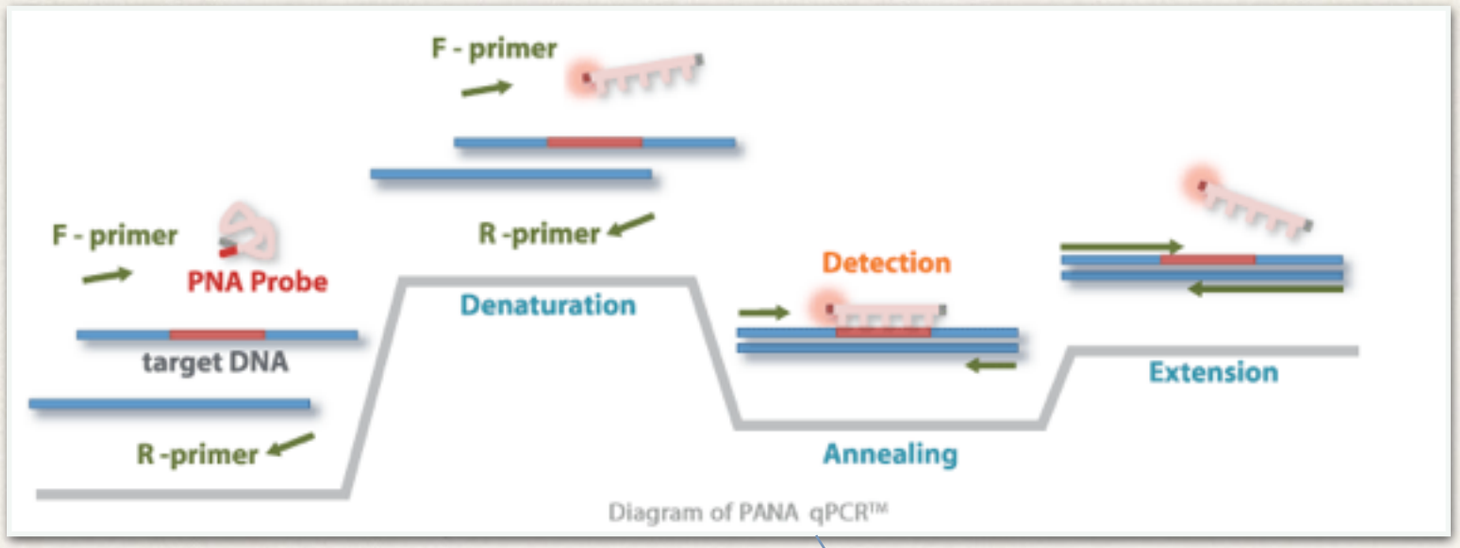
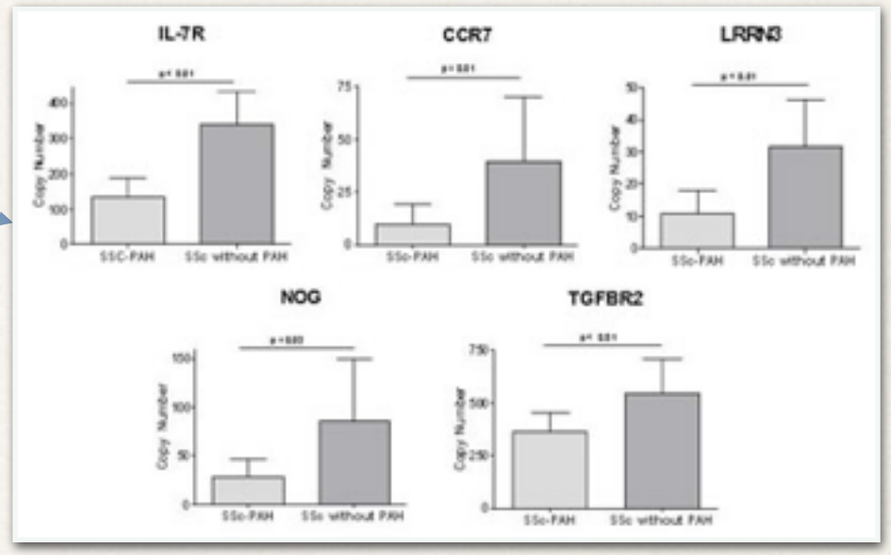


Diagram of PANA qPCR™



Analysing qPCR outcomes

Lecture Analysis of Variance by Dr Maartje Klapwijk

22 October 2014

Personal Background

Since 2009 Insect Ecologist at SLU

Climate Change and other anthropogenic effects on interaction between predator and prey

Mostly field based research

Teaching

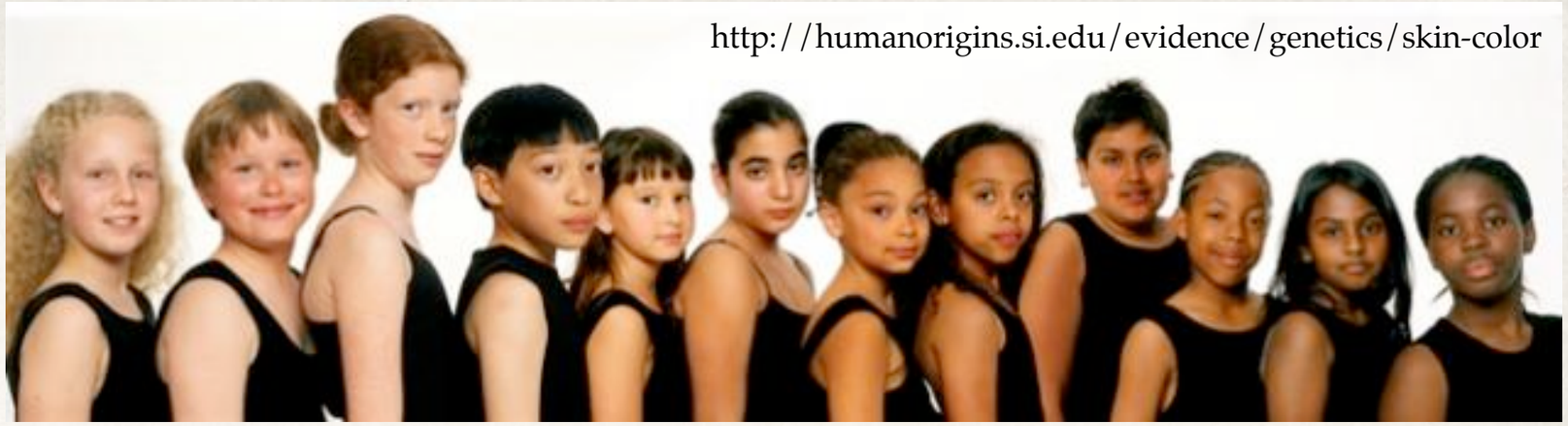
Basic Statistics and ANOVA during PhD

Experimental design and ANOVA at SLU

Program

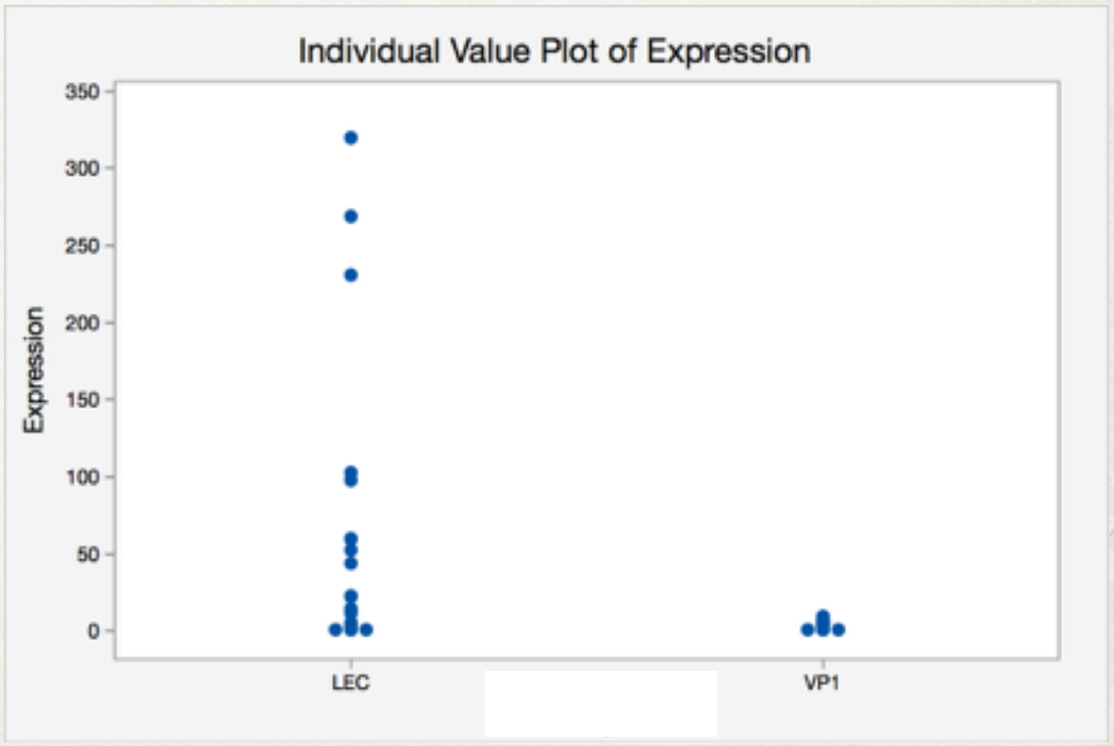
- Descriptive statistics
- t - test
- one-way ANOVA
- General Linear Model

Variability



Spread of data points in the data set

Anova measures and compares variability within and between groups

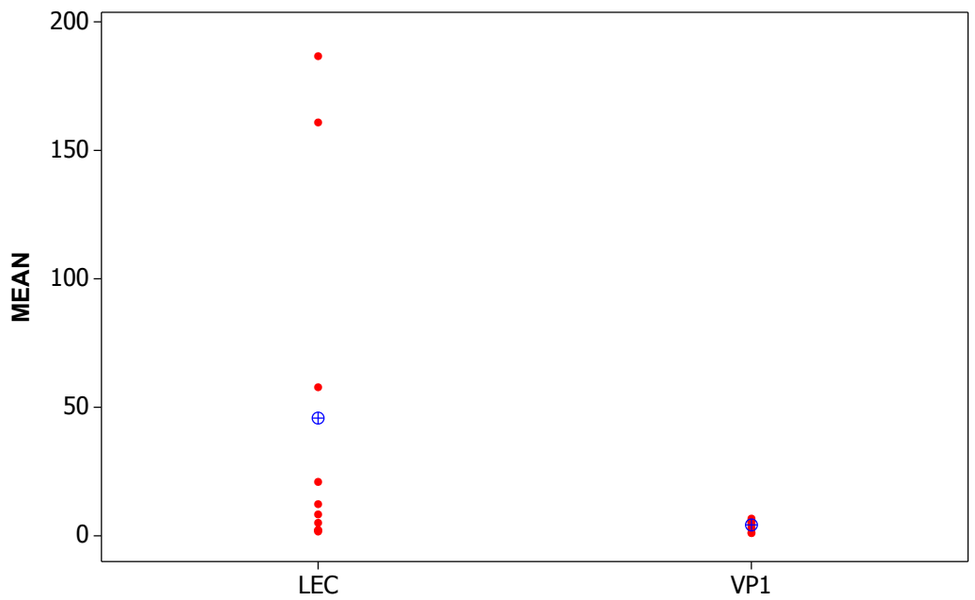


Statistics to describe data

- Mean
- Variance
- Standard deviation
- Standard Error of the Mean
- Confidence intervals

The mean

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 \ldots\ldots + x_n}{n}$$

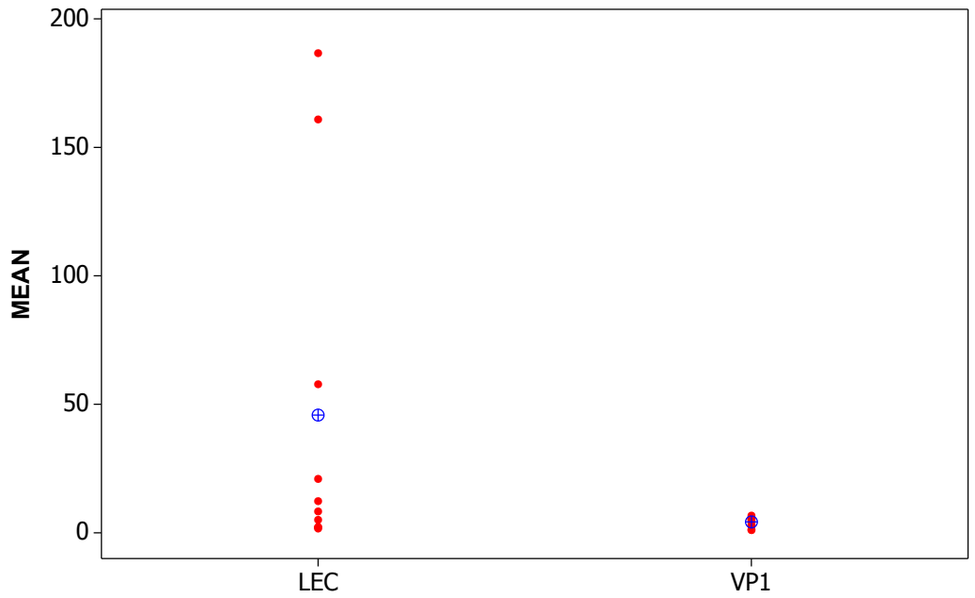


Variance

$$Var(x) = \sigma^2 = \sum [(x - \bar{x})^2]$$

Var (LEC) = 4864.40

Var (VP1) = 4.006



Standard Deviation

$$s = \sigma = \sqrt{\sigma^2}$$

Var (LEC) = 4864.40 ⟶ *s* (LEC) = 69.7

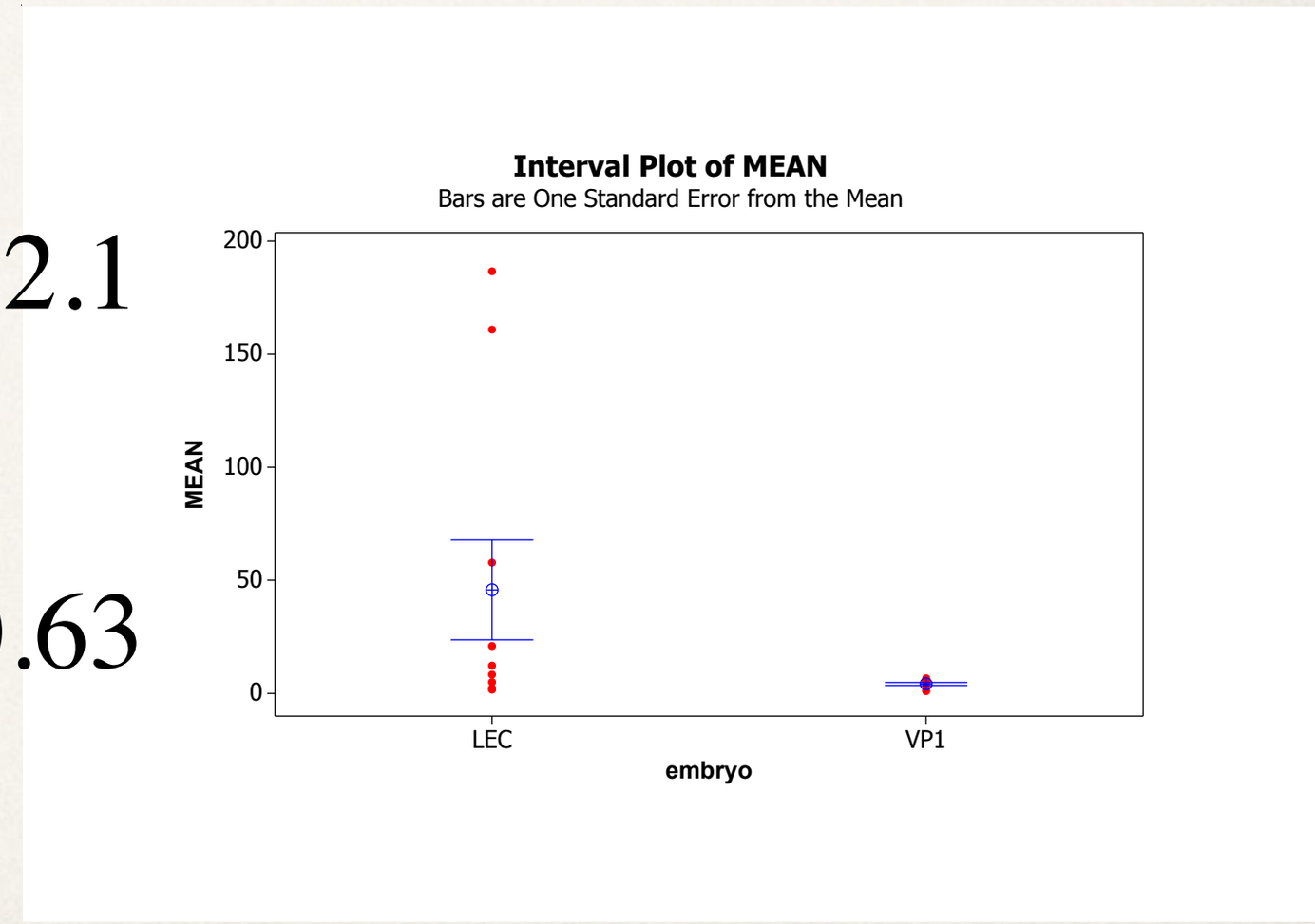
Var (VP1) = 4.006 ⟶ s (VP1) = 2.001

Standard Error of the Mean

$$SEM = \frac{\sigma}{\sqrt{n}}$$

$$SEM(LEC) = \frac{69.7}{\sqrt{10}} = 22.1$$

$$SEM(VP1) = \frac{2.001}{\sqrt{10}} = 0.63$$



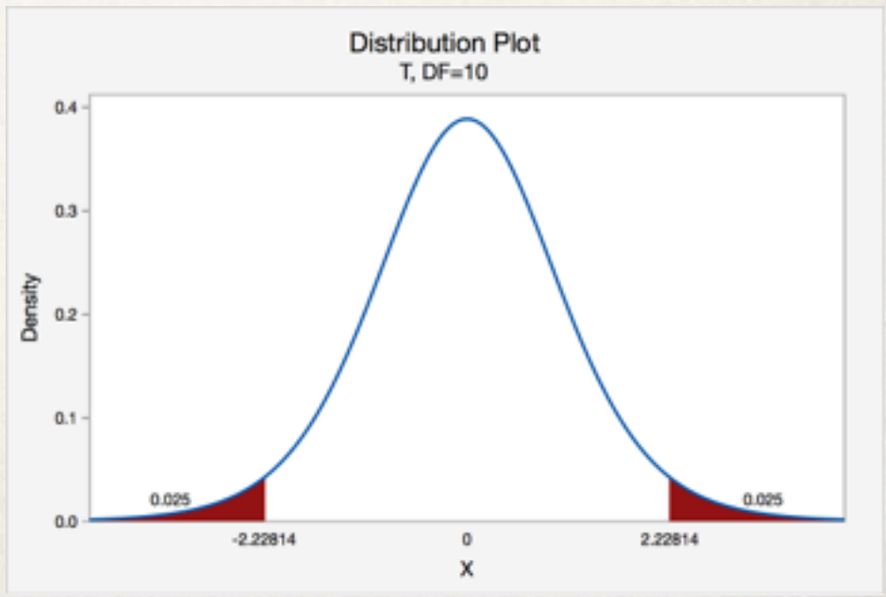
Confidence intervals

$$CI = \bar{x} \pm t_{\alpha/2,d.f} \times SEM$$

Critical t value for $\alpha = 0.05$ for 95% confidence interval

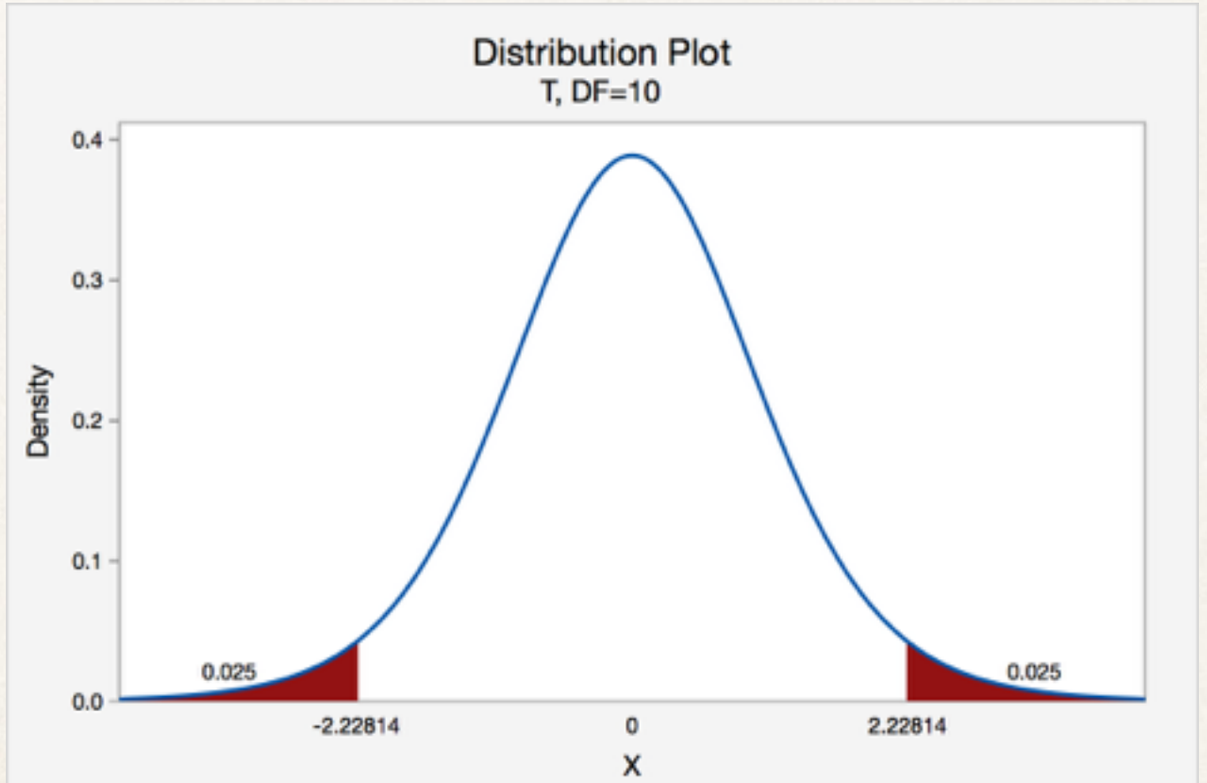
95% CI for LEC = 49.28-22.1*2.2;49.28+22.1*2.2
[0.66; 97.9]

95% CI for VP1=



Probability distributions

What is probability?
The statistical number of outcomes considered divided by the number of all outcomes



Simple 1-sample t-test

Question: How likely is it that a data point comes from a collection of data with a known mean and standard deviation?

Descriptive Statistics

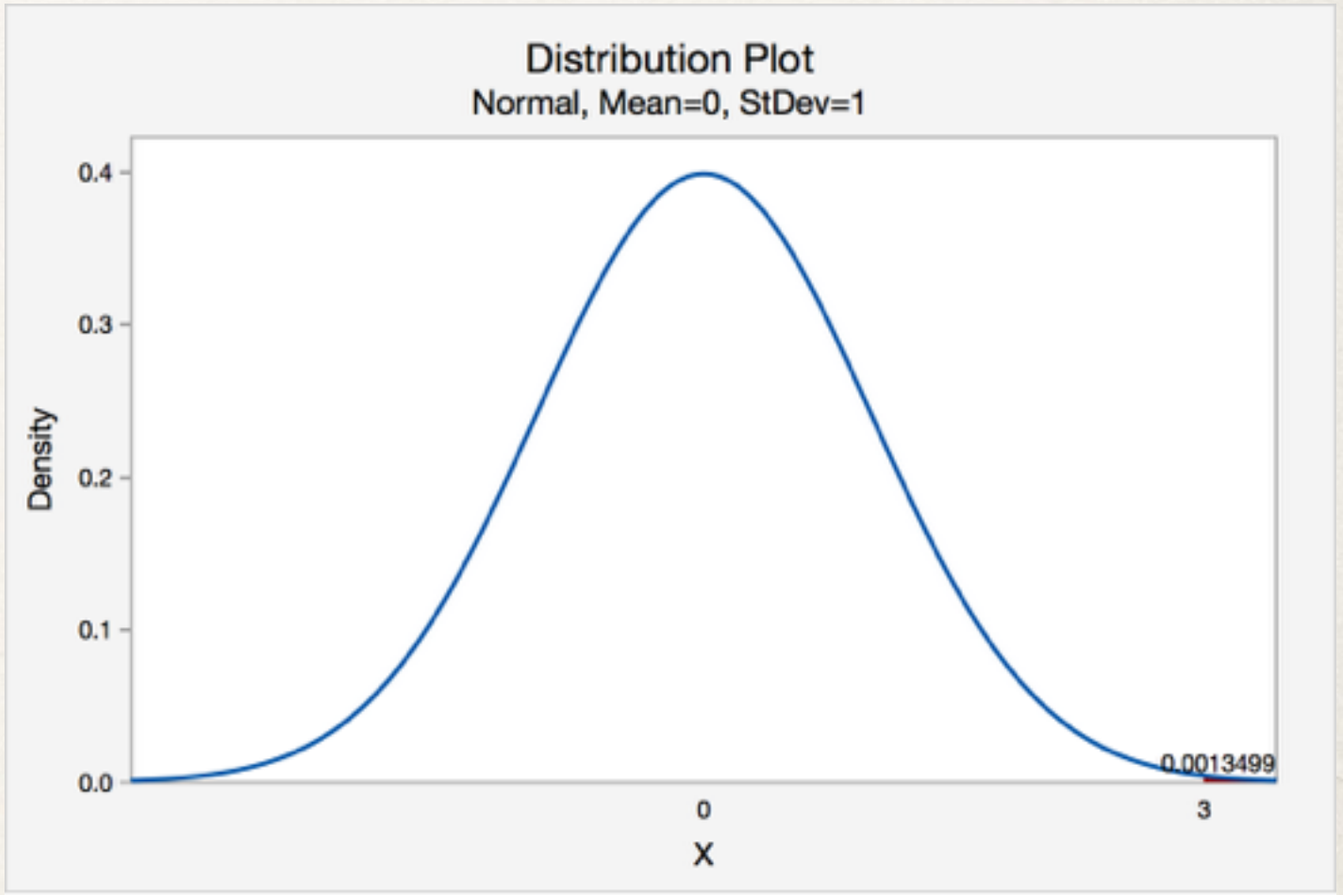
N	Mean	StDev	SE Mean	95% CI for μ
10	0.2426	1.1328	0.3582	(-0.5678, 1.0529)

μ: mean of Normal

Test

Null hypothesis	H_0: μ = 3
Alternative hypothesis	H_1: μ ≠ 3

T-Value	P-Value
-7.70	<0.0001



2-sample t-test

Question: How likely is it that two datasets come from a collection of data with a known mean and standard deviation?

Similar to the one-sample t-tests only it compares two datasets

Example:
Two groups of people have been measured for height.
Question: is the height in group A similar to the height in group B?

The 2-sample t-test compares the mean and variance

When would the test show that there is a difference?

Parametric model assumptions

- Independence of samples and error
- Normality of error
- Equal variances
- additivity of treatments

Non-parametric models

When the data does not comply to the model assumptions one cannot use parametric models

Non-parametric models use other ways of establishing differences

Independence

Data-points are independent if knowing the error of one or a subset of data-points provides no knowledge of the error of any others

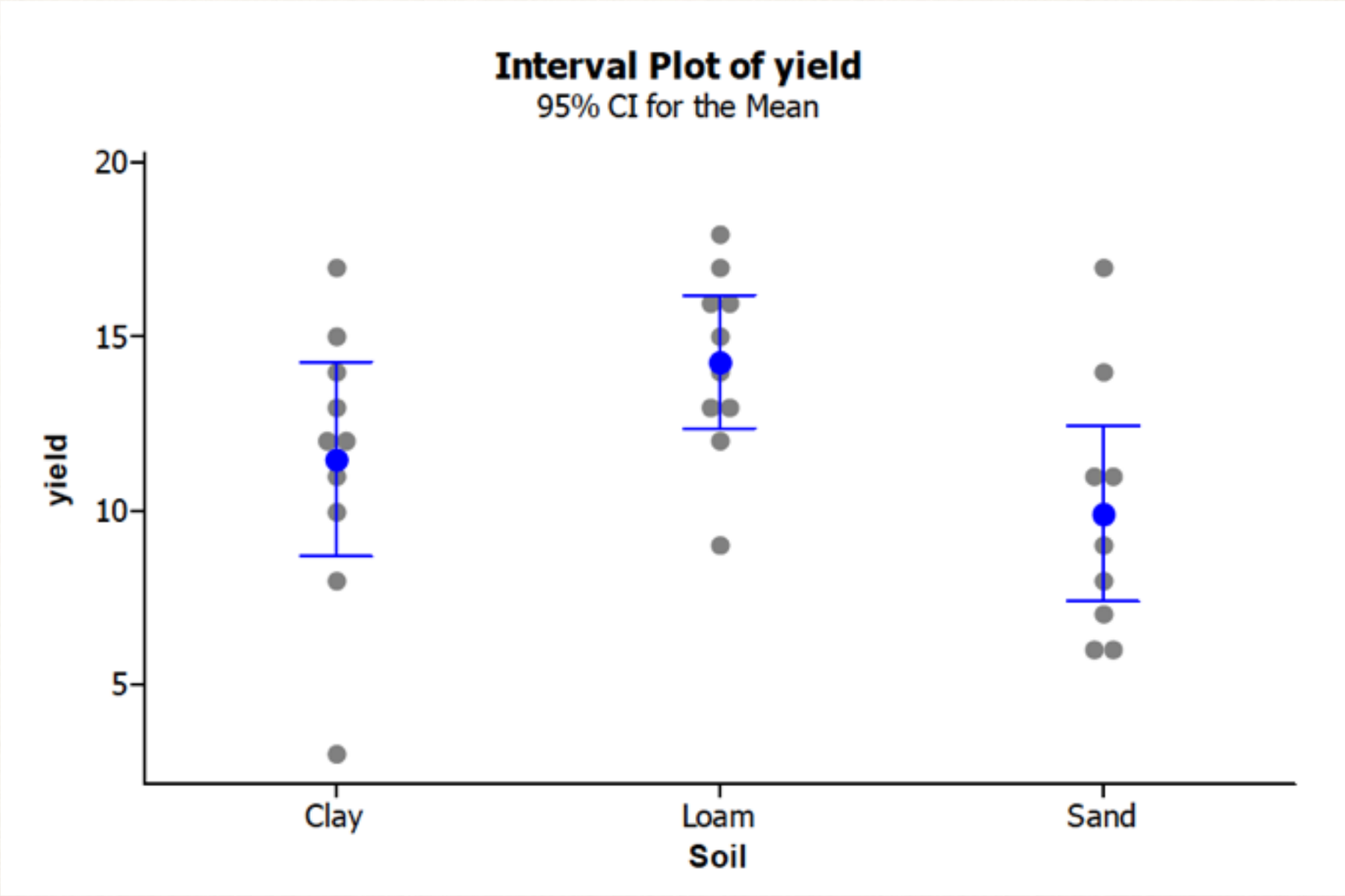
Normality of error

Refers to the shape of the distribution of the residuals around the model

Residuals = observed values - the fitted values

Residuals

= Treatment mean - observed values



Equal variances

This assumption refers to the shape of the distribution of the residuals

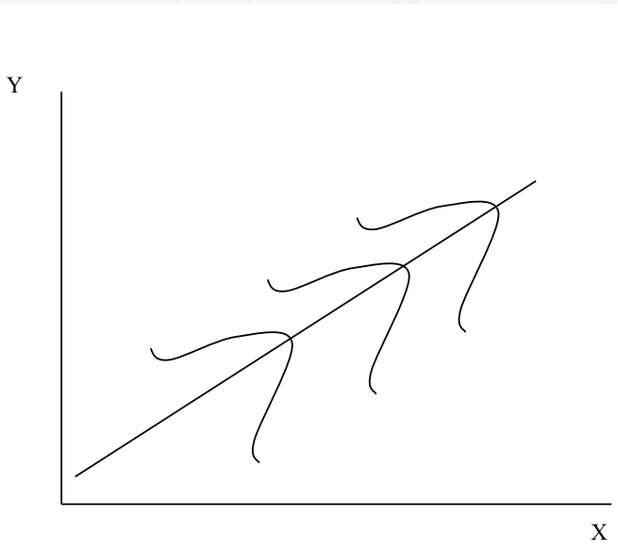


Figure 9.1a

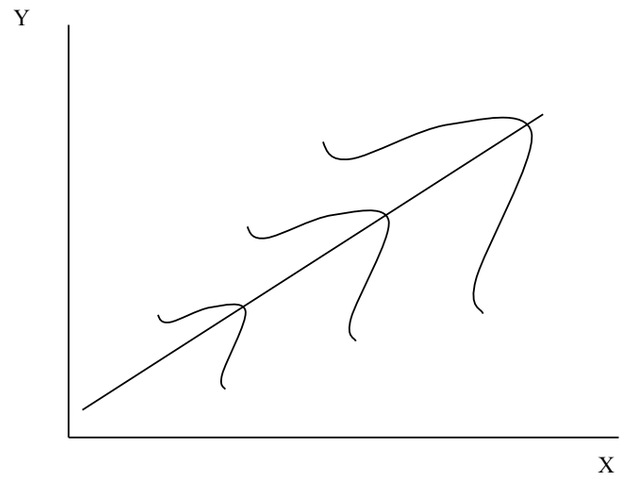


Figure 9.1b

Figure 9.1c

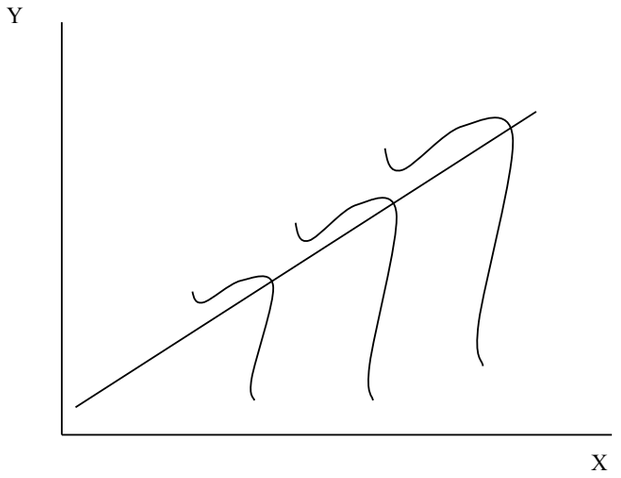


Figure 9.1d

Linearity/Additivity

GLM estimates a linear relationship between the response variables and the explanatory variables

$$y = a + bx + \varepsilon$$

$$y = x1 + x2 + \varepsilon$$

one-way ANOVA

Question: Are my treatment groups different?

H1: They are different

H0: They are not different

Again comparison of the mean and variances

Total Sums of Squares

The total sums of squares represent the total variability within the dataset

$$SST = \sum y^2 - \frac{\left(\sum y\right)^2}{kn}$$

k=number of levels in a treatment
n=number of replicates in each level

Treatment Sums of Squares

The treatment sums of squares are a representation of the variability of the data within the groups of the treatment

$$SSA = \frac{\sum C^2}{n} - \frac{\left(\sum y\right)^2}{kn}$$

C=treatment total

this is the sum of all *n* replicates within a given level

Error Sums of Squares

Error = unexplained variation in the data set

$$SSE = SST - SSA$$

Our data

We have done experiment where we grew a crop on 3 different types of soil

Variable	Soil	Mean	Variance
yield	Clay	11,50	15,39
	Loam	14,300	7,122
	Sand	9,90	12,54

Before the analysis

(http://www.slideshare.net/lssblackbelt/test-for-equal-variances)

Test for Equal Variances: yield versus Soil

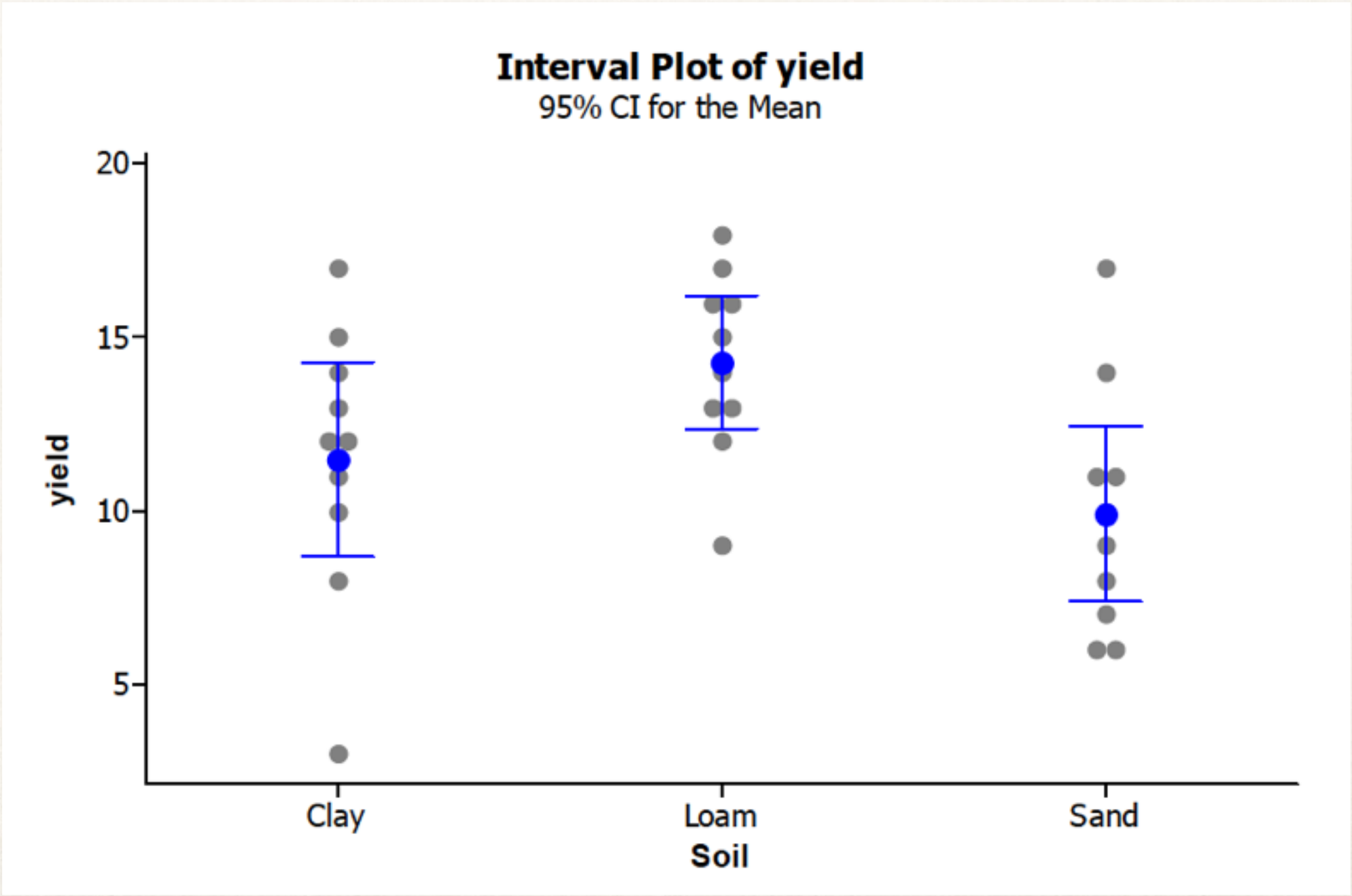
95% Bonferroni confidence intervals for standard deviations

Soil	N	Lower	StDev	Upper
Clay	10	2,49902	3,92287	8,34818
Loam	10	1,70010	2,66875	5,67932
Sand	10	2,25627	3,54181	7,53727

Bartlett's Test (Normal Distribution)
Test statistic = 1,28; p-value = 0,528

Levene's Test (Any Continuous Distribution)
Test statistic = 0,25; p-value = 0,781

Interval plot for yield vs soil



oneway ANOVA

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

Degrees of Freedom = n - 1

$$SSA = \frac{\sum C^2}{n} - \frac{\left(\sum y\right)^2}{kn}$$

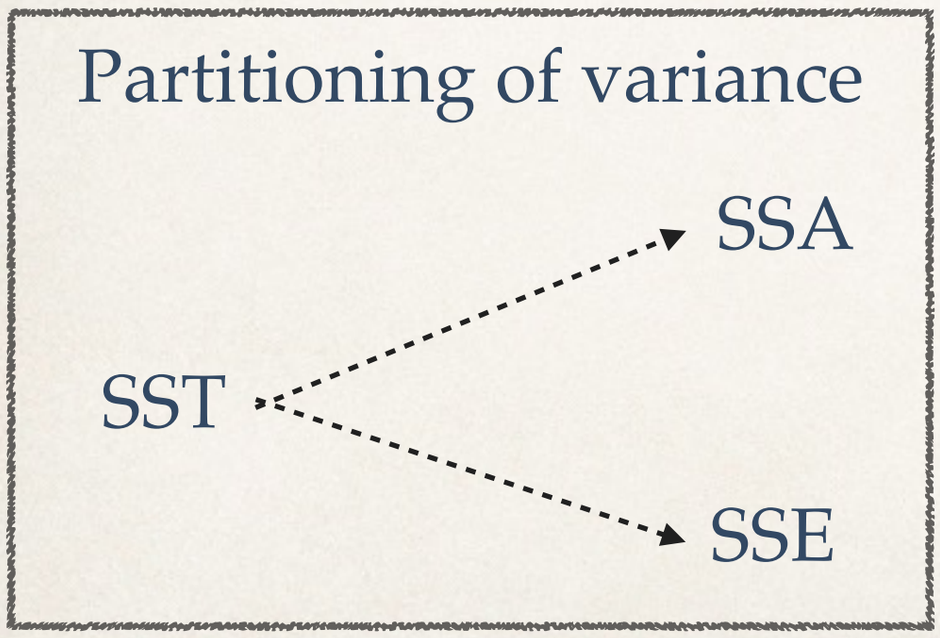
$$SST = \sum y^2 - \frac{\left(\sum y\right)^2}{kn}$$

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

Sums of Squares



SST = sum((yield-mean(yield))^2)
= 414.7

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

Error = (sum((sand-mean(sand))^2))+(sum((loam-mean(loam))^2))
+(sum((clay-mean(clay))^2)) = 315.5

Thus follows:

SST(414.7)-SSE(315.5)=SSA(99.2)

Within group variability

Between group variability

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

Mean sum of squares = weighted SS for degrees of freedom

$MS_{soil} = 99.2/2 = 49.6$

$MS_{error} = 315.5/27 = 11.7$

One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%

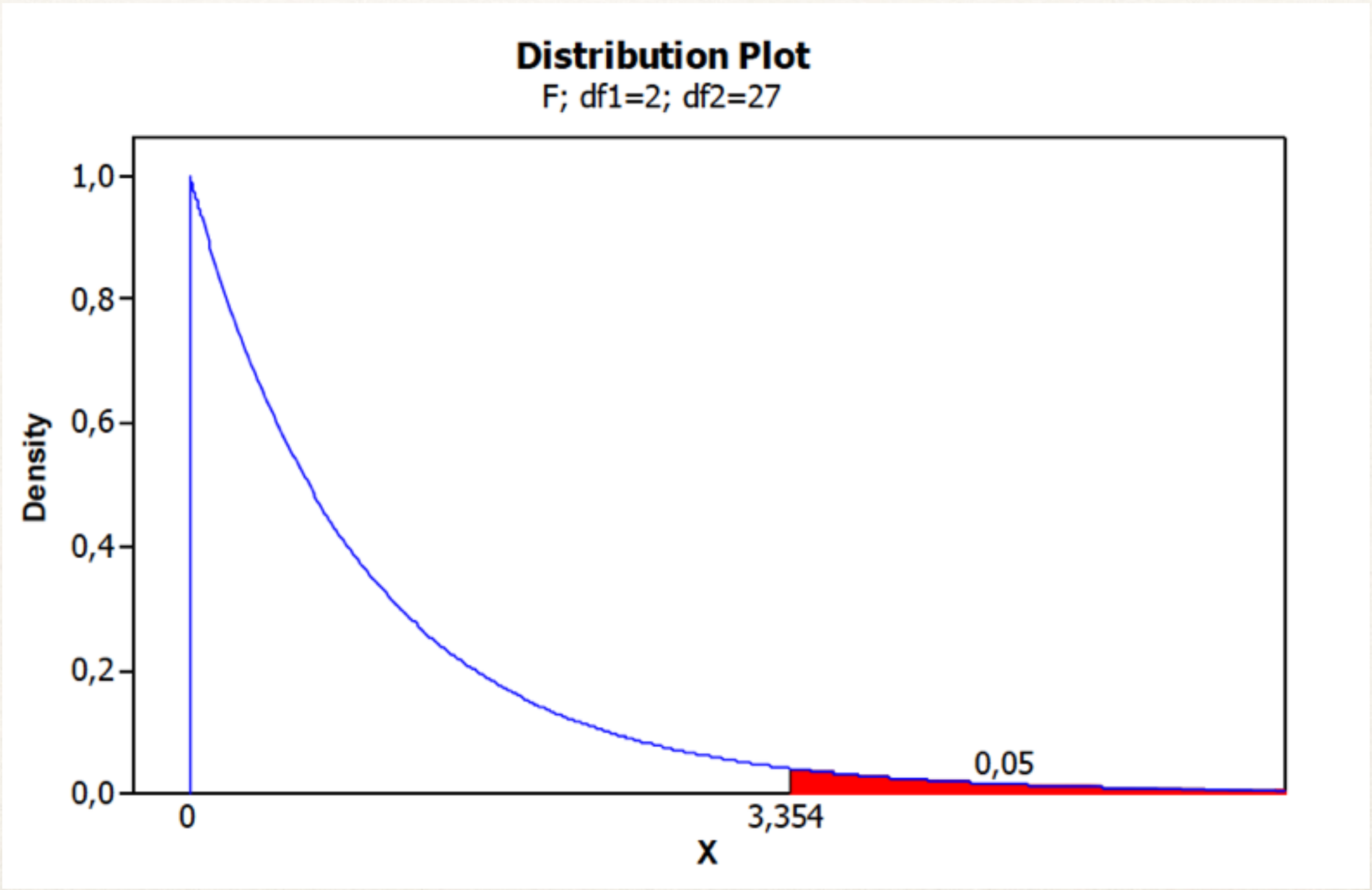
Mean sum of squares = weighted SS for degrees of freedom

$MS_{soil} = 99.2/2 = 49.6$

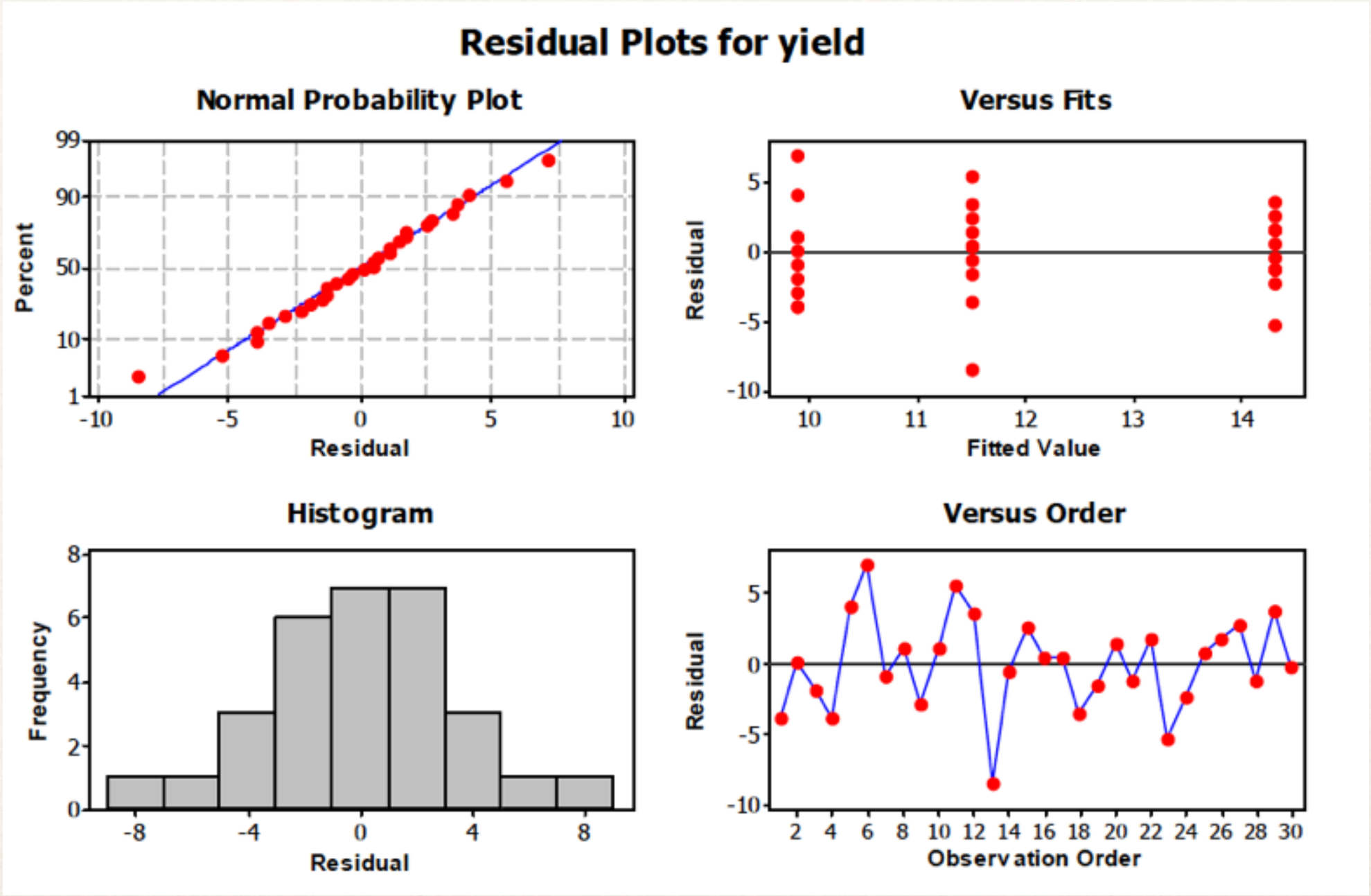
$MS_{error} = 315.5/27 = 11.7$

$MSA/df = F = 49.6/11.7 = 4.24$

F -distribution



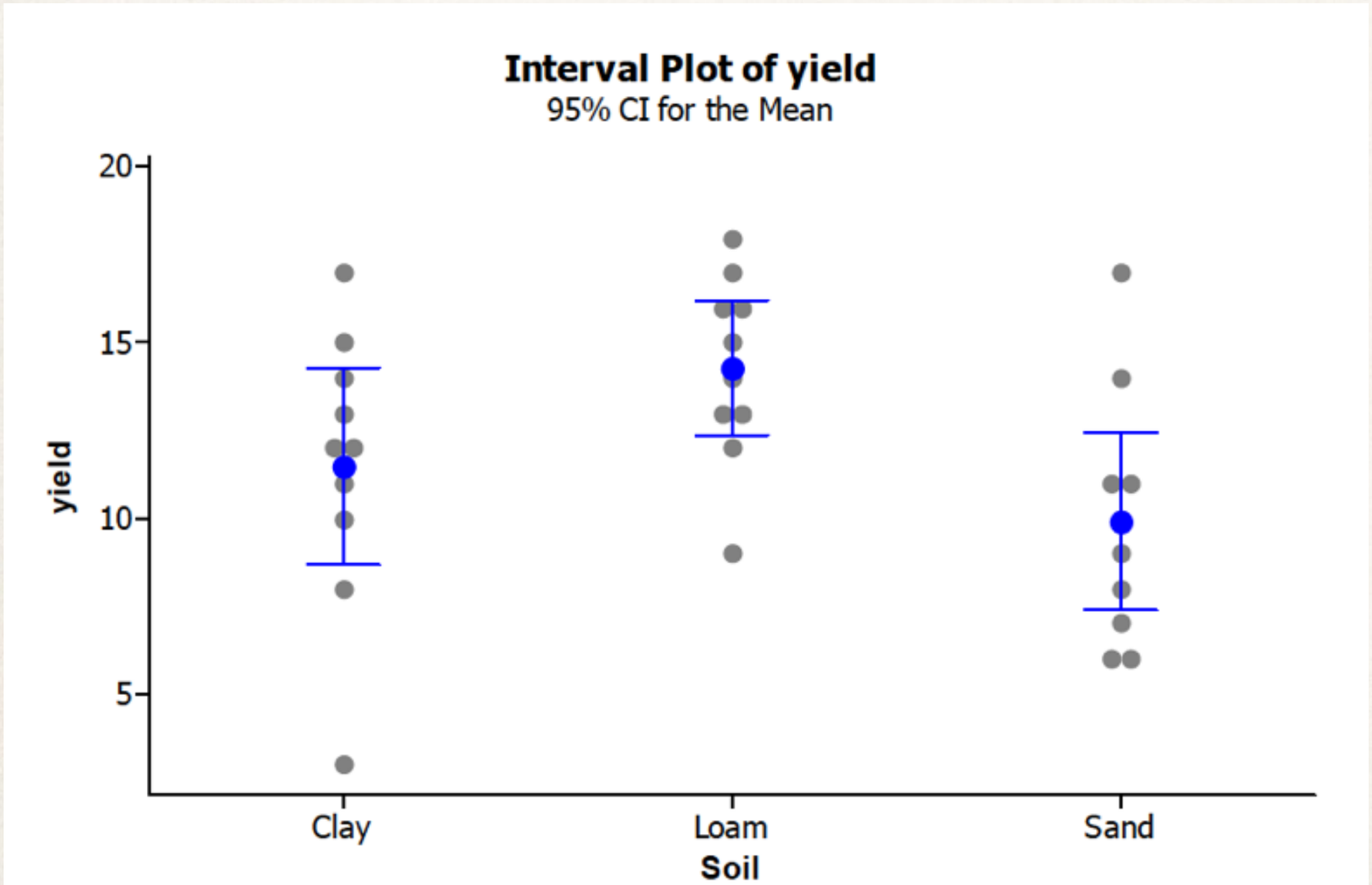
Diagnostics



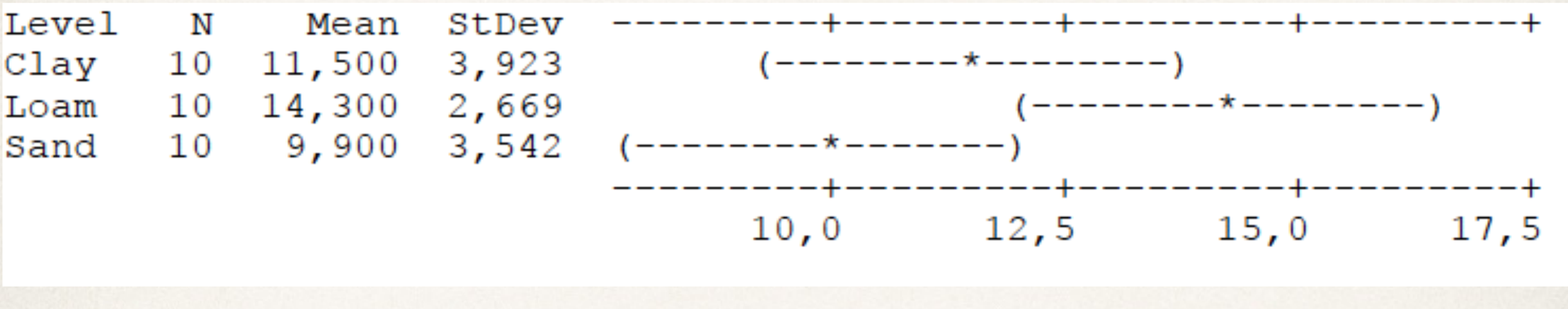
One-way ANOVA: yield versus Soil

Source	DF	SS	MS	F	P
Soil	2	99,2	49,6	4,24	0,025
Error	27	315,5	11,7		
Total	29	414,7			

S = 3,418 R-Sq = 23,92% R-Sq(adj) = 18,29%



Where are the differences?



Tukey test for differences

Grouping Information Using Tukey Method

Soil	N	Mean	Grouping
Loam	**10**	**14,300**	**A**
Clay	10	11,500	A B
Sand	**10**	**9,900**	**B**

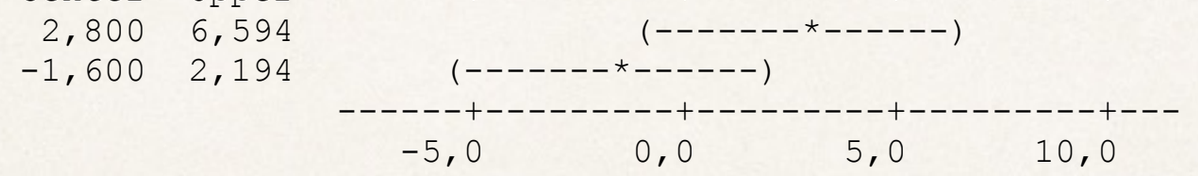
Means that do not share a letter are significantly different.

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Soil

Individual confidence level = 98,04%

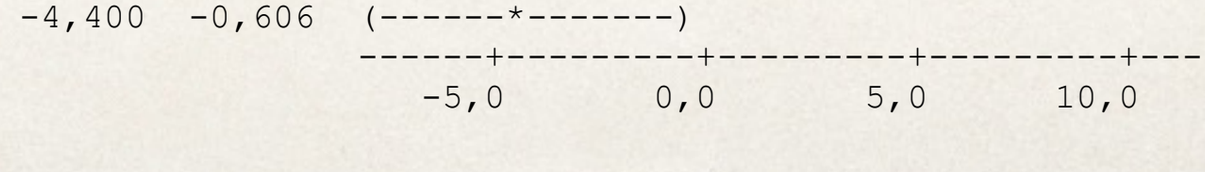
Soil = Clay subtracted from:

Soil	Lower	Center	Upper
Loam	-0,994	2,800	6,594
Sand	-5,394	-1,600	2,194



Soil = Loam subtracted from:

Soil	Lower	Center	Upper
Sand	-8,194	-4,400	-0,606



General Linear Models

Advantages:

- **handles unbalanced designs**
- can include continuous variables (covariates)
- can include interactions
- can include random effects

Assumptions and underlying rational the same as ANOVA

General Linear Model (GLM)

Moving to some qPCR data

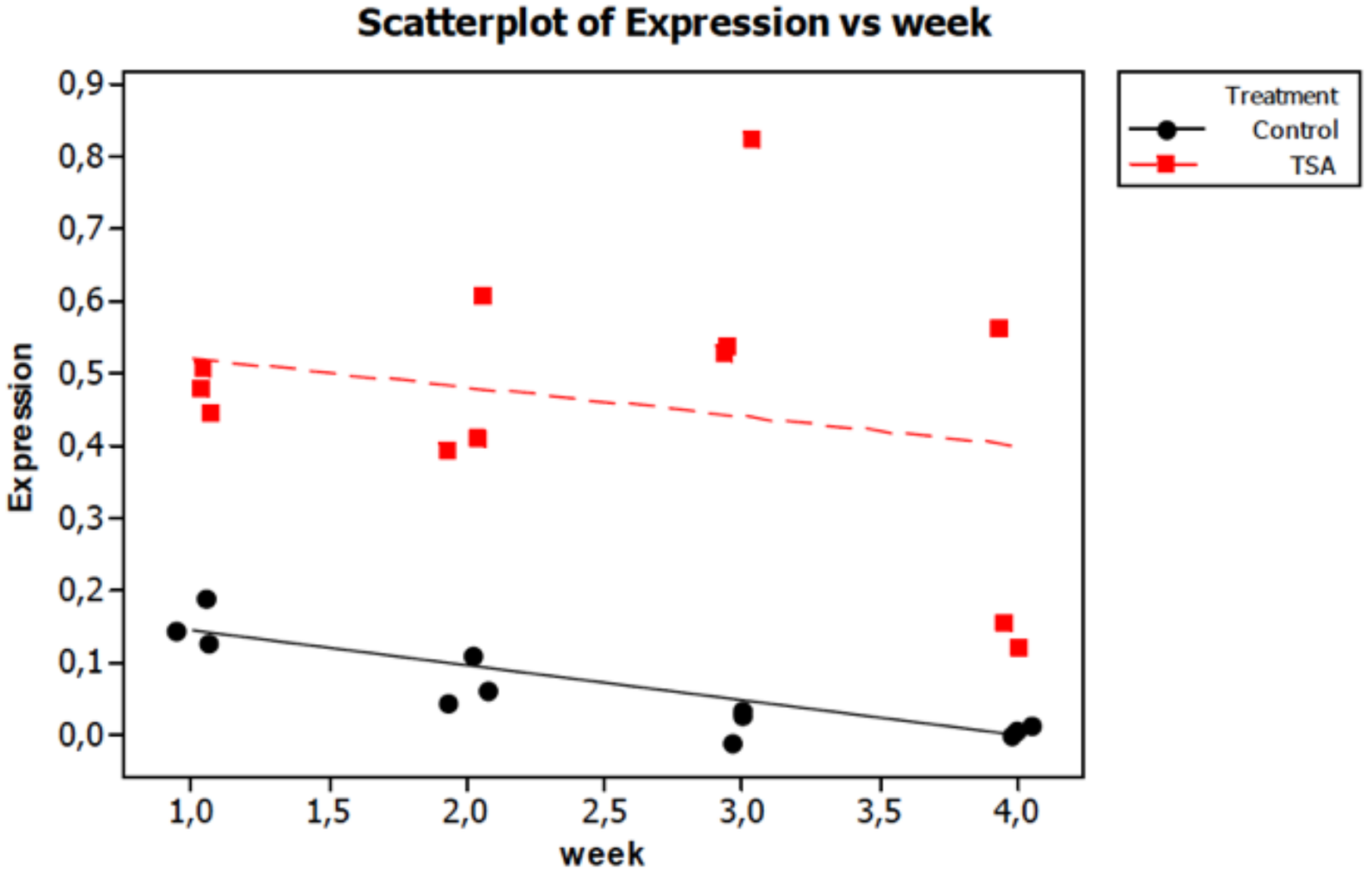
Treatment: 10μM TSA added to the growth medium
Control: growth medium

Expressions are measured every week for 4 weeks

Data collected:
3 technical replicates (constitutes one biological replicate)
3 biological replicates
4 weeks of expressions

resulting in 72 replicates (but only 24 biological replicates)

Experiment VP1 - 10μM TSA



Response:
Expression

Explanatory:
Treatment
- control
- treatment

Week (continuous)

Model formula

Expression= treatment + week+treatment*week+ ε

GLM output - ANOVA table

General Linear Model: Expression versus Contorl

Factor	Type	Levels	Values
Ctrl	fixed	2	Control; Treatment

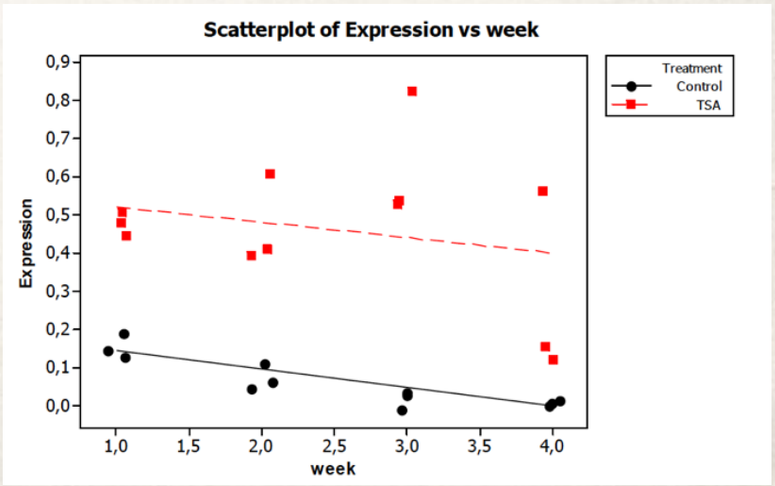
Analysis of Variance for Expression, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Treatment	**1**	**0,90704**	**0,13409**	**0,13409**	**7,83**	**0,011**
week	1	0,05940	0,05940	0,05940	3,47	0,077
Trtm*week	1	0,00061	0,00061	0,00061	0,04	0,852
Error	20	0,34245	0,34245	0,01712		
Total	23	1,30949				

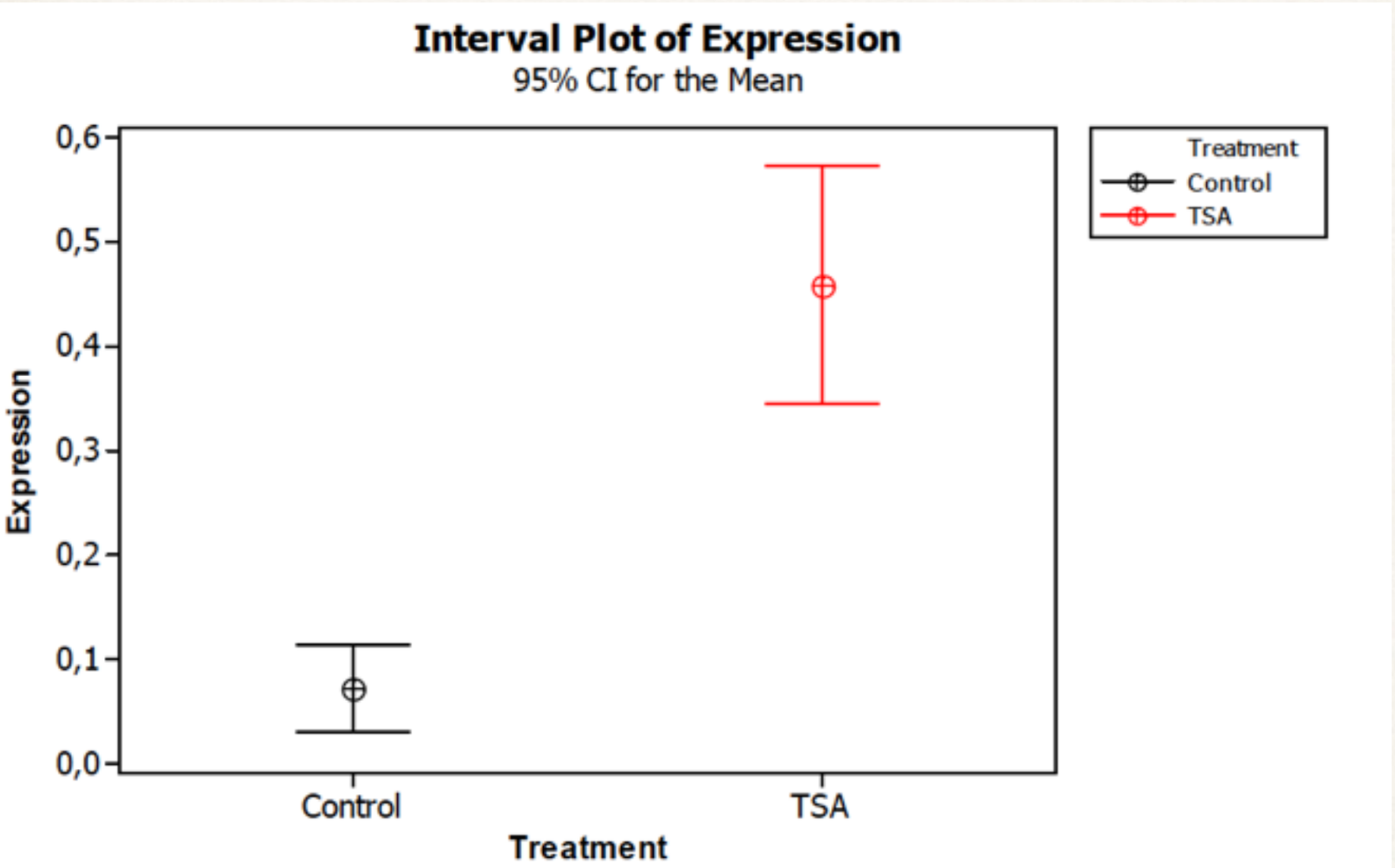
S = 0,130852 R-Sq = 73,85% R-Sq(adj) = 69,93%

GLM output - coefficients

Term	Coef	SE Coef	T	P
Constant	0,37669	0,06543	5,76	0,000
Treatment				
Control	-0,18309	0,06543	-2,80	0,011
week	-0,04450	0,02389	-1,86	0,077
week*Trtm				
Control	-0,00453	0,02389	-0,19	0,852



Results:

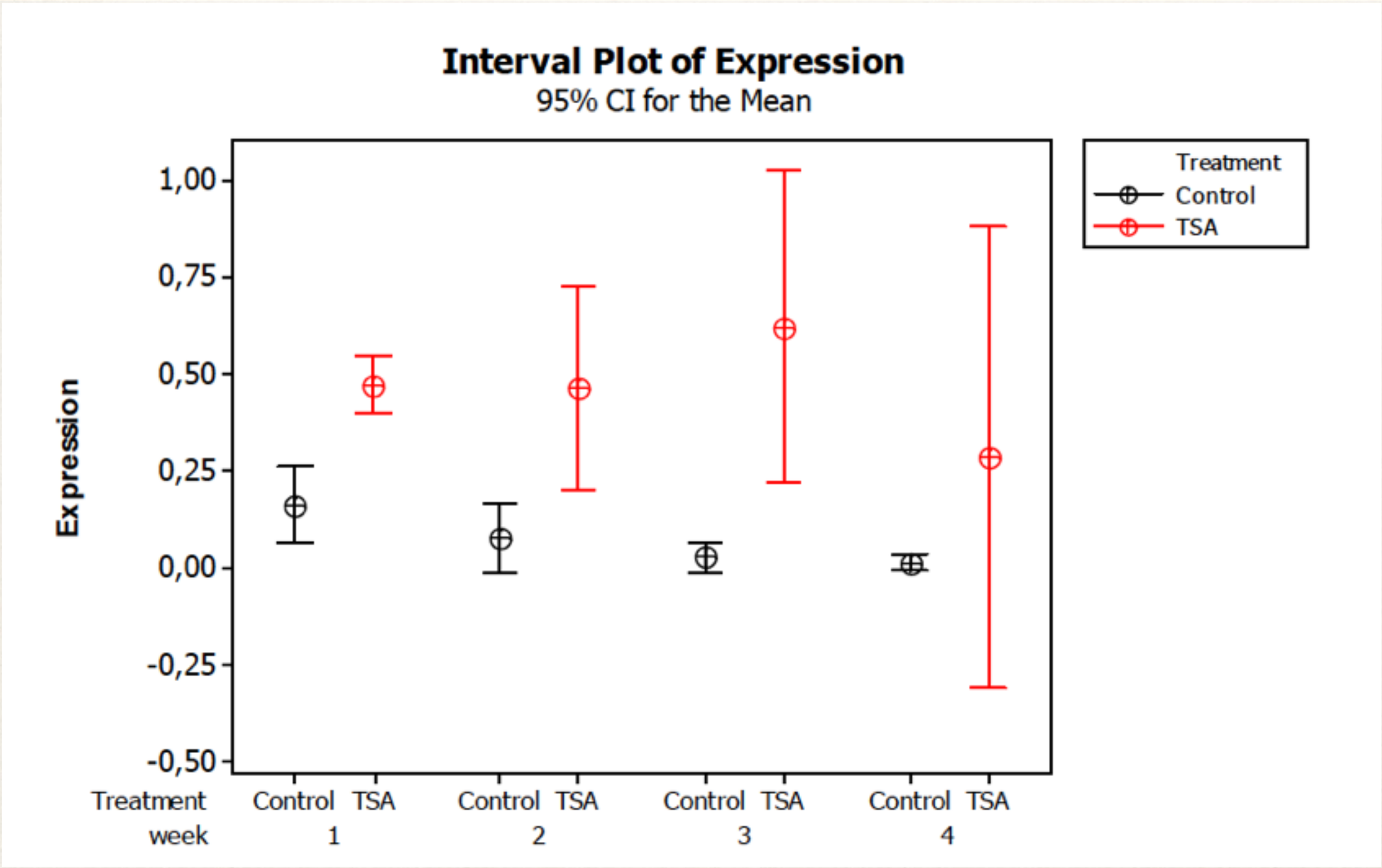


Only treatment influences gene expression

Mean of the control is -0.18 lower than the overall mean.

Mean of treatment is 0.18 higher than the overall mean

GLM - week = category



GLM - week = category

General Linear Model: Expression versus Ctrl; week

Factor	Type	Levels	Values
Ctrl	fixed	2	Control; Treatment
week	fixed	4	1; 2; 3; 4

Analysis of Variance for Expression, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Ctrl	1	0,90704	0,90704	0,90704	73,98	0,000
week	3	0,11467	0,11467	0,03822	3,12	0,056
Ctrl*week	3	0,09161	0,09161	0,03054	2,49	0,097
Error	16	0,19618	0,19618	0,01226		
Total	23	1,30949				

GLM - week = category

Term		Coef	SE Coef	T	P
Constant		0,26545	0,02260	11,74	0,000
Ctrl					
Control		-0,19440	0,02260	-8,60	0,000
week					
1		0,05113	0,03915	1,31	0,210
2		0,00565	0,03915	0,14	0,887
3		0,05779	0,03915	1,48	0,159
Ctrl*week					
Control	1	0,04049	0,03915	1,03	0,316
Control	2	0,00088	0,03915	0,02	0,982
Control	3	-0,10060	0,03915	-2,57	0,021

Some more general points

- Handling technical replicates
- Sample size estimation
- Ethics of data handling