Reviewing statistics for journals

Musts, Mays, Misses, and Musings

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

Adam Flöhr & Jan-Eric Englund

Musts

– What must always be included

Mays

– What one would like to see

Misses

– Checks for mistakes and fraud

Musings

– Something on the current role of hypothesis testing and p-values

Musts: *Statistical Analysis* in *M & M*

- Statistical analysis is the only fully reproducible part of a study
- The description should be clear enough for *you* to redo the analysis
- The description should be self-explanatory
  - Should not require detective work
  - There is a difference between a *best guess* and a clear description

Example of method description

All data were statistically treated with [a statistical program] *to analyse variance in a multifactoral* [sic] *ANOVA.*

When differences were detected, a Least Significant Difference (LSD) multiple range test was used to compare them.

Non-descriptive.

To compare means, or

test for treatment differences

Well…

All data were statistically treated with [a statistical program] *to analyse variance in a multifactoral* [sic] *ANOVA.*

When differences were detected, a Least Significant Difference (LSD) multiple range test was used to compare them.

multifactorial

Unclear model.
Interactions?

Compare what?

Unclear method.
LSD and the MRT
are different tests.

Checks of model assumptions

- All statistical tests rely on model assumptions

- t-tests and ANOVA models
  - Normal distribution (or reliance on Central Limit Theorem)
  - Equal variance within groups (homoscedasticity)
  - Independent observations

- There must at least be a mention of applied assumption tests

Exact p-values

- Results of hypothesis tests must be presented with the *exact* p-value, not just significance at selected levels.
- Standard report of a test should include
  - 1. Type of test (could be mentioned in general in Statistical Analysis)
  - 2. Degrees of freedom (if applicable)
  - 3. Test statistic
  - 4. p-value

…showed an increase (t(10) = 2.34, p = 0.041)

- … but may vary between journals

Scripts

- For analysis done in scripted languages (Python, R, SAS), scripts must be published

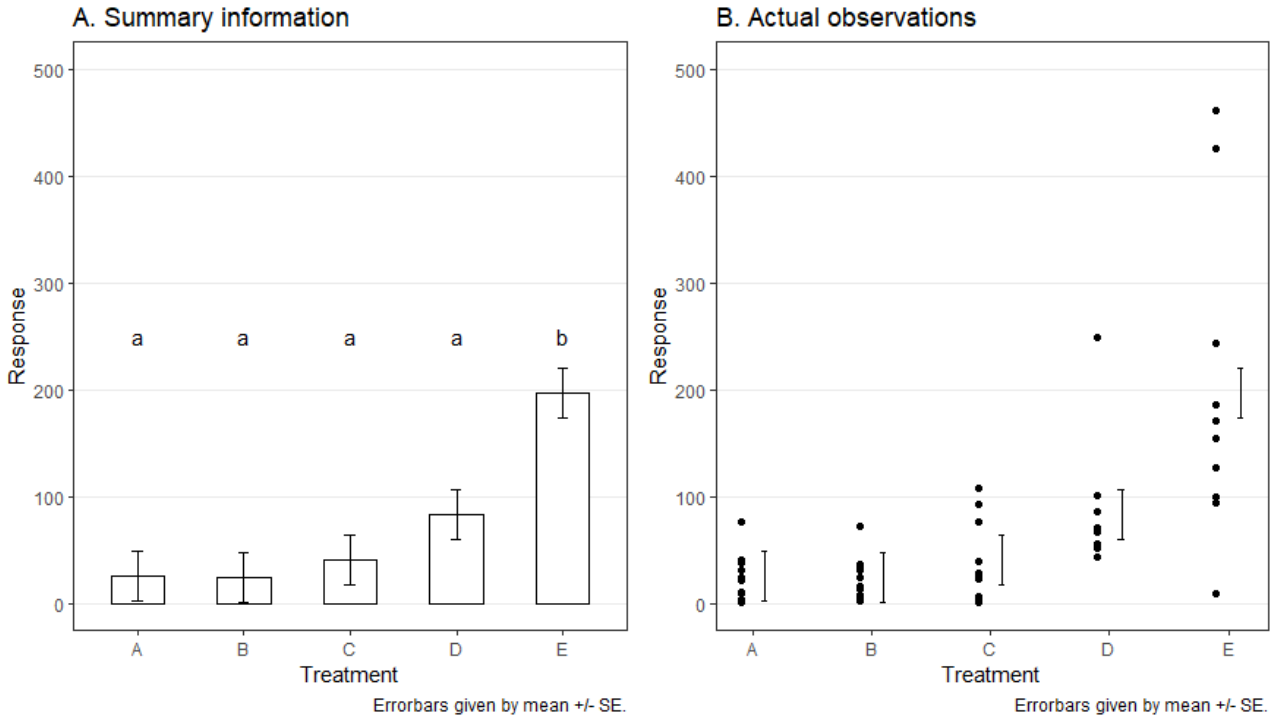
- Many alternatives to making code available
  - Article appendix in the journal
  - Free online repositories (GitHub)
  - University personell pages

Mays: Figures illustrating actual data

- Graphics offer higher information density

Clear non-normality

and heteroscedasticity



Effect sizes

- The estimated magnitude of effect *must* be given in *some* form
- Typically presented as estimated means per treatment
- Consider including some standard measure, such as Cohen's d

$$d = \frac{\widehat{\mu_1} - \widehat{\mu_2}}{s}$$

- Simplifies comparisons between studies

Misses: Hypotheses

- Null Hypothesis Significance Tests (NHST)
  - The idea with NHST is to
    first formulate the null hypothesis and the alternative hypothesis,
    then collect the data.
  - HARKing*:* **H**ypothesizing **A**fter the **R**esults are **K**nown.

- Sometimes you don't want to reject the null hypothesis!
  - When you test if the data is normally distributed.
  - When you test if the variances are homogenous (= homoscedastic).
  - Perhaps also when you test for interaction.

Misses: Null hypothesis

- What is the null hypothesis in Bartlett's test?

RESULTS

Analysis of variance

Bartlett's test suggested ($\chi^2 = 28{\cdot}2$, $P < 0{\cdot}05$) that error variances were homogeneous. The combined ANOVA

- It is not possible to check what is wrong because the degrees of freedom is not given.

Misses: p-value calculations

- Standardized result presentation

$$t(10) = 2.34,\ p = 0.041$$

- Easy to check if the p-value matches the t-value
  - You can see if it is a one-sided or two-sided.

One-sided or two-sided test?

Length of hospitalization was defined as day of surgery to day of discharge. These data were assumed to be only ordinal because surgery was performed at different times of day and discharge times were somewhat different. The records showed that patients with window views of the trees spent less time in the hospital than those with views of the brick wall: 7.96 days compared with 8.70 days per patient [Wilcoxon matched-pairs signed-ranks analysis, $T(17) = 35$, $z = 1.965$, $P = 0.025$].

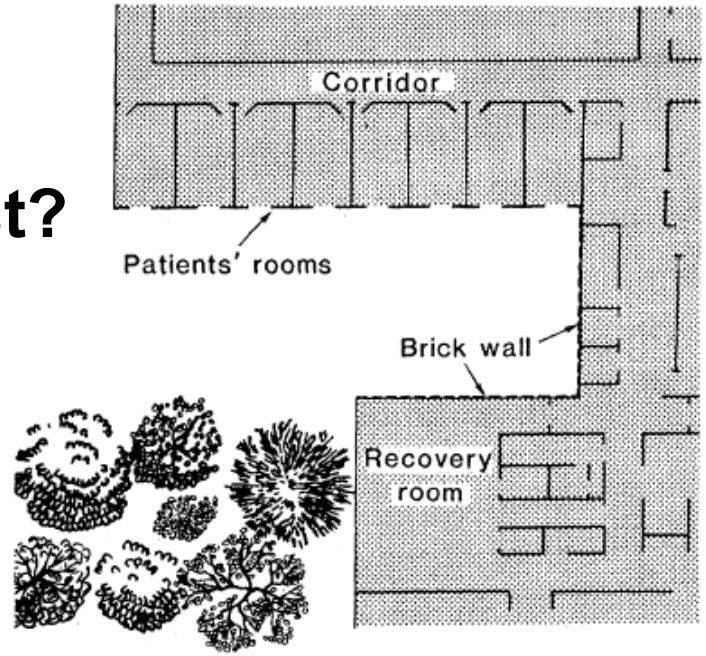
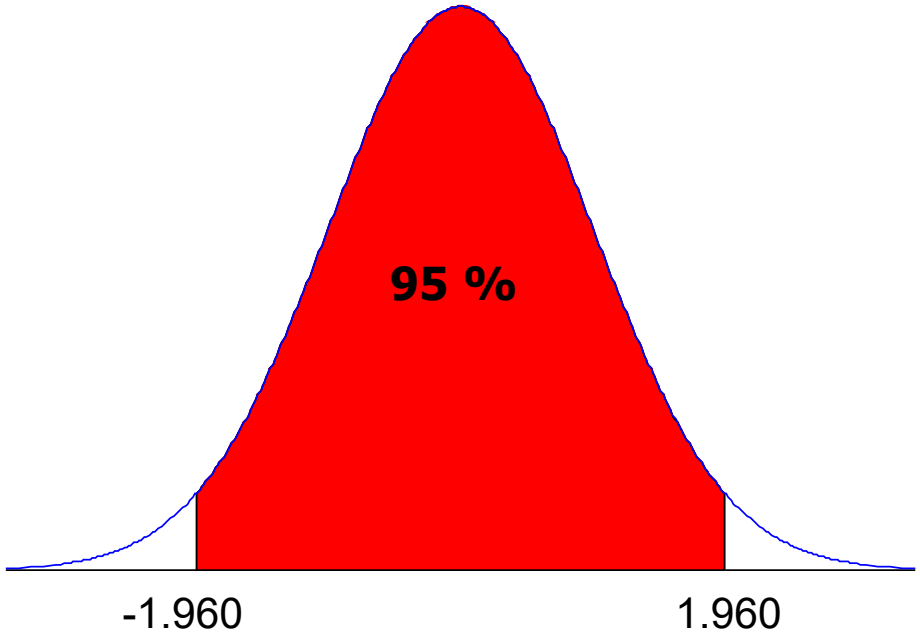


Fig. 1. Plan of the second floor of the study hospital showing the trees versus wall window views of patients. Data were also collected for patients assigned to third-floor rooms. One room on each floor was excluded because portions of both the trees and wall were visible from the windows. Architectural dimensions are not precisely to scale.

Reference (6128 citations!):

Ulrich (Science, 1984):
View Through a Window May Influence Recovery from Surgery.

Standardized normal distribution denoted *z*



Conclusion: The test in the article is one-sided!

Summary statistics

- Check for inconsistency between summary statistics and data
- Mean value calculations from sums and number of observations
- Visual checks if data shown in plots

Summary statistics: the GRIM test

- For discrete data and small sample sizes, not all decimal values are possible!
  - Example:

    If *n* = 7 and the mean is given to one decimal, the decimal must be 0, 1, 3, 4, 6, 7 or 9.

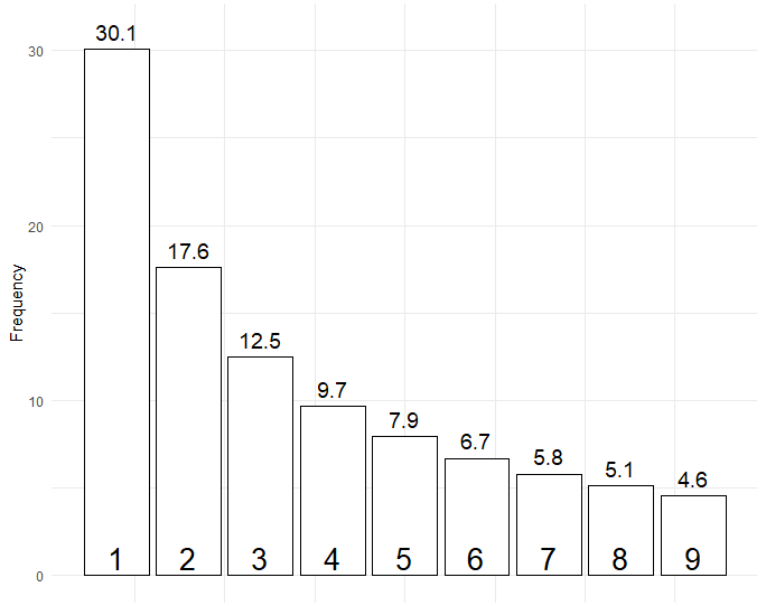
- Inconsistencies are (perhaps) common
  - Brown & Heathers, 2016, *The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology*
  - 36 faulty articles in a sample of 71

Falsified data

- Given raw data, the distribution of values might indicate fraud

- Histograms on subgroups
  - Uniform distributions indicate computer-generated data

- Benford's law
  - In large datasets with wide range, the first non-zero digit follows a specific distribution



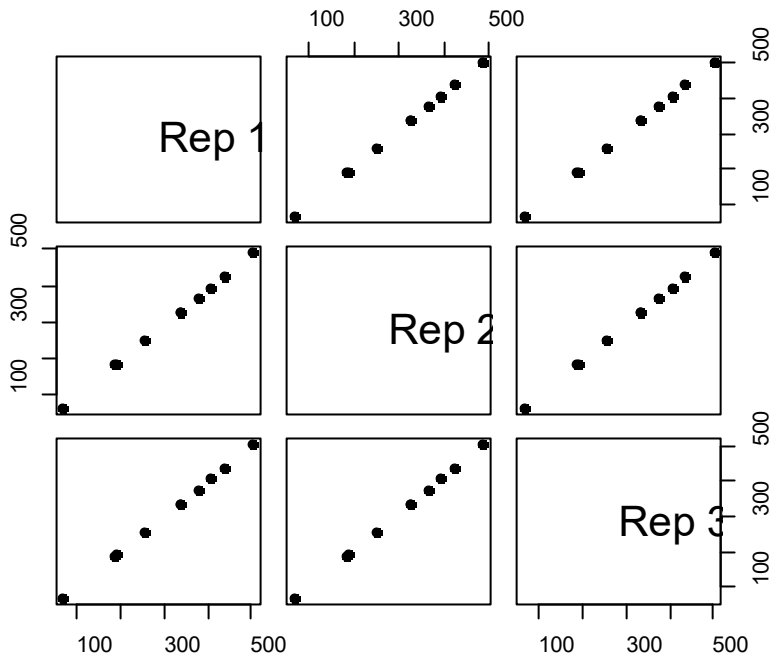
Statisticians can help you!

A true data set from a laboratory with three technical replicates (part of the data for this illustration)

Rep 1	Rep 2	Rep 3
503.94	488.82	501.96
191.51	185.76	189.53
337.27	327.15	335.29
437.27	424.15	435.29
376.66	365.36	374.68
255.75	248.08	253.77
336.95	326.84	334.97
407.27	395.05	405.29
187.27	181.65	185.29
66.97	64.96	64.99

Statisticians can help you!

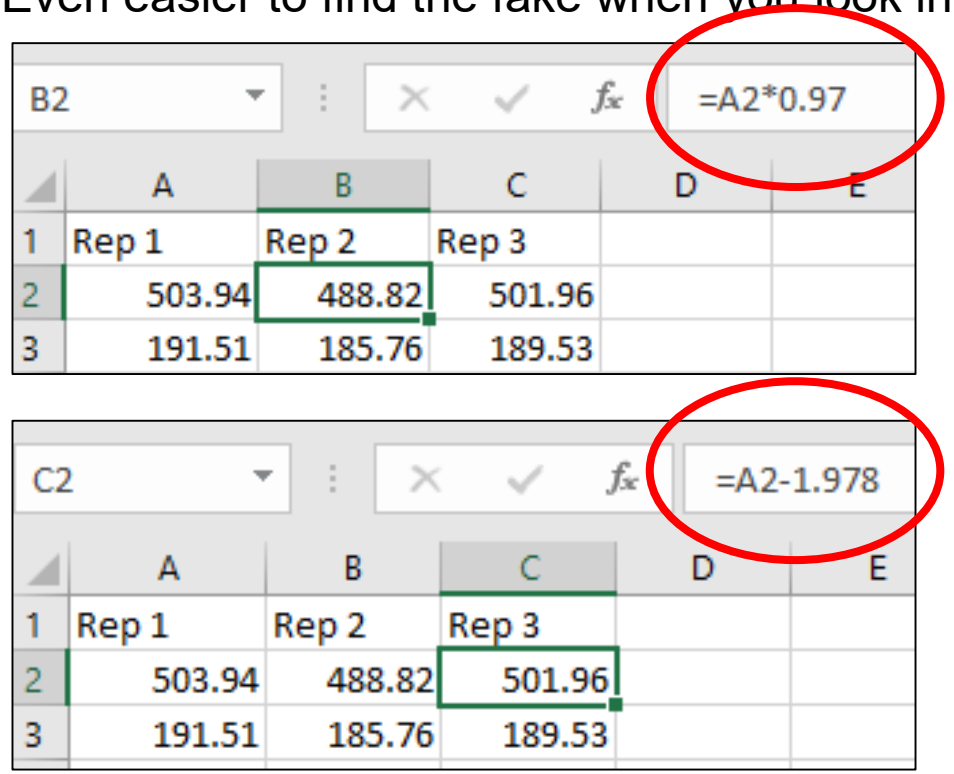
A plot to see the variation in the data:



??????

Statisticians can help you!

Even easier to find the fake when you look in the Excel sheet with the data:



More is needed to deceive the statistician…

The Reproducibility Crisis

- Is Mendel’s experiment correct?
  - Fisher asserted that "the data of most, if not all, of the experiments have been falsified so as to agree closely with Mendel's expectations“
- Reproducibility of 100 studies in psychological science (2015):
  - Overall, 36% of the replications yielded significant findings (*p* value below 0.05) compared to 97% of the original studies that had significant effects.

Driving factors

- File drawer problem
  - Positive results are more likely to be written.
- Publication bias
  - Significant results are more likely to be published.

    Wikipedia: ”Statistically significant results are three times more likely to be published than papers with null results.”

    Furthermore: Significant results are more likely to be cited.
- Multi-response experiments
  - In larger experiment with multiple response variables,

    some significance is very likely (*HARK!*)

The state of significance tests

- The American Statistical Association statement on p-values (2016)
  - A clarification on the use of p-values
- The American Statistician vol 73 (2019)
  - Special issue on p-values and statistical significance

The six ASA statements for p-values

1. p-values can indicate how incompatible the data are with a specified statistical model.
2. p-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

American Statistician vol 73

- 43 discussion articles on the use of and alternative to NHST

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

- Evaluate based on experimental setup, not the outcome

- Power analysis before data collection

Still not significant

What to do if your p-value is just over the arbitrary threshold for 'significance' of p=0.05?

Still not significant

So if your p-value remains stubbornly higher than 0.05, you should call it 'non-significant' and write it up as such. The problem for many authors is that this just isn't the answer they were looking for: publishing so-called 'negative results' is harder than 'positive results'.

a borderline significant trend (p=0.09)

a certain trend toward significance (p=0.08)

a clear tendency to significance (p=0.052)

a clear trend (p<0.09)

a clear, strong trend (p=0.09)

a considerable trend toward significance (p=0.069)

a decreasing trend (p=0.09)

a definite trend (p=0.08)

slightly non-significant (p=0.10)

slightly significant (p<0.1)

ally significant (p=0.059)

significance (p>0.10)

weakened..significance (p=0.06)

weakly non-significant (p=0.07)

weakly significant (p=0.11)

weakly statistically significant (p=0.0557)

well-nigh significant (p=0.11)

Statistics in practice

Three blocks and eight precrops.

The response is the difference between two treatments.

(Data from Maria Ernfors)

Precrop	Block	A	B	Plants
Majs	1	13680	12385	104
Havre	1	10831	10742	112
Korn_Raj	1	13760	12911	116
Korn	1	8678		92
Hostraps	1	4805	7206	76
Hostrag	1	9446	10115	100
Hostvete	1	4449	6727	104
Lucern	1	7793	8095	116
Majs	2	13363	12160	120
Havre	2	13943	11648	108
Korn_Raj	2	12223	9254	88
Korn	2	7177	7531	112
Hostraps	2	9207	8470	96
Hostrag	2	8169	8575	112
Hostvete	2	9491	10370	100
Lucern	2	10160	10931	108
Majs	3	11385	8730	100
Havre	3	9595	5520	124
Korn_Raj	3	8850	7772	100
Korn	3	10323	10302	132
Hostraps	3	8893	9291	96
Hostrag	3	8493	8738	96
Hostvete	3	9927	9605	112
Lucern	3	10398	12399	116

By a mistake by me when copying the data, the missing value for B was copied as 92 due to a missing tab.

Does is matter?

Certainly it does!

Precrop	Block	A	B	Plants
Majs	1	13680	12385	104
Havre	1	10831	10742	112
Korn_Raj	1	13760	12911	116
Korn	1	8678	92	
Hostraps	1	4805	7206	76
Hostrag	1	9446	10115	100
Hostvete	1	4449	6727	104
Lucern	1	7793	8095	116

Musts
Mays
Misses
Musings

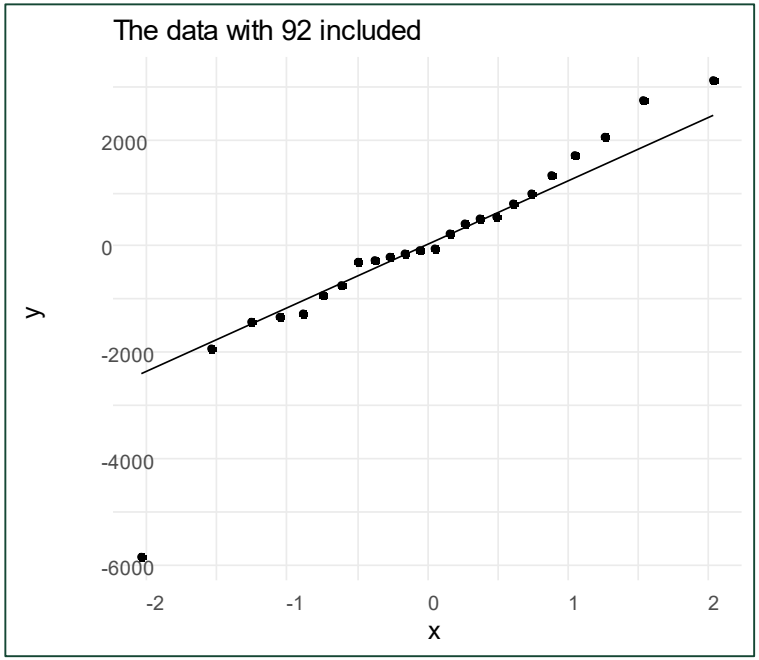
Statistics in practice

```
Type III Analysis of Variance Table with Kenward-Roger's method
          Sum Sq Mean Sq NumDF DenDF F value Pr(>F)
Precrop 51306916 7329559     7    14  1.6018 0.2144
```

```
        Shapiro-Wilk normality test

data:  resid(mod.wrong)
W = 0.9039, p-value = 0.02606
```

The model with normal distribution is not correct and the result should perhaps not be published.



Precrop	Block	A	B	Plants
Majs	1	13680	12385	104
Havre	1	10831	10742	112
Korn_Raj	1	13760	12911	116
Korn	1	8678	NA	92
Hostraps	1	4805	7206	76
Hostrag	1	9446	10115	100
Hostvete	1	4449	6727	104
Lucern	1	7793	8095	116

Musts
Mays
Misses
Musings

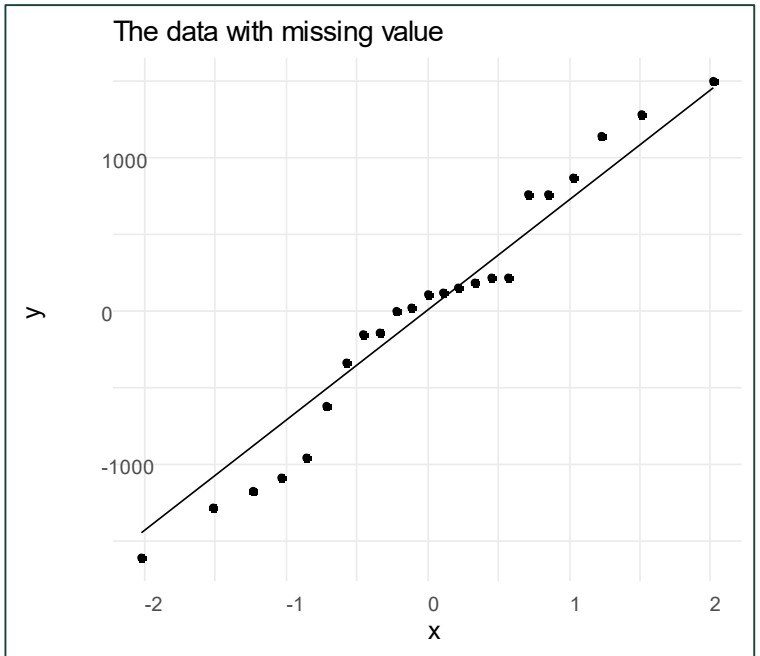
Statistics in practice

```
Type III Analysis of Variance Table with Kenward-Roger's method
          Sum Sq Mean Sq NumDF  DenDF F value   Pr(>F)
Precrop 37176567 5310938     7 13.048  4.6361 0.008345 **
```

```
        Shapiro-Wilk normality test

data:  resid(mod.correct)
W = 0.96391, p-value = 0.5467
```

The correct data has a model where normal distribution is OK and the result is ** significant according to R.



Precrop	Block	A	B	Plants
Majs	1	13680	12385	104
Havre	1	10831	10742	112
Korn_Raj	1	13760	12911	116
Korn	1	8678	**4500**	
Hostraps	1	4805	7206	76
Hostrag	1	9446	10115	100
Hostvete	1	4449	6727	104
Lucern	1	7793	8095	116

Musts
Mays
Misses
Musings

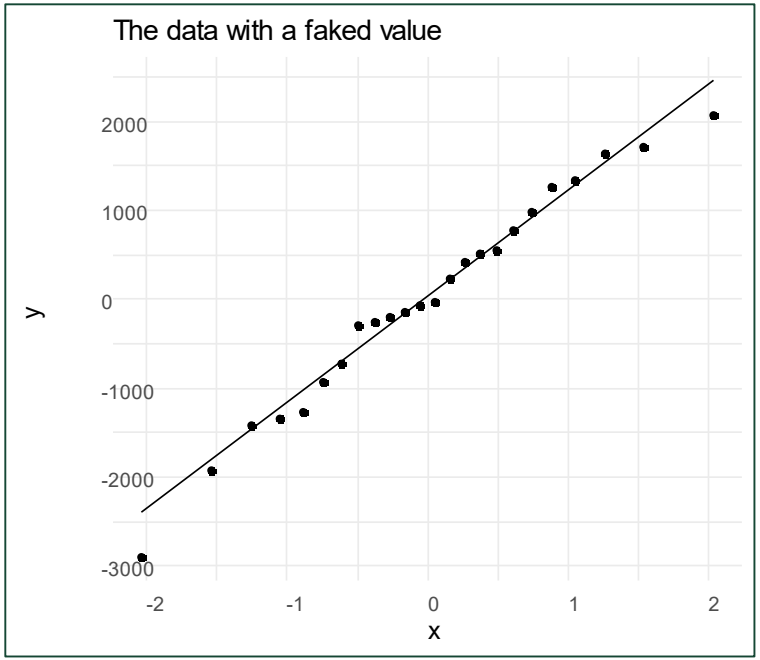
Statistics in practice

```
Type III Analysis of Variance Table with Kenward-Roger's method
          Sum Sq Mean Sq NumDF DenDF F value  Pr(>F)
Precrop 38403965 5486281     7    14  2.5279 0.06622 .
```

```
        Shapiro-Wilk normality test

data:  resid(mod.faked)
W = 0.97946, p-value = 0.885
```

If the inserted value should be 4500 and not 92, the model is OK and the result should be announced as no significant difference between the precrops.



Conclusion – What is the problem?

- From the Material and Methods it is impossible to reproduce the experiment.
  - No discussion about the method used but only on significance and results.
  - Scripts are not published.
  - Statistical packages can produce different results.
- The reviewer reviews only the parts within his/hers research area.
- Too much focus on the number of published papers.
- ***Not enough time and credit for the reviewer!***

Thank you, and

beware out there!

Adam.Flohr@SLU.SE

Jan-Eric.Englund@SLU.SE