Data analysis: principles

Ben Bolker

12:22 28 June 2015

More important than statistical philosophy

- Good experimental design (replication, randomization, independence, control, interspersion, adequate power)

  > The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data. (Tukey 1986)

- Sensible, well-posed questions
  - if you want to know if a variable is "important", or what model is "best", you need to know what you mean by that
- Knowledge of the system
- Strong signals will always be detectable; weak signals will never be
- Better analyses should be (within limits)
  - more powerful
  - better at disentangling (unavoidably) messy data
  - more interpretable
  - more convenient, faster, easier (cf. O'Hara and Kotze (2010) vs. Ives (2015))
- No free lunches

Philosophies

- don't look for a single philosophy (Gigerenzer and Marewski 2015)
- in many cases different philosophies give similar answers; differences should be understandable (e.g. Ludwig (1996))

Frequentist

- classic, well-tested
- much maligned
- Fisherian (strength of evidence) vs. Neyman-Pearson (decision-theoretic)
- null-hypothesis significance testing
- objective (?)

Bayesian

- basic idea:
  - it's easy to compute the probability of the *data* happening given the model (parameters etc.): *likelihood*
  - if we want to compute the probability of a *model* (parameters), we need to use **Bayes' Rule**
  - ... this in turn means we need to specify *prior probabilities* i.e., what did we think before we saw the data?
  - often bend over backwards to use *weak* or *uninformative* priors
- for strong data, simple cases, get nearly identical answers to freq.
- easier to incorporate prior knowledge (McCarthy 2007)
- easier to incorporate uncertainty (Ludwig 1996)
- easy=easy; medium=hard; hard=possible
- have to specify priors
- convenience/pragmatic/computational Bayesians: cf. Lele et al
- more natural statement of confidence ...
- **but** ... 'calibrated Bayesianism' (Gelman, de Valpine)
- frequentist approaches (de Valpine 2003; Sólymos 2010; Ponciano et al. 2009)

Computational

permutation testing

- similar to rank-based non-parametrics (Mann-Whitney, Wilcoxon, Spearman correlations ...)
- robust
- only gives p-values (usually)
- e.g. current *phylogenetic overdispersion* methods (Cavender-Bares et al. 2009)
- combine with parametric models for robust p-values

information theoretic/algorithmic (Breiman 2001)

- interested in prediction
- large data sets; data mining
- cross-validation etc.
- information-theoretic approaches loosely fall in this category (fitting is based on frequentist tools, inference is prediction-based)

Last thoughts on philosophy

- most of the statisticians I respect are agnostic about philosophies (e.g. Andrew Gelman:

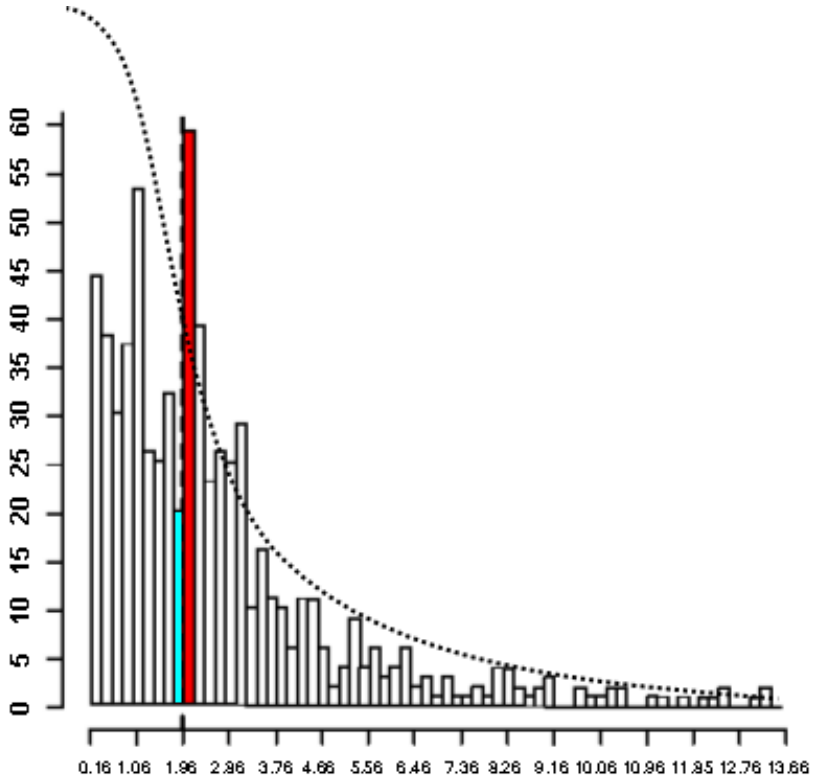
> "I have no problem with non-Bayesians: those statisticians who for whatever combination of theoretical or applied reasons prefer not to use Bayesian methods in their own work"

- good statisticians choose good tools *and* get good results; makes it harder to tell if the tools or the person is what's powerful (the *methodological attribution problem*, Gelman (2010)).
- Crome (1997):

> Perhaps the average user of significance tests, without knowing it, smears him- or herself over the three major statistical schools [Fisherian frequentist, Neyman-Pearson frequentist, Bayesian], and disobeys the rules of each ...

More on practice

- People (including scientists) are lazy (or think they have more important things to do) and prefer simple rules.
- *Some* rules of thumb for filtering noise are helpful ($p = 0.05$ is more or less a historical accident (Dallal 2015): Fisher: "in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas."
- Listening to music makes you younger: (Simmons, Nelson, and Simonsohn 2011)
- Graphical evidence for *p*-value abuse:



(Gerber and Mahotra 2008; Drum 2006)

- Need to decide in advance if you are trying to *confirm (test) hypothesis*, *predict* future outcomes or *explore* the data looking for interesting patterns:
  different rules/procedures in each case!

References

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199–215. http://www.jstor.org/stable/2676681.

Cavender-Bares, Jeannine, Kenneth H. Kozak, Paul V. A. Fine, and Steven W. Kembel. 2009. "The Merging of Community Ecology and Phylogenetic Biology." *Ecology Letters* 12 (7): 693–715. doi:10.1111/j.1461-0248.2009.01314.x.

Crome, Francis H. J. 1997. "Researching Tropical Forest Fragmentation: Shall We Keep on Doing What We're Doing?" In *Tropical Forest Remnants: Ecology, Management, and Conservation of Fragmented Communities*, edited by W. F Laurance and R. O Bierregard, 485–501. Chicago, IL: University of Chicago Press.

Dallal, Gerard E. 2015. "Why P=0.05?" Accessed June 28. http://www.jerrydallal.com/LHSP/p05.htm.

de Valpine, Perry. 2003. "Better Inferences from Population-Dynamics Experiments Using Monte Carlo State-Space Likelihood

Methods." *Ecology* 84 (11): 3064–77.

Drum, Kevin. 2006. "The Washington Monthly." http://www.washingtonmonthly.com/archives/individual/2006_09/009531.php.

Gelman, Andrew. 2010. "Bayesian Statistics Then and Now." *Statistical Science* 25 (2): 162–65. doi:10.1214/10-STS308B.

Gerber, Alan S., and Neil Mahotra. 2008. "Publication Incentives and Empirical Research: Do Reporting Standards Distort the Published Results?" *Sociological Methods and Research* 37 (1): 3–30. http://polmeth.wustl.edu/retrieve.php?id=640.

Gigerenzer, Gerd, and Julian N. Marewski. 2015. "Surrogate Science: The Idol of a Universal Method for Scientific Inference." *Journal of Management* 41 (2): 421–40. http://jom.sagepub.com/content/41/2/421.short.

Ives, Anthony R. 2015. "For Testing the Significance of Regression Coefficients, Go Ahead and Log-Transform Count Data." *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12386.

Ludwig, Donald. 1996. "Uncertainty and the Assessment of Extinction Probabilities." *Ecological Applications* 6 (4): 1067–76.

McCarthy, M. 2007. *Bayesian Methods for Ecology*. Cambridge, England: Cambridge University Press.

O'Hara, Robert B., and D. Johan Kotze. 2010. "Do Not Log-Transform Count Data." *Methods in Ecology and Evolution* 1 (2): 118–22. doi:10.1111/j.2041-210X.2010.00021.x.

Ponciano, José Miguel, Mark L. Taper, Brian Dennis, and Subhash R. Lele. 2009. "Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning." *Ecology* 90 (2): 356–62. http://www.jstor.org/stable/27650990.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66. doi:10.1177/0956797611417632.

Sólymos, Péter. 2010. "Dclone: Data Cloning in R." *The R Journal* 2 (2): 29–37. http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Solymos.pdf.

Tukey, John W. 1986. "Sunset Salvo." *The American Statistician* 40 (1): 72–76.