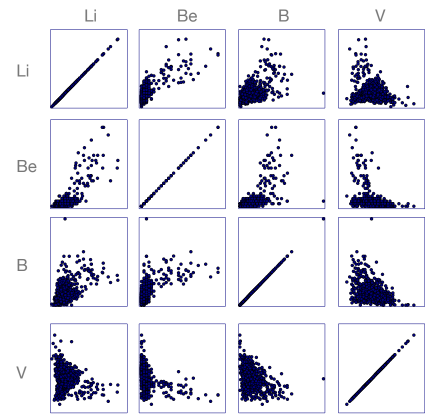
Data analysis and Geostatistics - lecture V

Statistical testing

Correlation coefficients - indicator of covariance

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



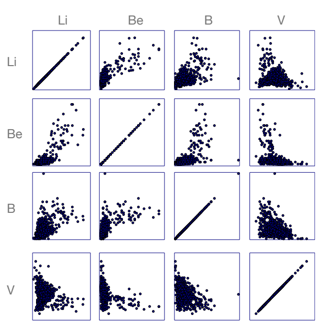
	Li	logBe	B	logV
Li	1	0.7	0.5	-0.3
logBe	0.7	1	0.6	-0.5
B	0.5	0.6	1	-0.4
logV	-0.3	-0.5	-0.4	1

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%

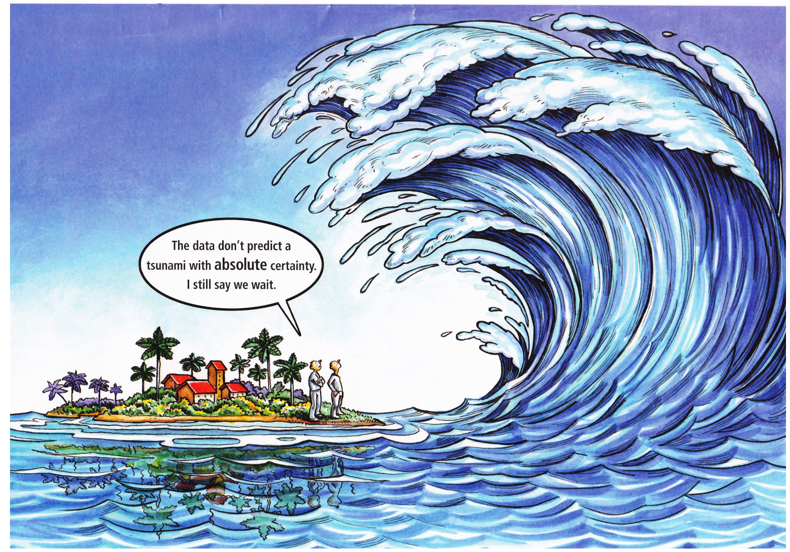
However, in statistics we cannot make such statements as we can never be 100% sure: provide a confidence level: alpha

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 α-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 ≠ 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 ≠ 0, cannot reject the absence of correlation

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

α = 0.05

r = -0.34

$t_{calc} = -1.73$

$t_{0.05;23} =$

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

$t_{calc} = -1.73$

$t_{0.05;23} = 1.71 = -1.71$

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 - significance of r

what if we want to be more certain ?

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 = 0.025

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

we can now only conclude that we have no reason to reject the absence of correlation, which is clearly not as strong a statement

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 !

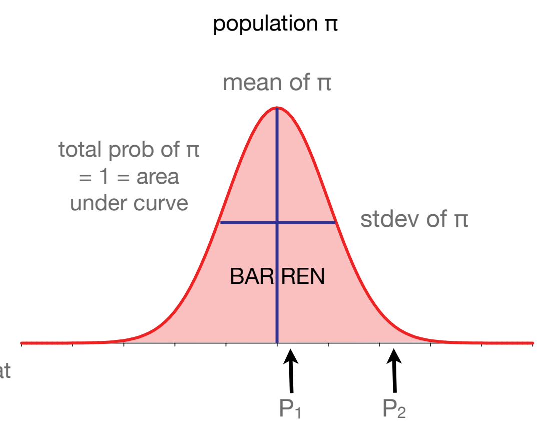
Statistical testing - confidence levels

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).

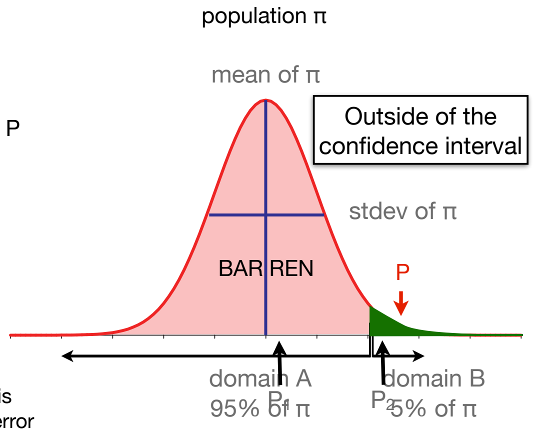
when P belongs to population π
the prospect is barren

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

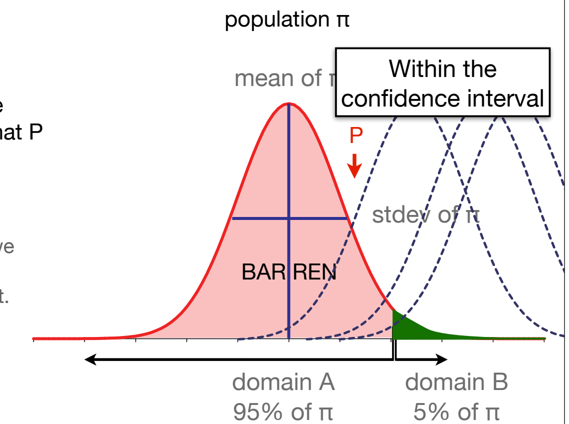
when P exceeds π: $$$$

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: $$$$

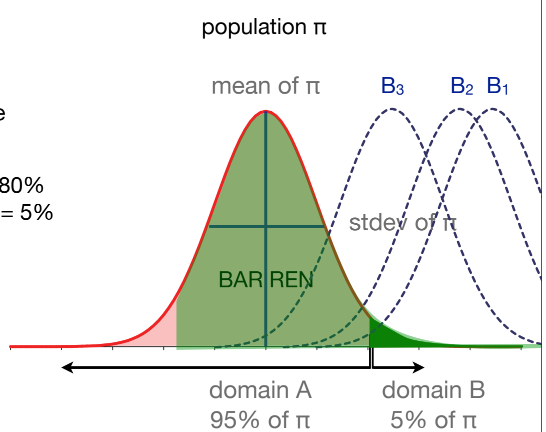
The cut-off level is controlled by the confidence level alpha and varies:

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 everything, but are commonly wrong

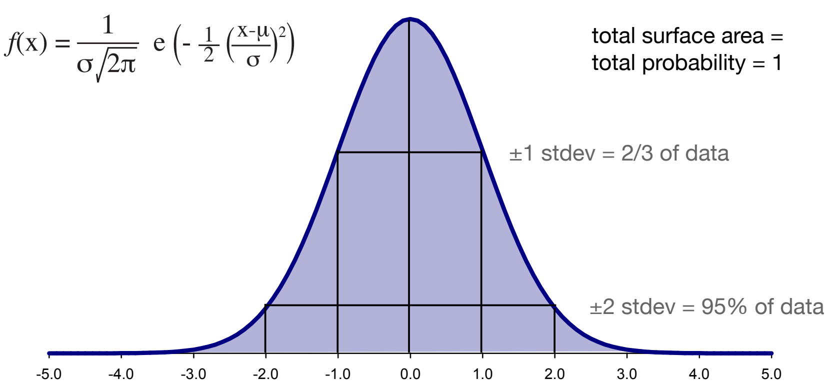


Statistical testing - the steps

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

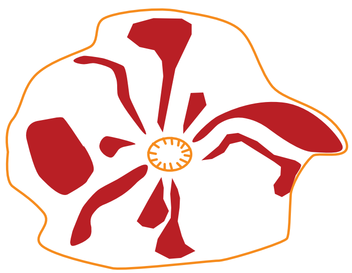
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}$$



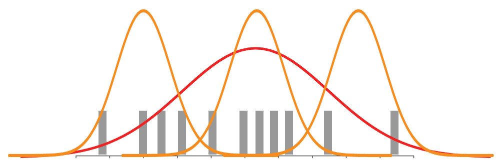
Statistical testing - testing against the normal dist.

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



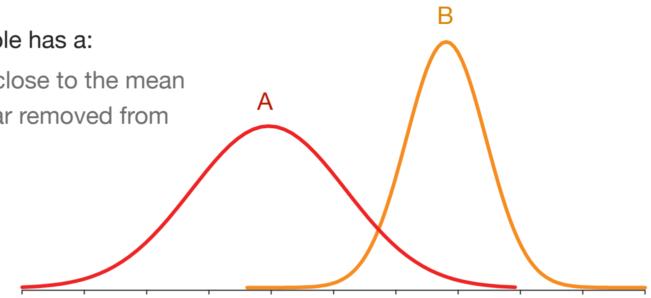
Statistical testing - population probability

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

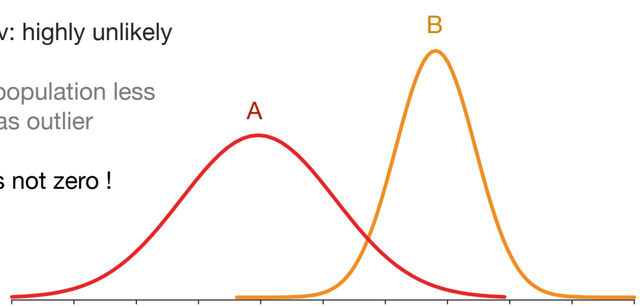
Statistical testing - population probability

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.

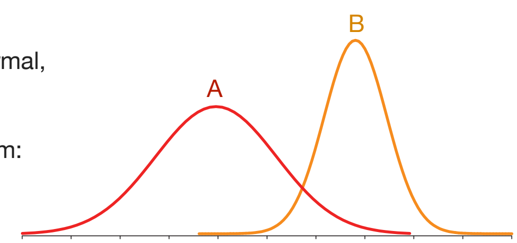
Statistical testing - population probability

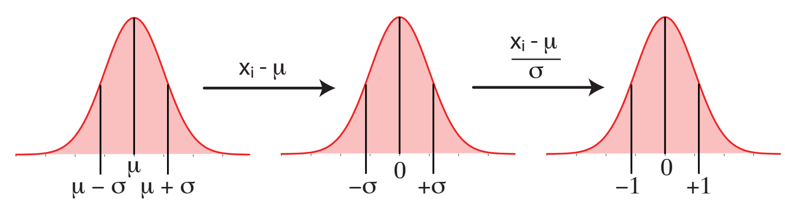
to use Gaussian probabilities, have to standardize populations

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

convert them to standardized form:

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



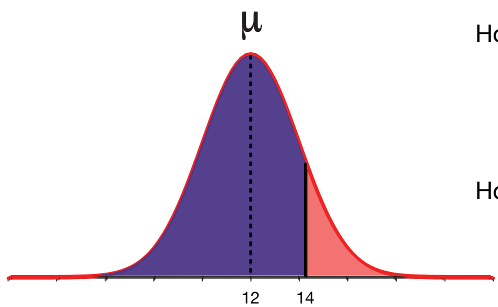


Statistical testing - population probability

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%

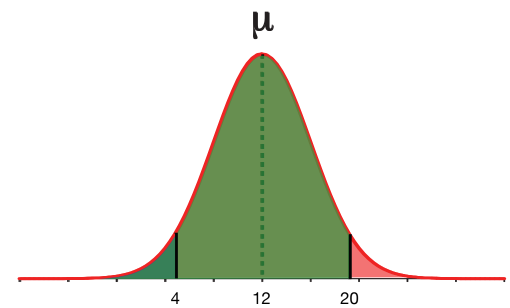
Given a population with a mean of 12 and a standard deviation of 8

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

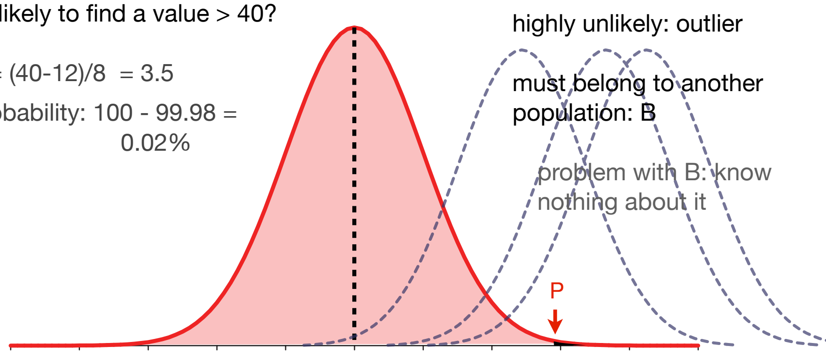
Statistical testing - population probability

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 α
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

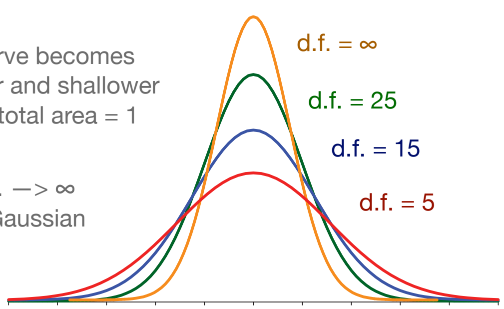
rarely know the population mean and stdev, rather sample stats

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

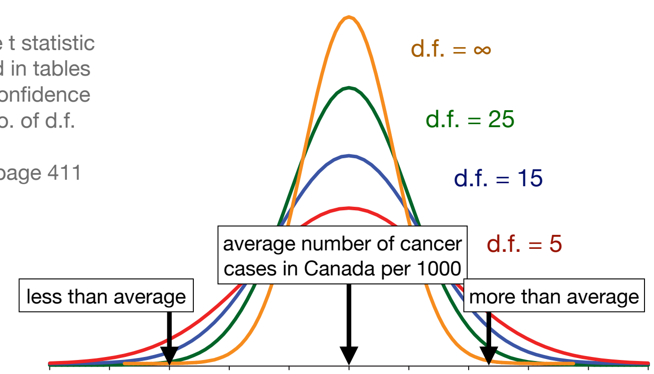
The curse of low sample numbers

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

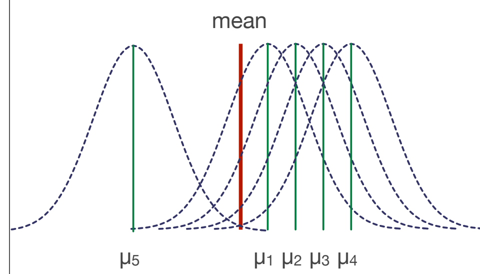
has the added advantage that we can ignore distribution (multi-modality......)

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 μ_1: prob high
from μ_2: prob lower
from μ_3: prob lower
from μ_4: prob low

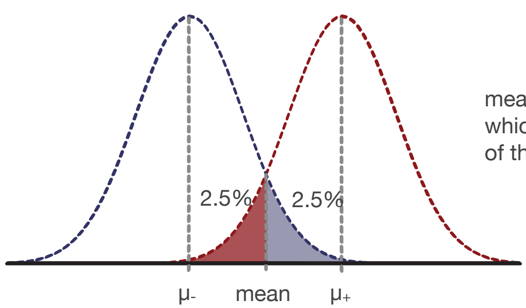
at some value of μ, will exceed the confidence level: too unlikely to come from a population with this mean: this is the upper μ

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

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
α = 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

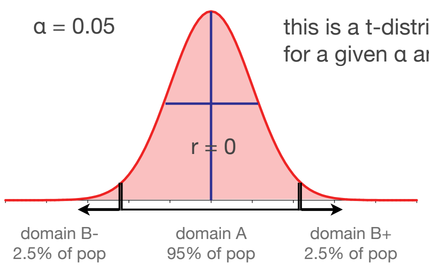
So, let's now return to the correlation coefficient:

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

α = 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

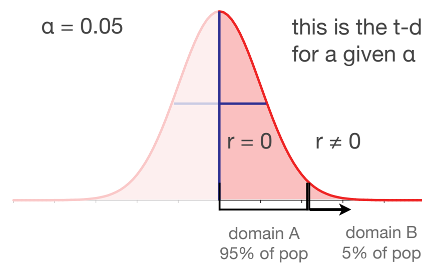
Testing the significance of the correlation coefficient:

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

α = 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
α = 0.05 or 0.025
r = -0.34

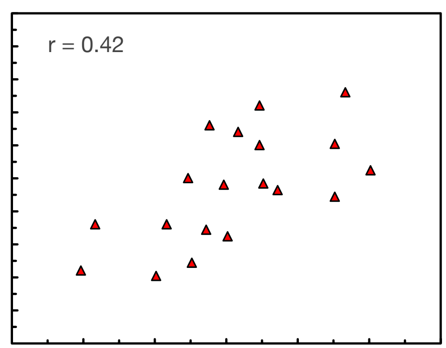
t_{calc} = -1.73
$t_{0.05;23}$ = 1.71 = -1.71
$t_{0.025;23}$ = 2.07 = -2.07

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 X

at n = 10; t = 1.31
$t_{0.05;8}$ = 1.860 X

at n = 25; t = 2.22
$t_{0.05;23}$ = 1.717 V