# Data analysis and Geostatistics - lecture VI

t-test of means, ANOVA and goodness-of-fit

## 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

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

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

13, 17, 15, 23, 27, $\qquad$ $\bar{x} = 21.3$ $\quad$ $s = 5.52$

29, 18, 27, 20, 24 $\qquad$ $n = 10$ $\quad$ $s_e = 1.75$

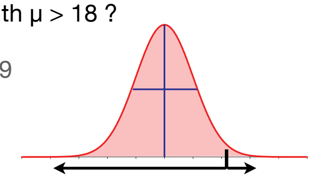
is it possible that this set is from a population with $\mu > 18$ ?

$H_0$; $\mu \leq 18$

$H_A$; $\mu > 18$

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

$t_{0.05;9} = 1.83$



## Statistical testing - comparing means

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

two sets of sandstone samples with the following porosities:

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

$n = 10$ $\qquad\qquad\qquad$ $n = 10$

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

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

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

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: $\quad$ $t_i = (\bar{x}_1 - \bar{x}_2) / SE$ $\qquad$ $H_0$; $\mu_1 = \mu_2$

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

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

$n = 10$

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

$n = 10$

$s_p = 5.18$

$s_e = 2.32$

$t_{calc} = 1.03$

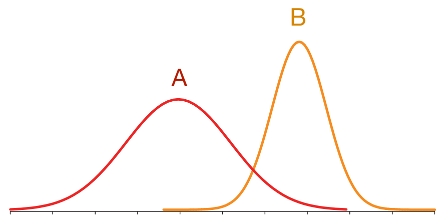
$df = n_1 + n_2 - 2$ (why?)

$t_{0.05;18} = 1.734$

## 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

## 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

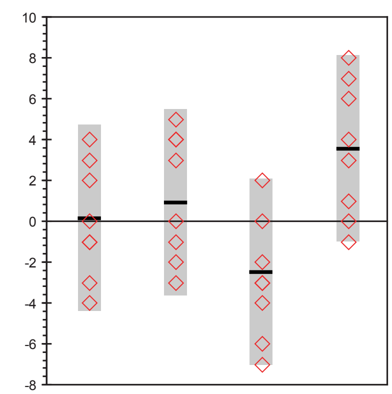
for differences to be significant, the variance within each unit has to be smaller than the variance between the units

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 $= 0.47$

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

## 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:

$s_e^2 = s^2 / n \;\; \text{->} \;\; s^2 = n \cdot s_e^2$

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:

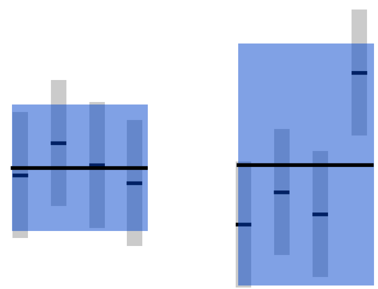
$H_0$; $\mu_1 = \mu_2 = \mu_3 = \mu_4$

$H_A$; 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 $s_{between} < s_{within}$ : all the same

$s_{between} > s_{within}$ : different at level α

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)

# 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

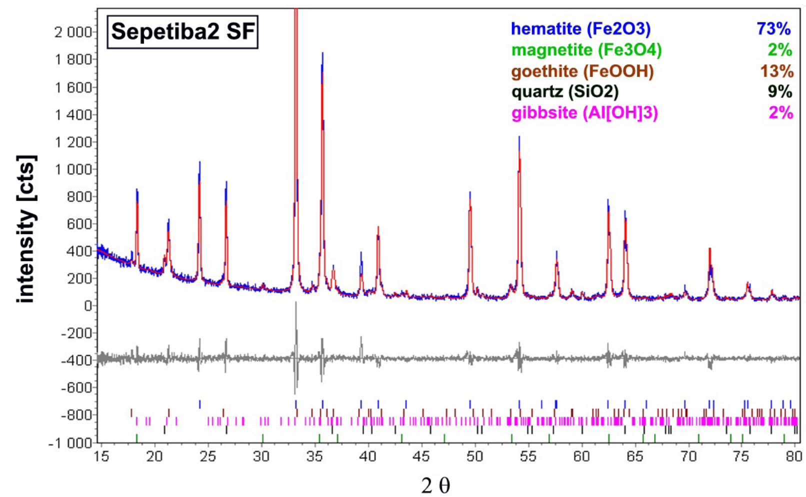
# Testing of “goodness-of-fit”

## 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

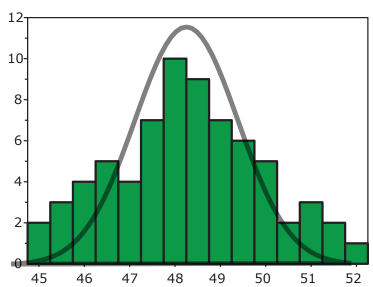



# Testing of “goodness-of-fit”

## 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

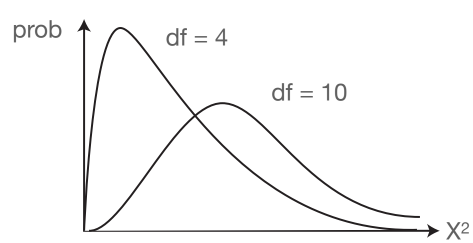
# Testing of “goodness-of-fit”

## 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

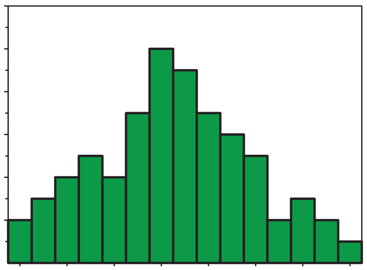



see table 2.6 on page 413

## 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

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