Data analysis and Geostatistics - lecture IV

correlation and the significance of the correlation coefficient

Propagation of uncertainties in equations

uncertainties are cumulative: need to add uncertainties from all sources:

- add/subtract: $w = ax + by - cz$ with $x \pm s_x$, $y \pm s_y$, $z \pm s_z$

$$s_w^2 = (as_x)^2 + (bs_y)^2 + (cs_z)^2$$

- multiply/divide: $w = x^a y^b z^{-c}$ with $x \pm s_x$, $y \pm s_y$, $z \pm s_z$

$$(s_w/w)^2 = (as_x/x)^2 + (bs_y/y)^2 + (cs_z/z)^2$$

Propagation of uncertainties in equations

uncertainties are cumulative: need to add uncertainties from all sources:

general formula: $w = f(x_i \ldots . x_n)$

$$s_w = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial w}{\partial x_i}\right)^2 \cdot s_{x_i}^2}$$

Propagation of uncertainties in equations

Examples:

$D_{bulk} = D_{ol} \cdot X_{ol} + D_{cpx} \cdot X_{cpx} + D_{mt} \cdot X_{mt} + \ldots$

assume there is only uncertainty on the mole fraction X
assume there is uncertainty on both D and X

Rb-Sr dating formulation;

$$t = \lambda^{-1} \ln \left\{ \frac{{}^{87}Sr/{}^{86}Sr - ({}^{87}Sr/{}^{86}Sr)_0}{{}^{87}Rb/{}^{86}Sr} \right\}$$

$\lambda \pm s_\lambda$
${}^{87}Sr/{}^{86}Sr \pm s_{SrSr}$
$({}^{87}Sr/{}^{86}Sr)_0 \pm s_{SrSr0}$
${}^{87}Rb \pm s_{Rb}$
${}^{86}Sr \pm s_{Sr}$

Standard error on the mean - SE

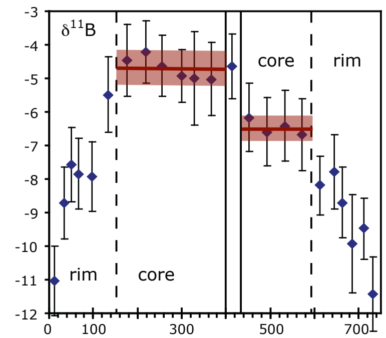
Propagated uncertainties often result in an overestimate of the actual uncertainty on the calculated property: always better off to obtain a direct estimate from duplicates where possible

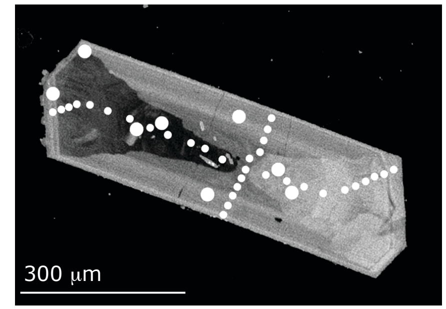
d^{11}B on the ion-probe:
- uncertainty on standardization
- uncertainty on bias/matrix correction
- uncertainty on each analysis

total uncertainty = 0.75‰ + 0.50‰ + 1.25‰ = 2.5‰ on each d^{11}B value

Standard error on the mean - SE

individual uncertainties versus uncertainty on the mean





All values in one domain are higher than in the other: no coincidence

Standard error on the mean - SE

As long as samples form homogeneous group: can calculate the mean and associated uncertainty on this mean

This is of course similar to repeat analysis of the same material

This property is called the standard error on the mean (SE) and is calculated from:

$$SE^2 = \frac{s^2}{n}$$

where n is the number of samples, so the more samples the smaller the SE

A great feature of the standard error on the mean is that it is completely independent of the shape of the host distribution

Central limit theorem

means from any distribution will tend to a normal distribution at increasing n, and so will the SE

So, when 5 geologists all sample the same set of rivers, their means will be normally distributed, whatever the original distribution was

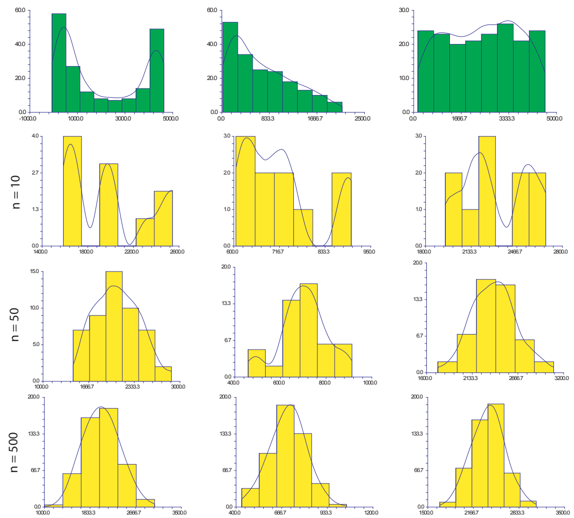
and this fit will improve with increasing number of geologists

This is clearly very useful and provides a method to deal with difficult or unknown distributions

So let's test this !

the spread in the means is smaller than the spread in the original data: have obtained a more precise estimate of the population mean !

The central limit theorem



Correlation: quantifying element relationships

So far we have been treating variables as isolated properties, where one variable is not linked in any way to another. However, many variables are linked and we can use this link or correlation between them.

Plotting relationships; x-y scatter plots and scatterplot matrices

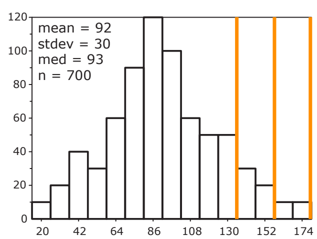
Correlation analysis; how to characterize correlations in numerical and non-numerical data, quantify the “degree of correlation”, and how to test if correlations are real

Regression analysis; quantitative formulation of correlation (y = ax + b), which allows for interpolation and extrapolation beyond the input data

Why are correlations important ?

The conc. of a heavy metal in soils from all over Europe:

determine the natural background so you can set pollution criteria



nice continuous distribution of the data;
can describe it with a mean/median and stdev/IQR

conclusion;
spread is large in the data, but there are no clear signs of pollution

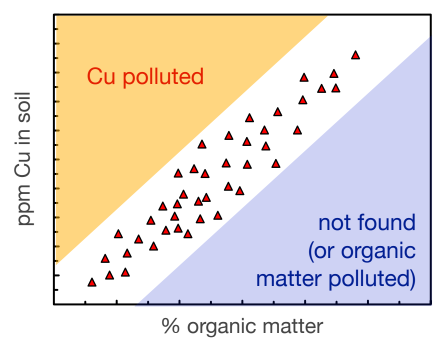
however; some samples were from heavily polluted sites, so why don’t they jump out in the total data set?

unlikely to be one background value: will depend on soil type, composition etc

Why are correlations important ?

The content of a heavy metal in soils from all over Europe:

the organic matter content of the soil completely controls the concentration of this heavy metal:



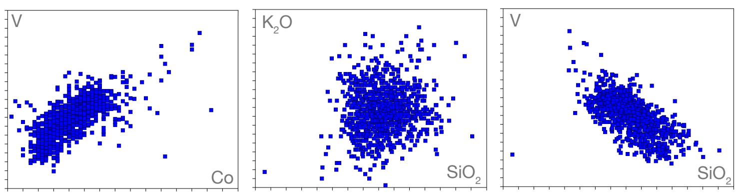
any soil with high organic matter content will have a natural enrichment

pollution will be an enrichment beyond that caused by organic matter

Plotting correlations: x-y scatterplots

Plotting variables against each other in x-y scatterplots is a very fast way to look for correlations between variables, and the sense of this correlation: is it positive (one enhances the other) or negative (one suppresses the other)

positive correlation | no correlation | negative correlation

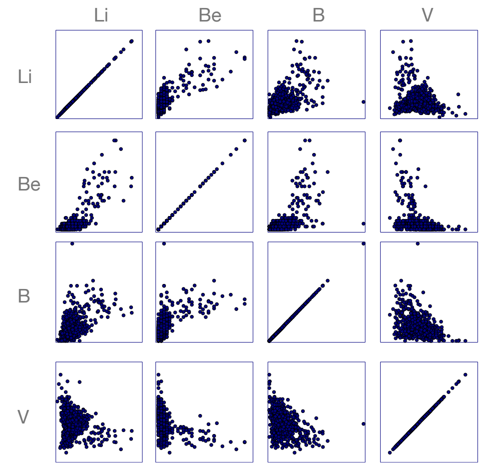


high Co is associated with high V
same source or process

high SiO_2 is associated with low V
same source or process

Plotting correlations: x-y scatterplot matrix

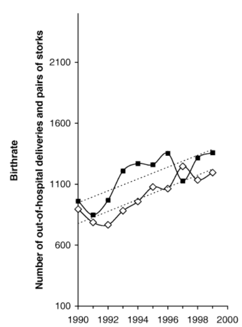
Most statistical software packages allow you to plot scatterplots in a matrix



Scatterplot matrices are a good way to quickly eyeball a dataset. Not only shows correlations, but also cases of multi-modality

Spurious correlations

Correlations are not proof of a link between the variables or of causation and you should always use common sense when interpreting correlations.



Worse, correlations may be introduced by data processing, e.g. closure

The correlation coefficient - closure effects

Correlation is sensitive to closure issues resulting from forcing values to a specified sum

These data are quite common in geology; weight % data for bulk rock analyses or EMP mineral analyses, % of a unit in a core, etc

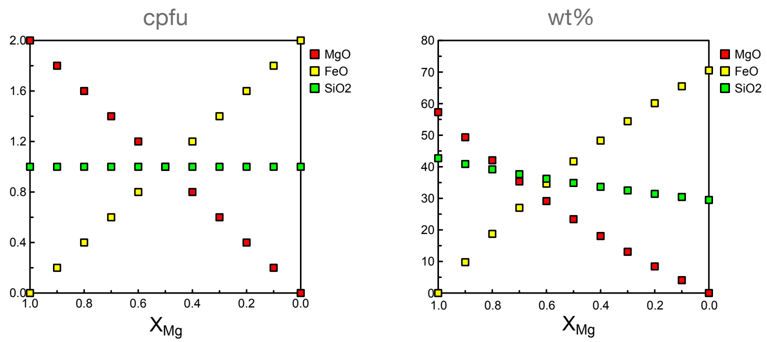
closure: when one element goes up, the others have to go down to satisfy a 100% sum

this mainly affects the major elements as changes in trace elements normally won't change the sum significantly

This introduces apparent correlation where there is none

Closure - examples: normalization of olivine data

Choice of normalization in mineral analysis



Looking at correlations that are generated by a mathematical transformation of your data —> an artefact !

Closure - examples: Kawah Ijen

Hydrothermal alteration of rocks - leaching

leaching removes everything except SiO_2:
silica appears to be added during hydrothermal alteration

Closure by leaching

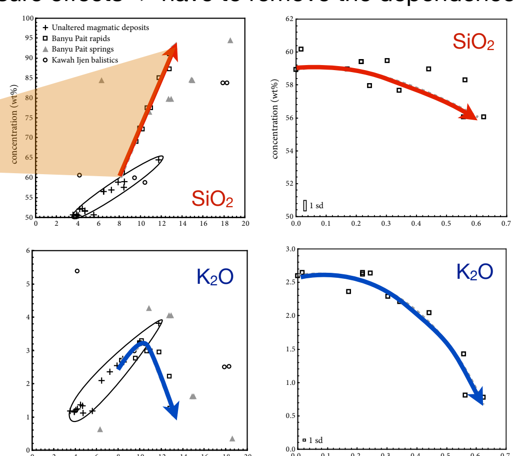
Acid leaching results in removal of all elements except SiO_2:

	wt%	wt%	re-norm to 100%	wt%	wt%	wt%	wt%
SiO_2	60	60		62.5	65.2	68.2	71.4
Al_2O_3	15	13.5		14.1	13.0	11.9	10.7
MgO	4	3.6		3.8	3.5	3.2	2.9
FeO	11	9.9		10.3	9.6	8.8	7.9
CaO	10	9		9.4	8.7	8.0	7.1
leaching	0%	10%		10%	20%	30%	40%

Results in residual enrichment and artificial correlations

Closure - how to deal with it

To correct or check for closure effects -> have to remove the dependence on the fixed total

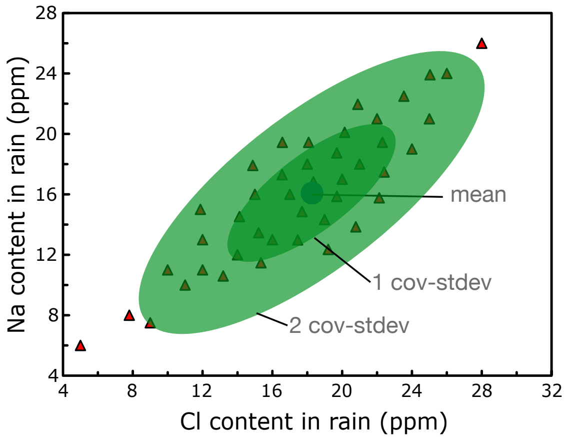


For example:

- use a ratio instead of variables separately
- normalize to a conservative elements

Covariance and correlation in variables

covariance is equivalence of variance in univariate case



Error propagation and covariance

covariance - the degree of correlation between the variables:

$$\mathrm{cov}_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$ compare with normal variance $$\frac{\sum (x_i - \bar{x})^2}{n-1}$$

when covariance is high: strong correlation between variables

However, inconveniently, cov_{xy} depends on the actual values of x and y

compare it against the variance in x and variance in y

or, in other words, determine how much of the total variance can be explained by covariance

The correlation coefficient

the correlation coefficient describes the amount of variance explained by covariance between variables:

$$r = \frac{\mathrm{cov}_{xy}}{s_x s_y}$$

when covariance close to variance: r -> 1

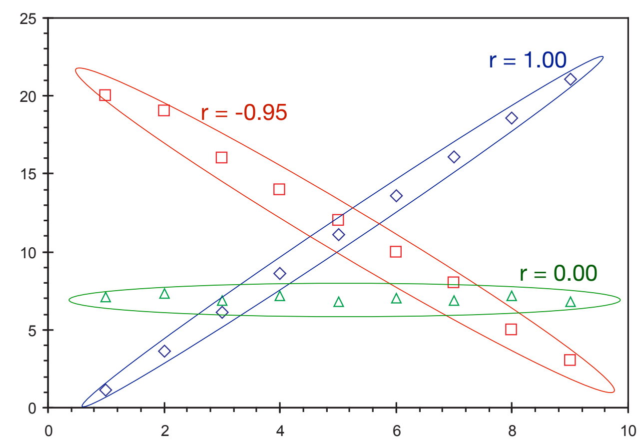
when variance >> covariance: r -> 0

So what do values of r mean;

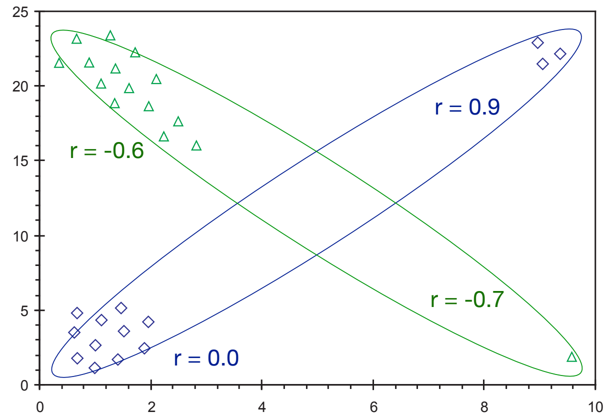
r = -1 perfect negative correlation between variables
r = +1 perfect positive correlation between variables
r = 0 no correlation: the variables are independent

This r value is known as the Pearson correlation coefficient
(not the same as R^2)

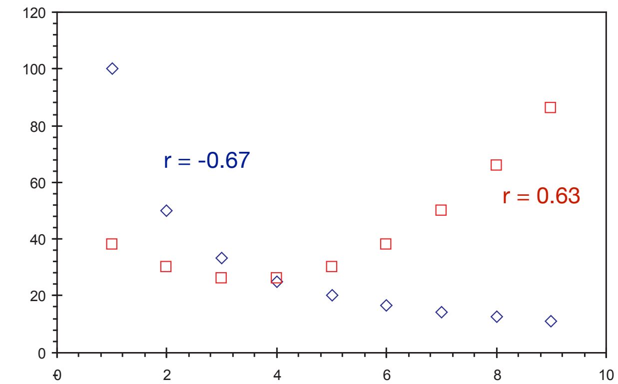
Examples of correlations - the good



Examples of correlations - the bad

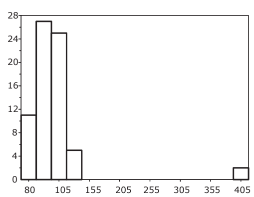


Examples of correlations - the ugly

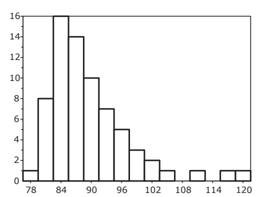


The correlation coefficient - data distribution

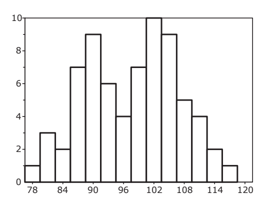
The correlation coefficient places strict constraints on the distribution of the input data: **normal for all vars**



outliers:
remove



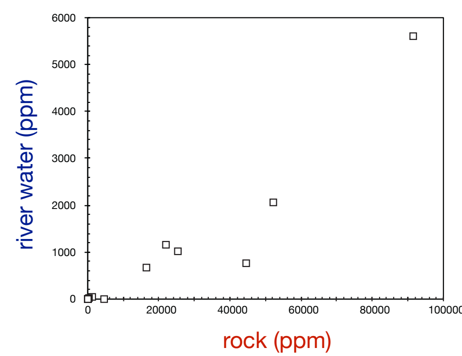
lognormal:
log transform

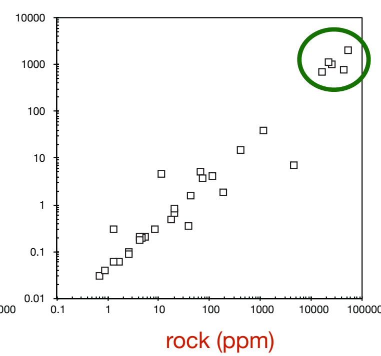


bimodal:
split up

Log-normal transformation

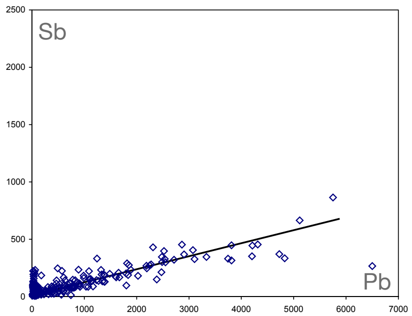
Easy to work with log-transformed data; just calculate the log of each value

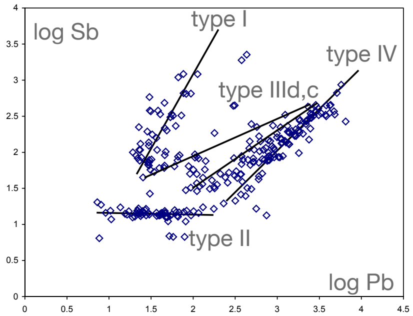




Concentrations in rocks vary from low ppt to wt%: difficult to compare all of them in one diagram in linear space, but works well after log-transform

The correlation coefficient - lognormal data





lognormal data exaggerate correlations at high concentration and can hide correlation at lower concentration + correlation coefficient is overestimated

Data transformations

The logarithmic transform is not the only data transformation that is useful. Others include:

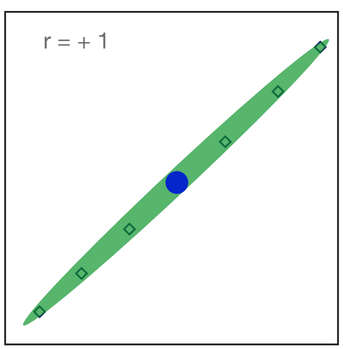
- Reciprocal: $^1/_x$
- Square root: $\sqrt{x}$
- Angular transformation: sin (x)

The important thing to note here is that such a transformation does not make any change to the data. At any point, you can transform the data and any derived properties back into linear space.

and you should !

Correlation and covariance

covariance requires a normal distribution in both variables



Perfect trend, x-y covariance equals variance in x and y

However; neither variable in this case is normally distributed because data points are equally spaced

This dataset **fails** the requirements for the Pearson correlation coefficient

Switch to a robust estimator of correlation: the Spearman correlation coefficient

The correlation coefficient - rank data

not all data can easily be transformed to a normal distribution: rank statistics

x_8 = 20 = 1
x_2 = 31 = 2
x_7 = 46 = 3.5
x_4 = 46 = 3.5
x_3 = 50 = 5
x_6 = 52 = 6
x_1 = 56 = 7
x_5 = 64 = 8

The rank correlation coefficient is known as the Spearman correlation coefficient and is calculated as follows;

$$r' = 1 - \left\{ \frac{6 \sum (R(x_i) - R(y_i))^2}{n\,(n^2-1)} \right\}$$

The Spearman r is a robust estimator, because it is not sensitive to outliers:

whether x_5 equals 64, 640 or 6400, its rank remains unchanged

However, lost some information: instead of an actual value, now only use its rank

Correlation and covariance

Why worry about normal distribution of variables ?

- In previous example, the covariance was obvious, but what if r = -0.4 ?
  - deviations from normality can easily introduce or hide the correlation between variables
- When requirements for Pearson r (or any stat property) are not met, the obtained value becomes meaningless
  - r = -0.9 describes the same amount of correlation for every combination of normally distributed variables, but this is not the case for variables deviating from normality.
  - lose your ability to compare: statements lose their strength

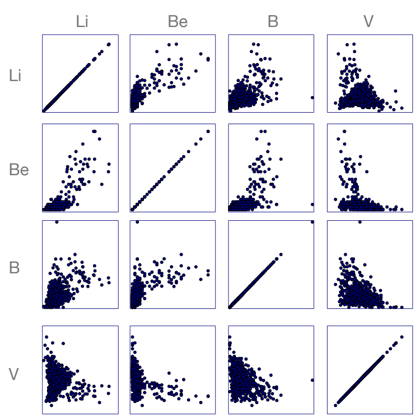
The importance of meeting method prerequisites

Why worry about method prerequisites ?

Your statistical argument loses all its value when the method prerequisites are not met. In the best scenario, by sheer luck it doesn't matter, but in general it leads to a wrong interpretation/conclusion. Occasionally, it has major implications

Correlation coefficients matrices

to quickly data mine large data sets: make a correlation coefficient matrix



	Li	logBe	B	logV
Li	1	0.7	0.5	-0.3
logBe	0.7	1	0.6	-0.5
B	0.5	0.6	1	-0.4
logV	-0.3	-0.5	-0.4	1

Correlation coefficients matrices

to quickly data mine large data sets: make a correlation coefficient matrix

	SiO2	Al2O3	Fe2O3	CaO	MgO	K2O	log_Mn	log_Ti	log_P	Li	log_Be	B	log_V	log_Cr	log_Co	log_Ni	log_Cu
SiO2	1	-0.5	-0.8	-0.3	-0.6	0.1	-0.5	-0.5	0.0	0.1	0.2	0.2	-0.7	-0.5	-0.6	-0.5	-0.4
Al2O3	-0.5	1	0.3	-0.3	0.1	0.2	0.2	0.3	0.0	0.2	0.1	0.1	0.4	0.2	0.2	0.2	0.2
Fe2O3	-0.8	0.3	1	0.4	0.7	-0.2	0.5	0.7	0.0	-0.3	-0.4	-0.4	0.8	0.5	0.6	0.5	0.4
CaO	-0.3	-0.3	0.4	1	0.6	-0.2	0.3	0.4	0.3	-0.1	-0.2	-0.2	0.3	0.3	0.3	0.3	0.2
MgO	-0.6	0.1	0.7	0.6	1	-0.3	0.3	0.6	0.0	-0.1	-0.3	-0.3	0.7	0.7	0.7	0.7	0.4
K2O	0.1	0.2	-0.2	-0.2	-0.3	1	0.0	-0.4	0.2	0.0	0.1	0.3	-0.3	-0.3	-0.2	-0.2	-0.1
log_Mn	-0.5	0.2	0.5	0.3	0.3	0.0	1	0.2	0.1	-0.1	-0.1	-0.1	0.3	0.2	0.4	0.3	0.3
log_Ti	-0.5	0.3	0.7	0.4	0.6	-0.4	0.2	1	0.0	-0.9	-0.4	-0.6	0.8	0.6	0.5	0.6	0.4
log_P	0.0	0.0	0.0	0.3	0.0	0.2	0.1	0.0	1	0.2	0.2	0.1	-0.1	-0.1	-0.1	-0.1	0.0
Li	0.1	0.2	-0.3	-0.1	-0.1	0.0	-0.1	-0.9	0.2	1	0.7	0.5	-0.3	0.0	-0.1	0.0	-0.1
log_Be	0.2	0.1	-0.4	-0.2	-0.3	0.1	-0.1	-0.4	0.2	0.7	1	0.6	-0.5	-0.1	-0.2	-0.2	-0.1
B	0.2	0.1	-0.4	-0.2	-0.3	0.3	-0.1	-0.6	0.1	0.5	0.6	1	-0.4	-0.1	-0.1	-0.1	0.0
log_V	-0.7	0.4	0.8	0.3	0.7	-0.3	0.3	0.8	-0.1	-0.3	-0.5	-0.4	1	0.7	0.6	0.6	0.5
log_Cr	-0.5	0.2	0.5	0.3	0.7	-0.3	0.2	0.6	-0.1	0.0	-0.1	-0.1	0.7	1	0.7	0.8	0.4
log_Co	-0.6	0.2	0.6	0.3	0.7	-0.2	0.4	0.5	-0.1	-0.1	-0.2	-0.1	0.6	0.7	1	0.8	0.5
log_Ni	-0.5	0.2	0.5	0.3	0.7	-0.2	0.3	0.6	-0.1	0.0	-0.2	-0.1	0.6	0.8	0.8	1	0.5
log_Cu	-0.4	0.2	0.4	0.2	0.4	-0.1	0.3	0.4	0.0	-0.1	-0.1	0.0	0.5	0.4	0.5	0.5	1

Correlation coefficients matrices - significance

But are these r values meaningful?

In statistical terms: are they significantly different from r = 0

there will be a critical r value above which it is significant

	SiO2	Al2O3	Fe2O3	CaO	MgO	K2O	log_Mn	log_Ti	log_P	Li	log_Be	B	log_V	log_Cr	log_Co	log_Ni	log_Cu
SiO2	1	-0.5	-0.8	-0.3	-0.6	0.1	-0.5	-0.5	0.0	0.1	0.2	0.2	-0.7	-0.5	-0.6	-0.5	-0.4
Al2O3	-0.5	1	0.3	-0.3	0.1	0.2	0.2	0.3	0.0	0.2	0.1	0.1	0.4	0.2	0.2	0.2	0.2
Fe2O3	-0.8	0.3	1	0.4	0.7	-0.2	0.5	0.7	0.0	-0.3	-0.4	-0.4	0.8	0.5	0.6	0.5	0.4
CaO	-0.3	-0.3	0.4	1	0.6	-0.2	0.3	0.4	0.3	-0.1	-0.2	-0.2	0.3	0.3	0.3	0.3	0.2
MgO	-0.6	0.1	0.7	0.6	1	-0.3	0.3	0.6	0.0	-0.1	-0.3	-0.3	0.7	0.7	0.7	0.7	0.4
K2O	0.1	0.2	-0.2	-0.2	-0.3	1	0.0	-0.4	0.2	0.0	0.1	0.3	-0.3	-0.3	-0.2	-0.2	-0.1
log_Mn	-0.5	0.2	0.5	0.3	0.3	0.0	1	0.2	0.1	-0.1	-0.1	-0.1	0.3	0.2	0.4	0.3	0.3
log_Ti	-0.5	0.3	0.7	0.4	0.6	-0.4	0.2	1	0.0	-0.9	-0.4	-0.6	0.8	0.6	0.5	0.6	0.4
log_P	0.0	0.0	0.0	0.3	0.0	0.2	0.1	0.0	1	0.2	0.2	0.1	-0.1	-0.1	-0.1	-0.1	0.0
Li	0.1	0.2	-0.3	-0.1	-0.1	0.0	-0.1	-0.9	0.2	1	0.7	0.5	-0.3	0.0	-0.1	0.0	-0.1
log_Be	0.2	0.1	-0.4	-0.2	-0.3	0.1	-0.1	-0.4	0.2	0.7	1	0.6	-0.5	-0.1	-0.2	-0.2	-0.1
B	0.2	0.1	-0.4	-0.2	-0.3	0.3	-0.1	-0.6	0.1	0.5	0.6	1	-0.4	-0.1	-0.1	-0.1	0.0
log_V	-0.7	0.4	0.8	0.3	0.7	-0.3	0.3	0.8	-0.1	-0.3	-0.5	-0.4	1	0.7	0.6	0.6	0.5
log_Cr	-0.5	0.2	0.5	0.3	0.7	-0.3	0.2	0.6	-0.1	0.0	-0.1	-0.1	0.7	1	0.7	0.8	0.4
log_Co	-0.6	0.2	0.6	0.3	0.7	-0.2	0.4	0.5	-0.1	-0.1	-0.2	-0.1	0.6	0.7	1	0.8	0.5
log_Ni	-0.5	0.2	0.5	0.3	0.7	-0.2	0.3	0.6	-0.1	0.0	-0.2	-0.1	0.6	0.8	0.8	1	0.5
log_Cu	-0.4	0.2	0.4	0.2	0.4	-0.1	0.3	0.4	0.0	-0.1	-0.1	0.0	0.5	0.4	0.5	0.5	1