A Selection Bias Conflict and Frequentist Versus Bayesian Viewpoints *

Micha Mandel

The Hebrew University of Jerusalem, Israel, 91905

and

Yosef Rinott

The Hebrew University of Jerusalem, Israel, 91905

and LUISS Rome, Italy

May 18, 2009

*Micha Mandel is Lecturer, Department of Statistics, The Hebrew University, Jerusalem 91905, Israel (email: msmic@huji.ac.il) Yosef Rinott is Professor, Department of Statistics and Center for the Study of Rationality, The Hebrew University, Jerusalem 91905, Israel (email: rinott@mscc.huji.ac.il), and LUISS, Rome. The work of Yosef Rinott was partially supported by Israel Science Foundation grant 473/04. We thank Ester Samuel-Cahn, Daniel Yekutieli, Manor Askenazi, and Marco Scarsini for their helpful comments.

Abstract

In many branches of modern science, researchers first study or mine large data sets, and then select the parameters they estimate and the data they use and publish. Such data-based selection complicates formal statistical inference. An example discussed here for the purpose of illustration, is that of pharmaceutical companies that typically conduct many experiments but may publish only selected data. The selection often depends on the outcomes of the experiments since naturally there is interest in potentially useful drugs, and it is in general unclear how it should affect inference. Is this effect the same for the company and the public? Does it matter if they are Bayesian or Frequentist? Should the company reveal all experiments it conducts, and if so, how should this change the conclusions?

This note discusses these questions in terms of a simple example of a sequence of binomial experiments conducted by a pharmaceutical company, where results are published only if the number of 'failures' is small. We do not suggest that this example corresponds to reality in the pharmaceutical industry, nor in science in general; our goal is to elaborate on the importance and difficulties of taking selection into account when performing statistical analysis.

KEYWORDS: confidence interval, credible set, Binomial model, decision theory, meta analysis, publication bias

1 INTRODUCTION

A recent paper by Psaty and Kronmal (2008) summarizes the events that led the pharmaceutical company Merck & Co., Inc. to withdraw the drug Rofecoxib (trademark Vioxx) from the market. The paper provides an extraordinary opportunity to compare the internal analyses of a pharmaceutical company to the reported analyses, and it exemplifies well the concerns appearing in Psaty and Kronmal's abstract that *Sponsors have a marketing interest to represent their products in the best light. This approach conflicts with scientific standards that require the symmetric and comparable reporting of safety and efficacy data. Selective reporting of the results of clinical trials can misrepresent the risk-benefit profile of drugs.*

This conflict of interests triggered editors of leading medical journals to announce that they would refuse to publish drug research sponsored by pharmaceutical companies unless the studies are registered in a public database from the outset. This initiative was reported in an article in the Washington Post on Sept 10 2004 (http://www.smh.com.au/articles/2004/09/09/1094530773888.html). The article explains the logic of this act by pointing out that *More than two-thirds of studies of anti-depressants given to depressed children, for instance, found the medications were no better than sugar pills, but companies published only the positive trials. If all the studies had been registered from the start, doctors would have learned that the positive data were only a fraction of the total.* For earlier references that suggest the creation of registries of research studies in connection with selection and meta analysis, see, for example, Iyengar and Greenhouse (1988) and references therein.

More generally, it is well known that scientists and scientific journals tend to submit and publish significant results, while the existence of other results remains unknown. This problem, known as publication bias (Rothstein, Sutton, and Borenstein 2005), is similar to problems that arise in models having a large number of parameters (e.g., DNA and similar data), where inference is published only on a few parameters selected by the data, usually on those corresponding to the 'statistically significant' findings. However, methods for controlling error probabilities in multiple testing and estimation (e.g., Benjamini and Yekutieli 2005) require data which due to selection is not published. Iyengar and Greenhouse (1988) and Cleary and Casella (1995) provide a truncation model formulation that has many common features with the present paper, and certain solutions to the publication bias problem under specific assumptions, as well as numerous references.

The purpose of this paper is to shed some light on the difficulty of trying to take data selection or publication bias into account, and the dangers of not doing so, by means of a very simple binomial example. We formulate our discussion in terms of a conflict between a pharmaceutical company and a regulator in Phase I studies, only as an illustration of a conflict of interests that may arise in the presence of selection bias. We do not claim to know how severe this conflict is in reality. Furthermore, Phase I clinical trials are followed by Phase II and Phase III studies, which provide certain control over the drugs or doses selected in Phase I.

In Section 2 we explain why in certain situations, selection bias implies that frequentist confidence intervals of a given level are constructed differently according to the possibly conflicting points of view of the company performing the study and the public that uses them. Section 3 approaches the problem created by this selection bias conflict using the Bayesian paradigm. Although Bayesian analysis is post-data in nature, it can solve the difficulties of selection bias appearing in the frequentist paradigm only under rather special and questionable assumptions discussed in Section 3. Section 4 extends the selection criterion of Sections 2 and 3, and Section 5 concludes with a comparative discussion of frequentist and Bayesian analyses under publication bias. It also discusses the ability of the aforementioned initiative of registering studies in a public database to solve the concerns that arise by selection of results.

2 A BINOMIAL MODEL

Pharmaceutical companies perform Phase I toxicity studies on drugs and doses on a regular basis in order to screen out drugs that have significant adverse effects. Because of competition, they disclose only the minimum information necessary to get their drugs approved. The mission of public agencies such as the FDA is to ensure that approved drugs are indeed safe for use.

Consider such a pharmaceutical company and suppose that for each experiment it conducts, the company statistician computes a confidence interval for θ, the probability of an adverse reaction to the drug. Each interval is based on the number of toxic cases in the given experiment, which is a Binomial random variable $X \sim \text{Bin}(n, \theta)$, where n is the number of patients in the experiment, assumed fixed.

Typically, there is a maximal probability of adverse reaction, θ^M, corresponding to the Maximum Tolerated Dose (MTD), which is the maximal value of θ acceptable for a given disease. A drug would be accepted if all values in a computed confidence interval lie below θ^M. For severe diseases, such as certain types of cancer, this value may be up to $\theta^M = 0.2$ or 0.25.

There are several methods for constructing $1-\alpha$ level confidence intervals for θ, see, e.g., Brown, Cai and DasGupta (2001) for a review. Here we consider the conservative method of Clopper and Pearson (1934) that, for $X = x$, solves for θ the equations $P_{\underline{\theta}}(X \geq x) = \alpha/2$ and $P_{\overline{\theta}}(X \leq x) = \alpha/2$ and defines the confidence interval $CP(X) = [\underline{\theta}\,,\ \overline{\theta}\,]$. The function $P_\theta(X \leq x)$ is decreasing in θ so that multiple solutions do not exist. However, for $x = 0$, the first equality has no solution and the interval uses 0 as its left limit. Likewise, for $x = n$, the right limit of the interval is 1.

For example, for $n = 20$ and $\alpha = 0.05$, this yields for $X = 7$ the interval $CP(7) = [0.1539, 0.5922]$. The interval $CP(7)$ includes high toxicity levels that under normal circumstances are not acceptable, hence the drug will not be approved. Here we consider a scenario in which drugs or doses associated with experiments resulting in such a high number of cases of adverse reactions are deemed useless by the company, and therefore such experiments are not disclosed to the public. Thus, only successful experiments, say with $X \leq c$ adverse reactions for some c, and their associated confidence intervals are published and no information on unsuccessful experiments, including their number and outcomes, is revealed. For simplicity, we first discuss the case of $c = 1$. We elaborate on the choice of c in Section 4, where it is also shown that the main issues discussed below are not specific to a particular choice of this parameter.

Continuing the example of $n = 20$ and $\alpha = 0.05$, if $c = 1$ then only $X = 0$ or 1 would be considered. The Clopper-Pearson intervals are $CP(1) = [0.0013, 0.2487]$ and $CP(0) = [0, 0.1684]$, possibly suggesting an acceptable level of toxicity.

Suppose a regulator claims that the selection policy of the company distorts the results, and, however large θ is, the outcomes $X = 0$ or 1 will occur eventually. And so, the regulator says, in order to achieve good results, all the company has to do is to perform sufficiently many experiments. Therefore, he insists that published intervals should be computed on the basis of the random variable X^* having the distribution of $X|X \leq 1$, i.e.,

a Bernoulli variable with probability

$$P_\theta(X^* = 1) = P_\theta(X = 1 | X \leq 1) = \frac{n\theta}{1 + (n-1)\theta} \equiv \theta^*(\theta) \equiv \theta^*. \tag{1}$$

For $\alpha < 1/2$, it is easy to verify that a $1 - \alpha$ level confidence interval for θ^* is

$$CI^*(X^*) := \begin{cases} [0, 1-\alpha] & X^* = 0 \\ [\alpha, 1] & X^* = 1 \ , \end{cases} \tag{2}$$

an interval satisfying $P_\theta(CI^*(X^*) \ni \theta^*) \geq 1 - \alpha$ for all possible values of θ. For $X^* = 0$ and $X^* = 1$, $CI^*(X^*)$ is equivalent to $0 \leq \theta^*(\theta) \leq 1 - \alpha$, and $\alpha \leq \theta^*(\theta) \leq 1$, respectively. Rearrangement as intervals for θ yields confidence intervals denoted by $CI(X^*)$, which for $X^* = 0$ and 1 are

$$CI(0) = \left[0, \frac{1-\alpha}{\alpha n + 1 - \alpha}\right], \qquad CI(1) = \left[\frac{\alpha}{(1-\alpha)n + \alpha}, 1\right]. \tag{3}$$

Thus, the resulting regulator's confidence intervals are (for $n = 20$, $\alpha = 0.05$): $CI(0) = [0, 0.4872]$ and $CI(1) = [0.0026, 1]$; they are strikingly different from those of the company statistician. The regulator can claim that the data prove nothing, since for $X^* = 1$, the confidence interval covers almost the whole range $[0, 1]$, and for $X^* = 0$ the probability of adverse reaction could be almost as high as 0.5, a level that could never be tolerated. Thus the company may claim that all published experiments indicate an acceptable level of toxicity, whereas the regulator claims that the published data lead to no such conclusions.

In general, as n increases, the interval $CI(1)$ includes large values of θ and becomes wider, eventually approaching $[0, 1]$. This may surprise a reader who expects that one success out of n indicates a small θ. However, taking the selection criterion $X \leq 1$ into account, the result is not surprising: for $\theta > 0$ and increasing n, an outcome of $X^* = 1$ occurs with probability approaching 1, and hence it is hardly informative, resulting in a large confidence interval that contains 1, rather than proving a small θ. The interval $CI(0)$ equals $[0, 1 - \alpha]$ for $n = 1$ and its upper limit decreases with n, as expected. Thus, for large n, an experiment that results in 0 adverse reactions may indeed prove the safety of the treatment even to the regulator. However, under commonly acceptable values of θ, the probability of 0 adverse reactions, that by (1) is bounded by $\frac{1-\theta}{n\theta}$, becomes small rather fast as n increases.

In order to better understand the differences and the conflict between the company and the regulator, we consider the following probabilistic model. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ be a

sequence of probabilities $0 < \theta_i < 1$, and let $X_i \sim \text{Bin}(n, \theta_i)$, $i = 1, 2, \ldots$, be independent binomial experiments. Let $T_0 = 0$ and define successively $T_j = \min\{i > T_{j-1} : X_i \leq 1\}$, $j = 1, 2, \ldots$, the indices of 'successful' experiments; we assume that $\boldsymbol{\theta}$ is such that $P_{\boldsymbol{\theta}}(T_j < \infty) = 1$ for all j.

The company statistician computes a confidence interval for each θ_i such that,

$$P_{\boldsymbol{\theta}}(CP(X_i) \ni \theta_i) \geq 1 - \alpha \qquad \text{for all } \boldsymbol{\theta} \in (0,1)^\infty. \tag{4}$$

This guarantees (see Berger 1985 pp 22-23) $\liminf_{N\to\infty} N^{-1} \sum_{i=1}^{N} I\{CP(X_i) \ni \theta_i\} \geq 1 - \alpha$ with probability 1, for all $\boldsymbol{\theta} \in (0,1)^\infty$, where $I\{E\}$ is the indicator of the event E. For $\alpha = 0.05$, for example, it can be loosely interpreted as a rule that in the long run allows 5% of all intervals, published or not, to miss the target parameters. Thus, the company uses a criterion that concerns all intervals, but discloses only selected ones.

The regulator is aware only of the experiments that are published, namely, the successful ones, and suggests the criterion

$$P_{\boldsymbol{\theta}}(CI(X_{T_j}) \ni \theta_{T_j}) \geq 1 - \alpha \qquad \text{for all } \boldsymbol{\theta} \in (0,1)^\infty, \tag{5}$$

that takes into account the distribution of the observations X_{T_j}. The regulator's view is that the confidence intervals must reflect the selection made by the company.

The conflict becomes very apparent if one imagines a dishonest company that repeats the same experiment with the very same dose (and hence the same θ) until a success ($X \leq 1$) is obtained, and calculates confidence intervals ignoring the fact that the probability θ is the same in all experiments. For example, suppose that the true probability of a toxic reaction is $\theta = 0.25$ so that the number of toxic cases follows a $\text{Bin}(20, 0.25)$ law. Under this model, the probability of a successful experiment is $0.75^{20} + 20 \times 0.25 \times 0.75^{19} \approx 0.0243$, so that on average, for every 40 experiments performed, the company will achieve one successful experiment and publish its corresponding confidence interval. This published interval includes only θ values below 0.25 as we saw above.

3 A BAYESIAN PERSPECTIVE

Failing to agree on a frequentist confidence measure, the company and the regulator may consult the Bayesian school. The two sides of the conflict hope for a useful advice from a Bayesian, knowing that Bayesians, unlike frequentists, analyze the observed data regardless

of the design (see Ferguson 1967 Chapter 7.2, and Berger 1985 Chapter 7.7 for a comprehensive discussion). A relevant example in our context is the Stopping Rule Principle in sequential analysis: *the stopping rule [terminating the data collection] should have no effect on the final reported evidence about θ obtained from the data* (Berger 1985 p. 502). This principle is espoused by Bayesians, but not necessarily by frequentists. A simple instance would be a sequence of N Bernoulli experiments which ends with one failure (see Lindley and Phillips 1976 for a similar example). A frequentist must know the experimental design: was N fixed in advance, or is it a Geometric random variable, whereas in Bayesian analysis the nature of N turns out to have no effect on the calculations.

Returning to the sequence of toxicity experiments, let $\mathbf{X}$ denote the whole data and let $T_j(\mathbf{x}) = t$ be the index of the j-th successful experiment. For any prior $\mathbf{\Pi}$ of the sequence $\boldsymbol{\theta}$, it is easy to see that the posterior distribution of $\theta_{T_j}|\mathbf{X} = \mathbf{x}$ is equal to the posterior of $\theta_t|\mathbf{X} = \mathbf{x}$, and as above, the reason for stopping is irrelevant. Therefore, if the regulator and the company were to use the same data and agree on the prior model, then their inference for θ_{T_j} would be the same in spite of the selection.

However, in the situation considered here, the regulator and the company do not have the same data, hence their inferences may be different as they condition on different events. Indeed, here the regulator observes X_t only for those t such that $T_j = t$, whereas the company has the values of X_i for all i. Thus, the company should base its inference on $\theta_t|\mathbf{X} = \mathbf{x}$, while the regulator should base his inference on $\theta_t|\{X_{T_1}, X_{T_2}, \ldots\}$.

In what follows, we consider two extreme models for the joint distribution of $\boldsymbol{\theta}$, a model of independence and a model of strong dependence, in order to demonstrate how the implications of Bayesian inference on the company vs. regulator conflict may be sensitive to the choice of a prior.

Independent Parameters. With the lack of better information on the dependence structure, a convenient though arguable assumption on the prior which we consider first is that $\theta_1, \theta_2, \ldots$ are independent identically distributed with a (marginal) prior law Π and density π. We also use the standard assumption that the experiments are designed so that $X_1, X_2, \ldots|\boldsymbol{\theta}$ are independent and $X_i|\boldsymbol{\theta} \sim \text{Bin}(n, \theta_i)$.

Under this model, X_i is sufficient for θ_i in the sense that $P(\theta_i \in I|X_1 = x_1, X_2 = x_2, \ldots) = P(\theta_i \in I|X_i = x_i)$ for any interval I, and therefore the posterior of $\theta_t|\mathbf{X} = \mathbf{x}$ is equal to that of $\theta_t|X_t = x_t$. Hence, Bayesian inference of the regulator ought to be the

same as that of the company (if they agree on the marginal prior Π), and it is unaffected by the selection of data and parameters.

For example, for the widely used flat prior $U(0,1)$, the posterior is $\text{Beta}(x+1, n-x+1)$ and the 95% equal-tails credible intervals for $n = 20$, are [0.0012, 0.1611] and [0.0117, 0.2382] for $X = 0$ and $X = 1$, respectively. These intervals are quite similar to those obtained and published by the company statistician using the frequentist procedure above. Therefore, the company may be indifferent to whether the Frequentist or the Bayesian paradigm is adopted, and may prefer the latter simply because it solves the conflict with the regulator.

Strong Dependence. The results of the previous model rely on the prior $\mathbf{\Pi}$ which assumes independence of the θ_i's. Next we contrast the independence assumption with the extreme case in which $\theta_i \equiv \theta$ for all i and $\theta \sim \Pi$. Such a model may be of interest to a suspicious regulator who aims to protect against a dishonest company that repeats the very same experiment until obtaining a successful result (this model is similar to that discussed at the end of Section 2).

When $\theta_i \equiv \theta$ for all i, there is exactly one draw from the prior distribution Π that determines the parameters of all the binomial experiments. For simplicity consider now inference based on $X^* \equiv X_{T_1}$ whose conditional probability $P(X^* = 1|\theta)$ is given in (1). The Bayesian $1-\alpha$ credible interval for θ is an interval I whose posterior probability $P(\theta \in I|X^* = x) = \int_I \pi(\theta|X^* = x)d\theta = 1-\alpha$.

For $\Pi = \text{U}(0,1)$, the posterior *distribution functions* for $X^* = 0$ and $X^* = 1$ are

$$\Pi(\theta|X^* = 0) = \int_0^\theta \frac{(1-u)\Big/[1+(n-1)u]du}{\int_0^1 (1-v)\Big/[1+(n-1)v]dv} = \frac{n\log\{1+(n-1)\theta\} - (n-1)\theta}{n\log(n) - (n-1)},$$

and

$$\Pi(\theta|X^* = 1) = \int_0^\theta \frac{nu\Big/[1+(n-1)u]du}{\int_0^1 nv\Big/[1+(n-1)v]dv} = \frac{(n-1)\theta - \log(1+(n-1)\theta)}{n-1-\log(n)},$$

(using $\int v/(a+bv)dv = b^{-2}\{bv - a\log(a+bv)\} + const$). These distributions are depicted in Figure 1 together with the prior distribution. The limits of the $1-\alpha$ level equal-tails credible intervals are the solutions for θ of $\Pi(\theta|X^* = x) = \alpha/2$ and $\Pi(\theta|X^* = x) = 1-\alpha/2$, see Figure 1. For $\alpha = 0.05$ and $n = 20$, these are $[0.0029, 0.6982]$ and $[0.0620, 0.9778]$ for $X^* = 0$ and 1, respectively. The intervals are considerably different from those obtained

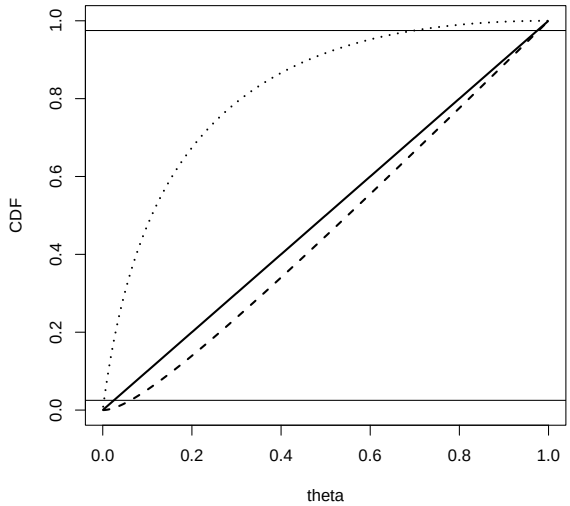

Figure 1: Distribution functions under the model $\theta_i \equiv \theta$ for all i; $n = 20$: solid line - the prior U(0,1), dashed line - the posterior for $X^* = 1$, dotted line - the posterior for $X^* = 0$. The 95% equal-tails credible intervals are defined by the points where the posterior distributions cross the horizontal $y = 0.025$ and $y = 0.975$ lines.

under the independence model above, and are much more similar to the confidence intervals of the frequentist regulator given after (3).

If the company insists on a prior consisting of independent θ_i's while the regulator assumes the same θ in all experiments, then the regulator and the company are in conflict. In fact, from the discussion in the previous section it is seen that the company and the regulator will have the same posterior distribution only when X_t is sufficient for θ_t. Thus, the Bayesian paradigm appears to solve the conflict only by imposing rather strict and questionable assumptions. Without such assumptions the conflict that arose under the frequentist paradigm remains unresolved. A similar phenomenon has been recently demonstrated by Senn (2008) for a normal model.

It is shown in the Appendix that when $\theta_i \equiv \theta$ the posterior distribution of θ given X^* is stochastically decreasing in n when observing either one or zero successes. This implies that smaller values of θ would be predicted as n increases. It is interesting to note that when the experiment results in one success, the posterior distribution is stochastically larger than the prior (see Figure 1), and converges to it; thus observing one success indicates to the Bayesian that θ is larger than initially believed. This is somewhat similar to the behavior of the confidence intervals of (3), whose limits decrease with n. The limiting behavior of

the posterior, discussed in details in the Appendix, is somewhat similar to that observed by Cleary and Casella (1995), who study p-values under a Bayesian publication bias model with a flat prior.

4 The General Binomial Case

So far, we have examined the publication bias problem for general n, and the case $c = 1$, and gave numerical examples for $n = 20$. The following discussion is aimed at showing that the selected example of $n = 20$ and the selection criterion of $X \leq c = 1$ is not a case of selection bias on our part. Indeed, the nature of the results does not depend on c or n. In this section we consider the regulator-company conflict only in frequentist terms; some Bayesian examples are given in the Appendix.

Let us consider how a company might choose its selection criterion c. Suppose that a toxicity level of at most $\theta^M = 0.25$ is acceptable, as may be the case in drugs for cancer in progressive stages, and the company decides to publish only those treatments that are considered safe, that is, experiments whose right limit of the associated Clopper-Pearson confidence interval is smaller than 0.25. For a fixed n, the selection criterion of $X \leq c$ is determined by $c = \max\{x : P_{0.25}(X \leq x) \leq \alpha/2\}$, i.e., the largest integer such that observing any $X \leq c$ gives rise to a confidence interval with right limit smaller than 0.25. For example, for $\alpha = 0.05$ and $n = 100$, a simple calculation shows that $CP(16) = [0.094, 0.247]$ and $CP(17) = [0.102, 0.258]$, hence results are published by the company only if $X \leq c = 16$. For the case of $n = 20$ mentioned above, this criterion gives $c = 1$.

Being aware of the company's policy, the regulator considers a published experiment X_c^* as having the distribution of $X \mid X \leq c$, and calculates a frequentist confidence interval by

$$CI_c(x) = \{\theta : \alpha/2 \leq P_\theta(X_c^* \leq x)\} \cap \{\theta : \alpha/2 \leq P_\theta(X_c^* \geq x)\}. \tag{6}$$

For a fixed n, this is indeed an interval as $P_\theta(X_c^* \leq x)$ can be shown to be non-increasing in θ (see Appendix A), and for $X_c^* = x$, the left and right limits are the solutions for θ of the equations $P_\theta(X_c^* \geq x) = \alpha/2$ and $P_\theta(X_c^* \leq x) = \alpha/2$, respectively. As in the Clopper-Pearson intervals, the left limit for $X_c^* = 0$ is 0 and the right limit for $X_c^* = c$ is 1.

Since $P_\theta(X_c^* \leq x) = P_\theta(X \leq x)/P_\theta(X \leq c) > P_\theta(X \leq x)$ for all θ and $c < n$, the right

limit of the confidence interval $CI_c(x)$ is always larger than the right limit of the Clopper-Pearson interval $CP(x)$ that does not take selection into account. Thus, it may well happen that the company's intervals are below $\theta^M = 0.25$, say, whereas the regulator's intervals contain values larger than θ^M. The disagreement on the right limit of the confidence interval is clear-cut when $x = c$. In this case, by the choice of c, the company's confidence interval has a right limit $< \theta^M = 0.25$, whereas the regulator's interval's right limit is 1. On the other hand, the disagreement is often negligible for small values of x, as the example below shows. It is also interesting to note that the left limit of $CI_c(x)$ is larger than the corresponding left limit of $CP(x)$ since $P_\theta(X_c^* \geq x) = P_\theta(X \geq x | X \leq c) \leq P_\theta(X \geq x)$.

Figure 2 compares the confidence intervals $CP(x)$ and $CI_c(x)$ for $5 \leq x \leq 16$, where $\alpha = 0.05$, $n = 100$, and $c = 16$. For $x < 5$ the intervals are practically identical. The figure shows the dramatic effect of selection on the intervals for values close to c. The company may claim that all the experiments it publishes prove acceptable toxicity, while the regulator would approve only those experiments that result in $x \leq 11$ adverse reactions, because only these intervals lie below $\theta^M = 0.25$. Here, the conflict between the company and the regulator is in the range $12 \leq x \leq 16$ adverse reactions.

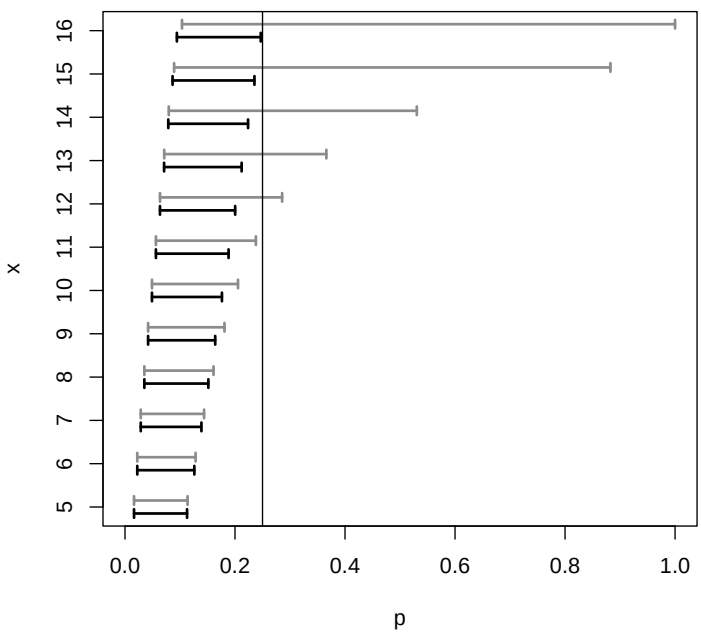

Figure 2: Comparison of confidence intervals with (gray) and without (black) taking selection into account for the model $n = 100$ and c=16.

Does this conflict arise frequently? Are experiments with fewer than 12 adverse reactions common or rare? The answer, of course, depends on the real unknown probability of an adverse reaction, θ. If θ is rather small compared to $\theta^M = 0.25$ of the current example, then the conflict arises rarely: in the above example, if $\theta = 0.05$, then $P_\theta(X_c^* \leq 11) = 0.996$, and therefore almost all experiments will indicate low toxicity of the treatment regardless of whether selection is taken into account or ignored. However, if the true probability of an adverse effect is close to the accepted probability, then the conflict arises for most published experiments: in the above example, if $\theta = 0.2$, then $P_\theta(X_c^* \leq 11) = 0.065$, hence the conflict arises for almost 95% of the published experiments. Since efficacy of a treatment is most often inversely related to safety, one may expect that effective drugs tend to have doses with a probability of an adverse reaction close to the maximal tolerable one θ^M. Therefore, here $\theta = 0.2$ is much more relevant than $\theta = 0.05$, and the conflict is expected to arise often.

5 Concluding Remarks

For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results. This citation is taken from the abstract of a well-known article by Rosenthal (1979) on publication bias. In this note, we consider the extreme version of the "file drawer problem", as defined by Rosenthal, and highlight the difficulties in constructing confidence measures for parameters that are selected on the basis of the observed data.

In the binomial example described here, the frequentist confidence intervals that take selection into account almost always contain both low and high probabilities of toxicity and are hardly informative, indicating the difficulties involved in the analysis of selected data. The most popular frequentist approach for dealing with selection is a correction for multiple tests, which requires knowledge of the total number of experiments performed. Such methods are useless when the number of experiments is unknown. Furthermore, even if it were known, it would be very large in the situations considered here, and a correction would necessarily lead to very wide confidence intervals under any multiple comparison

method.

Bayesians assume a prior distribution for the parameters (probabilities of an adverse reaction to the drug) and their inference on these parameters is sensitive to the choice of the prior. The most common criticism of the Bayesian paradigm concerns the subjective choice of the prior and the way it affects inference (see Berger, 1985 for a discussion). The claim that for large sample size the effect of the prior diminishes may not be true when selection is involved, and the role of the prior is much more important, as demonstrated in Section 3 (see also Dawid 1994, Mandel and Rinott 2007, Senn 2008, and the last paragraph of the Appendix). In particular, Bayesian inference can ignore selection (of data and parameters) only when observed data are sufficient for the observed parameters, e.g., when the parameters are independent. This is a very strong assumption that should not be overlooked.

It is interesting to ask if and how the initiative mentioned in Section 1, of registering experimental data from the outset, can help in dealing with the problem of publication bias. There is definitely a need to prevent misrepresentation of results by selecting and reporting only the successful ones. A purely scientific approach is to avoid selection and publish all experiments, regardless of their outcome, in a unified manner. But then, who is going to publish or read the *Annals of Insignificant Results*?

Perhaps the good news is that the worst case assumption that all drugs have an unacceptable toxicity level is most likely not true, and there exist effective and relatively non-toxic drugs. Both the frequentist and the Bayesian will in general rank drugs from less to more toxic in a similar order, and the disagreement is on the criteria for the final decision of which drugs to accept, rather than their relative quality. A frequentist who applies a reasonable multiple comparison criterion and a Bayesian whose prior assumes that good drugs do exist, are likely to discover them in real life situations, when the samples are large enough.

Appendix

A. Monotonicity of $P_\theta(X_c^* \leq x)$

Here we prove Monotonicity of $P_\theta(X_c^* \leq x)$ which assures that the confidence intervals of (6), based on X_c^*, are indeed intervals. This monotonicity property will also be used in

part B of the Appendix.

Lemma .1. *Let X_c^* be a random variable having the conditional distribution of $X|X \leq c$. Then $P_\theta(X_c^* \leq x)$ is non-increasing in θ, that is, X_c^* is stochastically increasing in θ.*

Proof. Let Y and X be random variables having a common support and probability functions (or densities) g and f. Y is said to be larger than X in the likelihood ratio order, denoted $Y \geq_{lr} X$, if g/f is nondecreasing on the common support. It is well known that $Y \geq_{lr} X$ implies $Y \geq_{st} X$, that is, Y is stochastically larger than X (see, e.g., Lehmann 1991, p. 85).

It is easy to see that $Y \geq_{lr} X$ implies $Y|\{Y \in A\} \geq_{lr} X|\{X \in A\}$ and hence $Y|\{Y \in A\} \geq_{st} X|\{X \in A\}$ for any subset A of the support. Thus, in order to prove that $P_\theta(X_c^* \leq x) = P_\theta(X \leq x|X \leq c)$ is non-increasing in θ, it is enough to show that if $\theta_1 < \theta_2$, $X \sim \text{Bin}(n, \theta_1)$ and $Y \sim \text{Bin}(n, \theta_2)$, then $Y \geq_{lr} X$. This simple fact is left as an exercise. □

B. On posterior distributions

The following propositions describe several simple but interesting properties of the posterior distributions under the model $\theta_i = \theta$ for all i.

Let $X|\theta \sim \text{Bin}(n, \theta)$ and $\theta \sim \Pi$, any prior, and for $x = 0, \ldots, c$ let $\Pi_n(\cdot|x)$ be the posterior cdf of $\theta|X_c^* = x$, where X_c^* has the distribution of $X|X \leq c$, for some fixed c.

Proposition .1. *The sequence of distributions $\{\Pi_n(\cdot|x)\}_{n>c}$ is stochastically decreasing in n for $0 \leq x \leq c$, that is, $\Pi_{n-1}(\theta|x) \leq \Pi_n(\theta|x)$ for all $0 < \theta < 1$ and all $n > c$.*

Also, for any $n > c$, $\Pi_n(\cdot|x)$ is stochastically larger than Π for $x = c$, and smaller for $x = 0$.

Proof. We prove the stronger likelihood ratio order. The posterior density associated with $\Pi_n(\theta|x)$ is

$$\pi_n(\theta|x) = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}/F_\theta(c;n)}{\int_0^1[\binom{n}{x}t^x(1-t)^{n-x}/F_t(c;n)]\pi(t)dt}\pi(\theta), \tag{7}$$

where $F_\theta(c;n) = P_\theta(X \leq c)$ denotes the $\text{Bin}(n, \theta)$ distribution function at c. Direct calculations show that $\pi_n(\theta|x)/\pi_{n-1}(\theta|x) \propto (1-\theta)F_\theta(c;n-1)/F_\theta(c;n)$, and the first part of the proposition will be proved if we show that the last expression is non-increasing in θ. Writing $(1-\theta)F_\theta(c;n-1) = \sum_{x=0}^{c}\binom{n}{x}\theta^x(1-\theta)^{n-x}(n-x)/n$, it is readily seen that

$(1-\theta)F_\theta(c;n-1)/F_\theta(c;n) = E_\theta\{(n-X_c^*)/n\}$. The problem reduces to showing that $E_\theta(X_c^*)$ increases in θ, which follows from Lemma .1.

For the second part of the proposition note that by (7), $\pi_n(\theta|0)/\pi(\theta) \propto P_\theta(X_c^* \leq 0)$, which by Lemma .1 is non-increasing in θ, proving likelihood ratio order. Similarly, for $x = c$ we use the relation $\pi_n(\theta|c)/\pi(\theta) \propto 1 - P_\theta(X_c^* \leq c-1)$. □

Proposition .2. *For $0 \leq x \leq c$ we have:*
(i) If $E_\Pi(\frac{1}{\theta})^{c-x} = \infty$, then $\lim_{n\to\infty} \Pi_n(\theta \mid x) = I\{\theta \geq 0\}$, the cdf of a r.v. degenerate at 0.
(ii) If $E_\Pi(\frac{1}{\theta})^{c-x} < \infty$, then $\lim_{n\to\infty} \Pi_n(\theta \mid x) \propto \int_0^\theta (\frac{1-t}{t})^{c-x}\pi(t)dt$.

Proof. For $\theta > 0$ we obtain by writing $F_t(c;n)$ in (7) explicitly and straightforward cancellations

$$1 - \Pi_n(\theta|x) = \int_\theta^1 \pi_n(t|x)dt = \frac{\int_\theta^1 \left(\sum_{k=0}^c (\frac{1-t}{t})^{x-k}[\binom{n}{k}/\binom{n}{c}]\right)^{-1}\pi(t)dt}{\int_0^1 \left(\sum_{k=0}^c (\frac{1-t}{t})^{x-k}[\binom{n}{k}/\binom{n}{c}]\right)^{-1}\pi(t)dt}.$$

For $k < c$, $\binom{n}{k}/\binom{n}{c} \to 0$ monotonically as $n \to \infty$. Therefore, each of the integrands (with respect to π) above, converges to $(\frac{1-t}{t})^{c-x}$. By monotone convergence, the integral in the numerator converges to $\int_\theta^1 (\frac{1-t}{t})^{c-x}\pi(t)dt$ which is always finite. The integral in the denominator converges to $\int_0^1 (\frac{1-t}{t})^{c-x}\pi(t)dt$. If the latter integral diverges, the resulting distribution $\Pi_n(\theta|x)$ is clearly degenerate at 0, and *(i)* follows. Otherwise, *(ii)* obtains. □

We conclude by discussing some interesting implications of Proposition .2. For $x = c$ the condition in *(ii)* holds trivially, and we have convergence to the prior, that is, $\lim_{n\to\infty} \Pi_n(\theta \mid c) = \Pi(\theta)$ for all θ. Focusing for simplicity on the case of $c = 1$, we see that observing $X^* = 1$ for large n, leads a Bayesian to stick to his prior. The frequentist also sees this observations as almost non informative, as reflected by the interval (3), which converges to $[0, 1]$.

By (3) again, the confidence intervals of a frequentist who observes $X^* = 0$ will converge to $[0, 0]$. A Bayesian with a prior satisfying the expectation condition in *(i)* will have a posterior that converges to a distribution concentrated at zero, thus agreeing with the frequentist on the confidence interval for large n. This joint conclusion makes a lot of sense. For arbitrarily large n, one may expect to observe $X^* = 0$ only for correspondingly small θ's.

On the other hand, consider a Bayesian whose prior assigns small enough measure to small values of θ, so that the expectation condition in *(ii)* holds, e.g. Beta(2,1). By

Proposition .1, the posterior π_n satisfies $\pi_n(t) \geq_{st} C(\frac{1-t}{t})^{c-x}\pi(t)$, where C is a normalizing constant. Obviously, the latter density is not concentrated at 0 and this Bayesian may not exclude the possibility of positive values of θ. Thus, observing $X^* = 0$ even for huge values of n, he will always consider some positive θ's plausible. For example, for $c = 1$, a prior of Beta(2,1) implies $\Pi(.25) = 0.06$, making small values of θ seem quite unlikely. The posterior, given $X^* = 0$, for any n, is stochastically larger than the Beta(1,2) distribution by Propositions .1 and .2. It turns out that this posterior satisfies $\Pi_n(0.25 \mid X^* = 0) < 0.5$ for all n, and a Bayesian who observes $X^* = 0$ would not reject the possibility that $\theta > 0.25$ for any n. The limiting 0.95 equal tails Bayesian credible interval is $[0.0126, 0.8419]$ and the shortest credible interval in this case is $[0, 0.7764]$. Both intervals are in contrast to the frequentist's confidence intervals of (3), whose left and right limits converges to 0 with n.

Finally, we remark on the sensitivity of Bayesian credible intervals to the choice of prior in the present setup. These intervals can be arbitrarily different in cases (i) and (ii) of Proposition .2. However, if π_1 satisfies (i) and π_2 satisfies (ii) then the ε-contaminated prior (see Berger 1985 Section 4.7) $\varepsilon\pi_1 + (1-\varepsilon)\pi_2$ is arbitrarily close as a distribution to π_2 for small ε, but it always satisfies (i).

References

[1] Benjamini, Y. and Yekutieli, D. (2005). "False Discovery Rate Controlling Confidence Intervals for Selected Parameters" (with discussion), *Journal of the American Statistics Association*, 100, 71-93.

[2] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer Series in Statistics.

[3] Brown, L. D., Cai, T., and DasGupta, A. (2001). "Interval Estimation for a Binomial Proportion" (with discussion), *Statistical Science*, 16, 101-133.

[4] Cleary, R. and Casella, G. (1995). "An Application of Gibbs Sampling to Estimation in Meta-Analysis: Accounting for Publication Bias," *Journal of Educational and Behavioral Statistics*, 22, 141-154

[5] Clopper, C. J., and Pearson, E. S. (1934). "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404-413.

[6] Dawid, A. P. (1994). "Selection Paradoxes of Bayesian Inference," in *Multivariate Analysis and its Applications* (Vol. 24), eds. T. W. Anderson, K. A.-T. A. Fang and I. Olkin, Philadelphia, PA: IMS.

[7] Ferguson, T. S. (1967) *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press.

[8] Iyengar, S., and Greenhouse, J. B. (1988). "Selection Bias and the File Drawer Problem," *Statistical Science*, 3, 109-117.

[9] Lehmann, E. L. (1991). *Testing Statistical Hypotheses*, Second Edition, Wadsworth & Brooks/Cole.

[10] Lindley, D. V., and Phillips, L. D. (1976). "Inference for a Bernoulli Process (a Bayesian View)," *The American Statistician*, 30, 112-119.

[11] Mandel, M., and Rinott, Y. (2007). "On Statistical Inference Under Selection Bias," Discussion Paper #473, Center for The Study of Rationality, The Hebrew University of Jerusalem.

[12] Psaty, B. M., and Kronmal, R. A. (2008). "Reporting Mortality Findings in Trials of Rofecoxib for Alzheimer Disease or Cognitive Impairment: A Case Study Based on Documents From Rofecoxib Litigation," *Journal of the American Medical Association*, 299, 1813-1817.

[13] Rosenthal, R. (1979). "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86, 638-641.

[14] Rothstein, H. R. Sutton, A. J. and Borenstein, M. (Editors) (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, Wiley.

[15] Senn, S. (2008). "A Note Concerning a Selection"Paradox" of Dawid's," *The American Statistician*, 62, 206-210.