One-way ANOVA

(STATS546.2: Applied analysis of variance (work in progress))

Anonymous

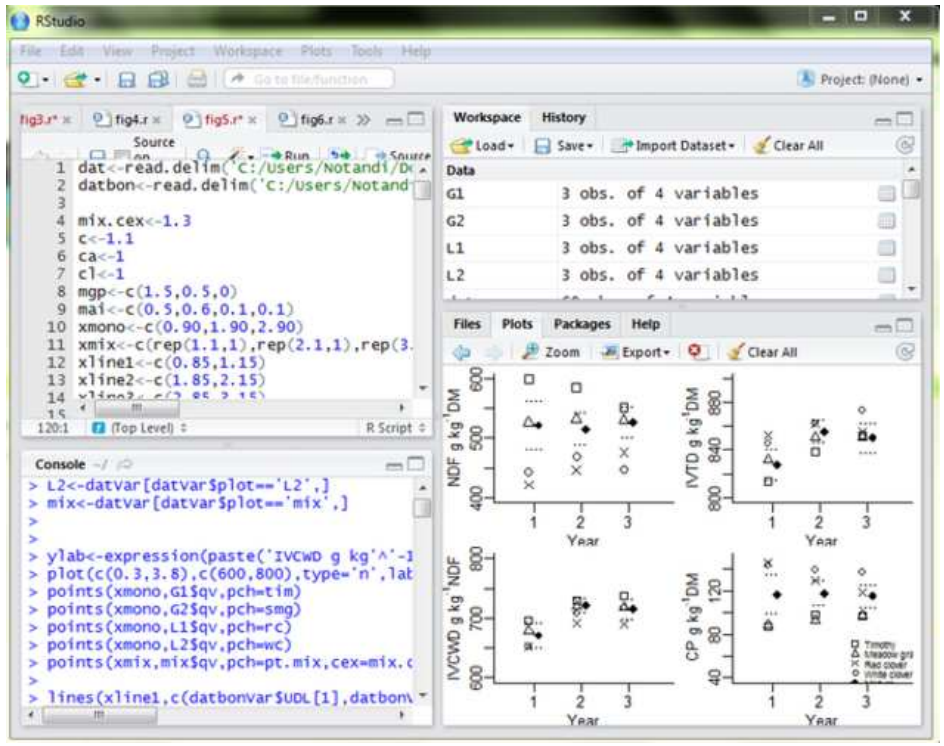
June 28, 2013

Introduction to ANOVA

1. One-way ANOVA and assumptions
2. Multiple comparisons and two way ANOVA
3. Multi-factor ANOVA and interactions
4. Random and mixed effect models
5. ANCOVA and nested designs
6. MANOVA and repeated measures
7. Examples from experiments and observations

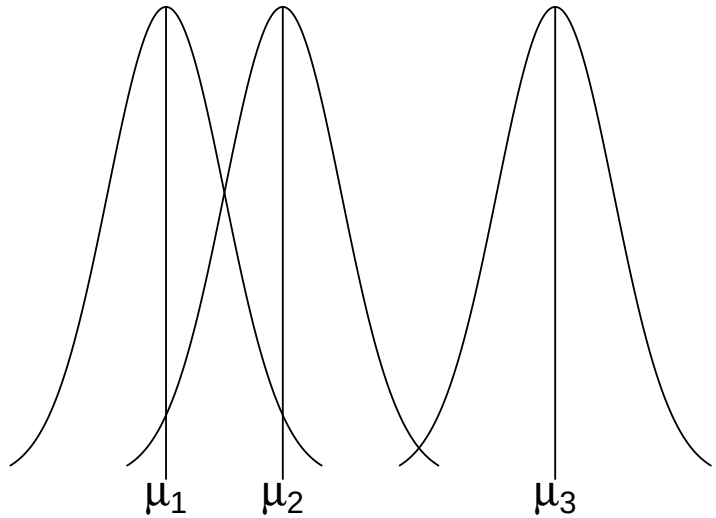
Statistical softwares

R
SAS
SPSS



T-test vs ANOVA

We use t-test if we want to test if two populations have the same mean but we use ANOVA if we want to test if three or more populations have the same mean.



Terminology

Response variable is the thing we measure, e.g. length, temperature, harvest etc.

Factor is the effect we are investigating, e.g. diet, drugs, fertilizer etc.

Levels of the factor are e.g. drug a, drug b, drug c.

Subject is the experimental unit, e.g animal, plot, human, plant etc.

Where does the name Analysis of variance come from

When we do analysis of variance we are partitioning the variation in the response variable into what is explained by the factors and what is unexplained.

In ANOVA we can examine the relative contribution of factors to the total variation in the response variable.

We can also test if there is a difference in the factor level means.

Rat diet experiment

3 diet groups:

1. Fat
2. Carbohydrates
3. Protein

Fat	Carbo	Protein
53	62	60
49	66	67
59	61	58
51	58	56
56	66	63
		64

The one-way model

The one-way ANOVA model is on the form:

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{1}$$

or

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{2}$$

Fitting the ANOVA model

The parameters are estimated with the method of least squares

$$S = \sum_i \sum_j (y_{ij} - \mu_i)^2$$

$$\hat{\mu}_i = \bar{y}_{i\cdot}.$$

Example Rat diet experiment

The parameters are estimated by calculating the mean in each group.

$\hat{\mu}_1 = \bar{y}_{1\cdot} = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \frac{53+49+59+51+56}{5} = 53.6$

$\hat{\mu}_2 = \bar{y}_{2\cdot} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} = \frac{62+66+61+58+66}{5} = 62.6$

$\hat{\mu}_3 = \bar{y}_{3\cdot} = \frac{1}{n_3} \sum_{j=1}^{n_3} y_{3j} = \frac{60+67+58+56+63+64}{6} = 61.3$

Hypothesis

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_i = 0$$

$$H_1 : \text{Not all } \alpha_i = 0$$

The ANOVA table

Table: ANOVA table

Source	df	SS	MS	F
Factor A	$g-1$	$SSA = \sum n_i(\bar{y}_i. - \bar{y}..)^2$	$MSA = SSA/(g-1)$	$F =$
Error	$N-g$	$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_i.)^2$	$MSE = SSE/(N-g)$	
Total	$N-1$	$SSTOT = \sum_i \sum_j (y_{ij} - \bar{y}..)^2$		

Sums of squares

Calculation of sum of square in the rat diet experiment

$$SSTOT = \sum_i \sum_j (y_{ij} - \bar{y}..)^2 = 435.4$$
$$SSA = \sum_i \sum_j (y_{ij} - \bar{y}_i.)^2 = 241.7$$
$$SSE = \sum n_i (\bar{y}_i. - \bar{y}_.)^2 = 193.7$$

R squared

R-squared is a measure of the proportion of the total variation in the response variable that is explained by the factors.

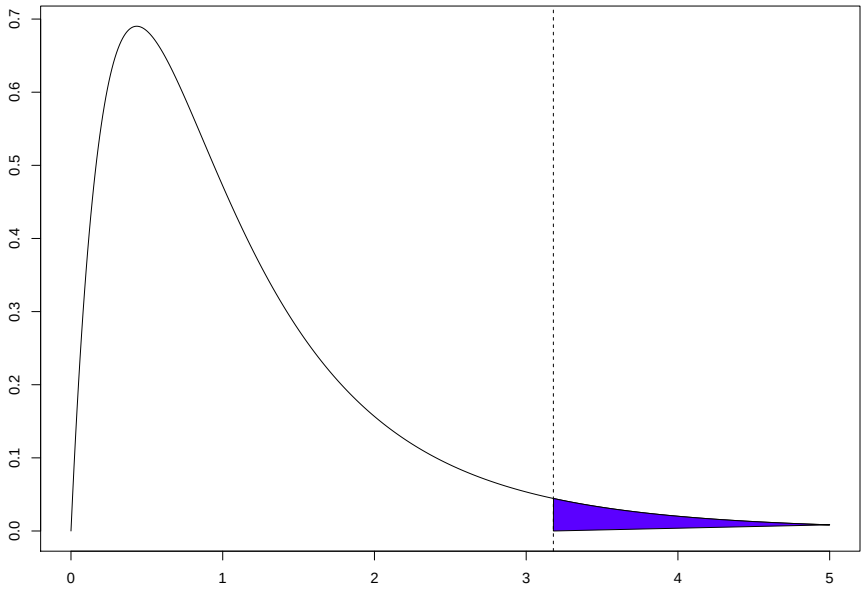
$$R^2 = \frac{SSA}{SSTOT}$$

F-test

F-test to test if there is a difference between treatment group means.
If the F-value from the ANOVA table is larger than the quantile from the F-distrubution with df_1 and df_2 degrees of freedom.

$$F > F_{df_1, df_2, 1-\alpha}$$

The null hypothesis is rejected and it is concluded that there is a significant difference between the treatment group means.



The F-table

$df_2 \backslash df_1$	1	2	3	4	5	6	7	8
9								
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147
7	5.591	4.737	4.347	4.12	3.972	3.866	3.787	3.726
8	5.318	4.459	4.066	3.838	3.687	3.581	3.5	3.438
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.23
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948
12	4.747	3.885	3.49	3.259	3.106	2.996	2.913	2.849
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767
14	4.6	3.739	3.344	3.112	2.958	2.848	2.764	2.699
15	4.543	3.682	3.287	3.056	2.901	2.79	2.707	2.641
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591
17	4.451	3.592	3.197	2.965	2.81	2.699	2.614	2.548
18	4.414	3.555	3.16	2.928	2.773	2.661	2.577	2.51
19	4.381	3.522	3.127	2.895	2.74	2.628	2.544	2.477
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447

Example - Iris

Is there a difference in petal length and width between iris species?

Output from R

```
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  437.1  218.55    1180 <2e-16 ***
Residuals   147   27.2    0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

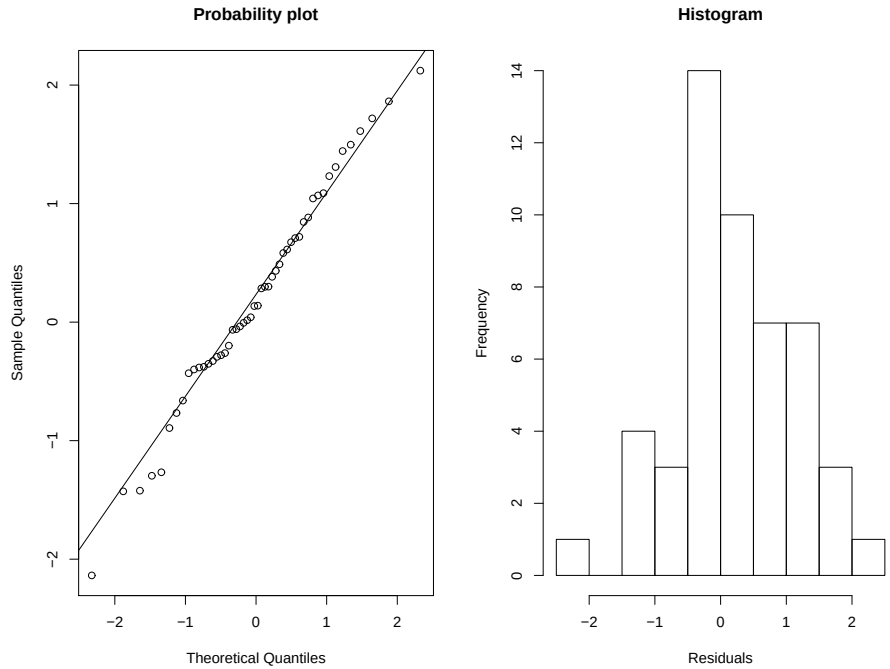
Assumptions

- Random sampling
- Independent measures or observations
- Normal distribution
- Equal variances

Normality

To see if the normal assumption is satisfied the residuals can be plotted to graphs, either a propability plot or a histogram.

In the probability plot the residuals should follow a straight line and the histogram should be in a bell-shape.



Test for normality

There are few test available to test for normality

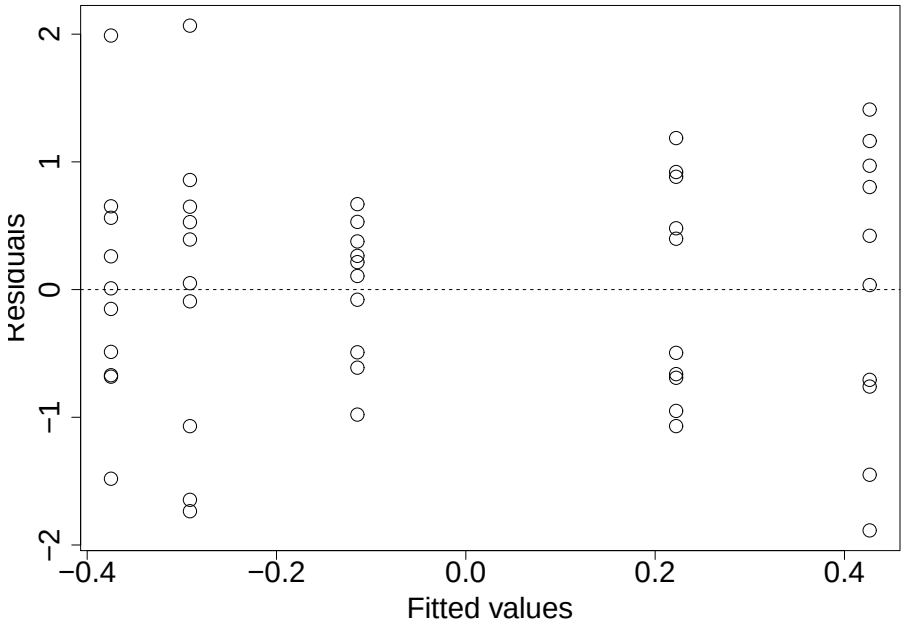
Anderson-Darling test

Kolmogorov–Smirnov test

Shapiro–Wilk test

Constant variance

The spread of the residual can be seen by plotting the residuals by the fitted values.



Test for constant variance

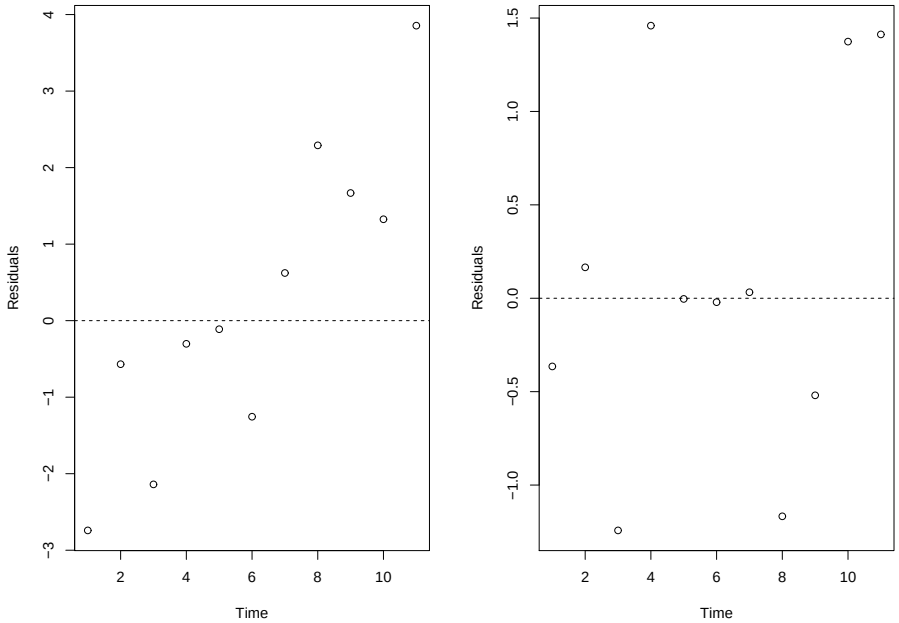
Hartley's test

Bartlett's test

Brown-Forsythe test

Independence

The design of the experiment has to be in that way that the measurements are independent, i.e. one observation has no effect on another one. In some cases there is a time factor in the experiment and can it be used to check if the observations are independent bu looking at a residual sequence plot.



Test for independence

Durbin–Watson statistic

What happens when assumptions do not hold

The parameters estimates are still valid but confidence intervals and significance test may not be.

ANOVA is robust to the departure of normality but is more sensitive to non-constant variance, especially when there are unequal sample sizes.

If the observations are not independent, as can be when there are repeated observation on the same subject, the variance can be underestimated and type I error rate gets inflated.

Transformations

When the assumptions do not hold it can be useful to transform the response variable.

$log(y)$

$\sqrt{(y)}$

$1/y$

Example - electrofishing 1

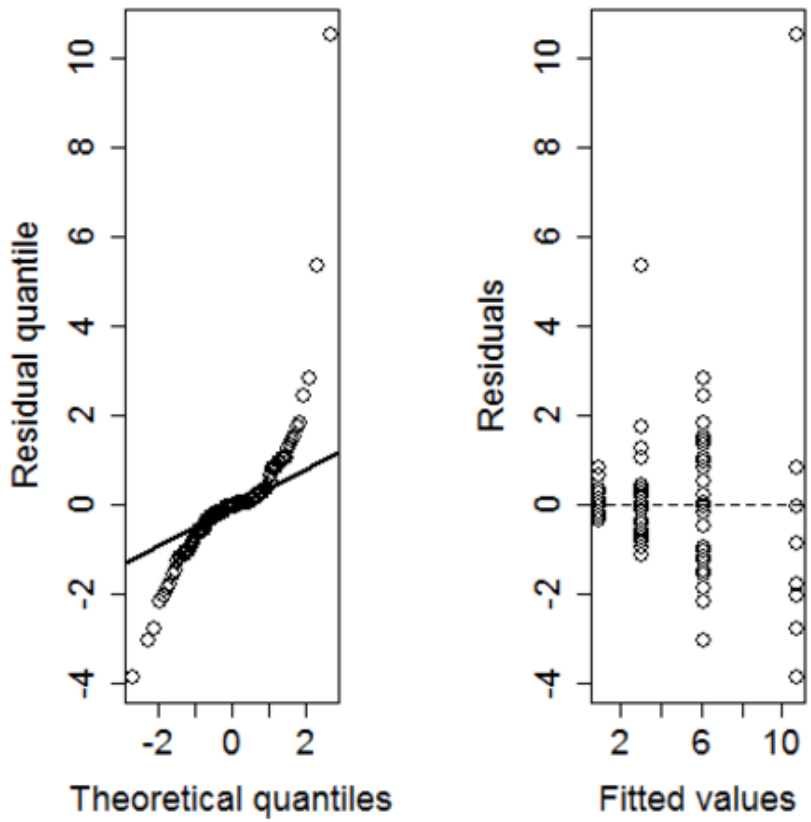
Juvenile fish is caught by electrofishing.

Want to test if weight of salmon is different between age-groups.

Caught salmon is in it first year up to ones in their fourth year.

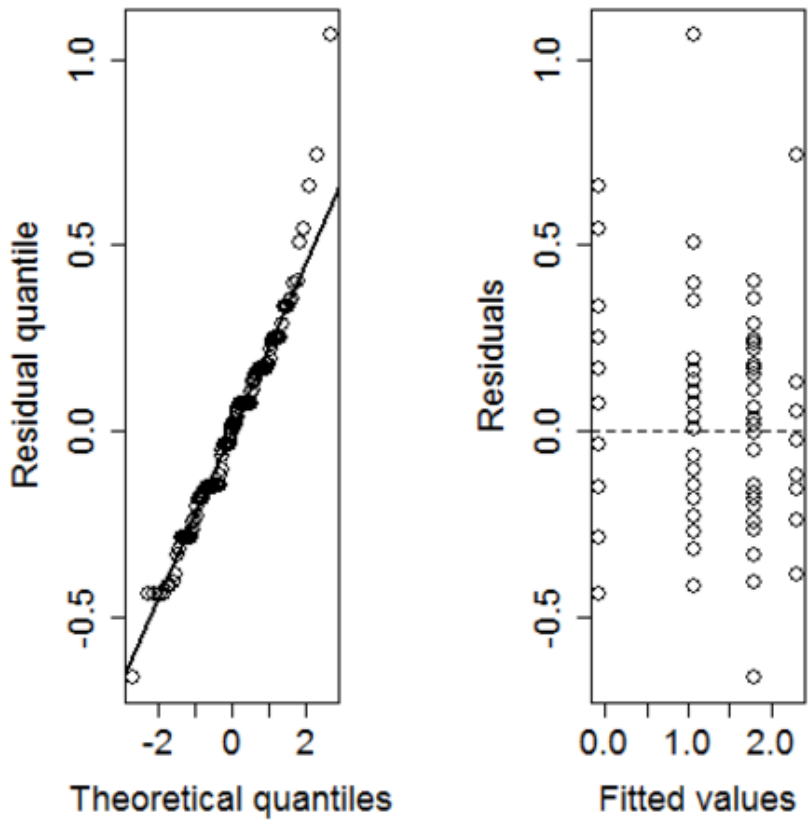
Example - electrofishing 2

The residuals do not follow a normal distribution and do not have constant variance.



Example - electrofishing 3

After log transforming weight the residuals now follow a normal distribution and the have a constant variance.



Example - electrofishing 4

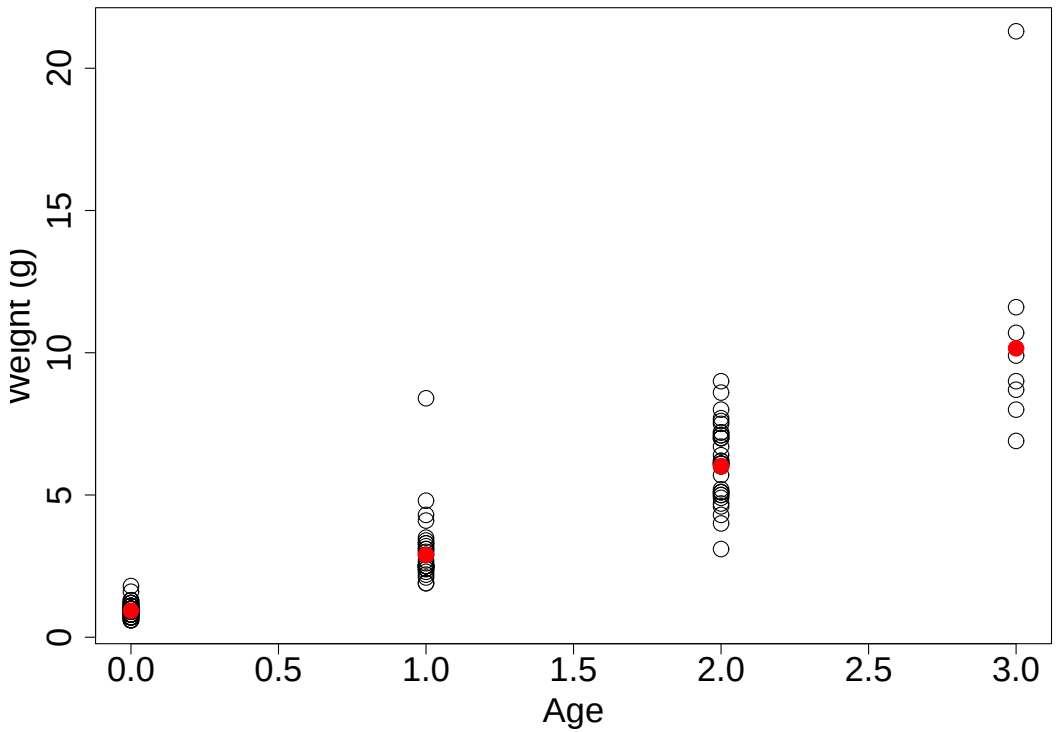


Figure: The data along with geometric mean as red dots

```
                  Df Sum Sq Mean Sq F value Pr(>F)
factor(dat1$age)   3 105.05   35.02   551.1 <2e-16 ***
Residuals        134   8.51    0.06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

Kruskal Wallis test

If the normality assumption can not be satisfied a non-parametric test called Kruskall-Wallis test can be applied

The only prerequisites is that all the groups have identical distributions.

Example - GMO Sugar 1

Experiment that compared an unmodified wild type sugarcane with three different genetically modified forms.

The measurements are weights of sugar that were obtained by breaking down the cellulose.

Example - GMO sugar 2

The research question is: Is there a difference in sugar production between the sugarcane varieties?
The ANOVA model is: $weight_{ij} = \mu + trt_i + \epsilon_{ij}$
and the hypotheses are:
$H_0 :_1 = trt_2 = trt_3 = trt_4 = 0$
H_1 : Not all $trt_i = 0$

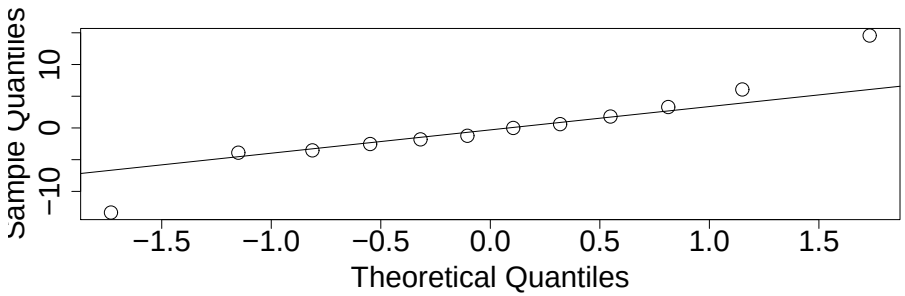
Example - GMO sugar 3

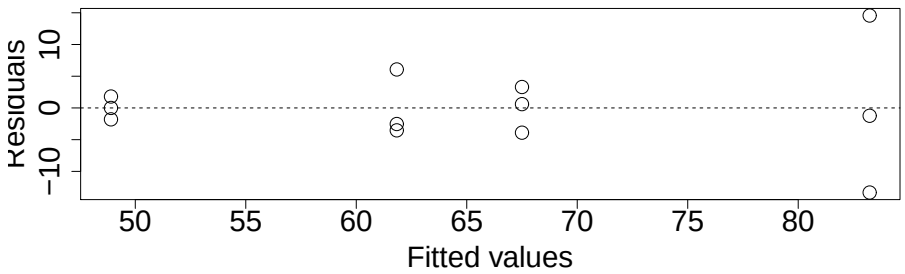
ANOVA is done in R and the assumptions are checked.

The variance does not seem to be constant.

Increasing with increasing fitted values.

Transformations are required.



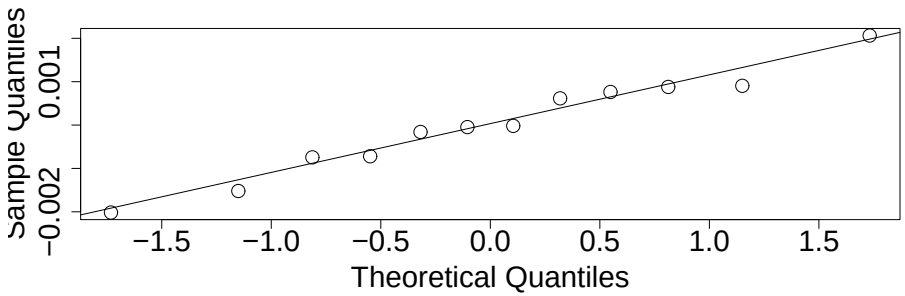


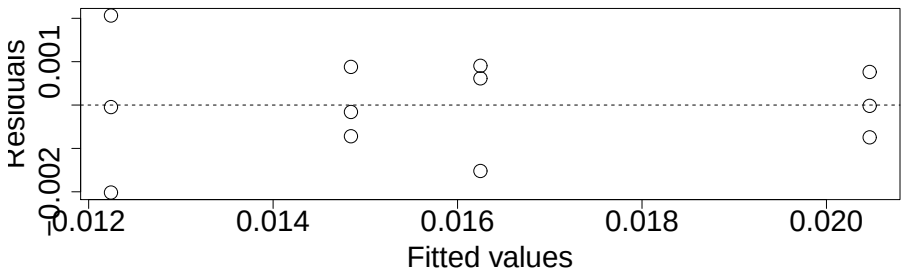
Example - GMO sugar 4

The transformations $log(y)$, $\sqrt{y}$ and $1/y$ were used.

The best one was $1/y$.

The variance looks a little bit better now, also the residuals seem to resemble the normal distribution better.





Example - GMO sugar 5

ANOVA table from R.

```
             Df     Sum Sq   Mean Sq F value  Pr(>F)
trt           3 1.064e-04 3.548e-05   19.86 0.00046 ***
Residuals     8 1.429e-05 1.790e-06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-value is 19.86 and associated P-value is 0.00046 which is lower than 0.05 so we conclude there is a significant difference between the treatment group means of sugar weight.