Statistical Hypothesis Testing Formulations for U.S. Environmental Regulatory Standards for Ozone

Mary Lou Thompson Lawrence H. Cox Paul D. Sampson

David C. Caccia

NRCSE

Technical Report Series

NRCSE-TRS No. 059

November 13, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

STATISTICAL HYPOTHESIS TESTING FORMULATIONS FOR U.S. ENVIRONMENTAL REGULATORY STANDARDS FOR OZONE

Mary Lou Thompson[1,2], Lawrence H. Cox[3], Paul D. Sampson[1], and David C. Caccia[1]

[1] National Research Center for Statistics and the Environment

Box 354323

University of Washington

Seattle, WA 98195

USA

mlt@biostat.washington.edu

[2] Department of Biostatistics

Box 357232

University of Washington

[3] US Environmental Protection Agency

National Exposure Research Laboratory (MD-75)

Research Triangle Park, NC 27711

Abstract

Environmental regulatory standards are intended to protect human health and environmental welfare. Current standards are based on scientific and policy considerations but appear to lack rigorous statistical foundations. We examine current and proposed U.S. environmental regulatory standards for ozone from the standpoint of their formulation and performance within a statistical hypothesis testing framework. We illustrate that the standards can be regarded as representing constraints on a percentile of the ozone distribution, where the percentile involved depends on the defined length of ozone season and the constraint is stricter in regions with greater variability. A hypothesis testing framework allows consideration of error rates (probability of false declaration of violation and compliance) and we show that the existing statistics on which the standards are based can be improved upon in terms of bias and variance. Our analyses also raise issues relating to network design and the possibilities of defining a regionally based standard that acknowledges and accounts for spatial and temporal variability in the ozone distribution.

1 Introduction

The U.S. Clean Air Act (CAA) and Amendments (42 USC Sec. 7401 et seq.) direct the Administrator of the US Environmental Protection Agency (EPA) to identify pollutants which "may reasonably be anticipated to endanger public health and welfare" and to set air quality criteria for them. CAA Sec. 109 directs the Administrator to propose and promulgate primary (public health) and secondary (public welfare) National Ambient Air Quality Standards (NAAQS) for such pollutants, and to review these standards periodically based on available scientific evidence and revise them if necessary. Accordingly, in 1970 EPA established NAAQS for six criteria pollutants: carbon monoxide, ozone, particulate matter, nitrous oxides, sulfuric oxides and lead, and has reviewed them periodically since then. A recent NAAQS review resulted in promulgation in 1997 of revisions to the standards for ozone (US Federal Register 1997).

Although there has been considerable focus in the general scientific and policy domains on the definition and implementation of standards, there has been relatively little in the way of statistical contribution to this arena. Carbonez et al. (1999) have considered and evaluated the formulation of the U.S. Drinking Water Regulations in statistical terms. We consider a number of different possible formulations of the US ozone standards and evaluate and compare their characteristics from a statistical perspective based on the framework of Barnett and O'Hagan (1997) in a report to the U.K. Royal Commission on Environmental Pollution. Barnett and O'Hagan distinguish between *realizable* standards where it is possible to determine without uncertainty if the standard is satisfied—as in the case of standards defined explicitly in terms of sample measurements—and *ideal* standards that

are not directly realizable, as is the case for standards defined in terms of distributional parameters. [Note: we would have preferred the term "idealized" as many standards satisfying this characterization are not likely to be judged "ideal" in the common sense of the word.]

A statistical assessment of environmental standards addresses two fundamental issues. Standards defined in terms of distributional parameters acknowledge, at least implicitly, the fundamental fact of *variation* expressed in a probability distribution. In the case of most of the U.S. ideal air quality standards, variation in time is acknowledged, but issues of variation in space are not explicitly addressed. The second fundamental concept is statistical *uncertainty* in estimation of distributional parameters. Barnett and O'Hagan incorporate this concept in their definition of a *statistically verifiable ideal standard* as consisting of an ideal standard together with a "standard for statistical verification of the ideal standard," typically expressed as a level of assurance of compliance with the ideal that must be demonstrated. As we elaborate below, this level of assurance might be defined, for example, in terms of statistical quality criteria such as specified levels for type I and type II errors in a hypothesis testing framework.

Our aim in this paper is to discuss ideal standards that are implicit in, or at least compatible with, current US air quality standards (which are specified as realizable standards) and then to propose and evaluate statistically verifiable approaches for these ideals. We consider an ideal standard as having the following components: an underlying random variable reflecting some measure of the pollutant of interest, one or more parameters of the distribution of this random variable which will be the focus of the standard, and some constraint or threshold that the standard places on the value of the parameter(s).

In Section 2 we describe the current US ozone standards, in Section 3 we present a general hypothesis testing framework for statistically verifiable ideal standards, and in Section 4 we propose ideal standards corresponding to the US realizable ozone standards. Section 5 presents explicit expressions for the parameters on which the standards in Section 4 are based, under the assumption that the corresponding daily ozone maxima are independently and identically distributed. While these assumptions (which apply to days in the ozone season and across sites in a region) are not entirely reasonable in practice, they provide a tractable theoretical framework from which we can make some important general points. We aim to illustrate that by specifying an ideal standard in terms of parameters of a distribution, one is in a position to develop statistically more rigorous and efficient estimates of the underlying parameters than those effectively specified in the implementation of the US ozone standard. A hypothesis testing framework enables evaluation of different realizable standards related to the same ideal. This means, for instance, that one can compare standards on issues such as probabilities of correctly affirming compliance and the probability of falsely declaring compliance.

In Section 6 we present, for each of the ideal standards, the compliance statistics mandated by the US regulations and the corresponding test statistics a statistician would choose. We use data from the South Coast Air District in Southern California (CARB) and Chicago region 67 (AIRS) to suggest reasonable parameter values for a simple model for ozone measurements, and then proceed in Section 7 to compare the proposed statistics under this model. In Section 8 we address the issue of the plausibility of the assumptions and we conclude with further discussion in Section 9.

2 US ozone standards

The current characterization of an air quality standard, as explained in the issue of the Federal Register announcing the proposed revisions to the standards, involves specification of an averaging time (e.g. one hour or eight hours), a concentration, and the form of the air quality statistic to be used as a basis for judging compliance (Federal Register 1997). A daily exceedance under the existing ozone standard is defined as a maximum daily 1-hour average ozone concentration in excess of 0.12 parts per million (ppm) ozone at designated monitoring sites—the "1-hour standard." Exceedance at a designated monitoring site under the proposed revised standard occurs when the daily maximum 8-hour average concentration exceeds 0.08 ppm—the "8- hour standard." Revision of the form of the chosen air quality statistic led to changing the criterion for violating the standard from "expected annual number of exceedances of the standard at any designated monitoring site over a consecutive 3-year period greater than one" for the 1-hour standard to "average at any designated monitoring site over any consecutive 3-year period of the fourth highest annual daily maximum 8-hour concentrations greater than the daily exceedance value (0.08ppm)" for the 8-hour standard (US Federal Register 1997).

The federal court recently ruled that the "construction of the Clean Air Act on which EPA relied in promulgating the [8-hour ozone standard] effects an unconstitutional delegation of legislative power" and instructed EPA to develop a construction of the Act that satisfies the constitutional requirement of the Delegation of Powers Act (U.S. Court of Appeals for the District of Columbia, 97-1440 and 1441, American Trucking Associations, Inc. et al. vs. U.S. Environmental Protection Agency). A concurrent revision of the particulate

matter NAAQS was interpreted similarly. In essence, the court found that EPA failed to articulate an "intelligible principle," based on available scientific evidence, on which to base its selection of the 0.08 ppm 8-hour ozone standard as the National Ambient Air Quality Standard for the protection of public health. Borrowing from this language, we are concerned here with intelligible principles based on statistical science for verifying that a standard has or has not been met. We do not address scientific bases for the choice of an 8-hour standard, or any other criterion, over a 1-hour standard.

The "form of the air quality statistic" for the U.S. 1-hour ozone standard, as discussed in the Federal Register (1997), corresponds to an ideal standard expressed in terms of a parameter of the probability distribution of daily maximum of hourly ozone measurements (at designated monitoring sites): that the *expected number* of annual exceedances of the 0.12 ppm threshold shall not exceed one. However, its implementation is formulated as a realizable standard: that the *average number* of exceedances over a three year period must not exceed one *at all designated sites* (40 CFR Part 50, Appendix H). In effect this form of the standard applies the law of large numbers to the case $n = 3$ stating that, since expected values are close to averages, the standard is violated if there are at least four violations in three consecutive years at any site. The 8-hour standard is also a realizable standard, defined in terms of the value of the average of certain order statistics for measured concentrations at a prescribed set of monitoring locations.

3 A hypothesis testing framework

Hypothesis testing provides one possible approach to developing a statistically verifiable ideal standard. Consider an ideal standard requiring that the value of some parameter θ not exceed a threshold, c_U, say. This threshold may be regarded as representing the level that the standard was designed to protect against. In developing an associated statistically verifiable ideal standard in a hypothesis testing framework one might specify in addition a lower threshold, c_L, say, below which it is believed that there is no adverse effect. The associated statistically verifiable ideal standard could be constituted by specifying the null hypothesis $H_0 : \theta > c_U$ and requiring the implementation of the standard to achieve prescribed type I and type II error rates, where the type II error rate is with regard to misclassification at the lower threshold c_L. In specifying the null hypothesis as representing violation of the standard we follow the proposal by Guttorp (2000) rather than the formulation suggested by Barnett & O'Hagan (1997). In the context of US air pollution regulations, this appears to us to be the appropriate orientation of the null hypothesis as the Clean Air Act (CAA Section 109 (b) (1)), states that the more serious error is to declare *false compliance*, i.e., to subject citizens to exceedances when in fact a region has been declared in compliance. Thus it seems appropriate to set a limit on the probability of false declaration of compliance as a type I error. The specification of a parameter θ (e.g., a percentile of a distribution) and thresholds c_L and c_U are very difficult problems involving consideration of scientific evidence and public policy issues. These are beyond the scope of the analysis here.

Note that the approach of placing demands on the assurance of compliance/violation

as expressed in terms of type I and II hypothesis testing errors requires both statistical modeling and analysis that will permit realistic assessment of these errors and also an environmental sampling design that will make it possible to achieve these error goals. Arbitrarily low type I and type II error rates may not be achievable by any sampling design, and this must be considered in the evaluation of any proposed environmental standard protocols.

While US air quality standards acknowledge variation in the distribution of ozone measurements, they do not consider statistical uncertainty and any quantification of assurance of compliance (with the ideal standard). From a statistical hypothesis testing point of view, both the 1-hour and 8-hour realizable standards specify the critical value for the test statistic (e.g. average fourth highest annual order statistic over three years for the 8-hour standard) as the border between the null and alternative hypotheses (0.08 ppm ozone for the 8-hour standard). The performance properties (types I and II error rates) of the associated statistical tests are not very good (see Section 6). Our main focus in this paper is to assess the characteristics of realizable standards in a hypothesis testing framework with a view towards proposing statistically verifiable ideal standards. In the discussion that follows we assume that the exceedance levels specified by the U.S. EPA, 0.12 and 0.08, represent harmful levels that the EPA wishes to protect against—i.e., they are not set with the aim of providing protection from exceedances at some higher level.

4 Ideal standards for US ozone

To consider the standards in a hypothesis testing framework, we express their formulation in terms of a parameter (vector) θ which summarizes aspects of the ozone behavior over the

region. We assume that ozone observations are available for K days in a year.

4.1 1-hour standard

For the 1-hour ozone standard, the parameter of interest is the expected number of exceedances of a critical level, $c_U = c_1$:

$$\theta_1 = E(\sum_{k=1}^{K} W_k) = E(W_.),$$

where $W_k = 1$ if $X_k > c_1$, $W_k = 0$ otherwise, and where X_k denotes the daily maximum 1-hour average ozone concentration at a given location on day k.

The implementation of the standard can be formulated statistically in terms of the hypotheses $H_0 : \theta_1 > 1$ versus $H_1 : \theta_1 \leq 1$. In the specification of the 1-hour U.S. ozone standard, the regulatory threshold c_1 is set at .12 ppm. The current realizable standard associated with this ideal standard corresponds to a particular choice of test statistic and test critical level. As noted above, utilizing this hypothesis testing formulation as the basis for a statistically verifiable ideal standard requires specification of type I and II error rates for the hypothesis tests and a further threshold (associated with the type II error rate), c_L, say, which is a level below which no hazard is believed to exist.

4.2 8-hour standard

Unlike the 1-hour standard, the 8-hour standard is not explicitly expressed in terms of a parameter of a distribution. A statistical formulation consistent with the proposed implementation of the standard is in terms of the expected value of an order statistic:

$$\theta_2 = E(U),$$

where $U = X^*_{[K-3]}$, i.e. the 4th highest annual order statistic of the set of daily maximum 8-hour average ozone concentrations $X^*_k, k = 1, ...K$ at a given location on day k. Here the hypotheses are $H_0 : \theta_2 > c_2$ versus $H_1 : \theta_2 \leq c_2$. In the specification of the 8-hour standard, the regulatory threshold $c_U = c_2$ is set at .08 ppm.

An alternative formulation in terms of exceedances, and so in the same framework as θ_1, is:

$$\theta_2^{'} = E(\sum_{k=1}^{K} V_k) = E(V_.),$$

where $V_k = 1$ if $X^*_k > c_2$, $V_k = 0$ otherwise. Here the hypotheses of interest are $H_0^{'} : \theta_2^{'} > 3$ versus $H_1^{'} : \theta_2^{'} \leq 3$.

The current realizable 8-hour standard may be viewed as representing a particular test statistic and critical level associated with the hypothesis H_0. Both versions of an ideal standard may be expanded to statistically verifiable ideal standards by consideration of error rates within a hypothesis testing framework as discussed above.

Note that for θ_1 and $\theta_2^{'}$, the threshold is incorporated in the definition of the random variable underlying the standard, whereas in θ_2 the threshold is specified in the null hypothesis. We show in Section 5 that the two formulations θ_2 and $\theta_2^{'}$ imply very similar constraints on the distribution of X^*. The essential difference relates to the choice of parameter on which to base the standard, whether it should be an expected value (as in expected number of exceedances) or the percentile of a distribution, which is arguably the intention in the formulation of the 8-hour standard (Federal Register 1997). We return to this point in Section 5.3.

5 Independently identically distributed case

For mathematical simplicity, we make the assumption that daily maximum 1-hour and 8-hour measurements of ozone concentration are independently and identically distributed over space and time. Although this assumption is unlikely to hold in practice, an examination of the structure of the standards and the characteristics of the associated estimates in this setting allows us to make important general points. The plausibility of the assumptions and the consequences of their violation are examined in Section 8. We are interested here in comparing the implications of the above hypotheses across the different standards and in evaluating the characteristics of the estimates of θ suggested in the current ("realizable") implementation of the standard with other, more statistically rigorous, estimates.

In the following we revise our definitions of X and X^* to assume that they represent monotone transformations, $g()$, of daily maximum 1-hour and 8-hour ozone concentrations for which the sample distributions are approximately Gaussian: $X_k \sim N(\mu, \sigma^2)$ and $X_k^* \sim N(\mu^*, \sigma^{*2})$. Analysis of data from Chicago and Southern California suggest squareroot transformations for both the maximum 1-hour and 8-hour concentrations (see Section 6).

5.1 1-hour standard

In the identically (but not necessarily independently) distributed setting, the expected exceedances parameter θ_1 may be expressed as:

$$\theta_1 = E(W_\cdot) = \sum_{k=1}^{K} P(W_k = 1) = \sum_{k=1}^{K} (1 - P(X_k \leq g(c_1))) = K(1 - \Phi((g(c_1) - \mu)/\sigma).$$

Here $H_0 : \theta_1 > 1$ is equivalent to $H_0 : \mu + \sigma\Phi^{-1}((K-1)/K) > g(c_1)$, where $c_1 = 0.12$ ppm. This is the ideal standard expressed in terms of the parameters of the distribution of X and

implies a constraint on the $(\frac{K-1}{K})$'th percentile of the distribution of X.

5.2 8-hour standard

Using standard results for the approximate moments of order statistics in the i.i.d. setting (Rice 1995), a first order approximation to the expected fourth highest order statistic parameter for the 8-hour standard, θ_2, may be expressed as:

$$\theta_2 \approx g^{-1}(\mu^* + \sigma^* \Phi^{-1}((K-3)/(K+1))).$$

Here $H_0 : \theta_2 > c_2$ is approximately equivalent to $H_0 : \mu^* + \sigma^* \Phi^{-1}((K-3)/(K+1)) > g(c_2)$ where $c_2 = 0.08$ ppm.

Arguments analogous to those for θ_1 above can be used to show that, assuming identical, but not necessarily independent distributions, the related expected exceedance parameter is:

$$\theta_2^{'} = K(1 - \Phi((g(c_2) - \mu^*)/\sigma^*).$$

Here $H_0^{'} : \theta_2^{'} > 3$ is equivalent to $H_0^{'} : \mu^* + \sigma^* \Phi^{-1}((K-3)/K) > g(c_2)$, which for large K will be approximately the same as the H_0 associated with θ_2. Hence, although the 8-hour standard is formulated in terms of order statistics rather than exceedances, an exceedance-based approach corresponds to a very similar formulation of the ideal standard. In particular, for lengths of ozone season (K) corresponding to the current U.S. regulatory framework, the two ideal standards are for practical purposes interchangeable.

5.3 Percentile based standard

In the case of identically distributed ozone measurements, both of the above standards relate to particular percentiles of the ozone distribution. One might hence also consider a more general formulation of standards based on percentiles, e.g.,

$$\theta_3 = F^{-1}(p)$$

where F is the cumulative distribution function of X or X^*. Here one might specify $H_0 : \theta_3 > c_3$ versus $H_1 : \theta_3 \leq c_3$ where

$$\theta_3 = \mu + \sigma\Phi^{-1}(p).$$

and H_0 may hence be expressed as $H_0 : \mu + \sigma\Phi^{-1}(p) > c_3$.

The 1 and 8-hour standards are simply special cases of this general formulation where the percentile depends on the number of monitoring days in a year, K. The use of a standard based on θ_3 could impose the same percentile, regardless of the number of monitoring days. The appeal of a percentile-based standard is limited, however, to settings where the assumption of identical distributions (over time and space) is plausible. In the more general setting with temporally and spatially varying distributional parameters, the notion of a percentile is not well defined and a standard based on an expected value (such as expected number of exceedances) appears preferable.

Note that the expressions for θ_1, θ_2' and θ_3 hold exactly under monotone transformation (with the corresponding transformation of the threshold), but that given above for θ_2 is a first order approximation to the expected value of the 4th order statistic of the untransformed concentrations.

6 Parameter Estimation in the i.i.d case

To assess the implications of hypotheses across the different formulations of the standards, we consider data from the South Coast Air District in Southern California (CARB) and Chicago region 67 (AIRS). We found both the daily maximum 1-hour and 8-hour values to be approximately normally distributed on a square root scale. Table 1 summarizes the 1-hour and 8-hour daily (square root ppm) mean levels and variability for the 10 sites in each region with highest variability for the period 1989-1991. In Southern California, the ozone regulatory season designated by the EPA stretches over the entire year (K=365), whereas in Chicago it is April 1- October 31 (K=214) (http://www.epa.gov/oar/oaqps/greenbk/o3season.html).

Note that on the square root scale, the ozone exceedance or critical levels are $g(c_1) = \sqrt{.12} = .346$ and $g(c_2) = \sqrt{.08} = .283$, respectively, for the 1-hour and 8-hour standards. Clearly the ozone levels in the California region over this period were higher and more variable than those in Chicago. As is to be expected, the 8-hour maximum has a lower mean level, but is only slightly less variable than the 1-hour maximum.

Table 1: Sample means and standard deviations of 1989-91 squareroot ozone pooled over 10 sites in Southern California and Chicago

Region	Standard	μ	σ
Southern California	1-hour	.245	.065
(K=365)	8-hour	.206	.060
Chicago	1-hour	.217	.044
(K=214)	8-hour	.202	.042

We assume that the implementation of the standard in a particular region is based

on daily measurements X_{ijk} (or X^*_{ijk}) at $i = 1, 2, ..., I$ sites; over $j = 1, 2, ..., J$ years; for $k = 1, 2, ..., K$ days within each year. It should be noted that the EPA regulations require evaluation of violation at each designated site in a region and the region is declared to be in violation if *any* site in the region is in violation. This implies that compliance is judged by assessing the *maximum* of site-specific test statistics across sites within a region.

For the purposes of the following calculations X_{ijk} and X^*_{ijk} are assumed to be i.i.d. across sites as well as over time. With spatial dependence, the effective number of sites in a region will clearly be less than the nominal number of sites. We discuss this issue in more detail in Section 8, but to make the comparisons here more plausible we base them on 5 effective sites rather than 10.

6.1 The 1-hour ozone standard

The realizable 1-hour standard in its current formulation specifies the following estimator (and test statistic) for θ_1:

$$\hat{\theta}_1 = \max_i \frac{1}{3} \sum_{j=1}^{3} W_{ij.},$$

where W_{ijk} indicates exceedances as defined in Section 4. We would reject H_0 (region declared in compliance) whenever $\hat{\theta}_1 \leq 1$. The distribution of $W_{i..}$, the number of exceedances over 3 years at site i, for parameter values in the neighborhood of the compliance region, is Binomial$(3K, 1 - \Phi(\frac{g(c_1)-\mu}{\sigma}))$. Thus $\hat{\theta}_1$ involves the *maximum* (over I sites) of independent Binomial random variables.

The maximum likelihood estimator of θ_1 in this setting is:

$$\hat{\theta}^*_1 = K(1 - \Phi(\frac{g(c_1) - \hat{\mu}}{\hat{\sigma}})),$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the usual estimates of the sample mean and standard deviation of the $X_{ijk}, i = 1,2,...I; j = 1,2,3; k = 1,2,...,K$. According to first-order Taylor series expansions, $\hat{\theta}_1^*$ is approximately unbiased with

$$Var(\hat{\theta}_1^*) \approx \frac{K\phi^2(\frac{g(c_1)-\mu}{\sigma})}{3I} + \frac{K^2\phi^2(\frac{g(c_1)-\mu}{\sigma})(\frac{g(c_1)-\mu}{\sigma})^2}{2I(3K-1)}.$$

Note that $\hat{\theta}_1^*$, is based on statistics *averaging* over the sites i, rather than the *maximum* operator in $\hat{\theta}_1$ above.

To compare the characteristics of these two statistics we consider a hypothetical example with a plausible *effective* number of independent sites. Consider, for instance, a region with 5 sites, an ozone season of 365 days where, on the square root scale, $\mu = .16$ (corresponding to a median of .026 ppm) and $\sigma = .065$, i.e. a region that in variability and duration of ozone season corresponds to Southern California. The true value of θ_1 (i.e. the expected number of exceedances of the .12 ppm threshold over the region) in such a region would be .75. So, this is a region that is well in compliance.

It is of interest, then, to explore the impact of using the two test statistics $\hat{\theta}_1$ and $\hat{\theta}_1^*$ to assess the hypothesis that the region is in compliance. In this setting, using the above representation in terms of Binomial random variables, we compute $E(\hat{\theta}_1) = 1.37$ and $Var(\hat{\theta}_1) = .207$. Hence the expected value of the estimate recommended by the current standard is almost double the true value. In contrast, the estimate $\hat{\theta}_1^*$ is approximately unbiased and has $Var(\hat{\theta}_1^*) = .0053$, allowing a far more precise estimate of θ_1. The probability of this region being falsely declared in violation (type II error) is essentially zero using test statistic $\hat{\theta}_1^*$, but is .66 using $\hat{\theta}_1$.

6.2 The 8-hour ozone standard

Here the realizable 8-hour standard in its proposed form specifies the following estimator of θ_2:

$$\hat{\theta}_2 = \max_i \frac{1}{3} \sum_{j=1}^{3} g^{-1}(X^*_{ij[K-3]}),$$

where $X^*_{ij[K-3]}$ is the fourth highest order statistic (on the transformed scale) at site i in year j. This implementation requires H_0 to be rejected whenever $\hat{\theta}_2 \leq .08$.

An alternative estimator of θ_2 based on the approximation in Section 5 and maximum likelihood estimators of μ and σ is given by:

$$\hat{\theta}^*_2 = g^{-1}(\hat{\mu}^* + \hat{\sigma}^* \Phi^{-1}((K-3)/(K+1)))$$

As in the case of the 1-hour standard, $\hat{\theta}_2$ is biased and has greater variance than $\hat{\theta}^*_2$.

There are two points to emphasize in these comparisons. First, the naive statistics implied by the existing realizable standards will be biased because the implementation of the standards forces an assessment of the *maximum* of the site-specific statistics. From a statistical perspective, the intent of the spatial maximum operator in the definitions of $\hat{\theta}_1$ and $\hat{\theta}_2$ is not clear. If the naive statistics $\hat{\theta}_1$ and $\hat{\theta}_2$ were simply based on the average over all sites rather than the maximum, they would be unbiased but would still have larger variance than $\hat{\theta}^*_1$ and $\hat{\theta}^*_2$. So, our second point is that, having specified a parameter and associated hypothesis test on which the ideal standard is to be based, the choice of corresponding test statistic should be based on statistical considerations which will allow optimization of the properties of the statistical test.

7 Implications for Hypothesis Testing

The results above allow one to explore different hypothesis testing scenarios that might be plausible in the Southern California and Chicago settings. For a region with 5 sites and the variability observed in the Chicago region, Figure 1 shows, corresponding to the test critical value (1 or .08 for the 1-hour and 8-hour standards, respectively), the probability of the region being declared in violation, as a function of the (square root) ozone mean level in standard deviation units from the null mean, μ_0, i.e. $\delta = \frac{\mu - \mu_0}{\sigma}$ where σ is taken from Table 1. Probabilities are plotted for the 1-hour EPA statistic, $\hat{\theta}_1$, 8-hour EPA statistic, $\hat{\theta}_2$, 1-hour MLE, $\hat{\theta}_1^*$, and 8-hour MLE, $\hat{\theta}_2^*$. Binomial and Normal distributions were used to compute probabilities for all but $\hat{\theta}_2$, which required Monte Carlo calculation. Deviations corresponding to the observed Chicago 1-hour and 8-hour means are indicated with arrows. Assuming that these were true means they show low probability of declaring Chicago in violation of the 1-hour standard and almost certain declaration of violation under the 8-hour standard.

From this perspective, neither the 1-hour nor 8-hour standard is uniformly more strict; in fact there is very little difference between the ML estimators $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$. Considering the EPA statistics $\hat{\theta}_1$ and $\hat{\theta}_2$, $\hat{\theta}_2$ is associated with fewer false declarations of violation only when one is well within the compliance region (say, $\delta < -0.2$), but is more likely than $\hat{\theta}_1$ to result in a false declaration of violation for values of the mean in a neighbourhood of the compliance region.

When standards are based on different random variables (e.g. X and X^*) it is not possible to compare their strictness analytically (without modeling the association between

the random variables). Empirical comparisons between the current 1-hour and 8-hour realizable standards have been made (Saylor *et al.* 1998). Prior to revising the ozone standard, EPA compared the potential impact of the 1-hour and 8-hour standards for each county in the U.S. by simulating their performance on historical data (US EPA 1996, Table 2). Based on both the number of counties and total population affected, for policy purposes the 8-hour standard appeared stricter. For scientific purposes, however, analytical or numerical comparison of proposed standards based on their distributional properties should be considered.

Figure 2 shows operating characteristic (OC: probability of declaring violation) curves for $\hat{\theta}_1$, $\hat{\theta}_1^*$, $\hat{\theta}_2$ and $\hat{\theta}_2^*$, for regions with variability and length of season corresponding to Southern California and Chicago and I = 1 and 5 sites per region. Four panels illustrate the probabilities separately for the 1-hour and 8-hour standards and show the probability of the statistic exceeding the specified threshold (and hence the region being declared in violation) expressed both in terms of the parameter θ (bottom axis) and the median (top axis) ozone level in ppm (i.e. the square of the mean on the square-root scale). As for Figure 1, probabilities for the 8-hour EPA statistic, $\hat{\theta}_2$, were computed by Monte Carlo. The test critical values (1 and .08 for the 1- and 8-hour standards, respectively) are indicated on each graph. Observed median ozone concentrations are indicated by arrows for the Chicago plots; observed median ozone concentrations for Southern California exceed the range of the horizontal axes, but are such that declaration of violation is close to certain for both 1-hour and 8-hour standards in this region.

For 5 effective sites, the probability of declaring violation for the naive statistics $\hat{\theta}_1$ and $\hat{\theta}_2$, is reasonable at the specified threshold, but just below the threshold there is clearly

a high probability of false declaration of violation. It is also clear that the probability of false declaration of violation increases with the number of monitoring sites, i.e. a region is punished for having more sites. Again, this is a result of a region being declared in violation if *any* site within the region is in violation. Although this calculation depends greatly on the assumption of independent sites, this intuitively plausible result holds in general.

The ML statistics $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$ should obviously be implemented at different (lower) thresholds if the goal is to have high probability of declaring violation in the regions defined by the current and revised standards. For instance, calculations show that in a hypothetical region with 5 sites and the variability of Southern California, $\hat{\theta}_1^*$ should be compared against a threshold of 0.85 (as opposed to 1) and $\hat{\theta}_2^*$ against a threshold of .079 (as opposed to .08) to ensure a 95% chance of being declared in violation when the true median is 1 or .08 respectively. The steeper MLE OC curves indicate better discrimination between compliance and violation for an appropriately chosen threshold.

As discussed above, one possible form of a statistically verifiable ideal standard is based on an ideal standard in a hypothesis testing setting where a second, lower, "safe" threshold is specified (e.g. background levels) and the implementation of the standard is required to satisfy specified type I and II error rates relative to these thresholds. The receiver operating characteristic (ROC) curves in Figure 3 show the trade-off between type I and type II error as the critical level of the test is varied, for the 8-hour EPA and MLE realizable standards in a region with 5 sites and with the variability and length of season observed in Chicago. The Figure shows the probability of being declared in violation given $\theta_2 = c_U$=.080 ("true positive") versus probability of being declared in violation given $\theta_2 = c_L$=.078 ("false positive") for varying critical levels "c". As for Figure 1, probabilities for the 8-hour EPA

statistic, $\hat{\theta}_2$, were computed by Monte Carlo. We see, for example, that in this setting, while 5% type I and II error rates would be achievable using the maximum likelihood estimate, this is not possible using the naive statistic. This example illustrates that the choice of critical level for the implementation of a standard should also include statistical considerations and need not equal either of the specified upper ("unsafe") and lower ("safe") thresholds.

Specification of the standards in a statistical framework allows more efficient estimation of the parameters of interest and hence construction of statistical tests (implementation of the standards) with more desirable characteristics.

8 Plausibility of the assumptions

There are two main sets of assumptions on which the calculations in this paper are based. The first deals with the probabilistic structure of the data, namely that the observations are independent and identically distributed; and the second deals with the distribution of the data, namely that square roots of daily maxima are normally distributed. In this section we discuss the validity of these assumptions, and the consequences for more realistic regional ozone models.

The independence assumption has two facets: temporal and spatial. Analysis of the California and Chicago data indicates that an AR(2)-model would capture most of the temporal dependence. The observed AR(2) structure implies the likelihood of ozone *episodes*, yet the current formulations of ozone realizable standards do not distinguish between one exceedance per year for each of three years and three exceedances in a single year (and no exceedances in the other two years). The spatial dependence in these two regions is

substantial: if one site violates the standard, the probability that nearby sites are also in violation is high. Correlations between measurements at sites within the region, even after correcting for the seasonality, are on the order of 0.8 (range 0.72 to 0.90) in the Chicago data.

The effect of spatial dependence on the distributional characteristics of the test statistics depends on the form of the statistic being considered and is different for the naive and non-naive statistics considered here. To explore this issue, we considered simulations comparing the distributions of the statistics $\hat{\theta}_1, \hat{\theta}_1^*, \hat{\theta}_2$ and $\hat{\theta}_2^*$ for settings where there were 1-10 independent sites (i.e. no spatial dependence) with that where there were a nominal 10 sites but where between sites correlation was .8. For the purposes of this analysis, we assumed temporal independence and identical distributions (spatially and temporally) based on the summary statistics in Table 1. The distribution of the naive statistics $\hat{\theta}_1$ and $\hat{\theta}_2$, based on 10 correlated sites resembled that based on 5 independent sites rather than the nominal 10. For the statistics $\hat{\theta}_1^*$ and $\hat{\theta}_2^*$, the reduction in *effective* number of sites was more extreme and the distribution of these statistics based on 10 correlated sites resembled that based on 1-2 independent sites.

The assumption of temporally identical distributions fails to take into account the seasonal structure of ozone time series. There is a pronounced seasonal effect in both the data sets we consider. A better model would include a seasonally varying mean and variance. However, our analysis of data from these two regions suggests that the definition of an "ozone season" is fairly arbitrary in that the seasonal pattern of California suggests a similar ozone season to the Chicago region. Arguments could be made in favour of a uniform shorter ozone season.

To address the assumption of an identical distribution spatially (i.e. that the distribution is the same at each site) we also considered simulations based on a nominal 10 sites as above, but allowing varying site-specific means, variances and between-site covariances. The choices of these parameters were based on the empirical estimates for California and Illinois. Perhaps not surprisingly, our analyses indicate that the naive estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ will be dominated by sites with more extreme means and/or variances, so that the distribution with a nominal 10 sites (with or without independence across sites) resembles that of 2 to 3 sites. This is a consequence of the use of the *maximum* operator in the form of these statistics and emphasizes again that the use of the maximum does not have good statistical properties. The insight gained from the above exploration into *effective* numbers of sites relative to the nominal number also has implications for network design and sampling.

The distributional assumption appears viable for the data sets considered in this study. Cox et al. (1998), looking at different data, found that the square root of 8-hour daily maxima followed a normal distribution, while the 1-hour daily maxima were better described by a lognormal distribution. Other workers have generally used a square root transformation for purposes of meteorological adjustment and trend estimation (Thompson et al. 2000).

9 Discussion

We did not intend in this paper to conduct a formal modeling exercise for regional ozone. Rather we have considered a simplified setting for the purpose of making general statistical points about setting and evaluating regulatory standards. The ideas that we have developed regarding a hypothesis testing framework and the evaluation of the performance of statistics

associated with realizable standards are applicable more generally than just to the setting of ozone standards. The challenge of how to address temporal and spatial correlation should be addressed in the formulation of the standard and the choice of statistics for its implementation.

One of the many issues that warrants further consideration is the choice of the form of the ideal standard, by which we mean the underlying random variable(s), the parameter(s) (of the random variables) on which the ideal standard is based and the threshold that the standard associates with the parameter(s). Another form of standard with different apparent health connections is, for instance, the *area over threshold* standard based on cumulative hourly ozone exposure for periods exceeding a certain threshold (see, e.g., Leadbetter (1991, 1994)). In principle, although perhaps less feasibly in practice, a standard might be specified in the form of a statistically verifiable ideal standard and its implementation (choice of sampling network, realizable standard) could be carried out separately over regions.

The present form of the US ozone ideal standards is defined at individual locations and implicitly assumed to apply everywhere. The associated realizable standards apply at monitoring sites which were located for specific purposes according to scientific judgement; the standards do not attempt a characterization of pollution and exceedances across a region in a statistically rigorous way. This is an issue which will be important, but statistically challenging, to address. The development of spatial standards appears to be an important line of research in this regard.

The regional implementation of a standard involves the choice of monitoring locations. Further work is needed here in network design and in exploring the consequences of size-based designs where monitors are located according to anticipated magnitudes of ozone

concentrations. The number of sites needed for a given precision (within the framework of the standard) will be a consideration. Our analyses suggest, for instance, that Chicago and Southern California could obtain the same information with fewer sites.

Statistical considerations will only be one component of the process that leads to the constitution of a standard. The choice of thresholds, for instance, will involve scientific and policy considerations. However, as we have discussed above, the critical value of an associated statistical test need not equal the regulatory threshold. The current thresholds for the US 1- and 8-hour ozone standards are based on a "knife-edge" principle, with the intention of equity across regions: e.g. the notion that tests at the hypothesized mean of one exceedance per year conducted in two different regions perform the same regardless of the respective standard deviations. The analyses above have illustrated that such equity is not achieved in the current implementation of the standards. In practice, as we illustrate above, given two regions with the same mean ozone level, both the 1-hour and 8-hour standards can be *stricter* , viz., identify more violations of the standard, for the region with higher variability (σ) and/or longer season (K).

In summary, we have demonstrated that, regardless of the science that determines the choice of the parameter and the associated thresholds on which the ideal standard is based, there is much to be gained by a statistical approach to the implementation of the standard. Within a hypothesis testing framework, statistical considerations can facilitate the choice of test statistic and critical level and will allow quantification of error rates. The current choice of test statistics in the existing ozone standards have been shown to be problematic in terms of their statistical properties. If standards are to be developed which treat different regions equitably, further work is needed on issues of network design and the development

of statistically verifiable ideal standards that acknowledge spatial and temporal variability.

Acknowledgement

The authors acknowledge helpful discussions with Peter Guttorp and Ronit Nirel. The information in this article has been funded wholly or in part by the United States Environmental Protection Agency under cooperative agreement CR825173-01-0 with the University of Washington. It has been subjected to Agency review and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. The opinions expressed are those of the authors; no Agency policy or position should be inferred.

References

[1] AIRS: http://www.epa.gov/airs/

[2] Barnett, V., and O'Hagan, A. (1997). Setting environmental standards: the statistical approach to handling uncertainty and variation. London: Chapman & Hall.

[3] CARB: http://www.arb.ca.gov/homepage.htm

[4] Carbonez, A. El-Shaarawi, A.H., and Teugels, J.L. (1999). Maximum microbiological contaminant levels. *Environmetrics* **10**: 79-86.

[5] Cox, L.H., Guttorp, P., Sampson, P.D., Caccia, D.C., and Thompson, M.L. (1998). A preliminary statistical examination of the effects of uncertainty and variability on environmental regulatory criteria for ozone. In: *Environmental statistics: analysing data for environmental policy.* Wiley, Chichester, 122-143.

[6] Guttorp, P. (2000). Setting environmental standards: a statistician's perspective. National Research Center for Statistics and the Environment, Technical Report, Number 48.

[7] Leadbetter, R.L. (1991). On a basis for "peaks over threshold" modeling. *Statistics and Probability Letters* **12**: 357-362.

[8] Leadbetter, R.L. (1994). On high level exceedance modeling and tail inference. *Journal of Statistical Planning and Inference* **45**: 247-260.

[9] Rice, J.A. (1995). Mathematical statistics and data analysis. 2nd Edition. Belmont: Duxbury Press.

[10] Saylor, R.D., Chameides, W.L. and Cowling, E.B. (1998) Implications of the new ozone National Ambient Air Quality Standards for compliance in rural areas. *Journal of Geophysical Research - Atmos* **103**: 31137-31141.

[11] Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P. and Sampson, P.D. (2000) A review of statistical methods for the meteorological adjustment of tropospheric ozone. *Atmospheric Environment*. To appear.

[12] US EPA. (1996). Fact Sheet. Washington DC: USEPA Office of Communications, Education and Public Affairs. Press Release, November 27, 1996.

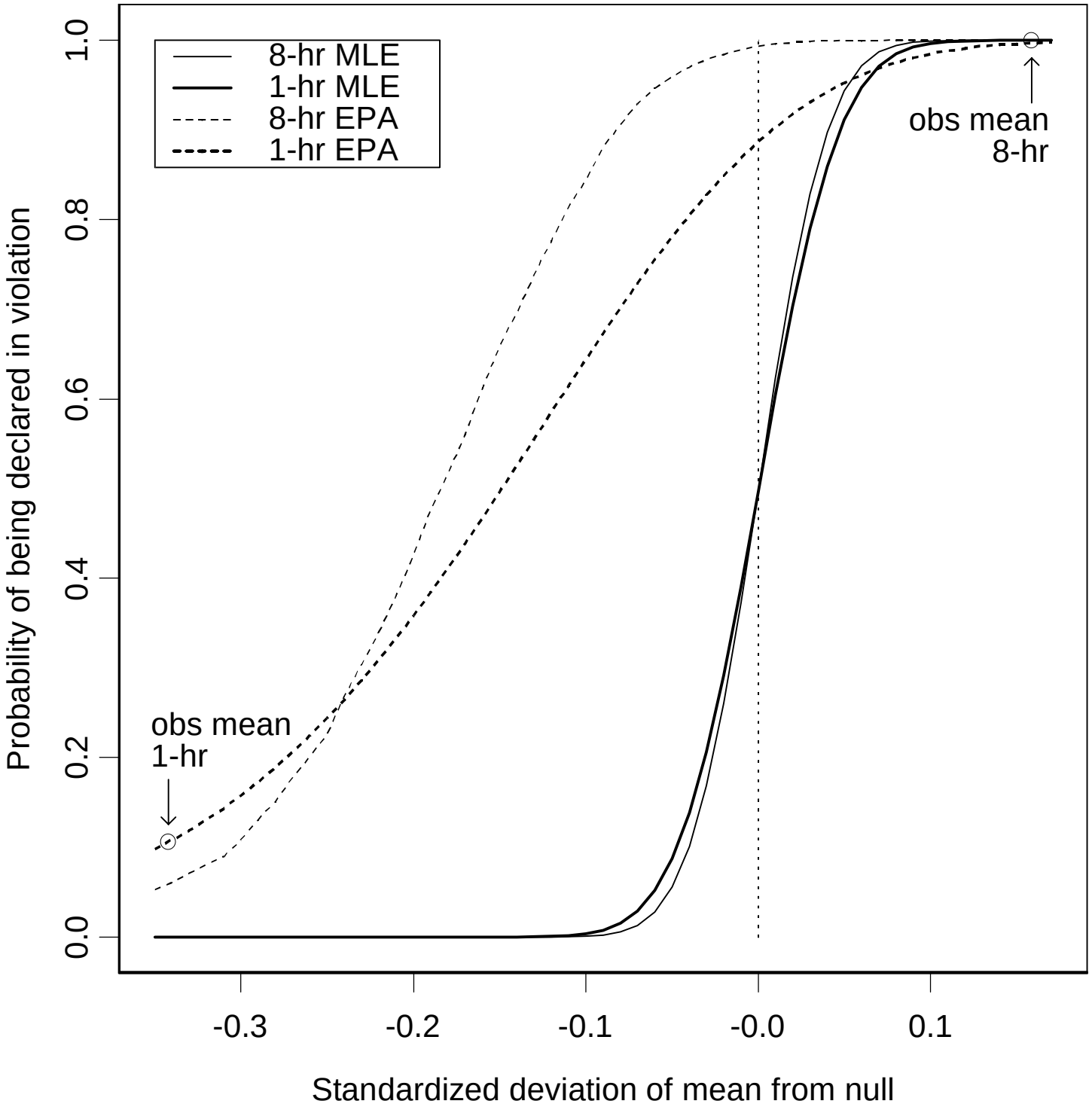
[13] US Federal Register. (1997). National ambient air quality standards for ozone: final rule. US Federal Register, vol 62, no. 138, July 18, 1997.

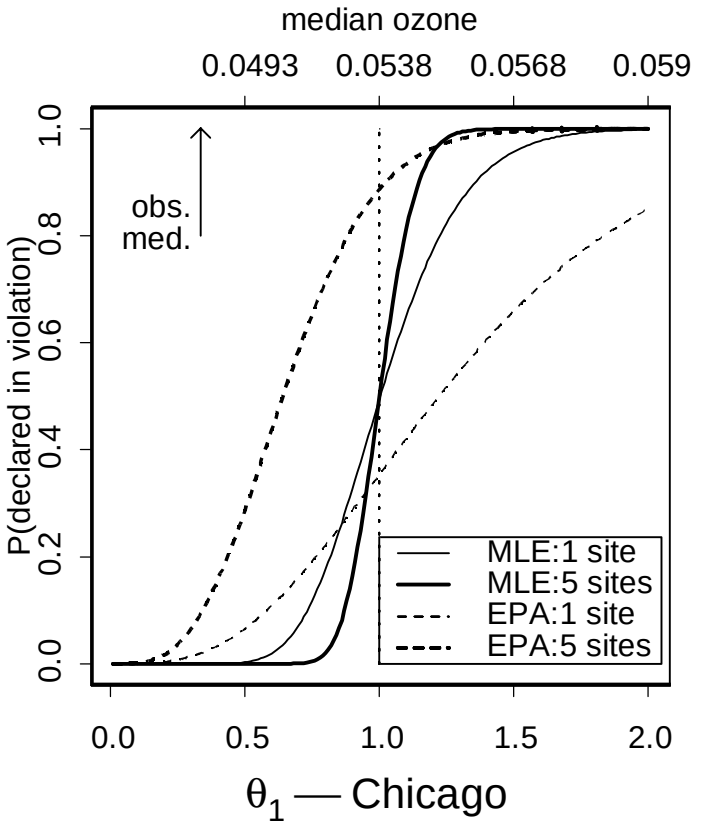
Figure Legends

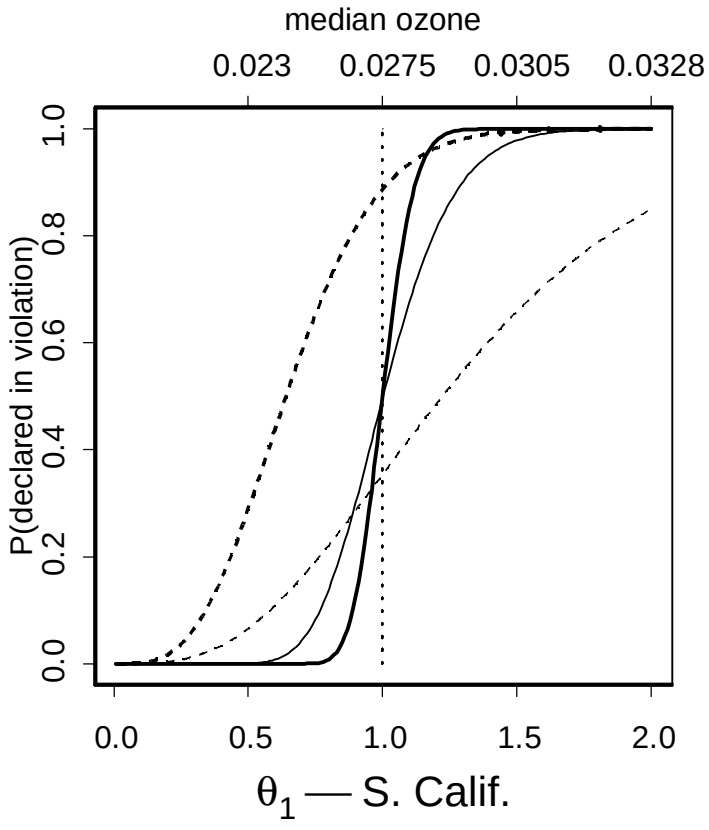
Figure 1. Probability of being declared in violation as a function of standardized deviation of the (square root) ozone mean concentration from the null mean corresponding to the test critical value (1 exceedance per year for the 1-hour standard, .08 ppm average concentration for the 8-hour standard). Standard deviations used for determination of the null mean and standardization of deviations from the null mean are the empirical values for Chicago from Table 1. Deviations corresponding to observed Chicago 1-hour and 8-hour means are indicated with arrows.

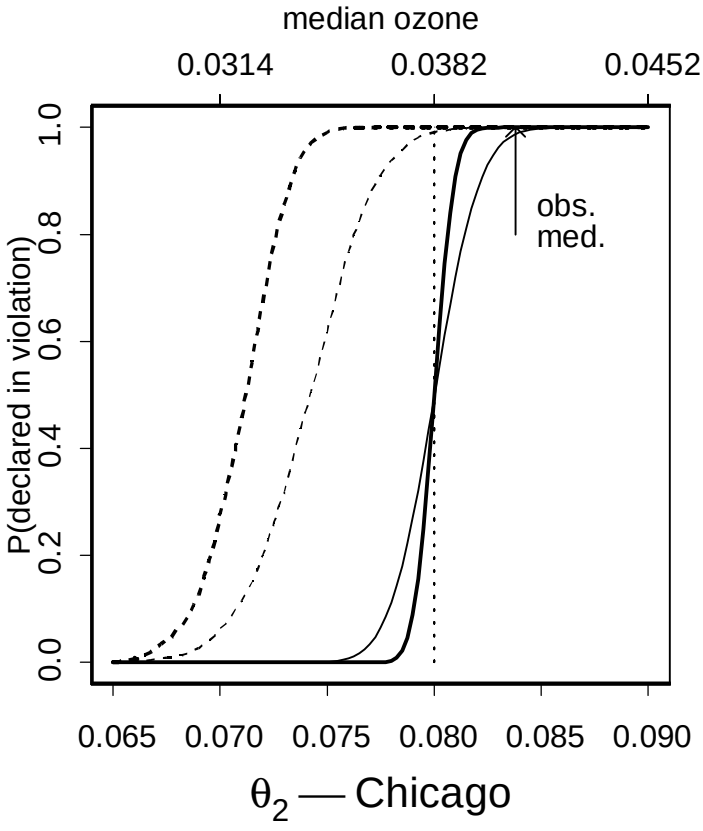
Figure 2. Operating characteristic (OC) curves: the four panels illustrate the probability of being declared in violation for the 1-hour and 8-hour standards as a function of the true value of θ and corresponding median ozone concentration under scenarios of 1 and 5 monitoring sites, for the length of season and variability of the Chicago and Southern California monitoring data from Table 1. Observed median ozone concentrations are indicated by arrows in the Chicago plots, those for Southern California exceed the range of the horizontal axes.

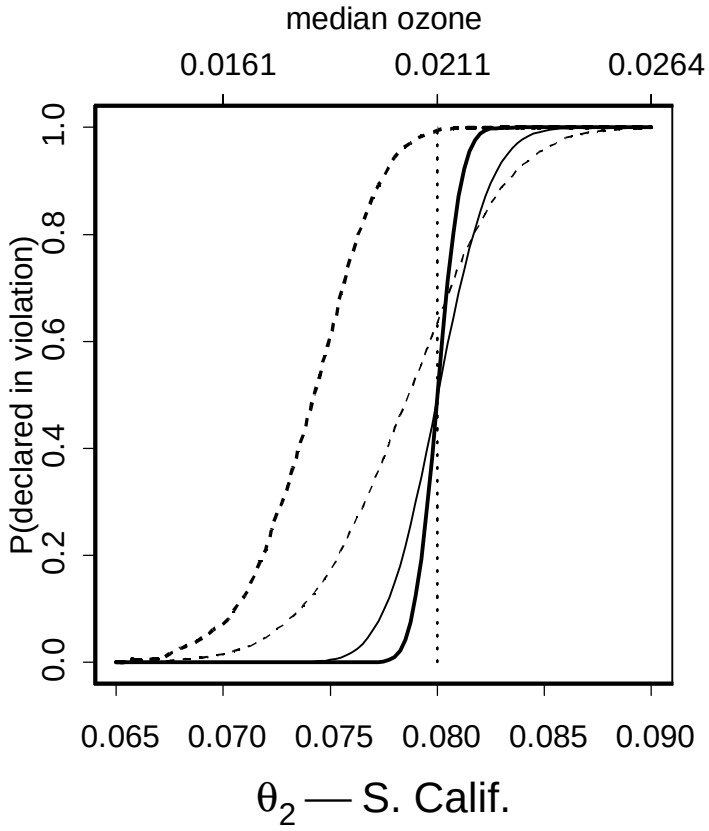
Figure 3. Receiver operating characteristic (ROC) curves: probability of being declared in violation given θ_2=.080 ("true positive") versus probability of being declared in violation given θ_2=.078 ("false positive") for varying critical levels "c" for the 8-hour EPA and MLE statistics assuming a region with 5 sites and the variability and length of season for Chicago from Table 1.

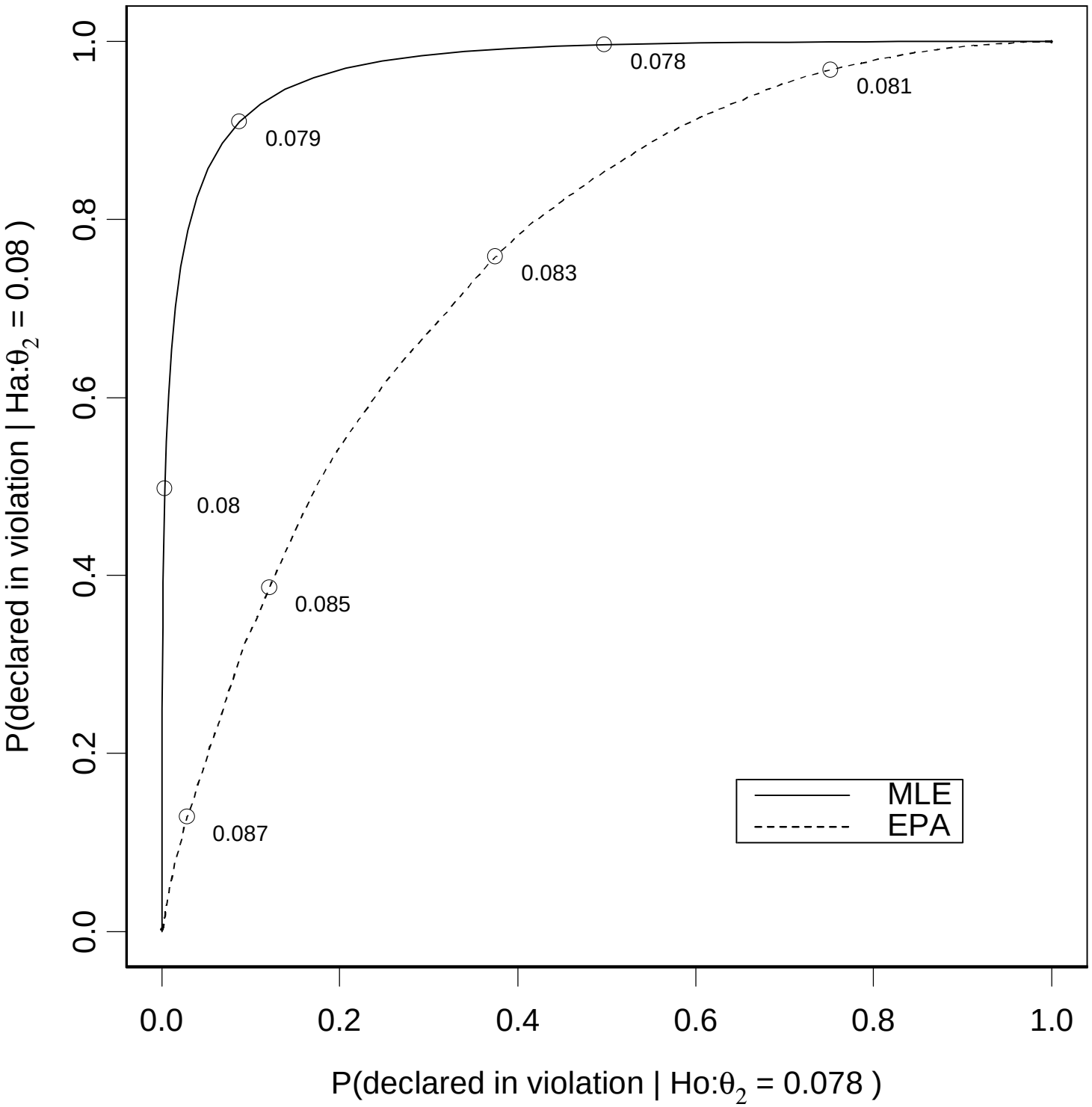

8-hr MLE
1-hr MLE
8-hr EPA
1-hr EPA
obs mean
8-hr
obs mean
1-hr
Probability of being declared in violation
0.0
0.2
0.4
0.6
0.8
1.0
-0.3
-0.2
-0.1
-0.0
0.1
Standardized deviation of mean from null


median ozone
0.0493
0.0538
0.0568
0.059
P(declared in violation)
0.0
0.2
0.4
0.6
0.8
1.0
obs.
med.
MLE:1 site
MLE:5 sites
EPA:1 site
EPA:5 sites
0.0
0.5
1.0
1.5
2.0
θ1 — Chicago


median ozone
0.023
0.0275
0.0305
0.0328
P(declared in violation)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
θ1 — S. Calif.


median ozone
0.0314
0.0382
0.0452
P(declared in violation)
0.0
0.2
0.4
0.6
0.8
1.0
obs.
med.
0.065
0.070
0.075
0.080
0.085
0.090
θ2 — Chicago


median ozone
0.0161
0.0211
0.0264
P(declared in violation)
0.0
0.2
0.4
0.6
0.8
1.0
0.065
0.070
0.075
0.080
0.085
0.090
θ2 — S. Calif.


1.0
0.8
0.6
0.4
0.2
0.0
P(declared in violation | Ha:θ2 = 0.08)
0.078
0.081
0.079
0.083
0.08
0.085
0.087
MLE
EPA
0.0
0.2
0.4
0.6
0.8
1.0
P(declared in violation | Ho:θ2 = 0.078)