Data analysis and Geostatistics:

On the use of statistical techniques in the Earth Sciences

Practical matters

Goal of the course is to convince you that (geo-)statistical techniques provide a useful and powerful tool to analyze geological data

- the course consists of lectures (Thu 10:00-12:00) and practical sessions (Tue 9:30-12:30) in which statistical tools will be applied in exercises and to a geochemical mapping data set from BC
- book: Introduction to geological data analysis - Swan & Sandilands
- additional resources online: www.statsoft.com/textbook/stathome.html
- software: spreadsheet programs, statistics packages - PAST, specific statistics software - Fuzzy and Surfer, and spatial geochemistry tools - PAST
- examination: written midterm (20%), written final exam (40%) and group data analysis project (40%). Late submissions are not accepted.
- full course details available on: eps.mcgill.ca/~hinsberg (see teaching)

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."

Course practicalities - Corona pandemic specifics

Lectures

My preferred style of teaching is to explain the course content visually on a blackboard and to teach interactively, neither of which is suited to online teaching.

To mitigate the limitation of online teaching to some extent, I will teach this course in a virtual classroom with presentation projection and a whiteboard, and I expect you all to participate in the lectures and ask questions when topics are unclear. I will also implement Q&A sessions for topics that are well described in the textbook, rather than teaching these online.

Labs

You can only learn statistics by doing it and applying the tools we discuss in the lectures in exercises and the group project are core components of this course.

In the practical work, I strongly encourage working together and discussing the exercises and the statistical tools. If we are forced to do this online for an extended period of time, we will do the labs in a Teams setting to still have this opportunity for discussion.

Course practicalities - McGill in-person lab rules

"Preventing the spread of COVID-19 on campus: The University has planned all on-campus activities for the Winter 2022 term in a manner that follows strict safety protocols that adhere to all public health directives. Because this course contains some in-person components, students are reminded of the University's health guidelines.

All individuals on our campuses are required to wear a mask or face covering at all times when in any indoor shared space, including, but not limited, to classrooms, Teaching and Study Hubs, labs, hallways, elevators, and bathrooms. There are a very few exceptions to this rule – learn more on the Health Guidelines page. It is also essential that individuals practice physical distancing, good hand hygiene, and cough etiquette.

You are not to come to class or campus if you have any of the symptoms described on the following website.

Students will also need their McGill ID card to enter the teaching spaces"

in-person labs will start the week of the 24th at the earliest

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

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....

Another challenge is that we are particularly bad in estimating probabilities or in understanding a given probability;

If the probability of the birth of boy or girl is exactly 50%, what is the chance that a family with 4 kids has 2 boys and 2 girls?

a. $^1/_4$ b. $^3/_8$ c. $^1/_2$

What is the probability to die in a car accident in a given year?

a. 0.02% b. 1.08% c. 3.95%

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"

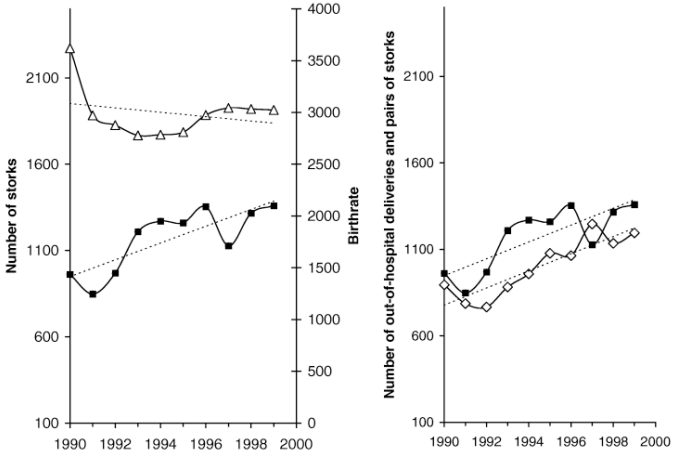
"Facts are stubborn, but statistics are more pliable"

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

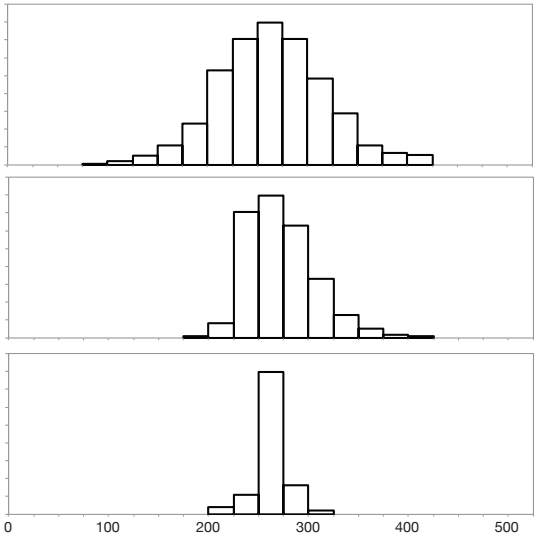
1. When analyzing data, and applying statistical tools, the simpler technique is generally the best, because it makes for the strongest point and is the easiest to explain and defend to your audience

   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"

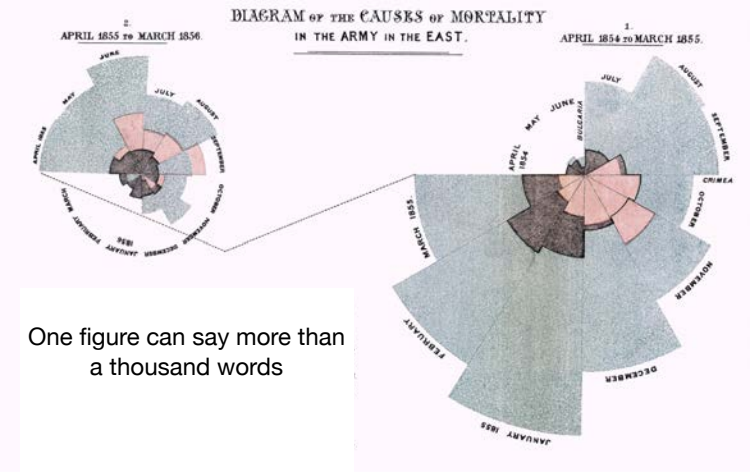
Graphical representation of data



"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"



One figure can say more than a thousand words

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

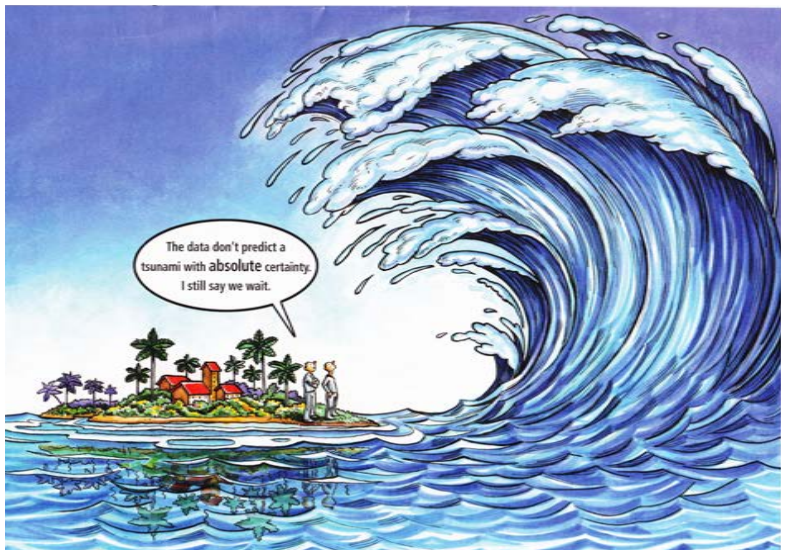
  This is a field in itself and will be limited here to its control on the confidence level of statistical statements = "statistical proof"

  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

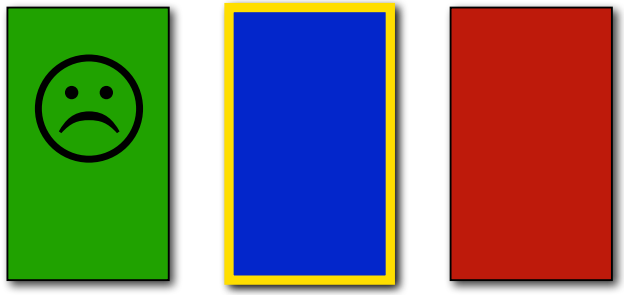
  In geochemistry generally 95% is chosen: in 1 out of 20 cases we’re wrong!
  In oil exploration closer to 10%, whereas in space exploration 99.99%.

  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)

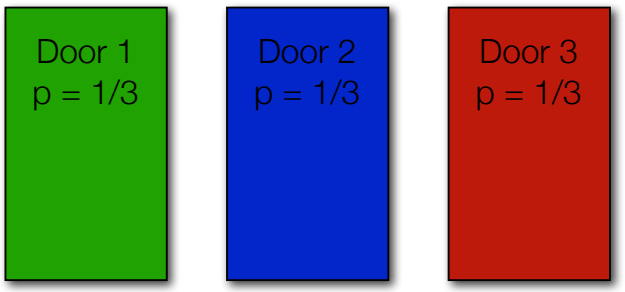
Quiz on television in the 80s;



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

Three door problem (Monty Hall problem)

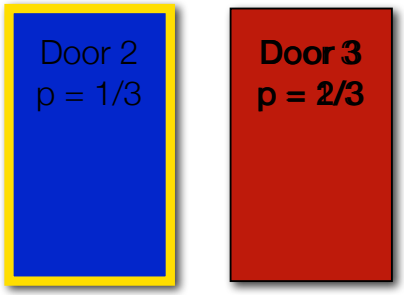
Quiz on television in the 80s - initially all doors have the same probability



Three door problem (Monty Hall problem)

Quiz on television in the 80s;

Door 1
p = 1/3



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

My philosophy:

no need to be an expert in statistics, but you need to know what's available, where to find it and how to apply it

simple techniques are always to be preferred (also easiest to convince your audience)

Qualitative overview of the most important statistical techniques in geology

- mean, median and mode - a general description of your data set

  however; means and outliers don't mix -> use medians

  hours of tv per week: 2,4,6,8,10 mean = 6, med = 6
  2,4,6,8,60 mean = 16, med = 6

Data analysis and Geostatistics - an overview

- spread of the data - one value is not enough to describe a data set

  e.g. spread in jeans sizes

  most common value for spread: standard deviation or variance
  however, stdev is calculated using the mean -> sensitive to outliers

  use a percentile range instead: P_{10} - P_{90} or P_{25} - P_{75}

  median and interquartile range are *robust*, mean and stdev are not

  a central value + spread only describe data sets that represent one process, value or event:

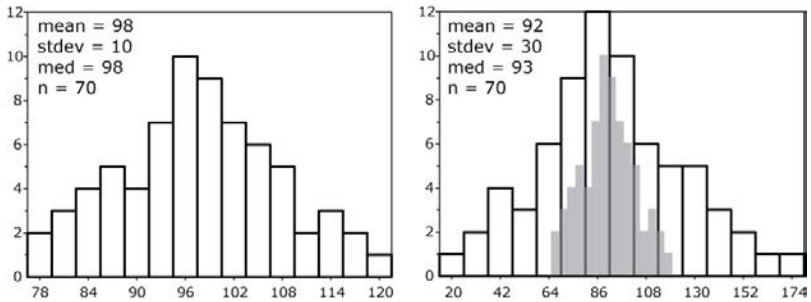
  % feldspar in different samples of a granite - there is a characteristic value for the granite, but each sample with have a slightly different number of crystals

Data analysis and Geostatistics - an overview

- histograms and boxplots - visualization of data sets

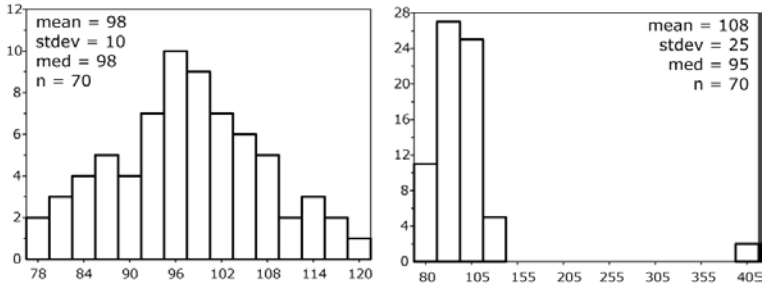
gives much more information than just central value + spread

most common plot in statistics: histogram - counts vs. value classes



Data analysis and Geostatistics - an overview

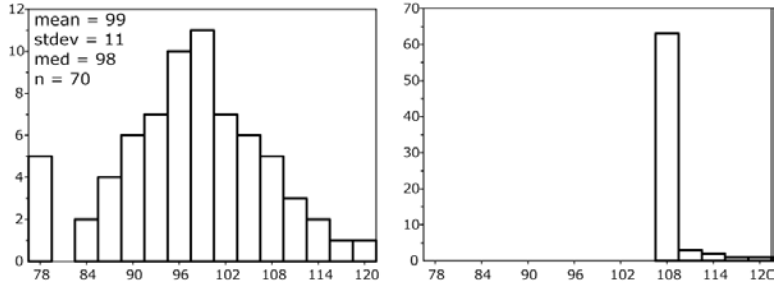
the impact of an outlier



never discard an outlier outright - could be extremely important

Data analysis and Geostatistics - an overview

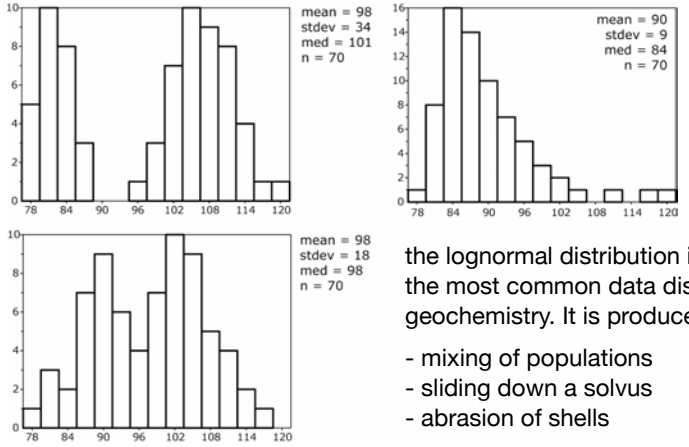
the influence of the detection limit



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

Data analysis and Geostatistics - an overview

data distributions - modality and skewness



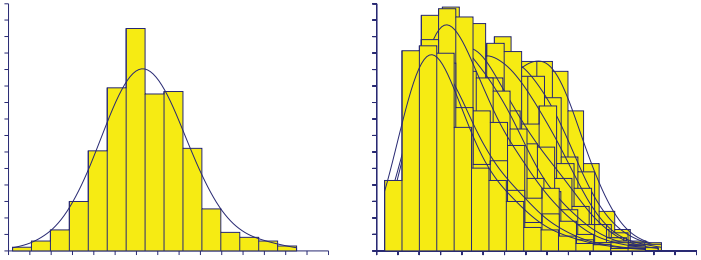
the lognormal distribution is probably the most common data distribution in geochemistry. It is produced by;

- mixing of populations
- sliding down a solvus
- abrasion of shells

Skewed data and the lognormal distribution

Abrasion of shells in the surf of a beach, or rock fragments going into the crusher of a mill: start out with a normal distribution

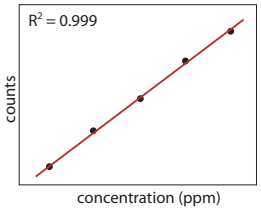
There is a certain possibility that a given shell or fragment breaks and samples can only become smaller: trends to a log-normal distribution



Data analysis and Geostatistics - an overview

This is all very basic stuff, but it can be quite powerful and may be all you need

Moreover, the distribution of data is crucially important for more advanced statistical analysis, most of which require your data to be normally distributed. And geological data rarely are.....



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 - $y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + ... + c$

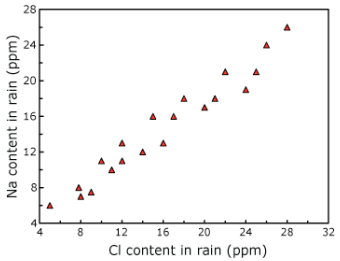
such models are for example used to estimate the viscosity, thermal conductivity, density, etc. of magmas, with a_n = fractional property and x_n = magma composition

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

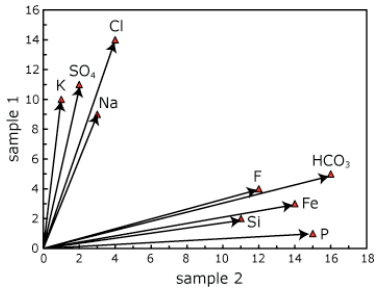
obvious in 2D, but most geochemical data sets are more than 10D

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

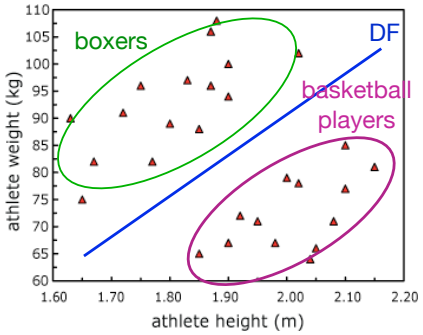
2 factors:
Si,F,P,Fe,HCO_3 - weathering
Na,K,Cl,SO_4 - evapotranspiration

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

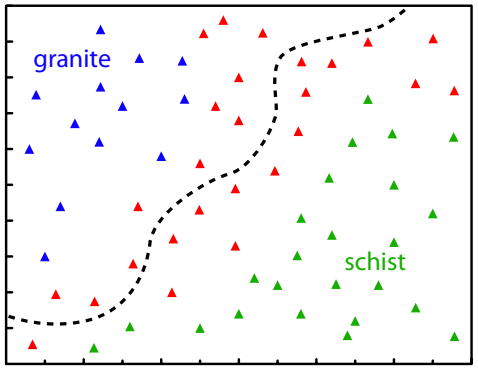
no separation in either variable, but can be separated using a combina-tion: the discriminant function

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 unknows to classify them

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

Multi-variate techniques - an overview

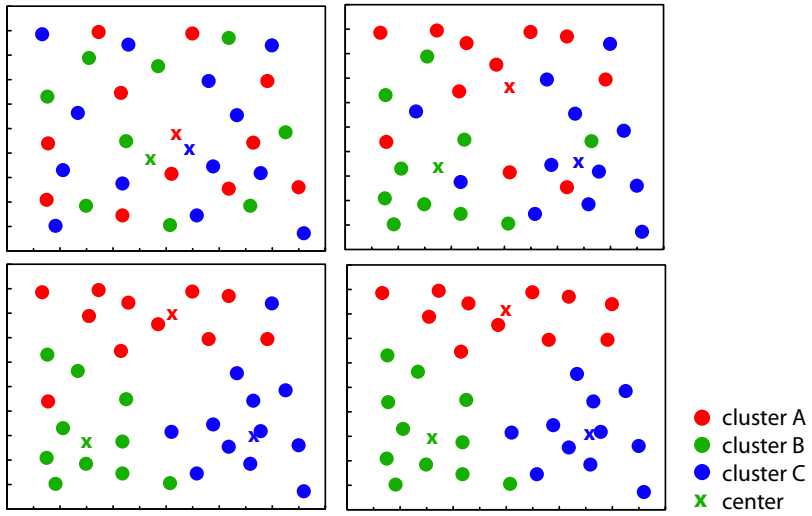
- cluster analysis - split data into homogenous groups

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"