Experiment Design for Computer Sciences (01CH740)

Topic 08 - Multiple Comparison – Anova and Post-Hoc testing

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

June 18th, 2020

Version 2020.1 (Updated June 19, 2020)

Announcement 1: Modification on Class Schedule

There was an extension for Grade submission in the graduate school, so I have revised the class schedule and deadlines:

- 6/19 – Topic 8: Multiple Comparisons
- 6/26 – Topic 7: Sample Sizes
- 6/27 – (Saturday) Cancelled
- 7/3 – Open Question Session
- 7/20 – Report 3 submission Deadline
- 7/26 – Grades Announcement

Announcement 2: Third Report and Grading

Third Report Topic

Like the first and second report, in the third report you have to design, run and analyse an experiment.

For report 3, you must include in your experiment design:

- Calculation of sample sizes (next lecture);
- Comparison of multiple samples (this lecture);

Grading

Originally, I was expecting to have some extra time after the 3rd report deadline to allow students to fix their reports if they wanted to increase their grades.

Because this is not possible, I will change the grading criteria: The final grade will be the **average of the two best reports.**

However, You must still submit all three reports to pass the course.

Announcement 3: Topics in Computational Sciences

TWINS CODE: 0AL5402 / 01CH751

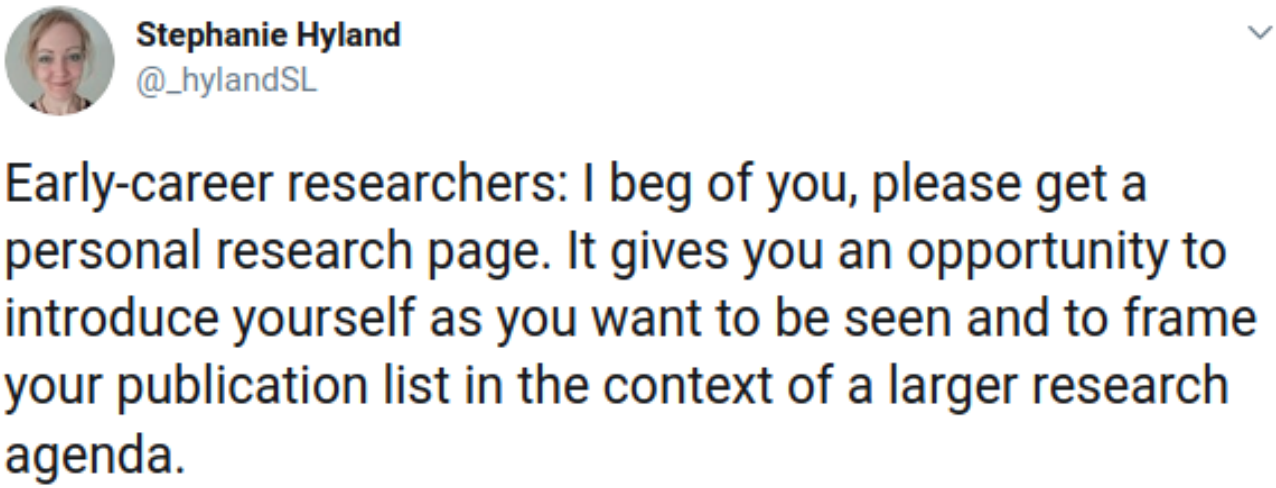
This is an intensive course (1 credit, about 12 hours of lectures), that covers several topics given by professors from different areas:

- Information Access And Knowledge Extraction from News Archives (Prof. Adam Jatowt)
- From Exponential Growth to Power Laws: The How and Why of Modeling the COVID pandemic (Prof. Stephen Turnbull)
- Clustering Methods and Applications (Prof. Ye Xiucai)
- Test Case Geeration: past, present and Future (Prof. Simona Vasilache)
- Bioinformatics: A platform to Understand Biological Systems Using Computational Techniques (Prof. Bakku Kumar)

Research Talk: Take control of your online presence!

Have you ever tried to google your own name?

People who might want to interact with you in the future will usually search for your name online (future professors, students, collaborators, employers...)



I say the same: When I am contacted by a student or researcher, first thing I do is to check out their online presence.

Research Talk: Take control of your online presence!

What is an academic webpage?

An academic webpage does not need to be anything complicated. Some basic information to introduce you as a researcher is all you need:

- Your institution and position ("I am a master student at X lab at the University of Tsukuba");
- Your "research agenda" (I am studying X and Y. In the future, I want to make Z happen.)
- A few research achievements:
  - Published articles;
  - Unpublished public manuscripts (arxiv);
  - blogs/github/social media with research opinions;

(**Opinions are important!**);

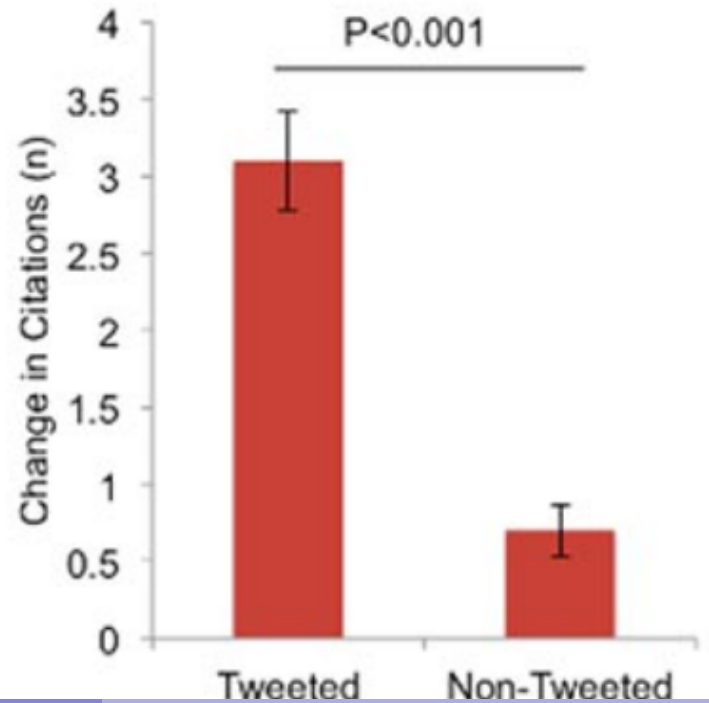
Congratulations, you are now on the map!

Research Talk: Take control of your online presence!

Is social media really that important?

Jessica Luc et al. *"Does Tweeting Improve Citations? One Year Results From the TSSMN Prospective Randomized Trial"*, Annals of Thoracic Surgery, June 2020, https://doi.org/10.1016/h.athoracsur.2020.04.065

- From 112 articles of the same journal, 1/2 were tweeted by a large account, the other half not;
- Citation counts for both groups were observed for one year after the experiment period;
- Read the paper for more details on experiment design. What are the take aways from this study for you?



Topic 08 – Multiple Comparisons

Multiple Comparison Scenarios

There are many situations in which we are interested in comparing multiple samples from possibly different populations at the same time. For example:

- **Parameter Tuning**: We want to test multiple settings of a different parameter (should this network have two, three or four layers?);
- **Comparison of Multiple Algorithms**: You are comparing your proposed algorithm against a set of different algortihms from the state of the art;
- Other situations...

The **t-test** that we studied in the last lecture can be used for hypotheses about one or two samples, so how do we test these cases?

A common mistake: Repeated Testing

One (wrong) solution that we see sometimes in the literature is to do "multiple pairwise testing": Out of 6 methods, test A against B, A against C, A against D, ... etc. And report the result for each comparison.

What is the problem with that?

Remember that each hypothesis test has an associated risk of **TYPE I Error**. In other words, there is a chance that the test will reject the null hypothesis, even if the null hypothesis is true. (We called this parameter α in previous classes).

When we repeat the same test several times, we open ourselves to **compound probabilities**.

A common mistake: Repeated Testing

Compound Probabilities

- Probability of Type I error on one test with ($\alpha = 0.05$):
  $1 - 0.95 = 0.05$
- Probability of Type I error on TWO tests with ($\alpha = 0.05$):
  $1 - 0.95 \times 0.95 = 0.09$
- Probability of Type I error on SIX tests with ($\alpha = 0.05$):
  $1 - 0.95^6 = 0.26$
- Probability of Type I error on TWENTY tests with ($\alpha = 0.05$):
  $1 - 0.95^{20} = 0.64$

See also: `https://xkcd.com/882/`

Lecture Outline

To avoid these problems, we need to use specific techniques for experiment designs involving multiple comparisons.

In this lecture, we will study:

- ANOVA: A statistical test to detect differences in **sets** of samples;
- Post-hoc comparisons: Techniques for making statistical inferences after the ANOVA test.

Of course, there are other statistics and techniques not covered in this lecture.

Comparison of multiple means

Introduction

In previous lectures, 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.

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.[1]

[1] 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.

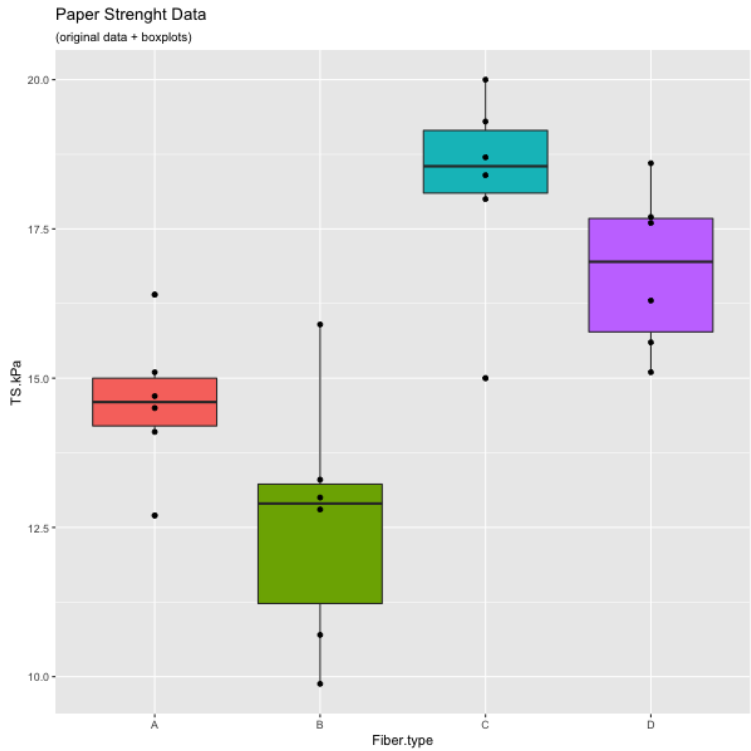
The minimum difference of practical meaning is defined as $5kPa$, and a reasonable upper estimate for the standard deviation of this process is $\hat{\sigma} = 6kPa$. Desired error levels are defined as $\alpha = 0.1$ and $\beta = 0.2$.

Example: paper manufacturing

Exploratory data analysis

It is always a good idea first to perform exploratory data analysis.

As we are interested in the differences between the four wood types, we plot each of them as a boxplot and observe the differences.




```
> 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()
```

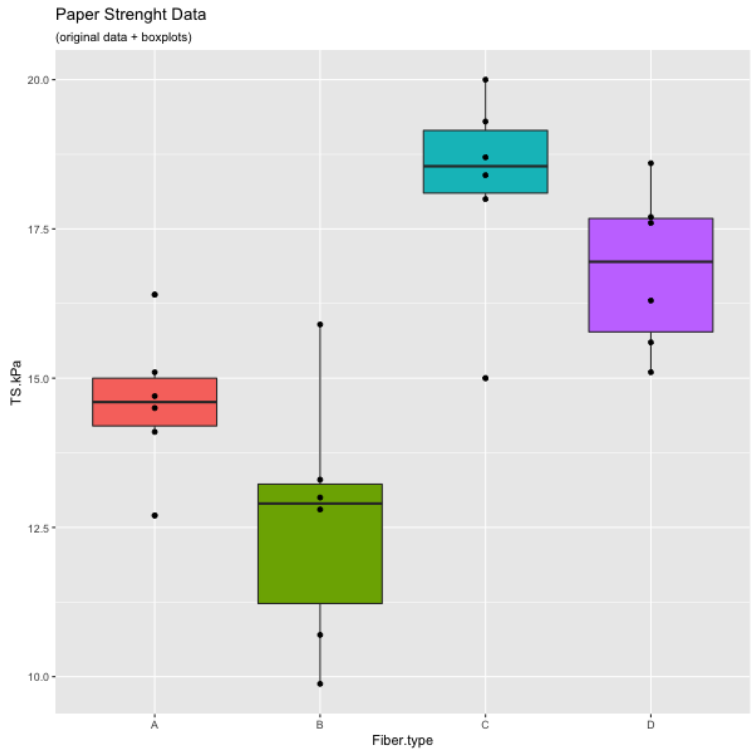
Example: paper manufacturing

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);

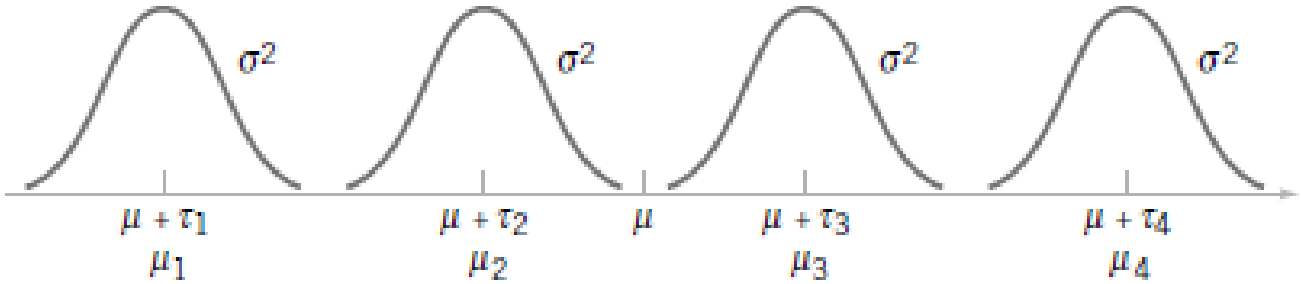
In the derivation of the statistical test for the existence of differences in the group means, we will employ the effects model, and initially consider a few assumptions about the residuals:

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

Example: paper manufacturing

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*.

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:

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

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

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

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"

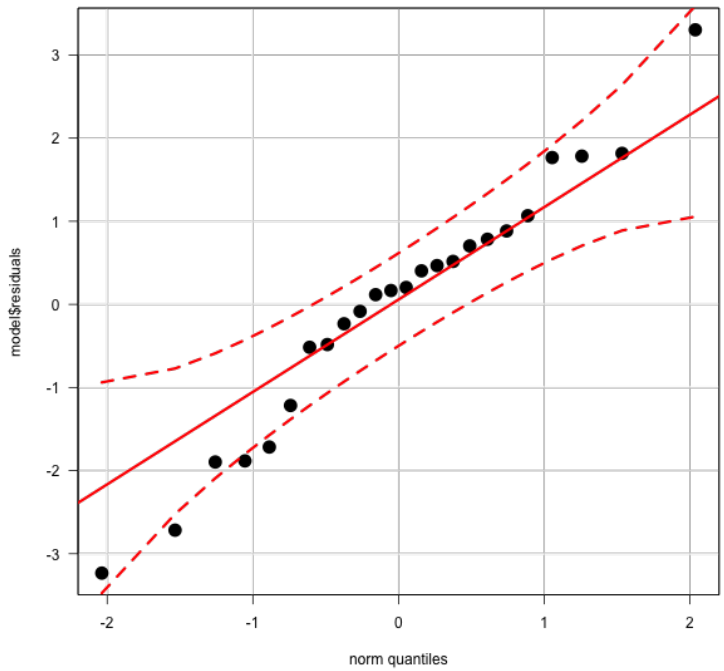
Assumptions – Model Validation

Normality Assumption

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)
```



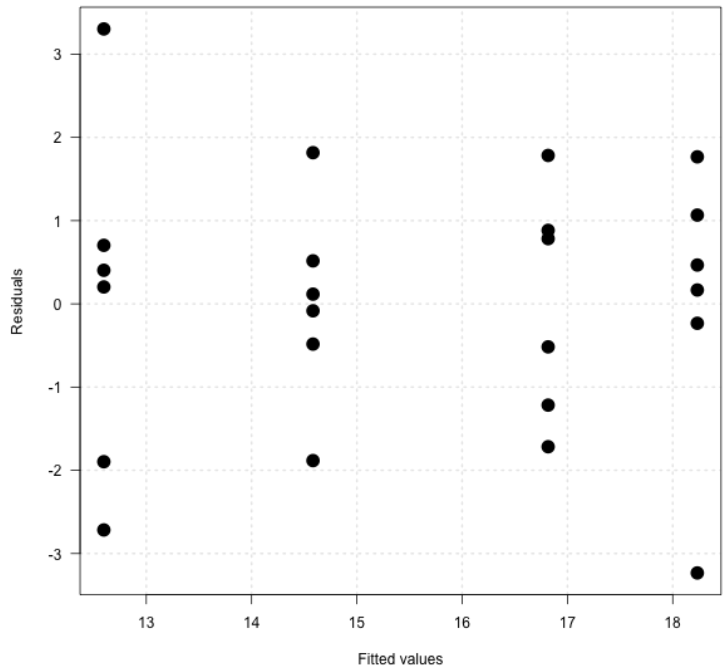
The ANOVA is relatively robust to moderate violations of normality, as long as the other assumptions are verified (or the sample size is large enough).

Assumptions – Model Validation

Homoscedasticity Assumption (similar variances)

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

Independence Assumption

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[1];

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[2]) are a common entry point for researcher biases into the analysis, and should be dealt with very carefully.

[1] 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).

[2] **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 does each level compare to the grand mean?*
- *How do the levels compare to each other (all vs. all)?*

Multiple comparisons

MHT considerations

The multiple comparisons performed after an ANOVA are essentially composed of a series of t-tests for the difference between two population means, with some slight modifications;

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;

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

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[3].

Two of the most common (and most conservative) are the Bonferroni and the Šidák corrections. Assuming K planned comparisons, the Bonferroni method tests each individual hypothesis with:

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

while the Šidák correction uses:

$$\alpha_{adj} = 1 - \left(1 - \alpha_{family}\right)^{1/K}$$

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

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.

The sample size calculations for this case can follow the same equations used for the t test for two independent samples, but with the α value corrected for multiple hypotheses and the number of degrees of freedom of the reference distribution equal to those of the residual term in the ANOVA, i.e., $a(n-1)$.

Multiple comparisons

All vs. all

For performing *all vs. all* multiple comparisons, a common alternative is to use Tukey's *Honest Significant Difference* (HSD) approach. This method is generally chosen because it provides a slightly higher power when compared to the Bonferroni correction[4].

A simple approach is to calculate the sample size using Bonferroni-corrected α-values (for simplicity), and performing the tests using Tukey's HSD corrections (for increased power):

$$n \approxeq 2 \left(\frac{\left(t_{(\alpha_{adj}/2)} + t_{\beta} \right) \hat{\sigma}}{\delta^*} \right)^2$$

with $a(n-1)$ degrees of freedom for the t variables.

[4] The difference is due to Tukey's approach using a modified value for the t_{β} term in the power calculations. See Mathews (2011) and Montgomery (2010) for details.

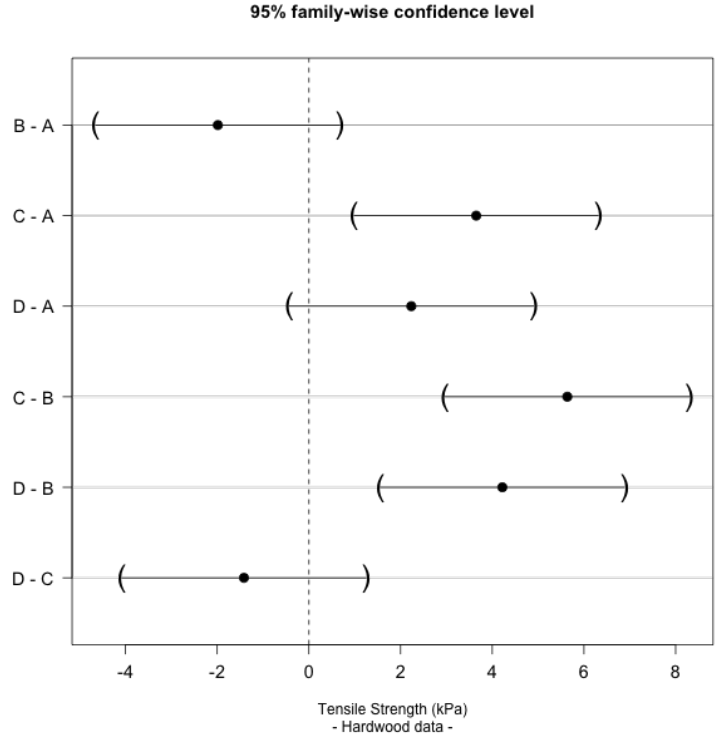
Multiple comparisons

All vs. all

```
> 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:
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. Each test can again be performed (at least in principle) using the t_0 test statistic:

$$t_0^i = \frac{\bar{y}_i - \bar{y}_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_0}\right)}}$$

Multiple comparisons

All vs. one

There are two main approaches to calculating sample size for *all vs. one* comparisons:

- Balanced design;
- Optimal allocation of units.

With a balanced design (that is, all levels have the same number of observations), the calculation of n follows the same approach as the *all vs. all* comparisons, but correcting α for only $a - 1$ comparisons.

For the optimal allocation of units, an unbalanced design is used.

Multiple comparisons

All vs. one - optimal allocation

As several levels will be compared against the single control group, the relative importance of the latter is greater and therefore it should have a larger sample size.

To maximize the power of this multiple comparisons procedure, the sample size of the control group should be:

$$n_0 = n_i\sqrt{K}$$

where n_i is the common sample size for the non-control levels:

$$n_i \approxeq \left(1 + \frac{1}{\sqrt{K}}\right)\left(\frac{(t_{(\alpha_{adj}/2)} + t_\beta)\hat{\sigma}}{\delta^*}\right)^2$$

[2]

[2] A good free software for doing sample size calculations and power analysis in nontrivial contexts such as this one is G*Power 3, `http://www.gpower.hhu.de/`. It is also relatively simple to implement these calculations in R.

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.

The control group sample size n_0 calculated assuming that Bonferroni-corrected t-tests will be used is slightly overestimated in relation to the required n_0 for Dunnett's test, but in practice the differences are small enough not to matter;

Multiple comparisons

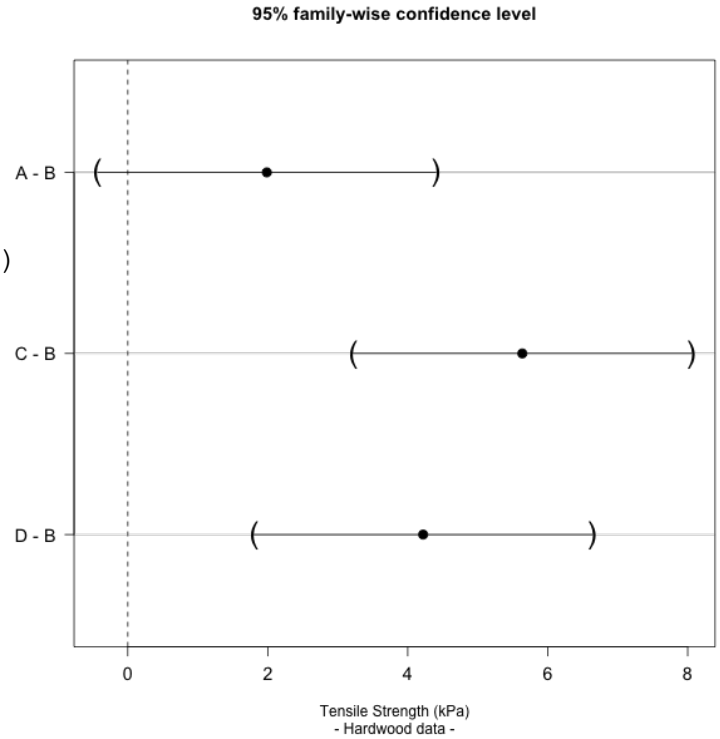
All vs. one

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 final considerations

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 usually of limited use since researchers frequently want to know where the detected differences lie.

There are a myriad of approaches for post-ANOVA multiple comparisons[5], but in general the formulas for sample size calculation will follow the ideas outlined above: correct the α value to account for type-I error inflation and calculate *n* based on formulas for two-sample *t* tests.

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

Class Summary

About these Slides

These slides were made by Claus Aranha, 2020. You are welcome to copy, re-use and modify this material.

These slides are a modification of "Design and Analysis of Experiments (2018)" by Felipe Campelo, used with permission.

Individual images in some slides might have been made by other authors. Please see the references in each slide for those cases.

Image Credits I

[Page 7] image from Jessica Luc et al.:
https://doi.org/10.1016/h.athoracsur.2020.04.065
[Page 13] Paper Mill Image from http://goo.gl/xYVW0M
[Page 19] Image from Montgomery&Runger (2010), Ch. 13