

*Experimental Design and Statistical Methods*



*Workshop*

# COMPARISON OF MEANS OF SEVERAL RANDOM SAMPLES. ANOVA

**Jesús Piedrafita Arilla**

[jesus.piedrafita@uab.cat](mailto:jesus.piedrafita@uab.cat)

*Departament de Ciència Animal i dels Aliments*



# Items

- One-way ANOVA (completely random design):
  - Principles for partitioning variation
  - ANOVA table and F test
  - Conditions of applicability
    - Normality in the residuals and transformations
    - Equality of variances
  - Comparisons of means
    - Multiple comparisons
    - Pre-planned comparisons
  - Power
  - Sample size
  - Matrix version
- Basic commands
  - `bartlett.test`, `leveneTest`
  - `tapply` - `cbind`
  - `aov` – `anova`, `summary lm`
  - `layout` – `plot`
  - `tukeyHSD`
  - `contrasts`
- Libraries
  - `car`
  - `agricolae`

# Inferences on more than two samples

This session is devoted to present methods to compare location parameters (means) of more than two samples through parametric procedures.

When we have more than two samples, **ANOVA techniques are preferable over several pair-wise comparisons** because of two main reasons:

1. **We get a better estimate of the within group (residual) variance.**
2. **The probability of false positives (Type I errors) is lower.** The *Experiment-wise error rate*, i.e., the probability of finding a significant result by chance when  $c$  comparisons are made is  $1 - (1 - \alpha)^c$ ,  $\alpha$  being the *Comparison-wise error rate*.

Comparisons	Prob. Not Significant	Prob. Significant
1	$(1 - \alpha)$	$1 - (1 - \alpha) = \alpha$
2	$(1 - \alpha)^2$	$1 - (1 - \alpha)^2$
.	.	.
.	.	.
$c$	$(1 - \alpha)^c$	$1 - (1 - \alpha)^c$

# One-way ANOVA (1) – Completely Random Design

Suppose we have 30 young bulls assigned randomly to three treatments (feed additives) and we measure carcass conformation on a SEUROP scale.

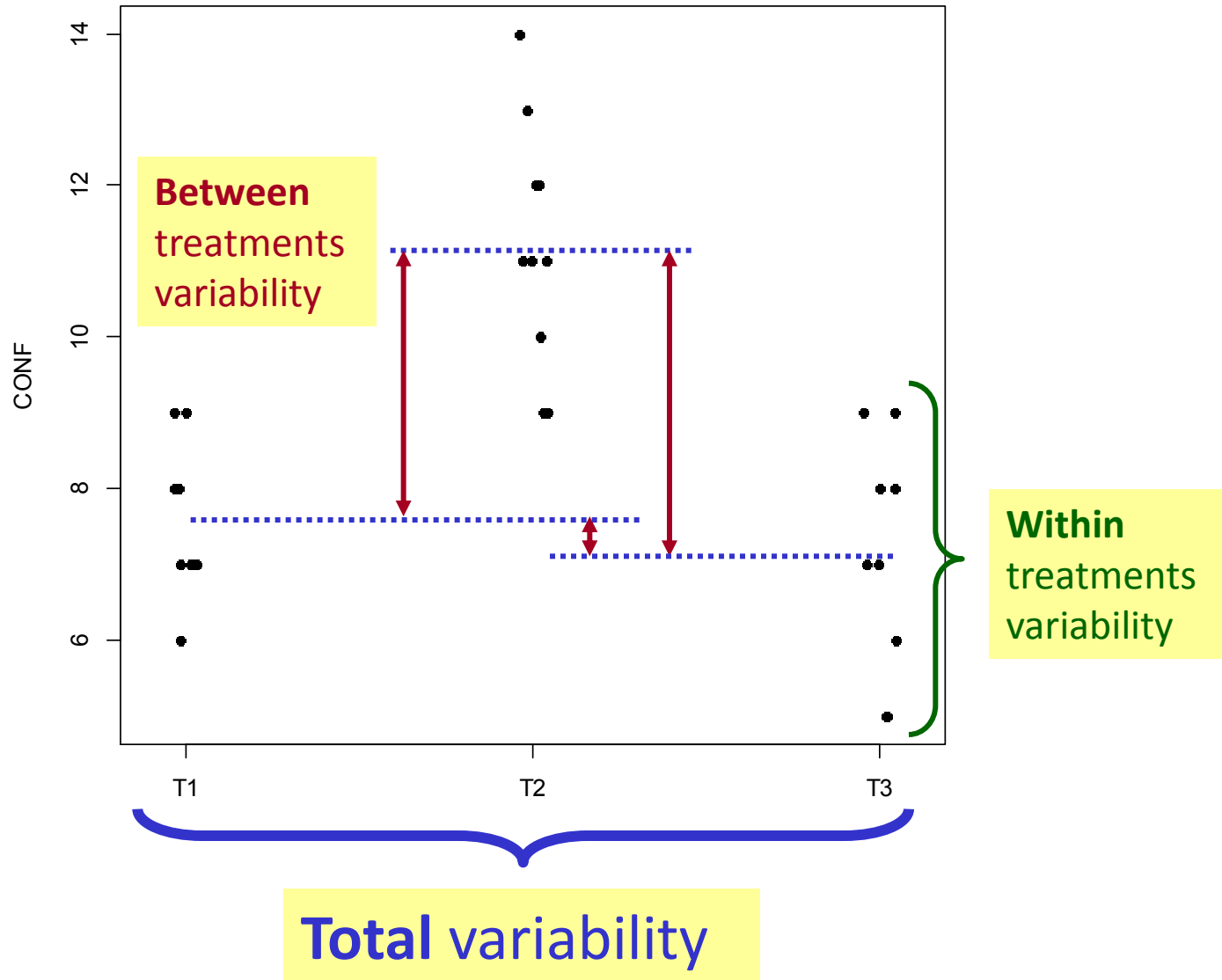
$\tau_i$	T1	T2	T3
	7	9	8
	8	13	7
	9	12	8
	7	11	5
	6	14	6
	9	11	9
	8	10	8
$y_{ij}$	7	12	5
	8	9	7
	7	11	9



with intermediate scores (+/-)

¿Are significantly different the means of the three treatments,  $H_0 : \bar{y}_1 = \bar{y}_2 = \bar{y}_3$  ?

# One-way ANOVA (2)



# One-way ANOVA (3)

## Model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where    i)  $1 \rightarrow t$   
          j)  $1 \rightarrow n$ , balanced design  
               $1 \rightarrow n_i$ , unbalanced design

↓  
**Level of a factor (fixed effect)**

Mean of level  $i$       Overall mean

$$y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

## Sums of squares (SS)

By squaring and summing these quantities, we arrive after some algebra at the computing formulas of the next slide

↓  
**Total SS,  $SS_T$**

↓  
**Within groups (residual),  $SS_W$**

↓  
**Between groups,  $SS_B$**

**! We make a partition of the variation!**

# One-way ANOVA (4) (for a balanced design)

Source of variation	Degrees of freedom	Sum of squares	Mean squares	$F$
Between groups	$t-1$	$SS_B$	$SS_B / (t-1)$	$MS_B / MS_W$
Within groups (error)	$t(n-1)$	$SS_W$	$SS_W / (tn-t)$	
Total	$tn-1$	$SS_T$	$SS_T / (tn-1)$	

$F = \frac{(SS_B / \sigma^2) / (t-1)}{(SS_W / \sigma^2) / (tn-t)}$

$$SS_B = \frac{\sum_i y_{i.}^2}{n} - \frac{y_{..}^2}{tn}$$

$$SS_W = \sum_{ij} y_{ij}^2 - \frac{\sum_i y_{i.}^2}{n}$$

$$SS_T = SS_B + SS_W \quad \Rightarrow \quad SS_T = \sum_{ij} y_{ij}^2 - \boxed{\frac{y_{..}^2}{tn}} \quad \text{Correction term}$$

# One-way ANOVA (5), contrast of hypothesis

The **null hypothesis** and the **alternative hypothesis** can be stated as:

$H_0: \mu_1 = \mu_2 = \mu_3$ , the population means are equal


$H_1: \mu_i \neq \mu_{i'},$  for at least one pair  $(i, i')$ , the means are not equal

An **equivalent formulation** of the hypothesis is:

$H_0: \tau_1 = \tau_2 = \tau_3$ , there is no difference among treatments

$H_1: \tau_i \neq \tau_{i'},$  for at least one pair  $(i, i')$ , a difference among treatments exist

It can be shown that the **expectations of the mean squares** are:

$E(MS_W) = \sigma^2$    $MS_W$  is an unbiased estimator of  $\sigma^2$

$$E(MS_B) \begin{cases} = \sigma^2 & \text{if } H_0 \\ > \sigma^2 & \text{if not } H_0 \end{cases} \quad \text{red arrow} \quad E(MS_B) = \sigma^2 + \frac{n \sum_i \tau_i^2}{t-1}, \quad \text{for } \sum_i \tau_i = 0$$

$$F = \frac{E(MS_B)}{E(MS_W)} = \frac{\sigma^2 + \frac{n \sum_i \tau_i^2}{t-1}}{\sigma^2} > 1 \quad \text{if not } H_0$$



# One-way ANOVA (6), our example

Source of variation	Degrees of freedom	Sum of squares	Mean squares	F
Between groups	$3 - 1 = 2$	$2350.4 - 2253.333 = 97.067$	$97.067 / 2 = 48.533$	$48.5333 / 1.911 = 25.40$
Within groups (error)	$3 * (10 - 1) = 27$	$2402 - 2350.4 = 51.6$	$51.6 / 27 = 1.911$	
Total	$3 * 10 - 1 = 29$	$2402 - 2253.33 = 148.667$	$148.667 / 29 = 5.126$	

$$\sum_{ij} y_{ij}^2 \Rightarrow 7^2 + 8^2 + 9^2 \dots + 5^2 + 7^2 + 9^2 = 2402$$

$$\frac{\sum_i y_i^2}{n} \Rightarrow \frac{76^2 + 112^2 + 72^2}{10} = 2350.4$$

$$\frac{y_{..}^2}{tn} \Rightarrow \frac{(7 + 8 + 9 + \dots + 5 + 7 + 9)^2}{3 \times 10} = 2253.333$$

$$25.40 > F_{2,27} (=3.35)$$



**Reject  $H_0$ , means are significantly different ( $p < 0.05$ ), BUT which ones?**  
***Pre-planned or Multiple comparisons***

## Protocol to develop ANOVA

1. Import the data (reorganize levels of the factor, if needed)
2. Assess normality/homogeneity of variance using boxplots
3. Assess homogeneity of variance assumption
4. Test  $H_0$  that population group mean are all equal – perform analysis of variance
5. Examine the ANOVA table
6. Make a residual plot
7. Perform post-hoc tests
8. Compute the power of the ANOVA test

# Conditions of applicability of ANOVA

The conditions of applicability (additivity and normal errors identically distributed with common variance  $\sigma^2$ ) must be met when we want to make some inference, such as the estimation of the confidence interval or some hypothesis testing. There are **two main conditions** to be checked:

1. **Normality of errors.** We check this in the residuals, our best estimate of the errors. The analysis of variance, however, is little sensitive (robust) to the non normality of the populations under study. In practice it is enough to avoid the use of the ANOVA when the samples deviate heavily from the normal distribution or the distribution of the samples is very different, mainly in small samples.
2. **Homogeneity of within group variances.** It can be tested through the Bartlett  $\chi^2$  test, among others. Its importance is relatively secondary when sample sizes are equal (balanced designs).

If the conditions of applicability are not met, we can use some **transformations**.

# Testing for homogeneity of within group variances (1)

```
> bartlett.test(CONF~TRT, CONFBEEF)
```

Bartlett test of homogeneity of variances

data: CONF by TRT

Bartlett's K-squared = 2.3132, df = 2, p-value = 0.3145



The null hypothesis of homogeneity of within treatment variance is not rejected

## Be careful:

The Bartlett test is too sensitive to the deviations to normality and may indicate non normality instead of variance heterogeneity

```
> 1-pchisq(2.3132, 2)  
[1] 0.3145538
```

## Testing for homogeneity of within group variances (2)

```
> #Levene test  
> library(car)  
> leveneTest(y = CONF, group = TRT)
```

Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	0.809	0.4558
	27		

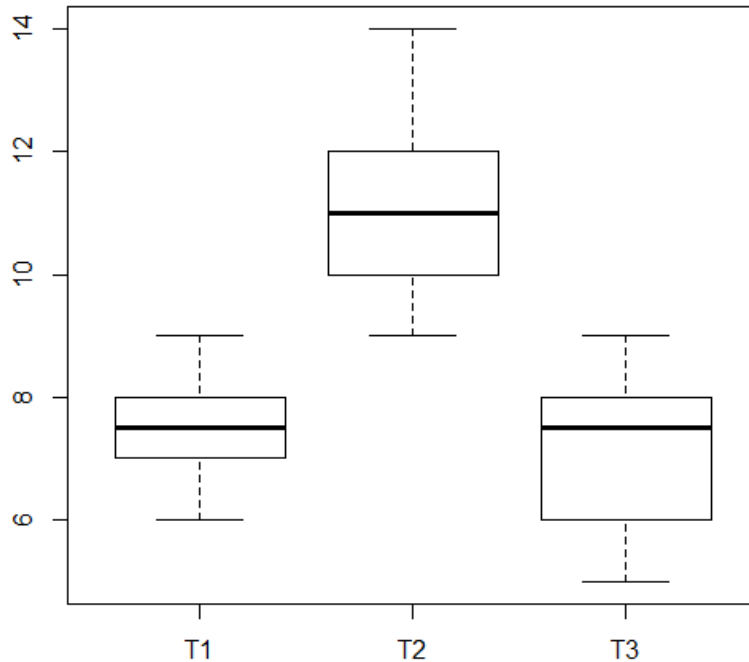


The null hypothesis of homogeneity of within treatment variance is not rejected

```
> 1-pf(0.809,2 27)  
[1] 0.4558065
```

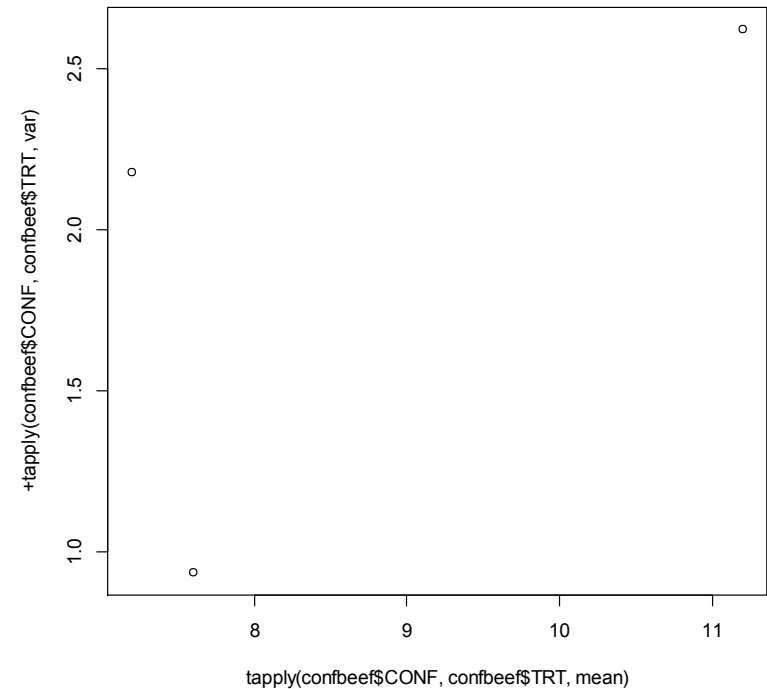
# Working ANOVA with R (1)

```
> boxplot(CONF~TRT, CONFBEEF)
```



No obvious violations of normality and homogeneity of variance (boxplots not asymmetrical and do not vary greatly in size)

```
> plot(tapply(CONF, TRT, mean),  
+      tapply(CONF, TRT, var))
```



No obvious relationship between group (treatment) mean and variance

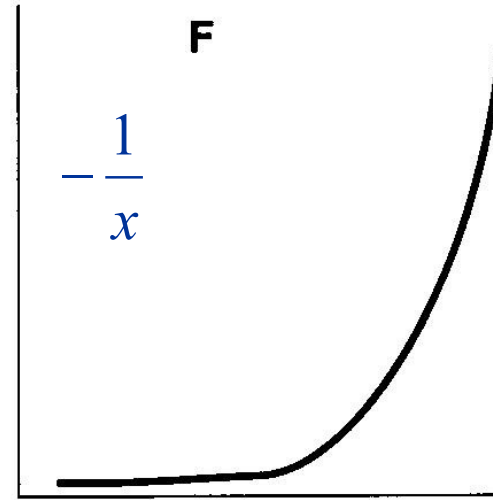
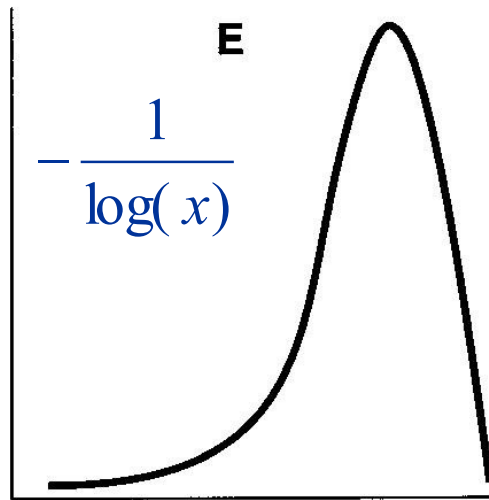
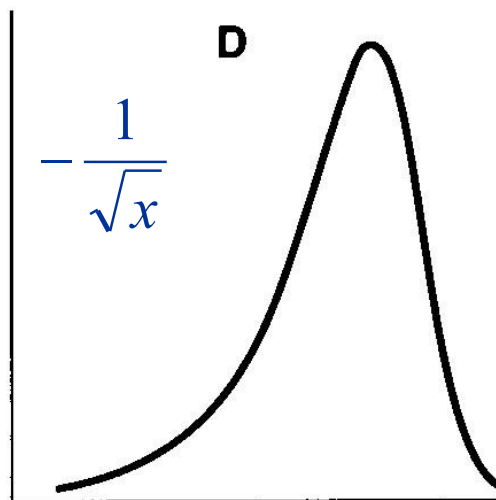
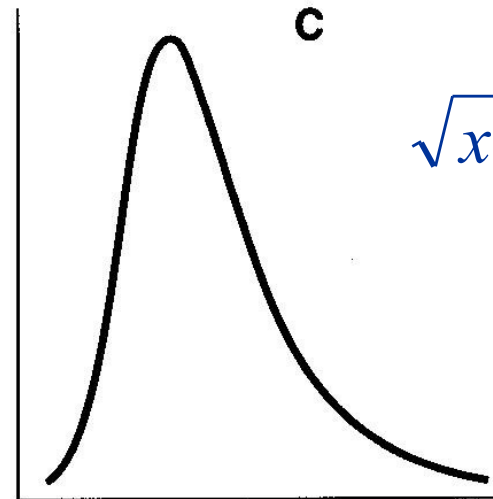
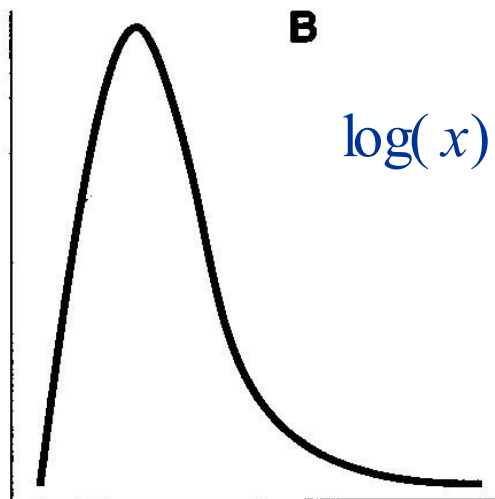
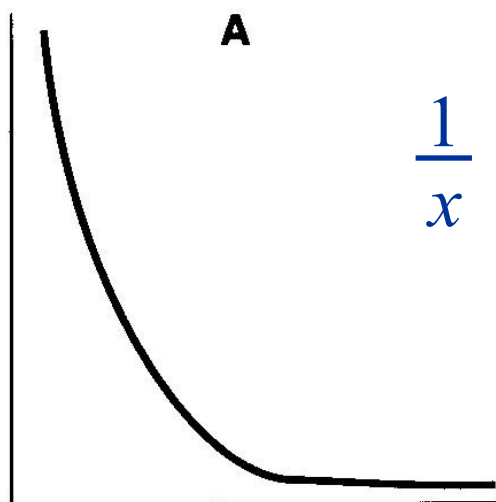
# Transformations (1)

The objective is to normalise the distribution and to stabilise the variances. If this procedure does not give satisfactory results Non Parametric Tests can be used.

1. **Logarithmic.** When the treatments have a multiplicative effect, i.e., when they increment or decrement the measurements in a percent and not in a fixed quantity.
2. **Root square.** For data consisting of integers coming from counting (ticks on a cow). It tends to equalise  $\sigma^2$ .
3. **Angular** (arcsine). Data are the number of individuals with some particular characteristic (percentages and proportions). Equalises  $\sigma^2$ .
4. **Probit.** For percentage data, like mortality. It is used in pharmacology.
5. **Box-Cox.** A general methodology to transform data.

**To present a true mean value of data in the linear scale it is necessary to reconvert the transformed mean. The standard deviation in this case is of no value and you should compute confidence limits of the transformed data and then convert these to the linear scale.**

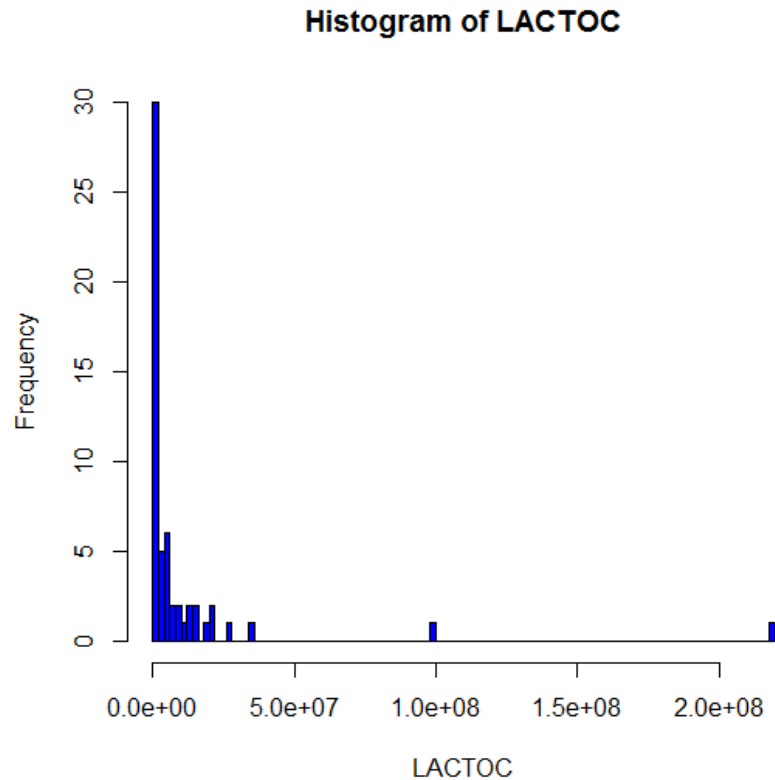
## Transformations (2)



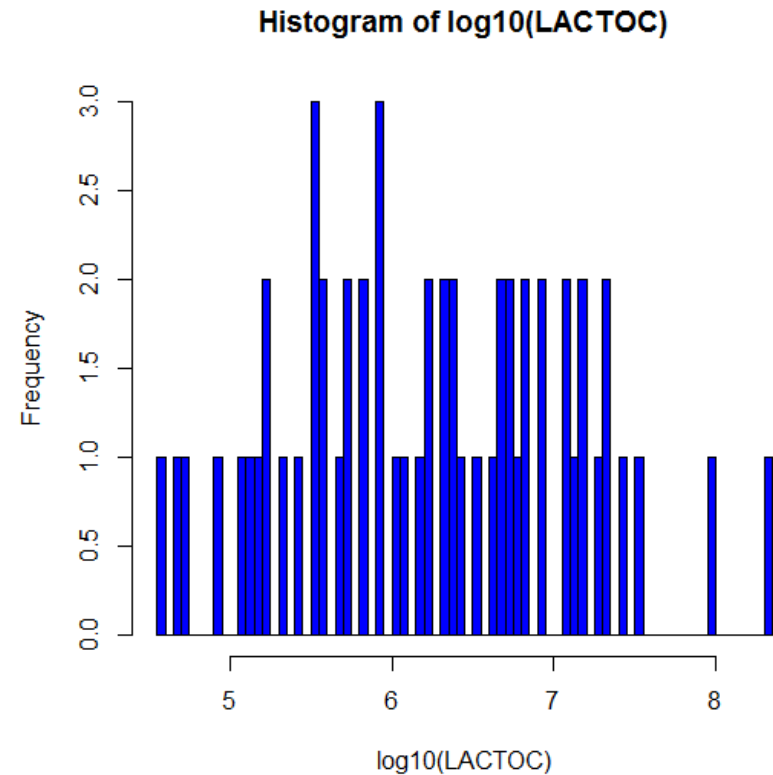


# Transformations (example)

```
>hist(LACTOC,col="blue",breaks=100)
```



```
>hist(log10(LACTOC),col="blue",breaks=100)
```



## Working ANOVA with R (2)

```
> CONFBEEF.AOV<-aov(CONF~TRT, CONFBEEF)
> anova(CONFBEEF.AOV)
```

Analysis of Variance Table

Response: CONF

	Df	Sum Sq	Mean Sq	F value
TRT	2	97.067	48.533	25.395
Residuals	27	51.600	1.911	

```
> 1-pf(25.395,2,27)
[1] 6.250862e-07
```

Pr(>F)
6.25e-07 ***

---

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

This is a typical ANOVA table that includes the  $p$ -value for the significance of the  $F$  value. This allows us to **reject the overall null hypothesis** of no influence of the factors included in the model. In this case we reject  $H_0$  of equality among all treatments.

## Working ANOVA with R (2b)

```
> CONFBEEF.LM<-lm(CONF~TRT, CONFBEEF)
> anova(CONFBEEF.LM)
```

### Analysis of Variance Table

Response: CONF

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TRT	2	97.067	48.533	25.395	6.25e-07 ***
Residuals	27	51.600	1.911		

---

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

Observe that we have used **lm** instead of **aov**. **lm** stands for Linear Model, a more general procedure than **aov**.

## Working ANOVA with R (3)

```
> CONFBEEF.AOV<-aov(CONF~TRT, CONFBEEF)
> summary.lm(aov(CONF~TRT))
```

```
Call: aov(formula = CONF ~ TRT)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-2.2	-0.6	-0.2	0.8	2.8

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	7.6000	0.4372	17.385	3.44e-16	***
TRTT2	3.6000	0.6182	5.823	3.38e-06	***
TRTT3	-0.4000	0.6182	-0.647	0.523	

Residual standard error: 1.382 on 27 degrees of freedom  
Multiple R-squared: 0.6529, Adjusted R-squared: 0.6272  
F-statistic: 25.4 on 2 and 27 DF, p-value: 6.25e-07

*R-squared* is a measure of the fit of the model,  $SS_{Model} / SS_T$ , and ranges from 0 to 1. It quantifies the proportion of (the variability) of the dependent variable explained by the model (independent variables), in this case 65.29%.

*Residual standard error* is an estimate of the within group standard deviation.

# Working ANOVA with R (3b)

```
> CONFBEEF.LM<-lm(CONF~TRT, CONFBEEF)
> summary(CONFBEEF.LM)
```

Call:

```
lm(formula = CONF ~ TRT, data = CONFBEEF)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.2	-0.6	-0.2	0.8	2.8

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	7.6000	0.4372	17.385	3.44e-16	***
TRTT2	3.6000	0.6182	5.823	3.38e-06	***
TRTT3	-0.4000	0.6182	-0.647	0.523	

---

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

Residual standard error: 1.382 on 27 degrees of freedom

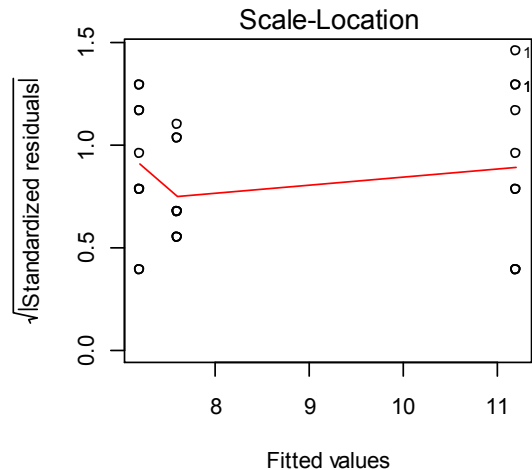
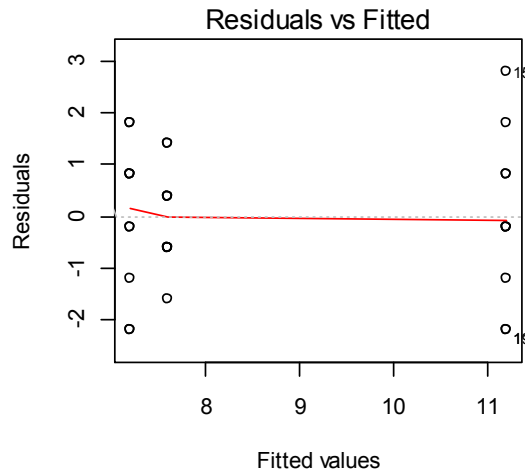
Multiple R-squared: 0.6529, Adjusted R-squared: 0.6272

F-statistic: 25.4 on 2 and 27 DF, p-value: 6.25e-07

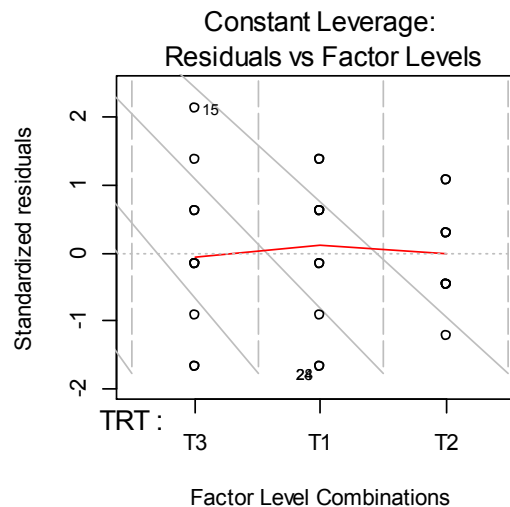
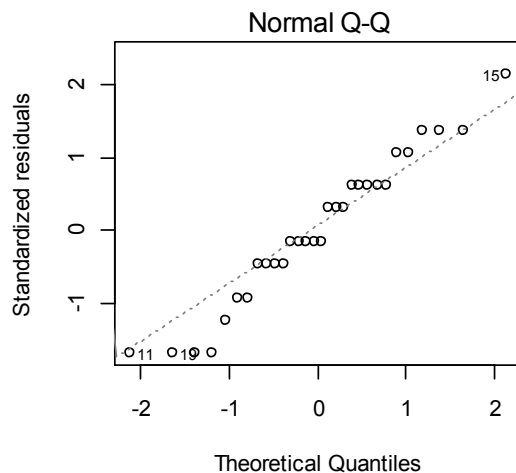
# Working ANOVA with R (4). Analysis of residuals

```
> layout(matrix(c(1,2,3,4),2,2))  
> plot(CONFBEEF.AOV)
```

$$\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{\tau}_i$$



No obvious violation of homogeneity of variance: no wedge shape in residuals.



No obvious violation of normality: Q-Q plot of residuals is linear.

Cook's D values meaningless in ANOVA.

# Multiple comparisons

If  $H_0$  is not rejected, it is not necessary or appropriate to further analyse the problem, although the researcher must be aware of the possibility of a Type II error.

If  $H_0$  is rejected, then we must question which treatment or treatments caused a differential effect, that is, between which groups is the significant difference found.

For  $t$  treatments, there is a total of  $\binom{t}{2}$  pair-wise comparisons of means. For each comparison there is the possibility of making Type I or Type II errors.

Looking at the experiment as a whole, the probability of making an error in conclusion is defined as the Experiment-wise Error Rate (EER).

There are many procedures of pair-wise comparisons of means: Bonferroni, Duncan, Dunnet, LSD, Scheffé, Student-Newman-Keuls, **Tukey**, among others.

## Working ANOVA with R (5)

Before comparing the means, we may want to print a table with them:

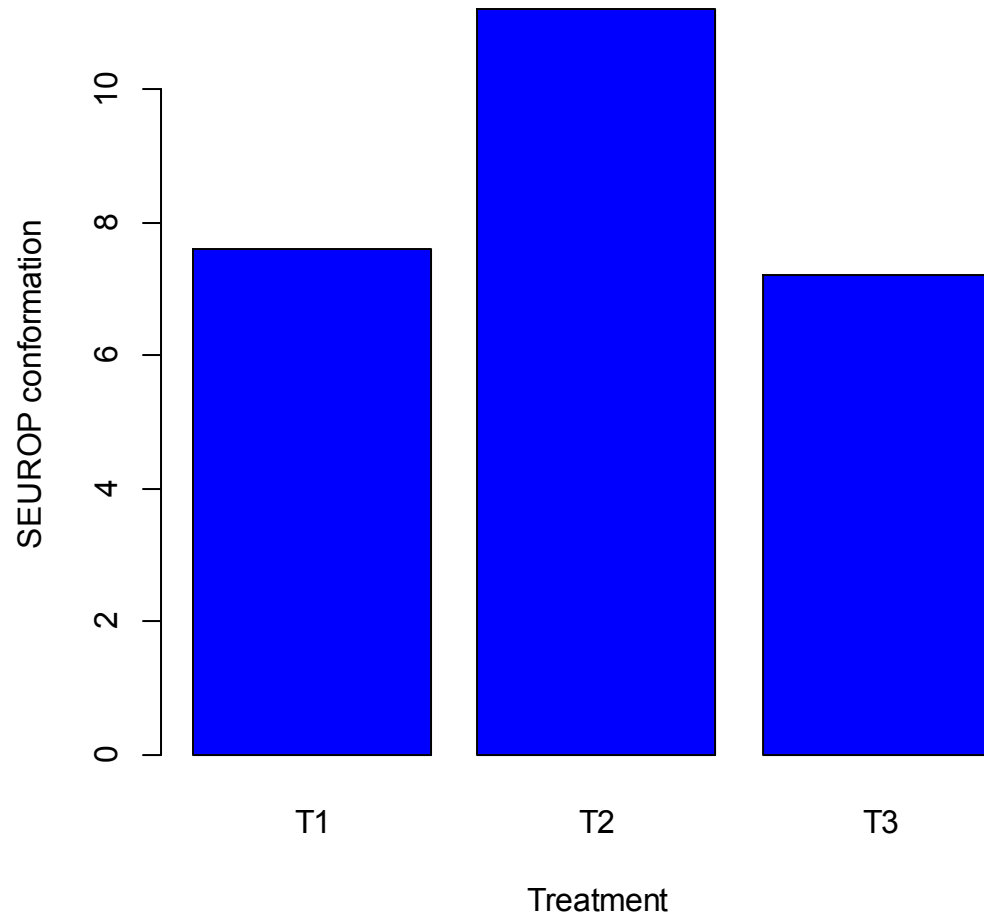
```
> M<-tapply(CONF, TRT, length)
> P<-tapply(CONF, TRT, mean)
> R<-tapply(CONF, TRT, sd)
> cbind(N=M, Mean=P, Std.dev=R)
```

	N	Mean	Std.dev
T1	10	7.6	0.9660918
T2	10	11.2	1.6193277
T3	10	7.2	1.4757296



## Working ANOVA with R (5b)

```
> barplot(P, xlab="Treatment", ylab="SEUROP conformation",  
+ col="blue")
```



# Working ANOVA with R (5c)

A function to include SE in the graphic (balanced design)

#Standard errors in a graphic (seBars function) from R-book p 216

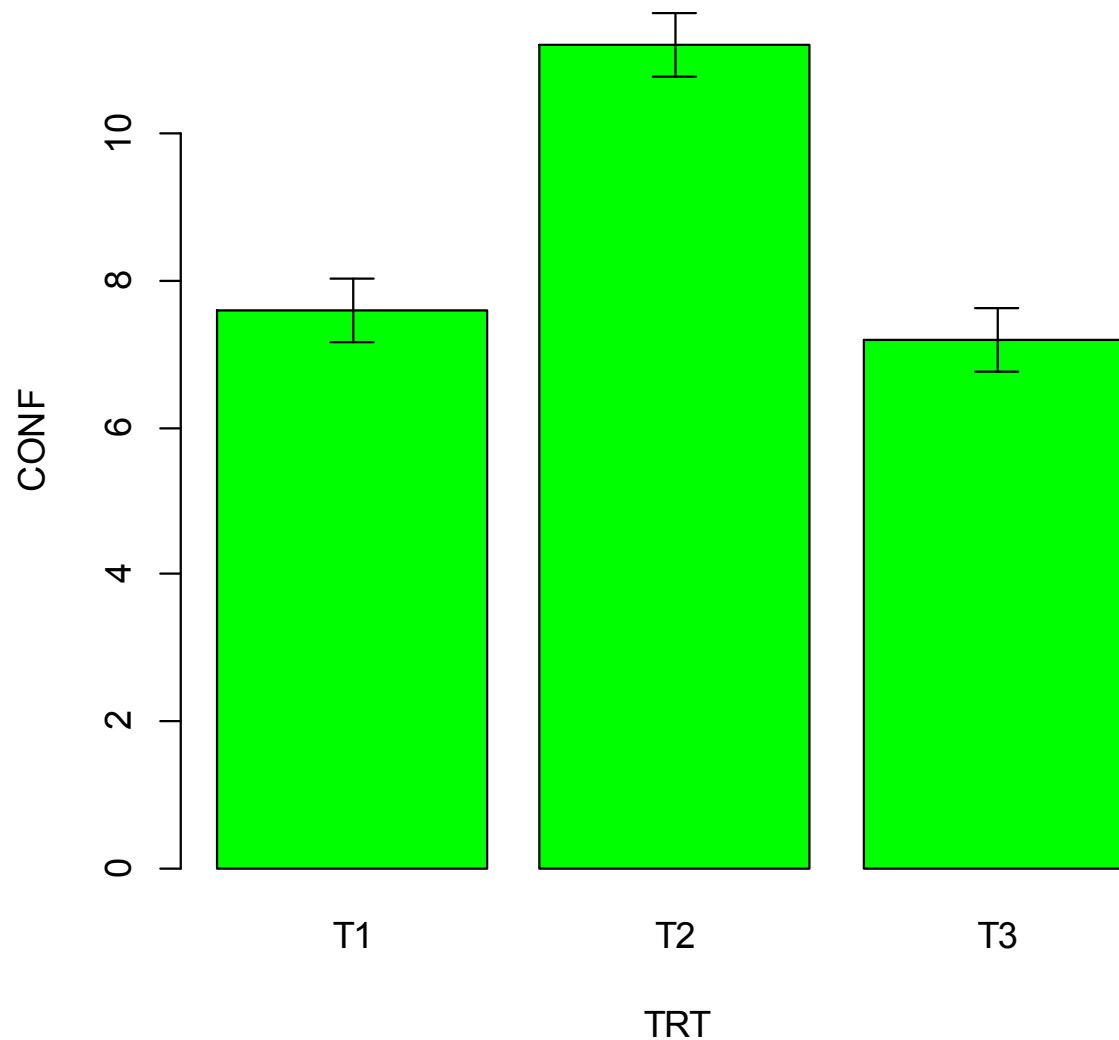
```
seBars<-function(x,y){
  model<-lm(y~factor(x))
  reps<-length(y)/length(levels(x))
  sem<-summary(model)$sigma/sqrt(reps)
  m<-as.vector(tapply(y,x,mean))
  upper<-max(m)+sem
  nn<-as.character(levels(x))
  xs<-barplot(m,ylim=c(0,upper),names=nn,col="green",
    ylab=deparse(substitute(y)),xlab=deparse(substitute(x)))
  for (i in 1:length(xs)){
    arrows(xs[i],m[i]+sem,xs[i],m[i]-sem, angle=90,code=3,length=0.1)}
}
```

seBars(TRT,CONF) # This executes the function for TRT and CONF

To get the 95% confidence intervals in the error bars write in this case

**sem\*qt(.975,10)** instead of **sem**

## Working ANOVA with R (5d)



## Working ANOVA with R (6)

```
> CONFBEEF.TUKEY<-TukeyHSD(CONFBEEF.AOV,"TRT")  
> CONFBEEF.TUKEY
```

**Tukey** multiple comparisons of means  
95% family-wise confidence level

Fit: aov(formula = CONF ~ TRT, data = CONFBEEF)

\$TRT

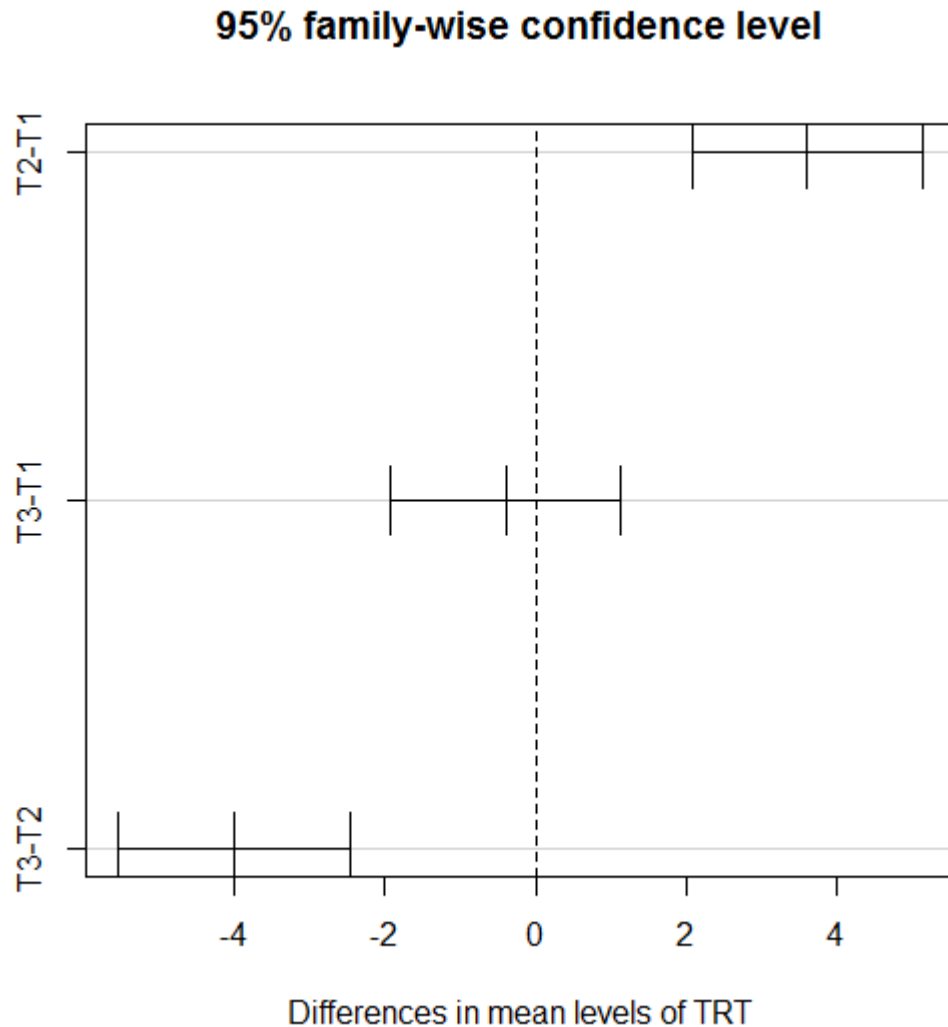
	diff	lwr	upr	p adj
T2-T1	3.6	2.067122	5.132878	0.0000099
T3-T1	-0.4	-1.932878	1.132878	0.7956248
T3-T2	-4.0	-5.532878	-2.467122	0.0000018

Treatment T2 differs significantly from treatments T1 and T3, which do not differ between them.

Observe that the confidence interval of the difference T3-T1 does not include 0.

# Working ANOVA with R (7)

```
> plot(CONFBEEF.TUKEY)
```



This is a graphic way to see the results in the previous slide. Only the confidence interval of T3-T1 overlaps 0 (i.e., their means do not differ).

# Working ANOVA with R (8)

```
> library(agricolae)
> HSD.test(CONFBEEF.AOV, "TRT")
```

Study:

HSD Test for CONF

Mean Square Error: 1.911111

TRT, means

	CONF	std.err	r	Min.	Max.
T1	7.6	0.3055050	10	6	9
T2	11.2	0.5120764	10	9	14
T3	7.2	0.4666667	10	5	9

alpha: 0.05 ; Df Error: 27

Critical Value of Studentized Range: 3.506426

Honestly Significant Difference: 1.532878

Means with the same letter are not significantly different.

Groups, Treatments and means

a	T2	11.2
b	T1	7.6
b	T3	7.2

## Pre-planned comparisons

Multiple comparison is the more frequent alternative. But we can be interested a priori in a limited number of comparisons.

Imagine that T2 would have been a standard diet and T1 and T3 experimental diets. We can be interested in testing if the experimental diets differ from the standard diet first, and later if differences exist between the two experimental diets.

This type of comparisons are called **contrasts**. Contrasts are linear functions of the solution's vector, in which the sum of coefficients (weighted by their sample size) must be 0. Two contrasts ( $c_{ij}$  and  $c_{ik}$ ) are **orthogonal** (independent) when satisfy:

$$\sum_{i=1}^t n_i c_{ij} c_{ik} = 0$$

If  $n_i$  is the same for all groups, this factor can be ignored. Generally, a design with  $t$  levels in a factor can be partitioned to a  $(t-1)$  orthogonal contrasts. It can be shown that orthogonal contrasts control Type I error.

# Working ANOVA with R (9)

We can use **contrasts** to test the difference between two or several groups (or combinations of them), written after the **anova** calculus.

**Contrasts** are linear combinations of levels that add to 1.

```
> contrasts(TRT) <- cbind(c(-0.5, 1, -0.5), c(1, 0, -1))
> round(crossprod(contrasts(TRT)), 2)
```

```
      [,1] [,2]
[1,]  1.5   0
[2,]  0.0   2
```

Independent contrasts

This tests the orthogonality of contrasts. Non diagonal elements must be 0.

```
> summary(CONFBEEF.AOV, split=list(TRT = list ("T2 vs
(T1+T3)/2" = 1, "T1 vs T3" = 2)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
TRT	2	97.07	48.53	25.395	6.25e-07	***
TRT: T2 vs (T1+T3)/2	1	96.27	96.27	50.372	1.25e-07	***
TRT: T1 vs T3	1	0.80	0.80	0.419	0.523	
Residuals	27	51.60	1.91			

---

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

T2 is different from the mean of T1 and T3; T1 and T3 do not differ.



# Power in the One-way ANOVA

The **power** of a test is the probability that a false null hypothesis is correctly rejected or a true difference is correctly declared different.

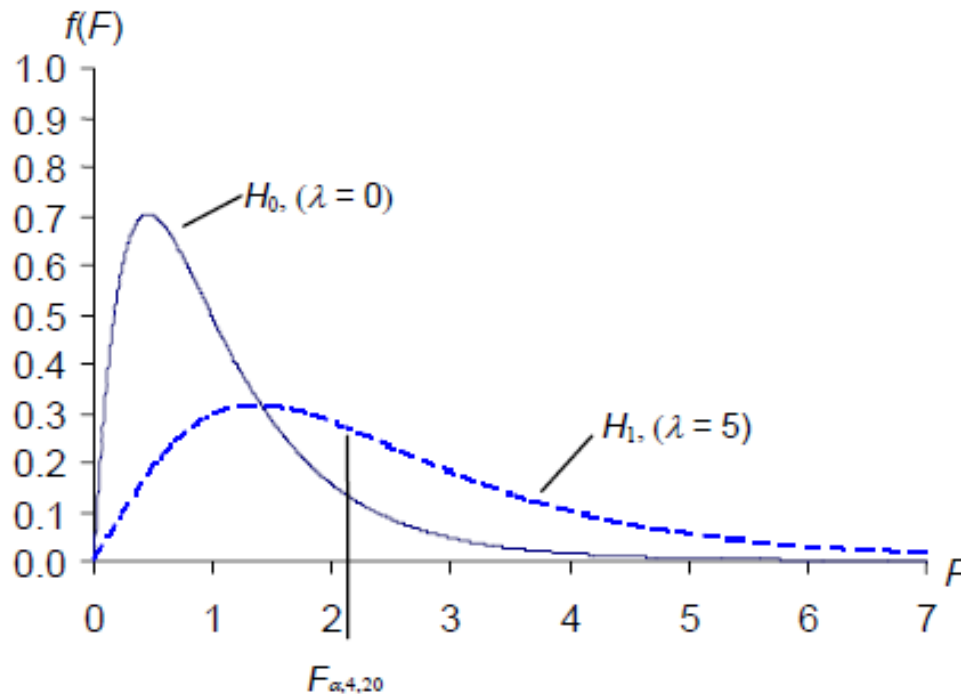
Under the  $H_0$ , the  $F$  statistic has a central  $F$  distribution with  $(t-1)$  and  $(tn-t = N-t)$  degrees of freedom. When at least one treatment effect is nonzero, the  $F$  test statistic follows a non-central  $F$  distribution with noncentrality parameter  $\lambda = SS_B / MS_W$ , and degrees of freedom as before. Then

$$\text{Power} = P(F > F_{\alpha, t-1, N-t} = F_{\beta})$$

using a non-central  $F$  distribution for  $H_1$ . The figure in the next slide represents the relationship between significance and power.

1. If a more stringent  $\alpha$  is chosen (critical value shifted to the right), the power will decrease.
2. A larger  $SS_B$  and a smaller  $MS_W$  means larger  $\lambda$ , the noncentrality curve is shifted to the right and this augments the area under this curve (power) to the right of the critical value.

# Relationship between significance and power



**Figure 11.4** Significance and power of the  $F$  test. Under  $H_0$  the  $F$  statistic has a central  $F$  distribution and under  $H_1$  it has a noncentral  $F$  distribution. The distributions with 4 and 20 degrees of freedom and noncentrality parameters  $\lambda = 0$  and 5 are shown. The critical value for an  $\alpha$  level of significance is  $F_{\alpha,4,20}$ . The area under the  $H_0$  curve to the right of the critical value is the level of significance ( $\alpha$ ). The area under the  $H_1$  curve to the right of the critical value is the power ( $1 - \beta$ ). The area under the  $H_1$  curve on the left of the critical value is the type II error ( $\beta$ ).

(Kaps and Lamberson, 2004)

# Power of our example

Following the example of K&L, p. 229, we can write in R:

```
> #Power in a CRD-one way ANOVA
> ACONFBEEF <- anova(lm(CONF~TRT, CONFBEEF))
> DFB <- ACONFBEEF[["Df"]][1]
> DFW <- ACONFBEEF[["Df"]][2]
> SSB <- ACONFBEEF[["Sum Sq"]][1]
> MSW <- ACONFBEEF[["Mean Sq"]][2]
> LAMBDA=SSB/MSW; ALPHA=0.05
> FCRIT=qf(1-ALPHA,DFB,DFW)
> POWER=1-pf(FCRIT,DFB,DFW,LAMBDA)
> cbind(Sig.level=ALPHA, DF.between=DFB, DF.within=DFW,
+ POWER=POWER)
```

Observe how particular values of the ANOVA table are extracted

	Sig.level	DF.between	DF.within	POWER
[1,]	0.05	2	27	0.9999942

**This value is bigger than the 0.8 usually required.**

The power of the tests was very high because of the differences among means of treatments (*TRT MS*) were much higher than the differences between observations of the same group (*Residual MS*).

## Power of our example (easier)

```
> power.anova.test(groups=3, n=10,  
between.var=48.533, within.var=1.911,  
sig.level=0.05)
```

Balanced one-way analysis of variance power calculation

```
groups = 3  
n = 10  
between.var = 48.533  
within.var = 1.911  
sig.level = 0.05  
power = 1
```

NOTE: n is number in each group

## Sample size in a One-way ANOVA (programme)

In a previous lesson, the number of replications necessary for a test of difference between two means was given. With more than two means, the level of significance must be adjusted for multiple comparisons.

An alternative is to compute sample size from power calculations:

```
> for (n in 2:20) {  
+   DFW[n] <- (DFB+1)*(n-1)  
+   FCRIT[n]<-qf(1-ALPHA,DFB,DFW[n])  
+   POWER[n]<-1-pf(FCRIT[n],DFB,DFW[n],LAMBDA)  
+     if(POWER[n]>=.80 && n<10){  
+       print(paste("n =", as.numeric(n), " ", "Power =",  
round(POWER[n], digits=7)), quote=FALSE)  
+     }  
+     else if(POWER[n]>=.80 && n>=10){  
+       print(paste("n =", as.numeric(n), " ", "Power =",  
round(POWER[n], digits=7)), quote=FALSE)  
+     }  
+ }
```

## Sample size in a One-way ANOVA (output)

[1]	n = 2	Power = 0.9378067
[1]	n = 3	Power = 0.9984039
[1]	n = 4	Power = 0.999802
[1]	n = 5	Power = 0.9999425
[1]	n = 6	Power = 0.9999741
[1]	n = 7	Power = 0.9999851
[1]	n = 8	Power = 0.9999901
[1]	n = 9	Power = 0.9999927
[1]	n = 10	Power = 0.9999942
[1]	n = 11	Power = 0.9999953
[1]	n = 12	Power = 0.999996
[1]	n = 13	Power = 0.9999965
[1]	n = 14	Power = 0.9999968
[1]	n = 15	Power = 0.9999971
[1]	n = 16	Power = 0.9999974
[1]	n = 17	Power = 0.9999975
[1]	n = 18	Power = 0.9999977
[1]	n = 19	Power = 0.9999978
[1]	n = 20	Power = 0.9999979

The required sample size (= 2) is actually a very low number, but it was expected if we remember that RMSE was very low and the differences among groups very high

# A matrix view of ANOVA (1)

Suppose we have the following data corresponding to the birth weight (kg) of the progeny of three sires:

S1	S2	S3
45	32	35
47	40	37
46		39

$$\begin{bmatrix} 45 \\ 47 \\ 46 \\ 32 \\ 40 \\ 35 \\ 37 \\ 39 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \end{bmatrix}$$

Incidence or design matrix

One way ANOVA design

$$y_{ij} = \mu + s_i + \varepsilon_{ij}$$

Fixed effect

In matrix terms

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

**Fixed model,**  
although  $\boldsymbol{\varepsilon}$  is  
random

We are interested in estimating the  
value of the elements of vector  $\mathbf{b}$

## A matrix view of ANOVA (2)

Normal equations:  $\mathbf{X}'\mathbf{X}\hat{\mathbf{b}} = \mathbf{X}'\mathbf{y} \Rightarrow \hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 3 & 2 & 3 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 45 \\ 47 \\ 46 \\ 32 \\ 40 \\ 35 \\ 37 \\ 39 \end{bmatrix} = \begin{bmatrix} 321 \\ 138 \\ 72 \\ 111 \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix}$$

$$\begin{array}{l} \text{Total number} \rightarrow \\ \text{Obs. Sire 1} \rightarrow \end{array} \begin{bmatrix} 8 & 3 & 2 & 3 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \end{bmatrix} = \begin{bmatrix} 321 \\ 138 \\ 72 \\ 111 \end{bmatrix} \Rightarrow \begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \end{bmatrix} = \begin{bmatrix} 8 & 3 & 2 & 3 \\ 3 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 321 \\ 138 \\ 72 \\ 111 \end{bmatrix}$$

But the determinant of  $\mathbf{X}'\mathbf{X} = 0$ : singular, it has not an inverse.



## A matrix view of ANOVA (3)

A convenient generalized inverse of  $\mathbf{X}'\mathbf{X}$  is  $\mathbf{G}$ :

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}$$

Observe that we can take the  $\mathbf{A}$  square submatrix with rank = rank( $\mathbf{X}'\mathbf{X}$ ) from different positions of  $\mathbf{X}'\mathbf{X}$ , and then put it in the corresponding position in  $\mathbf{G}$ . This can give different generalized inverses.

And then,

$$\begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 321 \\ 138 \\ 72 \\ 111 \end{bmatrix} = \begin{bmatrix} 0 \\ 138/3 \\ 72/2 \\ 111/3 \end{bmatrix} \leftarrow$$

This is in fact a solution vector for  $\mathbf{b}$ , not an estimator of  $\mathbf{b}$ . The vector depends upon the generalized inverse calculated.

## A matrix view of ANOVA (4)

It is possible to obtain a vector of predicted values from  $\hat{\mathbf{b}}$ ,  $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ . The square sum of the deviations of observed  $\mathbf{y}$ 's to their predicted values is the error or residual sum of squares ( $SS_e$ ):

$$SS_e = \sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2 = \cdots = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} \quad (\text{Searle, 1982})$$

**This sum of squares is invariant to any generalized inverse that we use to estimate  $\hat{\mathbf{b}}$**

Furthermore, the total (corrected) sum of squares is:

$$SS_T = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = \cdots = \mathbf{y}'\mathbf{y} - N\bar{y}^2$$

Then, we can write the following ANOVA table:

Source of Variation	d.f.	Sums of Squares	F
Between groups (Fitting of the model)	$t-1$	$SS_B = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - N\bar{y}^2$	$\frac{SS_B/(t-1)}{SS_e/(N-t)} = \frac{MS_B}{MS_e}$
Within groups (error)	$t(n-1)$	$SS_e = \mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y}$	
Total	$tn-1$	$SS_T = \mathbf{y}'\mathbf{y} - N\bar{y}^2$	$(N = tn)$

**This is the way in which statistical packages like R work**

## A matrix view of ANOVA (5)

```
> birth.aov<-aov(BIRTH.W~SIRE)
```

```
> anova(birth.aov)
```

Analysis of Variance Table

Response: BIRTH.W

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SIRE	2	166.88	83.438	9.933	0.01813 *
Residuals	5	42.00	8.400		

```
> summary.lm(aov(BIRTH.W~SIRE))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	46.000	1.673	27.490	1.19e-06 ***
SIRES2	-10.000	2.646	-3.780	0.0129 *
SIRES3	-9.000	2.366	-3.803	0.0126 *

Residual standard error: 2.898 on 5 degrees of freedom  
 Multiple R-squared: 0.7989, Adjusted R-squared: 0.7185  
 F-statistic: 9.933 on 2 and 5 DF, p-value: 0.01813

Observe that the solutions for  $\mathbf{b}$  in this table are different from the solutions obtained with our generalized inverse. R equals to 0 the first level of each factor.

## References

Kaps M., Lamberson W.R. 2004. *Biostatistics for the Animal Sciences*. CABI, Wallingford.