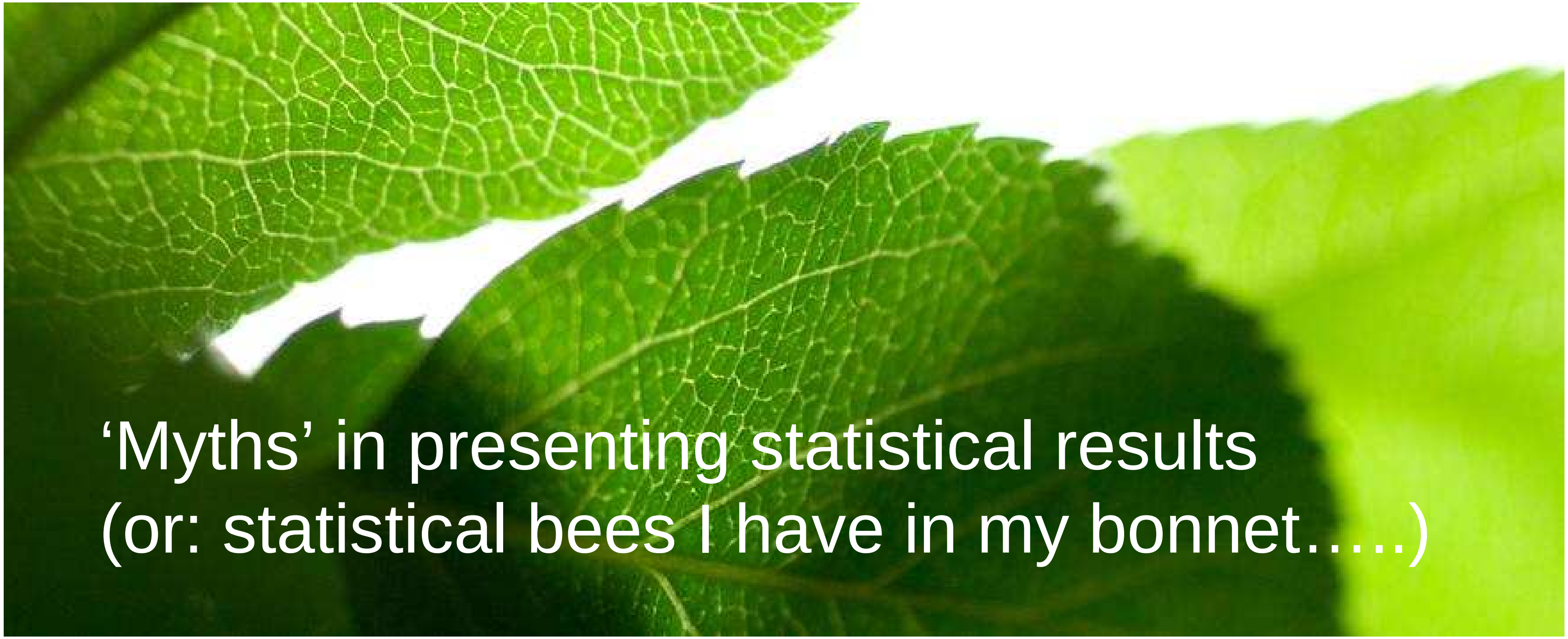
Plant & Food RESEARCH
RANGAHAU AHUMĀRA KAI

The New Zealand Institute for Plant & Food Research Limited



‘Myths’ in presenting statistical results (or: statistical bees I have in my bonnet…..)

Ruth Butler

PSYLLIDS: A busy table: Psyllid development

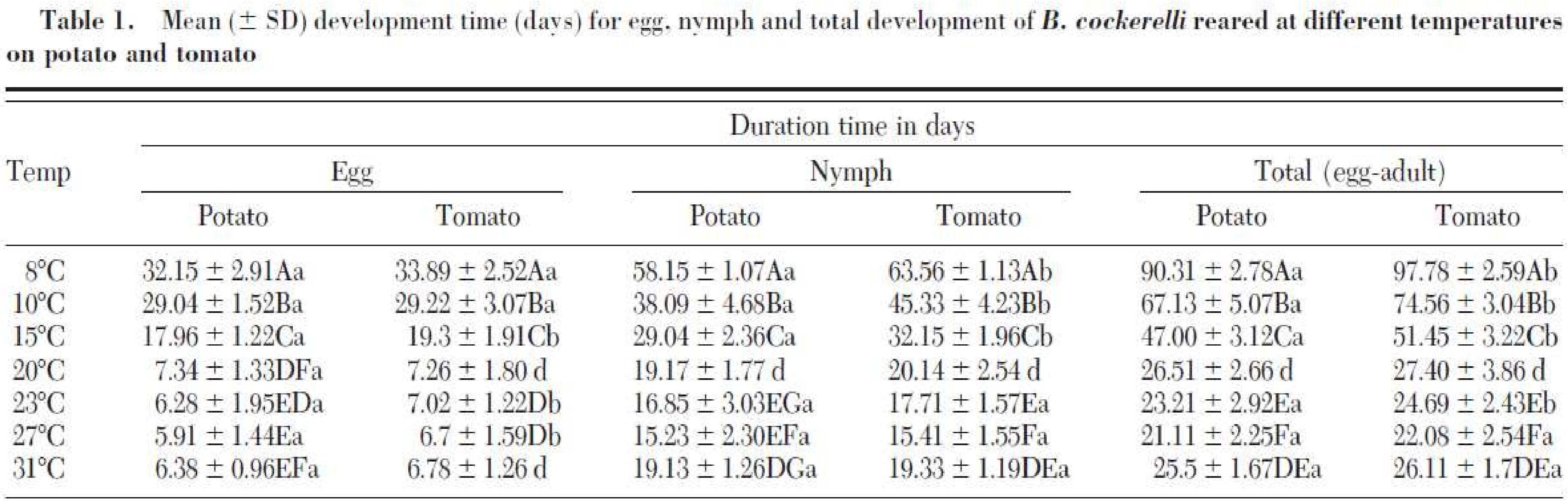


Table 1. Mean (± SD) development time (days) for egg, nymph and total development of ***B. cockerelli*** reared at different temperatures on potato and tomato

Temp	Duration time in days					
	Egg		Nymph		Total (egg-adult)	
	Potato	Tomato	Potato	Tomato	Potato	Tomato
8°C	32.15 ± 2.91Aa	33.89 ± 2.52Aa	58.15 ± 1.07Aa	63.56 ± 1.13Ab	90.31 ± 2.78Aa	97.78 ± 2.59Ab
10°C	29.04 ± 1.52Ba	29.22 ± 3.07Ba	38.09 ± 4.68Ba	45.33 ± 4.23Bb	67.13 ± 5.07Ba	74.56 ± 3.04Bb
15°C	17.96 ± 1.22Ca	19.3 ± 1.91Cb	29.04 ± 2.36Ca	32.15 ± 1.96Cb	47.00 ± 3.12Ca	51.45 ± 3.22Cb
20°C	7.34 ± 1.33DFa	7.26 ± 1.80 d	19.17 ± 1.77 d	20.14 ± 2.54 d	26.51 ± 2.66 d	27.40 ± 3.86 d
23°C	6.28 ± 1.95EDa	7.02 ± 1.22Db	16.85 ± 3.03EGa	17.71 ± 1.57Ea	23.21 ± 2.92Ea	24.69 ± 2.43Eb
27°C	5.91 ± 1.44Ea	6.7 ± 1.59Db	15.23 ± 2.30EFa	15.41 ± 1.55Fa	21.11 ± 2.25Fa	22.08 ± 2.54Fa
31°C	6.38 ± 0.96EFa	6.78 ± 1.26 d	19.13 ± 1.26DGa	19.33 ± 1.19DEa	25.5 ± 1.67DEa	26.11 ± 1.7DEa

Means within the same temp and the same insect stage followed by the same small letter are not significantly different. Means followed by the same capital letter on the same plant in the same column are not significantly different (P<0.05, Tukey HSD).

2 plant species x 7 temperatures: factorial

2 replicates (20 insects per rep, mean per rep analysed)

3 variables analysed (Egg, Nymph, Egg+Nymph)

Mean±SD plus Tukey letters (apparently, ANOVA not done)

Tran, L.T., Worner, S.P., Hale, R.J. & Teulon, D.A.J. 2012. Estimating Development Rate and Thermal Requirements of Bactericera cockerelli (Hemiptera: Triozidae) Reared on Potato and Tomato By Using Linear and Nonlinear Models. *Environmental Entomology* ***41(5), 1190-1198.***

BEES: Typical barchart: Zinc nutrition and honey bees

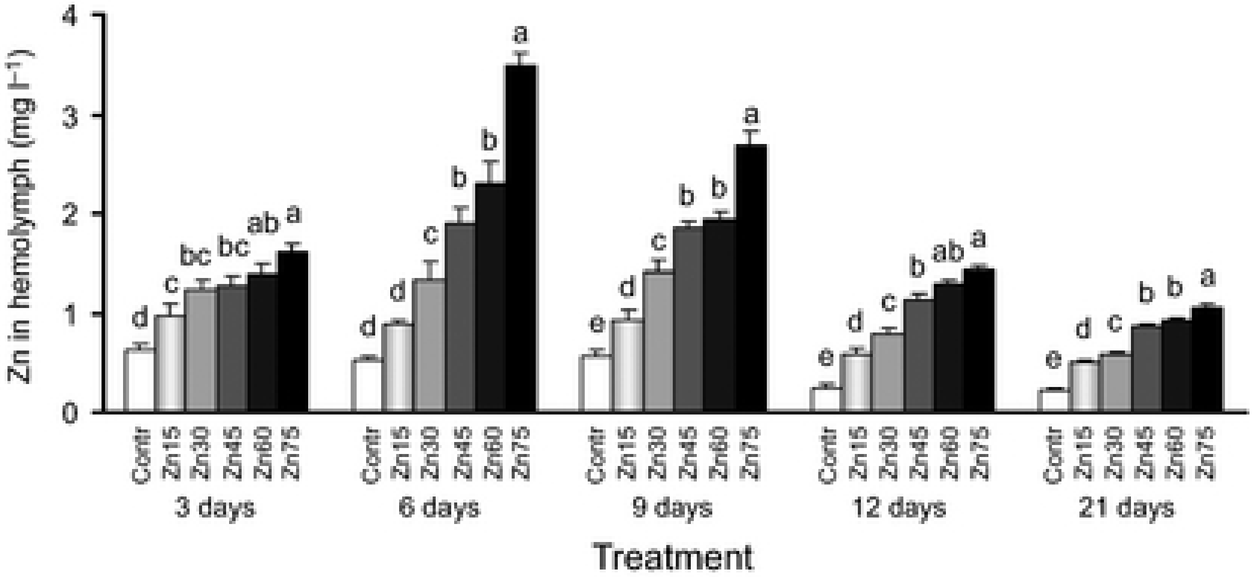


Figure 2. Mean (+ SEM; n = 3) Zn concentration in the hemolymph of captive reared worker bees at 3, 6, 9, 12, and 21 days of age measured in three samples of 40 bees from each treatment (see Table 1). Means within treatment capped with different letters are significantly different (Duncan's test: P<0.05).

6 Zinc levels x 5 times: factorial, two quantitative factors
3 replicates

ANOVA for each time separately, then Duncan's letters
Individually calculated means and s.e.m.

Entomologia Experimentalis et Applicata
Volume 156, Issue 3, pages 201-210, 29 JUL 2015 DOI: 10.1111/eea.12342
http://onlinelibrary.wiley.com/doi/10.1111/eea.12342/full#eea12342-fig-0002

BEES: Textual summary from the paper:

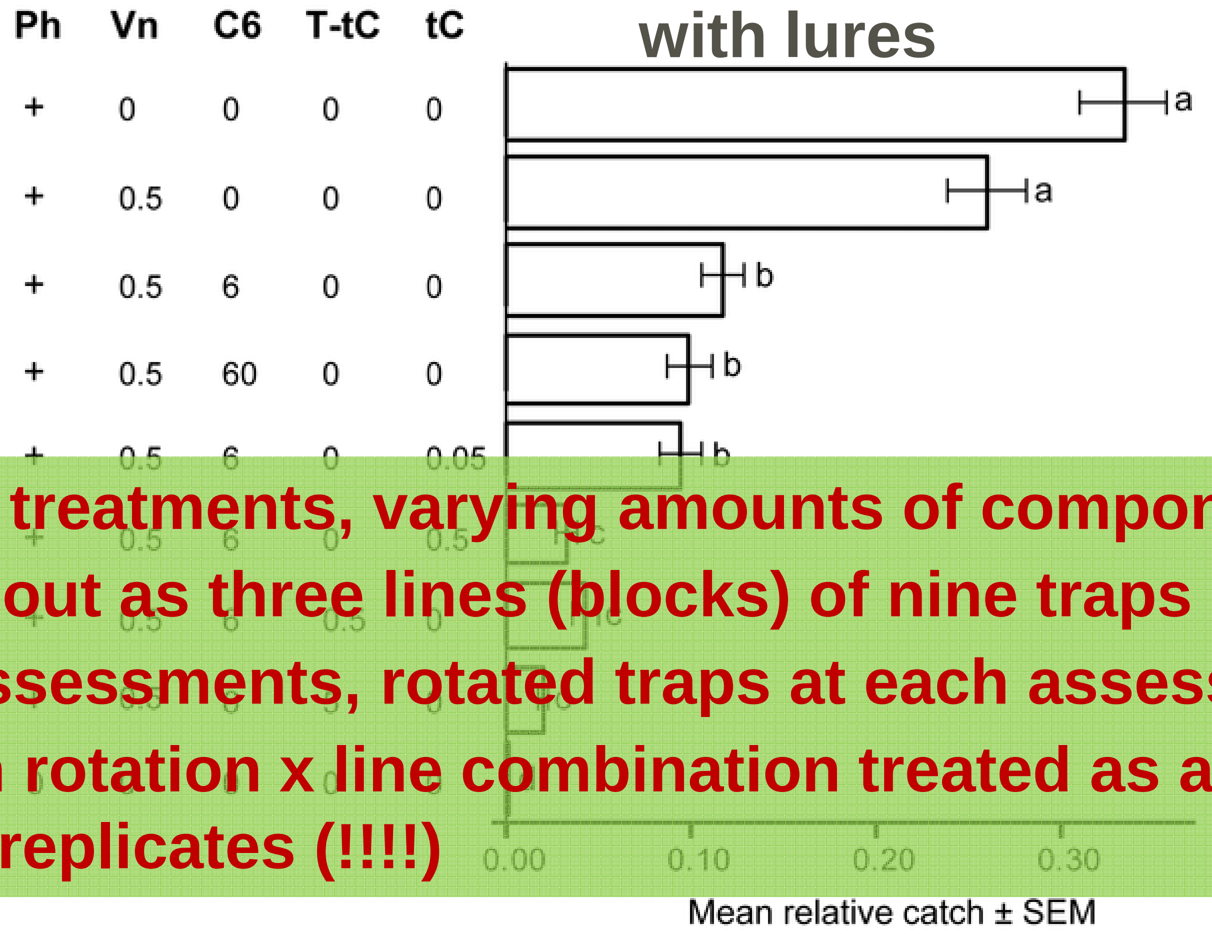
No difference in Cu/Zn-SOD activity among treatment groups was apparent in 3-day-old bees (ANOVA: $F_{5,12} = 1.12$, $P>0.05$; Figure 3A), but the Cu/Zn-SOD activity of 6- and 9-day-old bees on the Zn30 diet was higher than that of any other treatment group (6-day-old bees: $F_{5,12} = 15.28$; 9-day-old bees: $F_{5,12} = 5.70$, both $P<0.05$; Figure 3A). The Cu/Zn-SOD activity of 12- and 21-day-old bees in the Zn30 treatment group was also higher than that of those in the Zn60 and Zn75 treatment groups (12-day-old bees: $F_{5,12} = 5.22$; 21-day-old bees: $F_{5,12} = 3.15$, both $P<0.05$). Peak Cu/Zn-SOD activity in Zn30 treatment group was observed at 6, 9, 12, and 21 days. Lower Cu/Zn-SOD activity was observed in treatment groups receiving ≥60 mg kg^{-1} Zn.

‘Not significant’=‘No difference’!

F statistic to 2 decimal places BUT P value only >0.05 !
(P is 0.40)

Also: no real discussion of *trends* vs Zinc level

BEETLES: Another Typical barchart: Beetle trapping, traps with lures



Nine treatments, varying amounts of components Ph, Vn etc

Laid out as three lines (blocks) of nine traps

27 assessments, rotated traps at each assessment (!)

Each rotation x line combination treated as a replicate: 3x27 = 81 replicates (!!!!)

Figure 4..[snip]. Bars show mean relative catch per replicate ±1 standard error based on n = 81 (n = 80 for Vn, n = 79 for C6 at 60 mg/day) replicates. [snip]. Bars with same letters are not significantly different (ANOVA on arcsin√(relative catch) followed by Dunnet's T3, p<0.05).

Relative catch: trap catch/ total catch for 'replicate'

ANOVA of arcsin(sqrt(relative catch)) then Dunnet's

Graph: raw means plus individually calculated s.e.m. plus letters from analysis

Unelius CR, Schiebe C, Bohman B, Andersson MN, Schlyter F (2014) Non-Host Volatile Blend Optimization for Forest Protection against the European Spruce Bark Beetle, Ips typographus. PLoS ONE 9(1): e85381. doi:10.1371/journal.pone.0085381
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0085381

Good thing:

- Captions actually describe everything including how the error bars were calculated…

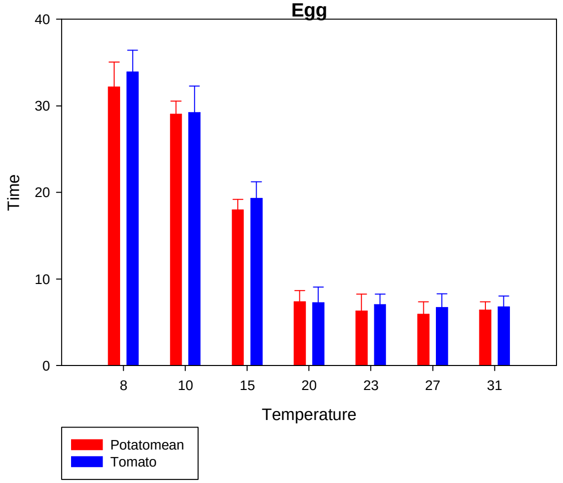
Bad things:

- Barcharts!- poor for showing treatment trends
- sd/ s.e.m./letters obscure quantitative response relationships, especially so in table
- Conclusions dominated by the ‘statistics’, little description of the *sizes* of effects or trends.
- Confusion between data summary and formal analysis. Beetles especially bad (letters *do not* relate to the means presented)

A very common process:

1. Individually calculate means, ‘s.e.m’ (SD/√n) or (rarely) SD
2. ‘Analyse’ data
   - Mostly ignore assumptions, except maybe ‘normality’
   - Sometimes transform data or use a non-parametric ‘test’
   - Often ignore/ poorly account for trial design (blocking, treatment structure)
   - Confuse pseudo replication with real replication
3. Get significance stars/ letters/ ‘n.s’.
   - Completely discard all other analysis results/ output
4. Combine 1 with 3 in the presentation
5. Report ‘it was significant’/ ‘it was not significant’

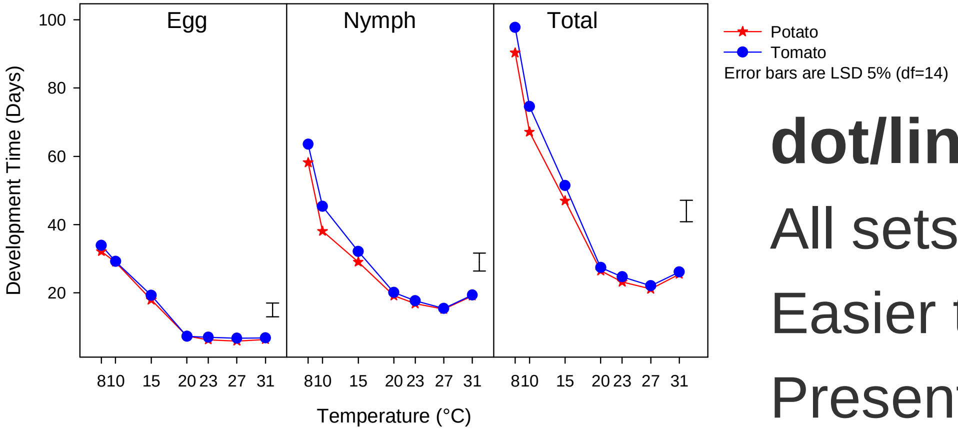
PSYLLIDS: Some suggestions:



Barchart:

Better than table. BUT:

- takes lots of space for one stage
- not easy to compare the trends.
- s.e.m. bars not that useful….



dot/line plot:

All sets shown in not much more space.

Easier to see trends, compare plants.

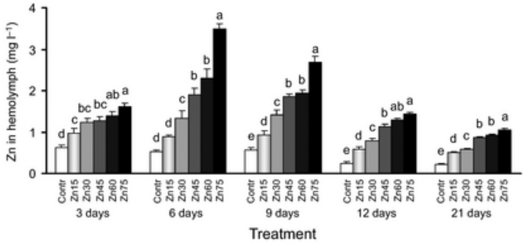
Presentation consistent with analysis

BEES and BEETLES: Some suggestions

BEES:

2-way factorial analysis & Response surface fitting

Line graphs



BEETLES:

So much wrong, hard to know what to do:

Design:

‘trap rotation’ ‘assessments->replicates’

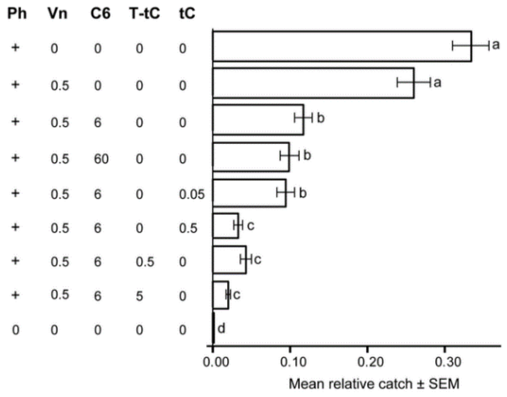
Analysis:

adjusted for block BEFORE analysis

difference between means of arcsin(√data) says NOTHING about difference between means of data!!

should have used methods appropriate for counts

ignored treatment structure …..



Things to consider: *my* bees

- WHY so common to mix summary of raw data (means, individually calculated s.e.m.) with analysis ‘results’ (p, stars, letters)?
- IF appropriate to use ANOVA, why not present a single bar/error plus means/estimates?
- Why the fixation with bar-charts?
- What is this obsession with letters anyway?
  - not appropriate for a factorial structure
  - not appropriate for quantitative factors
  - mostly just add clutter without useful information

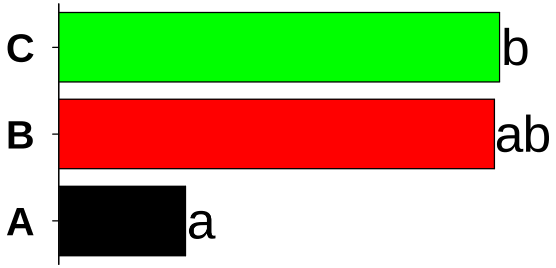
More of my bees…

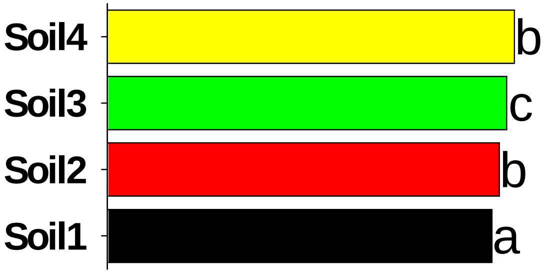
- Why the very accurate information for the test statistics (F etc) but only P>0.05 or P<0.05?
  - P>0.05 DOES NOT mean ‘no difference’/ the same!!
  - P<0.05 DOES NOT mean ‘real difference’!!
- Why do people think ‘statistics’ ↔ ‘is it different’/ ‘letters’?
- Why so little discussion of trends & patterns?

Reasons not to use Letters/ Multiple range tests

- Nothing ‘magic’ about p=0.05.
  - Letters convert sliding scale of difference into ‘falling off a cliff’ (C. Triggs)
  - p-values are only estimates (get p=0.051; ‘real’ p might be 0.049!)

- Letters can be highly unhelpful:



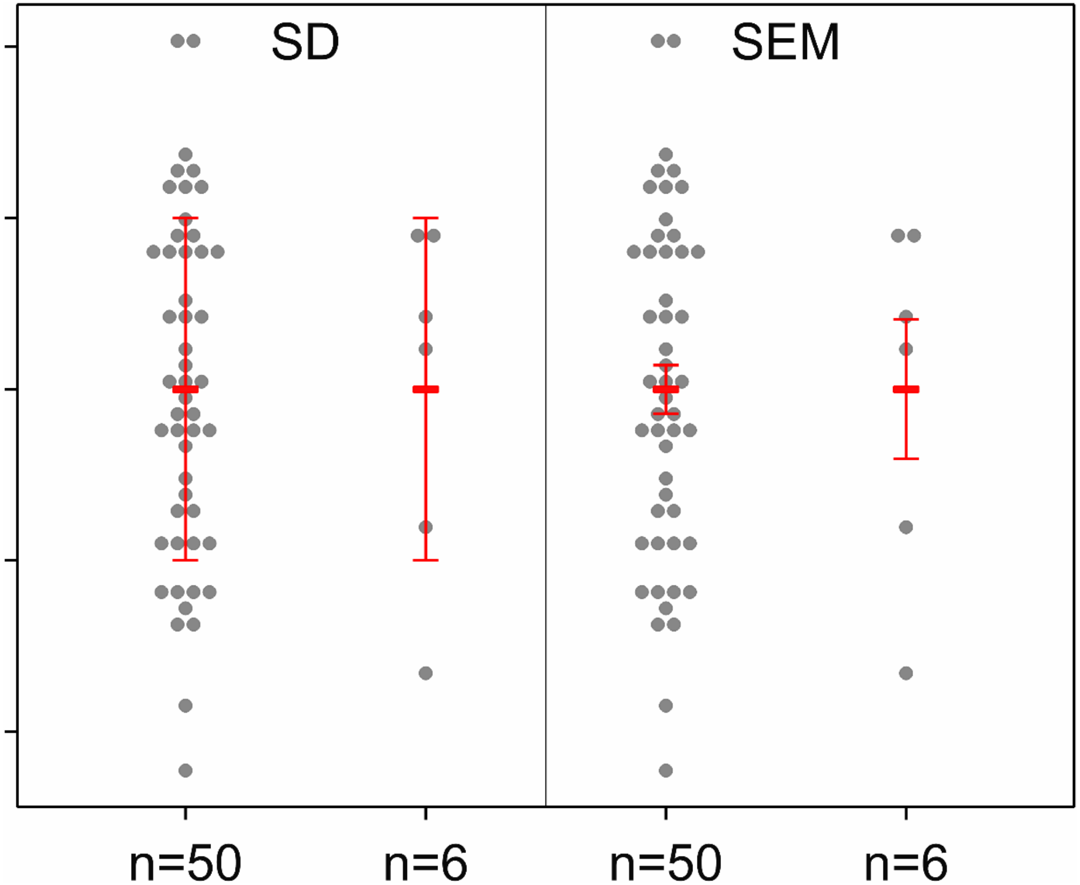


- Should not be used if there is structure-> comparisons need to be made according to the structure

…Letters / Multiple range tests

- Should not be used with quantitative factors
- Multiple comparison ‘corrections’: mostly just about moving the cut-off for deciding what is interesting. How do you choose the ‘correct’ one?
- Fisher thought that the ‘cut off’ should be decided on the basis of context- so sometimes p=0.1 might be a good choice, sometimes a smaller p. (1926)

Comparison of error bars



SD:

shows spread of data

not affected by n

SEM

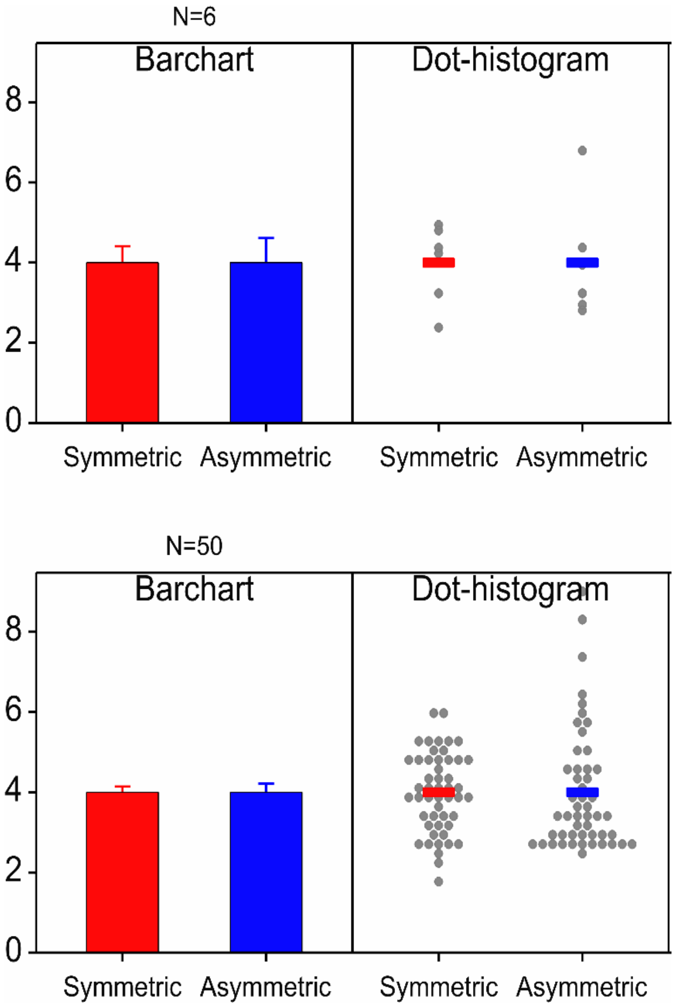
makes assumptions about data

affected by n

So WHY is SEM so popular?

Adapted from:
The Journal of Physiology
Volume 589, Issue 8, pages 1861-1863, 14 APR 2011 DOI: 10.1113/jphysiol.2011.205062
http://onlinelibrary.wiley.com/doi/10.1113/jphysiol.2011.205062/full#f2

Barcharts vs Dot-histograms



Barchart:

obscures: # values, distribution of data

Bar dominates: width, colour, shading can distract

Only important feature: top of bar

Error bars: often not properly described

Bars don't show relationships/ trends well

Dot-histogram:

each data-point shown

can add info like means

Adapted from:
The Journal of Physiology
Volume 589, Issue 8, pages 1861-1863, 14 APR 2011 DOI: 10.1113/jphysiol.2011.205062
http://onlinelibrary.wiley.com/doi/10.1113/jphysiol.2011.205062/full#f1

Where do people learn to do this ~~rubbish~~ sort of thing?

- Has it always been like this?- when did this sort of results presentation become so ubiquitous?
- Fisher vs Neyman/ Pearson -> confusions as to the meaning of significance testing?
- Has Excel contributed to the prevalence of Bar-charts (with or without errors)?
- Why do people think ‘statistics’ ⇔ ‘is it different’?
- Why do so many people only use ‘letters’, p values or ‘stars’ from an analysis and ignore all other results (estimates, s.e.s etc)?

…Where do people learn to do this ~~rubbish~~ sort of thing?

- what is the source of such ‘rules’ as:
  - ‘it’s categorical so you have to do a bar chart’!!
  - You can’t join points unless the x-axis is quantitative (even if ordered)!!
- Why do so many people think you cannot describe an effect unless it has been ‘tested’?
- Contrary to popular belief, Significant ≠ Biologically Important

Some potential answers:

- Senior scientists
  - prevalent belief: scientists can & should do all the analyses
- Poor university teaching (often by non-statisticians)
- Too few applied statisticians
- Software
  - ‘Real Statistics, Real Easy’ : SPSS ad
  - Default output
- …… any other suggestions??

- But mostly: US

Some nice quotes

- ‘To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of’ (Fisher, 1938).
- ‘The choice of how to express the data is very important and should not be made solely on the basis of habit or convention. Always inspect the data in its raw form’ (Lew, 2007).
- ‘To conclude ‘this shows that there is no difference’ here is to make perhaps one of the commonest errors in biology. A useful summary phrase is ‘absence of evidence is NOT evidence of absence’ ’ (Altman & Bland, 1995)

Bibliography

Fisher, Ronald Aylmer. 1926. “The Arrangement of Field Experiments.” Journal of the Ministry of Agriculture GB 33: 503–13

Gardner, M.J. & Altman, D.G. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ 292(6522), 746-750.

Lang T, Altman D, . (20013) Basic statistical reporting for articles published in clinical medical journals: the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook, European Association of Science Editors. http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf

Robbins, N.B. (2013) Creating More Effective Graphs, Chart House

Stern, Coe, Allan, Dale eds; (2004) Good Statistical Practice for Natural Resources Research, CABI publishing (Chapter 21).

Cleveland, W.S. (1994) The elements of Graphing Data, AT&T Bell Labs

Maindonald, J.H. & Cox, N.R. 1984. Use of statistical evidence on some recent issues of DSIR agricultural journals. New Zealand Journal of Agricultural Research 27, 597-610.

Maindonald, J. 1992. Statistical design, analysis, and presentation issues. New Zealand Journal of Agricultural Research 35, 121-141.

Goodman, S. 2008. A Dirty Dozen: Twelve P-Value Misconceptions. Seminars in hematology. Pp. 135-140.

Anderson, D.R., Burnham, K.P. & Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64(4), 912-923.

Dawkins, H.C. 1983. Multiple comparisons misused: Why so Frequently in Response-Curve studies? Biometrics 39, 789-790.

Saville, D.J. & Rowarth, J.S. 2008. Statistical Measures, Hypotheses, and Tests in Applied Research. Journal of Natural Resources and Life Sciences Education 37, 74 - 82.

Cumming, G., Fidler, F. & Vaux, D.L. 2007. Error bars in experimental biology. J. Cell Biol. 177(1), 7-11.

Drummond, G.B. & Vowler, S.L. 2011. Show the data, don't conceal them. *The Journal of Physiology* ***589(8), 1861-1863.***

Altman, D.G. & Bland, J.M. 1995. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* ***311(7003), 485.***

Plus many more….