Experiment Design for Computer Sciences (0AL0400)

Statistical Inference III – Non-Normality, and Multiple Sample testing

Claus Aranha
caranha@cs.tsukuba.ac.jp

University of Tsukuba, Department of Computer Sciences

Version 2021.1 (Updated May 25, 2022)

Part I – Dealing with Non-normal Data

Non Normality

What is non-normality?

- Until now we studied test statistics which assume that the **estimator** calculated from the sample comes from a normal distribution (or close enough).
- In some cases, this assumption **does not hold**. In this condition, how can we perform the statistical analysis of the results?

Non Normal data make everything different

Weight Loss Example

A researcher is examining two different diets, **Diet A** and **Diet B**, and wants to compare the weight loss by people following one diet or the other. They obtained the following data:

```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
```

As you can see, Diet A has one big outlier that makes the data not normal. How much does this affect the statistical test?

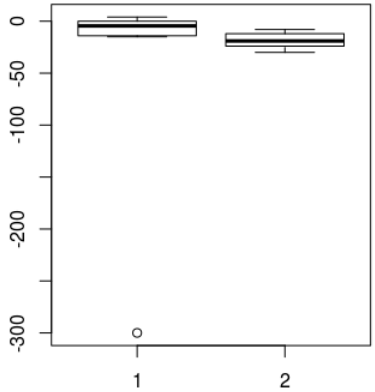
Note!

Remember that in real research we need to ask ourselves: Why does this outlier exist? Is it an error in the experiment? An error in the data input? A new discovery? This is part of research!

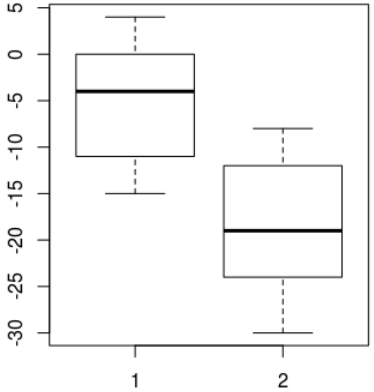
Non Normal data make everything different

The visualization of the data is very different with and without the outlier.

Data with Outlier



Data without Outlier



Non Normal data make everything different

The outlier influences the result of the t-test

The standard T-test does not indicate a difference between these samples, and even suggests that the mean of the first sample is lower!

```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
t.test(diet.a,diet.b)

##  Welch Two Sample t-test
## data:  diet.a and diet.b
## t = -0.53945, df = 9.1048, p-value = 0.6025
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -82.9774  50.9774
## sample estimates:
## mean of x mean of y
##     -34.5     -18.5
```

Non Normal data make everything different

What can we do about it?

- Should we find and remove outliers?
  - If the outlier is an experimental error, it makes sense to remove it;
  - Sometimes, the outlier is a important effect that needs to included in the analysis;
- It is also possible to use statistical methods that are not sensitive to the outlier.

Non Normal data make everything different

Non-parametric methods

Non-parametric tests use statistics that come from non-parametric distributions.

In this case, a non-parametric test will indicate the difference between the **location shift** of the two samples (i.e., the first sample has smaller observations than the second).

```
diet.a <- c(4,3,0,-3,-4,-5,-11,-14,-15,-300)
diet.b <- c(-8,-10,-12,-16,-18,-20,-21,-24,-26,-30)
wilcox.test(diet.a,diet.b)

##  Wilcoxon rank sum test
##
## data:  diet.a and diet.b
## W = 82, p-value = 0.01469
##
## alternative hypothesis: true location shift is not equal to 0
```

Examples of Non-Normal Data

There are many different ways that data can violate the assumption of normality:

- *Special Observations in the Data*:
  - Outliers, data collection errors;
  - Absolute limits in the data (measuring time);
- *Extreme Non-Normal Distributions:*
  - Power Distribution, Cauchy Distribution, etc.
- *Ordinal Data:*
  - Ordinal data is data that can be ordered and compared by some criteria, but you cannot apply traditional algebra on it (ex: subjective scores); ←Important!
- *Completely Non-numerical data*:
  - categorical data, class data, etc; (ex: colors)

Non-Normal Data Example: Random Processes

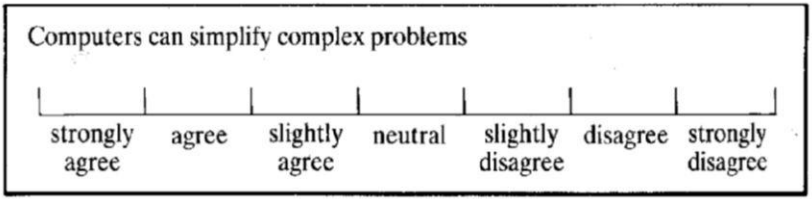
Random processes in nature, such as plant growth or shell formation, are often seen to follow a normal distribution or bell curve.

On the other hand, **random artificial processes not always follow a normal**.

- In general, Pseudo Random Number Generators will use an **Uniform Distribution**. Because of the CLT, aggregations of these results will tend to normal distributions.
- On the other hand, random social processes will often show **Power Distributions** (salaries, social networks) or **Binomial Distributions** (queues);
- It is important to study and understand the process being researched to know its characteristics;

Non Normal Data Example: Likert Data

Likert data is the format often collected from surveys and interview questions.



Why can't we treat likert data directly as numerical? A few reasons:

- Values outside of the 0-5 range have no meaning;
- Algebra on likert data has no meaning (Example: Neutral+Disagree=???)
- The difference between levels is not clear.
  - Is "slightly agree" closer to "agree" or closer to "neutral"?

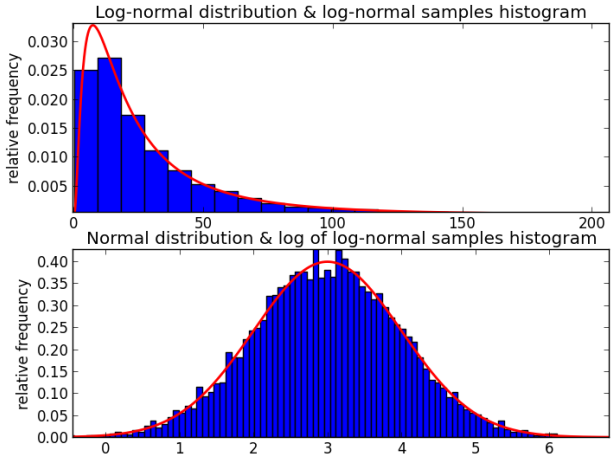
Strategies for working with Non-Normal data

When our data does not follow the normality assumption, there are many different strategies that we can apply, depending on the type of data, and the type of normality violation:

- **Do Nothing**
  - We can just remove the outliers that break the normality assumption;
  - We can trust that test will be robust for small deviations of normality;
- **Transform the Data**
  - Transformation of the data can restore the normal property to data;
- **Non parametric Testing**
  - Some statistical tests do not assume normal data (in exchange for smaller power);
- **Look for a new statistics textbook**
  - There are entire books dedicated to the analysis of non-normal data.

Data Transformation

Using Log Transformation to transform from Lognormal to Normal distribution




```
# R Example of log transform:

# example lognormal data
z <- exp(rnorm(200, -2, 0.4))

# Log transformation
y <- log(z)

# Normal estimators
# from lognormal data:
mu.hat <- mean(y)
sigma.hat <- sd(y)
```

Data Transformation

Skew Transformation

A strong skew in the sampling distribution can be a larger problem for the standard statistical tests. It is possible to remove these through data transformations:

- For left skewed data:
  - square root, cube root, log
- For right skewed data:
  - square root (constant $-x$), cube root (constant $-x$)

Attention: The logarithm of 0 and negative data is not defined. If your data includes 0 or negatives, you may need to add a constant before the transformation.

Data Transformation

Be careful when transforming data

- Pay attention when describing the analysis on a paper or report:
  - When you talk about the analysis, you need to explain the transformation used;
  - When you discuss the results, you must consider the transformed data, as well as the original data;
  - In particular, the **Meaningful difference** must be discussed on the original data;
- Beware that the hypotheses may not be equivalent!
  - Example: The lognormal mean includes the variance. But the transformed lognormal mean does not. In this case, the null hypothesis is only equivalent when the variance of the transformed distribution is equal!

Bootstrapping

Using the CLT to make the data more normal

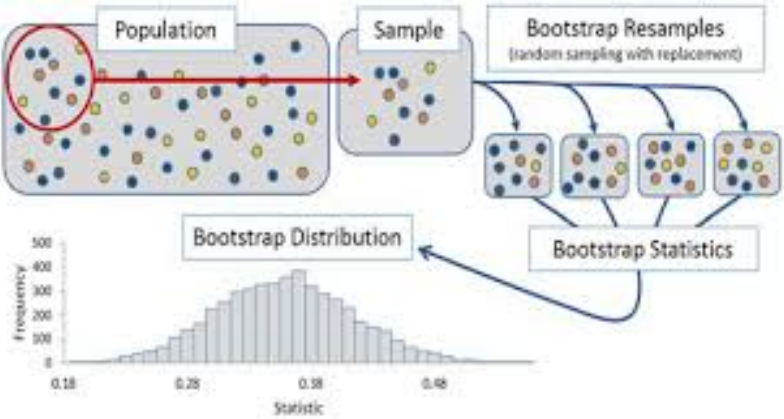
The Bootstrapping procedure is used to obtain an approximation of the "sample mean distribution" from the sample data.

By the Central Limit Theorem, the sample mean distribution of a random variable will usually follow a normal distribution; even when the underlying distribution of observation values is not normal;

So how can we use this to transform non-normal data?

The Bootstrapping Procedure

- Take an initial sample with m observations;
- Create n **bootstrap samples** by selecting $m_b < m$ from the initial sample n times;
- Calculate the mean of each bootstrap sample.



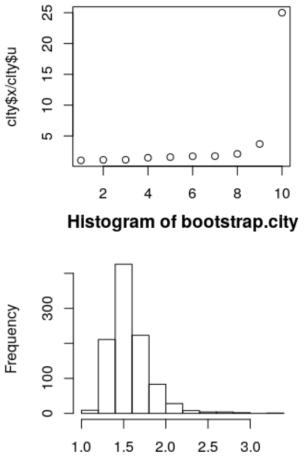
Generating Bootstrapped Data

R Package "boot" for bootstrapping, confidence intervals, and tests.

```
% Non-normal data:
> city
    1    2    3    4    5    6    7    8    9   10
u 138   93   61  179   48   37   29   23   30    2
x 143  104   69  260   75   63   50   48  111   50

% We are interested in the ratio of u and x
> ratio <- function(d, w) sum(d$x * w) / sum (d$u * w)

% Using library boot to create bootstrapped data:
> library(boot)
> bootstrap <- boot(city, ratio, R = 999, stype = "w")
> bootstrap.city <- bootstrap[[2]]
```



Non Parametric Tests

Non-parametric Tests involve statistics that do not assume normality from the population distribution.

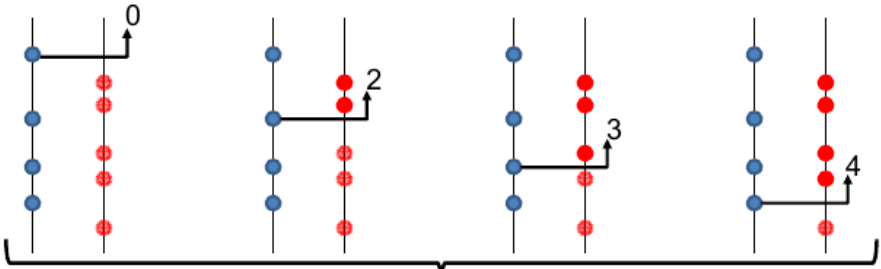
Weak assumptions about the population, however, causes the non-parametric tests to be less strong than parametric ones. Also, usually non-parametric statistics usually do not calculate the distance between the parameter estimate and the hypothesis values.

- Wilcoxon Signed Rank Test (1 sample)
- Wilcoxon Ranked Sum Test / Mann-whitney Test (2 samples)
- Kruskall-Wallis Test (multiple samples)

Mann-Whitney U-test

Unpaired test for two samples

1. Calculate a *U* value.

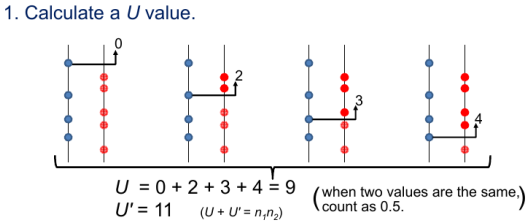


$U = 0 + 2 + 3 + 4 = 9$

$U' = 11$ $\quad (U + U' = n_1 n_2)$

(when two values are the same, count as 0.5.)

Mann-Whitney U-test



- Choose the smaller value of U or U'
- Null Hypothesis: **Both samples come from the same distribution**
- Under the null hypothesis, for big enough n_1 and n_2, U follows roughly a normal distribution with mean $\frac{n_1 n_2}{2}$ and variance $\frac{n_1 n_2 (n_1+n_2+1)}{12}$
- Calculate the test statistic z, and find the p-value from the α-percentile in the z distribution.

Wilcoxon Signed Rank Test

data of 2 groups		# of winnings and losses	
173	174	-	+
143	137	+	-
158	151	+	-
156	143	+	-
176	180	-	+
165	162	+	-

- The Wilcoxon test takes the relative difference between pairs (positive or negative)
- Null hypothesis: **Positive and Negative signs are equally likely**
- The overall number of signs is compared against a binomial distribution under the Null hypothesis.

Wilcoxon Signed Rank Test

R code example

```
## Hollander & Wolfe (1973), 29f.
## Hamilton depression scale factor measurements in 9 patients with mixed anxiety
## and depression, taken at the first (x) and second (y) visit after initiation
## of a therapy (administration of a tranquilizer).

# Data:
% x <- c(1.83,  0.50,  1.62,  2.48, 1.68, 1.88, 1.55, 3.06, 1.30)
% y <- c(0.878, 0.647, 0.598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29)

# Running the test:
% wilcox.test(x, y, paired = TRUE, alternative = "greater")

Wilcoxon signed rank test
data:  x and y
V = 40, p-value = 0.01953
alternative hypothesis: true location shift is greater than 0
```

Recommended Reading

- Kristin L Sainani, "Dealing with Non-Normal Data."
  https://onlinelibrary.wiley.com/doi/full/10.1016/j.pmrj.2012.10.013
- Feng et al., "Log transformation and its implication for data analysis."
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/
- Bommae Kim, "Should I always Transform My Variables to Make them Normal?"
  https://data.library.virginia.edu/normality-assumption/

Part II – Multiple Sample Testing

Multiple Comparison Examples

There are many situations in which we are interested in comparing more than two samples at the same time, to test if they belong to the same population or not.

- **Parameter Tuning**: We want to test multiple sets of parameters for one algorithm (ex: Compare a network with N_1, N_2, N_3 or N_4 layers);
- **Comparison of Multiple Algorithms**: Compare a proposed algorithm with four different algorithms from the state-of-the-art;

Can we use the **t-test** from previous class in this case?
(compare A vs B, A vs C, A vs D, etc...)

Multiple Comparison

A common mistake: Repeated Testing

A common mistake is to perform "multiple pairwise testing": 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 every test has a probability **TYPE I Error**. (test parameter α).

When we repeat the same test many times, these errors are **multiplied!**.

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/`

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

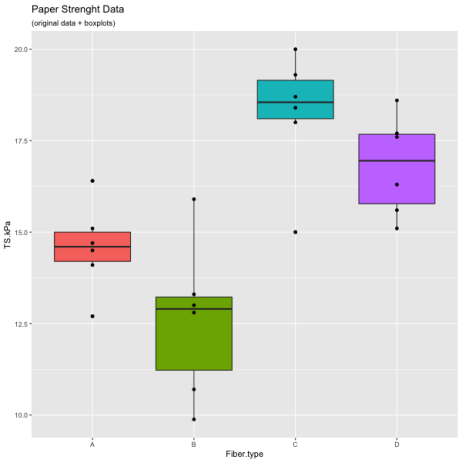
Exploratory data analysis

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 = "paper_strength.csv",
+                     header = TRUE, sep    = ",")

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



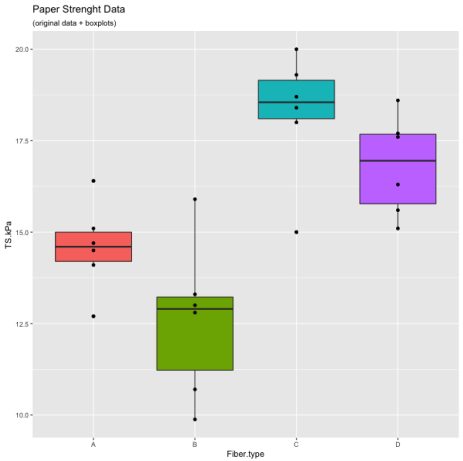
Example: paper manufacturing

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:

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

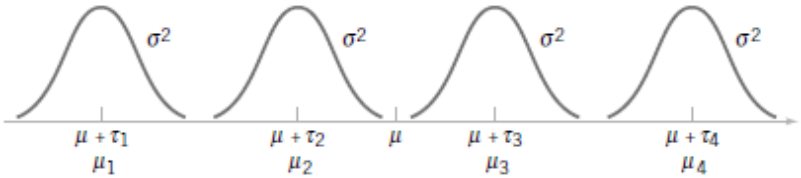
where μ is the overall mean, τ_i represents the effect of the i-th level, and ϵ_{ij} is the residual (random error, or unmodeled variability);

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:

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

Example: paper manufacturing

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:

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

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 gives a quantity known as *mean squares*.

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

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

The expected values of these quantities are:

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

The Fixed Effects Model

Development

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

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, $E\left[MS_{Levels}\right] = E\left[MS_E\right] = \sigma^2$. *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:

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

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

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"

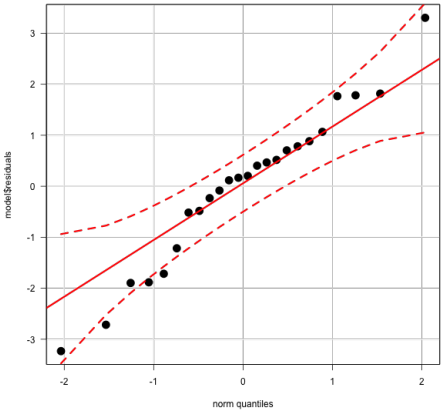
Assumptions – Model Validation

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)
```



Assumptions – Model Validation

What if the residuals do not follow a normal distribution?

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).

If the sample size is not large, or the other assumptions cannot be verified, then a non-parametric test of multiple samples should be considered.

There are several tests, but you should begin here:

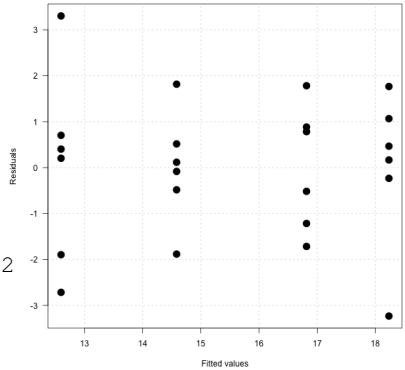
- Unpaired non-parametric test for multiple samples: Kurskal-Wallis test
- Paired non-parametric test for multiple samples: Friedman test

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 long 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

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.

About these Slides

These slides were made by Claus Aranha, 2022. 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 following references for those cases.

Image Credits I

[Page 29] Paper Mill Image from `http://goo.gl/xYVW0M`
[Page 34] Image from Montgomery&Runger (2010), Ch. 13