

Data analysis and Geostatistics

Short Course on the use of statistical techniques
in the geosciences



Vincent van Hinsberg • McGill University



The damaging impact of means

The “average” human: male, 25-30 years old, 76 kg, 1.77 m tall, caucasian

This model controls:

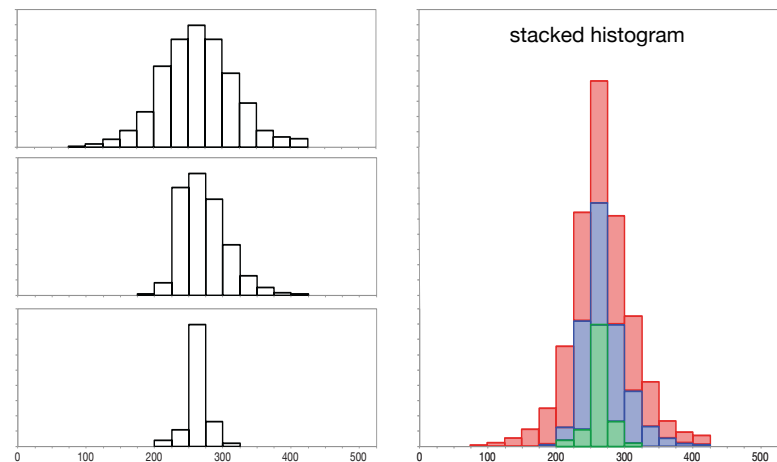
- Car crash-test dummies
- Office temperature
- Police officer's safety vests
- Gas masks
- Height of desks, shelves, cupboards, etc
- Exposure limits for chemicals
- Size of gadgets, including phones
- Size of tools, bricks, notebooks, etc etc etc

Women are 47% more likely to be seriously injured in a car crash, 71% more likely to be moderately injured and 17% more likely to die, which can be directly related to car design (Guardian, Feb 23 2019).

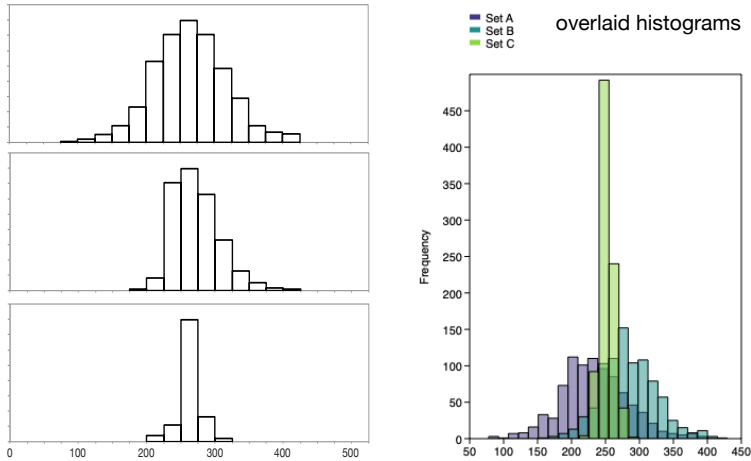
Geotop Short Course in Data Analysis and Geostatistics
Part 4. Univariate data visualization and comparison



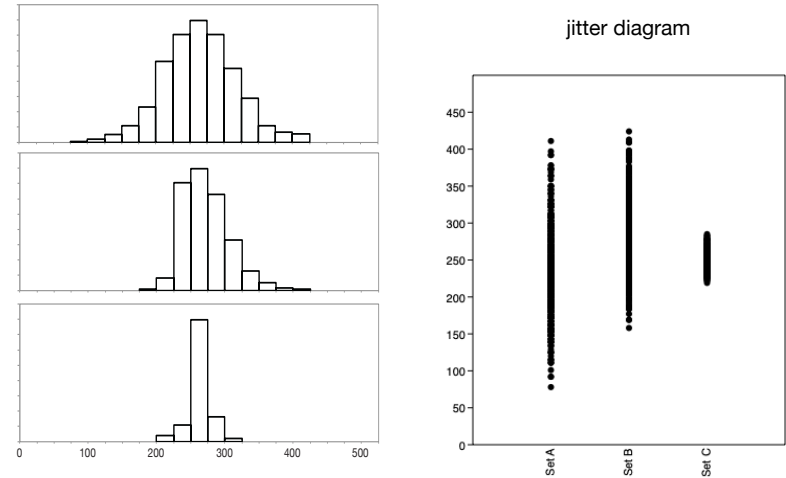
Graphical representation of data - comparisons



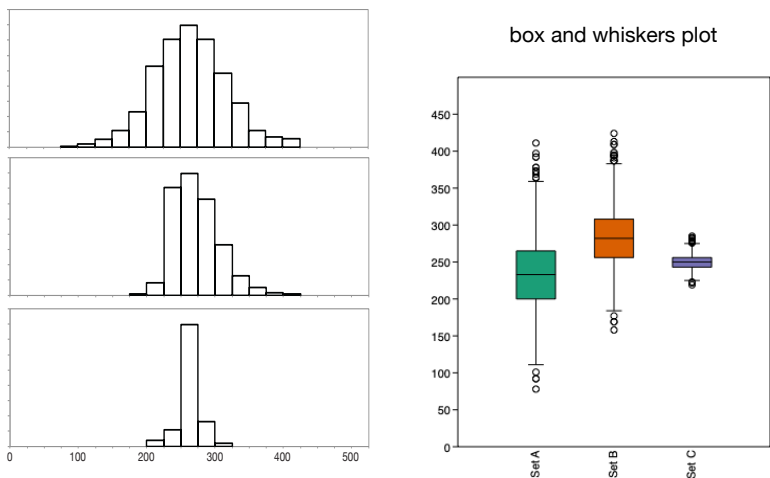
Graphical representation of data - comparisons



Graphical representation of data - comparisons



Graphical representation of data - comparisons



Box and whiskers plots

histograms are not the only way to show the distribution of a data set

- stem and leaf diagrams
- box and whiskers plots - extremely useful in data comparisons:

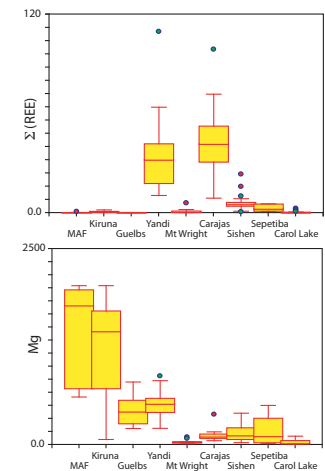
- extreme value
- outlier
- spread

extreme values:
outside the 3x IQR
box - unlikely part of distribution

outliers: within a box defined by 3x the IQR - part of a normal distribution

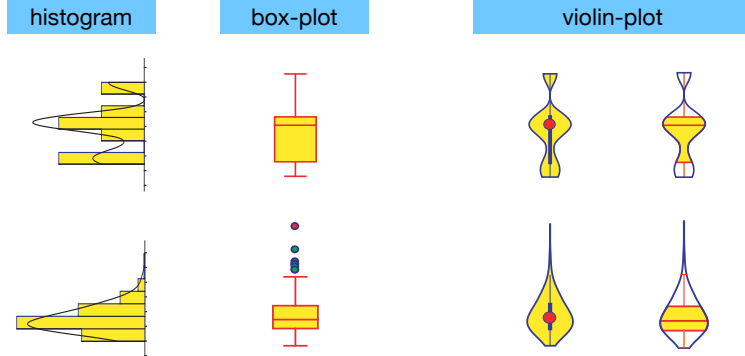
- P₇₅
- median
- P₂₅
- spread

spread: generally defined as 1.5x the Interquartile Range

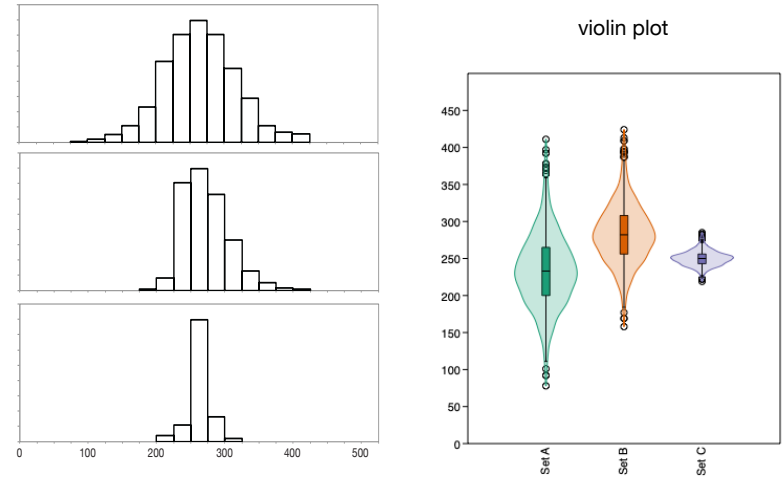


Even better: violin plots

Histograms and box and whisker plots assume a continuous data distribution: you do lose some information → problem for multi-modal datasets

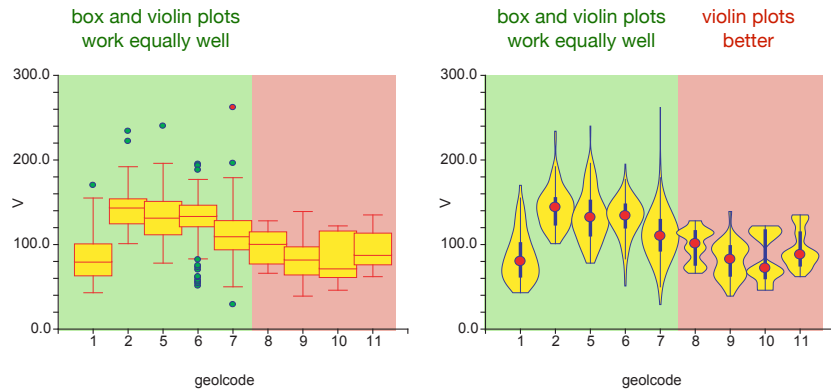


Graphical representation of data - comparisons



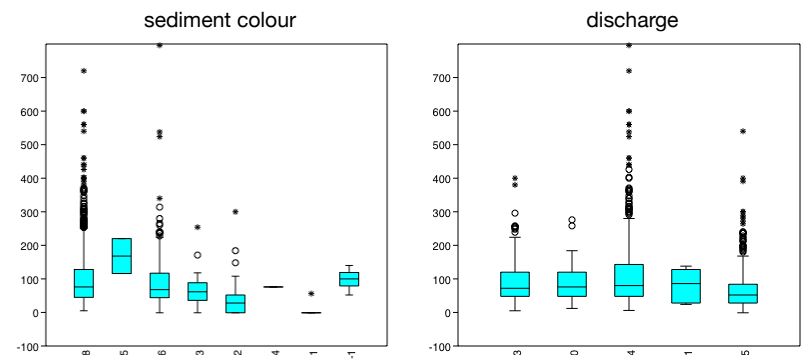
Even better: violin plots

Histograms and box and whisker plots assume a continuous data distribution: you do lose some information → problem for multi-modal datasets



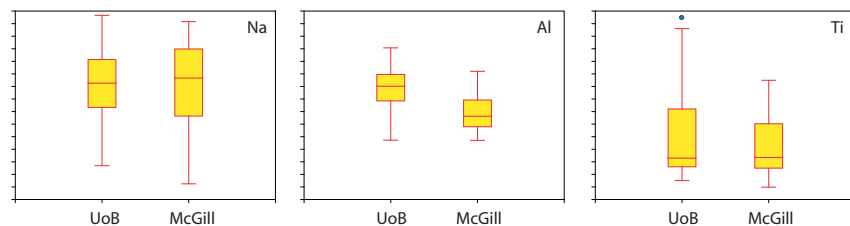
Compare by different criteria (grouping variable)

By defining a number of grouping variables you can use box plots to quickly see if any of these have significant control on your dataset:



Comparison of data sets - quality control

EMP data for a tourmaline crystal measured at different labs:



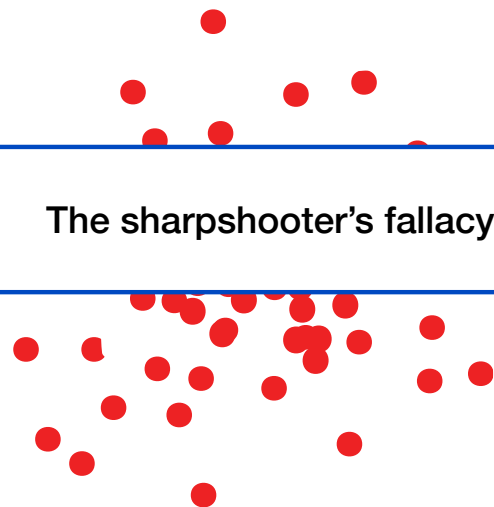
Systematic offset between the labs for Al: Which data are better? How to deal with this offset? Can it be corrected for? Etc...

Geotop Short Course in Data Analysis and Geostatistics Part 5. Precision, trueness and accuracy



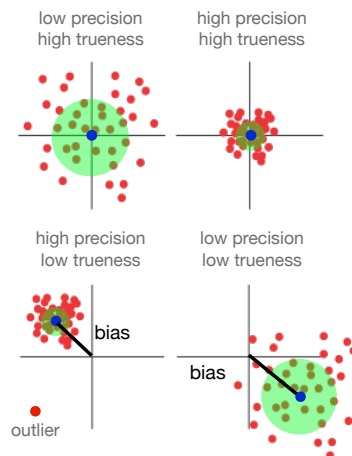
Wyatt Earp's stable door - what was he aiming for ?

The sharpshooter's fallacy



Statistical considerations

Not much liked, but very important in research to reach correct conclusions



Precision: related to *random error*. Reproducibility of analyses.

Trueness: closeness to the actual value: the opposite of *systematic error* or *bias*

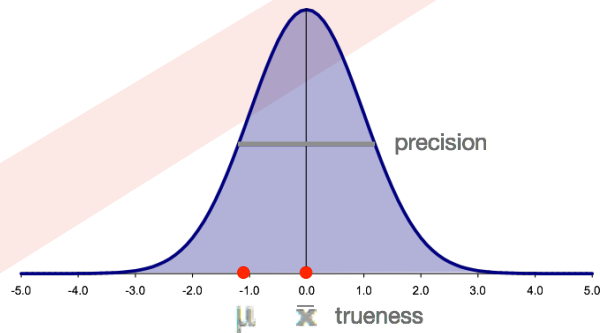
Outliers: individual measurement values which considerably differ from the mean value.

Trend: A data set shows a trend when the chronologically ordered values move steadily downwards or upwards

Gross errors: Gross errors result from human mistakes, or have their origins in instrumental or computational errors.

Statistical considerations

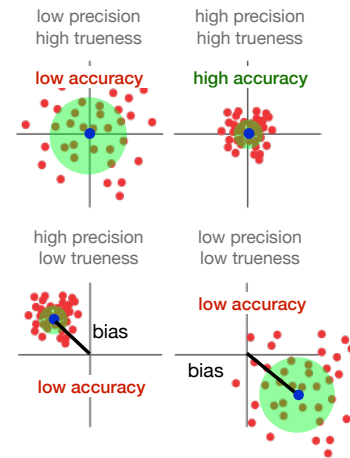
Or when represented for the univariate case:



low precision: large spread in the data, stdev is large (distribution wide and flat)
 low trueness: deviation in mean from true mean: bias

Statistical considerations

Not much liked but very important in research to reach correct conclusions

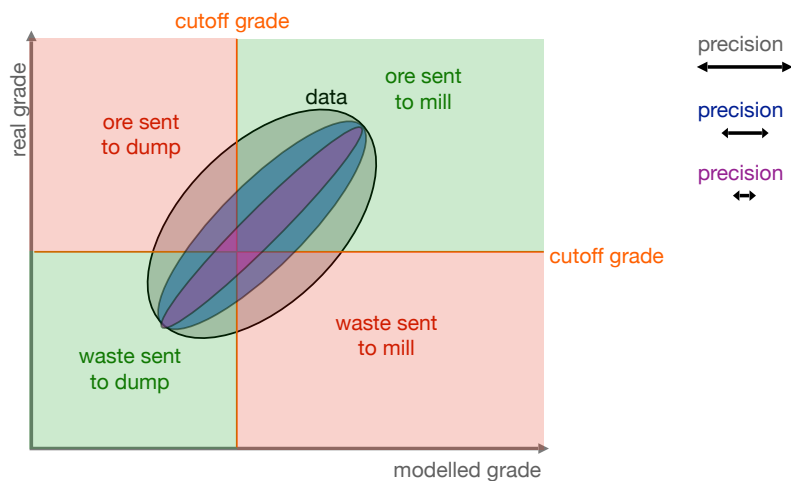


Precision: related to *random error*. Reproducibility of analysis.

Trueness: closeness to the actual value: the *opposite* of systematic error or bias

Accuracy: a combination of trueness and precision with good agreement between the measured value and its true value with a small uncertainty on the measured value

The need for accuracy and precision



How to determine trueness and precision

Precision Duplicates: repeat analyses of a sample (NOT a standard)
 If your precision is unacceptable: need to identify the source (field, lab, sample prep)
 If from the field: take more samples, if from sample prep: improve the method (avoid contamination), if from lab: better technique (pH meter instead of pH paper)

What is an unacceptable value? 100% uncertainty on 1 ppb is maybe not that problematic, because it still means the concentration is at the ppb-level

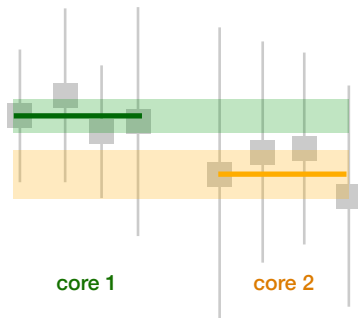
Trueness Analysis of reference materials (NOT standards)

A problem in trueness is introduced after sampling and can be related to sample prep contamination, incomplete digestion, standardization problems, etc.

If it is highly systematic and reproducible, it can be corrected for using a reference material (bias correction). Best practice is to hold back one or more reference materials to check that the corrected data are now "true"

Precision - standard of the mean

Uncertainty on individual points is larger than uncertainty on a group of points



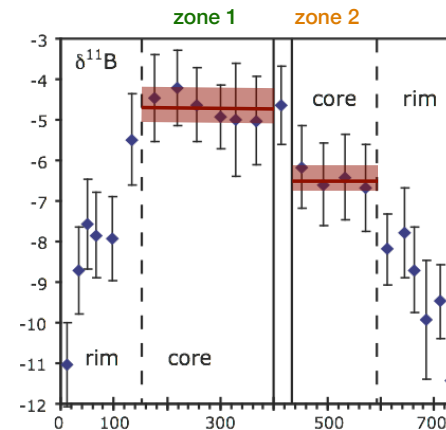
At first glance it would appear that the uncertainty for values in core 1 and core 2 prohibits differentiating between them

However, all values in one core are higher than in the other: this is no coincidence

Calculating the mean for each core and the associated uncertainty on this mean shows that they can be differentiated

Precision - standard of the mean

Uncertainty on individual points is larger than uncertainty on a group of points



We can also calculate an estimate for how certain we can be that these are really different, although we have to phrase this by the inverse:

probability that they are the same

Will do this on Monday

Precision - standard of the mean

The formula for standard error of the mean shows how a required precision for a reported result can be attained by repeating the analysis n times:

$$\text{standard error of the mean} \quad s_e^2 = s_x^2 / n \quad s_e = s_x / \sqrt{n}$$

If the precision of an analysis is 4% and if for the precision of the end result a value of 2% is required, then the analysis has to be repeated 4 times because:

$$s_e^2 = 2^2 = s_x^2 / n = 4^2 / 4 = 16 / 4 \rightarrow s_e = 2$$

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

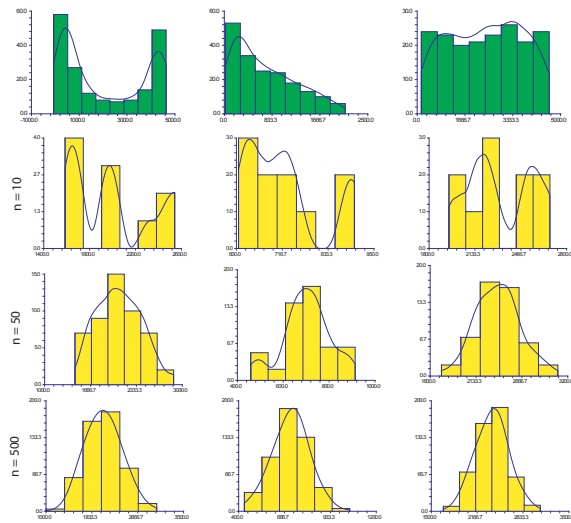
and this fit will improve with increasing number of geologists

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



The importance of precision: data rounding

A survey of MSc and PhD grad students at McGill gave the following results when asked how you decide how many significant digits you report (whether to report 0.05 or 0.053 or 0.0531):

1. this is fixed for a given instrument/type of data
2. this is specified by the journal I submit my data to
3. I would look this up by looking at a published data table
4. always use 2
5. this is free for me to choose
6. Excel sets this for me

So how do you decide this ?

precision

Data reporting: rounding

How you report values dictates their meaning, and specifies precision even if you do not report this.

- 5.41 means that you know that this value is between 5.40 and 5.42
- 5.4 means that you know that this value is between 5.3 and 5.5

Conversely, precision dictates significant values and choosing how many to use is straightforward and fixed:

- 10% stdev: 8.12 has to be reported as 8, because stdev ± 0.8 , but 0.12 would be reported as 0.12, because stdev ± 0.01

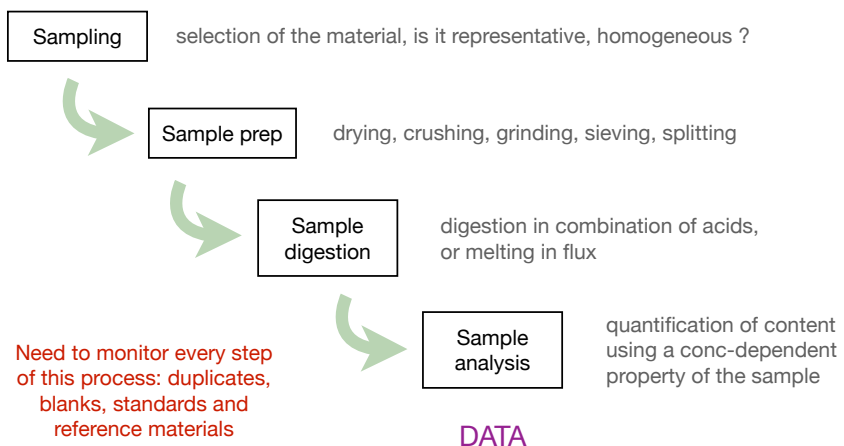
A separate rounding has to be determined for each value based on its precision

Geotop Short Course in Data Analysis and Geostatistics Part 6. Data quality assessment and control



Steps in data acquisition

SAMPLE



Monitoring the data acquisition process

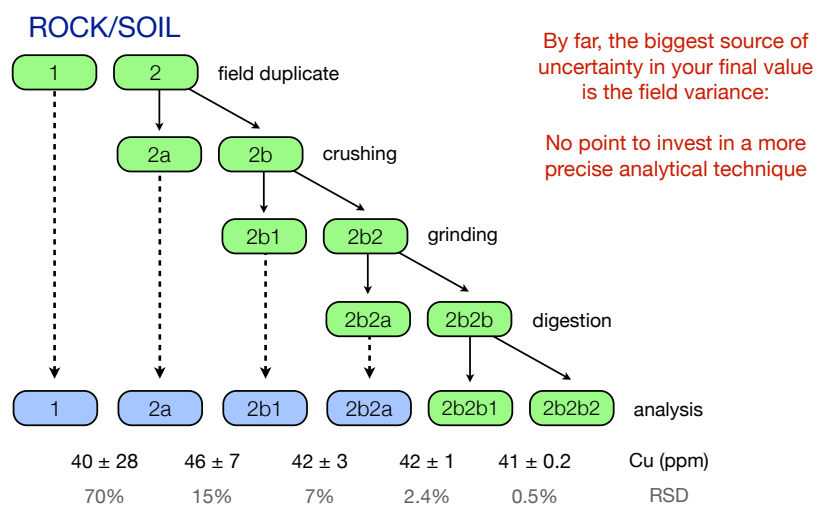
Duplicate	Two samples are taken from the same material using the identical methods, but not necessarily at the same time. A duplicate should be taken at each manipulation step.
Blank	A control sample that is passed through the complete procedure, or a subset of the procedure, to identify contamination and inter-sample carry-over. For rock samples this is often clean quartz.
Standard	A material of known concentration used to calibrate the analytical instrument that is used to obtain the data. Standards are normally supplied by the lab.
Reference material	A material of certified composition used to assess the trueness of analyses. This material should be similar to your sample material and provided by you. In fact, these samples are normally not disclosed to the lab beforehand.

Monitoring the data acquisition process

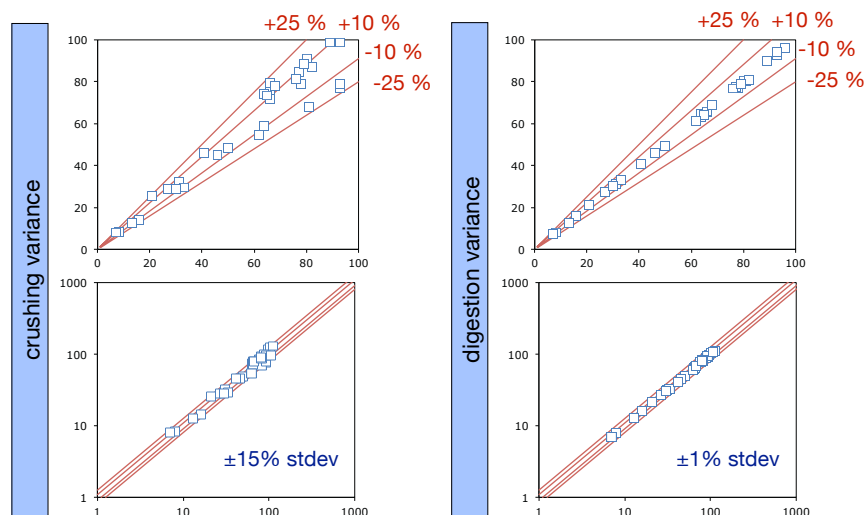
Duplicate A duplicate allows you to assess the spread in the data at each step of the data acquisition procedure. This is not error, but uncertainty on the data value. In fact, field variability is commonly the largest source of uncertainty, and this is an inherent sample property

Sampling	Field duplicate: two samples are taken from the same material, often a little apart. This is generally the largest variability
Sample prep	Duplicate after each step: crushing, grinding, splitting. The variability is expected to decrease in subsequent steps
Sample digestion	Lab duplicate: two aliquots of the same sample are independently digested and subsequently analysed.
Sample analysis	Analytical duplicate: a sample solution is analysed twice, but not sequentially. Duplicates should be interspersed with regular samples

Monitoring the data acquisition process



Monitoring the data acquisition process: duplicates



Monitoring the data acquisition process

Duplicate	Two samples are taken from the same material using the identical methods, but not necessarily at the same time. A duplicate should be taken at each manipulation step.
Blank	A control sample that is passed through the complete procedure, or a subset of the procedure, to identify contamination and inter-sample carry-over. For rock samples this is often clean quartz.
Standard	A material of known concentration used to calibrate the analytical instrument that is used to obtain the data. Standards are normally supplied by the lab.
Reference material	A material of certified composition used to assess the accuracy of analyses. This material should be similar to your sample material and provided by you. In fact, these samples are normally not disclosed to the lab beforehand.

Monitoring the data acquisition process: blanks

We can expect that duplicate variance decreases along the sample preparation procedure. If it does not, it could indicate a contamination issue: use blanks

Blank A blank is a sample of known composition (generally clean) that is used to identify and quantify contamination in the data acquisition procedure. Ideally, the blank behaves in the same way as your sample.

Blanks are often passed through the complete procedure, but if problems are found, blanks for each step are needed to locate the source



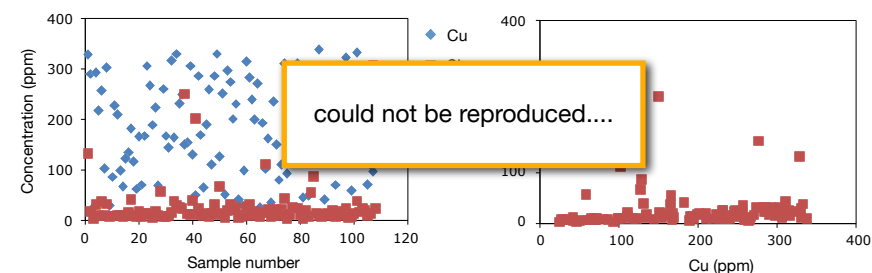
When using a ball mill, material can get stuck to the mill and get carried over to the next sample. This is especially problematic when going from a mineralised sample to a distant sample

image source: Wikipedia

Monitoring the data acquisition process: blanks

It can be very difficult to identify contamination and carry-over effects in your data: Need to include blanks in your procedure. Most labs routinely add blanks, but rarely provide the data: ask!

An exciting, but unexpected Antimony anomaly:



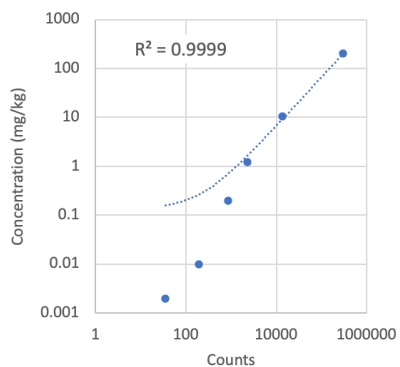
example courtesy of Gerben Mol

Monitoring the data acquisition process: standards

Standards are used to convert the raw output of an analytical instrument (often in counts) into concentrations by means of a calibration curve.

ICP-MS multi-standard calibration

Counts	Conc
35	0.002
192	0.01
837	0.2
2282	1.2
13260	10.4
300118	200.1

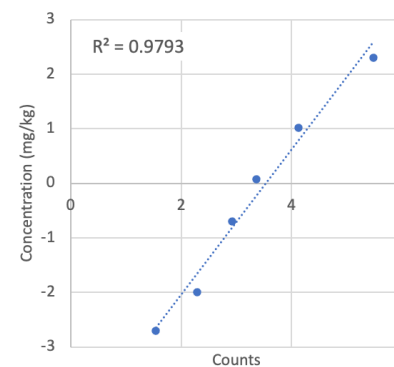


Monitoring the data acquisition process: standards

Standards are used to convert the raw output of an analytical instrument (often in counts) into concentrations by means of a calibration curve.

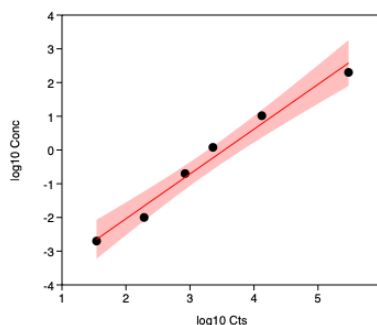
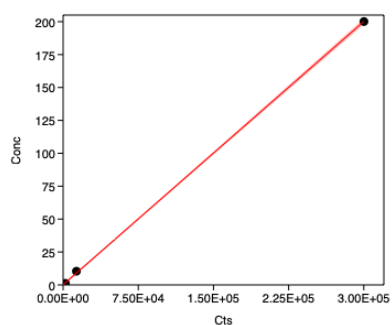
ICP-MS multi-standard calibration

Counts	Conc
35	0.002
192	0.01
837	0.2
2282	1.2
13260	10.4
300118	200.1



Monitoring the data acquisition process: standards

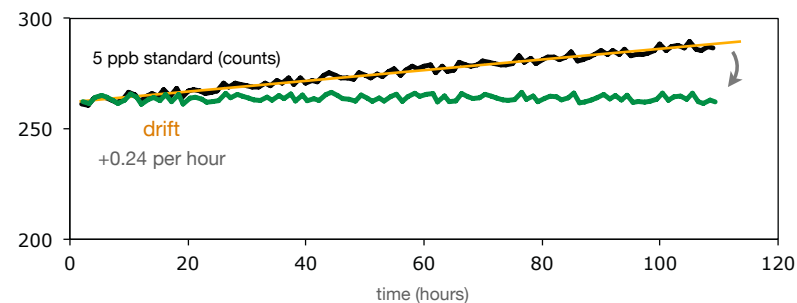
Standards are used to convert the raw output of an analytical instrument (often in counts) into concentrations by means of a calibration curve.



Monitoring the data acquisition process: standards

Analytical instruments drift over time: temperatures fluctuate during the day, power usage varies, instruments warm up, tubing degrades, detectors wear, etc etc etc

Requires monitoring of drift. Common practice: monitor the middle standard. If it deviates, a re-standardisation is applied. Can also correct post-analyses:

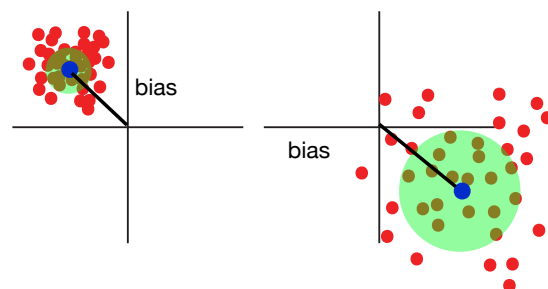


Monitoring the data acquisition process

Duplicate	Two samples are taken from the same material using the identical methods, but not necessarily at the same time. A duplicate should be taken at each manipulation step.
Blank	A control sample that is passed through the complete procedure, or a subset of the procedure, to identify contamination and inter-sample carry-over. For rock samples this is often clean quartz.
Standard	A material of known concentration used to calibrate the analytical instrument that is used to obtain the data. Standards are normally supplied by the lab.
Reference material	A material of certified composition used to assess the accuracy of analyses. This material should be similar to your sample material and provided by you. In fact, these samples are normally not disclosed to the lab beforehand.

Monitoring the data acquisition process: SRMs

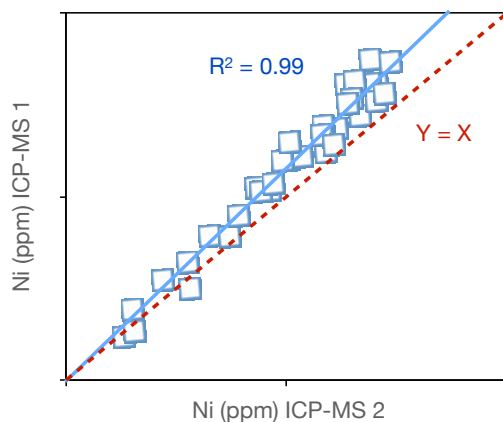
Data values are determined by comparing counts on an unknown - the sample, against the calibration curve as obtained from standards. We make the inherent assumption that the calibration curve is correct. **Needs to be verified: SRMs**



How do we know the correct value, i.e. the trueness? **Standard Reference Materials**

Comparing analytical techniques

We commonly obtain data from multiple instruments, either deliberately to check results, or because the lab changed over to a new instrument



Conclusion:
great agreement between the data with a R^2 of 0.99!

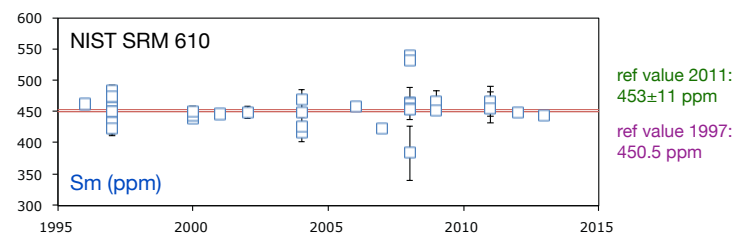
In reality:
big problem in the dataset with a clear bias: ICP-MS 1 is always too high!

Not regression fit that is needed here as a check, but goodness-of-fit to the model $Y = X$

Monitoring the data acquisition process: SRMs

A SRM is a material, either natural or manufactured, of which composition is known, most commonly from analyses in a variety of different certified labs using a diversity of analytical methods and instruments.

- SRMs are generally only certified for a number of elements
- Compositions can change as more analyses become available

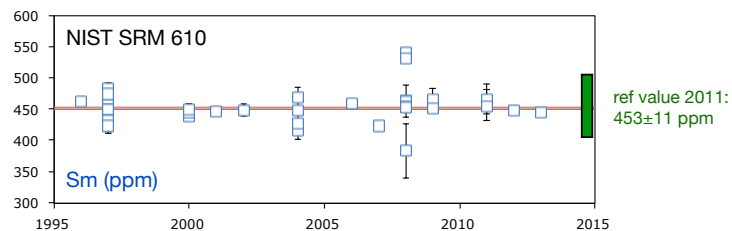


- Data depositories of SRM values are a great resource: GeoREM website: <http://georem.mpch-mainz.gwdg.de>

Monitoring the data acquisition process: SRMs

A SRM is a material, either natural or manufactured, of which composition is known, most commonly from analyses in a variety of different certified labs using a diversity of analytical methods and instruments.

- SRM concentrations have an associated uncertainty: can **never** obtain a trueness greater than the uncertainty on the SRM value. However, you can achieve a precision that is better.



- SRMs are not always homogeneous: can receive a bad batch

Monitoring the data acquisition process: SRMs

A SRM is a material, either natural or manufactured, of which composition is known, most commonly from analyses in a variety of different certified labs using a diversity of analytical methods and instruments.

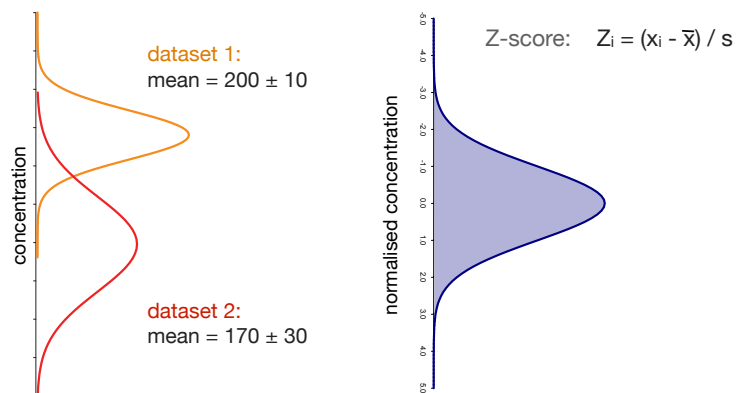
- SRMs should be as similar as possible to your sample material
- SRM should also have a similar concentration range. In most cases you need more than 1 -> choose them to cover your sample's range
- SRM allow for assessment of trueness, but also bias correction



- SRMs have a limited shelf life, and may settle during transport

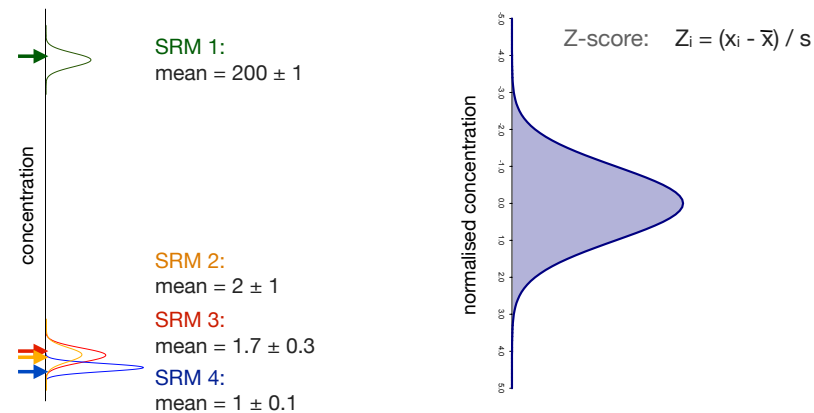
Data levelling using SRMs

If the same SRMs have been measured in multiple datasets, you can level these data perfectly, because these are the same samples. Moreover, their data should have a normal distribution: can use Z-scores for levelling:



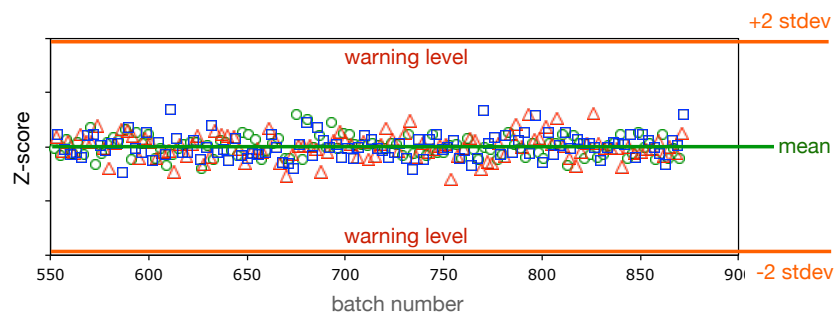
Quality control using multiple SRMs

SRMs should cover the compositional range in your samples and this means that it can be a challenge to visually show all SRMs in one time series. Could log-transform but there is a better way: plot Z-scores



Monitoring the data acquisition process: SRMs

Identifying problems with the accuracy of your data:



An elegant way to check all your SRMs at the same time, is to plot the Z-score of each value: this scales SRMs with different absolute concentrations and stdev

Monitoring the data acquisition process - example

Dataset of mineral analyses with duplicates and a large set of reference materials.



Re-analysed the same standard ~every 50 points: **monitor of drift**

Analysed 6 SRMs ~every 50 points: **monitor of accuracy, and any compositional dependence**

Every time a standard or SRM was measured, it was measured twice: **analytical uncertainty**

Analysed for SiO₂, TiO₂, Al₂O₃, MgO, FeO, CaO, Na₂O, and total REE.

Precision of standardisation is ± 2% for all elements except the REE at ± 35% relative

Monitoring the data acquisition process - example

First determine the analytical uncertainty and reduce the duplicates to one value: they represent two estimates of the value for a given sample and should not be treated as two samples !

sample no	SiO ₂	TiO ₂	(sample1 - sample2) ²		average SiO ₂
			SiO ₂	TiO ₂	
109	79.46	0.11	0.325	0.0004	78.95 wt%
110	78.89	0.13			root (sum (deviations))
					1.37
131	78.82	0.10	0.000	0.0018	RSD SiO ₂
132	78.81	0.06			1.37 * 100% / 78.95 = 1.7%
149	78.68	0.05	0.017	0.0022	
150	78.81	0.09			
189	79.40	0.14	0.068	0.0046	RSD TiO ₂
190	79.66	0.07			0.10 * 100% / 0.09 = 106%
211	77.87	0.08	1.464	0.0006	
218	79.08	0.10			
			sum	1.87	0.010

Monitoring the data acquisition process - example

First determine the analytical uncertainty and reduce the duplicates to one value: they represent two estimates of the value for a given sample and should not be treated as two samples !

sample no	CaO	(sample1 - sample2) ²		average CaO
		CaO		
109	0.10	0.000		0.14 wt%
110	0.10			root (sum (deviations))
				0.45
131	0.15	0.005		RSD CaO
132	0.08			0.45 * 100% / 0.14 = 313%
149	0.08	0.000		
150	0.08			
189	0.52	0.194		
190	0.08			
211	0.11	0.000		
218	0.12			
		sum	0.20	

looks like an outlier !

Outliers

Outliers are a relatively common occurrence in geo-data. They refer to values that are **extreme relative** to the other data.

They can be the result of analytical problems or mistakes. However, they can equally be an extreme value that occurs purely by chance.

What is the procedure to deal with a suspected outlier ?

If you have reasons to suspect the data value (you spilled some of the sample, the instrument acted up, etc): **do not use it**

If you have no reason to suspect the datapoint: **repeat the analysis and see if the value can be reproduced**

If this is impossible: **test if the data value can belong to data set made up of your other analyses using a statistical test**

Outlier identification when spread unknown

If you do not have a measure of the spread in your data, which often happens if you have a limited set of values and can therefore not calculate a good estimate of stdev, you can use the Dixon Q-test to identify outliers:

$Q = (x_n - x_{n-1}) / (x_n - x_1)$, where:

x_n your suspected outlier
 x_{n-1} its nearest value
 x_1 the value furthest away

Critical value of Q		
n	$\alpha = 0.10$	$\alpha = 0.05$
3	0.89	0.94
4	0.68	0.77
5	0.56	0.64
6	0.48	0.56
7	0.43	0.51

In the test: if $Q_{\text{calc}} > Q_{\text{critical}}$

we identify the value as an outlier

alpha = confidence level

Assumption: normal distribution

Outlier identification when spread unknown

If you do not have a measure of the spread in your data, which often happens if you have a limited set of values and can therefore not calculate a good estimate of stdev, you can use the Dixon Q-test to identify outliers:

$Q = (x_n - x_{n-1}) / (x_n - x_1)$, where:

x_n your suspected outlier
 x_{n-1} its nearest value
 x_1 the value furthest away

In the test: if $Q_{\text{calc}} > Q_{\text{critical}}$

we identify the value as an outlier

n	$\alpha = 0.10$
3	0.89
4	0.68
5	0.56
6	0.48
7	0.43

sample	var(CaO)
109	0.000
110	
131	0.005
132	
149	0.000
150	
189	0.194
190	
211	0.000
218	

0.194 your suspected outlier
 0.005 its nearest value
 0.000 the value furthest away

$Q_{\text{calc}} = 0.974 \rightarrow n = 5 \rightarrow Q_{\text{crit}, \alpha=0.10} = 0.56 \rightarrow$ **outlier**

Outlier identification when spread known

If you have a lot of other duplicates, you can estimate the spread in the distribution to which your duplicates belong by calculating the stdev: **the more duplicates, the closer the sample stdev is to the population stdev**

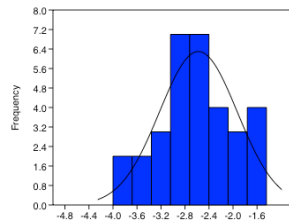
Assuming that you have a normal distribution, and that you have sufficient duplicates to obtain a good estimate of mean and standard deviation, you can use the probabilities of the normal distribution to calculate the chance that your suspected outlier is part of the distribution

this takes us back to Z-scores

Monitoring the data acquisition process - example

Steps in a Z-score approach to outlier identification

Step 1. Determine the data distribution for the duplicate deviation for CaO



The data distribution is lognormal:
will use a log-transform of the data

Step 2. Select the confidence interval that you are going to use

Confidence level of 95%: if the value has a probability of occurrence of less than 5% -> classify it as an outlier

Monitoring the data acquisition process - example

Cannot just discard a datapoint: have to show that it is an outlier.

Step 3. Calculate the Z-score for the offending duplicate pair

sample no	CaO	(sample1 - sample2) ² CaO
109	0.10	0.000
110	0.10	
131	0.15	0.005
132	0.08	
149	0.08	0.000
150	0.08	
189	0.52	0.194
190	0.08	
211	0.11	0.000
218	0.12	

$$\log(0.194) = -0.71$$

distribution (logged):

$$\text{mean} = -2.58$$

$$\text{stdev} = 0.65$$

$$\text{Z-score} = (-0.71 - -2.58) / 0.65 = 2.88$$

Monitoring the data acquisition process - example

Table 1: Probability values for the standardized normal distribution.

Z	Two-sided	One-sided	Z	Two-sided	One-sided	Z	Two-sided	One-sided
0.00	1.000	0.500	1.61	0.107	0.054	2.41	0.016	0.008
0.05	0.960	0.480	1.62	0.105	0.053	2.42	0.016	0.008
0.10	0.920	0.460	1.63	0.103	0.052	2.43	0.015	0.008
0.15	0.881	0.440	1.64	0.101	0.051	2.44	0.015	0.007
0.20	0.842	0.421	1.65	0.099	0.050	2.45	0.014	0.007
0.25	0.803	0.401	1.66	0.097	0.049	2.46	0.014	0.007
0.30	0.764	0.382	1.67	0.095	0.048	2.47	0.014	0.007
0.35	0.726	0.363	1.68	0.093	0.047	2.48	0.013	0.007
0.40	0.689	0.345	1.69	0.091	0.046	2.49	0.013	0.006
0.45	0.653	0.326	1.70	0.089	0.045	2.50	0.012	0.006
0.50	0.617	0.309	1.71	0.087	0.044	2.51	0.012	0.006
0.55	0.582	0.291	1.72	0.085	0.043	2.52	0.012	0.006
0.60	0.549	0.274	1.73	0.084	0.042	2.53	0.011	0.006
0.65	0.516	0.258	1.74	0.082	0.041	2.54	0.011	0.006
0.70	0.484	0.242	1.75	0.080	0.040	2.55	0.011	0.005
0.75	0.453	0.227	1.76	0.078	0.039	2.56	0.011	0.005
0.80	0.424	0.212	1.77	0.077	0.038	2.57	0.010	0.005
0.85	0.395	0.198	1.78	0.075	0.038	2.58	0.010	0.005
0.90	0.368	0.184	1.79	0.074	0.037	2.59	0.010	0.005
0.95	0.342	0.171	1.80	0.072	0.036	2.60	0.009	0.005
1.00	0.317	0.159	1.81	0.070	0.035	2.61	0.009	0.005
1.01	0.313	0.156	1.82	0.069	0.034	2.62	0.009	0.005
1.02	0.308	0.154	1.83	0.067	0.034	2.63	0.009	0.005
1.03	0.303	0.152	1.84	0.066	0.033	2.64	0.009	0.005
1.04	0.298	0.149	1.85	0.064	0.032	2.65	0.009	0.005
1.05	0.294	0.147	1.86	0.063	0.031	2.66	0.008	0.004
1.27	0.204	0.102	2.08	0.038	0.019	2.88	0.004	0.002

$$\text{Z-score} = 2.88$$

$$p = 0.002 \rightarrow 0.2\%$$

Monitoring the data acquisition process - example

Cannot just discard a datapoint: have to show that it is an outlier.

Step 3. Calculate the Z-score for the offending duplicate pair

sample no	CaO	(sample1 - sample2) ² CaO
109	0.10	0.000
110	0.10	
131	0.15	0.005
132	0.08	
149	0.08	0.000
150	0.08	
189	0.52	0.194
190	0.08	
211	0.11	0.000
218	0.12	

$$\log(0.194) = -0.71$$

distribution (logged):

$$\text{mean} = -2.58$$

$$\text{stdev} = 0.65$$

$$\text{Z-score} = (-0.71 - -2.58) / 0.65 = 2.88$$

probability of occurrence is 0.2%

prob < 5%: is indeed an outlier

Monitoring the data acquisition process - example

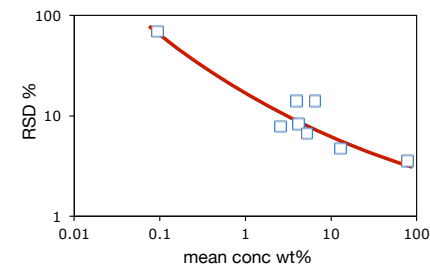
First determine the analytical uncertainty and reduce the duplicates to one value: they represent two estimates of the value for a given sample and should not be treated as two samples !

sample no	CaO	(sample1 - sample2) ² CaO	average CaO
109	0.10	0.000	0.11 wt% ← also changes
110	0.10		root (sum (deviations)) 0.07
131	0.15	0.005	RSD CaO $0.45 * 100\% / 0.14 = 313\%$
132	0.08		RSD CaO $0.07 * 100\% / 0.11 = 63\%$
149	0.08	0.000	
150	0.08		
189	0.52	0.194	
190	0.08		
211	0.11	0.000	
218	0.12		
	sum	0.005	

Monitoring the data acquisition process - example

First determine the analytical uncertainty and reduce the duplicates to one value: they represent two estimates of the value for a given sample and should not be treated as two samples !

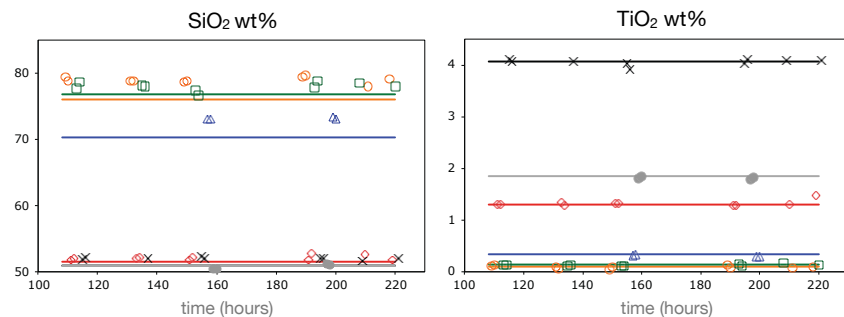
	RSD %	mean conc wt%
SiO ₂	3.5	78.95
TiO ₂	67	0.09
Al ₂ O ₃	4.8	13.1
MgO	8.3	4.2
FeO	14	6.4
CaO	6.7	5.3
Na ₂ O	14	3.9
REE	7.7	2.5



low concentrations have higher uncertainty

Monitoring the data acquisition process - example

We can now check the trueness using 6 SRMs that were measured for this dataset



SiO₂ is consistently overestimated: **bias**

TiO₂ is spot-on !

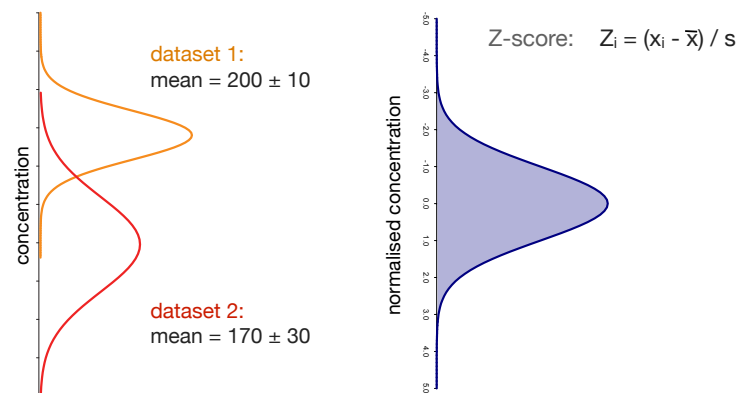
To obtain the final corrected dataset: would shift the SiO₂ concentrations to match the certified values for the SRMs

Geotop Short Course in Data Analysis and Geostatistics Part 7. Combining and levelling datasets

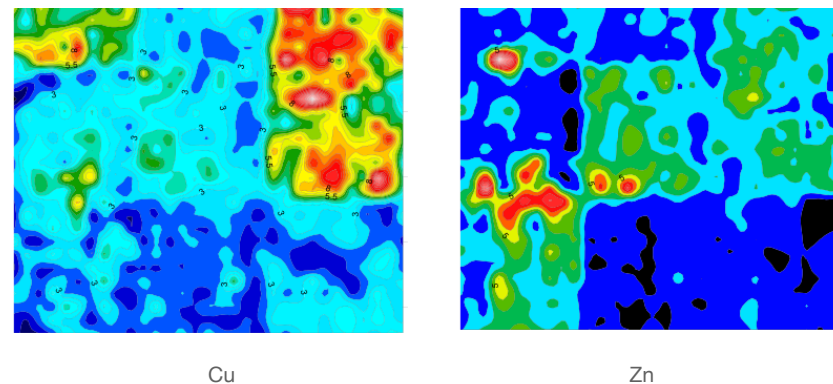


Data levelling using SRMs

If the same SRMs have been measured in multiple datasets, you can level these data perfectly, because these are the same samples. Moreover, their data should have a normal distribution: can use Z-scores for levelling:



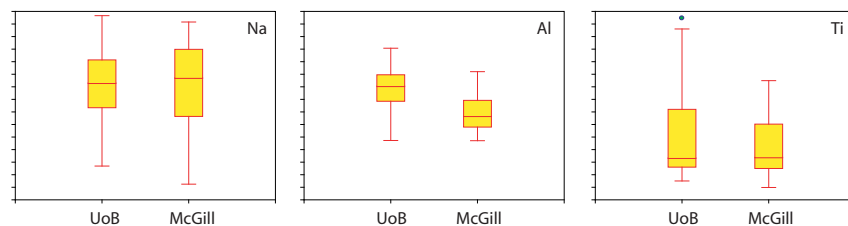
Data levelling



Data levelling

It is very common that you need to combine datasets. However, samples may have been prepared differently and analysed by different techniques in different labs, leading to each set having a different data distribution, mean/median and spread.

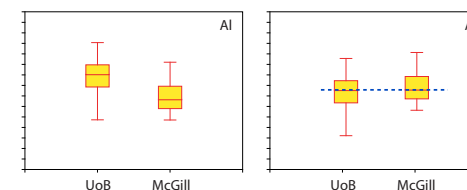
This can introduce spurious anomalies into your data: **data need to be levelled first**



Data levelling

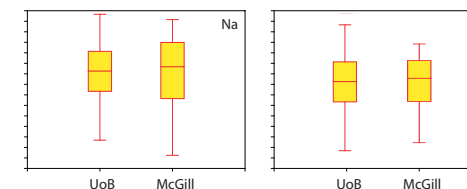
- shift to same mean or median, or ratio to the mean or median (data spread remains different)

median = robust, whereas mean is affected by outliers



- normalize using Z-score (both value and spread are matched between datasets)

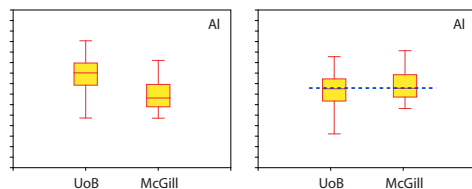
a robust equivalent also exists using the median and mean-average-deviation (robust Z-score levelling) or using ranks instead of data (Gauss levelling)



Data levelling - mean or median shift

- shift to same mean or median, or ratio to the mean or median (data spread remains different)

median = robust, whereas mean is affected by outliers

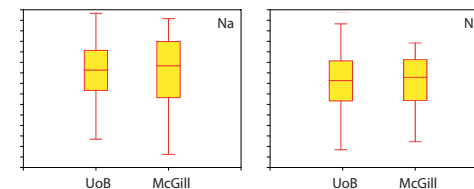


Data:	UoB	McGill	level to mean		level to median		levelled to UoB	
			UoB	McGill	UoB	McGill	mean	median
	10.2	4.3	2.1	-1.0	2.3	-0.9	7.1	7.0
	8.4	5.8	0.3	0.5	0.5	0.6	8.6	8.5
	6.7	5.2	-1.4	-0.1	-1.2	0.0	8.0	7.9

mean	8.1	5.3	x - mean		x - median			
median	7.9	5.2						

Data levelling - Z-score levelling

- normalize using Z-score (both value and spread are matched between datasets)



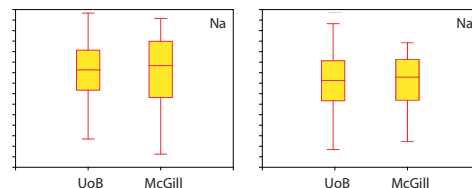
Data:	UoB	McGill	Z-score level		levelled to UoB	
			UoB	McGill	UoB	McGill
	102	45	-0.23	-0.88	102	79
	94	158	-0.46	0.54	94	129
	125	68	0.43	-0.59	125	89

mean	110	115	0	0	110	110
stdev	35	80	1	1	35	35

$$Z_i = (x_i - \mu) / \sigma$$

Data levelling - robust Z-score levelling

- normalize using Z-scores calculated from the median and MAD which are robust alternatives to mean and stdev



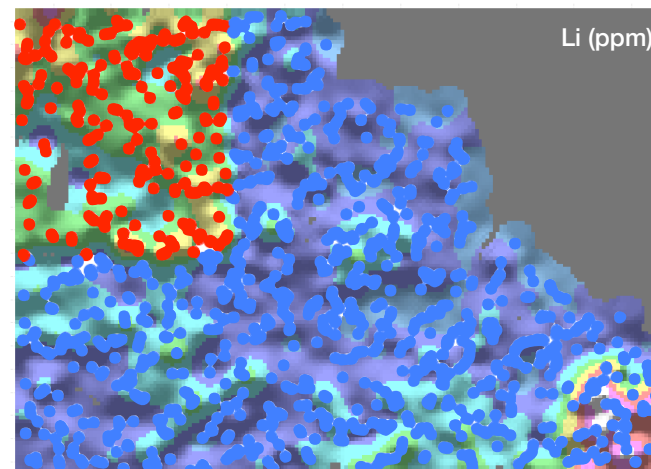
Data:	UoB	McGill	robust Z-score level		levelled to UoB	
			UoB	McGill	UoB	McGill
	102	45	-0.15	-0.92	102	87
	94	158	-0.55	0.97	94	124
	125	68	1.00	-0.53	125	94

median	105	100	0	0	105	105
MAD	20	60	1	1	20	20

$$Z_i = (x_i - \text{med}) / \text{MAD}$$

Data levelling

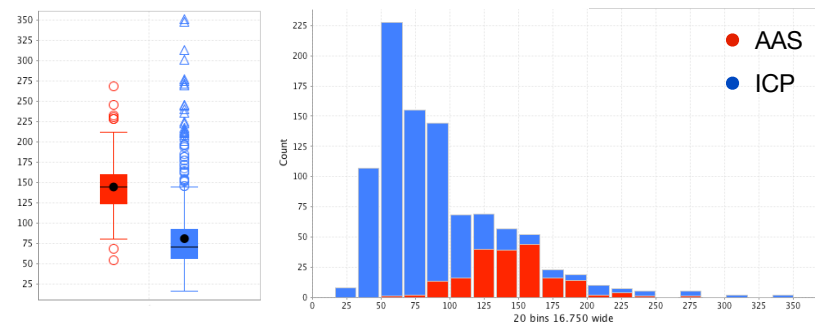
When mixing data sources: have to make sure they fit together



- AAS
- ICP

Data levelling

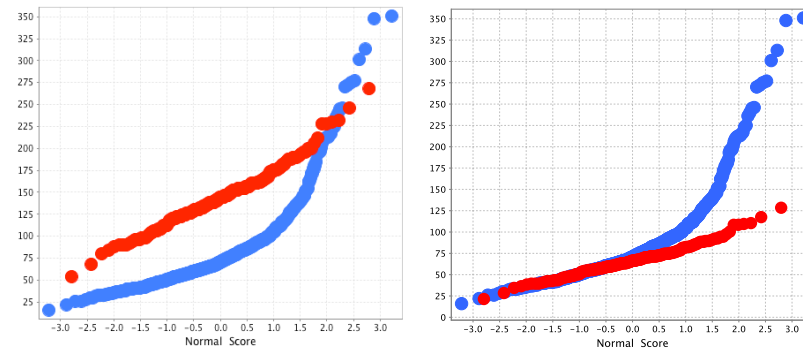
The two datasets are clearly different, both in concentration and in their data distribution: they do not sample the same geology in the same proportion!



Need smart data levelling that deals with this, as well as with variations in the characteristics (e.g. stdev) of each technique

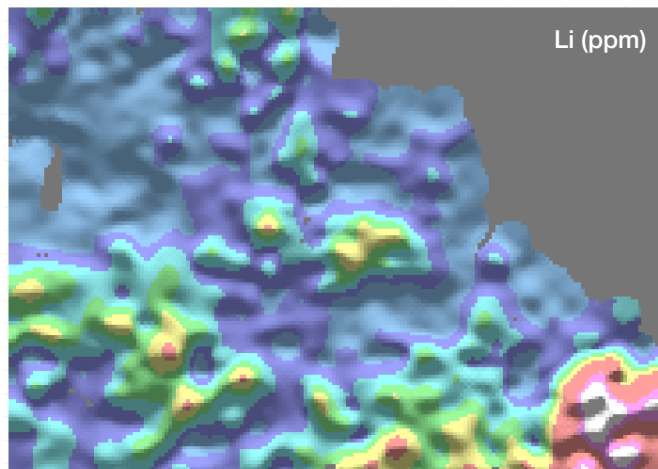
Data levelling

This is the data levelling result for the Li data using robust Z-score levelling. The sets now overlap nicely, but their markedly different distribution has been preserved



Data levelling

When done right, the datasets fit together smoothly and you can interpret them together



Geotop Short Course in Data Analysis and Geostatistics
Part 7. Bivariate data analysis



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

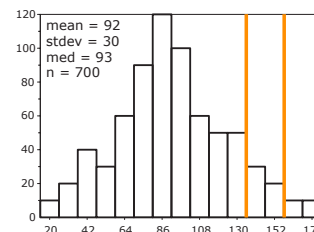
Correlation analysis; how to characterize correlations in numerical and non-numerical data, quantify the “degree of correlation”, and how to test if correlations are real

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

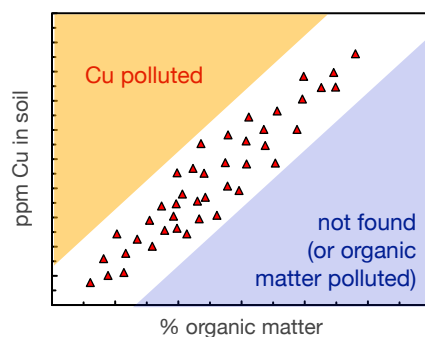
however; some samples were from heavily polluted sites, so why don't they jump out in the total data set?

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:

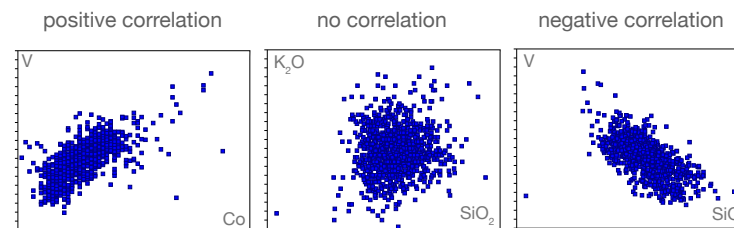


any soil with high organic matter content will have a natural enrichment

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)

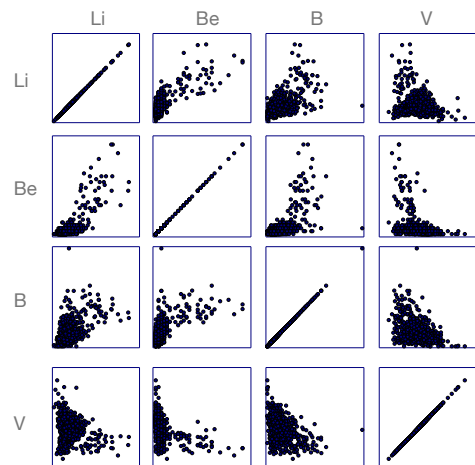


high Co is associated with high V
same source or process

high SiO₂ 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

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

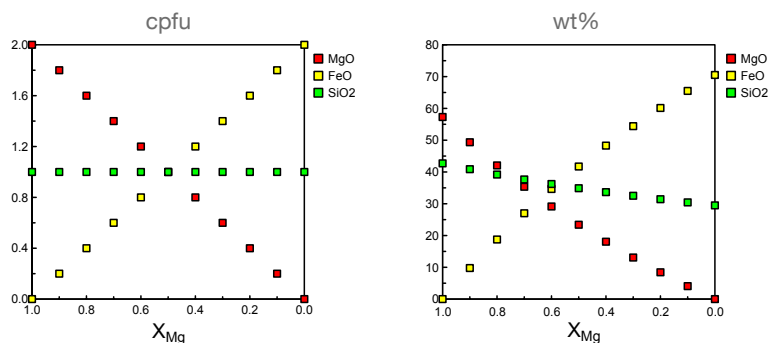
closure: when one element goes up, the others have to go down to satisfy a 100% sum

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 by leaching

Acid leaching results in removal of all elements except SiO₂:

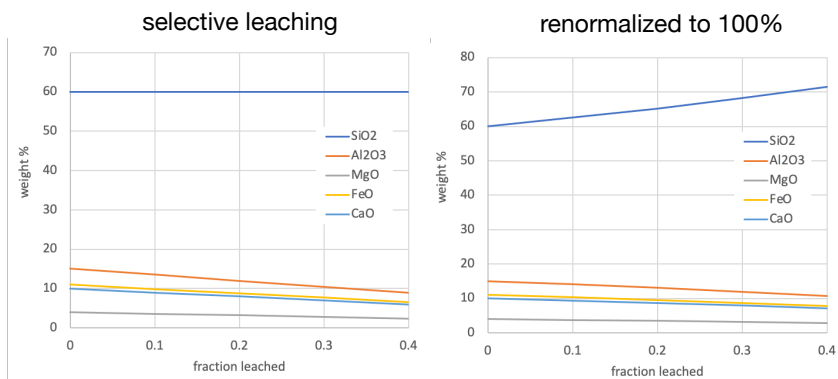
	wt%	wt%	wt%	wt%
SiO ₂	60	60	62.5	65.2
Al ₂ O ₃	15	13.5	14.1	13.0
MgO	4	3.6	3.8	3.5
FeO	11	9.9	10.3	9.6
CaO	10	9	9.4	8.7
leaching	0%	10%	10%	20%
			30%	40%
			68.2	71.4
			11.9	10.7
			3.2	2.9
			8.8	7.9
			8.0	7.1

re-norm
to 100%

Results in residual enrichment and artificial correlations



Closure by leaching

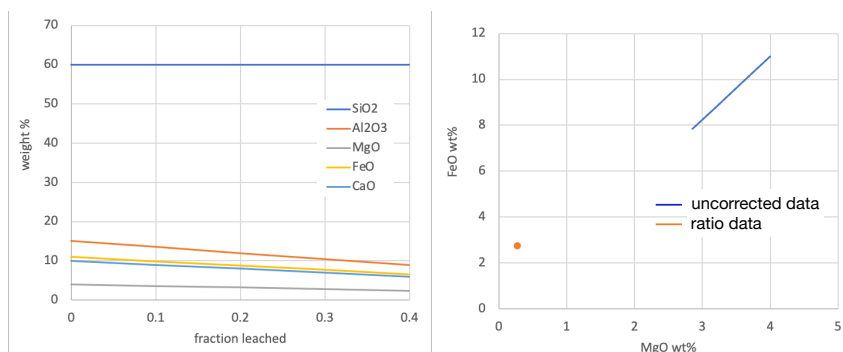


SiO₂, which is immobile, appears to be progressively added

Closure by leaching

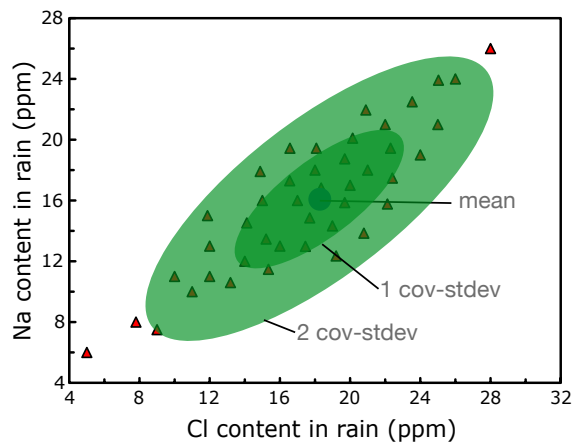
normalize data to an immobile element, in this case SiO₂ but more commonly Zr, Hf, Ti, etc

use ratios: both variables are affected by closure, but this cancels out in a ratio



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:

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

normal variance

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:

$$r = \frac{\text{COV}_{xy}}{S_x S_y}$$

when covariance close to variance: $r \rightarrow 1$

when variance \gg covariance: $r \rightarrow 0$

So what do values of r mean;

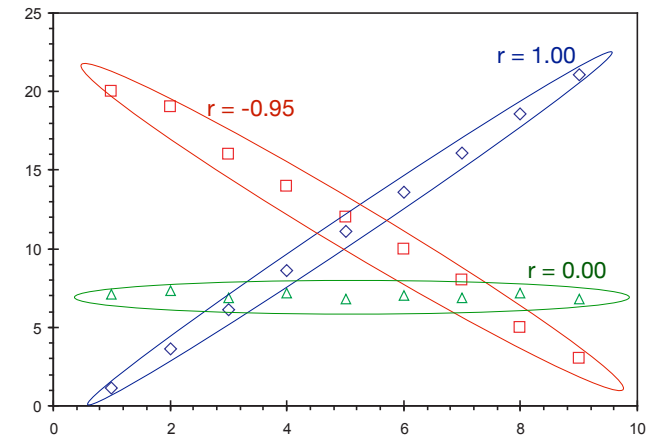
$r = -1$ perfect negative correlation between variables

$r = +1$ perfect positive correlation between variables

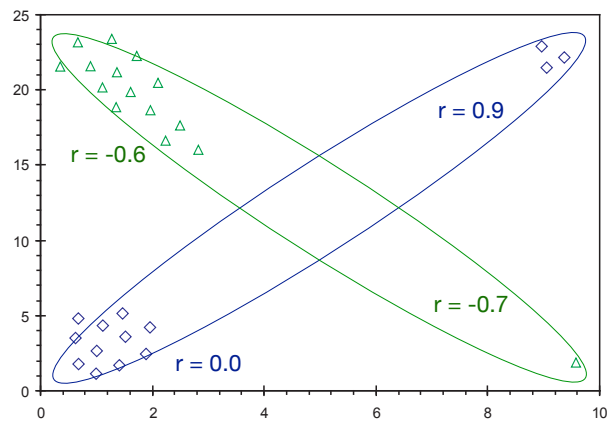
$r = 0$ no correlation: the variables are independent

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

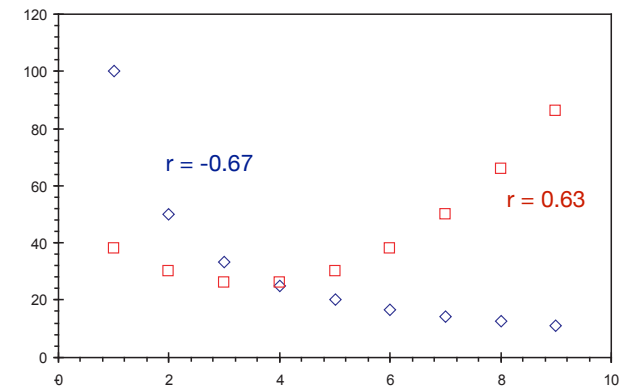
Examples of correlations - the good



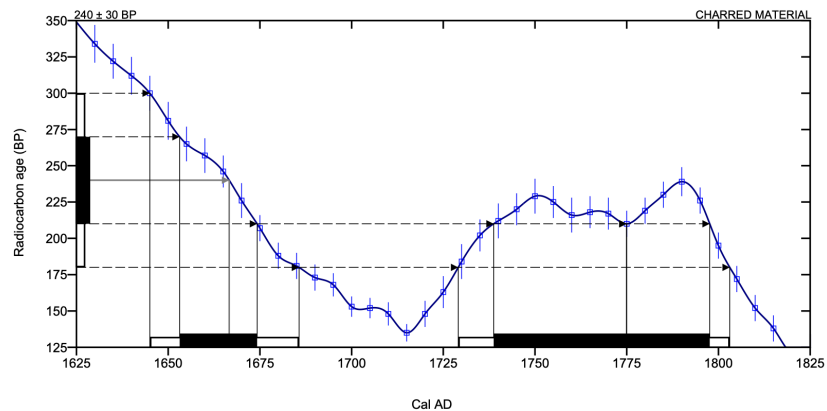
Examples of correlations - the bad



Examples of correlations - the ugly

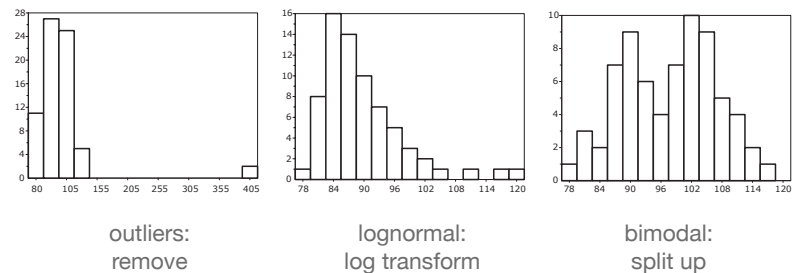


Examples of correlations - the ugly



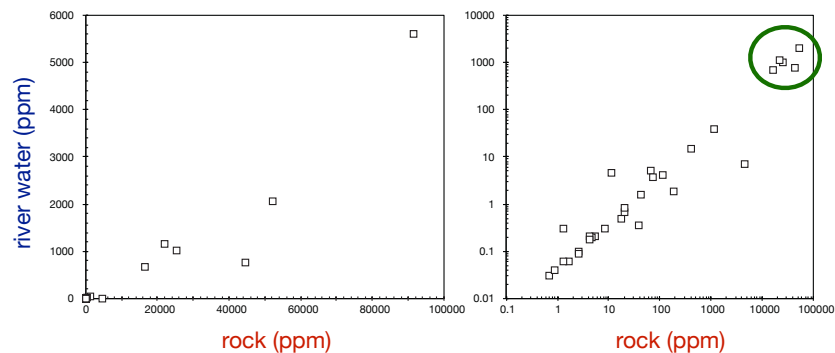
The correlation coefficient - data distribution

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



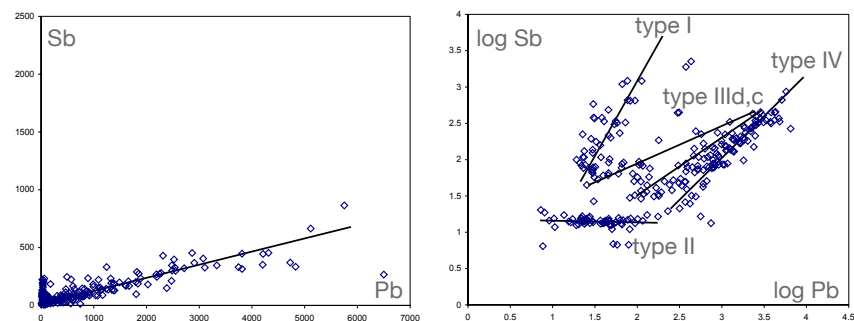
Log-normal transformation

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:

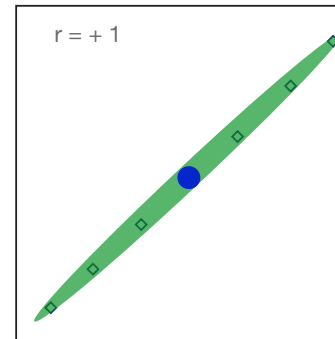
- Reciprocal: $1/x$
- Square root: \sqrt{x}
- Angular transformation: $\sin(x)$

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

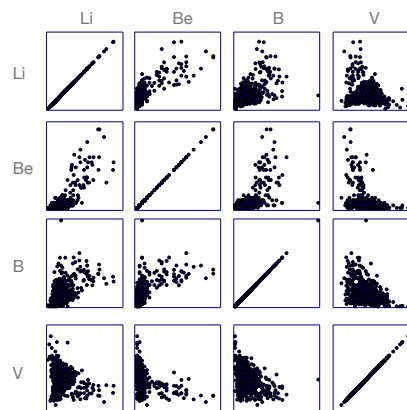
The importance of meeting method prerequisites

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	0.0	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	
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	
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	0.2	
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	0.4	
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	-0.1	
log_Mn	-0.5	0.2	0.5	0.3	0.3	0.0	1	0.2	0.1	-0.1	-0.1	0.3	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	
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	0.0	
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	
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	
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	0.0	
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	0.5	
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	
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	
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	
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	

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	0.0	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	
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	
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	0.2	
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	0.4	
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	-0.1	
log_Mn	-0.5	0.2	0.5	0.3	0.3	0.0	1	0.2	0.1	-0.1	-0.1	0.3	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	
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	0.0	
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	
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	
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	0.0	
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	0.5	
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	
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	
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	
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	

Statistical testing: the student-t test of r

What values of r are meaningful for a given confidence level

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

When calculated $t >$ critical t
significant correlation


t depends on the number of samples and the desired confidence interval

- ▶ the more samples, the smaller the uncertainty on your r-value
less uncertainty on deciding whether something is significant
- ▶ the confidence level governs how strong your statements will be:
95% - wrong conclusion in 1 out of 20 cases
98% - wrong in 1 out of 50 cases

Statistical testing: the student-t test of r

So in the case of our correlation analysis:

cannot test the presence of correlation but we can test for the **absence** of correlation between the variables:

$r = 0$  reject, $r \neq 0$, so there **is** a correlation between the vars
accept, at this confidence interval there is no significant correlation between the variables

hypotheses: H_0 : hypothesis to be tested $r = 0$
 H_a : alternative hypothesis $r \neq 0$

In most cases you will be testing the negative conclusion; there is no correlation, there is no difference between two groups, etc.

Statistical testing: the student-t test of r

an example of significance testing of the correlation coefficient:

Our hypotheses: H_0 : $r = 0$, if true, no significant correlation
 H_a $r \neq 0$, cannot reject the absence of correlation

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

Let's say:
 $r = -0.34$
 $n = 25$

$t_{\text{calc}} = -1.73$
 $t_{\text{critical}} = -1.71$

t_{calc} exceeds t_{critical} -> reject H_0

in this example we can reject the H_0 : so we can make the strong statement that at 95% confidence, there is a significant correlation between the vars

what if we want to be more certain ?

Statistical testing: the student-t test of r

an example of significance testing of the correlation coefficient:

Our hypotheses: H_0 : $r = 0$, if true, no significant correlation
 H_a $r \neq 0$, cannot reject the absence of correlation

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

Let's say:
 $r = -0.34$
 $n = 25$

$t_{\text{calc}} = -1.73$
 $t_{\text{critical}} = -1.71$ at 95%
 $t_{\text{critical}} = -2.07$ at 97.5%

t_{calc} does not exceed t_{critical} ->
we **cannot** reject H_0

we can now only conclude that we have no reason to reject the absence of correlation, which is clearly not as strong a statement

Have entered the field of statistical testing....

Day 2 - topics covered



- Visual data comparison (box-and-whiskers plots, violin plots)
- Precision, trueness and accuracy
- QA/QC of a sample and data collection campaign (duplicates, blanks, standards and reference materials)
- Levelling in combining data sources
- covariance and correlation
- the correlation coefficient

