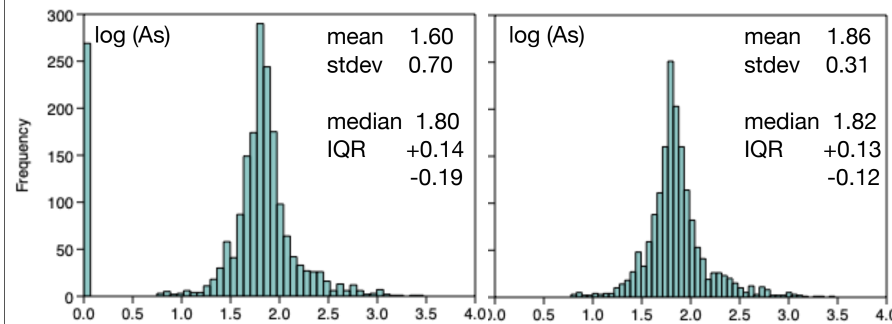
# Data analysis and Geostatistics

Short Course on the use of statistical techniques in the geosciences

Vincent van Hinsberg • McGill University

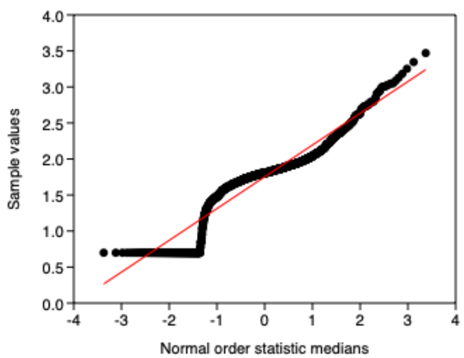
## Missing values

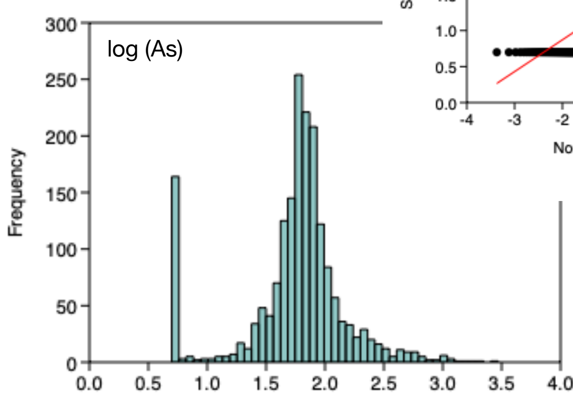
d.l. values are assigned a 0 | d.l. values are removed




## Missing values

in database as ²/₃ x d.l.

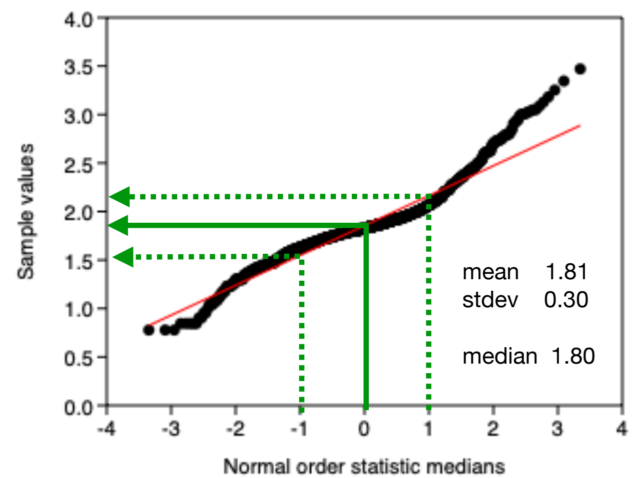


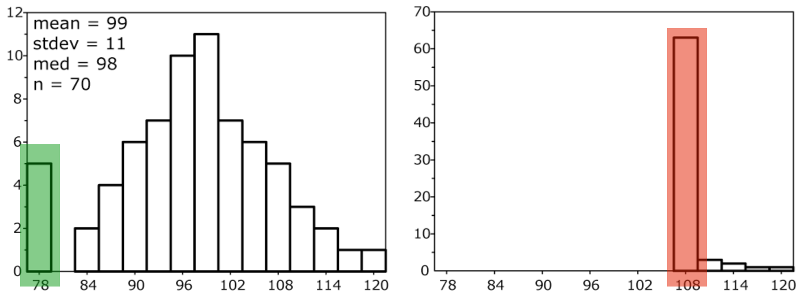

mean 1.75
stdev 0.44

median 1.80
IQR +0.14
-0.16

## Missing values



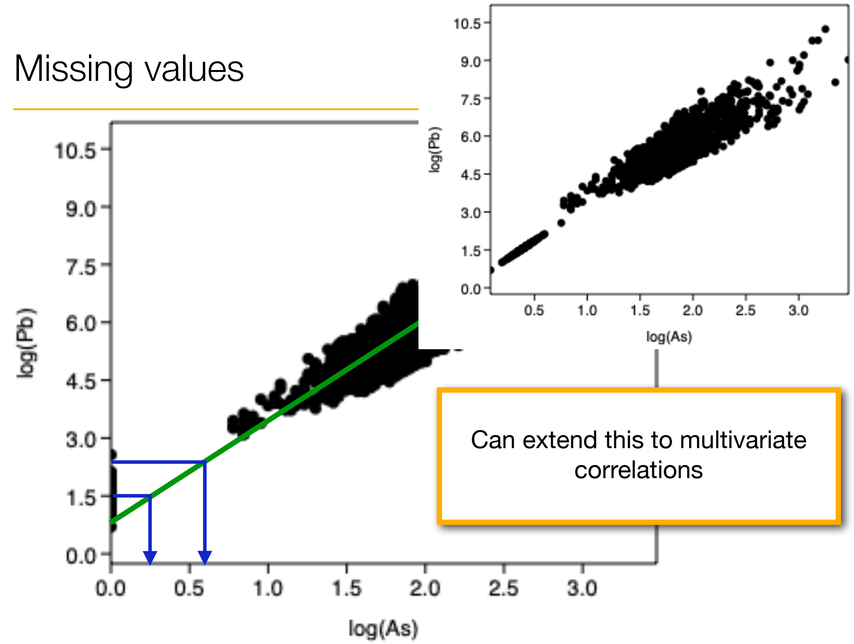
## Missing values




acceptable

hopeless

## Missing values



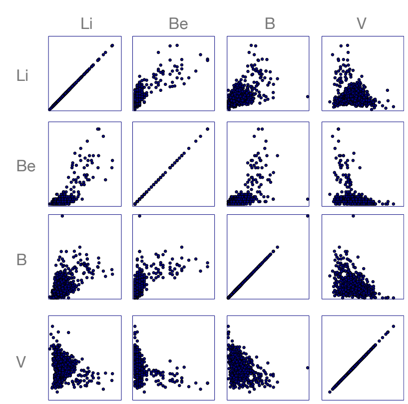

Geotop Short Course in Data Analysis and Geostatistics

## Testing the significance of the correlation coefficient

geotop

## Correlation coefficients - indicator of covariance

Pearson —> normally distributed data, Spearman —> all others



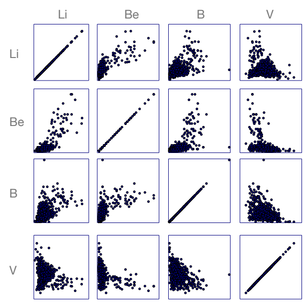

| | Li | logBe | B | logV |
|---|---|---|---|---|
| Li | 1 | 0.7 | 0.5 | -0.3 |
| logBe | 0.7 | 1 | 0.6 | -0.5 |
| B | 0.5 | 0.6 | 1 | -0.4 |
| logV | -0.3 | -0.5 | -0.4 | 1 |

# Correlation coefficients matrices - significance

## But are these r values meaningful?

In statistical terms: are they significantly different from r = 0

there will be a critical r value above which it is significant




| | Li | logBe | B | logV |
|---|---|---|---|---|
| Li | 1 | 0.7 | 0.5 | -0.3 |
| logBe | 0.7 | 1 | 0.6 | -0.5 |
| B | 0.5 | 0.6 | 1 | -0.4 |
| logV | -0.3 | -0.5 | -0.4 | 1 |

# Statistical testing: the student-t test of r

## What values of r are meaningful for a given confidence level

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

When calculated t > critical t
significant correlation

t depends on the number of samples and the desired confidence interval

- the more samples, the smaller the uncertainty on your r-value
less uncertainty on deciding whether something is significant

- the confidence level governs how strong your statements will be:
  95% - wrong conclusion in 1 out of 20 cases
  98% - wrong in 1 out of 50 cases

**Have entered the field of statistical testing....**

# Statistical testing - confidence intervals

## So why do we do statistical testing ?

In general you want to make a statement about your data:

these variables are correlated
the stdev on the mean of this samples set is 10%

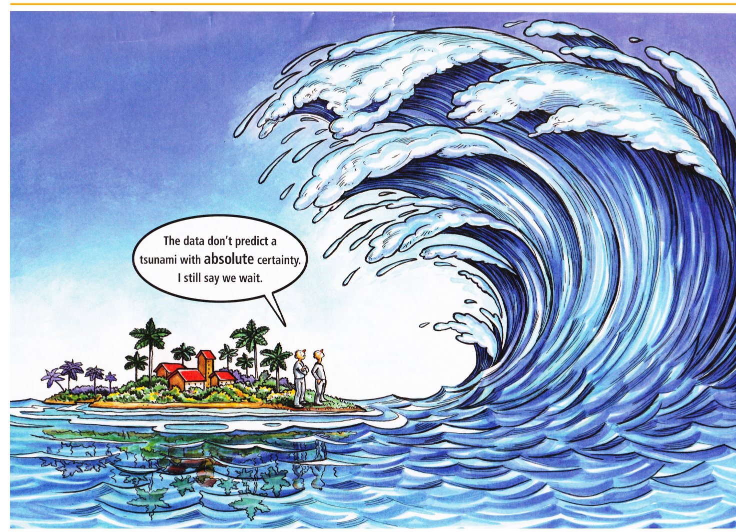
However, in statistics we cannot make such statements as we can never be 100% sure: provide a confidence level: alpha

alpha is up to the researcher to select! There are no “accepted” values and the choice depends strongly on the specific circumstances.

e.g. when mining sector is up: alpha ~ 0.80
down: alpha ~ 0.05

why? at low alpha rarely wrong, but you don’t find much.
at high alpha, will find everything, but are commonly wrong

# Confidence levels




Union of concerned scientists

## Statistical testing - hypothesis testing

### So in the case of our correlation analysis:

Setting the confidence interval at alpha = 0.05 (or 5%);
if we conclude that there is a correlation: will be wrong in 5% of cases

but how do we conclude this?

Hypothesis testing: test at $\alpha$-level
- reject
- accept

In statistics, we cannot prove anything, can only disprove things!

have to choose your hypotheses carefully

## Statistical testing - hypothesis testing

### So in the case of our correlation analysis:

cannot test the presence of correlation but we can test for the **absence** of correlation between the variables:

$r = 0$
- reject, $r \neq 0$, so there **is** a correlation between the vars
- accept, at this confidence interval there is no significant correlation between the variables

hypotheses: $H_0$: hypothesis to be tested $r = 0$
$H_a$: alternative hypothesis $r \neq 0$

In most cases you will be testing the negative conclusion; there is no correlation, there is no difference between two groups, etc.

## Statistical testing - hypothesis testing

### When testing hypotheses there are 4 possible outcomes;

| | $r = 0$ | $r \neq 0$ |
|---|---|---|
| reject $H_0$ | type I error | OK |
| accept $H_0$ | OK | type II error |

type I error: we conclude there is a correlation where there is in fact none: this is the confidence interval we select: alpha

type II error: no reason to reject $H_0$, so we conclude $r = 0$, whereas in reality there is a correlation between the variables: beta

## Statistical testing - hypothesis testing

### we can only **disprove** statements in stats, so only a rejection of $H_0$ results in a strong conclusion

we're willing to accept a number of incorrect rejections and control that with the confidence interval we choose ( beforehand of course! )

but if we cannot reject our $H_0$, the conclusion is weak: there is clearly a possibility that the statement is wrong, but we have no control over that: type II error

mining company: $H_0$: prospect = barren
$H_a$: prospect $\neq$ barren $$$$

| | barren | non-barren |
|---|---|---|
| reject $H_0$ | alpha | $$$$ |
| accept $H_0$ | OK | beta |

## Statistical testing - degrees of freedom

### statistical tests depend on the number of samples

However,
when testing we're always working with a sample and not the full population

this means;
the parameter that we are testing has been derived from our dataset
it has been estimated from the same data that we use to test it

cannot use all the data, because then we would be using data double

Corrected by using the **degrees of freedom** instead:

degrees of freedom (d.f.) are the no of observations or data remaining after estimating the parameter(s) to be tested

## Statistical testing - degrees of freedom

### some examples;

1) the standard deviation;

5 data points: $n = 5$
determine the mean of this dataset: $\sum(x_i)/n$
now determine the variance: $\sum\{(x_i - \text{mean})^2\}$

this uses the mean that we estimated from the data, therefore only 4 independent values: $x_5 = 5^*\text{mean} - x_1 - x_2 - x_3 - x_4$

so we have 4 degrees of freedom:

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

## Statistical testing - degrees of freedom

### some examples;

2) testing of the correlation coefficient

calculated from both the mean in x and the mean in y, so to derive the correlation coefficient, two degrees of freedom have already been consumed:

test against n - 2 degrees of freedom

$$r = \frac{cov_{xy}}{s_x s_y} \qquad t = r\sqrt{\frac{n - 2}{1 - r^2}}$$

## Statistical testing - significance of r

### an example of significance testing of the correlation coefficient:

$t = r\sqrt{\frac{n - 2}{1 - r^2}}$ with d.f. = n - 2 and $t_{\alpha;d.f.}$

Our hypotheses: $H_0$: $r = 0$, if true, no significant correlation
$H_a$ $r \neq 0$, cannot reject the absence of correlation

Let's say: $n = 25$, so d.f. = 23
$\alpha = 0.05$
$r = -0.34$

$t_{calc} = -1.73$

$t_{0.05;23} =$

When calculated t > critical t
significant correlation

# Statistical testing - significance of r

an example of significance testing of the correlation coefficient:

n = 25, so d.f. = 23;  α = 0.05

| df | alpha = 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | 0.0005 |
|---|---|---|---|---|---|---|---|
| **1** | 3.078 | 6.314 | 12.71 | 31.82 | 63.66 | 318.3 | 636.6 |
| **5** | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 6.869 |
| **10** | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 4.587 |
| **15** | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 4.073 |
| **16** | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 4.015 |
| **17** | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 3.965 |
| **18** | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 3.922 |
| **19** | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 3.883 |
| **20** | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 3.850 |
| **21** | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 3.819 |
| **22** | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 3.792 |
| **23** | 1.319 | **1.714** | 2.069 | 2.500 | 2.807 | 3.485 | 3.768 |
| **24** | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 3.745 |
| **25** | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 3.725 |

# Statistical testing - significance of r

an example of significance testing of the correlation coefficient:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ with d.f. = n - 2 and $t_{\alpha;d.f.}$

Our hypotheses: $H_0$: r = 0, if true, no significant correlation

$H_a$ r ≠ 0, cannot reject the absence of correlation

Let's say: n = 25, so d.f. = 23

α = 0.05

r = -0.34

$t_{calc}$ = -1.73

$t_{0.05;23}$ = 1.714 = -1.714

$t_{calc}$ exceeds $t_{0.05;23}$ -> **reject $H_0$**

in this example we can reject the $H_0$: so we can make the strong statement that at the 5% confidence level, there is a significant correlation between the vars

# Statistical testing - the steps

1. Define a hypothesis to test

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

# Statistical testing - the steps

1. Define a hypothesis to test

   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 desired level of confidence

   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

so what do you choose ? depends very much on the situation

identifying cheating schoolteachers: you have to be very certain !

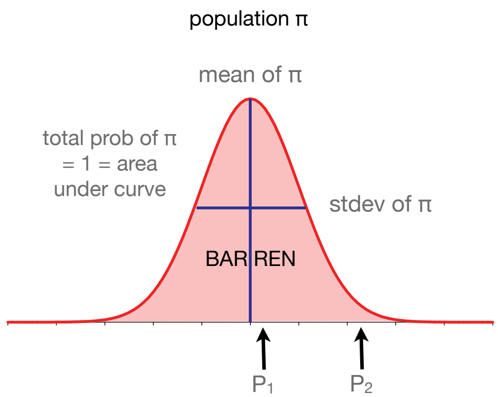
# Statistical testing - confidence levels

For example: a mining company measures a property P (for example As content).

when P belongs to population π
the prospect is barren

when P exceeds π: $$$$

so, what does π look like ?




At $P_1$: probability is high that this measured value belongs to the population π: barren

At $P_2$: probability is much lower that this measured value belongs to the population π: $$$$ more likely

# Statistical testing - confidence levels

For example: a mining company measures a property P (for example As content).

when P belongs to population π
the prospect is barren

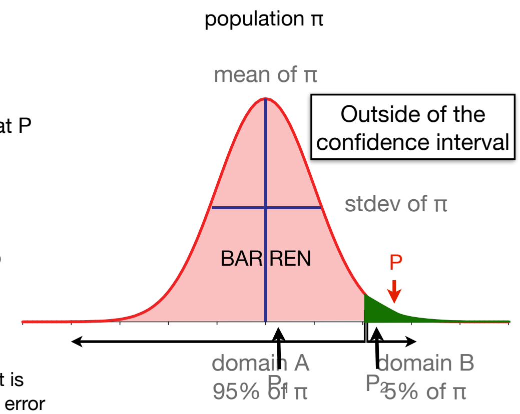
when P exceeds π: $$$$

the confidence level specifies the domain(s) of π where we reject that P belongs to π, i.e. the cutoff level

Let's set alpha = 5%

If P has a value in the green domain: we assume that it does not belong to the red, barren distribution, but comes from a separate distribution that describes the ore deposit

However, there is a 5% chance that it is still part of the red distribution: type I error




# Statistical testing - confidence levels

For example: a mining company measures a property P (for example As content).

when P belongs to population π
the prospect is barren

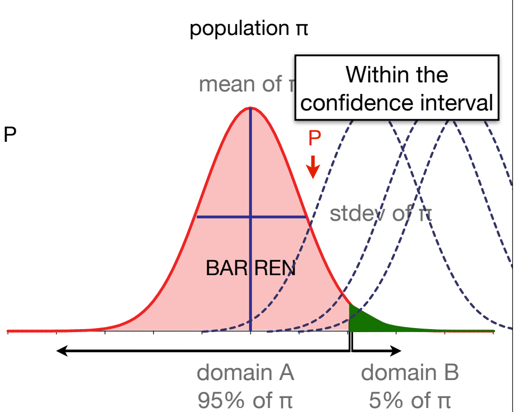
when P exceeds π: $$$$

the confidence level specifies the domain(s) of π where we reject that P belongs to π, i.e. the cutoff level

Let's set alpha = 5%

If P has a value in the red domain: we assume that it belongs to the red, barren distribution and will not drill it.

However, there is a chance that it is part of the ore distribution, because we don't know what it's distribution looks like: type II error




# Statistical testing - confidence levels

For example: a mining company measures a property P (for example As content).

when P belongs to population π
the prospect is barren

when P exceeds π: $$$$

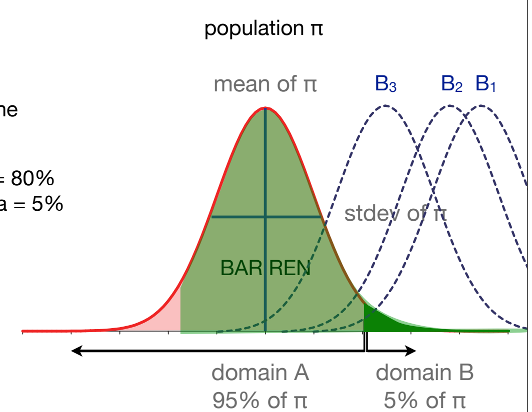
The cut-off level is controlled by the confidence level alpha and varies:

when mining is doing well: alpha = 80%
when mining is under stress: alpha = 5%

why?

at low alpha rarely wrong, but you don't find much.

at high alpha, will find every-thing, but are commonly wrong



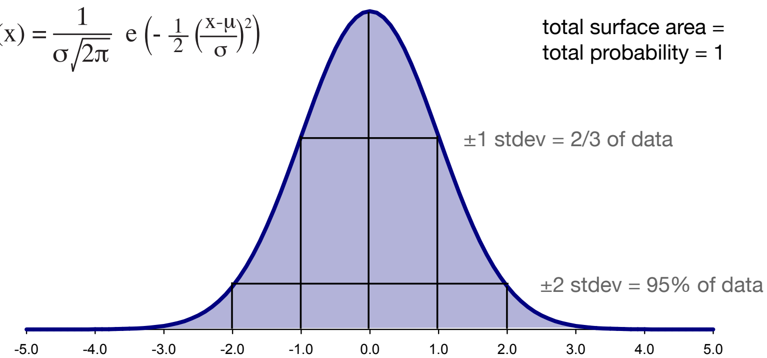
## Statistical testing - the steps

3. Compare the test property against a certain probability distribution

the expected distribution defines the probability of finding a certain observation: can find these values in tables, for example the normal and student-t distributions

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

total surface area = total probability = 1

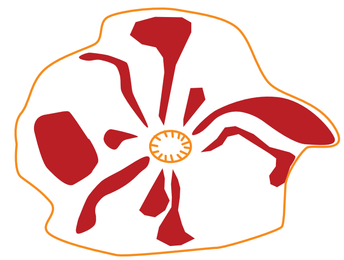



## Statistical testing - testing against the normal dist.

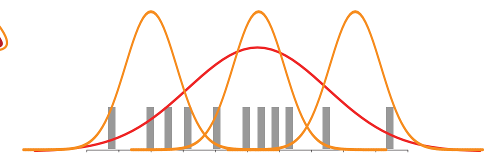
every value or data point is derived from a population

So, for a set of measurements:

- all same population
- all different population
- grouped in populations



each population has a mean and stdev



composition

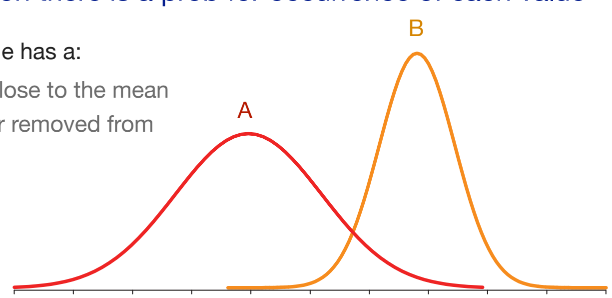
## Statistical testing - population probability

within a population there is a prob for occurrence of each value

every random sample has a:

high prob of being close to the mean

low prob of being far removed from the mean




this probability is known if:

- normally distributed
- mean is known
- variance is known

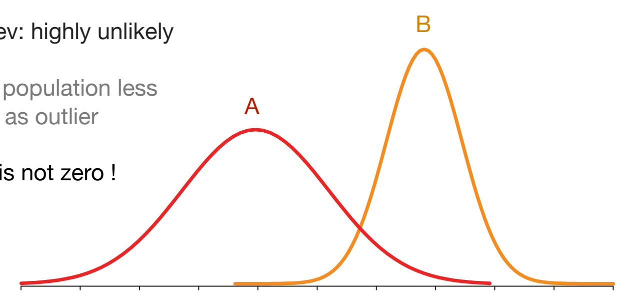
## Statistical testing - population probability

outliers are values that have a low probability of occurrence

values beyond 3 stdev: highly unlikely

prob of belonging to population less than 0.5%: regarded as outlier

However, possibility is not zero !

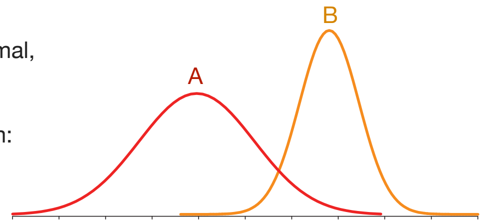



Identical for populations A and B, but a given deviation from the mean will be less likely to be an outlier in case A where the spread is larger.

## Statistical testing - population probability

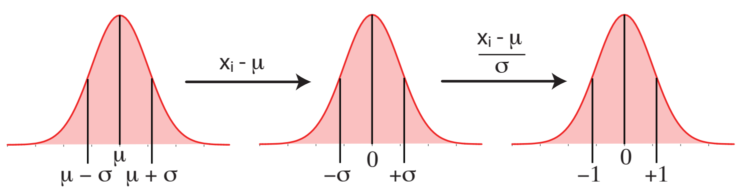
### to use Gaussian probabilities, have to standardize populations

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




convert them to standardized form:

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

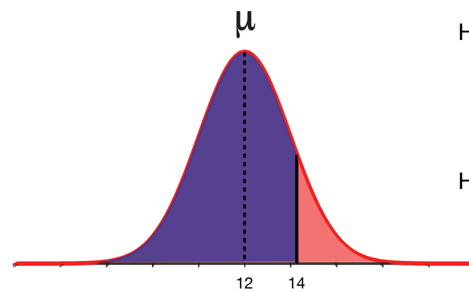



## Statistical testing - population probability

### Z-scores: standardized normal distribution

can use the probabilities of the Gaussian distribution to determine the probability for a given value to occur:

see table 2.2 on page 409




How likely to find a value < 14 ?

$Z = (14-12)/8 = 0.25$

probability of $Z = 0.25$: 59%

How likely to find a value > 14 ?

probability of 100-59 = 41%

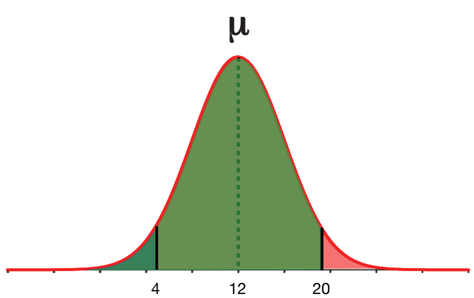
Given a population with a mean of 12 and a standard deviation of 8

## Statistical testing - population probability

### Z-scores: standardized normal distribution

can use the probabilities of the Gaussian distribution to determine the probability for a given value to occur:

see table 2.2 on page 409




How likely to find a value that is between 4 and 20 ?

$Z_4 = (4-12)/8 = -1$
$Z_{20} = (20-12)/8 = +1$

probability of $Z = -1$: 15.9%
probability of $Z = +1$: 84.1%

so prob = 84.1-15.9 = 68.2%

Given a population with a mean of 12 and a standard deviation of 8

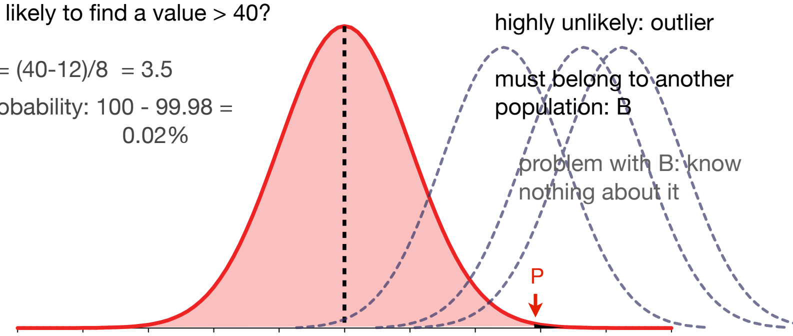
## Statistical testing - population probability

### can be used as a criterion to classify a data point as an outlier

How likely to find a value > 40?

$Z = (40-12)/8 = 3.5$

probability: 100 - 99.98 = 0.02%




highly unlikely: outlier

must belong to another population: B

problem with B: know nothing about it

## Statistical testing - population probability

### So to summarize these observations:

if we can exclude something from population A:

strong statement, exceeds our specified threshold of $\alpha$
will be wrong sometimes, but at least we know and can control it

if we cannot exclude something from population A:

there is still a possibility that it belongs to another population (e.g. B),
but because we know nothing of B, cannot specify the prob of this
weak statement:

type II errors are worse

you know your chances of failure, but not those of success...

## Statistical testing - population probability

### what if we know the properties of the other pop as well ?

for the ore sample example:

population A: $\mu = 60$, population B: $\mu = 130$
population P: $\mu = 110$, $SE_{A,B} = 20$ (SE because comparing means)

$Z_i = (\mu_P - \mu) / SE$ at $\alpha = 0.05$: $-1.96 < Z < 1.96$

1) hypothesis: P part of A $H_0$; $\mu_P = \mu_A$

$Z = 2.5$, so it exceeds Z range: rejected

2) hypothesis: P part of B $H_0$; $\mu_P = \mu_B$

$Z = -1.0$, so it is within Z range: accepted

## Statistical testing - population probability

### what if we know the properties of the other pop as well ?

another example:

a well-established fossil population has length $\mu = 14.2 \pm 4.7$ mm
now a researcher finds a mean of 30 mm from $n = 10$

can these belong to the same population?

hypotheses: $H_0$; $\mu_{new} = \mu$
$H_A$; $\mu_{new} \neq \mu$

$Z = (\mu_{new} - \mu) / (\sigma/\sqrt{n})$ at $\alpha = 0.05$: $-1.96 < Z < 1.96$

$Z = (30-14.2)/(4.7/\sqrt{10}) = 10.63$

## Statistical testing - the t-distribution

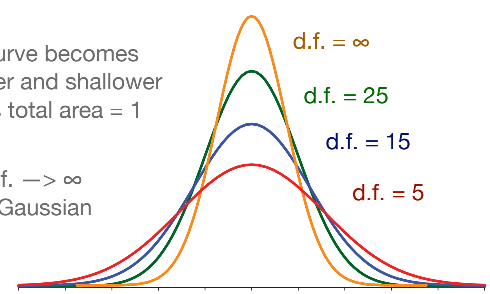
### rarely know the population mean and stdev, rather sample stats

In the previous examples we presumed to know the mean and stdev of the population, but in reality we rarely do: estimate these from a sample

so, the test distribution should have a larger uncertainty and this has to depend on the number of samples (degrees of freedom): the t-distribution

curve becomes wider and shallower as total area = 1

as d.f. —> ∞
t —> Gaussian




values for the t statistic can be found in tables for a given confidence level and no. of d.f.

e.g. 2.4 on page 411

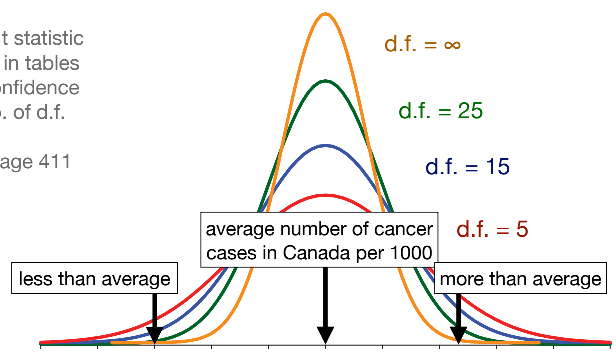
## The curse of low sample numbers

The t-distribution elegantly shows the effect of small sample numbers on the probability of finding extreme values:

the probability of finding a certain value depends on the number of samples: less samples means (ironically) a higher probability

values for the t statistic can be found in tables for a given confidence level and no. of d.f.

e.g. 2.4 on page 411




## Statistical testing - t-distribution testing

testing against the t-distribution is identical to that of Z-scores

$t = (\bar{x}-\mu) / (s/\sqrt{n})$ using the means and SE, so independent of the type of distribution !

Normally we do not test individual values against the t-distribution, but rather the mean derived from a sample against the mean of the population we think these values come from

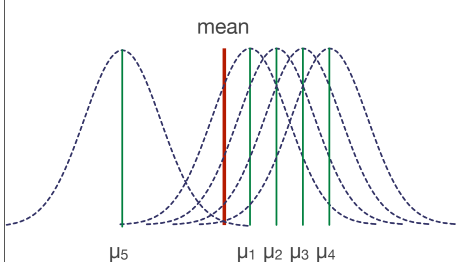
has the added advantage that we can ignore distribution (multi-modality......)

Can use it for two very useful properties:

- the confidence interval for a value or property by extracting μ
- required sample size for specified confidence by extracting n

## Statistical testing - t-distribution testing

Commonly, a company needs to guarantee certain specifications for a product. For example, that the concentration of the ore element is at a certain level, or the concentration of a contaminant below a certain level. Missing such targets can be very costly. So how do you decide what is a good, as in achievable, level ?




suppose $\bar{x}$ from $\mu_1$: prob high
from $\mu_2$: prob lower
from $\mu_3$: prob lower
from $\mu_4$: prob low

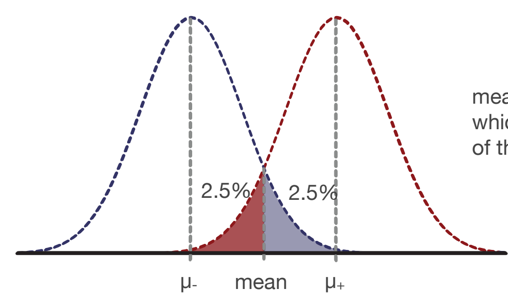
at some value of μ, will exceed the confidence level: too unlikely to come from a population with this mean: this is the upper μ

similarly, will reach a lower μ when working down from the mean

The confidence interval on the mean represent the range from this lower to the upper population mean for a given confidence level.

## Statistical testing - t-distribution testing

the confidence interval for a mean at a given confidence level:




mean can belong to populations for which the probability of occurrence of this mean is more than 0.5α

formulae: $\mu_+ = \bar{x} + t_{\alpha;df} \cdot \frac{s}{\sqrt{n}}$ $\mu_- = \bar{x} - t_{\alpha;df} \cdot \frac{s}{\sqrt{n}}$

# Statistical testing - t-distribution testing example 1

## the confidence interval for the concentration of phosphorus in iron ore

Say we are required to supply iron ore with a bulk phosphorus content of less than 250 ppm, or the company has to pay a fine. The mean P content that you have determined is 215 ± 30.8 ppm based on 8 samples.

the specifics: our mean: 215 ± 30.8 ppm from n = 8
the limit: 250 ppm
desired confidence: 95%

d.f. = n - 1
$\alpha = 0.05$
$t_{\alpha;df} = 2.365$

What is the 95% confidence interval on the bulk concentration?

$$\mu_+ = \bar{x} + t_{\alpha;df} \cdot \frac{s}{\sqrt{n}} \qquad \mu_+ = 215 + 2.365 \cdot \frac{30.8}{\sqrt{8}}$$

$$\mu_- = \bar{x} - t_{\alpha;df} \cdot \frac{s}{\sqrt{n}} \qquad \mu_- = 215 - 2.365 \cdot \frac{30.8}{\sqrt{8}}$$

189 < mean < 241

ok

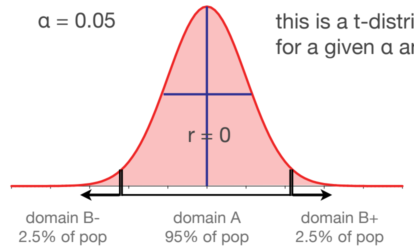
# Statistical testing - significance of r

## So, let’s now return to the correlation coefficient:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ with d.f. = n - 2 and $t_{\alpha;d.f.}$

Our hypotheses: $H_0$: r = 0, if true, no significant correlation: domain A
$H_a$ r ≠ 0, cannot reject the absence of correlation: B

$\alpha = 0.05$




this is a t-distribution for a given α and d.f.

when r plots in domain B: prob of it belonging to the population r = 0 is lower than our threshold α:

reject r = 0 and therefore conclude that the variables are correlated

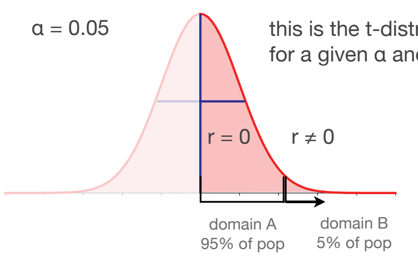
# Statistical testing - significance of r

## Testing the significance of the correlation coefficient:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ with d.f. = n - 2 and $t_{\alpha;d.f.}$

Our hypotheses: $H_0$: r = 0, if true, no significant correlation: domain A
$H_a$ r ≠ 0, cannot reject the absence of correlation: B

$\alpha = 0.05$




this is the t-distribution for a given α and d.f.

when r plots in domain B: prob of it belonging to the population r = 0 is lower than our threshold α:

reject r = 0 and therefore conclude that the variables are correlated

# Statistical testing - significance of r

## What values of r are meaningful for a given confidence level

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$ with d.f. = n - 2 and $t_{\alpha;d.f.}$

Our hypotheses: $H_0$: r = 0, if true, no significant correlation
$H_a$ r ≠ 0, cannot reject the absence of correlation

Let’s say: n = 25, so d.f. = 23
$\alpha$ = 0.05 or 0.025
r = -0.34

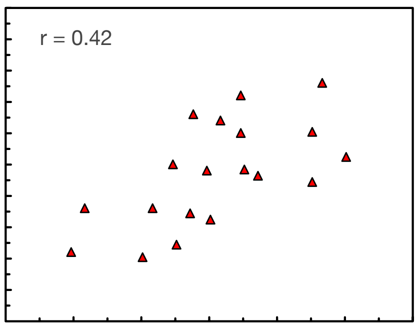
$t_{calc} = -1.73$
$t_{0.05;23} = 1.71 = -1.71$
$t_{0.025;23} = 2.07 = -2.07$

$t_{calc}$ exceeds $t_{0.05;23}$ -> reject $H_0$
$t_{calc}$ doesn’t exceeds $t_{0.025;23}$ -> cannot reject $H_0$

## Statistical testing - significance of r

### The effect of degrees of freedom (n) on the significance:

e.g. a data set like this:




is this correlation significant at α = 0.05 and the following n ?

at n = 5; t = 0.80

$t_{0.05;3}$ = 2.353 ✗

at n = 10; t = 1.31

$t_{0.05;8}$ = 1.860 ✗

at n = 25; t = 2.22

$t_{0.05;23}$ = 1.717 V

## Statistical testing - probability of a value

### Z- and t-test can be used to determine the prob of a value

Commonly use the mean to avoid problems associated with deviations from normality, plus uncertainty on mean is smaller: stronger statements

$Z_i = (\mu_C - \mu) / SE$ $t_i = (\bar{x} - \mu) / SE$

e.g. given 10 sandstone samples with the following porosities:

13, 17, 15, 23, 27, 29, 18, 27, 20, 24

$\bar{x}$ = 21.3 s = 5.52

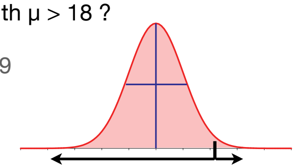
n = 10 $s_e$ = 1.75

is it possible that this set is from a population with μ > 18 ?

$H_0$; μ ≤ 18
$H_A$; μ > 18

$t_{calc}$ = (21.3-18)/1.75 = 1.89

$t_{0.05;9}$ = 1.83 V



## Statistical testing - comparing means

### What if we repeat this sampling and want to compare them?

two sets of sandstone samples with the following porosities:

$\bar{x}$ = 21.3 $s^2$ = 30.46 $\bar{x}$ = 18.9 $s^2$ = 23.21

n = 10 n = 10

are they from the same population? $H_0$; $\mu_1 = \mu_2$

$H_A$; $\mu_1 \neq \mu_2$

$t_i = \{(\bar{x}_1 - \mu_1) - (\bar{x}_2 - \mu_2)\} / SE$ for $\mu_1 = \mu_2$ : $t_i = (\bar{x}_1 - \bar{x}_2) / SE$

but what error do we use ? That of set 1 or that of set 2 ?

will have to use a combination of both, in the proportion to the number of samples in each set: more samples: stronger control on error

## Statistical testing - pooled standard deviation

### combined standard deviation is called the pooled stdev - $s_p$

$$s_p^2 = \frac{(n_1 - 1)\cdot s_1^2 + (n_2 - 1)\cdot s_2^2}{(n_1 - 1) + (n_2 - 1)} \qquad s_e^2 = s_p^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

add the variance in proportion to the df in each set: if $n_1 > n_2$, $s_1$ will dominate the pooled stdev and vice versa

So, in this example: $t_i = (\bar{x}_1 - \bar{x}_2) / SE$

$H_0$; $\mu_1 = \mu_2$

$H_A$; $\mu_1 \neq \mu_2$

$\bar{x}$ = 21.3 $s^2$ = 30.46

n = 10

$\bar{x}$ = 18.9 $s^2$ = 23.21

n = 10

$s_p$ = 5.18

$s_e$ = 2.32

$t_{calc}$ = 1.03

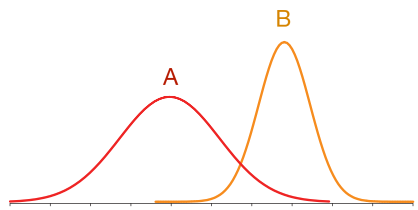
df = $n_1 + n_2$ - 2 (why?)

$t_{0.05;18}$ = 1.734 ✗

# Requirements for t-test

## When conducting a t-test, you assume the following:

1. samples have been taken randomly
   so if sampled by two geologists: no preference in what they sampled
2. sample sets normally distributed
   if not: use the means and $s_e$
3. sample sets have equal variance
   so $\sigma_1 = \sigma_2$




Of these, the third is the most crucial. If we have a marked deviation from equality of variance: have to switch to another test (rank-based)

so how do we determine if the data fulfill this requirement ?

# The F - test

## To determine the (in)equality of the variance in two datasets:

Test the ratio of the variance against the F - distribution

if it exceeds a critical F at your chosen α: not equal
if it doesn't: no reason to assume that the variances are different

So what are the hypotheses for this test ?

$H_0$; $\sigma_1 = \sigma_2$

$H_A$; $\sigma_1 \neq \sigma_2$

Testing always works in exactly the same way: you have a probability distribution, be it the Z-, t- or F-distribution. If your calculated value for Z, t or F exceeds the probability level α: reject $H_0$

$F = (s_1)^2 / (s_2)^2$

depends on the df of both set 1 and set 2

what is the df in this case ?

see table 2.5, p 412

# The F - test

## So for our sandstone porosity example:

Did we meet all the requirements of the t-test ?

$\bar{x} = 21.3$ $s^2 = 30.46$
n = 10

$\bar{x} = 18.9$ $s^2 = 23.21$
n = 10

$F = (s_1)^2 / (s_2)^2$

by convention: $s_1 > s_2$

$F = 30.46 / 23.21$
$= 1.31$

from table 2.5:

$F_{0.05;9;9} = 3.18$

hypotheses for the F - test are:

$H_0$; $\sigma_1 = \sigma_2$

$H_A$; $\sigma_1 \neq \sigma_2$

so ?

no reason to reject $H_0$ as the calculated F value does not exceed the $F_{0.05;9;9}$

$\sigma_1 = \sigma_2$

# Mann-Whitney test for non-normal data

A t-test uses **mean** and **standard deviation** and can thus only be applied to data that fit the normal distribution, or that can be mathematically transformed to a normal distribution.

To test equality of datasets that are not normally distributed, we can use the robust equivalent: the Mann-Whitney test.

Instead of using the mean, as in the t-test, we compare medians, which are robust. And we use the rank of a value, rather than its actual value.

We subsequently calculate the Mann-Whitney statistic for our datasets and compare this to tabulated critical values to reach our conclusion

## Mann-Whitney test for non-normal data

are two sets of data from the same population?

$H_0$; $med_1 = med_2$

$H_A$; $med_1 \neq med_2$

| dataset A conc Cu | dataset B conc Cu | value rank (dataset A) | value rank (dataset B) |
|---|---|---|---|
| 20 | 19 | 4 | 3 |
| 14 | 34 | 2 | 8 |
| 25 | 28 | 5 | 6 |
| 32 | 41 | 7 | 10 |
| 11 | 36 | 1 | 9 |

$n_A = 5$

$n_B = 5$

$T = \sum R(A_i) - n_A \cdot (n_A+1) / 2$

$T = 19 - 5 \cdot (5+1) / 2 = 4$

$T_{critical}$ ($df$ = 5,5) = 2 to 4
at confidence level = 5%

cannot reject the null hypothesis: from same population

## An extension of the t-test

The approach breaks down when there are a large number of data sets to compare

Need to do a t-test and a F-test for each combination:

| | | | | |
|---|---|---|---|---|
| $\bar{x}_1 = \bar{x}_2$ | t - test | | $\sigma_1 = \sigma_2$ | F - test |
| $\bar{x}_2 = \bar{x}_3$ | t - test | & | $\sigma_1 = \sigma_3$ | F - test |
| $\bar{x}_1 = \bar{x}_3$ | t - test | | $\sigma_2 = \sigma_3$ | F - test |

For three data sets this is still doable, but if you have five, there are already 10 combination of sample means and stdevs that you need to test

and at $\alpha$ = 0.10, on average one of these would give you a significant difference purely by chance !

Better to switch to another type of testing: analysis of variance - ANOVA

Geotop Short Course in Data Analysis and Geostatistics

## Analysis of Variance - ANOVA

## Analysis of variance - ANOVA

ANOVA may seem daunting, but conceptually it is not difficult

e.g. in northern Spain, metamorphism has overprinted all evidence of depositional environment in a series of limestones. However, you wonder if the $\delta^{13}C$ signature may still preserve this information

need to determine first of all if there are differences between these marbles and only then see if you can link them to environment

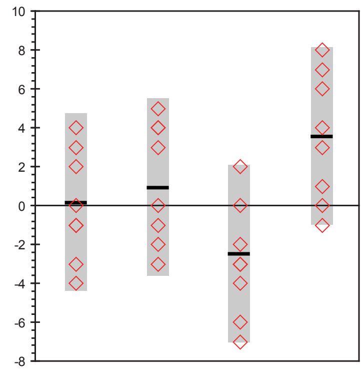
for differences to be significant, the variance within each unit has to be smaller than the variance between the units

otherwise your possible signal is lost in the noise

# Analysis of variance - ANOVA

## The analytical data for the four marble units:

| | unit 1 | unit 2 | unit 3 | unit 4 |
|---|---|---|---|---|
| | -3 | +3 | -3 | +4 |
| | +3 | -1 | -6 | +7 |
| | -1 | -2 | -2 | -1 |
| | -1 | +4 | +2 | +1 |
| | +4 | 0 | -3 | +6 |
| | -4 | -3 | -4 | +3 |
| | +2 | +5 | 0 | 0 |
| | 0 | +4 | -7 | +8 |
| mean | 0 | 1.25 | -2.88 | 3.5 |
| $s^2$ | 8 | 9.6 | 8.7 | 11.1 |
| n | 8 | 8 | 8 | 8 |
| SS | 56 | 67.5 | 60.9 | 78 |




difference between needs to exceed difference within

# Analysis of variance - ANOVA

## So, let's analyze the variance in this data-set - 3 types;

1. total variance in the data

lump all the samples together into one big sample and calculate the variance in the full data set:

n = 8 + 8 + 8 + 8 = 32 d.f. = n - 1 = 31 mean = 0.47

$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{df} = 13.9$ $SS_{TOT} = \Sigma (x_i - \bar{x})^2 = 432$

# Analysis of variance - ANOVA

## So, let's analyze the variance in this data-set - 3 types;

2. within variance of the data set

the spread in each unit combined in a pooled variance in proportion to the df of each sample set (in this case equal for each unit):

$$s_p^2 = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2 + (n_3-1)\cdot s_3^2 + (n_4-1)\cdot s_4^2}{(n_1-1)+(n_2-1)+(n_3-1)+(n_4-1)}$$

df = n - 1 = 7
SS = $s^2 \cdot$ df

$$s_p^2 = \frac{SS_1 + SS_2 + SS_3 + SS_4}{df_1 + df_2 + df_3 + df_4} = \frac{\sum SS_i}{\sum df_i}$$

$$SS = \sum (x_i - \bar{x})^2$$

$s^2$ = (56.0 + 67.5 + 60.9 + 78.0) / (7 + 7 + 7 + 7) = 262.4 / 28 = 9.4

# Analysis of variance - ANOVA

## So, let's analyze the variance in this data-set - 3 types;

3. between variance of the data set

the variance in between the units - we can calculate that from the variance on their means:

$s_e^2 = s^2 / n$ -> $s^2 = n \cdot s_e^2$

df = m - 1 = 3
SS = $s^2 \cdot$ df

$$s_e^2 = \frac{SS}{df} = \frac{\sum (\bar{x}_i - \bar{x}_{tot})^2}{m-1}$$

$s_e^2 = 21.2 / 3 = 7.1$

$s^2 = n \cdot s_e^2 = 8 \cdot 7.1 = 56.5$

in SS notation: $\frac{21.2 \times 8}{3} = \frac{169.6}{3}$

## Analysis of variance - ANOVA

### We can also summarize this information in a table:

| | sum of squares | d.f. | variance |
|---|---|---|---|
| between | 169.6 | 3 | 56.5 |
| within | 262.4 | 28 | 9.4 |
| total | 432 | 31 | 13.9 |

note: conservation of sum of squares and degrees of freedom
SS very useful property, conservation of df makes sense (I hope)

from this it is already clear that the variance between the units is much larger than that within each unit, or the total variance of the data:

suggests that there is indeed a significant difference between these units

## Analysis of variance - ANOVA

### The hypotheses for this example and what to test:

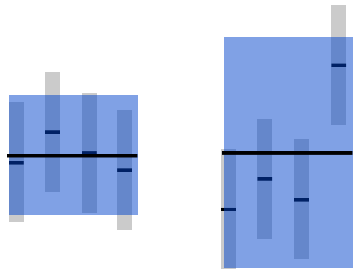
$H_0$; $\mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_A$; one of these is not equal, because derived from other pop

assumptions are equal to those of the t-test: variance is the same

if $H_0$ = true; the variance between units is indistinguishable from that within each unit, so no difference between units

if $H_0 \neq$ true; the variance within each unit will not change, but variance between them and the total variance will increase and exceed within var



## Analysis of variance - ANOVA

### So how do we test our hypotheses ?

if $s_{between} \leq s_{within}$ : all the same

$s_{between} > s_{within}$ : different at level $\alpha$

test this with the F-test: $F = s^2_{between} / s^2_{within}$ at df 3 and 28
$\alpha = 0.05$

critical F ~ 3

calculated F = 6

so, in this case the F exceeds the critical F:

reject the $H_0$ that there are no significant differences between the units:

can segregate them based on $\delta^{13}C$

## ANOVA - Analysis of variance

In previous example: only interested in the differences between the units

one variable: one-way ANOVA

However, we may be interested in more than one variable

### ANOVA can be extended to as many variables as you like

differences between the 4 marble units

differences between the laboratories that analyzed the samples

differences between the geologists who sampled them

## ANOVA - Analysis of variance

### An example: 4 geologists determined the Cu content in 3 units:

Is the Cu content different in the different units?

Is there any difference between the geologists?

| | geologist | | | |
|---|---|---|---|---|
| formation | I | II | III | IV |
| 1 | 30 | 70 | 30 | 30 |
| 2 | 80 | 50 | 40 | 70 |
| 3 | 100 | 60 | 80 | 80 |

2 null-hypotheses:

$H_0$; $\mu_I = \mu_{II} = \mu_{III} = \mu_{IV}$

$H_0$; $\mu_1 = \mu_2 = \mu_3$

$H_A$; one of these is not equal

## ANOVA - Analysis of variance

### Should assess the variance at the same time, because both variables will affect the variance and the data are the same

Hypothesis 1; $s^2_{between\ geol} > s^2_{within}$

Hypothesis 2; $s^2_{between\ units} > s^2_{within}$

$s^2_{within}$ is the variance inherent in the data: not explained by diff in unit or geologist: residual

| | sum of squares | degrees of freedom | variance |
|---|---|---|---|
| between units | $SS_A$ | 3-1 | $s^2_A$ |
| between geol | $SS_B$ | 4-1 | $s^2_B$ |
| within/residual | $SS_R$ | (4-1)·(3-1) | $s^2_R$ |
| total | $SS_{TOT}$ | (4·3)-1 | $s^2_{TOT}$ |

## ANOVA - Analysis of variance

### Input the data into PAST with two factors: unit and geologist

| | sum of squares | degrees of freedom | variance | F-ratio | F-crit |
|---|---|---|---|---|---|
| between geol | 3200 | 2 | 1600 | 4 | 5.14 |
| between units | 600 | 3 | 200 | 0.5 | 4.76 |
| within/residual | 2400 | 6 | 400 | | |
| total | 6200 | 11 | | | |

From this it is clear that the variance between units is smaller than the within variance, but this is not true for the variance between geologists

However, at $\alpha = 5\%$, neither exceeds the critical probability: all are the same

## ANOVA - Analysis of variance

### Input the data into PAST with two factors: unit and geologist

| | sum of squares | degrees of freedom | F-ratio | F-crit | p (same) |
|---|---|---|---|---|---|
| between geol | 3200 | 2 | 4 | 5.14 | **0.08** |
| between units | 600 | 3 | 0.5 | 4.76 | **0.70** |
| within/residual | 2400 | 6 | | | |
| total | 6200 | 11 | | $\alpha =$ | **0.05** |

Can also change the question, at what probability are they the same or what is the confidence of my conclusion that they are the same ?

Most stats software, including PAST, provides this information as well (and sometimes only this information)

## ANOVA - Analysis of variance

### another example: water composition of 5 rivers in 4 seasons

| | river | | | | |
|---|---|---|---|---|---|
| seasons | I | II | III | IV | V |
| winter | 10 | 80 | 4 | 60 | 19 |
| spring | 20 | 60 | 19 | 40 | 34 |
| summer | 2 | 12 | 20 | 80 | 12 |
| autumn | 4 | 28 | 17 | 50 | 20 |

Are there any significant differences between the rivers ?

Are there significant differences between the seasons ?

Can extend this in ANOVA given more grouping variables and data;
Are there any differences between years ?
Are there differences with depth, width, ice-cover?

## Rank testing of differences of the mean

### To conduct an ANOVA test we have to fulfill the same requirements as for the t-test:

most important of these is equality of variance:

$\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$

What if this condition is not met ?

Have to switch to robust testing: i.e. rank testing:

Mann-Whitney test < - > t-test

Kruskal-Wallis test < - > ANOVA

to find out more about these and how to apply them: 4.2.2 and 4.2.3

## Geotop Short Course in Data Analysis and Geostatistics
## Goodness-of-fit

geotop

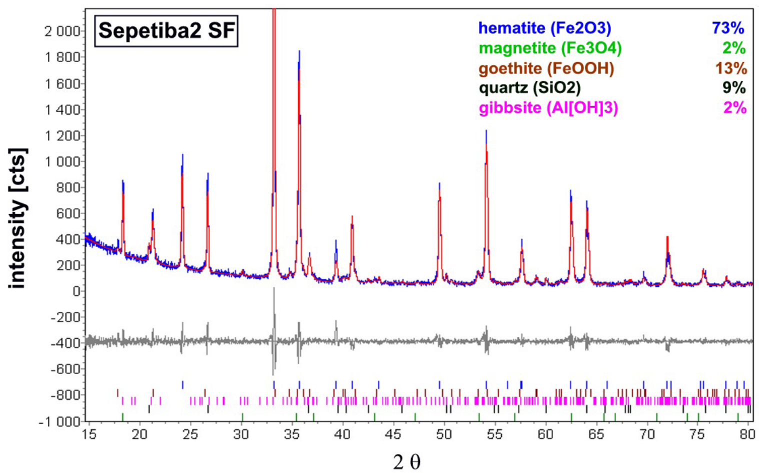
## Testing of "goodness-of-fit"

### in a lot of cases we want to compare curves, not values

Some examples;

- are my data normally distributed ?
  is there a significant difference between my data distribution and that of the normal distribution
- does my model accurately represent the data ?
  is there a significant difference between my predicted data values and the observed ones
- can my minerals/species explain the observed spectrum ?
  is there a significant difference between my predicted spectrum and the observed one

## Fit between measured and predicted spectrum

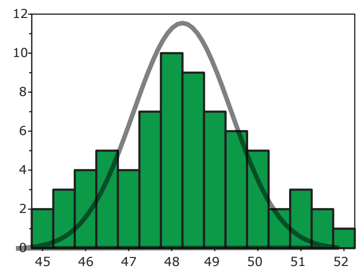



## Testing of “goodness-of-fit”

### comparison of curves: predicted and observed values

the cumulative discrepancy between the predicted and observed values is a measure of the goodness-of-fit

if this exceeds a critical value: can reject the fit that we are testing




this is the Chi-squared ($X^2$) test:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

with $O_i$ = observed value of i
and $E_i$ = predicted value of i

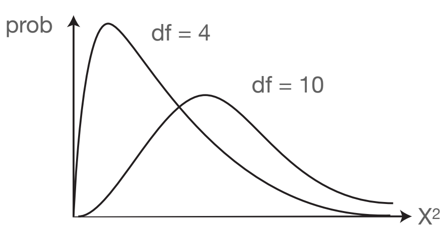
## Testing of “goodness-of-fit”

### The Chi-squared distribution

The Chi-squared test has a very easy formulation and can be applied equally to parametric and non-parametric data (i.e. it is robust)

as in all other tests we then compare our calculated Chi-squared to a tabulated critical value for a given confidence level to reach our conclusion

in this case we test against the Chi-squared distribution

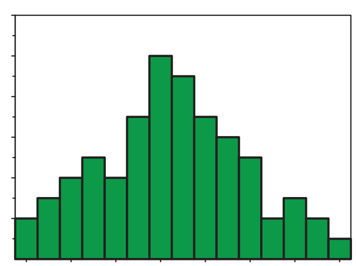



see table 2.6 on page 413

## Testing of “goodness-of-fit”

### An example: testing of normality of a data set

Does the following datas ste show significant deviation from normality ?



$\bar{x}$ = 2.58 s = 0.0195 n = 50

requirements for testing:

- more than 5 samples per class
- more than 3 classes
- convert data to Z-scores

we will convert the histogram into 4 classes and shift the data with x-μ/σ

## Testing of "goodness-of-fit"

### Deriving the observed and expected occurrence of data:

| Z class | observed | | prob. | expected |
|---|---|---|---|---|
| < -1 | 6 | can now determine the probability for each Z class from the normal distribution | 0.16 | 7.93 |
| -1 to 0 | 20 | | 0.34 | 17.07 |
| 0 to +1 | 18 | | 0.34 | 17.07 |
| > +1 | 6 | | 0.16 | 7.93 |
| N | 50 | | 1.00 | 50 |

Can then use these data to calculate the Chi-squared value: 1.494

Now need to know the critical value at say a confidence level of 0.05:

what is the number of df for this test ?

df = no. of classes - parameters required to describe the pop ($\bar{x}$,s) - N = n - 3

$X^2_{0.05:1}$ = 3.84 : calc does not exceed it : no reason to reject normality

## Testing of "goodness-of-fit"

### Calculating the confidence interval on the stdev using the $X^2$

The Chi-squared distribution is derived from the Z-scores:

$$\chi^2_{df} = \sum_i^{df} \frac{(x_i - \mu)^2}{\sigma^2} = \sum_i^{df} z_i^2$$

and because of this relation we can use it to determine the confidence interval on the stdev or variance:

$$\frac{(n-1)s^2}{\chi^2_{1-\frac{1}{2}\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{1}{2}\alpha}}$$

So, for a confidence level of 90%, or α = 0.10, this becomes:

$$\frac{(n-1)s^2}{\chi^2_{0.95}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{0.05}}$$

## Testing of "goodness-of-fit"

### An example of the confidence interval for the stdev:

a standard has been analyzed 20 times: s = 0.8%

What is the confidence interval for the standard deviation of this technique at α = 5% ?

$$\frac{(n-1)s^2}{\chi^2_{1-\frac{1}{2}\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\frac{1}{2}\alpha}}$$

s = 0.8%
n = 20
df = 20-1 = 19

$$\frac{19 \cdot 0.8^2}{\chi^2_{0.975}} < \sigma^2 < \frac{19 \cdot 0.8^2}{\chi^2_{0.025}} \qquad \frac{19 \cdot 0.8^2}{32.9} < \sigma^2 < \frac{19 \cdot 0.8^2}{8.91} \qquad 0.61 < \sigma < 1.17$$

## Day 3 - topics covered

- How to deal with missing data
- Statistical testing; hypotheses, confidence intervals, prob. distribution
- Z and t probability tests
- Comparing groups
- ANOVA
- Goodness-of-fit