Data analysis and Geostatistics - lecture II

Univariate statistics and the use of data descriptors

Previous lecture

previous lecture: overview of statistical techniques and concepts

data descriptors - median, mode, mean, stdev, IQR

data visualization - histograms, box plots

data analysis - separation (DFA) and clustering
- process recognition (PCA & factor)
- curve fitting (regression)

also: impartiality of the observer -> influences your confidence intervals

lack of appropriate control group in the Earth Sciences

start with core techniques and then move on to more complicated stats

Data and nomenclature

what are data and why do we gather them ?

a datum is a measurement of a property on a sample...

where
property can be density, length, ppm Ca, thermal conductivity
sample can be a rock, soil, water, plant

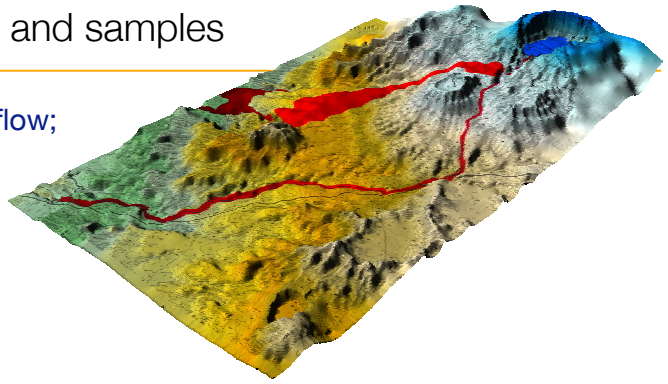
...intended to give us a value for the material where this sample came from

we are actually not interested in the composition of the sample, but rather in the composition of the source of this sample

brings us to the concept of a statistical population and sample

Populations and samples

given a lavaflow;



The complete lava flow is the population - if you want to know the exact composition of the population, you have to analyze it in its entirety

obviously impossible:

instead: analyze a representative sample of this population - estimate the properties of the host population

Estimating the properties of a population

From a set of samples, we can estimate the properties of the population, such as its characteristic value (mean or median) and the spread in values.

spread ≠ error !

A jeans shop will have a mean size, but it will also stock a spread of sizes

This mean + spread is the shop's estimate of the jeans sizes for its clientele population

Populations and samples

A representative sample has to cover all data characteristics of the host population:

- its central value (mean, median)
- the spread in the data (stdev, IQR)
- the data distribution (lognormal, modality)
- the relations with other variables

invariably it requires more than one *geological sample* to obtain a representative *statistical sample*

- the number of samples depends on the characteristics of the host population, but also on the sampling technique employed, the sample treatment and the analytical technique

e.g. granite vs. basalt, spot samples vs. mixtures, soil vs. stream sediments, mixing of crushed or milled rocks, field variance vs. lab variance

Populations and samples

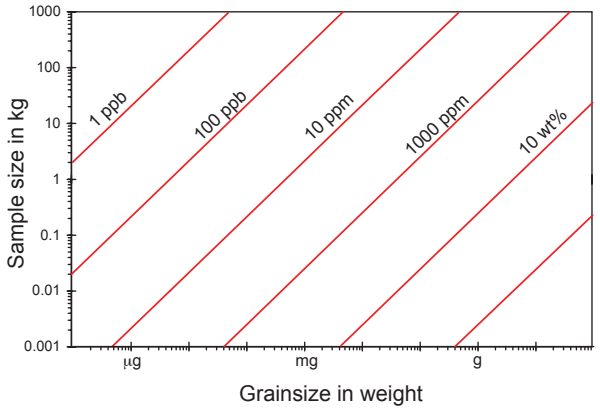
The statistics on national health suggest that 1 out of every 4 Americans, or 1 out of every 5 Canadians will suffer from a certain type of illness in their lifetime

This means at least 2 in the current Geostats cohort….

Is this reasoning correct ?

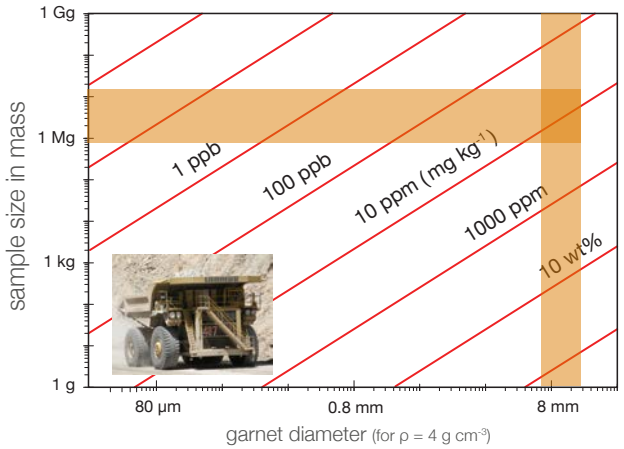
Representative samples

If you know something about your material you can estimate the number of samples you will need to get a representative sample of the population



Required sample size for a representative sample

10 cm

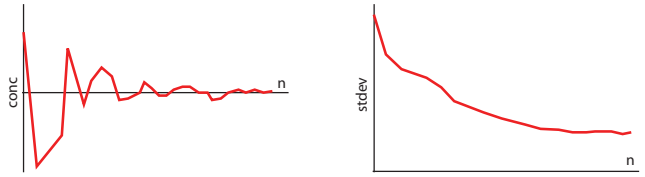


we would essentially need a sample bigger than the complete outcrop

Populations and samples

In geology we generally no longer have the population at our disposal
e.g. due to erosion, weathering and alteration

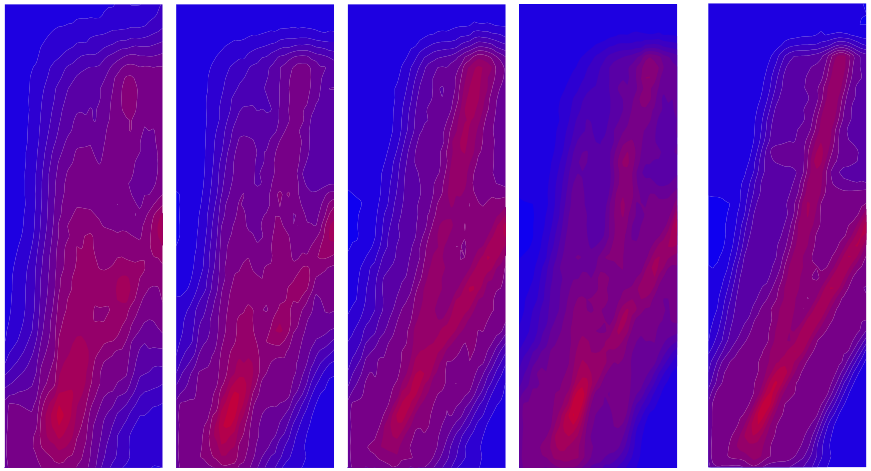
all the more important to make sure that your sample is representative



increasing number of samples -> when data characteristics no longer change -> representative sample

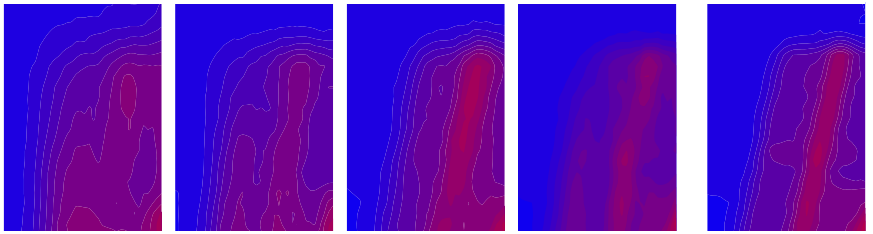
can estimate this if you know something of your samples: pilot sampling

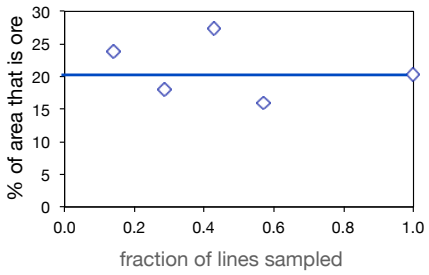
Infill drilling: vein-type deposit with halo

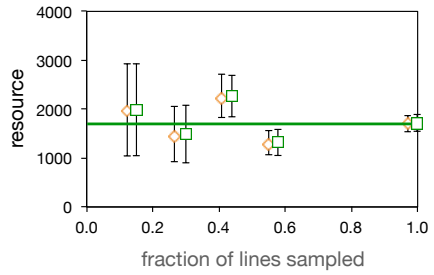


1 out of 7 lines sampled | 2 out of 7 lines sampled | 3 out of 7 lines sampled | 4 out of 7 lines sampled | 7 out of 7 lines sampled

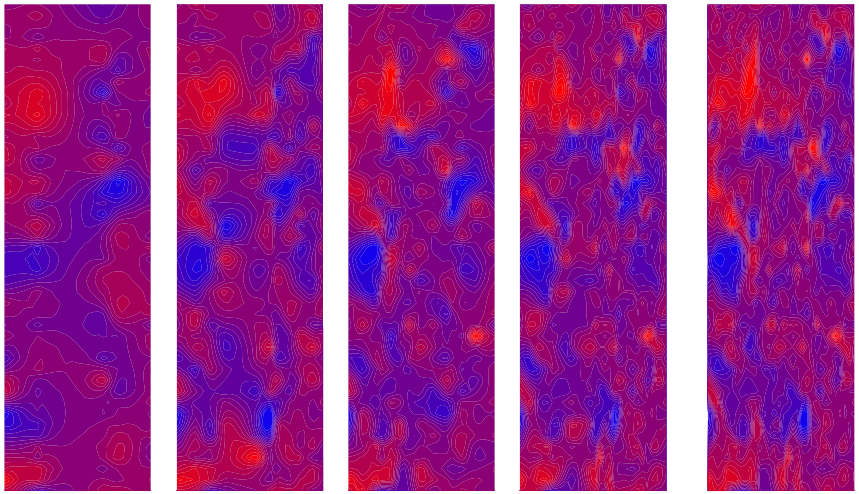
Infill drilling: vein-type deposit with halo





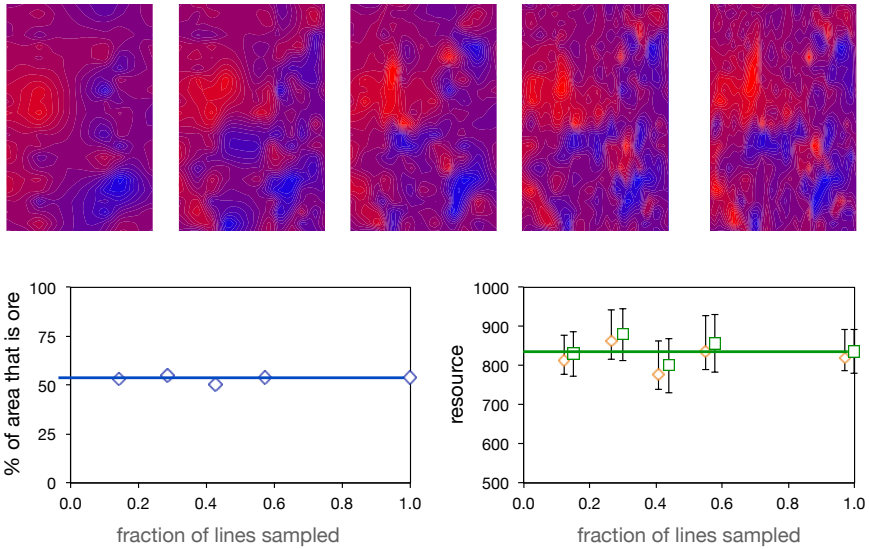


Infill drilling: disseminated deposit



1 out of 7 lines sampled | 2 out of 7 lines sampled | 3 out of 7 lines sampled | 4 out of 7 lines sampled | 7 out of 7 lines sampled

Infill drilling: disseminated deposit



Types of data

Not all data are equal in “quality” and this requires specific stats for some

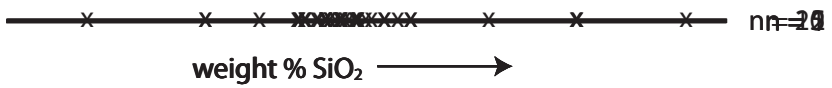
- ratio scale data: the most versatile of all. They have a natural zero point. e.g. charge, weight, length, concentration
- interval scale data: the intervals between the values are constant, but they do not have a natural zero point. e.g. °F, °C
- closed data: the data sum to a specified value. e.g. wt%, % of a core. Note the closure problem in these.
- ordinal scale data: the intervals between the values are not constant. e.g. Moh’s hardness scale of minerals
- discrete data: only certain values are allowed, mostly the integers. e.g. number of grains in a sample. Not ppm !
- categorical data non-numerical observations. e.g. colour, presence/absence of a feature in a fossil.

Ways to analyze your data

- univariate: each variable is analyzed separately: data distribution, central value and data spread/uncertainty
- bivariate: two variables are analyzed together to look for correlation or separation of data - regression
- multivariate: more than 2 variables are analyzed together. Generally difficult to visualize data and results
- spatial statistics: variation of variables in space, either 1D (well logs), 2D and 3D (topography) or >3D, but some have to be spatial !
- time series: variation of variables along a time progression

We will start with univariate techniques - the distribution of data

Univariate statistics



repeated analyses of the same sample, or a variety of samples from the same host population, will not return an identical value due to:

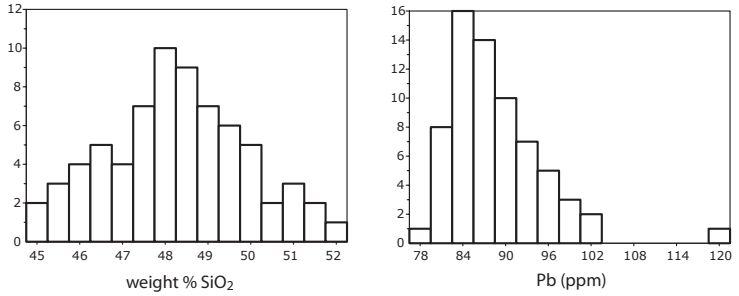
sample heterogeneity: % olivine in each sample
lava heterogeneity: layering or phase separation
analytical uncertainty: not every ion makes it to the detector

analytical uncertainty ~ error, but heterogeneity is a property of the host population and is not error -> both result in uncertainty on your estimate of the central population value

e.g. average Pb content of all Canadian rocks

Data visualization

To understand your data: plot their distribution!



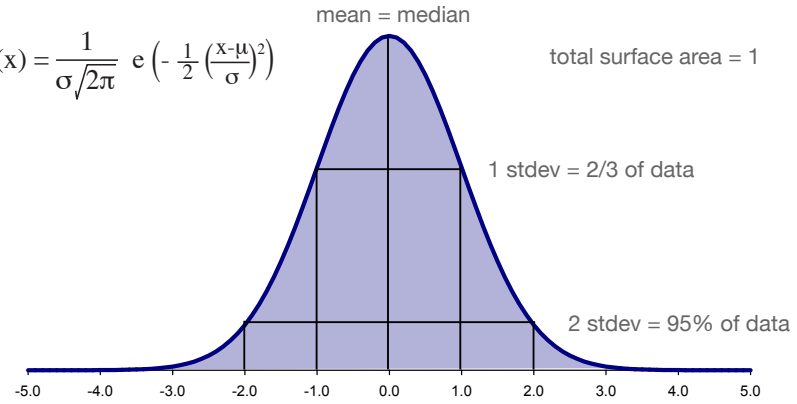
For these distribution you can now define a central value and spread:

$$\mu = \frac{\Sigma (x_i)}{n} \quad \sigma^2 = \frac{\Sigma (x_i-\mu)^2}{n} \qquad \bar{x} = \frac{\Sigma (x_i)}{n} \quad s^2 = \frac{\Sigma (x_i-\bar{x})^2}{n-1}$$

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

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



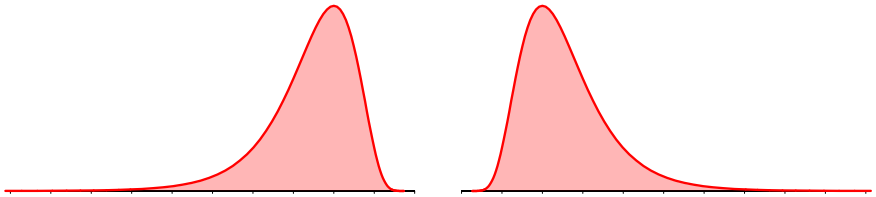
Normality in data sets

normal distribution only when looking at one phenomenon, when all variation is averaged out, or when one phenomenon is dominant

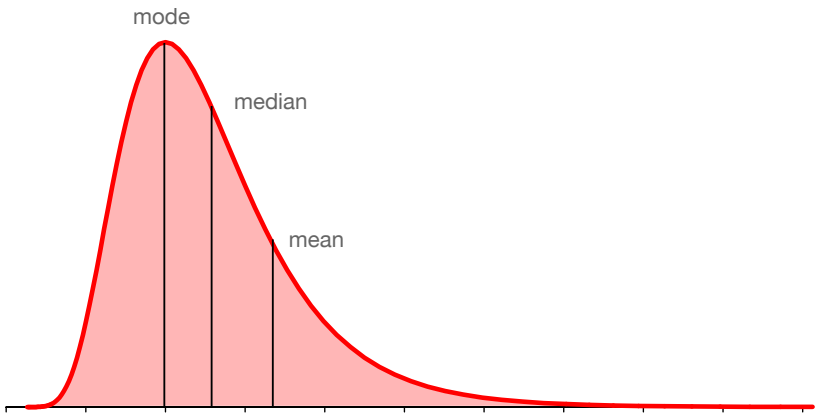
So: in most cases in geology -> deviations from normality

skewness is especially common and leads to the lognormal distribution
now median is not equal to mean:

negative skew: mean-median < 0, tail to the left (low values)
positive skew: mean-median > 0, tail to the right (high values)

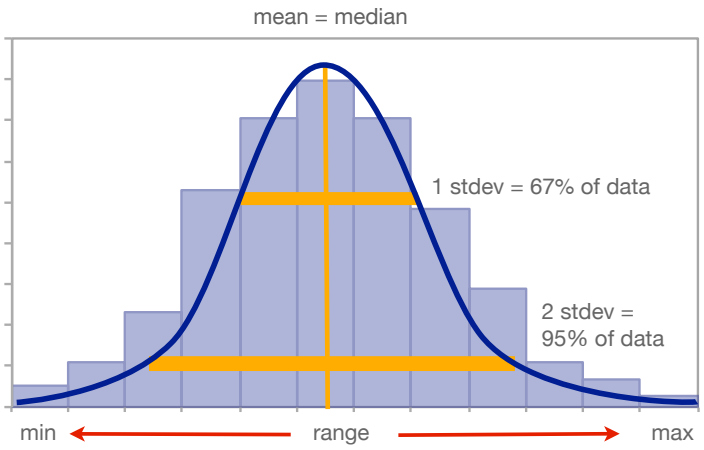


Log-normal distribution



Standardized data descriptors

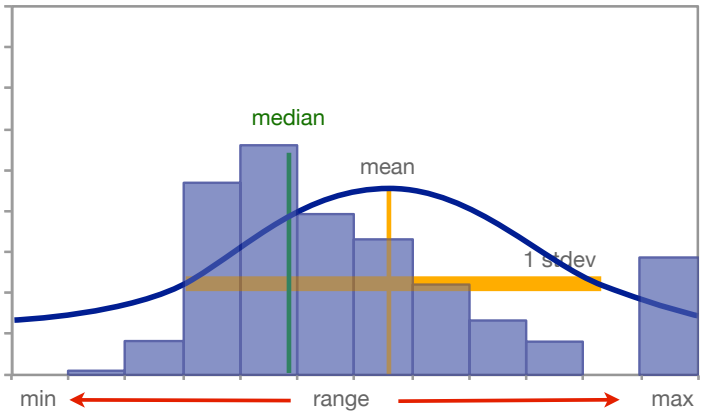
If your data describe a phenomenon with one central value and random disturbances around this value: will trend to a normal distribution



Standardized data descriptors

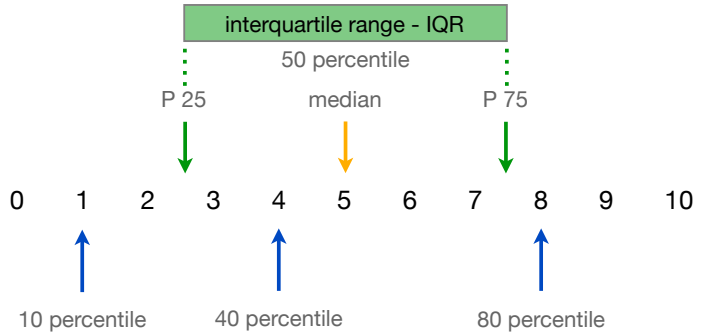
Unfortunately, many data sets are not normally distributed

the range in the data is identical, but the data data distribution has changed



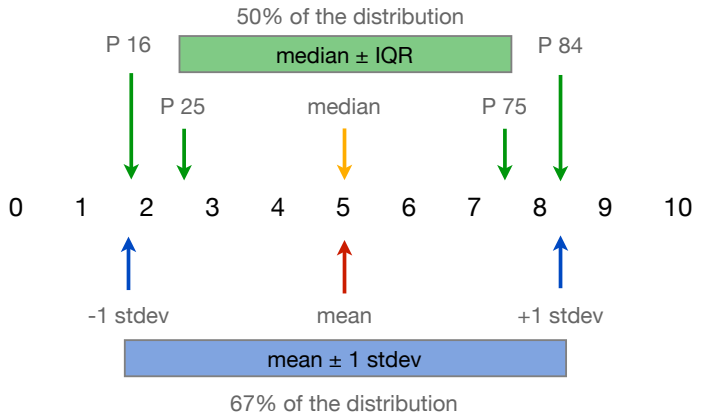
Robust descriptors: median and percentiles

The median - middle value - is a robust indicator that is not influenced by outliers. Now need an estimator of the spread: the interquartile range IQR



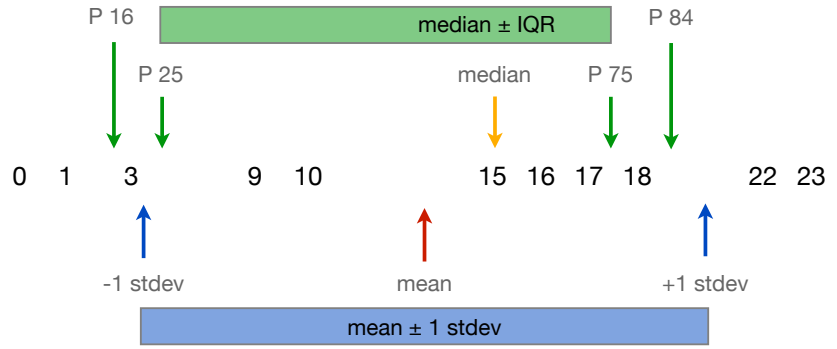
Robust descriptors: median and percentiles

The median - middle value - is a robust indicator that is not influenced by outliers. Now need an estimator of the spread: the interquartile range IQR



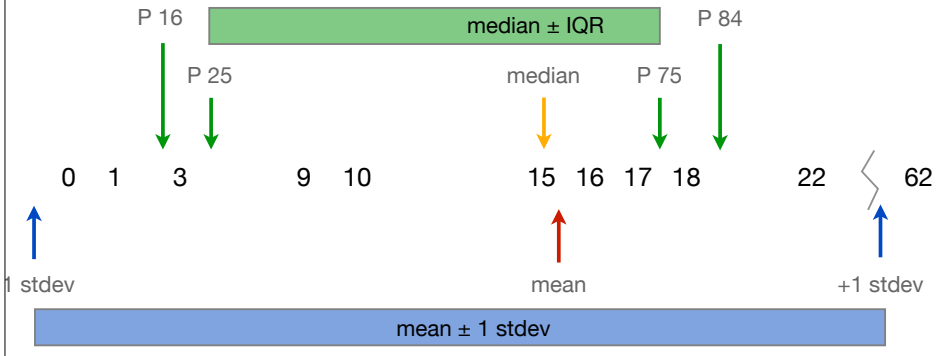
Robust descriptors: median and percentiles

The median - middle value - is a robust indicator that is not influenced by outliers. Now need an estimator of the spread: the interquartile range IQR



Robust descriptors: median and percentiles

The median - middle value - is a robust indicator that is not influenced by outliers. Now need an estimator of the spread: the interquartile range IQR



mean - 1 stdev < minimum value !

Hours of Netflix watched per week

Median + IQR (or P_{84}-P_{16}) is a robust indicator of characteristic value + spread, whereas mean ± stdev is not-robust and sensitive to outliers:

Hours of Netflix watched per week for a group of students:

2,4,6,8,10 mean = 6, median = 6

2,4,6,8,60 mean = 16, median = 6

Including the stdev and P_{84}-P_{16} indicators of spread:

2,4,6,8,10 mean = 6 ± 3, median = 6 -3,+3

2,4,6,8,60 mean = 16 ± 25, median = 6 -3,+21

This says that $^2/_3$ of the data fall between -9 and +41 in the case of the mean. Although true, this does not describe the data well at all !

What is an outlier ?

Outliers are extreme values in a dataset, but these are not necessarily caused by things like measurement errors: outliers are NOT faulty data, although they can be

Better definition (from Banett and Lewis): a set of data that are inconsistent with the remainder of the dataset. In exploration geochemistry one of the tasks is in fact the identification of outliers.

Outliers are always defined relative to a data distribution, because they are values that are not expected for that given data distribution.

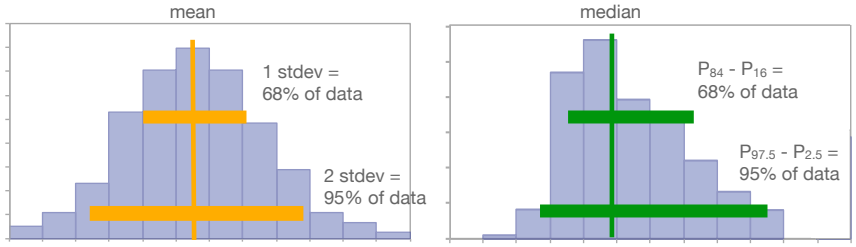
A data distribution essentially tells you the probability of finding a given value. This is useful, because it allows us to designate a value as an outlier:

for example, if the probability of that value occurring in my distribution is less than 1%, I will classify the value as an outlier

However, this depends on the distribution of the data !

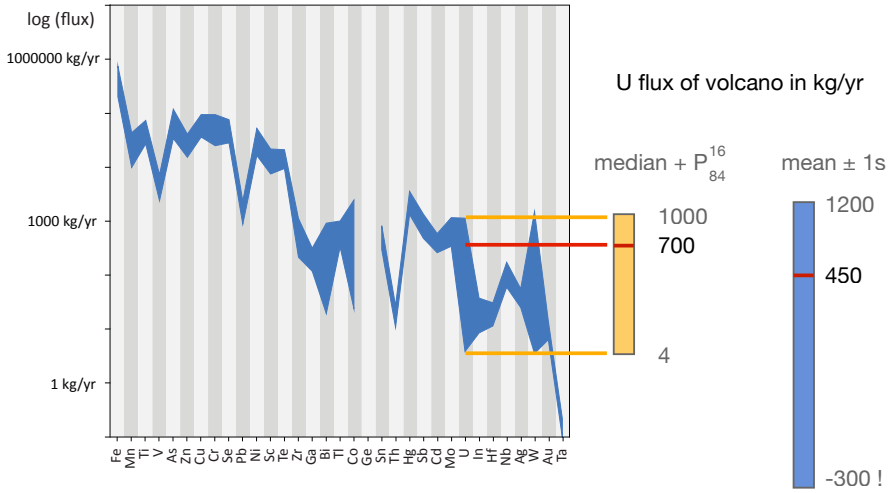
Summarizing your data

By reporting a dataset's characteristic value and its spread as mean ± stdev, the reader has an expectation of the data distribution, which is only correct for a normal distribution. Median + IQR is generally more appropriate and correct.



For a non-normal distribution, spread is generally asymmetric when using median + percentiles. This immediately gives information on the distribution of the data !

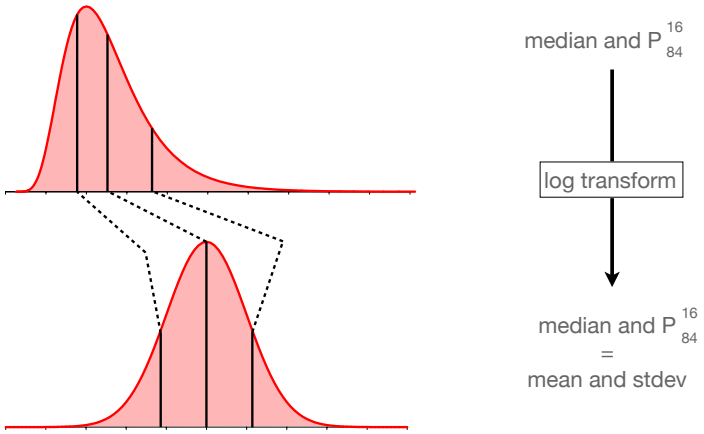
Ratio and logarithmic data in spider-diagrams



Not only are median + IQR robust, these properties also represent real values

Log-normal transformation

most statistical techniques cannot deal with a lognormal distribution -> transform it to a normal distribution



Benefits of a robust indicator - example

The name "robust" refers to these parameters not being sensitive to outliers or to addition of small sets of data: the value stays the same. This is in sharp contrast to the mean, for example, which changes with every value added

Ni (ppm)				
34				
55				
23				
25				
31				
65				
39				
45				
41				
43				
mean =	40	91	167	157
stdev =	±13	±169	±308	±291
median =	40	41	42	41
P_{25} =	-10.5	-10	-10	-9.5
P_{75} =	+7.5	+14	+20.5	+20

Median absolute deviation - MAD

Sometimes it is impractical to have a lower and upper uncertainty on the median, and one characteristic value for robust spread is needed: MAD

The median absolute deviation is the robust equivalent of the stdev.

MAD = the median of the absolute deviations from the data median

Pb content (ppm)	|deviation| from median	sorted |deviation|	
10	10	0	
10	10	0	
20	0	10	
20	0	**10**	← MAD
40	20	20	
60	40	40	
90	70	70	

Standard deviation and MAD differ by a scaling factor. For the normal distribution, this scaling is stdev = 1.4826 • MAD

Confidence level on your data descriptors

It is very useful to know what the confidence is on your central value and its spread: How much is my mean likely to shift if I collect more data, assuming that my pilot study is representative?

If you know your data distribution, this can be calculated exactly. However, in geochemistry, we generally estimate the distribution from the data we have.

Ni
34
55
23
25
31
65
39
45
60

Bootstrapping:

subsampling your dataset

calculating the parameters on these subsets

resulting spread: confidence level

Ni	Ni	Ni	Ni	
34	23	31	39	
55	25	65	45	
23	31	39	60	etc
Ni	Ni	Ni	Ni	
34	23	31	39	
23	31	39	60	
31	39	60	55	etc

Bootstrapping - example with PAST

Ni
34
55
23
25
31
65
39
45
60
42
48
24
31
55
39
36
51
47
53
2500

Other deviations from normality

Many other deviations from normality:

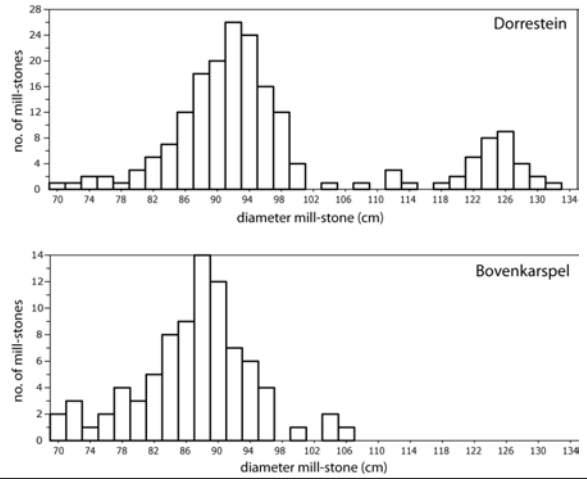
- outliers - need robust estimators
- bimodality or multimodality - data set will have to be split
- kurtosis - steepness of the distribution

One way to check for normality is the cumulative frequency plot

Multi-modal datasets: Why is there multi-modality?

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

Size of discarded medieval millstones in two locations in the Netherlands



Other deviations from normality

Many other deviations from normality:

- outliers - need robust estimators
- bimodality or multimodality - data set will have to be split
- kurtosis - steepness of the distribution

One way to check for normality is the cumulative frequency plot

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