12.119 Geochemical Analysis of Environmental Materials Spring 2006

Precision, Accuracy and Quality Control

When we interpret geochemical data, we must keep the "significance" of the numbers in mind at all time. Given two numbers (say, an EPA "action level" and a measurement of an element in a water sample), is the measurement above or below the limit. If the sample measures "29 ppb" and the action level is "30 ppb", is the water safe to drink? If you were informed that the measurement has an uncertainty of 2 ppb, would your answer change?

Some terms to be familiar with:

Bias	Standard Deviation
Accuracy	Standard Error
Precision	Pooled Standard Deviation
"Internal" Precision	One-sigma, two-sigma, three-sigma
"External" Precision	
Reproducibility	
	Small-number statistics
Error	t-statistic
Systematic error	
Random error	Detection Limit
Correlated errors	
Significant figures	Counting Statistics
Error Propagation	Shot Noise
Mean (average)	Variance
Median	Regression
Gaussian distribution (bell curve)	Linear Regression
Population	Correlation
Sample population	
degrees of freedom	Significant figures

Standard Deviation, Pooled Standard Deviation, and Standard Error

Mean:

$$x_{mean} = \frac{\sum_i x_i}{n}$$

Standard Deviation of the mean:

$$\sigma = \sqrt{\frac{\sum_i (x_i - x_{mean})^2}{n-1}}$$

where

x_i are the individual observations
x_{mean} is the average of the individual observations
n is the number of points

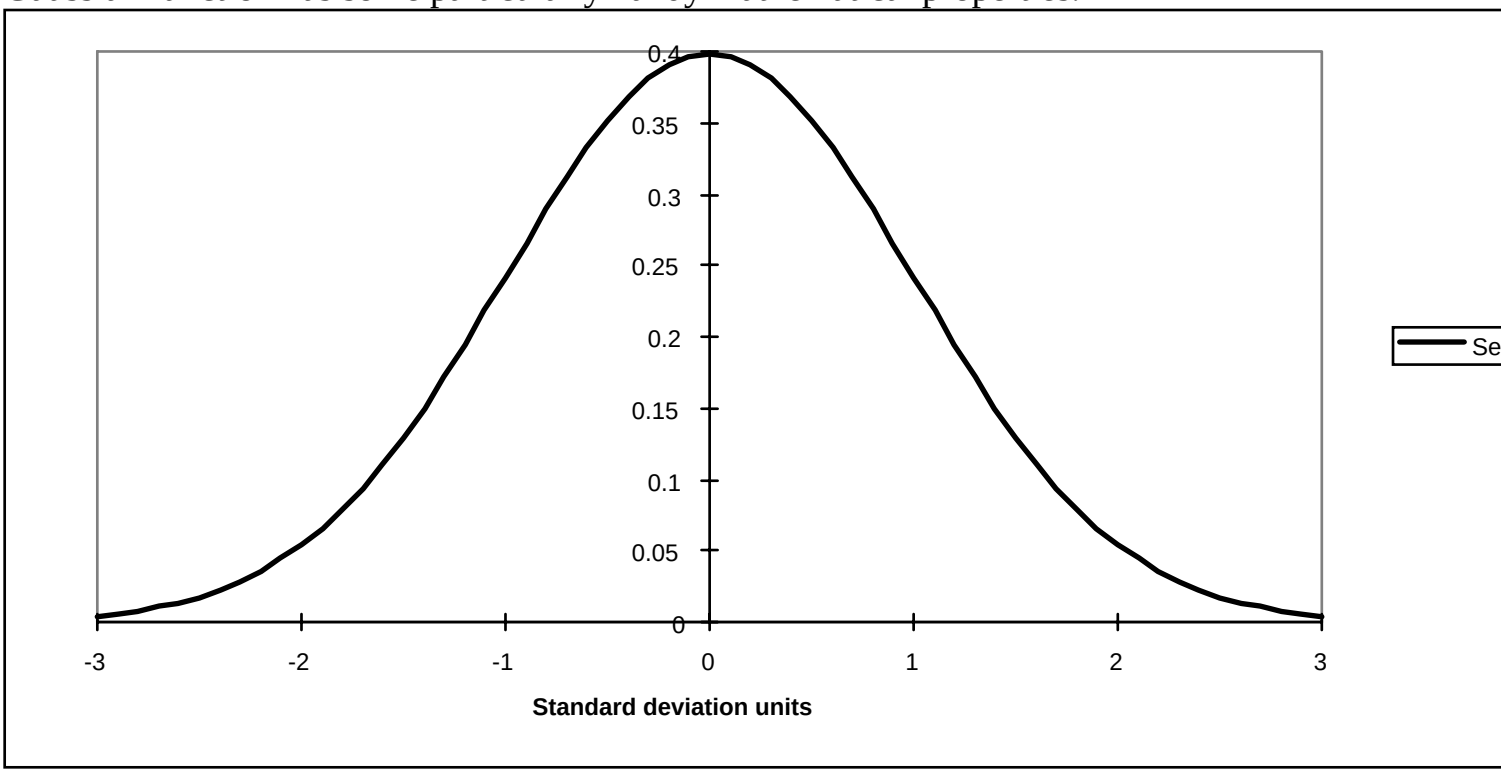
The term n-1 is introduced to compensate for the loss of degrees of freedom. When you made your n observations, you had n degrees of freedom; in other words, none of the numbers was literally determined by any of the other numbers. However, when you calculate a mean you not only have the n observations, you have one more number (something for nothing; get 11 numbers for the price of 10...). Since you can't have something for nothing, you have to give up one degree of freedom because the average is not independent of the n observations. Stating it another way, if I told you that I had ten numbers and gave you nine of them and the average, you could calculate the tenth.

The squaring and square roots derive to the concept of the **Gaussian Distribution** (as known to physicists; statististicians and mathematicians call it the "normal distribution" and social scientists call it the "bell curve"), in which the frequency distribution of multiple observations of a variable ***x*** that has a mean value of ***μ*** is distributed according to the equation:

$$P(x) = \frac{1}{\sigma\sqrt{\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

The basic idea is that small errors are more probable than large errors. For some processes (e.g. radioactive decay or ion counting), we can prove that the deviations should follow this relation. For some other processes, we know that the deviations should follow a different probability density function (PDF). Most of the time, we

don’t know for sure what the distribution is. But except for exceptional cases requiring very high accuracy of error estimates, we typically just assume that the relationship is Gaussian, because most other relationships are similar to it, and the Gaussian function has some particularly handy mathematical properties.



If you integrate under a bounded range of the curve, you get the percentage of observations that will fall within those bounds. e.g., the integral of the bell curve within the bounds ±1σ is 71%; within the bounds ±2 σ is 95%; within the bounds ±3 σ is 99%.

In order to get a good estimate of the standard deviation, you need many measurements. For example, if n=10, then the standard deviation is estimated only to within about 20%.

In real life, you rarely analyze even as many as ten replicates of samples. Often, you may want to do samples in duplicate (or triplicate at most) to be sure that some error hasn't occurred. In this case, you can estimate the pooled standard deviation: given a set of replicate analyses of different samples, the pooled standard deviation estimates the error by pooling the statistics.

$$PSD = \sqrt{\frac{\sigma_1^2 f_1 + \sigma_2^2 f_2 + \ldots}{f_1 + f_2 + \ldots}}$$

where f_i is the number of degrees of freedom for sample 1 (e.g., for triplicate analyses of sample 1, $f_1 = 3\text{-}1 = 2$; for duplicate analyses of sample 2, $f_2 = 2\text{-}1 = 1$). This approach assumes that all of the measurements have the same σ, so it cannot be

applied to cases where the error varies between measurements. For example, in some analyses, the error is a constant percentage of the signal. In that case, you can pool samples that fall within some range of values.

Propagation of errors:

The basic idea is that if a (final) number is computed from other (primary) numbers, an uncertainty in the primary numbers propagates to uncertainty in the final number. The simplest way to estimate the effect of error in one of the primary numbers on the final number is just to do the calculation 3 times:

Suppose $y = f(x)$ and that x has an uncertainty σ_x. Then calculate:

$$y(x+\sigma_x),\ y(x),\ y(x-\sigma_x)$$

Then $y = y(x) + y(x+\sigma_x)-y(x) - y(x-\sigma_x)-y(x)$

If the relationship between y and x is non-linear (e.g. $y=e^x$) and the error is relatively large, the difference between the estimate and the upper error bound may differ from that of the lower error bound. This method does not conform to strict statistical principles, but it is better than nothing when you have a formula that doesn't obey the simple rules below.

If a final number is calculated from several primary numbers, then a more formal error analysis is necessary. A decision is also required: Are the errors uncorrelated (no relationship between the magnitudes of the errors between the primary variables) or correlated? Most error analyses assume that errors are uncorrelated (although this is not necessarily so). In this event, however, it is not possible to do the calculation as we have done it above because when one error in variable 1 zigs up, the error in variable 2 may be zigging down.

The total error **du** is equal to the partial derivatives of **u** with respect to each variable times the error attributable to each variable **dx**. In other words, if **u** is the result of a calculation involving the three variables **x**, **y**, and **z**, then the uncertainty in **u** can be calculated from the partial derivatives:

$$du = \left(\frac{\partial u}{\partial x}\right)_{y,x} dx + \left(\frac{\partial u}{\partial y}\right)_{x,z} dy + \left(\frac{\partial u}{\partial z}\right)_{x,y} dz$$

Generally, most calculations of this sort can be simplified by assuming that the errors are small compared to the numbers themselves (often but not always true).

Approximations using Taylor expansions are employed, leading to the following expressions, where **u** and **v** are numbers with uncertainties $\boldsymbol{\sigma_u}$ and $\boldsymbol{\sigma_v}$

First, if you simply multiply a number **u** with an uncertainty $\boldsymbol{\sigma_u}$ by an exact constant **a** with no error (e.g. π is an exact constant), then the error in the product **au** is:

$$\sigma_{au} = a\sigma_u$$

For addition, multiplication, and division:

If $x = u \pm v$, then $\sigma_x^2 = \sigma_u^2 + \sigma_v^2 \pm 2\sigma_{uv}^2$

If $x = \frac{u}{v}$, then $\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} - 2\frac{\sigma_{uv}^2}{uv}$

If $x = uv$, then $\frac{\sigma_x^2}{x^2} = \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} + 2\frac{\sigma_{uv}^2}{uv}$

Practically, as long as the errors are a small percentage of the numbers themselves and are not correlated between variables, the final terms of these equations (cross products) can be neglected, resulting in the simpler expressions:

If $x = u \pm v$, then $\sigma_x^2 \approx \sigma_u^2 + \sigma_v^2$

If $x = \frac{u}{v}$, then $\frac{\sigma_x^2}{x^2} \approx \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2}$

If $x = uv$, then $\frac{\sigma_x^2}{x^2} \approx \frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2}$

In most cases you will be able to use these approximations in your calculations for this course. If you have to use more complicated functions (e.g. logs or exponentials), you can approximate the error in f(x) by calculating $f(x-\sigma_x)$ and $f(x+\sigma_x)$.

Curve Fits and Small Number Statistics

Although it is not always the only right thing to do, it is common (and for Gaussian error distributions, statistically correct) to do a least-squares fit to data. The idea is that given some function f(x) that the data should conform to, the "best" fit is one that minimizes the standard deviation of the observations relative to the curve fit.

If you are using observations to fit a presumed function f(x) with coefficients $\mathbf{a_i}$ that are determined by the fit (e.g., fitting a straight line with intercept $\mathbf{a_1}$ and slope $\mathbf{a_2}$ to a series of paired x,y observations) and want to see how well the function fits the data, you can calculate the standard deviation of the fit:

$$\text{Standard deviation (s.d.)} = \sqrt{\frac{\sum_i \left(y_i - f(x_i)\right)^2}{n - m}}$$

where m is the number of coefficients in your curve fit (e.g., if you estimate a slope and intercept, then your degrees of freedom is reduced to n-2).

For example, if we have a series of paired observations $\mathbf{(x_i,y_i)}$ which we are fitting to a line $\mathbf{a_1 + b_1 x}$, then we calculate:

$$S = \sum_i \left[y_i - \left(a_1 - a_2 x_i \right) \right]^2$$

We want a_1 and a_2 such that S is minimized. So we take the derivatives of S with respect to a_1 and a_2 and set them to zero. This gives us two equations in two unknowns, and so we can solve for a_1 and a_2. This gives the usual least squares formula built into calculators. Note that this line may not always be the truly "best" line, because it assumes that the **x** variable has NO error and that all of the error is in the **y** variable. In the more general case where both **x** and **y** have errors, a more complicated treatment is necessary [the "York" method: Earth Planet. Sci. Lett. 5:320-324 (1969)].

For large n, there are standard statistical procedures for calculating the error on the estimates for derived quantities such as slope and intercept of a straight line. Because of the aforementioned "luck of the draw" effects, the error of derived quantities such as the slope and intercept is often underestimated. For example, if you have only three paired observations (x_1,y_1) and (x_2,y_2), they will define a straight line with only one degree of freedom. The estimated error for slope and intercept of a least-squares fit calculated from the large-n statistical formulas will underestimate the true error, so it is necessary to adjust these estimates using a **t-statistic**, a multiplier that compensates for the small number statistics. For example, for a straight line fit for n=3, t=6.3 (but t declines rapidly for larger n).

Counting Statistics

Sometimes we measure properties by counting individual events. For example: How many atoms of a radioactive isotope have decayed and been counted by a Geiger counter? How many atoms have passed through a mass spectrometer and hit the detector? How many photons have been detected by a photomultiplier tube?

There is a fundamental law of statistics that applies whenever we sample a larger population by randomly by counting. If we have counted **n** atoms, then the one-sigma uncertainty is $\sqrt{\mathbf{n}}$. For example, suppose we were to count the number of radioactive decays occurring for a 1 minute interval, and then repeat this process many times. If on average we find **n** decays per minute, the standard deviation of our count data for the many 1 minute intervals would be $\sqrt{\mathbf{n}}$.

Quality Control in the Analysis of Geological Materials: Part I

In addition to fundamental instrumental, methodological and statistical errors that affect the analysis (that will apply whenever a measurement is made), further errors will creep into any effort to analyze materials over an extended period. For example: the standards employed may be unstable or erroneous pipettes and balances may be misused or improperly calibrated. In general, few laboratories can consistently analyze a well-known standard as an unknown and obtain results over months and

years with a standard deviation as low as that obtained from replicate determinations for a shorter period. In order to maintain the best reproducibility over long time periods and assess the true analytical reproducibility, quality-control efforts are required.

Government agencies have devised various analytical quality control procedures known by names such as “GLP” (good laboratory practices), "CLP" (chemical laboratory practices), etc.. These are rigidly enforced upon any laboratory that does work that is required by federal regulations. If you go to work in such a laboratory, you will want to become familiar with these procedures because they are required by law. Most academic labs do not follow these procedures. They are not required to do so by law, and in any event most academic laboratories feel that their own quality control efforts are superior to the rigid CLT protocols because they are tailored to the application at hand, rather than being monolithically imposed to suit everything. The CLT protocols are probably successful in imposing some minimum level of competency on environmental consulting firms, but in some cases they require an inappropriate error-ridden old-fashioned method to the exclusion of a modern method that gives more reliable numbers.

One of the keys to maintaining good quality control is the maintenance of a thorough laboratory book. A good lab book starts with bound pages (so they don't fall out from wear-and-tear) of high quality acid-free paper. Entries into this book should be dated and made either with a permanent ink (carbon-black is best) or with similarly reliable computer output (these days, that generally means black laser-printer output, although old-fashioned dot-matrix printers with carbon ribbons are also satisfactory. But the laser printer paper used would have to be acid free ideally, and it hardly ever is). Thermal paper and (often but not always) ink-jet prints are not stable and will fade with time, exposure to light, and being taped over...). Any procedure performed on a sample is described in complete detail. For example:

> 2.15 mg of reagent grade $CaCO_3$ was weighed out onto weighing paper on a tared Mettler P61 balance and transferred into a calibrated 500 ml glass volumetric flask (V=499.1 ml). 100 ml of water distilled from a quartz still was added to the flask, the solid was stirred into the solution, and 10 ml of concentrated analytical grade HNO_3 added dropwise so as to minimize fizzing and spattering. Another 380 ml of distilled water was added. After cooling to a room temperature of 20°C, the solution was diluted to the mark and thoroughly mixed.

If a procedure is repeated over and over, you can simply refer to an earlier description of the method and give the essential changing details such as the sample weights. Save the raw data from the instrument in the lab book in a permanent form, and describe the calibration and calculation procedures.

One modern problem is that data is often collected by computer systems; how does one properly archive this data when data storage systems, programs, and protocols are rapidly changing. A multiple-method backup system is usually best, including:

(1) Paper output for primary data, if it can be done reasonably (sometimes, a plot is better than a long list of numbers, e.g. in chromatography).

(2) Save the primary data on digital medial in as simple a computer representation as possible; e.g., rather than saving the data table (only) as an Excel©, SigmaPlot©, or MassLynx© file, save it as a flat delimited ASCII files. The reason for doing this is that ASCII files have been around for decades and all computers can read ASCII files. There aren't any current versions of Excel, for example, that will read "Excel 1.5" files.

For the uninitiated, a "flat delimited ASCII file" has the following properties:

(1) It only contains ASCII characters (which are ultimately one-byte numbers between 0 and 127); basically, ASCII characters are English letters and numbers, along with a few punctuation marks, symbols, and "control characters" such as "tab" and "end of paragraph".

(2) It is "delimited" if it is essentially a list in row-and-column format, with the individual "cell" entries listed in order starting from the upper left corner, across the columns to the right, then the second row, etc. down to the lower left. The individual numbers in a row are separated by a "delimiter" such as "tab", "comma", "semicolon", "space" – or – the columns have a fixed number of entries per row ("fixed width") e.g. a 7 digit number in row 1, a five digit number in row 2, etc – and the rows are terminated by and "end of paragraph" mark.

These days, most commercial software has a "save as…" option that will allow you to choose "text" file and perhaps give you a choice as to the method of delimitation.

(3) Make at least two copies of the files in separate directories on your hard disk (in case of file corruption), and make copies on at least two separate removable media storage systems stored in different places from your computer (and better yet, from each other. The reasoning behind this advice is: (a) file corruption on hard disks is the most common cause of data loss, (b) hard drive failure is the second most common cause of data loss, (c) removable media can be unreliable (I have had CD-R disks that simply sat in a box without ever being touched go from readable to unreadable over a period of a few years), and (d) when the fire, flood, tornado, hurricane, earthquake, or meteor strike occurs, do you want to still have access to your data?

Fundamentals of Correlation Analysis

I. Correlation

A. Correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right) = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$

where S_x = std. dev. of $x = (\Sigma(x_i - \bar{x})^2/n-1)^{1/2}$

and S_y = std. dev. of $y = (\Sigma(y_i - \bar{y})^2/n-1)^{1/2}$

i.e. S_x and S_y relate the deviations of points from the average relative to the "range" (actually std. dev.) of the observations.

B. Variance and Covariance

1. Variance (of **x** is denoted S_x^2; variance of **y** is denoted S_y^2) is a measure of the scatter of values of a variable about its mean:

$$S_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 / n$$

2. Covariance (of **x** and **y**) expresses the relationship between two variables (a measure of the scatter of values of points in a plane relative to the mean value):

$$S_{xy}^2 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / n$$

3. r^2 is "the variance of **Y** accounted for by it's covariance with **x**" (usually expressed in % units). In other words, the Covariance divided by the product of the Variances.

C. Relation to linear regressions (x on y; y on x; others)

1. Common linear regression of **y** on **x**: **y = A + Bx**

Let $\mathbf{S = \Sigma\,(y_i - A - B\,x_i)^2}$

Set $\partial S / \partial A = 0$; $\partial S / \partial B = 0$; solve for **A** and **B**.

2. Matrix math solution of Linear Regression:

for eq'n $\mathbf{A\underline{x} = \underline{b}}$ (m eq'ns, n unknowns),
if columns of A are linearly independent,
then:

$$\underline{x} = (A^T A)^{-1} A^T \underline{b}$$

a. For example, for the simple linear regression

$$\mathbf{y = C + Dx,}$$

where we want to fit pairs of data

$$\mathbf{x_i,\ y_i}$$

we want to find

$$\mathbf{C,\ D}$$

that mimimize

$$\mathbf{\Sigma\ [y_i - (C + Dx_i)]^2}$$

In matrix form, we write the equation **y = C + Dx** as:

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

i.e. $\mathbf{A\ \ x\ =\ b}$

Similarly, to solve the equation $y = A + Bx + Cx^2$:

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & \\ . & . & \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

b. Simple matrix formulas also allow you to compute the estimated uncertainties of the regression coefficients and the correlation coefficients.

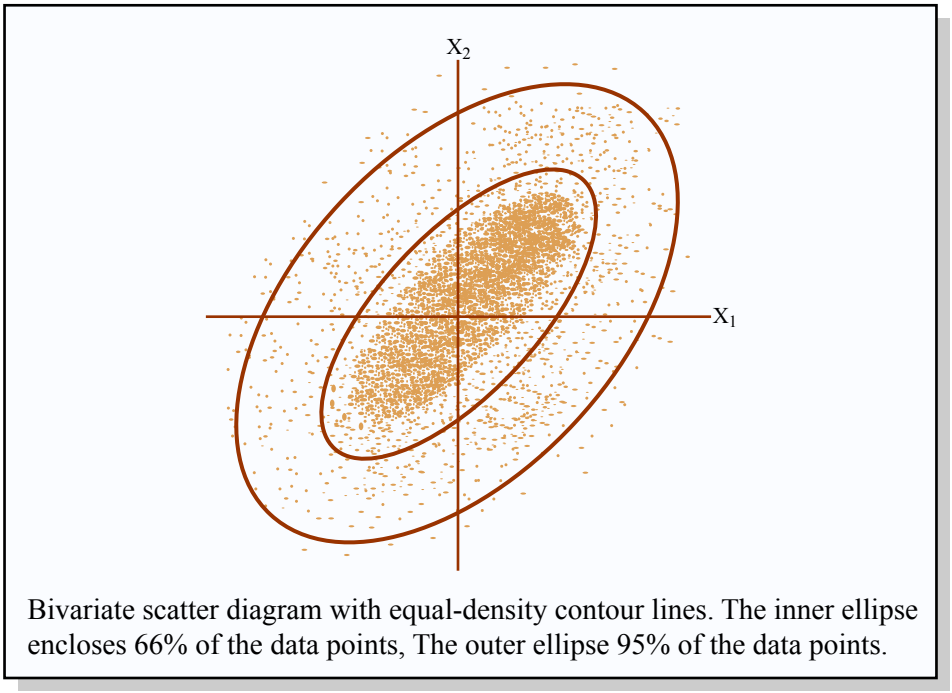
II. Correlation in n dimensions

A. Multiple linear regression

B. r-matrix (later we will refer to this as the matrix Σ)

Property	1	2	3	4	5
1	1.00	0.86	0.45	0.83	0.45
2	0.86	1.00	0.74	0.23	0.64
3	0.45	0.74	1.00	0.78	0.57
4	0.83	0.23	0.78	1.00	0.39
5	0.45	0.64	0.57	0.39	1.00

(ρ_{ij})



Bivariate scatter diagram with equal-density contour lines. The inner ellipse encloses 66% of the data points, The outer ellipse 95% of the data points.

Figure by MIT OCW.

source: Joreskog et al. Geological Factor Analysis (1976)

C. Factor analysis: construction of a few “artificial variables” that caption statistical essence of a large data set.