Atmospheric Environment Vol. 10, pp. 941–950. Pergamon Press 1976. Printed in Great Britain.

ON FREQUENCY DISTRIBUTIONS OF AIR POLLUTANT CONCENTRATIONS

KENNETH E. BENCALA and JOHN H. SEINFELD*

Department of Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, U.S.A.

(*First received* 21 *January* 1976 *and in final form* 2 *June* 1976)

Abstract—Observed frequency distributions of air pollutant concentration levels are critically analyzed with respect to their statistical description. It is demonstrated that several common distributions can be used to fit observed data, one of which is the popular log-normal distribution. The observation that concentration distributions for all averaging times are approximately log-normal can be explained if the short averaging time data are themselves assumed to be log-normally distributed. The near log-normality of pollutant concentration frequency distributions can be explained on the basis of the near log-normality of wind speed distributions, although this explanation does not establish that wind speed distributions are solely responsible for observed concentration distributions. It is concluded that pollutant concentration frequency distributions are the result of complex phenomena and cannot be predicted exactly, but that the approximate log-normal character of the distributions is useful from a practical point of view and can be understood qualitatively on the basis of the relation between wind speed and concentration.

INTRODUCTION

Most current United States federal air quality standards are stated in terms of the yearly frequency of violation of a specified concentration level for a given averaging time. For example, the 1-h average carbon monoxide concentration may only exceed 35 ppm once during the year. This is equivalent to specifying that the 1-h average CO concentration may exceed 35 ppm only 0.011% of the time. Therefore, if the frequency distribution for hourly average CO levels in a particular urban area could be predicted, then it could be ascertained what degree of emission control would be required to enable meeting of the air quality standard.

There has been interest for some time in the frequency distributions of air pollutant concentrations. Larsen (1971) carried out a comprehensive analysis of the data collected in the Continuous Air Monitoring Program (CAMP) for the years 1962–1968. The CAMP data contain measurements of seven pollutants (CO, NO, NO_2, NO_x, oxidant, SO_2, and hydrocarbons) in eight cities. Readings were recorded every five minutes and then averaged over time periods ranging from ten minutes to one year. Based on his analysis of the data Larsen concluded that regardless of pollutant, city, or averaging time urban concentration frequency distributions could be universally represented as the log-normal. (The log-normal is a two parameter distribution and is represented by a straight line on log-probability paper.) For example, Fig. 1 shows frequency distributions for one hour averaged CO concentration in various cities. Figure 2 shows frequency distributions for CO concentrations in Chicago for various averaging times. Quantitative explanations of why the distributions tend to be log-normal are incomplete, and there is currently no way of predicting how the distributions will shift if emission levels are changed.

The objectives of this work are as follows. First, we wish to consider the question of whether or not

* To whom correspondence should be addressed.

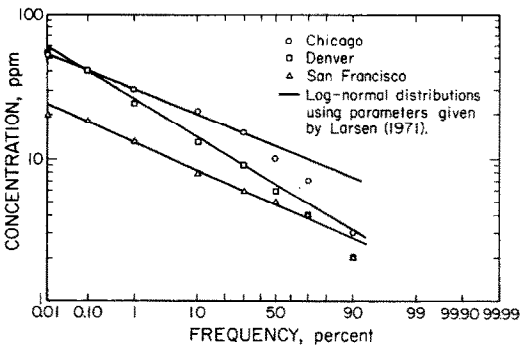

Fig. 1. 1 h average CO concentration distributions; CAMP data (1962–1968).

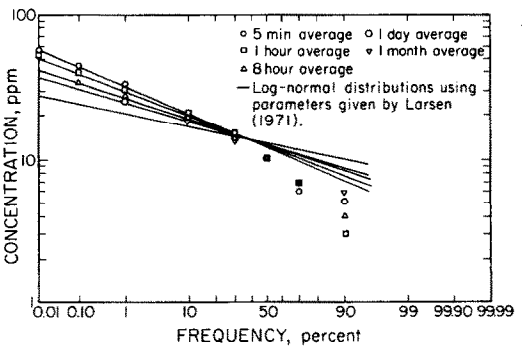

Fig. 2. CO concentration distributions in Chicago for various averaging times; CAMP data (1962–1968).

among common distributions the log-normal does, in fact, provide the best fit for pollutant concentration data. Specifically, we desire to see if other available statistical distributions provide a better fit to a selected sample of data. Second, we wish to analyze the effect of averaging time on the frequency distribution of air pollutant concentrations. We seek specifically to understand the observation that frequency distributions tend to be log-normal regardless of averaging time. Finally, we desire to analyze the possible physical reasons for the near log-normality of air pollutant data.

Our overall aim is to attempt to shed some additional light on the issue of air pollutant concentration frequency distributions. From the outset it must be recognized that these distributions are the result of many complex phenomena, and that we cannot expect to be able to predict them exactly in a given situation. Nevertheless, if the extent of validity of the distributions can be understood at least in a semi-quantitative manner, the use of the distributions can be facilitated.

REPRESENTATION OF CONCENTRATION DATA

Although the log-normal distribution has generally been used to represent air pollution concentration frequency distributions, there are other common distributions which resemble the log-normal and are candidates for representing the data (Lynn, 1974; Mage and Ott, 1975; Pollack, 1975). It is of interest to examine how well different distributions actually fit air quality data. Because chemical reaction behavior may affect the form of concentration distributions, we confine our attention to CO. Our purpose in this section is to examine the ability of a variety of statistical distributions to fit selected air quality data.

The statistics of a set of air quality data may be analyzed in terms of either its probability density function (pdf) or its distribution function. The distribution function is an integral function of the pdf, thus for the purpose of evaluating the fit of mathematical forms, comparison of the data to the pdf provides a much more stringent test than comparison to the distribution function. Unfortunately, a plot of the pdf is more difficult to interpret than is a plot of the more commonly used distribution function; therefore, we consider the distribution function.

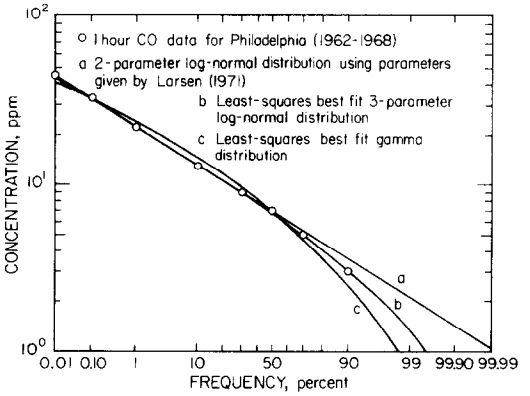

Fig. 3. CO concentration distributions in Philadelphia with lines representing the two-parameter log-normal, the three-parameter log-normal and the gamma distributions.

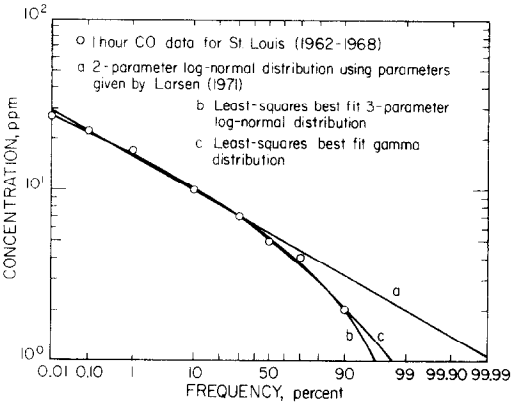

Fig. 4. CO concentration distributions in St. Louis with lines representing the two-parameter log-normal, the three-parameter log-normal and the gamma distributions.

In attempting to determine which parameter values in a particular mathematical form provide the best fit of that form to the data one has considerable latitude. We selected an unweighted least-squares criterion for determining the parameters in the distributions which best fit the data. (However, an implicit weighting arises in that relatively more data points were available at high concentration values than at low.)

Table 1 presents four potentially applicable pdf's. Figures 3 and 4 show Philadelphia and St. Louis CO data plotted against three of these distributions, the two-parameter and three-parameter log-normal and the gamma. The parameters used for the two-parameter log-normal plots are those given by Larsen, while for the other two the least-squares best fit parameters are used. On log-probability coordinates the Weibull and gamma distributions are concave while the three-parameter log-normal can be either concave or convex depending on the sign of the third parameter, δ. Of the eight data sets investigated, six are to some degree concave while two (Los Angeles and Washington) are slightly convex. Table 2 suggests possible physical interpretations of these deviations from linearity.

Table 3 presents a comparison of the sum of squares error in fitting the distribution functions to the eight data sets. In the least-squares sense the three-parameter log-normal is superior to the two-parameter distributions. The added flexibility afforded by a third parameter accounts for this. In comparing Figs. 3 and 4 with the quantitative measures of goodness of fit given in Table 3, we see that in most cases the two-paramter log-normal distribution provides a useful, if not excellent, approximation to the data, but that in that some cases it is possible to find other two-parameter distributions which provide better fits.

Table 1. Common probability density functions

Name	$p_X(x)$		
2-parameter log-normal	$(\sqrt{2\pi}\ x \ln \beta)^{-1} \exp\left\{-\frac{(\ln x - \ln \alpha)^2}{2 \ln^2 \beta}\right\}$		
		$\delta > 0$	$\delta < 0$
3-parameter log-normal	$\int_{-\delta}^{0} [\sqrt{2\pi}(t+\delta)\ln\beta]^{-1} \exp\left\{-\frac{(\ln(t+\delta) - \ln\alpha)^2}{2\ln^2\beta}\right\} dt$	$x = 0$	
	$[\sqrt{2\pi}(x+\delta)\ln\beta]^{-1} \exp\left\{-\frac{(\ln(x+\delta) - \ln\alpha)^2}{2\ln 2\beta}\right\}$	$x > 0$	$x > -\delta$
Weibull	$\alpha x^{\beta} \exp\left\{-\frac{\alpha x^{\beta+1}}{\beta+1}\right\}$		
Gamma	$\frac{x^{\alpha}}{\Gamma(\alpha+1)\beta^{\alpha+1}} \exp\left(-\frac{x}{\beta}\right)$		

Table 2. Implication of the shape of a distribution function represented on log-probability coordinates relative to a straight line (the log-normal)

	Concave shape	Convex shape
At high concentrations	Fewer days of high concentration. An upper limit in concentration suggested.	More days of higher concentration.
At low concentrations.	More days of low concentration.	Fewer days of low concentration. A lower limit (background) in concentration suggested.

Given the complex dynamic nature of urban air pollution, no one distribution will always be the best or even an adequate representation of the data, however, the two-parameter log-normal distribution is clearly a useful distribution both for describing data and for establishing an understanding of air pollutant statistics.

EFFECT OF AVERAGING TIME ON FREQUENCY DISTRIBUTIONS

Ambient data are generally reported in terms of time averaged values. The time averaged value of concentration C centered at time t and over an averaging period T can be represented as

$$\bar{C}_T(t) = \frac{1}{T}\int_{t-T/2}^{t+T/2} C(\eta)\, d\eta. \quad (1)$$

Typical values of T in air pollution applications range from 5 min to 1 y.

It has been observed that air pollutant concentration distributions approximate the log-normal regardless of averaging time, and that the median concentration is proportional to the averaging time raised to a power (Larsen, 1971). Based on this empirical observation, the standard geometric deviation σ_{gb} for one averaging time T_b can be related to the standard geometric deviation σ_{ga} for another averaging time T_a by

$$\sigma_{gb} = \sigma_{ga}^{V}, \quad (2)$$

where

$$V = \left(\frac{\ln(T/T_b)}{\ln(T/T_a)}\right)^{1/2}, \quad (3)$$

Table 3. Sum of squares error in fitting the distributions in Table 1 to 1-h average CO CAMP data, 1962–1968*

City	Two-parameter log-normal Larsen (1971) values	Best fit	Three-parameter log-normal	Weibull	Gamma
Los Angeles	0.35	0.12	0.03	1.08	0.48
Philadelphia	0.15	0.07	0.01	0.55	0.22
Denver	0.87	0.20	0.03	0.36	0.20
San Francisco	0.76	0.56	0.14	0.30	0.17
Cincinnati	0.64	0.32	0.31	1.14	0.83
St. Louis	1.25	0.44	0.04	0.13	0.04
Washington	0.31	0.08	0.05	0.78	0.57
Chicago	7.24	1.17	0.04	0.08	0.20

* Error based on reduced variate.

where T is the total period over which data are available (usually one year).

It is of interest to examine if this empirical observation can be explained strictly on the basis of the properties of log-normality distributed random variables. Thus, we ask—if the raw data averaged over a period of, say, 5 min are *assumed* to be log-normally distributed, will the data averaged over longer periods continue to be log-normally distributed.

Let $X(t)$ represent the average value of the concentration over the time period from $t - \tau$ to t. Thus, $X(t)$ is considered as the "raw" data, with an inherent averaging time τ, attributable to instrument function. The record of raw data then can be represented as the sequence, $X(t_1), X(t_2), \ldots, X(t_n)$, where $X(t_1)$ is the value in the interval $[t_1, t_1 + \tau]$, $X(t_2)$ is the value in the interval $[t_1 + \tau, t_1 + 2\tau] = [t_2, t_2 + \tau]$, etc. Now, the average concentration over the double interval $t - 2\tau$ to t is written as $Z(t) = \frac{1}{2}[X(t - \tau) + X(t)]$. The series $Z(t_2), \ldots, Z(t_n)$ then represents the sequence of concentrations averaged over periods of length 2τ. The basic problem we wish to consider is—Assuming that $X(t)$ is log-normally distributed, determine the probability density function of $Z(t)$. Thus we seek to relate the statistics of time-averaged concentrations to those of the raw data which are used to construct the averages. We carry out the analysis for an averaging period twice the length of the fundamental averaging period on which the raw data are based. In practice, averaging periods greater than twice the basic average are used. For instance, daily averages are computed from 1-h averages. The salient features of the averaging process will, however, be elucidated with a period twice the length of the basic period.

We assume that the first order density function of $X(t)$, $p_X(x;t)$, is log-normal,

$$p_X(x;t) = \frac{1}{\sqrt{2\pi}\, x \ln \sigma_{gx}} \exp\left\{-\frac{(\ln x - \ln \mu_{gx})^2}{2 \ln^2 \sigma_{gx}}\right\} \tag{4}$$

and that the second order density function $p_X(x, t - \tau; y, t)$ is the following joint log-normal density,

$$p_X(x, t - \tau; y, t) = [2\pi xy \ln^2 \sigma_{gx} \sqrt{1 - r^2}\,]^{-1} \cdot \exp\left\{\frac{(\ln x - \ln \mu_{gx})^2 - 2r(\ln x - \ln \mu_{gx})(\ln y - \ln \mu_{gx}) + (\ln y - \ln \mu_{gx})^2}{2(1 - r^2) \ln^2 \sigma_{gx}}\right\}, \tag{5}$$

where μ_{gx} and σ_{gx} are respectively the geometric mean and the standard geometric deviation of $X(t)$, and r is a correlation parameter.

The first order density function of $Z(t)$ can be determined from the relation

$$P_Z(z;t) = 2\int_0^{2z} P_X(x, t - \tau; 2z - x, t)\, dx. \tag{6}$$

The distribution function $F_Z(z;t)$ is related to $P_Z(z;t)$ by

$$F_Z(z;t) = \int_0^z P_Z(\eta;t)\, d\eta. \tag{7}$$

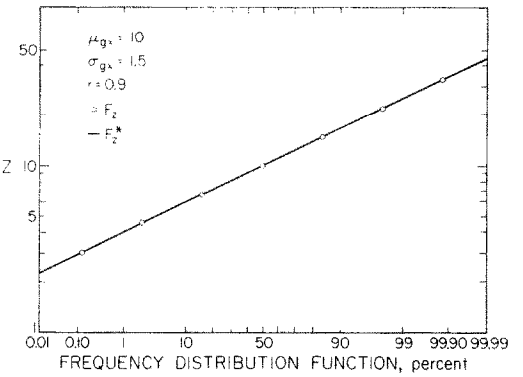

Fig. 5. Comparison of F_Z to F_Z^* for a typically low value of σ_{gx}.

Using (5–7), we obtain F_Z as

$$F_Z(z;t) = \frac{1}{2\sqrt{2\pi} \ln \sigma_{gx}} \int_0^{2z} x^{-1} \times \exp\left(\frac{-(\ln x - \ln \mu_{gx})^2}{2 \ln^2 \sigma_{gx}}\right) \cdot \left\{1 + \operatorname{erf}\left(\frac{\ln(2z - x) - r \ln x + (r - 1) \ln \mu_{gx}}{\sqrt{2(1 - r^2)} \ln^2 \sigma_{gx}}\right)\right\} dx. \tag{8}$$

In summary, F_Z is the distribution function of the random variable $Z(t)$, which is the average of two log-normally distributed random variables. We seek to determine how close the distribution of $Z(t)$ approximates a log-normal distribution. Therefore, we compare F_Z from (8) with F_Z^*, a log-normal distribution function with the same mean and variance as F_Z. Comparisons of F_Z and F_Z^* were made in 15 cases: $\sigma_{gx} = 1.18, 1.5, 2, 4, 10$ and $r = 0, 0.5, 0.9$. (The range of σ_{gx} for typical air pollution data is between 1 and 4.) Figures 5 and 6 show comparison of F_Z and F_Z^*. Qualitatively we note that F_Z compares closely with F_Z^*, particularly at large values of z. In the Appendix we discuss in more detail the comparison between F_Z and F_Z^*.

We can conclude from the results shown in Figs. 5 and 6 that the distribution of the average of two correlated log-normal variables approximates a log-

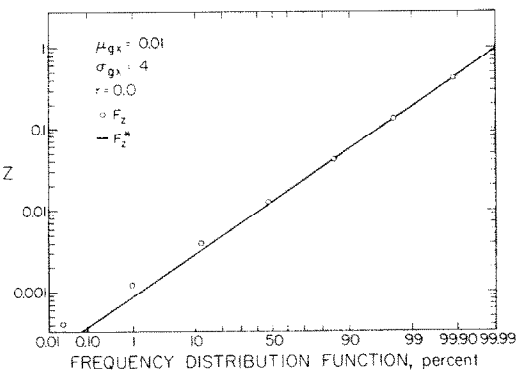

Fig. 6. Comparison of F_Z to F_Z^* for a typically high value of σ_{gx}.

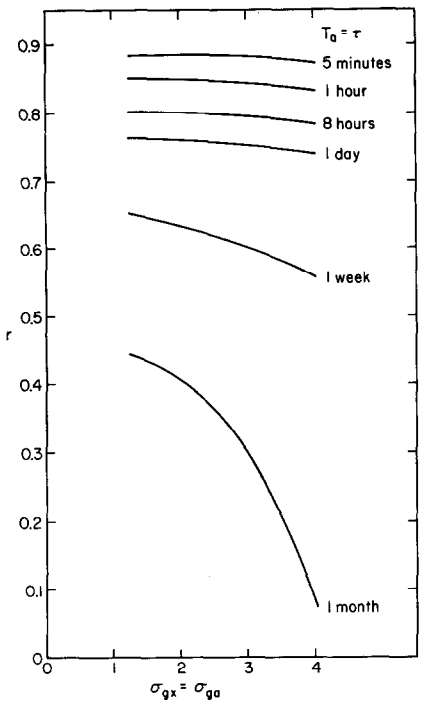

Fig. 7. Value of the correlation parameter r necessary for the standard geometric deviation of F_Z^* to equal σ_{gb} plotted as a function of the standard geometric deviation of the concentration distribution for averaging time T_a.

normal distribution. We now return to the observation (2). If we let $T_a = \tau$ and $T_b = 2\tau$, then we ask—Given $\sigma_{gx} = \sigma_{ga}$, can we, by appropriate choice of the correlation parameter r, have the standard geometric deviation of F_Z^* equal σ_{gb}. Figure 7 shows that for the values of σ_{gx} and τ of interest such an r exists. Although the correlation parameter r is not known for actual data, the results of this section indicate that the observation (2) can be explained simply as a consequence of the near log-normality of the short-averaging time concentrations.

Others have also investigated the effects of averaging time on concentration frequency distributions. There does not appear to be any other work which predicts both the form and parameters of time averaged concentration distributions, however, other relationships for measures of the variance in the distribution versus averaging time have been developed. Saltzman (1970) presented an empirical relationship for the standard geometric deviation, while Shoji and Tsukatani (1973) and Larsen and Peterson (1974) developed relationships based on assumed spectral properties of the concentration time series.

ANALYSIS OF FREQUENCY DISTRIBUTIONS

Up to this point we have shown that several distributions are capable of representing a number of air pollutant concentration frequency distributions. The log-normal distribution, while not, in fact, a perfect representation of the data is a convenient one because the parameters of the distribution can be easily determined from a log-probability plot of the data. Assuming log-normality of the basic data, we then showed that time averages formed from the original data would essentially preserve the log-normality of the data. We now come to the crucial question— why do the concentrations tend to be approximately log-normally distributed in the first place. This section is devoted to an attempt to propose possible explanations for this observed phenomenon.

Because of the approximate universality of pollutant frequency distributions we would expect the principal factors affecting urban concentration frequency distributions also to exhibit universality. That is, the fundamental characteristics of such a factor relative to its influence on urban pollutant concentration must be approximately the same for all pollutants in all cities. Conversely, a factor which is fundamentally different either from city to city or in its effect on different pollutants does not appear to have a major influence on frequency distributions.

Certainly there exists no characteristic similarity over all cities or all pollutants among either the spatial distribution or strength of sources. Similarly, wind direction distributions can be argued not to be a major factor in air pollutant frequency distributions. The concentration of a pollutant near the center of a uniform area source will not be sensitive to wind direction, whereas the concentration at a position near a single point source will be very sensitive to wind direction. Thus the impact of wind direction on concentration can range from negligible to significant and certainly varies from location to location. By such reasoning one is led to the conclusion that the two factors most likely to influence air pollutant frequency distributions are wind speed and mixing height. Increases in both factors will lead to a decrease in concentration.

A number of studies have been carried out investigating the correlation between wind speed, mixing depth, and air pollutant concentrations. Schmidt and Velds (1969) calculated a correlation coefficient between yearly average SO_2 concentrations in Rotterdam and wind speed of -0.97. Marsh and Withers (1969) found significant correlation between 6 h average SO_2 concentrations in Reading and wind speed but not between vertical turbulence and concentration. Similarly, in analyzing hourly ozone data in New York City, Bruntz *et al.* (1974) determined close correlation with wind speed, and were not able to improve the correlation by including mixing height. While certainly not proving that wind speed is the sole factor governing pollutant frequency distributions, these studies do indicate at least that wind speed/concentration correlations are as one might expect. Because of the difficulty in measuring mixing depths relative to wind speeds, fewer studies exist wherein mixing depth/concentration correlations were computed. Undoubtedly, mixing depth does play a role in determining pollutant frequency distributions.

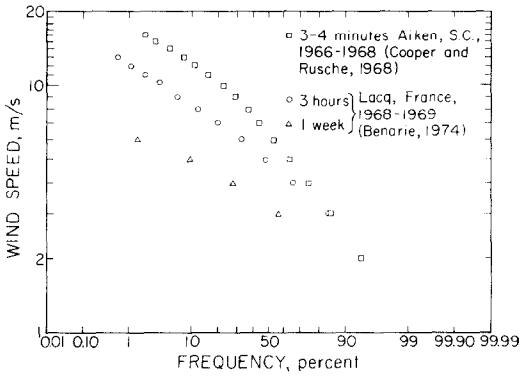

Fig. 8. Wind speed distributions in two cities. Data from Lacq shown for two averaging times.

From the above qualitative arguments we have arrived at the conclusion that wind speed and mixing depth should be the major factors influencing air pollutant frequency distributions. Following this supposition, we need to be more quantitative, that is to determine in what manner wind speed and mixing depth influence frequency distributions. The remainder of this section is devoted to an attempt to explain observed frequency distributions on a fundamental basis. In doing so, we restrict our attention to non-chemically reacting pollutants.

Influence of wind speed on instantaneous concentrations

In this subsection we wish to consider the influence of wind speed on the distribution of instantaneous pollutant concentrations. To begin, we need some notion as to observed frequency distributions of wind speed. Figure 8 shows frequency distributions of wind speed from Lacq, France and Aiken, SC. In both cases we note that the wind speed is approximately log-normally distributed. Why wind speeds seem to be log-normally distributed and even, in fact, if the log-normal is the best distribution to describe wind speeds are questions not central to our purpose here. We seek only to ascertain the consequences of this approximate log-normality in determining pollutant frequency distributions.

We begin in some sense with the most basic problem, that of determining the effect of wind speed variations on the instantaneous concentration of a pollutant released into that wind field. Clearly, attacking this problem on the basis of a three-dimensional urban flow is impossible. Therefore, we need to isolate the key elements of the problem in a much simpler hypothetical situation, which, nevertheless, retains the basic physics. Such a situation is embodied in a one-dimensional flow into which a pollutant is steadily emitted at a plane.

The fundamental equation describing the instantaneous concentration C of an inert atmospheric species is the continuity equation,

$$\frac{\partial C}{\partial t} + \nabla \cdot \mathbf{U}C = \mathscr{D}\nabla^2 C \tag{9}$$

where $\mathbf{U}$ is the instantaneous wind velocity vector, and $\mathscr{D}$ is the molecular diffusivity of the species in air. Molecular diffusion is normally neglected when (9) is applied to atmospheric species, the result being the so-called advection equation,

$$\frac{\partial C}{\partial t} + \nabla \cdot \mathbf{U}C = 0. \tag{10}$$

Because the wind speed $\mathbf{U}$ is a random variable, the instantaneous concentration C is a random variable.

As noted above, let us consider a one-dimensional flow (in the x-direction) into which a pollutant is steadily emitted at the $x = 0$ plane at a rate $S\ \mathrm{g\,cm^{-2}\,s^{-1}}$. The wind velocity in the x-direction is taken to be a function of time only, i.e. $U_x = U(t)$. This case, although highly simplified, exhibits the basic features of the situation in which there are three velocity components.

The instantaneous concentration is described by the one-dimensional form of (10),

$$\frac{\partial C(t,x)}{\partial t} + U(t)\frac{\partial C(t,x)}{\partial x} = 0 \tag{11}$$

subject to

$$C(0,x) = 0 \tag{12}$$

$$C(t,0) = \frac{S}{U(t)}, \quad t > 0. \tag{13}$$

The solution of (11-13) is

$$C(t,x) = \left.\frac{S}{U(t')}\right|_{x = \int_{t'}^{t} U(t'')\,dt''}. \tag{14}$$

Equation (14) relates the concentration at any position x and time t to the source strength S and the wind speed. We recall that $U(t)$ is a random variable, and therefore that $C(t,x)$ is a random variable. Given the pdf of $U(t)$, $p_U(u;t)$, we wish to determine the pdf of C, $p_C(c;t,x)$. It is advantageous to assume that there are only a finite number l of possible wind speeds, with the maximum wind speed denoted by u_l. In addition, we assume that a given wind speed persists for a time Δt. If Δx is the distance a fluid element moves in Δt corresponding to the slowest wind speed u_1, i.e. $u_1 = \Delta x/\Delta t$, and if the wind speeds obey the relations, $u_i = iu_1$, $i = 1,2,\ldots,l$, then the probability of observing concentration c_i at time $t_n = n\Delta t$ and position $x_m = m\Delta x$, $p_C(c_i; t_{n,x_m})$, is given by

$$x_m \geq x_1 \text{ and } t_1 \leq t_n < \infty$$

$$p_C(c_i; t_n, x_m) = \sum_{j=1}^{l} p_U(u_j; t_n) p_C(c_i; t_{n-1}, x_{m-j}) + \begin{cases} p_U(u_i; t_n) & \text{if } x_m \leq i\Delta x \\ 0 & \text{if } x_m > i\Delta x \end{cases} \quad i = 1,2,3,\ldots,l \tag{15a}$$

$$-\infty < x_m \leq 0 \text{ or } t_n \leq 0$$

$$p_C(c_i; t_n, x_m) = 0 \qquad i = 1,2,\ldots,l. \tag{15b}$$

The first term in (15a) represents the probability of a fluid element with concentration c_i being

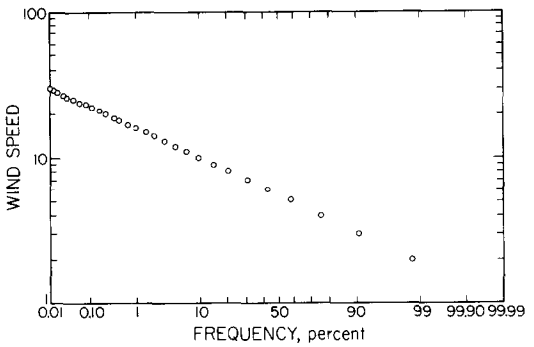

Fig. 9. Assumed wind speed distribution for one-dimensional advection problem.

advected to location x_m while the second term represents the probability of 'fresh' emissions reaching x_m. As l is increased, this formulation yields an approximation to the continuous density $P_C(c;t,x)$. The solution was evaluated for the case $l = 30$, $\Delta t = 1$ and $u_i = 30$. The wind speed was assumed to be log-normally distributed in accordance with observation. Its frequency distribution is shown in Fig. 9. The resulting frequency distribution for the concentration at $x = 30$, $t = 30$, $C(30,30)$, is shown in Fig. 10. The result is a close approximation to a log-normal distribution.

In summary, this situation of a steadily emitting source in a one-dimensional flow in which the wind speed is log-normally distributed leads to an instantaneous concentration that is approximately log-normally distributed. This example, while clearly highly idealized, does, however, demonstrate rigorously that log-normality of wind speed does lead to log-normality of instantaneous concentration. Of course, in an atmospheric flow other phenomena will influence concentration distributions, and so this example does not necessarily establish the cause of approximate log-normality in air quality data.

Influence of wind speed and mixing depth on mean concentrations

In the previous subsection we considered the effect of wind speed distribution on the instantaneous concentration of a species. Monitoring data reflect the instantaneous concentration at a point. This instantaneous concentration is a random quantity because of the turbulent nature of the atmosphere. When air pollutant concentrations are analyzed from a theoretical point of view, only the mean concentration $\langle c \rangle$ can be predicted, where $C = \langle c \rangle + c'$. (The mean concentration $\langle c \rangle$ is a function of location and time but is theoretically the result of an ensemble average.) Virtually all mathematical models of air pollutant behavior are concerned with the prediction of $\langle c \rangle$ (Lamb and Seinfeld, 1973).

There exist a large number of urban air pollution models depending on the dynamic nature (steady vs unsteady), type of source (point, line, area), number of spatial dimensions, meteorological assumptions, boundary conditions, etc. In these models the inputs are usually specified as known, or deterministic, quantities (such as wind speed and mixing depth). However, in attempting to assess the effect of the variability of these inputs on the predicted mean concentration, one can propose to allow the inputs to assume a *distribution* of values and determine the resulting *distribution* of values of the mean concentration. In particular, we are interested, of course, in the effect of the variability of wind speed and mixing depth on the distribution of mean concentrations.*

Box model. It is sometimes assumed that the mean concentration in a local region of an urban area can be represented by the simple box model relationship,

$$\langle c \rangle = \frac{k}{uh}, \tag{16}$$

where u is the mean wind speed, h is the mixing depth, and k is an empirical proportionality constant. If we allow u and/or h to be random variables, then the distribution of $\langle c \rangle$ can be computed directly from (16). Knox and Lange (1974) employed this approach by assuming a frequency distribution for u from hourly average readings and computing the frequency distribution of $\langle c \rangle$. In the cases presented, the approximate log-normality of the observation was reproduced.

We have investigated the effect of assuming different forms for the wind speed distribution on the concentration distribution using (16). In addition to the log-normal, the Weibull, gamma, uniform, and bounded distributions were investigated (Table 4). In all cases we assumed that the mean wind speed was the same and for the two-parameter distributions that the wind speed variance was also the same. The wind speed densities along with the resulting concentration densities are given in Table 4. In most cases the high concentration end of the distribution could be approximated by a straight line (although not the one

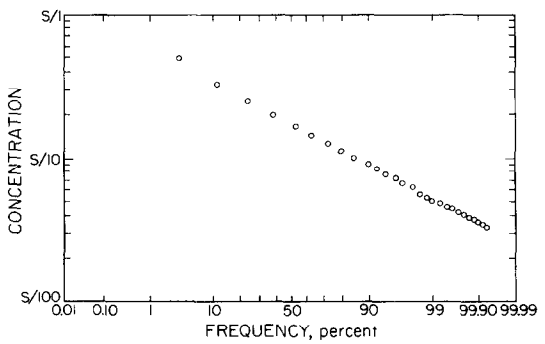

Fig. 10. Calculated concentration distribution corresponding to wind speed distribution in Fig. 9.

* It is reasonable to suppose that the distribution of *mean concentrations* over a time period of the order of several months to one year is not significantly different than the distribution of *instantaneous concentrations* over the same time period. Therefore, this approach should not be wholly inapplicable to the basic issue of analyzing ambient monitoring data.

Table 4. Relationship between probability density functions for wind speed and concentration when $c = k/u$

	Pdf for wind speed $p_u(u)$	Pdf for concentration $p_c(c)$
Log-normal	$(\sqrt{2\pi}\, u \ln\beta)^{-1} \exp\left\{-\frac{(\ln u - \ln\alpha)^2}{2\ln^2\beta}\right\}$	$(\sqrt{2\pi}\, c \ln\beta)^{-1} \exp\left\{-\frac{(\ln c - \ln k/\alpha)^2}{2\ln^2\beta}\right\}$
Weibull	$\alpha u^\beta \exp\left\{-\frac{\alpha u^{\beta+1}}{\beta+1}\right\}$	$\alpha k^{\beta+1} c^{-(\beta+2)} \exp\left\{-\frac{\alpha}{\beta+1}\left(\frac{k}{c}\right)^{\beta+1}\right\}$
Gamma	$\frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)} u^\alpha \exp\left(-\frac{u}{\beta}\right)$	$\frac{k^{\alpha+1} c^{-(\alpha+2)}}{\beta^{\alpha+1}\Gamma(\alpha+1)} \exp\left(-\frac{k}{\beta c}\right)$
Uniform	α^{-1}	$\frac{k}{\alpha c^2}$
Bounded	$-\alpha^{-1} \ln\left(\frac{\alpha - u}{\alpha}\right)$	$\frac{-k}{\alpha c^2} \ln\left(\frac{\alpha - k/c}{\alpha}\right)$

found by assuming the original log-normal distribution for wind speeds). It is reasonable to conclude that with the simple model of (16) a variety of wind speed distributions could actually exist and still the log-normal would be an adequate approximation to the concentration distribution. The explanation, of course, lies in the inverse relationship between concentration and wind speed, as pointed out previously by Benarie (1969, 1971, 1974) and Pollack (1975). Similarly, the inverse relationship between concentration and mixing depth indicates that the same correspondence between mixing depth and concentration is to be expected.

Gaussian plume model. For conditions of (1) a continuous point source located at $(0,0,z_0)$ and (2) wind speed constant and direction aligned with the x-axis, the ground-level mean concentration $\langle c(x,y,0)\rangle$ can be estimated by the familiar Gaussian plume equation,

$$\langle c(x,y,0)\rangle = \frac{S}{\pi\sigma_y\sigma_z u} \exp\left\{\frac{-y^2}{2\sigma_y^2}\right\} \exp\left\{\frac{-z_0^2}{2\sigma_z^2}\right\}. \tag{17}$$

Let us assume that the wind speed u is distributed according to a log-normal density with parameters μ_u and σ_u. As expected, we then obtain $P_{\langle c\rangle}(\langle c\rangle)$ from (17) as log-normal,

$$p_{\langle c\rangle}(\langle c\rangle) = \frac{1}{\sqrt{2\pi}\langle c\rangle \ln\sigma_c} \times \exp\left\{\frac{-(\ln\langle c\rangle - \ln\mu_c)^2}{2\ln^2\sigma_c}\right\}, \tag{18}$$

where $\sigma_c = \sigma_u$ and

$$\mu_c = \frac{S}{\pi\sigma_y\sigma_z\mu_u} \exp\left\{\frac{-y^2}{2\sigma_y^2}\right\} \exp\left\{\frac{-z_0^2}{2\sigma_z^2}\right\}. \tag{19}$$

Eddy diffusion model. Finally, we consider two-dimensional, steady-state diffusion as described by the atmospheric diffusion equation,

$$u(z)\frac{\partial\langle c\rangle}{\partial x} = \frac{\partial}{\partial z}\left(K(z)\frac{\partial\langle c\rangle}{\partial z}\right), \tag{20}$$

where $u(z) = u_1 z^\alpha$ and $K(z) = K_1 z^\beta$. The solutions of (20) for ground-level crosswind line and area sources are, (Monin and Yaglom, 1971; Lebedeff and Hameed, 1975)

$$\langle c(x,0)\rangle = \frac{Slp}{u_1\Gamma(q)}\left(\frac{u_1}{K_1 x p^2}\right)^q \tag{21}$$

and

$$\langle c(x,0)\rangle = \frac{S_a}{K_1(1-\beta)\Gamma(q)}\left(\frac{u_1}{K_1 x p^2}\right)^{\beta-1/p}, \tag{22}$$

where $p = \alpha - \beta + 2 > 0$, $q = (\alpha+1)/(\alpha-\beta+2)$, and $0 \le p < 1$. Equations (21 and 22) can be written in the form $\langle c(x,0)\rangle = A u_1^\gamma$. If we now assume u_1 to be log-normally distributed, we find $p_{\langle c\rangle}(\langle c\rangle)$ to be given by (18) with $\sigma_c = \sigma_{u_1}^{|\gamma|}$ and $\mu_c = A\mu_{u_1}^\gamma$. Thus, the eddy diffusion model, while more detailed than the Gaussian model, still predicts a log-normal concentration distribution provided that the wind speed is log-normally distributed.

Summary

In this section we have attempted to shed some light on the fundamental question of why urban air pollutant concentration frequency distributions tend to be log-normal. Starting from information available from statistical correlations between wind speed and mixing depth (primarily wind speed), we investigated

how both instantaneous and mean concentrations might be influenced by wind speed and mixing depth (primarily wind speed) variability. In all cases we found that if wind speeds are nearly log-normally distributed then resulting concentrations will be nearly log-normally distributed. In fact, other distributions, namely the gamma and Weibull, are capable of producing nearly log-normal concentrations. This result does not, of course, establish that wind speeds are the primary influence on concentration distributions in the atmosphere, since other effects are most certainly influential. Nevertheless, the results here are convincing of the role of wind speed in pollutant frequency distributions.

It is interesting to note that in the three simple models considered for the mean concentration, the standard geometric deviation is independent of the source strength and the geometric mean varies linearly with it. Thus, if one were using (17, 21, or 22) to predict mean concentrations for source emission changes, only the intercept of the concentration distributions plotted on log-probability paper would change, not the slopes.

CONCLUSIONS

Air pollutant concentration frequency distributions are the result of complex phenomena. The direct prediction of these distributions does not appear to be possible. Observed data are generally represented as log-normal, although other common statistical distributions are capable of representing the data as well as or better than the log-normal. The log-normal is convenient because the mean and variance can be easily determined from a log-probability plot of the data. The fundamental question of why concentration distributions tend to be approximately log-normal cannot be answered unequivocally. The persistence of log-normality for all averaging times can be explained if the raw data are themselves log-normally distributed. The log-normality of the raw data, i.e. the instantaneous concentrations, can be shown to result if wind speed is log-normally distributed. Conventional models for mean concentrations, such as the Gaussian plume and eddy diffusion, can also lead to log-normality for concentration distributions if the wind speeds are log-normal. It is shown how these models may be used to estimate the shift in the distribution resulting from source emission level changes.

Acknowledgement—This work was supported by National Science Foundation Grant ENG71-02486.

REFERENCES

Benarie M. (1969) Calculation of the amount and the nuisance of pollutants emitted from a point source. *Atmospheric Environment* **3,** 467–473.

Benarie M. (1971) About the validity of the log-normal distribution of pollutant concentrations. Paper SU-18D, *Proceedings of the 2nd International Clean Air Congress* (Edited by H. M. Englund and W. T. Berry). Academic Press, New York.

Benarie M. (1974) The use of the relationship between wind velocity and ambient pollutant concentration distributions for the estimation of average concentrations from gross meteorological data. Paper 5, *Proceedings of the Symposium on Statistical Aspects of Air Quality Data.* Environmental Protection Agency Pub. No. 650/4-74-038.

Bruntz S. M., Cleveland W. S., Kleiner G. and Warner J. L. (1974) The dependence of ambient ozone on solar radiation, wind, temperature, and mixing height. *Symposium on Atmospheric Diffusion and Air Pollution,* Santa Barbara, CA. American Meteorological Society, Boston, MA.

Cooper R. E. and Rusche B. C. (1968) The SRL meteorological program and off-site close calculations. E. I. DuPont de Nemours and Co., Savannah River Laboratory, Aiken, SC., Rept. DP-1163.

Knox J. B. and Lange R. (1974) Surface air pollutant concentration frequency distributions: Implications for urban modeling. *J. Air Pollut. Control Ass.* **24,** 48–53.

Lamb R. G. and Seinfeld J. H. (1973) Mathematical modeling of urban air pollution: General theory. *Environ. Sci. Technol* **7,** 253–261.

Larsen R. I. (1971) A mathematical model for relating air quality measurements to air quality standards. Environmental Protection Agency Pub. No. AP-89.

Larsen S. E. and Petersen E. L. (1974) Statistical description of air pollution concentration, averaging time and frequency. *Symposium on Atmospheric Diffusion and Air Pollution,* Santa Barbara, CA. American Meteorological Society, Boston, MA.

Lebedeff S. A. and Hameed S. (1975) Steady state solution of the semi-empirical diffusion equation for area sources. (Preprint).

Lynn D. A. (1974) Fitting curves to urban suspended particulate data. Paper 13, *Proceedings of the Symposium on Statistical Aspects of Air Quality Data.* Environmental Protection Agency Pub. No. 650/4-74-038.

Mage D. T. and Ott W. R. (1975) An improved statistical model for analyzing air pollution concentration data. Paper No. 75-51.4, *68th Meeting of the Air Pollution Control Association,* Boston, MA.

Marsh K. J. and Withers V. R. (1969) An experimental study of the dispersion of the emissions from chimneys in Reading—III. The investigation of dispersion calculations. *Atmospheric Environment* **3,** 281–302.

Mitchell R. L. (1968) Permanence of the log-normal distribution. *J. Opt. Soc. Am.* **58,** 1267–1272.

Monin A. S. and Yaglom A. M. (1971) *Statistical Fluid Mechanics.* MIT Press, Cambridge, MA.

Pollack R. I. (1975) *Studies of pollutant concentration frequency distributions.* Environmental Protection Agency Pub. No. 650/4-75-004.

Saltzman B. E. (1970) Significance of sampling time in air monitoring. *J. Air Pollut. Control Ass.* **20,** 660–665.

Schmidt F. H. and Velds C. A. (1969) On the relation between changing meteorological circumstances and the decrease of sulphur dioxide concentration around Rotterdam. *Atmospheric Environment* **3,** 455–460.

Shoji H. and Tsukatani T. (1973) Statistical model of air pollutant concentration and its application to the air quality standards. *Atmospheric Environment* **7,** 485–501.

APPENDIX. THE DISTRIBUTION OF THE AVERAGE OF TWO CORRELATED LOG-NORMAL VARIATES

Equations (7 and 8) give the distribution function, F_Z, of the average of two correlated log-normal variates. Figures 5 and 6 show typical comparisons of F_Z to F_Z^*, the log-normal distribution function with the same mean

Table A.1. Comparison of the actual difference $F_Z(z)-F_Z^*(z)$ with the first correction term $E(z)$ for the distribution of the average of two log-normal variates

σ_{gx}	μ_{gx}	r	z	$E(z)$	$F_Z(z)-F_Z^*(z)$
1.5	10	0.0			
			4.32	-0.413×10^{-4}	-0.104×10^{-3}
			7.76	-0.816×10^{-3}	-0.950×10^{-3}
			10.4	0.231×10^{-3}	0.791×10^{-3}
			13.9	0.829×10^{-3}	0.696×10^{-3}
			25.0	-0.129×10^{-3}	0.107×10^{-3}
		0.9	3.07	-0.930×10^{-8}	-0.749×10^{-7}
			6.76	-0.300×10^{-6}	-0.549×10^{-6}
			10.0	-0.778×10^{-7}	0.330×10^{-6}
			14.9	0.255×10^{-6}	-0.183×10^{-4}
			32.9	-0.345×10^{-7}	-0.195×10^{-6}
4.0	0.01	0.0	0.397×10^{-3}	-0.222×10^{-6}	-0.111×10^{-2}
			0.411×10^{-2}	-0.232×10^{-4}	-0.416×10^{-1}
			0.132×10^{-1}	-0.573×10^{-4}	-0.249×10^{-1}
			0.425×10^{-1}	-0.501×10^{-4}	0.519×10^{-2}
			0.440	0.519×10^{-5}	-0.494×10^{-5}
		0.9	0.181×10^{-3}	-0.246×10^{-10}	-0.484×10^{-5}
			0.271×10^{-2}	-0.271×10^{-8}	-0.795×10^{-3}
			0.105×10^{-1}	-0.727×10^{-8}	-0.557×10^{-4}
			0.405×10^{-1}	-0.771×10^{-8}	0.378×10^{-5}
			0.606	0.218×10^{-9}	0.330×10^{-7}

and variance as F_Z. Qualitative comparison such as these indicate that F_Z is itself nearly log-normal. Following Mitchell's (1968) analysis of the sum of n independent log-normal variates, we investigate analytically the near log-normality of F_Z in this Appendix.

The pdf, p_Z, of the average of two correlated log-normal variates can be expressed as an orthogonal polynomial expansion in terms of p_Z^*, the pdf of the log-normal density with the same mean and variance as p_Z, i.e.

$$p_Z(z) = b_0 u_0(z) p_Z^*(z) + b_1 u_1(z) p_Z^*(z) + \ldots, \qquad (A.1)$$

where the u_n are the orthogonal polynomials of order n defined by

$$\int u_n(\eta) u_m(\eta) p_Z^*(\eta)\, d\eta = \begin{cases} 1 & m = n \\ 0 & m \neq n \end{cases} \qquad (A.2)$$

and the b_n are constants

$$b_n = \int u_n(\eta) p_Z(\eta)\, d\eta. \qquad (A.3)$$

Because p_Z and p_Z^* have identical means and variances (A.1) becomes

$$p_Z(z) = p_Z^*(z) + b_3 u_3(z) p_Z^*(z) + b_4 u_4(z) p_Z^*(z) + \ldots \qquad (A.4)$$

In this form we see that p_Z is the long-normal density, p_Z^*, plus a "correction" consisting of an infinite series of polynomials multiplying p_Z^*. If we truncate (A.4) after the first correction term and (1) if the truncated expansion is a good approximation to p_Z and (2) if the correction term is small, then we see analytically that p_Z is in fact nearly log-normal. Similarly we can integrate the truncated expansion and form the same conclusions about the distribution function F_Z. The expansion is then

$$F_Z(z) \approx F_Z^*(z) + E(z), \qquad (A.5)$$

where

$$E(z) = \int_0^z b_3 u_3(\eta) p_Z^*(\eta)\, d\eta. \qquad (A.6)$$

The resulting integrated expression for $E(z)$ is

$$E(z) = \left\{\frac{\mu_{3z} - \mu_{3z}^*}{(\alpha_{1z})^3}\right\} \left\{\frac{\begin{array}{r} -\rho^6 N_0(z) + \rho^2(\rho^4 + \rho^2 + 1) N_1(z) \\ -(\rho^4 + \rho^2 + 1) N_2(z) + N_3(z) \end{array}}{\rho^{12}(\rho^2 - 1)^3(\rho^2 + 1)(\rho^4 + \rho^2 + 1)}\right\} \qquad (A.7)$$

where

$\mu_{3z} \equiv$ third central moment p_Z

$\mu_{3z}^* \equiv$ third central moment of p_Z^*

$\alpha_{1z} \equiv$ first moment about the origin of p_Z

$$\rho^2 \equiv \frac{\text{mean of } p_Z^*}{\text{median of } p_Z^*} = \exp\{\ln^2 \sigma_{gz}^*\} \qquad (A.8)$$

$$N_k(z) = \frac{1}{2}\left\{\text{erf}\left[\frac{1}{\sqrt{2}}\left(\frac{\ln z - \ln \mu_{gz}^*}{\ln \sigma_{gz}^*} - k \ln \sigma_{gz}^*\right)\right] + 1\right\},$$

$\mu_{gz}^* \equiv$ geometric mean of p_Z^*

$\sigma_{gz}^* \equiv$ standard geometric deviation of p_Z^*

or in terms of the parameters of the initial second order density function (equations 3 and 4)

$$\frac{\mu_{3z} - \mu_{3z}^*}{(\alpha_{1z})^3} = \tfrac{1}{8}[\exp(\ln^2 \sigma_{gx}) - \exp(r \ln^2 \sigma_{gx})]^3$$

$$\mu_{gz}^* = \mu_{gx}\left(\frac{\exp(\ln^2 \sigma_{gx})}{\frac{1}{2}\exp(\ln^2 \sigma_{gx}) + \frac{1}{2}\exp(r \ln^2 \sigma_{gx})}\right)^{1/2} \qquad (A.9)$$

$$\sigma_{gz}^* = \exp\{\ln[1/2 \exp(\ln^2 \sigma_{gx}) + 1/2 \exp(r \ln^2 \sigma_{gx})]\}^{1/2}.$$

For the 15 cases mentioned earlier, E was evlauated and compared to the actual difference between F_Z and F_Z^*. The results of four of these cases are given in Table A.1. These results are typical in that they show that (1) for low values of σ_{gx} the difference between F_Z and F_Z^* is small and E is a good estimate of this difference and (2) for relatively high values of σ_{gx} the difference between F_Z and F_Z^* is still small but E underestimates this difference.