Reading and Interpretation of the Ten Statistical Commandments of Chairman Alroy

Peter D Smits [1] [2]

[1]School of Biological Sciences
Monash University

[2]Committee on Evolutionary Biology
University of Chicago

August 7, 2012

John Alroy

- quantitative paleobiologist currently at Macquarie
  - diversity curves
  - PaleoDB
- teaches quantitative methods in (paleo)biology workshop

Background

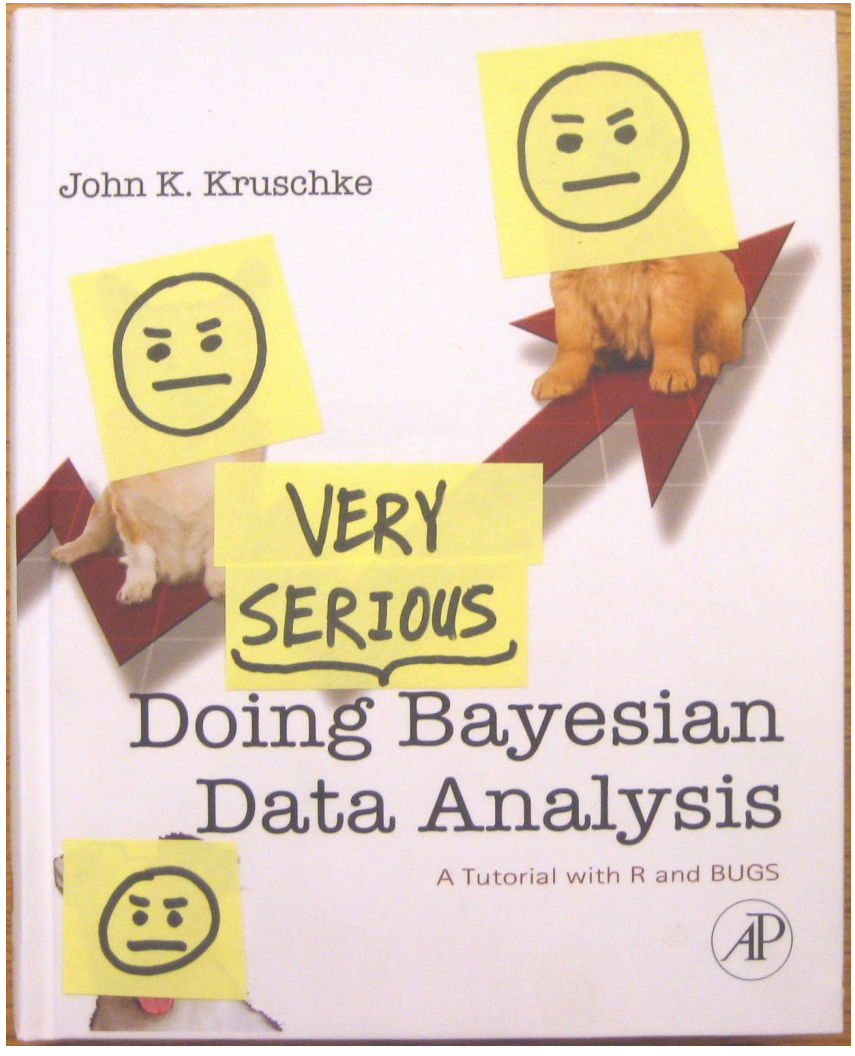
John Alroy is (in)famous for getting extremely angry about the use of statistical methods in paleontology (and biology in general.)

Up until about 2-3 years ago he had a list of 10 statistical commandments on his NCEAS website. Currently he only lists an abbreviated version.

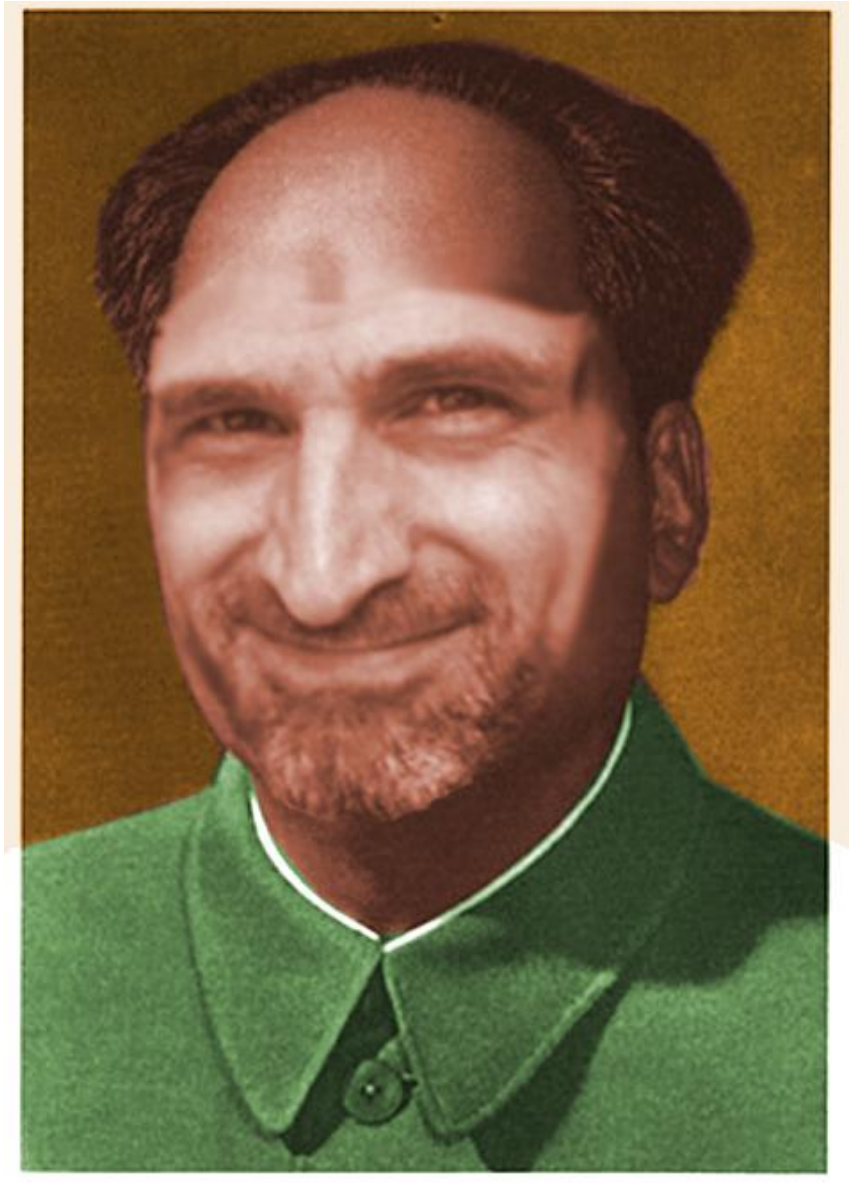
They are one of the few things I have displayed at my desk.

While the content is serious, the presentation is parody.

The other picture at my desk. . .



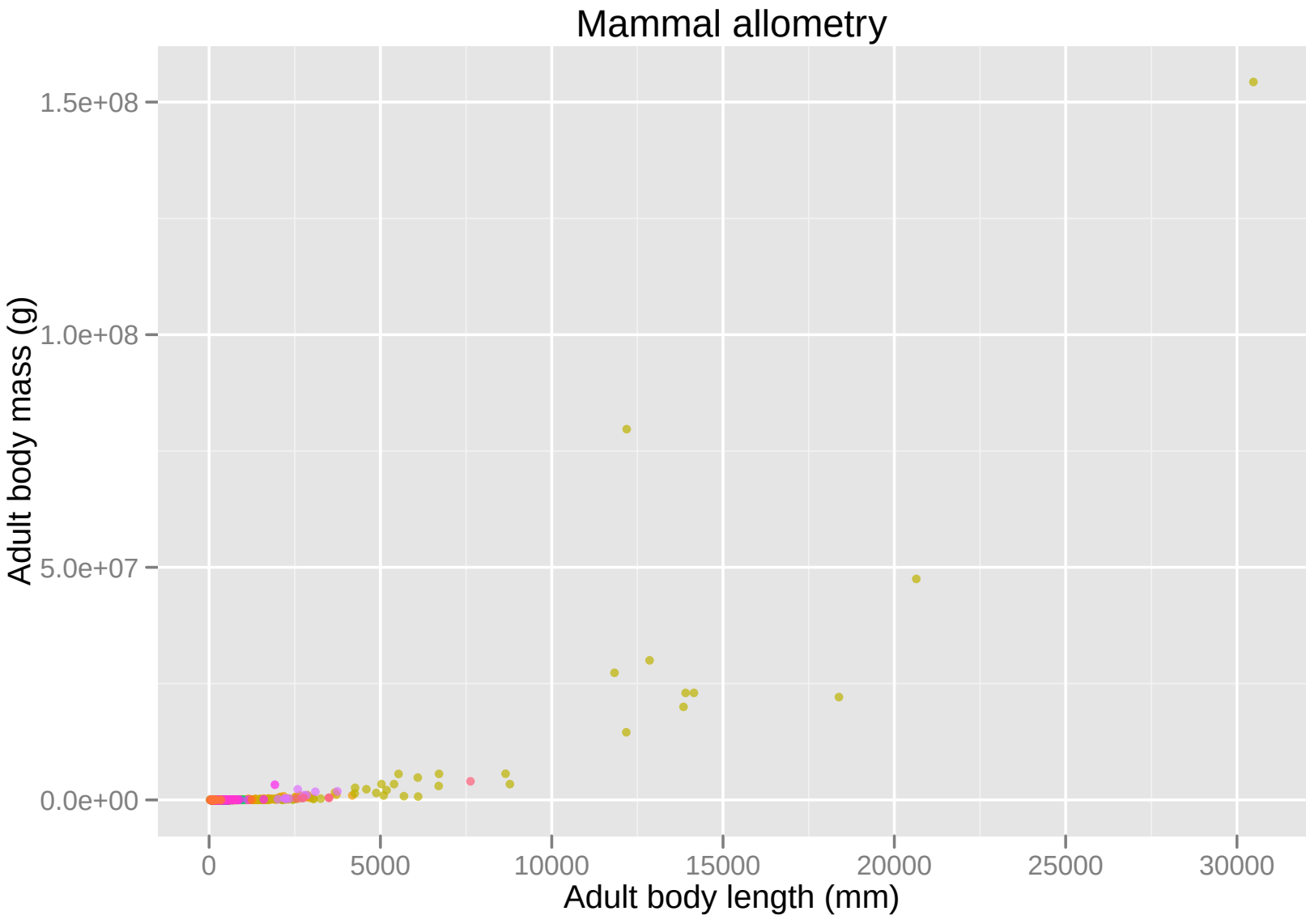
And Our Beloved Chairman Alroy said . . .



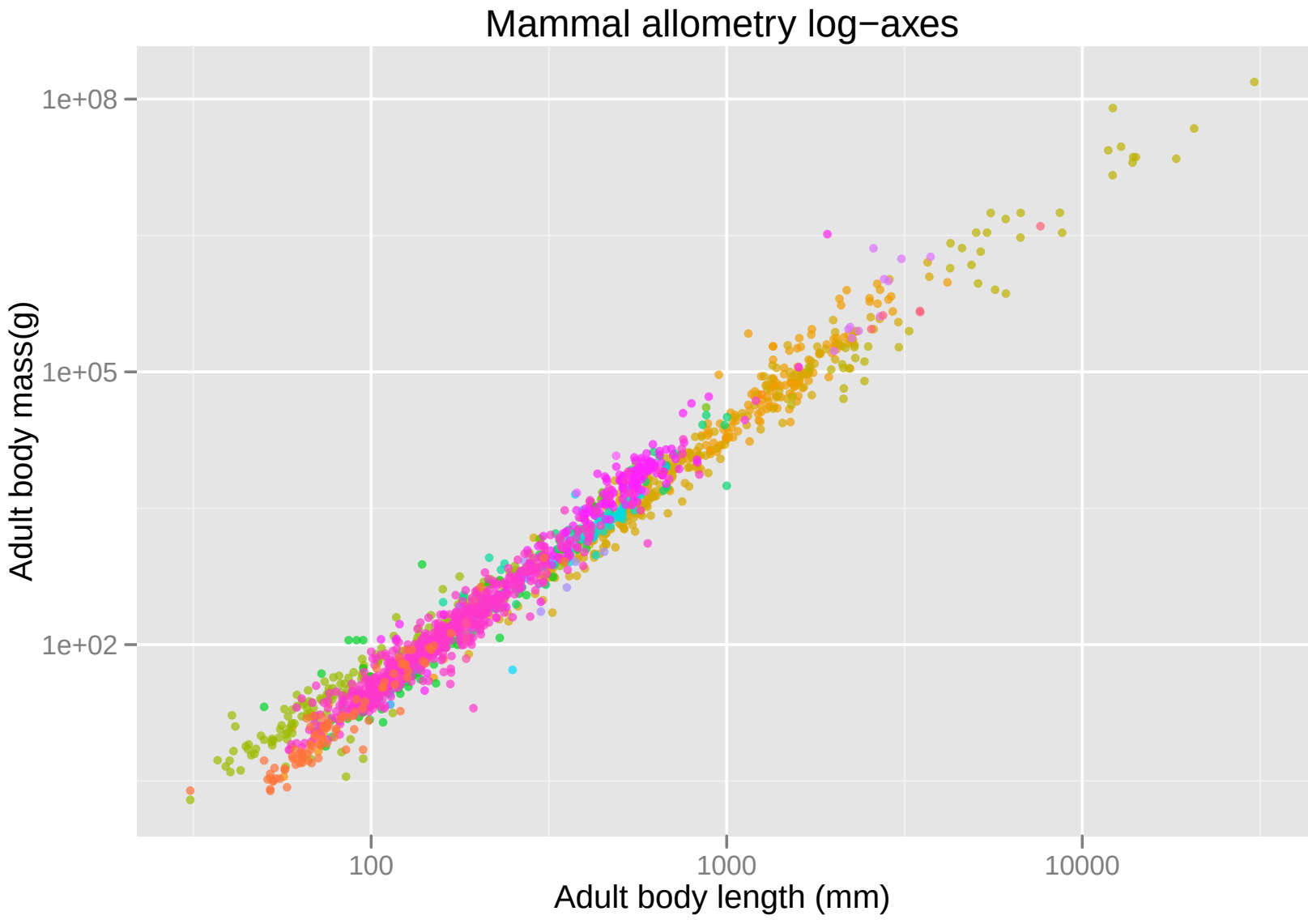
Commandment 1

Thou shalt log thy data! We live in a multiplicative world, which means our data live in a log world. Always log any data with a lower zero bound, unless there's also an upper bound, in which case that shalt perform a logit transformation. *Log until proven linear, and be holy.*

Monotonic transform



Monotonic transform



Information and log

Imagine we have a device that can give us one of three items at a time:
A B C

Everytime we get something from the device, we receive information and our uncertainty decreases.

We can say this device as ‘‘uncertainty 3’’

Information and log

Now imagine a second device that gives one of two items at a time (uncertainty 2):
Y Z

If we put both devices together, we have six possibilities:
AY AZ BY BZ CY CZ

Even though we have two "devices."

The easy way to just do this is take the log of the uncertainty of both "devices." Can contiue from here to explain Shannon's entropy (I won't).

$\log 3 + \log 2 = \log 6$

Commandment 2

Thou shalt run non-parametric tests! If the parametric and non-parametric tests come out the same, thou hast lost nothing. If they don't, the data are non-normal, the parametric test is wrong, and thou shalt use the non-parametric result. Spearman, Mann-Whitney, and Kolmogorov-Smirnov are the Holy Trinity (or Quintinity, or whatever). Worship them!

Non-parametric tests

Parametric tests have a number of assumptions (normality). Even the basic Pearson's product-momment of correlation (r) assumes normality.

Frequently, these assumptions are either un-tested (homogeneity of variance) or violated (independence).

Very rarely in real systems can most tests be applied. Life doesn't always fit in that box.

Non-parametric tests avoid (most) of the problems revolving around distributions.

Non-parametric tests

Non-parametric statistics, specifically the tests, come in two flavours

- distribution free
- non-parametric (I love tautologies)

Some examples...

Non-parametric tests

Spearman rank order coeffecient rho (ρ) and Kendall's tau (τ)

- similar to r but is based on ranks and, thus, distribution free
- choice of ρ or τ depends mostly on sample size

Wilcox signed-rank test and Mann-Whitney U

- similar to t-test but is based on ranks and is for difference of medians (not means).
- Wilcox is an alternative to paired t-test
- Mann-Whitney is an alternative to two-sample t-test.

Kolmogorov-Smirnov test

- Did a sample come from a specific probability distribution (one-sample) or did two samples come from the same probability distribution?
- the two-sample is really unique and incredibly useful

Resampling methods

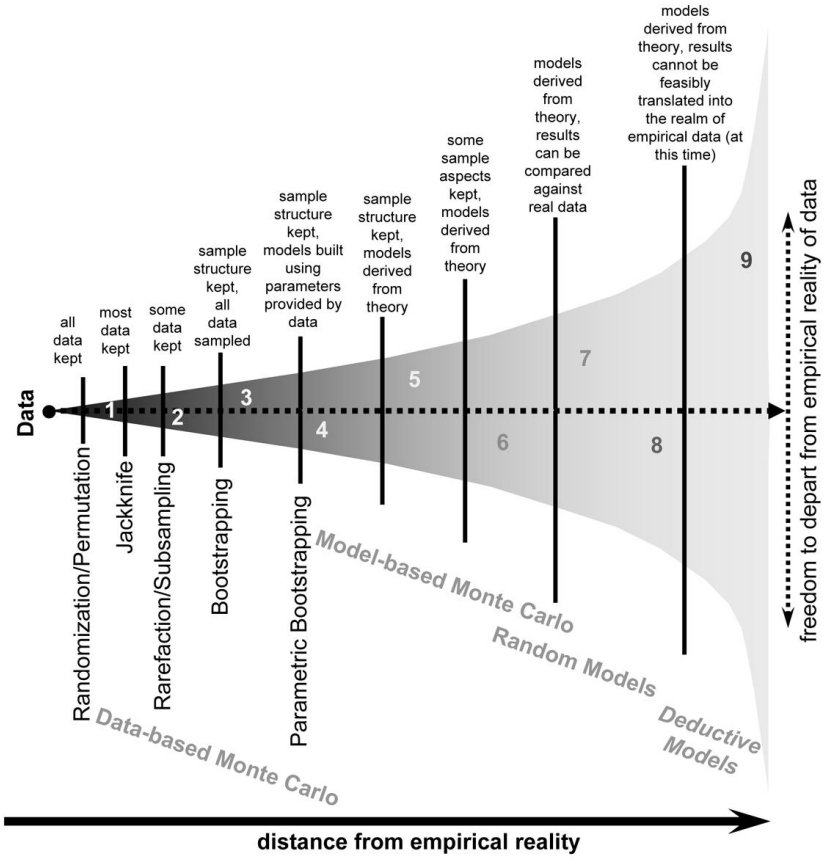
Missing from the commandment are resampling methods.

Resampling methods (can) have even fewer assumptions than non-parametric methods!

In general, you just assume your sample is representative of the population which we do this almost implicitly.

Resampling methods are various classes of Monte Carlo methods (randomized) and are generally considered computer intensive.

Resampling methods



from Kowalewski and Novack-Gottshall 2010

Commandment 3

Thou shalt disdain p-values! $p = 0.05$ is a heathen idol, and ANOVAs are for those who have not yet seen the light, still dwelling in the darkness of obsessive frequentist hypothesis testing. Remember, *if thou hast enough data anything will turn significant, no matter how small the difference.* And the ‘‘significance level’’ is whatever thou choosest it to be, not what someone tells thee it should be. So, describe data, don't just test data. Don't merely ask *whether* there's a significant difference, ask *what* is the difference, *why* is there a difference, and *have I confidence* in that difference?

Definition

A p-value is the probability of a result given that the null is true.
Many misconceptions and problems

- statement of $P(x|H0)$, not $P(H1|x)$ or $P(H0|x)$
- trivial null hypotheses abound
- false significant/not significant binary
- may reflect large sample size rather than anything meaningful
- etc. . .

p-values

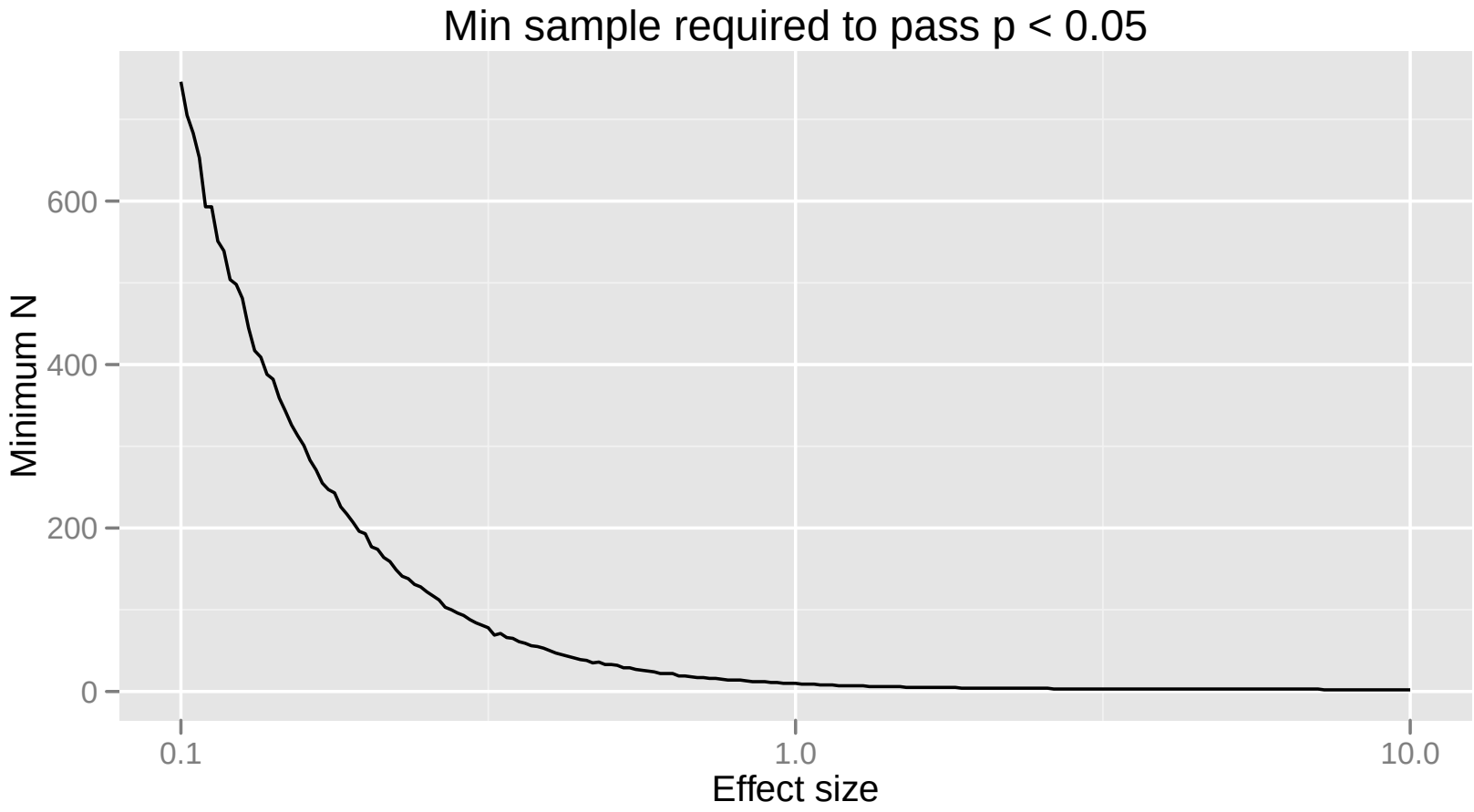
p-values are the target of near constant attack.

Part of this has to do with confusing Fisher's original definition and Neyman-Pearson Type I and Type II error rates (I won't discuss this).

John Myles White (sociology grad-student and consumate Bayesian) has been producing a nice series about the problems of the NHST paradigm that touches on the above points: worker effort and effect size.

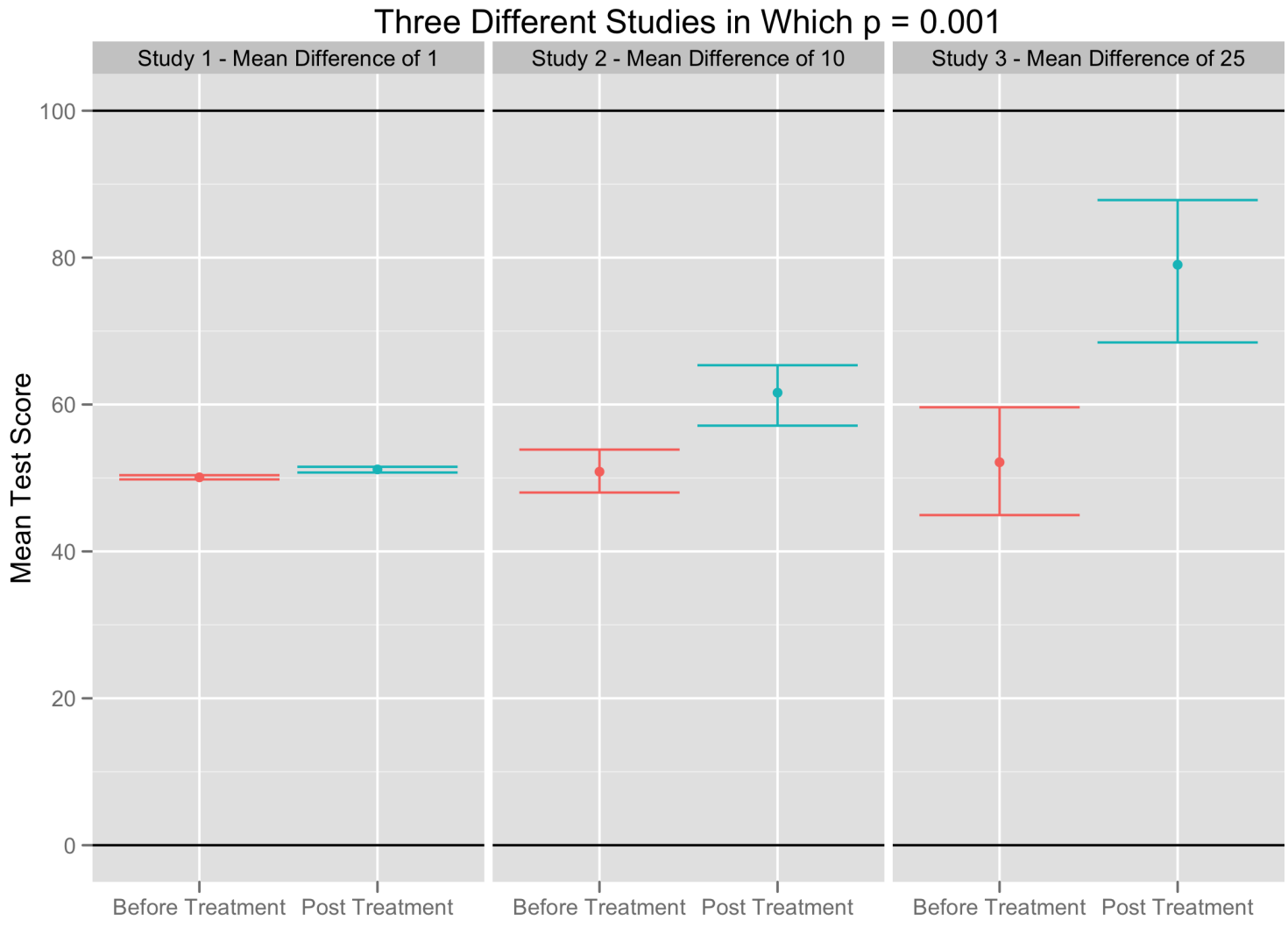
worker effort

If effect size > 0, there exists a large enough sample to be "significant."



from JWM's blog, code on github

Reduction of dimensionality



from JWM's blog

Comrade Polly

Believe what Chairman Alroy says about *p*-values! But you can still use ANOVA to test for differences in means, though believe what our Beloved Chairman says about asking what is the difference, why is it, and do I believe it!

Take home message

"Significance" is a product of interpretation, not of rigid rules.

Statistical inference is concerned with the "art of approximation" (Akaike).

Commandment 4

Thou shalt worship the almighty power! Despite the preceding commandment, accepting the null hypothesis is a vile, ungodly thing. Always make sure thou hast the statistical power and *a small enough difference relative to what thou carest about* to argue that a difference doesn't matter (not just that it isn't "significant"). When in doubt, find a power calculator on the web and do a proper power analysis.

Commandment 5

Thou shalt abhor tiny little time series! All too often people are seduced by ‘‘trends’’ of two or three data points, damning themselves to eternal hellfire. The two-tailed probability of a flawless ‘‘trend’’ with six points is 0.0625 (!). ‘‘Before’’ and ‘‘after’’ comparisons are no better than a single coin flip, unless the points in each category have significantly different averages. Coincidences are often coincidences: if (say) the biggest extinction happened in the same interval as the biggest climate change, and there are ten intervals, well, $p = 0.10$. So, demand that a time series analysis include a healthy number of data points, at least a dozen or a score or a cubit.

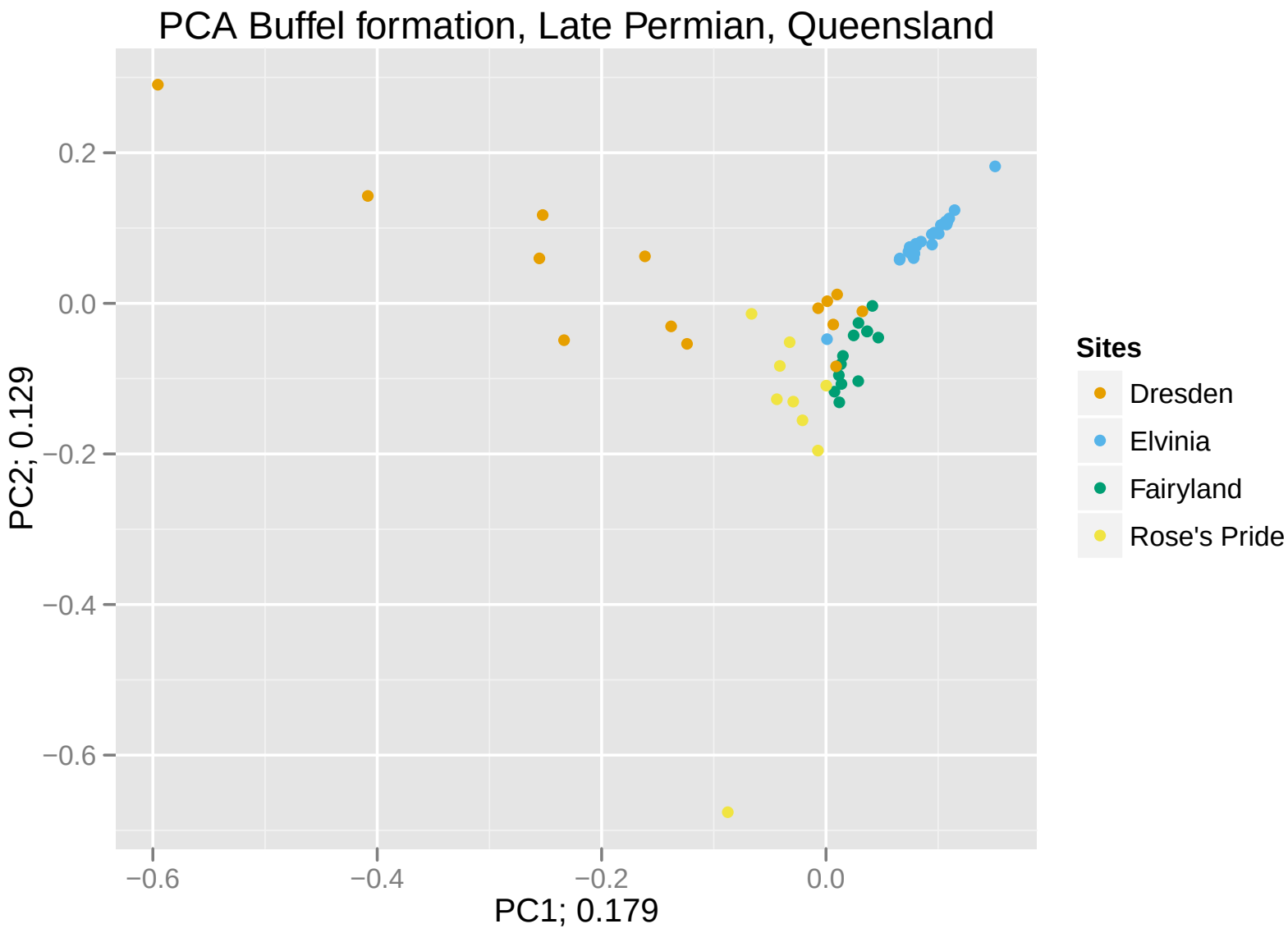
Commandment 6

Thou shalt difference thy data! Time series data are almost always autocorrelated (and thou shalt test for that). Still, people insist on interpreting ‘‘trends’’ shared by pairs of time series as meaningful cross-correlations, even though autocorrelation makes finding these demonic things *the null hypothesis!* Even random walks produce such patterns! FEAR YE SINNERS! The easiest and most powerful way to remove the autocorrelation is to take first differences. So, the next time thou wantest to correlate population growth with the rate of sea-floor spreading - and people will - *difference thy %#$% data*.

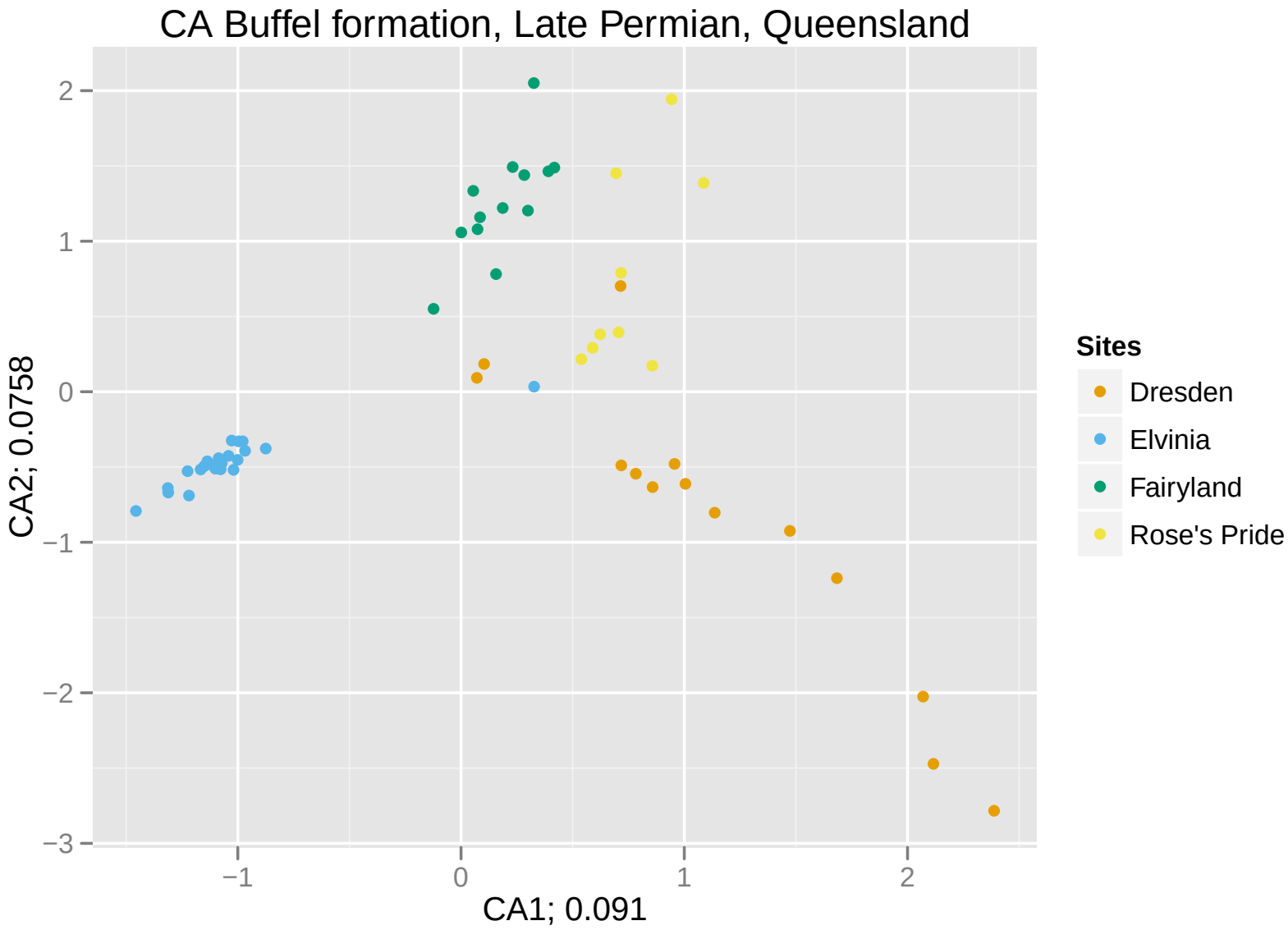
Commandment 7

Thou shalt not play with PCA! Principal components analysis assumes linear responses of observed variables to underlying variables, but most ecological data show modal responses. Vain mortal, what power grants thee the right to assume linearity? Correspondence analysis can handle both kinds of responses and works wonderfully on modal data (we won't mention that nasty little arch effect...).

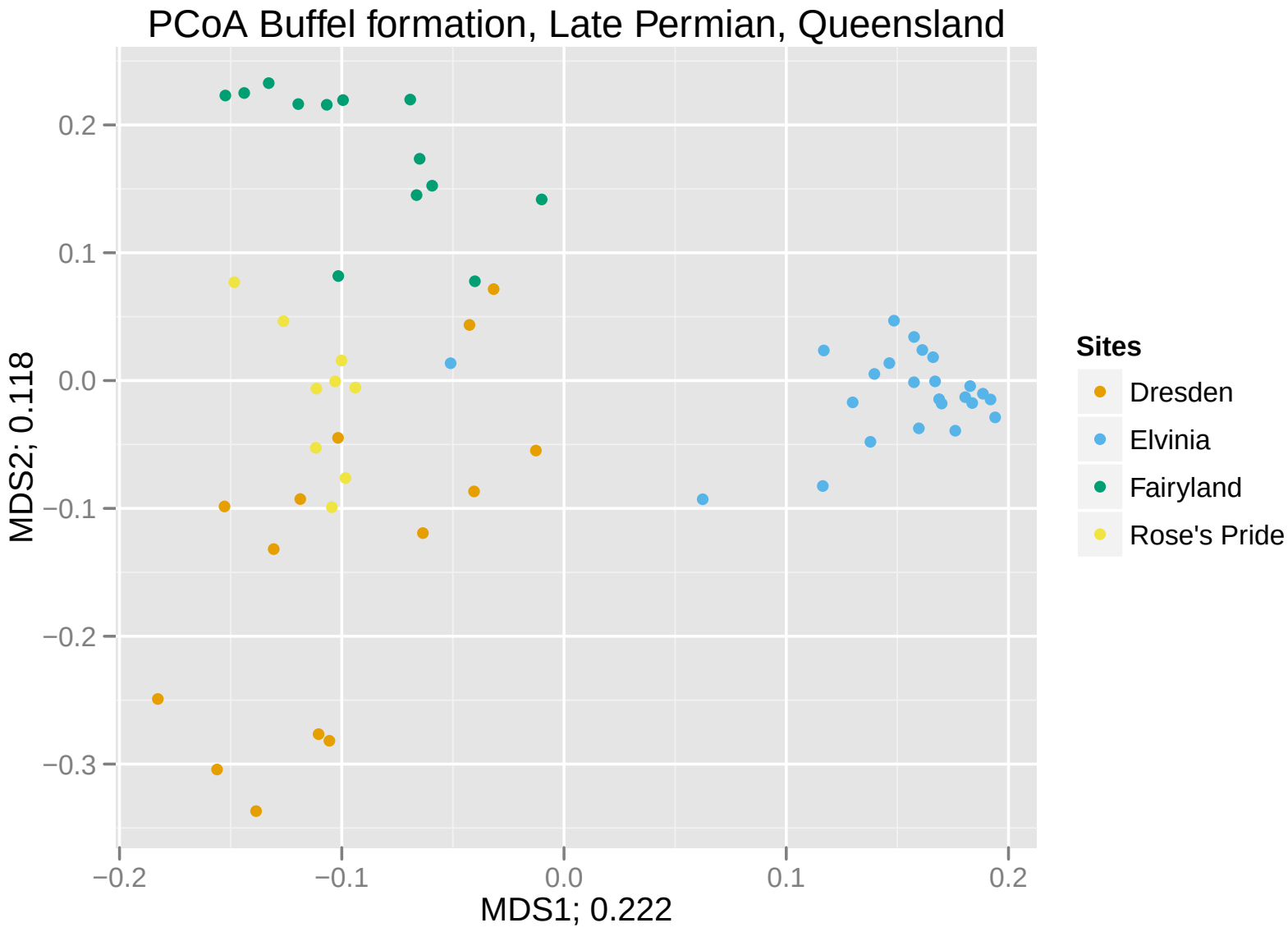
PCA


PCA Buffel formation, Late Permian, Queensland
PC2; 0.129
PC1; 0.179
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
Sites
Dresden
Elvinia
Fairyland
Rose's Pride

CA


CA Buffel formation, Late Permian, Queensland
CA2; 0.0758
CA1; 0.091
2
1
0
−1
−2
−3
−1
0
1
2
Sites
Dresden
Elvinia
Fairyland
Rose's Pride

PCoA


PCoA Buffel formation, Late Permian, Queensland
MDS2; 0.118
MDS1; 0.222
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.2
−0.1
0.0
0.1
0.2
Sites
Dresden
Elvinia
Fairyland
Rose's Pride

Ordinations

Remember, look at the higher axes!

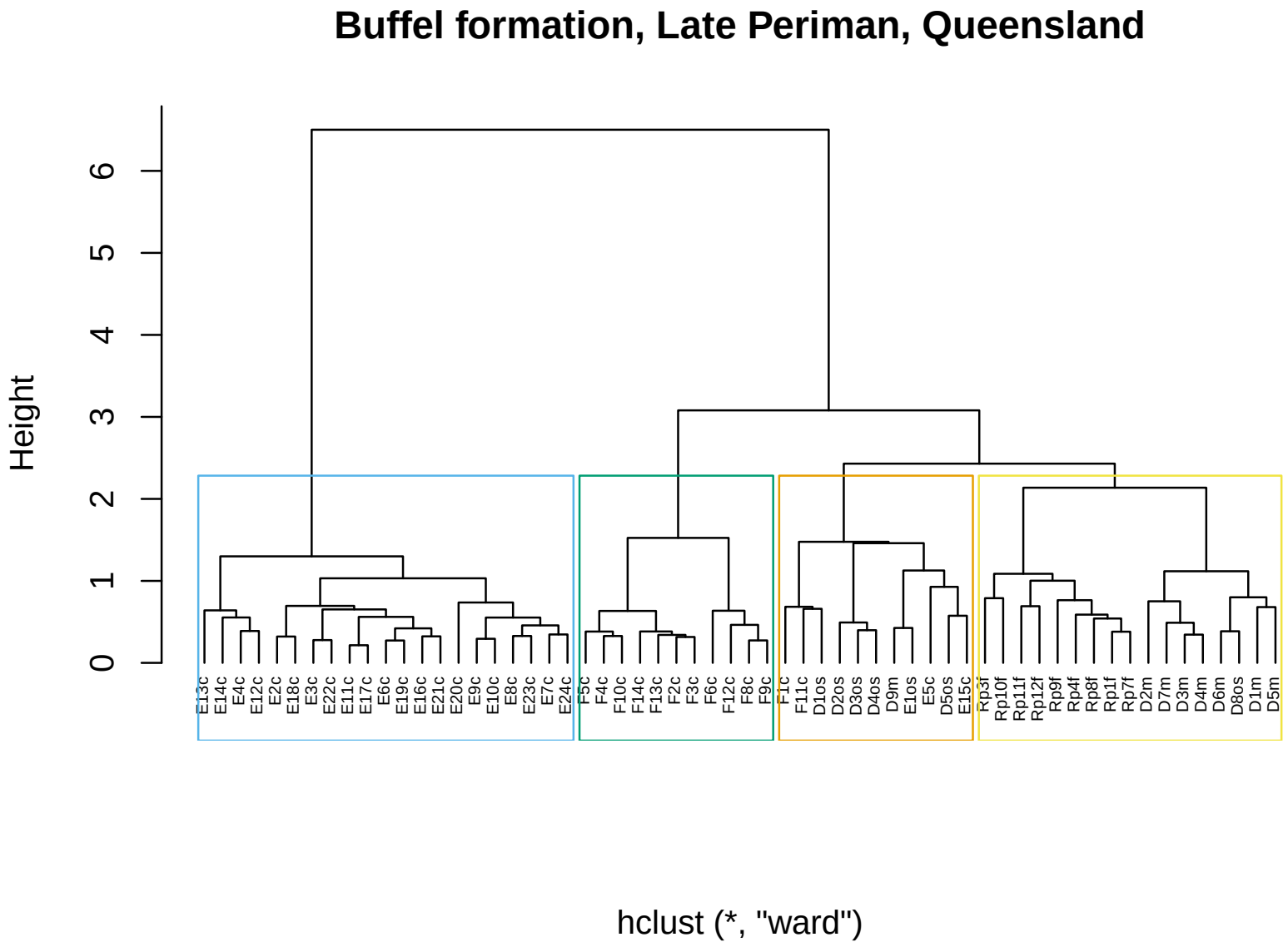
I have not have done that here, but there is always more information beyond the first and second axes. Explore them and learn the patterns in your data better

Don't use detrended correspondence analysis as your default. It destroys depictions of complex multidimensional gradients. **Just don't do it.**

Commandment 8

Thou shalt not cluster shamelessly! The world is full of fuzziness and apostasy, not cool, clean Platonic categories. But cluster analysis imposes categories on data regardless of whether they're gradiential. If the clusters are really there, thou shalt see them as a ray of divine light in the shadowy purgatory of a multivariate ordination space. So why bother?

Clustering


Buffel formation, Late Periman, Queensland
Height
0
1
2
3
4
5
6
E13c
E14c
E4c
E12c
E2c
E18c
E3c
E22c
E11c
E17c
E6c
E19c
E16c
E21c
E20c
E9c
E10c
E8c
E23c
E7c
E24c
F5c
F4c
F10c
F14c
F13c
F2c
F3c
F6c
F12c
F8c
F9c
F1c
F11c
D1os
D2os
D3os
D4os
D9m
E1os
E5c
D5os
E15c
Rp3f
Rp10f
Rp11f
Rp12f
Rp9f
Rp4f
Rp8f
Rp1f
Rp7f
D2m
D7m
D3m
D4m
D6m
D8os
D1m
D5m
hclust (*, "ward")

Commandment 9

Thou shalt stand awe-struck before the shining brilliance of the G-test! Chi-square this, chi-square that. The G is easier to compute, it doesn't blow up as easily because of small values, it depends on the awesome power of the log transform, it stands for "GOD," and most importantly it's a maximum likelihood ratio...

G-test now appears in Sokal and Rohlf's ‘‘Biometry’’

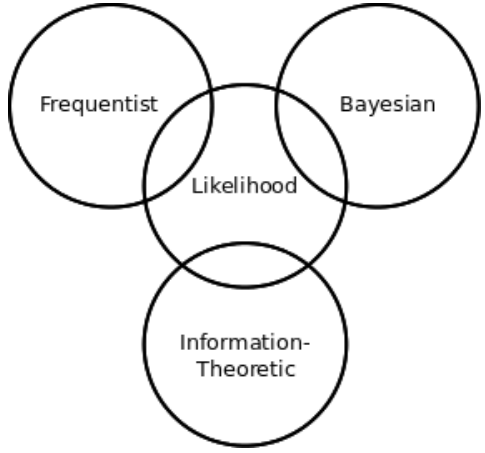
Also, this makes more sense in the context of...

Commandment 10

Thou shalt sing the praises of likelihood, not "fit"! Anyone can design another fit statistic. Why minimize the sum of squares instead of the sum of cubes or just the sum of differences? None of this has a theoretical basis without a notion of probability, and specifically of likelihood. After all, that's what the divine theologian Popper said.

Likelihood

Fundamental part of all statistical inference.



Measure of how much support our sample gives for some parameter estimate. Technically always a function, but we are normally most interested in the *maximum* likelihood.

$$L(\hat{\theta} \mid x, model)$$

AIC

Approximation of the Kullbach-Leibler divergence for unknown reality. ‘‘What?’’ I hear you say...

Incidentally, residual sum squares are a maximum likelihood estimator when the residuals are normally distributed . . .

K-L divergence

K-L divergence is the ‘‘distance’’ (not technically a distance) between a reference probability distribution f (also known as reality) and an approximating probability distribution g.

$$I(f, g) = \sum_{i=1}^{k} p_i \cdot \log(\frac{p_i}{\pi_i})$$

$$I(f, g) = \int f(x) \log(\frac{f(x)}{g(x|\theta)})$$

Two of the three fundamental formula in information theory along with Shannon's entropy.

back to AIC

Akaike showed that the relative expected K-L divergence and the log of the maximum likelihood are related.

$AIC = -2\log(L(\hat{\theta} \mid x)) + 2K$

The lowest AIC among the candidate models is the ‘‘closest’’ to the unknown reality, with some caveats.

Likelihood is extremely powerful.

Hallelujah!

but wait, there's more

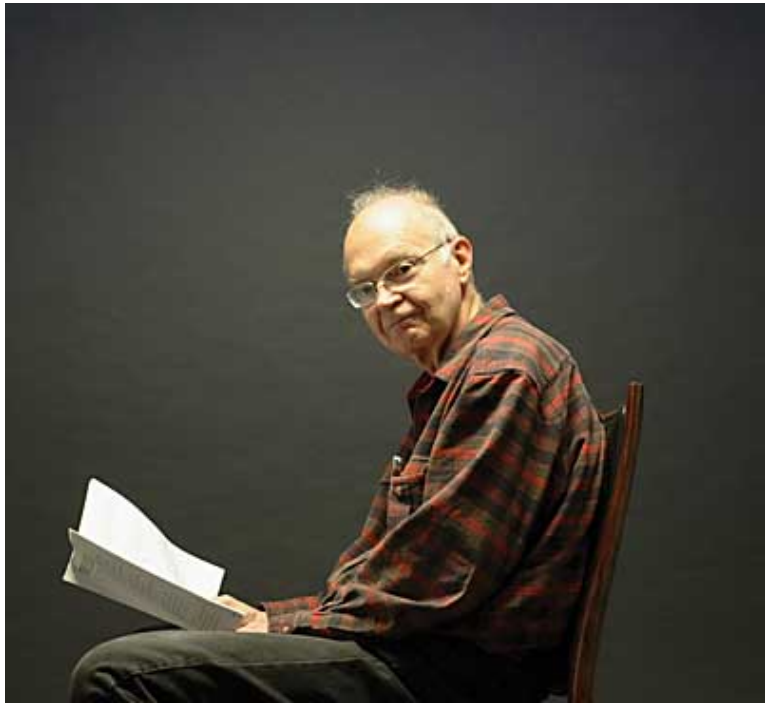
Peter's first words of advice

Release your source! Never underestimate the necessity of reproducible research and analysis. Using a markup language like LaTeX and literate programming tools you can embed code in reports, papers, and websites so that anyone, including yourself, can reproduce the exact results with minimal effort. So find a source repository and release your source to the world. The less time people expend trying to reproduce your work means more time spent extending and imporving your initial offering. By making your steps more open, you are more accountable for your work and thus a better scientist.

Literate programming?

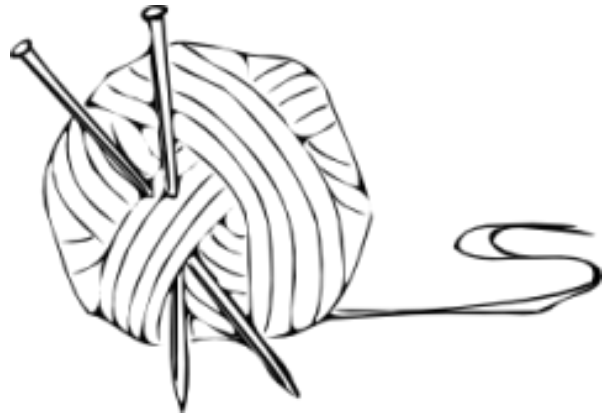
Donald Knuth:

> *The main idea is to regard a program as a communication to human beings rather than a set of instructions to a computer.*



Tools for R

- Sweave
  - LaTeX
  - old-school
  - mandatory for writing R help files
- knitr
  - LaTeX and markdown
  - shiny new thing designed for dynamic report generation
  - used to write this presentation



Source repository?

Server to host your source and use version control information at the same time.

I like github.

Other options exist like sourceforge, Google Code, R-Forge, etc.

Thanks!

- Alistair Evans
- Roger Close
- G10 and all the grad students
- Ross and Patrick
- PBDB instructors
  - **John Alroy**
  - Gene Hunt
  - Tom Olszewski
  - Peter Wagner
  - P. David Polly