

Sampling, Modeling and Measurement Error in Inference from Clinical Text

Bob Carpenter

Columbia University, Department of Statistics

LingPipe, Inc.

ICML 2011 Workshop:

Learning from Unstructured Clinical Text

Cancer Clusters

(Gelman et al., *Bayesian Data Analysis*)

Highest kidney cancer death rates

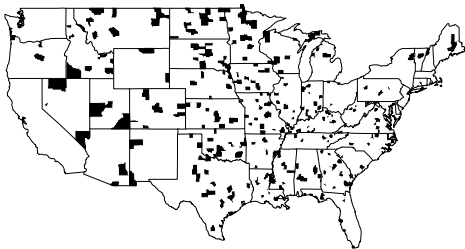


Figure 2.7 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.8 for discussion.*

Non-Cancer Clusters

Lowest kidney cancer death rates

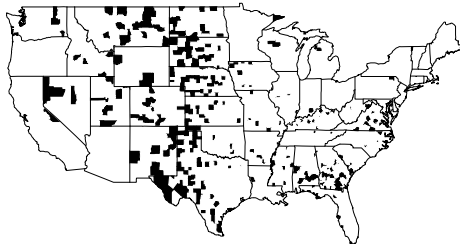


Figure 2.8 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.7.*

Part I. Sampling Error

What's Going On with the Plains?

- Nothing
- Less populous counties in plains states
- Apparent effect due to **sampling error**
- Small samples are more variable
- Want (super)population inference for county
- Population data itself based on measurement...

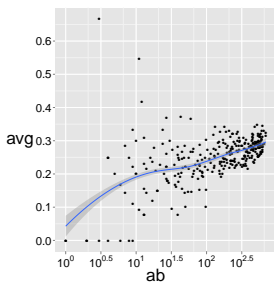
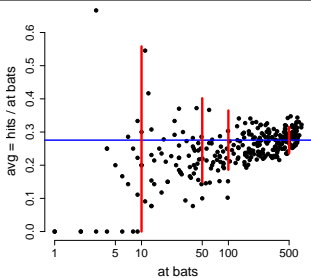
How Variable are Small Samples?

- Focus of frequentist hypothesis testing (e.g., p -values)
- Consider frequentist confidence intervals
 - these ignore shape and skew of posterior
- Consider estimating a mean by averaging N samples
- Posterior interval and confidence intervals have width

$$\propto \frac{1}{\sqrt{N}}$$

- Assuming (a) finite variance, and (b) i.i.d. samples

Batting Average (2006 AL Position Players)



- *Blue Line*: League Average; *Red Lines*: Binomial 95% Intervals
- Sampling bias: at-bats & average correlated (cf., sicker patients)
- More noise due to variance of small samples (cf., rare diseases)

Sampling Bias

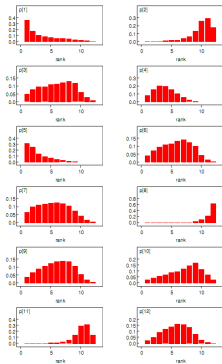
- Traditional tests presuppose items chosen at random
- Many sampling biases creep into real samples
 - Self-selection bias in surveys
 - Selecting patients likely to respond well to treatments
 - Selecting “easy to process” records (e.g., small, well formatted, etc.)
 - For causal studies, not balancing or recording confounding factors (e.g., age, smoking history, severity of condition, etc.)
 - Selecting records from a limited time period during the year
- Active learning a common source

Significance Fishing

- Common to measure many possible effects at once
- Choose most “significant” predictors
 - E.g., choosing best performing system in an evaluation
 - E.g., finding most predictive coefficients in a regression
- Often “adjusted” for false discovery rate (i.e., multiple comparisons)
- Sampling “best” predictors biased to overestimate effects
- Rookie-of-the-year to sophomore slump
 - Further samples closer to mean than initial estimate

Pediatric Surgery Deaths (BUGS, Examples I)

Hospital	No of ops	No of deaths
A	47	0
B	148	18
C	119	8
D	810	46
E	211	8
F	196	13
G	148	9
H	215	31
I	207	14
J	97	8
K	256	29
L	360	24



- Data and Posterior Comparison with Hierarchical Priors

Part II. Measurement Error

Kinds of Measurement Error

- Error may arise in predictors (i.e., features) or predictions (e.g., outcomes of a classification)
 - distinction only relevant for controlled experiment
- Traditional measurement error arises from many common sources
 - Error from limited accuracy (e.g., measurement of height, blood pressure, etc.)
 - Error from quantization (e.g., height reported to 1 inch; age reported to 1 year; condition reduced to ICD-9 code)
 - Error from censoring (e.g., age truncated to 18 for adults)
- Error may be biased (e.g., scale averages +0.1kg)

Measurement Error in Raw Clinical Text

- Natural language is inherently vague and ambiguous
 - One person's "very bad" is another's "flesh wound"
 - Even ICD-9 is imprecise and varies in granularity
- Adds a great deal of noise

Example: History of Present Illness

- Patient's original perception of events is noisy
- Patient's recollection of events is noisy
- Patient's language for describing conditions is imprecise compared to clinician's
- Patient may be lying (e.g., drug history)
- Patient may not even be a mentally competent adult native speaker

Example: HPI (cont.)

- Clinician can misunderstand patient for above reasons
- Clinician subject to distraction or inattention
- Clinician bias (e.g., sexual, racial stereotypes)
- Transcription errors add more noise through brainos, typos, faulty memory, etc.
- May enter wrong item in wrong fields, even wrong patient
- Dictation and (automatic) transcription is error prone
- Fraud, or just copying swaths of records from day to day

But it's Worse with NLP

- With NLP-based inferences, further noise
- Almost all biased
 - e.g., only find find common or easy UMLS terms
 - e.g., bias toward terms like training data
 - e.g., mismatch in genre between train test (e.g., Genia blood oncology named entities versus clinical note data)
 - e.g., negative bias with misspelled terms, or adjectives sprinkled into UMLS terms, or new names not in training data
 - e.g., positive bias of fraction 1/2 reported as January 2

Correcting for Predictive Biases

- Good news is we can correct (in some cases)
- Suppose two category classifier for clinical notes
- Categories = smoking history / no smoking history
- Train a binary classifier over a labeled training set
- Evaluate sensitivity and specificity (i.e., bias and accuracy)
- Adjust predictions for estimated classifier bias and accuracy
- Apply by patient probabilistically or collectively

Correcting for Predictive Biases (cont)

- A technique that's been around since at least 1970
- Sensitivity = $TP / (TP + FN)$
- Specificity = $TN / (TN + FP)$
- Prevalence = fraction of smokers in population
- $\Pr[\text{classifier+}] = \text{sens} * \text{prev} + (1 - \text{spec}) * (1 - \text{prev})$
- estimated prev
= $(\Pr[\text{class+}] + \text{spec} - 1) / (\text{sens} + \text{spec} - 1)$

Errors in Corpus Annotation

- Corpora used for training themselves noisy and biased
 - Unclear or misunderstood coding standard
 - Annotators who don't know everything
 - Annotators with systematic bias (e.g., overrate or underrate everything, skew to extremes, hew to middle)
 - Annotator inattention or sloppiness
- Typical solution to censor data to agreements
 - this biases evaluations tremendously
- Or to adjudicate ties (measure this accuracy/bias)

Dawid and Skene's Annotator Models (1979)

- Have multiple annotators label overlapping items
- Estimate annotator accuracy and bias from data
- “Gold standard” data not necessary (but helps, even if noisy)
- Adjust predictions with estimated accuracy and biases
- Leads to much cleaner corpora
- Posterior of corpus is still not certain
 - Propagate uncertainty with weighted training (Smyth 1995)
 - Jointly estimate annotators, labels and classifier (Raykar et al. 2010)

“Adjusting” for Measurement Error

- Posterior uncertainty larger
- Frequentist: mondo calculus to compute how much
 - without adjustments, overstating confidence
- Bayesian Approach
 - without adjustments, posteriors underdispersed (too narrow)
 - propagate uncertainty through the model
 - e.g., Use weighted training data (Smyth 1995)
 - e.g., Jointly estimate annotators, labels, and classifiers (Raykar et al. 2010)

Part III. Model Specification Error

All of Our Models are Wrong

- Our models are just approximations
- Including frequentist models (i.e., still “subjective”)
- Every model provides examples, common ones being
 - false normality (or large sample) assumptions
 - false (conditional) independence assumptions
 - missing predictors (e.g., time series)
 - missing interactions (e.g., geriatric & female)
 - data not independent, identically distributed (iid)
 - errors not random noise (skewed, based on predictors like time)

Common Errors in Models of Language

- Independence of words, phrases, etc.
 - Leads to overconfident predictions (e.g., naive Bayes, HMMs, language models)
 - Leads to underdispersed models (e.g., don't expect to see name or other terms repeated as often as they are)
- Interaction of features ignored, especially long-distance

Imbalanced Training and Test Data

- Training data distributed one way, test data another
 - e.g., “Balanced” training sets
 - e.g., non-stationary (changing topics/language over time)
 - Genre or style mismatch (e.g., MEDLINE vs. clinical notes, *WSJ* vs. blogs)
- Post-stratification can adjust
 - train based on balanced data set
 - use population demographics to adjust predictions
 - e.g., model may predict a 20% chance male, rural, high-school student smoking, and know number of such individuals in the population

Diagnosing Model-Specification Errors

- Similar in applied Bayesian and frequentist studies
 - see texts by Friedman, Gelman et al., or anything with “applied” in the title
- Check model predictions
 - Check fits, parameter uncertainty and outliers
 - Simulate from the model
 - Error analysis on held-out data

Simulating from Model (Shannon 1948)

- Character and token n -gram simulation from model
 - **Char 0)** XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.
 - **Char 1)** OCRO HLI RGWR NMIELWIS EU LL NBNE-
SEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.
 - **Char 2)** ON IE ANTSOUTINYS ARE T INCTORE
ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE
AT
 - **Char 3)** N NO IST LAT WHEY CRATICT FROURE
BIRS GROCID PONDENOME OF DEMONSTURES
OF THE

- **Word 1)** REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE
- **Word 2)** THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT

- Decent local syntactic fit
- Poor semantic fit

IV. Bayesian Modeling

Statistics about

- Analyzing data sets
- Prediction of new unseen data
- Causal scientific inference
- All require reasoning under uncertainty
- Rewards for true/false positives and true/false negatives not symmetric
- Need decision theory

Prediction with Point Estimates

- Most common approach in NLP
- Predict new data y' given training data y
- Point estimate parameter(s) $\hat{\theta}$ given data y (e.g., MLE)
- Predict using $p(y'|y) = p(y'|\hat{\theta})$
- Does not model uncertainty estimating θ
- Overconfident predictions result

Bayesian Uncertainty Propagation

- Bayesian models allow probability statements about unobservable parameters θ
 - Priors: $p(\theta)$
 - Posteriors: $p(\theta|y)$ where y is observed data
 - Joint distribution: $p(\theta, y)$
- Posterior predictive distribution uses posterior $p(\theta|y)$
- Predictive distro $p(y'|y) = \int p(y'|\theta) p(\theta|y) d\theta$
- Coherent way to propagate uncertainty (i.e., no Dutch book)

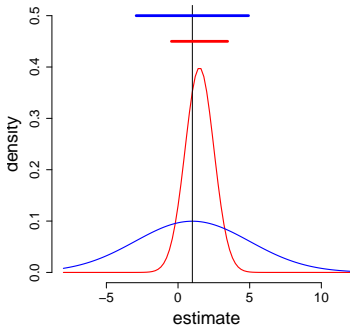
V. Questions and Discussion

VI. Estimation Bias

Estimation Bias and Variance

- So far, we're considering estimator variance due to sampling the population
- Error is difference between estimate and true value
- Bias is expected error
- Estimators can be systematically biased
 - E.g., sample variance underestimates variance by $(N - 1)/N$
- Can reduce error to trade bias for variance (e.g., regularization)

Example of Bias Reducing Expected Error



- Vertical black bar at 1.0 is true value
- Red estimator Norm(1.5, 1) biased
- Blue estimator Norm(1, 2) unbiased
- Bars are 95% intervals for estimators