Neyman and Stochastic Models

By L. LeCam
University of California, Berkeley

"Le raisonnement est manifestement sans valeur; mais il n'en est pas moins remarquable que la formule put servir à Gompertz pour représenter la loi de survie de la table de Carlisle, entre 10 et 60 ans."

Henri Galbrun, 1924

1. Introduction

Neyman is well known for his fundamental contributions to the theory of Statistics, but he started his statistical career by a series of papers on applications of statistics. These were applications to agricultural experiments. Later on he was to spend much time and effort on applications to various scientific domains. We shall be concerned here with a particular aspect of Neyman's interests, namely the building and use of stochastic models. He did that in very many fields, from Astronomy to Zoology following a consistent philosophy throughout: Given a question about a particular phenomenon, Neyman tried to visualize the "mechanism" underlying the phenomenon. He then translated this vision into mathematical assumptions and formulas. *After* devising the stochastic model to his satisfaction, he would derive the statistical methods appropriate to the case at hand.

This procedure sets Neyman apart from the more typical "applied statistician" who has learned a large number of statistical methods and applies to the particular problem one of the procedures in his tool kit. It meant that Neyman had to learn about the field and consult with experts in it.

Of course when trying to decipher the mechanism behind a particular phenomenon, one cannot be sure that one has really caught what makes it work. Neyman's philosophy was that one should try anyway and that it is better to catch a bit of the mechanism than to use sundry "interpolation formulas". When pressed on this point he would occasionally quote H. Galbrun who wrote about the Gompertz-Makeham derivation of a distribution for human survival.

Here we shall sketch Neyman's construction of a few particular stochastic models. It will be seen that, even in very different fields the models are often related to each other.

2. The theory of clusters

In 1939, Neyman published a paper entitled “On a new class of contagious distributions applicable in entomology and bacteriology”. He describes the origin of the problem as that of comparing effects of insecticides by counting larvae in plots in a field. The paper is notable for its main contribution but also for several discussion of the role of models in Statistics.

Neyman notes that observed distributions of counts of larvae cannot be fitted by a Poisson distribution. There are too many empty plots and too few with one larva. He goes on to say that it is not difficult to find one of the reasons:

“Larvae are hatched from eggs which are being laid in so-called ‘masses’. After being hatched they begin to travel in search of food. Their movements are slow...”

He then goes on to say that “A similar explanation may apply to microorganisms counted in single squares of a haemacytometer or to colonies on parallel plates. However, here the situation is not as clear as in the case of larvae”. Later, on the same page, he says:

“Owing to the fact that the cause of the contagiousness of the distribution of larvae is clear... Consequently, if the theoretical distributions that we shall deduce fit the empirical ones, we shall be more or less justified in assuming that we guessed the actual machinery of movements of larvae. On the other hand, if the same theoretical distributions appear also to fit satisfactorily empirical counts of bacteria then in respect of these applications it will be safer to consider that we were lucky enough to find a sufficiently flexible interpolation formula”.

The passage set very clearly a philosophy that Neyman would apply in many other domains: If the “machinery” is clear and the formula fits one is *more or less* justified in assuming that our guess at the machinery was correct. However, if the machinery is not clear, we just have an “interpolation formula”.

The actual assumptions made in the derivation are as follows:

(A) A larva born at (ξ, η) and surviving at the time of observation will be found a random location according to a density $f(x-\xi, y-\eta)$.

(B) The mass of eggs located at (ξ, η) will furnish a random number S of survivers with $P[S=n]=p(n)$.

(C) The various larvae are asocial and behave independently.

(D) There are N masses of eggs on the field. They are distributed there independently and uniformly.

Eventually, Neyman fixes the average number of masses of eggs per unit area to a value m and lets the number of unit areas in the field go to infinity.

It is possible to put the assumptions (A) to (D) together and obtain a formula as follows

Let M be the measure that counts how many larvae are in any set and let γ be an arbitrary positive bounded measurable function, perhaps with compact support. Look at the random variable $Z = \int \gamma(x) M(dx)$. One can easily see that $E \exp\{-Z\}$ is equal to $[\int G[h(\xi)]d\xi]^N$ where $G(u) = \sum_n u^n p(n)$ and where $h(\xi) = \int \exp\{-\gamma(x)\} f(x|\xi)dx$. Here for simplicity of notation we have abbreviated $f(x-\xi, y-\eta)$ to $f(x|\xi)$, x and ξ being points in the plane.

It is seen that this expression depends on two arbitrary functions: The generating function G and the dispersal function $f(x|\xi)$. To obtain specific formulas, Neyman makes various choices for G and f, explaining that the choices are really arbitrary, because not much is known about either G or f. The choice where G is Poisson and f is uniform over an area A gives the "contagious distributions of type A".

It is interesting to note that at the start of his argument and later in the conclusion, Neyman mentions that numbers of larvae in adjacent plots are dependent random variables. However he does not attempt to compute any correlations. They can readily be obtained from the generating function written above.

Some 25 years later Neyman, with Elizabeth L. Scott, was going to consider a mechanism for the spread of epidemics in some respects analogous to the spread of the larvae. It starts with a population of susceptibles distributed geographically and a population of infected people who, after an incubation period, become infectious. In the meantime they have travelled to some random geographical location and they infect the susceptibles they meet there. The difference between the larvae model and the epidemic one is that the newly infected people will become infectious after the required incubation period. They have travelled in the meantime and the process feeds on itself and can continue indefinitely.

Neyman and Scott give formulas for the generating functions of the number of infected people to be found in any finite system of disjoint regions in the habitat. This is done for the n^{th} generation in the epidemic process as well as for an epidemic continuously nurtured by mutations of the pathogens.

Neyman and Scott point out two unexpected results of that study. One is that the probability of an epidemic getting out of hand in a small community is just the same as the probability of the epidemic becoming explosive in the entire habitat. The other is that vaccinating at random a proportion θ of the population reduces the expected total size of the epidemic by a factor $(1-\theta)/\theta$.

Another paper of Neyman, with Grace E. Bates, deals with another aspect of "contagion". It is about accident proneness. It had been observed that in many instances the number of accidents occuring in a given period of time in a well defined population (such as bus drivers in London) has a distribution that is well fitted by a negative binomial. O. Lundeberg and W. Feller observed that the same negative binomial can be obtained from two entirely different mechanisms. One is that different individuals have different proneness to accidents, but a proneness that does not change in time. A different mechanism would be that all individuals start alike but that each accident disturbs the situation and makes the affected individual more susceptible to further accidents. The particular disturbance of proneness was imitated from a paper of G. Pólya. The basic assumption states that if an individual has by time S incured m accidents, then the probability $P_{m,0}(S,T)$ of having zero accidents in the interval $(S,T]$ is such that

$$\frac{dP_{m,0}(S,T)}{dT} = -\lambda \frac{1+\mu m}{1+\nu S}$$

where λ, μ and ν are numerical coefficients. Grace E. Bates and Neyman argue that if one counts the number of accidents in several intervals $I_1, I_2, \ldots, I_k$ then the mixture possibility with individuals of different proneness can be distinguished from the Pólya contagion process if $k \geq 2$, *except* for the odd case where $\nu = \lambda\mu$.

The problem was of much interest to the U.S. Air Force who was trying to predict the probability that a pilot would have severe accidents by recording his past number of minor accidents. Neyman did not get to try his model on actual Air Force data because as a result of the political climate of the time he was denied clearance to look at the data.

Turning his attention from larvae to galaxies, Neyman, around 1950 was to argue that galaxies occur in clusters, probably by a mechanism similar to the one he had guessed in 1939 for haemacytometer counts or bacteria on Petri plates: They might attract each other. That such is the case for galaxies is certainly true. He embarked, with E.L. Scott in a long study of the spatial distribution of galaxies. They assumed that clusters occur around "centers" distributed Poisson wise uniformly, gave each "center" a random number of galaxies and placed them independently of one another at random distances from the center. This simple clustering model seemed to fit reasonably well the observed counts of galaxies in the sky. However, when Neyman and Scott produced a simulated appearance of the sky from their model they were very surprised: The actual pictures of the sky were a lot more lumpy than what their simulation had produced. A picture of a piece of that simulation appears in Scientific American (Sept. 1956).

One had to cluster more. Neyman and Scott proceeded to do so, introducing recursively what they called n^{th} order clustering. The first order clustering consists of the system we just described with "centers" around which one places galaxies. An n^{th} order clustering process proceeds in the same way but the "centers" instead of being taken Poissonwise are themselves taken from an $(n-1)^{\text{th}}$ order clustering process. Even with only a 2nd-order clustering process the simulated appearance of the sky looked much more like the real thing. Neyman and Scott pursued that study for many years, working in particular on the abundance of different types of galaxies (elliptical, spirals, etc.).

This clustering picture was the one almost universally used by astronomers till the late seventies. By that time many more observations were available, in particular for estimates of the distance of various galaxies. Due to the work of G.D. Abell and others we have now a different view of the organization of the universe; It looks more like a mass of (empty) soap bubbles with galaxies strewn around the places were bubbles touch each other. As said above, that picture became available only shortly before Neyman's death. One wonders what he would have done with it had he had a chance.

The Neyman-Scott model of the spatial distribution of galaxies was motivated by the fact that galaxies attract each other. Nowadays several cosmologists go further. They attribute the bumpy appearance of the skies to minute local fluctuations in the radiation soup that followed the Big Bang and are trying to argue from the basic physical principles of quantum theory. The fact that to get something similar to the actual sky they have to assume that 90 to 99% of the matter in the universe is "dark" and invisible certainly should incite some caution.

3. Carcinogenesis

The collaboration between Neyman and Scott on cosmology extended from around 1950 to Neyman's death in 1981. Another collaboration of long duration had to do with carcinogenesis. This started around 1957, then around 1960 Neyman spent several months at the National Institutes of Health in Washington D.C. There he met M.B. Shimkin, who with collaborators, had been studying the experimental induction of lung tumors in mice by injection of urethane.

Urethane is a water soluble fairly simple chemical that had a history of use as a veterinary anesthetic. Shimkin and Polissar had carried out experiments where mice were sacrificed at various times after urethane injection. They found not only frankly cancerous cells but also modified cells occurring in what they called "hyperplastic loci". They adduced that these may be precursors of cancer cells. For more information on this point see the article by M.B. Shimkin *et al* in the 4th volume of the Fifth Berkeley Symposium.

Neyman, who had earlier (unpublished) proposed a two stage theory of carcinogenesis seized the opportunity to test a theory on actual experimental data.

Multistage theories of carcinogenesis have a long history. A short summary is given by P. Armitage and R. Doll in the 4th volume of the Proceedings of the Fourth Berkeley Symposium.

The particular model considered by Neyman and Scott is one in which

i) The growth of both benign and concerns tumors are described by birth and death processes. For benign tumors the process is subcritical. For cancer it is supercritical.

For injection at time zero, let $Df(t)$ represents the amount of carcinogen present in the time at time t for a function f such that $\int_0^\infty f(t)dt = 1$. Then:

ii) Cells will suffer first order mutations according to a Poisson process of intensity proportional to $Df(t)$.

iii) Cells that are daughters of first order mutants and their descendants can suffer a second mutation with intensity $a + bDf(t)$. If so they become cancerous, subject to the supercritical branching process.

Neyman and Scott also incorporate in the model a provision for counting errors, with small clones more likely to be missed than larger ones.

Some of the conclusions derived are as follows. Let X be the tumor count at the end of the experiment. Then, if excretion of the carcinogen is rapid the expectation of X for a single injection is linear in the dose D. Under the one stage hypothesis, or if $b = 0$ the expectation of X is always proportional to the dose D and does not depend on the function f.

On the contrary, for a two-stage model, the expectation of X will depend on f.

These conclusions, when compared to experimental results, favor a two-stage model with $b > 0$.

It should be noted that Neyman and Scott express this conclusion very cautiously. They had obtained the collaboration of Dr. Margaret White and colleagues at the Donner Laboratory, Berkeley. Dr. White performed many experiments on mice, including some where the time pattern of excretion of urethane was determined. One of the conclusions is that the rate of excretion decreases when the dose is increased. This, if taken into account in the Neyman Scott model would further enhance the parabolic shape of the dose-response relation.

Many other conclusions were tested including the effect of giving a total dose D but in a fractionated protocol. This for adult mice, decreases the tumor yield, but may increase it for young mice. The model can fit these observations at least qualitatively, but the case is not closed.

4. Struggle for existence

In a totally different domain Neyman and Scott became interested in the experiments carried out by Thomas Park on the struggle for existence. Park had two species of flour beetles Tribolium castaneum and Tribolium confusum. They were bred in small vials with monthly replacement of the flour and counts of all present, eggs, larvae, pupae and adults. The two species at adult stage could only be distinguished under the microscope. Park had surmised that, since the two species were very much alike, competition would be severe.

One of his observations was that if bred separately each one of the two species survived “indefinitely” that being of the order of 30 years in Park’s experiments. They established stable populations. If, however, two or four beetles of each species were placed in the same vial then, within a year or so, there was only one species left in the population.

It was not always the same and the proportion of vials where castaneum won over confusum depended on temperature and humidity conditions.

Flour beetles behave in a most disgraceful manner: They cannibalize their eggs and pupae, be they of their own species or another.

Neyman, Park and Scott built models of such interacting population with different fertility and voracity. The model was a stochastic version of the deterministic struggle for life of Volterra and Lotka. As is well known, the resulting equations are not solvable in any analytic manner.

Neyman, with Park and Scott give a report in the Third Berkeley Symposium, vol. 3.

The total population in Park's vials was of the order of 400 beetles per 8 grams of flour. It would be interesting to find solutions of the equations on a high speed computer.

5. Radiation

Neyman was concerned for a long time with effects of radiation. Together with Prem S. Puri he devised models of the action of radiation on cells in culture. One irradiate the cells, separate them and plate them on Petri-dishes. The "survivors" form colonies that can be counted. Some of the colonies are disorganized and aberrant. They represent cells that underwent a malignant mutation in the process.

One of the typical results of such experiments can be described as follows. One plots the logarithm of the surviving fraction as function of the radiation dose in rads or Grays. If the radiation operates in such a way that any single hit on the nucleus of the cells will kill it, the dose response curve would be a straight line. This is what is usually observed for high LET irradiation, say, by neutrons or accelerated heavy ions.

On the contrary low LET radiation such as X-rays or gamma rays, or electrons, produces a different type of dose response curves. They have a "shoulder". That is the response is curved and concave, looking straighter for large doses that kill a high proportion of the cells.

Neyman and Puri attempted to construct a stochastic model that would accomodate both possibilities. See Proceeding of the Royal Soc. of London 1981. The model has a mechanism for the induction of lesions and for their repair or misrepair.

Neyman and Puri do get a shoulder but that is at the cost of representing the length of time T the cell was irradiated as $T = D/\rho$ where D is the total dose and ρ is the dose rate. Later Yang and Swenberg were to modify the model. They introduced a different definition of "death". The Neyman Puri cells could survive with an indefinite number of unrepaired lesions (that did not become "lethal"). The Yang-Swenberg cells die if they have unrepaired lesions. Also Yang and Swenberg modified the initiation of lesion process in a manner analogous to the Kellerer-Rossi idea: The effect of incoming radiation becomes more severe as the lesions accumulate. Such a model does produce shoulders without any difficulty. However the Yang-Swenberg model, just as the Neyman-Puri one uses a linear mechanism for repair. That is, lesions are repaired independently of each other. There is considerable evidence that such is not the case. C. Tobias and colleagues have proposed mechanisms where a form of repair is an interaction of two lesions. A stochastic, Markov chain version, of Tobias model was investigated by N. Albright. One could insert that in the Yang-Swenberg model, but then the equations are impossible to solve. Further investigation by R. Sachs and others are likely to shed some light on what is really happening, but, at the time of this writing the case is not closed.

6. Conclusion

The above are just a few of Neyman's contributions to applications of Statistics. They have been chosen mostly because the stochastic models he used had a reasonable background of "mechanism" behind them instead of mere "interpolating formulas". He clearly prefered the "mechanisms". Some of Neyman's other contributions are very good indeed but they do not quite carry that flavor.

One can cite, for instance, his thesis of 1923, partly translated in Statistical Science of November 1990. In it Neyman considers fictitious quantities called $U_{i,k}$ which would be the true yield of variety i if grown on plot k. He allows for measurement errors and random variability but bases his estimates and calculations of variances on an assumption that the assignment of varieties to plots is completely random. This is a model in a sense but it could also be taken as an instruction to the experimenter to "randomize", but Neyman himself says that he did not mean it that way. He attributes to R.A. Fisher the idea that one must randomize. For the analysis Neyman introduces "fertility gradients" with polynomial form. The paper is now considered a classic.

Another part of Neyman's work was his discussion of sampling human populations. The methods he introduced have now been adopted almost everywhere.

He did much more, but it would be too long to report it here. Still one can say that Neyman had a particular flair for those domains of science where sound statistical thinking would be useful. As far as I know he was one of the first statisticians to look at applications of statistics in molecular biology. This is reflected in a large volume of the Sixth Berkeley Symposium.

He was always full of energy and ideas and “imprinted” them on his students in courses or in individual contacts.

References

Feller, W. (1949). "On the theory of stochastic processes, with particular references to applications". *Proc. Berkeley Symp. on Math. Stat. and Proba.* University of California Press.

Galbrun, H. (1924). *Traité du calcul des probabilités et de ses applications*, pas Emile Borel, Tome III fasc 1, Assurance sur la vie, Calcul des primes par Henri Galbrun. Paris, Gauthier Villars (see page 44).

Lundeberg, O. (1940). *On random processes and their application to sickness and accident insurance.* Uppsala, Almquist and Wiksells.

Neyman, J. (1923). "Justification of applications of the calculus of probability to the solution of certain questions of agricultural experimentation (Polish, German summary)." *Polish Agric. Forest Journ.* **10**, 1-51.

Neyman J. (1939). "On a new class of contagious distributions, applicable in entomology and bacteriology." *Ann. Math. Stat.* **10**, 35-57.

Neyman J. (1952). "Contribution to the theory of accident proneness II. True or false contagion" (with Grace E. Bates). *Univ. of Calif. Publ. in Statist.* **1**, 255-276.

Neyman J. (1956). "Struggle for existence. The Tribolium model: Biological and statistical aspects" (with Thomas Park and E.L. Scott). *Proc. Third Berkeley Symp. on Math. Stat. and Proba.* **4**, 41-79. University of California Press.

Neyman J. (1981). "A hypothetical stochastic mechanism of radiation affects in single cells" (with Prem S. Puri). *Proc. Royal Soc. London* B **213**, 139-160.

Tobias, C.A. (1985). "The repair-misrepair model in radiobiology. Comparison with other models." *Radiation Research* **104** #2 S577-S595.

Yang, G.L. and Swenberg, C.E. (1991). "Stochastic modeling of dose-response for single cells in radiation experiments." *Math. Scientist* **16**, 46-65.