Lecture 27. Air pollution statistics.

Objectives:

1. Statistical analysis of air pollution data.

1. Statistical analysis of air pollution data.

In pollution science statistical methods are applied to quantify and evaluate data derived from observations as well as for mathematical models.

The basis of control of air pollutants consists essentially of two different approaches:

1. Use observations in field studies to develop the empirical source-receptor relationships (for instance, using statistical receptor models);
2. Application of various mathematical chemical transport models that relate the concentration of the secondary pollutants formed to the initial primary pollutant concentrations (NOTE: these models will be discussed in Lectures 29-30).

Statistics is a branch of mathematics that deals with the collection, analysis, and presentation of observations expressed as numbers.

Experiment is a process by which a measurement or observation is obtained. The result of the experiment is called **sample point**, and the set of possible outcomes is called the **sample.**

Sample value is the arithmetic mean defined by

$$\mathbf{x_{mean} = \Sigma\, x_i / n}$$

where x_i is the i-th observation; x_1 , x_2 , x_3,….are the observed values, and n is the number of values.

We are often interested in how much the sample is spread about the mean. The spread, or **dispersion**, from the average is given by the **sample variance s^2** defined by

$$s^2 = \Sigma (x_i - x_{mean})^2 / (n-1)$$

Standard deviation = s (or the square root of the sample variance)

Frequency histogram is a graph in which we plot the measured values against numbers of observations.

Example: The following mass concentrations, q, of PM10 (in $\mu g/m^3$) were measured in Los Angeles:

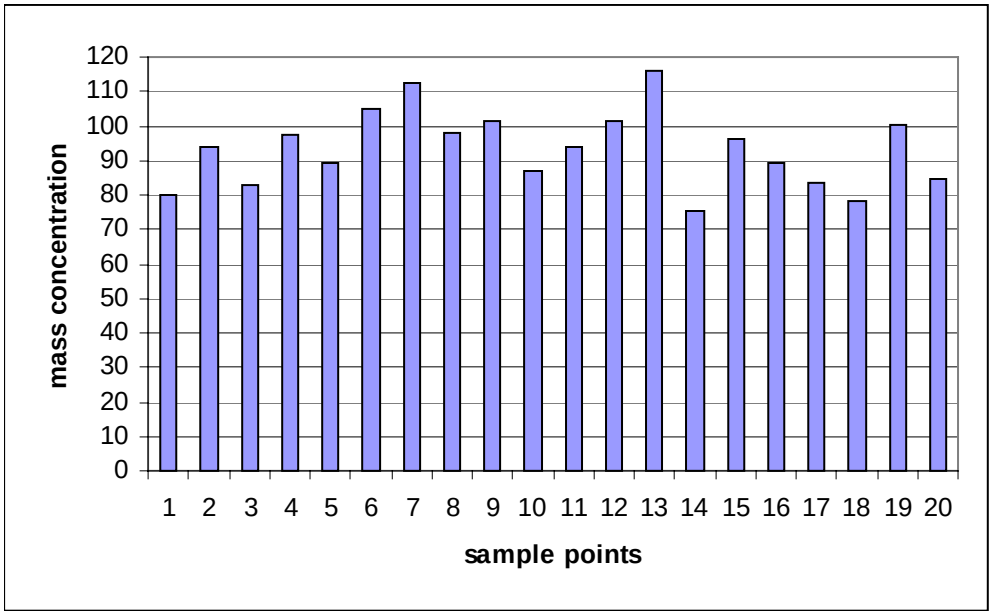
80.2	105.2	94.2	89.2
94.1	112.4	101.7	83.5
100.2	98.2	116.1	112.4
97.3	101.5	118.9	100.3
89.0	87.2	96.5	84.8

This sample has 20 sample points (thus n = 20)

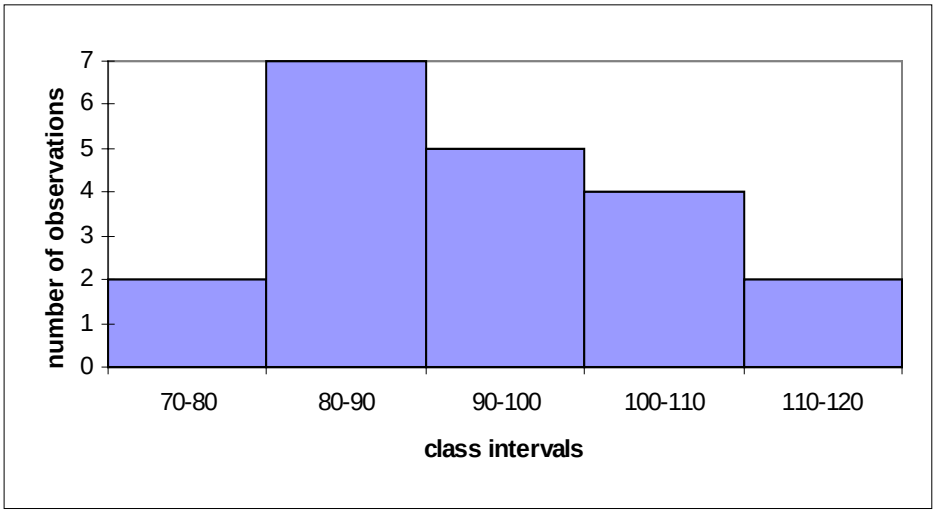
To find the average (or mean) concentration of PM10 we must find the sample mean:

$q_{mean} = \Sigma q_i / n = \Sigma q_i / 20 = 1867.8/20 = 93.39$ ($\mu g/m^3$)

Graphical representation of experimental data:



Frequency histogram for the example above:



- Frequency histogram illustrates the distribution of values versus class intervals, and it shows how often (or frequency) the measured values occur in a given class interval.

If there are two or more variables in the sample and one depends on the other, a regression analysis is often used.

Regression analysis is a way of fitting an equation to a set of data to describe the relationship between the variables.

Linear regression is used when the relationship between variables is close to linear. For two variables (or samples) $X=\{x_i\}$ and $Y=\{y_i\}$ we have:

$$\mathbf{Y = a + bX}$$

where

$\mathbf{b} = \Sigma (x_i - x_{mean}) (y_i - y_{mean}) / (n-1) s_x^2$

$\mathbf{a} = y_{mean} - bx_{mean}$

here s_x is the standard deviation of X, and n is the number of sample points.

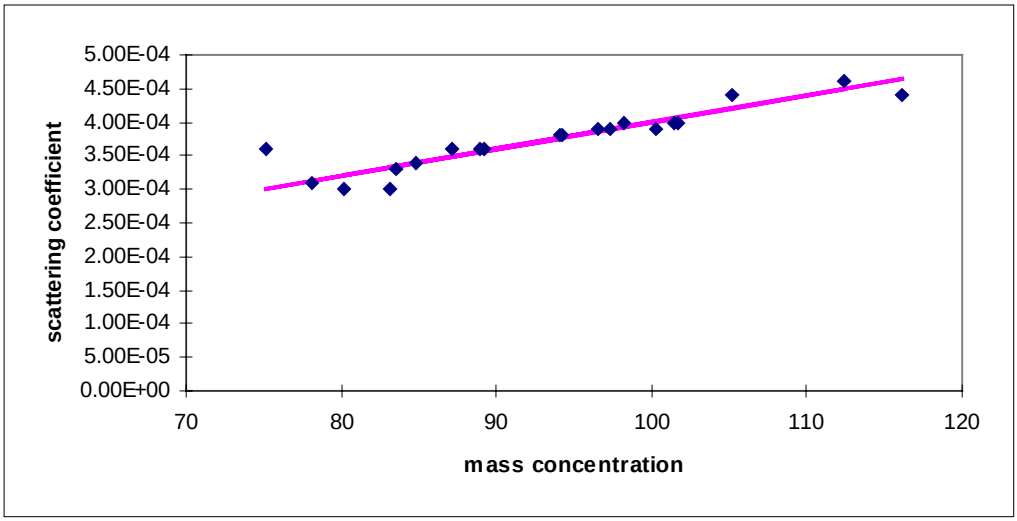
Example: measurements of PM10 mass concentration (see example above) were performed simultaneously with measurements of aerosol scattering coefficient ε_{sc}.

The following data ε_{sc} (m^{-1}) were collected

3.0E-04	4.4E-04	3.8E-04	3.6E-04
3.8E-04	4.6E-04	4.0E-04	3.3E-04
3.0E-04	4.0E-04	4.4E-04	3.1E-04
3.9E-04	4.0E-04	3.6E-04	3.9E-04
3.6E-04	3.6E-04	3.9E-04	3.4E-04

Let's apply the regression analysis for measurements PM10 and ε_{sc}.

First, will plot one variable against the other. We see that we can draw a straight line such that the points are all reasonably close to the line.



We can find the dependence of ε_{sc} (m^{-1}) on mass concentration PM10 ($\mu g/m^3$) by calculating a regression line (in other words, we need to calculate the values of slope **b** and intercept **a**). We have a = $3.12\ 10^{-18}$ and b = $4\ 10^{-6}$, thus

$\varepsilon_{sc} = 3.12\ 10^{-18} + 4\ 10^{-6} q$ or approximate relationship: $\varepsilon_{sc} / q = 4\ 10^{-6}$

- If we need to quantify how closely the straight line fits the data points a correlation coefficient is calculated.

Correlation coefficient, r, is defined by

$$r = \Sigma (x_i - x_{mean}) (y_i - y_{mean}) / (n-1)\ s_x\ s_y$$

where s_x and s_y are the standard deviations for sample X and for sample Y. The square r^2 is called the **coefficient of determination.**

NOTE: larger value of r^2 indicates a better fit. If X and Y follow a perfect linear relationship, r^2 will be exactly 1. In general, r^2 is from 0 to 1; and r is from –1 to 1. The sign of r (the same as for slope b) indicates whether the two variables increase together or are inversely related.

- Because the number of observations is often limited various theoretical frequency distributions are used (such as normal distribution, binomial distribution, and Poison distribution).

Normal distribution (or Gauss distribution) is the most widely used theoretical distribution, which gives the frequency (or density) distribution by

$$f(x) = \exp[-(x - \mu)^2/ (2 \sigma^2)] / \{ (2\pi)^{1/2} \sigma \}$$

where μ is the population mean, and σ is the population variance. Here population is a collection of all possible sample points.

NOTE: Although μ and σ defined for the population almost the same as x_{mean} and s defined for the sample, there is an important difference: x_{mean} and s depend on the samples taken and only approximate theoretical μ and σ which are assumed to be known exactly. This means that as the number of samples increases, x_{mean} and s should get closer and closer to μ and σ.

<u>For normal distribution:</u>

about 68% (or about 2/3) of the values will lie between $\mu - \sigma$ and $\mu + \sigma$;

about 95% of the values lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.