

Chemometrics

A chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant information by analyzing chemical data.

1. Development of method (chemistry)

Selection

Optimization

Sample preparation

2. Determination

Measurement

3. Data interpretation

Data acquisition

Signal Chemical Information

Chemical Information User information

Figure 3 from text

Precision and Accuracy

Precision: refers to the tightness of a set of measurements (or estimators) and it is measured by standard statistical measures of spread, i.e. standard deviation, interquartile range, mean absolute deviation from the mean, etc.

Accuracy: refers to the distance of a set measurements (or estimators) from the true value and it is measured by standard statistical measures of accuracy such as Mean Squared Error and mean absolute deviation from the true value.

Sources of error: Must be understood from a global view as well as microscopic view. Improvements in over all accuracy are important, but tremendous improvements in unimportant parts may be unimportant.

Variability (Spread or measures of sample precision)

Range Difference between largest and smallest measurement

Percentiles are used to calculate the interquartile range and other measures of spread

Interquartile range is the difference between the upper and lower quartiles

Variance of a set of measurements is denoted by

$$s^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2$$

Standard deviation is $\sqrt{s^2} = s$.

(This is not an unbiased estimate of)

Approximate value for S is Range/4.

Better Estimates

3/4 interquartile range is substituted for S.

For normally distributed data

68% of measurements are within 1 std of mean

95% of measurements are within 2 std of mean

Gauss' inequality

Chebychev inequality

Standard error of the sample mean = $\frac{S}{\sqrt{n}}$ and it is estimated by $\frac{S}{\sqrt{n}}$

.

Measures of location

Most descriptions include location and size.

The first numerical description of your data will be its location.

Mode = value (or values) that occur most often.

Median = Middle value of the data, i.e. half the data is bigger and half the data is smaller.

$$\text{Mean} = \text{average} = \frac{1}{n} \sum_{i} y_i = \bar{y}$$

Descriptive Properties

Mean most often used and understood.

It is taught in grammar school.

The mean is extremely sensitive mistakes in data collection.

Median is taught later in grammar school

than mean. It is less well understood and often confused with mean.

It is not very sensitive to mistakes in data collection.

Mode is not taught till statistics courses and can give odd descriptions of location.

Probability distributions:

If measurements can be modeled as occurring on a continuous scale then they are often modeled by a density function, $f(x)$.

The density $f(x)$ gives the probability(long run percent) of measurements occurring in an interval A by the formula

$$P(A) = \int_A f(x) dx .$$

If measurements can be better modeled as occurring on a discrete scale then it has a probability mass function $g(x)$ that gives the probability(long run percent) of measurements occurring in an interval A by the formula

$$P(A) = \sum_{a_j \text{ in } A} g(a_j) .$$

Mean and variance of populations:

Quadrature formulas:

$$\text{Mean} \quad \sum_{i=1}^n b_i X_i = \sum_{i=1}^n b_i \mu_i$$

$$\text{Variance} \quad \sum_{i=1}^n b_i X_i^2 = \sum_{i=1}^n b_i^2 \mu_i^2$$

$$\text{Normal density } f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

Central Limit Theorem

Confidence Intervals:

For the mean or for the true value???

Assumptions and their importance in science.

Population

$$\text{Discrete measurement} \quad \mu = \sum_{i=0}^{\infty} x_i g(x_i)$$

$$\text{Continuous measurement} \quad \mu = \int_0^{\infty} g(x) dx$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 g(x_i)$$

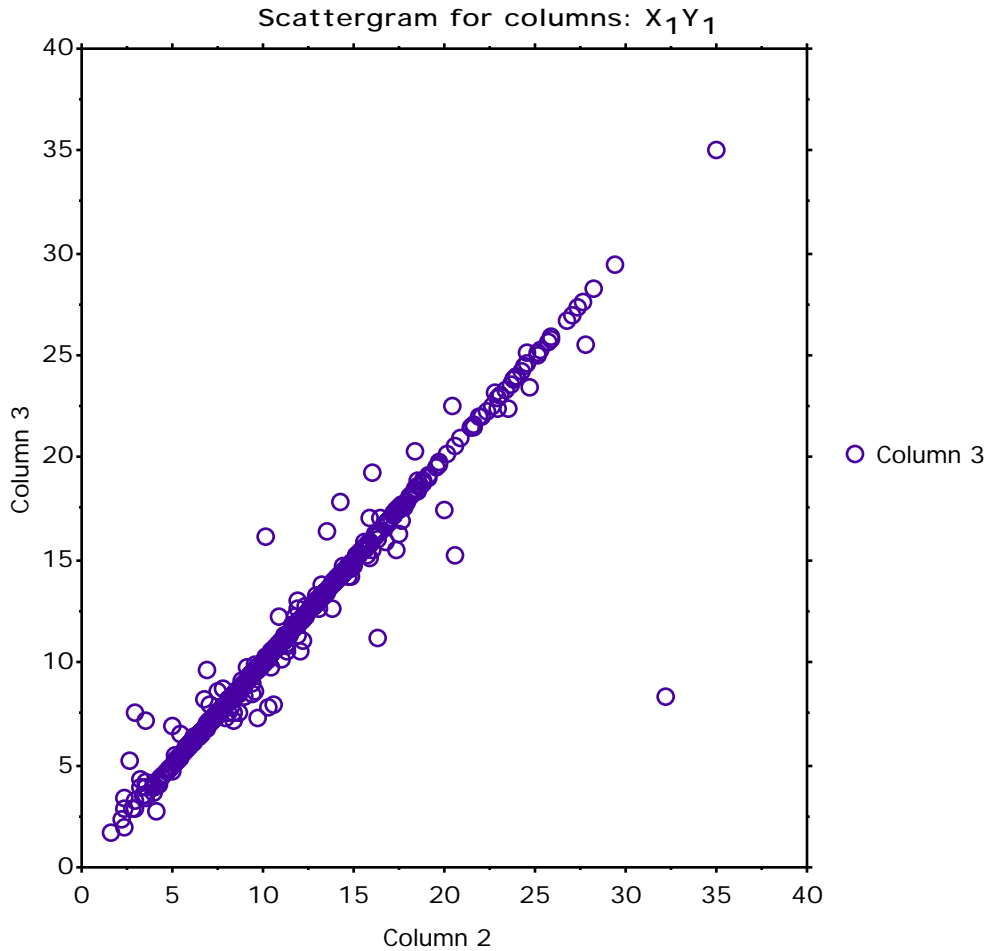
$$s^2 = \int_0^{\infty} (x - \mu)^2 g(x) dx$$

Measuring Laboratory Bias

Done using interlaboratory comparisons:

Use standard reference materials (SRM's) and a basic ANOVA model(we'll get into this later)

Youden Plots: Each laboratory measures two SRM's one is plotted on the x axis and one on the y axis. If a big correlation exists between x and y there is said to be a method bias otherwise not.



Relationship between precision and concentration levels:

For most chemical measurements, concentration and precision are related. Usually the greater the concentration the greater the standard deviation of the measurements.

Data from "AU for 221"

