AVERAGING TIME AND MAXIMA FOR AIR POLLUTION CONCENTRATIONS

RICHARD E. BARLOW
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

The Public Health Service, and now the Environmental Protection Agency (EPA), has operated a Continuous Air Monitoring Program (CAMP) since January 1962 (see Larsen [7]). Under CAMP, air pollutant concentrations are punched automatically into a computer tape every five minutes. Air pollutants which are being monitored include carbon monoxide, various hydrocarbons, nitric oxide, nitrogen dioxide, total oxidants (chiefly ozone), and sulfur dioxide. Monitoring stations are located in Chicago, Cincinnati, Los Angeles, New Orleans, Philadelphia, San Francisco, and Washington, D.C. Measurements are recorded in parts per million (ppm), parts per hundred million (pphm), and parts per billion (ppb), and micrograms per cubic meter ($\mu g/m^3$). For example, oxidant, a chief constituent of smog, is considered undesirable if its concentration reaches or exceeds 0.1 ppm.

In San Francisco, the Bay Area Air Pollution Control District (BAAPCD) publishes, on a monthly basis, daily average high hour oxidant values as well as daily peak oxidant values. Carbon monoxide values are similarly recorded. However, sulfur dioxide values (in ppb) and particulate values ($\mu g/m^3$) are recorded only as 24 hour averages. Averaging times vary widely because of the nature of the pollutant and the monitoring system used. For example, particulate matter is measured by the high volume sampler. In this device, air is blown through a filter which is then weighed after 24 hours. In the San Francisco Bay Area, particulate readings tend to be made (for a 24 hour period) every other day and occasionally every third day. Particulate readings are recorded at nine locations in the Bay Area and there are wide variations in the data due to location.

Table I, taken from the pamphlet "Air Pollution and the San Francisco Bay Area" provides a summary of the main air pollutants, the 1969 California state standards for these pollutants and reasons for controlling their concentrations.

For purposes of evaluating air quality, it is important to know the probability of maximum pollutant concentrations exceeding state standards which are

This research has been partially supported by the Office of Naval Research under Contract N00014-69-A-0200-1036 and the National Science Foundation under Grants GP-29123 and GK-23153 with the University of California. Reproduction in whole or in part is permitted for any purpose of the United States Government.

TABLE I

AIR QUALITY STANDARDS—BAY AREA 1969

Substance	State standard	Objective
Oxidant	0.1 ppm for 1 hour	To prevent eye irritation and possible impairment of lung function in people with respiratory problems. Also to prevent damage to vegetation
Carbon monoxide	20 ppm for 8 hours	To prevent carboxyhemoglobin levels greater than 2%
Sulfur dioxide	0.04 ppm for 24 hrs. (particulate $>100\ \mu g/m^3$)	To prevent possible increase in chronic respiratory disease and damage to vegetation
	0.5 ppm for 1 hour (regardless of particulate)	To prevent possible alteration in lung function; also odor prevention
Particulate matter	$60\ \mu g/m^3$ ann. geom. mean	To improve visibility and prevent acute illness when present with about 0.05 ppm sulfur dioxide.
	No single 24 hour sample to exceed $100\ \mu g/m^3$	
Visibility reducing particles	Visibility of not less than 10 mi. when relative humidity is less than 70%	To improve visibility
Nitrogen dioxide	0.25 ppm for 1 hour	To prevent possible risk to public health and atmospheric discoloration
Hydrogen sulfide	0.03 ppm for 1 hour	To prevent odor

stated for various averaging times. We use extreme value theory to determine the limiting distribution of maximum air pollutant concentrations as a function of averaging time. Bounds on the location parameter of the corresponding extreme value distribution are used to evaluate air quality. These are then used to evaluate suspended particulate data.

2. Larsen's results

R. I. Larsen and co-workers in a series of papers [6], [7] analyzed three years of gaseous air pollutant data, from December 1961 to December 1964, for the seven cities previously mentioned. They stated [6] two main conclusions:

(1) Concentrations are approximately lognormally distributed for all pollutants in all cities for all averaging times.

(2) The median concentration (50th percentile) is proportional to averaging time to an exponent (and thus plots as a straight line on logarithmic paper).

Figure 1, [6], is a plot of three years' data for Washington, D.C. illustrating empirically Larsen's second assertion.

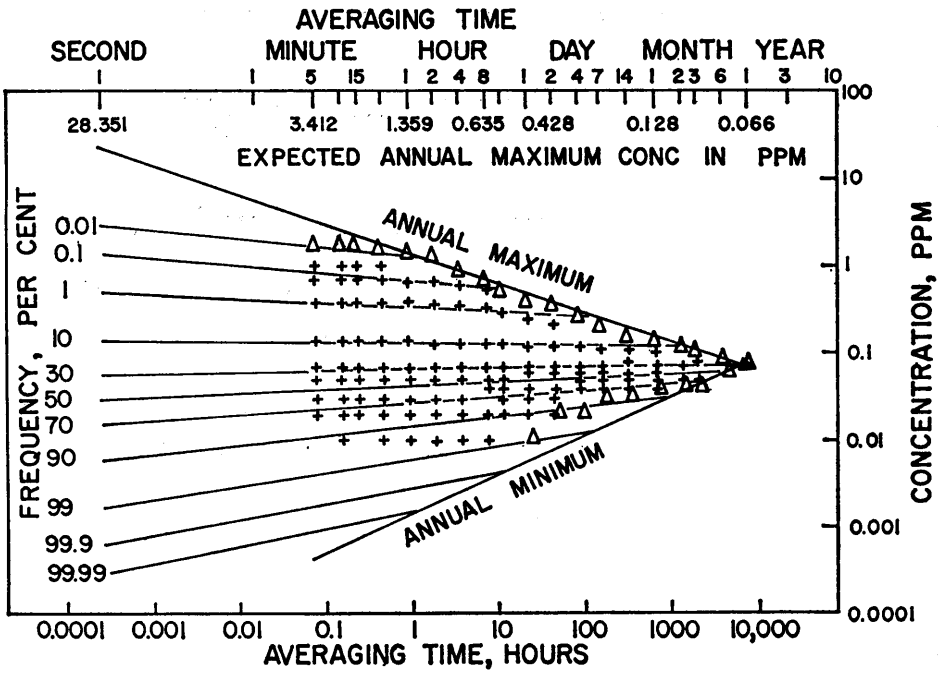

FIGURE 1

Concentration *versus* averaging time and frequency for nitrogen oxides in Washington from 12/1/61 to 12/1/64.

Since sums of (independent) lognormal distributed random variables are not distributed as lognormal random variables, Larsen's first result might be considered suspect. Histograms of air pollution concentration data are highly skewed, much as are life test data plots. One suspects that the data might as easily be fit with gamma or Weibull distributions as with a lognormal distribution. The randomness in air pollutant concentrations results mainly from meteorological phenomena. For this reason, the observations will not be independent. However, observations averaged over long time periods and within a given season of the year may be considered to be statistically independent and identically distributed.

Recently, N. D. Singpurwalla has interpreted Larsen's results using extreme value theory [11]. However, he again approximates sums of lognormal random variables by lognormal random variables.

3. A mathematical model based on extreme value theory

Suppose n observations $x_1, x_2, \cdots, x_k, x_{k+1}, \cdots, x_n$ are taken, say over a year or a season. We assume, for now, that observations are independent with distribution F. For Larsen [7], F is the lognormal distribution. He estimates parameters from the empirical distribution. Let $x_1 + x_2 + \cdots + x_k$ have distribution F_k so that F_k is the k-fold convolution of F with itself. Consider averages of length k

$$\frac{x_1 + x_2 + \cdots + x_k}{k}, \frac{x_{k+1} + \cdots + x_{2k}}{k}, \cdots, \frac{x_{n-k+1} + \cdots + x_n}{k} \tag{3.1}$$

where $k \ll n$. Let

$$\eta_{k,n} = \max\left\{\frac{x_1 + x_2 + \cdots + x_k}{k}, \cdots, \frac{x_{n-k+1} + \cdots + x_n}{k}\right\}. \tag{3.2}$$

We are interested in the behavior of $\eta_{k,n}$ as a function of the averaging time, k. Larsen [6] estimates the median of $\eta_{k,n}$ (say $M_{k,n}$) by

$$M_{k,n} = Ck^{-b} \tag{3.3}$$

where $b > 0$ is tabulated [6] as a function of various one hour standard geometric deviations. These values were apparently computed empirically from data.

Fix k and let $\alpha_{k,n} > 0$ and $\beta_{k,n}$ be a sequence of norming constants such that

$$\lim_{n\to\infty} P\left[\frac{\eta_{k,n} - \beta_{k,n}}{\alpha_{k,n}} \leqq x\right] \tag{3.4}$$

exists and is nondegenerate. Gnedenko [4] showed that, for nonnegative random variables, there are only two possible limiting distributions (up to scale and location) namely

$$\Lambda(x) = \exp\{-e^{-x}\} \qquad -\infty < x < \infty \tag{3.5}$$

and

$$\Phi_\alpha(x) = \begin{cases} 0 & x \leqq 0 \\ \exp\{-x^{-\alpha}\} & x > 0 \end{cases} \qquad \alpha > 0. \tag{3.6}$$

The lognormal, Weibull, gamma and most other commonly used distributions lie in the domain of attraction of Λ; that is, there exist constants $\alpha_{k,n} > 0$, $\beta_{k,n}$ for these distributions such that (3.4) equals $\Lambda(x)$. Marcus and Pinsky [9] give necessary and sufficient conditions for a distribution to lie in the domain of attraction of Λ. For some $\alpha > 0$, F belongs to the domain of attraction of Φ_α if and only if $1 - F(x) = x^{-\alpha}L(x)$ where

$$\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1 \tag{3.7}$$

for each $x > 0$. (See [3], pp. 270–272). Intuitively, Λ seems the more reasonable limiting distribution for air pollution concentrations and we make that assumption henceforth. In particular, we assume that F_k belongs in the domain of attraction of Λ.

Let $G(x) = 1 - e^{-x}$ for $x \geqq 0$ and $R_k(x) = G^{-1}F_k(x)$. Gnedenko [4] (see [9]) showed that the norming constants could be expressed as

$$\beta_{k,n} = \frac{R_k^{-1}[\log n]}{k} \tag{3.8}$$

and

$$\alpha_{k,n} = \frac{R_k^{-1}[1 + \log n] - R_k^{-1}[\log n]}{k} \tag{3.9}$$

Hence for large n,

$$P[\eta_{k,n} \leqq x] \sim \Lambda\left[\frac{x - \beta_{k,n}}{\alpha_{k,n}}\right] \tag{3.10}$$

where $\beta_{k,n}$ and $\alpha_{k,n}$ are given by (3.8) and (3.9). $\beta_{k,n}$ is the location parameter and also approximately the 37th percentile of $\Lambda[(x - \beta_{k,n})/\alpha_{k,n}]$. Since $\alpha_{k,n}$ is typically small relative to $\beta_{k,n}$, the latter provides a convenient way of summarizing $\eta_{k,n}$. The distribution of Λ is tabulated by Owen [10].

The main difficulty in using $\beta_{k,n}$ occurs in computing the convolution F_k. In the case where

$$F(x) = \int_0^x \frac{u^{\lambda-1}e^{-u/\theta}}{\theta^\lambda \Gamma(\lambda)}\, du \tag{3.11}$$

that is, the gamma distribution, then, of course, F_k is again a gamma distribution and there is no problem in computing R_k. For n large and $k \ll n$, Gurland [5] has approximated $\beta_{k,n}$ as

$$\beta_{k,n} \sim \frac{1}{k}\left[\theta \log\left(\frac{n}{\left(\frac{\theta}{k}\right)^{\lambda k-1}\Gamma(\lambda k)}\right) + \theta(k\lambda - 1)\log\left(\frac{\theta}{k}\log n\right)\right] \tag{3.12}$$

where $\Gamma(\cdot)$ is the gamma function. Hence,

$$\beta_{k,n} \sim \frac{\theta}{k}\log n = C_n k^{-1} \tag{3.13}$$

for large n. Since the right tail of the gamma distribution behaves like the exponential distribution, (3.13) is not surprising. If we let

$$F(x) = 1 - \exp\left\{-\left(\frac{x}{\delta}\right)^{1/b}\right\} \qquad \text{for} \quad x \geqq 0 \tag{3.14}$$

then $\beta_{k,n}$ behaves like k^{-b} in a sense to be made precise. Of course, $b = 1$ corresponds to the exponential distribution. The lognormal distribution has an $R(x)$ which is first convex and then concave over adjacent intervals.

4. Bounds on $\beta_{k,n}$

We wish to obtain bounds on $\beta_{k,n}$ for distributions other than the gamma distribution. To motivate this discussion, consider the Weibull distribution

(3.14). Unfortunately, numerical methods are necessary to compute convolutions of Weibull distributions. For this distribution,

$$R(x) = G^{-1}F(x) = \left(\frac{x}{\delta}\right)^{1/b}. \tag{4.1}$$

Note that for $0 < b < 1$, R is convex while for $b > 1$, R is concave. Many other distributions useful in life testing, especially, have the property that R is convex for certain parameter values and concave for others. [A distribution F for which R is convex (concave) is called IFR (DFR) in Barlow and Proschan [2].] It is easily seen that if $R(0) = 0$ and R is convex, then R is *superadditive*, that is,

$$R(x + y) \geqq R(x) + R(y) \qquad \text{for} \quad x, y \geqq 0. \tag{4.2}$$

This weaker property is sufficient to provide one sided bounds on $R_k(x)$ and hence on $\beta_{k,n}$.

THEOREM 4.1. *If F is continuous, $F(0) = 0$ and R is superadditive (subadditive), then*

$$R_k(x) \leqq (\geqq) G^{-1}\Gamma_k[R(x)] \qquad x \geqq 0, \tag{4.3}$$

where $\Gamma_k(x) = 1 - e^{-x}\left[\sum_{j=0}^{k-1} \frac{x^j}{j!}\right]$ *for* $x \geqq 0$ *is the gamma distribution and* $G \equiv \Gamma_1$.

PROOF. Assume R is superadditive so that R^{-1} is subadditive, that is,

$$R^{-1}(x + y) \leqq R^{-1}(x) + R^{-1}(y) \qquad \text{for} \quad x, y \geqq 0. \tag{4.4}$$

Let $Y_j, j = 1, 2, \cdots, n$, be i.i.d. random variables with distribution $G(x) = 1 - e^{-x}$ for $x \geqq 0$. Then $X_j = R^{-1}(Y_j), j = 1, 2, \cdots n$, are i.i.d. with distribution F. Since R^{-1} is subadditive,

$$\sum_{j=1}^{n} R^{-1}(Y_j) \geqq R^{-1}\left(\sum_{j=1}^{n} Y_j\right). \tag{4.5}$$

Then

$$\begin{aligned} 1 - F_n(x) &= P\left[\sum_{j=1}^{n} X_j > x\right] \geqq P\left[R^{-1}\left(\sum_{j=1}^{n} Y_j\right) > x\right] \\ &= P\left[\sum_{j=1}^{n} Y_j > R(x)\right] \\ &= e^{-R(x)} \sum_{j=0}^{n-1} \frac{[R(x)]^j}{j!} = 1 - \Gamma_k[R(x)]. \end{aligned} \tag{4.6}$$

Hence, $R_k(x) \leqq G^{-1}\Gamma_k[R(x)]$ as claimed. The proof for the subadditive case is similar. Q.E.D.

The proof for R convex was first noted by Erwin Straub [12]. Note that (4.3) is an equality if $k = 1$ or if $R(x) = ax$ for some $a > 0$.

The following theorem provides additional bounds.

THEOREM 4.2. *If F is continuous, $F(0) = 0$ and R is convex (concave), then*

$$R_k(x) \geqq (\leqq) G^{-1}\Gamma_k\left[kR\left(\frac{x}{k}\right)\right] \qquad \text{for} \quad x \geqq 0. \tag{4.7}$$

The proof is due to Straub [12].

We can now state useful bounds on the extreme value location parameter, $\beta_{k,n}$.

COROLLARY 4.3. *If F is continuous, $F(0) = 0$ and R is convex (concave), then*

$$\frac{1}{k} R^{-1}\Gamma_k^{-1}G(\log n) \leqq (\geqq)\ \beta_{k,n} \leqq (\geqq)\ R^{-1}\left[\frac{1}{k}\Gamma_k^{-1}G(\log n)\right]. \tag{4.8}$$

PROOF. Equation (4.8) follows from (4.3), (4.7) and repeated use of the fact that $(AB)^{-1}(x) = B^{-1}A^{-1}(x)$ where $AB(x)$ means $A[B(x)]$; that is, functional composition. Q.E.D.

From (3.12), we see that for large n and $k \ll n$

$$\Gamma_k^{-1}G(\log n) \sim \left[\log\left(\frac{n}{\Gamma(k)}\right) + (k-1)\log\log n\right] = c_{k,n} \tag{4.9}$$

so that for R convex (concave)

$$\frac{1}{k} R^{-1}[c_{k,n}] \leqq (\geqq)\ \beta_{k,n} \leqq (\geqq)\ R^{-1}\left[\frac{c_{k,n}}{k}\right]. \tag{4.10}$$

EXAMPLE. If $F(x) = 1 - \exp\{-(x/\delta)^{1/b}\}$, then $R^{-1}(y) = \delta y^b$ and

$$\begin{aligned} \beta_{k,n} &\leqq \delta(c_{k,n})^b k^{-b} \sim \delta[\log n]^b k^{-b} \\ \beta_{k,n} &\geqq \delta(c_{k,n})^b k^{b-1} \sim \delta[\log n]^b k^{b-1} \end{aligned} \tag{4.11}$$

if $0 < b < 1$. For the Weibull distribution, we see the connection between the power of the averaging time and the shape parameter of the distribution.

5. Bounds on $\alpha_{k,n}$

We could also provide bounds on

$$\alpha_{k,n} = \frac{R_k^{-1}(1 + \log n) - R_k^{-1}(\log n)}{k} \tag{5.1}$$

by using Theorems 4.1 and 4.2. However, the bounds will be less elegant than the bounds on $\beta_{k,n}$. Typically, $\alpha_{k,n}$ will be small and probability will tend to be concentrated around the location parameter $\beta_{k,n}$. When R is convex, an upper bound on $\alpha_{k,n}$ is available.

THEOREM 5.1. *If F is continuous with mean θ, $F(0) = 0$ and R is convex, then*

$$\alpha_{k,n} \leqq \frac{1}{k^2\theta} \qquad \textit{for large } n. \tag{5.2}$$

PROOF. If R is convex, then R_k is also convex by Theorem 5.1 on p. 36 of [2]. Since $R_k^{-1}(x)$ is concave, it crosses the ray $x/k\theta$ at most once, and from above. $R_k(x)$ crosses $x/k\theta$ exactly once since F_k and $G(x) = 1 - \exp\{-x/k\theta\}$ have the same mean, namely $k\theta$. Hence for large values of x, the slope of R_k^{-1} is less than the slope of $x/k\theta$. It follows that

$$\alpha_{k,n} = \frac{R_k^{-1}[1 + \log n] - R^{-1}[\log n]}{k} \leqq \frac{1}{k^2\theta}. \tag{5.3}$$

Q.E.D.

6. An example using suspended particulate data

As was mentioned in the introduction, suspended particulates are averaged over a 24 hour period. The California state standard for particulate matter is 60 $\mu g/m^3$ annual geometric mean. (The geometric mean is used because of the lognormal distribution assumption.) The state standard also specifies that no single 24 hour sample is to exceed 100 $\mu g/m^3$. A severe pollution episode occurs if average particulate values remain high for several days in succession. Hence, an interesting question to ask is, what is the probability that the maximum of, say, three day averages over the course of several seasons will exceed any specified amount? An alternative approach is to ask for upper bounds on $\beta_{k,n}$, the location parameter of $\eta_{k,n}$. Table II shows 24 hour average particulate measurements recorded in $\mu g/m^3$ for San Jose, California during November, 1969.

TABLE II

24 HOUR PARTICULATE MEASUREMENTS FOR SAN JOSE, CALIFORNIA NOVEMBER, 1969

Day in November 1969		Suspended particulates ($\mu g/m^3$)
S	1	140
Sunday	2	.
M	3	.
T	4	122
W	5	.
T	6	39
F	7	.
S	8	.
Sunday	9	.
M	10	.
T	11	129
W	12	.
T	13	147
F	14	.
S	15	31
Sunday	16	.
M	17	.
T	18	.
W	19	132
T	20	124
F	21	.
S	22	.
Sunday	23	.
M	24	.
T	25	158
W	26	.
T	27	105
F	28	.
S	29	140
Sunday	30	.

They were taken from the BAAPCD Contaminant and Weather Summary. The particulate "season" in the San Francisco Bay Area is roughly September through December, and November, 1969, was unusually high. Notice that only 11 out of 30 days were actually recorded. The mean value for this month was 115 $\mu g/m^3$.

Assuming a Weibull distribution for particulate values during this month, we found the linear invariant estimates for δ and b using tables computed by Nancy Mann [8]. For this particular data, we found $\hat{\delta} = 127.74$ and $\hat{b} = .2727$. In this case,

$$R(x) = \left(\frac{x}{\hat{\delta}}\right)^{1/b} \tag{6.1}$$

is convex. There is no *a priori* reason, however, why b should lie in [0,1]. Letting $n = 270$ days corresponding to three seasons of 90 days each, we computed the upper bounds on $\beta_{k,n}$ for $k = 1, 2, \cdots, 7$, shown in Table III. The lower bounds on $\beta_{k,n}$, however, were unreasonably low in view of the data at hand.

TABLE III

UPPER BOUNDS ON $\beta_{k,n}$ ($n = 270$)

k	($\mu g/m^3$)
1	176.25
2	145.89
3	130.68
4	120.76
5	113.61
6	108.07
7	103.62

The author would like to acknowledge Professor Nozer Singpurwalla for bringing this air pollution problem to his attention and for making available to him preprints of his papers.

REFERENCES

[1] "Air pollution and the San Francisco bay area," Bay Area Air Pollution Control District Publications, San Francisco, California, 1970.

[2] R. E. BARLOW and F. PROSCHAN, *Mathematical Theory of Reliability*, New York, Wiley, 1965.

[3] W. FELLER, *An Introduction to Probability Theory and its Applications*, Vol. 2, New York, Wiley, 1966.

[4] B. V. GNEDENKO, "Sur la distribution limite du terme maximum d'une série aléatoire," *Annals of Mathematics*, Vol. 44 (1943), pp. 423–453.

[5] J. GURLAND, "Distribution of the maximum of the arithmetic mean of correlated random variables," *Ann. Math. Statist.*, Vol. 26 (1955), pp. 294–300.

[6] R. I. LARSEN, "A new mathematical model of air pollutant concentration averaging time and frequency," *J. Air Poll. Cont. Assoc.*, Vol. 19 (1969), pp. 24–30.
[7] R. I. LARSEN and C. E. ZIMMER, "Calculating air quality and its control," *J. Air Poll. Cont. Assoc.*, Vol. 15 (1965), pp. 565–572.
[8] N. R. MANN, "Tables for obtaining the best linear invariant estimates of parameters of the Weibull distribution," *Technometrics*, Vol. 9 (1967), pp. 629–645.
[9] M. MARCUS and M. PINSKY, "On the domain of attraction of $e^{-e^{-x}}$," *J. Math. Anal. Appl.*, Vol. 28 (1969), pp. 440–449.
[10] D. B. OWEN, *Handbook of Statistical Tables*, Addison-Wesley, p. 962.
[11] N. D. SINGPURWALLA, "Extreme values from a lognormal law with applications to air pollution problems," *Technometrics*, to appear.
[12] E. STRAUB, "Application of reliability theory to insurance," ASTIN Colloquium, Randers, Denmark (1970).