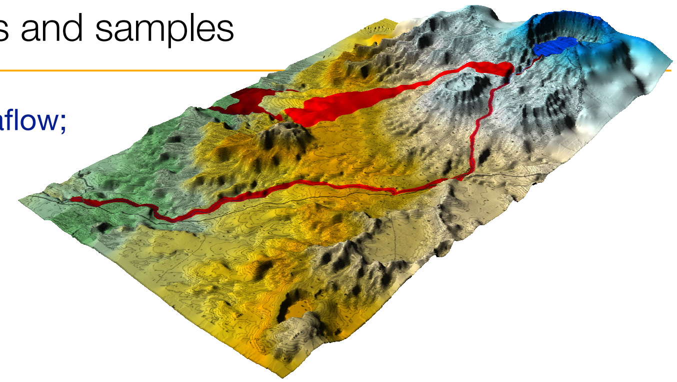
Data analysis and Geostatistics - lecture VIII

A quick review

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

Populations and samples

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

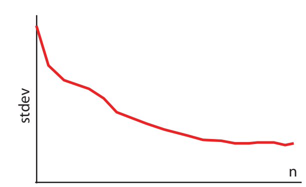
This model controls:
- Car crash-test dummies
- Office temperature
- Police officer’s safety vests
- Gas masks
- Height of desks, shelves, cupboards, etc
- Exposure limits for chemicals
- Size of gadgets, including phones
- Size of tools, bricks, notebooks, etc etc etc

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

Populations and samples: representative samples

In geology we generally no longer have the population at our disposal and it is therefore critical to ensure that your sample is representative

Samples are estimates of the properties of the host population

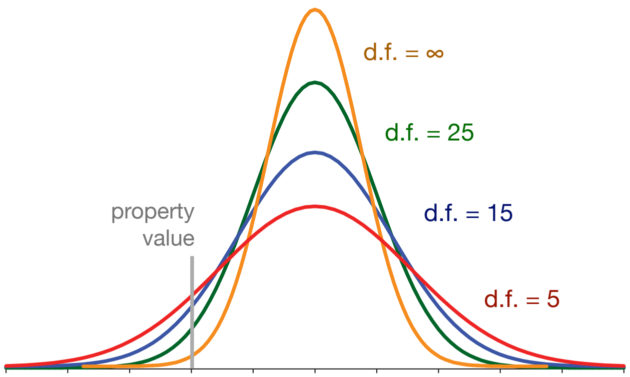


Increasing the number of samples leads to better estimates of the true population values and characteristics. This is commonly done by first conducting a pilot study and when the characteristics no longer change: representative sample

Sample size and probabilities

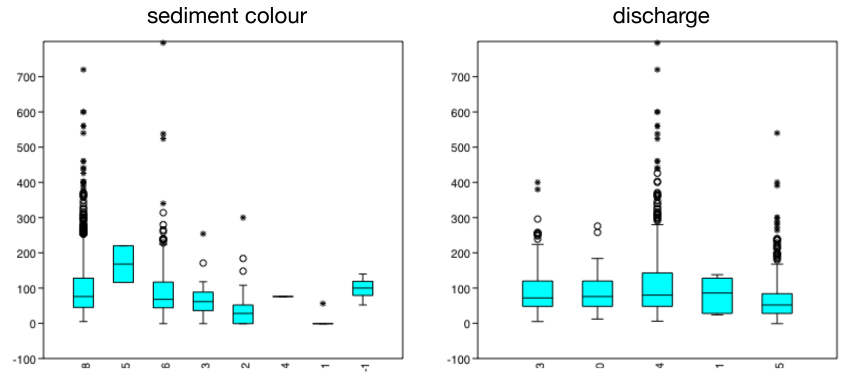
Sample size is also a critical parameter in evaluating the probability of encountering a certain value (data value, test statistic, etc)

For small numbers of samples, the uncertainty on the property you are interested in is larger: it is a less precise estimate of the characteristics of the population



Comparing properties - visually

One of the commonest uses of statistics is to determine whether two things are the same: two groups of samples, two regression models, two geological units, etc



Comparing properties - testing

In statistical testing, you quantify the confidence of your interpretations/statements

1. Define a hypothesis to test: a mutually exclusive H_0 and H_A

   in statistics only a hypothesis rejection is a strong statement: have to choose your hypothesis carefully (example: white swans - black swans)

2. Decide on a confidence level

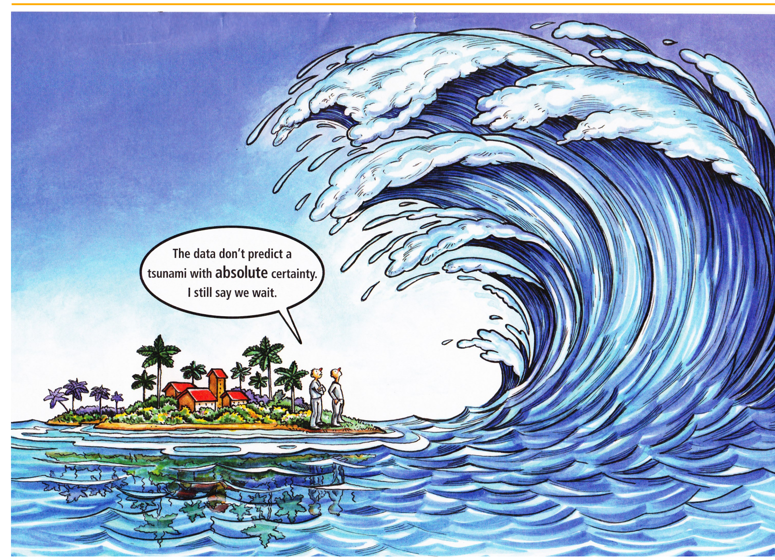
   you cannot be 100% certain, because the chance of an unlikely event is small, but never zero: have to select a level of confidence that fits your research question

   at α = 5%, you accept to reach the wrong conclusion in 1 out of 20 cases
   at α = 2%, it is 1 out of 50 cases

3. Determine the probability distribution to test against

   this is the expected behaviour of the property that you are testing and provides the required probabilities to determine whether to accept or reject your H_0

Confidence levels



Union of concerned scientists

Statistical testing - type I and type II errors

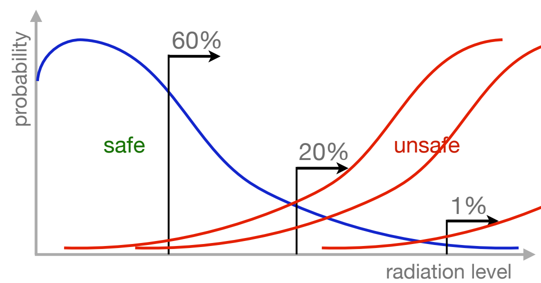
From the example midterm: The nuclear safety commission uses a very high alpha value when determining whether a safe radiation level is exceeded. Why?

hypotheses:

H_0: radiation level = safe

H_A: radiation level ≠ safe

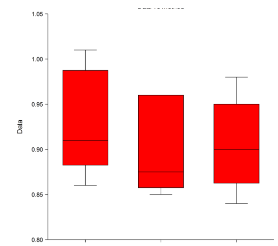
	safe	non-safe
reject H_0	false alarm	alarm
accept H_0	OK	disaster



ANOVA analysis

sample	method	value
1	ICP-MS	0.96
2	ICP-MS	0.96
3	ICP-MS	0.85
4	ICP-MS	0.86
5	ICP-MS	0.86
6	ICP-MS	0.89
1	INAA	0.94
2	INAA	0.98
3	INAA	0.87
4	INAA	0.84
5	INAA	0.87
6	INAA	0.93
1	AAS	0.98
2	AAS	1.01
3	AAS	0.86
4	AAS	0.9
5	AAS	0.89
6	AAS	0.92

Is there a difference between the methods and/or between samples: Can be evaluated sequentially (one-way ANOVA) or simultaneously (two-way ANOVA or MANOVA)



between the methods

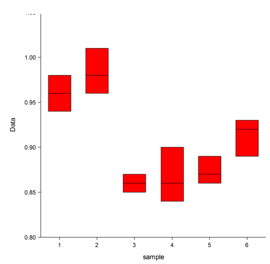
Analysis of Variance Table and F-Test

Model Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Reject Equal Means? (α=0.05)	Power (α=0.05)
Between (method)	2	0.002877778	0.001438889	0.4971	0.61796	No	0.11665
Within (Error)	15	0.04341667	0.002894444				
Adjusted Total	17	0.04629444					
Total	18						

ANOVA analysis

sample	method	value
1	ICP-MS	0.96
2	ICP-MS	0.96
3	ICP-MS	0.85
4	ICP-MS	0.86
5	ICP-MS	0.86
6	ICP-MS	0.89
1	INAA	0.94
2	INAA	0.98
3	INAA	0.87
4	INAA	0.84
5	INAA	0.87
6	INAA	0.93
1	AAS	0.98
2	AAS	1.01
3	AAS	0.86
4	AAS	0.9
5	AAS	0.89
6	AAS	0.92

Is there a difference between the methods and/or between samples: Can be evaluated sequentially (one-way ANOVA) or simultaneously (two-way ANOVA or MANOVA)



between the samples

Analysis of Variance Table and F-Test

Model Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Reject Equal Means? (α=0.05)	Power (α=0.05)
Between (sample)	5	0.04082778	0.00816556	17.9244	0.00003	Yes	1.00000
Within (Error)	12	0.005466667	0.0004555556				
Adjusted Total	17	0.04629444					
Total	18						

ANOVA analysis

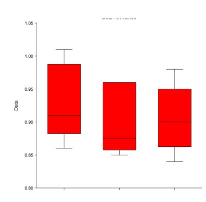
sample	method	value
1	ICP-MS	0.96
2	ICP-MS	0.96
3	ICP-MS	0.85
4	ICP-MS	0.86
5	ICP-MS	0.86
6	ICP-MS	0.89
1	INAA	0.94
2	INAA	0.98
3	INAA	0.87
4	INAA	0.84
5	INAA	0.87
6	INAA	0.93
1	AAS	0.98
2	AAS	1.01
3	AAS	0.86
4	AAS	0.9
5	AAS	0.89
6	AAS	0.92

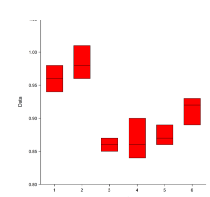
Is there a difference between the methods and/or between samples: Can be evaluated sequentially (one-way ANOVA) or simultaneously (two-way ANOVA or MANOVA)

Analysis of Variance Table for Data

Source Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level
A: sample	5	0.04082778	0.008165556	31.54	0.000008*
B: method	2	0.002877778	0.001438889	5.56	0.023821*
AB	10	0.002588889	0.0002588889		
S	0	0			
Total (Adjusted)	17	0.04629444			
Total	18				

* Term significant at alpha = 0.05





ANOVA analysis

sample	method	value
1	ICP-MS	0.96
2	ICP-MS	0.96
3	ICP-MS	0.85
4	ICP-MS	0.86
5	ICP-MS	0.86
6	ICP-MS	0.89
1	INAA	0.94
2	INAA	0.98
3	INAA	0.87
4	INAA	0.84
5	INAA	0.87
6	INAA	0.93
1	AAS	0.98
2	AAS	1.01
3	AAS	0.86
4	AAS	0.9
5	AAS	0.89
6	AAS	0.92

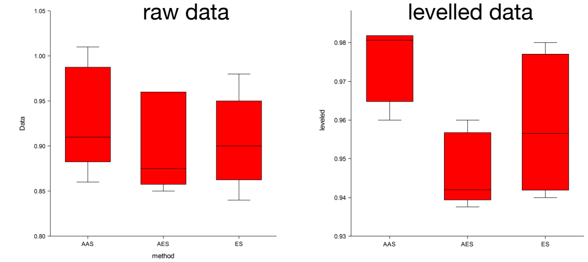
Is there a difference between the methods and/or between samples: Can be evaluated sequentially (one-way ANOVA) or simultaneously (two-way ANOVA or MANOVA)

Sample	Method	Date				Z-score	leveled
1	ICP-MS	0.96	mean	0.96		0.00	0.96000
1	INAA	0.94	stdev	0.02		-1.00	0.94000
1	AAS	0.98				1.00	0.98000
2	ICP-MS	0.96	mean	0.98		-0.93	0.94146
2	INAA	0.98	stdev	0.03		-0.13	0.95735
2	AAS	1.01				1.06	0.98119
3	ICP-MS	0.85	mean	0.86		-1.00	0.94000
3	INAA	0.87	stdev	0.01		1.00	0.98000
3	AAS	0.86				0.00	0.96000
4	ICP-MS	0.86	mean	0.87		-0.22	0.95564
4	INAA	0.84	stdev	0.03		-0.87	0.94254
4	AAS	0.90				1.09	0.98182
5	ICP-MS	0.86	mean	0.87		-0.87	0.94254
5	INAA	0.87	stdev	0.02		-0.22	0.95564
5	AAS	0.89				1.09	0.98182
6	ICP-MS	0.89	mean	0.91		-1.12	0.93758
6	INAA	0.93	stdev	0.02		0.80	0.97601
6	AAS	0.92				0.32	0.96641

ANOVA analysis

sample	method	value
1	ICP-MS	0.96
2	ICP-MS	0.96
3	ICP-MS	0.85
4	ICP-MS	0.86
5	ICP-MS	0.86
6	ICP-MS	0.89
1	INAA	0.94
2	INAA	0.98
3	INAA	0.87
4	INAA	0.84
5	INAA	0.87
6	INAA	0.93
1	AAS	0.98
2	AAS	1.01
3	AAS	0.86
4	AAS	0.9
5	AAS	0.89
6	AAS	0.92

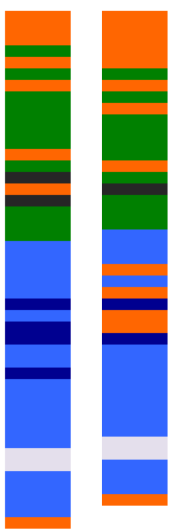
Is there a difference between the methods and/or between samples: Can be evaluated sequentially (one-way ANOVA) or simultaneously (two-way ANOVA or MANOVA)



Analysis of Variance Table and F-Test

Model Term	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Reject Equal Means? (α=0.05)	Power (α=0.05)
Between (method)	2	0.002541473	0.001270736	8.4404	0.00350	Yes	0.92218
Within (Error)	15	0.002258309	0.0001505539				
Adjusted Total	17	0.004799782					
Total	18						

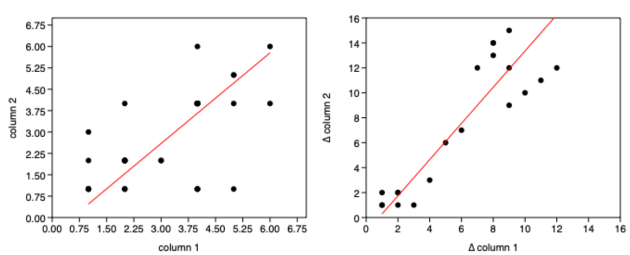
Timeseries analysis



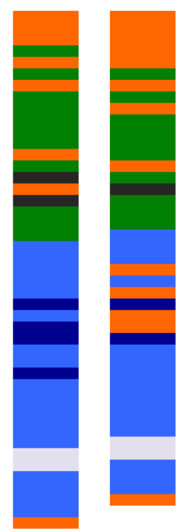
Given these two stratigraphic columns;

- Are these time sequences correlated ?
- Are the lithology transitions correlated ?

- Does each column have a non-random sequence ?



Timeseries analysis



Given these two stratigraphic columns;

- Are these time sequences correlated ?
- Are the lithology transitions correlated ?

- Does each column have a non-random sequence ?

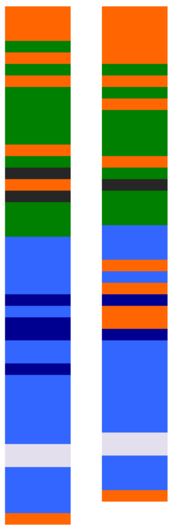
Column 1

	1	2	3	4	5	6
1	2	4	1	0	0	0
2	3	**6**	1	1	0	0
3	1	1	0	0	0	0
4	0	0	0	**12**	3	1
5	0	0	0	3	1	0
6	0	0	0	1	0	1

Column 2

	1	2	3	4	5	6
1	5	4	0	1	2	0
2	3	**5**	1	1	0	0
3	0	1	0	0	0	0
4	3	0	0	**11**	0	1
5	1	0	0	1	0	0
6	0	0	0	1	0	1

Timeseries analysis



Given these two stratigraphic columns;

- Are these time sequences correlated ?
- Are the lithology transitions correlated ?

- Does each column have a non-random sequence ?

Column 1

	1	2	3	4	5	6
1	2	4	1	0	0	0
2	3	**6**	1	1	0	0
3	1	1	0	0	0	0
4	0	0	0	**12**	3	1
5	0	0	0	3	1	0
6	0	0	0	1	0	1

Expected

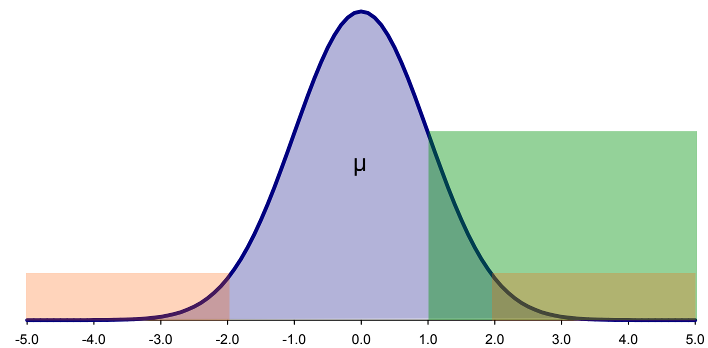
	1	2	3	4	5	6
1	3	3	0	**4**	1	1
2	3	**2**	0	**4**	0	0
3	0	0	0	0	0	0
4	**4**	**4**	0	**5**	1	1
5	1	0	0	1	0	0
6	1	0	0	1	0	0

$p = 0.003$

One-sided and two-sided probability distribution

Given a mean cancer rate in Montreal,

1. What is the probability of finding an occurrence of more than 1 stdev **higher** than the mean cancer rate —> one-sided (only higher values count)
2. What is the probability of finding an occurrence more than 2 stdev **away** from the mean —> two-sided (both higher and lower values count)



One-sided and two-sided t-test

Testing the equality of two sample sets with the student-t test:

1. H_0: $\bar{x}_1 = \bar{x}_2$ H_A: $\bar{x}_1 \neq \bar{x}_2$ This is a two-sided test, because we can reject H_0 when $\bar{x}_1 - \bar{x}_2 < 0$ or $\bar{x}_1 - \bar{x}_2 > 0$
2. H_0: $\bar{x}_1 > \bar{x}_2$ H_A: $\bar{x}_1 \leq \bar{x}_2$ This is a one-sided test, because we can reject H_0 only if $\bar{x}_1 - \bar{x}_2 \leq 0$