# THE ANALYSIS OF PUBLISHED SIGNIFICANT RESULTS

by

M.J. Bayarri
University of Valencia
and Purdue University

and

M.H. DeGroot
Carnegie Mellon University

Technical Report # 91-21

Department of Statistics
Purdue University

April 1991

$\star$

# THE ANALYSIS OF PUBLISHED SIGNIFICANT RESULTS

by

M.J. Bayarri
University of Valencia
and Purdue University

and

M.H. DeGroot
Carnegie Mellon University

## ABSTRACT

In many areas of applied science, testing hypotheses is the most widely used statistical technique, if not the only one. Indeed, it is a well known fact that the vast majority of papers that appear in the applied scientific journals involve the testing of hypotheses. The overabuse of this technique has naturally led to an overappreciation of statistically significant findings, which has a number of undesirable consequences. First, many scientific journals encourage the publication of significant results in their publication policies, either explicitly or implicitly, so that articles with statistically significant results are more likely to get published. Also, the scientists themselves may regard their findings as useless unless they result in statistical significance and therefore not even submit them for publication. As a consequence, publications are highly biased towards reporting significant results and the question arises as to what can be concluded from such results. We will present a Bayesian analysis of this type of data. We use selection models to model the behavior of published significant results and show how significant results may in fact give very strong support to the null hypothesis. The selection mechanisms that may be give rise to the available data are explicitly considered.

Key Words and phrases: file–drawer problem, meta–analysis, selection mechanisms, selection models, statistical significance, weighted distributions.

## 1. Introduction

The usefulness of statistical methods in the analysis of experimental data is widely recognized in most areas of applied research. In spite of the great variety of statistical methods that have been developed for use in various types of problems, tests of hypotheses or tests of significance have become by far the most widely–used statistical methodology in published reports of research. Studies illustrating this assertion have been reported by Sterling (1959), Bakan (1966), and Bozarth and Roberts (1972) for the psychology literature, Zellner (1980) for economics, and DeGroot and Mezzich (1985) for psychiatry. It was found in these studies that in the vast majority of papers that made use of statistical methods, the culmination of the analysis was the performance of some significance tests and the reporting of the observed $p$–values.

Statisticians have only themselves to blame for this deplorable situation in which $p$–values are senselessly reported, with an asterisk or two, or even three, indicating that they are less than 0.10, 0.05, or 0.01, without any regard to the sample size or whether the "statistically significant" deviations from the null hypothesis are of any practical significance. In their zeal to have formal statistical methods applied to substantive problems, statisticians have allowed researchers to adopt the testing methodology in its most naive form in which experimental outcomes are regarded simply as significant or not significant, without further insight or interpretation. The shortcomings of this misuse of statistical methods has been discussed by several authors, including Cornfield (1975), Greenwald (1975), DeGroot (1980), Pocock (1980), and Salsburg (1985).

A natural consequence of this heavy emphasis on significance tests and the reporting of $p$–values is that statistical significance has become a primary goal of much of the applied research that is performed, with the result that most of the studies published in the professional literature report significant results. Thus, not only do journal editors tend to encourage the publication of articles in which statistical significance has been obtained (Melton, 1962; Greenwald, 1975), but also many experimenters themselves regard their results as being useless unless the results are statistically significant and will not even submit them for publication.

Discussion of this problem has reached beyond the scientific literature and has appeared in regular newspapers. In his column "The Doctor's World," in the *New York Times*, Altman (1986) writes:

> It verges on the scandalous, in the minds of many critics of modern science, that negative results are so seldom published...But pride and careers in science rarely benefit much from negative results. The Nobel Prize is rarely awarded for producing crucial negative information, no matter how skillful the scientist or how time–consuming the work.
>
> There are about 23,000 medical and scientific periodicals in the world, and most have a strong bias for publication of studies reporting positive results. Occasionally the editors of these journals publish studies with negative results, but the articles usually refute claims of prior positive results or are subjects of critical importance.
>
> The tendency to emphasize positive results skews the scientific process...

Altman goes on to point out that scientists who would find it useful to know certain negative results may needlessly repeat the work of others. Moreover, because of the close relationship between publications and research grants, scientists may be "more likely to submit safer proposals for grant support, ones they expect to yield positive results, instead of more imaginative, bolder experiments with a greater chance of failure." These concerns pertain to the information that we do not receive because of the experiments that are not performed or not published. In this paper, our central concern is with the appropriate interpretation of the information that we do receive. The policy of giving favorable treatment to the publication of statistically significant results can lead to substantial selection bias that must be taken into account. (See Begg and Berlin, 1988, for discussions and references.)

In this paper, we address the problem of properly analyzing results from experiments that are subject to this type of publication bias due to statistical significance. We use selection models to describe the behavior of published significant results. In Section 2

we introduce the formulation of weighted distributions and selection models and discuss briefly Bayesian inference for the latter. In Section 3 they are applied to the modelling of published significant results. In Section 4 the important question of whether we should take into account the selection mechanisms is addressed. A general formulation is given in section 5 and two particular selection mechanisms that are of interest when studying published significant results are further elaborated in Sections 6 and 7. Finally, in section 8, the results of previous sections are applied to the analysis of published significant results, including in a natural way the "file drawer" effect.

## 2. SELECTION MODELS AND WEIGHTED DISTRIBUTIONS

We will use weighted distributions, and in particular selection models, to describe the statistical behavior of published significant results. In this section, we discuss these models briefly. A more extended description can be found in Bayarri and DeGroot (1987,1988) together with many examples and references.

In many practical situations, a random sample from the population of interest is not available so that special models that incorporate the bias of the observations must be developed. Weighted distributions and selection models occur naturally in contexts in which the probability (or density) that a particular observation enters the sample gets multiplied by some (non–negative) weight function, or in contexts in which data is available just from some selected portions of the population.

Assume that an observable random variable $X$ is distributed over the population of interest according to the (generalized) density $g(x|\theta)$, and that it is desired to make inferences about $\theta$. The usual statistical analysis assumes that a random sample from $g(x|\theta)$ is available. Assume instead that the probability or density of observing a particular value $x$ of $X$ gets multiplied by the (non–negative) weight $w(x)$ so that the observed sample is, in fact, a random sample from

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

Distributions represented by (2.1) were called weighted distributions or weighted versions of $g(x|\theta)$ by Rao (1965) who first unified the formulation of these models, although their

use can be traced to Fisher (1934). Good surveys on the topic are Patil (1984) and Rao (1985).

We will be especially concerned with a special case of weighted distributions called selection or truncation models. In these models, observations can be obtained just on a subset $S$ of the sample space, so that the weight function $w(x)$ is simply the indicator function of the so–called selection set $S$,

$$w(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{if } x \notin S. \end{cases} \tag{2.2}$$

Thus, a selection sample (a random sample from (2.1) with the above weight function) is a random sample from the density:

$$\begin{aligned} f(x|\theta) &= \frac{w(x)g(x|\theta)}{\Pr(X \in S|\theta)} \quad \text{if } x \in S, \\ &= 0 \qquad \text{otherwise.} \end{aligned} \tag{2.3}$$

In this paper, we deal mainly with the case in which the selection set $S$ is the upper tail of $g(x|\theta)$ containing all values $X \geq \tau$, where $\tau$ is a fixed value, so that the joint density of a selection sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is

$$f(\boldsymbol{x}|\theta) = \frac{\prod_{i=1}^{n} g(x_i|\theta)}{[1 - G(\tau|\theta)]^n}, \tag{2.4}$$

where $G$ is the c.d.f. corresponding to $g(\cdot)$.

We will particularly consider the case where $g(x|\theta)$ is the pdf of a normal distribution with mean $\theta$ and variance 1 and just a single observation is obtained from $f(x|\theta)$. The Bayesian analysis, although straightforward, can not be carried out in closed form due to the dependence of (2.4) on $\theta$ through $G(\tau|\theta)$. In general, Bayesian analysis is simpler when carried out under conjugate families. In this case, however, the choice of an appropriate prior distribution for $\theta$ is not clear. The natural election of a conjugate prior density for $\theta$ would be directly based on the form of the likelihood function for $\theta, \ell(\theta)$, which is given by

$$\ell(\theta) \propto \frac{\phi(x - \theta)}{1 - \Phi(\tau - \theta)}, \tag{2.5}$$

and would therefore be given by

$$\xi(\theta) \propto \frac{\exp[-h(\theta - m)^2/2]}{[1 - \Phi(\tau - \theta)]^a}, \tag{2.6}$$

where the hyperparameters $h, m$, and $a$ are specified constants and $\Phi$ is the c.d.f. of the standard normal distribution. Although the form of any conjugate prior is related to the experiment to be performed, the explicit dependence of (2.6) on $\tau$ can be very inappropriate, and this will indeed be the case when selection models are used to analyze statistical significant reports, since $\tau$ will then be fixed by routine statistical practice and will not convey any information about $\theta$.

If relative simplicity of calculation is desired, it seems more appropriate to use a prior distribution which is conjugate with respect to the original model $g(x|\theta)$, so that in this case we would simply choose a normal prior distribution for $\theta$. This election is in fact equivalent to taking $a = 0$ in (2.6) thus eliminating the undesired dependence on $\tau$, but not the difficulties in the posterior analysis, which remains essentially unchanged.

If the prior distribution for $\theta$ is normal with mean $m_0$ and precision (inverse of the variance) $h_0$, then the posterior p.d.f. $\xi(\theta|x)$ is of the form (2.6) with $a = 1$ and

$$m = \frac{h_0 m_0 + x}{h_0 + 1}, \qquad h = h_0 + 1. \tag{2.7}$$

It should be noted that the expressions for $m$ and $h$ are the same as those that would be obtained for the hyperparameters of the conjugate posterior distribution for $\theta$ were $x$ an observation from the unrestricted density. We will also use this distribution to give expressions for the posterior mean and variance. Specifically, for any function $\psi(\theta)$, let $E^u[\psi(\theta)]$ denote expectation with respect to the normal distribution with mean $m$ and precision $h$ as given in (2.7), that is, with respect to the posterior distribution that would be obtained if the value $x$ were observed in the unrestricted model. Furthermore, let

$$s(\theta) = [1 - \Phi(\tau - \theta)]^{-1}. \tag{2.8}$$

Then, the mean and variance of $\theta$ under the posterior p.d.f. $\xi(\theta|x)$ obtained when $x$ is observed in the selection model are:

$$\begin{aligned} E(\theta|x) &= E^u[\theta s(\theta)]/E^u[s(\theta)] \\ \mathrm{Var}(\theta|x) &= \{E^u[\theta^2 s(\theta)]/E^u[s(\theta)]\} - [E(\theta|x)]^2. \end{aligned} \tag{2.9}$$

The expectations in (2.9) have to be evaluated numerically or otherwise approximated in some fashion. A simple approximation is the one obtained when the delta method is applied to each of the factors in (2.9), yielding:

$$\frac{E^u[s(\theta)]}{s(m)} \approx 1 + \frac{1}{2hM(\tau-m)}\left[\frac{2}{M(\tau-m)} - (\tau-m)\right]$$
$$\frac{E^u[\theta s(\theta)]}{s(m)} \approx 1 + \frac{1}{2hM(\tau-m)}\left[m\left(\frac{2}{M(\tau-m)} - (\tau-m)\right) - 2\right], \qquad (2.10)$$

where $M(\cdot)$ denotes Mill's ratio (inverse of the hazard–rate function for the standard normal distribution) defined by

$$M(x) = \frac{1-\Phi(x)}{\phi(x)}. \qquad (2.11)$$

## 3. MODELING PUBLISHED SIGNIFICANT RESULTS

We will use a very simple (unrealistically so) example so that the technicalities will not obscure the main points that we are trying to put forward. The framework will be that of one–sided tests of hypotheses on the mean of a normal distribution with known variance.

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 $\mu$ and known variance $\sigma^2$ and the uniformly most powerful test is used for testing

$$H_0: \mu \leq 0$$
$$H_1: \mu > 0, \qquad (3.1)$$

at some level $\alpha$. In this case, the distribution $g(t|\theta)$ of the test statistic

$$T = \frac{\sqrt{m}}{\sigma}\,\overline{Y}_m, \qquad (3.2)$$

is normal with mean $\theta = \frac{\sqrt{m}}{\sigma}\,\mu$ and variance 1, where $\overline{Y}_m$ represents, as usual, the sample mean in a given experiment. The restriction to equal sample sizes $m$ and (known) variances

$\sigma^2$ is, of course, quite unrealistic and it is used here just to ease the presentation. A similar argument would apply to more realistic settings and the concerns that we will raise would remain virtually unchanged.

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 concluded from this experiment? Do the published significant results provide strong evidence against $H_0$?

The first reaction of most readers is 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 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:

$$f(t|\theta) = \frac{\phi(t-\theta)}{1-\Phi(\tau-\theta)} \qquad \text{for } t \geq \tau, \tag{3.3}$$

where $\phi$ and $\Phi$ are the p.d.f. and c.d.f. of the standard normal distribution, and $\tau = \Phi^{-1}(1-\alpha)$. In the analysis presented here, assume that the experiment is performed repeatedly until a significant result is obtained and published. Without this assumption, the probability of observing a significant result must be taken into account. This point will be discussed further in the next section.

When a statistically significant result is analyzed using the model (3.3) it may very well provide strong support for the null hypothesis. To see this, it is enlightening to calculate the maximum likelihood estimator of $\theta$ for a given observed value of $T$. It can be shown from (3.3) that the maximum likelihood estimator $\hat{\theta}$ is the unique solution to the equation

$$(t-\hat{\theta})M(\tau-\hat{\theta}) = 1, \tag{3.4}$$

where Mill's ratio $M(\cdot)$ is given in (2.11). The values of $\hat{\theta}$ for $\alpha = 0.01, 0.05$ and $0.10$ are given in Table 1. Of course, the most widely used criteria 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 appear published in the scientific journal, 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 $\alpha$.

| $\hat{\theta}$ | $\alpha = 0.01$ | | $\alpha = 0.05$ | | $\alpha = 0.10$ | |
|---|---|---|---|---|---|---|
| | $t$ | $p$ | $t$ | $p$ | $t$ | $p$ |
| 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 $\theta$ for various observed values $t$ and $p$.

Since $\hat{\mu} = \sigma\hat{\theta}/\sqrt{m}$ it follows from (3.1) 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 $H_0$. Moreover, a published $p$–value only slightly smaller than the effective $\alpha$ can be regarded as supporting $H_0$ rather than rejecting $H_0$. In fact, as the $p$–value approaches $\alpha$ in this selection model, $\hat{\theta} \to -\infty$.

From a practical point of view, it would be nice if we could find a rough, easy to compute, threshold value $t_0$ for the observed $t$ such that the null hypothesis is supported ($\hat{\theta}$ negative) for observed values of $T \leq t_0$. Of course, it follows from (3.4) that the exact value for $t_0$ is $t_0 = 1/M(\tau)$. On the other hand, it is a cute empirical fact that $\hat{\theta}$ is negative when $p$, the $p$–value associated with the observed $t$ is, roughly, greater than (0.4) $\alpha$. We found the explanation for this fact in conversations with Satish Iyengar and it goes as follows:

It is well known (see, i.e., Feller, 1968, Vol. 1) that for large $\tau$:

$$\frac{1}{M(\tau)} \approx \tau + \frac{1}{\tau} + \ldots \tag{3.5}$$

Therefore, the $p$–value associated to $t_0 = 1/M(\tau)$ can be approximated by

$$p = 1 - \Phi(t_0) \approx 1 - \Phi\left(\tau + \frac{1}{\tau}\right). \tag{3.6}$$

But (see, i.e. Feller, 1968, Vol. 1)

$$\lim_{\tau \to \infty} \frac{1 - \Phi(\tau + \frac{1}{\tau})}{1 - \Phi(\tau)} = \frac{1}{e} = 0.3679. \tag{3.7}$$

Since $1 - \Phi(\tau) = \alpha$ we have that, as $\tau$ goes to $\infty$ (or $\alpha$ goes to 0),

$$P \approx \frac{1}{e}\alpha = (0.3679)\alpha. \tag{3.8}$$

Thus, for small values of $\alpha$, the MLE $\hat{\theta}$ will be negative when the observed $p$–value $p$ is greater than $\alpha/e$.

## 4. WORRYING ABOUT WHAT YOU DON'T SEE

Assume that our hypothetical published experiment reports an observed value of $t = 1.844$ (which would also be the reported MLE $\hat{\theta}$ of $\theta$), with associated $p$–value 0.0326. Data is thus declared significant at level $\alpha = 0.05$.

Nevertheless, being aware of the publication bias effect, we do not take these quantities at face value, but use instead the selection model $f(t|\theta)$, carrying out the analysis of the preceding section and finding $\hat{\theta} = -3$ (see table 1), an estimate of $\theta$ which is at least three standard deviations away from the parameter values in $H_1$. We thus conclude that the data is, in fact, providing very strong support to $H_0$.

When using selection models, however, it has to be kept in mind that we are in fact making the implicit assumption that, regardless of the true value of $\theta$ and regardless of any other unpublished experimental outcomes, experimentation is continued by at least one experimenter until a significant result is obtained and published. Under different conditions, the analysis should proceed differently.

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 $g(t|\theta)$. In this case, we should conclude that $\hat{\theta} = 1.844$ and that the data does provide evidence against $H_0$. Here we know that if the experiment had yielded a non–significant result, we would not have seen any published report at all and thus the lack of a report would have given us information about $\theta$ as well.

Intermediate conclusions between these two extremes can be obtained by taking into consideration the selection rule that produced the observed selection sample. The first and most evident conclusion is that what we do *not* get to observe, namely the number of performed experiments, can decisively influence the statistical analysis of the reported data.

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, ch. 14) and Iyengar and Greenhouse (1988) among others. The effect does not pertain only to the analysis of significant reports, but it is more general. Indeed, when dealing with selection models, it can happen that the way in which the selection sample was produced, that is, the selection mechanism generating the observed sample, has a decisive influence on the statistical analysis of data and, therefore, it might have to explicitly be taken into consideration. In the remainder of the paper, we will briefly present a general discussion of the selection mechanisms that might have generated a selection sample and we will apply them afterwards to the particular case of reported significant results.

## 5. SELECTION MECHANISMS

Consider again the general formulation in which a selection sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ of size $n$ is obtained from some underlying density $g(x|\theta)$ restricted to the selection set $S$. In general, both $\theta$ and $S$ are unknown, with $\theta \in \Omega$ and $S \in \mathcal{S}$. When we model the joint density of the selection sample $\boldsymbol{x}$ by

$$f(\boldsymbol{x}|\theta, S) = \frac{\prod_{i=1}^{n} g(x_i|\theta)}{[\Pr(S|\theta)]^n} \text{ for } x_i \in S,\ i = 1, 2, \ldots, n, \tag{5.1}$$

we are implicitly assuming that the number $n$ of observations was fixed in advance or, in any event, that the observed $n$ contains no information about $\theta$ or $S$.

In general, however, depending on the selection mechanism used to generate $\boldsymbol{x}$, $n$ itself may contain useful information. For example, in many problems a small value of $n$ might indicate that observations in $S$ are hard to obtain, which in turn might imply that $\Pr(S|\theta)$ is small. In recognition of this fact we will consider $n$ as the realized value of an observable random variable and it will be explicitly introduced in the notation. Thus, although once $\boldsymbol{x}$ is observed $n$ is automatically also observed, it will be convenient to denote our data by the pair $(\boldsymbol{x}, n)$.

To present a general discussion of the possible selection mechanisms giving rise to $(\boldsymbol{x}, n)$, it is useful to think in terms of a (usually fictional) investigator that somehow produces a selection sample from the underlying density $g(x|\theta)$ and a statistician who can only observe the selection sample $(\boldsymbol{x}, n)$. Accordingly, assume that the investigator observes a sequential random sample $y_1, y_2, \ldots$ from the underlying $g(y|\theta)$. Also, for $i = 1, 2, \ldots$, let $z_i$ be one or zero according to whether $y_i$ is to be or not included in the selection sample, that is:

$$z_i = I_S(y_i) = \begin{cases} 1 & \text{if } y_i \in S, \\ 0 & \text{if } y_i \notin S. \end{cases} \tag{5.2}$$

We will let $\delta$ denote the stopping rule used by the investigator to determine when to stop sampling, about which we will make the following assumptions: i) $\delta$ depends on $y_1, y_2, \ldots$ only through the observed values of $z_1, z_2, \ldots$ ii) $\delta$ is such that with probability one sampling will eventually terminate. When sampling terminates, let $N$ denote the total number of observations $y_i$, let $n = \sum_{i=1}^{N} z_i$ be the number of observations $y_i$ in $S$, and denote their observed values by $x_1, x_2, \ldots, x_n$. The statistician only observes $(\boldsymbol{x}, n)$ and in general doesn't know either the stopping rule $\delta$ used nor all the values $y_1, y_2, \ldots, y_N$ obtained. Since, for given $\theta$ and $S$, $z_1, z_2, \ldots$ is a sequence of Bernoulli trials and since $\delta$ depends only on $z_1, z_2, \ldots$, what we are in fact assuming is that $\delta$ is a binomial sampling plan of the general type described by Girshick, Mosteller and Savage (1946), and DeGroot (1959).

Once $(\boldsymbol{x}, n)$ is observed, the general goal of the statistician is to compute the joint posterior density $p(\theta, S|\boldsymbol{x}, n)$ which, by Bayes theorem, is given by

$$p(\theta, S|\boldsymbol{x}, n) \propto p(\boldsymbol{x}|n, \theta, S)p(n|\theta, S)p(\theta, S). \tag{5.3}$$

Here, as in the rest of the paper, the symbol $p$ is used to denote an arbitrary density without any implications that it is the same for all variables. We will now discuss the first two factors on the right–hand side of (5.3).

If $\delta$ were known, then the statistician would compute the conditional posterior density

$$p(\theta, S|\boldsymbol{x}, n, \ \delta) \propto p(\boldsymbol{x}|n, \theta, S, \delta)p(n|\theta, S, \delta)p(\theta, S|\delta). \tag{5.4}$$

Because $\delta$ is a binomial sampling plan, it follows that $p(\boldsymbol{x}|n, \theta, S, \delta)$ is given by the right–hand side of (5.1). That is, given the observed value $n$, the observations $\boldsymbol{x}$ form a selection sample irrespective of the particular $\delta$ used. Therefore,

$$p(\boldsymbol{x}|n, \theta, S, \delta) = p(\boldsymbol{x}|n, \theta, S) = f(\boldsymbol{x}|\theta, S) = \frac{\prod_{i=1}^{n} g(x_i|\theta)}{[\Pr(S|\theta)]^n}, \tag{5.5}$$

gives also the expression for the first factor on the right–hand side of (5.3).

In general, the statistician is uncertain about the stopping rule $\delta$ that was used to produce $(\boldsymbol{x}, n)$. Let $p(\delta|\theta, S)$ denote the statistician's prior distribution over some available set $\Delta$ of possible sampling plans (as our notation indicates, we allow the possibility that the statistician's prior beliefs on $\delta$ depend on his or her prior beliefs on $\theta$ and $S$.) Then:

$$p(\theta, S|\boldsymbol{x}, n) = \sum_{\delta \in \Delta} p(\theta, S, \delta|\boldsymbol{x}, n) \propto p(\boldsymbol{x}|n, \theta, S) \left\{ \sum_{\delta \in \Delta} p(n|\theta, S, \delta)p(\delta|\theta, S) \right\} p(\theta, S), \tag{5.6}$$

and the second factor in (5.3) is in general given by

$$p(n|\theta, S) = \sum_{\delta \in \Delta} p(n|\theta, S, \delta)p(\delta|\theta, S). \tag{5.7}$$

If $p(n|\theta, S)$ above does not depend on $(\theta, S)$, then (5.3) can be expressed as

$$p(\theta, S|\boldsymbol{x}, n) \propto p(\boldsymbol{x}|n, \theta, S)p(\theta, S),$$

so that the particular selection mechanism used can be ignored and data is analyzed according to the selection model. A note of caution is in order. Notice that our use of the term *ignorable* in this context results in an interpretation that differs from the usual one that is given to the term. Here that the selection mechanism is *ignorable* means that data is analyzed according to the selection model.

We next discuss two particular sampling plans that will be used to describe possible selection mechanisms that might occur in the publishing of significant results scenario: sampling plans in which the number of observations in $S$ is fixed and sampling plans in which the total number of observations is fixed. (A more detailed discussion of selection mechanisms in this context can be found in Bayarri and DeGroot, 1990.)

## 6. PLANS WITH FIXED $n$

Consider first inverse binomial sampling plans $\delta$ in which sampling is continued until a fixed number of observations $n$ lying in $S$ have been obtained. In this set–up, a particular sampling plan $\delta$ can in fact be characterized by the fixed value $n$ that it requires, and $\Delta$ can be taken to be the set of all positive integers. Accordingly, sampling plans in this section will be denoted by $\delta_n$ (and sometimes simply by $n$).

If $\delta_{n_0}$ is a given sampling plan that specifies sampling until $n_0$ observations in $S$ have been obtained, then $p(n|\theta, S, \delta_{n_0})$ in (5.7) is a degenerate distribution concentrated in the single value $n = n_0$, therefore

$$p(\theta, S|\boldsymbol{x}, n, \delta_{n_0}) \propto p(\boldsymbol{x}|n_0, \theta, S)p(\theta, S) \qquad \text{for } n = n_0, \tag{6.1}$$

and the selection mechanism is ignored.

In general, the value $n_0$ to be used in the sampling plan may not be known by the statistician until the selection sample has actually been observed. In this case, he or she would assess $p(\delta_n|\theta, S)$ for $n = 1, 2, 3, \ldots$, and once $n_0$ is observed,

$$p(\theta, S|\boldsymbol{x}, n_0) \propto p(\boldsymbol{x}|n_0, \theta, S)p(\theta, S)p(\delta_{n_0}|\theta, S), \tag{6.2}$$

so that if the probability that the statistician assigns to the investigator fixing the actual, observed value $n_0, p(\delta_{n_0}|\theta, S)$, does not depend on $\theta, S$ (as it is very frequently assumed to

be the case), then the analysis is again based on $p(\boldsymbol{x}|n_0, \theta, S)$ and the selection mechanism is ignored.

In the situation that we are considering in this section, the total number $N$ of observations taken by the investigator (the "file drawer" size in the context of published significant results) is usually not known to the statistician, so that he or she may want to learn about it. In this case, the goal is to compute the joint posterior distribution

$$p(\theta, S, N|\boldsymbol{x}, n) \propto p(\boldsymbol{x}|n, \theta, S, N)p(n|\theta, S)p(N|n, \theta, S)p(\theta, S), \tag{6.3}$$

where $p(\boldsymbol{x}|n, \theta, S, N) = p(\boldsymbol{x}|n, \theta, S)$ is given by the selection model (5.5). Also, if we let $P = \Pr(S|\theta)$, then the probability that $N$ observations are needed to achieve $n$ of them in $S$ is given by

$$p(N|n, \theta, S) = \binom{N-1}{n-1} P^n (1-P)^{N-n}, \; N = n, n+1, \ldots \tag{6.4}$$

and if we make the usual assumption that $p(n|\theta, S)$ does not depend on $(\theta, S)$, then the joint posterior (6.3) can be expressed as

$$p(\theta, S, N|\boldsymbol{x}, n) \propto \frac{\prod_{i=1}^{n} g(x_i|\theta)}{P^n} \binom{N-1}{n-1} P^n (1-P)^{N-n} p(\theta, S), \tag{6.5}$$

from which a marginal posterior distribution for $N$ can be derived.

It should be stressed that, as $N$ is *not* observed, the mere introduction of it into the formulation does not provide any information about $(\theta, S)$ and should not change the inferences about $(\theta, S)$. This is indeed true in a Bayesian analysis of the problem since inferences about $(\theta, S)$ will be based on the marginal joint posterior density for $(\theta, S)$ which, from (6.5),is given by

$$p(\theta, S|\boldsymbol{x}, n) \propto \frac{g_n(\boldsymbol{x}|\theta)}{P^n} p(\theta, S), \tag{6.6}$$

and thus it is irrelevant whether or not we wish to learn about $N$ as well. This may not be the case when a likelihood analysis is carried out. In fact, if learning about $N$ is not of interest, then a likelihood analysis would be based on the likelihood function

$$\ell(\theta, S; \boldsymbol{x}, n) \propto \frac{g_n(\boldsymbol{x}|\theta)}{P^n} \tag{6.7}$$

whereas if $N$ is also of interest, the likelihood function would be

$$\ell(\theta, S, N; \boldsymbol{x}, n) \propto g_n(\boldsymbol{x}|\theta)\binom{N-1}{n-1}(1-P)^{N-n} \tag{6.8}$$

and the likelihood analysis based on these two different likelihood functions will usually differ. In particular, the MLE of $\theta$ and $S$ will be different depending on which likelihood function, (6.7) or (6.8), is used. Thus, the mere consideration of the possibility of learning about the unobserved $N$ can change the inferences (here the MLE) about $\theta$ and $S$, in flagrant contradiction with the likelihood principle.

## 7. PLANS WITH FIXED $N$

In this section we will consider sampling plans $\delta$ in which the number of observations $N$ to be drawn from the underlying density $g(x|\theta)$ is fixed in advance. In other words, suppose that the investigator draws a random sample of fixed size $N$ from $g(x|\theta)$ and then reports to the statistician those observations among these $N$ that happen to fall in $S$. Any such $\delta$ can be characterized by the $N$ that it specifies; because of the importance that these plans will have in the next section, we will abuse notation a little so that a particular sampling plan $\delta_N$ which specifies sampling till $N$ observations are obtained will simply be denoted by $N$.

For any such sampling plan $N$, the joint posterior of $(\theta, S)$ can be expressed as

$$p(\theta, S|\boldsymbol{x}, n, N) \propto p(\boldsymbol{x}|n, \theta, S)p(n|\theta, S, N)p(\theta, S). \tag{7.1}$$

But the probability of obtaining $n$ observations in $S$ is, for any given $\theta, S$ and $N$,

$$p(n|\theta, S, N) = \binom{N}{n}P^n(1-P)^{N-n}, \tag{7.2}$$

where, as before, $P = \Pr(S|\theta)$. Thus,

$$p(\theta, S|\boldsymbol{x}, n, N) \propto g_n(\boldsymbol{x}|\theta, S)(1-P)^{N-n}p(\theta, S), \tag{7.3}$$

so that the selection mechanism can not be ignored. This is not a surprising conclusion, since with this selection mechanism, the observed value of $n$ does carry information about $P = \Pr(S|\theta)$.

In general, the statistician is uncertain about the actual $N$ used. We will see that in this situation the selection mechanism can sometimes be ignored. This is quite remarkable, since we have just seen that when $N$ is known, and for all values of $N$, it can *never* be ignored.

When $N$ is uncertain, the statistician quantifies this uncertainty by $p(N|\theta, S)$ so that

$$p(\theta, S|\boldsymbol{x}, n) \propto p(\boldsymbol{x}|n, \theta, S) \left\{ \sum_{N=n}^{\infty} p(n|\theta, S, N) p(N|\theta, S) \right\} p(\theta, S), \tag{7.4}$$

and according to the comments in previous sections, the selection mechanism can be ignored (and data analyzed according to the selection model) if

$$p(n|\theta, S) = \sum_{N=n}^{\infty} \binom{N}{n} P^n (1-P)^{N-n} p(N|\theta, S) \tag{7.5}$$

does not depend on $(\theta, S)$. We will next present two examples of priors $p(N|\theta, S)$ for which this result holds.

In the first one, the conditional distribution of $N$ is assumed to be Poisson with mean $\lambda/P$, where $\lambda$ is a fixed and known constant. This is not unreasonable, since the less likely the set $S$ is, the larger the mean of this distribution is taken to be. That is, the statistician believes that, since the investigator is to report only observations in $S$, then the investigator is likely to fix large values of $N$ when observations in $S$ are difficult to get. It follows from (7.5) and the assumption just made that

$$\begin{aligned} p(n|\theta, S) &= \sum_{N=n}^{\infty} \frac{1}{n!(N-n)!} \left( \frac{1-P}{P} \right)^{N-n} e^{-\frac{\lambda}{P}} \lambda^N \\ &= \frac{\lambda^n}{n!} e^{-\lambda} \sum_{k=0}^{\infty} \frac{1}{k!} \left[ \left( \frac{1-P}{P} \right) \lambda \right]^k e^{-\frac{1-P}{P}\lambda}, \end{aligned} \tag{7.6}$$

so that $p(n|\theta, S)$ is a Poisson distribution with mean $\lambda$, and since it does not depend on $(\theta, S)$, the selection mechanism is to be ignored.

In the second example, $p(N|\theta, S)$ will itself not depend on $(\theta, S)$, and will be taken to be an improper prior distribution. Specifically, suppose that

$$p(N|\theta, S) \propto \frac{1}{N} \text{ for } N = 1, 2, \ldots \tag{7.7}$$

Then, for any positive integer $n$, it follows from (7.5) that

$$p(n|\theta, S) \propto \frac{1}{n} \sum_{N=n}^{\infty} \binom{N-1}{n-1} P^n (1-P)^{N-n} = \frac{1}{n} \tag{7.8}$$

which is also improper and does not depend on $(\theta, S)$. (Notice, however, that an improper uniform prior for $N$ will not produce a similar result). Thus, again, the selection mechanism is ignored.

We will finish the paper by applying the two particular sampling plans discussed in the last two sections to our example of a published significant result.

## 8. SELECTION MECHANISMS FOR A PUBLISHED SIGNIFICANT RESULT

Consider again the example described in Sections 3 and 4, in which we read the published report of an experiment in which a one–sided test of hypothesis on the mean of a normal distribution with known variance was carried out. Based on the $N(\theta, 1)$ distribution for the test statistic $T$, data was declared significant yielding a small $p$–value. We also assumed that non–significant results did not get published so that, as readers of the journal, we can only observe values of $T$ such that $T \geq \tau = \Phi^{-1}(1-\alpha)$; $\tau$ was 1.96 for $\alpha = 0.05$. Since we are considering just one published experiment, the size of the selection sample in this example is $n = 1$. The two extreme conclusions reached in Section 4, as well as a wide variety of intermediate ones, can now very easily be expressed in terms of specific sampling plans that could have been used to produce the observed report.

For example, if we believe that the experiment under consideration was performed repeatedly, possibly by different experimenters in different laboratories around the world, until a significant result was obtained and published, then the selection mechanism is based on an inverse binomial sampling plan with $n_0 = 1$. Hence, as explained in Section 6, the selection mechanism can be ignored and the published result should be analyzed according to the selection model (3.3), so that the posterior distribution for $\theta$, as given by (6.1), is

$$p(\theta|t, n_0 = 1) \propto \xi(\theta) \frac{\phi(t-\theta)}{1-\Phi(\tau-\theta)}, \tag{8.1}$$

where $\xi(\theta)$ is the prior distribution for $\theta$. The assumptions made could be appropriate for

instance to analyze experiments that are cheap and easy to carry out, not requiring overly specialized skills or equipment,...,etc.

On the other hand, if we believe that the published significant result corresponds to the only experiment of this type that has been performed, then the selection mechanism is based on a sampling plan of fixed total sample size $N = 1$. In this case, it follows from the discussion in Section 7 that the selection mechanism can not be ignored so that the selection model (3.3) is no longer appropriate to analyze $t$. In fact, it follows from (7.3) that the posterior distribution for $\theta$ is in this case

$$p(\theta|t, n = 1, N = 1) \propto \phi(t - \theta)\xi(\theta) \tag{8.2}$$

so that the analysis of the observed value of $t$ is based on the original, underlying distribution $N(\theta, 1)$, implying that $t$ and its associated $p$–value should be taken at face value, as common sense suggested. The assumptions made could be appropriate if, for instance, the experiment under consideration is very expensive, or takes a very long time to complete, or one that involves a large number of scientists or very sophisticated instrumentation,...,etc.

The file drawer problem can also be described in terms of a sampling plan of fixed total sample size $N$. $N$ would here be the total number of performed experiments and its value can not be observed. A Bayesian approach requires the specification of a prior distribution $p(N|\theta, \tau)$, so that the posterior distribution for $\theta$, as given by (7.4) is

$$p(\theta|t, n = 1) \propto \phi(t - \theta)\xi(\theta) \sum_{N=1}^{\infty} N[\Phi(\tau - \theta)]^{N-1}\, p(N|\theta, \tau). \tag{8.3}$$

The analysis will be based on the selection model $\phi(t - \theta)/[1 - \Phi(\tau - \theta)]$ if $p(N|\theta, \tau)$ is such that

$$\sum_{N=1}^{\infty} N[1 - \Phi(\tau - \theta)][\Phi(\tau - \theta)]^{N-1}\, p(N|\theta, \tau) \tag{8.4}$$

does not depend on $\theta$, as it is in the case when $p(N|\theta, \tau)$ corresponds to a Poisson distribution with mean $\lambda/\Phi(\tau - \theta)$ for any fixed $\lambda$, or when it is taken to be inversely proportional to $N$. Otherwise, the full analysis based in (8.3) has to be carried out.

If it is not desired to carry out the full Bayesian analysis, we strongly recommend that, at least, conditional posterior distributions as given by (7.3), namely

$$p(\theta|t, n = 1, N) \propto \phi(t - \theta)[1 - \Phi(\tau - \theta)]^{N-1}\xi(\theta), \tag{8.5}$$

or their associated quantities of interest (estimators, HPD regions,...,etc.) be computed for a "likely" range of values of $N$ to tentatively investigative how sensitive our conclusions are to the value of $N$. If they are sensitive, a fully Bayesian analysis, with maybe different types of priors for $N$, is very advisable.

## ACKNOWLEDGEMENTS

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

## REFERENCES

Altman, L.K. (1986). Negative results: a positive viewpoint. *New York Times*, Tuesday, April 19, 1986.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin* **66**, 432–437.

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

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. (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.), 211–227. Amsterdam:Elsevier Science Publisher B.V. (North–Holland).

Begg, G.B. and Berlin, J.A. (1988). Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society A* **151**, 419–463.

Bozarth, J.D., and Roberts, R.R. (1972). Signifying significant significance. *American Psychologist* **27**, 774–775.

Cornfield, J. (1975). A statistician's apology. *Journal of the American Statistical Association* **70**, 7–14.

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. (1959). Unbiased sequential estimation for binomial populations. *Annals of Mathematical Statistics* **30**, 80–101.

DeGroot, M.H. (1980). Remarks on the statistical analysis of hypotheses. In *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 204–206.

DeGroot, M.H. and Mezzich, J.E. (1985). Psychiatric statistics. In *A Celebration of Statistics: The ISI Centenary Volume* (A.G. Atkinson, S.E. Fienberg, eds.), 145–165. New York:Springer–Verlag.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (3rd ed.). New York:John Wiley and Sons.

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

Girshick, M.A., Mosteller, F. and Savage, L.J. (1946). Unbiased estimates for certain binomial sampling problems with applications. *Annals of Mathematical Statistics* **17**, 13–23.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin* **88**, 359–369.

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

Iyengar, S. and Greenhouse, J.B. (1988). Selection models and the file drawer problem.

*Statistical Science* **3**, 109–135.

Melton, A.W. (1962). Editorial. *Journal of Experimental Psychology* **64**, 553–557.

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

Pocock, S.J. (1980). The role of statistics in medical research. *British Journal of Psychiatry* **137**, 188–190.

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.

Salsburg, D.S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician* **39**, 220–223.

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.

Zellner, A. (1980). Statistical analysis of hypotheses in economics and econometrics. In *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 199–203.