Data analysis and Geostatistics

Short Course on the use of statistical techniques in the geosciences

Vincent van Hinsberg • McGill University

Goal of this short course

This course aims to convince you that (geo-)statistical techniques provide a useful and powerful tool to analyze geological data

Six days is not enough time to make you a stats expert. Instead, the aim is to make you comfortable with a wide variety of techniques, from basic to advanced, so you know what is available and how to interpret results

key ideas:

1. If a simple technique suffices, more advanced techniques will only do harm
2. A single figure conveys more than a thousand words
3. Concepts are more important for the stats user than theory
4. There is a major role for robust techniques in analysing geo-data

Practical matters

Short Course	Short Course for credit
lecture	lecture
lecture	lecture
lab	lab
	written exam (50%)
	group data project (50%)

Practical matters

- the course consists of **lectures in the morning** that discuss concepts, theory and tools, and **practical sessions in the afternoon** in which you get to apply these statistical tools to geo-datasets
- **book:** Introduction to geological data analysis - Swan & Sandilands
- **additional resources online:** https://www.ncss.com/software/ncss/ncss-documentation
- **software:** spreadsheet programs and the statistics package PAST
- **examination:** written final exam (50%) and data analysis project (50%).
- **full course details available on:** eps.mcgill.ca/~hinsberg/intro/Teaching.html

Practical matters - topics covered

data description: mean - median - mode, histograms, normality, outliers, modality, box and whiskers plots, stem and leaf diagrams

measures of uncertainty: sources of uncertainty, range, standard deviation, variance, inter-quartile range, error propagation

missing values: common problem in geology and generally ignored - real missing values vs. detection limits, and how to deal with missing values

statistical testing: hypotheses, confidence levels, value and rank testing, Z-, t-, Chi-squared, Kolmogorov-Smirnov, Mann-Whitney tests

regression & correlation: Scatter diagrams, Pearson & Spearman correlation coefficients, significance of correlation, curve fitting, (non-)linear models

multivariate techniques: sum of squares methodology, discriminant function analysis, principle component & factor analysis, cluster analysis

spatial data analysis: spatial distribution of data, 3D visualization (isolines, bubble plots, trend surfaces), semi-variograms, kriging

Course practicalities - Schedule

	day 1 Wed, Mar 13	day 2 Thu, Mar 14	day 3 Mon, Mar 18	day 4 Wed, Mar 20	day 5 Thu, Mar 21	day 6 Thu, Mar 28
9:15 - 11:00	Introduction	Distributions QA/QC and levelling	Statistical testing I: the basics	Timeseries analysis	Discriminant analysis and clustering	Vector methods II: PCA, FA, PLS
11:00-11:15	break	break	break	break	break	break
11:15 - 13:00	Data and univariate data descriptors	Bivariate data and correlation analysis	Statistical testing II: tests and ANOVA	Regression analysis and curve fitting	Vector methods I: PCA, FA, PLS	Spatial data analysis and kriging
13:00 - 14:00	lunch	lunch	lunch	lunch	lunch	lunch
14:00 - 17:00	**Lab 1:** Distributions and data descriptors	**Lab 2:** Distributions and data descriptors	**Lab 3:** Distributions and data descriptors	**Lab 4:** Distributions and data descriptors	**Lab 5:** Distributions and data descriptors	**Lab 6:** Open forum
Lecture room	FDA 348	FDA 348	FDA 232	FDA 348	FDA 348	FDA 348
Lab room	FDA 348	FDA 315	FDA 232	FDA 348	FDA 315	FDA 315

Course practicalities - McGill policy statements

"In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded."

"McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures" (see www.mcgill.ca/students/srr/honest/ for more information).

"© Instructor-generated course materials (*e.g.*, handouts, notes, summaries, exam questions, etc.) are protected by law and may not be copied or distributed in any form or in any medium without explicit permission of the instructor. Note that infringements of copyright can be subject to follow up by the University under the Code of Student Conduct and Disciplinary Procedures."

"Work submitted for evaluation as part of this course may be checked with text matching software within myCourses."

"In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change."

"You are reminded of your responsibility in ensuring that this content and associated material are not reproduced or placed in the public domain. This means that it can be used for your educational purposes, but you cannot allow others to use it by putting it up on the Internet or by giving it or selling it to others who may also copy it and make it available. Please refer to McGill's Guidelines for Instructors and Students on Remote Teaching and Learning for further information."

Geotop Short Course in Data Analysis and Geostatistics

Part 1. An introduction

geotop

Before we start....

Lots of strong opinions on statistics and data analysis:

“Fools can figure and figures can fool”

“The only use of statistics is in politics”

“You can prove anything with statistics”

“You have lies, you have damned lies, and you have statistics”

“Facts are stubborn, but statistics are more pliable”

Unfortunately, most people are not sufficiently familiar with statistics to spot its abuse and they therefore dismiss its proper use in analyzing data

This has become a particular issue during the pandemic

Before we start....

Lots of strong opinions on statistics and data analysis:

“Fools can figure and figures can fool”

“The only use of statistics is in politics”

“You can prove anything with statistics”

“You have lies, you have damned lies, and you have statistics”

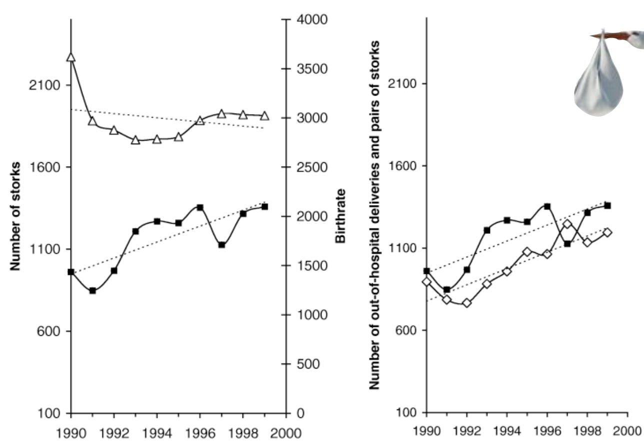
“Facts are stubborn, but statistics are more pliable”

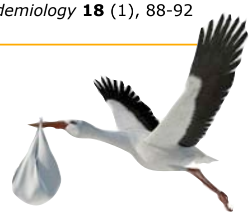
Unfortunately, most people are not sufficiently familiar with statistics to spot its abuse and they therefore dismiss its proper use in analyzing data

- Theory of the stork
- Anderson’s lion
- White and black swans

Theory of the Stork

Paediatric & Perinatal Epidemiology **18** (1), 88-92





Before we start....

Lots of strong opinions on statistics and data analysis:

“Fools can figure and figures can fool”

“The only use of statistics is in politics”

“You can prove anything with statistics”

“You have lies, you have damned lies, and you have statistics”

“Facts are stubborn, but statistics are more pliable”

Unfortunately, most people are not sufficiently familiar with statistics to spot its abuse and they therefore dismiss its proper use in analyzing data

- Theory of the stork
- Anderson’s lion
- White and black swans

Two key concepts to start with

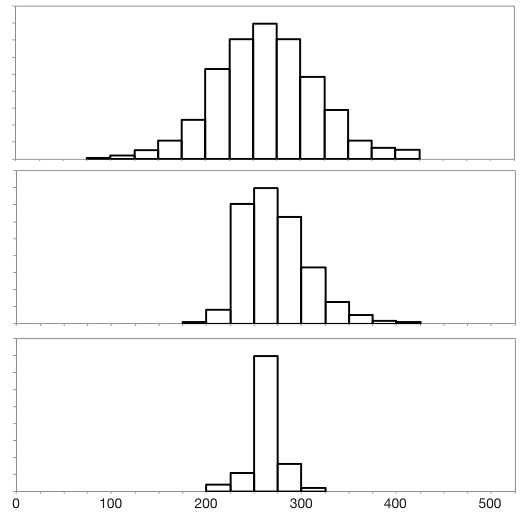
1. When analyzing data, and applying statistical tools, the simpler technique is generally the best, because it makes for the strongest point and is the easiest to explain and defend to your audience

   if a mean and standard deviation do the trick, why go further ?

2. A figure says more than a thousand words: graphical representations of data and results are always easier to interpret and convey

   “The results of the survey indicate that unit A contains 234 ppm of Ga with a range from 38 to 445 and 50 ppm standard deviation, unit B contains 283 ± 40 ppm with a range from 180 to 448, and unit C has a range from 200 to 300 with a mean of 250 and 10ppm standard deviation”

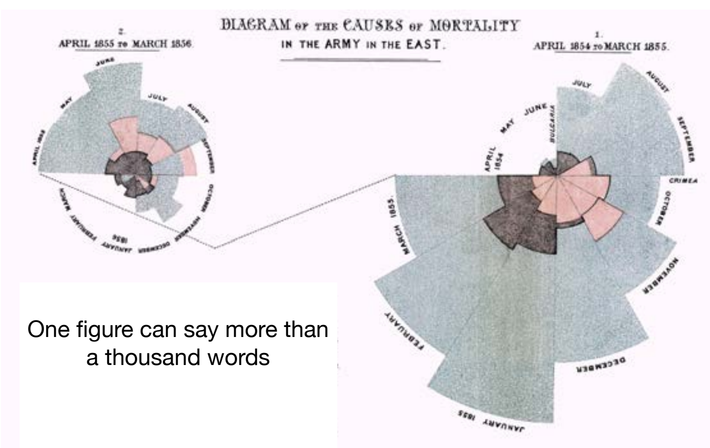
Graphical representation of data



“The results of the survey indicate that unit A contains 234 ppm of Ga with a range from 38 to 445 and 50 ppm standard deviation, unit B contains 283 ± 40 ppm with a range from 180 to 448, and unit C has a range from 200 to 300 with a mean of 250 and 10ppm standard deviation”

Florence Nightingale - Crimean war

"To understand God's thoughts we must study statistics, for these are the measure of His purpose"



Three main fields in statistics

- Data analysis - data appraisal and data mining

  should be the first step in any data analysis exercise

  - was my sampling ok?
  - what about analyses? appropriate? accurate?
  - what do the values mean?
  - are there any outliers and what do they mean?

  Commonly, this is all you need to do, but for some reason it is generally skipped (e.g. kriging when lowest Pb content is well above intervention value)

- Probability analysis - confidence of statistical statements

- Statistical testing and modeling - process recognition and quantification

Three main fields in statistics

- Data analysis - data appraisal and data mining
- Probability analysis - confidence of statistical statements

  This is a field in itself and will be limited here to its control on the confidence level of statistical statements = “statistical proof”

  what is the chance that my correlation is purely coincidental and do I accept this probability

  In geochemistry generally 95% is chosen: in 1 out of 20 cases we are wrong! In oil exploration closer to 10%, whereas in space exploration 99.99%.

- Statistical testing and modeling - process recognition and quantification

Confidence levels - catching cheating teachers

Union of concerned scientists

Three main fields in statistics

- Data analysis - data appraisal and data mining
- Probability analysis - confidence of statistical statements

  This is a field in itself and will be limited here to its control on the confidence level of statistical statements = “statistical proof”

  what is the chance that my correlation is purely coincidental and do I accept this probability

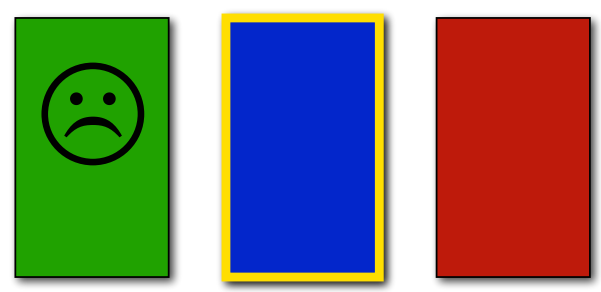
  In geochemistry generally 95% is chosen: in 1 out of 20 cases we’re wrong! In oil exploration closer to 10%, whereas in space exploration 99.99%.

  Note: because we generally study events after they have happened, we’re not an impartial observer -> this changes the probabilities!

- Statistical testing and modeling - process recognition and quantification

Three door problem (Monty Hall problem)

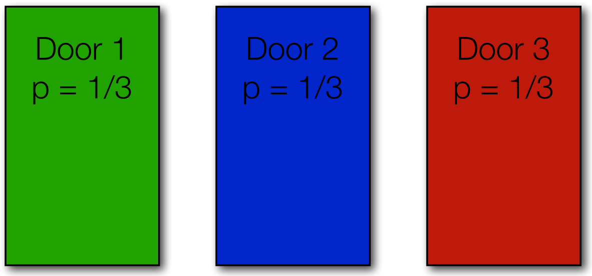
Quiz on television in the 80s;



are you better off swapping doors or does it not make any difference ?

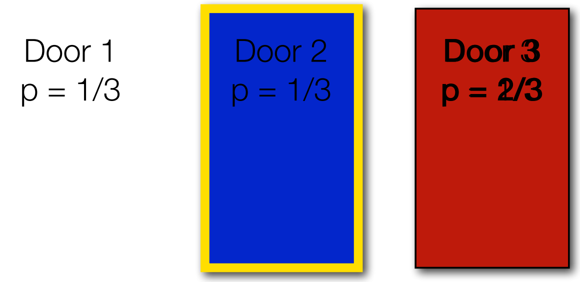
Three door problem (Monty Hall problem)

Quiz on television in the 80s - initially all doors have the same probability



Three door problem (Monty Hall problem)

Quiz on television in the 80s;



In geology: location of factory will suggest location of pollution in sampling area and colour of a rock can suggest something about composition

Three main fields in statistics

- Data analysis - data appraisal and data mining
- Probability analysis - confidence of statistical statements
- Statistical testing and modeling - process recognition and quantification

  This is a huge field with everything from basic tests to extremely complex methods for process recognition and variance analysis.

  will cover a selection to look at relations between variables, identification of processes, modeling of data and testing using data distributions

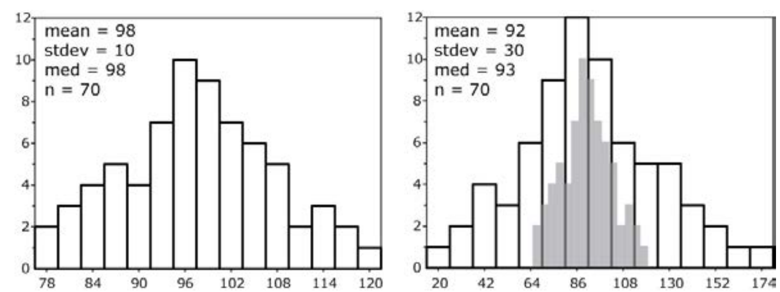
Of all these techniques, data analysis is the most important, especially in geology:

no control-group: garbage in = garbage out

Data analysis and Geostatistics - an overview

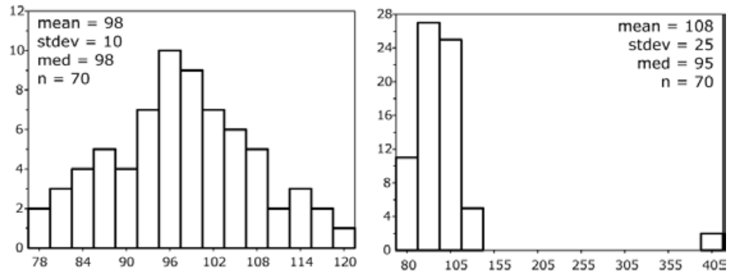
- Visualization of data distributions and data descriptors

Histograms, boxplots, summary statistics of central value + spread



Data analysis and Geostatistics - an overview

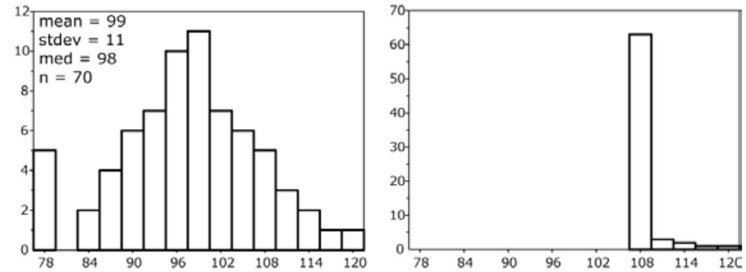
the impact of an outlier



never discard an outlier outright - could be extremely important

Data analysis and Geostatistics - an overview

the influence of the detection limit



values below the detection limit are not zero, so can not be ignored

Multi-variate techniques - an overview

- regression - quantitative description of the relation between two variables

in arid settings, the conductivity is strongly correlated with Cl content due to evaporation

can be described by Cl = a * EC + b

This allows you to estimate one variable from another or a set of others:
multiple regression - $y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + ... + c$

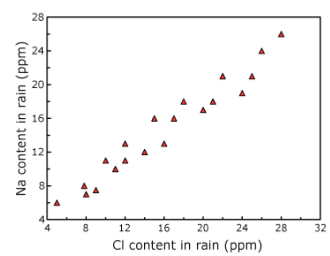
such models are for example used to estimate the viscosity, thermal conductivity, density, etc. of magmas, with a_n = fractional property and x_n = magma composition

Multi-variate techniques - an overview

- factor analysis or PCA - search for directions of most variance

similar to regression analysis, but here we do not know beforehand what relations to expect - can eventually quantify them with a regression fit

main uses: - data reduction
- process identification



although plotted in 2D this is clearly a 1D data set along a factor or principle component that is a combination of Na and Cl -> allows reduction of data

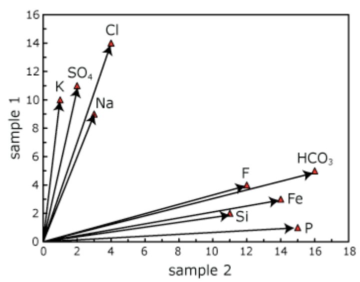
obvious in 2D, but most geochemical data sets are more than 10D

Multi-variate techniques - an overview

Process identification - looking for the trends in the data

from psychology: derive the variables of interest from trends in data for many other variables

in geology: which variables show the same behaviour? Can point to an underlying process



groundwater in an arid region of Portugal on fractured granite bedrock

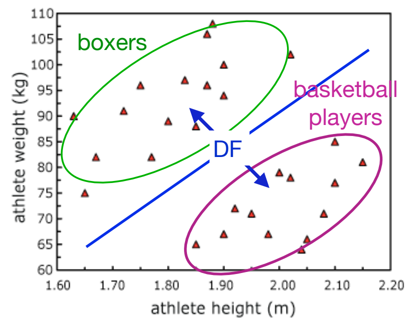
2 factors:
Si,F,P,Fe,HCO_3 - weathering
Na,K,Cl,SO_4 - evapotranspiration

can represent this also in object-space

Multi-variate techniques - an overview

- discriminant function analysis

not always looking for directions in data set, but rather a function to separate
this allows you to separate your data and classify unknowns



group of athletes: boxers and basket-ball players

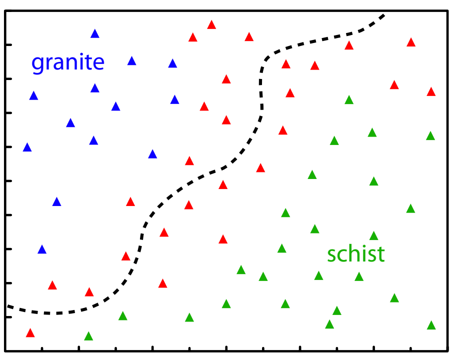
no separation in either variable, but can be separated using a combination: the discriminant function

knowing this DF, we can now apply it to a group of unknowns to classify

Multi-variate techniques - an overview

most statistical techniques can only be applied to homogenous groups:

have to separate your data set into such groups -> DFA



geological boundary mapping in tropical terrain / soil classification:

use a set of knowns to derive a discriminating function and apply this to unknowns to classify them

e.g. differentiate between schist and gneiss based on Si, U, C, X_{Mg}

Multi-variate techniques - an overview

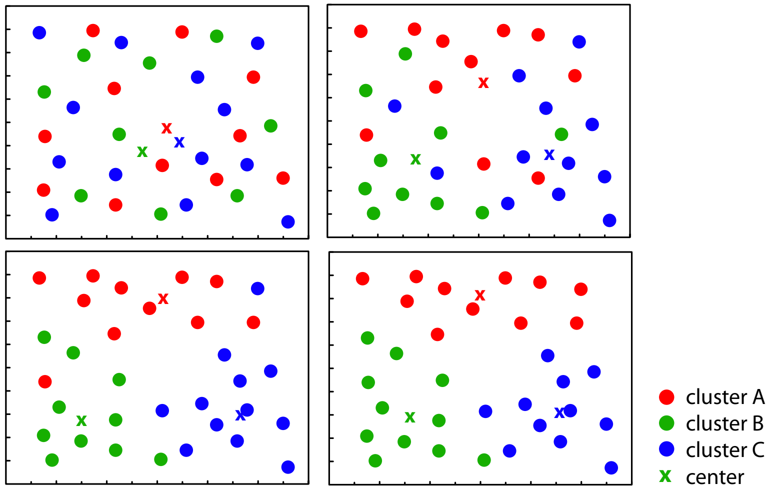
- cluster analysis - split data into homogenous groups

Discriminant function analysis can only be applied when groups are known, but in most geological examples, the groups and number of groups are not known beforehand ---> cluster analysis instead

in cluster analysis, similar samples are grouped by minimizing the deviation of each sample to its cluster mean, in multi-dimensions
these groups can generally not be visualized directly as the separation is based on a combination of many variables

the most versatile is fuzzy clustering where cluster centra are sought iteratively and cluster assignment can vary during the routine as cluster centra move around

Fuzzy c-means clustering - inverse distance



Data analysis and Geostatistics - an overview

This brief overview already covers some of the most advanced statistical techniques used in the Earth Sciences and although they are mathematically complex and have strict requirements for proper implementation, they are not difficult to understand conceptually

All strive to bring order to the data chaos by converting it into a form that can be analyzed and interpreted using the most basic statistical tools, without the loss of any information!

"Most people use statistics as a drunkard uses a lamppost: for support rather than illumination"

Geotop Short Course in Data Analysis and Geostatistics

Part 2. What are data ?

geotop

Data and nomenclature

In this section we will look at isolated variables (**univariate data**) and the tools to visualise data, describe them and quantify their characteristics.

what are data and why do we gather them ?

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

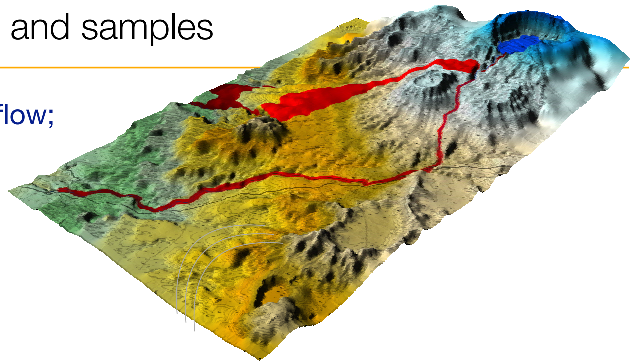
where
property can be density, length, ppm Ca, thermal conductivity
sample can be a rock, soil, water, plant

...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

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 and use that to 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

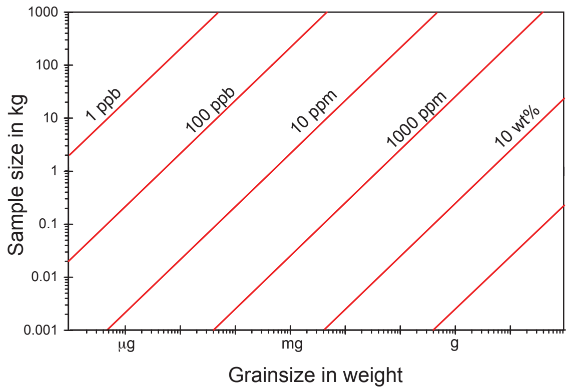
The statistics on national health suggest that 1 out of every 4 Americans, or 1 out of every 5 Canadians will suffer from a certain type of illness in their lifetime

This means at least 6 in the current Geostats cohort….

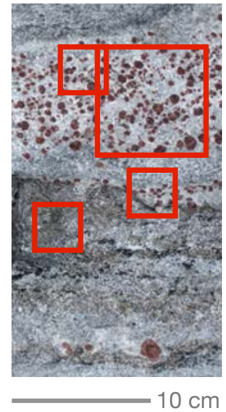
Is this reasoning correct ?

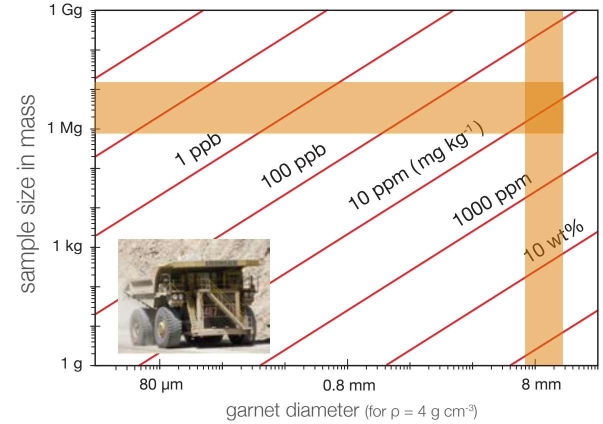
Representative samples

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



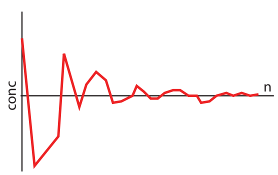


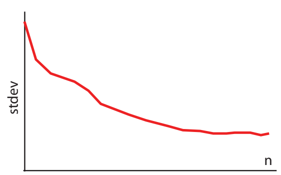
we would essentially need a sample bigger than the complete outcrop

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

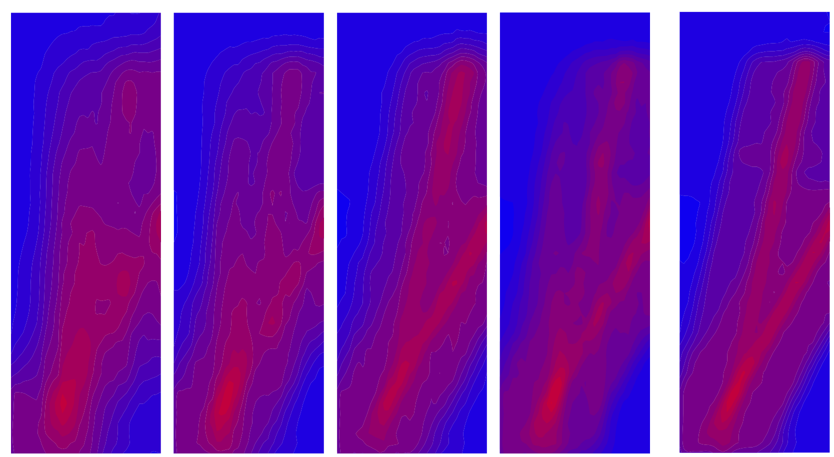




increasing number of samples -> when data characteristics no longer change -> representative sample

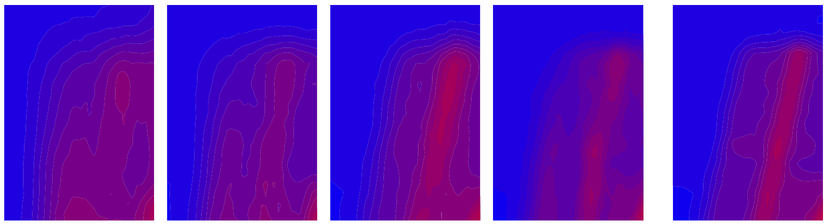
can estimate this if you know something of your samples: pilot sampling

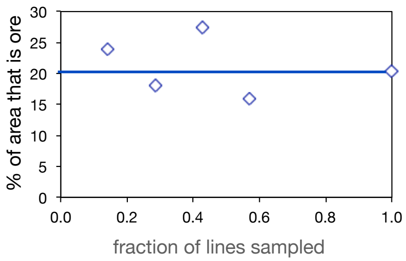
Infill drilling: vein-type deposit with halo

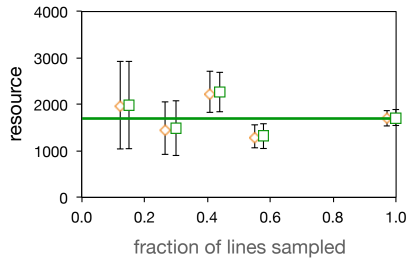


1 out of 7 lines sampled | 2 out of 7 lines sampled | 3 out of 7 lines sampled | 4 out of 7 lines sampled | 7 out of 7 lines sampled

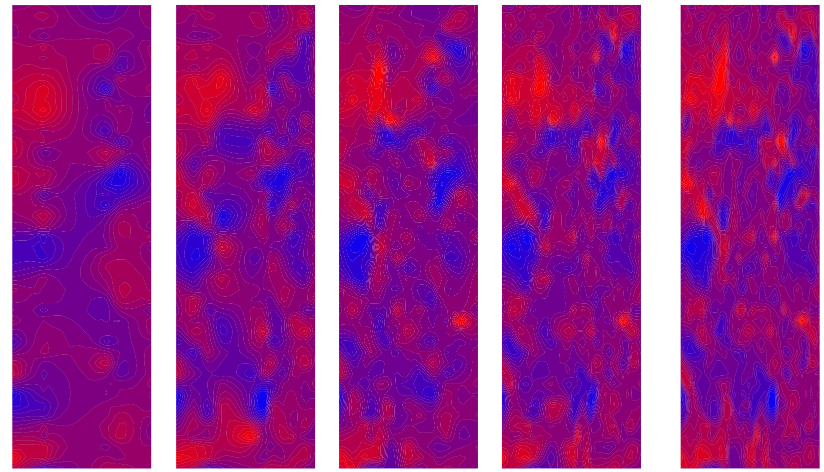
Infill drilling: vein-type deposit with halo



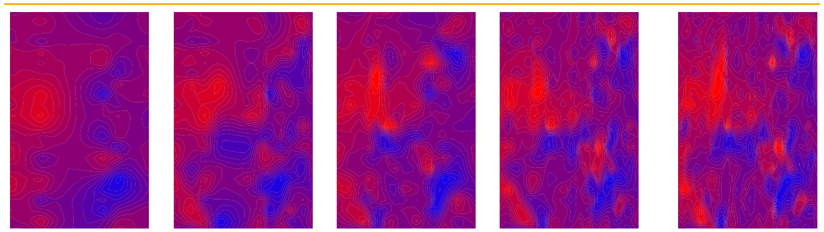


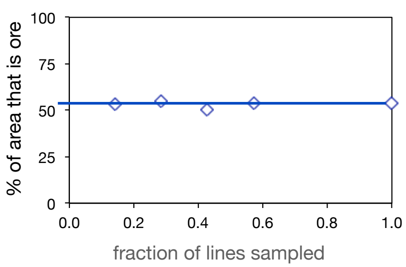


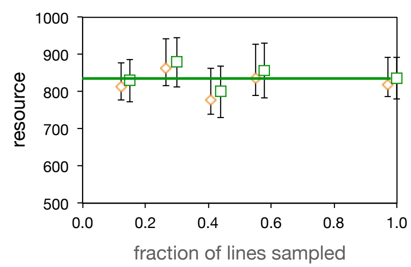
Infill drilling: disseminated deposit



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. ºF, ºC
- 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

Geotop Short Course in Data Analysis and Geostatistics

Part 3. Data distributions and descriptors

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:

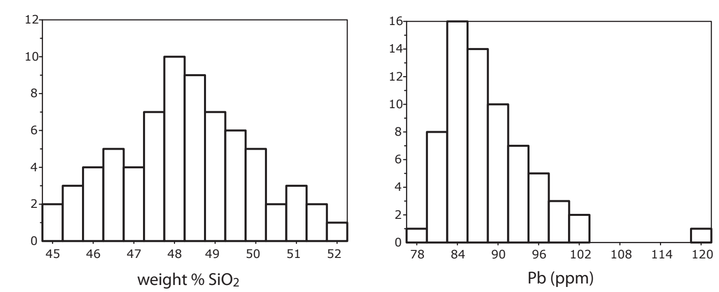
sample heterogeneity: % olivine in each sample
lava heterogeneity: layering or phase separation
analytical uncertainty: not every ion makes it to the detector

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:

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

Data distributions in the geosciences

We are all most familiar with the normal or Gaussian distribution:

But what does this mean ?

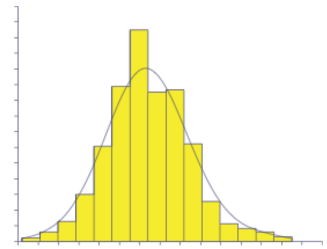
And is this what we should expect for geological and geochemical data ?

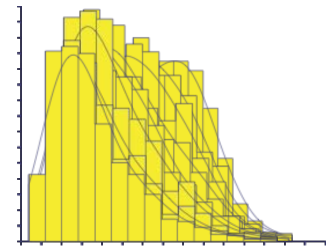
A normal distribution represents one process or one material: we can expect to find a single central value + random noise around this

Data distributions in the geosciences

We all love the normal or Gaussian distribution, but its occurrence in the geosciences is actually quite rare.

Example: a normal distribution of rock fragments is going into a crusher. Each piece has an equal probability of being fragmented. How does the size distribution evolve ?



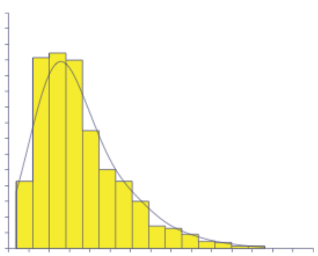


The normal distribution evolves into a lognormal distribution !

Data distributions in the geosciences

A lognormal or skewed-to-the-right distribution is much more likely for geo-data, because we commonly mix multiple sources or processes:

Multiple sediment sources in a basin, different geological basement units in a mapping area, alteration overprint on primary geology, local contamination of groundwater, etc etc



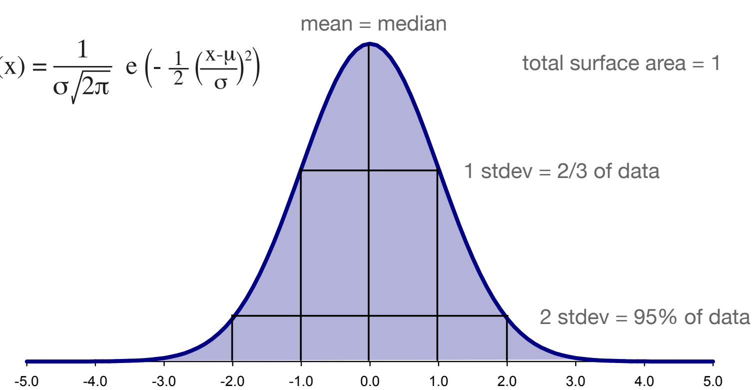
statistical parameters designed for the normal distribution do not work well in this case !

Should use robust parameters

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

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



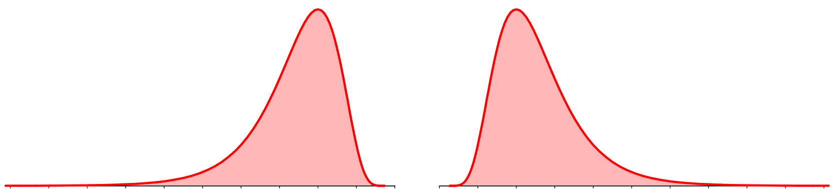
Normality in data sets

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

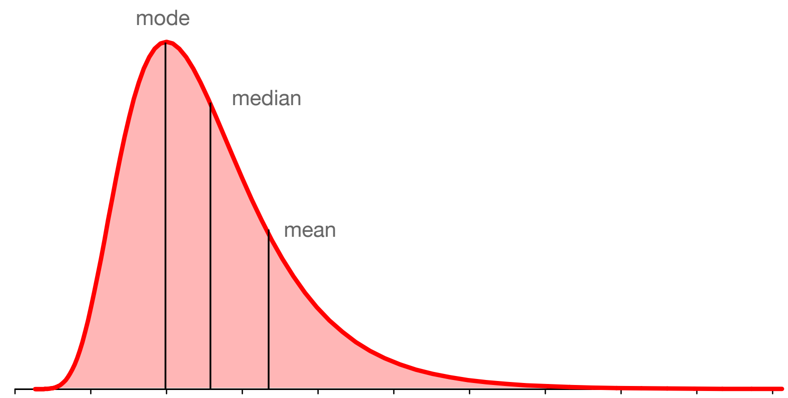
So: in most cases in geology -> deviations from normality

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)

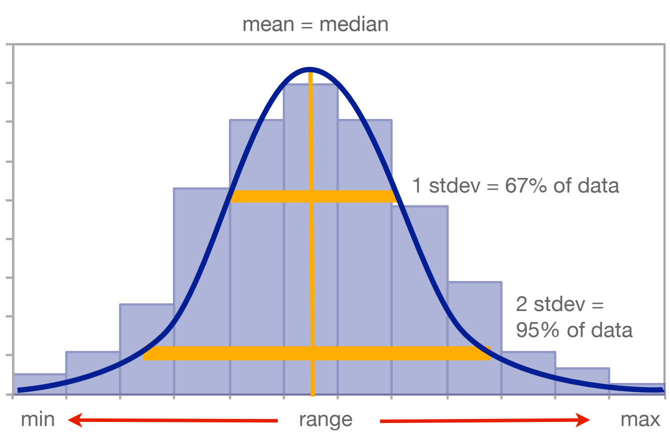


Log-normal distribution



Standardized data descriptors

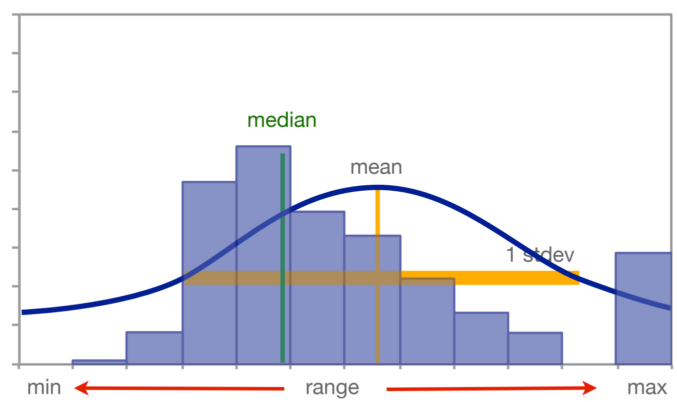
If your data describe a phenomenon with one central value and random disturbances around this value: will trend to a normal distribution



Standardized data descriptors

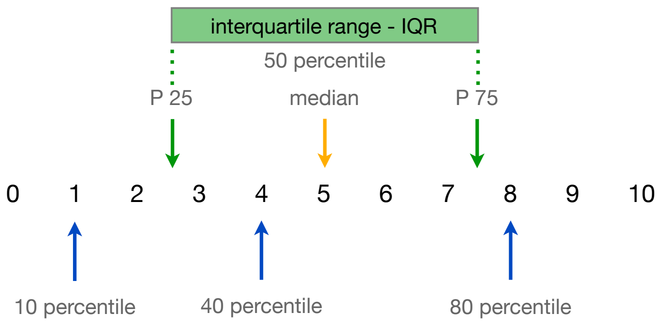
Unfortunately, many data sets are not normally distributed

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



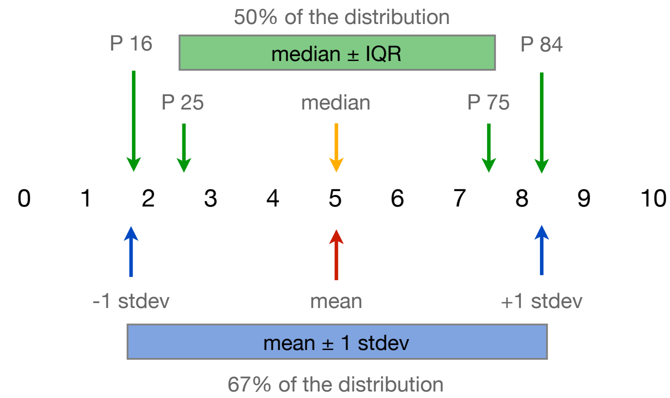
Robust descriptors: median and percentiles

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



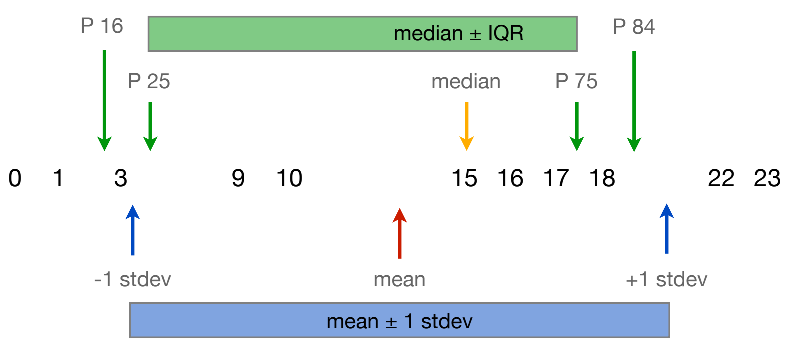
Robust descriptors: median and percentiles

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



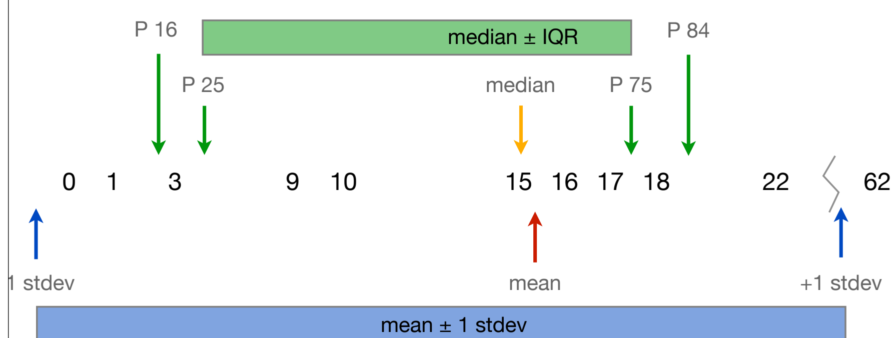
Robust descriptors: median and percentiles

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

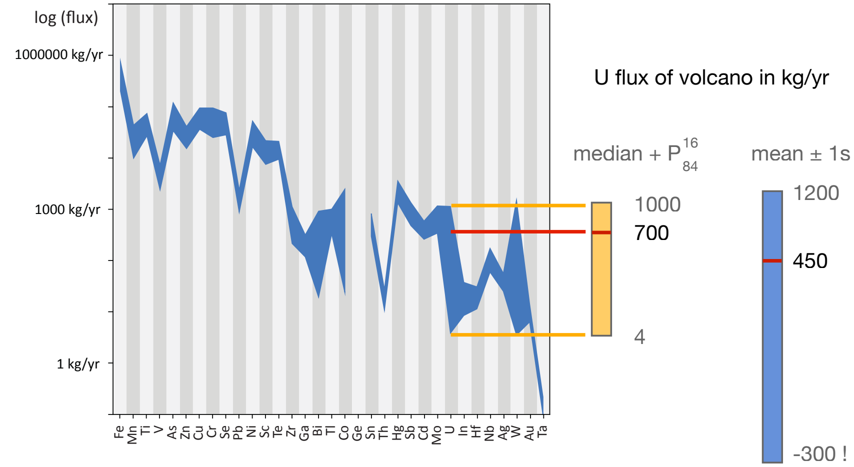
Including the stdev and P_{84}-P_{16} indicators of spread:

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 !

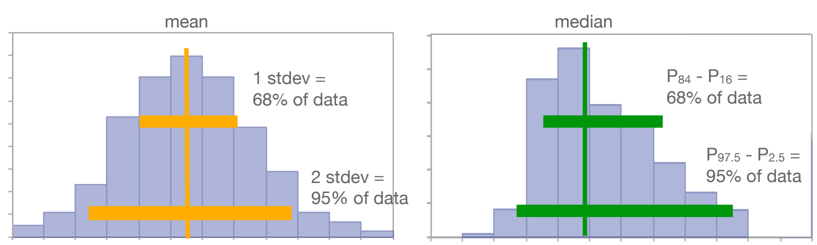
Ratio and logarithmic data in spider-diagrams



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

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.

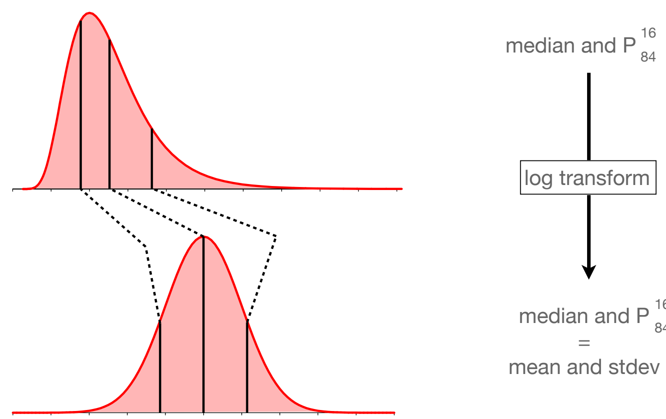


For a non-normal distribution, spread is generally asymmetric when using median + percentiles. This immediately gives information on the distribution of the data !

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	etc
23	31	39	60	

Ni	Ni	Ni	Ni	
34	23	31	39	
23	31	39	60	etc
31	39	60	55	

Bootstrapping - example with PAST

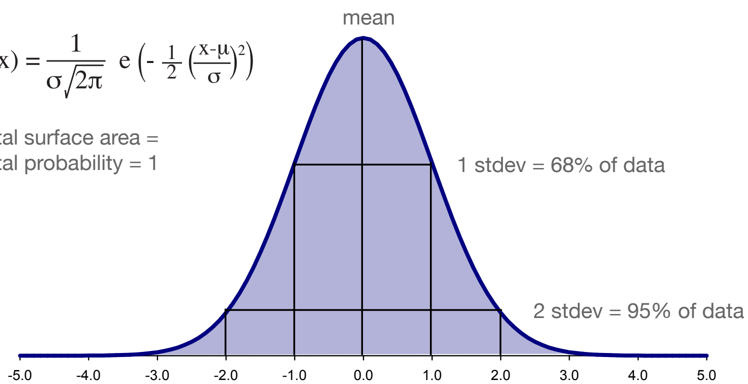
Ni
34
55
23
25
31
65
39
45
60
42
48
24
31
55
39
36
51
47
53
2500

Testing for normality

A normal data distribution is very useful, because we know its properties best

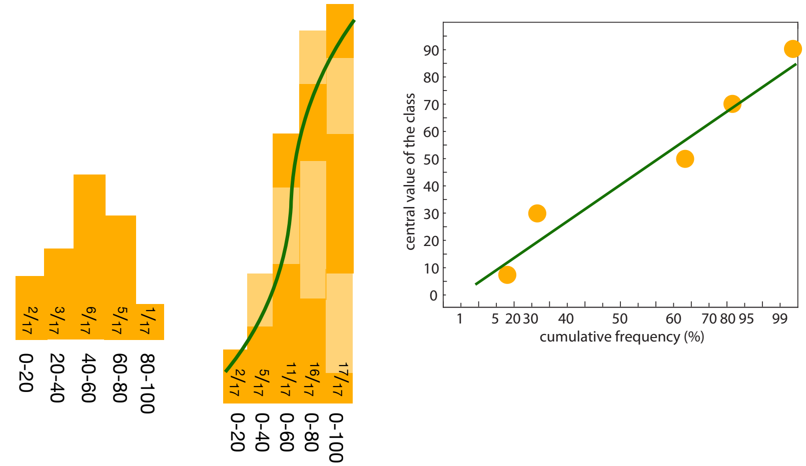
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

total surface area =
total probability = 1



Testing for normality: cumulative frequency plot

Probability plot allows for identifying deviations from normality and multi-modality



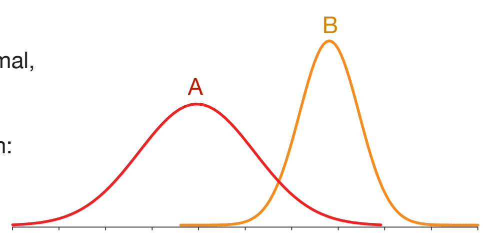
Testing for normality

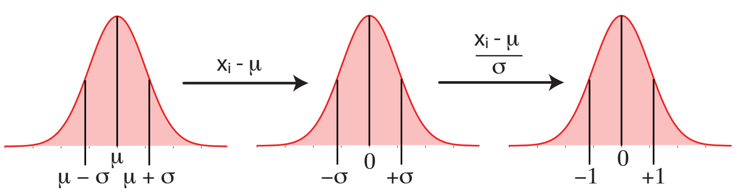
Any normally distributed data can be converted into the standard Gaussian by using a Z-score transformation:

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

convert them to standardized form:

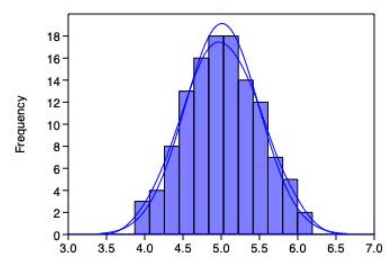
Z-score: $Z_i = (x_i - \mu) / \sigma$

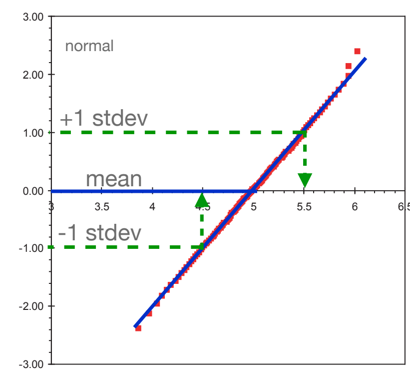




Testing for normality: cumulative frequency plot

data distribution and cumulative frequency diagrams

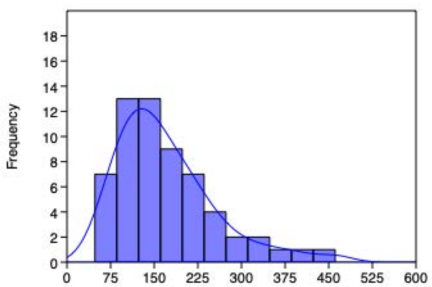


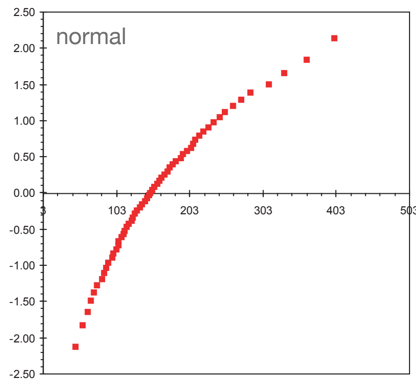


mean = 5.0 ± 0.5
median = 5.0 -0.5 +0.5

Testing for normality: cumulative frequency plot

data distribution and cumulative frequency diagrams

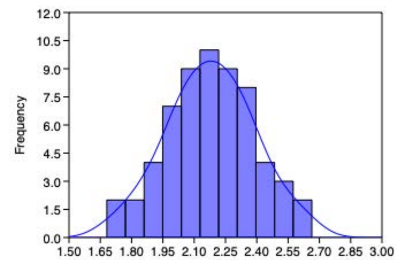


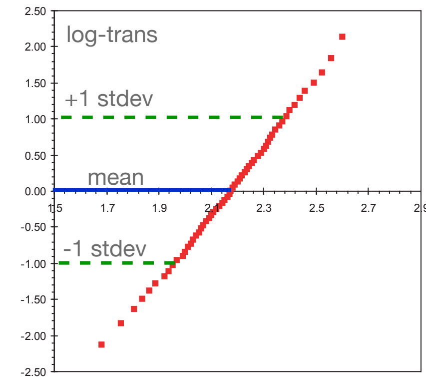


mean = 170 ± 85
median = 151 -56 +89

Testing for normality: cumulative frequency plot

data distribution and cumulative frequency diagrams





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

Deviations from normality

There are many possible deviations from a normal distribution

skewness —> robust estimators or data transformation

outliers —> need robust estimators

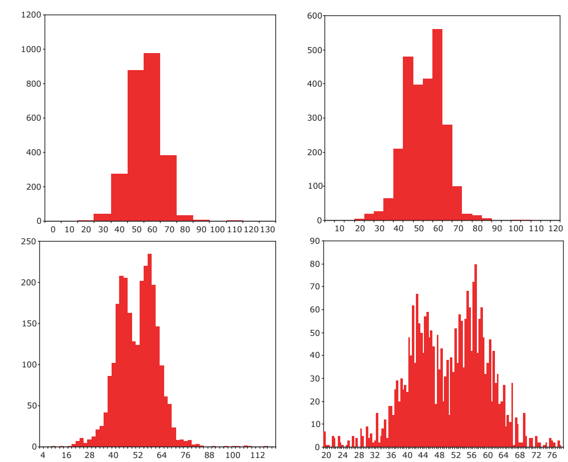
bimodality or multimodality —> data set will have to be split

kurtosis —> steepness of the distribution, can be an indication of selective data inclusion

Data distributions need to be inspected before you start analyzing your data, because these do matter

Histograms provide a quick data distribution view

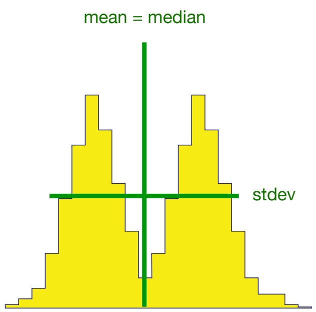
dependence of histograms on choice of classes - there are no rules

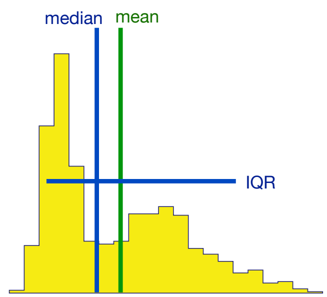


Multi-modal datasets: need to split them up

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

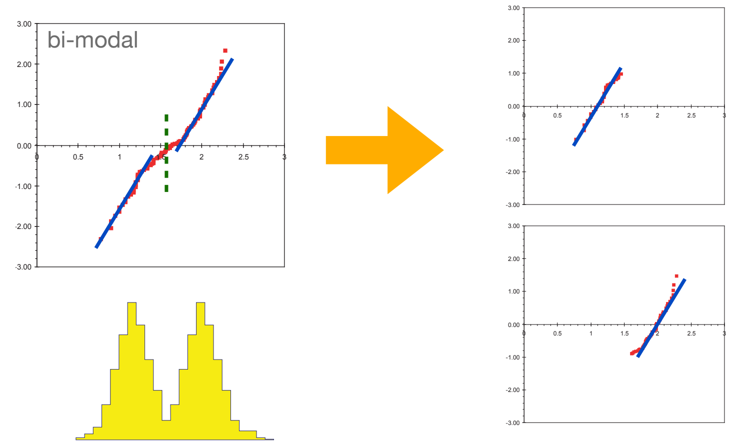
to interpret such datasets you will need to split them up, otherwise you look at a mixed signal and the mean or median you calculate is a meaningless data descriptor. Moreover, neither stdev nor IQR will capture the spread in the data.



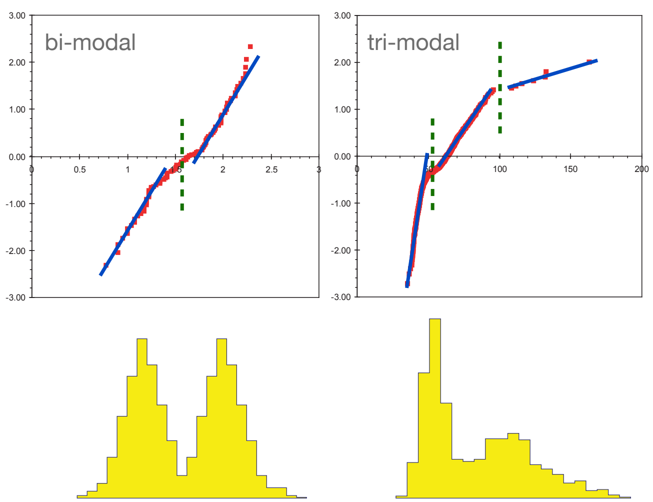


How to deal with multi-modal data sets

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

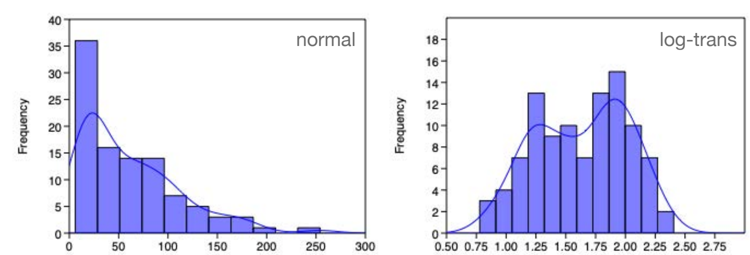


How to deal with multi-modal data sets



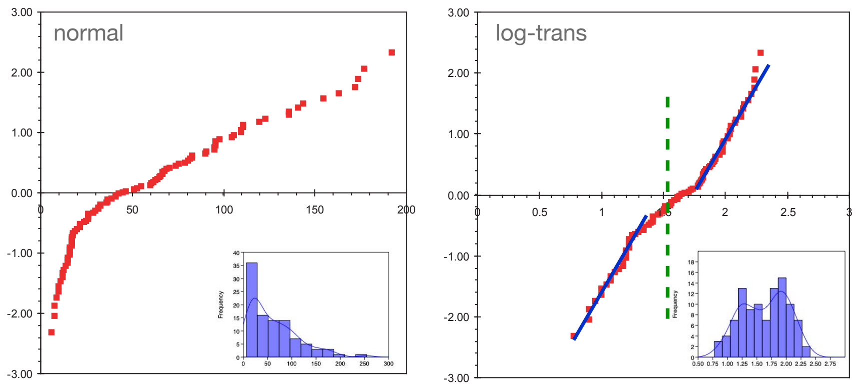
Data distributions may have multiple deviations

Dataset that is both log-normal and bimodal



Data distributions may have multiple deviations

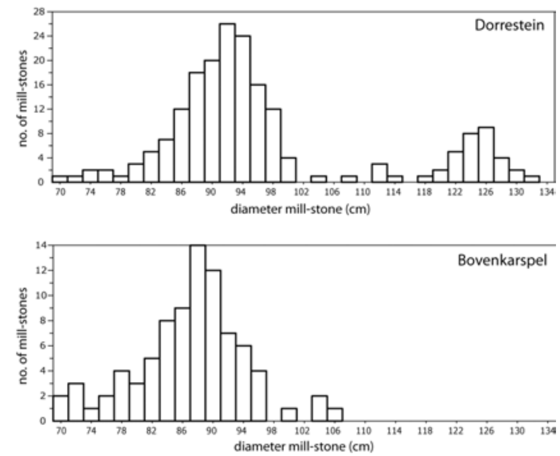
Dataset that is both log-normal and bimodal



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

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

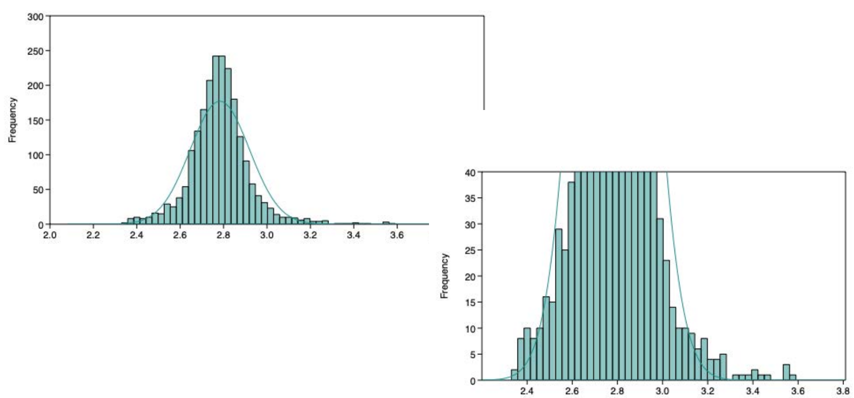
Size of discarded medieval millstones in two locations in the Netherlands



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

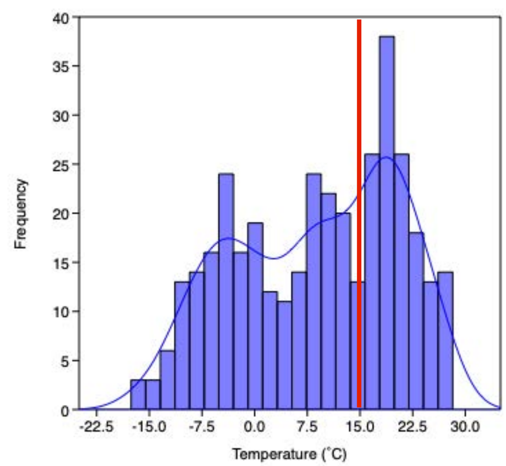
Multi-modal datasets: datasets that represent multiple samples or processes

Presence of ore deposits



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

Day 1 - topics covered

- What are data and data distributions. What data distributions can we expect for different types of geo-data.
- How can we visualize data distributions
- What parameters can be used to summarize data for different distributions