A massively significant seminar on non-significance

Unit of Applied Statistics
Dept. of Biosystems and Technology
Swedish University of Agricultural Sciences (SLU)

Adam Flöhr & Jan-Eric Englund

Content

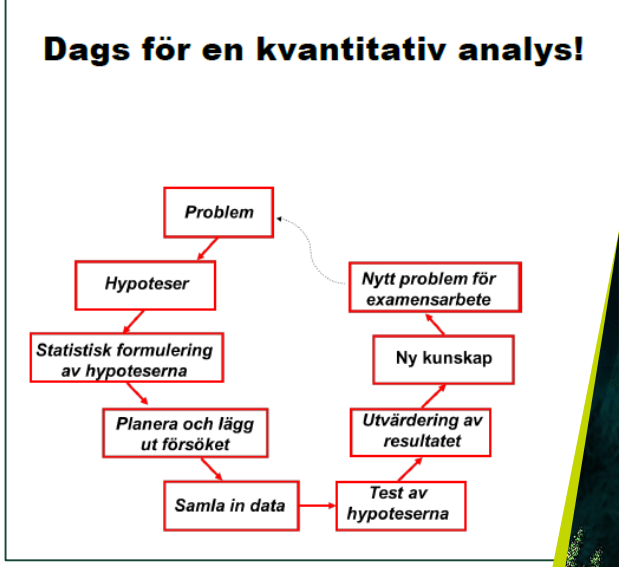
- Statistics@SLU and PhD courses
- Time for a quantitative analysis
- A lot of tests (Adam)
- Not so many tests (Jan-Eric)

Statistics@SLU

SLU:s statisticians (most of them) at a joint meeting in Alnarp in June 2022.

Statistics@SLU

- Åsa Lankinen is the new representative in the steering group.
- The mission is to help employees at SLU with statistical problems.
- Statistics@SLU should also coordinate PhD courses in statistics.
- The manager is Claudia von Brömssen from Ultuna,
  deputies are Magnus Ekström from Umeå and Jan-Eric Englund.



Time for a quantitative analysis

Best Practice project to help students without statistical background to write their theses.

The project have five headings:

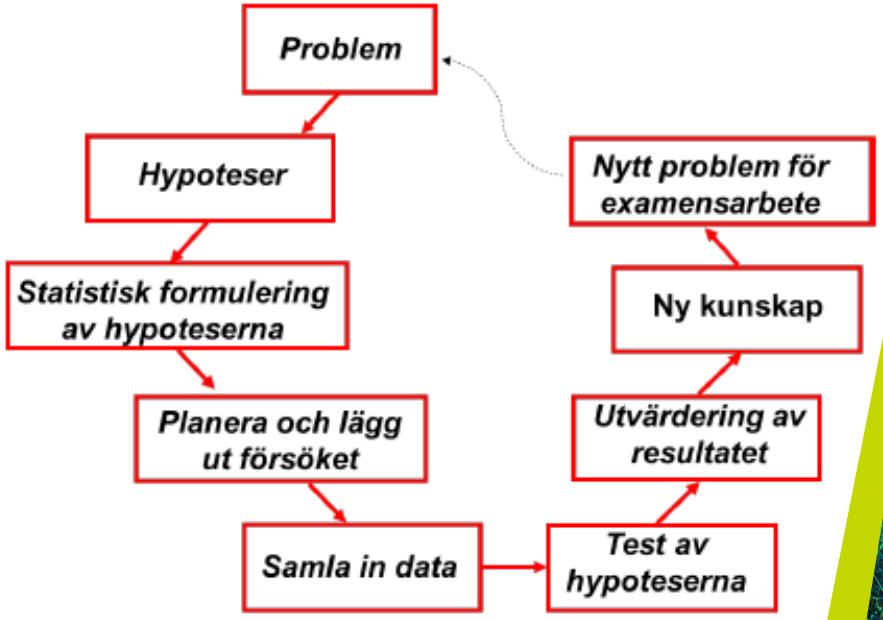
- Define the problem
- Experimental Design
- Collection of data
- Data analysis
- ***Scientifically based conclusions***

NOTE: Everything is in Swedish!

Time for a quantitative analysis

The idea is to help the student to write the thesis based on quantitative data. The front page describes the different steps.

- Problem
- Hypotheses
- Statistical formulation of the hypotheses
- Collect the data
- Test the hypotheses
- Evaluation of the result
- New ideas for new thesis



Last time we spoke (spring 2021)

- Strong reliance on null hypothesis significance testing
- Non-significant results go unpublished
- Effects sizes are exaggerated
- Results are not reproducible

- The American Statistical Association

  In sum, "statistically significant" — don't say it and don't use it.

- Alternative measures of effect or confidence intervals
- Evaluate based on experimental setup, not the outcome

Multiple testing

- The significance level is the probability of rejecting a true null hypothesis
- Multiple testing, higher overall significance level
- The probability of false discoveries is the *family-wise error rate (FWER)*

No of tests	Overall significance
1	5%
2	$1 - 0.95^2 = 9.75\%$
10	$1 - 0.95^{10} = 40.1\%$

So many things to do

- We can easily justify many tests on a single dataset
- Take a simple experiment with four treatments and two response variables
- Data transforms: log, squares and square-roots, ratios or differences between variables, interval classes, excluding extremes, excluding zeroes, drop treatment groups, merge treatment groups, divide by values in control group
- Tests: one-way Anova, Ancova, pairwise post-hoc comparisons, non-parametric tests, pairwise non-parametric tests, ordinal models after classification, correlation and regression between variables, logistic after 0/1-classification of either response

P-value adjustments

- Controlling the false discovery rate (FDR)
  - The proportion of significant results which are false positives
- Controlling the family-wise error rate (FWER)
  - The probability of *some* zero-effect being significant

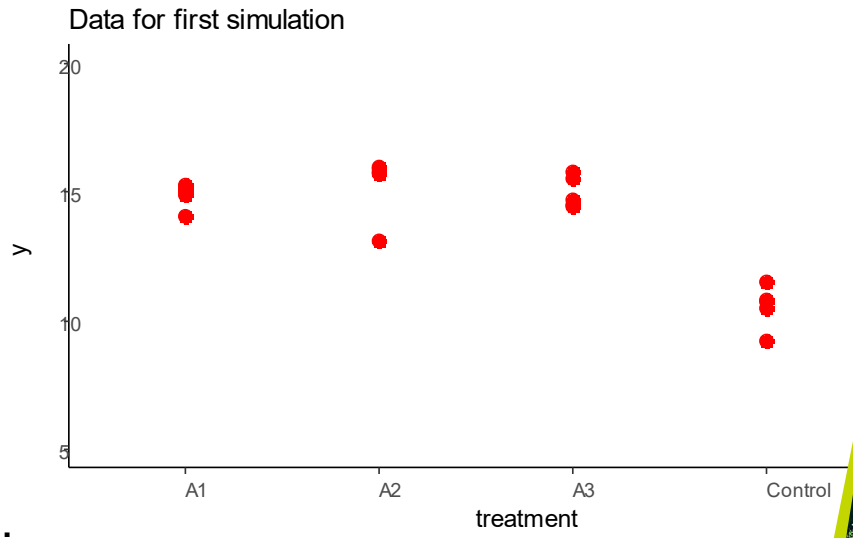
	Declared non-significant	Declared significant	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
	$m - R$	R	m

False discovery rate

- The Benjamini-Hochberg procedure
  1. Perform m tests, each results in a p-value
  2. Order by p-value. Let i be an index of the order
  3. Find the largest p-value which is smaller than $\frac{i}{m}\alpha$
  4. Reject all hypotheses with p-values below the value from (3)

Benjamini Y, Hochberg Y (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, Series B. 57 (1): 289–300. MR 1325392.

- This will give an overall tests with FDR at most equal to α
- Typically, at most five percent of discoveries will be false positives



Example 1 (and only)

There is a standard method called Control.

There are three new treatments, labelled A_1, A_2 and A_3.

A_1, A_2 and A_3 are in fact identical, but better than the Control.

A completely randomized design with five replicates per treatment.

The yields are (simulated data to illustrate my ideas!):

Control:	**10.8, 10.7, 11.5, 10.5, 9.2**	**mean: 10.56 (True value: 10)**
A_1:	**15.2, 14.0, 14.9, 15.3, 15.0**	**mean: 14.89 (True value: 15)**
A_2:	**16.0, 13.1, 15.8, 15.7, 15.9**	**mean: 15.30 (True value: 15)**
A_3:	**14.7, 15.5, 14.5, 15.8, 14.4**	**mean: 14.99 (True value: 15)**

What is the problem?

With ***one*** null hypothesis this is the table illustrating wrong decisions when the significance level is 5%.

	Reject null hypothesis	**Do not reject null hypothesis**
Null hypothesis is true	5%	95%
Null hypothesis is not true	Power	1 – Power

From the example:

If we *only* consider whether Control = A_1 = A_2 = A_3, the risk is 5% that the conclusion from the experiment is that there is a difference if there is no difference.

What is the problem?

Now consider the ***six*** pairwise comparisons with null hypotheses

- Control = A_1
- Control = A_2
- Control = A_3
- $A_1 = A_2$
- $A_1 = A_3$
- $A_2 = A_3$

There are two different situations:

- All null hypotheses are true ($\Rightarrow$ Control = $A_1 = A_2 = A_3$)
- There is at least one false null hypothesis.

What is the problem?

If ***all*** null hypothesis are true, use the previous table with a small modification:

	Reject *at least* one of the true null hypotheses	**Do not reject any of the true null hypotheses**
All null hypotheses are true	5%	95%
There are at least one false null hypothesis	What about the false null hypotheses? Are they rejected?	What about the false null hypotheses? Are they rejected?

From the example:

- True null hypothesis: $A_1 = A_2$, $A_1 = A_3$ and $A_2 = A_3$
- False null hypothesis: Control = A_1, Control = A_2 and Control = A_3

Confidence interval

Make confidence intervals for the differences between *all* pairs.

The probability that ***all*** confidence intervals cover the true value should be at least 95% to satisfy FWER.

Remember:

If there is no difference between two treatments, the confidence interval for the difference should cover 0, but with FWER we also guarantee for the false null hypotheses.

From the example:

95% probability that the intervals for $A_1 - A_2$, $A_1 - A_3$, $A_2 - A_3$ covers 0 **and** Control $- A_1$, Control $- A_2$ and Control $- A_3$ covers 5.

Hypothesis testing

The probability that ***at least one*** true null hypothesis is rejected is smaller than 5% and don’t bother about the false null hypothesis.

It seems more effective to use hypothesis testing, but sometimes you need confidence intervals in your analysis.

Note: We can’t identify and don’t know the number of true null hypotheses.

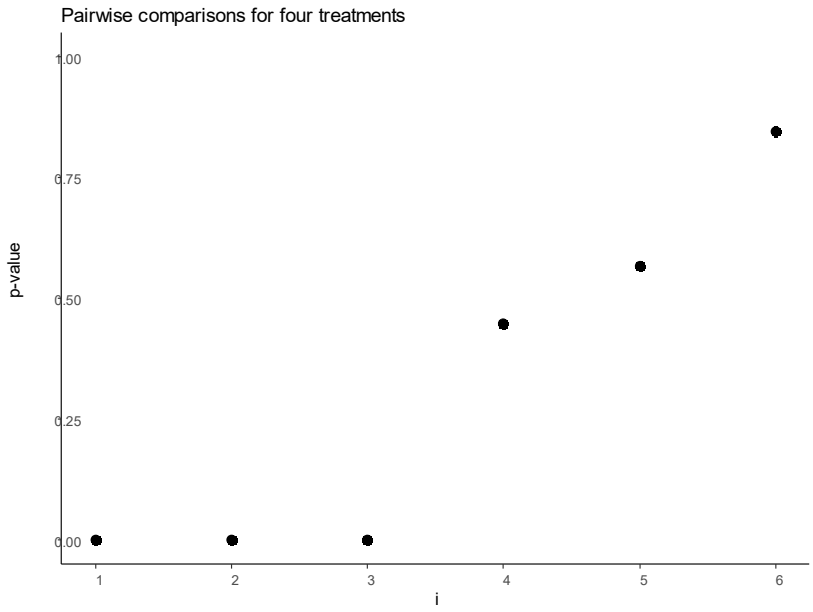
Some of the solutions for *m* tests

- **Bonferroni:** Use 0.05/*m* as the significance level.
  ⇒ Low power.
  Confidence intervals with *simultaneous* confidence level 95%.
  Can adjust p-values in general to prevent mass significance.
- **Tukey**: Use the computer to find the levels for the p-value.
  ⇒ Lower power if there are false null hypotheses, but well-known.
  Confidence intervals with *simultaneous* confidence level 95%.
- **Holm**: An alternative from a paper from 1979 by Sture Holm.
  ⇒ Only for hypothesis testing.
  The guarantee is only for the true null hypotheses.
  Can adjust p-values in general to prevent mass significance.
  Not one of the alternatives in SAS but available by p.adjust in R.

Graphical illustration for four treatments

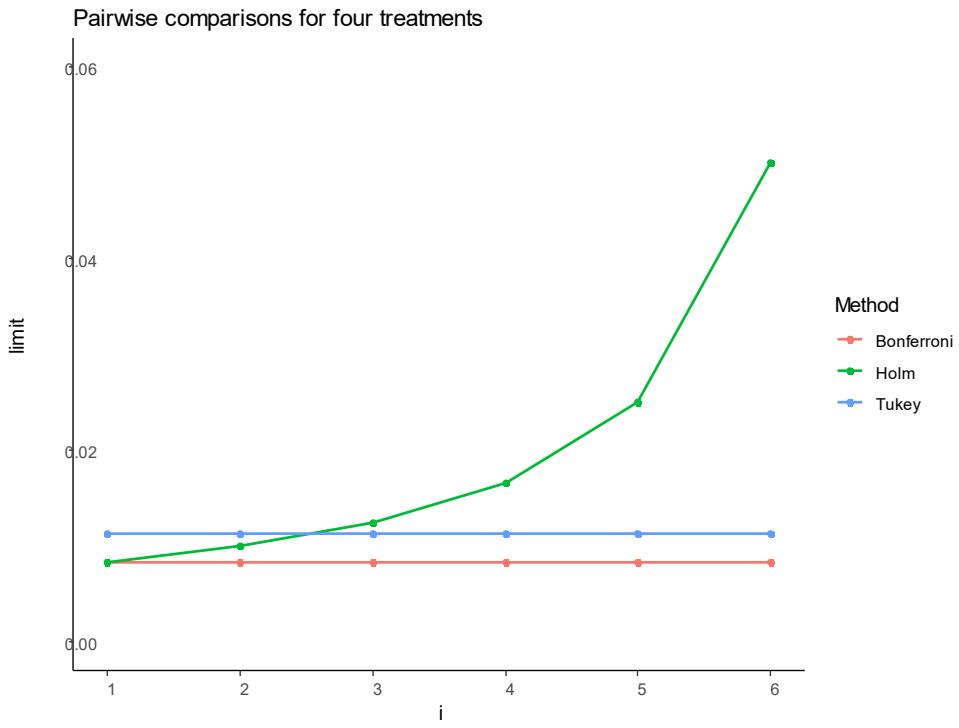
Sort the p-values from the smallest to the largest:

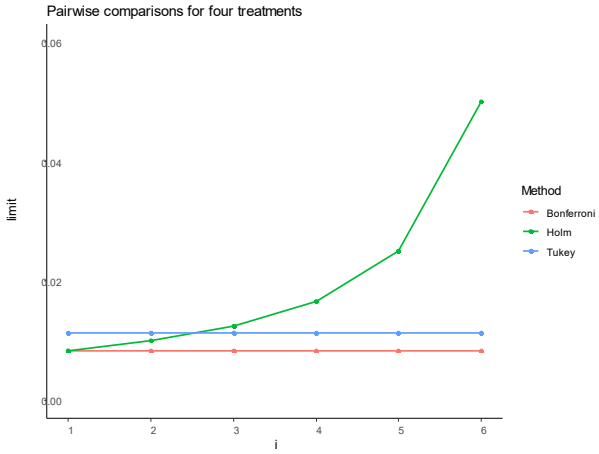
1. Control = A_2 p-value $1.4 \cdot 10^{-7}$
2. Control = A_3 p-value $3.4 \cdot 10^{-7}$
3. Control = A_1 p-value $4.6 \cdot 10^{-7}$
4. $A_1 = A_2$ p-value 0.45
5. $A_2 = A_3$ p-value 0.57
6. $A_1 = A_3$ p-value 0.85



Graphical illustration for four treatments

Limits for the p-values; note the scale on the second axis!





The test procedures

Bonferroni: Use 0.05/6 = 0.0083 as the limit for all p-values.

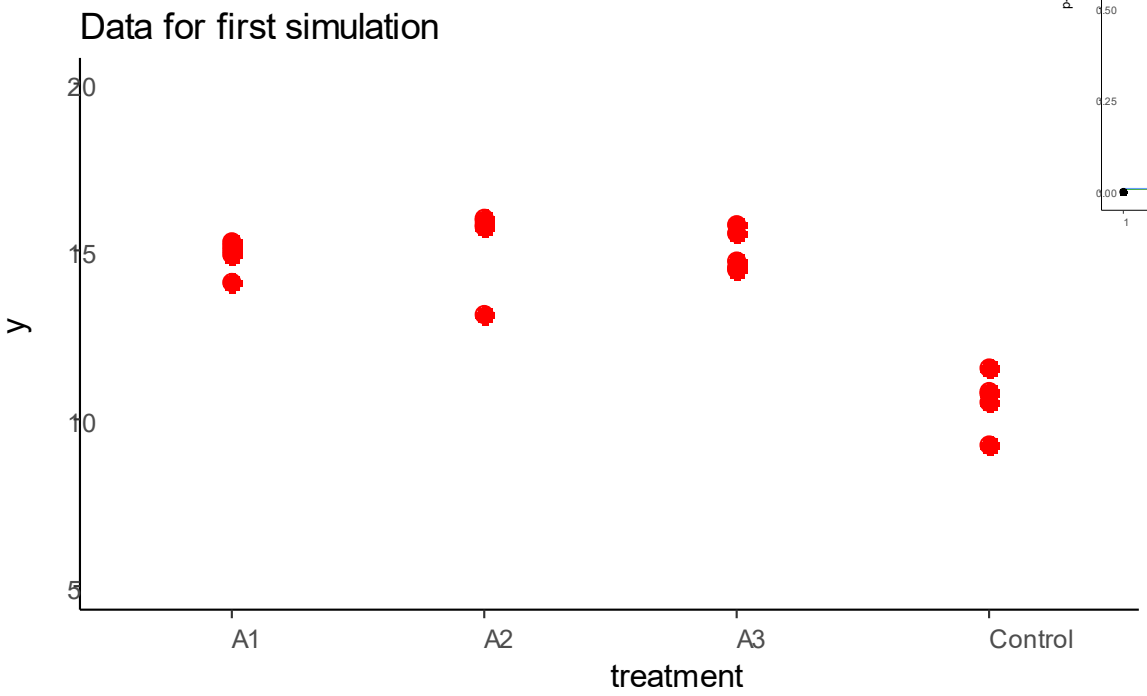
Tukey: Find a general limit from a table or the computer.
With four treatments and six comparisons the limit is 0.0113.

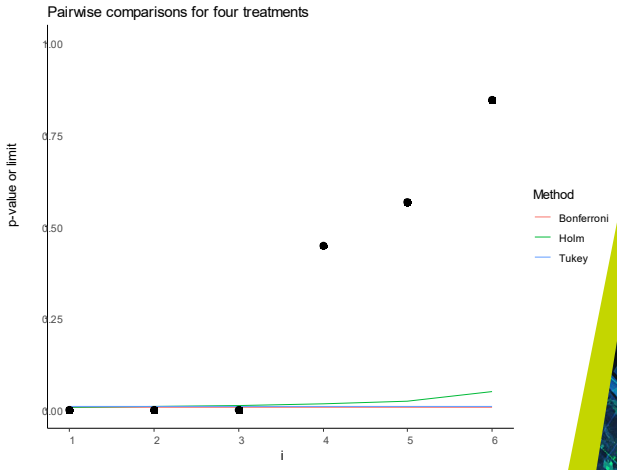
Holm:

- Start with Bonferroni's 0.05/*m* for the smallest p-value.
- If the p-value is larger, stop the process and all tests are *not* rejected,
  If the p-value is smaller, reject this test and continue the process.
- Continue with Bonferroni's test for *m*–1 tests, that is 0.05/(*m*–1).
- If the p-value is larger, stop the process and say that this and all remaining tests are *not* rejected.
  If the p-value is smaller, reject the test and continue the process.
- … Stop when you cannot reject or when all tests are rejected.

Simulation with Control (10 000 simulations)

- Mean of Control: 10. Mean of all Treatments: 15
- Common standard deviation 1,
  Normal distributions, 10 000 simulations

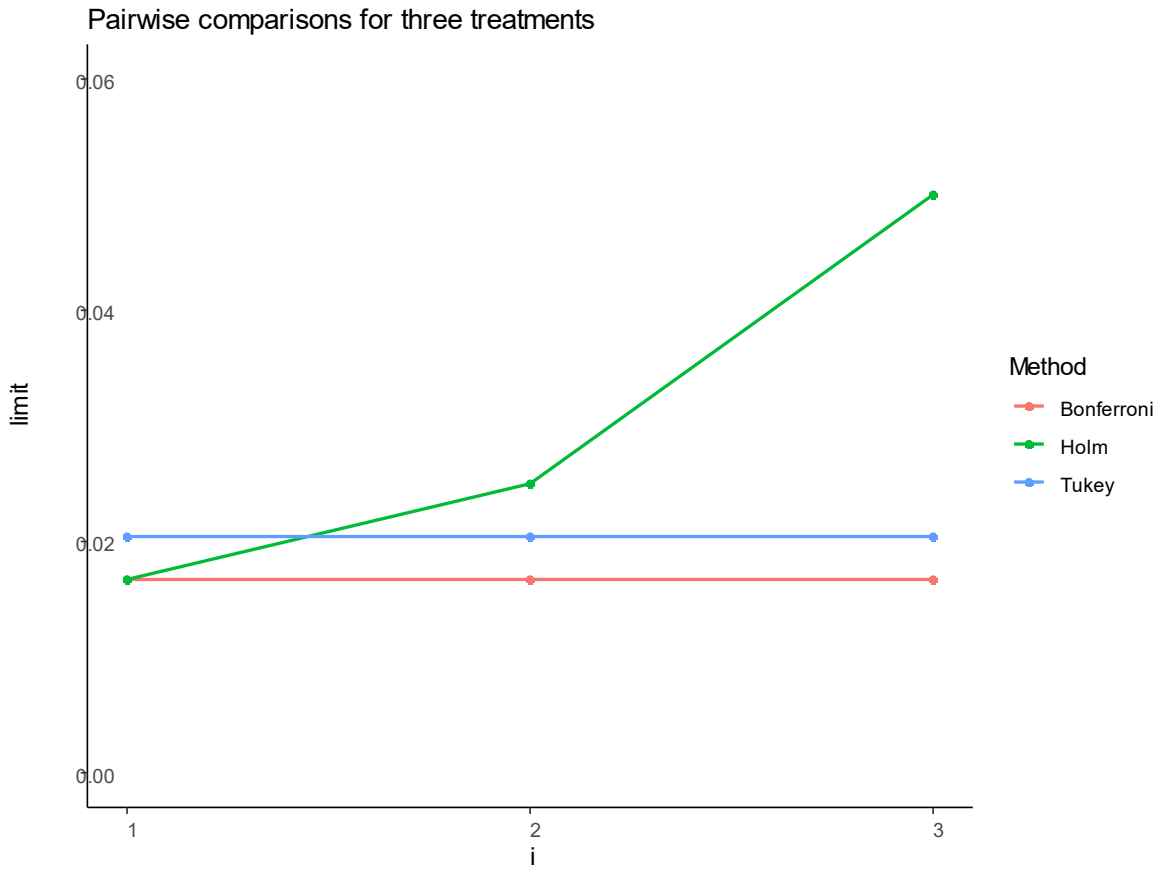




Simulation with Control (10 000 simulations)

- **Bonferroni:** Use 0.05/6 = 0.0083 as the significance level in all tests.
  At least one significant difference between A_1, A_2 and A_3 in **2.4%** of the simulations.
- **Tukey**: Use 0.0113 as the significance level in all tests (from a table).
  At least one significant difference between A_1, A_2 and A_3 in **3.1%** of the simulations.
- **Holm**:
  At least one significant difference between A_1, A_2 and A_3 in **4.3%** of the simulations.
  $\Rightarrow$ Best?

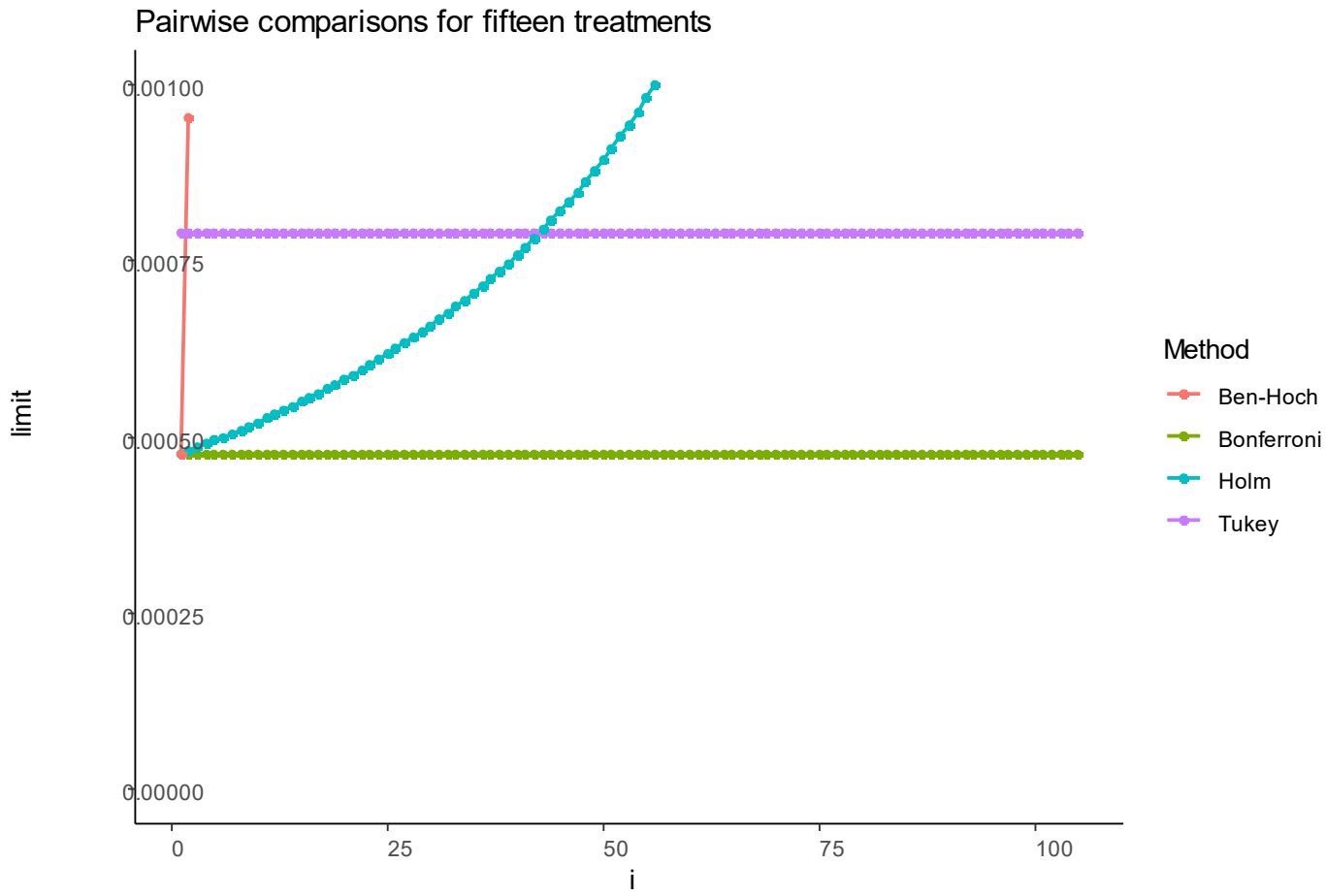
Graphical illustration without Control



Simulation without Control (10 000 simulations)

- **Bonferroni:** Use 0.05/3 = 0.0167 as the significance level in all tests.
  At least one significant difference between A_1, A_2 and A_3 in **4.27%** of the simulations.
  $\Rightarrow$ OK, but only because there are just three treatments!
- **Tukey:** Use 0.0205 as the significance level in all tests (from a table).
  At least one significant difference between A_1, A_2 and A_3 in **5.21%** of the simulations.
  $\Rightarrow$ OK
- **Holm:**
  At least one significant difference between A_1, A_2 and A_3 in **4.27%** of the simulations.
  $\Rightarrow$ OK

Graphical illustration for 15 treatments (105 pairs)



Conclusion

- If the inference is of importance, use a family wise correction (FWER).
- If screening is of importance, use false discovery rate (FDR)
- The control matters – Treatment comparisons will depend on dropping or keeping the control.
- Methodology depends on scientific field and computer package.

Thank you for your attention

Adam.Flohr@SLU.SE

Jan-Eric.Englund@SLU.SE