

# Multivariate Statistics in Ecology and Quantitative Genetics

## **Some remarks on balanced design and on parameter transformations.**

Dirk Metzler

[http://evol.bio.lmu.de/\\_statgen](http://evol.bio.lmu.de/_statgen)

26. Juni 2016

# Contents

Balanced Design

Randomized Balanced Block Design

Type I and Type II ANOVA

Transforming the Data

Hypothetical study: 100 LMU students were selected to participate in a 10km footrace. To motivate the participants, each participant got a release of the tuition fees, and this reward was better, the faster the students ran.

Hypothetical study: 100 LMU students were selected to participate in a 10km footrace. To motivate the participants, each participant got a release of the tuition fees, and this reward was better, the faster the students ran.

The aim of the study was to assess how the sportiveness depended on gender and smoking behavior. Thus, the students were subdivided into four groups:

	male	female	$\Sigma$
smoker	18	9	27
non-smoker	30	43	73
$\Sigma$	48	52	100

Hypothetical study: 100 LMU students were selected to participate in a 10km footrace. To motivate the participants, each participant got a release of the tuition fees, and this reward was better, the faster the students ran.

The aim of the study was to assess how the sportiveness depended on gender and smoking behavior. Thus, the students were subdivided into four groups:

	male	female	$\Sigma$
smoker	18	9	27
non-smoker	30	43	73
$\Sigma$	48	52	100

(Smoking seems to be gender-specific,  $p = 0.026$ , Fisher's exact test)

```
> t.test(runtime[smoking=="s"],runtime[smoking=="n"])
```

Welch Two Sample t-test

data: runtime[smoking == "s"] and runtime[smoking == "n"]

t = 0.1102, df = 60.611, p-value = 0.9126

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-7.522165 8.399714

sample estimates:

mean of x mean of y

91.06888 90.63010

```
> drop1(lm(runtime~smoking+sex),test="F")
```

Single term deletions

Model:

```
runtime ~ smoking + sex
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(F)
<none>			20570	538.64		
smoking	1	1078.7	21648	541.75	5.087	0.02635 *
sex	1	18548.6	39118	600.92	87.469	3.356e-15 ***

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

In another (hypothetical) survey, a balanced design was used, that is, equal numbers of students were selected for the four groups:



In another (hypothetical) survey, a balanced design was used, that is, equal numbers of students were selected for the four groups:

	male	female	$\Sigma$
smoker	25	25	50
non-smoker	25	25	50
$\Sigma$	50	50	100

In another (hypothetical) survey, a balanced design was used, that is, equal numbers of students were selected for the four groups:

	male	female	$\Sigma$
smoker	25	25	50
non-smoker	25	25	50
$\Sigma$	50	50	100

Balanced design, but no representative sampling!

```
> drop1(lm(runtime~smoking+sex),test="F")
```

Single term deletions

Model:

```
runtime ~ smoking + sex
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(F)	
<none>			23691	552.77			
smoking	1	3084.3	26776	563.01	12.628	0.0005889	***
sex	1	10648.1	34339	587.89	43.597	2.158e-09	***

---

Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1

```
> t.test(runtime[smoking=="s"],runtime[smoking=="n"])
```

Welch Two Sample t-test

data: runtime[smoking == "s"] and runtime[smoking == "n"]

t = 2.9669, df = 94.736, p-value = 0.003808

alternative hypothesis: true difference in means is not e

95 percent confidence interval:

3.674649 18.539956

sample estimates:

mean of x mean of y

101.1723 90.0650

Note that the linear model commands

```
summary(lm(runtime~smoking+sex))
```

and

```
drop1(lm(runtime~smoking+sex),test="F")
```

are neither restricted to representative sampling nor to balanced design.

## But how to interpret the group means?

Representative sampling:

```
> mean(runtime[sex=="male"])
[1] 76.99001
> mean(runtime[sex=="female"])
[1] 103.4488
> mean(runtime[smoking=="s"])
[1] 91.06888
> mean(runtime[smoking=="n"])
[1] 90.6301
```

Balanced design:

```
> mean(runtime[sex=="male"])
[1] 85.29967
> mean(runtime[sex=="female"])
[1] 105.9376
> mean(runtime[smoking=="s"])
[1] 101.1723
> mean(runtime[smoking=="n"])
[1] 90.065
```

## But how to interpret the group means?

Representative sampling:

```
> mean(runtime[sex=="male"])
[1] 76.99001
> mean(runtime[sex=="female"])
[1] 103.4488
> mean(runtime[smoking=="s"])
[1] 91.06888
> mean(runtime[smoking=="n"])
[1] 90.6301
```

Balanced design:

```
> mean(runtime[sex=="male"])
[1] 85.29967
> mean(runtime[sex=="female"])
[1] 105.9376
> mean(runtime[smoking=="s"])
[1] 101.1723
> mean(runtime[smoking=="n"])
[1] 90.065
```

In the balanced design, smokers are overrepresented (compared to reality), and females are overrepresented among the smokers and underrepresented among the non-smokers.

Let  $i$  be the index for the row of a data table. The data are subdivided into groups and  $G_i$  is the group row  $i$  (or patient  $i$ ) belongs to; e.g.  $G_i$  can be the treatment of patient  $i$ . Let  $Y_i$  be the response variable, e.g. the blood pressure of patient  $i$ . We can apply an anova to check whether  $Y$  depends on  $G$ , and the model behind it is:

$$Y_i = b_{G_i} + \varepsilon_i$$

where the  $\varepsilon_i$  are assumed to be independent and normally distributed with expectation 0, and all  $\varepsilon_i$  have the same variance  $\sigma^2$ . During the ANOVA we estimate the influence  $b_{G_i}$  of the group on  $Y_i$  by the group mean  $\widehat{b}_g$ . Thus, the residuals  $r_i := Y_i - \widehat{b}_{G_i} \approx Y_i - b_{G_i} = \varepsilon_i$  should be approximately normally distributed.



More than one factor can play a role. For example we may take into account that the blood pressure  $Y_i$  of a patient may depend on the sex  $S_i$  of the patient. In this case the model behind the anova takes the form

$$Y_i = b_{G_i} + c_{S_i} + \varepsilon_i.$$

$b_{G_i}$  depends only on the treatment group and  $c_{S_i}$  only on the sex of the female. If we also want allow in *interaction* between the treatment and the sex, we need another variable  $d_{G_i, S_i}$  that may depend on both:

$$Y_i = b_{G_i} + c_{S_i} + d_{G_i, S_i} + \varepsilon_i.$$

This makes possible, for example, that a certain treatment has a stronger effect for males than for females.

A *balanced design* means, that the sample size are the same for each combination of factors. E.g. 10 males and 10 females in each treatment group. Some ANOVA-based method will only work for balanced designs. Therefore, it is preferable to use a balanced design when planning an experiment. If the data, however, are observations from nature, the “design” is usually unbalanced and this has to be taken into account in the analysis.

One of the methods for which you need a balanced design is Tukey's HSD (honest significant differences). From an ANOVA it computes confidence intervals for the pairwise differences between the group means with multiple-testing correction (see slides on ANOVA in the EES&MEME basic statistics course).

# Contents

Balanced Design

Randomized Balanced Block Design

Type I and Type II ANOVA

Transforming the Data

The npk dataset from the MASS<sup>1</sup> library: Yield of peas that grew with or without application of nitrogen (N), phosphate (P), and potassium (K).

---

<sup>1</sup>Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth edition. Springer

The npk dataset from the MASS<sup>1</sup> library: Yield of peas that grew with or without application of nitrogen (N), phosphate (P), and potassium (K).

The pease grew on 6 different fields (“blocks”), each of which was subdivided into four parts with different treatments.

---

<sup>1</sup>Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth edition. Springer

The npk dataset from the MASS<sup>1</sup> library: Yield of peas that grew with or without application of nitrogen (N), phosphate (P), and potassium (K).

The pease grew on 6 different fields (“blocks”), each of which was subdivided into four parts with different treatments.

We compensate for effects of the block and randomize within and between the blocks.

---

<sup>1</sup>Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth edition. Springer

The npk dataset from the MASS<sup>1</sup> library: Yield of peas that grew with or without application of nitrogen (N), phosphate (P), and potassium (K).

The pease grew on 6 different fields (“blocks”), each of which was subdivided into four parts with different treatments.

We compensate for effects of the block and randomize within and between the blocks.

Balanced design: Each treatment appears three times.

---

<sup>1</sup>Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth edition. Springer

Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
PK 49.5	N 59.8	P 62.8	N 62	NP 52	NK 57.2
NP 62.8	NPK 58.5	NPK 55.8	NPK 48.8	— 51.5	NP 59
— 46.8	K 55.5	N 69.5	K 45.5	NK 49.8	PK 53.2
NK 57	P 56	K 55	P 44.2	PK 48.8	— 56

- ▶ Note the balance within the blocks: Any substance appears twice in each block.
- ▶ Cannot estimate triple interaction N:P:K because it is confounded with block differences.



```
> (npk.aov <- aov(yield~block + N*P*K,data=npk))
```

```
Call:
```

```
  aov(formula = yield ~ block + N * P * K, data = npk)
```

```
Terms:
```

	block	N	P	K	N:P	N:K	P:K
Sum of Squares	343.2950	189.2817	8.4017	95.2017	21.2817	33.1350	0.4817
Deg. of Freedom	5	1	1	1	1	1	1
	Residuals						
Sum of Squares	185.2867						
Deg. of Freedom	12						

```
Residual standard error: 3.929447
```

```
1 out of 13 effects not estimable
```

```
Estimated effects may be unbalanced
```

```
> summary(npk.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
block	5	343.29	68.659	4.4467	0.015939	*
N	1	189.28	189.282	12.2587	0.004372	**
P	1	8.40	8.402	0.5441	0.474904	
K	1	95.20	95.202	6.1657	0.028795	*
N:P	1	21.28	21.282	1.3783	0.263165	
N:K	1	33.13	33.135	2.1460	0.168648	
P:K	1	0.48	0.482	0.0312	0.862752	
Residuals	12	185.29	15.441			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Giving  $p$  values for variables that are also involved in interaction terms makes sense only if the design is balanced. It refers to a coefficient that is averaged over the different cases of interaction.

From the R manual page of “aov”:

*“ ‘aov’ is designed for balanced designs, and the results can be hard to interpret without balance: beware that missing values in the response(s) will likely lose the balance.”*

The command `drop1(lm(...), test='F')` does not assume a balanced design and therefore does not report  $p$  values for variables that are involved in interactions.

```
> drop1(lm(yield~block +(N+P+K)*(N+P+K),data=npk),test="F")
Single term deletions
```

```
Model:
```

```
yield ~ block + (N + P + K) * (N + P + K)
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(F)	
<none>			185.29	73.052			
block	5	343.30	528.58	88.211	4.4467	0.01594	*
N:P	1	21.28	206.57	73.662	1.3783	0.26317	
N:K	1	33.14	218.42	75.001	2.1460	0.16865	
P:K	1	0.48	185.77	71.115	0.0312	0.86275	
---							

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Contents

Balanced Design

Randomized Balanced Block Design

Type I and Type II ANOVA

Transforming the Data

Be careful with the interpretation of ANOVA tables! The R command `anova`, applied to a single model gives a so-called “Type I Anova”, where each line take only the variables in the lines above into account. Example: Chill coma recovery times measured by different persons on different days for different fly lines.

```
> anova(model4)
```

```
Analysis of Variance Table
```

```
Response: log(ccrt)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
line	1	1.2224	1.22238	13.1486	0.0003812	***
day	11	2.8471	0.25883	2.7841	0.0023769	**
person	1	0.0850	0.08504	0.9147	0.3402393	
[...]						

For example, the p-value 0.0023769 tells how much better the model with line and day can explain the data compared to a model that only takes line into account. Thus, the values assigned to variables depend on the input order.

If you use the R command `drop1` with the option `test="F"`, you get a so-called "Type II Anova", in which each line shows the influence of one variable, given the estimates of *all* other variables.

```
> drop1(model4, test="F")
[...]
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(F)
<none>			15.618	-418.91		
line	1	0.05860	15.677	-420.23	0.6304	0.428338
day	11	2.47080	18.089	-414.18	2.4161	0.008177 **
person	1	0.08504	15.703	-419.92	0.9147	0.340239

For example, the  $p$ -value 0.008177 says that a model that takes line, day and person into account explains the data significantly better than a model that uses only line and person.

Back to the footrace example with non-balanced design:

```
> summary(aov(runtime~sex+smoking))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	17473.7	17473.7	82.400	1.316e-14	***
smoking	1	1078.7	1078.7	5.087	0.02635	*
Residuals	97	20569.7	212.1			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> summary(aov(runtime~smoking+sex))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
smoking	1	3.8	3.8	0.0179	0.8939	
sex	1	18548.6	18548.6	87.4693	3.356e-15	***
Residuals	97	20569.7	212.1			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



But for the dataset with *balanced design* (for which aov is more appropriate) the input order does not matter even for Type I anova:

```
> summary(aov(runtime~sex+smoking))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sex	1	10648.1	10648.1	43.597	2.158e-09	***
smoking	1	3084.3	3084.3	12.628	0.0005889	***
Residuals	97	23691.2	244.2			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> summary(aov(runtime~smoking+sex))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
smoking	1	3084.3	3084.3	12.628	0.0005889	***
sex	1	10648.1	10648.1	43.597	2.158e-09	***
Residuals	97	23691.2	244.2			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Contents

Balanced Design

Randomized Balanced Block Design

Type I and Type II ANOVA

Transforming the Data

It is often important to rescale (i.e. transform) the data. For example, if a comparison between fitted values (group means) and the residuals show that the larger values have larger standard deviations, this may mean that the random error is rather multiplicative than additive (as it should be). In this case, a log transform may help. Sometimes, there is a good explanation why a certain transformation should be applied. Sometimes the Box-Cox-Transform can help, which can take various shapes, depending on a parameter to be optimized. Other transformations are also possible, not only for the target variable but also for explanatory variables in regression models.

Back to the example with chill coma recovery times with simulated data motivated by



N. Svetec, A. Werzner, R. Wilches, P. Pavlidis, J.M. Alvarez-Castro, K.W. Broman, D. Metzler, W. Stephan (2011) Identification of X-linked quantitative trait loci affecting cold tolerance in *Drosophila melanogaster* and fine mapping by selective sweep analysis.  
*Molecular Ecology* **20**(3):530-544

```
> fly <- read.table("CCRT.txt",h=T)
> str(fly)
'data.frame': 182 obs. of 7 variables:
 $ line   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1
 $ day    : Factor w/ 12 levels "May10","May11",...: 12 12
 $ box    : int    4 4 4 4 4 4 4 4 4 4 ...
 $ ISO    : int    2 2 2 2 2 2 2 2 2 2 ...
 $ day.no: int    12 12 11 12 12 12 11 11 12 11 ...
 $ person: Factor w/ 2 levels "A","B": 2 2 1 2 2 2 1 1 2
 $ ccrt   : int    41 52 37 16 33 37 19 45 41 39 ...
```

```
> drop1(model,test="F")
```

```
Single term deletions
```

```
Model:
```

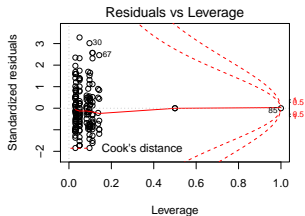
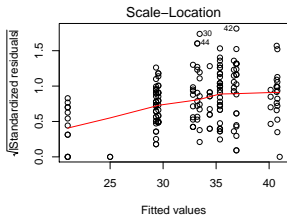
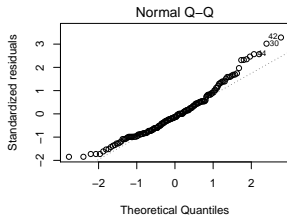
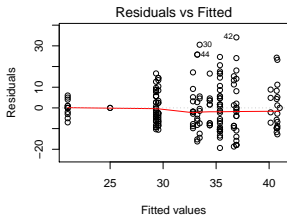
```
ccrt ~ line + box + day + person
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(F)
<none>			19046	874.41		
line	1	58.22	19105	872.97	0.5135	0.47460
box	0	0.00	19046	874.41		
day	10	2300.77	21347	875.17	2.0294	0.03318 *
person	1	98.55	19145	873.35	0.8693	0.35250

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Transforming the Data



```
> model2 <- lm(log(ccrt)~line+box+day+person,fly)
> drop1(model2,test="F")
```

Single term deletions

Model:

```
log(ccrt) ~ line + box + day + person
```

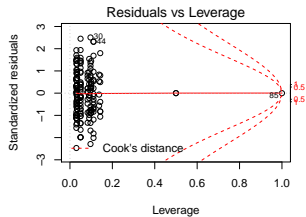
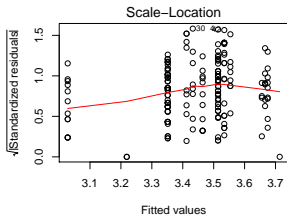
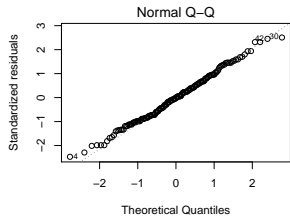
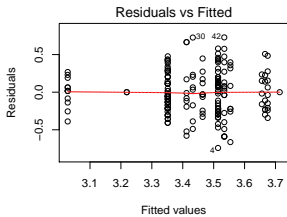
	Df	Sum of Sq	RSS	AIC	F value	Pr(F)	
<none>			15.618	-418.91			
line	1	0.05860	15.677	-420.23	0.6304	0.428338	
box	0	0.00000	15.618	-418.91			
day	10	2.45864	18.077	-412.30	2.6446	0.005096	**
person	1	0.08504	15.703	-419.92	0.9147	0.340239	

---

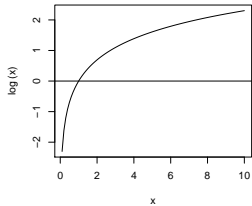
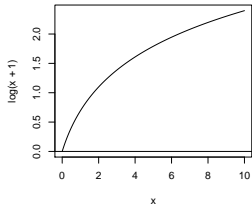
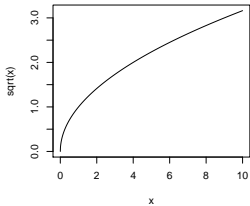
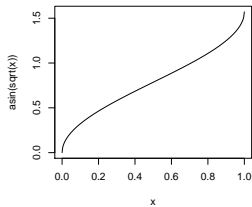
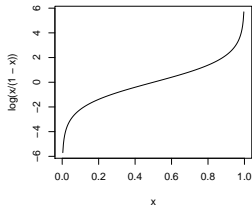
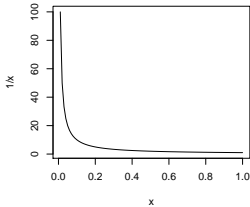
Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1



# Transforming the Data

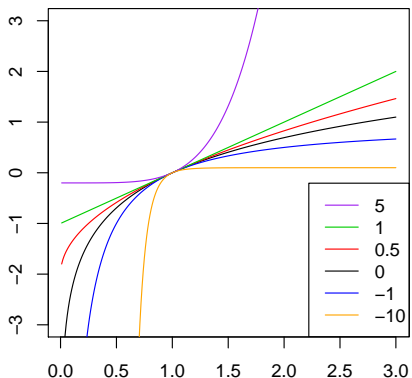


# Popular Transformations

**log****log(1+x)****sqrt(x)****arcsin(sqrt(x))****logit****1/x**

# Box-Cox-Transformations

Box-Cox transformations



$$f_{\lambda}(x) = \begin{cases} \log(x) & \text{if } \lambda = 0 \\ (x^{\lambda} - 1) / \lambda & \text{if } \lambda \neq 0 \end{cases}$$

```
boxcox(ccrt~line+box+day+person,data=fly)
```

