Statistician's (Personal) Summary

Richard Lockhart

Simon Fraser U

Phystat-ν 2016 Fermilab

September 21, 2016

Outline

- This has been a lot of fun for me.
- What issues did we see?
- What kinds of things do I like?
- Random remarks on specific talks.
- Calibrated Bayes.

Wilk's theorem

- Wilk's theorem came up many times.
- The theoretical framework is this.
  1. Log-likelihood is a sum of a 'large' number of terms.
  2. Some local maximum of log-likelihood is 'close' to true parameter value.
  3. Gradient vanishes at that local maximum.
  4. Two term Taylor expansion of log-likelihood.
  5. Quadratic in ball surrounding true value, max over ball at MLE.
  6. Dimension of parameter vector fixed, not big compared to say n.
- For parameters on boundary ($\sin^2(\theta) = 0$) points 2, 3, and 5 likely to fail.

Example 1 of 2: many nuisance parameters; Neyman-Scott

- X_i, Y_i pair of measurements of θ_i; total of n pairs.
- Gaussian errors, common SD σ.
- Test famous theory that $\sigma = 1$.
- Log-likelihood is

$$\ell(\mu_1, \ldots, \mu_n; \sigma) = -\frac{\sum_i \left[(X_i - \theta_i)^2 + (Y_i - \theta_i)^2\right]}{2\sigma^2} - 2n \log(\sigma)$$

- MLEs: $\hat{\theta}_i = (X_i + Y_i)/2$ and $\hat{\sigma^2} = T/(2n)$ where

$$T = \sum_i \left[(X_i - \hat{\theta}_i)^2 + (Y_i - \hat{\theta}_i)^2\right] = \frac{1}{2} \sum_i (X_i - Y_i)^2$$

- Log-likelihood ratio test statistic simplifies to

$$\Lambda \equiv 2\left[\ell(\hat{\mu}_1, \ldots, \hat{\mu}_n; \hat{\sigma}) - \ell(\hat{\mu}_1, \ldots, \hat{\mu}_n; 1)\right] = \frac{T}{2} - n - 2n \log(T/(2n))$$

- When $\sigma = 1$, null is right, $\Lambda \to \infty$, not χ^2 as $n \to \infty$.

Example 2 of 2: Mixture models

- Observations $X_1, \ldots, X_n$ from density

$$\frac{\theta}{\sigma_1\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-\theta}{\sigma_2\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}$$

- Five parameters. Not *identifiable* (I forget who already said this).
- Log-likelihood has n places where it is infinite.
- But: if true π not 0, 1 and two normal distributions not equal then there is *local maximum* where theory applies.
- Quality of quadratic approximation not uniform.
- Wilks theorem valid for some hypotheses; not for $\pi = 0$.

Mean parameters in variance: Bob Cousins

- Thermoluminescence dating of sand dunes.
- Photon count N_i when sample from sand dune core heated.
- Sand core irradiated with dose D_i. Modelled as

$$\begin{aligned} N_i &= f(D_i + D_0, \theta)(1 + \sigma\epsilon_i) \\ &= f(D_i + D_0, \theta) + \sigma f(D_i + D_0, \theta)\epsilon_i \end{aligned}$$

  with independent mean 0 standard deviation 1 noise ϵ_i.
- Variability proportional to mean.
- D_0 proportional to burial duration of sample.

Three estimation schemes

- Least squares: Minimize
$$\sum_i \frac{(N_i - f(D_i + D_0, \theta))^2}{f^2(D_i + D_0, \theta)}$$
- Weird Least squares: Minimize
$$\sum_i \frac{(N_i - f(D_i + D_0, \theta))^2}{N_i^2}$$
- Iteratively reweighted least squares. Start with $\sigma_{(0)} = 1$. Minimize
$$\sum_i (N_i - f(D_i + D_0, \theta))^2$$
to get initial values of $D_{0,0}$ and θ_0.
- Get $k+1$st set of parameters from kth by minimizing
$$\sum_i \frac{(N_i - f(D_i + D_0, \theta))^2}{f^2(D_i + D_{0,k}, \theta_k)}.$$

Lessons from the project

- All three methods different even in large samples.
- Only third method is ‘right’.
- Equations solved by estimates are the objects to study mathematically.
- Can also do low-noise asymptotics with n fixed.
- Glenn Berger, Jen-ni Kuo and L in *Nuclear Tracks etc*.

Example 3 of 2: von Mises

- $X_1, \ldots, X_n$ random angles in $[0, 2\pi)$.
- von Mises density:

$$\frac{1}{2\pi I_0(\kappa)} e^{k \cos(x-\theta)}.$$

- Two parameters: κ and θ.
- Test null $\kappa = 0$; uniform density on circle.
- Wilk's theorem applies. 2 degrees of freedom.
- Problem is that polar co-ordinate transformation is singular at origin.
- Rewrite density in form

$$\frac{1}{2\pi I_0(\sqrt{\tau_1^2 + \tau_0^2})} e^{\tau_1 \cos(x) + \tau_2 \sin(x)}$$

  with $\tau_1 = \kappa \cos(\theta)$ and $\tau = \kappa \sin(\theta)$.
- Null is $\tau_1 = \tau_2 = 0$ in interior of parameter space – the plane.
- Same thing happened in Scott Oser's talk.

Unfolding

- Very useful for presentation, for evaluation of images by eye.
- Natural to provide estimate of interpretable quantity.
- So unfolding is clearly worthy of study.
- Less clear to me that you should unfold before analysis.
- Historically statisticians favoured transforming data so that assumptions like Gaussian, linear models, constant errors were nearly met.
- We have moved to modelling just writing down likelihoods.
- The folding matrix is there in the likelihood of course.
- Is my transformation analogy relevant?

Decision Theory, Likelihood Ratios

- Data $\mathbf{Y}$.
- Two possible densities, f (null) and g (alternative).
- Test function T; notation $\mathbf{T} = T(\mathbf{Y})$.
- Level is $\mathrm{E}_f(\mathbf{T})$.
- Power is

$$\mathrm{E}_g(\mathbf{T}) = \mathrm{E}_f\left[\mathbf{T}\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right] = \text{Level} + \mathrm{Cov}_f(\mathbf{T}, LR).$$

where

$$\text{LR} = \text{Likelihood Ratio} = \frac{g(\mathbf{Y})}{f(\mathbf{Y})}.$$

Primitive Frequency Theory Ideas

- Neyman and Pearson were advocates of worst case analysis.
- Real Bayes is average case analysis.
- Calibrated Bayes is sort of in between perhaps.
- Hypothesis testing and confidence sets can certainly be combined.

Separate hypotheses

- Example 1. Sample $X_1, \ldots, X_n$ either from Normal$(-1, 1)$ or from N$(1, 1)$.
- Example 2. Sample from f or g.
- Example 3. Sample either from Gamma distribution or from Weibull distribution.
- Example 4. X, Y independent normals, means μ_x, μ_y. Hypothesis A: $\mu_x \leq 0, \mu_y \leq 0$; Hypothesis B $\mu_x \geq 0, \mu_y \geq 0$.

Two simple hypotheses $\mu = -1$ or $\mu = 1$

- Total error rate minimized by: pick $\mu = -1$ iff sample mean negative.
- Total error rate is

$$2P(\text{Normal} > \sqrt{n}) \sim 2\frac{\exp -n/2}{\sqrt{2\pi n}}$$

  which goes to 0 pretty fast.
- For non-normal models details of tails of law of $\bar{X}$ matter.
- Called *large deviations* regime.
- Rates specific to distributions, total error rate stunningly low.
- Generally not a good approximation.

Two simple hypotheses f or g

- Log-likelihood ratio is

$$\Lambda = \sum \log \{f(X_i)/g(X_i)\}$$

- Study when g is true density. Maybe CLT applies?
- If so $\Lambda \sim N(\mu, \sigma^2)$ approximately. But

$$\mathrm{E}_g \left[\frac{f(X)}{g(X)} \right] = \int \frac{f(x)}{g(x)} g(x)\, dx = \int f(x)\, dx = 1$$

so

$$\mu = -\sigma^2/2$$

- AND if f is true density then

$$\Lambda \sim N(\sigma/2, \sigma^2).$$

- Notice symmetry about 0; equal variances, means $\pm\sigma^2/2$.
- Basic ingredient in version of large sample theory called *contiguity*.

Asymptotic methods in statistics

- For fixed f, g stuff on previous slide is nonsense.
- Can be real if at least one of f or g varies with n so that f, g get closer together.
- Neyman made calculations that way.
- Usually in parametric models; null θ_0, alternative $\theta_0 + \gamma/\sqrt{n}$.
- Has impact in Wilks Theorem failures.

Separate Hypotheses, example 4

- Test lower left quadrant against upper right quadrant of plane.
- Log-likelihood ratio 0 in lower left quadrant.
- Compares X, Y to $0, 0$ when (X, Y) in first quadrant.
- When $X > 0$, $Y < 0$ compare X, Y to $0, Y$.
- But what is the distribution when $\mu_x \leq 0$, $\mu_y \leq 0$?
- Worst case analysis: $\mu_x = \mu_y = 0$; mixture of χ^2.
- But when you see $Y << 0$ should you say corner wrong?

Parametric Bootstrapping

- Most commonly statisticians seem to favour parametric bootstrap to attach P-value to statistic.
- If you have to do parametric bootstrapping to compute the P-value for a test statistic T then the real test statistic is the P-value.
- Neyman and Pearson say: pick a test statistic and a critical value and reject if the statistic is larger than the critical value.
- The critical value doesn't get to change depending on the data.
- But *lots* of statistical procedures don't work that way!
- Different way to say this: it is the critical region which matters. Which data sets lead to rejection? How uniform is P for different parameter values.

Why asymptotic (large sample) methods?

- Statistic T_n. Distribution depends on θ.
- Compute things like $P(T_n > t|\theta)$.
- Make approximation

$$P(T_n > t|\theta) \approx \lim_{n\to\infty} P(T_n > t|\theta).$$

- Sometimes limit is discontinuous in θ.
- But the thing being approximated is not, for any finite n.
- Cries out for different analysis in neighbourhood of discontinuity.
- I have some examples in my work.
- Example 4 above is like that.
- Aixin Tan's work like this, I believe.

LEE, 5σ, error rate control

- Plots of local *P*-values dipping below 3×10^{-7} are not 5σ effects in a statistician's mind.
- A global *P*-value like that would be.
- Era of automated searches magnifies LEE.
- False Discovery Rate work relevant.
- Ioannidas (2005) *PLoS Medicine*. Why Most Published Research Findings Are False
- Much attention in statistical community to error rates in published work.
- Much attention (Kuffner) to adjusting inference to account for process of tuning analysis.
- Schizophrenia in statistical community.
- Blind analysis methods relevant.

Some remarks

- In goodness-of-fit there is an alternative hypothesis.
- It is vague – described by the test statistic, essentially.
- 'Which one is the correct one?' No statistician ever answers this question credibly.
- Most of us don't think the question has an answer.
- Questions like: what are the weaknesses of this method? Is there a clearly better method than this?

On off calibrated Bayes example

- Y is a measurement of background plus signal, $b + s$.
- Formal: X is a measurement of a background b
- Take X ∼Poisson(b), and Y ∼Poisson($b + s$).
- Null hypothesis $H_o : s = 0$ is 'composite' unless background b is known so that X is useless.
- Log-likelihood is

$$\ell(b, s) = X \log(b) + Y \log(b + s) - 2b - s$$

Neyman Pearson lemma

- So now suppose $b = 20$ is known without error.
- And suppose the signal is either $s = 0$ or $s = 5$.
- Neyman says we need a rule for discovery.
- A rule is like declare discovery if $Y \in D$ for a set D (*Discovery* set or *Critical Region* or *Rejection Region*.
- Neyman Pearson approach minimize

$$\beta = P(Y \notin D | s = 5)$$

subject to constraint

$$\alpha = P(Y \in D | s = 0) = \alpha_0$$

Decision Theory

- Neyman and Pearson used Lagrange multipliers; minimize

$$\beta + \lambda\alpha$$

  then adjust α so solution also satisfies constraint.
- This is the same as minimizing

$$\frac{1}{1+\lambda}\beta + \frac{\lambda}{1+\lambda}\alpha \equiv (1-\pi)\beta + \pi\alpha$$

- So Neyman and Pearson's solution is Bayesian but with the prior adjusted to satisfy constraints.
- Sweeping discreteness under the rug.
- For $b = 20$ find, for 5σ stringency, $D = \{y \geq 47\}$.
- So if we see 47 or more events we declare discovery.

Unknown background

- But if b is unknown then the Neyman Pearson lemma doesn't help.
- A likelihood ratio test is the ancient suggestion; it has, usually, no totally compelling justification
- Compare $b, s = 0$ to b, s for 'most likely' values under each possible assumption.
- If $s = 0$ guess $b = \hat{b}_0 = (X + Y)/2$; most likely value in null hypothesis.
- Alternative: if $Y < X$ then $\ell(b, s)$ is maximized at $\hat{s}_1 = 0$ and $\hat{b}_1 = (X + Y)/2$ while for $Y \geq X$ the maximum is at $\hat{s} = Y - X$ and $\hat{b} = X$.
- So the likelihood ratio is

$$\text{Deviance drop} = 2\left[\ell(\hat{b}_1, \hat{s}_1) - \ell(\hat{b}_0, 0)\right]$$

- Algebra skipped. In limit $b \to \infty$ statistic has a χ^2_1 distribution.

Calibrated Bayes, decision theory

- Recast problem from Hypothesis Testing and Confidence Interval to variable level confidence set.
- Output of inference is set S of possible values for s.
- Loss is sum of penalties for errors:

$$\begin{aligned} L(s,S) = & \ell_0 1(s=0) 1(0 \notin S) && \text{null incorrectly rejected} \\ & + C_0 1(s \neq 0) 1(0 \in S) && \text{null incorrectly accepted} \\ & + \ell_s 1(s \neq 0) 1(s \notin S) && \text{alternative not covered} \\ & + \int_{t>0} C_t 1(t \in S)\, dt. && \text{overcoverage on alternative} \end{aligned}$$

Prior distributions

- Idea is to choose a *proper* prior on s:

$$\pi(0)\delta(s) + \pi(s) \text{ continuous prior on } s > 0.$$

- Imagine theory suggests order of 50 events expected.
- Perhaps continuous part exponential distribution prior

$$\pi(s) = \exp\{-s/50\}/50$$

- Posterior Bayes risk of set S is

$$r_\pi(S|X,Y) = \pi(0|X,Y)\ell_0 1(0 \notin S) + (1 - \pi(0|X,Y))C_0 1(0 \in S) + \int_{S^c} \pi(s|X,Y)\ell_s\, ds + \int_S C_s\, ds$$

- For positive s put $s \in S$ if

$$\pi(s|X,Y)\ell_s > C(s) \quad \equiv \quad \pi(s|X,Y)\frac{\ell_s}{C_s} > 1$$

- Put $0 \in S$ if

$$\frac{\pi(0|X,Y)}{1 - \pi(0|X,Y)}\frac{\ell_0}{C_0} > 1.$$

Calibration

- Wiggle ℓ_s and C_s *or* $\pi(s)$ to control coverage probabilities.
- Can insist

$$P(0 \in S|s = 0) = 1 - P(\text{Normal}(0, 1) > 5)$$

  and

$$P(s \in S|s) = 0.95 \text{ for } s > 0.$$

- Like Feldman-Cousins – points added in order determined by posterior.
- So have 5σ for discovery, 95% for confidence.
- Calibration removes the bulk of the influence of the prior.
- Have done this for Higg's discovery on Monte Carlo data with much more complex Bayesian model for background.
- Calibration by importance sampling down to *global* $\alpha = 10^{-4}$.
- Use Integrated Nested Laplace Approximations (INLA); Rue and Martino.
- All laptop stuff. Only one channel, 5 production modes summed. Could do much better with a computer.

Our interval for $X = 100$, varying Y; 3σ, 95%