Empirical Probability Generating Function: an Overview

Miguel Nakamura & Víctor Pérez-Abreu

Tech. Rept. I-92-2 (CIMAT/PE) 120

Received: June 25, 1992 Approved: July 20, 1992

EMPIRICAL PROBABILITY GENERATING FUNCTION: AN OVERVIEW[1]

Miguel Nakamura

and

Víctor Pérez-Abreu

Centro de Investigación en Matemáticas
Apdo. Postal 402, Guanajuato, Gto. 36000, México

Summary

A convenient approach to the statistical analysis of distributions for counts is possible using the empirical probability generating function. In this paper we give an overview of recent results and show the usefulness and advantages of this methodology. On one hand, there are some stochastic models in which the probability generating function arises naturally and therefore it is consistent with nature to use its empirical counterpart. On the other hand, this statistical tool has demonstrated to be useful in the study of classical statistical problems of distributions for counts, especially in exploratory data analysis, rapid multi-parameter estimation and testing the goodness of fit. Our recommendation is to make allowance for the empirical probability generating function when dealing with statistical inference for discrete distributions.

Key words: Discrete distribution, Distribution for counts, Empirical probability generating function, Estimation, Goodness of fit, Random counts, Statistical transform methods.

[1] Research partially supported by CONACYT Grant 20E9105

1.- Introduction.

The use of inference methods based on statistical transforms has been considered by several authors (see Feuerverger & McDunnough, 1984 and Csörgó & Mason, 1989). For example, in dealing with inferences for stable laws, the empirical characteristic function appears as a fundamental tool since the characteristic functions of such laws are tractable while the corresponding densities or distributions functions are not (see Prakasa Rao, 1987, Ch. 8 and references therein). On the other hand, the empirical moment generating function has also proved to be a very useful transform in the study of several statistical problems as shown, for example, in the interesting papers by Read (1981), Epps et. al (1982), Feuerverger (1988), Csörgó & Teugels (1990) and Baringhaus & Henze (1991).

When dealing with inference for distributions for counts, the empirical probability generating function (epgf) appears as the natural statistical transform to be considered. It has been used in several contexts by Kocherlakota & Kocherlakota (1986, 1990), Kemp & Kemp (1988), Marques & Pérez-Abreu (1989), Rueda, Pérez-Abreu and O'Reilly (1991), Baringhaus & Henze (1992) and Nakamura & Pérez-Abreu (1991), amongst others. Although all three statistical transforms are closely related, and one can be obtained from the others, the probability generating function (pgf) possesses convenient features not shared by the Fourier or Laplace Transform such as being a real valued continuous analytic function which always exists in the range [0,1]. However, the main reason for the utilization of the pgf is its simplicity and the potential for use of its empirical counterpart in statistical analysis of distributions for counts.

In this paper we give an overview of several applications of the epgf to methods of statistical inference for distributions for counts. Section 2 presents the epgf and some of its main properties. Section 3 deals with the epgf as a tool for exploratory data analysis. Section 4 considers the use of the epgf in the problems of estimation of parameters, hypothesis testing, goodness of fit and change-point estimation for distributions for counts. Section 5 presents two applications, one related to a cumulative damage model and the other to the estimation of the distribution function of a maximum. Both models share the feature that the pgf appears as an intrinsic tool whose properties may be exploited.

2.- The empirical probability generating function.

Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution F over 0, 1, 2, ... , with corresponding probabilities p_k $k = 0, 1, 2, \ldots$. The *empirical probability generating function* is defined as

$$\phi_n(t) \doteq (1/n) \sum_{l=1}^{n} t^{X_l},$$

for $t \in [0,1]$. This transform of the empirical distribution function F_n, is an estimator of

$$\phi(t) = E(t^{X_1}) = \sum_{k=0}^{\infty} p_k \, t^k \qquad |t| \le 1 ,$$

the *probability generating function* associated to F. The relation $M(t) = \phi(e^t)$ between the moment generating function M(t) and the probability generating function $\phi(t)$ gives the corresponding relationship between their empirical counterparts $M_n(t)$ and $\phi_n(t)$. Both are special cases of general statistical transforms (see Feuerverger & McDunnough, 1984) of the form $\int g_t(x)\, dF(x)$ with kernels $g_t(x)$ equal to e^{tX} and t^x respectively. While $\phi(t)$ always exists for $t \in [0,1]$, M(t) might not. Both $\phi_n(t)$ and $\phi(t)$ are analytic (and therefore continuous) functions on $t \in [0,1]$. The driving idea for constructing methods for statistical inference which are suitable for a distribution for counts F, is based on the fact that $\phi_n(t)$ is an estimator of $\phi(t)$. The efficiency of some of these methods is then obtained from Feuerverger & McDunnough (1984). A pleasant feature of working with $\phi_n(t)$ is that it is a continuous real valued function, while the empirical distribution function F_n or the empirical density are not. This allows us to adopt as a basic framework the well-known Banach space of continuous real valued functions C[0,1]. The reason for mentioning C[0,1] and not C[-1,1] is that in some of what follows we shall also consider other useful transforms like $Y(t) = \log(\phi(t))$ and its empirical counterpart $Y_n(t) = \log(\phi_n(t))$.

For each fixed t, $\phi_n(t)$ is an unbiased estimator of $\phi(t)$. Moreover, by the law of large numbers $\phi_n(t)$ is a consistent estimator for $\phi(t)$, and by the central limit theorem $n^{1/2}(\phi_n(t) - \phi(t))$ converges in distribution to a Gaussian random variable with zero mean and variance $\sigma^2(t) = \{\phi(t^2) - (\phi(t)^2)\}$. More interesting limiting results can be obtained in functional form (uniformly in t) in the space C[0,1]. Thus, for example (see Feuerverger, 1988 and Nakamura & Pérez-Abreu, 1991), it holds almost surely that

$$\sup_{0\le t\le 1} |\phi_n(t) - \phi(t)| \longrightarrow 0$$

and

$$\sup_{0<\varepsilon\le t\le 1} |Y_n(t) - Y(t)| \longrightarrow 0 \ . \tag{2.1}$$

Furthermore, if $f^{(k)}$ denotes the k-derivative of a function f, it follows from Feuerverger (1988) that almost surely

$$\sup_{0<t<1} |\phi_n^{(k)}(t) - \phi^{(k)}(t)| \longrightarrow 0$$

and

$$\sup_{0<t<1} |Y_n^{(k)}(t) - Y^{(k)}(t)| \longrightarrow 0 .$$

The limiting behavior in C[0,1] of the empirical probability generating processes $Z_n(t) = n^{1/2}\{\phi_n(t) - \phi(t)\}$ has been studied in different situations. For a random sample the sequence of processes Z_n converges weakly in C[0,1] to a zero mean Gaussian process with covariance function

$$K(s,t) = \phi(st) - \phi(s)\phi(t). \qquad (2.2)$$

Marques & Pérez-Abreu (1989) consider weak convergence of $Z_n(t)$ when $X_1, X_2, \ldots,$ is a sequence of stationary dependent integer valued random variables, dealing as well with the case of multivariate discrete distributions. Assuming that $X_1, X_2, \ldots$ are independent and identically distributed random variables, Feuerverger (1988) studies weak convergence in C[0,1] of $Z_n^{(k)}(t)$ and $n^{1/2}\{Y_n^{(k)}(t) - Y^{(k)}(t)\}$, while Rueda et. al (1991) consider weak convergence when observations come from a distribution having pgf $\phi(t,\theta)$ and the parameter θ is unknown and must be estimated. In the latter case $\hat{Z}_n(t) = n^{1/2}\{\phi_n(t) - \phi(t,\hat{\theta}_n)\}$ converges weakly in C[0,1] to a continuous zero mean Gaussian process with covariance function given by

$$R(s,t) = K(s,t) + (I(\theta) -2)) \frac{\partial}{\partial\theta} \phi(s,\theta) \frac{\partial}{\partial\theta} \phi(t,\theta) \ ,(2.3)$$

where $\hat{\theta}_n$ is the maximum likelihood estimator of θ and $I(\theta)$ is the Fisher information of $p_k = p_k(\theta)$, $k = 0, 1, 2, \ldots$.

Other asymptotic properties which have been studied include the work by Csörgó & Mason (1989) where bootstrapping of the epgf $\phi_n(t)$ and other statistical transforms are considered. All the above limiting results are important to establish asymptotic properties of any statistical procedures based on the epgf $\phi_n(t)$.

The preceding ideas may be defined as well for multivariate discrete distributions. Let $\underline{X} = (X_1,\ldots,X_r)$ be an r-dimensional discrete random vector with multivariate probability generating function $\underline{\phi}(\underline{t}) = E(t_1^{X_1} \dots t_r^{X_r})$, $\underline{t} = (t_1,\dots,t_r)$, and let $\underline{X}^n = (X_1^n,\dots,X_r^n)$, $n \geq 1$, be a sequence independent random vectors from this distribution. The *multivariate empirical probability generating function* of the first n observations is defined as

$$\underline{\phi}_n(\underline{t}) = (1/n) \sum_{i=1}^{n} t_1^{X_1^n} \dots t_r^{X_r^n} .$$

Limiting results similar to the univariate case can be obtained for $\underline{\phi}_n(\underline{t})$ (see Marques & Pérez-Abreu, 1989).

3.- Exploratory data analysis.

In this section we discuss a graphical method considered previously by Nakamura & Pérez-Abreu (1991) based on the plot of $Y_n(t) = \log(\phi_n(t))$ on [0,1] which may be useful in preliminary analysis of count data. Supplementary examples are provided in this section for illustration. This exploratory method is especially useful in identifying feasible models for random counts as well as recognizing possible outlying observations or the homogeneity or independence of several samples. This graphical scheme has appealing features in that it involves a plot of a continuous function instead of a function which has jumps occurring at observed data points (as in the case of probability plots and cumulative distributions), it does not require parameter estimation, and the occurrence of ties in data is unimportant.

The driving idea in this section is the result (2.1), which asserts that if the model is correct, $Y_n(t)$ should be close to Y(t), since it is a consistent estimator of Y(t) in C[0,1]. In this section observations X_1, X_2, ... are not assumed to be independent. It is sufficient to consider them as coming from a strictly stationary ergodic sequence, since (2.1) still holds in this case.

3.1 *Identifying distributions for counts.*

The following facts are useful in identifying a possible model for random counts when plotting $Y_n(t)$ for $t \in [0,1]$ (see Nakamura & Pérez-Abreu, 1991). For a Poisson distribution with mean λ, $Y(t) = \lambda(t-1)$, which is a straight line with intercept $-\lambda$ and zero at $t = 1$. If the Poisson model is correct, the intercept term in the plot of $Y_n(t)$ yields a preliminary estimate of $-\lambda$. For a binomial distribution $Y(t)$ is a concave function while for a negative binomial or general mixtures of Poisson distributions, the shape of $Y(t)$ is always convex. For a truncated distribution for counts, that is, when $p_0 = 0$, $Y(t)$ diverges to $-\infty$ as t converges to zero. In particular, for a truncated Poisson distribution, this behavior near $t = 0$ is like the function $\log(t)$ and close to $t = 1$ it resembles a straight line. The shape of $Y(t)$ under a heavy-tailed distribution is convex near $t = 1$ while concave near $t = 0$. This is the case of distributions with long positive tails, such as the logarithmic series or zeta distributions (see Johnson & Kotz, 1969 p. 166 and p. 240 respectively). Figure 2.1 plots the described behavior of $Y(t)$ of selected distributions for counts of some of the types specified above. As an illustration, in Figure 2.2 we have plotted $Y_n(t)$ for two different sets of previously analyzed data. On the one hand, we have the yearly death by horsekicks in the Prussian army recorded by von Bortkiewicz (1898) over the twenty years 1875-1894 (observations are 3, 5, 7, 9, 10, 18, 6, 14, 11, 9, 5, 11, 15, 6, 11, 17, 12, 15, 8 and 4). As the figure suggests there is no evidence that the data come from a Poisson distribution but rather from a truncated Poisson model. In the same plot we consider the plot of $Y_n(t)$ for counts of characteristic subduction earthquakes on Mexico's Pacific coast (from Jara & Rosenblueth, 1988) over periods of ten years between 1806 and 1985. This plot indicates that data cannot support the hypothesis that this variety of earthquakes results from a Poisson process. Rather, they display a mixture of Poisson or extra Poisson behavior. Actual observations are 1, 2, 0, 2, 1, 0, 3, 0, 2, 4, 5, 0, 6, 3, 1, 3, 2 and 7.

3.2 *Detecting outlying observations.*

A way of looking for an outlying observation consists in evaluating the variation of $Y_n(t)$ (as a function of t) by a leave-one-out procedure. That is, leave out X_i, construct $Y_n(t)$ from the remaining $n - 1$ observations and show all resulting n curves on the same plot. The curve that exhibits a small (large) change of $Y_n(t)$ in $0 < t < 1$ along with a large (small) change for $t > 1$, indicates that the observation left out is possibly a large (small) outlier. This procedure is clearly extended if we are interested in looking for more than one possible discordant observation.

As an example of small outlier, in Figure 2.3 we applied the method just described to data on quality inspection of a manufactured item presented in Barnett & Lewis (1983, p. 201). Observations are 5, 4, 4, 5, 4, 1, 4, 5, 3, and 4. The leave-one-out procedure for $Y_n(t)$ indicates that the value 1 is a small outlier. A discordancy test for this value conducted by Barnett & Lewis (1983) gives weak evidence for regarding it as a discordant observation from a binomial random sample.

For an example of possible upper outliers, we consider Figure 2.4 for the data on Household Size on a Housing Allowance Demand Experiment (Hoaglin & Tukey, 1985, Table 9.4). Observations are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 with corresponding frequencies 210, 315, 292, 176, 125, 57, 38, 18, 6, 1, 0, 1. The leave-one-out procedure does not indicate the presence of any single upper outlying observation.

3.3 *Analysis of k-samples.*

The plot of t against Y(t) may also be useful in the statistical k-sample analysis of general distributions for counts, as tests of homogeneity or trends and independence. The first approach is simply plotting, on the same graph, the corresponding $Y_n(t)$ for each of the k samples. For example, Figure 2.5 shows the graphs of $Y_n(t)$ for two sets of data presented in Stewart & Campbell (1970), who studied the claim experience of a single car in the entire driving population of North Carolina over a four-year period and published the distribution of drivers by number of accidents during two consecutive two-year periods. As indicated by the plot, there is evidence that the number of claims over the two periods do not arise from the same distribution. In fact the figures suggest that one conforms to a Poisson distribution while the other to a mixture of Poissons.

An exploratory procedure to test for independence of two random discrete samples can be conducted as follows using the multivariate version of the epgf. Let $\phi_1(t_1)$ and $\phi_2(t_2)$ be the marginal pgf's and $\underline{\phi}(\underline{t})$ be the joint pgf, and let $\phi_1^n(t_1)$, $\phi_2^n(t_2)$ and $\underline{\phi}^n(\underline{t})$ be their empirical counterparts. Under the hypothesis of independence $\underline{\phi}(\underline{t}) = \phi_1(t_1)\ \phi_2(t_2)$ for all $\underline{t} = (t_1,t_2)$, so a three dimensional plot of $\log(\underline{\phi}^n(\underline{t})) - \log \phi_1^n(t_1) + \log\ \phi_2^n(t_2)$ should reveal that this function is identically zero.

4.- Procedures in statistical inference based on the empirical probability generating function.

4.1 *Estimation and testing.*

Let $X_1, \dots, X_n$ be a random sample from a discrete distribution $F(x,\theta)$ over 0, 1, 2, ... , having pgf $\phi(t,\theta)$ where the true value of $\theta \in \Theta$ is unknown.

A general framework for estimation phrased in terms of statistical transforms is given by Feuerverger & McDunnough (1984); it covers the pgf as a special case, although most investigations concern the empirical characteristic function and the empirical moment generating function as specific cases (see the aforementioned paper for a rich list of references). In their context, Θ is a real, open interval, and one class (of three presented by Feuerverger & McDunnough) of estimates of θ based on the pgf is obtained in the following way. Fix $\underline{t} = (t_1, \dots, t_k) \in [-1,1]^k$, define

$$g(\underline{t},\theta) = (\phi(t_1,\theta), \dots, \phi(t_k,\theta))$$

$$g_n(\underline{t}) = (\phi_n(t_1), \dots, \phi_n(t_k)) ,$$

and estimate θ by solving

$$\underline{d}\ g^T(\underline{t},\theta) = \underline{d}\ g_n^T(\underline{t})$$

where $\underline{d}$ is a 1xk vector of constants. There are conditions under which a sufficiently extended grid $t_1, \dots, t_k$ yields an estimator of θ with arbitrarily high relative efficiency (provided $\underline{d}$ is chosen optimally). A continuous version of the previous estimating equation, in which all values of t instead of a finite subset of them play a role in estimation, is given by

$$\int_{-1}^{1} \phi(t,\theta)\ dH(t) = \int_{-1}^{1} \phi_n(t)\ dH(t).$$

See Feuerverger & McDunnough (1984) for details, as well as other classes of estimators based on statistical transforms. In the same paper, a Wald-type test of the hypothesis H_0: $\theta = \theta_0$ is displayed, motivated by the asymptotic normality of $\underline{g}_n(\underline{t})$. The test statistic in the context of pgf's, is

$$D_Q^2 = n(\underline{g}_n(\underline{t}) - \underline{g}(\underline{t},\theta_0))Q(\underline{g}_n(\underline{t}) - \underline{g}(\underline{t},\theta_0))^T$$

where Q is a nonnegative definite constant matrix which may be selected optimally.

With the intention of obtaining rapid estimates of θ, Kemp & Kemp (1988) also propose approaches based on the pgf. Rapid, in the sense that iterations are not needed to obtain an estimate, which may then be used, say, as an initial value for iterative methods. The simplest form for an estimator of a k-dimensional parameter $\underline{\theta}$ based on the pgf consists in fixing $t_1, \ldots, t_k \in [-1,1]$ and establishing the simultaneous equations

$$\phi_n(t_i) = \phi(t_i,\underline{\theta}),\ i = 1, \ldots, k$$

which are then solved to obtain the vector $\hat{\underline{\theta}}_n$. Kemp & Kemp focus on k = 2, and show that certain limiting choices of t_1 and t_2 lead to other rapid estimation methods such as the method of moments, the mean-and-zero-frequency method, or the method of even points. More generally, other possible estimating equations based on the pgf may be obtained by setting

$$\phi_n^{(m)}(t_i) = \phi^{(m)}(t_i,\underline{\theta}),\ i = 1, \ldots, k.$$

With regard to the selection of m and the t's, it appears that choices are possible which lead to estimates with high relative efficiency (with respect to maximum likelihood), but these are not necessarily good for all distributions; in fact, generally speaking, even if the family of distributions is fixed, high relative efficiency is not obtained uniformly over all regions of Θ for any one choice of m and $t_1, \dots, t_k$. This makes it necessary to study particular families of distributions one at a time and to check the performance of methods by simulation. Other types of estimating equations also considered in Kemp & Kemp (1988) are ones obtained by considering

$$(d/dt)\log\{\phi(t,\theta)\} = (d/dt)\log\{\phi_n(t)\}.$$

An immediate advantage of any of these methods is that they may be applied to problems where maximum likelihood is not operative, say for instance, if a closed form of the density is not available, or if solving likelihood equations is difficult. The first example in Section 5 is an instance of this latter case. Advantages are also apparent in the analysis of multivariate count data.

4.2 *Goodness of fit for discrete distributions*

The issue of testing the fit of a discrete distribution seems to be relatively underdeveloped when compared with the amount of literature devoted to the continuous case (see D'Agostino & Stephens, 1986). A general perspective of this problem for discrete distributions is offered by the epgf. It was first used in this context by Kocherlakota & Kocherlakota (1986) and recently explored for the Poisson case by Rueda et. al (1991), Nakamura & Pérez-Abreu (1991) and Baringhaus & Henze (1992).

A quick method to test the composite hypothesis H_0 that $X_1,\dots,X_n$ originates from a general (possibly multivariate or multiparametric) discrete distribution $F(x,\underline{\theta})$ with pgf $\phi(t) = \phi(t,\underline{\theta})$ ($\underline{\theta}$ unknown) is based on the following concept. As in the rapid estimation approach, consider a fixed number of t's, say $\underline{t} = (t_1,\dots,t_k)$. Then, under the null hypothesis, if $\hat{\underline{\theta}}$ denotes the maximum likelihood estimator of $\underline{\theta}$, the random vectors $\underline{Z}_n(\underline{t}) = n^{1/2}\,(\phi_n(t_1) - \phi(t_1,\hat{\underline{\theta}}),\dots,\phi_n(t_k) - \phi(t_k,\hat{\underline{\theta}}))$ converge to a k-valued normal distribution with mean zero and covariance matrix $Q(\underline{\theta}) = [q(t_i,t_j)]$, where

$$q(t_i,t_j) = \phi(t_i t_j,\underline{\theta}) - \phi(t_i,\underline{\theta})\phi(t_j,\underline{\theta}) - \sum_{l=1}^{k}\ \sum_{m=1}^{k}\sigma_{lm}\,\frac{\partial}{\partial\theta_l}\phi(t_i,\underline{\theta})\,\frac{\partial}{\partial\theta_m}\,\phi(t_j,\underline{\theta})$$

and $\Sigma = \{\sigma_{ij}\}$ is the inverse of the Fisher information. When $\underline{\theta}$ is known this expression simplifies to $q(t_i,t_j) = \phi(t_i t_j,\underline{\theta}) - \phi(t_i,\underline{\theta})\phi(t_j,\underline{\theta})$.

Thus, to test the hypothesis H_0, Kocherlakota & Kocherlakota (1986) suggest a test based on the statistic $\underline{Z}_n(\underline{t})Q^{-1}(\hat{\underline{\theta}})\underline{Z}_n^T(\underline{t})$, and reject if $\underline{Z}_n(\underline{t})Q^{-1}(\hat{\underline{\theta}})\underline{Z}_n^T(\underline{t}) > \chi^2_{k;1-\alpha}$, the 100(1-α) percent point in the χ^2 distribution with k degrees of freedom. The special case of $\underline{\theta}$ known follows in a straightforward manner. From a simulation study Kocherlakota & Kocherlakota (1986) conclude that it is convenient to use a small number for k as well as values of t close to zero. Nakamura & Pérez-Abreu (1991) find that in the Poisson case, although this procedure is not consistent, it should not be disregarded, since it behaves well against distributions which do not display upper heavy tails. The test might be convenient as a preliminary test in multivariate and multi-parametric situations. Inspired on an idea developed in Epps et. al (1982), Kocherlakota & Kocherlakota (1990) have used this technique in the case of the weighted binomial distribution.

To avoid the dependence on the number of selected t's as well as their specific values, Rueda et. al (1991) propose a continuous extension of the previous method, which resembles the Cramér von-Mises statistic. The proposed test statistic is

$$d_n(\hat{\underline{\theta}}) = n \int_0^1 (\phi_n(t) - \phi(t,\hat{\underline{\theta}}))^2 \, dt, \tag{4.1}$$

which, under the null hypothesis, asymptotically possesses the distribution of $d = \int_0^1 (Z(t))^2 \, dt$, where $Z(t)$ is a Gaussian process with covariance given by (2.3) (or by (2.2) if $\underline{\theta}$ is known). The distribution of d is then tabulated as in Durbin (1973) using numerical methods to solve an eigenvalue problem and to invert a characteristic function. There is the disadvantage that this distribution depends on the parameter $\underline{\theta}$. However, this is a general procedure that still has to be explored for many discrete distributions. The particular goodness of fit test of the Poisson (λ) distribution was considered in Rueda et. al (1991), in which case, (4.1) takes on the value

$$(1/n) \sum_{i,j=1}^{n} 1 / (X_i + X_j + 1) - 2 \sum_{i=1}^{n} [(-1)^{X_i+1} \frac{X_i!}{\lambda^{X_i+1}} e^{-\lambda}$$
$$+ \sum_{j=1}^{n} (-1)^j X_i! / \{(X_j - j)! \lambda^{j+1}\}] + (n / 2\lambda)(1 - e^{-2\lambda}).$$

Using characteristic properties of the Poisson pgf, recently two "Poissonness tests" have been considered. Here, H_0 is the composite hypothesis that the model is a Poisson distribution. On one side, using the distinguishing fact that for a Poisson distribution $\lambda\phi(t) = \phi^{(1)}(t)$, Baringhaus and Henze (1992) proposed the test statistic

$$D_n = \int_0^1 (\bar{X} \phi_n(t) - \phi_n^{(1)}(t))^2 \, dt =$$
$$(1/n) \sum_{i,j=1}^{n} [\bar{X}^2/(X_i+X_j+1) - X_i X_j/(X_i+X_j-1)] - (n - C_n^2/n)\bar{X},$$

where $C_n = \sum_{i=1}^{n} I(X_i=0)$. The limiting distribution of D_n is derived by the last named authors and it turns out to be a weighted infinite sum of independent χ_1^2 random variables, which depends on the parameter λ. Baringhaus and Henze (1992) prove that this is a consistent test for alternative distributions with finite first moment.

On the other hand, Nakamura & Pérez-Abreu (1991) consider a test motivated by the graphical method of Section 3.1. In assessing the linearity of $Y_n(t) = \log(\phi_n(t))$ as a function of t, they obtain the statistic $T_n^* = nT_n/\bar{X}^{1.45}$, where

$$T_n = (1/n^4) \sum_{i,j,k,l=1}^{n} X_i(X_i-X_j-1)X_k(X_k- X_l- 1)I_{\{X_i+X_j = X_k+X_l\}}.$$

The test rejects for $T_n^* > q_{1-\alpha}$, where the percentile values $q_{1-\alpha}$ of the limiting distribution of T_n^* were computed in Nakamura & Pérez-Abreu (1991). This distribution is also a weighted infinite sum of independent χ_1^2 random variables and has the particular feature that it seems to be independent of λ. In the latter work a simulation study was conducted to compare the behavior of this test with other well-known Poissonness tests. The conclusion is that, although not the most powerful for any particular alternative, the test $\tilde{T}_n^*$ maintains relatively high power against a wider range of alternatives than any of the other tests considered.

All of the procedures explained above show that the empirical probability generating function is potentially useful in constructing goodness-of-fit tests for discrete distributions. We expect that more work on the subject will be developed for distributions for counts other than the Poisson distribution.

4.3 *Change Point Estimation.*

Let $X_1,\ldots,X_n$ be independent random variables, such that for an unknown *change point parameter* θ, $X_1,\ldots,X_{[n\theta]}$ have discrete distribution F and $X_{[n\theta]+1},\ldots,X_n$ a discrete distribution G, where $F \neq G$ are unknown. Following an idea presented in Carlstein (1988), an easy nonparametric procedure to estimate θ can be constructed using the epgf. The method is as follows: For each $\lambda \in T_n = \{1/n,2/n,\dots,(n-1)/n\}$, define the pre-$\lambda$ and post-λ empirical probability generating functions as

$$^{\lambda}\phi_n(t) = (1/[n\lambda]) \sum_{i=1}^{[n\lambda]} t^{X_i},$$

$$\phi_n^{\lambda}(t) = (1/[n(1-\lambda)]) \sum_{i=[n\lambda]+1}^{n} t^{X_i}, \quad t \in [0,1] .$$

Consider the distance

$$d_n(\lambda) = \lambda(1-\lambda) \int_0^1 (^{\lambda}\phi_n(t) - \phi_n^{\lambda}(t))^2 t^2 \, dt =$$

$$\lambda(1-\lambda) \left\{ 1/(n\lambda)^2 \sum_{i=1}^{[n\lambda]} \sum_{j=1}^{[n\lambda]} 1/(X_i+X_j+3) - 2/(n^2\lambda(1-\lambda)) \sum_{i=1}^{[n\lambda]} \sum_{j=[n\lambda]+1}^{n} 1/(X_i+X_j+3) \right.$$
$$\left. + 1/(n^2(1-\lambda)^2) \sum_{i=[n\lambda]+1}^{n} \sum_{j=[n\lambda]+1}^{n} 1/(X_i+X_j+3) \right\} .$$

Then, the change point estimator $\hat{\lambda} \in T_n$ of θ is the one such that $d_n(\hat{\lambda}) = \max \{d_n(\lambda); \lambda \in T_n\}$. Apart from the fact that this is a very intuitive and easy to compute estimator, it is consistent and a Monte Carlo experiment (Pérez-Abreu, 1989) has shown that this estimator is as good as the best estimator presented in Carlstein (1988).

5.- Two examples.

We give a brief account of two situations in which a pgf emerges inherently in a statistical problem. A few ideas discussed in Section 4 may therefore be applied.

5.1 A cumulative damage model.

Bogdanoff & Kozin (1985) describe stochastic models for cumulative damage based on certain Markov Chains. A simple version of their model involves a finite-state chain having transition probability matrix

$$\begin{bmatrix} p_1 & q_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & p_2 & q_2 & 0 & \dots & 0 & 0 \\ \dots & & & & & & \\ 0 & 0 & 0 & 0 & \dots & p_b & q_b \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

where $1 > p_1 > p_2 > \ldots > p_b > 0$. In their description, pgf's are used as standard tools for finding and/or identifying discrete distributions, in particular, for the distributions of waiting times between states. For estimation of parameters in the model, they recommend either the method of moments, or maximum likelihood, which is more difficult to implement. Having explicit expressions for the pgf's, it seems natural to consider one of the estimation methods based on the epgf described in Section 4.1. If appropriately selected, the epgf-based method may provide higher efficiency than the method of moments, it would be more easily computed than maximum likelihood, and it may still have high relative efficiency with respect to maximum likelihood (Pérez, 1991 and Kemp & Kemp, 1988). Since engineers involved with cumulative damage are really interested in hazard functions or cumulative distribution functions for times to failure, a matter of interest would be to investigate the differences in the methods when estimating these functions.

5.2 Nonparametric estimation of the distribution of a maximum.

The probability generating function also arises naturally in Hydrology, in the estimation of the distribution function of the annual maximum level of a river, or in the more general problem of nomination sampling (Boyles & Samaniego, 1986). Let (X_1,N_1), (X_2,N_2), ..., (X_n,N_n) be a random sample from the bivariate vector (X,N) where X has distribution F and N has a distribution for counts with probability generating function $\phi(t) = \sum_{k=0}^{\infty} t^k P_k$. The observations X_i's (annual maxima) are independent random variables which are maxima of N_i (levels above a certain amount) independent random variables. Then, the distribution function of each X_i is given by

$$F_M(x) = \sum_{k=0}^{\infty} F^k(x)P_k = \phi(F(x)).$$

Boyles & Samaniego (1986) introduce a nonparametric estimator, say $\hat{F}_n(x)$, of F(x). Then, using the empirical probability generating function ϕ_n one obtains the following nonparametric estimation of the distribution of the maximum $F_M(x)$ based on the nomination sampling $(X_1,N_1),(X_2,N_2),\ldots,$ (X_n,N_n):

$$\hat{F}_M(x) = \hat{\phi}_n(\hat{F}_n(x)).$$

References

Baringhaus L. & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Ann. Inst. Statist. Math.* 43, 551-564.

Baringhaus L. & Henze, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statist. Probab. Letters* 13, 269-274.

Barnett, V. & Lewis, T. (1983). *Outliers is Statistical Data.* 2nd Edition, Wiley, New York.

Bogdanoff, J. L. & Kozin, F. (1985) *Probabilistic Models of Cumulative Damage*, Wiley, New York.

Bortkiewicz, L. V. (1898). *Das Gesetz der Kleinen Zahlen,* Teubner, Leipzig.

Boyles, R. A. & Samaniego, F. J. (1986). Estimating a distribution function based on nomination sampling. *J. Amer. Statist. Assoc.* 81, 1039-1045.

Carlstein, E. (1988). Nonparametric change-point estimation. *Ann. Statist.* 16, 188-197.

Csörgó, S. & Mason, D. (1989). Bootstrapping empirical functions. *Ann. Statist.* 17, 1447-1471.

Csörgó, S. & Teugels, J. L. (1990). Empirical Laplace transform and approximation of compound distributions. *J. Appl. Probab.* 27, 88-101.

D'Agostino, R. B. & Stephens, M. E. (1986). *Goodness of Fit Techniques,* Marcel Dekker, New York.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample*

Distribution Function. Regional Conf. Series in Appl. Math. 9, SIAM.

Epps, T. W., Singleton, K. J. & Pulley, L. B. (1982). A test of separated families of distributions based on the empirical moment generating function. *Biometrika* 69, 391-399.

Feuerverger A. (1988). On the empirical saddlepoint approximation. *Biometrika* 76, 457-464.

Feuerverger A. & McDunnough P. (1984). On statistical transform methods and their efficiency. *Canadian J. Statist.* 12, 303-317.

Hoaglin, D. C. & Tukey, J. W. (1985) Checking the Shape of Discrete Distributions. In *Exploratory Data Tables, Trends, and Shapes.* Eds. D. C. Hoaglin, F. Mosteller & J. W. Tukey, Wiley, New York.

Jara, J. M. & Rosenblueth, E. (1988). Probability distributions of times between Characteristic Subduction Earthquakes. *Earthquakes Spectra* 4, 499-529.

Johnson, N. L. & Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions,* Houghton Mifflin, Boston.

Kemp, C. D. & Kemp, A. W. (1988). Rapid estimation for discrete distributions. *The Statistician* 37, 243-255.

Kocherlakota, S. & Kocherlakota, K. (1986). Goodness of fit tests for discrete distributions. *Commun. Statist. Theory Meth.* A 15, 815-829.

Kocherlakota, S. & Kocherlakota, K. (1990). Tests of hypothesis for the weighted binomial distribution. *Biometrics* 46, 645-656.

Marques, M. & Pérez-Abreu, V. (1989). Law of large numbers and central limit theorem of the empirical probability generating function of stationary random sequences and processes. *Aportaciones Matemáticas, Soc. Mat. Mex.* 4, 100-109.

Nakamura, M. and. Pérez-Abreu V. (1991). An empirical probability generating function approach for testing a Poisson model. Unpublished manuscript.

Pérez, F. J. (1991). Modelos de Daño Acumulado basados en Cadenas de Markov. Unpublished bachelor's degree paper.

Pérez-Abreu, V. (1989). Estimación noparamétrica del tiempo de cambio para distribuciones discretas. *Aportaciones Matemáticas*, Soc. Mat. Mex. 6, 289-301.

Prakasa Rao, B.L.S. (1987). *Asymptotic Theory of Statistical Inference.* Wiley, New York.

Read, R. R. (1981). Representation of certain covariances matrices with applications to asymptotic efficiency. *J. Amer. Statist. Assoc.* 76, 148-154.

Rueda, R., Pérez-Abreu, V. & O'Reilly, F. (1991). Goodness of fit for the Poisson distribution based on the probability generating function. *Commun. Statist. Theory Meth.* A 20, 3093-3110.

Stewart & Campbell (1970). The statistical analysis between past and future accidents and violations.

Fig.2.1: Log of Probability Generating Function

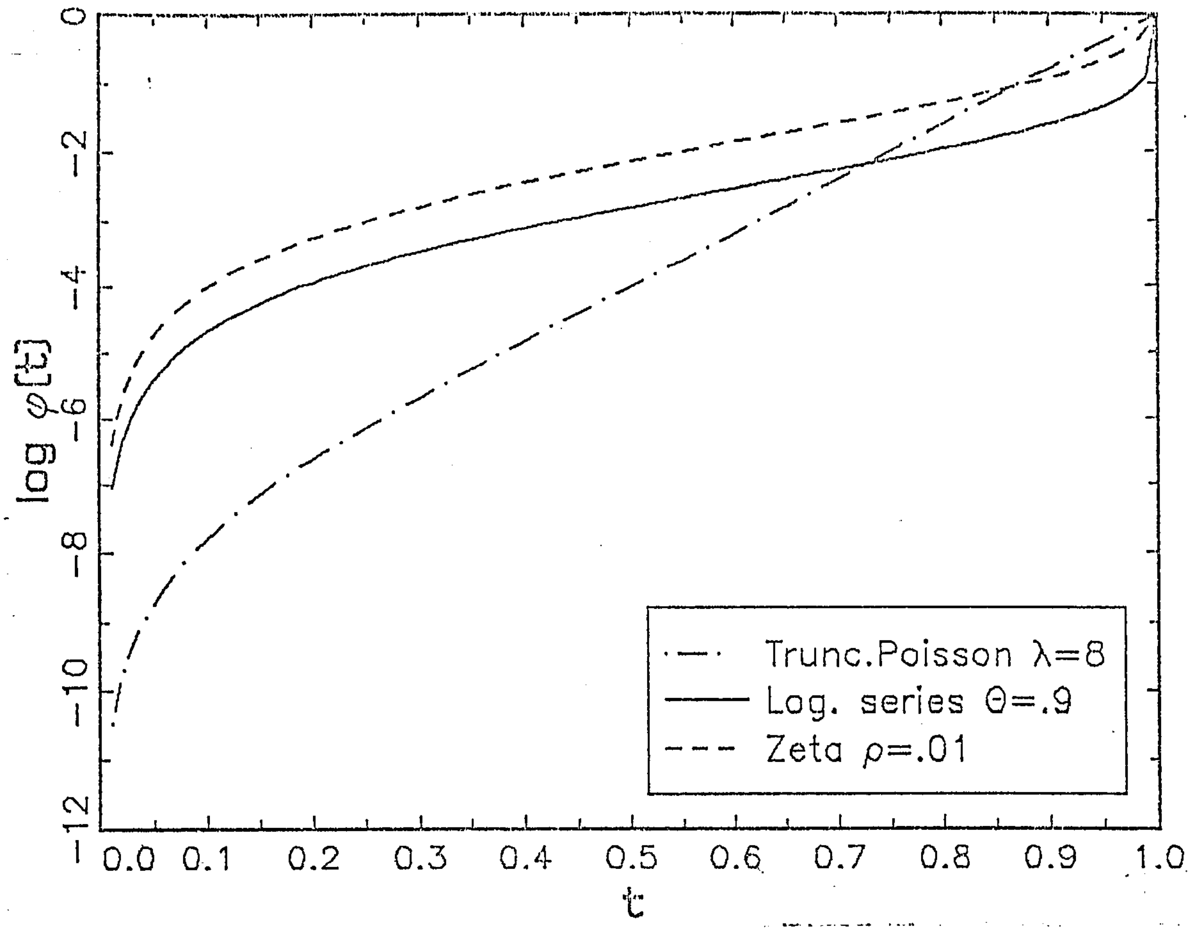


Fig. 2.2a: Horsekick Data

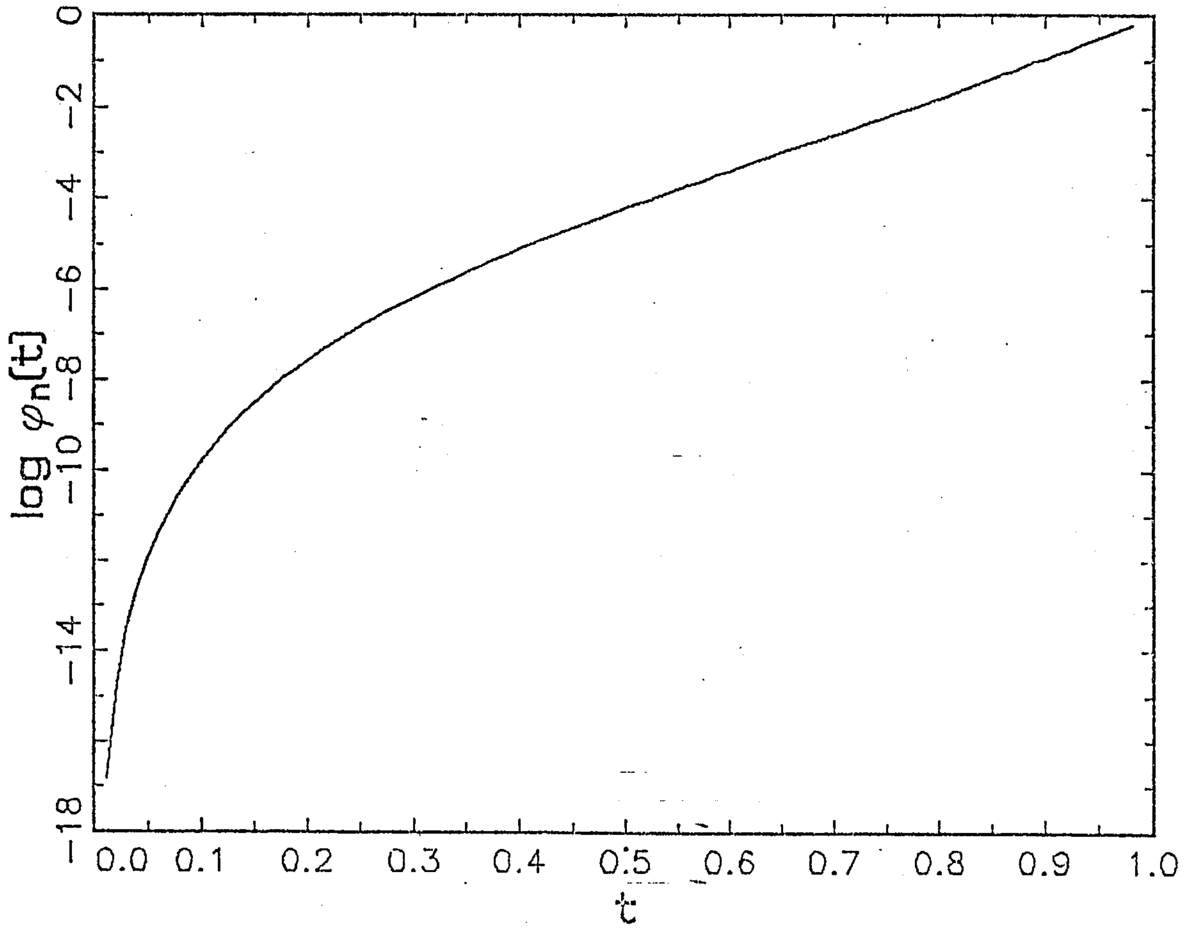


Fig. 2.2b: Earthquake Data

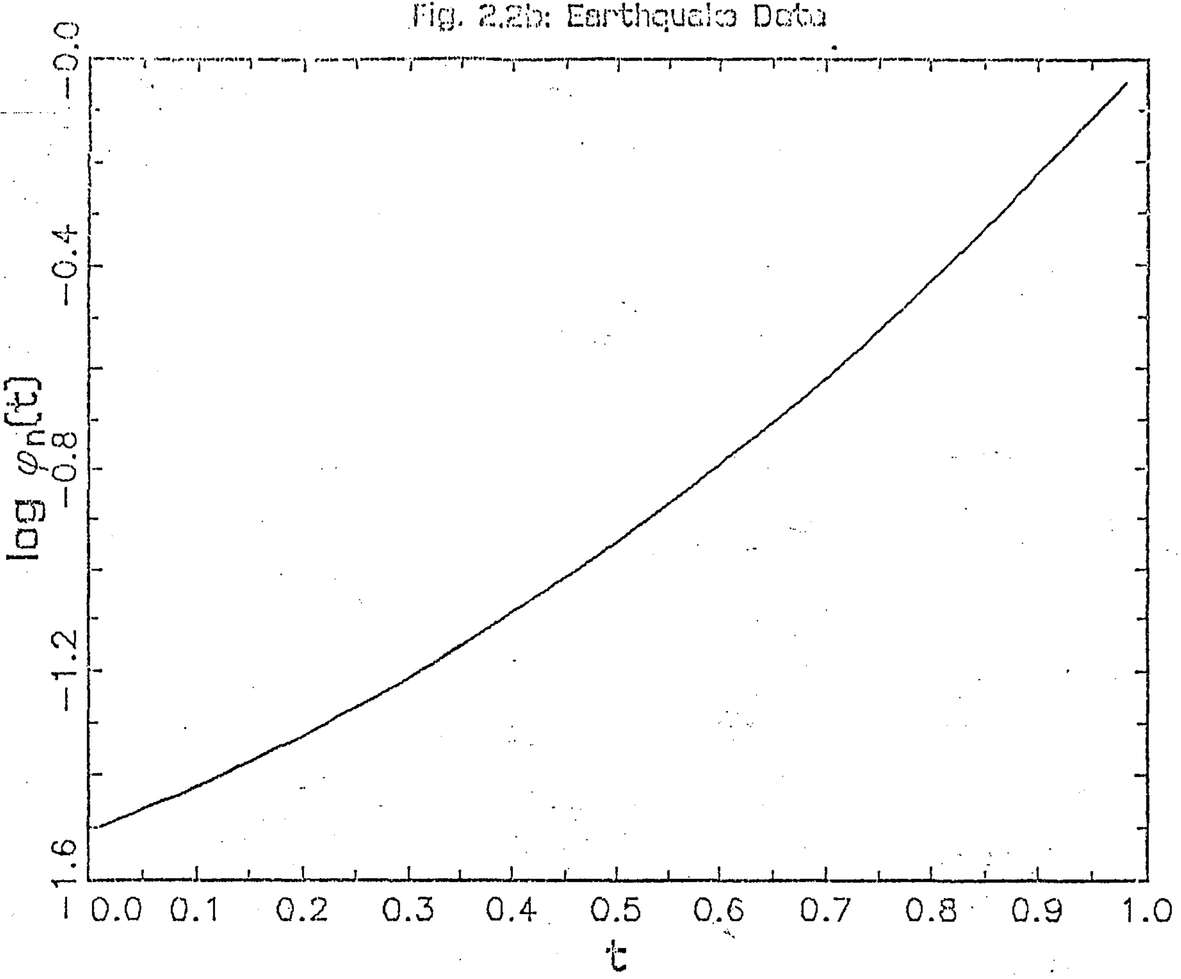


Fig.2.3: Quality Inspection Data

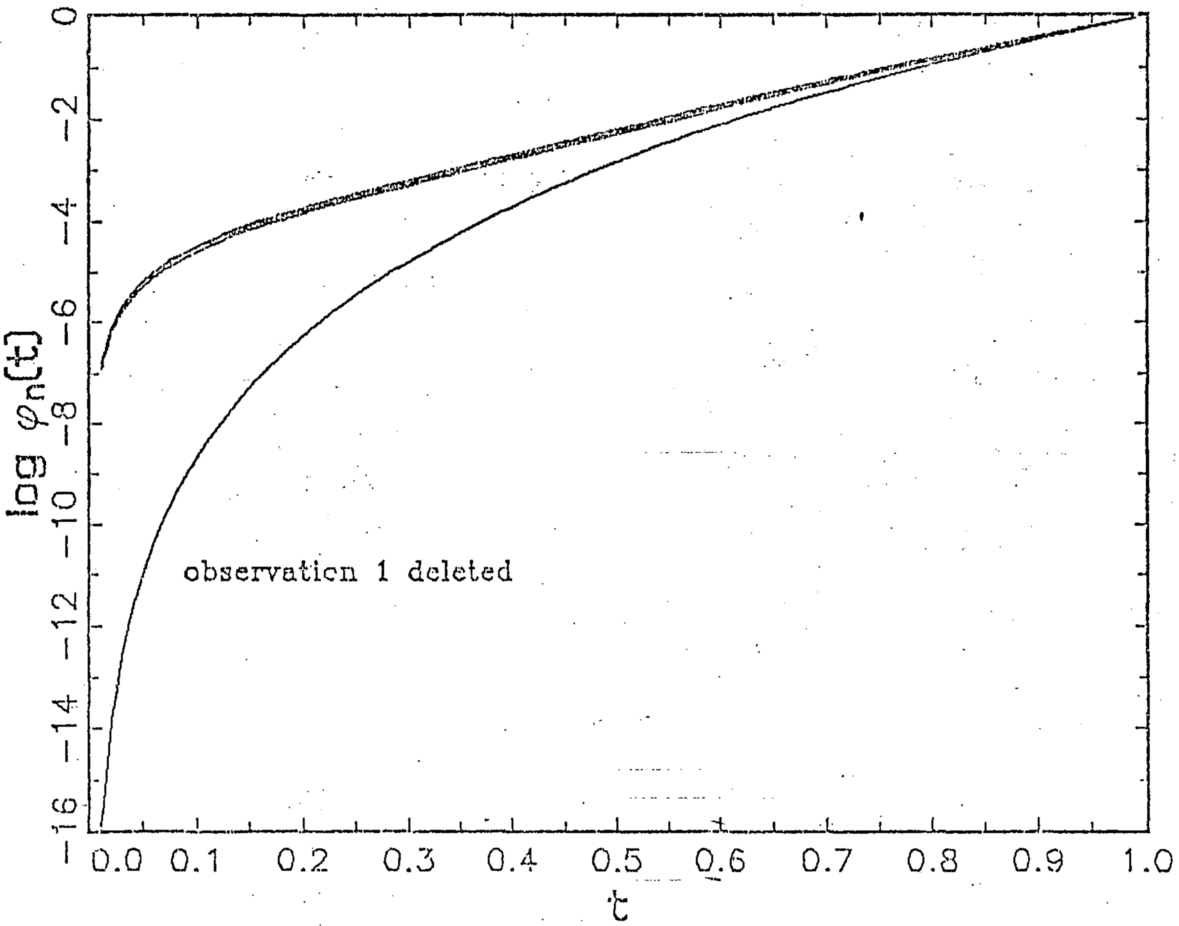


Fig.2.4: Data on Household allowance demand

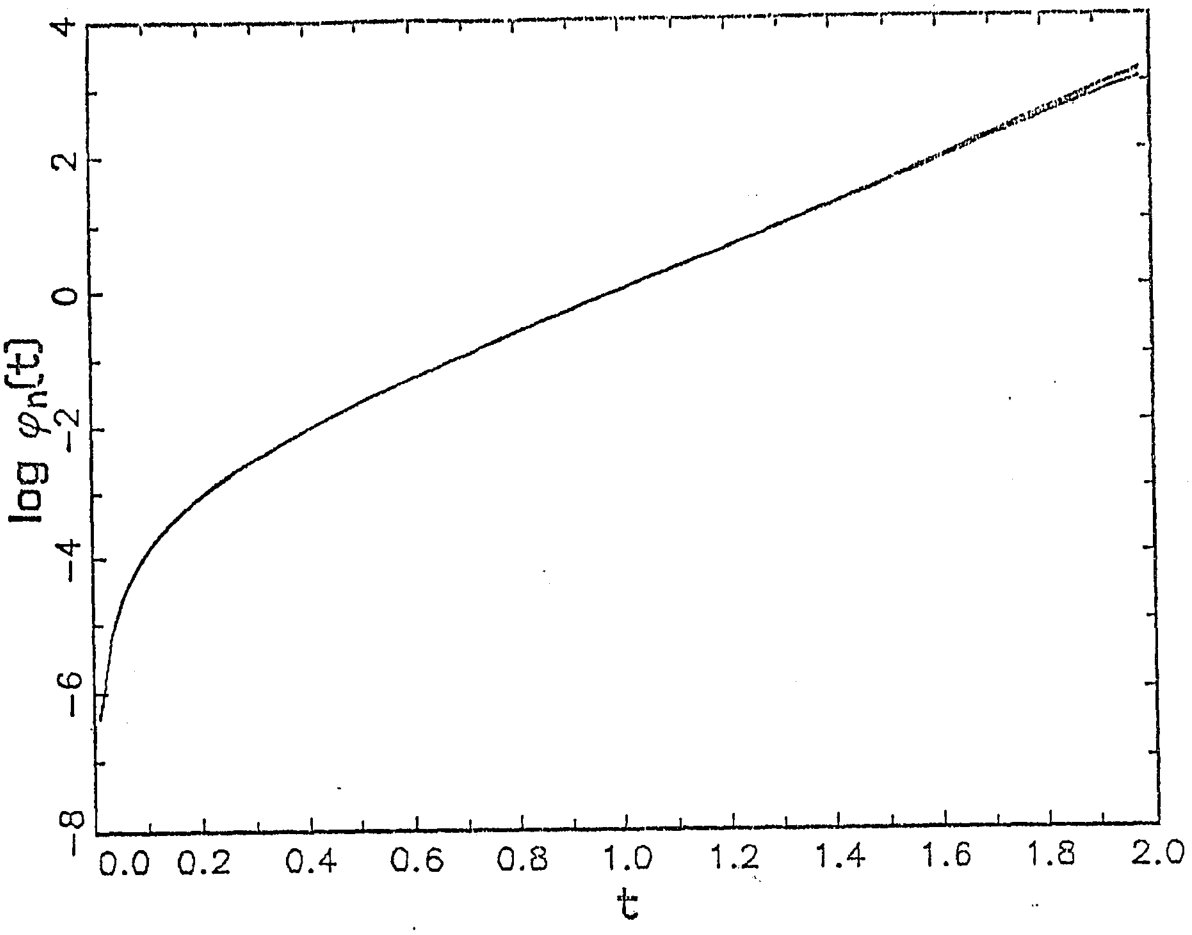
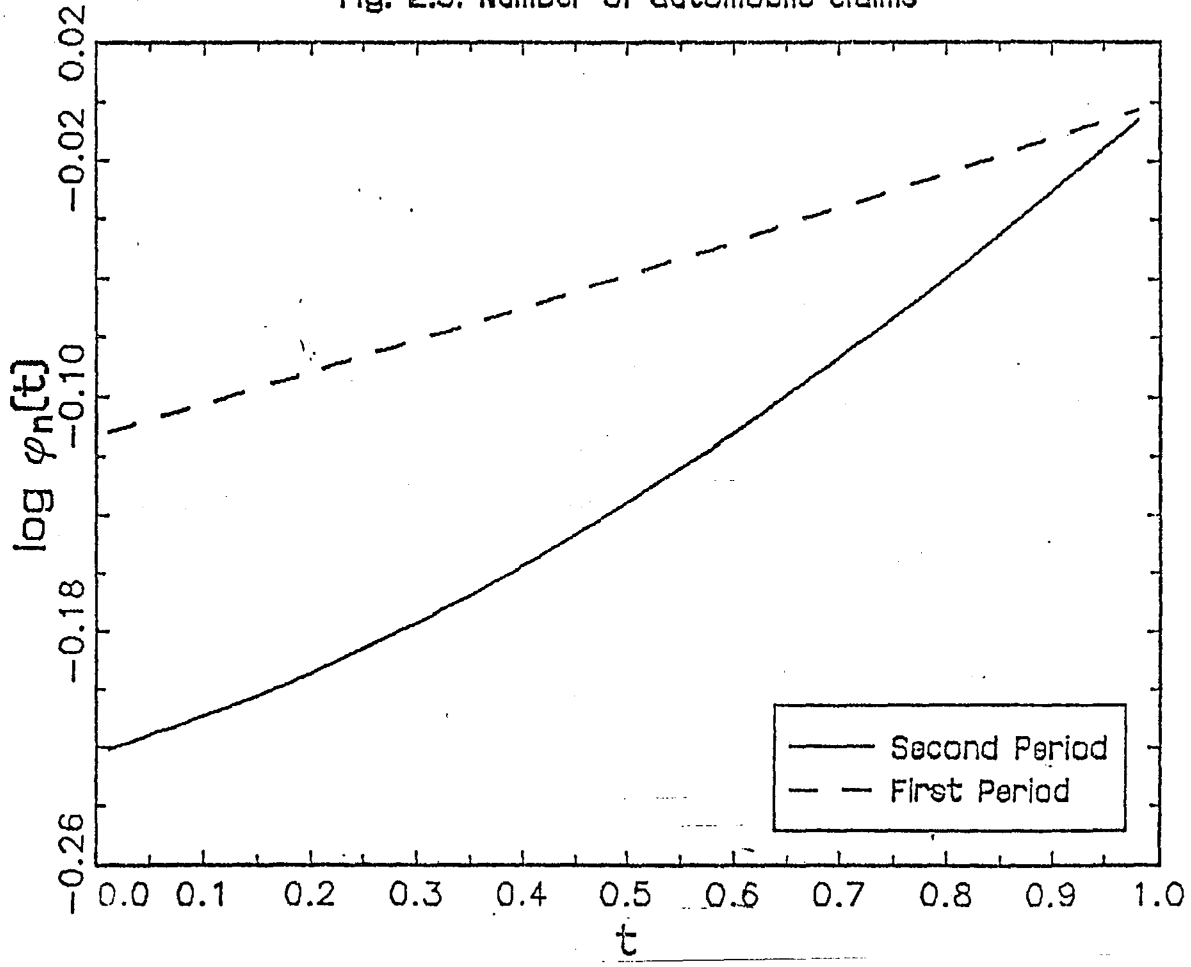


Fig. 2.5: Number of automobile claims


log φ_n(t)
t
Second Period
First Period