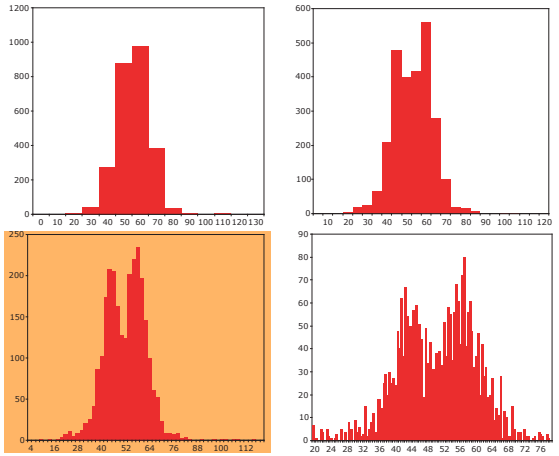
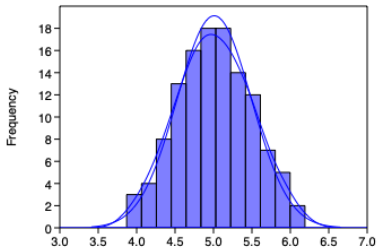
Review of practical exercises - 1.1

dependence of histograms on choice of classes - there are no rules

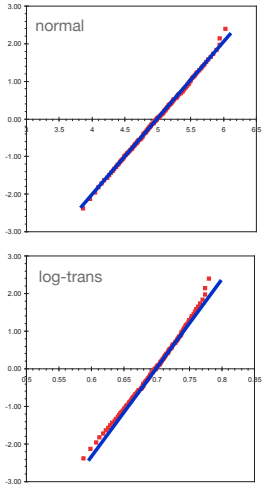


Review of practical exercises - 1.2

data distribution and cumulative frequency diagrams

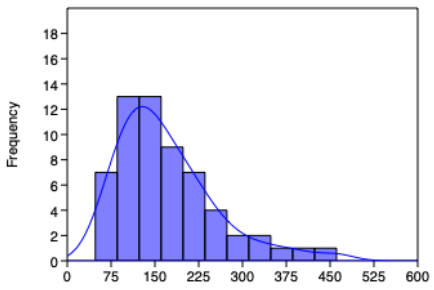


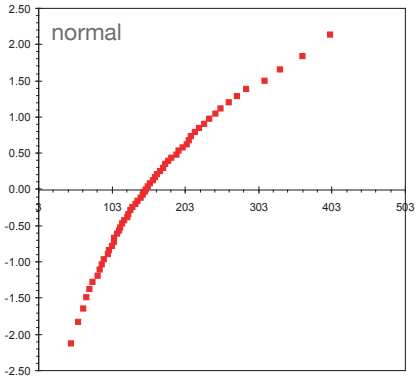
mean = 5.0 ± 0.5
median = 5.0 -0.5 +0.5



Review of practical exercises - 1.3

data distribution and cumulative frequency diagrams

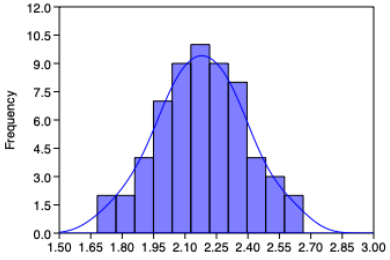


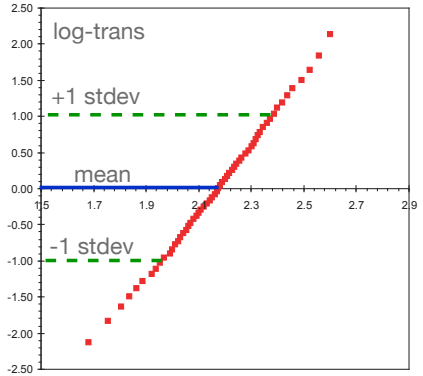


mean = 170 ± 85
median = 151 -56 +89

Review of practical exercises - 1.3

data distribution and cumulative frequency diagrams

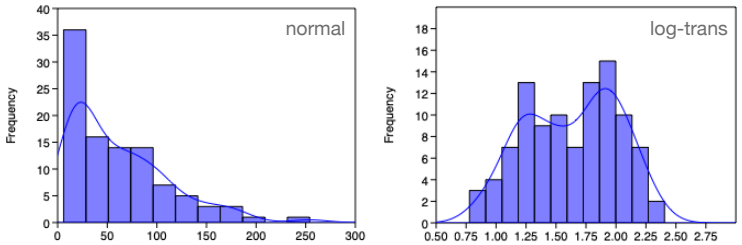




mean = 170 -58 +95
median = 151 -56 +89

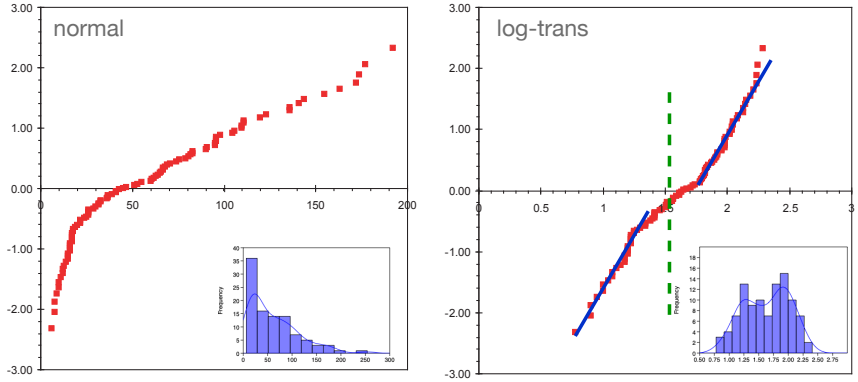
Review of practical exercises - 1.4

data distribution and cumulative frequency diagrams



Review of practical exercises - 1.4

data distribution and cumulative frequency diagrams



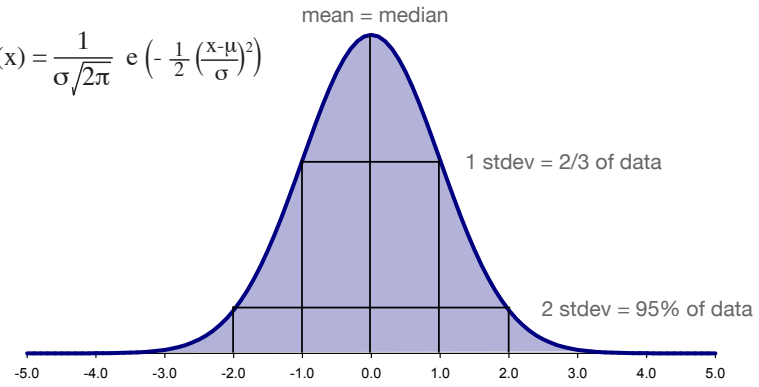
Data analysis and Geostatistics - lecture II

Assessing data quality and merging datasets

The normal or Gaussian distribution

If your data describe a phenomenon with one central value and variance around it due to many different disturbances: will trend to normal at high n

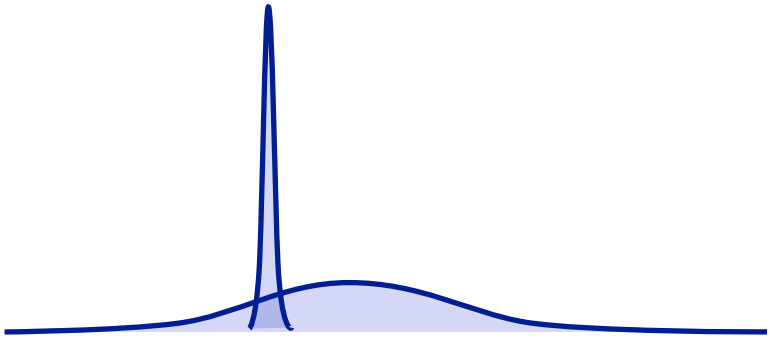
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



The normal or Gaussian distribution

In sample statistics, any property is an estimate with an associated uncertainty, where the uncertainty becomes less as more samples are obtained.

Re-phrased in terms of probabilities: a dataset with large uncertainty has a broad probability distribution



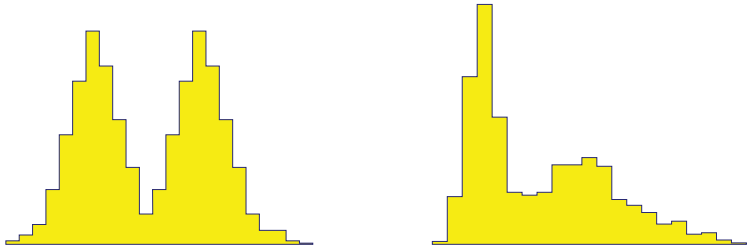
Multi-modal datasets: need to split them up

Multi-modal datasets: datasets that represent multiple samples or processes

to interpret such datasets you will need to split them up, otherwise you look at a mixed signal. But how to split up a dataset; where to put the boundary ?

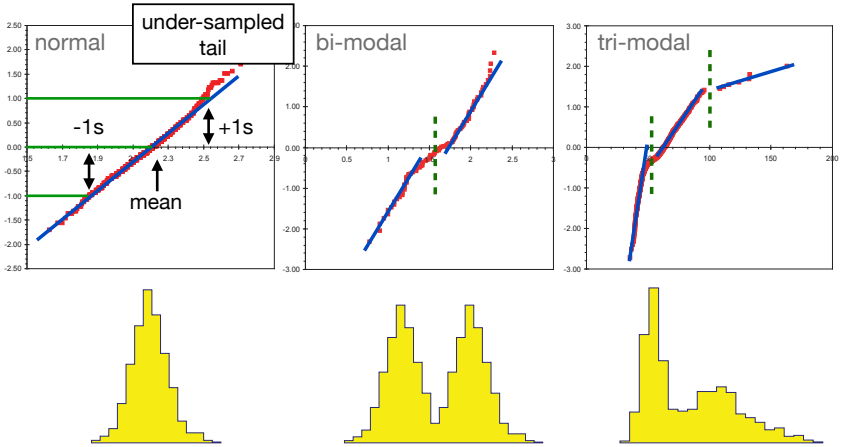
Individual distributions in a multi-modal dataset are likely to overlap

Probability plots allow you to determine where to split a dataset



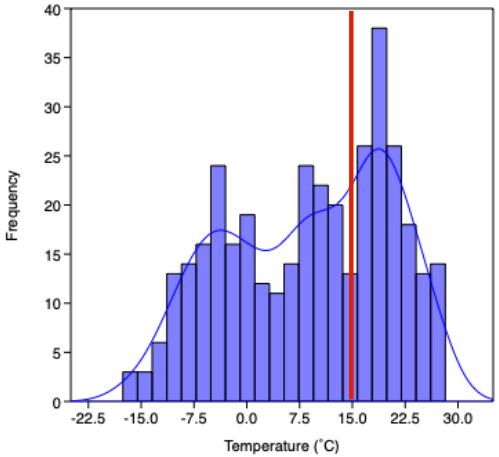
How to deal with multi-modal data sets

Have to split up the data set into groups: probability plots



The importance of data distribution

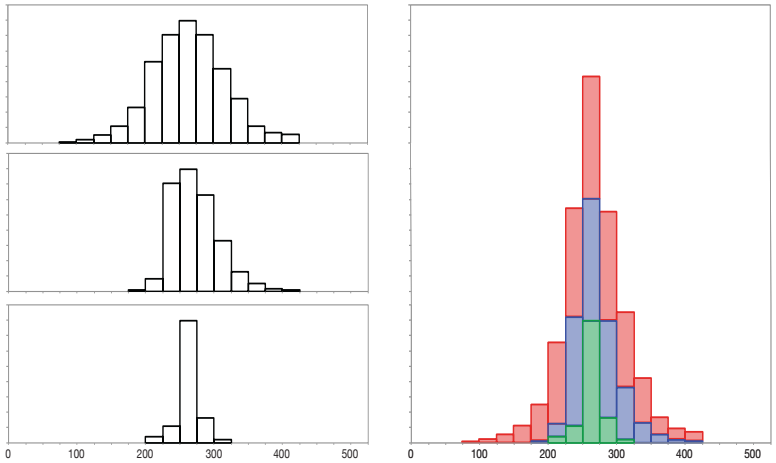
In Canada, the volume at the gas station is normalized to a temperature of 15˚C.



In 2021, the temperature in Montreal was < 15˚C for 224 days (61% of the year)

The same is true for Toronto and Vancouver

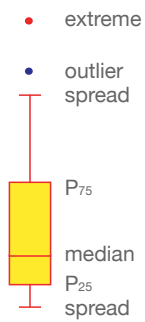
Graphical representation of data - comparisons



Box and whiskers plots

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

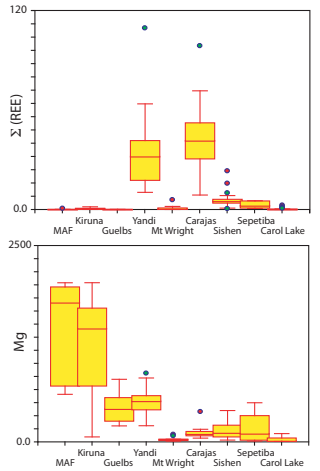
- stem and leaf diagrams
- box and whiskers plots - extremely useful in data comparisons:



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

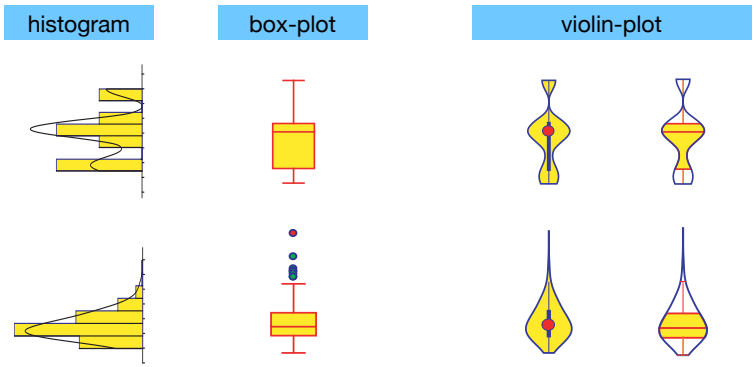
outliers: within a box defined by 3x the IQR - part of a normal distribution

spread: generally defined as 1.5x the Interquartile Range



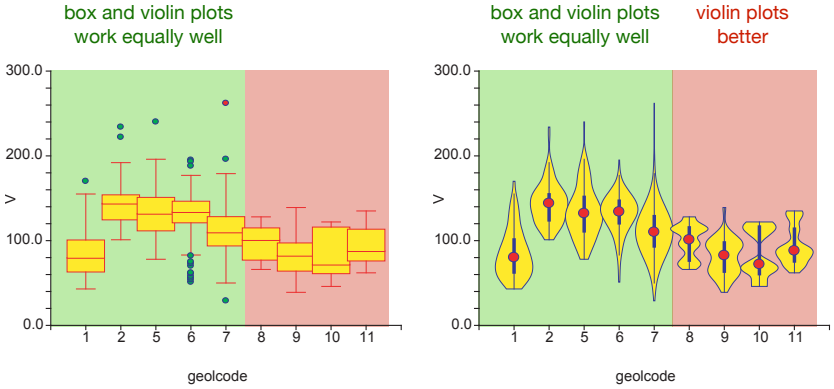
Even better: violin plots

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



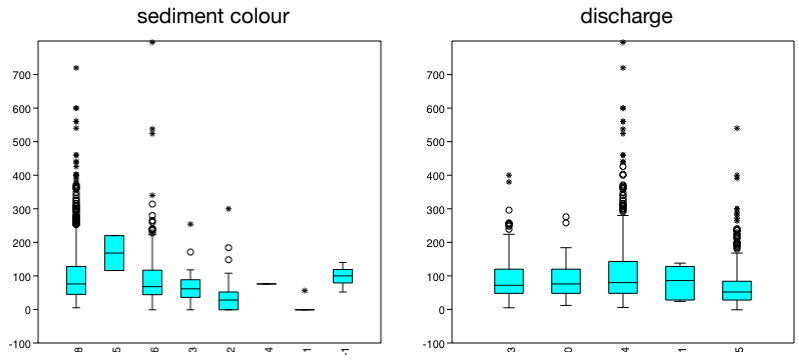
Even better: violin plots

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



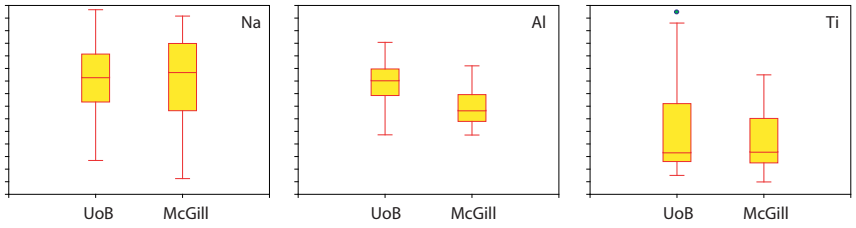
Compare by different criteria (grouping variable)

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:



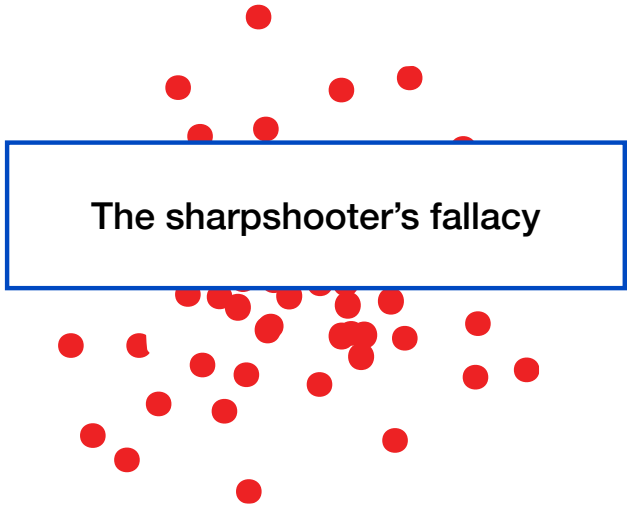
Comparison of data sets - quality control

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…

Wyatt Earp's stable door



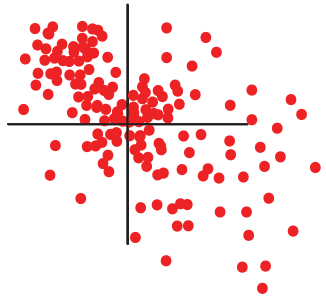
Precision and accuracy in data

what analyzing data, two principles are of crucial importance

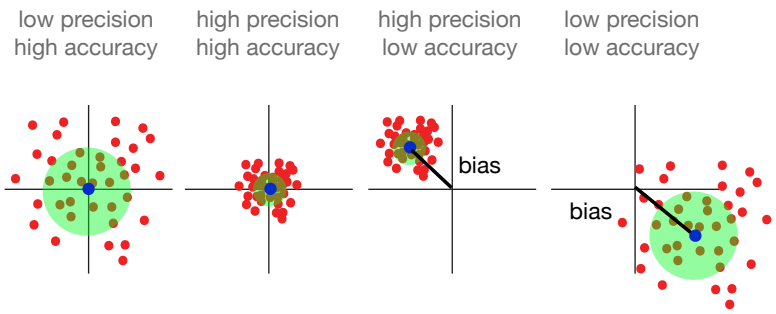
precision and accuracy:

the spread in the data and the deviation from the true value

the error bulls-eye:



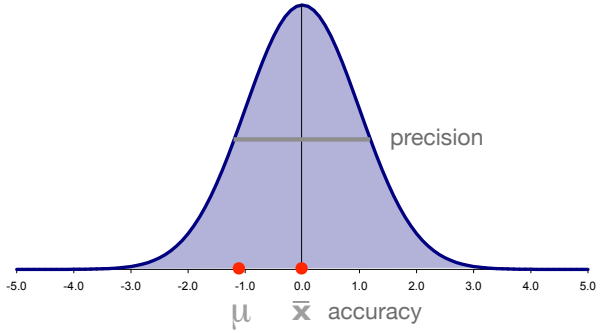
Precision and accuracy



low precision: large spread in the data, stdev is large relative to mean
low accuracy: deviation in mean from true mean - bias

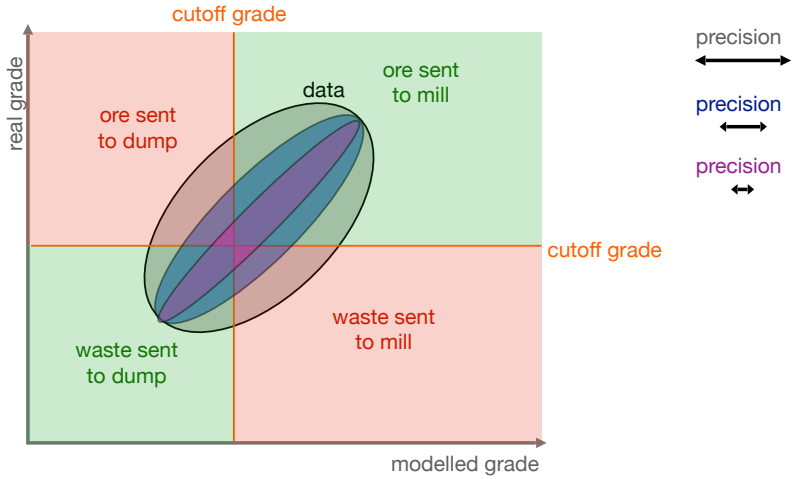
Precision and accuracy

Or when represented for the univariate case:



precision: the spread in the data - the width of the distribution
accuracy: the deviation between the sample and the population mean

The need for accuracy and precision



Precision and accuracy

how to deal with precision and accuracy ?

Whereas spread in the data is perfectly acceptable (and unavoidable), a bias in the data is not !

to improve the precision: more samples or a more precise analytical technique (pH meter instead of pH paper)

to remove a bias in data: analyze secondary standards and normalize your data to these (SRMs)

Accuracy is also of crucial importance when you want to compare data as accuracy issues can easily be mistaken for real differences

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

Quantifying the precision

Assuming that any accuracy issues have been dealt with, we're mainly interested in quantifying the precision in data analysis

So how to determine the uncertainty on your data?

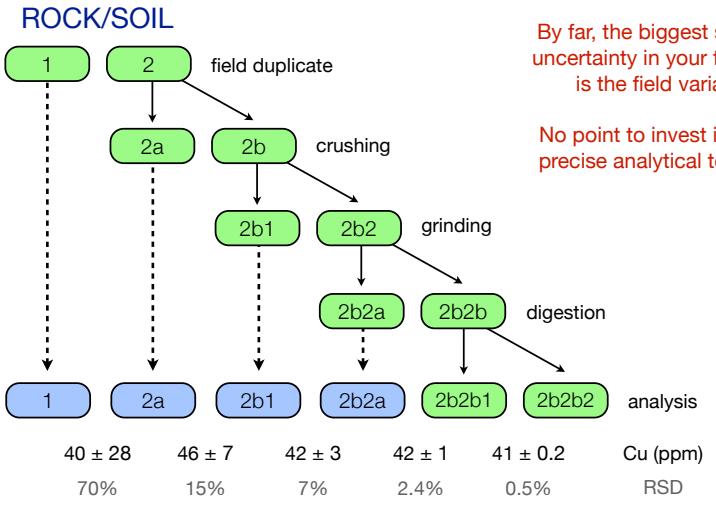
analyze 100 samples of the *same* lavaflow:

very good estimate of your precision, but you've probably also been fired.....

normally, we analyze a large number of different lavaflows: gives you an idea of spread in composition but not of precision

Instead, analyze a set of duplicate samples

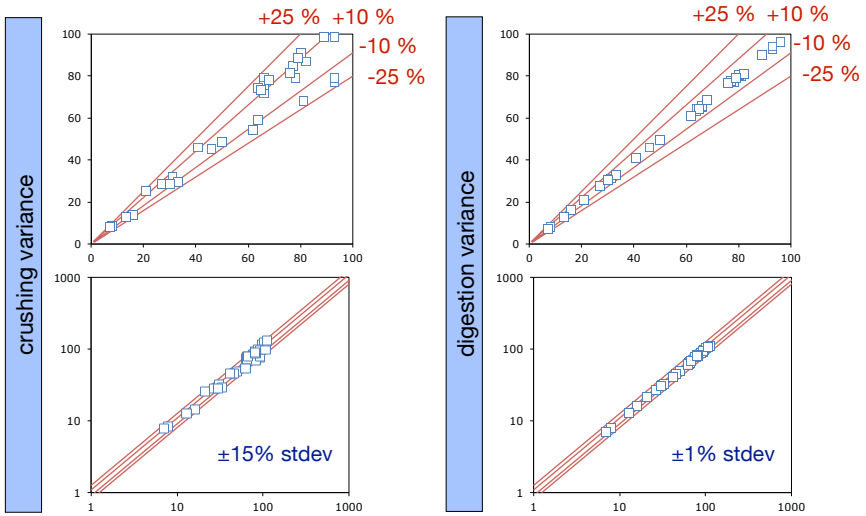
Monitoring the data acquisition process



By far, the biggest source of uncertainty in your final value is the field variance:

No point to invest in a more precise analytical technique

Monitoring the data acquisition process: duplicates



Count statistical uncertainty

Duplicates are not the only way to get an idea of precision

In the Earth Sciences a good portion of analytical techniques uses some form of counting:

microprobe - no. of counts at specified wavelength -> concentration
mass spec - no. of counts at specified mass -> concentration

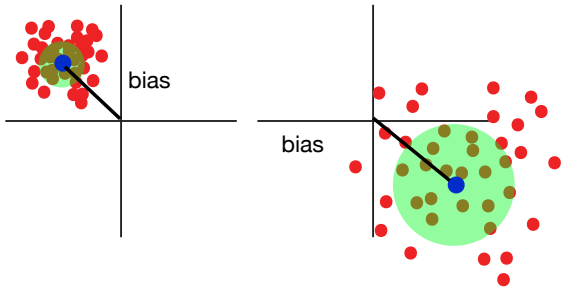
similar for XRF, XRD, AAS, and many more

CSE; count statistical error = $\sqrt{\text{counts}}$

Note: this is only a measure of the analytical uncertainty and may strongly overestimate the true precision in your samples !

Quantifying accuracy: 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

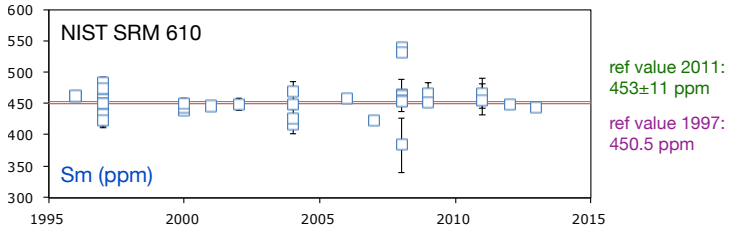


How do we know the correct value, i.e. the accuracy? Standard Reference Materials

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

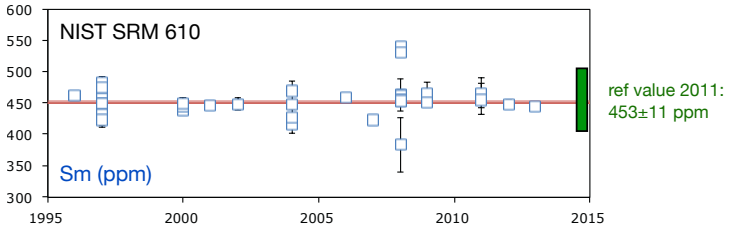


- Data depositories of SRM values are a great resource: GeoREM website: http://georem.mpch-mainz.gwdg.de

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.

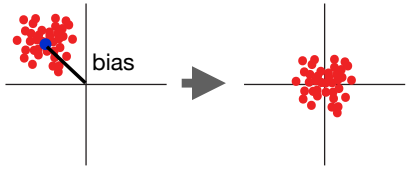


- SRMs are not always homogeneous: can receive a bad batch

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.

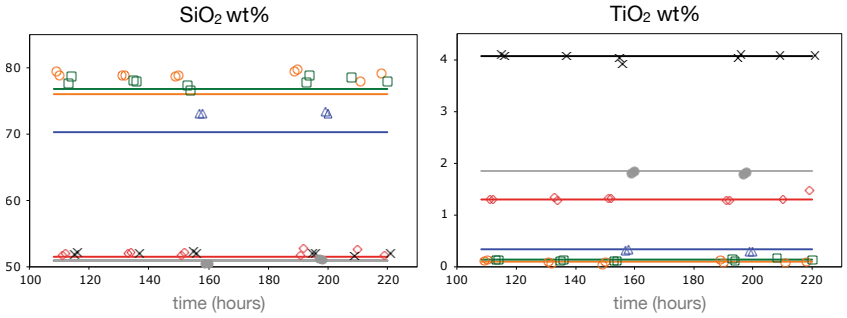
- SRMs should be as similar as possible to your sample material
- SRM should also have a similar concentration range. In most cases you need more than 1 -> choose them to cover your sample's range
- SRM allow for assessment of trueness, but also bias correction



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

Monitoring the data acquisition process - example

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



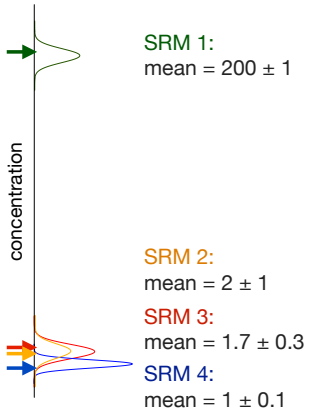
SiO_2 is consistently overestimated: bias

TiO_2 is spot-on !

To obtain the final corrected dataset: would shift the SiO_2 concentrations to match the certified values for the SRMs

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

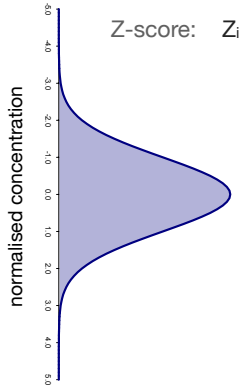


SRM 1:
mean = 200 ± 1

SRM 2:
mean = 2 ± 1

SRM 3:
mean = 1.7 ± 0.3

SRM 4:
mean = 1 ± 0.1



Z-score: $Z_i = (x_i - \bar{x}) / s$

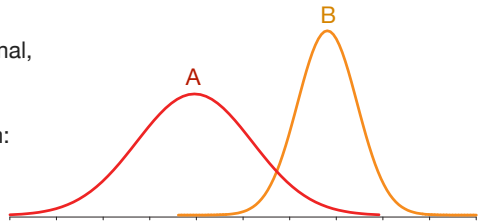
Z-scores

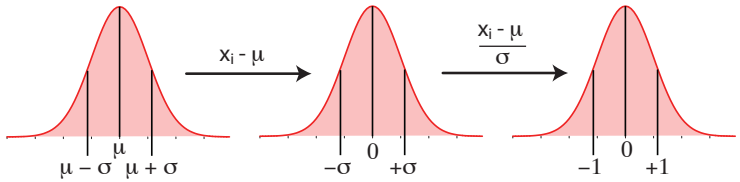
Z-score transformation for normally distributed data

populations A and B are both normal, but different in shape:

convert them to standardized form:

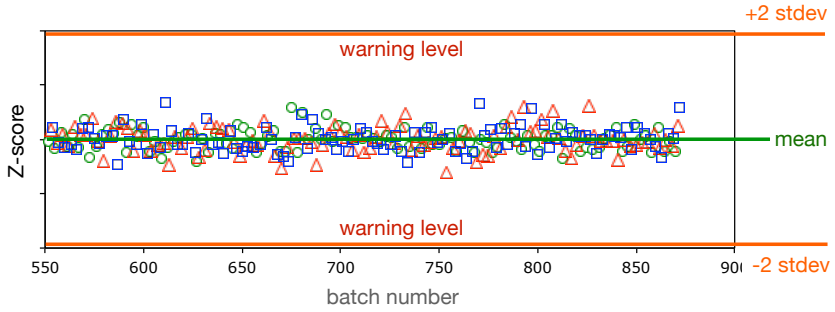
Z-score: $Z_i = (x_i - \mu) / \sigma$





Monitoring the data acquisition process: SRMs

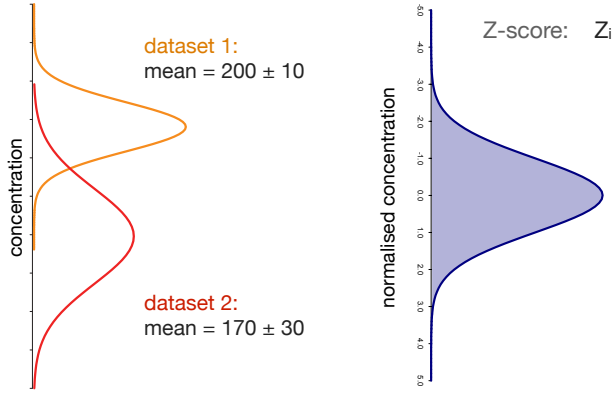
Identifying problems with the accuracy of your data:



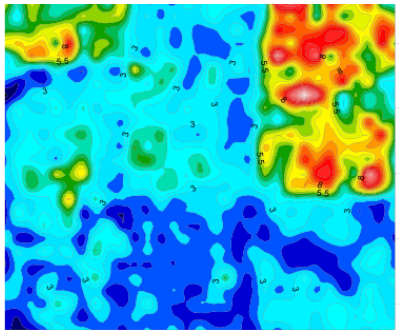
An elegant way to check all your SRMs at the same time, is to plot the Z-score of each value: this scales SRMs with different absolute concentrations and stdev

Data levelling using SRMs

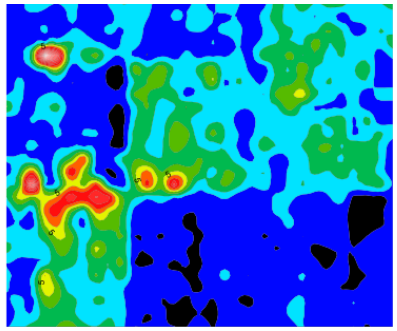
If the same SRMs have been measured in multiple datasets, you can level these data perfectly, because these are the same samples. Moreover, their data should have a normal distribution: can use Z-scores for levelling:



Data levelling



Cu

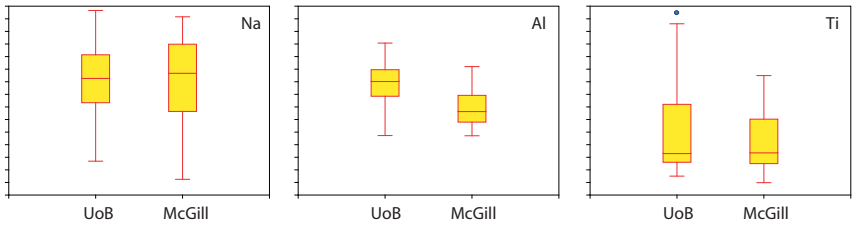


Zn

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.

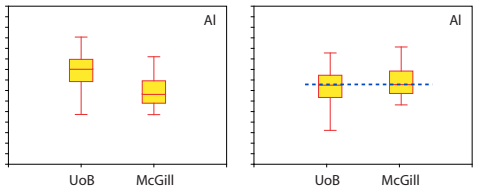
This can introduce spurious anomalies into your data: data need to be levelled first



Data levelling

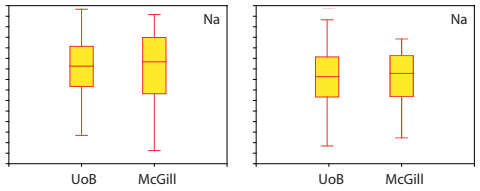
- shift to same mean or median, or ratio to the mean or median (data spread remains different)

  median = robust, whereas mean is affected by outliers



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

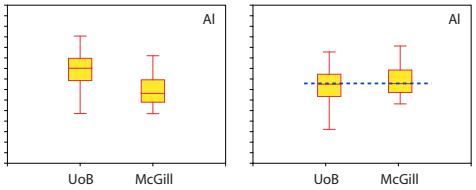
  a robust equivalent also exists using the median and mean-average-deviation (robust Z-score levelling) or using ranks instead of data (Gauss levelling)



Data levelling - mean or median shift

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

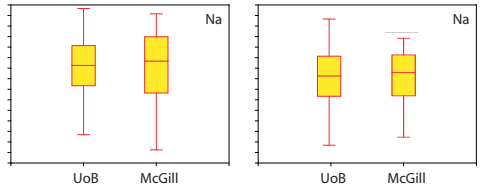
  median = robust, whereas mean is affected by outliers



			level to mean		level to median		levelled to UoB	
Data:	UoB	McGill	UoB	McGill	UoB	McGill	mean	median
	10.2	4.3	2.1	-1.0	2.3	-0.9	7.2	7.0
	8.4	5.8	0.3	0.5	0.5	0.6	8.7	8.5
	6.7	5.2	-1.4	-0.1	-1.2	0.0	8.1	7.9
	…	…	…	…	…	…	…	…
mean	8.1	5.3						
median	7.9	5.2		x - mean		x - median		

Data levelling - Z-score levelling

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



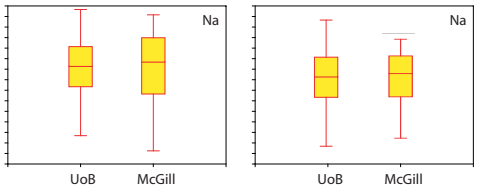
				Z-score level		levelled to UoB	
Data:	UoB	McGill		UoB	McGill	UoB	McGill
	102	45	113	-0.23	-0.88	102	79
	94	158		-0.46	0.54	94	129
	125	68		0.43	-0.59	125	89
	…	…		…	…	…	…
mean	110	115		0	0	110	110
stdev	35	80		1	1	35	35

Z-score: $Z_i = (x_i - \mu) / \sigma$

50

Data levelling - robust Z-score levelling

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



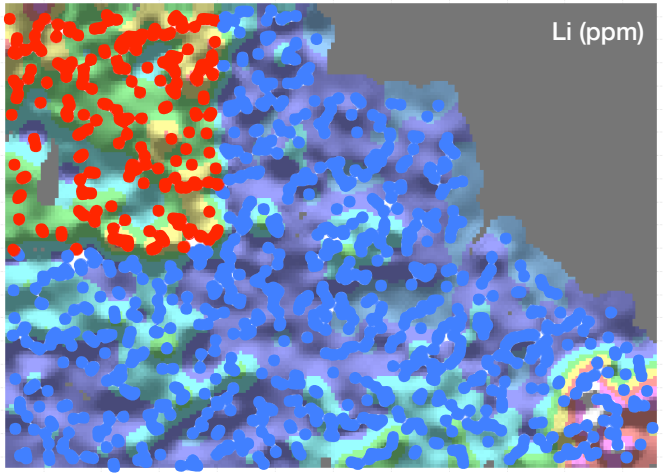
Data:	UoB	McGill
	102	45
	94	158
	125	68
	…	…
median	105	100
MAD	20	60

Z-score: $Z_i = (x_i - med) / MAD$

robust Z-score level		levelled to UoB	
UoB	McGill	UoB	McGill
-0.15	-0.92	102	87
-0.55	0.97	94	124
1.00	-0.53	125	94
…	…	…	…
0	*0*	105	*105*
1	*1*	20	*20*

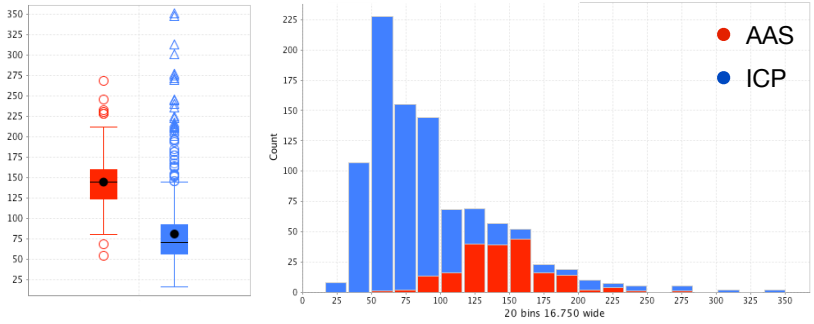
Data levelling

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



Data levelling

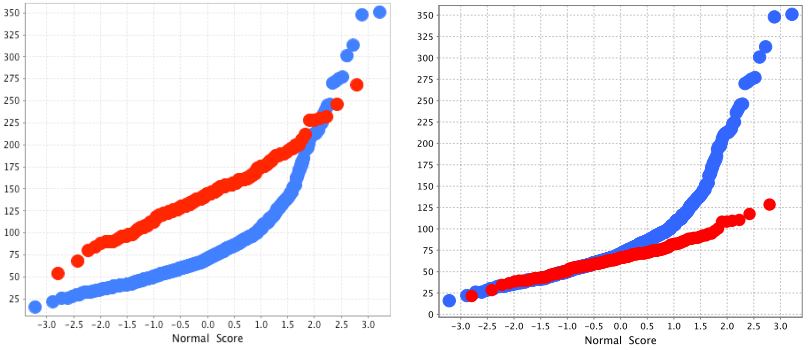
The two datasets are clearly different, both in concentration and in their data distribution: they do not sample the same geology in the same proportion!



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