UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Design and Analysis of Experiments

10 - Analysis of Variance

Version 2.12.2018b

Felipe Campelo

http://orcslab.cpdee.ufmg.br/

Graduate Program in Electrical Engineering

Belo Horizonte
October 2018

"The attempt to correlate all the known phenomena, and to bind them together into one consistent whole, led to the deduction of new facts, which, when duly tested by experiment, became parts of the growing body, and, themselves, opened up fresh questions, to be answered in their turn by experiment."

Hertha Ayrton
1854–1923
British engineer, mathematician and physicist

Image by George Doutsiopoulos:

https://www.behance.net/gallery/56913255/STEM-Epic-Heroes-Engineering-Heroes-(part-1)

Comparison of multiple means

Introduction

In the previous sections, we have (hopefully) developed a solid understanding of the main concepts associated with comparing the means of two groups;

There are many cases, however, in which one may want to perform inference about differences of the means of multiple populations;

We will develop the main concepts and ideas related with this kind of test by examining a simple example, related to a paper manufacturing operation.

Image: http://goo.gl/xYVW0M

Example: paper manufacturing

Problem definition

Tensile strength (TS) is an important characteristic for certain types of paper for industrial use;

A reasonable conjecture is that this characteristic is influenced by the kind of wood fiber used in the manufacturing process.

The process engineer wants to investigate whether four different wood fibers result in papers with relevant differences of TS, using a pilot plant as his experimental unit.

Example adapted from Montgomery & Runger (2010), Ch. 13.

Example: paper manufacturing

Problem definition

Suppose that the total budget allocated for the experiment allows only six production runs for each kind of wood fiber.

Under these specifications, the experiment has a single experimental *factor* (*wood fiber*) with $a = 4$ *levels* (fiber types *A*, *B*, *C* and *D*) and $n = 6$ *replicates* at each level.

The response variable will be the tensile strength of paper (measured, e.g., in kPa). The engineering team is interested in finding out whether any fiber type leads to an increase in the mean TS value of the paper.

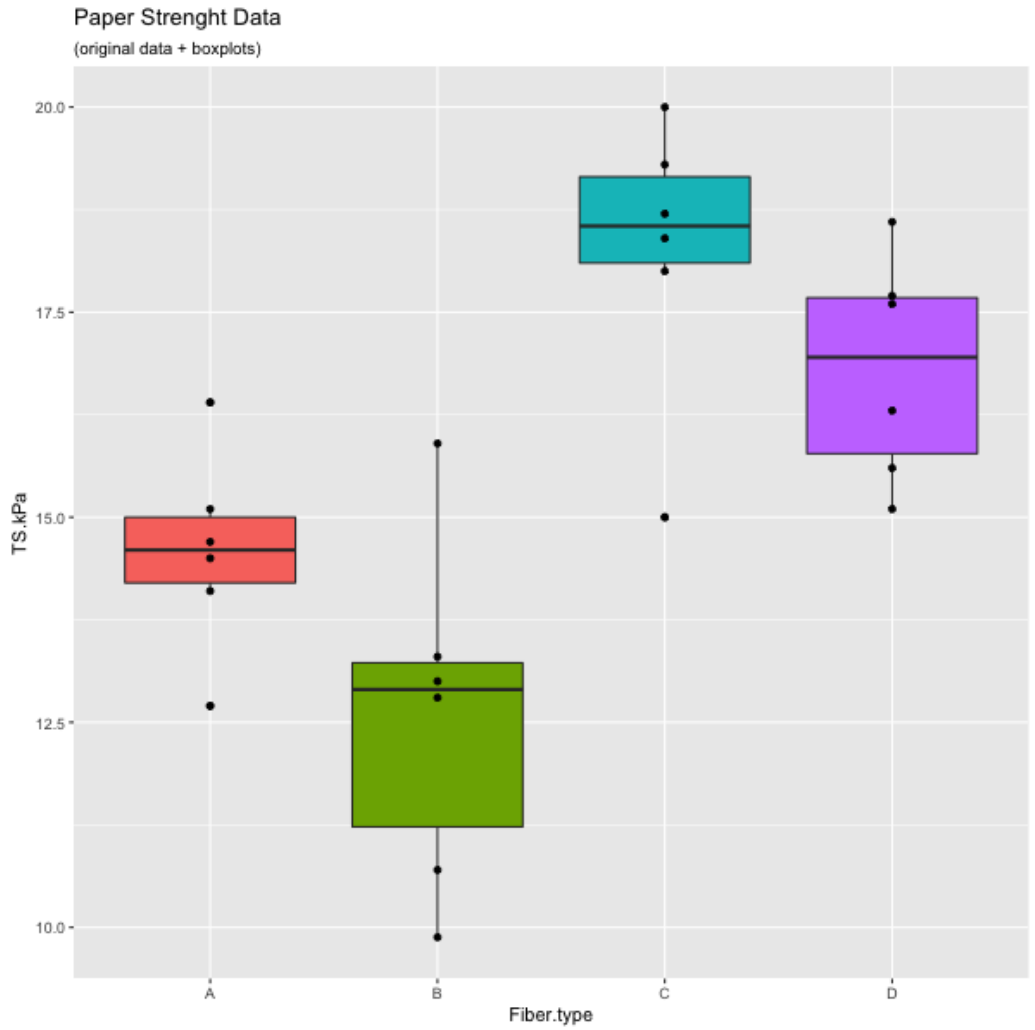
The minimum difference of practical meaning is defined as $5kPa$, and a reasonable upper estimate for the standard deviation of this process is $\hat{\sigma} = 6kPa$. Desired error levels are defined as $\alpha = 0.1$ and $\beta = 0.2$.

Example: paper manufacturing

Exploratory data analysis

```
> paper <- read.table(file   = "../data files/paper_strength.csv",
+                     header = TRUE,
+                     sep    = ",")

> library(ggplot2)
> ggplot(paper,
+     aes(x    = Fiber.type,
+         y    = TS.kPa,
+         fill = Fiber.type)) +
+  geom_boxplot() +
+  geom_point()
```



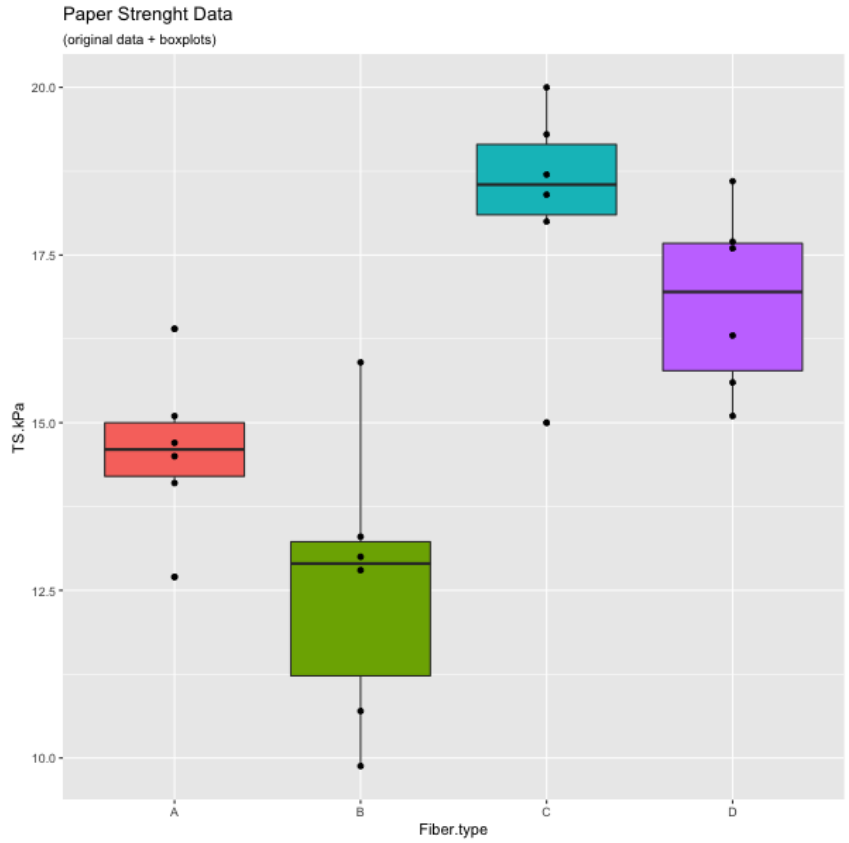
Example: paper manufacturing

Exploratory data analysis

The boxplot suggests the existence of differences among factor levels;

Besides, we can also observe a small variability in the spread of different levels; some suggestion of asymmetry in level *B*; and a possible outlier in level *C*.

These characteristics will need to be taken into account during the analysis.



Example: paper manufacturing

Statistical model

This data can be described by a linear statistical model of the form:

$$y_{ij} = \underbrace{\mu_i + \epsilon_{ij}}_{\text{Means model}} = \underbrace{\mu + \tau_i + \epsilon_{ij}}_{\text{Effects model}} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

where μ is the overall mean, τ_i represents the effect of the i-th level, and ϵ_{ij} is the residual (random error, or unmodeled variability);

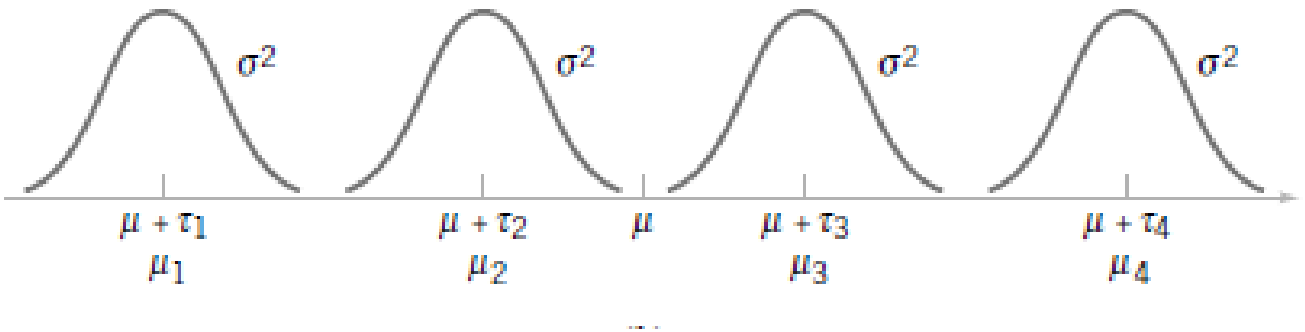
In the derivation of the statistical test for the existence of differences in the group means, we will employ the effects model, and initially consider a few assumptions about the residuals:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}, \quad \text{with } \epsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$$

Example: tensile strength

Statistical model

If these assumptions are correct, the populations are expected to be distributed as:



Since we are interested in testing our data for differences in the mean values of each population, the test hypotheses can be described as:

$$\begin{cases} H_0 : \tau_i = 0, \quad \forall i \in \{1, 2, \ldots, a\} \\ H_1 : \exists\, \tau_i \neq 0 \end{cases}$$

If data collection is performed in random order, under constant experimental conditions, we have a *completely randomized design*.

Image: Montgomery&Runger (2010), Ch. 13

The Fixed Effects Model

Definition

This approach to modeling the mean effects of specific factor levels is known as the *fixed effects model*;

This approach is appropriate to testing hypotheses in situations when factor levels are arbitrarily defined by the experimenter;

For these cases, the inference is made over the mean values for each level, and cannot be extended to similar levels that were not tested (e.g., other types of wood fiber);

Other situations may require different kinds of modeling, such as *random* or *mixed effects models*, but these will not be explored here.

The Fixed Effects Model

Development

As mentioned earlier, we will use the *effects model* for describing the development of the statistical test:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, a \\ j = 1, \ldots, n \end{cases}$$

where treatment effects are seen as deviations from the grand mean μ.

By construction, we have that:

$$\sum_{i=1}^{a} \tau_i = 0;$$

The Fixed Effects Model

Development

The total variability of the data can be expressed by the *total sum of squares*, which represents the sum of the squared deviations between each observation and the overall sample mean:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$$

With some relatively simple algebra, the SS_T can be divided into two terms, representing the within-group and the between-group variability:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2 = \underbrace{n\sum_{i=1}^{a}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2}_{SS_{Levels}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2}_{SS_E}$$

where $\bullet$ indicates the summation over an index, and $\bar{}$ indicates an averaging operation.

The Fixed Effects Model

Development

Dividing the sums of squares by their respective number of degrees of freedom yields quantities known as *mean squares*.

The relevant means squares for our test will be the *levels mean square* and the *residual mean square*:

$$MS_E = \frac{SS_E}{a(n-1)} \qquad MS_{Levels} = \frac{SS_{Levels}}{a-1}$$

The expected values of these quantities are:

$$E[MS_E] = \sigma^2 \qquad E[MS_{Levels}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

The Fixed Effects Model

Development

$$E[MS_E] = \sigma^2 \qquad E[MS_{Levels}] = \sigma^2 + \frac{n\sum_{i=1}^{a}\tau_i^2}{a-1}$$

Notice that MS_E is an unbiased estimator for the common variance of the residuals, while MS_{Levels} is biased by a term that is proportional to the squared values of the τ_i coefficients.

However, under H_0 we have that $\tau_i = 0$ for all i, that is, $E[MS_{Levels}] = E[MS_E] = \sigma^2$. *But only if the null hypothesis is true*.

The Fixed Effects Model

Development

It can be shown that, if H_0 is true, the statistic

$$F_0 = \frac{MS_{Levels}}{MS_E}$$

is distributed according to an F distribution with $a - 1$ degrees of freedom for the numerator and $a(n - 1)$ for the denominator. The usual notation is $F_{(a-1),a(n-1)}$

If H_0 is false, the expected value of MS_{Levels} is larger than that of MS_E, which results in larger values of F_0 and defines the critical region for our test:

Reject H_0 at the α significance level if
$f_0 > F_{1-\alpha;(a-1),a(n-1)}$

Example: paper manufacturing

Computational analysis

```
> my.model <- aov(TS.kPa ~ Fiber.type,
+                 data = paper)
>
> summary.aov(my.model)

Df Sum Sq Mean Sq F value   Pr(>F)
Fiber.type   3 110.77   36.92   13.62 4.56e-05 ***
Residuals   20  54.24    2.71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *ANOVA table* provides information on the sources of variation, together with their corresponding *d.o.f.*, sums of squares and mean square values. The table also informs the values of the test statistic and the corresponding p-value of the test ($Pr(> F)$).

In this case, the p-value ($p = 4.56 \times 10^{-5}$) suggests the rejection of the null hypothesis in favor of the alternative. But what does that mean?

Example: paper manufacturing

Computational analysis

Recall the null and alternative hypotheses for the ANOVA:

$$\begin{cases} H_0 : \tau_i = 0, \quad \forall i \\ H_1 : \exists\, \tau_i \neq 0 \end{cases}$$

The rejection of the null hypothesis leads to the conclusion that *there is at least one level with an effect significantly different from zero*. But which one?

For this analysis to be complete, we still need to answer two questions:

- Can we verify the assumptions of the test?
- Which means are different from which, and by how much?

Assumptions

Model validation

As mentioned earlier, the ANOVA model is based on three assumptions on the behavior of the residuals:

- *Independence*;
- *Homoscedasticity*, i.e., equality of variances across groups;
- *Normality*;

The residuals of the model can be easily obtained as:

$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - (\hat{\mu} + \hat{\tau}_i) = y_{ij} - \bar{y}_{i\bullet}$$

or extracted directly from the fitted object in R using `my.model$residuals`.

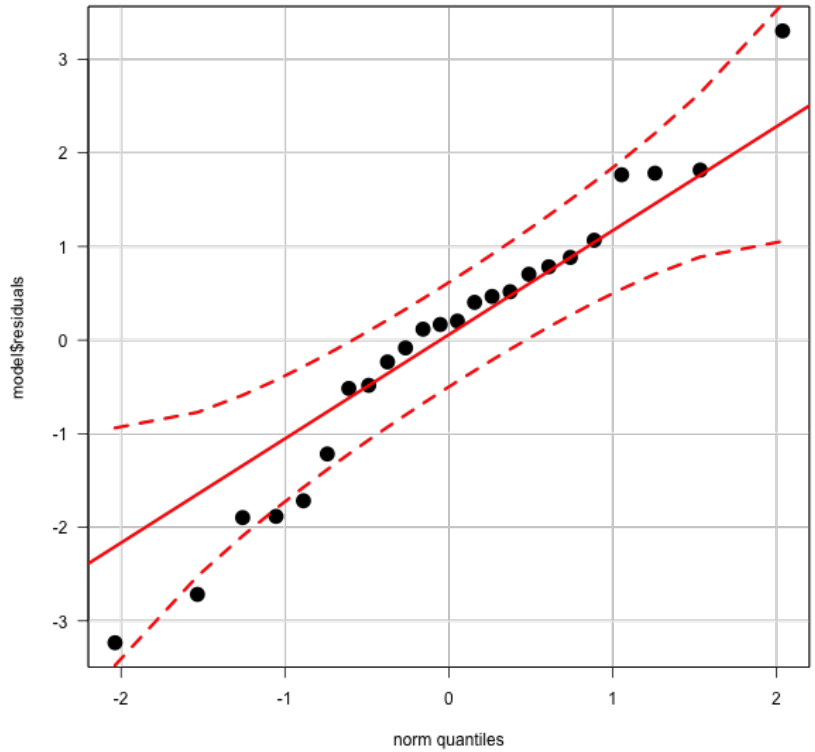
Assumptions

Model validation

The normality assumption can be tested using the Shapiro-Wilk test coupled with a normal QQ plot of the residuals.

```
> shapiro.test(model$residuals)
Shapiro-Wilk normality test
data:  my.model$residuals
W = 0.9722, p-value = 0.7225

> library(car)
> qqPlot(my.model$residuals,
pch = 16,
lwd = 3,
cex = 2,
las = 1)
```



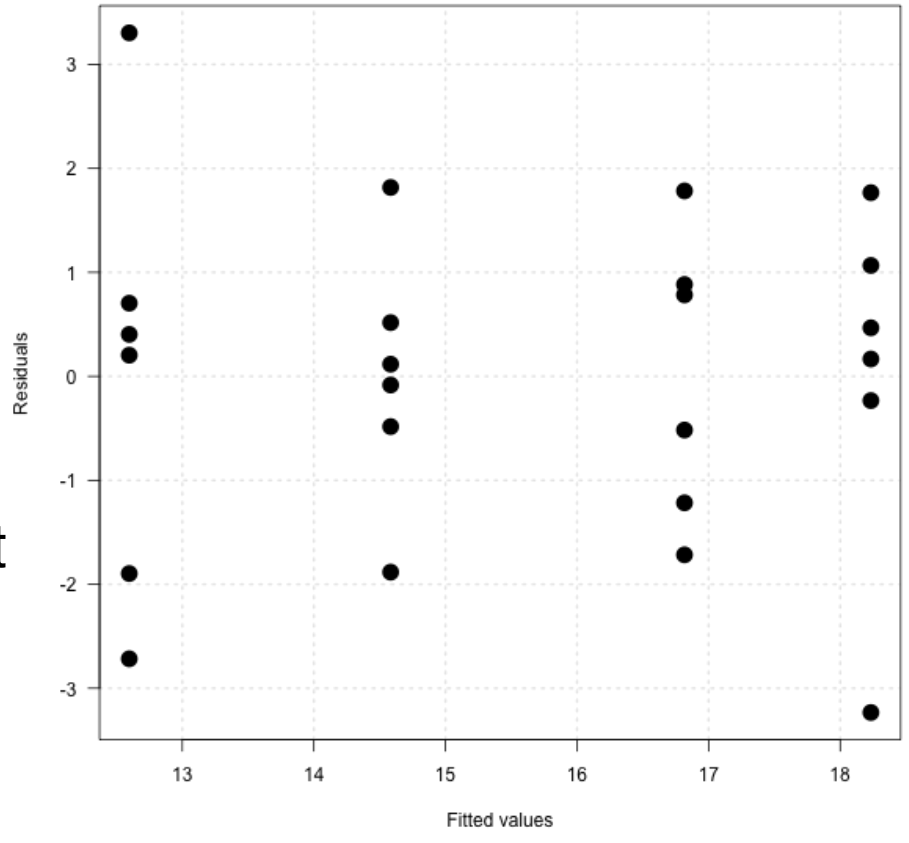
The ANOVA is relatively robust to moderate violations of normality, as long as the other assumptions are verified (or the sample size is large enough).

Assumptions

Model validation

The homoscedasticity assumption can be verified by the Fligner-Killeen test, together with plots of residuals by fitted values:

```
> fligner.test(TS_kPa~Hardwood, data = paper)
Fligner-Killeen test of homogeneity of
variances
data:  data:  TS.kPa by Fiber.type
Fligner-Killeen:
med chi-squared = 1.0622, df = 3,
p-value = 0.7862
>
> plot(x = my.model$fitted.values,
+      y = my.model$residuals)
```



ANOVA is relatively robust to modest violations of homoscedasticity, as far as the sample is *balanced*.

Assumptions

Model validation

As usual, the independence assumption should be guaranteed (to the best of the experimenter's knowledge) on the design phase, as well as on the analysis. This includes avoiding pseudoreplication and ordering effects, among others.

To test for serial correlations, we can use the Durbin-Watson test, but that only really makes sense if the data is presented to the DW test ordered by an unmodelled and possibly influential variable (such as by order of data collection).

The ANOVA can be quite sensitive to violations of independence. Randomization and attention to possibly influential factors can help avoiding violations of this assumption.

Multiple comparisons

The need for multiple comparisons

If the ANOVA assumptions are verified (i.e., if we have solid grounds for trusting the result of the test), we usually need to determine *which* levels of the factor are significantly different[a];

Whenever possible, the planning of which comparisons will be after an analysis of variance procedure should be defined *a priori*. Post-hoc definition of hypotheses (a.k.a. HARKing[b]) are a common entry point for researcher biases into the analysis, and should be dealt with very carefully.

---

[a] Of course this is only necessary if we rejected H_0 in the original test. For more on how to proceed with nonsignificant results, see Ellis(2010).

[b] **H**ypothesizing **A**fter the **R**esults are **K**nown. See Kerr(1998).

Multiple comparisons

Types of comparisons

The planning of multiple comparisons must be guided by the technical question underlying the experiment.

Whenever possible, the researcher should opt to perform the smallest number of comparisons needed to adequately answer his or her question. This will require the smallest sample size, or result in the largest power for a given experimental setup.

Usual questions involve (but are not limited to):

- *How does one level compare to the others?*
- *How do the levels compare to each other (all vs. all)?*

Multiple comparisons

All vs. all

Pairwise comparisons of the *all vs. all* type appear whenever we are simply interested in detecting which levels are significantly different from which, without any prior information or special interest in one specific level or ordering.

In these cases, the number of comparisons is $K = a(a-1)/2$, where a is the number of levels.

To perform *all vs. all* multiple comparisons, a common approach is to use Tukey's *Honest Significant Difference* (HSD) approach. This method provides a slightly higher power than performing multiple t-tests with adjusted values of α.

Multiple comparisons

All vs. all

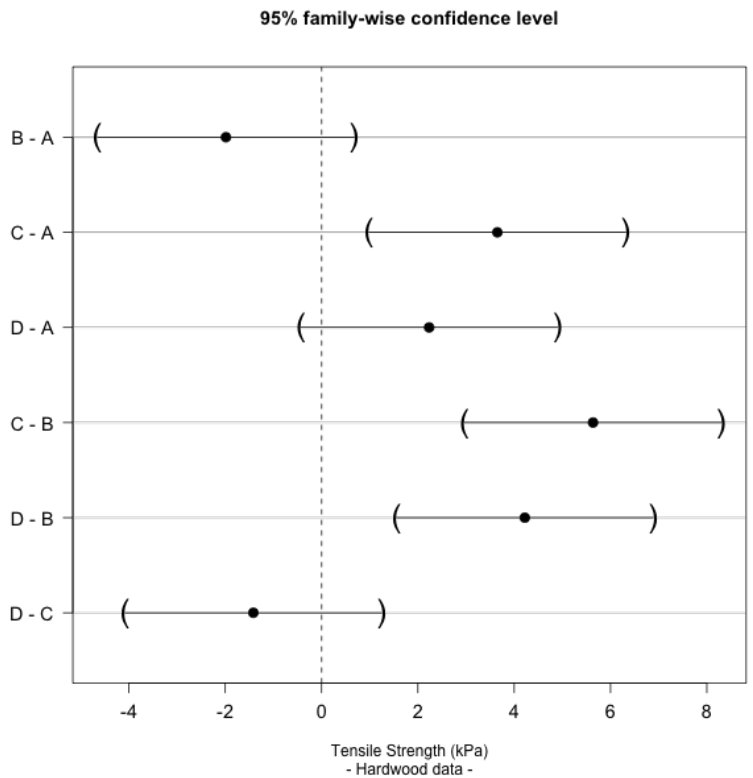
Suppose that was the case for our wood fiber example:

```
> library(multcomp)
> mc1    <- glht(my.model, linfct = mcp(Fiber.type = "Tukey"))
> mc1_CI <- confint(mc1, level = 0.95)

Simultaneous Confidence Intervals
Multiple Comparisons of Means: Tukey Contrasts
95% family-wise confidence level

Linear Hypotheses:
           Estimate lwr     upr
B - A == 0 -1.9867  -4.6478  0.6745
C - A == 0  3.6500   0.9889  6.3111
D - A == 0  2.2333  -0.4278  4.8945
C - B == 0  5.6367   2.9755  8.2978
D - B == 0  4.2200   1.5589  6.8811
D - C == 0 -1.4167  -4.0778  1.2445

> plot(mc1_CI)
```



Multiple comparisons

All vs. one

Pairwise comparisons of the *all vs. one* type usually emerge in the context of comparing levels against a control:

- Comparison of a proposed approach vs. existing ones;
- Comparison of different approaches vs. a standard one (or a placebo-like group);

In these cases, the number of comparisons is $K = a - 1$, where a is the number of levels.

Multiple comparisons

All vs. one - **Dunnett's test**

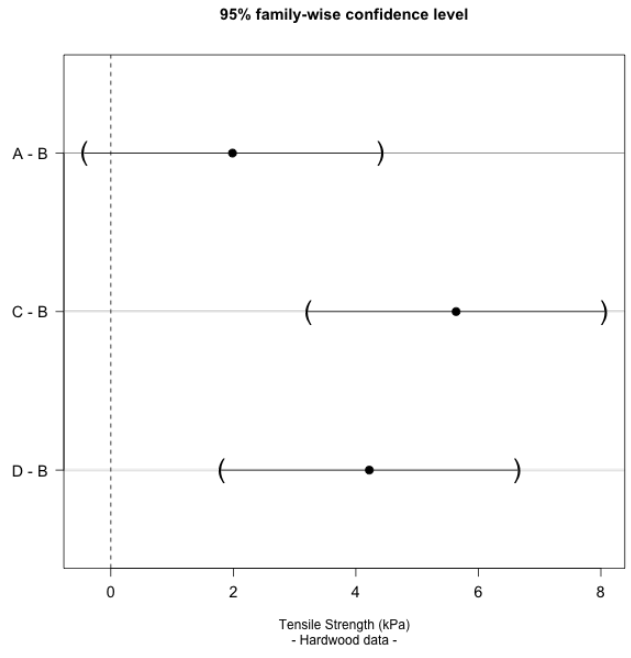
As in the case of *all vs. all* comparisons, there is a test that is usually employed for its superior sensitivity: Dunnett's test.

Assuming that in our example the *B* level is the standard one, against which the other ones are to be compared:

```
> paper$Fiber.type <- relevel(paper$Fiber.type, ref = "B")
> model2           <- aov(TS.kPa ~ Fiber.type, data = paper)
> mc2              <- glht(model2, linfct = mcp(Fiber.type = "Dunnett"))
> mc2_CI           <- confint(mc2, level = 0.95)
```

```
Simultaneous Confidence Intervals
Multiple Comparisons of Means: Dunnett Contrasts
95% family-wise confidence level
Linear Hypotheses:
Estimate lwr     upr
A - B == 0  1.9867  -0.4275  4.4008
C - B == 0  5.6367   3.2225  8.0508
D - B == 0  4.2200   1.8058  6.6342
```

```
> plot(mc2_CI)
```



Multiple comparisons

Some considerations on sample sizes

The kind of comparisons that are to be performed after an ANOVA should be planned in advance, as it influences your data collection and sample size calculations. There are of course sample size formulas for the pure ANOVA, but these are often of limited use since they require the specification of counterintuitive parameters (such as MRES in terms of the ratio of within-groups to across-groups variances).

There are a myriad of approaches for post-ANOVA multiple comparisons[e], but in general the formulas for sample size calculation will follow the ideas outlined below: correct the α value to account for type-I error inflation and calculate *n* based on formulas for two-sample *t* tests.

[e] Check Hothorn *et al.* (2008) for an idea on how varied this can get.

Multiple comparisons

Some considerations on sample sizes

A usual, simpler approach for sample size calculations in the context of the one-way ANOVA is to perform the estimation based on the multiple comparisons one wishes to perform in case an effect is detected.

To perform these calculations we need to account for the fact that we will be testing multiple hypotheses on the same data, and correct the type-I error rate accordingly.

Multiple comparisons

MHT considerations

The multiple comparisons performed after an ANOVA can be thought of as a series of t-tests (with some slight modifications) on the difference between two population means;

If the assumptions of the ANOVA are verified, we already have some information about the data: we know, for instance, that the groups are homoscedastic, and that their common variance is estimated by MS_E, with $a(n-1)$ degrees of freedom;[f].

We also know that the probability of a type-I error on **each test** is given as α. If we want to to keep our overall error rate controlled at a given level while we perform multiple tests on the same data set, we will need to correct the α value used for each test.

[f] This value is the one that should be used in sample size calculations. There are ways to estimate sample sizes for the omnibus ANOVA, but we'll usually want to pinpoint the differences anyway.

Multiple comparisons

MHT corrections

There are a number of ways of adjusting the α value of the pairwise comparisons in order to maintain the *familywise error rate* (FWER) at a controlled level[c].

Two of the most common (and most conservative) are the Bonferroni and the Šidák corrections, which have different formulations but tend to provide very similar results.

Assuming K planned comparisons, the Bonferroni method tests each individual hypothesis with:

$$\alpha_{adj} = \frac{\alpha_{family}}{K}$$

[c] The methods presented here work well when the number of comparisons is relatively small. For more on MHT, see Schaffer(1995)'s discussion on controlling the False Discovery Rate.

Multiple comparisons

All vs. all

A simple approach is to calculate the sample size using Bonferroni-corrected α-values (for simplicity), and performing the actual pairwise tests using Tukey's or Dunnett's methods (for increased power).

An approximate formula to calculate balanced sample sizes is:

$$n \approxeq 2 \left[\left(t_{\alpha_{adj}/2}^{a(n-1)} + t_{\beta}^{a(n-1)} \right) d^* \right]^2$$

where d^* is the (standardized) MRES for the comparison of two groups.

More on sample sizes

Sample size formulas for ANOVA

If one is interested in calculating the required sample size directly for the ANOVA procedure (without worrying about the eventual post-hoc testing), the formulas are almost as simple as those used for the t tests.

Essentially, the power/sample size calculations for the ANOVA boil down to the equality:

$$F_{(1-\alpha)} = F_{\beta;\phi}$$

with both F distributions having $(a-1)$ degrees of freedom in the numerator and $a(n-1)$ in the denominator. The noncentrality parameter ϕ is given by:

$$\phi = \frac{n\sum_{i=1}^{a}\tau_i^2}{\hat{\sigma}^2}$$

More on sample sizes

Sample size formulas for ANOVA

To ilustrate the sample size calculation procedure, imagine an experimental design with $a = 4$, $\alpha = 0.05$, an upper estimate of the within-groups standard deviation $\sigma = 7$, and suppose that the researcher wants to be able to detect whether any two means present differences of magnitude $\delta^* = 12$ with power $(1 - \beta) = 0.8$.

Under these conditions, we can assume a hypothetical (conservative) scenario where we have two levels biased symmetrically about the grand mean, and all the others equal to zero:

$$\tau = \left\{ -\frac{\delta^*}{2}, \frac{\delta^*}{2}, 0, 0 \right\}$$

More on sample sizes

Sample size formulas for ANOVA

For this case we have a noncentrality parameter:

$$\phi = \frac{4\left(6^2 + 6^2 + 0 + 0\right)}{7^2} = 5.88$$

Which allows us to calculate the required sample size by iterating on n until:

$$F_{(1-\alpha)} \leq F_{\beta;\phi}$$

More on sample sizes

Sample size formulas for ANOVA

Doing it manually:

```
> a       <- 4
> alpha   <- 0.05
> sigma   <- 7
> delta   <- 12
> beta    <- 0.2
>
> tau <- c(-delta/2, delta/2, rep(0, a - 2)) # define tau vector
> n   <- 2
> while (qf(1 - alpha, a - 1, a*(n - 1)) >
+        qf(beta, a - 1, a*(n - 1), n*sum(tau^2)/sigma^2)) n <- n + 1
> print(n)
[1] 9
```

Using `power.anova.test()`:

```
> vartau <- var(tau)
> power.anova.test(groups = 4, between.var = vartau,
+                  within.var = sigma^2, sig.level = alpha,
+                  power = 1 - beta)$n
[1] 8.463358
```

Bibliography

Required reading

1. D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers*, Ch. 13. 5th ed., Wiley, 2010.
2. N.L. Kerr, *HARKing: Hypothesizing After the Results are Known*, Personality and Social Psychology Reviews 2(3): 196–217, 1998.

Recommended reading

1. P. D. Ellis, *The Essential Guide to Effect Sizes*. 1st ed., Cambridge, 2010.
2. J.P. Schaffer, *Multiple Hypothesis Testing*, Annual Reviews on Psychology 46, 561–584, 1995.
3. P. Mathews, *Sample Size Calculations: Practical Methods for Engineers and Scientists*. Ch. 8, 1st ed., MMB, 2010.
4. T. Hothorn, F. Bretz, P. Westfall, *Simultaneous Inference in General Parametric Models*. Biometrical Journal 50(3), 346–363, 2008.

About this material

Conditions of use and referencing

Please reference this work as:

Felipe Campelo (2018), *Lecture Notes on Design and Analysis of Experiments*.
Online: https://github.com/fcampelo/Design-and-Analysis-of-Experiments
Version 2.12. Creative Commons BY-NC-SA 4.0.

```
@Misc{Campelo2018,
  title={Lecture Notes on Design and Analysis of Experiments},
  author={Felipe Campelo},
  howPublished={\url{https://github.com/fcampelo/Design-and-Analysis-of-Experiments}},
  year={2018},
  note={Version 2.12. Creative Commons BY-NC-SA 4.0.},
}
```