A "BAD" VIEW OF WEIGHTED DISTRIBUTIONS AND SELECTION MODELS

by

M.J. Bayarri
University of Valencia
and Purdue University

and

M.H. DeGroot
Carnegie Mellon University

Technical Report # 91-12

Department of Statistics
Purdue University

March 1991

$\star$

A "BAD" VIEW OF WEIGHTED DISTRIBUTIONS AND SELECTION MODELS

by

M.J. Bayarri
University of Valencia
and Purdue University

and

M.H. DeGroot
Carnegie Mellon University

ABSTRACT

Weighted distributions and selection models arise when a random sample from the entire population of interest cannot be obtained or it is wished not to do so. Instead, the probability or density that a particular value enters the sample gets multiplied by a non-negative weight function (weighted distributions) that may, in particular, adopt the form of the indicator function of some selection set (selection models). Bayarri and DeGroot have studied in a number of papers many interesting statistical issues arising when dealing with these models. This paper will be a brief, unifying review of their joint work on the subject.

1. INTRODUCTION

Over the years, friends and colleagues have often asked me when and how did Morris DeGroot and myself get first interested in weighted distributions and selection models. The answer is very simple: since the very first day of my one–year visit to Carnegie Mellon University during the academic year 1985/86. Morrie was showing me CMU around, introducing me to faculties there, explaining how could I get books and secretarial work and the alike, and while so doing we were chatting in an informal way about various statistical problems of the kind Morrie used to produce and that always led to interesting, stimulating, and lively discussions among the people around. One such problem was in fact a question: "Do you think that, given the choice, we should go for a sample that has been obtained in some selected subset of the sample space, maybe some set in which observations are very difficult to obtain (as in the tails), or should we always stick to the usual random sample over the whole population?" Nor that I had the singlest clue at the moment, but that was in fact the starting point of a long and productive (and for me extremely interesting) joint work in the subject; indeed, some few months later Morrie gave our first joint talk at a meeting on weighted distributions. The name BAD (Bayarri and DeGroot) was given to us by Nozer Singpurwalla in his discussion of our paper presented at the third Valencia Meeting, and we did enjoy it ever since.

In our work on selection models and weighted distributions, we treated different statistical aspects. Our first incursions in the subject appear in BAD (1987a, 1988), where we formulate the problem, give several examples, and discuss inferences in various selection models.

We got interested very soon, indeed since the very beginning, in the question of whether a selection sample (that is, a random sample obtained in some subset of the sample space) could be more informative than a random sample over the whole population. Our first results pertain to the comparison of the information in a random sample from the whole population with the information in a selection sample according to the criteria of Fisher information and pairwise sufficiency. These results appear in BAD (1987b). We were working in this subject while Morrie was visiting OSU (Ohio State University) and

got Prem Goel also interested in comparing selection experiments with unrestricted ones; we had many interesting conversations on the topic. Eventually, a by–product of our study of Fisher information in selection models got elaborated, investigated, took entity by itself and became a simple and interesting new property of discrete distributions, as appears in Bayarri, DeGroot, and Goel (1989).

In the meantime, we came across with various striking statistical features that appear when analyzing selection models; also, our favorite example (that of properly analyzing statistically significant results that are published in the scientific literature) got studied from different point of views and discussed with friends and colleagues. (I collected part of our results in BAD, 1991.) This example also got us interested in the issue of whether we should worry about observations that we do not observe, in apparent contradiction with the likelihood principle (at the time we were also pursuing some work on the likelihood function and likelihood principle), which led us to explicitly consider the selection mechanisms that could have produced the actual selection sample. All these topics are considered in BAD, 1990.

In the last paper that we jointly wrote in the subject (BAD, 1989), we turned back to our primary question, namely, whether weighted samples can be more informative than random samples. In that paper we moved from selection models to more general weighted distributions, and we definitely abandoned Fisher information and explored sufficiency of experiments. We wrote that paper while Morrie was fighting against cancer. Along all those years, our work did benefit from comments and conversations with many friends, all of whom would make too long a list to appear here; I would nevertheless like to especially acknowledge James Berger, Prem Goel, Joel Greenhouse and Satish Iyengar, with all of whom we spent many hours in fruitful and interesting discussions in the subject and whose remarks directly influenced, improved or motivated some of our results.

The need for writing a review paper often arose, but we never got it written. This intends to be a hopefully good review of BAD's work. The main goal has been to stress the ideas and inferential aspects more than the specific results. Thus, derivation of posterior distributions, proofs, ..., etc. are not given here; also, technical details have been kept to a

minimum.

The paper has 5 sections the first of which is this introduction. In section 2 we present the BAD approach to weighted distributions and selection models, give examples, and motivate their use. In section 3 some interesting statistical aspects that appear when carrying out a Bayesian analysis of selection models are discussed. Section 4 is devoted to the analysis of published significant results and section 5 to the comparison of the information in weighted samples with the information in a random sample.

2. WEIGHTED DISTRIBUTIONS AND SELECTION MODELS

Suppose that the random variable (or random vector) X is distributed over some population of interest according to the (generalized) density $g(x|\theta)$, and that it is desired to make inferences about θ. The usual statistical analysis assumes that a random sample $X_1, \ldots, X_n$ from $g(x|\theta)$ can be observed. There are many situations, however, in which a random sample might be too difficult or too expensive or impossible to obtain and statistical models has to then be developed to incorporate the non–randomness or bias in the observations. (We will see later that in some situations, even if a random sample can be obtained, the experimenter may decide not to do so, since a carefully biased sample may be more informative about θ than a random sample.) *Weighted distributions* arise when the probability (or density) that the value x enters the sample gets distorted so that it is multiplied by some (non–negative) weight function $w(x)$, which in turn may involve some unknown parameters. Thus, the observed data is random sample from the following weighted version of $g(x|\theta)$:

$$f(x|\theta) = \frac{w(x)\ g(x|\theta)}{E_\theta[w(X)]}, \tag{2.1}$$

where the expectation in the denominator is just a normalizing constant so that $f(x|\theta)$ integrates to 1. Distributions with densities of the form (2.1) are called *weighted distributions* by Rao (1965) who first unified the available results, although their use can be traced to Fisher (1934). Good surveys in the topic are Patil (1984) and Rao (1985).

As an example, suppose that the number of talks X that participants in a given statistical meeting would attend has distribution $g(x|\theta)$ and that, while in the meeting, we

are interested in inferring about θ. In all statistical meetings a list of participants is usually available, so that the obvious way to get a random sample $X_1, X_2, \ldots, X_n$ from $g(\cdot|\theta)$ would be to select n participants at random from that list and ask them how many conferences are they attending. This procedure can, nevertheless, be inconvenient, cumbersome and time consuming, since the particular participants thus selected have to then be located and interviewed. An easier and simpler way to get observations X_i would be to select one talk (or several talks) at random and then n people among those attending the talk. Notice, however, that in this case the more talks a participant attends (the larger the value of X) the more likely it is for him or her to be present in the sample. In other words, each observation is *size–biased* in the sense that its density is given by (2.1) where $w(x)$ is an increasing function of x. Under certain equilibrium conditions, it is standard to assume that $w(x) = x$; this particular case is sometimes referred to in the literature as *length–biased* sampling. A more general formulation would be to take $w(x) = x^\tau$ and to treat τ as an unknown parameter. Examples can be found in the references above as well as in Patil and Rao (1977) and Patil (1978). Mahfoud and Patil (1982) relate properties of the original distribution $g(x|\theta)$ and its weighted version $f(x|\theta)$.

Often, as commented above, the weight function $w(x)$ is uncertain and sometimes this uncertainty can be easily modeled by introducing a parameter τ into the function $w(x)$. Thus, a general formulation of (2.1) is

$$f(x|\theta,\tau) = \frac{w(x|\tau)g(x|\theta)}{\int w(x|\tau)g(x|\theta)dx}, \tag{2.2}$$

where g is restricted to be a density whereas the only restriction on the function w is that it has to be non–negative and such that the integral in the denominator is finite. It is however worth emphasizing the fact that both functions w and g enter (2.2) in a completely symmetric fashion.

Various interesting conclusions then follow. First, the restriction that g is a density can be removed as long as it is a non–negative function such that the integral in (2.2) is finite. Hence, the distinction between weight function and the unweighted data–generation process becomes meaningless and they should receive the same statistical treatment. In

particular, θ and τ can be seen to have identical status in (2.2). A Bayesian analysis of these models, in which a joint distribution for θ and τ gets revised, is more in accordance with this fact and seems therefore more sensible than the usual non–Bayesian approach in which the learning processes about θ and τ are regarded very differently (see, e.g., Patil and Rao, 1977, and references there). Each observation does carry information about θ and τ and it is impossible to learn about one of them without simultaneously learning about the other.

A particular case of weighted distributions to which BAD gave special emphasis are the the so–called *selection models* (or truncation models) in which observations are obtained only from some "selected" portions of the population of interest.

For example, suppose that the distribution of a certain vector X of characteristics in a given population is represented by the density $g(x|\theta)$ and that individuals for whom the value of X lies in some set S manifest a certain disease. Assume that we are interested in the distribution of X in the whole population but that X is only measured for individuals who manifest the disease because the observation of X is expensive, painful, or even dangerous (involving perhaps some sort of surgery).

In this example, inferences about θ and S have to be based on data obtained from people who manifest the disease because X will be observed for these persons in the course of their treatment. The density of x for such a person is given by (2.1) where the weight function $w(x)$ is of the form:

$$w(x) = \begin{cases} 1, & \text{if } x \in S, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

When selection occurs, data consist of observations of a random sample $X_1, \ldots, X_n$ of the selection model $f(x|\theta, S)$ which, from (2.1) and (2.3) is given by

$$f(x|\theta, S) = \frac{g(x|\theta)}{\Pr(X \in S|\theta)} \qquad \text{for } x \in S, \tag{2.4}$$

and $f(x|\theta, S) = 0$ otherwise. A random sample from (2.4) will be called *selection sample* and the set S, *selection set.* Moreover, if S is unknown and can be characterized by some

parameter τ, then τ will be called *selection parameter.* The name "selection models" in this context is due to Fraser (1957, 1966) although the term "selection" was used in a more general setting by Tukey (1949).

Selection samples are common in scientific work and in our daily lives. For example, a random sample from a uniform distribution over an unknown subset of the real line or the plane can be regarded as a selection sample (provided the restriction of g being a density is indeed removed). This and related problems are treated in DeGroot and Eddy (1983). Problems of selection bias in a regression context have been widely studied in the econometrics literature (see, e.g., Amemiya, 1984; Heckman, 1976; Little, 1985, and references there). In the rest of the paper we will restrict ourselves to problems in which $g(x|\theta)$ is a univariate distribution. We will let G denote the d.f. corresponding to g.

The simplest selection model in this context occurs when selection is from a known density $g(x)$ and S is one of the tails of the distribution, say the upper tail containing all the values $x \geq \tau$, where τ is uncertain. In this case, (2.4) reduces to:

$$f(x|\tau) = \frac{g(x)}{1 - G(\tau)} \qquad \text{for } x \geq \tau, \tag{2.5}$$

and $f(x|\tau) = 0$ otherwise.

Other simple yet very frequent selection models occur when θ in (2.4) is unknown but the selection set S is known. This is, for example, the case when, in industrial settings, detailed measurements are made only on items which are within certain specified limits. One interesting particular case arises when selection occurs in one of the tails, say the upper one, of the distribution. In this case, the selection model (2.4) is given by:

$$f(x|\theta) = \frac{g(x|\theta)}{1 - G(\tau|\theta)} \qquad \text{for } x \geq \tau, \tag{2.6}$$

and $f(x|\theta) = 0$ otherwise, where τ is a specified value.

Selection models of this type arise naturally when sampling from historical records. In such problems, observations from the entire historical population are usually not available and inferences have to be based on observations that were recorded just for some selected

group. For example, estimation of the distribution of the height of the population might have to be based on the historical observations of the heights of members of the army, for which there was a known minimum required height τ (Wachter and Trussell, 1982).

Other practical problems in which τ is known arise in the analysis of data reported to reinsurance companies, which only receive claims that exceed τ, where the value of τ is fixed by agreement with the original insurance company. The reinsurance company, however, is often interested in making inferences about the entire distribution of claims. Still another example is provided by situations in which measurements are made with instruments of a given sensitivity so that some items cannot be recorded; for example, in astronomy, only sufficiently bright objects can be observed. Sampling from a truncated binomial or Poisson distribution in which the zero class is missing ($\tau = 1$) is another selection model widely treated in the literature (see, e.g. David and Johnson, 1952; Irwin, 1959; Cohen, 1960; Dahiya and Gross, 1973; Sanathanan, 1977; Blumenthal and Sanathanan, 1980; and the survey paper by Blumenthal, 1981); more generally, the lowest $\tau - 1$ classes might be missing, where τ is a known non–negative integer. BAD's favorite example of selection models with known τ occurs when only experimental results that are found to be "statistically significant" are published in the scientific literature; we will turn back to this example in section 4.

Finally, consider selection models in which both θ and S are unknown but there exists a known relationship between them, as for instance selection models in which $\Pr(X \in S|\theta)$ is known. When X is univariate and S is the upper tail of the distribution having fixed probability α, then the density for this selection model is

$$f(x|\theta) = \frac{1}{\alpha}\, g(x|\theta) \qquad \text{for } x \geq G^{-1}(1-\alpha|\theta), \tag{2.7}$$

and $f(x|\theta) = 0$ otherwise. An illustration of this type of problem is provided by a simplified version of the medical example mentioned at the beginning of this section, namely when individuals manifest a certain disease whenever the value of an associated latent random variable X exceeds some critical threshold τ, and it is known that the proportion of individuals in the population who have the disease is α. If, on the basis of the values of X

observed in a random sample of patients having the disease it is desired to make inferences about the distribution of X in the whole population, then model (2.7) should be used.

Selection samples also occur in much more general settings of our daily lives. Consider the news that is reported on newspaper or on television: because of space and time restrictions, it is a selection sample from the events of the day. Another entertaining example is provided by what we can wrongly perceive as consecutive "perfect" predictions, as described in DeGroot (1986, chapter 1) and in BAD (1988).

3. INADEQUACIES OF "ROUTINE" PRIOR DISTRIBUTIONS

In this section we will show that, when dealing with selection models, the naive use of "routine" type of prior distributions can be inadequate. By "routine" priors we mean the ones that are usually assumed in much of Bayesian work, namely non–informative priors or conjugate priors.

Selection models provide very simple examples where improper priors *always* lead to improper posteriors, no matter which observations or how many observations have been obtained. Consider the case in which selection occurs in the upper tail of a completely specified $g(x)$, so that the selection model is given by (2.5). Then, if $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a selection sample from (2.5), the likelihood function $\ell(\tau)$ is

$$\ell(\tau) \propto [1 - G(\tau)]^{-n} \qquad \text{for } \tau \leq \min(x_1, \ldots, x_n) \tag{3.1}$$

and $\ell(\tau) = 0$ otherwise. Based in (3.1), a natural choice for a conjugate prior density for τ is

$$\xi_1(\tau) \propto [1 - G(\tau)]^{a} \qquad \text{for } \tau \leq b \tag{3.2}$$

where the hyperparameters a and b are constants to be specified. A distribution of this form may be appropriate in problems in which the support of $g(\cdot)$ is bounded on the left, so the possible values of τ are also bounded on the left. However, if the possible values of τ are unbounded from the left, then the density $\xi_1(\tau)$ will be improper for every real number a. Although Bayesians are used to improper prior distributions, this particular one is completely useless, since it is conjugate, and hence the posterior distribution will

also be improper, for all values of n and all values of $x_1, x_2, \ldots, x_n$. Thus if it is desired to carry out an analysis under a conjugate distribution, a different form for the conjugate prior has to be selected.

There are, indeed, many different families of prior distributions that are closed under sampling. For example, any density of the form

$$\xi(\tau) \propto h(\tau)[1 - G(\tau)]^a \qquad \text{for } \tau \leq b \tag{3.3}$$

will have this property, where h could be *any* function such that the right–hand–side of (3.3) is integrable over τ. In particular, h could be taken to be of the form

$$h(\tau) = \begin{cases} 0 & \text{for } \tau < c \\ 1 & \text{for } \tau \geq c \end{cases} \tag{3.4}$$

so as to specify a lower bound c for the possible values of τ. However, a more suitable choice in this case is to simply take h to be g. The resulting prior density:

$$\xi_2(\tau) \propto g(\tau)[1 - G(\tau)]^a \qquad \text{for } \tau \leq b \tag{3.5}$$

is proper for all values of a and the Bayesian analysis is straightforward.

It should be emphasized that it is not the particular form of (3.2) which makes the posterior improper. As a matter of fact, *any* improper density $\xi_u(\tau)$ for which the area under the lower tail is infinite will result in posterior distributions that are improper for all n and all $x_1, x_2, \ldots, x_n$. Moreover, if $\xi(\tau)$ is a proper prior for which exactly the first k moments exist ($k = 0, 1, \ldots$) and for which higher moments do not exist because the integral $\int_{-\infty}^{c} |\tau|^{k+1} \xi(\tau) d\tau$ is infinite for all c, then the corresponding posterior distribution $\xi(\tau|\underset{\sim}{x})$ will also have exactly k moments for all n and all $\underset{\sim}{x}$ (see BAD, 1990).

Simple examples outside the selection models framework in which the election of the prior determines whether the posterior will be proper or improper, and its number of moments are provided by sampling binomial data with both parameters unknown (see Kahn, 1987), and the simple mixture model $f(x|\theta) = \lambda g_0(x) + (1-\lambda) g(x|\theta)$, where $g_0(x)$ is some completely specified density. It can be immediately seen that if the prior distribution

for θ is improper, so will be the posterior, and that if the prior is proper with exactly k moments, the posterior distribution will also have exactly k moments.

In the examples just presented, we might say that data have little influence on the posterior distribution in the sense that they cannot make proper an improper prior. Selection models also provide examples of the opposite situation for which, under conjugate analysis, the posterior distribution is not *at all* influenced by the hyperparameters of the prior after just a finite number of observations are obtained (see BAD, 1990).

Still another way in which conjugate prior distributions can work poorly with selection models is by exhibiting an inappropriate dependence on the experiment to be performed. Assume, for example that the underlying density $g(x|\theta)$ is a member of an exponential family so that

$$g(x|\theta) = r(x)s(\theta)\exp\{u(x)v(\theta)\}, \tag{3.6}$$

and that the value x is observed only if $x \geq \tau$, where τ is a known value. In this case, the density $f(x|\theta)$ is given by (2.6), and it follows that the likelihood function $\ell(\theta)$ for θ based on the selection sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ is given by

$$\ell(\theta) \propto \frac{[s(\theta)]^n \exp\{v(\theta)\sum_{i=1}^{n} u(x_i)\}}{[1 - G(\tau|\theta)]^n}. \tag{3.7}$$

Thus, a conjugate prior density for θ based directly on the form of $\ell(\theta)$ is

$$\xi(\theta) \propto \frac{[s(\theta)]^a \exp\{bv(\theta)\}}{[1 - G(\tau|\theta)]^c}, \tag{3.8}$$

where the hyperparameters a, b and c are specified constants. Although the form of any conjugate prior distribution is always related to the experiment to be performed, the form of (3.8) seems especially unsuitable because of its explicit dependence on the value of τ. It is true that in some cases τ and θ are related as, for example, in the sampling from historical records where τ was the minimum required height to join the army; obviously τ should be assumed to be related to the mean height of the population. In most problems, however, the fixed value of τ conveys no information about θ: for example, it would usually be unsuitable to assume, in the astronomy example, that the mean number of stars in regions of interest is related to the sensitivity of the observing instrumental.

It seems more appropriate in these problems to use a prior distribution that is conjugate with respect to the unrestricted model, which is equivalent to using the value $c = 0$ in (3.8). With this choice we have eliminated the dependence of the prior distribution on τ without unduly complicating the statistical analysis that remains essentially unchanged.

As a final example of inadequate conjugate prior distributions consider the selection model mentioned in the previous section in which selection occurs in the upper tail ($x \geq \tau$) of $g(x|\theta)$, and both θ and τ are unknown but $\alpha = \Pr(X \geq \tau|\theta)$ is known. The density for this selection set is given by (2.7), and the likelihood function for θ based on the selection sample $x_1, \ldots, x_n$ is

$$\ell(\theta) \propto \prod_{i=1}^{n} g(x_i|\theta) \qquad \text{for } \theta \in \Omega(\alpha, t), \tag{3.9}$$

where $\Omega(\alpha, t)$ is the set of all values of θ such that $G(t|\theta) \geq 1 - \alpha$ and $t = \min\{x_1, \ldots, x_n\}$.

A conjugate prior density based directly on the form of $\ell(\theta)$ given in (3.9) would require that the range $\Omega(\alpha, a)$ of possible values of θ be dependent not only on a hyperparameter a whose value can be specified arbitrarily, but also on α. For reasons similar to those presented above, it seems desirable to eliminate the dependence of the prior on α. Thus, in problems where there is a suitable family of conjugate prior distributions for the unrestricted model $g(x|\theta)$, it is usually more appropriate to use this family for the selection model as well. In effect, this type of prior is equivalent to choosing the value $a = \infty$ for the hyperparameter a in $\Omega(\alpha, a)$.

4. SELECTION MODELS FOR SIGNIFICANT RESULTS

As mentioned in section 2, an important example of selection models with known value of τ occurs when only experimental results that are found to be "statistically significant" are published in the scientific literature. This situation is very common both because the editors of some journals encourage the publication of articles in which statistical significance has been obtained, and because many experimenters themselves regard their results as being useless unless the results are statistically significant and will not even submit them for publication.

The unfortunate effects that these policies can have on scientific learning have been discussed by several authors and some references are given in BAD (1987a, 1991). The careful modeling of publication bias should be especially important when a meta–analysis (combining the results from different experiments) is carried out, since meta–analyses are usually based on published experiments. Unfortunately, this is seldom the case. An interesting modeling for meta–analysis can be found in Iyengar and Greenhouse (1988).

To present our discussion of this problem, we will use a very simple (unrealistically simple) example so as to highlight the main points we would like to stress. Real situations are seldom so clear–cut and the effects of publication bias might be less dramatic than the ones we will find here. The framework will be that of one–sided test of hypotheses on the mean of a normal distribution with known variance, and we will restrict ourselves to the analysis of a single published significant result. The concerns we will rise here would also apply, of course, to more complicated scenarios and to a full meta–analytic analysis.

Assume that independent experiments are carried out by the same or different experimenters around the world. In each of them a random sample $y_1, y_2, \ldots, y_m$ of size m is taken from a normal distribution with unknown mean μ and known variance σ^2 and the uniformly most powerful test at some level α is used for testing

$$H_0: \mu \leq 0 \text{ versus } H_1: \mu > 0. \tag{4.1}$$

In this case, the distribution $g(t|\theta)$ of the test statistic

$$T = \frac{\sqrt{m}}{\sigma} \quad \overline{Y}_m, \tag{4.2}$$

is normal with mean $\theta = \sqrt{m}\ \mu/\sigma$ and variance 1, where $\overline{Y}_m$ represents, as usual, the sample mean in a given experiment.

The restrictions to equal (and known) variances σ^2, equal sample sizes m and equal means μ in all the experiments are, of course, quite unrealistic and are used here just to ease the presentation. In a more realistic setting, the sample sizes m would be different in different experiments, the variances σ^2 would be unknown and, as the means μ, should be allowed to vary from experiment to experiment, although in a meta–analysis they would

obviously be related, relation that could maybe be modeled by using a hierarchical model. Arguments similar to the ones we will present could also be derived to apply to these realistic settings, and the nature of the concerns that we will rise would remain essentially unchanged.

Let's come back to our simplified situation and assume that the results of one such experiment appear published in some scientific journal rejecting the null hypothesis H_0 and declaring the data significant because they yield a small p–value. Assume also that only experimental results that are found "statistically significant" get published.

What should be conclude from such an experiment? The usual reaction is, of course, to take the distribution $g(t|\theta)$ of the test statistic T, its observed value t and its associated p–value $\Pr(T \geq t|\theta = 0)$ at face value and thus conclude that the data is indeed significant. But as readers of the journal, we will not learn the results of this experiment unless they lead to the rejection of H_0; that is, unless $T \geq \Phi^{-1}(1-\alpha)$. In this situation, the density of any value of T that we will actually get to observe is not simply $g(t|\theta)$, but is given by the selection model:

$$f(t|\theta) = \frac{\phi(t-\theta)}{1-\Phi(\tau-\theta)} \qquad \text{for } t \geq \tau, \tag{4.3}$$

and $f(t|\theta) = 0$ for $t < \tau$, where ϕ and Φ are the p.d.f. and d.f. respectively, of the standard normal distribution, and $\tau = \Phi^{-1}(1-\alpha)$.

To see how a "significant" value of T when properly analyzed can support the null hypothesis H_0, it is enlightening to calculate the maximum likelihood estimate for θ.

It can be shown from (4.3) that the maximum likelihood estimator $\hat{\theta}$ is the unique solution of the equation:

$$(t-\hat{\theta}) \;\; M(\tau-\hat{\theta}) = 1, \tag{4.4}$$

where Mills' ratio $M(\cdot)$ is defined by

$$M(x) = \frac{1-\Phi(x)}{\phi(x)}. \tag{4.5}$$

The values of $\hat{\theta}$ for $\alpha = 0.01$, 0.05, and 0.10 and various observed values of t are given in Table 1. Of course, the most widely used criterion for statistical significance is $\alpha = 0.05$.

In Table 1 we also give the p–values corresponding to the observed values of t as they would be reported in the literature, that is, calculated from the standard normal distribution. Since only "significant" values of T can be observed, all the p–values must be less than α.

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
$\hat{\theta}$	t	p	t	p	t	p
3	3.424	0.0003	3.175	0.0008	3.095	0.0010
2	3.017	0.0013	2.586	0.0049	2.403	0.0081
1	2.792	0.0026	2.249	0.0124	1.985	0.0236
0	2.665	0.0038	2.063	0.0195	1.755	0.0397
−1	2.588	0.0048	1.955	0.0253	1.625	0.0526
−2	2.538	0.0056	1.888	0.0305	1.546	0.0611
−3	2.503	0.0062	1.844	0.0326	1.496	0.0675
−5	2.453	0.0071	1.789	0.0368	1.434	0.0759

Table 1: Maximum likelihood estimate of θ for various observed values t and p.

Since $\hat{\mu} = \sigma\hat{\theta}/m^{1/2}$, it follows from (4.1) that negative values of $\hat{\theta}$ support the null hypothesis H_0. The basic conclusion to be drawn from the discussion in this section is that even observed values of t that appear to be highly significant can yield maximum likelihood estimates that are very negative and give strong support to H_0. For example, it can be seen from Table 1 that when $\alpha = 0.05$, a p–value as small as 0.033 yields $\hat{\theta} = -3$, an estimate that is at least 3 standard deviations away from the parameter values in H_1. Similar behavior is found for $\alpha = 0.01$ and $\alpha = 0.10$.

These results emphasize the fact that under current statistical practice, a published p–value only slightly smaller than the effective α can only be regarded as supporting H_0 rather than rejecting H_0. In fact, as the p–value approaches α in this selection model $\hat{\theta} \to -\infty$.

In order to reach this conclusion we have assumed that the experiment was performed repeatedly (not necessarily by the same experimenter, but also by different experimenters working in the same problem) until one significant report was obtained and published. The number N of performed experiments does not provide further information about θ and it should not change our inferences about θ even if we explicitly introduce it into the

analysis. We will turn to this point later on.

Under different conditions, the analysis should proceed differently. For example, suppose at the other extreme that we know beforehand that there is just a single experimenter who performs the experiment just once. In this case, when we read his or her published report of a significant result, we know that this is the actual outcome of the only experiment that was performed and therefore it should be accepted at face value and analyzed using the density $g(t|\theta)$. We know that if the experiment had yielded a non–significant result, we would not have seen any published report at all. This lack of a report would, therefore, have given us information about θ as well.

Intermediate conclusions between these two extreme ones can be obtained by taking into consideration the number N of performed experiments. Problems of this type, dubbed the "file drawer problem" by Rosenthal (1979) are also considered by Sterling (1959), Dawid and Dickey (1977), Hedges and Olkin (1985, chap. 14), and Iyengar and Greenhouse (1988), among others. Of course, this is only one example of a more general phenomenon that can be present when analyzing selection samples: the fact that the way the selection sample was actually produced, that is, the selection mechanism generating the observed sample, can have a decisive influence in the statistical analysis of the data and sometimes it has to explicitly be taken into consideration. In BAD (1990), selection mechanisms are considered and discussed in the general scenario of selection models, as well as conditions under which the selection mechanisms can be ignored in the analysis of data, either because it does not provide additional information about θ, or because even if it does, the particular form of the prior distribution makes it ignorable when making inferences about θ. We will content ourselves here with our particular example.

Suppose, in the spirit of this discussion, that we observe n published results $\boldsymbol{t} = (t_1, \ldots, t_n)$ all of them significant, and we are uncertain about the total number N of experiments that have been performed. Included in the $N - n$ unpublished results there might be both significant and non–significant outcomes. Under the Bayesian approach to this problem, a joint generalized density $p(\boldsymbol{t}, n, N, \theta)$ is specified, which we may factor in

the following convenient way:

$$p(\underset{\sim}{t}, n, N, \theta) = p(\underset{\sim}{t}|n, N, \theta)p(n|N, \theta)p(N|\theta)p(\theta). \tag{4.6}$$

In (4.6), we are using the symbol p to denote an arbitrary density without any implication that it is the same one for all variables. It should be emphasized that $\underset{\sim}{t}$ and n are observed, and N and θ are unobservables. Hence, $p(N|\theta)p(\theta)$ represents the joint prior density of N and θ. We assume that τ is fixed and known throughout the analysis, and hence it is omitted from the notation although (4.6) can be generalized to the situation in which τ is unknown or varies from experiment to experiment.

Next, we make the basic assumption that for the observed value of n, the observed significant results $t_1, \ldots, t_n$ form a random sample from the selection model given by (4.3). Therefore,

$$\begin{aligned} p(\underset{\sim}{t}|n, N, \theta) = p(\underset{\sim}{t}|n, \theta) &= \prod_{i=1}^{n} f(t_i|\theta) \\ &= \frac{\prod_{i=1}^{n} \phi(t_i - \theta)}{[1 - \Phi(\tau - \theta)]^n} \quad \text{for } t_i > \tau \quad (i = 1, \ldots, n). \end{aligned} \tag{4.7}$$

The two situations that we are discussing are special cases of the model (4.6) under the assumption (4.7): In the first situation, it is known beforehand that n will have the fixed value $n = 1$. Thus, (4.6) becomes

$$f(t_1|\theta)p(N|\theta)p(\theta), \tag{4.8}$$

where $f(t_1|\theta)$ is given by (4.3) and the joint density of t_1 and θ simply reduces to

$$p(t_1, \theta) = f(t_1|\theta)\left[\sum_{N=1}^{\infty} p(N|\theta)\right] p(\theta) = f(t_1|\theta)p(\theta). \tag{4.9}$$

It follows from (4.9) that analysis of this published significant result is based on the selection model $f(t_1|\theta)$, so that it might very well be providing strong evidence in favor of H_0, as discussed previously.

In the second situation, it is known beforehand that $N = 1$. In this case, the probability of obtaining $n = 1$ is

$$p(n = 1|N = 1, \theta) = 1 - \Phi(\tau - \theta). \tag{4.9}$$

Hence, from (4.9) and the fact that $p(N = 1|\theta) = 1$, the joint density (4.6) becomes

$$f(t_1|\theta)[1 - \Phi(\tau - \theta)]p(\theta) = g(t_1|\theta)p(\theta) = \phi(t_1 - \theta)p(\theta), \tag{4.10}$$

where $g(t_1|\theta)$ is the original, unrestricted density of T. It follows from (4.10) that the observed significant result t_1 should be analyzed at face value as common sense suggested.

In the file–drawer problem, none of the two situations above is usually assumed. Instead, it is assumed that there is an unknown number N of performed experiments and subjective knowledge about N is required from the reader in order to properly analyze the significant result. From a classical point of view, a value of N that would reverse the conclusion (that is, that would make the result non–significant) is computed (called the *file–safe sample size*) and the reader is asked to compare that with his or her subjective opinion about N so as to conclude whether or not the significant data supports H_0. From a Bayesian point of view, the reader is asked to specify his or her prior distribution for $N, p(N|\theta, \tau)$ that, in general, might depend on θ and τ. Inferences about θ will be based in the posterior distribution:

$$p(\theta|t, n = 1) \propto \phi(t - \theta)p(\theta) \sum_{N=1}^{\infty} N[\Phi(\tau - \theta)]^{N-1} p(N|\theta, \tau). \tag{4.11}$$

It can be seen from (4.11) that the analysis would simply be based on the selection model $f(t|\theta) = \phi(t - \theta)/[1 - \Phi(\tau - \theta)]$ if $p(N|\theta, \tau)$ is such that

$$\sum_{N=1}^{\infty} N[1 - \Phi(\tau - \theta)][\Phi(\tau - \theta)]^{N-1} p(N|\theta, \tau) \tag{4.12}$$

does not depend on θ, as it is the case when $p(N|\theta, \tau)$ corresponds to a Poisson distribution with mean $\lambda/\Phi(\tau - \theta)$, for any fixed λ, or when it is taken to be inversely proportional to $1/N$ (which results in an improper prior). Otherwise the full analysis based on (4.12) has to be carried out. For details, see BAD (1990, 1991).

If it is not wished to carry out a fully Bayesian approach, we strongly recommend, at least, that conditional posterior distributions of the form

$$p(\theta|t, n = 1, N) \propto \varphi(t - \theta)[1 - \Phi(\tau - \theta)]^{N-1} p(\theta) \tag{4.13}$$

or their associated quantities of interest (estimators, HPD regions,...,etc.) be computed for a likely range of values of N to tentatively investigate how sensitive our conclusions are to the value of N. If they are very sensitive, a full Bayesian analysis has to be carried out (with maybe different types of priors for N).

5. INFORMATION IN SELECTION MODELS AND WEIGHTED DISTRIBUTIONS

One constant of BAD work on selection models and weighted distributions over the years have been their interest in comparing the information about θ provided by a random sample from the underlying density $g(x|\theta)$ with the information in a selection sample or, more generally, in a random sample from a weighted version $f(x|\theta)$ of $g(x|\theta)$. We will see that a selection sample is *not* necessarily less informative about θ than an unrestricted random sample, and that in some statistical problems we are better off with a biased sample than with a random sample.

In order to compare both types of experiments it is useful to elaborate further the notation so far employed. We will continue to restrict ourselves to univariate problems. In this section, we will use X to denote a random variable whose distribution is the original, underlying $g(x|\theta)$ and will use Y to denote random variables whose distribution is characterized by some weighted version $f(y|\theta)$ of $g(x|\theta)$, as given in (2.1), where w will be some specific weight function. Also, $\mathcal{E}_X$ will denote the statistical experiment in which an observation X is to be obtained, and $\mathcal{E}_Y$ is defined similarly.

There are many different ways to measure and compare the information in statistical experiments. Probably the most common one in Statistics is based on Fisher information. We shall let $I_X(\theta)$ and $I_Y(\theta)$ denote Fisher information for the experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$, respectively, under the standard regularity conditions. If $I_X(\theta) \geq I_Y(\theta)$ for all values of θ, then (and abusing language a little) the unrestricted experiment $\mathcal{E}_X$ will always yield at least as much information about θ as the weighted experiment $\mathcal{E}_Y$. In this case, we write $\mathcal{E}_X \succeq_F \mathcal{E}_Y$. The relation $\mathcal{E}_Y \succeq_F \mathcal{E}_X$ is defined analogously. Comparison of some weighted and unweighted distributions according to this criterion can be found in BAD (1987b) and

Patil and Taillie (1987).

In spite of its widespread use in statistics, Fisher information does not have any clear–cut operational interpretation in statistical decision theory. More decision oriented although more restrictive, are the concepts of information based on the theory of the comparison of statistical experiments as developed originally by Blackwell (1951, 1953). The experiment $\mathcal{E}_X$ is said to be *sufficient* for the experiment $\mathcal{E}_Y$, denoted $\mathcal{E}_X \succeq \mathcal{E}_Y$, if there exists a stochastic transformation of X to a random variable $Z(X)$ such that, for each value of θ, the random variables $Z(X)$ and Y have identical distributions. The relation $\mathcal{E}_Y \succeq \mathcal{E}_X$ is defined in an analogous way.

Sufficiency of experiments is a very restrictive ordering, in the sense that it is relatively rare for one experiment to be regarded as more informative than another one. Nevertheless, when this does occur, every decision maker would prefer the more informative experiment, regardless of his or her prior distribution and utility function, since the relation $\mathcal{E}_X \succeq \mathcal{E}_Y$ holds if any only if for every decision problem involving θ and every prior distribution for θ, the expected Bayes risk from $\mathcal{E}_X$ is no greater than that from $\mathcal{E}_Y$.

A somehow less restrictive partial order that still makes use of the basic concept of sufficiency is based in the notion of pairwise sufficiency. An experiment $\mathcal{E}_X$ is said to be *pairwise sufficient* for the experiment $\mathcal{E}_Y$, denoted $\mathcal{E}_X \succeq_2 \mathcal{E}_Y$, if for *every* pair of values $\theta_1, \theta_2, \mathcal{E}_X$ is sufficient for $\mathcal{E}_Y$ when the parameter space is restricted to contain just the two values θ_1 and θ_2. The relation $\mathcal{E}_Y \succeq_2 \mathcal{E}_X$ is defined in an analogous way.

Clearly, if $\mathcal{E}_X \succeq \mathcal{E}_Y$ then $\mathcal{E}_X \succeq_2 \mathcal{E}_Y$. Also, the relationship $\mathcal{E}_X \succeq \mathcal{E}_Y$ implies a similar ordering in terms of Fisher information; i.e., if $\mathcal{E}_X \succeq \mathcal{E}_Y$ then $\mathcal{E}_X \succeq_F \mathcal{E}_Y$. However, the converse relations does not necessarily hold. Moreover, since the Fisher information can be obtained from the Kullback–Liebler information by considering pairs of values of θ that are arbitrarily close to each other, it can be shown that if $\mathcal{E}_X \succeq_2 \mathcal{E}_Y$ then $\mathcal{E}_X \succeq_F \mathcal{E}_Y$. Some of these relations and other properties of the comparison of experiments are described in Stein (1951), Stone (1961), Kullback (1968), Torgersen (1970, 1972, 1976), Hansen and Torgersen (1974), and Goel and DeGroot (1979).

In general, we will be interested in the information in weighted and random samples, and not only in the information in single observations. Nevertheless, it will be sufficient for us to consider the simple experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$ since it can be shown that, if any of the relations above between $\mathcal{E}_X$ and $\mathcal{E}_Y$ holds, then the same relation holds between the experiment in which a random sample $X_1, \ldots, X_n$ is obtained from the underlying density $g(x|\theta)$ and the experiment in which a random sample $Y_1, \ldots, Y_n$ is obtained from the weighted density $f(y|\theta)$.

We will next compare $\mathcal{E}_X$ and $\mathcal{E}_Y$ for various underlying models $g(x|\theta)$ and various weight functions and will show that the ordering of experiments can be very sensitive to the weight function used.

5.1. Selection from an exponential family

First consider the question of whether we gain or lose Fisher information when a selection sample rather than random sample is obtained from a distribution belonging to an exponential family. Thus, we are assuming that $g(x|\theta)$ is of the form $g(x|\theta) = r(x)s(\theta)\ \exp\{u(x)\ v(\theta)\}$, and that Y is observed only if Y lies in some selection set S. Hence, in this case, $w(y)$ is just the indicator function of the set S.

The relation $\succsim_F$ is very easy to characterize in this situation in terms of the function $t(\xi) = \Pr[X \in S|\theta = v^{-1}(\xi)]$. Indeed $\mathcal{E}_X \succsim_F \mathcal{E}_Y$ holds if and only if $t(\xi)$ is a log–concave function of ξ, and $\mathcal{E}_Y \succsim_F \mathcal{E}_X$ if and only if $t(\xi)$ is log–convex.

As an example, assume that the underlying distribution is gamma with a known value of the shape parameter α and with unknown scale parameter θ, and that selection occurs in the upper tail $y \geq \tau$. In this case, if $\alpha > 1$, then $\mathcal{E}_X \succsim_F \mathcal{E}_Y$ and if $0 < \alpha < 1$, then $\mathcal{E}_Y \succsim_F \mathcal{E}_X$. In words, if $g(x|\theta)$ is a J–shaped gamma, then according to this criterion, we would rather have all our observations in the upper tail since all we can possibly get from also having observations in the interval $(0, \tau)$ is a loss of Fisher information. On the other hand, if $g(x|\theta)$ is a bell–shaped gamma density, then a random, unrestricted sample provides greater Fisher information about θ than a selection sample from the upper tail. When $\alpha = 1$, the distribution of both X and $Y - \tau$ is exponential with parameter θ and

both experiments are equivalent (each one is sufficient for the other one).

Next, we compare the binomial and Poisson distributions with their zero–truncated versions in which the zero class is missing and observations are restricted to be positive. Hence, $\mathcal{E}_X$ will denote the experiment in which an unrestricted observation X is taken from one of these two families, and $\mathcal{E}_Y$ will denote the experiment in which an observation Y is taken from the corresponding zero–truncated version. Then a straightforward calculation shows that $\mathcal{E}_X \gtrsim_F \mathcal{E}_Y$; as a matter of fact $\mathcal{E}_X \gtrsim_2 \mathcal{E}_Y$ (as pointed to us in independent personal communications by Prem K. Goel and Erik N. Torgersen). However, it can be explicitly shown (BAD, 1987b) that in spite of these relations, it is *not* true that $\mathcal{E}_X \gtrsim \mathcal{E}_Y$, thus providing an interesting example of experiments in which the parameter space is an open subset of the real line and $\mathcal{E}_X \gtrsim_2 \mathcal{E}_Y$ but it is not true that $\mathcal{E}_X \gtrsim \mathcal{E}_Y$. It also follows that since $\mathcal{E}_X \gtrsim \mathcal{E}_Y$ does not hold, there must exist a decision problem and prior distribution for θ such that the Bayesian expected risk is smaller with the selection sample than with the random sample.

5.2. The "Big Three" and the "CV" property

This subsection is a bit an aside from the general argument in this section and it just pretends to mention a cute fact we came across in our study of the Fisher information in truncated versions of standard discrete distributions. The initial "cute fact" was further elaborated and investigated and became a general property of discrete random variables, not necessarily longer related to the selection models scenario. The results can be found in Bayarri, DeGroot and Goel (1989).

Assume here that X is a random variable whose distribution belongs to any one of the "Big 3" families of discrete distributions: binomial, Poisson and negative binomial, and that Y is its zero–truncated version. When deriving by "brute force" $I_X(\theta)$ and $I_Y(\theta)$ for the experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$, we surprisingly found that, for all the "Big 3", the ratio of $I_Y(\theta)$ to $I_X(\theta)$ could be expressed as:

$$\frac{I_Y(\theta)}{I_X(\theta)} = \frac{\Pr(X \geq 2|\theta)}{\Pr(X \geq 1|\theta)]^2}. \tag{5.1}$$

This was the starting "cute" fact. Further elaboration took into account two other general

facts.

The first fact is that, when $g(x|\theta)$ is a member of an exponential family of the form given by (3.5) with $u(x) = x$, and Y gets observed only if it lies in a fixed selection set S, then it can be shown that

$$\frac{I_Y(\theta)}{I_X(\theta)} = \frac{\text{Var}(Y|\theta)}{\text{Var}(X|\theta)}. \tag{5.2}$$

The second general fact is that if X is any discrete random variable taking values on the non–negative integers and whose distribution has mean μ and variance σ^2, and if Y is its zero truncated version, then

$$\frac{\text{Var}(Y)}{\text{Var}(X)} = \frac{1 - p_0 - \frac{\mu^2}{\sigma^2}p_0}{(1-p_0)^2}, \tag{5.3}$$

where $p_0 = \Pr(X = 0)$ is assumed to be < 1.

Both general facts did apply to our statistical situation, so that putting together (5.1), (5.2) and (5.3) we found that the "Big 3" satisfy the "CV property" that is defined as follows:

Suppose that a random variable X has a discrete distribution on the non–negative integers with finite variance, and let

$$\begin{aligned} p_i &= \Pr(X = i) \qquad \text{for } i = 0, 1, 2, \ldots \\ \mu &= E(X), \text{ and } \sigma^2 = \text{Var}(X). \end{aligned} \tag{5.4}$$

Then the distribution of X is said to have the *CV property* if

$$\frac{p_0}{p_1} = \frac{\sigma^2}{\mu^2}. \tag{5.5}$$

The name *CV property* given to (5.5) was inspired by the fact that the right–hand side of (5.5) is the square of the *coefficient of variation* (the CV) of the distribution of X.

Although isolated examples of distributions satisfying the CV property can be constructed, we do not know of any other widely–used family of discrete distributions apart from the "Big 3" all of whose members satisfy it. (In particular, it is *not* satisfied by discrete uniform, hypergeometric, nor beta–binomial distributions.) For a dipper treatment of this property the reader is referred to Bayarri, DeGroot and Goel (1989).

5.3. Normal Selection

We now come back to comparing information in random and selection samples by considering various selection sets for an underlying normal distribution.

We begin by considering an example in which the experiment based on a selection sample is pairwise sufficient for the analogous experiment based on an unrestricted random sample. Consider a population in which X has a normal distribution with a known mean, which we take to be 0, and an unknown precision θ. (The precision of a normal distribution is the reciprocal of its variance). For $i = 1, 2$, let $\mathcal{E}_i$ denote the experiment in which a single observation y_i is obtained by restricting X to the tails $X \geq \tau_i$ and $X \leq -\tau_i$, where $\tau_i > 0$ is given and fixed. (It should be noted that this experiment is equivalent to one in which X is restricted to the upper tail only.) It can be shown in this case that if $\tau_1 < \tau_2$ then $\mathcal{E}_2 \gtrsim_2 \mathcal{E}_1$, that is the experiment with a smaller selection set is pairwise sufficient for the experiment with a larger selection set. It follows that a selection sample from the upper tail of a normal distribution with known mean and unknown precision is pairwise sufficient for a selection sample of the same size from the upper tail with a smaller truncation point. Moreover, a selection sample from the upper tail is pairwise sufficient for a random sample from the entire distribution.

It is striking that the reverse relationships holds when the mean is unknown and the precision is known. To see this, assume now that X has a normal distribution with unknown mean θ and known precision. For $i = 1, 2$, let $\mathcal{E}_i$ denote the experiment in which a single observation Y_i is obtained by restricting X to the upper tail $X \geq \tau_i$, where $-\infty < \tau_1 < \tau_2 < \infty$. Here it can be shown that $\mathcal{E}_1 \gtrsim_2 \mathcal{E}_2$. In particular, by letting $\tau_1 \to -\infty$, it follows that a random sample from the entire normal distribution is pairwise sufficient for a selection sample from the upper tail.

We do not know whether the relation $\gtrsim_2$ can be replaced by the stronger relation $\gtrsim$ in these problems.

If X has a normal distribution with unknown mean θ and known precision, and the observation Y is obtained by restricting X to the two tails $X \leq \tau_1$ and $X \geq \tau_2$, then it

can be shown that neither $\mathcal{E}_X \succsim_F \mathcal{E}_Y$ nor $\mathcal{E}_Y \succsim_F \mathcal{E}_X$, and therefore the stronger relations $\succsim_2$ and $\succsim$ cannot hold either.

5.4. Other weight functions

In this subsection we will present some results on sufficiency obtained when various weight functions are applied to standard statistical distributions, which we will use to demonstrate the wide variety of effects that weighing can have in the ordering of experiments. In this and the next subsection this ordering refers to sufficiency in all the examples.

Assume first that the distribution of X is exponential and that the distribution of Y is a weighted version of it. We will see that all the possibilities in the ordering of the experiments $\mathcal{E}_X$ and $\mathcal{E}_Y$ are attained for particular weight functions $w(x)$. Indeed, if we have exponential weight $w(x) = e^{-ax}(a > 0)$ then $\mathcal{E}_X \succsim \mathcal{E}_Y$, and a random sample is thus preferred to a "weighted" sample. On the other hand, for size–biased weights of the form $w(x) = x^a(a > 0)$ the relation is reversed and we have in fact $\mathcal{E}_Y \succsim \mathcal{E}_X$; notice that, if we are in such statistical situations, we should be very glad of having size–biased samples. Finally, as we have already mentioned, we have equivalent experiments $\mathcal{E}_X \approx \mathcal{E}_Y$ when the weight function is the indicator function of any set of the form $x \geq a$, that is, when Y represents selection from the upper tail.

We have just presented an example in which the original and the weighted experiments were equivalent experiments. A less known example of the same type of behavior is the following: Assume that X has a normal distribution with mean θ and variance 1 and that the distribution of Y is its weighted version with exponential weight of the form $w(x) = e^{ax}$. Then $\mathcal{E}_X$ and $\mathcal{E}_Y$ are equivalent experiments.

If the underlying distribution is a Poisson distribution, we can find a very simple family of weight functions that results in different orderings of the original and weighted experiments. Indeed, if the distribution of X is a Poisson distribution with parameter θ and Y is its weighted version with a weight function of the form $w(x) = a^x$ for some specific value of a, then if $a < 1$, a random sample is preferred since $\mathcal{E}_X \succsim \mathcal{E}_Y$, while the

relation is reversed when $a > 1$, in which case we have $\mathcal{E}_Y \gtrsim \mathcal{E}_X$.

Last, we present an example to demonstrate that the effect of the *same* weight function on the information about different parameters of a multiparameter distribution can be dramatic. Assume that the distribution of the underlying X is Gamma with shape parameter α and scale parameter β and that the distribution of Y is a generalized size–biased version of it, that is, with $w(x) = x^a$ for some a. Then, if α is known (and β unknown), $\mathcal{E}_Y \gtrsim \mathcal{E}_X$, while if β is known (and α unknown), $\mathcal{E}_X \gtrsim \mathcal{E}_Y$.

5.5. Size–biased sampling

We finally explore the consequences of using the widely used size–biased weight in the distributions belonging to the "Big 3" families of distributions. Interestingly enough, the effect on each of them is different. Thus, assume that the distribution of X is one of the "Big 3": binomial, Poisson and negative binomial, and that Y is its size–biased version with $w(x) = x$. Then, if the distribution of X is Binomial, random samples are better, since $\mathcal{E}_X \gtrsim \mathcal{E}_Y$; if the distribution of X is Poisson, then both experiments are equivalent; if the distribution of X is negative binomial, then size–biased samples are better since $\mathcal{E}_Y \gtrsim \mathcal{E}_X$.

5.6. The criminal example

We would like to finish the paper with still another admittedly unrealistic example (but a very cute one) that shows how what at first might look as a very bad sample might in fact provide more information about the parameter of interest than a random sample that would anyway be impossible to get.

The general problem is to study the total population of criminals by sampling from just the criminals who are in jail. (We cannot, obviously, get a random sample from the whole population of criminals.)

We assume that each criminal commits crimes according to a Poisson process with his or her own rate of X crimes per year. We also assume that the density of X over the total

population of criminals is exponential with parameter θ, that is:

$$g(x|\theta) = \theta e^{-\theta x} \tag{5.6}$$

and that we want to make inferences about θ.

Suppose that each time a crime is committed, there is a fixed probability q that the criminal get caught and sent to jail. We will sample from criminals who have been sent to jail during the year. Then, the distribution $f(x|\theta)$ for the rates of crimes X of criminals who are in jail will be a weighted version of the original $g(x|\theta)$ where the weight function $w(x)$ is simply the probability that a criminal who commits crimes at rate x is sent to jail during the year, that is:

$$\begin{aligned} w(x) &= \Pr(\text{criminal is sent to jail during the year}) \\ &= 1 - \sum_{n=0}^{\infty} \Pr(\text{commits } n \text{ crimes and does not get caught}) \\ &= 1 - \sum_{n=0}^{\infty} \frac{e^{-x}x^n}{n!}(1-q)^n \\ &= 1 - e^{-qx}. \end{aligned} \tag{5.7}$$

Therefore, the distribution of X over the population of criminals who are in jail is given by:

$$f(y|\theta) \propto (1 - e^{-qy})g(y|\theta) = \frac{\theta(\theta+q)}{q}(1 - e^{-qy})e^{-\theta y}. \tag{5.8}$$

An easy calculation shows that

$$\begin{aligned} I_X(\theta) &= \frac{1}{\theta^2}, \\ I_Y(\theta) &= \frac{1}{\theta^2} + \frac{1}{(\theta+q)^2}. \end{aligned} \tag{5.9}$$

Since $I_Y(\theta) > I_X(\theta)$ for all θ, it follows rather surprisingly that, as far as Fisher information is concerned, it is *more* informative to sample from criminals who are in jail than to sample from the whole population of criminals.

Another interesting fact about the family of weights $\{w(y) = (1 - e^{-qy}),\ 0 < q < 1\}$ is worth mentioning. First notice from (5.9) that as q decreases to 0, $I_Y(\theta)$ increases to

$2/\theta^2$. Also, it can be seen from (5.8) that as $q \to 0$, $f(y|\theta) \to \theta^2 y e^{\theta y}$, but this is just the ordinary size–biased version ($w(x) = x$) of the exponential density $g(x|\theta)$ and we have seen that in this case, $\mathcal{E}_Y$ is actually sufficient for $\mathcal{E}_X$. In words, we have a family of weights for the exponential distribution that provides increasing Fisher information as $q \to 0$ and that in the limit leads to full sufficiency.

Acknowledgements

This work was supported in part by the Spanish Ministry of Education and Science under D.G.I.C.Y.T. grant number BE91–038.

References

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics* **24**, 3–61.

Bayarri, M.J., and DeGroot, M.H. (1987a). Bayesian analysis of selection models. *The Statistician* **36**, 137–146.

Bayarri, M.J., and DeGroot, M.H. (1987b). Information in selection models. In *Probability and Bayesian Statistics* (R. Viertl, ed.), 39–51. New York: Plenum Press.

Bayarri, M.J., and DeGroot, M.H. (1988). A Bayesian view of weighted distributions and selection models. In *Accelerated Life Testing and Expert's Opinions in Reliability* (C.A. Clarotti and D.V. Lindley, eds.), 70–82. Amsterdam: North–Holland.

Bayarri, M.J., and DeGroot, M.H. (1989). Comparison of experiments with weighted distributions. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.), 185–197. Amsterdam: Elsevier Science Publishers B.V. (North–Holland).

Bayarri, M.J., and DeGroot, M.H. (1990). Selection models and selection mechanisms. In *Bayesian and Likelihood Methods in Statistics and Econometrics* (S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, eds.) Amsterdam: Elsevier Science Publishers B.V. (North–Holland).

Bayarri, M.J., and DeGroot, M.H. (1991). The analysis of published significant results. In *Rassegna de Metodi Statistici ed Applicazioni* (W. Racugno, ed.). Bologna: Pitagora

(in press).

Bayarri, M.J., DeGroot, M.H., and Goel, P.K. (1989). Truncation, information, and the coefficient of variation. In *Contributions to Probability and Statistics* (L.J. Gleser, M.D. Perlman, S.J. Press and A.R. Sampson, eds.), 412–428. New York: Springer–Verlag.

Blackwell, D. (1951). Comparison of experiments. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 93–102. Berkeley, California: University of California Press.

Blackwell, D. (1953). Equivalent comparison of experiments. *Annals of Mathematical Statistics* **24**, 265–272.

Blumenthal, S. (1981). A survey of estimating distributional parameters and sample sizes from truncated samples. In *Statistical Distributions in Scientific Work* (C. Taillie, G.P. Patil, and B. Baldessari, eds.), Vol. 5, 75–86. Dordrecht: Reidel.

Blumenthal, S. and Sanathanan, L.P. (1980). Estimation with truncated inverse binomial sampling. *Communications in Statistics* **A9**, 997–1017.

Cohen, A.G. (1960). Estimation in truncated Poisson distributions when zeros and some ones are missing. *Journal of the American Statistical Association* **55**, 342–348.

Dahiya, R.G., and Gross, A.J. (1973). Estimating the zero class from a truncated Poisson sample. *Journal of the American Statistical Association* **68**, 731–733.

David, F.N., and Johnson, N.L. (1952). The truncated Poisson. *Biometrics* **8**, 275–285.

Dawid, A.P., and Dickey, J.M. (1977). Likelihood and Bayesian inference from selectively reported data. *Journal of the American Statistical Association* **72**, 845–850.

DeGroot, M.H. and Eddy, W.F. (1983). Set–valued parameters and set–valued statistics, in *Recent Advances in Statistics* (M.H. Rizvi, J.S. Rustagi, and D. Siegmund, eds.), 175–195. New York: Academic Press.

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

Fraser, D.A.S. (1952). Sufficient statistics and selection depending on the parameters. *Annals of Mathematical Statistics* **23**, 417–425.

Fraser, D.A.S. (1966). Sufficiency for selection models. *Sankhya A* **28**, 329–334.

Goel, P.K. and DeGroot, M.H. (1979). Comparison of experiments and information measures. *Annals of Statistics* **7**, 1066–1077.

Hansen, O.H. and Torgersen, E.N. (1974). Comparison of linear normal experiments. *Annals of Statistics* **2**, 367–373.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475–492.

Hedges, L.V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* Orlando, Florida: Academic Press.

Irwin, J.O. (1959). On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. *Biometrics* **15**, 324–326.

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem. *Statistical Science* **3**, 109–135.

Kahn, W.D. (1987). A cautionary note for Bayesian estimation of the binomial parameter *n*. *American Statistician* **41**, 38–40.

Kullback, S. (1968). *Information Theory and Statistics.* New York: Dover.

Little, R.J.A. (1985). A note about models for selectivity bias. *Econometrica* **53**, 1469–1474.

Patil, G.P. (1984). Studies in statistical ecology involving weighted distributions. In *Statistics: Applications and New Directions*, 475–503. Calcutta: Indian Statistical Institute.

Patil, G.P. and Taillie, C. (1987). Weighted distributions and the effects of weight functions on Fisher information. Technical Report, Center for Statistical Ecology and Environ-

mental Statistics, Department of Statistics, Pennsylvania State University.

Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions* (G.P. Patil, ed.), 320–333. Calcutta: Statistical Publishing Society.

Rao, C.R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In *A Celebration of Statistics: The ISI Centenary Volume* (A.G. Atkinson and S.E. Fienberg, eds.), 543–569. New York: Springer–Verlag.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* **86**, 638–641.

Sanathanan, L.P. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association* **72**, 669–672.

Stein, C. (1951). Information and the comparison of experiments. Unpublished report. University of Chicago.

Sterling, T.D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association* **54**, 30–34.

Stone, M. (1961). Non–equivalent comparisons of experiments and their use for experiments involving location parameters. *Annals of Mathematical Statistics* **32**, 326–332.

Torgersen, E.N. (1970). Comparison of experiments when the parameter space is finite. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **16**, 219–249.

Torgersen, E.N. (1972). Comparison of translation experiments. *Annals of Mathematical Statistics* **43**, 1383–1399.

Torgersen, E.N. (1976). Comparison of statistical experiments. *Scandinavian Journal of Statistics* **3**, 186–208.

Tukey, J.W. (1949). Sufficiency, truncation and selection. *Annals of Mathematical Statistics* **20**, 309–311.

Wachter, K.W. and Trussell, J. (1982). Estimating historical heights. *Journal of the American Statistical Association* **77**, 279–293.