CHAPTER 24

Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent?

C. Radhakrishna Rao

Abstract

The concept of weighted distributions can be traced to the study of the effects of methods of ascertainment upon the estimation of frequencies by Fisher in 1934, and it was formulated in general terms by the author in a paper presented at the First International Symposium on Classical and Contagious Distributions held in Montreal in 1963. Since then, a number of papers have appeared on the subject. This paper reviews some previous work, points out, through appropriate examples, some situations where weighted distributions arise, and discusses the associated methods of statistical analysis.

Weighted distributions occur in a natural way in specifying probabilities of events as observed and recorded by making adjustments to probabilities of actual occurrence of events taking into account methods of ascertainment. Failure to make such adjustments can lead to wrong conclusions.

1. Sample and Population

Statisticians are often required to work with data provided by customers and answer questions raised by them. The questions relate to a real or hypothetical population which gave rise to the data, usually referred to as a "sample from a population." The role of statistical methodology is to extract the relevant information from a given sample to answer specific questions about the parent population (which the sample is presumed to represent). For this purpose, it is necessary to identify all possible samples that can be observed from a population (sample space) and to provide a stochastic model for attaching probabilities to different sets of samples (specification). The link between sample and population is specification, and the question "What population does a sample represent?" is technically equivalent to "What is the appropriate specification?" Wrong specification can lead to invalid inference, which is sometimes referred to as the third kind of error, the first two being the

Key words and phrases: damage models, nonresponse, probability sampling, quadrat sampling, size biased sampling, truncation, weighted distributions.

A. C. Atkinson et al. (eds.), *A Celebration of Statistics*

familiar two kinds of errors associated with the Neyman–Pearson theory of testing of hypotheses.

How does a statistician decide on an appropriate specification for a given sample? Unfortunately, there is not much discussion of this basic question in statistical literature, although any statistical inference presupposes some kind of specification. The present paper is primarily addressed to the problem of specification based on how a sample is drawn from a population. It illustrates through live examples in some areas of applied research the use of what are called weighted distributions in choosing the appropriate specification and the associated statistical methodology.

The problem of specification is not a simple one. A detailed knowledge of the procedure actually employed in acquiring data is an essential ingredient in arriving at a proper specification. The situation is more complicated with field observations and nonexperimental data, where nature produces events according to a certain stochastic model, which are observed and recorded by investigators. There does not always exist a suitable sampling frame for observing events and applying the classical sampling theory. In practice, it is not always possible to observe and record all events which occur. For instance, certain events may not be observable by the method we follow and therefore be missed in the record (truncated, censored, and incomplete samples). Or an event may be observable only with a certain probability depending on the characteristics of the event, such as its conspicuousness and the procedure employed to observe it (unequal probability sampling). Or an event may change in a random way by the time or during the process of observation, so that what comes on record is a modified event (damage models). Sometimes, events produced under two or more different mechanisms with unspecified relative frequencies get mixed up and brought into the same record (outliers, contaminated samples). In all these cases, the specification for the original events (as they occur) may not be appropriate for the events as they are recorded (observed data) unless it is suitably modified.

In a classical paper, Fisher (1934) demonstrated the need for such adjustment in specification depending on the way the data are ascertained. In extending the basic ideas of Fisher, the author (Rao, 1965) introduced the concept of a weighted distribution as a method of adjustment applicable to many situations. In the present paper we discuss, through live examples, some procedures for making adjustments in specification based on methods of ascertaining data.

Although I have mentioned only field observations which are collected without the help of a suitable sampling frame, I must emphasize that similar problems of specification arise with data collected in large scale sample surveys and also with data acquired through field and laboratory experiments. Survey practitioners are faced with problems of incomplete frame, which raise questions of the representativeness of a sample for a given population (see Kruskal and Mosteller, 1980, and references therein); nonresponse, which raises questions of repeated visits to sampled units; replacement of nonresponding units by others with possibly similar characteristics, and imputation of values (Fienberg and Tanur, 1983; Fienberg and Stasny, 1983; Rubin, 1976, 1980); and nonsampling errors, which raise questions about their recognition, detection, and measurement, and lead to making adjustments in expressing the precision of estimates (Mahalanobis, 1944; Mosteller, 1978). Similarly, in the design of experiments, difficulties in random allocation of treatments and choice of controls in field trials, pooling of evidence from different experiments conducted over space and time, and missing values (dropouts) introduce additional uncertainties in statistical inference and the interpretation of results for practical use or policy purposes (for typical problems and references see Fienberg, Singer, and Tanur, 1985, Chapter 12 in this volume; Neyman, 1977).

2. Truncation and Censoring

Some events, although they occur, may be unascertainable, so that the observed distribution is truncated to a certain region of the sample space. An example is the frequency of families with both parents heterozygous for albinism but having no albino children. There is no evidence that the parents are heterozygous unless they have an albino child, and the families with such parents and having no albino children get confounded with normal families. The actual frequency of the event "zero albino children" is thus not ascertainable. Adjustment to the probability distribution applicable to observable events in such a case is simple.

In general, if $p(x, \theta)$ is the pdf (probability density function), where θ denotes unknown parameters, and the rv X is truncated to a specified region $T \subset \mathscr{X}$, the sample space, then the pdf of the truncated random variable X^w is

$$p^w(x, \theta) = \frac{w(x, T)p(x, \theta)}{u(T, \theta)}, \tag{2.1}$$

where $w(x, T) = 1$ if $x \in T$ and $= 0$ if $x \notin T$, and $u(T, \theta) = E[w(X, T)]$. If $x_1, \ldots, x_n$ are independent observations subject to truncation, then the likelihood is

$$\frac{p(x_1, \theta) \cdots p(x_n, \theta)}{[u(T, \theta)]^n}. \tag{2.2}$$

In some cases we may have independent observations $x_1, \ldots, x_n$ arising from a truncated distribution in addition to a number m (and not the actual values) of observations falling outside T. Then the likelihood is

$$\frac{(n + m)!}{m!} p(x_1, \theta) \cdots p(x_n, \theta)[1 - u(T, \theta)]^m. \tag{2.3}$$

A more complicated case is the following. Suppose that we have a measuring device which records the time at which a bulb fails. If we are experimenting with n bulbs in a life testing problem using a measuring device which may itself fail at a random time, then the observations would be of the type

$$x_1, \ldots, x_{n_1}, n_2, n_3, \tag{2.4}$$

where $x_1, \ldots, x_{n_1}$ are the lifetimes of n_1 bulbs recorded before an unknown time point T at which the measuring device failed, n_2 is the number of bulbs that failed between T and T_0, the known time at which the experiment was terminated, and n_3 is the number of bulbs still burning after T_0. Let

$$w_1(T, \theta) = P(x \leq T), \qquad w_2(T, \theta) = P(T < x \leq T_0),$$
$$w_3(T, \theta) = 1 - w_1(T, \theta) - w_2(T, \theta).$$

Then the likelihood based on the data (2.4) is

$$\frac{n!}{n_2! n_3!} p(x_1, \theta) \cdots p(x_{n_1}, \theta) [w_2(T, \theta)]^{n_2} [w_3(T, \theta)]^{n_3}, \tag{2.5}$$

where T is unknown as well as the basic parameters θ. Inference on T and θ based on (2.5) does not seem to have been fully worked out, but could be developed on standard lines.

The expressions (2.2), (2.3), and (2.5) are simple examples of weighted distributions, whose general definition is given in Section 3.

3. Weighted Distributions

In Section 2, we have considered situations where certain events are unobservable. But a more general case is where an event that occurs has a certain probability of being recorded (or included in the sample). Let X be a rv with $p(x, \theta)$ as the pdf, and suppose that when $X = x$ occurs, the probability of recording it is $w(x, \alpha)$, depending on the observed value x and possibly also on an unknown parameter α. Then the pdf of the resulting rv X^w is

$$p^w(x, \theta, \alpha) = \frac{w(x, \alpha) p(x, \theta)}{E[w(X, \alpha)]}. \tag{3.1}$$

Although in deriving (3.1) we chose $w(x, \alpha)$ such that $0 \leq w(x, \alpha) \leq 1$, we can define (3.1) for any arbitrary nonnegative weight function $w(x, \alpha)$ for which $E[w(X, \alpha)]$ exists. The distribution (3.1) obtained by using any nonnegative weight function $w(x, \alpha)$ is called (see Rao, 1965) a weighted version of $p(x, \theta)$. In particular, the weighted distribution

$$p^w(x, \theta) = \frac{|x| p(x, \theta)}{E[|x|]}, \tag{3.2}$$

where $|x|$ is the norm or some measure of size of x, is called the size biased distribution. When x is univariate and nonnegative, the weighted distribution

$$p^w(x, \theta) = \frac{xp(x, \theta)}{E(X)} \tag{3.3}$$

is called length (size) biased distribution. For example, if X has the logarithmic series distribution

$$\frac{\alpha^r}{-r\log(1-\alpha)}, \qquad r = 1, 2, \ldots, \tag{3.4}$$

then the distribution of the size biased variable is

$$\alpha^{r-1}(1-\alpha), \qquad r = 1, 2, \ldots, \tag{3.5}$$

which shows that $X^w - 1$ has a geometric distribution. A truncated geometric distribution is sometimes found to provide a good fit to an observed distribution of family size (Feller, 1968). But, if the information on family size has been ascertained from school children, then the observations will have a size biased distribution. In such a case a good fit of the geometric distribution to the observed family sizes would indicate that the underlying distribution of family size is, in fact, a logarithmic series.

Table 1 gives a list of some basic distributions and their size biased forms. It is seen that the size biased form belongs to the same family as the original distribution in all cases except the logarithmic series [see Rao (1965), Patil and Ord (1975), Janardhan and Rao (1983) for characterizations and examples of size biased distributions].

An example of weighted distributions arises in sample surveys when unequal probability sampling or pps (probability proportional to size) sampling is employed. A general version of the sampling scheme involves two rv's X and Y with pdf $p(x, y, \theta)$ and a weight function $w(y)$ which is a function of y only, giving the weighted pdf

$$p^w(x, y, \theta) = \frac{w(y)p(x, y, \theta)}{E[w(Y)]}. \tag{3.6}$$

In sample surveys we obtain observations on (X^w, Y^w) from the pdf (3.6) and draw inference on the unknown parameter θ.

It is of interest to note that the marginal pdf of X^w is

$$p^w(x, \theta) = \frac{w(x, \theta)p(x, \theta)}{E[w(X, \theta)]}, \tag{3.7}$$

which is a weighted version of $p(x, \theta)$ with the weight function

$$w(x, \theta) = \int p(y \mid x)w(y)\, dy, \tag{3.8}$$

which may involve the unknown parameter θ.

Table 1. Certain Basic Distributions and Their Size-Biased Forms

Random variable (rv)	pf (pdf)	Size-biased rv
Binomial, $B(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$1+B(n-1,p)$
Negative binomial, $\mathrm{NB}(k,p)$	$\binom{k+x-1}{x}q^x p^k$	$1+\mathrm{NB}(k+1,p)$
Poisson, $\mathrm{Po}(\lambda)$	$e^{-\lambda}\lambda^x/x!$	$1+\mathrm{Po}(\lambda)$
Logarithmic series, $L(\alpha)$	$\{-\log(1-\alpha)\}^{-1}\alpha^x/x$	$1+\mathrm{NB}(1,\alpha)$
Hypergeometric, $H(n,M,N)$	$\binom{n}{x}\frac{M^x(N-M)^{n-x}}{N^n}$	$1+H(n-1,M-1,N-1)$
Binomial beta, $\mathrm{BB}(n,\alpha,\gamma)$	$\binom{n}{x}\frac{\beta(\alpha+x,\gamma+n-x)}{\beta(\alpha,\gamma)}$	$1+\mathrm{BB}(n-1,\alpha,\gamma)$
Negative binomial beta, $\mathrm{NBB}(k,\alpha,\gamma)$	$\binom{k+x-1}{x}\frac{\beta(\alpha+x,\gamma+k)}{\beta(\alpha,\gamma)}$	$1+\mathrm{NBB}(k+1,\alpha,\gamma)$
Gamma, $G(\alpha,k)$	$\alpha^k x^{k-1}e^{-\alpha x}/\Gamma(k)$	$G(\alpha,k+1)$
Beta first kind, $B_1(\delta,\gamma)$	$x^{\delta-1}(1-x)^{\gamma-1}/\beta(\delta,\gamma)$	$B_1(\delta+1,\gamma)$
Beta second kind, $B_2(\delta,\gamma)$	$x^{\delta-1}(1+x)^{-\gamma}/\beta(\delta,\gamma-\delta)$	$B_2(\delta+1,\gamma-\delta-1)$
Pearson type V, $\mathrm{Pe}(k)$	$x^{-k-1}\exp(-x^{-1})/\Gamma(k)$	$\mathrm{Pe}(k-1)$
Pareto, $\mathrm{Pa}(\alpha,\gamma)$	$\gamma\alpha^\gamma x^{-(\gamma+1)},\ x\geq\alpha$	$\mathrm{Pa}(\alpha,-1)$
Lognormal, $\mathrm{LN}(\mu,\sigma^2)$	$(2\pi\sigma^2)^{-1/2}x^{-1}\exp-\left(\frac{\log x-\mu}{\sigma\sqrt{2}}\right)^2$	$\mathrm{LN}(\mu+\sigma^2,\sigma^2)$

An extensive literature on weighted distributions has appeared since the concept was formalized in Rao (1965); it is reviewed with a large number of references in a paper by Patil (1984) with special reference to ecological work. Reference may also be made to two earlier contributions by Patil and Rao (1977, 1978), and Patil and Ord (1976) which contain reviews of previous work and details of some new results.

In the next sections, we consider several examples where weighted distributions are used in the analysis of data.

4. Are Only Small Skulls Well Preserved?

The following problem arose in the analysis of cranial measurements. A sample of skulls dug out from ancient graves in Jebel Moya, Africa, consisted of some well-preserved skulls and the rest in a broken condition (see Mukherji, Trevor, and Rao, 1955). On each well-preserved skull it was possible to take four measurements, C (capacity), L (length), B (breadth), and H (height), while on a broken skull only a subset of L, B, and H and *not* C could be measured. The observed data, thus, consisted of samples from a four variate population with several observations missing. There were some sets with all the four measurements C, L, B, H, and some with one or two or three of the measurements L, B, and H only. The problem was to estimate the mean values of C, L, B, and H in the *original* population of skulls from the recovered fragmentary samples. In a number of papers which appeared in the early issues of *Biometrika*, it was the practice to estimate the unknown population mean value of any characteristic, say C, by taking the mean of all the available measurements on C. An alternative to this, which is often recommended, is to compute maximum likelihood estimates of the unknown mean values, variances, and covariances by writing down the likelihood function based on all the available data assuming a four variate normal distribution for C, L, B, and H and using the derived marginal distribution for an incomplete set of measurements. This is based on the *assumption* that each skull admitting all the four measurements or any subset of the four can be considered as a random sample from the *original* population of skulls. Is this assumption valid?

It is common knowledge that a certain proportion of the original skulls get broken, depending on the length of time and depth at which they lay buried. Let $w(c)$ be the probability that a skull of capacity c is not broken, and $p(c, \theta)$ be the pdf of C in the original population. Then the pdf of C measured on well-preserved skulls is

$$\frac{w(c)p(c, \theta)}{E[w(C)]}. \tag{4.1}$$

If $w(c)$ depends on c, then the *observed* measurements on C cannot be considered as a random sample of C from the *original* population. Further, if $w(c)$ is a decreasing function of c, then there will be a larger representation of small skulls among the unbroken skulls, and therefore the mean of the available measurements on C will be an underestimate of the mean capacity of the original population.

Is there any evidence that $w(c)$ depends on c? To answer this question, the regression of C on L, B, and H (in terms of logarithms) was estimated from the data sets where all the four measurements were available and used to predict the mean capacity of broken skulls by substituting the observed averges $\bar{L}$, $\bar{B}$, and $\bar{H}$ of broken skulls in the regression equation. At least in two series of cranial measurements (see Rao and Shaw, 1948; Rao, 1973, p. 280), it was found that the average measured capacity of unbroken skulls was smaller than the estimated average capacity of broken skulls. This provided some evidence about the differential preservation of skulls, with smaller skulls having a higher chance of remaining unbroken.

This finding invalidates the assumption that skulls providing all four measurements constitute a random sample from the original population of skulls. The pdf associated with these measurements is more appropriately (4.1), which is a weighted version of the original pdf with an unknown weight function. Presumably, the pdf associated with observations on any subset of L, B, and H will again be a weighted pdf with a weight function depending on the degree of damage to a skull. The expression for the correct likelihood will then depend on the original pdf and the probabilities of different degrees of damage as assessed by subsets of measurements that can be taken on a skull, which are likely to be unknown. Is there a reasonable solution to the problem of estimation of mean values in a situation like the above?

There are several possibilities, of which the following procedure for estimating the mean of C appears to be a natural one. We use the complete sets of measurements, C, L, B, and H, on unbroken skulls to compute the regressions of C on different subsets of L, B, and H. Using the appropriate regression function, we estimate (predict) the missing value of C for each broken skull. Then an average is taken of all the measured and estimated values of C. Such an average is likely to be a valid estimate of the mean of C. The estimation is based on the assumption that the complete sets of measurements (C, L, B, H) can provide valid estimates of relationships like the regression functions of C on L, B, H and its subsets, although they are biased samples from the original population. Similar methods can be used to estimate the mean values of L, B, and H.

Paleontologists compare the characteristics of fossils of long bones and cranial material discovered in different parts of the world to trace the evolutionary history of hominids. Such studies based on physical measurements may be misleading, as the discovered fossils may not be representative samples from the original populations due to differential preservation of skeletal material. It is gratifying to note that attempts are being made to compare the fossils in terms of some basic chemical measurements which are not likely to be subject to the phenomenon of differential preservation.

5. Enquiry Through an Offspring

In genetic and sociopsychological studies it is the common practice to locate an abnormal individual and through him or her collect information on the status of brothers and sisters, parents, uncles, and aunts. From such data estimates are made of the incidence of abnormality in families by sex and parity of birth. A family is the basic unit whose characteristics may have a specified distribution. But our method of ascertainment gives unequal probabilities to families depending on the mechanism inherent in the selection of an abnormal family member. Thus, the distribution applicable to observed data on families is a weighted version of the distribution specified for the families. We consider some examples, discuss the nature of the problems involved in each case, and suggest possible solutions.

5.1. Too Many Males?

During the last few years, while lecturing to students and teachers in different parts of the world, I collected data on the numbers of brothers and sisters in the family of each individual in the audience. The results are summarized in Tables 2, 3, and 4. The data from the male respondents given in Tables 2 and 4 show that the ratio of B, the total number of brothers including the respondents, to $B+S$, the total number of brothers and sisters, is much larger than one-half in each case, indicating a preponderence of male children in the families of male members of my audiences.

The number of male children in a family of a given size has a binomial distribution, and this would have been the specification if the sib compositions had been ascertained from families selected at random from the population of families.

Table 2. Data on Male Respondents (Students)[a]

Place and year	k	B	S	$\frac{B}{B+S}$	$\frac{B-k}{B+S-k}$ [b]	χ^2
Bangalore (India, 75)	55	180	127	.586	.496	.02
Delhi (India, 75)	29	92	66	.582	.490	.07
Calcutta (India, 63)	104	414	312	.570	.498	.04
Waltair (India, 69)	39	123	88	.583	.491	.09
Ahmedabad (India, 75)	29	84	49	.632	.523	.35
Tirupati (India, 75)	592	1902	1274	.599	.484	.50
Poona (India, 75)	47	125	65	.658	.545	1.18
Hyderabad (India, 74)	25	72	53	.576	.470	.36
Tehran (Iran, 75)	21	65	40	.619	.500	.19
Isphahan (Iran, 75)	11	45	32	.584	.515	.06
Tokyo (Japan, 75)	50	90	34	.725	.540	.49
Lima (Peru, 82)	38	132	87	.603	.519	.27
Shanghai (China, 82)	74	193	132	.594	.474	.67
Columbus (USA, 75)	29	65	52	.556	.409	2.91
College St. (USA, 76)	63	152	90	.628	.497	.01
Total	1206	3734	2501	.600	.503	0.14

[a] k = number of students, B = total number of brothers including the respondent, S = total number of sisters.

[b] Estimate of π under size biased binomial distribution.

Table 3. Data on Female Respondents (Students)

Place and year	k	B	S	$\frac{S}{B+S}$	$\frac{S-k}{B+S-k}$	χ^2
Lima (Peru, 82)	16	37	48	.565	.464	.36
Los Banos (Philippines, 83)	44	101	139	.579	.485	.18
Manila (Philippines, 83)	84	197	281	.588	.500	.00
Bilbao (Spain, 83)	14	19	35	.576	.525	.10
Total	158	354	503	.587	.493	.11

Table 4. Data on Male Respondents (Professors)

Place and year	k	B	S	$\frac{B}{B+S}$	$\frac{B-k}{B+S-k}$	χ^2
State College (USA, 75)	28	80	37	.690	.584	2.53
Warsaw (Poland, 75)	18	41	21	.660	.525	2.52
Poznan (Poland, 75)	24	50	17	.746	.567	1.88
Pittsburgh (USA, 81)	69	169	77	.687	.565	2.99
Tirupati (India, 76)	50	172	132	.566	.480	0.39
Maracaibo (Venezuela, 82)	24	95	56	.629	.559	1.77
Richmond (USA, 81)	26	57	29	.663	.517	0.03
Total	239	664	369	.642	.535	3.95

In the case of the data reported in Tables 2 and 4, a male student is located first and the sib composition in his family is ascertained; in such a case, each family included in the sample has at least one male child, which indicates a departure from the full binomial distribution. What population then does our sample represent? It is clear that the effective population is the *subset* of families having a male child of a particular description, such as a specified age group and qualifications which gave him a chance of being included in the enquiry. Rao (1977) argued that the distribution of brothers and sisters in such a subset of families of a given size is likely to be size biased binomial, so that the probability of r brothers and $n - r$ sisters in a family of size n is the weighted binomial

$$\frac{r}{E(r)}\binom{n}{r}\pi^r(1-\pi)^{n-r} = \binom{n-1}{r-1}\pi^{r-1}(1-\pi)^{n-r}, \tag{5.1.1}$$

where π is the probability of a male child. Under this hypothesis we find that

$$E\left(\frac{B-k}{B+S-k}\right) = \pi, \tag{5.1.2}$$

where k is the number of male respondents, so that $(B-k)/(B+S-k)$ is an estimate of π, and

$$\frac{[B - k - (B + S - k)\pi]^2}{(B + S - k)\pi(1 - \pi)} \tag{5.1.3}$$

has an asymptotic chi-square distribution on 1 degree of freedom. Similar results hold for the data from female respondents in Table 3. It is seen from the chi-square values in Tables 2 and 3 that the data collected from the students are consistent with the hypothesis of a size biased binomial with $\pi = \frac{1}{2}$. (Actually the chi-squares are too small. This needs an investigation).

The situation is somewhat different in Table 4, relating to data from the professors. The estimated π is more than one-half in each case, and the chi-square values are high. This implies that the weight function appropriate for these data is of a higher order than r, the number of brothers. A possible sociological explanation for this is that a person coming from a family with a larger number of brothers tends to acquire better academic qualifications to compete for jobs.

The following example on observed sex ratio is illuminating. In a survey of fertility and mortality, Dandekar and Dandekar (1953) gave the distribution of brothers (excluding the informant), sisters, sons, and daughters as reported by 1115 "male heads," contacted through households chosen with equal probability for each household. It may be observed that in a survey of this type, a family with r brothers gets a chance nearly proportional to r, and the conditions for a weighted binomial with $w(r) = r$ hold for the number of brothers in a family. Yet we find from Table 5 that the total number of brothers, 1325 (excluding the informants) is far in excess of the number of sisters, 1014, giving

$$\chi^2 = \frac{(1325 - 1014)^2}{1325 + 1014} = 41.35,$$

which is very high on 1 degree of freedom. Is the theory of the size biased binomial wrong?

Table 5. Distribution by Age of Brothers, Sisters, Sons, and Daughters[a]

Age group	Brothers	Sisters	Sons	Daughters
0–4	5	10	357	348
5–9	27	31	330	354
10–14	63	62	305	226
15–19	87	85	208	190
20–24	155	100	167	130
25–29	181	130	85	63
30–34	156	130	29	33
35–39	179	123	18	16
40–44	146	105	13	5
Rest	336	228	21	10
Total	1325	1014	1533	1375

[a] Dandekar and Dandekar (1953).

It is clear from Table 5 that the disproportionate sex ratio is confined to the age groups above 15–19 years, and the same phenomenon seems to occur in the case of sons and daughters. There is perhaps an underreporting of sisters and daughters who are married off, due to a superstitious custom of not including them as members of the household. Underreporting of female members is a persistent feature in data on fertility and mortality collected in developing countries.

5.2. Albinism

In studies of the inheritance of rare diseases, it is more convenient to collect family data by first locating an affected individual and then enquiring about the status of each of his or her brothers and sisters. While the different categories of children classified by disease, sex, etc., may have a multinomial distribution among families of given size, the numbers as ascertained do not have the same distribution, due to unequal probabilities of selection of families. In the previous section we have encountered a situation where the probability of selection of a family was proportional to the number of male children. In the case of rare diseases, the probability of selection of a family through an affected child may be not a linear but a more complicated function of the number of affected children. In this section we propose a model for selection probabilities in such cases and develop the appropriate methodology.

Let π_1 be the probability that a male child is an albino, and π_2 that a female child is an albino. Then the probability that a family of n children has r_1 males of whom t_1 are albinos and r_2 females of whom t_2 are albinos is

$$p(r_1, t_1; r_2, t_2) = \binom{n}{r_1}\left(\frac{1}{2}\right)^n \binom{r_1}{t_1}\pi_1^{t_1}\phi_1^{r_1-t_1}\binom{r_2}{t_2}\pi_2^{t_2}\phi_2^{r_2-t_2}, \qquad (5.2.1)$$

where $\phi_1 = 1 - \pi_1$ and $\phi_2 = 1 - \pi_2$, and the probability of a child being a male or a female is taken as one-half.

There are a number of ways in which we can introduce probabilities of selection of affected families. We consider some models which are extensions of those suggested by Fisher (1934) and Haldane (1938).

Introducing α and $\beta = 1 - \alpha$ as relative probabilities of observing a male or a female albino, we may consider a mixture of two size biased distributions,

$$p^w(r_1, t_1; r_2, t_2) = \left(\frac{2\alpha t_1}{n\pi_1} + \frac{2\beta t_2}{n\pi_2}\right)p(r_1, t_1; r_2, t_2), \qquad (5.2.2)$$

as the appropriate distribution of the observed vector (r_1, t_1, r_2, t_2). If we have data on (r_1, t_1, r_2, t_2) from N ascertained families, we can write down the likelihood using the expression (5.2.2) and estimate the parameters α, π_1, and π_2. Alternatively, we can use the method of moments, using the statistics $\sum t_1$, $\sum t_2$, and $\sum r_1$ to estimate the unknown parameters.

If $\pi_1 = \pi_2 = \pi$, the expression (5.2.2) reduces to

$$\frac{2}{n\pi}(\alpha t_1 + \beta t_2) p(r_1, t_1; r_2, t_2), \tag{5.2.3}$$

and the estimates of α and π can be obtained from the equations

$$\begin{aligned} \bar{t}_1 &= \alpha + \frac{\pi}{2k}\sum(n_i - 1), \\ \bar{t}_2 &= \beta + \frac{\pi}{2k}\sum(n_i - 1), \end{aligned} \tag{5.2.4}$$

where k is the number of families, n_i is the number of children in the ith family, and $\bar{t}_1$ and $\bar{t}_2$ are the average numbers of male and female albino children in a family.

Another model is as follows. Let ρ_1 and ρ_2 be the probabilities of observing a male and a female albino respectively. Then the probability that a family with n children having t_1 male albinos and $r_1 - t_1$ normal males, and t_2 female albinos and $r_2 - t_2$ normal females, is investigated s_1 times by observing a male albino and s_2 times by observing a female albino is

$$\binom{t_1}{s_1}\rho_1^{s_1}(1-\rho_1)^{t_1-s_1}\binom{t_2}{s_2}\rho_2^{t_2}(1-\rho_2)^{t_2-s_2}p(r_1, t_1; r_2 t_2). \tag{5.2.5}$$

Since a family is not investigated unless at least one of t_1 and t_2 is different from zero, the effective distribution for the observed data is (5.2.5) normalized by the quotient $1 - (1 - \rho)^n$, where $\rho = (\rho_1\pi_1 + \rho_2\pi_2)/2$. The method of estimation of ρ_1, ρ_2, π_1, and π_2 when we have the additional information on the number of times each family is investigated is discussed in detail in Rao (1965).

In case a family is investigated only once although more than one abnormal child in the family is observed, the appropriate distribution is

$$\frac{[1 - (1-\rho_1)^{t_1}(1-\rho_2)^{t_2}]p(r_1, t_1; r_2, t_2)}{1 - (1-\rho)^n}, \tag{5.2.6}$$

where $\rho = (\pi_1\rho_1 + \pi_2\rho_2)/2$. If $\rho_1 = \rho_2 = \rho$ and $\pi_1 = \pi_2 = \pi$, then the expression (5.2.6) reduces to

$$\frac{1 - (1-\rho)^{t_1+t_2}}{1 - (1-\pi\rho)^n}\frac{n!}{t_1!(r_1 - t_1)!t_2!(r_2 - t_2)!}\left(\frac{\pi}{2}\right)^{t_1+t_2}\left(\frac{\phi}{2}\right)^{n-t_1-t_2}. \tag{5.2.7}$$

If sex is ignored, then(5.2.7) becomes

$$\frac{1 - (1-\rho)^t}{1 - (1-\pi\rho)^n}\frac{n!}{t!(n-t)!}\pi^t\phi^{n-t}, \tag{5.2.8}$$

where $t = t_1 + t_2$, which is the expression used by Haldane (1938).

We have considered three different models (5.2.2), (5.2.5), and (5.2.6) for the probability of selection of a family. In the case where we have information only on the number r of abnormal children in a family of size n without any sex distinction, we may consider the weighted binomial distribution

$$\frac{w(r)}{E[w(r)]}\binom{n}{r}\pi^r\phi^{n-r}, \tag{5.2.9}$$

where $\phi = 1 - \pi$, with three possible alternatives for $w(r)$:

$$w(r) = r \tag{5.2.10}$$

$$= r^\alpha \qquad (\alpha \text{ unknown}) \tag{5.2.11}$$

$$= 1 - (1 - \rho)^n \qquad (\rho \text{ unknown}). \tag{5.2.12}$$

The maximum likelihood method of estimating α and π under the model (5.2.9), (5.2.11) is discussed in Rao (1965), and of ρ and π under the model (5.2.9), (5.2.12) in Haldane (1938). To demonstrate the relevance of the weight function (5.2.11), we compare in Table 6 the observed data on frequencies of albino children in families of different sizes with the expected values under the two different weight functions $w(r) = r$ and $w(r) = r^{1/2}$, choosing $\pi = \frac{1}{4}$. It is seen that the weight function $w(r) = r^{1/2}$ provides a better fit.

For a general discussion of the type of problems discussed in this section, and a few other models for selection probabilities, the reader is referred to Stene (1981) and other references mentioned in that paper. For estimation of α and π in the model (5.2.9), (5.2.11), reference may be made to Rao (1965).

5.3. Alcoholism, Family Size, and Birth Order

Smart (1963, 1964) and Sprott (1964) examined a number of hypotheses on the incidence of alcoholism in Canadian families using the data on family size and birth order of 242 alcoholics admitted to three alcoholism clinics in Ontario. The method of sampling is thus of the type discussed in Sections 5.1 and 5.2.

One of the hypotheses tested was that "larger families contain larger numbers of alcoholics than expected." The null hypothesis was interpreted to imply that the observations on family size as ascertained arise from the weighted distribution

$$np(n)/E(n), \qquad n = 1, 2, \ldots, \tag{5.3.1}$$

where $p(n)$, $n = 1, 2, \ldots$, is the distribution of family size in the general population. Smart and Sprott used the distribution of family size as reported in the 1931 census of Ontario for $p(n)$ in their analysis. It is then a simple matter to test whether the observed distribution of family size in their study is in accordance with the expected distribution (5.3.1).

It may be noted that the distribution (5.3.1) would be appropriate if we had chosen individuals (alcoholic or not) at random from the general population (of individuals) and ascertained the sizes of the families to which they belonged. But it is not clear whether the same distribution (5.3.1) holds if the

Table 6. Observed and Expected Frequencies of Albino Children for Each Family Size n

No. of albinos	$n=2$ Observed	Expected[a] (1)	Expected[a] (2)	$n=3$ Observed	Expected[a] (1)	Expected[a] (2)	$n=4$ Observed	Expected[a] (1)	Expected[a] (2)	$n=5$ Observed	Expected[a] (1)	Expected[a] (2)
1	31	30.00	32.37	37	30.93	35.81	22	21.10	26.07	25	19.00	24.93
2	9	10.00	7.63	15	20.63	16.88	21	21.09	18.43	23	25.31	23.50
3				3	3.44	2.30	7	7.03	5.02	10	12.65	9.59
4							0	0.78	0.48	1	2.81	1.85
5										1	0.23	0.13

No. of albinos	$n=6$ Observed	Expected[a] (1)	Expected[a] (2)	$n=7$ Observed	Expected[a] (1)	Expected[a] (2)	Total Observed	Expected[a] (1)	Expected[a] (2)
1	18	12.58	17.46	16	8.21	11.98	149	121.82	148.62
2	13	20.96	20.58	10	16.37	16.94	96	114.36	103.98
3	18	13.98	11.20	14	13.64	11.53	47	50.74	39.64
4	3	4.66	3.23	5	6.06	4.43	9	14.31	10.00
5	0	0.77	0.48	1	1.51	0.99	1	2.51	1.61
6	1	0.05	0.03	0	0.20	0.12	1	0.25	0.15
				0	0.01	0.01	0	0.01	0.01

[a] (1) for $w_r = r$; (2) for $w_r = r^{1/2}$.

enquiry is restricted to alcoholic individuals admitted to a clinic, as assumed by Smart and Sprott. This could happen, as demonstrated below, under an interpretation of their null hypothesis that the number of alcoholics in a family has a binomial distribution (like failures in a sequence of independent trials), and a further assumption that every alcoholic has the same independent chance of being admitted to a clinic.

Let π be the probability of an individual becoming an alcoholic, and suppose that the probability that a member of a family becomes an alcoholic is independent of whether another member is alcoholic or not. Further let $p(n)$, $n = 1, 2, \ldots$, be the probability distribution of family size (whether a family has an alcoholic or not) in the general population. Then the probability that a family is of size n and has r alcoholics is

$$p(n)\binom{n}{r}\pi^r\phi^{n-r}, \qquad r = 0, \ldots, n; \qquad n = 1, 2, \ldots, \tag{5.3.2}$$

where $\phi = (1 - \pi)$. From (5.3.2), it follows that the distribution of family size in the general population, given that a family has at least one alcoholic, is

$$\frac{(1 - \phi^n)p(n)}{1 - E(\phi^n)}, \qquad n = 1, 2, \ldots. \tag{5.3.3}$$

If we had chosen households at random and recorded the family sizes in households containing at least one alcoholic, then the null hypothesis on the excess of alcoholics in larger families could be tested by comparing the observed frequencies with the expected frequencies under the model (5.3.3). However, under the sampling scheme adopted, the weighted distribution of (n, r),

$$p^w(n, r) = rp(n)\binom{n}{r}\frac{\pi^r\phi^{n-r}}{\pi E(n)}, \tag{5.3.4}$$

is more appropriate. If we had information on the family size n as well as on the number of alcoholics (r) in the family, we could have compared the observed joint frequencies of (n, r) with those expected under the model (5.3.4).

From (5.3.4), the marginal distribution of n alone is

$$np(n)/E(n), \qquad n = 1, 2, \ldots, \tag{5.3.5}$$

which is used by Smart and Sprott as a model for the observed frequencies of family sizes. It is shown in (5.3.3) that in the general population, the distribution of family size in families with at least one alcoholic is

$$\frac{(1 - \phi^n)p(n)}{1 - E(\phi^n)},$$

which reduces to (5.3.5) if ϕ is close to unity. In other words, if the probability of an individual becoming an alcoholic is small, then the distribution of family size as ascertained is close to the distribution of family size in families with at least one alcoholic in the general population. This is not true if ϕ is not close to unity.

Smart and Sprott found that the distribution (5.3.5) did not fit the observed frequencies, which had heavier tails. What conclusions can we draw from this test? It is seen that the weighted distribution (5.3.5) is derived under two hypotheses. One is that the distribution of family size in the subset of families having at least one alcoholic in the general population is of the form (5.3.3) which is implied by the original null hypothesis posed by Smart. The other is that the method of ascertainment is equivalent to pps sampling of families, with probability proportional to the number of alcoholics in a family. The rejection of (5.3.5) would imply the rejection of the first of these two hypotheses if the second is assumed to be correct. There are no *a priori* grounds for such an assumption, and in the absence of an objective test for this, some caution is needed in accepting Smart's conclusions.

An alternative to (5.3.4) is obtained by assuming that each alcoholic has a chance θ of being admitted to a clinic independently of other alcoholic family members. In such a case, the probability that a family of size n has r alcoholics and a member has been admitted to a clinic is

$$p(n)\binom{n}{r}\pi^r\phi^{n-r}(1-(1-\theta)^r). \tag{5.3.6}$$

The marginal distribution of n with the normalizing factor is then

$$p(n)\frac{1-(1-\pi\theta)^n}{E(1-(1-\pi\theta)^n)}, \qquad n = 1, 2, \ldots. \tag{5.3.7}$$

The distribution (5.3.7) involves one unknown parameter $\pi\theta$ which needs to be estimated in fitting to the observed frequencies of family sizes. Some examples of distributions of the type (5.3.7) have been considered by Barrai, Mi, Morton, and Yasuda (1965). The distribution (5.3.7) is close to (5.3.5) if $\pi\theta$ is small.

We may also consider a more complicated model by assuming different probabilities π_1 and π_2 respectively for a male and a female becoming alcoholic, and also different probabilities θ_1 and θ_2 for male and female alcoholics being referred to a clinic. In such a case, the probability of inclusion of a family of size n with r_1 males, s_1 male alcoholics, r_2 females, and s_2 female alcoholics is

$$p(n)\binom{n}{r}\left(\frac{1}{2}\right)^n\binom{r_1}{s_1}\pi_1^{s_1}\phi_1^{r_1-s_1}\binom{r_2}{s_2}\pi_2^{r_2}\phi_2^{r_2-s_2}(1-(1-\theta_1)^{s_1}(1-\theta_2)^{s_2}), \tag{5.3.8}$$

where $\phi_1 = 1 - \pi_1$ and $\phi_2 = 1 - \pi_2$. This gives the marginal distribution of n as

$$p(n)\frac{1-2^{-n}(2-\pi_1\theta_1-\pi_2\theta_2)^n}{E(1-2^{-n}(2-\pi_1\theta_1-\pi_2\theta_2)^n)}, \tag{5.3.9}$$

which again involves one unknown parameter, $(\pi_1\theta_1 + \pi_2\theta_2)/2$. The marginal distribution of r_1 and r_2 obtained from (5.3.8) is

$$p(n)\binom{n}{r_1}\left(\frac{1}{2}\right)^n \frac{1-(1-\pi_1\theta_1)^{r_1}(1-\pi_2\theta_2)^{r_2}}{E(1-2^{-n}(2-\pi_1\theta_1-\pi_2\theta_2)^n)}, \tag{5.3.10}$$

where $n = r_1 + r_2$. If $\pi_1\theta_1$ and $\pi_2\theta_2$ are small, then (5.3.10) becomes

$$p(n)\binom{n}{r_1}\left(\frac{1}{2}\right)^n \frac{r_1\pi_1\theta_1 + r_2\pi_2\theta_2}{2^{-1}(\pi_1\theta_1+\pi_2\theta_2)E(n)}. \tag{5.3.11}$$

If we had the joint frequencies of males and females in the observed families of alcoholics, we could have fitted distributions of the type (5.3.10) and (5.3.11) to test the null hypothesis of larger numbers of alcoholics in larger families.

It is seen from (5.3.10) and (5.3.11) that the distribution of (r_1, r_2) will not be symmetric unless $\pi_1\theta_1 = \pi_2\theta_2$. This may result in an excess of males or females in observed families. Such an effect (with an excess of males) can be seen in similar data studied by Freire-Mala and Chakraborty (1975) and Rao, Mazumdar, Waller, and Li (1973); these authors have not, however, commented on this phenomenon.

Another hypothesis considered by Smart was that the later-born children have a greater tendency to become alcoholic than the earlier-born. The method used by Smart may be somewhat confusing to statisticians. Some comments were made by Sprott criticizing Smart's approach. We shall review Smart's analysis in the light of the model (5.3.4). If we assume that birth order has no relationship to becoming an alcoholic, and the probability of an alcoholic being referred to a clinic is independent of the birth order, then the probability that an observed alcoholic belongs to a family with n children and r alcoholics and has given birth order $s \le n$ is, using the model (5.3.4),

$$\frac{rp(n)}{nE(n)}\binom{n}{r}\pi^{r-1}\phi^{n-r}, \qquad s = 1, \dots, n, \quad r = 1, \dots, n, \quad n = 1, 2, \dots. \tag{5.3.12}$$

Summing over r, we find that the marginal distribution of (n, s), the family size and birth order, applicable to the observed distribution, is

$$p(n)/E(n), \qquad s = 1, \dots, n, \quad n = 1, 2, \dots, \tag{5.3.13}$$

where it may be recalled that $p(n)$, $n = 1, 2, \dots$, is the distribution of family size in the general population. Smart gave the observed bivariate frequencies of (n, s), and since $p(n)$ was known, the expected values could have been computed and compared with the observed. But, he did something else.

From (5.3.13), the marginal distribution of birth rank is

$$\frac{1}{E(n)}\sum_{i=r}^{\infty} p(i), \qquad r = 1, 2, \dots. \tag{5.3.14}$$

Smart's (1963) analysis in his Table 2 is an attempt to compare the observed distribution of birth ranks with the expected under the model (5.3.14) with $p(i)$ itself estimated from data using the model (5.3.1).

A better method is as follows: from (5.3.13) it is seen that for given family size, the expected birth order frequencies are equal as computed by Smart (1963) in Table 1, in which case individual chi-squares comparing the expected and observed frequencies for each family size would provide all the information about the hypothesis under test. Such a procedure would be independent of any knowledge of $p(n)$. But it is not clear whether a hypothesis of the type posed by Smart can be tested on the basis of the available data without further information on the other alcoholics in the family, such as their ages, sexes, etc.

Table 7 reproduces a portion of Table 1 in Smart (1963) relating to families up to size 4 and birth ranks up to 4. It is seen that for family sizes 2 and 3, the observed frequencies seem to contradict the hypothesis, and for family sizes above 3 (see Smart's Table 1), birth rank does not have any effect. It is interesting to compare the above data with a similar type of data (Table 8) collected by the author on birth rank and family size of the staff members in two departments at the University of Pittsburgh. It appears that there are too many earlier-borns among the staff members, indicating that becoming a professor is an affliction of the earlier born! It is expected that in data of the kind we are considering there will be an excess of the earlier born without implying an implicit relationship between birth order and a particular attribute, especially when it is age dependent.

Table 7. Distribution of Birth Rank s and Family Size n[a]

	$n = 1$		2		3		4	
s	O	E	O	E	O	E	O	E
1	21	21	22	16	17	13.3	11	11.75
2			10	16	14	13.3	10	11.75
3					9	13.3	13	11.75
4							13	11.75

[a] Smart (1963, Table 1). O = observed, E = expected.

Table 8. Distribution of Birth Rank s and Family Size $n \leq 4$ Among Staff Members (University of Pittsburgh)

s	$n = 1$	2	3	4
1	7	14	9	6
2		6	4	2
3			2	0
4				0

6. Quadrat Sampling with Visibility Bias

For estimating wildlife population density,quadrat sampling has been found generally convenient. Quadrat sampling is carried out by first selecting at random a number of quadrats of fixed size from the region under study and ascertaining the number of animals in each. The following assumptions are made:

A$_1$. Animals are found in groups within each quadrat, and the number of animals X in a group follows a specified distribution.

A$_2$. The number of groups N within a quadrat has a specified distribution.

A$_3$. The number of groups within a quadrat and the numbers of animals within groups are independent.

Let the method of sampling be such that the probability of sighting (or recording) a group of x animals is $w(x)$. If X^w and N^w represent the rv's of the number of animals in a group and number of groups within a quadrat as ascertained, then we have the following results:

(i)

$$P(N^w = m \mid N = n) = \binom{n}{m} w^m (1 - w)^{n-m}, \tag{6.1}$$

where

$$w = \sum_1^\infty w(x)\, P(X = x) \tag{6.2}$$

is the visibility factor (or the probability of recording a group).

(ii)

$$P(N^w = m) = \sum_{n=m}^\infty \binom{n}{m} w^m (1 - w)^{n-m} P(N = n). \tag{6.3}$$

(iii)

$$P(N^w = m, X_1^w = x_1, \ldots, X_m^w = x_m) = w^{-m} P(N^w = m) \prod_{i=1}^m w(x_i)\, P(X = x_i). \tag{6.4}$$

(iv) Let $S^w = X_1^w + \cdots + X_m^w$. Then

$$P(S^w = y) = \sum_{m=1}^\infty P(N^w = m) P(S^w = y \mid m) \tag{6.5}$$

and

$$P(S^w = y \mid m) = \sum_{\Sigma x_i = y}^{\infty} \frac{w(x_1)}{w} \cdots \frac{w(x_m)}{w} P(X_1 = x_1) \cdots P(X = x_m). \tag{6.6}$$

The formulae listed above are useful in many practical situations. Usually the sighting probability is of the form

$$w(x) = 1 - (1 - \beta)^x. \tag{6.7}$$

For some applications, the reader is referred to papers by Cook and Martin (1974) and Patil and Rao (1977, 1978).

7. Waiting Time Paradox

Patil (1984) reported a study conducted in 1966 by the Institute National de la Statistique et de l'Economie Appliquee in Morocco to estimate the mean sojourn time of tourists. Two types of surveys were conducted, one by contacting tourists residing in hotels and another by contacting tourists at frontier stations while leaving the country. The mean sojourn time as reported by 3000 tourists in hotels was 17.8 days, and by 12321 tourists at frontier stations was 9.0. Suspected by the officials in the department of planning, the estimate from the hotels was discarded.

It is clear that the observations collected from tourists while leaving the country correspond to the true distribution of sojourn time, so that the observed average 9.0 is a valid estimate of the mean sojourn time. It can be shown that in a steady state of flow of tourists, the sojourn time as reported by those contacted at hotels has a size biased distribution, so that the observed average will be an overestimate of the mean sojourn time. If X^w is a size biased random variable, then

$$E(X^w)^{-1} = \mu^{-1} \tag{7.1}$$

where μ is the expected value of X, the original variable. The formula (7.1) shows that the harmonic mean of the size biased observations is a valid estimate of μ. Thus the harmonic mean of the observations from the tourists in hotels would have provided an estimate comparable with the arithmetic mean of the observations from the tourists at the frontier stations.

It is interesting to note that the estimate from hotel residents is nearly twice the other, a factor which occurs in the waiting time paradox (see Feller, 1966; Patil and Rao, 1977) associated with the exponential distribution. This suggests, but does not confirm, that the sojourn time distribution may be exponential.

Suppose that the tourists at hotels were asked how long they had been staying in the country up to the time of enquiry. In such a case, we may assume that the pdf of the rv Y, the time a tourist has been in a country up to the time of enquiry, is the same as that of the product $X^w R$, where X^w is the size biased version of X, the sojourn time, and R is an independent rv with a uniform distribution on [0, 1]. If $F(x)$ is the distribution function of X, then the pdf of Y is

$$\mu^{-1}[1 - F(y)]. \tag{7.2}$$

The parameter μ can be estimated on the basis of observations on Y, provided the functional form of $F(y)$, the distribution function of the sojourn time, is known.

It is interesting to note that the pdf (7.2) is the same as that obtained by Cox (1962) in studying the distribution of failure times of a component used in different machines from observations on the ages of the components in use at the time of investigation.

8. Damage Models

Let N be a rv with probability distribution, p_n, $n = 1, 2, \ldots$, and R be a rv such that

$$P(R = r \mid N = n) = s(r, n). \tag{8.1}$$

Then the marginal distribution of R truncated at zero is

$$p'_r = (1 - p)^{-1} \sum_{n=r}^{\infty} p_n s(r, n), \qquad r = 1, 2, \ldots, \tag{8.2}$$

where

$$p = \sum_{1}^{\infty} p_i s(0, i). \tag{8.3}$$

The observation r represents the number surviving when the original observation n is subject to a destructive process which reduces n to r with probability $s(r, n)$. Such a situation arises when we consider observations on family size counting only the surviving children (R). The problem is to determine the distribution of N, the original family size, knowing the distribution of R and assuming a suitable survival distribution.

Suppose that $N \sim P(\lambda)$, i.e., distributed as Poisson with parameter λ, and let $R \sim B(\cdot, \pi)$, i.e., binomial with parameter π. Then

$$p'_r = e^{-\lambda\pi} \frac{(\lambda\pi)^r}{r!} \frac{1}{1 - e^{-\lambda\pi}}, \qquad r = 1, 2, \ldots. \tag{8.4}$$

It is seen that the parameters λ and π get confounded, so that knowing the distribution of R, we cannot find the distribution of N. Similar confounding occurs when N follows a binomial, negative binomial, or logarithm series distribution. When the survival distribution is binomial, Sprott (1965) gives a general class of distributions which has this property. What additional information is needed to recover the original distribution? For instance, if we know which of the observations in the sample did not suffer damage, then it is possible to estimate the original distribution as well as the binomial parameter π.

It is interesting to note that observations which do not suffer any damage have the distribution

$$p_r^u = cp_r\pi^r, \tag{8.5}$$

which is a weighted distribution. If the original distribution is Poisson, then

$$p_r^u = e^{-\lambda\pi}\frac{(\lambda\pi)^r}{r!}\frac{1}{1-e^{-\lambda\pi}}, \tag{8.6}$$

which is same as (8.4). It is shown in Rao and Rubin (1964) that the equality $p_r^u = p_r'$ characterizes the Poisson distribution.

The damage models of the type described above were introduced in Rao (1965). For theoretical developments on damage models and characterization of probability distributions arising out of their study, the reader is referred to Alzaid, Rao, and Shanbhag (1984).

9. Nonresponse: The Story of an Extinct River

Sample survey practitioners define nonresponse as a missing observation or nonavailability of measurements on a unit included in a sample. It is clear that if the missing values can be considered as a random sample from the population under survey, then the observed values constitute a representative sample of the whole population (Rubin, 1976). Usually this is not the case, and special techniques are developed in sample surveys to cope with such situations.

In general, nonresponse poses serious issues, such as the problem of broken skulls not providing direct measurements on capacity (see Section 4 of this paper). More complex cases are as follows.

For instance, we may try to estimate the underground resources in a given region by making borings at a randomly chosen set of points and taking some measurements. But it may so happen that borings cannot be made at some chosen points, for example because of the presence of rocks. The measurements at such points may be of a different type from the rest, in which case the observed sample will not be a representative sample from the whole region.

Such a problem arose in an investigation by geologists at the Indian Statistical Institute to estimate the mean direction of flow of an extinct river of geological times in a certain region (see Sengupta, 1966; J. S. Rao and Sengupta, 1966). The geologists collected a series of observations on direction cosines of flow (two dimensional vector data), which seemed ideal for an application of Fisher's (1953) distribution and the associated theory for estimation of the mean direction of flow. Then the question arose as to what the hypothetical population was from which the observations could be considered as a random sample. The measurements on direction cosines could not be made at any chosen point, but only at certain points where there were outcrops. The geologists walked along the region under exploration and made measurements wherever they came across outcrops. If the outcrops had been uniformly distributed over space, then it might have been possible to define a population of which the observations made by the geologists could be a representative sample. The locations at which observations were made, when plotted on a topographical map of the region, showed an unequal distribution of outcrops in different areas of the region, indicating the nonrandom nature of the occurrence of outcrops. In such a case the estimate of mean direction assuming that each observation is an independent sample with a common expectation will be biased. In order to minimize the bias in estimation, the following method of estimation was adopted. A square lattice was imposed on the topographical map, and the measurements in each grid were replaced by their average. Then a simple average of these averages was taken as an estimate of the mean direction of flow. This estimate differed somewhat from the average of all the measurements and was considered to have less bias.

This study points out the need for a reexamination of the data on directions of rock magnetism collected by geologists and analysed by Fisher (1953), who developed a special theory for that purpose. If the outcrops at which measurements of direction are possible are not uniformly distributed over space, then there will be some difficulty in interpreting the observed mean direction as an estimate of some specific parameter.

10. Conclusions

Some of the broad conclusions that emerge from the discussion of the live examples in the paper are as follows:

Specification, or the choce of a model, is of great importance in data analysis. An appropriate specification for given data can be arrived at on the basis of past experience, information on the stochastic nature of events, a detailed knowledge of how observations are ascertained and recorded, and an exploratory analysis of current data itself using graphical displays, preliminary tests, and cross validation studies.

Inaccuracies in specification can lead to wrong inference. It is therefore worthwhile to review the data under different possible specifications (models) to determine how variant the conclusions could be.

What population does an observed sample represent? What is the widest possible universe to which the conclusions drawn from a sample apply? The answers depend on how the observations are ascertained and what the deficiencies in data are in terms of nonresponse, measurement errors, and contamination.

Every data set has its own unique features which may be revealed in an initial scrutiny of data and/or during statistical analysis, which may have to be taken into account in interpreting data. Routine data analysis based on textbook methods or software packages can be misleading.

Generally in scientific investigations, a question cannot be answered without knowing the answers to several other questions. It often pays to analyse the data to throw light on a broader set of relevant and related questions.

What data should be collected to answer a given question? Lack of information on certain aspects may create undue complications in applying statistical methods and/or restrict the nature of conclusions drawn from available data. Attempts should be made to collect information on concomitant variables to the extent possible, whose use can enhance the precision of estimators of unknown parameters, and provide broader validity to statistical inference.

Acknowledgements

The work is supported by the Air Force Office of Scientific Research under Contract F49620-82-K-0001. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Bibliography

Alzaid, A. H., Rao, C. R., and Shanbhag, D. N. (1984). *Solutions of Certain Functional Equations and Related Results on Probability Distributions*. Technical Report. Univ. of Sheffield.

Barrai, I., Mi, M. P., Morton, N. E., and Yasuda, N. (1965). "Estimation of prevalence under incomplete selection." *Amer. J. Hum. Genet.*, **17**, 221–236.

Cook, R. D. and Martin, F. B. (1974). "A model for quadrat sampling with visibility bias." *J. Amer. Statist. Assoc.*, **69**, 345–349.

Cox, D. R. (1962). *Renewal Theory*. London: Chapman and Hall.

Dandekar, V. M. and Dandekar, K. (1953). *Survey of Fertility and Mortality in Poona District*. Publication No. 27. Poona, India: Gokhale Institute of Politics and Economics.

Feller, W. (1966). *Introduction to Probability Theory and its Applications, Vol. 2*, New York: Wiley.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications, Vol. 1* (3rd edn.) New York: Wiley.

Fienberg, S. E., Singer, B., and Tanur, J. M. (1985). "Large scale social experimentation." In this volume, Chapter 12.

Fienberg, S. E. and Stasny, E. A. (1983). "Estimating monthly gross flows in labor force participation." *Survey Methodology*, **9**, 77–98.

Fienberg, S. E. and Tanur, J. M. (1983). "Large scale social surveys: perspectives, problems and prospects." *Behavioral Sci.*, **28**, 135–153.

Fisher, R. A. (1934). "The effect of methods of ascertainment upon the estimation of frequencies." *Ann. Eugen.*, **6**, 13–25.

Fisher, R. A. (1953). "Dispersion on a sphere." *Proc. Roy. Soc. London Ser. A*, **217**, 295–305.

Freire-Mala, A. and Chakraborty, R. (1975). "Genetics of archeiropodia." *Ann. Hum. Genet. London*, **39**, 151–161.

Haldane, J. B. S. (1938). "The estimation of the frequency of recessive conditions in man." *Ann. Eugen, (London)*, **7**, 255–262.

Janardhan, K. G. and Rao, B. R. (1983). "Lagrange distributions of the second kind and weighted distributions." *SIAM J. Appl. Math.*, **43**, 302–313.

Kruskal, W. and Mosteller, F. (1980). "Representative sampling IV: The history and the concept in statistics, 1815–1939." *Internat. Statist. Inst. Rev.*, **48**, 169–195.

Mahalanobis, P. C. (1944). "On large scale sample surveys." *Philos. Trans. Roy, Soc. Ser. B*, **231**, 329–451.

Mosteller, F. (1978). "Errors: Nonsampling errors." In W. H. Kruskal and J. M. Tanur (eds.), *The International Encyclopedia of Statistics.* New York: Free Press, 208–229.

Mukherji, R. K., Trevor, J. C., and Rao, C. R. (1955). *The Ancient Inhabitants of Jebel Moya.* London: Cambridge U. P.

Neyman, J. (1977). "Experimentation with weather control and statistical problems generated by it." In P. R. Krishnaiah (ed.), *Applications of Statistics.* Amsterdam: North-Holland, 1–26.

Patil, G. P. and Ord, J. K. (1976). "On size-biased sampling and related form-invariant weighted distributions." *Sankhyā B*, **38**, 48–61.

Patil, G. P. (1984). "Studies in statistical ecology involving weighted distributions." In *Statistics: Applications and New Directions*, Calcutta: Indian Statistical Institute, 478–503.

Patil, G. P. and Rao, C. R. (1977). "The weighted distributions: A survey of their applications." In P. R. Krishnaiah (ed.), *Applications of Statistics*, Amsterdam: North Holland, 383–405.

Patil, G. P. and Rao, C. R. (1978). "Weighted distributions and size biased sampling with applications to wildlife populations and human families." *Biometrics*, **34**, 179–189.

Rao, B. R., Mazumdar, S., Waller, J. H., and Li, C. C. (1973). "Correlation between the numbers of two types of children in a family." *Biometrics*, **29**, 271–279.

Rao, C. R. (1965). "On discrete distributions arising out of methods of ascertainment." In G. P. Patil (ed.), *Classical and Contagious Discrete Distributions*, Calcutta: Statist. Publ. Soc., 320–333. Reprinted in *Sankhyā A*, **27**, 311–324.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edn. New York: Wiley.

Rao, C. R. (1977). "A natural example of a weighted binomial distribution." *Amer. Statist.*, **31**, 24–26.

Rao, C. R. (1975). "Some problems of sample surveys." *Suppl. Adv. Appl. Probab.*, **7**, 50–61.

Rao, C. R. and Rubin, H. (1964). "On a characterization of the Poisson distribution." *Sankhyā A*, **25**, 295–298.

Rao, C. R. and Shaw, D. C. (1948). "On a formula for the prediction of cranial capacity." *Biometrics*, **4**, 247–253.

Rao, J. S. and Sengupta, S. (1966). "A statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranhita-Godavari Valley." *Sankhyā B*, **28**, 165–174.

Rubin, D. B. (1976). "Inference and missing data." *Biometrika*, **63**, 581–592.

Rubin, D. B. (1980). *Handling Nonresponse in Sample Surveys by Multiple Imputations.* A Census Bureau Monograph. Washington.

Sengupta, S. (1966). "Studies on orientation and imbrication of pebbles with respect to cross-stratification." *J. Sed. Petrology*, **36**, 362–369.

Smart, R. G. (1963). "Alcoholism, birth order, and family size." *J. Abnorm. Soc. Psychol.*, **66**, 17–23.

Smart, R. G. (1964). "A response to Sprott's 'Use of chi square'." *J. Abnorm. Soc. Psychol.*, **69**, 103–105.

Sprott, D. A. (1964). "Use of chi square." *J. Abnorm. Soc. Psychol.*, **69**, 101–103.

Sprott, D. A. (1965). "Some comments on the question of identifiability of parameters raised by Rao." In G. P. Patil (ed.), *Classical and Contagious Discrete Distributions.* Calcutta: Statist. Publ. Soc., 333–336.

Stene, Jon (1981). "Probability distributions arising from the ascertainment and analysis of data on human families and other groups." In C. Taille, G. P. Patil and B. Baldessari (eds.), *Statistical Distributions in Scientific Work, Vol. 6*. Dordrecht: Reidel, 51–62.