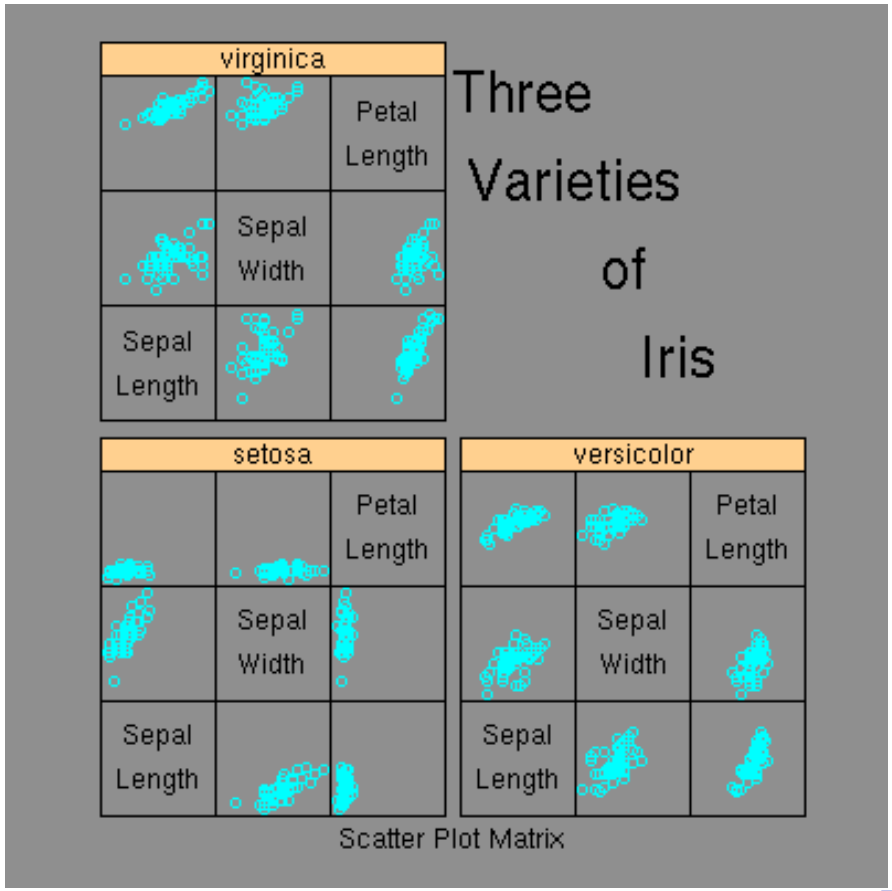
ANalysis Of VAriance & Co, with R

Olivier Flores



Background I

The **AN**alysis **O**f **V**ariance studies relationships between a *quantitative observed* variable (*response*, y: height, weight,...) and one or more *qualitative explicative* variable (*factors*, F,...: species, treatment...).

$\rightsquigarrow$ Aims at predicting the *mean* response across *classes* defined by (combinations of) the factor(s):

- 1 factor: One-Way Anova,
- 2 factors: Two-Way Anova

Background II

- ANOVA is a special case of *linear modelling*:

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \ i = 1, \ldots, n$$

In matrix terms, this would be written:

$$y = \mathbf{X}\beta + \epsilon,$$

where the y is the response vector, $\mathbf{X}$ is the *model matrix* or *design matrix*, β is the vector of *model parameters*, ϵ is the vector of *error terms* or *residuals*

Background III

- Example:
  One-Way ANOVA: Suppose F has p levels, $L_F = L_1, \ldots, L_p$

$$\begin{aligned} y &= a_{L_F} + \epsilon && (p \text{ parameters}) \\ \text{or } y &= \mu + \alpha_{L_F} + \epsilon && (p+1 \text{ parameters}) \end{aligned}$$

- Data are structured as:

y	F_1	$\ldots$
$\vdots$	$\vdots$	$\vdots$

- Requires at least 5 observations per class

Hypotheses

The ANOVA relies on hypotheses regarding statistical *units* (= observations) and classes:

- units are independent
- error is normally distributed in the classes
- variances on error are equal (σ^2)

Principle I

Decompose **total** variance, S^2 (measured on all classes) in **inter-classes** (= explained, V_E) and **intra-classes** (= residual, V_R) variances:

- $\overline{y_k} = \frac{1}{n_k} \sum_{i_k=1}^{n_k} y_{i_k}$, empirical mean of class k, with n_k number of observations in class k ,
- $V_k = \frac{1}{n_k} \sum_{i_k=1}^{n_k} (y_{i_k} - \overline{y_k})^2$, empirical variance of class k
- $\overline{X}$, empirical mean of total sample.

Then:

- $V_E = \sum_{k=1}^{p} \frac{n_k}{n} V_k$, with n total number of observations $(= \sum_{k=1}^{p} n_k)$,
- $V_R = \sum_{k=1}^{p} \frac{n_k}{n} (\overline{X_k} - \overline{X})^2$,

We have:

$$S^2 = V_E + V_R$$

Principle II

Given hypotheses regarding normality and homosedasticity,
V_E and V_R follow $\mathcal{X}^2$ laws, and their ratio follows a Fisher law:

- $ESS = n\frac{V_E}{\sigma^2} \sim \mathcal{X}^2(n-p)$
- $RSS = n\frac{V_R}{\sigma^2} \sim \mathcal{X}^2(p-1)$
- $r = \frac{V_E/(p-1)}{V_R/(n-p)} \sim \mathcal{F}(p-1, n-p)$

ESS: explained sums of squares,

RSS: residual sum of squares

When r is large, differences are large across classes.

Principle III

The **null hypothesis** tested by the ANOVA is the equality of empirical means:

$$\mathcal{H}_0 : \overline{X_1} = \cdots = \overline{X_k}$$

against alternative:

$$\mathcal{H}_1 : \text{there are at least two different means}$$

The analysis rejects $\mathcal{H}_0$ when the weighted ratio r is significantly large compared to the quantiles of the law $\mathcal{F}(p-1, n-p)$.

Principle IV

If $\mathcal{H}_0$ is rejected, two-by-two comparison of means (multiple comparisons):

- **Tukey test**:
  $\mapsto$ for balanced designs (= same number of observations in all classes) or mildly unbalanced designs (correction included).
  Also a plot method to visualize the results:
  `> plot(TukeyHSD(`x`,...))`, where x is a fitted ANOVA model.

- **t-tests with corrections for multiple comparisons**:
  `> pairwise.t.test(`y`, `*factor*`, `*p.adjust.method*` = method,...)`
  `method` can take several values, including `"bonferroni"`.
  The default is `"holm"` which should be preferred.

  If the response is a **proportion**:
  `> pairwise.prop.test(`y`, `*factor*`, `*p.adjust.method*` = method,...)`

Verify hypotheses I

1. Normality:
   Graphical diagnostics is sufficient
   > `qqnorm(data)`
   > `qline(data)`

   If not enough, try testing with:
   - **Kolmogorov-Smirnov** test: avoid, `ks.test` in R,
   - **Shapiro-Wilks** test: better but not very good when ties in data (measures with same value): `shapiro.test` in R,
   - **D'Agostino-Pearson** test: not implemented in R but routine available on internet

Verify hypotheses II

If non-normal samples, try non-parametric methods (independent of assumptions regarding distributions).

Within R, routines available in the packages:

- `pairwise.wilcox.test`: produces pairwise Wilcoxon test on ranks,
- `npmc`: Non-Parametric Multiple Comparison, provides simultaneous rank test procedures for the one-way layout,
- `multcomp`: Mutliple Comparisons.

Verify hypotheses III

2. Homosedasticity:

   If not verified, variable transformation can help to stabilize the variance.

$$
\begin{aligned}
y &\rightarrow \log(y) \\
y &\rightarrow \sqrt{(y)} \\
y &\rightarrow \arcsin(\sqrt{(y)}) \ (\textit{for proportions})
\end{aligned}
$$

Two-Way ANOVA I

- Aim: predict a quantitative variable (*response*, y) by two qualitative variables (*factors*, A and B),
- Hypotheses and principle similar to the One-Way ANOVA.

  In addition:
  - Additivity and interaction
  - Orthogonality

Two-Way ANOVA II

1. Two models are possible:
   - Additive model: $y_i = \mu + a_i + b_j + \epsilon$
   - Model with interaction: $y_i = \mu + a_i + b_j + c_{ij} + \epsilon$

   Estimation of the interaction term requires *repetition* in the data (replicates).

2. Othogonality:
   An orthogonal design is advantageous:
   - the sums of squares are additive,
   - the order of the factors is not important for the estimation
   - parameters are not correlated

   . . . which is not true if the design is not orthogonal.

   A **sufficient condition of orthogonality** is: $n_{ij} = \text{constant}$

Two-Way ANOVA III

Condition for orthogonality:

- n_{ij} : number of observations in the ith class of A and jth class of B,
- $n_{i\bullet} = \sum_j n_{ij}$: number of observations in the ith class of A,
- $n_{\bullet j} = \sum_i n_{ij}$: number of observations in the jth class of B,
- $n = \sum_{i,\ j} n_{ij}$: total number of observations

The experiment design is orthogonal $\Leftrightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$

Example I

```
> library(datasets)
> summary(fm1 <- aov(breaks ~ wool + tension,
+ data = warpbreaks))
            Df Sum Sq Mean Sq F value   Pr(>F)
wool         1  450.7   450.7  3.3393 0.073614 .
tension      2 2034.3  1017.1  7.5367 0.001378 **
Residuals   50 6747.9   135.0
--
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Marginal effect of factor *wool* and significant effect of factor *tension*

Example II

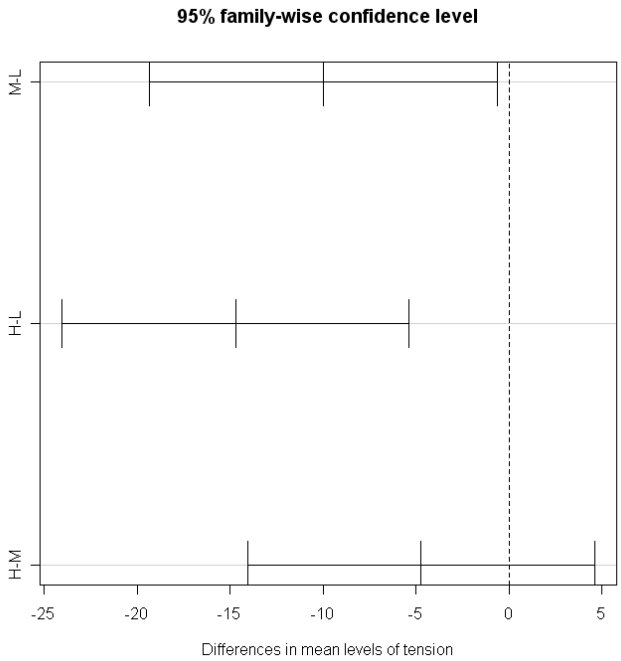
```
> TukeyHSD(fm1, "tension", ordered = TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = breaks ~ wool + tension, data = warpbreaks)

$tension
         diff        lwr      upr     p adj
M-H  4.722222 -4.6311985 14.07564 0.4474210
L-H 14.722222  5.3688015 24.07564 0.0011218
L-M 10.000000  0.6465793 19.35342 0.0336262
```

Example III

```
> plot(TukeyHSD(fm1, "tension"))
```



Significant differences between classes M and L, and H and L at 95%.
No significant difference between classes H and M.

ANalysis of COVAriance I

General objective:
Predict a quantitative variable y with a qualitative variable A and a quantitative variable x:

- additive model: $y_i = \mu + a_i + bx + \epsilon$
- interaction model: $y_i = \mu + a_i + b_i x + \epsilon$

Several purposes:

1. Increase precision in an experiment
2. Control for an extraneous variable
3. Compare regressions among several groups

ANalysis of COVAriance II

Orthogonality is verified if x takes the same values for each class in A.

If normality is not verified, possible to try a *Nonparametric analysis of covariance*:
`sm.ancova` from package **sm** (CRAN)

Multivariate Analysis of variance and Covariance I

General objective:
Investigate the effects of one (or more) factor(s) on a multivariate response:

> y is a **matrix** of dependent variables,
> A is an explicative factor.

The MANOVA answers the question:

Are there differences in the vectors of mean responses in y across classes of A?

A significative multivariate test indicates *some* effect of the factor on the vector of means per class.

Multivariate Analysis of variance and Covariance II

Advantage over ANOVA:

$\longrightarrow$ limited *Type-1*[1] error compared to univariate ANOVA conducted independently.

Consequently, MANOVA can reveal differences not discovered by ANOVA tests

Multivariate Analysis of variance and Covariance III

Limits:

- Considering several dependent variables decreases the number of degrees of freedom in the analysis (less powerful tests),
- The dependent variable should be largely uncorrelated (if not, tests even less powerful)

Usually, MANOVA is followed by univariate ANOVA.

[1] Type 1- or α- or *false-positive*-error: rejecting the null hypothesis $\mathcal{H}_0$ whereas it is true

Type 2- or β- or *false-negative*-error: accepting the null hypothesis $\mathcal{H}_0$ whereas it is false

Models of analysis of variance in R I

Models are fitted with a unique function, *aov*:

```
aov(formula, data = NULL,...),
```

where:
- *formula* describes the model to be fitted,
- *data* is the data frame containing variables used in the formula (optional)

Models of analysis of variance in R II

1 Formulae in R

The model formulae specify the *response* and the *columns of the model matrix* (one parameter estimated per column).

- One-Way ANOVA:
  $y \sim A$,
- Two-Way ANOVA:

$y \sim A + B$	without interaction
$y \sim A + B + A:B$ $y \sim A * B$	with interaction term, $A:B$
$y \sim B\%in\%A$ $y \sim A/B$	with factor B nested in factor A

- ANCOVA:
  $y \sim A/x$: linear regressions of y on x within the levels of A

Models of analysis of variance in R III

The MANOVA in R is produced by a call to the function `manova`:

```
> manova(y~A,...)
# all arguments of aov can be passed to manova
```

Models formulae are similar to the ANOVA:

- *Multiple* factors can be specified (Two-way MANOVA,...).
- MANCOVA can be realized with the same function, with a mixture of quantitative and qualitative independent variables.

Other specifications are possible:

$y \sim A - 1$ specifies a model *without intercept* term.

...

Models of analysis of variance in R IV

2 Contrasts

For factors, the model matrix depends on *contrasts* specification.

By default, contrasts are not orthogonal in R.

⇒ **Set orthogonal contrasts prior to analysis**:

```
> op = options(contrasts = c("contr.helmert",
"contr.poly"))
```

Models of analysis of variance in R V

3 Attributes of statistical models in R

Depend on the class of the model, but some are common

- *coefficients*: estimated regression coefficients,
- *fitted values*,
- *residuals*,
- *df.residual*: residual degrees of freedom,
- ...

All attributes can be accessed by:

object_name$attribute_name,

or using generic functions.

Models of analysis of variance in R VI

4 Generic functions on statistical models

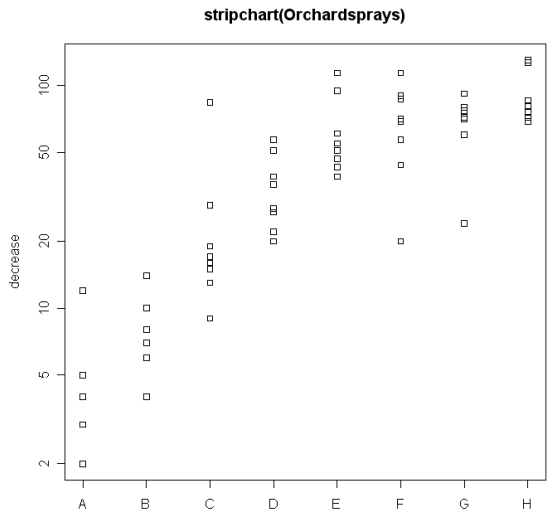
Used to display, extract or plot information:

- `coef, fitted, resid, formula`: extract coefficients, fitted values, residuals and model formula,
- `plot`: produces four plots for diagnostics
- `anova`: produce analysis of variance tables and **compare** models:
  `anova(model1, model2)`,
- `deviance`: residual sum of squares,
- `print`
- `summary`

Exploratory analysis I

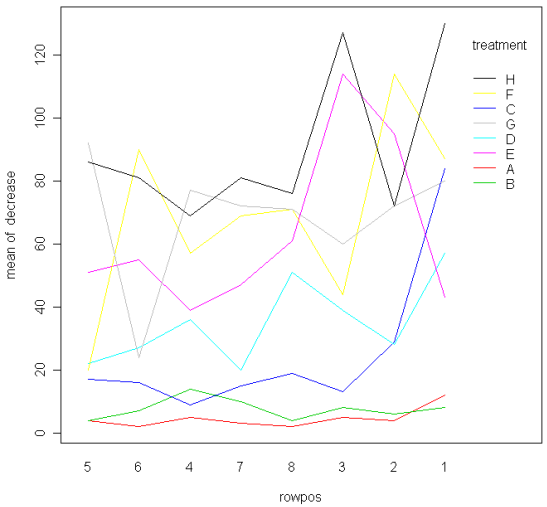
1. **Graphical procedures**
   - **Histogram**:
     `hist(x,...)`,
   - **Quantile-Quantile plot**:
     `qqlot(x,y)`; `qqnorm(y)`; `qqline(y)`,
   - **Boxplot**:
     `boxplot(X,...)`, or `boxplot(f,...)`,
     where X is a matrix of continuous variables, f a formula ($y \sim A$)

`stripchart` is an alternative for small data sets



Exploratory analysis II

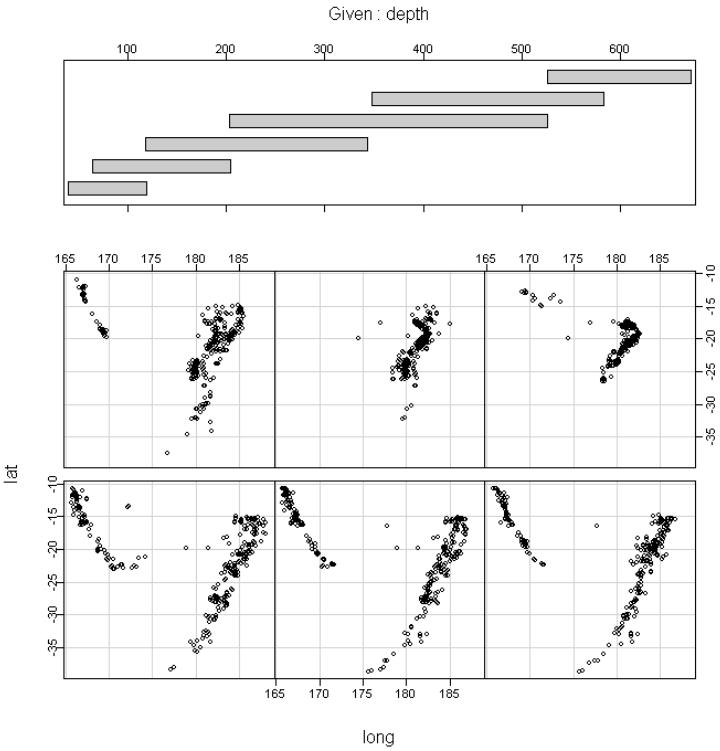
- **Interaction plot**:
  Useful in Two-Way ANOVA to investigate the presence of interaction between factors:
  `Interaction.plot(x.factor, trace.factor, response, fun = mean,...)`,
  - *x.factor*: a factor whose levels will form the x axis
  - *trace.factor* another factor whose levels will form the traces,
  - *response* a numeric variable giving the response,
  - *fun* the function to compute the summary (*mean* by default, *var*,...)



Non-parallel lines indicate some level of interaction

Exploratory analysis III

- **Coplot**:
  Useful in ANCOVA to display the response y as a function of the covariate x for each levels of A:
  `coplot(`$y \sim x|A$`,...)`



Also works with a numerical variable A, levels being replaced by intervals of A values

Exploratory analysis IV

2 Numerical procedures

- General statistical functions:
  `summary; mean; var;...`
- Numbers of observations per class:
  `> table(factor)`
  or crossed-tables:
  `> table(factor1, factor2)`
- Output of graphical functions:
  `> x=boxplot(f,...)`
- Statistics per class:
  The functions `by` and `aggregate` automatically compute statistics for subsets of data:
  `> by(data, INDICES, FUN,...)`
  `> aggregate(data, INDICES, FUN,...)`
  or combine functions *lapply* (or *sapply*) with *split*:
  `> sapply(split(y, A), mean)`

References

This presentation is largely inspired by the manual:

An Introduction to R.
Notes on R: A Programming Environment for Data Analysis and Graphics
Version 2.4.0, 2006-10-03
by W. N. Venables, D. M. Smith and the R Development Core Team

See also the book:

Introducing ANOVA and ANCOVA, a GLM approach
by Andrew Rutherford,
ISM series, SAGE Publications

Most of the graphical examples come from the help documentation on functions