North Holland Publishing Company (1977)

THE WEIGHTED DISTRIBUTIONS: A SURVEY OF THEIR APPLICATIONS

C. P. PATIL AND C. RADHAKRISHNA RAO

The Pennsylvania State University
University Park, Pennsylvania

The Indian Statistical Institute
New Delhi, India

1. INTRODUCTION

The concept of weighted distributions has been introduced and formalized recently, see for example, Rao (1965). Although the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized. This paper is intended to provide the needed focus to the notion of weighted distributions and the inference problems associated with them. To this end, some interesting and varied applications from a number of disciplines have been indentified. While they may look quite unrelated at the first glance, they are all indeed natural examples of the single underlying theme of weighted distributions.

To begin with, the paper defines and illustrates the concept of weighted distributions. Situations leading to weighted discrete distributions are discussed first. These relate to the analysis of family data, the problem of family size and alcoholism, the aerial survey involving visibility bias in wildlife ecology, and a discrete renewal system. The problems generating weighted continuous distributions refer to computer generation of random variables using rejection technique, renewal theory, continuous waiting time paradox, cell kinetics and early disease detection, forest products research, particle size distributions by thin section methods and low density traffic streams. Lastly, the weight size distributions used in small particle physics and sedimentology are recognized as the weighted distributions. So also are recognized the moment distributions in economics. It should be worthwhile to comment here that while the earlier examples involve weighted distributions essentially because of built in probability sampling at some stage in the problem, the last two examples can not be explained on the probability sampling basis. These involve a distinctly different recording mechanism responsible for generating a class of weighted distributions.

2. NOTATION AND TERMINOLOGY

Consider a natural mechanism generating a random variable X with p.f.

$f(x; \theta)$ where $\theta \in \Omega$, the parameter space. For drawing a random sample of observations on X, we have to use a method of selection which gives the same chance of including in the sample any observation produced by the original mechanism. But in practice it may so happen that the relative chances of inclusion of two observations x and y are $w(x)$: $w(y)$ where $w(\cdot)$ is non-negative valued function. Then the recorded X to be denoted by X_w has the *pdf*

$$f^w(x; \theta) = \frac{w(x)f(x; \theta)}{\omega} \qquad (2.1)$$

where $\omega = E[w(X)] = \int w(x)f(x; \theta)dx$ or $= \sum w(x)f(x; \theta)$ depending on whether X is continuous or discrete. Further, if $0 \leq w(x) \leq 1$, ω is the probability of including an observed value in the sample.

The distribution defined by (2.1) is called a weighted distribution with weight function w(x) which can be arbitrary. When an investigator collects a sample of observations produced by nature, according to a certain model, the original distribution may not thus be reproduced. The main interest in any investigation is, however, to determine the characteristics of the original distribution. Further, it also becomes important to assess the nature and amount of distortion caused in the determination of these characteristics in case the change in the underlying distribution due to sampling bias is ignored.

The following examples may help illustrate a few general situations involving "non-response" responsible for generating weighted distributions.

(i) *Truncation:* The distribution of a random variable truncated to a set T is a weighted distribution with weight function $w(x) = 1$ for $x \in T$ and zero elsewhere.

(ii) *Missing Data:* If a proportion $1-w(x)$, $0 \leq w(x) \leq 1$, out of the natural frequency of x is missing, that is if the probability of x missing is $1-w(x)$, the pdf to use for the analysis of the observed data is with the weight function $w(x)$.

(iii) *Damaged Observations:* Consider a damage model where an observation $X = x$ is reduced to y by a destructive process with *pdf* $d(y|x)$ (See Rao (1965). Then the probability that the observation $X = x$ is undamaged is $d(x|x)$, and the distribution of the undamaged observation is the weighted distribution with $w(x) = d(x|x)$. For example, under binomial survival model, $d(x|x) = \theta^x$, $0 < \theta < 1$. An investigator recording only undamaged observations will need to work with a corresponding weighted distribution.

The following table gives various weight functions used in literature that we have come across.

We note that the weight functions in the table are all monotone functions, either increasing or decreasing.

The assumptions of the implications of the relationships between the original distribution of x and the weighted distribution obtained using some weight function w(x) can generate interesting and useful characterization results. See, for example, Rao, Rao-Rubin (1964), Patil-Ord (1975), Gupta (1975), and sections 7 and 8 of this paper.

A TABLE OF WEIGHT FUNCTIONS USED

No	$x > 0$	$w(x)$	references
1	General	x	most of the references in this paper
2	discrete	x^a, $a > 0$	Rao (1965), Patil-Ord (1975)
3	continuous	x^a, $a > 0$	Brown (1972), Patil-Ord (1975)
4	discrete	$1-(1-\beta)^x$, $0 < \beta < 1$	Haldane (1938), Rao (1965) Neel-Schull (1966), Cook-Martin (1974)
5	discrete	$x + 1$	This paper
6	discrete	$x(x-1)\ldots(x-r+1)$	Patil-Ord (1975), Gupta (1975)
7	discrete	ϕ^x, $0 < \phi < 1$	Rao-Ruhin (1964), Rao (1965), Kemp (1973)
8	continuous	$e^{\omega x}$	Patil-Ord (1975)

PART I: DISCRETE MODELS AND SIZE-BIASED PROBABILITY SAMPLING

3. THE ANALYSIS OF FAMILY DATA

Various demographic and social studies involve family size and sex ratio as important factors which have some bearing on the main study. The following examples show as to how the weighted distributions arise here as a result of the size-biased probability sampling. This discussion is based on Rao (1965) and Neel and Schull (1966).

3.1. The Ratio and Weighted Binomials: The following data relate to brothers and sisters in families of 104 boys who were admitted to a postgraduate course at the Indian Statistical Institute.

Let us assume that in families of given size n, the probability of a family with x boys coming into our record is proportional to x. Also, suppose that the number of boys follow a binomial distribution with probability parameter π. Then

$$f(x;\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x},$$

$$w(x) = x,\ E[w(X)] = \omega = n\pi,$$

$$f^w(x;\pi) = \binom{n-1}{x-1} \pi^{x-1} (1-\pi)^{n-x},$$

$$E[X_w - 1] = (n-1)\pi,$$

$$E[X_w/n] = \pi + \frac{1-\pi}{n} > \pi,$$

$$E[(X_w-1)/(n-1)] = \pi.$$

* If k boys representing families of size $n_1, n_2, \ldots, n_k$ report $x_1, x_2, \ldots, x_k$ boys, $E[(\Sigma X_i - k)/(\Sigma n_i - k)] = \pi$, and an unbiased estimate of π is $\hat{\pi} = (\Sigma x_i - k))/(\Sigma n_i - k) = (414 - 104)/(726 - 104) = 310/622 \simeq \frac{1}{2}$. Whereas if one wrongly treats x_i's as observations on k randomly drawn families with at least one boy, i.e., as arising from a truncated binomial, then the estimate of π will have a serious upward bias.

TABLE III.I

based on Rao (1965), p. 325

Family Size	1	2	3	4	5	6	7	8	9	10	11	12	13	15	Total
No. of Families	1	6	6	13	12	7	14	11	12	8	6	5	2	1	104
Brothers:	1	8	12	34	34	29	59	50	54	46	32	31	16	8	414
Sisters:	0	4	6	18	26	13	39	38	54	34	34	29	10	7	312

3.2. Geometric Distribution as Weighted Log Series for Family Size:

A geometric distribution is sometimes found to provide a good fit to an observed distribution of family size, specially when data are obtained from children attending a school. But this may be the effect of sampling with families of large size having a higher chance of being recorded with w(x) proportional to x, and the actual distribution may well be logarithmic series (Rao, 1965). What we have is:

$$f(x;\theta) = \frac{1}{x}\ \theta^x/[-\log(1-\theta)] = a\theta^x/x \quad x = 1,2,\ldots$$

$$w(x) = x,\ E[w(X)] = \omega = a\theta/(1-\theta) \qquad (3.2)$$

$$f^w(x;\theta) = (1-\theta)\theta^{x-1} \qquad x = 1,2,\ldots$$

3.3. A Study of Albinism and Weighted Binomials: Fisher (1934), Haldane (1938), Rao (1965).

If we wish to study the distribution of albino children in families capable of producing such children, we may contact a large number of families and ascertain from each family the number of albinos. The families with at least one albino child provide a truncated binomial distribution and the probability of a child being an albino can be estimated from such a distribution. But this method of investigation is wasteful as the proportion of abnormal families is investigated. A convenient method in such a case is first to discover an albino child and and through it obtain the information about the family to which it belongs. But such a procedure may not give equal chance to all families in which albinos have occurred. The exact chance for a family with x albinos is that of detecting at least one of its albino children, which may be a function of x. The weight functions of the following form have been studied.

$$w_1(x) = 1 - (1-\beta)^x,\ 0 \le \beta \le 1$$

$$w_2(x) = x^\alpha \qquad (3.3)$$

Rao (1965) observes that $\alpha = 1/2$ seems to provide a good fit.

3.4. Analysis of Family Data in Human Heredity: Neel and Schull (1966), p. 211-229.

A primary object in collecting family data is to compare the proportion of affected children actually observed with some theoretical proportion based upon the type of mating and the suspected mode of inheritance. To this end, we may select families at random without reference to the type of offspring produced and study only the families which contain at least one affected child. If we are interested in a rare inherited trait, then a random selection of all families will necessarily lead to a preponderance of families in the data which could yield no information, since the majority of families would not possess the rare gene. Ascertainment through affected individuals may then be the only reasonable way of going about the data collection.

Neel and Schull consider binomial distributions with weight function

$$w(x) = 1 - (1-\beta)^x,\ 0 \le \beta \le 1 \qquad (3.4)$$

and estimate the sighting probability β to be 80/100, and use this estimate further in estimating the desired common probability parameter of the binomial distributions.

4. ALCOHOLISM AND FAMILY SIZE

In a social psychological study relating to alcoholism Smart (1964) tested various hypotheses on the basis of data obtained from 242 alcoholics treated in three alcoholism clinics in Ontario, Canada. One of the variables recorded is the size of the family to which an alcoholic belongs. What is the appropriate theoretical distribution with which the observed distribution can be compared?

It is clear that the method of ascertainment is such that it gives greater chance to families of larger size being represented in our sample. Then the question arises as what is the appropriate weight function to be applied to the actual distribution of family size in the population to make it comparable to the observed distribution.

Let p_n be the proportion of families of size n in the population, and suppose that each individual has a probability k, independently of the others, of becoming an alcoholic. Then the distribution of family size among families having at least one alcoholic is

$$p_n^* = \frac{1-(1-k)^n}{\sum 1-(1-k)^n p_n} p_n \qquad (4.1)$$

which is the limit as $k \to 0$, is of the form

$$p_n^* = \frac{n p_n}{\sum n p_n} \qquad (4.2)$$

It may be noted that the probability of a family being of size n and having r alcoholics is

$$p_{nr} = \binom{n}{r} k^r (1-k)^{n-r} p_n. \qquad (4.3)$$

Suppose our sampling mechanism is such that a family with r alcoholics gets a chance proportional to r of being selected. Then the probability for a family of size n with r alcoholics coming into our sample is proportional to

$$p_{nr}^* = r\, p_{nr} \qquad (4.4)$$

and the corresponding probability for a family of size n is

$$p_n^* = \frac{\sum r p_{nr}}{\sum r p_{nr}} = \frac{n p_n}{\sum n p_n} \qquad (4.5)$$

which is the same as (4.2). Thus the method of selection of a family with probability proportional to the number of alcoholics does not alter the distribution of family size among families with at least one alcoholic in the population.

Sprott (1964), following the work of Smart (1963), used the 1931 Canadian Census data for finding p_n and fitted the weighted distribution (4.5) to the observed distribution of family size. The observed and expected values are given in the Table.

TABLE

Frequency Distribution of Family Size as Observed on the 242 Alcoholics and Expected on the Assumed Hypothesis

Family Size	1	2	3	4	5	6	7	8	9	TOTAL
Observed	21	32	40	47	29	23	20	11	19	242
Expected	34.2	51.4	47.3	37.6	26.4	17.9	12.1	7.4	8.2	242

The value of χ^2 goodness of fit is 39.1 which is high for 8 degrees of freedom, so that the weight function $w(n) = n$ in (4.5) does not explain the observed data. There is still excess of observed families with larger size over the expected values indicating the possibility that the weight function is of a higher order of magnitude than the family size, which may have some sociological significance.

5. AERIAL SURVEY AND VISIBILITY BIAS

Visibility bias is a recognized problem in aerial survey techniques for estimating wildlife population density. This source of error is generally conceded to be the main cause of inaccurate aerial census data and depending on various contributing environmental factors can produce severely biased population density estimates.

The visibility bias is present because of the failure to observe some animals. Cook and Martin (1974) have presented a model for quadrat sampling of randomly occurring groups whose size follows a single parameter power series distribution when there is a probability $\theta > 0$ of missing single animals.

The sampling model is based on three main assumptions:

(a) that animals occur within quadrats in groups of varying size,

(b) that each animal has a probability θ of being observed, and

(c) that, conditional on observing at least one member of a group, the entire group is observed with certainty.

For the purposes of the present discussion, the main point of interest being to

estimate the mean group size μ, let the random group size X have *rdf* $f(x; \theta)$. It may be easily verified that the recorded group size X_w has pdf

$$f^w(x; \theta) = \frac{w(x; \beta)f(x; \theta)}{\sum_x w(x; \beta)f(x; \theta)} \tag{5.1}$$

which, for small β, can be approximated by $f^*(x; \theta) = xf(x; \theta)/\mu$, where $\mu = \sum xf(x; \theta)$, the mean group size.

Based on the data consisting of g group sizes observed, $x_1, x_2, \ldots, x_g$, the sample mean $\bar{x}$ overestimates μ as its expectation is $\mu + \sigma^2/\mu$ where σ^2 is the variance of X. We consider the following two models:

(i) If we take X-1 ~ Poisson (θ), that is,

$$f(x; \theta) = e^{-\theta}\theta^{x-1}/(x-1)!, \tag{5.2}$$

then the estimating equation for the maximum likelihood estimate of μ based on the corresponding $f^*(x; \theta)$ is

$$\bar{x}-1 = \hat{\mu} - \frac{1}{\hat{\mu}}. \tag{5.3}$$

(ii) if we take X ~ Zero-truncated Poisson (θ), that is,

$$f(x; \theta) = e^{-\theta}\frac{\theta^x}{x!}/(1-e^{-\theta}), \tag{5.4}$$

then the likelihood equation based on the corresponding $f^*(x; \theta)$ is

$$(\bar{x}-1)/(1-e^{-(\bar{x}-1)}) = \hat{\mu}. \tag{5.5}$$

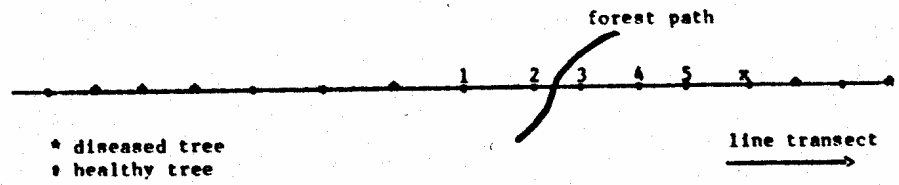
Cook and Martin (1974) report the following aerial moose census data collected over northeast of Minnesota in 1970. The number of groups observed is g = 113.

Group size x:	1	2	3	4	5	6
Observed pf:	.398	.407	.133	.044	.009	.009

We obtain $\bar{x} = 1.84$. Based on the model for X-1 ~ Poisson (θ), the maximum likelihood estimate of μ is 1.505, whereas for X ~ zero-truncated Poisson (θ), the maximum likelihood estimate of μ is 1.478.

6. FOREST DISEASE AND LINE TRANSECT SAMPLING

Consider a forest stand of trees. Suppose that an investigator is interested in estimating the proportion of diseased trees in the stand using a line transect for sampling. For logistic reasons, the investigator uses forest paths to get to the line transect and records the run length x of healthy trees on the transect across the path intersection. The following drawing may provide a schematic picture.



If L represents the random distance between successive trees on the line transect, the probability that the investigator records a healthy run length x is proportional to $w(x) = E[L_1 + L_2 + \ldots + L_{x+1}] = (x+1)E[L]$. Without loss of generality, one may take E[L] = 1 and w(x) = x+1.

Now, if X ~ Geometric (θ), that is,

$$\begin{aligned} &f(x; \theta) = (1-\theta)^x\theta, \quad x = 0,1,2,\ldots \\ &E[X] = \mu = (1-\theta)/\theta, \\ &E[X+1] = 1/\theta. \end{aligned} \tag{6.1}$$

Also,

$$\begin{aligned} &f_w(x; \theta) = (x+1)(1-\theta)^x\theta^2, \\ &E[X_w] = 2\cdot(1-\theta)/\theta = 2E[X], \\ &E\left[\frac{1}{X_w+1}\right] = \theta. \end{aligned} \tag{6.2}$$

Note that X_w ~ Negative Binomial (2, θ). To estimate μ on the basis of $x_1, x_2, \ldots, x_n$ observed at intersections of n paths with the transect, the best unbiased estimate for μ is $\frac{1}{2}\bar{x}$ and not $\bar{x}$ as seen from (6.2).

Further, based on these n observations on X_w, the best unbiased estimate for θ is given by $1/(\bar{x}+1)$. However, a random sample on X does not provide any unbiased estimate for θ.

7. DISCRETE WAITING TIME PARADOX

Feller (1966) states continuous waiting time paradox, and this is discussed in this paper in Part II on continuous models. A discrete version of it arises in the context of the line transect sampling on one side of the forest paths for the purposes of estimating incidence as discussed in the previous section. Using the terminology there, let the run length between diseased trees follow a geometric process with mean length unity, that is, let $p = \frac{1}{2}$, and $E[X] = q/p = 1$ be the expected length of the healthy patch on the line transect *across* a forest path. If V denotes the healthy patch length on the line transect on one side of a forest path, a question arises as to what is $E[V]$? Is $E[V] = 1/2$, because the intersection of the forest path with the line transect has a random location relative to the healthy patch? Or, is $E[V] = 1$. The initial reasoning leading to $E[V] = 1/2$ can be modified by observing that what is observed is not V, but its size-biased version V_w for which $E[V_w] = 1/2\ E[X_w] = 1/2 \cdot 2 = 1$.

More situations involving discrete waiting time paradox may be identified. Uppuluri and Patil (1976) discuss exact discrete analog of Feller's continuous waiting time paradox in connection with inferences about rare events.

Now, we raise two further questions. Is the discrete waiting paradox an exclusive property of the geometric process? Or, of the weight function $w(x) = x+1$? We have the following partial answers when X has a power series distribution (for definitions, see Patil and Joshi (1968)).

Theorem: Assume that X has a power series distribution with series function $f(\theta) = \sum a(x)\theta^x$. Let the weight function be $w(x) = x+1$ for the weighted version X_w of X. Then $E[X_w] = 2E[X]$ if and only if $f(\theta) = (1-\theta)^{-1}$, in which case X has the geometric distribution with parameter $1-\theta$.

Proof: We know that

$$E[X] = \theta \frac{d}{d\theta} [\log f(\theta)].$$

Further, we observe that X_w can be shown to have the power series distribution with series function $f_w(\theta) = \frac{d}{d\theta}[\theta f(\theta)]$. The solution of

$$\frac{d}{d\theta}[\log f_w(\theta)] = 2 \frac{d}{d\theta}[\log f(\theta)]$$

gives $f(\theta) = (1-\theta)^{-1}$.
The proof for 'if' part is obvious.

Theorem: Assume that X has a power series distribution with series function $f(\theta) = \sum a(x)\theta^x$. Let $[f(\theta)]^{-k} = \sum b(x, k)\theta^x$. Let $w(x) \neq 0$, and the corresponding weighted version of X be X_w. Then $E[X_w] = kE[X]$ if and only if $w(x) = b(x, k)/a(x)$.

In particular, if X has a geometric distribution, $E[X_w] = 2E[X]$ if and only if $w(x) = x + 1$.

Proof: We observe that X_w has power series distribution with series function $f_w(\theta) = \sum w(x)a(x)\theta^x$. The rest of the proof is straightforward.

PART II. CONTINUOUS MODELS AND SIZE-BIASED PROBABILITY SAMPLING

8. RENEWAL THEORY AND ITS APPLICATIONS

Cox (1962) considers the following problem: Suppose that we have a number of independent realizations of the same renewal process, for example a number of components of the same type in use on different machines. Suppose that to investigate the distribution of failure-time a survey is made at time t to obtain the ages of the components currently in use. The distribution of the observations will be that of U_t, the backward recurrence time.

Cox obtains the limiting pdf of U_t as $t \to \infty$ to be $\bar{F}(x)/\mu$, where $\bar{F}(x) = 1 - F(x)$ with $F(x) = \int_0^x f(x)dx$, the distribution function of the life length of a component with expected mean life length μ. He also obtains the same result alternatively using the weighted-distribution-argument resultant from what he calls length-biased sampling as follows: Consider for any renewal process a recurrence-time R defined in the following way. First, we take a sampling point chosen at random over a very long time interval. Then R is defined as the time measured from the sampling point forward to the next renewal. Entirely the same properties would hold for the time measured from the sampling point back to the previous renewal.

It is clear that one samples from a population of failure-times distributed according to f(x) while the probability of selection of any individual in the population is proportional to its length x. If X_w denotes the failure time of the component in whose life the sampling point falls, then X_w has the p.d.f. $f_w(x) = xf(x)/\mu$. Conditionally on $X_w = x_w$, the p.d.f. of R is rectangular over $(0, x_w)$, which leads to its marginal p.d.f. at x to be $\int_x^\infty \frac{1}{x_w} \cdot f_w(x)dx = \bar{F}(x)/\mu$.

Cox (1969) gives the following example for length-biased sampling. An idealized model of a textile yarn is an assembly of parallel fibers, with p.d.f. of fiber length f(x). The fiber left-ends are arranged at random along a line. Take a particular cross-section of the yarn, that is, a particular point on the line, and consider the fibers that intersect this cross-section. This is length-biased sampling; the p.d.f. of length of the fibers is $xf(x)/\mu$. For discussions on similar problems, reference may be made to papers by Coleman (1972), Daniels (1942), Palmer (1948), Kriens (1963), and Moran (1966, 1969).

It may be instructive to record a few more examples.

(1) Morrison (1973): Size-biased sampling has implications for certain survey research data that is gathered at an arbitrary point in time. When you ask a person, "When did you last purchase Product A?", or "When did you last attend a baseball game?", etc., you are obtaining particular outcomes of a random variable U_t where U_t is the time elapsed since the last purchase until the randomly selected interview time t.

(2) Robson (1975): In assessing the extent of utilization of the national parks and other recreational facilities, an investigator asks an individual present on such a location as to since when he has been there. The data recorded on a number of such surveyed individuals have the size-biased feature.

9. WAITING TIME PARADOX

Feller, p. 11-13, introduces the waiting time paradox as follows: "Buses arrive in accordance with a Poisson process, the expected time between consecutive buses being α^{-1} (=1, say). I arrive at an epoch t say noontime sharp. What is the expectation $E[W_t]$ of my waiting time W_t for the next bus? Two contradictory answers stand to reason:

(a) The lack of memory of the Poisson process implies that $E[W_t]$ should be independent of t, that is, $E[W_t] = E[W_0] = 1$.

(b) The epoch of my arrival is 'chosen at random' in the interval between two consecutive buses, and for reasons of symmetry $E[W_t] = \frac{1}{2}$.

Let X be the inter-arrival time between two consecutive buses. It is given that X has the *pdf* $f(x) = e^{-x}$ of a standard exponential with $E[X] = 1$. Then it can be shown that the inter-arrival time X^* between two consecutive buses that cover my epoch t has *pdf* $f^*(x) = xe^{-x}$, so that $E[X^*] = 2$.

It is clear that $E[W_t] = E[E[W_t|X^*]] = E[\frac{1}{2} X^*] = \frac{1}{2} \cdot 2 = 1$. We note that using X instead of X^* in the preceding equation leads to the wrong answer of $\frac{1}{2}$ discussed in (b) above.

Now, in what follows, we prove a theorem that characterizes exponential distribution within the setup of the continuous waiting time paradox resulting from $E[X^*] = 2E[X]$.

Theorem: Let the inter-arrival times have a distribution belonging to the linear exponential family defined by the *pdf* $f(x) = \exp[\theta x + a(x) - b(\theta)]$, where $E[X] = b'(\theta) < \infty$. Let X^* have *pdf* $f^*(x) = xf(x)/E[X]$. Then X is exponential if and only if $E[X^*] = 2E[X]$.

Proof: From Theorem, $E[X^*] = E[X] + V(X)/E[X]$, and therefore $E[X^*] = 2E[X]$ implies $V(X) = (E[X])^2$, which relation characterizes the exponential distribution within linear exponential family, as proved in Wani and Patil [10].

Now for the exponential distribution within the setup of the continuous waiting time paradox resulting from $E[X^*] = 2E[X]$, we characterize the weight function to be $w(x) = x$, as follows.

Theorem: Let X have *pdf* $f(x) = \lambda e^{-\lambda x}$. Let X^* have *pdf* $f^*(x) = \frac{w(x)f(x)}{E\,w(X)}$, where $w(x) > 0$ with $E[w(X)] < \infty$. Then $E[X^*] = 2E[X]$ if and only if $w(x) = x$.

Proof: The condition that $E[X^*] = 2E[X]$ in this simplifies to

$$\frac{\int x w(x) e^{-\lambda x} dx}{\int w(x) e^{-\lambda x} dx} = \frac{2}{\lambda}.$$

We note that LHS represents the mean value function for a linear exponential family with exponent parameter λ. Following Patil and Shorrock [7], the mean value function has to be of the gamma distribution with index parameter 2, implying that $w(x) = x$.

Before we conclude this section, we note that Kotz and Johnson (1974) characterize the exponential distribution to be the distribution of the inter-arrival time X by requiring the waiting time W_t to be identically distributed like X, when t is chosen at random.

10. CELL CYCLE ANALYSIS AND PULSE LABELING

The following is based on Takahashi (1966) and Zelen (1974). The idealized cell cycle model for a proliferating cell consists of four phases. After mitosis, a cell which is destined to proliferate enters a (i) pre-DNA synthesis phase called G_1, (ii) DNA-synthesis or S-phase, (iii) post-DNA synthesis or G_2 phase, and finally (iv) mitosis during which time cell division occurs.

The cell cycle of a population of cells may be studied by exposing the cells to pulse labeling. Only those cells that are in S-phase get labeled. Further, when a labeled cell divides, the label is generally passed on to the daughter cells.

Periodically, the samples of cells from the cell population are observed under a microscope. The information on the proportion of cells labeled and the proportion of labeled mitotic cells can be used to estimate important characteristics of the kinetics of the cell cycle. Of particular interest is the mean duration of the S-phase and the mean cell cycle time.

A difficulty is in the interpretation of the experimental data. A diversity of opinions have been expressed in this connection. The data of Defendi and Manson (1963) or of Tolani (1965), if evaluated by Quastler's revised method (1963) indicate nearly twice as large values for mean S-phase durations. Should Quastler's original method (1959) be true, van't Hof's estimates (1965) for mean S-phase duration must be halved. A standardization of interpreting method is

therefore needed to render the published data comparable on a common basis.

As Zelen observes, it is not generally realized that the radioactive labeling of cells is actually choosing a sample of cells by a length-biased sampling procedure. Cells with longer S-phase durations will have a higher probability of being labeled. Hence, the parameters obtained from labeling experiments are different from those of the target population of cells under investigation. It is both instructive and interesting to see that the following analysis using a limiting size-biased sampling argument can explain the puzzle of double and half mentioned above for the mean S-phase duration.

Let X and T denote the durations of the S-phase and the cell cycle. Let their joint *pdf* be f(x, t). Then, the observed x and t have the *pdf* given by

$$f^w(x, t) = \frac{xf(x,t)}{E[X]} \quad 0 \le x \le t \le \infty.$$

It follows that $E(X_w) = E(X)[1+C^2]$ and $E(T_w) = E(T) + \frac{Cov(X,T)}{E(X)}$ where C is the coefficient of variation of X. Assuming that the coefficient of variation of X is unity which is the case when X is exponential, we get

$$E[X_w] = 2E[X] > E[X].$$

Further, if X and T-X are uncorrelated,

$$E[T_w] = E[T] + E[X] > E[T].$$

Thus under the assumptions of X being exponential, and with X and T-X being uncorrelated, the mean duration of the S-phase at the label point will be larger than the population mean by a factor of two and the cell cycle length will have a bias of E[X].

11. EFFICACY OF EARLY SCREENING FOR DISEASE

Currently there is increasing emphasis on early detection programs which identify individuals who are unaware that they have a particular chronic disease. The hope is that earlier detection and treatment will result in an enhanced therapeutic benefit to the individual. Zelen (1974) provides an illuminating discussion of certain aspects of the problems involved. The following is taken from this discussion.

Consider an individual to be in one of three disease states, S_0: a disease free state; S_p: a pre-clinical state; S_c: a clinical state. In the state S_0, individuals do not have the disease, or have a form of the disease which cannot be found by the early detection program. The definition of the pre-clinical state is that the individual has the disease, but is unaware of this condition.

The early detection program is capable of identifying the individual as being in S_p. The clinical state is defined as one where the disease has become clinical and has been diagnosed.

Now consider a population of individuals having a particular chronic disease. There will be a probability distribution governing the duration of the pre-clinical disease. At a particular point in time, this population is examined under an early detection program, and those in S_p are identified. Thus an early detection program identifies individuals by a size-biased scheme. Hence, if a tumor is slow growing, the pre-clinical state is long, then the clinical course of the disease will tend to be long, resulting in longer survival of the individual. Thus, regardless of whether the therapy is enhanced by early detection, individuals so found will tend to live longer than the general population of individuals having this same disease. Consequently, evidence of longer survival for an earlier detected group of cases relative to a control group is not valid scientific evidence of the effectiveness of the screening program.

Under suitable assumptions, it can be shown that

$$E[R_w] = \frac{1}{2}\,(E[T] + V(T)/E[T]) \tag{11.1}$$

where T is the duration of S_p, and R is the true lead time gained by earlier diagnosis because of screening or the duration of the pre-clinical state until screening. R_w is the observed R under size-bias. The time gained by early diagnosis cannot be observed directly. Once the disease is diagnosed, its course is interrupted by treatment. However, the mean lead time can be estimated indirectly by

$$E[R_w] = E[T] - (A_c - A_p), \tag{11.2}$$

where A_p is the mean age of an individual when detected in S_p by screening, and A_c is the mean age at transition into S_c if not screened, (Zelen and Feinleib (1960)).

If T is exponential, (11.1) and (11.2) imply $A_p = A_c$, which says that the mean age of those detected early as well as of those who are routinely diagnosed without the benefit of an early detection program are identical. This apparent paradox arises because of two compensating features. Although those diagnosed by an early detection procedure are found earlier, the size-biased feature of the detection shows that the individuals tend to spend longer time in S_p. Furthermore, if the coefficient of variation T exceeds one, the mean age of individuals detected in an early detection program will even be higher than for those detected through routine medical care.

12. FOREST PRODUCTS RESEARCH

The present discussion of size-biased sampling in forest products research is based on Warren (1975). The general problem here concerns identifying a distribution from data collected by a method in which the probability of an individual being included in the sample is a function of that individual's size. A particular instance arises in examining the size of wood cells. A convenient method consists of measuring the size (cross-sectional area) of only those cells selected by a set of random points on a microscopic field. The probability of a cell being selected is then proportional to its cross-sectional area size, in which case $w(x) = x$. In effect, we have two distributions: (1) the underlying distribution of cell size about which we wish to make inferences, and (2) the distribution generated by our sampling mechanism.

The situations where size-biased sampling is conveniently employed commonly exhibit large numbers of small individuals, with declining numbers of large individuals. Under certain environmental conditions, the diameters of trees may have a distribution of this form, and the forest then be economically surveyed by an analogous system.

13. PARTICLE SIZE DISTRIBUTIONS BY THIN SECTION METHODS

Particle size statistics is an important subject area in several scientific fields such as physics, geology, agricultural engineering, etc. The following discussion is primarily based on Kendall and Moran (1963).

Consider a problem of determining the distribution of the sizes of particles embedded in an opaque medium from the measurement of the figures formed by their intersections with a random plane, or from the segments formed by their intersections with a random line. For example, Tallis (1970) considers a problem as follows: A large block of Swiss cheese was thinly sliced and one hundred slices were drawn at random with replacement. Diameters of each hole appearing in the slices were suitably obtained. It was required to estimate the distribution function of the diameters of the spherical airspaces in the cheese.

If R denotes the diameter of a sphere with *pdf* $f(r)$, the probability of this sphere being intersected by a random plane is proportional to r, and therefore the *pdf* of the diameter of a sphere cut by a random plane is

$$f_w(r) = rd(r)/\int_0^\infty rf(r)dr = rf(r)/\mu. \tag{13.1}$$

Actually, the circular section of the sphere with the plane is what is observed and its diameter x recorded. One obtains its density as

$$\phi(x) = x\int_x^\infty \left(\frac{1}{r} f_w(r)/\sqrt{(r^2-x^2)}\right) dr, \tag{13.2}$$

and gets the solution for $f(r)$ as

$$f(r) = -\frac{2r\mu}{\pi}\int_{x^2}^\infty (r^2-x^2)^{-1/2}\frac{d}{dx}(x^{-1}\phi(x))dx. \tag{13.3}$$

In his pioneering paper, Krumbein (1935) applied the above theory to the analysis of thin sections of sediments in petrography, assuming the particles to be spheres. His original theory neglected the fact that larger spheres have a larger probability of being included in the section. Later work by Greenman (1951), Rosenfelt et al. (1953), Packham (1955) also use the incorrect equation by wrongly using $f(r)$ where $f_w(r)$ given by (13.1) needs to be used. Interestingly, however, Krumbein obtained a closer fit to observed values with his simpler and incorrect theory when examining two samples of sand boty by this cross-section method and by actual measurement of the particle size after separation.

We note that if a random line (linear probe) was used instead of a random plane (planar probe), the diameter of the sphere so detected will the *pdf*

$$f_w(r) = r^2 f(r)/\int_0^\infty r^2 f(r)dr = r^2 f(r)/\mu_2'. \tag{13.4}$$

Further, if a sphere of volume v with *pdf* $f(v)$ is selected by a sampling mechanism with probability proportional to its surface area, the *pdf* of the observed v will be

$$f_w(v) = v^{2/3} f(v)/\int_0^\infty v^{2/3} f(v)dv. \tag{13.5}$$

14. AERIAL SURVEY IN TRAFFIC RESEARCH

We have discussed an aerial survey problem in wildlife ecology in Section 5. In quite a different way, a different kind of observational bias enters in the aerial survey of low density traffic streams. This discussion is based on Brown (1972).

Assume that vehicles enter the highway according to a non-homogeneous Poisson process of intensity λ. The vehicles choose velocities at random from a distribution of V with *pdf* $f(v)$. Assume that $E[1/V] < \infty$.

It turns out that the velocity of a vehicle in $[a,b]$ at time t, as $t \to \infty$, has its limiting *pdf* as

$$f_w(v) = \frac{1}{v} f(v) \Big/ \int_0^\infty \frac{1}{v} f(v)dv \qquad (14.1)$$

for all [a,b]. Moreover, the velocities of vehicles which lie in [a,b] at time t are independent. Thus if one records the velocities of vehicles in [a,b] at time t (for example through aerial photographs) for t large, he will be observing a random sample of Poisson size drawn from the *pdf* $f_w(V)$ and not the *pdf* f(v).

The main effect of this is that he will tend to observe an unduly high proportion of slower vehicles. This can be seen directly by comparing the two distribution functions corresponding to f(v) and $f_w(v)$. And this is intuitively plausible since slower vehicles take longer to traverse [a,b] and therefore tend to be around when the interval is samples. The problem, then, is to construct an estimate of the distribution function of f(v) based on a sample of Poisson size from the distribution function of $f_w(v)$ defined by (14.1).

15. REJECTION TECHNIQUE IN RANDOM VARIABLES GENERATION

In this section we show that the rejection technique used in the generation of random variables is simply a computer counterpart of the concept of weighted distributions. This discussion is based on Patil-Boswell-Friday (1975).

For, suppose we can generate a r.v. Y with pdf $f_Y(y)$, and we want to generate a r.v. X with pdf $f_X(x)$. The rejection technique consists of two parts: generate an observation on Y (=y, say), and accept y to be the realization of X with probability w(y). Proceed until a realization of Y is accepted. Clearly, in order for this method to work, Y has to assume every value that X can assume. Further,

$$f_X(x) = \frac{w(x)f_Y(x)}{\omega}$$

where ω is the probability that a random generation of Y is accepted. Thus, the acceptance probability w(x) is determined by

$$w(x) = \omega \cdot f_X(x)/f_Y(x)$$

implying that w(x) is proportional to the likelihood ratio of r.v. X on r.v. Y. Further, w(x) and ω are maximized by choosing $1/\omega = \sup f_X(x)/f_Y(x)$, which if bounded, the method works. If not bounded, w(x) > 1 for some x, and is no longer a probability for such x. What can be done in such a case is an interesting problem. The following approach may have some use in certain situations.

Let X have *pdf* f(x). Let the weight function be arbitrary, that is, $0 < w(x) < \infty$ with $E[w(X)] < \infty$. Let X_w have the *pdf* $f^w(x) = w(x)f(x)/E[w(X)]$.

If X_w has to be computer-generated from available X, X_w may be realized by using the formulation of

$$f_2(x) = w_1(x)f_1(x)/E_1[w_1(X)]$$

and

$$f^w(x) = w_2(x)f_2(x)/E_2[w_1(X)]$$

where $f_1(x) = f(x)$, $w_1(x) = w(x)/n(x) \leq 1$, and $w_2(x) = n(x) = [w(x)]$, the least integer not less than w(x).

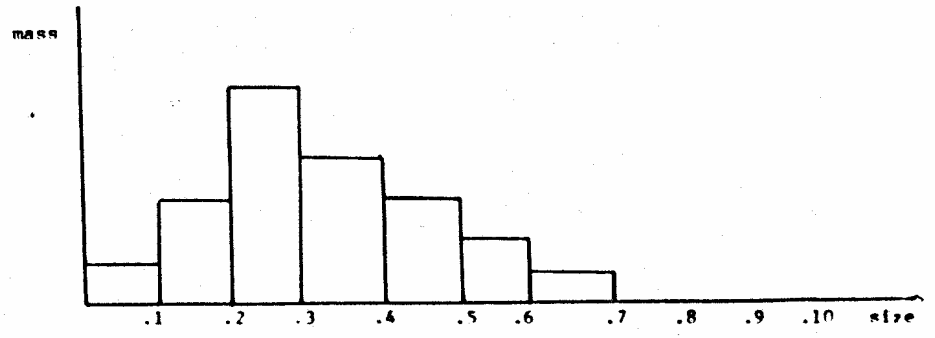
Thus, the available x corresponding to *pdf* f(x) is selected with probability $w_1(x)$, and the selected x is recorded $w_2(x)$ times. If necessary, one may computer-generate x also, and repeat the procedure a large number of times. The recorded values then constitute a simulated population of values for X_w. A random sample of desired size on X_w can be drawn as usual from this population of values of X_w.

PART III. WEIGHTED DISTRIBUTIONS AS MASS-SIZE DISTRIBUTIONS IN GEOLOGY AND ECONOMICS

16. WEIGHT-SIZE DISTRIBUTIONS IN A BREAKAGE PROCESS

Problems relating to small particle statistics continue to be important in physics, sedimentology, atmospheric science and air environmental research. This brief discussion here is based on Schultz (1975).

The distribution of particles formed in a physical breakage process is usually analyzed by mass rather than frequency. Consider, for example, a sieve analysis of crushed stone after a standard drum test, where the weights, or masses, of all particles retained by sieves of various sizes as shown in the following histogram.



With reference to results such as the above, Krumbein and Pettijohn (see Herdan (1960), p. 291) remarked, "There is one conspicuous manner in which the statistical data of coarse disperse systems differ from the conventional statistical data ... frequency in coarse disperse systems is usually expressed by weight instead of by number. No complete investigations of this ... have been made and the problem of weight versus number is still largely unsolved."

Let X be the diameter size with pdf $f(x; \theta)$ with $\mu_j' = E(x^j)$. It is interesting to note that the mass-size density is nothing but $f_w(x; \theta) = x^3 f(x; \theta)/\mu_3'$, the weighted pdf with weight function $w(x) = x^3$ of the random variable X_w, say. One may wish to refer to Herdan (1960) for several interesting features and problems related to the weight-size distributions. One problem is to estimate $\mu = E[X]$. It is easy to verify that μ has an alternative interpretation in terms of X_w given by $\mu = H(X_w^3)/H(X_w^2)$, where H stands for the harmonic mean. This is suggested to be an estimating equation for μ in applied literature. It is clear that the observed diameter mean will be estimating μ_4'/μ_3' and not μ in general.

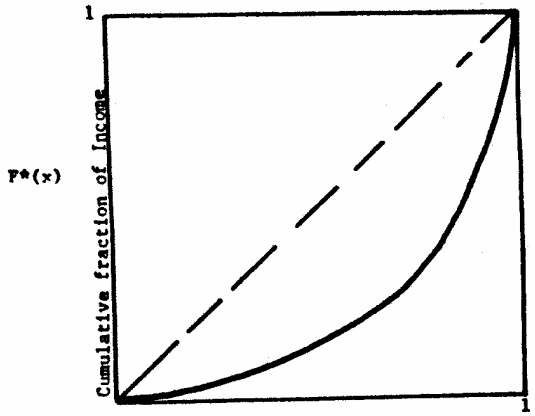
17. MOMENT DISTRIBUTIONS IN ECONOMICS

In the economics literature, weighted distributions with weight functions $w(x) = x^j$ are used. They have been called as moment distributions. The following discussion is based on Hart (1975), Klein (1962), and Ord (1975).

Income distributions at different points of time or location are frequently compared and ranked by some measure of inequality, with low values being preferred to higher values. An income distribution usually follows a typical skew pattern. A main descriptive feature of such a distribution is its degree of inequality. Still another type of tabulation and graph discussed below brings out the degree of inequality.

The Lorenz curve associated with a distribution gives a joint cumulation, both of the frequencies and of the variables being distributed. For example, the distribution of income may be cumulated to show the per cent of income received by the bottom tenth class of spending units, by the bottom two tenth classes, and so on. In other words, if $X \sim F(x)$, where X is the income of an individual having distribution function F(x) and if $F^*(x) = \int_0^x x\,dF(x)/E[X]$, the Lorenz curve is the graph of $F^*(x)$ on F(x). Clearly, $F^*(x) \leq F(x)$.

The diagonal line is the curve of equal distribution. The departure of the actual curve from the line of perfect equality shows the degree of inequality. A merit of the Lorenz curve technique is that it enables us to compare distributions in dissimilar units (currencies).



REFERENCES

[1] BERGERUD, A. T. and MANUEL, F. (1969). Aerial census of moose. *J. Wildlife Mmagement*, 33, 910-916.

[2] BLUMENTHAL, S. (1967). Proportional sampling in life length studies. *Technometrics*, 9, 205-218.

[3] BROWN, M. (1972). Low density traffic streams. *Advances in Appl. Probability*, 4, 177-192.

[4] COLEMAN, R. (1972). Sampling procedures for the lengths of random straight lines. *Biometrika*, 59, 415-426.

[5] COOK, R. D. and MARTIN, F. B. (1974). A model for quadrat sampling with "visibility bias". *J. Amer. Statist. Assoc.*, 69, 345-349.

[6] COX, D. R. (1962). *Renewal Theory*. Frome and Long: Butler & Tanner, Ltd.

[7] COX, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, (Johnson and Smith, Eds.). 289-316. University of North Carolina Press.

[8] DANIELS, H. E. (1942). A new technique for the analysis of fibre length distribution in wool. *J. Text. Inst.*, 33, 137-150.

[9] DAVID, H. A. (1973). Waiting time paradoxes and order statistics. *J. Amer. Statist. Assoc.*, 68, 743-745.

[10] EVANS, C. D., TROYER, W. A. and LENSINK, C. J. (1966). Aerial census of moose by quadrat sampling units. *J. Wildlife Mmagement*, 30, 767-776.

[11] FELLER, W. (1966). *Introduction to Probability Theory and Applications*, 2, 10-14. Wiley & Sons, New York.

[12] FISHER, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.*, 6, 13-25.

[13] GUPTA, R. C. (1975). Some characterizations of discrete distributions by properties of their moment distributions. *Comm. Statist.*, 4, 761-765.

[14] HALDANE, J. B. S. (1938). The estimation of the frequency of recessive conditions in man. *Ann. Eugen.*, 7, 255-262. London.

[15] HART, P. E. (1975). Moment distributions in economics: an exposition. *J. Roy. Statist. Soc. Ser. A.*, 138, 423-434.

[16] HART, P. E. (1976). The comparative statistics and dynamics of income distributions. *J. Roy. Statist. Soc. Ser. A.*, 139, 108-125.

[17] HERDAN, G. (1960). *Small Particle Statistics.* Elsevier, Amsterdam.

[18] KEMP, C. D. (1973). An elementary ambiguity in accident theory. *Accid. Anal. of Prev.*, 5, 371-373.

[19] KEMP, C. D. (1975). Models for visibility bias in quadrat sampling. Statistics Reports and Reprints, No. 22, University of Bradford.

[20] KENDALL, M. G. and MORAN, P. A. P. (1963). *Geometrical Probability.* Charles Griffin & Co., Ltd., London.

[21] KOTZ, S. and JOHNSON, N. L. (1974). A characterization of exponential distributions by a waiting time property. *Comm. Statist.*, 3, 257-258.

[22] KRIENS, J. (1963). The procedures suggested by de Wolff and van Heerden for random sampling in auditing. *Statist. Neerlandica*, 17, 215-231.

[23] MORAN, P. A. P. (1966). Measuring the length of a curve. *Biometrika*, 53, 359-364.

[24] MORAN, P. A. P. (1969). A second note on recent research in geometrical probability. *Advances in Appl. Probability*, 1, 73-89.

[25] MORRISON, D. G. (1973). Some results for waiting times with an application to survey data. *The American Statistician*, 27, 226-227.

[26] NEEL and SCHULL. (1966). *Human Heredity*, 211-227. University of Chicago Press, Chicago.

[27] NOLL, K. E. and PILAT, M. J. (1971). Size distribution of atmospheric giant particles. *Atmospheric Environment*, 5, 527-540.

[28] ORD, K. (1975). Statistical models for personal income distributions. In *Statistical Distributions in Scientific Work* (G. P. Patil, S. Kotz and K. Ord, eds.), 2, 151-158.

[29] PALMER, R. C. (1948). The dye sampling method of measuring fibre length distribution. *J. Text. Inst.*, 39, 8-22.

[30] PALMER, R. C. and DANIELS, H. E. (1947). The sampling problem in single fibre testing. *J. Text. Inst.*, 38, 94-100.

[31] PATIL, G. P., BOSWELL, M. T. and FRIDAY, D. S. (1975). Chance mechanisms in computer generation of random variables. In *Statistical Distributions in Scientific Work* (G. P. Patil, S. Kotz and K. Ord, eds.).

[32] PATIL, G. P. and JOSHI, S. W. (1968). *A Dictionary and Bibliography of Discrete Distributions.* Oliver and Boyd and Hafner Publishing Company, Edinburgh and New York.

[33] PATIL, G. P. and ORD, J. K. (1975). On size-biased sampling and related form-invariant weighted distributions. *Sankhyā* (To appear).

[34] PROROK, P. C. (1976). The theory of periodic screening I: a lead time and proportion detected. *Advances in Appl. Probability*, 8, 127-143.

[35] RAO, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions* (G. P. Patil, ed.), 320-332. Statistical Publishing Society, Calcutta. Also reprinted in *Sankhyā A*, 27, 311-324.

[36] RAO, C. R. and RUBIN, H. (1964). On a characterization of the Poisson distribution. *Sankhya*, 26, 295-298.

[37] ROBSON, D. S. (1975). A person communication.

[38] SCHEAFFER, R. L. (1972). Size-biased sampling. *Technometrics*, 14, 635-644.

[39] SCHOTZ, W. E. and ZELEN, M. (1971). Effect of length sampling bias on labelei mitotic index waves. *J. Theoret. Biol.*, 32, 383-404.

[40] SCHULTZ, D. M. (1975). Mass-size distributions: a review and a proposed new model. In *Statistical Distributions in Scientific Work* (G. P. Patil, S. Kotz and K. Ord, eds.), 275-288.

[41] SCHULTZ, D. M. and CROUSE, C. F. (1973). Random splittings: a model for a mass-size distribution. *S. African Statist. J.*, 7, 143-152.

[42] SIMON, R. (1975). Length biased sampling in the estimation of antigen frequencies. National Cancer Institute. A Preliminary Paper.

[43] SMART, R. J. (1963). Alcoholism, birth order, and family size. *J. Abnorm. Soc. Psychology*, 66, 17-23.

[44] SPROTT, D. A. (1964). Use of chi square. *J. Abnormal and Social Psychology*, 69, 101-103.

[45] TAKAHASHI, M. (1966). Theoretical basis for cell cycle analysis. II. Further studies on labeled mitosis wave method. *J. Theoret. Biol.*, 18, 195-209.

[46] TAKAHASHI, M. (1968). Theoretical basis for cell cycle analysis. I. Labeled mitosis wave method. *J. Theoret. Biol.*, 13, 202-211.

[47] TALLIS, G. M. (1970). Estimating the distribution of spherical and elliptical bodies in conglomerates from plane sections. *Biometrics*, 26, 87-104.

[48] TRUCCO, E. and BROCKWELL, P. J. (1968). Percentage labeled mitosis curves in exponentially growing cell populations. *J. Theoret. Biol.*, 20, 321-337.

[49] UPPULURI, V. R. R. and PATIL, S. A. (1976). Inferences about rare events. Oak Ridge National Laboratory. Technical Report. 1-6.

[50] VAN BELLE, G. and SCHNEIDERMAN, M. (1973). Some statistical aspects of environmental pollution and protection. *Int. Statist. Rev.*, 41, 315-331.

[51] WALLACH, R. V. and SICHEL, H. S. (1963). The measurement of mechanical strength and abradability of coke by means of statistical parameters. *J. Inst. Fuel.*, 36, 421-435.

[52] WARREN, W. G. (1975). Statistical distributions in forestry and forest products research. In *Statistical Distributions in Scientific Work*, (G. P. Patil, S. Kotz and K. Ord, eds.), 2, 369-384.

[53] WATSON, G. S. (1971). Estimating functionals of particle size distributions. *Biometrika*, 58, 483-490.

[54] ZELEN, M. (1971). Problems in the early detection of disease and the finding of faults. *Bull. Int. Statist. Inst.*, 38, 649-661.

[55] ZELEN, M. (1974). Problems in cell kinetics and the early detection of disease. *Reliability and Biometry. SIAM*, 701-726.