Good Smoothing

Jim Albert

Bowling Green State University

March 15, 2010

Outline

Introduction

- General problem in categorical data analysis is how to handle small counts.
- Wald confidence interval for a proportion

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

  does not work well for small n.
- $P(\text{interval covers } p)$ is not uniformly 0.95.

Ad-hoc solution

- Add small counts to data, and apply frequentist methods to the adjusted data.
- John Tukey suggested "starting" counts by $1/6$.
- Agresti and Coull suggest adding "2 successes and 2 failures" to data, and then apply Wald interval estimate.
- In contingency tables with zero counts, common to add $1/2$ to each cell.

Why not Bayes?

- Adding imaginary counts corresponds to prior information.
- Leads to a Bayesian analysis.
- I. J. Good was one of the first to discuss the choice of imaginary counts in smoothing categorical data.
- Famous 1967 paper by Good "A Bayesian significance test for multinomial distributions" discusses his general approach.

Good's Testing problem

- Observe $y = (y_1, \ldots, y_t)$ from multinomial distribution with sample size n and probabilities $p = (p_1, \ldots, p_t)$.
- Test hypothesis $H : p_1 = \ldots = p_t = \frac{1}{t}$
- Usual test procedure is Pearson's statistic:

$$X^2 = \sum_{j=1}^{t} \frac{\left(y_j - \frac{n}{t}\right)^2}{\frac{n}{t}}$$

which is asymptotically $\chi^2(t-1)$.

Motivation for Bayes

- Accuracy of chi-square approximation for small counts is questionable.
- Desirable to develop an "exact" Bayesian test free from asymptotic theory.
- Use procedure with confidence for all t and n.

Bayes factor

- Ratio of marginal densities under the hypotheses H and A (not H).
- Under H, have

$$m(y|H) = \frac{n!}{\prod_{j=1}^{t} y_j!}(1/t)^n.$$

- Under A, put prior $g(p)$ on p and have

$$m(y|A) = \frac{n!}{\prod_{j=1}^{t} y_j!}\int \prod_{j=1}^{t} p_j^{y_j} g(p)dp,$$

- Bayes factor $BF = m(y|A)/m(y|H)$.

How to choose prior under A, $g(p)$?

- "Johnson's postulate": Posterior mean for p_j should depend only on the multinomial count y_j (not other y_k).
- This postulate implies that

$$E(p_j|y) = \frac{y_j + k}{n + tk},$$

for some choice of "flattening constant" k.

- This implies that p has a symmetric Dirichlet distribution:

$$g(p|k) = \frac{\Gamma(tk)}{\Gamma(k)^t} \prod_{j=1}^{t} p_j^{k-1}.$$

Choice for flattening parameter k?

- Maximum likelihood estimate assumes $k = 0$.
- Uniform prior assumes $k = 1$.
- Jeffreys' prior assumes $k = 1/2$.
- Good argues that none of these are appropriate.

Assumes a hierarchical prior

- k given a density $\phi(k)$
- Prior for p is given by

$$g(p) = \int_0^\infty \frac{\Gamma(tk)}{\Gamma(k)^t} \prod_{j=1}^{t} p_j^{k-1} \phi(k) dk.$$

- Good uses a log Cauchy density for k.

Expression for Bayes factor

- Compare models: H: equiprobability, $A : p$ has symmetric Dirichlet with parameter k.
- Bayes factor in support of A is

$$BF(k) = \frac{m(y|A)}{m(y|H)} = t^n \frac{D(y+k)}{D(k)},$$

where $D(a)$ is the Dirichlet function.

- If k is assigned a density $\phi(k)$

$$BF = \int_0^\infty BF(k)\phi(k)dk.$$

Other test statistics

- Useful to plot $BF(k)$ as function of k (like a likelihood function).
- Alternative test statistic

$$BF_{max} = \max_k BF(k).$$

Provides estimate for the proportion vector p

- Estimate of p_j is

$$\hat{p}_j = \frac{y_j + \hat{k}}{n + t\hat{k}},$$

where $\hat{k}$ is posterior mode.

- Smooth rates $\{y_j/n\}$ towards equiprobability value $1/t$.

An example

- Counts of new visits to my book website during one week in March 2009.

Sun	Mon	Tue	Wed	Thu	Fri	Sat
14	25	16	11	22	12	6

- Want to test hypothesis that the probabilities are equiprobable.

$$H : p_1 = \ldots = p_7$$

Traditional approach

- The Pearson statistic $X^2 = 16.96$ (p-value $=$ 0.0094).
- If we view p-value as $P(H)$, and H and A have equal prior probabilities

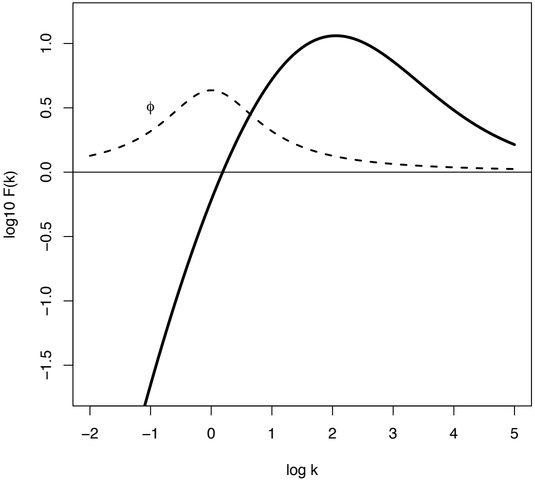
$$\log_{10} BF = -log_{10} BF = 2.23.$$

Good's approach

- Plot $\log_{10} BF(k)$ as function of $\log k$.
- Bayes factor maximized at $\log k = 2.05$ and

$$\log_{10} BF_{max} = 1.06$$

- Compare with evidence suggested by p-value.
- Compute BF by averaging BF_k over prior.



φ
log10 F(k)
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2
3
4
5
log k

Smoothed estimates at proportions

- Have $\hat{k} = \exp(2.05) = 7.8$.
- Bayes estimate at proportion is

$$E(p_j|y) = \frac{y_j + 7.8}{n + 7(7.8)}$$

Notable aspects of Good's approach

- Smoothing problem related to test of a model
- Degree of smoothing depends on agreement of data with model
- Effort to compare with frequentist methods
- New test statistics (like k_{max}) evolve from Bayesian model
- Advocated hierarchical priors

Applications

- Apply Good's smoothing strategy to some problems with small counts.
- Estimating a proportion.
- Estimating probabilities in a two-way contingency table.
- In each case, we will be smoothing counts towards a particular model.

Estimating a proportion

- Observe y from a binomial(n, p) distribution.
- When $y = 0$ or $y = n$, typical estimate y/n is undesirable.
- Can adjust estimate by applying beta(a, b) density.
- Let $\eta = a/(a + b)$, $K = a + b$.
- Smoothed estimate is $(y + K\eta)/(n + K)$.

Unknown K

- Suppose one can make intelligent guess at η.
- K unknown, assigned a log Cauchy density.
- Posterior density of $\log K$ is

$$g(\log K|y) \propto \frac{B(K\eta + y, K(1-\eta) + n - y)}{B(K\eta, K(1-\eta))} \frac{1}{(1 + (\log K)^2)}.$$

- Estimate $\log K$ by its posterior mode.

An example

- Sample size $n = 20$
- Guess at η is 0.5.
- Estimate for K is 0.6 at extreme values $y = 0, 20$.
- Estimate for K is 1.41 when $y = 10$.
- Bayesian procedure is "add 0.3 to 0.7 to number of successes and number of failures"
- Similar to "add a half count" rule of thumb.

Both K, η unknown

- Assign a vague prior: η assigned Jeffreys' prior, K assigned a log Cauchy density.
- Find posterior mode of joint density.
- Estimate of η shrinks proportion y/n towards 0.5.
- Get estimates that approximate "add a half count" rule of thumb.

Look at "add 2 successes and 2 failures" algorithm from Bayes perspective

- Algorithm says "add 2 + 2 pseudo counts" to data.
- Apply standard algorithm to adjusted data.
- Equivalent to assigning p a beta(2, 2) prior and estimating p from the posterior.
- Example: $y = 0$, $n = 10$, posterior is beta(2, 12).
- 90% interval estimate for p is (0.028, 0.316).

Since adding $2 + 2$ is arbitrary, better to use a hierarchical prior

- Construct a prior on (K, η) that reflects the desire to add 2 successes and 2 failures.
- Assign $\log K$ a Cauchy density with location $\log 4$ and scale 1 (want to add 4 observations).
- Assign η a beta prior with mean 0.5 and precision $K_0 = 80$ (want to divide pseudo counts equally between successes and failures).

Interval estimates for proportion p

- If $y = 0, n = 10$, 90% "hierarchical" interval estimate for p is (0.000, 0.336).
- The "add 2 + 2 interval" was (0.028, 0.316).
- Hierarchical interval is wider since it reflects uncertainty in adding 2 successes and 2 failures.

Smoothing a 2 by 2 table

- Observe independent counts $y_1 \sim B(n_1, p_1)$, $y_2 \sim B(n_2, p_2)$.
- Want to smooth counts in table

	Successes	Failures
Pop 1	y_1	$n_1 - y_1$
Pop 2	y_2	$n_2 - y_2$

Prior beliefs

- Suppose p_1, p_2 are assigned common beta(η, K) prior.
- We wish to add the "prior counts"

	Successes	Failures
Pop 1	$K\eta$	$K(1-\eta)$
Pop 2	$K\eta$	$K(1-\eta)$

- Assign vague priors to K, η.

Smoothed estimates

- Posterior mean of p_1 given by

$$\hat{p}_1 = \frac{y_1}{n_1}\left(1 - \frac{\hat{K}}{n_1 + \hat{K}}\right) + \hat{\eta}\frac{\hat{K}}{n_1 + \hat{K}},$$

- $\hat{\eta}$ is pooled estimate of proportions under "independence" model where $p_1 = p_2$
- estimate $\hat{K}$ reflects agreement of counts with independence model
- For table [0, 20; 20 0] (far from independence), $\hat{K} = 0.3$
- For table [10, 10; 10 10] (close to independence), $\hat{K} = 4.0$

Smoothing in a I by J table

- Observe Poisson counts $\{y_{ij}\}$ with means $\{\lambda_{ij}\}$
- Want to smooth towards log linear model $\log \lambda_{ij} = \log x_{ij}\beta$
- Ex: $\log \lambda_{ij} = \beta_0$ (smoothing towards constant frequencies)
- Ex: $\log \lambda_{ij} = \beta_0 + u_i + v_j$ (smoothing towards independence model)

Model

- λ_{ij} are independent Gamma$(\alpha, \alpha/\mu_{ij})$
- $\{\mu_{ij}\}$ satisfy the log-linear model

$$\log \lambda_{ij} = x_i \beta.$$

- α and β are independent with β distributed uniform, α distributed log Cauchy density with location $\log \mu$ and scale σ

Posterior Estimates

- Estimate at λ_{ij} given by

$$\hat{\lambda}_{ij} = \frac{y_{ij} + \hat{\alpha}}{1 + \hat{\alpha}/\hat{\mu}_{ij}},$$

- $\hat{\mu}_{ij}$ and $\hat{\alpha}$ are respectively posterior estimates at μ_{ij} and α
- estimate $\hat{\alpha}$ is the number of pseudo-counts added to each cell

An Example

Crosstabulation of student teachers rated by two supervisors.

		Rating of Sup 2		
		Auth	Dem	Perm
Rating of Sup 1	Auth	17	4	8
	Dem	5	12	0
	Perm	10	3	13

Posterior estimates

Clear pattern of dependence in the table; obtain only modest shrinkage of the counts towards independence ($\hat{\alpha} = 1.84$)

		Rating of Sup2		
		Auth	Dem	Perm
Rating of Sup 1	Auth	16.3	4.8	7.9
	Dem	5.5	10.2	1.3
	Perm	10.2	4	11.8

Bayesian smoothing of large tables

- Batting data collected for 487 nonpitchers in 2008 season.
- Simultaneously estimate performance for all hitters.
- Simultaneously estimate "situational effects" for all hitters. (Compare performance, say at home games versus away games.)
- Hard to interpret individual hitting measures due to varying sample sizes.
- Smoothing by exchangeable models is helpful.

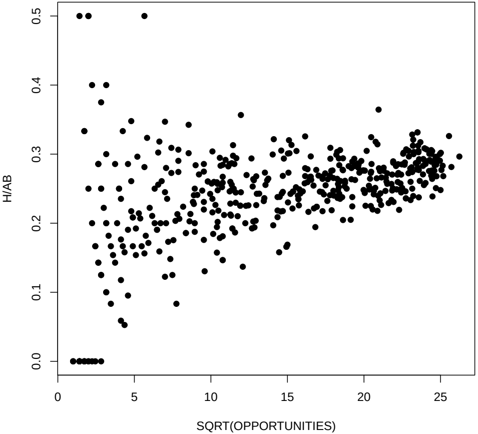
Smoothing model

- Observe independent $y_j \sim \text{binomial}(n_j, p_j)$
- Assume $p_1, \ldots, p_N$ random sample from beta(η, K)
- (η, K) assigned prior

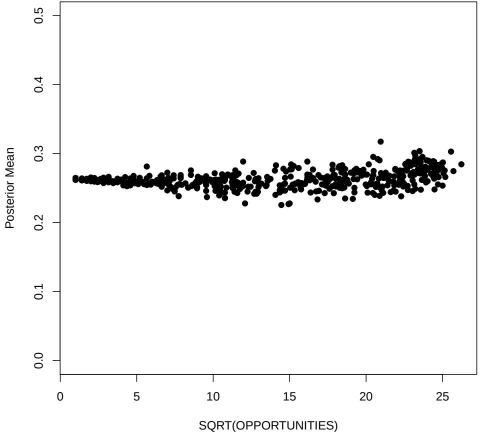
$$g(\eta, K) \propto \frac{1}{\sqrt{\eta(1-\eta)}} \frac{1}{(1+K)^2}.$$

- Estimate p_j by posterior mean.

Batting averages against the root sample sizes



Posterior means

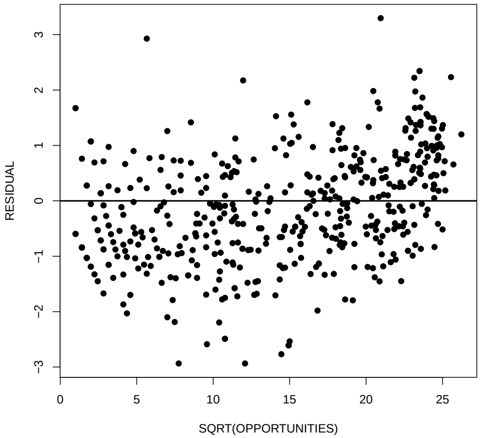


Looking further ...

- Is an exchangeable model appropriate?
- Unusual batting rates?
- Examine predictive residuals

$$r_j = \frac{y_j/n_j - \hat{\eta}}{\sqrt{\hat{\eta}(1-\hat{\eta})\left(1/n_j + 1/(\hat{K}+1)\right)}},$$

Residual plot



Estimating situational effects

- How do players perform in different situations?
- Obvious biases – players tend to play better at home, batters hit better against pitchers of the opposite arm
- Situational data for jth player:

	Hits	Outs
Home	s_{jH}	f_{jH}
Away	s_{jA}	f_{jA}

Exchangeable model

- Hits in two situations are independent binomial with parameters p_{jH} and p_{jA}
- Odds ratio for jth player

$$\alpha_j = \frac{p_{jH}/(1 - p_{jH})}{p_{jA}/(1 - p_{jA})}$$

- Assume $\alpha_1, ..., \alpha_N$ are iid $N(\mu, \sigma^2)$, μ, σ^2 are given vague priors

Example

- Have home/away data for 195 players
- Posterior estimate for μ is positive (batters tend to hit better at home)
- Posterior estimates of α_j shrink 82-93% towards overall mean
- Half of the estimates fall between 0.058 and 0.090
- Conclusion: players have essentially same hitter advantage at home vs away

Summing up

- Bayes is a natural way of handling small counts in a contingency table
- Good's approach based on a Bayesian test of an underlying model.
- Hierarchical priors are suitable for smoothing tables.
- These type of models are very suitable in looking for patterns in large collections of counts.