האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

ON STATISTICAL INFERENCE UNDER SELECTION BIAS

by

MICHA MANDEL and YOSEF RINOTT

Discussion Paper # 473 December 2007

מרכז לחקר הרציונליות
CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

On statistical inference under selection bias

Micha Mandel [a,*] , Yosef Rinott [b]

[a] *Department of Statistics, The Hebrew University of Jerusalem, Israel*

[b] *Department of Statistics, and The Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel*

15 December 2007

Abstract

This note revisits the problem of selection bias, using a simple binomial example. It focuses on selection that is introduced by observing the data and making decisions prior to formal statistical analysis. Decision rules and interpretation of confidence measure and results must then be taken relative to the point of view of the decision maker, i.e., before selection or after it. Such a distinction is important since inference can be considerably altered when the decision maker's point of view changes. This note demonstrates the issue, using both the frequentist and the Bayesian paradigms.

Key words:
Confidence interval, Credible set, Binomial model, Decision theory, Likelihood principle

1 Introduction

1.1 Background

Decision theory provides precise and well-defined criteria that quantify our confidence in data-based decisions such as the level of significance, p-value, confidence interval level, and Bayesian confidence interval (credible interval). Such criteria are predicated on a variety of assumptions and models. An important but problematic condition which is often made and rarely satisfied is

* Corresponding author.
Email addresses: msmic@huji.ac.il (Micha Mandel), rinott@huji.ac.il (Yosef Rinott).

that the data are not subject to selection or screening prior to their analysis, and that the hypotheses, models, and statistical methods are determined in advance, that is, before observing the data.

The problem has become more severe in recent years, where sometimes research starts with *data mining* and data selection, and is followed by formulation of hypotheses, and selection of statistical models and relevant tools for testing and estimation: an order of operations that violates the basic tenets of decision theory. An exceptional area that acknowledges selection bias to some degree is *meta-analysis*, where individual studies are combined to estimate an overall effect. Much of the effort in meta-analysis research is devoted to understanding and testing for the existence of selection bias due to use of published data only, and to making the necessary adjustments in inference. See, for example, Duval and Tweedie (2000) and Berger (1985, Problem 4 of Chapter 4).

In specific well-structured cases, decision-theoretic criteria can be evaluated in spite of violations of this type. For example, Olshen (1973) considers application of Scheffé simultaneous confidence intervals computed only on data that first lead to rejection of the null hypothesis of equality of the parameters in question to zero. He shows how to calculate the correct frequentist region in this two-stage problem. The reply in Scheffé (1977) to this procedure is enlightening, and can be summarized by the well-known frequentist mantra: one must decide on the criterion before analyzing the data. For other examples of confidence rules that are constructed conditionally on a result of a statistical test see Siegmund (1978), Meeks and D'Agostino (1983), Finner (1994) and Chiou and Han (1999).

Bayesian advocates may feel that these problems are unique to the frequentist school, since the Bayesian philosophy follows the Likelihood Principle, which enables post-data inference. This, however, is only partially true. It would be true if all decisions including the choice of the model, the prior, and the loss function were made in advance.

The purpose of this paper is to shed some light on the difficulties of trying to take data selection into account and the dangers of not doing so, by means of a very simple binomial example. In particular, we make the point that in certain situations of the kind described above, frequentist confidence levels may have to be taken as relative to the point of view of the user. Such measures may vary between the scientist performing the study, the statistician advising him, and the reader and user of the research results. We also characterize situations in which selection bias can be ignored by the Bayesian statistician and models under which the bias must be taken into account.

To crystalize the issues, we consider the following scenario, which is a prototype of the situations described above. A scientist comes to a statistician after conducting a binomial experiment, asking to construct a confidence interval for the probability of success p. The scientist says that she would not have come had she observed two or more successes, but came because fewer successes had been observed and she was not sure how to construct the confidence interval. Furthermore, she says that all scientists act in this way, that is, go to a statistical consultant only when they observe fewer than two successes; when observing two or more successes they construct the confidence interval by themselves. No further information regarding other scientists or experiments is provided.

In the sequel, we will distinguish between two possible population models:
Model I - there is an infinite number of independent binomial experiments $\mathrm{Bin}(n, p)$ with the **same probability of success** p.
Model II - there is an infinite number of independent binomial experiments $\mathrm{Bin}(n, p_i)$ with **different probabilities of success** p_i.
Under both models, the statistician observes an experiment from the subset of experiments that result in one or zero successes, not knowing the number of unreported experiments that might have been performed.

As an example, p_i could be the probability of an adverse reaction to a new drug at a given dose d_i given to n subjects in each of a number of experiments. Only "successful" experiments that end with at most one case of adverse reaction are brought to the statistician for analysis. *Model I* holds when all doses and hence all p_i's are equal; otherwise, we have *Model II*.

A formal description of the problem is as follows. Let $P = (p_1, p_2, \ldots)$ be a sequence of probabilities $0 < p_i < 1$, and let $X_i \sim \mathrm{Bin}(n, p_i),\ i = 1, 2, \ldots$, be independent binomial experiments, where under *Model I* $p_i \equiv p$. The random variable observed is X_T, where $T = \min\{i : X_i \leq 1\}$, and T is unknown to the statistician. We assume that P is such that $P_P(T < \infty) = 1$. The aim is to construct a $1 - \alpha$ level confidence interval $CI(X_T)$ for p_T. We remark that under *Model I*, the sequence P can be identified with the real number p and $p_T \equiv p$. However, under *Model II*, the parameter p_T is random.

2 A Frequentist Perspective

2.1 The Statistician's Viewpoint

For the frequentist statistician, the parameter P (a sequence of probabilities) is fixed but unknown, and the data is a realization of X_T obtained by the process and stopping rule defined above. He aims at constructing a confidence interval for p_T that satisfies

$$P_P(CI(X_T) \ni p_T) \geq 1 - \alpha \qquad \text{for all possible } P. \tag{1}$$

Thus, the statistician's criterion depends only on the experiments he observes, and the intervals he constructs are supposed to cover p_T for $1-\alpha$ of the times on average, no matter what the parameter is.

Under *Model I*, $p_i \equiv p$ and the confidence interval is based on

$$P_p(X_T = x) = P_p(X = x | X \leq 1) := P(X^* = x|p), \tag{2}$$

where $X^*|p$ has the distribution of $[X|X \in \{0,1\}]$ and $X \sim Bin(n,p)$.

Under *Model II*, p_T is a random coordinate of the parameter P and we have

$$P_P(CI(X_T) \ni p_T) = \sum_t P_P(CI(X_t) \ni p_t \,|\, T = t) P_P(T = t). \tag{3}$$

Now for $x = 0, 1$ we have

$$\begin{aligned} P_P(X_t = x | T = t) &= P_P(X_t = x | X_t \leq 1, X_1 > 1, ..., X_{t-1} > 1) \\ &= P_{p_t}(X_t = x | X_t \leq 1) := P(X^* = x|p_t). \end{aligned} \tag{4}$$

Therefore, if $CI(X_t)$ is constructed as a $1-\alpha$ confidence interval for p_t based on $X^*|p_t$, then (4) implies $P_P(CI(X_t) \ni p_t \,|\, T = t) \geq 1 - \alpha$, and (1) follows from (3).

Thus, for the frequentist statistician the distinction between *Model I* and *Model II* is unimportant. He provides a confidence interval for p_T corresponding to "successful" experiments, based on the variable X^* having a Bernoulli distribution with probability

$$P_p(X^* = 1) = P_p(X = 1 | X \in \{0,1\}) = \frac{np}{1 + (n-1)p} \equiv p^*. \tag{5}$$

For $\alpha < 1/2$, a $1-\alpha$ level confidence interval for p^*, that is, a confidence interval $CI(X^*)$ satisfying $P_p(CI(X^*) \ni p^*) \geq 1-\alpha$ for **all** possible values of p, is

$$CI_{p^*}(X^*) = \begin{cases} (0, 1-\alpha) & X^* = 0 \\ (\alpha, 1) & X^* = 1 \ . \end{cases} \tag{6}$$

For $X^* = 0$, $CI_{p^*}(X^*)$ is equivalent to the interval $0 < np/(1+(n-1)p) < 1-\alpha$, and for $X^* = 1$, the interval $CI_{p^*}(X^*)$ becomes $\alpha < np/(1+(n-1)p) < 1$, which after rearrangement as intervals for p reduce to

$$CI^*(0) = \Big(0, \frac{1-\alpha}{\alpha n + 1 - \alpha}\Big), \qquad CI^*(1) = \Big(\frac{\alpha}{(1-\alpha)n + \alpha}, 1\Big). \tag{7}$$

The interval $CI^*(0)$ equals $(0, 1-\alpha)$ for $n = 1$ and decreases with n as expected. However, $CI^*(1)$ includes large values of p and becomes wider as n increases, eventually approaching $(0,1)$. For example, for $\alpha = 0.05$ and $n = 20$, $CI^*(0) = (0, 0.4872)$ and $CI^*(1) = (0.0026, 1)$.

The result may surprise a scientist who expects that one success out of n indicates a small p. However, taking the selection into account, the result is not surprising: for $p > 0$ 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 p.

2.2 The Scientist's Viewpoint

For a scientist who conducts experiments many times until one or no successes are obtained, the confidence interval (7) is correct. It would clearly be wrong and dishonest to ignore the selection issue, and compute a confidence interval based on $X \sim Bin(n,p)$. Under *Model I* ($p_i \equiv p$), the best honest option for the scientist is of course to combine all the data he has on p, while taking into account any stopping rule.

However, let us look again at the problem from the viewpoint of a scientist who according to *Model II* conducts binomial experiments with different p_i's. Such a scientist does not use the interpretation (1), but instead would like a procedure that satisfies for all i

$$P_P(CI(X_i) \ni p_i) \geq 1-\alpha \qquad \text{for all } P \in (0,1)^\infty. \tag{8}$$

In terms similar to Berger (1985), she looks for a procedure that satisfies (with probability one) $\liminf_{N\to\infty} N^{-1} \sum_{i=1}^{N} I\{CI(X_i) \ni p_i\} \geq 1-\alpha$ for all $P \in (0,1)^\infty$, where $I\{E\}$ is the indicator function of the event E. She therefore

should compute $1-\alpha$ level confidence intervals based on $X \sim Bin(n,p)$ in each experiment, including those ending with 0 or 1 successes, so that on average $(1-\alpha)100\%$ of her intervals will contain the true p_i. For $X = 0$ or 1, these intervals concentrate on small values of p and are considerably different from the statistician's intervals (7). For example, for $\alpha = 0.05$ and $n = 20$, the intervals $CI_C(0) = (0, 0.1391)$, and $CI_C(1) = (0.0013, 0.2487)$ based on the exact method of Clopper and Pearson (1934), are considerably different from the intervals presented at the end of Section 2.1.

Thus, the scientist knows that in the long run she will miss the parameter in $\alpha 100\%$ of her intervals, but she is neutral as to who constructed those "bad" intervals, be it the statistician who advised her for the cases of $X = 0$ and 1 or herself.

2.3 The Statistician's Dilemma

Consider *Model II* and a scientist who comes to the statistician for advice only when $X = 0$ or 1, and otherwise computes the confidence intervals by herself. The statistician now faces a dilemma. If he wants to serve the client well, he should base his inference on X so that the scientist's interpretation of the result (8) will be correct. However, his inference (1), will be very poor if p is not small. For example, let $\alpha = 0.05$, $n = 20$, and $p_i > 0.25$ for all i. Then the statistician using the interval $CI_C(X)$ based on X described above will be wrong 100% of the time he is consulted (see calculations of CI_C for this case in Section 2.2). He should instead base his inference on X^* if he wants to guarantee intervals that cover the true parameter for an average of at least $(1-\alpha)100\%$ of the time he is consulted. The scientist's using intervals based on X rather than X^* will of course also be wrong if $X = 0$ or $X = 1$ and $p_i > 0.25$, but the proportion of these cases, $P_P(X_i \leq 1)$, is relatively small, less than 2.5% (when $p_i > 0.25$ for all i), and hence her total coverage probability is correct. The scientist and the statisticians have different goals, and conflicting interests: the scientists wants to ensure correct inference in $(1-\alpha)100\%$ of all experiments, and the statistician cares only about experiments in which he is consulted.

This conflict cannot be resolved in the frequentist framework. The statistician must decide whether to use the interpretation (8) that involves parameters he has not been asked about, or to use the interpretation (1) that may be less interesting for the scientist and provides counterintuitive intervals. The statistician sees data after selection, and using X^* he computes confidence intervals for parameters selected by use of the data. The fact that his resulting confidence intervals based on X^* are so different from those of the scientist who computes them using X, shows how critical selection can be. The Bayesian

statistician, as shown next, has far less conflict with the scientist.

3 A Bayesian Perspective

3.1 Model I

For the Bayesian there is a prior distribution π for the proportions p in the population. Under *Model I*, there is exactly one draw from the prior distribution π and all the binomial experiments are conditional on that draw. If all experiments are ignored until 0 or 1 is obtained, then the data is X^* whose conditional distribution $X^*|p$ is given in (5). The confidence (or credible) interval for p is an interval I whose posterior probability is $1-\alpha$, i.e., $P(p \in I|X^* = x) = \int_I d\pi(p|X^* = x) = 1-\alpha$.

Proposition 3.1 *Let $X|p \sim \mathrm{Bin}(n,p)$ and $p \sim \pi$, any prior, and for $x = 0, 1$ let $\pi_n(\cdot|x)$ be the posterior distribution of $p|X^* = x$. Then $\{\pi_n(\cdot|x)\}$ is a stochastically decreasing sequence of distributions for both $x = 0$ and $x = 1$.*

Also, for any n, $\pi_n(\cdot|x)$ is stochastically larger than π for $x = 1$ and smaller for $x = 0$.

Proposition 3.2 $\lim_{n\to\infty} \pi_n(\cdot|1) = \pi(\cdot)$.

Proposition 3.1 says that as n increases, one predicts smaller values of p when observing either one or zero successes. Proposition 3.2 now implies that for one success, the posterior approaches the prior from above. Naturally, after one observes $X^* = 1$ the prior is updated "upwards" and vice versa for $X^* = 0$.

To prove the propositions we need the following known lemma whose proof is given for the sake of completeness.

Lemma 3.1 *Consider a random variable $R \sim F$, and let $w(\cdot) \geq 0$ be a decreasing function such that $E_F w(R) = \int_{-\infty}^{\infty} w(r)dF(r) < \infty$ and $E_F w(R) > 0$. Let $F^w(t) = \int_{-\infty}^{t} w(r)dF(r)/E_F w(R)$. Then $F^w \leq_{\mathrm{st}} F$.*

Proof of Lemma. For t such that $w(t) \geq E_F w(R)$ we have

$$\int_{-\infty}^{t} \left[\frac{w(r)}{E_F w(R)} - 1\right] dF(r) \geq 0$$

so that $F^w(t) \geq F(t)$. For t such that $w(t) \leq E_F w(R)$ we have

$$\int_{t}^{\infty} \left[\frac{w(r)}{E_F w(R)} - 1\right] dF(r) \leq 0$$

so again $F^w(t) \geq F(t)$. $\square$

Proof of Proposition 3.1. Using Bayes formula,

$$d\pi_n(p|x) = d\pi(p)g(p,n,x)/E_\pi g(p,n,x) \text{ for } x = 0,1, \tag{9}$$

where $g(p,n,x) := P(X^* = x|p)/P(X^* = x) = \frac{npx+(1-p)(1-x)}{1+(n-1)p}$ (the expectation exists for all n since $|g(p,n,x)|$ is bounded by 1). It is easy to check that $d\pi_{n+1}(p|x) \propto w_n(p)d\pi_n(p|x)$, where $w_n(p) = 1 - 1/(\frac{1}{p} + n)$. Stochastic monotonicity follows by noticing that $w_n(p)$ is decreasing in p and applying Lemma 3.1 with R being the random variable p, and $F = \pi$.

For the second part of the proposition note that $g(p,n,x)$ is increasing in p for $x = 1$, and decreasing for $x = 0$ and any n. Using Lemma 3.1 and (9) we conclude that for any n the posterior $\pi_n(\cdot|x)$ is stochastically larger than π for $x = 1$ and smaller for $x = 0$. $\square$

Proof of Proposition 3.2. This follows from $g(p,n,1) \to 1$ when $n \to \infty$. $\square$

Let a be the left limit of the support of π. One might expect that $\lim_{n\to\infty} \pi_n(\cdot|0) = \delta_a$, the measure concentrated at a. However, since both $g(p,n,0) \to 0$ and $E_\pi g(p,n,0) \to 0$, this is not true in general. In fact, if π is discrete with atoms $0 < a_1 < a_2 < \ldots$ and respective masses $\pi_1, \pi_2, \ldots$, then

$$\frac{d\pi_n(a_i|0)}{d\pi_n(a_j|0)} = \frac{(1-a_i)/[1+(n-1)a_i]}{(1-a_j)/[1+(n-1)a_j]}\frac{\pi_i}{\pi_j} \xrightarrow[n\to\infty]{} \frac{1-a_i}{a_i}\pi_i \Big/ \frac{1-a_j}{a_j}\pi_j,$$

showing that the limit distribution is a weighted version of π with weight $(1-a_i)/a_i$ at a_i. This result can be generalized as follows:

Proposition 3.3 *If $E_\pi p^{-1} = \infty$, then $\lim_{n\to\infty} \pi_n(\cdot|0) = \delta_0$; otherwise $\lim_{n\to\infty} \pi_n((-\infty,t]\,|\,0) \propto \int_0^t \frac{1-v}{v} d\pi(v)$.*

Proof of Proposition 3.3. For $t > 0$

$$1 - \pi_n(t|0) = \int_t^1 \frac{\frac{1-p}{1+(n-1)p}d\pi(p)}{\int_0^1 \frac{1-v}{1+(n-1)v}d\pi(v)} = \int_t^1 \frac{\frac{1-p}{1/(n-1)+p}d\pi(p)}{\int_0^1 \frac{1-v}{1/(n-1)+v}d\pi(v)} \to \int_t^1 \frac{\frac{1-p}{p}d\pi(p)}{E_\pi(p^{-1}-1)}.$$

$\square$

For $\pi = \mathrm{U}(0,1)$, $\alpha = 0.05$, and $n = 20$, the equal tail credible intervals are $(0.0029, 0.6982)$ for $X^* = 0$ and $(0.0620, 0.9778)$ for $X^* = 1$. These should be compared to the prior credible interval (0.025,0.975).

3.2 Model II

Bayesian inference under *Model II* is totally different. While under *Model I* all experiments use the same p, under *Model II* the probabilities of success differ for different experiments. In principle, the prior distribution should be specified for the parameter vector P. However, since the index T is unobserved, inference is problematic unless further assumptions are made. We consider the case where $p_1, p_2, \ldots$ are independent identically distributed with prior law π.

Under the above model, the distribution of values of p in experiments that end with exactly x successes, is just the posterior of p given $X = x$. This rather obvious observation is shown as follows: for $x = 0, 1$ we have

$$\begin{aligned}
&P(p_T \in B \,|\, X_T = x) \\
&= \sum_t P(p_t \in B \,|\, X_t = x, X_1 > 1, \ldots, X_{t-1} > 1) P(T = t | X_T = x) \\
&= \sum_t P(p_t \in B \,|\, X_t = x) P(T = t | X_T = x) = P(p_1 \in B \,|\, X_1 = x),
\end{aligned}$$

where the equalities hold because $(p_1, X_1), (p_2, X_2), \ldots$ are independent and identically distributed. Thus, under *Model II*, the Bayesian statistician is unaffected by the selection and has no conflict with the scientist; i.e., he should consider the data as coming from the law of $X|p$ and not from $X^*|p$. For $\pi = \mathrm{U}(0, 1)$, $\alpha = 0.05$, and $n = 20$, for example, the credible intervals are (0.0012,0.1611) for $X = 0$ and (0.0117,0.2382) for $X = 1$, considerably different from those calculated under *Model I*.

Bayesian inference is known to be post data, and should not be affected by selection bias. However, it has just been shown, that this claim is true under *Model II*, but not under *Model I*. The crucial difference between the models is in the space the selection acts on. Under *Model I*, the probability of selection is $P(X \leq 1|p)$, whereas under *Model II* the probability of selection is $P(X \leq 1) = E_\pi P(X \leq 1|p)$, which is independent of p. The next section discusses this point in a more formal and general setting.

3.3 Selection Bias and Bayesian Analysis

Let Θ be the (marginal) sample space of θ and let $\mathcal{X}$ be the (marginal) sample space of X. Suppose that on a probability space $(\Theta \times \mathcal{X}, \mathcal{A}, \mathcal{P})$, the marginal density of θ is π and the conditional density of X given θ is f. Consider a sample of size k. Then, the sample space of (a generalized) *Model I* is $\Theta \times \mathcal{X}^k$ and that of (a generalized) *Model II* is $(\Theta \times \mathcal{X})^k$. Let $w : \mathcal{X} \mapsto \mathbb{R}^+$ be a

non-negative function; then the samples under *Models I* and *II* have densities

$$g_1(\theta, x_1, \ldots, x_k) = \pi(\theta) \prod_{i=1}^{k} \frac{w(x_i) f(x_i|\theta)}{\int w(u) f(u|\theta) du}, \tag{10}$$

and

$$g_2(\theta_1, \ldots, \theta_k, x_1, \ldots, x_k) = \prod_{i=1}^{k} \frac{w(x_i) f(x_i|\theta_i) \pi(\theta_i)}{\int\int w(u) f(u|\nu) du \, d\pi(\nu)}, \tag{11}$$

respectively, where in the binomial example $w(x) = 1$ for $x = 0, 1$, and 0 otherwise. Under *Model I*, the posterior is proportional to

$$\pi(\theta) \prod_{i=1}^{k} \frac{f(x_i|\theta)}{\int w(u) f(u|\theta) du},$$

which in general differs from

$$\frac{\pi(\theta) \prod_{i=1}^{k} f(x_i|\theta)}{\int \prod_{i=1}^{k} f(x_i|\nu) d\pi(\nu)},$$

the posterior with $w \equiv 1$, i.e., without selection bias.

Under *Model II*, the posterior is proportional to $\prod_{i=1}^{k} f(x_i|\theta_i)\pi(\theta_i)$ and it is equivalent to the posterior with $w \equiv 1$. Thus, for the Bayesian, selection bias under *Model II* makes no difference, but under *Model I* it cannot be ignored. Note the distinction between the two models where the posterior under the first is one dimensional while that of the second has k dimensions.

Models I and *II* look very similar in the one-dimensional case $k = 1$, where the sample spaces are equal. However, they differ because the selection is done on different spaces. Specifically,

$$g_1(\theta, x) = \pi(\theta) \frac{w(x) f(x|\theta)}{E\{w(X)|\theta\}} \quad , \quad g_2(\theta, x) = \pi(\theta) \frac{w(x) f(x|\theta)}{E\{w(X)\}}.$$

It follows that $\pi_1(\theta) = \pi(\theta)$, and $f_1(x|\theta) = w(x)f(x|\theta)/E\{w(X)|\theta\}$ are the marginal density of θ and the conditional density of $X|\theta$ under *Model I*, and $\pi_2(\theta) = E\{w(X)|\theta\}\pi(\theta)/E\{w(X)\}$, $f_2(x|\theta) = w(x)f(x|\theta)/E\{w(X)|\theta\}$ are the corresponding densities under *Model II*. Note that in both models the distribution of $X|\theta$ is the same, but in *Model II* the prior is biased. Roughly speaking, *Model I* generates θ from π and then generates X conditionally on the value of θ but with bias, while *Model II* generate pairs (θ, X) with bias according to X.

The different treatments of the two models stem from a very basic principle of the Bayesian paradigm, the Likelihood Principle. This principle states that the conclusions drawn from two random systems that produce proportional likelihoods must be the same (Cox and Hinkley 1974. pp. 39). Because the

denominator in (11) does not involve any unknown parameter, the likelihood of the generalized *Model II* is proportional to the likelihood of unbiased data, hence should be treated the same way. The likelihood (10) is not proportional to that of unbiased data, hence it leads to different conclusions.

Thus, unlike the frequentist who always treats the scientist as coming from *Model I*, and hence has a conflict with the scientist of *Model II*, the Bayesian has the ability to adjust his analysis to the specific population model to which the scientist belongs, (*I* or *II*), and to provide proper inference for each. Of course, he must know the scientist's model and whether selection tilted the distribution of (X, θ) or that of $X|\theta$, while the frequentist's inference is independent of that knowledge.

4 Concluding Remarks

In our binomial model, the frequentist scientist does not determine the interval before observing the data. Specifying the confidence region after observing the data seems illegitimate because in calculating the level of a procedure, a frequentist must take into account all possible outcomes; see (8). However, noticing that for a parameter θ and a random set $R(X)$,

$$P_\theta(\theta \in R(X)) = P_\theta(\theta \in R(X)|X \in A)P_\theta(X \in A)+P_\theta(\theta \in R(X)|X \in \bar{A})P_\theta(X \in \bar{A}),$$

where P_θ is the probability under θ, we see that a confidence region R that satisfies both $\inf_\theta P_\theta(\theta \in R(X)|X \in A) = 1-\alpha$ and $\inf_\theta P_\theta(\theta \in R(X)|X \in \bar{A}) = 1-\alpha$ is valid. One can calculate $R(X)$ over A only when the event $X \in A$ occurs. This is exactly the statistician's approach, where $A = \{0, 1\}$.

Another justification for constructing the confidence intervals after observing the data is due to Brown, Cai, and Dasgupta (2001), who advocate modifications of simple intervals around the boundaries (i.e., for values near 0 and n). A reasonable approach is to go to the statistician only if such modifications are needed. In this case, the statistician serves as an instrument (calculator) to calculate the needed intervals, and his own interpretation should be ignored.

Nowadays, most scientists have access to user-friendly and automated statistical softwares; hence most statistical analyses are conducted by non-statisticians. Commonly, statisticians are consulted when the data do not clearly prove anything, and when a scientist does not succeed in proving her point. In this sense, the life of a statistician is "tragic", since he often gets to see the data only when the results are inconclusive, and therefore his percentage of correct decisions falls drastically, and he must take drastic measures to prevent this from happening. The binomial case reveals how large the problem can be and how

important it is to define clearly the criterion behind the decision rule. We do not offer a solution to the statistician's "tragedy", but hope that the current example helps researchers appreciate the possible magnitude of the problem and encourage them to define clearly their decision criteria before interpreting their results.

Acknowledgments

The work of Yosef Rinott was partially supported by Israel Science Foundation grant 473/04. We thank Ester Samuel-Cahn for helpful comments on an early draft of this work.

References

[1] Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics.

[2] Brown, L.D., Cai, T., DasGupta, A., 2001. Interval estimation for a binomial proportion (with discussion). Statistical Science. 16, 101-133.

[3] Chiou P., Han C.P., 1999. Conditional interval estimation of the ratio of variance components following rejection of a pre-test. J. Statist. Comp. Simul. 63, 105-119.

[4] Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limit-sillustrated in the case of the binomial. Biometrika. 26, 404-413.

[5] Cox, D.R., Hinkley, D.v., 1974. Theoretical statistics. Chapman and Hall, London, UK.

[6] Duval, S., Tweedie, R., 2000. A non-parametric "Trim and Fill" method of accounting for publication bias in meta-analysis. J. Amer. Statist. Assoc. 95, 89-98.

[7] Finner, H., 1994. Two-sided tests and one-sided confidence-bounds. Ann. Stat. 22, 1502-1516.

[8] Meeks, S.L., D'Agostino, R.B., 1983. A Note on the use of confidence limits following rejection of a null hypothesis. Amer. Statist. 37, 134-136.

[9] Olshen, R.A., 1973. The conditional level of the F-Test. J. Amer. Statist. Assoc. 68, 692-698.

[10] Scheffé, H., 1977. A note on a reformulation of the S-Method of multiple comparison. J. Amer. Statist. Assoc. 72, 143-144.

[11] Siegmund, D., 1978. Estimation following sequential tests. Biometrika. 65, 341-349.