

Some ado about (mostly) nothing:
zero-dominated data

Alameda County Workshop on Avian Mortality at Altamont
Emeryville, CA
22 September 2009

Philip B. Stark

Disclaimer

I know nothing about bird-turbine interactions.

I work on applications that have zero-dominated and left-censored data, but all problems are different:

What I know might be useless to you.

That said, I'm eager to learn from you.

Why the zeros?

1. real values are mostly zero
2. small real values tend to be missed (left censoring)
3. measurement bias and truncation (measurements tend to be smaller than the truth, unless the truth is zero)
4. all of the above?

Most of my expertise is with (1).

I'll call (3) *shrinkage*.

What constraints are there?

Context: counts.

The fact that true counts are nonnegative is critical.

Some applications have upper bounds instead of lower bounds.

Some fields with zero-dominated and left-censored data

- Astronomy (extinction, “nearly black,” MACHOs)
- Auditing (financial, healthcare, elections)
- Geophysics (earthquake catalog incompleteness)

What do they do?

The questions differ, as do the constraints.

Astronomy: Some problems related to extinction, especially at high red-shift. Other problems, more concern about false non-zeros than false zeros. Often very low signal-to-noise. “Standard candles,” gravitational lensing, . . .

Auditing: financial, elections, healthcare: overstatements are rare, bounded. Presumption that audit reveals truth—no censoring.

Earthquake catalogs: models for earthquake frequency as a function of magnitude (Gutenberg-Richter). Some (partly circular) theoretical justification, but mostly empirical. Pretty good globally; weak locally. Bounds on total moment from geodetics.

What is the question?

Unbiased estimate of mortality? (censoring, shrinkage, background mortality will be problems)

Estimates of biomass, or numbers?

Low MSE estimate of mortality?

Lower confidence bound on mortality? (possible rigorously if the sample is random—unless background mortality matters)

Upper confidence bound? (not possible rigorously)

Mixture models

Mixture of a point mass at zero and some distribution on the positive axis. (Zero-inflated Poisson is like this.)

Alternatives are limitless:

1. observe $\max\{0, \text{Poisson}(\lambda_j) - b_j\}$, $b_j > 0$
2. observe $b_j \times \text{Poisson}(\lambda_j)$, $b_j \in (0, 1)$.
3. observe true count in area j with error ϵ_j , where $\{\epsilon_j\}$ are dependent, not identically distributed, nonzero mean

Where do the models come from?

- Why random?
- Why Poisson?
- Why independent from site to site? From period to period? From bird to bird? From encounter to encounter?
- Why doesn't chance of detection depend on size, coloration, groundcover, ...?
- Why do different observers miss carcasses at the same rate?
- What about background mortality?

What data are available?

Sampling schemes? In any sense random?

Follow-up studies to assess the incompleteness of recording?

Other detectors: sonar, radar, video, impact, suspended nets?
(pardon my naiveté!)

How far from the windmills do the carcasses land? What about injured birds?

Can “freshness” of the carcasses be used to estimate scavenger rates?

Complications at Altamont

1. Why is randomness a good model? Random is not the same as haphazard or unpredictable.
2. Why is Poisson in particular reasonable? Do birds in effect toss coins, independently, with same chance of heads, every encounter with a turbine? Is $\# \text{encounters} \times P(\text{heads})$ constant?
3. Why estimate the parameter of a distribution rather than actual mortality? Do we want to know how many birds died, or the value of λ in an implausible stochastic model?
4. Background mortality—varies by time, species, etc.

Complications, contd.

5. Are all birds equally likely to be missed? Smaller more likely than larger? Does coloration matter?
6. Nonstationarity (seasonal effects—migration, nesting, etc.; weather; variations in bird populations)
7. Spatial and seasonal variation in shrinkage due to ground-cover, coloration, illumination, etc.
8. Interactions and dependence.

Complications, contd.

9. Variations in scavenging. (Dependence on kill rates? Satiation? Food preferences? Groundcover?)
10. Birds killed earlier in the monitoring interval have longer time on trial for scavengers.
11. Differences or absolute numbers? (Often easier to estimate differences accurately.)
12. Same-site comparisons across time, or comparisons across sites?

Two very different situations:

1. Scientist creates randomness by taking a random sample, assigning subjects at random to treatment or control, etc.
2. Scientist invents (assumes) a probability model for data the world gives.

(1) allows sound inferences.

Inferences drawn in (2) are only as good as the assumptions.

Gotta check the assumptions against the world:
Empirical support? Plausible? Iffy? Absurd?

Recommendations

1. Concentrate on exploratory data analysis for now.
2. Make spot maps of found birds with turbine strings and characteristics identified.
3. Make cross-tabs by factors that might matter.
4. Think of ways to calibrate the number of missed carcasses and the background mortality, birds injured by one string that die near another, etc.
5. Forget about the Poisson model and other models. Extremely far removed from avian mortality.

Recommendations, contd.

6. Bayesian analysis is even more removed from the data and the original question. Uncertainties in Bayesian analysis can greatly understate the actual uncertainties. Sensitivity to the prior can be large.
7. Think about estimating differences between pairs that are matched on covariates that might affect the bias. E.g., changes in the same string in the same season across years. Estimating differences where biases are likely to be constant is probably more accurate than estimating an absolute number.

Freedman's Conservation of Rabbits Principle

If a rabbit is to be pulled from a hat, a rabbit must first be placed in the hat.

Assumptions put the rabbit in the hat.

Can't borrow a rabbit from an empty hat and restore the balance later with simulations, priors, etc.

No negative rabbits.

Further reading

Panel on Nonstandard Mixtures of Distributions, 1988. *Statistical models and analysis in auditing: A study of statistical models and methods for analyzing nonstandard mixtures of distributions in auditing*, National Academy Press, Washington, D.C., 91pp.

Kaplan, H.M., 1987. A Method of One-Sided Nonparametric Inference for the Mean of a Nonnegative Population, *The American Statistician*, 41, 157–158.

Fienberg, S.E., J. Neter and R.A. Leitch, 1977. Estimating total overstatement error in accounting populations, *J. Am. Stat. Assoc.*, 72, 295–302.

Kvanli, A.H., Y.K. Shen and L.Y. Deng, 1998. Construction of confidence intervals for the mean of a population containing many zero values, *J. Bus. Econ. Stat.*, 16, 362–368.

Further reading (contd)

Smieliauskas, W., 1986. A Note on a comparison of Bayesian with non-Bayesian dollar-unit sampling bounds for overstatement errors of accounting populations, *The Accounting Review*, 61, 118–128.

Stark, P.B. and D.A. Freedman, 2003. What is the Chance of an Earthquake? in *Earthquake Science and Seismic Risk Reduction*, F. Mulargia and R.J. Geller, eds., NATO Science Series IV: Earth and Environmental Sciences, v. 32, Kluwer, Dordrecht, The Netherlands, 201–213. (preprint: <http://statistics.berkeley.edu/~stark/Preprints/611.pdf>)

Further reading (contd)

Stark, P.B., 2009. Risk-limiting post-election audits: P -values from common probability inequalities, *IEEE Transactions on Information Forensics and Security*, accepted. (preprint: <http://statistics.berkeley.edu/~stark/Preprints/pvalues09.pdf>)