Confidence Intervals for the CDF from"noisy" iid samples

Daniel Yekutieli

Tel Aviv University

Luminy conference
"Mathematical Methods of Modern Statistics"

Problem description

Data: 30 day readmissions due to chronic obstructive pulmonary disease in VA hospitals (downloaded from Medicare website)

Random effects model

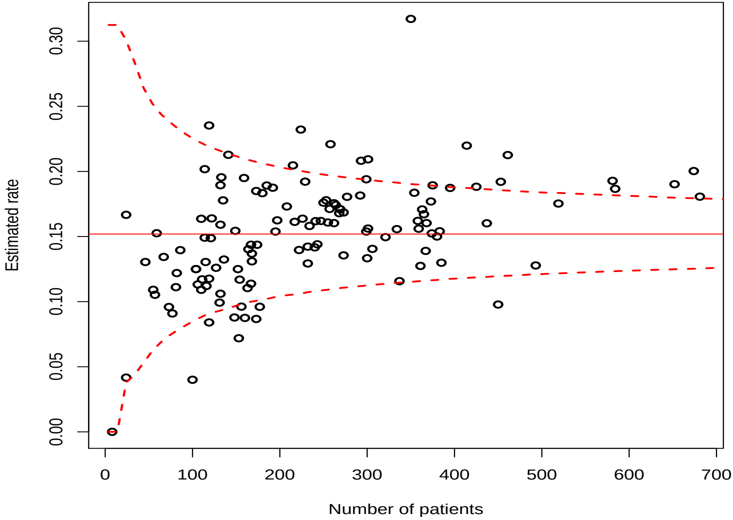
- Hospital index $i = 1 \cdots 129$
- n_i number of at risk patients on Hospital i
- x_i number of readmitted patients

$$X_i \sim Binomial(n_i, p_i)$$

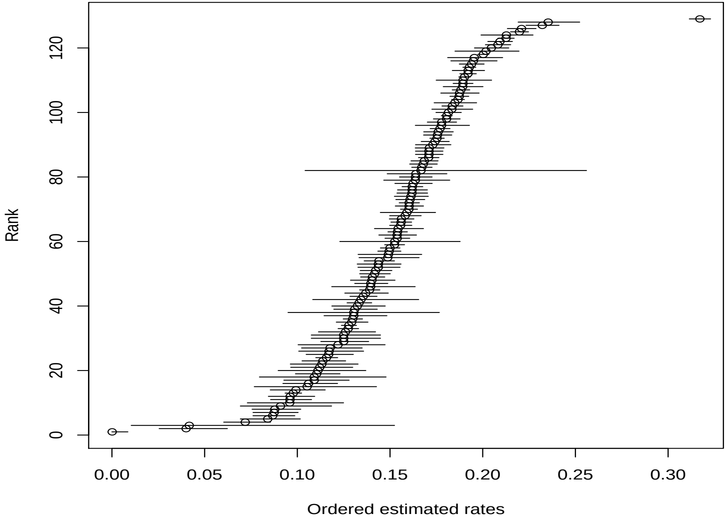
- p_i iid sample from distribution π

Goal: use "noisy" sample $x = (x_1 \cdots x_{129})$ to construct point-wise CI's for CDF and quantiles of π

Estimated readmission rates as a function of sample sizes



Ordered estimated readmission rate with 95% CI's



How do we construct CI's for π?

e.g. 95% CI for $CDF_\pi(0.16)$ of the form $[\hat{q}, 1]$:

1. For $q = 0.000, 0.0001, 0.0002, 0.0003, \ldots, 1$

$$\Omega_0(p) = \{\pi : \; CDF_\pi(0.16) \leq q\}$$

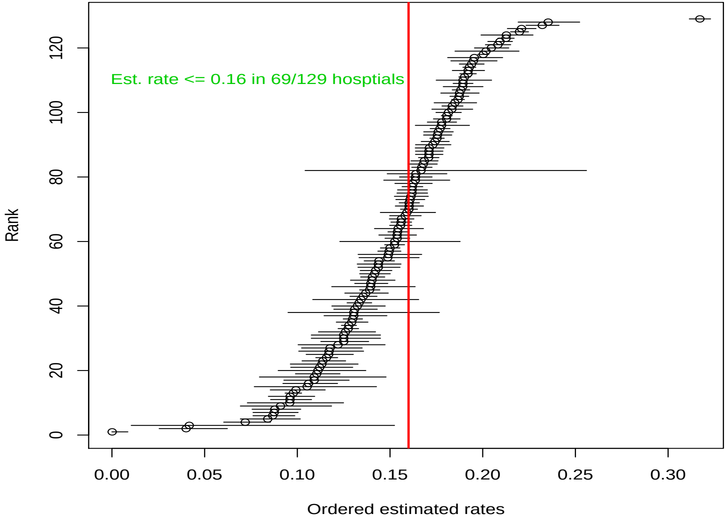
2. Run level 0.05 test

$$H_0(q) : \pi \in \Omega_0(q) \text{ vs. } H_1(q) : \pi \notin \Omega_0(q)$$

3. 95% CI is

$$\{\, p : H_0(q) \text{ has been accepted in level } \alpha \text{ test}\}$$

The count statistic used for eCDF



95% CI for $CDF_\pi(0.16)$ based on count statistic

e.g. let's test the null $H_0 : CDF_\pi(0.16) \leq 0.40$ where we reject H_0 for large $T(0.16) = \#\{i : x_i/n_i \leq 0.16\}$

- In the no-noise case ($p_i \equiv x_i/n_i$) for *any* null π we have

$$T(0.16) \leq Binom(129, 0.40)$$

  therefore 69 is a very LARGE count (p-value $= 0.0013$)

- For the noisy case consider null π for which $p_i = 0$ wprob 0.40 and $p_i = 0.1601$ wprob 0.60 for which

$$T(0.16) \approx 0.40 \cdot 129 + Binom(0.60 \cdot 129, 0.5)$$

  69 is actually a small count (p-value > 0.50)

95% CI for $CDF_\pi(0.16)$ based on count statistic (cont.)

Now, let's test the null $H_0 : CDF_\pi(0.16) = 0$

- In the noisy case for null π that assigns all the mass to $p_i = 0.1601$ we get

$$T(0.16) \sim Binom(129, 0.5)$$

  so 69 is greater than the mean but insignificant (p-value $= 0.241$)
- i.e. for the noisy case the 95% for $CDF_\pi(0.16)$ based on the count statistic is $[0, 1]$

“Bayesian” tests for composite hypotheses

General framework for testing composite null and alternative hypotheses presented in Yekutieli (2014) for testing Simpson’s Paradox

- The parameter is $\boldsymbol{\pi} \in \Omega$ with prior distribution is $\mathcal{D}(\pi)$
- the data is $\boldsymbol{X} = (X_1 \cdots X_k)$ and the likelihood is $\Pr(\boldsymbol{x}|\ \pi)$
- The null hypothesis is $H_1\ :\pi \in \Omega_0$ for $\Omega_0 \subseteq \Omega$
- The alternative hypothesis is $H_1\ :\pi \in \Omega_1$ for $\Omega_1 = \Omega - \Omega_0$
- For rejection region S, test $\mathcal{T}(S) := I(\boldsymbol{x} \in S)$ is a mapping $\mathcal{T}(S):\ \Omega \to \{0, 1\}$, with $\mathcal{T} = 1$ corresponding to rejecting H_0

Bayesian generalization of Neyman-Pearson tests

Testing is viewed as a classification problem with loss:

$$L(S;\lambda_1,\lambda_2) = \lambda_1 \cdot I(\boldsymbol{X} \in S, \pi \in \Omega_0) + \lambda_2 \cdot I(\boldsymbol{X} \notin S,\ \pi \in \Omega_1).$$

Classifier S that minimizes the average risk is

$$S^{Bayes}(\lambda_1,\lambda_2) = \{\boldsymbol{x} :\ \frac{\lambda_1}{\lambda_1+\lambda_2} \le \Pr(\pi \in \Omega_1 |\, \boldsymbol{x})\}$$

<u>Method:</u> order data sample space according to $\Pr(\pi \in \Omega_1 |\, \boldsymbol{x})$; use this ordering to sequentially enter data points in to S; set rejection threshold according to the significance level $\sup_{\pi_0 \in \Omega_0} \Pr(\boldsymbol{x} \in S|\pi_0)$

$\Rightarrow S^{Bayes}(\alpha)$ is Bayes classifier with significance level α

Mean Most Powerful tests

Properties

- The *mean significance level* of $\mathcal{T}(S)$ is $Pr(\boldsymbol{N} \in S | \, \boldsymbol{p} \in \mathcal{P}_0)$
- The *mean power* of $\mathcal{T}(S)$ is $Pr(\boldsymbol{N} \in S | \, \boldsymbol{p} \in \mathcal{P}_1)$
- $\mathcal{T}(S)$ is a *mean most powerful* test if all tests with less or equal mean significance level have less or equal mean power.
- Per construction, $S^{Bayes}(\alpha)$ is a mean most powerful test.

Relation to other approaches

- A generalization of likelihood ratio tests...
- The MMP statistic is equal to one minus the local FDR
- Proportional to a Bayes factor between models H_1 and H_0.

MMP test statistic for step functions

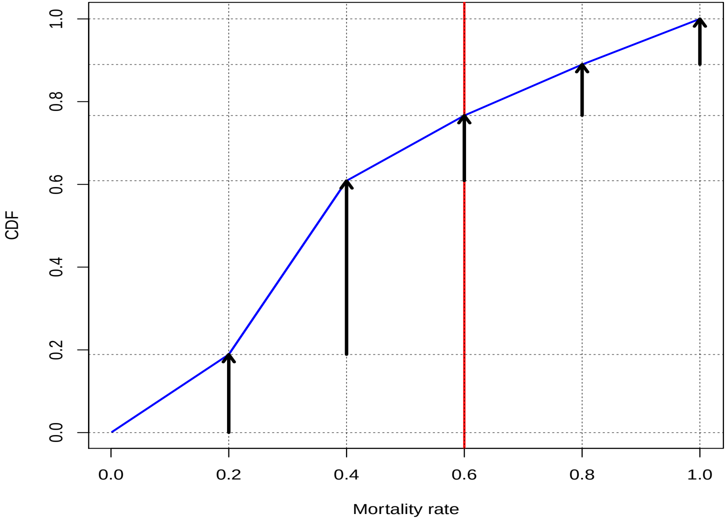
1. Partition $[0,1] = [a_0, a_1] \cup \cdots \cup [a_{I-1}, a_I]$ where $\exists a_{i_p} = p$
2. We consider distributions that are step function in this partition

$$\Omega^S = \{\pi_1 \cdot \frac{I(\theta \in [a_0, a_1])}{a_1 - a_0} + \cdots + \pi_I \cdot \frac{I(\theta \in [a_{I-1}, a_I])}{a_I - a_{I-1}} \; : \; \sum \pi_i = 1\}$$

3. We use the one-to-one correspondence between $\pi \in \Omega^S$ and $\vec{\pi} = (\pi_1 \cdots \pi_I)$ to define $\mathcal{D}(\pi)$ as the $Dirichlet(\vec{\alpha})$ density
4. Note $CDF_\pi(p)$ is $\pi_p = \pi_1 + \cdots + \pi_{i_p}$

Idea: derive MMP statistic for the step function sample space Ω^S and then use it to test the hypotheses regarding Ω

5-interval step function CDF for tests on $CDF_\pi(0.60)$



The MMP test in the no-noise case

Suppose we get to observe $p_1 \cdots p_K$

- Then we have $n_i = \#\{k : p_k \in [a_{i-1}, a_i]\}$, for which

$$\vec{\pi} | \vec{n} \sim Dirichlet(\vec{\alpha} + \vec{n})$$

- Thus for $n_p = n_1 + \cdots + n_{i_p}$,

$$\pi_p | \vec{n} \sim Beta(\alpha_p + n_p, (\alpha_+ - \alpha_p) + (I - n_p))$$

i.e. for any choice of step intervals and $\vec{\alpha}$ the MMP test is the binomial test that sorts the data sample space according to the count statistic

The MMP test in the general noisy case

- For $\delta_K = i$ iff $p_k \in [a_{i-1}, a_i]$, conditional on $\vec{\delta} = (\delta_1 \cdots \delta_K)$

$$\vec{\pi} | \; \vec{\delta} \; \sim \; Dirichlet(\vec{\alpha} + \vec{n}(\vec{\delta}))$$

  and the conditional statistic value $T(\vec{\delta})$ is the the CDF of a Beta
- The (unconditional) test statistic value is

$$T(\boldsymbol{x}) = \frac{\sum_{\vec{\delta}} T(\vec{\delta}) \cdot \Pr(\boldsymbol{x}|\vec{\delta}) \cdot \Pr(\vec{\delta})}{\sum_{\vec{\delta}} \Pr(\boldsymbol{x}|\vec{\delta}) \cdot \Pr(\vec{\delta})}$$

- and we also show that the statistic distribution is increasing in π

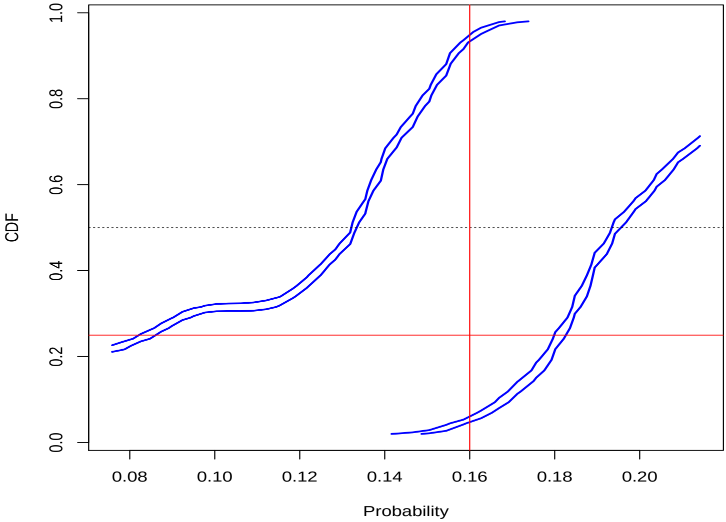
$\Rightarrow$ Therefore π_0^{max} that assigns mass q at p and mass $1-q$ at 1 yields largest test statistic distribution of all π_0 in $\Omega_0 = \{\pi : q \leq CDF_\pi(p)\}$

CI's for quantiles and the CDF of π

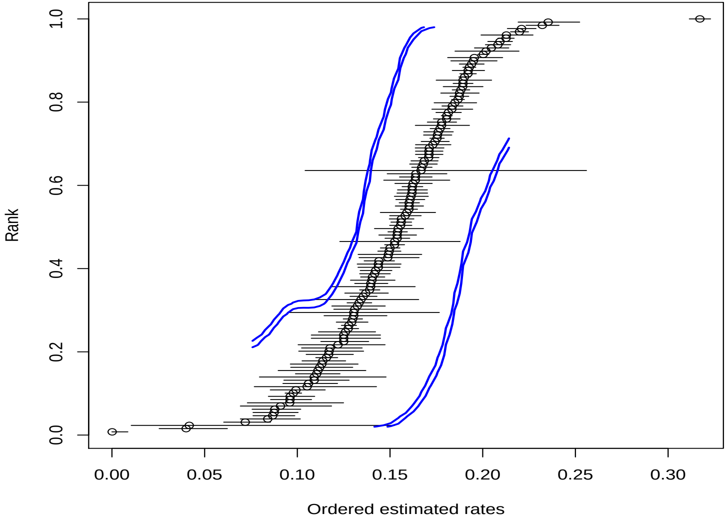
Algorithm:

1. Compute test statistic values for testing $H_0 : CDF_\pi(p) \leq q$ and $H_0 : CDF_\pi(p) \geq q$ for a grid of $q \in [0, 1]$ and $p \in [0, 1]$ values.
2. At each grid point, compute statistic values for N null data samples from π_0^{max}
3. Assess significance by proportion of null samples with test statistic values greater than observed test statistic values
4. point-wise CI's are 0.05 or 0.025 contours of the significance level surface

CI's for VA COPD readmission data



CI's for VA COPD readmission data (overlay)



Discussion

- It is very difficult to come up with good statistics for messy high dimensional tests
- The MMP approach produces automatic likelihood-based weighing of information from different hospitals
- Choice of Intervals and concentration parameter (and more generally the prior) is very important
- Are these good CI's? optimality for CI's? Consistency?

Thank you!