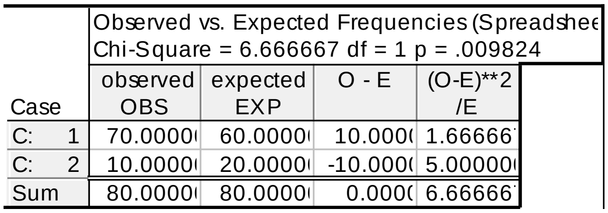
Univariate Methods

Warning: in many examples the number of replications is desperately low. This is just to keep the examples simple and small. In real problems, it is much better to have more replications. Also, majority of examples are imaginary, so the conclusions drawn are sound according to the data presented, but can contradict to the reality.

Goodness of fit (χ^2-test)

Example1: The expected Mendelian ratio in the second filial generation was 3:1. We observed 70 plants with dominant phenotype and 10 with recessive phenotype. Is there a significant difference between expected and observed ratio?

Put observed values to one variable (e.g. OBS), expected values to the other one (EXP). Use **Statistics > Nonparametrics**; ask for **Observed versus expected X.** You will get:



Observed vs. Expected Frequencies (Spreadshee
Chi-Square = 6.666667 df = 1 p = .009824

Case	observed OBS	expected EXP	O - E	(O-E)**2 /E
C: 1	70.00000	60.00000	10.0000	1.666667
C: 2	10.00000	20.00000	-10.0000	5.000000
Sum	80.00000	80.00000	0.0000	6.666667

The result of the test is significant, so we reject the null hypothesis that the observed frequencies come from the Mendelian 3:1 ratio, usually, we would write that “The observed frequencies differ significantly from the expected 3:1 ratio (χ^2=6,667, df=1, p=0,0098 [alternatively p<0,01]).

Example 2: Comparison with Hardy-Weinberg equilibrium:

Observed numbers of plant of genotypes in a sample from a population were (obtained by molecular analysis):

AA	20
Aa	40
Aa	10

First, estimate p(A) – i.e. the frequency of A allele in the population - from data: (2x20 + 40)/180 = 0.444 (90 individuals have 180 alleles)
Expected relative frequencies are p^2, 2pq, q^2
Expected number of AA is 0.444^2 x 90 = 17.777
Etc.
Note, df = number of categories – 1 – number of parameters estimated from the data (we estimated p) =
3 – 1 – 1 = 1

The number of df differs from that automatically provided by the program. You have to find the significance using **Probability calculator** in **Statistics.**

Contingency tables

Example 3: Effect of chilling on seed germination:

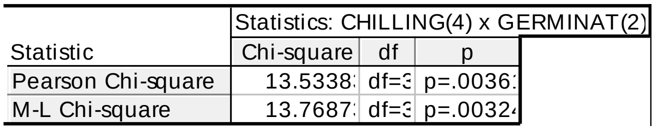
Four sets of 50 seeds were stored at four temperatures for 3 months: 20 ^oC, 4 ^oC, -4 ^oC and –20 ^oC. The germination was 30%, 40%, 60% and 60%. Each seed was treated so that it can be considered independent observation. The contingency table is (**number of cases, not percentages**):

Chilling type	Germinated	Not germinated
T=20 (chilling=1)	15	35
T=4 (chilling=2)	20	30
T=-4 (chilling=3)	30	20
T=-20 (chilling=4)	30	20

Enter data as (**file chilling.sta**):

	CHILLING	GERMINAT	FREQUE
1	1.000	1.000	15.000
2	1.000	0.000	35.000
3	2.000	1.000	20.000
4	2.000	0.000	30.000
5	3.000	1.000	30.000
6	3.000	0.000	20.000
7	4.000	1.000	30.000
8	4.000	0.000	20.000

Use **Statistics > Basic statistics** > ***Tables and Banners***
First, use FREQUE as weight, then, in the panel specify using *Specify tables* specify the grouping variables (i.e. CHILLING and GERMINAT) and in Options check *Expected frequencies* and *Pearson & M-L Chi-square* and ask for *Detailed two-way tables*.
You will get:



	Statistics: CHILLING(4) x GERMINAT(2)		
Statistic	Chi-square	df	p
Pearson Chi-square	13.5338	df=3	p=.0036
M-L Chi-square	13.7687	df=3	p=.0032

M-L is maximum likelihood Chi-square (G-test).

The null hypothesis, that the germination rate is independent of the chilling treatment was rejected (for both the tests, the p-values is smaller than 0.05, in fact, smaller than 0.01).

Other examples:

Example 4: 100 plots, 1m^2 each were randomly located in a plot and the occurrence of 2 species (*Cirsium* and *Agropyron*) was observed. In 20 plots, both species were found, in

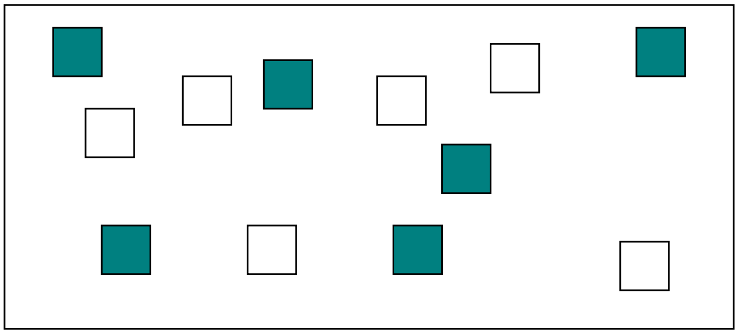
10 plots *Cirsium* only, in 20 plots *Agropyron* only, and in 50 plots none of the two species. Is the species' occurrence independent? (Possible ecological explanations: Passive and active associations).

Example 5: 50 male and 50 female plants of a dioecious species were marked in the field at the start of vegetation season. At the end of the season it was found that 40% of male plants are still alive, whereas only 22% of female plants. Is the survival rate of male and female plant different?

Comparison of two means

Note: two independent samples can be compared either by the t-test for independent samples or by one way ANOVA with two categories (**the results are identical**). In the t-test, we can have the one-sided (one-tailed) null hypothesis. (two-tailed H_0: $\mu_1=\mu_2$; one-tailed H_0:$\mu_1\leq\mu_2$ or $\mu_1\geq\mu_2$). For both methods, we expect homoscedascity (variances are equal). For t-test, we have the possibility of version with separate estimates of variance for each sample. The decision about one- or two-tailed test depends on our a-priori knowledge and intention of the test and has to be done before carrying out the test. **Note:** The other assumption is that the data come from a normal distribution. Nevertheless, what is really important is that the means have normal distribution. Consequently the test is very robust when the sample-size is large (follows from Central limit theorem).

Two independent samples (Control (open) vs. treatment (filled)):

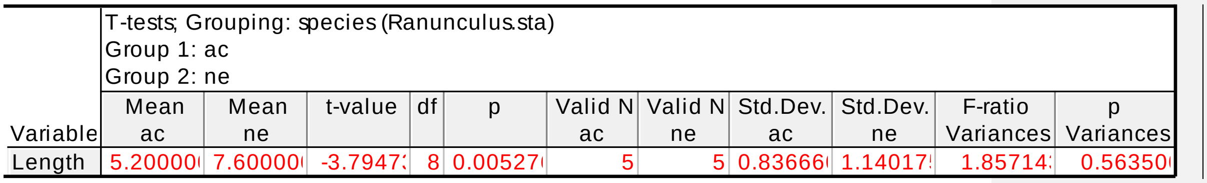


Example 6: Let's compare the length of petals in two *Ranunculus* species (*Ranunculus acer* a *R. nemorosus*). Five independent observations (Should be probably more!) are available in each **sample** (what is random independent observation and how to get it – relation of sample and population).
The values should be entered as follows

All the response values are in one variable (*length)* and the other variable (*species*) is classification of cases (tells us, to which species the observation belongs): (data are in the file **Ranunculus.sta**]

species	Length
ac	5
ac	6
ac	4
ac	6
ac	5
ne	7
ne	8
ne	9
ne	6
ne	8

Classification variable can be also a numeric one (say, 1 instead ac and 2 instead of ne) Use **Basic statistics** and *t-test for independent samples by groups*, species is the *Grouping* variable, Length is the *Response.* You will get results of t-test, and also of the F-test comparing the variances



T-tests; Grouping: species (Ranunculus.sta)
Group 1: ac
Group 2: ne

Variable	Mean ac	Mean ne	t-value	df	p	Valid N ac	Valid N ne	Std.Dev. ac	Std.Dev. ne	F-ratio Variances	p Variances
Length	5.20000	7.60000	-3.7947	8	0.00527	5	5	0.83666	1.14017	1.85714	0.56350

So, the differences in length are significant, and variances are not significantly different (which is fine, because this is the assumption of t-test) - otherwise, we would select the Option and ask for separate variance estimates.

If you are interested in one-tailed test, simply calculate P (one-tailed) = P(two-tailed)/2. (!!if the difference against the null hypothesis goes in the direction of alternative hypothesis).

Example 7:

Compare weight of seeds of two species (ten independent observations available for each species).

Weghts:

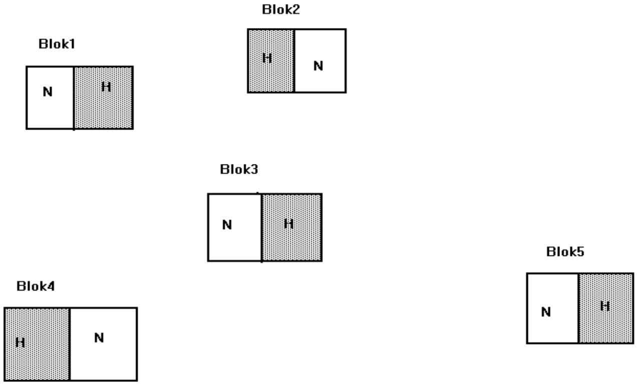
Species A: 15, 16, 17, 15, 16, 14, 15, 16, 19 , 19

Species B: 14, 13, 15, 13, 16, 14, 12, 11, 13, 15

Calculate the t-test, P-value for two-tailed test, SD, SEM (explain the difference), confidence interval, plot multiple box and whisker-plot.

Two dependent samples (paired t-test)

Example 8. Five blocks (the experiment was carried out in Czech republic, so the block is called blok) were diveded in two half, one fertilized (Nitrogen - N) and other was control (H).:



Biomass values in particular plots:

Block	1	2	3	4	5
Fertilized	23	25	36	19	22
Unfertilized	20	24	33	18	21

Does fertilizer have any effect? (Consider one-tailed test, when we want to test whether nitrogen is a limiting factor in the plot)

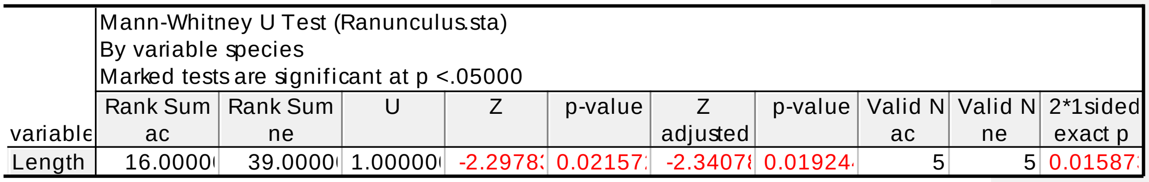
The data are entered in two variables. one variable for fertilized and one for unfertilized plot, each block is a case }so no need to have a variable for case). Ask for t-test for dependent samples. Results: t = 3.674235, df=4, p=0.021312

Other examples of paired observations: Comparison of bark thickness on northern and southern site of a tree: for each tree you have two values – one for southern, one for northern.

Comparison of students' weight before and after visit at parents' house.

Non-parametric counterparts:

t-test for independent samples: Mann-Whitney U test (in Statistica **Nonparametrics > comparing two independent samples, response and groupins as for t-test, ask for Mann-Whitney U test – with Ranunculus.sta data, you will get** .



	Mann-Whitney U Test (Ranunculus.sta) By variable species Marked tests are significant at p <.05000									
variable	Rank Sum ac	Rank Sum ne	U	Z	p-value	Z adjusted	p-value	Valid N ac	Valid N ne	2*1sided exact p
Length	16.0000	39.0000	1.00000	-2.2978	0.02157	-2.3407	0.01924	5	5	0.01587

Paired t-test (t-tesp for dependent samples) – Wilcoxon matched pairs test in Nonparametrics/- ask for two dependent variables – data are entered exactly as for paired t-test

Take care, when using the non-parametric test, you either test the hypothesis, that the distributions are identical (then there are no assumptions about distributions), or you test equality of means (or medians), but then you assume, that the distribution shape is identical, and test, whether the distributions differ in location.

Comparison of more than two means – ANOVA

ANOVA for two groups and t-test are identical; multiple t-test is not advisable, because the probability of Type I error is α in each of the t-tests, and consequently, probability of Type I error in at least one of the particular test is very high – this can lead to “statistical fishing”.

One-way ANOVA

(completely randomized design)

Example 9: Effect of soil type on plant height was tested in a pot experiment. 5 plants were grown in sandy soil, 5 plants in clay soil, and 5 plants in a peat soil. The final heights are in a table (in a way, how they should be entered for Statistica (i.e. grouping variable [= soil] and response [=height]) – **file soiltype.sta**:

(Note: soil type is a factor with fixed effect.)

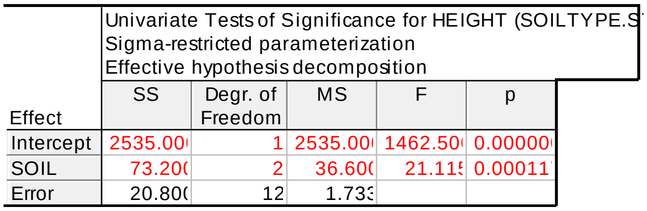
CASE	SOIL	HEIGHT
1	s	15.000
2	s	17.000
3	s	14.000
4	s	16.000
5	s	17.000
6	c	13.000
7	c	12.000
8	c	11.000
9	c	13.000
10	c	15.000
11	p	11.000
12	p	12.000
13	p	10.000
14	p	9.000
15	p	10.000

Use *Statistics > ANOVA/MANOVA > One way ANOVA.*
In very similar manner, you can use *Statistics > Advanced linear/nonlinear models > General linear models* **-** and then ask for *One way ANOVA* there. Its slightly longer, but you have more options there.

In the panel (click Variables):
Independent (factors): soil
Dependent: Height

Press **OK**, and in the next panel ask for **All effects**

You will get the ANOVA result table:

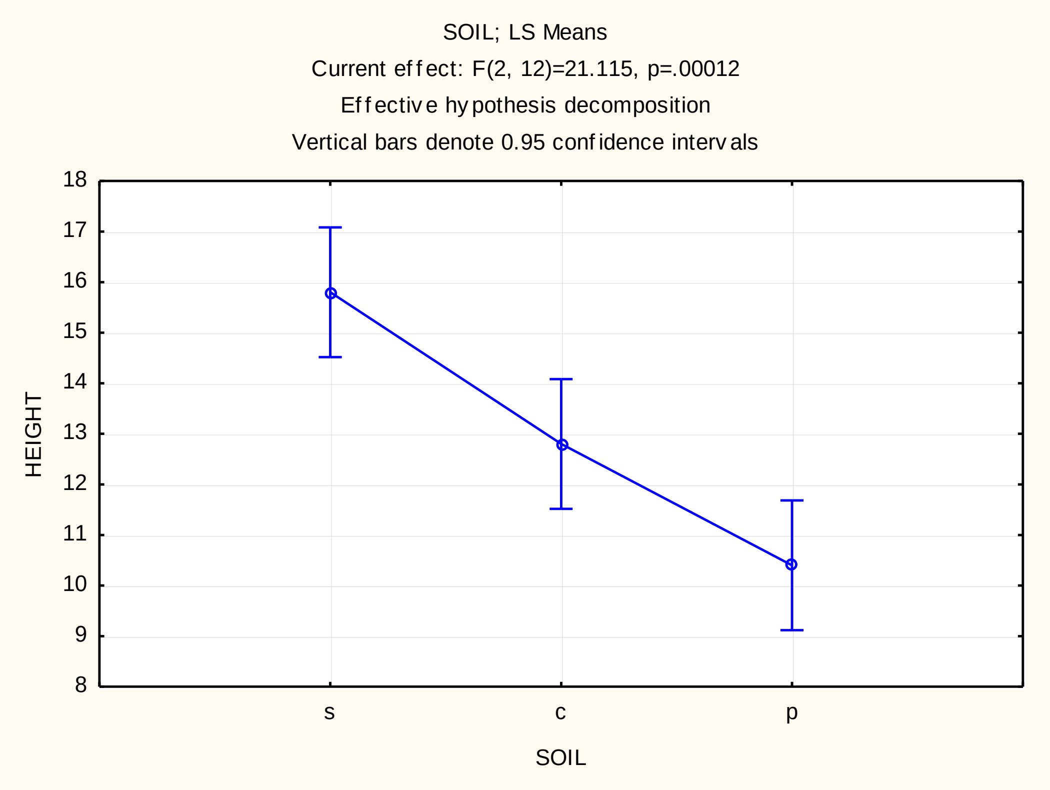


Univariate Tests of Significance for HEIGHT (SOILTYPE.S
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	2535.00	1	2535.00	1462.50	0.00000
SOIL	73.20	2	36.60	21.11	0.00011
Error	20.80	12	1.73		

As p=0.000117, we can conclude that the effect of soil type is highly significant.
Note that test of intercept has here no real meaning – in fact, it test the hypothesis, that the mean height of all the plants over all the groups is zero, which is a nonsense.

Reasonable graphical presentation can be obtained by *All effects/Graphs*



For multiple comparisons ask More results, Post hoc comparisons (unless you have a priori planned ones). Tukey is recommended.

Other examples:

Random factor (note that for the one-way ANOVA, the results are the same for fixed and random factors): Individuals from three clones of *Festuca rubra* were vegetatively propagated under identical conditions. Then, 5 tillers from each clone were grown, each in a separate pot, for 5 weeks and the number of tillers was calculated to find, whether there is effect of genetic variability (i.e. the difference between clones) on tillering.
Results (number of additional tillers from each of original 5 tillers):
Clone 1: 6,4,5,8,6
Clone 2: 2,3,2,4,3
Clone 3: 4,6,5,7,4

Probably, the multiple comparison is meaningless.
Probably, the square-root transformation can be useful.
When to use the log-transformation? When the data are log-normal, sd is linearly dependent on mean and effects are multiplicative.

Non-parametric counterpart: Kruskal-Wallis ANOVA (or median test). Use procedure **Nonparametrics/comparing multiple independent samples**., *Kruskal-Wallis*. Panel is similar to parametric test.

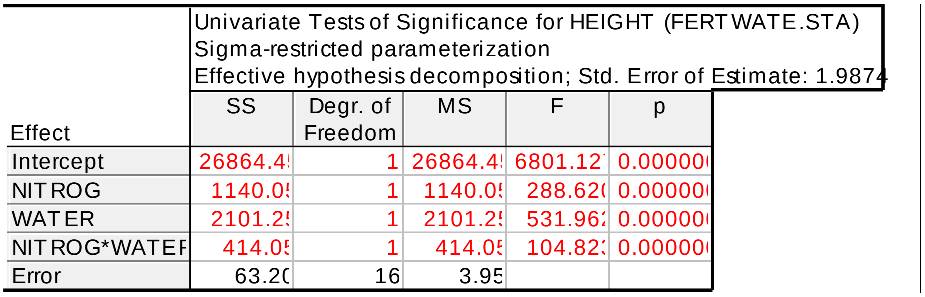
Two-way analysis of variance: factorial experimental design

Example10:
Effect of nitrogen and watering on plant height was studied in a pot experiment. Two levels of each factor were applied (normal – 0, increased – 1)
Enter each of independent factors into one variable (**file fertwate.sta**)

	Nitrog	Water	Height
1	0.000	0.000	23.000
2	0.000	0.000	25.000
3	0.000	0.000	24.000
4	0.000	0.000	26.000
5	0.000	0.000	19.000
6	0.000	1.000	32.000
7	0.000	1.000	37.000
8	0.000	1.000	34.000
9	0.000	1.000	35.000
10	0.000	1.000	36.000
11	1.000	0.000	29.000
12	1.000	0.000	28.000
13	1.000	0.000	29.000
14	1.000	0.000	31.000
15	1.000	0.000	30.000
16	1.000	1.000	57.000
17	1.000	1.000	59.000
18	1.000	1.000	62.000
19	1.000	1.000	58.000
20	1.000	1.000	59.000

Use *Statistics > ANOVA/MANOVA > Factorial ANOVA.*

In very similar manner, you can use *Statistics > Advanced linear/nonlinear models > General linear models* - and then ask for *Factoria ANOVA* there. We will use this second option here
You will use : Nitrog and Water are categorical predictorst, Height is dependent. After **All effects** you will get:

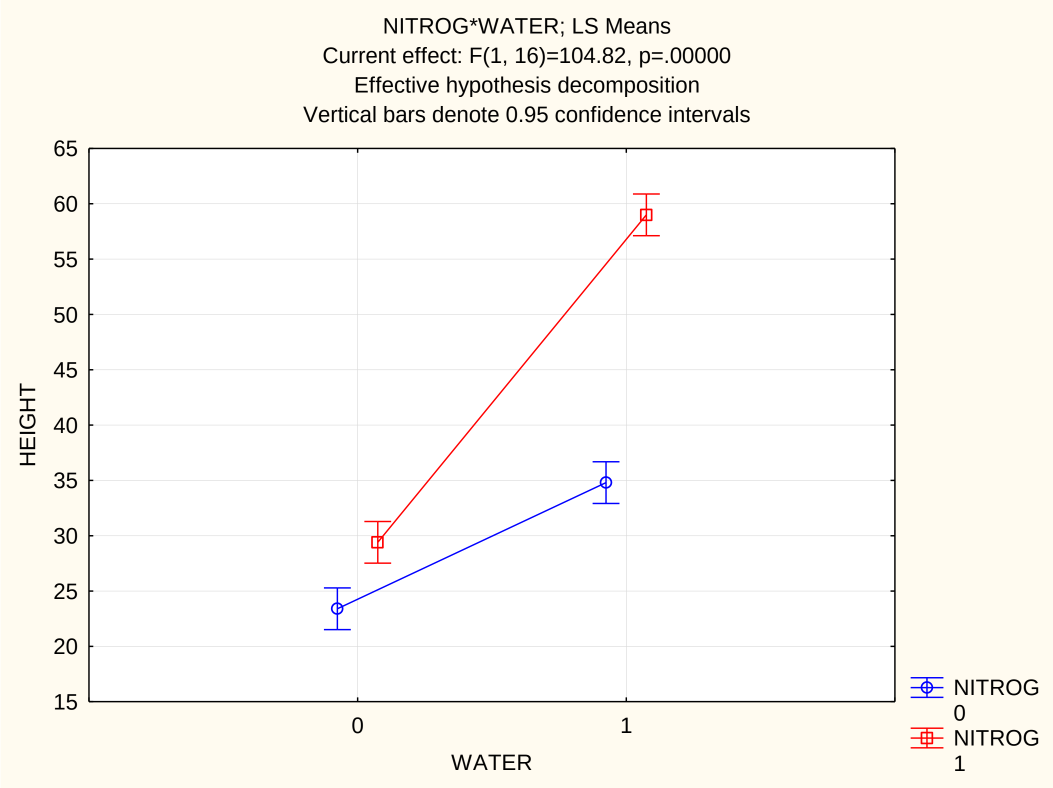


Univariate Tests of Significance for HEIGHT (FERTWATE.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 1.9874

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	26864.4	1	26864.4	6801.12	0.00000
NITROG	1140.0	1	1140.0	288.62	0.00000
WATER	2101.2	1	2101.2	531.96	0.00000
NITROG*WATE	414.0	1	414.0	104.82	0.00000
Error	63.2	16	3.95		

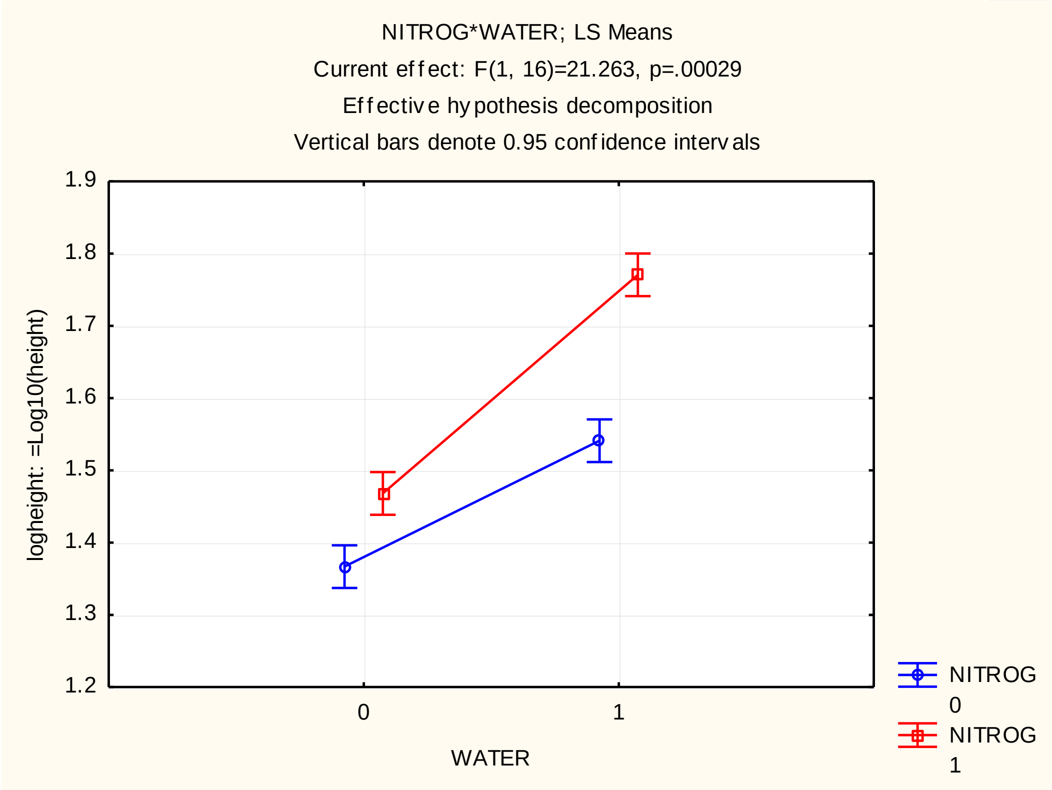
Note, that test of |Intercept is again meaningless

Meaning of interaction: the main effect are not additive; see the picture obtained form **All effects/graphs** after asking for interactions:



The lines are not parallel => effects are not additive. It is, however, a question, whether the additivity is a good null hypothesis. With height, one can think a multiplicativity as a good null hypothesis. Try with the log-transformed height. (After the log transformation, you test the null hypothesis of multiplicativity on the original (non-transformed) scale. You will get

	Univariate Tests of Significance for logheight (FERTWATE.STA) Sigma-restricted parameterization Effective hypothesis decomposition; Std. Error of Estimate: .03110				
Effect	SS	Degr. of Freedom	MS	F	p
Intercept	47.2247	1	47.2247	48821.2	0.00000
NITROG	0.1369	1	0.1369	141.5	0.00000
WATER	0.2843	1	0.2843	293.9	0.00000
NITROG*WATEF	0.0205	1	0.0205	21.2	0.00028
Error	0.0154	16	0.0009		

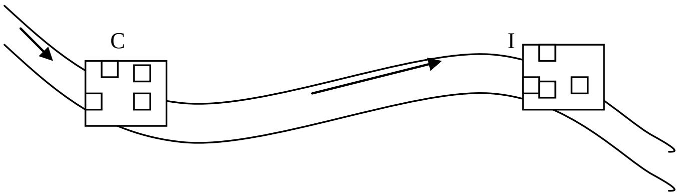


So, even after the log transformation, the interaction is significant, i.e. the common application of nitrogen and water increases height more than would be expected from the additive effect from each of the factors (the test with non-transformed values), but even more than would be expected from the multiplicative effect of the two factors (the significant interaction of the log-transformed data).

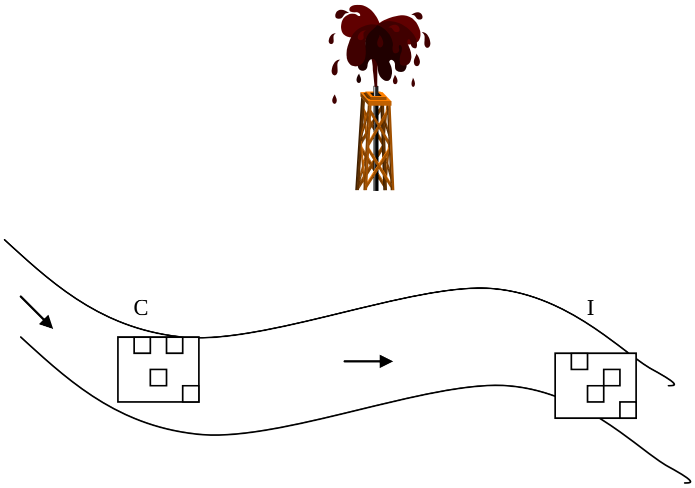
Note that log-transformation changes also the distributional characteristics of the variables (would be probably good to have a look on residuals disdtribution – in my view better than to perform formal test), but the changed semantics is, in my view, the most important matter.

Non-replicated BACI (Before After Control Impact)

Before:



After:



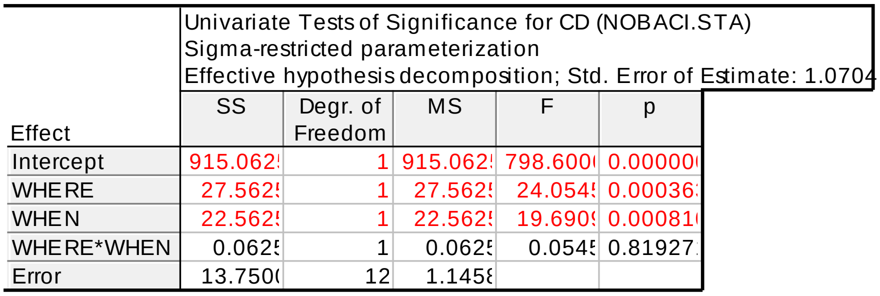
The response (e.g. content of Cd and Pb in algae, file noBACI.sta) is analyzed by two way analysis of variance. Main factors are WHEN (**B**efore and **A**fter impact) and

WHERE (above [**C**ontrol plot] and below [**I**mpact plot] the oil spill). The significant interaction is (with caution because of pseudoreplication) considered to be a proof of impact:

Data: (file **NOBACI.STA)**

	WHERE	WHEN	CD	PB
1	C	B	5.000	4.000
2	C	B	4.000	6.000
3	C	B	6.000	5.000
4	C	B	5.000	3.000
5	I	B	8.000	6.000
6	I	B	9.000	5.000
7	I	B	6.000	7.000
8	I	B	8.000	7.000
9	C	A	6.000	4.000
10	C	A	7.000	7.000
11	C	A	9.000	7.000
12	C	A	8.000	6.000
13	I	A	10.000	11.000
14	I	A	11.000	13.000
15	I	A	9.000	12.000
16	I	A	10.000	14.000

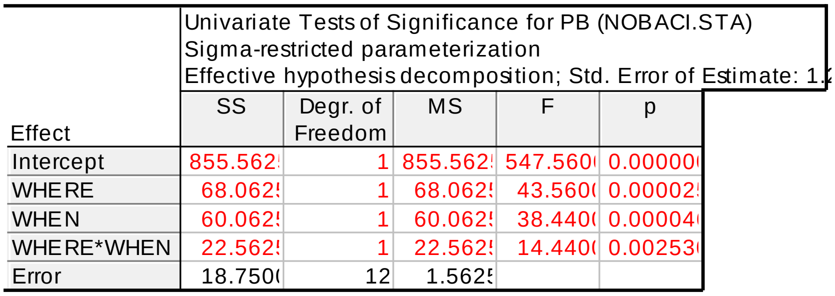
Results:



Univariate Tests of Significance for CD (NOBACI.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 1.0704

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	915.062	1	915.062	798.600	0.00000
WHERE	27.562	1	27.562	24.054	0.00036
WHEN	22.562	1	22.562	19.690	0.00081
WHERE*WHEN	0.062	1	0.062	0.054	0.81927
Error	13.750	12	1.145		

Cadmiun – the interaction is not significant

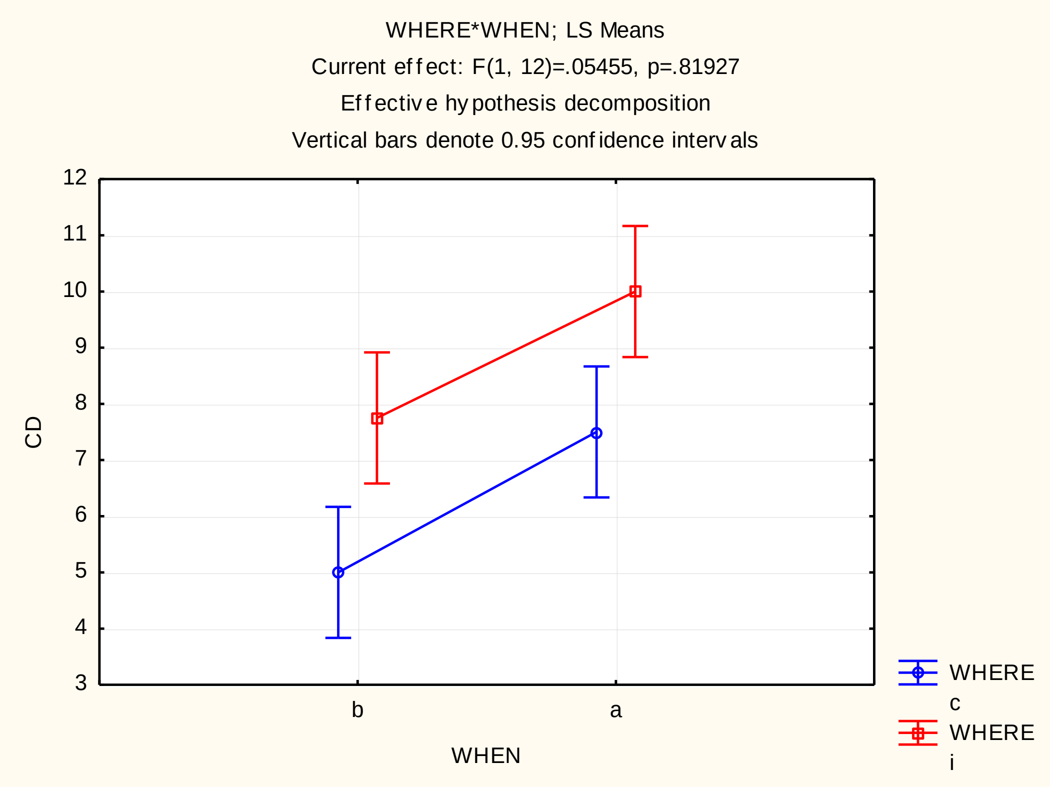


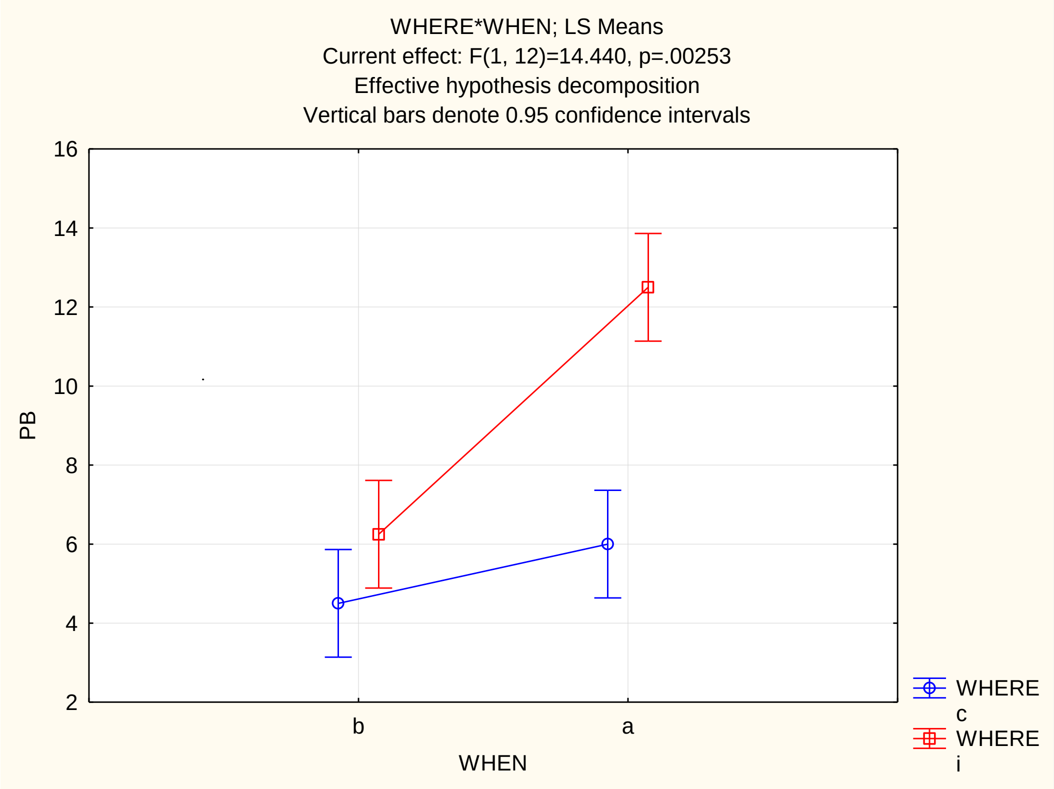
Univariate Tests of Significance for PB (NOBACI.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 1.

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	855.562	1	855.562	547.560	0.00000
WHERE	68.062	1	68.062	43.560	0.00002
WHEN	60.062	1	60.062	38.440	0.00004
WHERE*WHEN	22.562	1	22.562	14.440	0.00253
Error	18.750	12	1.562		

Lead (Pb) – the interaction is significant

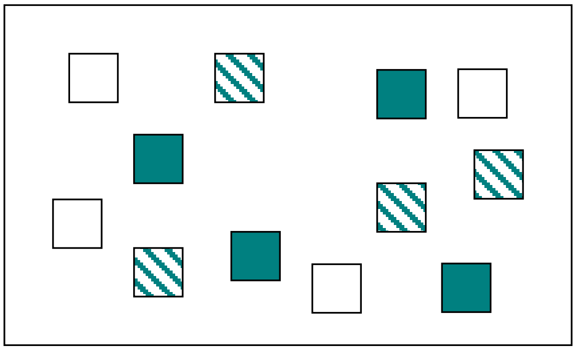
We have no reason to expect the effect on Cd (interaction is non-significant – accordingly, lines in graph are parallel), even when both main effects are significant. On the contrary, there is effect on Pb.



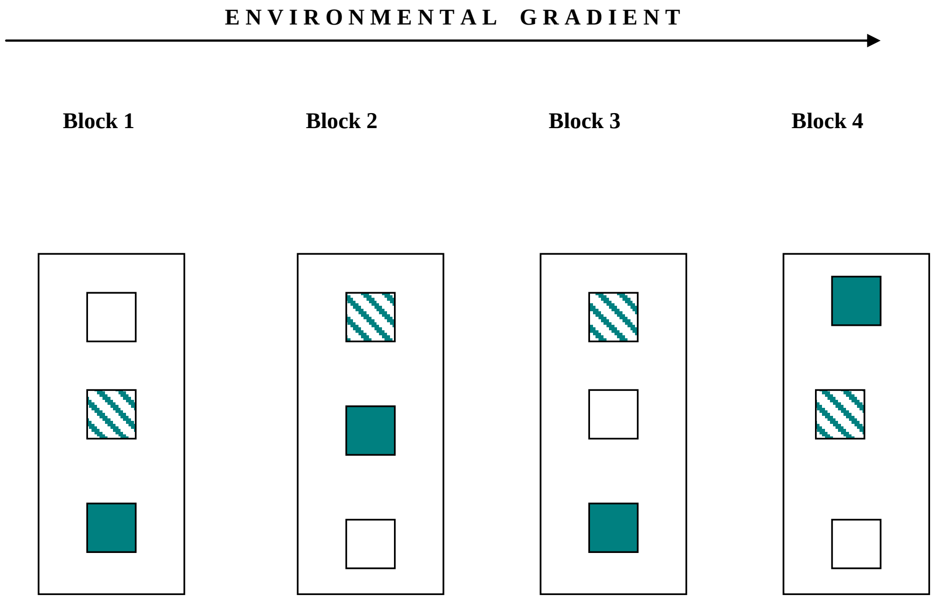

WHERE*WHEN; LS Means
Current effect: F(1, 12)=14.440, p=.00253
Effective hypothesis decomposition
Vertical bars denote 0.95 confidence intervals
16
14
12
10
8
6
4
2
PB
b
a
WHEN
WHERE c
WHERE i

Experimental design:

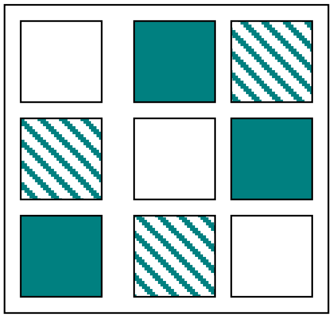
Completely randomized (correct)



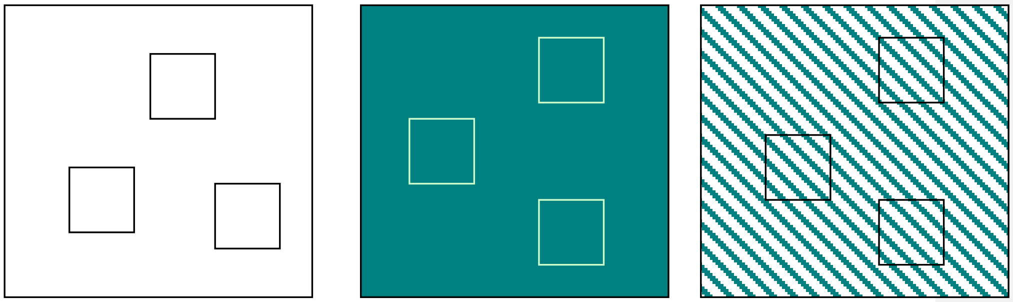
Randomized complete blocks (correct):



Latin square design (correct)



FALSE (Pseudoreplications!!!!)

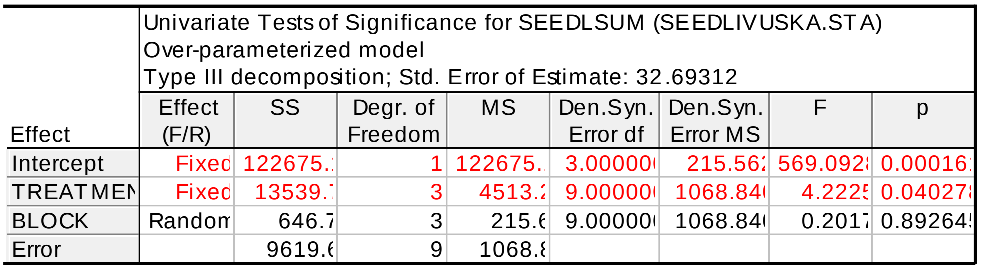


Randomized complete blocks: (Example 11, file seedlivuska.sta): In an experiment set in 4 randomized complete blocks, following treatments were used: control (1), litter removal (2), Nardus removal (3) and litter and moss removal (4). The response is number of seedling per 0.5m x 0.5m plot (seedlsum).

	TREATMEN	BLOCK	SEEDLSUM
rel1	1	1	95
rel2	2	1	91
rel3	3	1	64
rel4	4	1	107
rel5	1	2	88
rel6	2	2	70
rel7	3	2	51
rel8	4	2	180
rel9	1	3	44
rel10	2	3	57
rel11	3	3	55
rel12	4	3	173
rel13	1	4	94
rel14	2	4	99
rel15	3	4	53
rel16	4	4	80

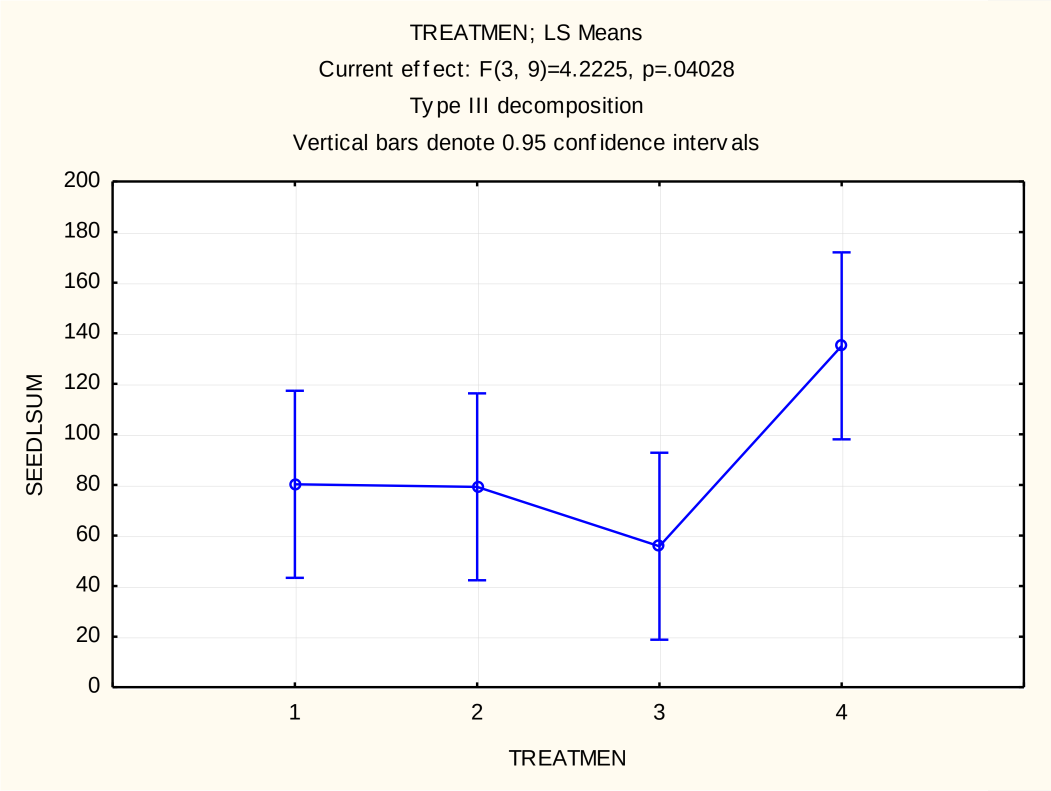
Analyzed by main effect ANOVA, (TREATMENT and BLOCK are main effect, interaction term is used as error term – of course, interaction cannot be tested. The same result is obtained, when you use in Statistica the factorial design, the interaction will not be tested)

In Statistica: *Statistics* > *Advanced linear/nonlinear models* > *General linear models*, select *Main effect ANOVA* and in *Options* specify BLOCK as a random effect factor

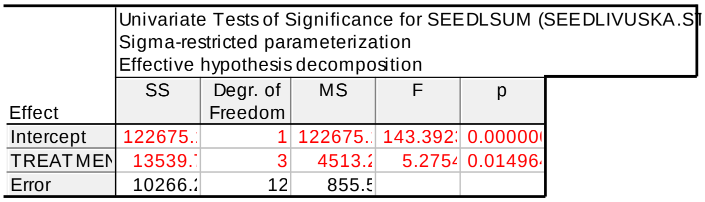


Univariate Tests of Significance for SEEDLSUM (SEEDLIVUSKA.STA)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: 32.69312

Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixed	122675.	1	122675.	3.00000	215.56	569.092	0.00016
TREATMEN	Fixed	13539.	3	4513.	9.00000	1068.84	4.222	0.04027
BLOCK	Random	646.7	3	215.6	9.00000	1068.84	0.2017	0.89264
Error		9619.	9	1068.				



If blocks do not differ among themselves (note that p-value for block is ~0.9), then block structure decreases the power of the test. In example above, the completely randomized design would yield:



Effect	SS	Degr. of Freedom	MS	F	p
	Univariate Tests of Significance for SEEDLSUM (SEEDLIVUSKA.ST Sigma-restricted parameterization Effective hypothesis decomposition				
Intercept	122675.	1	122675.	143.392	0.00000
TREATMEN	13539.	3	4513.	5.275	0.01496
Error	10266.	12	855.5		

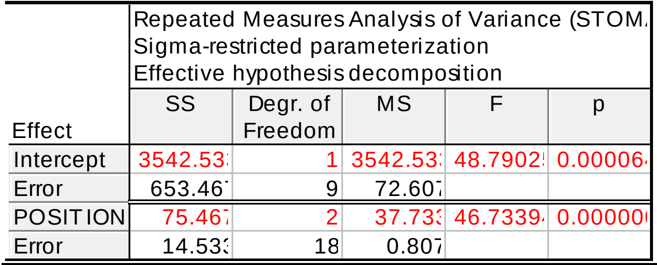
It is, however, not correct to drop the block only because it is not significant.
 Another possible data arrangement for randomized complete blocks is in the file STOMATA.STA.

Example12 (file stomata.sta):

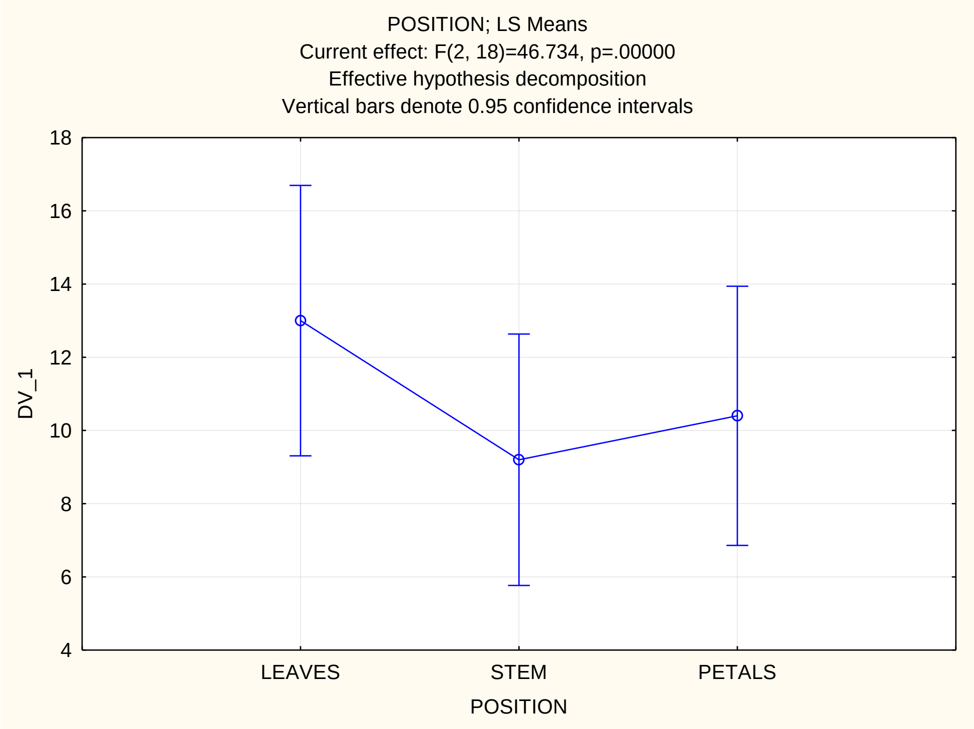
Stomatal densities on leaves, stem and petals were compared. 10 plants were used and for each plant, we have one value for leaves, one value for stem and one value for petals:

Plant	Leaves	Stem	petals
1	9	6	7
2	15	9	10
3	7	3	4
4	15	10	12
5	11	7	9
6	20	15	17
7	19	18	18
8	4	3	3
9	16	11	13
10	14	10	11

This means, each block (in this case, each plant is a block) is a ro, each variable (column) is one treatment level (in this case, position on a plant, i.e., not real treatment, just explanatory variable). In this case, specify *Statistics* > *ANOVA* > *Repeated measure ANOVA* – you will have three response variables (Leaves, Stem, Petals), and in the wizard, specify *Within effect* is *POSITION* with three levels – and you will get (*All effects)*

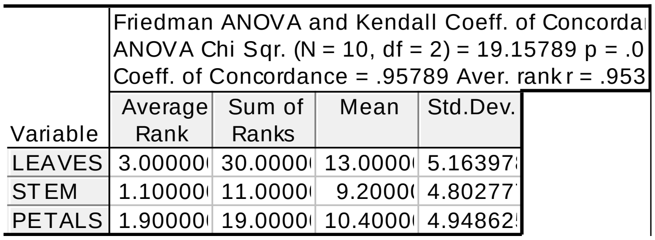


Effect	SS	Degr. of Freedom	MS	F	p
	Repeated Measures Analysis of Variance (STOM. Sigma-restricted parameterization Effective hypothesis decomposition				
Intercept	3542.53	1	3542.53	48.7902	0.00006
Error	653.46	9	72.60		
POSITION	75.46	2	37.73	46.7339	0.00000
Error	14.53	18	0.807		



Note, that in this case, the confidence intervals are calculated without taking into account the block structure – so these are correct confidence intervals, but do not take into account, that you have first filtered out the variability of individual plants using them as a block.

Non-parametric counterpart: Friedman test (in *Statistics > Nonparametrics/compare multiple dependenty samples (variables)*



Friedman ANOVA and Kendall Coeff. of Concorda
ANOVA Chi Sqr. (N = 10, df = 2) = 19.15789 p = .0
Coeff. of Concordance = .95789 Aver. rank r = .953

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
LEAVES	3.00000	30.0000	13.0000	5.16397
STEM	1.10000	11.0000	9.2000	4.80277
PETALS	1.90000	19.0000	10.4000	4.94862

Fixed and random effects

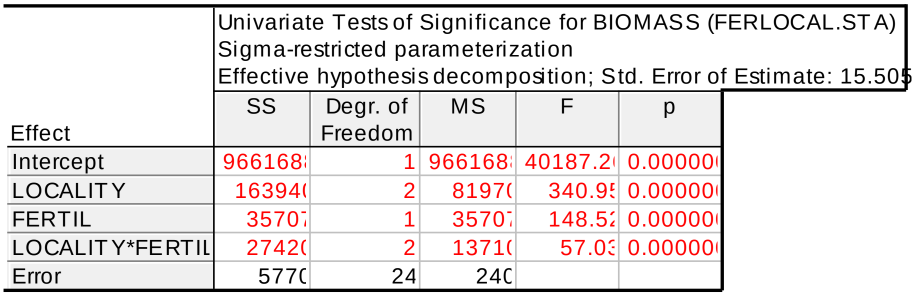
Example 13 (file ferlocal.sta): In three meadow localities, 5 control plots and 5 fertilized plots were established. The biomass at the end of the season was harvested, oven dried and weighted. Following results were obtained:

FERTIL BIOMASS
LOCALITY

1	0	510
1	0	520
1	0	525
1	0	545
1	0	500
1	1	600
1	1	610
1	1	620
1	1	610
1	1	605
2	0	400
2	0	420
2	0	410
2	0	405
2	0	430
2	1	520
2	1	570
2	1	560
2	1	520
2	1	550
3	0	680
3	0	670
3	0	650
3	0	660
3	0	670
3	1	670
3	1	650
3	1	630
3	1	645
3	1	670

Are there differences among localities? Is there any effect of fertilization? Is the fertilization effect the same at all the localities?

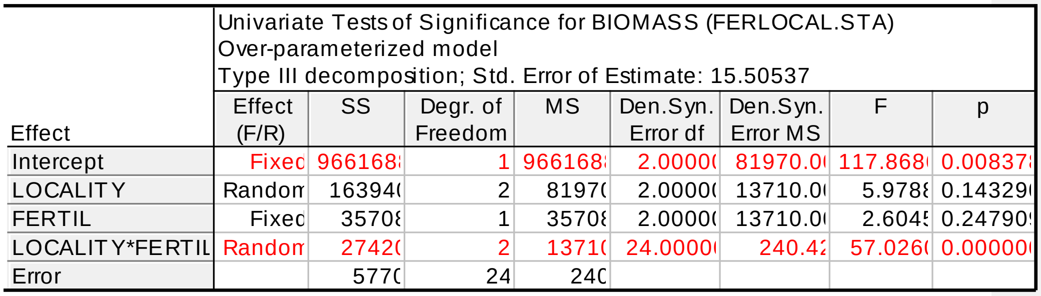
Use the factorial ANOVA (you need to use the Advanced linear/nonlinear models >general linear models, if you select ANOVA only, you are not able to specify the random effects|). Compare the results when locality is a **fixed** effect factor:



Univariate Tests of Significance for BIOMASS (FERLOCAL.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 15.505

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	966168	1	966168	40187.2	0.000000
LOCALITY	163940	2	81970	340.9	0.000000
FERTIL	3570	1	3570	148.5	0.000000
LOCALITY*FERTIL	27420	2	13710	57.03	0.000000
Error	5770	24	240		

And when locality is a **random** effect factor:



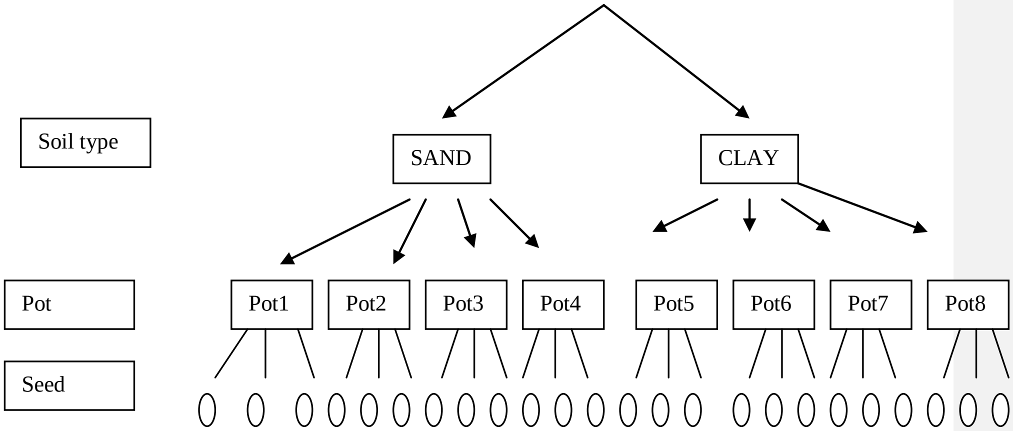
Univariate Tests of Significance for BIOMASS (FERLOCAL.STA)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: 15.50537

Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixed	966168	1	966168	2.00000	81970.00	117.8680	0.00837
LOCALITY	Random	163940	2	81970	2.00000	13710.00	5.978	0.143290
FERTIL	Fixed	3570	1	3570	2.00000	13710.00	2.604	0.24790
LOCALITY*FERTIL	Random	27420	2	13710	24.00000	240.4	57.0260	0.000000
Error		5770	24	240				

The results for the fixed factor (i.e. fertile) differ considerably (the results for the other two terms should be in my view identical, but are not – I am convinced that in this case, Statistica uses a wrong error term for the random effect). But, so as so, we are usually most interested in the fixed effect. There is difference in the meaning: when locality is a fixed factor, the results are to be generalized to the three localities only (i.e., on average, the fertilization increases biomass on the three localities). When the locality is a random factor, then the three localities are random sample from (potentially infinite) set of all possible localities; in this case we do not have enough evidence to say anything about the fertilization effect in the whole set (except that the effect is not the same in all the localities (significant interaction).

Hierarchical (nested) designs

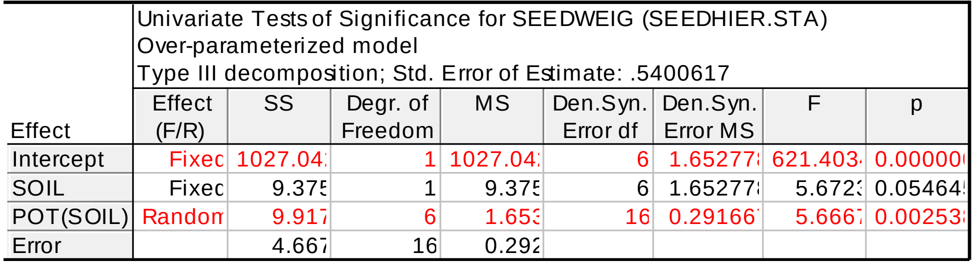
Simple hierarchy: Example 14: We study the effect of soil type on seed weight. We have four pots with sand and four pots with clay. From each plant, we weighted 3 seeds. The design was:



The data should be entered as follows (file **seedhier.sta**):

	SOIL	POT	SEEDWEIG
1	s	1	6
2	s	1	7
3	s	1	6
4	s	2	5
5	s	2	6
6	s	2	5
7	s	3	7
8	s	3	7
9	s	3	6
10	s	4	5
11	s	4	5
12	s	4	6
13	c	5	8
14	c	5	7
15	c	5	8
16	c	6	7
17	c	6	7
18	c	6	8
19	c	7	8
20	c	7	7
21	c	7	8
22	c	8	6
23	c	8	6
24	c	8	6

The analysis of variance has to reflect the hierarchical nature of the design: in particular, pot (a random factor) is nested the factor soil. So, use *Statistics* > *Advanced* *linear/nonlinear models* > *General linear models* and there, select Nested designs – there in *Between effect* specify that Pot is nested in soil (Soil is not nested) and finally, you have to state that pot is a factor with random effect (in Options). Ignore the warning. You will get:

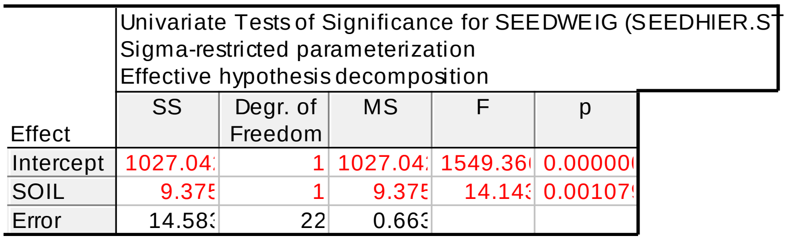


Univariate Tests of Significance for SEEDWEIG (SEEDHIER.STA)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: .5400617

Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixec	1027.04	1	1027.04	6	1.65277	621.403	0.00000
SOIL	Fixec	9.375	1	9.375	6	1.65277	5.672	0.05464
POT(SOIL)	Randon	9.917	6	1.653	16	0.29166	5.666	0.00253
Error		4.667	16	0.292				

It follows that (at α=0.05) we were not able to reject the null hypothesis that soil has no effect, but there is significant effect of the pot. Note, that for soil we have used as an error term MS for pot, not the residual MS. For testing the effect of soil, particular pots are the independent observations. The pots are tested against the residual (i.e. between seed within a pot) variability.

If we use (erroneously) the particular seeds as independent observations, we would get nicely significant differences between soil type:



Univariate Tests of Significance for SEEDWEIG (SEEDHIER.ST
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1027.04	1	1027.04	1549.36	0.00000
SOIL	9.375	1	9.375	14.14	0.00107
Error	14.58	22	0.66		

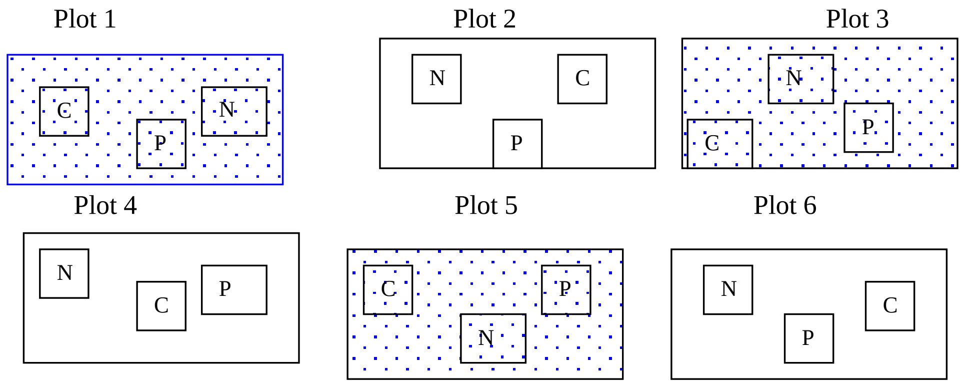
Unfortunately, this is **false analysis**, and tremendously underestimates the Type I error probability.

Split-plot design

Split-plot is sometimes used also for the simple hierarchy described above; here we will call split-plot the situation where there is a within-plot factor, effect of which is also tested.

Example 15:

The effect of fertilization was studied on 6 plots, 3 of them on limestone □ , and 3 of them on granit. ⬚ In each field following treatment were established: control (C), fertilized by Nitrogen (N) and fertilized by Phosphorus (P). The design looked like:

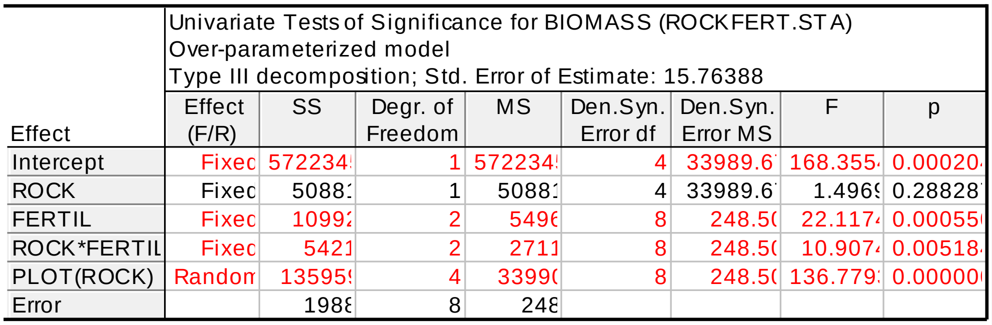


The response was total biomass in a plot. We are interested in following questions: Is there any difference between biomass on granit and limestone (test rock), is there any general effect of fertilization (test fertil), and the effect of fertilization the same on granit and on limestone (test interaction rock x fertil). Because of the hierarchical structure, we are not allowed to use the two-way analysis of variance, but we have to include the plot (1 to 6) as another factor, which is nested within rock.

The data should be entered as (file rockfert.sta):

	ROCK	FERTIL	PLOT	BIOMASS
1	g	C	1	625
2	g	N	1	688
3	g	P	1	645
4	l	C	2	455
5	l	N	2	482
6	l	P	2	520
7	g	C	3	695
8	g	N	3	756
9	g	P	3	740
10	l	C	4	420
11	l	N	4	460
12	l	P	4	499
13	g	C	5	460
14	g	N	5	488
15	g	P	5	456
16	l	C	6	520
17	l	N	6	590
18	l	P	6	650

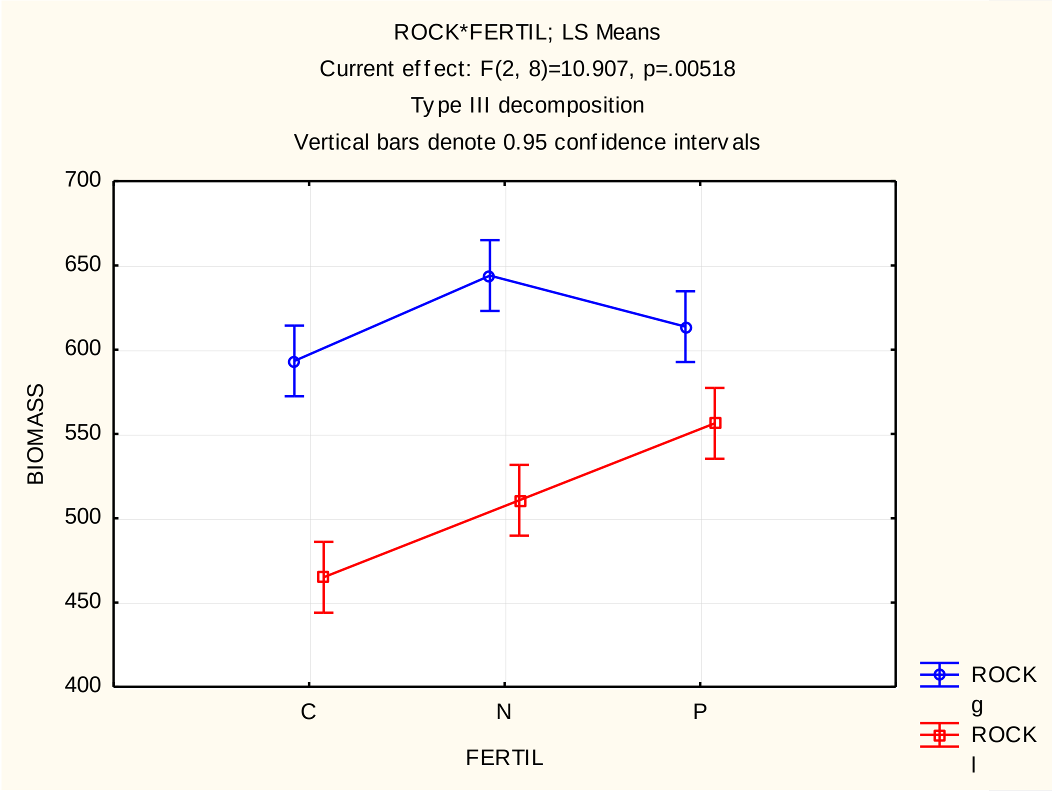
The independent variables are ROCK, FERTIL and PLOT, dependent is BIOMASS. Then (in general linear model (last line) in general linear models) state that you want to use as categorical predictors ROCK, FERTIL and PLOT, and in between effect (custom design) state that PLOT is nested within ROCK, you are also interested in the ROCK * FERTIL interaction, and that PLOT is a random factor (Options). The final results are:



Univariate Tests of Significance for BIOMASS (ROCKFERT.STA)
Over-parameterized model
Type III decomposition; Std. Error of Estimate: 15.76388

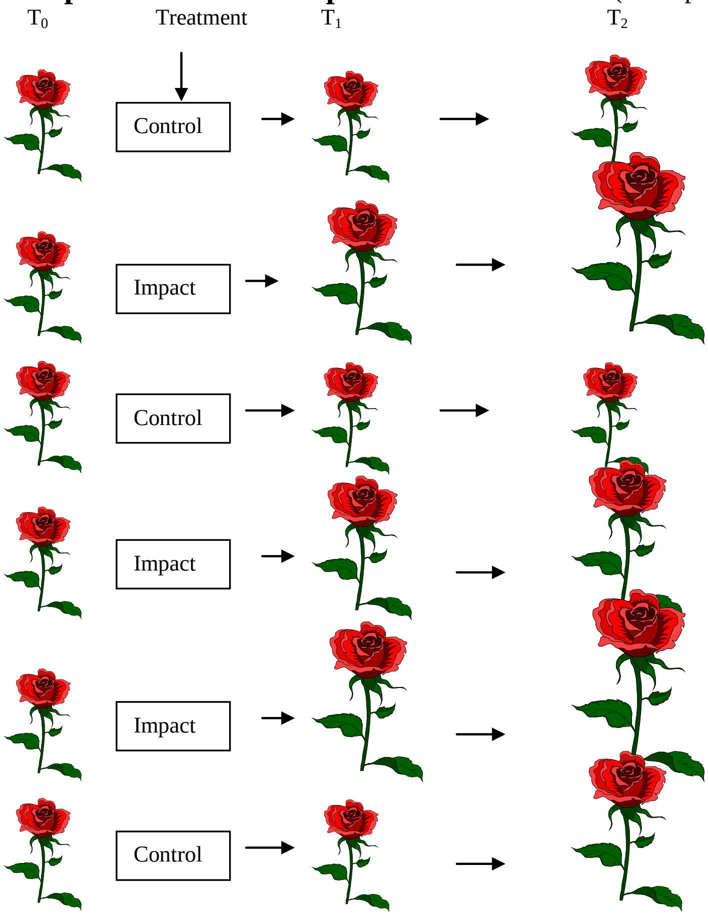
Effect	Effect (F/R)	SS	Degr. of Freedom	MS	Den.Syn. Error df	Den.Syn. Error MS	F	p
Intercept	Fixed	5722345	1	5722345	4	33989.67	168.3554	0.000204
ROCK	Fixed	50881	1	50881	4	33989.67	1.4969	0.288287
FERTIL	Fixed	10992	2	5496	8	248.50	22.1174	0.000550
ROCK*FERTIL	Fixed	5421	2	2711	8	248.50	10.9074	0.005184
PLOT(ROCK)	Random	135959	4	33990	8	248.50	136.7793	0.000000
Error		1988	8	248				

Note, that for the effect of ROCK ("main plot effect)", the PLOT MS is used as error in F calculation. We can conclude that on average, the biomass do not differ between limestone and granit, that the fertilization has a significant effect, and that the effect of fertilization is NOT the same on granit and limestone: this can be illustrated by a picture (use means/graph and plot interaction ROCK and FERTIL):



On limestone, the effect of phosphorus is higher than that of nitrogen, on granit, the reverse is true. The confidence intervals are here based on

Replicated BACI – Repeated measurement (Example 16)



H0: Growth is the same in control and impact group.

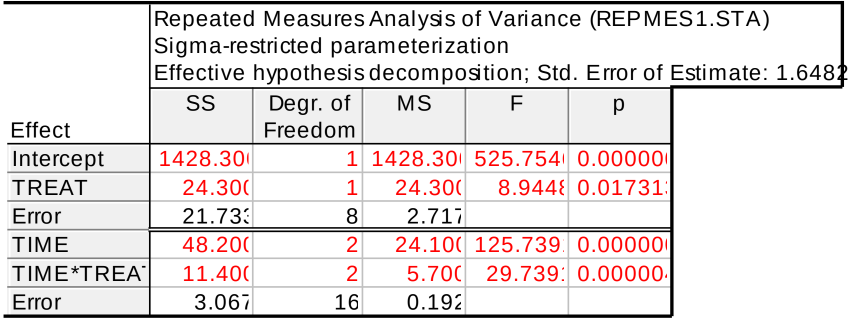
Use **repeated measurement ANOVA**, most important is the interaction between time and treatment. Example above: Control –without nutrient addition, Impact – nutrient addition.

Data are in the form (file repmes1.sta):

TREAT	OBS0	OBS1	OBS2
c	5	6	7
i	6	8	10
c	4	5	5
i	4	6	9
i	5	8	11
c	6	7	8
i	7	8	11
c	5	5	6
i	6	8	10
c	6	7	8

OBS0 – OBS2 are plant heights observed in times 0 to 2. *Statistica > ANOVA > Repeated measurement ANOVA* OR *Statistica > Advanced Linear/nonlinear…> General linear models > Repeated measurement ANOVA.* In the panel, TREAT is the only independent factor, dependent variables are OBS0, OBS1, OBS2. You must specify the *Within effect* **–will be something like TIME** – factor has 3 levels.

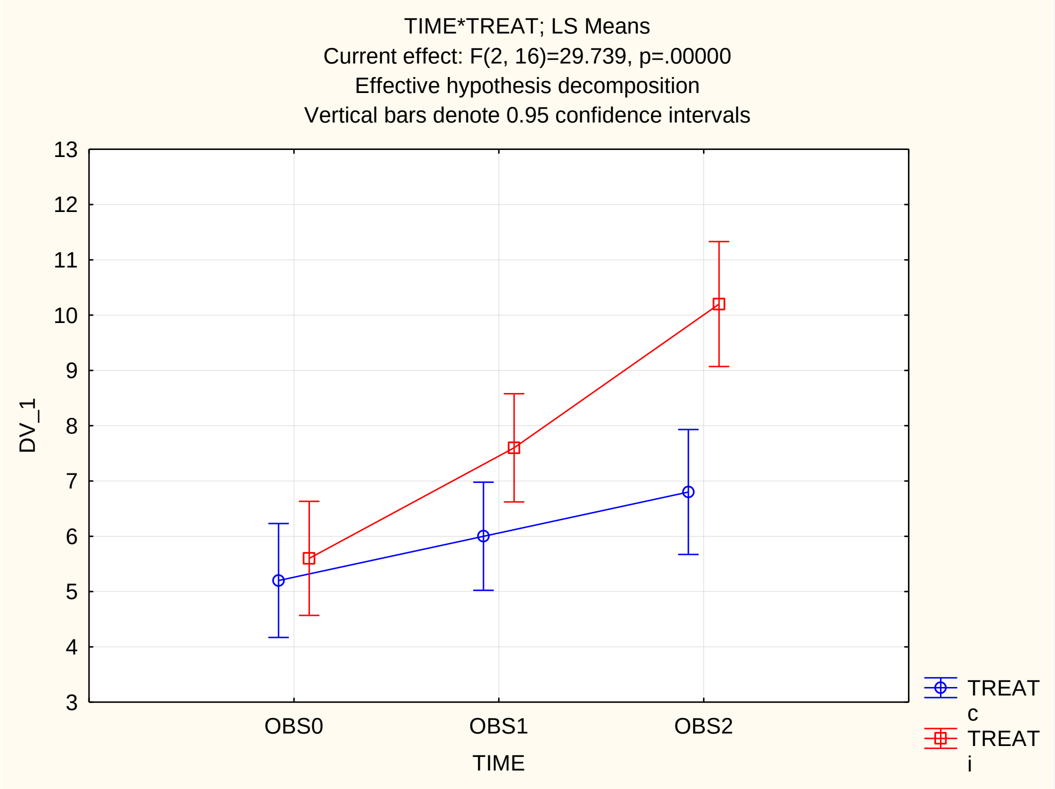
The resulting table is:



Repeated Measures Analysis of Variance (REPMES1.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 1.6482

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	1428.30	1	1428.30	525.754	0.00000
TREAT	24.30	1	24.30	8.944	0.01731
Error	21.73	8	2.717		
TIME	48.20	2	24.10	125.739	0.00000
TIME*TREAT	11.40	2	5.70	29.739	0.00000
Error	3.067	16	0.192		

It tells us that there are differences among treatments (1 - significant TREAT, P<0.05 only), which shows that the average plant size over time differs between control and impact groups, that the plant size changes with time (2 - TIME, of course, plants grow), and most important, there is significant interaction between TIME * TREAT, the development of the control and impact groups differ, P<<0.001). The last term, interaction is the most important one. It can be shown by a graph:



At the beginning, the plants were of roughly same size. With time, the groups start to differ and the difference increases with time.

Regression

Simple linear regression:

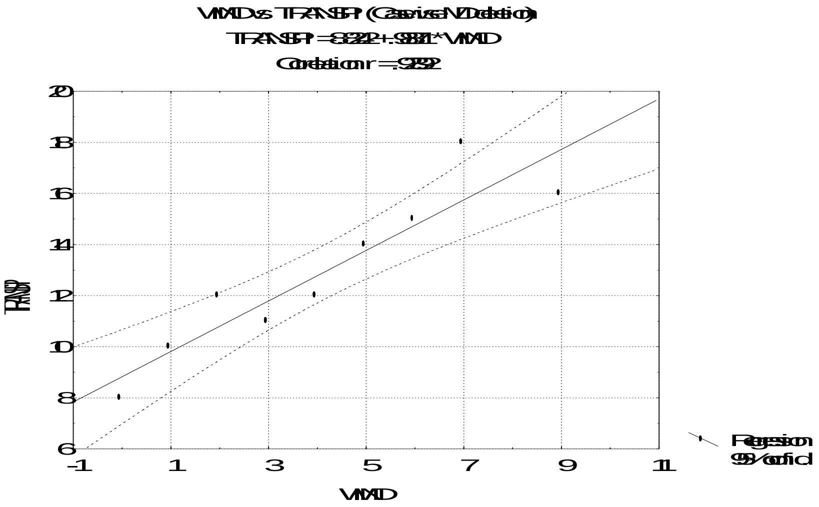
Example 16:Transpiration rate was measured as a response to changing wind velocity. Results:

WIND	TRANSPI
2	12
9	16
5	14
6	15
7	18
3	11
4	12
1	10
0	8

We can reasonably suppose that wind is independent and transpiration depends on wind. For regression, we have to expect that wind is measured without error, whereas transpiration as a response contains some random variability (the random variability is supposed to have normal distribution with zero mean).

Compare: correlation (two random variables) vs. regression.

Simple linear regression can be easily obtained from **Basic statistics, Correlation matrices:** ask for detailed table of results and you will get all parameters of regressions (both X on Y, and Y on X), correlation coefficient, and coefficient determination (R^2). 2D scatterplot gives:



Statistica plots confidence band (i.e. where the regression line lies); the other possibility is tolerance band (i.e. where the observations are)

Regression diagnostics is available in **Multiple regression.**
Transformations: NOTE: by transforming independent variable, you change the shape of the dependence only. By changing dependent variable, you change **both** the shape and distribution characteristics.

Example 17:

Population increases in an exponential way (we expect $N_t=N_0e^{rt}$). The population size increase is described:

Time	population size
0	5
1	7
2	10
3	14
4	19
5	27
6	39
7	50

Estimate r by regression.
Solution:
Log (base e) – transform the population size; you will get
$\ln(N_t) = \ln(N_0) + rt$.
Slope of the regression line $\ln(N_t)$ on t is the estimate of r.

Log-transformation of population size corresponds well to the expected log-normal distribution of population size as a random variable and variability increasing with the size.

Example 18. We expect the relationship between forest patch size (A) and number of vascular species in the patch (S) to be:
$S=c\ A^z$
Data on patch size and number of vascular plant species are (file specarea.sta):

AREA (ha)	Number of Species
2	44
5	60
8	70
12	85
6	57
17	97
23	105
90	129

Estimate the parameters (c, z) of the relationship.

Use log-transformation of both the variables. You will get
Log(S) = log (c) + z log(A), which can be calculated by regression.
Y = a + b. X

Remember:
Correlation coefficient: between –1 and +1, does not depend on units of measurement (describes how well are the variables correlated and the direction of the relationship)

Coefficient of determination (R^2): between 0 and +1. Proportion of variability in dependent variable explained by independent variable(s).
Regression coefficients – between -∞ and +∞, slope of the regression line, depends on the units, in which are the variables measured.

Terminology: independent and dependent variable or predictor and response.

More predictors (independent variables) – Multiple regression

Lets extend the example with transpiration: included in the file are variables TEMPER (temperature), HUMID (humidity) and SUN (sunshine – yes=1, no=0). As in many real cases, the predictors are correlated and probably influence each other.

First, run standard multiple regression with two predictors: WIND and TEMPER: Select WIND and TEMPER as independent and TRANSPI as dependent variable. After OK ask first for **Analysis of Variance.** You will get ANOVA of the complete model:

```
Analysis of Variance; DV: TRANSPI (windtran.sta)
            Sums of              Mean
            Squares    df      Squares       F        p-level
Regress.   73.32337     2     36.66168   39.52372   0.000351
Residual   5.565521     6     0.927587
Total      78.88889
```

The p-level corresponds to the null hypothesis, that there is no effect of any of the explanatory variables.
Then we can ask for **Regression summary:**

```
Regression Summary for Dependent Variable: TRANSPI
R= .96408046 R²= .92945114 Adjusted R²= .90593485
F(2,6)=39.524 p<.00035 Std.Error of estimate: .96311
                       St. Err.               St. Err.
            BETA       of BETA       B         of B        t(6)      p-level
Intercpt                          7.281043   0.836624   8.702885   0.000127
WIND      0.762415   0.125712    0.815877   0.134527   6.064788   0.000912
TEMPER    0.319223   0.125712    0.191351   0.075355   2.539328   0.044121
```

Here, we have for intercept and for particular explanatory variables BETA values (standardized regression coefficients), B (non-standardized regression coefficients) and corresponding t-tests with probability levels. B-values are those from the regression equation:

TRANSPI=7.28+ 0.82 WIND + 0.19 TEMP

Note, that the B values depend on units and consequently can not be used to compare the importance of particular predictors (the standardized coefficients can be used for this purpose). The t-tests test the effect of the corresponding predictor **in the presence of all the other predictors** (it is why they are called *partial* coefficients). Note that neither B nor p-level for WIND are the same as in simple regression.
As in many real cases, the test for intercept is non-sense: this is the test of null hypothesis that the constant (intercet) in the equation is zero, which would mean that there is no transpiration at zero temperature and no wind. (Note, that use of linear regression means some approximation, which will be probably reasonable one in the range of data, but should not be used for extrapolation.)

It might happen, that ANOVA is significant, and none of B values is significant: this happens if the predictors are correlated: it tells us that probably each of them separately explains significant portion of variability, but we are not able to say, which one is important. The predictor is redundant in the presence of the other one.
It also might happen, that one of the predictors is significant, but the total ANOVA is not. This is the case when one good predictor is accompanied by many predictors that do not explain anything.

Building model with stepwise linear regression
Is it better when the computer selects the subset of variables according to some algorithm, or should researcher interact during the procedure?
Danger of statistical fishing!

Do not forget about regression diagnostics (test of assumptions of regression – the residuals are expected to be independent of both predictors, and predicted values).

Violation of assumptions: solution: either transformation, or Generalized linear models.

Nonparametric (Rank) correlation coefficients (Spearman).

Comment [L1]: Old Statistica

ANCOVA – Analysis of Covariance

Use of categorical variables – equivalence of regression and anova: general linear models.
ANCOVA is often used, when we have continuous variable, effect of which should be filtered out first and then we want to compare groups. For example: in the previous example, we want to test the effect of sunshine on transpiration, but we know that transpiration depends also on wind. Use wind as covariate in ANOVA/MANOVA table.

The effect of categorical variable will be obtained by asking for *All effects*, effect of covariate from *Within cell regression.*

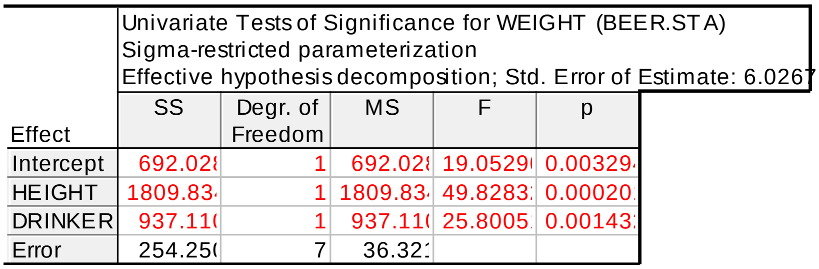
Other (not very botanical) example:

We were interested in the effect of regular beer-drinking (characterized as binary variable, yes-1, no – 0) on persons weight. It is reasonable to use the person’s height as covariable (file **BEER.STA)**:

Weight	Height	Drinker
80	180	0
60	170	0
70	165	1
90	185	0
95	182	1
105	185	1
90	195	0
111	190	1
70	180	0
100	205	0

Go again for *General linear models* and in the panel, as for Analysis of Covariance. Response is Weight, categorical predictor is Drinker, continuous predictor is Height.

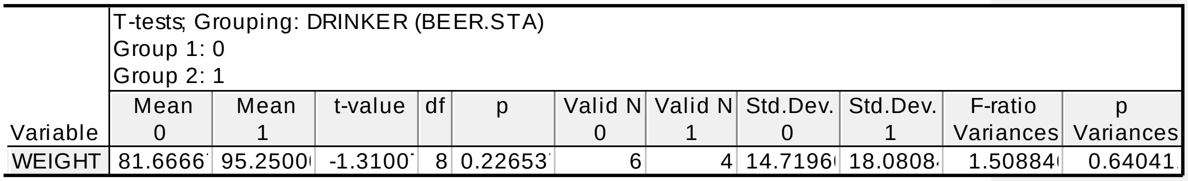
All effects gives you\



Univariate Tests of Significance for WEIGHT (BEER.STA)
Sigma-restricted parameterization
Effective hypothesis decomposition; Std. Error of Estimate: 6.0267

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	692.028	1	692.028	19.0529	0.00329
HEIGHT	1809.834	1	1809.834	49.8283	0.00020
DRINKER	937.11	1	937.11	25.8005	0.00143
Error	254.25	7	36.32		

So, the effect of height is significant (we expected this), and effect of drinking also. Note that t-test, comparing only drinkers with non-drinkers, is not significant



T-tests; Grouping: DRINKER (BEER.STA)
Group 1: 0
Group 2: 1

Variable	Mean 0	Mean 1	t-value	df	p	Valid N 0	Valid N 1	Std.Dev. 0	Std.Dev. 1	F-ratio Variances	p Variances
WEIGHT	81.6666	95.2500	-1.3100	8	0.22653	6	4	14.7196	18.0808	1.50884	0.64041

This is because of large unexplained variability in weights, which is due to differences in person heights.

Recommended reading:

Basic textbooks:

Zar J.H. Biostatistical Analysis. Prentice-Hall, Englewood Cliffs, NJ. (Second edition 1984, 3-rd edition 199*)

Sokal, R.R. & Rohlf, F.J. Biometry. Freeman & Comp. San Francico. [second ed. 1981, 3-rd ed. 1995)

Very useful reading:

Schneider, S.M. & Gurevitch, J. [eds] 1993. Design and Analysis of Ecological Experiments. Chapman & Hall, New York.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 54: 187-211.

For more advanced:

Mead, R. 1988. The desigh of experiments. Statistical principles for practical application. – Cambridge Univ. Press, Cambridge.

Underwood, A.J. 1997. Experiment in Ecology. The logical design and interpretation using analysis of variance. Cambridge Univ. Press.

Hairston, N.G. 1989. Ecological experiments. Purpose, design, and execution. Cambridge Univ. Press, Cambridge.