Appendix A

Statistical Techniques Employed in Atmospheric Sampling

A.1 Introduction

Proper use of statistics and statistical techniques is necessary for assessing the quality of ambient air sampling data. For a comprehensive discussion of the subject of data quality assessment (DQA), review EPA's technical assistance document, Guidance for Data Quality Assessment, Practical Methods of Data Analysis, EPA QA/G-9 (EPA/600/R-96/084), January 1998. This reference document provides practical demonstrations on how to use the data quality assessment (DQA) technique in evaluating environmental data sets and shows how to apply some graphic and statistical tools for performing DQA.

This chapter is intended as an introduction to statistics and statistical concepts and their use in analyzing ambient air sampling data. Topics addressed include: (a) Data Quality Objectives (DQO), (b) Data Plotting, (c) Measures of Central Tendency, (d) Measures of Dispersion, and (e) Distribution Curves. Although these topics are not simple, they can be understood and used by non-statisticians. If a detailed statistical analysis of data is required, it is recommended that an experienced statistician be consulted.

Students who could benefit from a review of basic mathematics in ambient monitoring are encouraged to access the EPA Air Pollution Training Institute course, SI 100: Mathematics Review for Air Pollution Control. This self-instruction course can be found at:

http://yosemite.epa.gov/oaqps/EOGtrain.nsf/DisplayView/SI_100_0-5?OpenDocument

In addition, the University of Illinois-Chicago, School of Public Health-Environmental and Occupational Health Division, has developed an Internet-based program entitled "Introduction to Environmental Statistics." This program is presented as a video series in three modules on topics which include interpreting monitoring data, sampling and analytical limitations and sample detection limits, and quality assurance and quality control. This program can be found at: http://www.uic.edu/sph/eohs_webcasts.htm.

It is important to note that the statistical calculations discussed in this Appendix are best and more easily performed by employing one of many commercially available computer-based statistical software packages.

A.2 The Data Quality Objectives (DQO) Process

While the Data Quality Objectives (DQO) Process is not a statistical technique *per se*, it is important because it helps to establish criteria for data quality and the development of data collection designs. DQOs provide the appropriate context for understanding the purpose of the ambient air sampling and analysis data collection effort. Also, they establish the qualitative and quantitative criteria for assessing the quality of the collected data set, based on the predefined intended use of data. Specific information on the Data Quality Objectives Process can be found in EPA document, "Guidance on Systematic Planning Using the Data Quality Objective Process" (EPA QA/G-4), at: http://www.epa.gov/quality/qs-docs/g4-final.pdf.

DQOs are qualitative and quantitative statements derived from the outputs of the first six steps of the DQO Process that encompass the following:

- Clarify the study objective.
- Define the most appropriate type of data to collect.
- Determine the most appropriate conditions from which to collect the data.
- Specify tolerable limits on decision errors which will be used as the basis for establishing the quantity and quality of data needed to support the decision.

The DQOs are then used to develop a scientific and resource-effective data collection design.

The Seven Steps of the DQO Process

Step 1: *State the Problem*. Concisely describe the problem to be studied. Review prior studies and existing information to gain a sufficient understanding to define the problem.

Step 2: *Identify the Goal of the Study*. Identify what questions the study will attempt to answer.

Step 3: *Identify Information Inputs*. Identify the information that needs to be obtained and the measurements that need to be taken to resolve the decision statement.

Step 4: *Define Boundaries of the Study*. Specify the time periods and spatial area to which decisions will apply. Determine when and where data should be collected.

Step 5: *Develop the Analytical Approach*. Define the statistical parameters of interest, specify the action level, and integrate the previous DQO outputs into a single statement that describes the logical basis for choosing among alternative actions.

Step 6: *Specify the Performance or Acceptance Criteria*. Define the decision maker's tolerable decision error rates based on a consideration of the consequences of making an incorrect decision.

Step 7: *Develop the Plan for Obtaining Data*. Evaluate information from the previous steps and generate alternative data collection designs. Choose the most resource-effective design that meets all DQOs.

Outputs of the DQO Process

The DQO Process leads to the development of a quantitative and qualitative framework for a study. Each step of the Process derives valuable criteria that will be used to establish the final data collection design. The first five steps of the DQO Process identify mostly qualitative criteria, such as what problem has initiated the study and what decision it attempts to resolve. These steps also define the type of data that will be collected, where and when the data will be collected, and a decision rule that specifies how the decision will be made.

The sixth step defines quantitative criteria expressed as limits on decision errors that the decision maker can tolerate.

The final step is used to develop a data collection design based on the criteria developed in the first six steps. The final product of the DQO Process is a data collection design that meets the quantitative and qualitative needs of the study.

A.3 Data Collection Design

A data collection design specifies the final configuration of the environmental monitoring or measurement effort required to satisfy the DQOs. It designates:

- the types and quantities of samples or monitoring information to be collected;
- where, when, and under what conditions they should be collected;
- what variables are to be measured; and
- QA/QC procedures to ensure that sampling design and measurement errors are controlled sufficiently to meet the tolerable decision error rates specified in the DQOs.

Data Plotting

Data is usually uninterpretable in the form in which it is collected. In this section, we shall consider the graphical techniques of summarizing such data so that the meaningful information can be extracted from it. There are two kinds of variables to which we assign data: continuous variables and discrete variables.

A continuous variable is one that can assume any value in some interval of values. Examples of continuous variables are weight, volume, length, time, and temperature. Most air pollution data are taken from continuous variables. Discrete variables, on the other hand, are those variables whose possible values are integers. Therefore, they involve counting rather than measuring. Examples of discrete variables are the number of sample stations, number of people in a room, and number of times a control standard is violated. Since any measuring

device is of limited accuracy, measurements in real life are actually discrete in nature rather than continuous, but this should not keep us from regarding such variables as continuous. When a weight is recorded as 165 pounds, it is assumed that the actual weight is somewhere between 164.5 and 165.5 pounds.

A.4 Graphical Analysis

Frequency Tables

Let us consider the set of data in Table A-1, which represents SO_2 levels for a given hour for 25 days. The first step in summarizing the data is to form a frequency table. A frequency table is a table prepared by dividing a data set into selected units or class intervals, then counting and inserting the number of points (frequency of occurrences) within the units or class intervals. Table A-2 is a frequency table prepared from the SO_2 data set given in Table A-1.

In constructing the frequency table, we have divided the 25 points in the data set into 11 class intervals with each interval being 15 units in length. The choice of dividing the data into 11 intervals was purely arbitrary. However, in dealing with data it is best to choose the length of the class interval such that 8 to 15 intervals will include all of the data under consideration. Deriving the frequency of occurrence column involves nothing more than counting the number of values in each interval. The relative frequency column is obtained by dividing the number of points or frequency of occurrences within a unit by the total number of events within the data set, which in this example is 25.

From observation of the frequency table, we can now see the data taking form. The values appear to be clustered between 25 and 85 ppb. In fact, nearly 80% are in this interval.

Table A-1. SO_2 levels.

Days	SO_2 Concentration (ppb)*
1	53
2	72
3	59
4	45
5	44
6	85
7	77
8	56
9	157
10	83
11	120
12	81
13	35
14	63
15	48
16	180
17	94
18	110
19	51
20	47
21	55
22	43
23	28
24	38
25	26

*ppb = parts per billion collected SO_2 levels.

Table A-2. Frequency table.

Class Interval (ppb)	Frequency of Occurrence (total 25)	Relative Frequency
25 - 40	4	4/25 = 0.16
40 - 55	7	7/25 = 0.28
55-70	4	4/25=0.16
70-85	4	4/25=0.16
85 - 100	2	2/25 = 0.08
100-115	1	1/25=0.04
115-130	1	1/25=0.04
130 - 145	0	0.00
145 - 160	1	1/25 = 0.04
160 - 175	0	0.00
175 - 190	1	0.04

The Frequency Polygon

The next step is to graph the information in the frequency table. One way of doing this would be to plot the frequency for the midpoint of each class interval. The solid line connecting the points of Figure A-1 forms a frequency polygon.

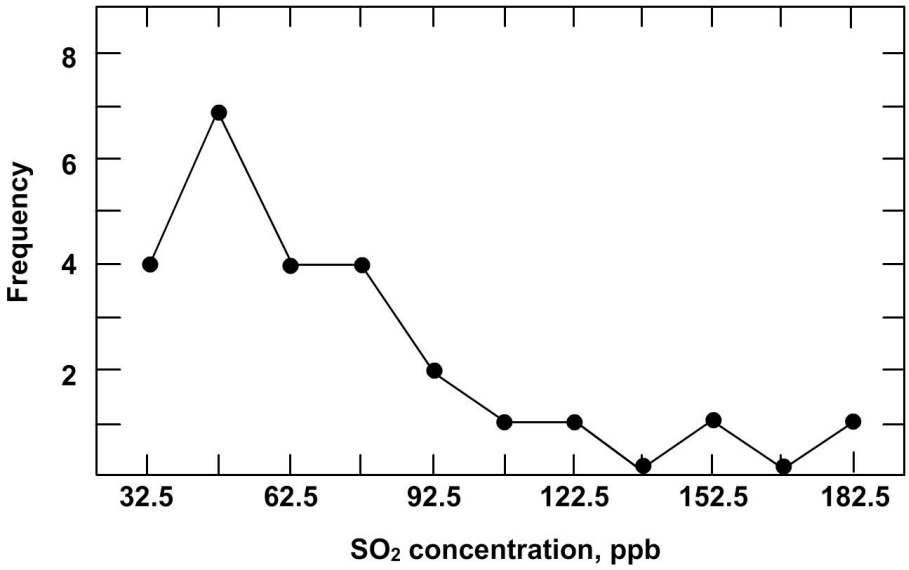

Figure A-1. Pollution concentration (midpoint of class interval) frequency polygon.

The Histogram

Another method of graphing the information would be by constructing a histogram as shown in Figure A-2. The histogram is a two-dimensional graph in which the length of the class interval is taken into consideration. The histogram can be a very useful tool in statistics, especially if we convert the given frequency scale to a relative scale so that the sum of all the ordinates equals one. This is shown in Figure A-3. Thus, each ordinate value is derived by dividing the original value by the number of observations in the sample, in this case, 25.

The advantage in constructing a histogram like this one is that we can read probabilities from it, if we can assume a scale on the abscissa such that a given value will fall in any one interval in the area under the curve in that interval. For example, the probability that a value will fall between 55 and 70 is equal to its associated interval's portion of the total area of intervals, which is 0.16.

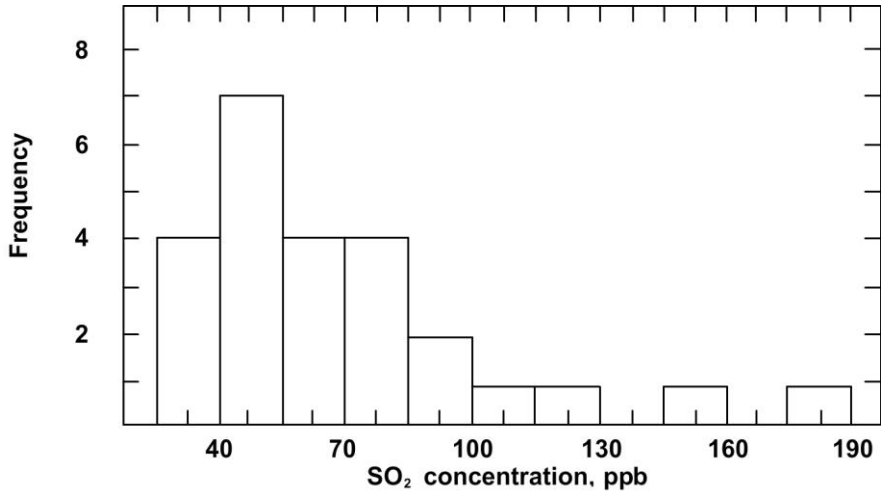

Figure A-2. Pollutant concentration histogram of frequency distribution curve.

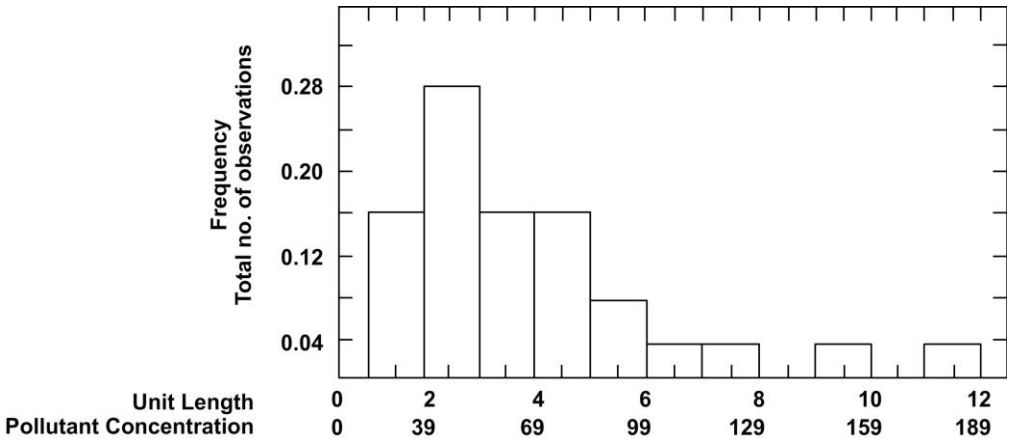

Figure A-3. Histogram of percent frequency distribution curve.

The Cumulative Frequency Distribution

Using the frequency table and histogram discussed above, we can construct a cumulative frequency table and curve as shown in Table A-3 and Figure A-4.

Table A-3. Cumulative frequency table.

SO_2 level		Cumulative frequency	Relative cumulative frequency
Under	40	4	0.16
”	55	11	0.44
”	70	15	0.60
”	85	19	0.76
”	100	21	0.84
”	115	22	0.88
”	130	23	0.92
”	145	23	0.92
”	160	24	0.96
”	175	24	0.96
”	190	25	1.00

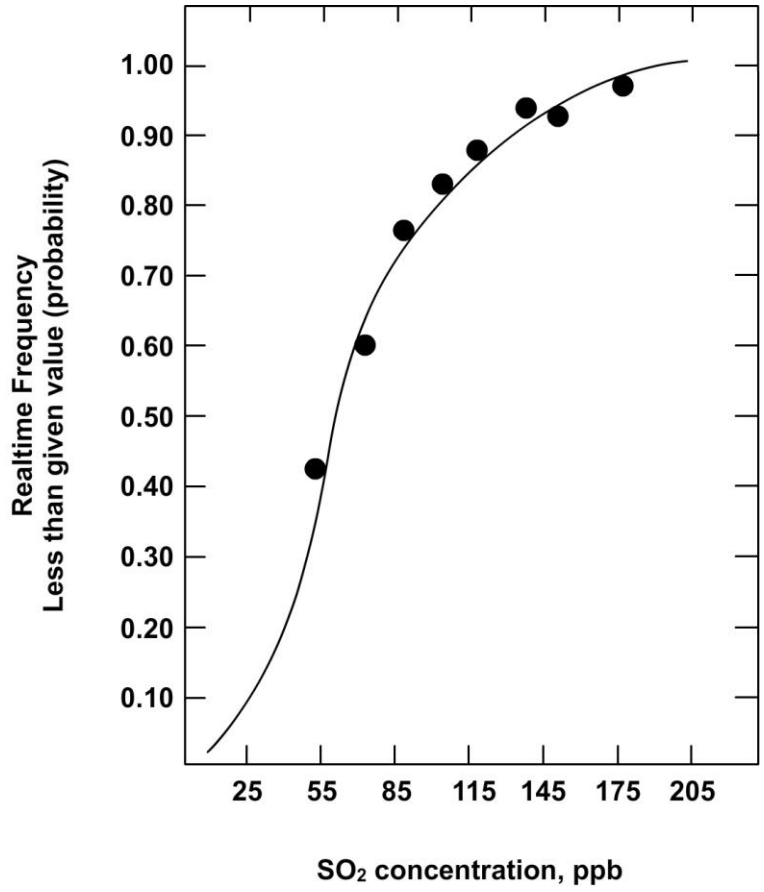

Figure A-4. Cumulative frequency distribution curve.

The cumulative frequency table gives the number of observations less than a given value. Probabilities can be read from the cumulative frequency curve or cumulative frequency table. For example, to find the probability that a value will be less than 85, we read up to the curve at the point $x = 85$ and across to the value 0.76 on the y-axis. An alternative way to use the table is to go to the row where the SO_2 level shows under 85, then go across to the relative cumulative frequency value of 0.76.

Distribution of Data

When we draw a histogram for a set of data, we are representing the distribution of the data. Different sets of data will vary in relation to one another and, consequently, their histograms will look different. In this chapter, we identify three characteristics that will distinguish the distributions of different sets of data. These are central location, dispersion, and skewness. These are characterized in Figure A-5. Curves A and B have the same central location, but B is more dispersed. However, both A and B are symmetrical and are, therefore, said not to be skewed. Curve C is skewed to the right and has a different central location than A and B. Mathematical measures of central location and dispersion will be discussed later.

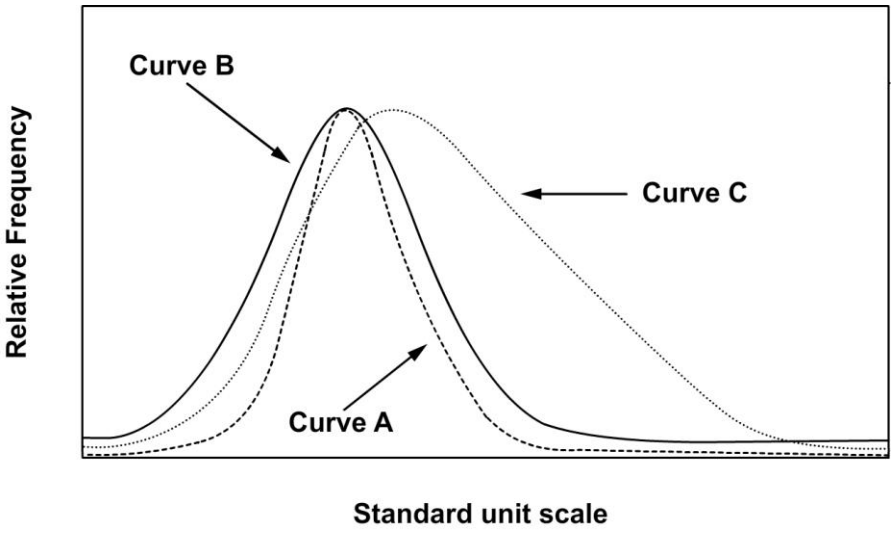

Figure A-5. Relative frequency distribution showing: Curve A and B both centrally located, Curve B being more dispersed than Curve A, and the skewness of Curve C.

Transformation of Data

In most statistical work, data that closely approximate a particular symmetrical curve, called the normal curve, are required. Both curves A and B in Figure A-5 are examples of normal curves. In dealing with skewed curves, such as C in the same figure, it is desirable to transform the data in some way so that a symmetrical curve resembling the normal curve is obtained. Referring to the frequency table (Table A-2) and histogram (Figure A-2) of the data used earlier, it

can be seen that for this set of data, the distribution is skewed (in the opposite direction as Curve C above), hence the data are not normally distributed.

The Logarithmic Transformation

One of the most successful ways of deriving a symmetrical distribution from a skewed distribution is by expressing the original data in terms of logarithms. The logarithms of the original data are given in Table A-4.

Arbitrarily dividing the logarithmic data into nine class intervals, each of 0.1 unit in length, we can prepare the logarithmic frequency table in Table A-5. As can be seen in Figure A-6, a frequency plot of the log transformed data more closely approximates a symmetrical curve than the arithmetic plot of the original data.

Table A-4. Logarithmic transformation.

Day	Pollutant conc. X	$Log_{10}X$
1	53	1.724
2	72	1.857
3	59	1.771
4	45	1.653
5	44	1.644
6	85	1.929
7	77	1.887
8	56	1.748
9	157	2.196
10	83	1.919
11	120	2.079
12	81	1.909
13	35	1.544
14	63	1.799
15	48	1.681
16	180	2.255
17	94	1.973
18	110	2.041
19	51	1.708
20	47	1.672
21	55	1.740
22	43	1.634
23	28	1.447
24	38	1.580
25	26	1.415

Table A-5. Logarithmic frequency table.

Class interval	Frequency of occurrence	Cumulative frequency	Relative cumulative frequency
1.4 - 1.5	2	2	0.08
1.5 - 1.6	2	4	0.16
1.6 - 1.7	5	9	0.36
1.7 - 1.8	6	15	0.60
1.8 - 1.9	2	17	0.68
1.9 - 2.0	4	21	0.84
2.0 - 2.1	2	23	0.92
2.1 - 2.2	1	24	0.96
2.2 - 2.3	1	25	1.00

Probability Graph Paper

Probability graph paper is used in the analysis of cumulative frequency curves; for example, the graph paper can be used as a rough test of whether the arithmetic or the logarithmic scale best approximates a normal distribution. The scale, arithmetic or logarithmic, on which the cumulative frequency distribution of the data is more nearly a straight line, is the one providing the better approximation to a normal distribution. Plotting the cumulative distribution curve of the data above on the two scales shows that the logarithmic scale yields the better fit (Figure A-6).

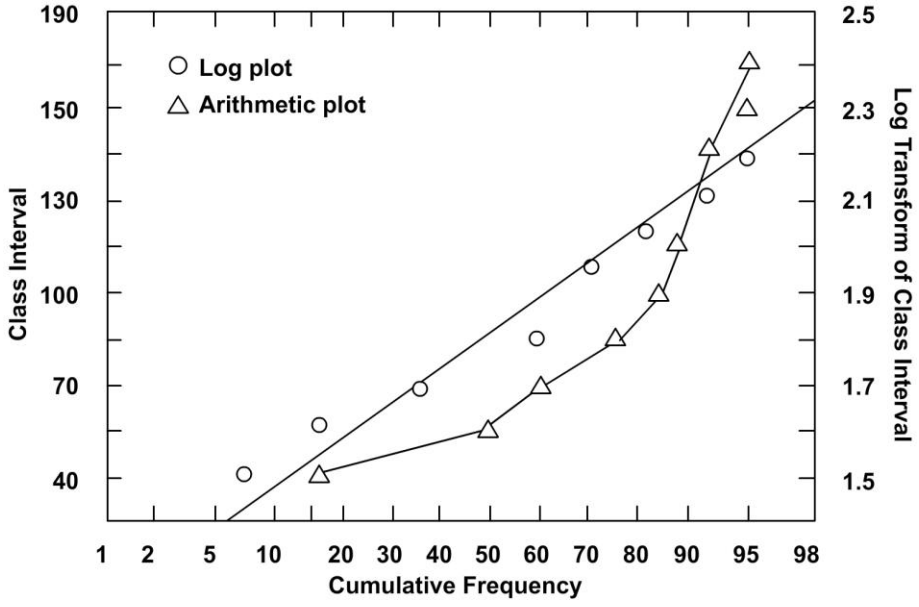

Figure A-6. Normalized data plot vs. non-transformed data.

These probability plots can be used, if the data are normally distributed, to estimate the mean and standard deviation of the data. The estimate of the mean, as will be shown later, is the 50th percentile point, and the estimation of the standard deviation is the distance from the 50th percentile to the 16th percentile. A percentile is a measure of the relative position of one of several observations in relation to all of the observations, and provides a measure of relative standing that is useful for summarizing data.

Least-Square Linear Regression

If the linear relationship between two variables is significant, a least-square linear regression line, or line of "best fit," may be drawn to represent the data. This relationship can then be used to determine the value of an unknown variable. For example, if the ambient air concentration is unknown, but linearly related to the response of an ambient air monitor, we can estimate the ambient air concentration based on an observed response from the air monitor. Algebraically, a straight line has the following form:

(Eq. A-1) $$y = mx + b$$

Where:
- y = *dependent variable plotted on the ordinate (y-axis)*
- x = *explanatory variable (independent variable) plotted on the abscissa (x-axis)*
- b = *the point at which the line intercepts the y-axis at* $x = 0$
- m = *slope, which shows how much of a change of 1 unit of* x *affects* y

Linear regression minimizes the vertical distance between all data points and the straight line (Figure A-7).

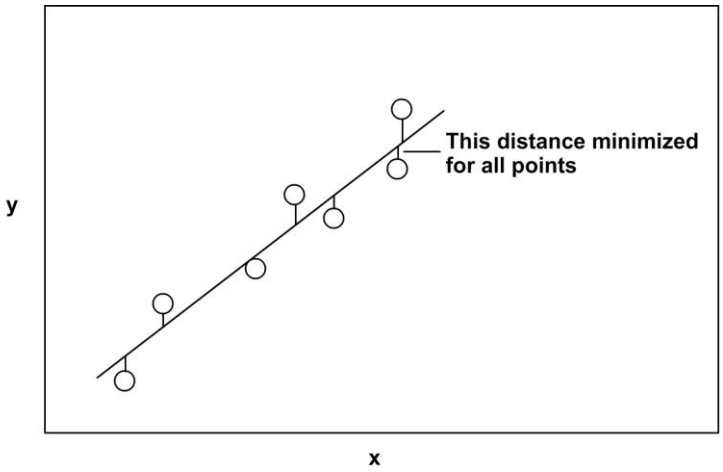

Figure A-7. Linear regression curve.

The constants m and b for the "least-square" line can be determined using the following equations:

(Eq. A-2)
$$m = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

(Eq. A-3)
$$b = \bar{y} - m\bar{x}$$

Where: n = *number of observations*
$\bar{y} = \sum y/n ; \bar{x} = \sum x/n$

Example Problem
Calibration of an ambient air analyzer is required before it can be used to provide reliable ambient air concentration measurements. A typical calibration consists of the introduction of known and certified standard concentrations, typically in parts per million (ppm) over the linear operational range of the instrument, and the recording of the corresponding response of the instrument in units such as volts. Based on the recorded responses and the known concentrations, a least-square linear relationship between the variables can be calculated and subsequently used to determine ambient concentrations based on the response of the analyzer. The following data were collected during a calibration of a chemiluminescent NO_X analyzer.

x = Concentration NO_x (ppm)	0.05	0.10	0.20	0.30	0.45
y = Instrument response (volts)	1.20	2.15	3.90	6.20	9.80

Values for m and b for the least-square or "best fit" line can be calculated from: $\sum x$, $\sum y$, $\sum x^2$, $\sum xy$, n, $\bar{y}$, and $\bar{x}$.

Solution:

$$\sum x = 0.05 + 0.10 + 0.20 + 0.30 + 0.45 = 1.1$$
$$\sum y = 1.20 + 2.15 + 3.90 + 6.20 + 9.80 = 23.25$$
$$\sum x^2 = (0.05)^2 + (0.10)^2 + (0.20)^2 + (0.30)^2 + (0.45)^2 = 0.345$$
$$\sum xy = (0.05)(1.20) + (0.10)(2.15) + (0.20)(3.90) + (0.30)(6.20) + (0.45)(9.80) = 7.33$$
$$n = 5$$

$$\bar{x} = \frac{\sum x}{n} = \frac{1.1}{5} = 0.22$$
$$\bar{y} = \frac{\sum y}{n} = \frac{23.25}{5} = 4.65$$

$$m = \frac{7.33 - \frac{(1.1)(23.25)}{5}}{0.345 - \frac{(1.1)^2}{5}} = \frac{2.22}{0.103} = 21.6$$

$$b = 4.65 - (21.6)(0.22) = -0.102$$

The equation for this calibration curve would be y = 21.6x – 0.102, where *y* (the instrument response in volts) is equal to the ambient concentration in ppm times the slope of the line which is 21.6, plus the y-intercept of *x*, which is 0.102.

To calculate ambient concentrations in ppm, we solve the equation for x:

$$x(ppm) = \frac{y - b}{m}$$

$$x(ppm) = \frac{y + 0.102}{21.6}$$

A.5 Measures of Central Tendency

Arithmetic Average, or Mean

A basic way of summarizing data is by the computation of a central value. The most commonly used central value statistic is the arithmetic average, or the mean. This statistic is particularly useful when applied to a set of data having a fairly symmetrical distribution. The mean is an efficient statistic in that it summarizes all the data in the set, and because each piece of data is taken into account in its computation. The formula for computing the mean is:

(Eq. A-4)

$$\overline{X} = \frac{X_1 + X_2 + X_3 ... + X_n}{n} = \frac{\sum X_i}{n}$$

Where:

$\overline{X}$ = *arithmetic mean*
X_i = i^{th} *measurement*
n = *total number of observations*

The arithmetic mean is not a perfect measure of the true central value of a given data set. Arithmetic means overemphasize the importance of one or two extreme data points. Many measurements of a normally distributed data set will have an arithmetic mean that closely approximates the true central value.

Example Problem
Calculate the mean of 3.0, 2.5, 2.2, 3.4, 3.2.

Solution:

$$\overline{X} = \frac{X_1 + X_2 + X_3 ... + X_n}{n} = \frac{\sum X_i}{n}$$

$$\overline{X} = \frac{3.0 + 2.5 + 2.2 + 3.4 + 3.2}{5}$$

$$\overline{X} = \frac{14.3}{5} = 2.86$$

Median

When a distribution of data is asymmetrical, such as that of Figure A-8, it is sometimes desirable to compute a different measure of central value. This second measure, known as the median, is simply the middle value of a distribution, or the quantity above which half the data lie and below which the other half of the data lie.

If *n* data are listed in their *order of magnitude* (from lowest to highest), the median is the [(*n*+1)/2] value. If the number of data is even, then the numerical data of the median is the value midway between the two data nearest the middle. The median, being a positional value, is less influenced by extreme values in a distribution than the mean.

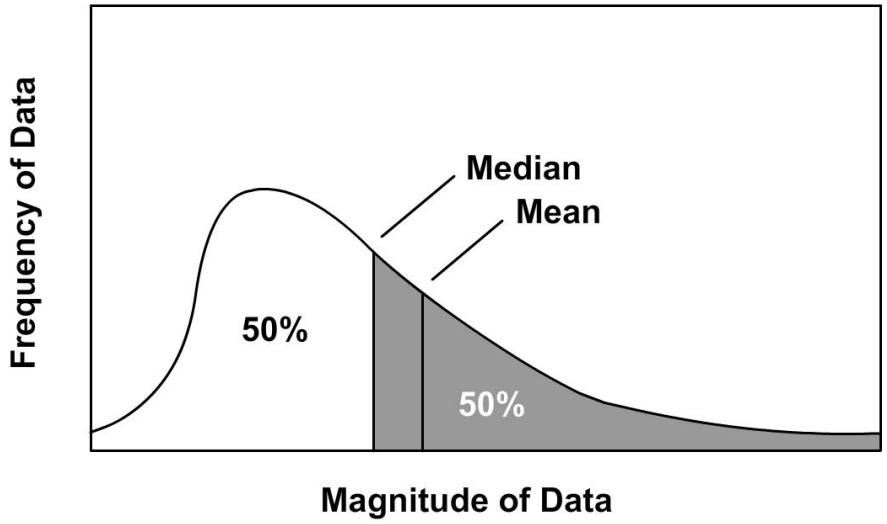

Figure A-8. Example of an asymmetrical distribution of data (median vs. mean).

Example Problem
Find the median of 22, 10, 15, 8, 13, 18.

Solution: The data must first be arranged in order of magnitude, such as:

8, 10, 13, 15, 18, 22

Since $n = 6$, the median is the $7/2 = 3.5$ value, thus the median is 14, or the value halfway between 13 and 15, since this data set has an even number of measurements.

Geometric Mean

Another measure of central tendency used in more specialized applications is the geometric mean (X_g). The geometric mean is defined by using the following equation:

(Eq. A-5) $$\overline{X}_g = \sqrt[n]{(X_1)(X_2)..(X_n)}$$

If scientific calculators are not available, a formula that more readily lends itself to a four-function calculator is:

$$Log_{10}\overline{X}_g = \frac{1}{n}\sum Log_{10}X_i$$

The formula is derived as follows.

$$Log_{10}\overline{X}_g = Log\left[\sqrt[n]{(X_1)(X_2)..(X_n)}\right] = Log\left[(X_1)(X_2)..(X_n)\right]^{1/n}$$

Where: *log is to base 10*

but $LogX^{1/n} = \frac{1}{n}LogX$

and $Log(X \times Y) = LogX + LogY$

Therefore:
$$Log\overline{X}_2 = \frac{1}{n}Log\left[(X1)(X2)(Xn)\right]^{1/n}$$

$$= \frac{1}{n}\left(Log\ X_1 + Log\ X_2 ... + Log\ X_n\right)$$

$$= \frac{1}{n}\sum_i Log\ X_i$$

The geometric mean is most often used for data whose causes behave exponentially rather than linearly, such as in the growth of bacteria, measurements that are ratios, or lognormal distributions.

In a distribution shaped like that of Figure A-8, the geometric mean, like the median, will yield a value closer to the main cluster of values than will the mean. The arithmetic mean is always higher than the geometric mean.

Example Problem
Calculate the geometric mean of 3.0, 2.5, 2.2, 3.4, 3.2.

Solution:

$$\overline{X}_{\tilde{x}} = \sqrt[5]{(3.0)(2.5)(2.2)(3.4)(3.2)} = 2.8$$

or

$$Log_{10}\overline{X}_{\tilde{x}} = \frac{1}{5}(0.477 + 0.398 + 0.342 + 0.531 + 0.505)$$

$$Log_{10}\overline{X}_{\tilde{x}} = 0.4506$$

$$\overline{X}_{\tilde{x}} = 10^{0.4506} = 2.8$$

A.6 Measures of Dispersion

Measures of central tendency are more meaningful if accompanied by information on measures of dispersion. Measures of dispersion describe how the data spread out from the center. Examples of measures of dispersion in a data set include the range, sample standard deviation, coefficient of variation, and the standard geometric deviation.

The Range

The easiest measure of dispersion of a set of data is the difference between the maximum and the minimum values in the set, termed the *range*. The range does not make full use of the information contained in the data, since only two of the data points are taken into account. Thus the range is a useful measure of variability for data sets of 10 or less.

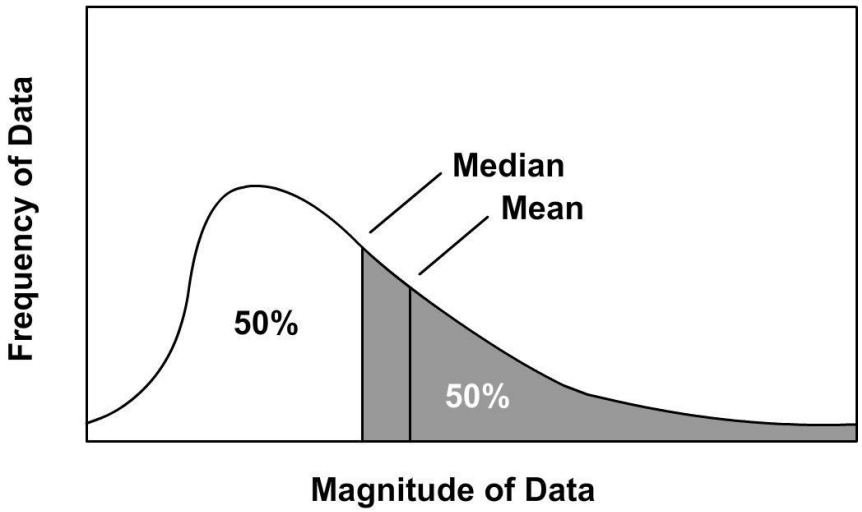

Figure A-9. Dispersion characteristic curves.

Standard Deviation

The most commonly used measure of dispersion, or variability, of sets of data is the standard deviation. Its defining formula is given by the expression:

(Eq. A-6)

$$s = +\sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

Where:

s = *the standard deviation (always positive)*
X_i = *i^{th} measurement*
$\overline{X}$ = *the mean of the data sample*
n = *the number of observations*

The expression $(X_i - \overline{X})$ shows how the deviation of each measurement from the overall mean is incorporated into the standard deviation.

An algebraically equivalent formula that makes computation much easier is:

$$s = +\sqrt{\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}}$$

where the variables are defined as above.

Example Problem: Standard Deviation
Using the data provided in the following table, calculate the standard deviation:

X_i	X_i^2
3.00	9
2.5	6.25
2.2	4.84
3.4	11.56
3.2	10.24
------	------
$14.31 \sum X_i$	$41.89 \sum X_i^2$

Solution:

$$s = +\sqrt{\frac{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}{n-1}}$$

$$s = +\sqrt{\frac{41.89 - \frac{(14.30)^2}{5}}{5-1}}$$

$$s = +\sqrt{\frac{41.89 - \frac{204.49}{5}}{5-1}}$$

$$s = +\sqrt{\frac{41.89 - 40.90}{4}}$$

$$s = +\sqrt{\frac{0.990}{4}}$$

$$s = +\sqrt{.248}$$

$$s = 0.498$$

Coefficient of Variation

The coefficient of variation (CV) is a unitless measure that allows the comparison of dispersion across several sets of data. It is the standard deviation divided by the sample mean. The CV is often used in environmental applications because variability (expressed as standard deviation) is often proportional to the mean.

(Eq. A-7) $$CV = s/\overline{X}$$

Where: s = *standard*
$\overline{X}$ = *sample mean*

Example Problem: Coefficient of Variation
Use the data presented in the previous example problem to solve for the CV.

$$CV = s/\overline{X}$$
$$CV = 0.498/2.86$$
$$CV = 0.174$$

Standard Geometric Deviation

Dispersion of skewed data such as lognormal distributions is measured by the standard geometric deviation. The standard geometric deviation is very similar to the standard deviation. The dispersion in the log of the measurements is measured by the *geometric* standard deviation instead of the dispersion of the measurements which would provide an *arithmetic* standard deviation. The log calculation normalizes the data to better approximate a normal distribution. The formula for calculating the standard geometric deviation is:

(Eq. A-8) $$s_z = antilog\left[\frac{\sum(logX_i) - \frac{\left(\sum logX_i\right)}{n}}{n-1}\right]^{1/2}$$

Where: *log is to the base 10*
s_z = *standard geometric deviation*
X_i = i^{th} *measurement*
X = *the mean of the sample*

The following formula is mathematically identical, yet it is much easier to use in calculation:

$$s_z = antilog\left[\frac{\sum(logX_i) - \frac{\left(\sum logX_i\right)}{n}}{n-1}\right]^{1/2}$$

Example Problem: Standard Geometric Deviation

Using the data provided in the following table, calculate the standard geometric deviation:

X_i	$\log X_i$	$(\log X_i)^2$
3.0	0.4771	.2276
2.5	0.3979	.1584
2.2	0.3424	.1173
3.4	0.5315	.2825
3.2	0.5051	.2552

$$\sum logX_i = 2.2541$$
$$\sum (logX_i)^2 = .0409$$
$$\left(\sum logX_i\right)^2 = 5.0810 \qquad \text{[i.e. } (2.2541)^2\text{]}$$

$$s_z = antilog\left[\frac{\sum (logX_i) - \frac{\left(\sum logX_i\right)}{n}}{n-1}\right]^{1/2}$$

$$s_z = antilog\left[\frac{1.0409 - \frac{5.0810}{5}}{5-1}\right]^{1/2}$$

$$s_z = antilog\left[\frac{1.0409 - 1.0162}{4}\right]^{1/2}$$

$$s_z = antilog\left[\frac{0.0247}{4}\right]^{1/2}$$

$$s_z = antilog[0.0062]^{1/2}$$

$$s_z = antilog[0.0786]^{1/2}$$

$$s_z = 1.1984 \text{or } 1.20$$

A.7 Distribution Curves

Distribution curves are graphical displays of the individual data points in a data set and are important because they can identify patterns and trends in data that might go unnoticed if the data were not plotted.

Many types of distribution curves exist: binomial, t, chi, F, normal, and lognormal are just a few of the existing distributions. However, in air pollution measurements, the normal and lognormal are the most commonly occurring ones. Thus, only these two will be discussed.

The Normal Distribution

One reason the normal (Gaussian) distribution is so important is that a number of natural phenomena are normally distributed or closely approximate it. In fact, many experiments when repeated a large number of times will approach the normal distribution curve. In its pure form, the normal curve is a continuous symmetrical, smooth curve shaped like the one shown in Figure A-10. Naturally, a finite distribution of discrete data can only approximate this curve. The normal curve has the following definite relations to the descriptive measures of a distribution.

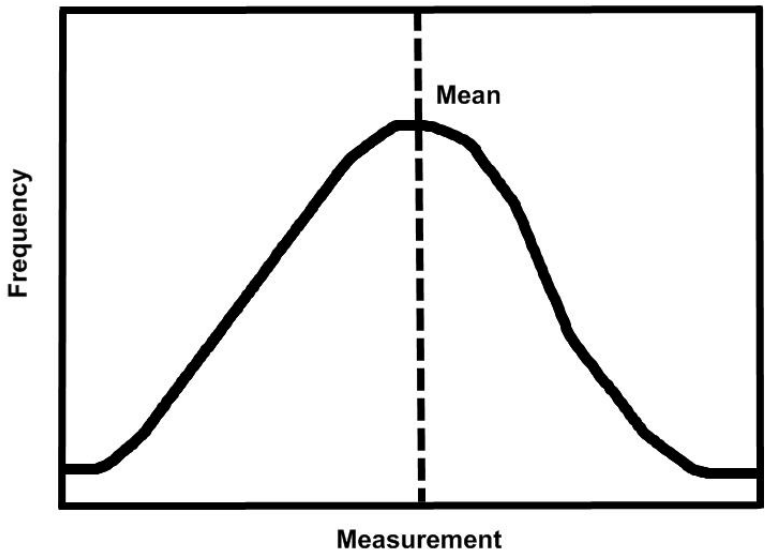

Figure A-10. Normal distribution curve.

The Mean and Median

The normal distribution curve is symmetrical; therefore, the mean and the median are equal and are found at the center of the curve. Recall that, in general, the mean and median of an asymmetrical distribution do not coincide.

The Range

The normal curve ranges along the x-axis from minus infinity to plus infinity. Therefore, the range of a normal distribution is infinite.

The Standard Deviation

The standard deviation, *s*, becomes a most meaningful measure when related to the normal curve. A total of 68.2% of the area lying under a normal curve is included by the part ranging from 1 standard deviation below to 1 standard deviation above the mean. A total of 95.4% lies $\pm$ 2 standard deviations from the mean and 99.7% lies within 3 standard deviations (Figure A-11). By using tables found in statistics texts and handbooks, one can determine the area lying under any part of the normal curve.

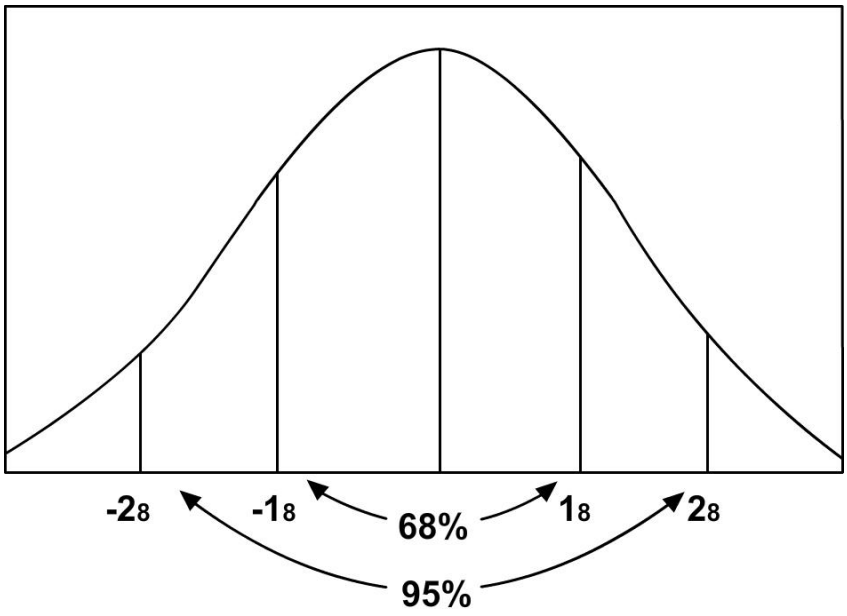

Figure A-11. Characteristics of the normal distribution.

These areas under the normal distribution curve can be given probability interpretations. For example, if an experiment yields a nearly normal distribution with a mean equal to 30 and a standard deviation of 10, we can expect about 68% of a large number of experimental results to range from 20 to 40, so that the probability of any particular experimental result's having a value between 20 and 40 is about 0.68.

In applying the properties of the normal curve to the testing of data readings, one can determine whether a change in the conditions being measured is shown

or whether only chance fluctuations in the readings are represented. For a well-established set of criterion data, a frequently used set of control limits is ± 3 standard deviations. That is, a special investigation of data readings trying these limits can be used to determine whether the conditions under which the original data were taken have changed. Since the limits of 3 standard deviations on either side of the mean include 99.7% of the area under the normal curve, it is very unlikely that a reading outside these limits is due to the conditions producing the criterion set of data. The purpose of this technique is to separate the purely chance fluctuations from the other causes of variation. For example, if a long series of observations of an environmental measurement yield a mean of 50 and a standard deviation of 10, then control limits will be set up as 50 ± 30 - in other words, ± 3 standard deviations, or from 20 to 80. So, a value of 81 would suggest that the underlying conditions have changed, and that a large number of similar observations at this time would yield a distribution of results with a mean different (larger) than 50.

This process of determining whether a value represents a significant change is closely related to the use of control charts. In setting up control limits, it is often necessary to divide the available data into subgroups and calculate the mean and standard deviations of each of these groups, making careful note of the conditions prevailing under each subgroup. In collecting data to establish control limits, as much information as possible should be gathered about the causes and conditions in effect during the period of obtaining a criterion set of data. Generally, the conditions during this period should be "normal," or as much in control as possible.

In the situation where one takes readings of some environmental quantity, the appearance of data beyond the control limits might suggest the starting of a new data grouping to further ascertain whether the underlying environmental variable has changed.

It should be kept in mind that the limits of ± 3 standard deviations are traditional rather than absolute. They have been found through experience to be very useful in many control situations, but each experimenter must decide what limits would be most suitable for a given purpose by determining what levels of probability would be needed to quantify acceptance and rejection bounds.

Lognormal Distributions

Lognormal distributions can best be demonstrated by means of an example:

If hourly sulfur dioxide concentrations are plotted against frequency of occurrence as in the Data Plotting Section, a skewed distribution would exist similar to the one in Figure A-12. Such a curve indicates that many concentrations are close to zero and that few are very high. Unlike temperature, sulfur dioxide concentrations are blocked on the left because values less than zero do not exist. Because numerous aids exist for normal distributions, it is desirable to normalize this type of distribution. By plotting the log of hourly SO_2 concentrations against the frequency of occurrence, a "bell-shaped" curve similar to Figure A-10 is obtained. By making this ample normalizing feature, all existing normal distribution tables can be used to make probability interpretations.

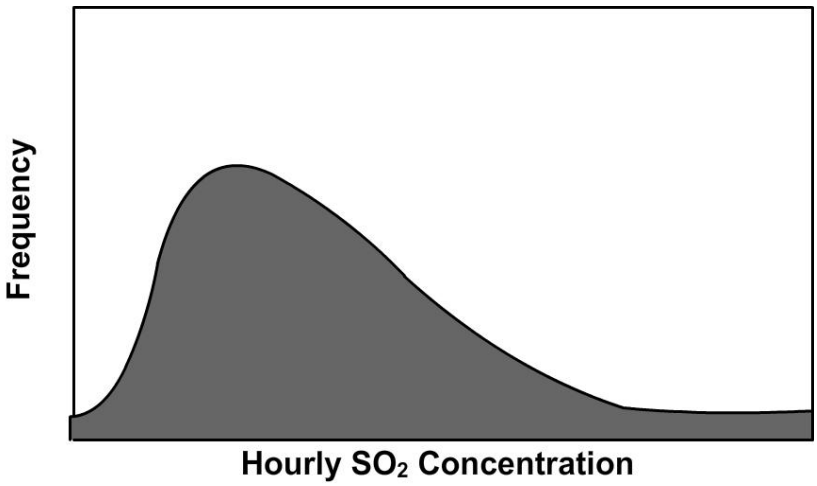

Figure A-12. Frequency vs. concentration of SO_2.

References

Ambient Monitoring Technology Information Center [Internet]. Available at: http://www.epa.gov/ttn/amtic

The Clean Air Act. 42 USC 85, 1990.

School of Public Health, Environmental and Occuapational Health Division, University of Illinois-Chicago. Course: Introduction to Environmental Statistics [Internet]. Available at: http://www.uic.edu/sph/eohs_webcasts.htm

U.S. Environmental Protection Agency. 1997 July 1. 40 CFR Pt. 50, Appendix A.

U.S. Environmental Protection Agency. 40 CFR Pt. 53.

U.S. Environmental Protection Agency. 40 CFR Pt. 58.

U.S. Environmental Protection Agency. 1977 Dec 14. 42 Fed. Reg. 1271-1289.

U.S. Environmental Protection Agency. 1978 Oct 5. 43 Fed. Reg., no. 194, pp. 46258-46261.

U.S. Environmental Protection Agency, Air Pollution Training Institute. Course SI 100: Mathematics Review for Air Pollution Control.

U.S. Environmental Protection Agency. Guidance on systematic planning using the Data Quality Objectives Process. EPA QA/G-4 (EPA/240/B-06/001).

U.S. Environmental Protection Agency. Quality assurance handbook for air pollution measurement systems. EPA 454/R-98-004.