показывается, что даже для случая, когда мы располагаем всей информацией, невозможно решить задачу выбора из двух объяснений и что не существует эксперимент, который помог бы принять решение, если для одних субъектов риск возрастает, а для других уменьшается. Более того, существует не только два объяснения происшествий, их число бесконечно.