Stopping rule paradox; Helicopter tour

Christopher A. Sims
Princeton University
sims@princeton.edu

November 11, 2019

Boy bias

- Consider a place where every family continues having children until a boy is born, then stops.

- What about estimating the probability of a boy (assuming that biologically the probability is constant and independent across births) as $r = n_b/(n_b + n_g)$?

- This is obviously frequentist biased if the probability p of a boy is .5: $E[r] = .5 \times 1 + .25 * .5 + .125 * 1/3 + \cdots > .5$, whereas if families were simply of random size, or all the same size, we would have $E[r] = .5$.

- But the likelihood function is just $p^{n_b}(1-p)^{n_g}$ in either case. $(\mathrm{Beta}(n_b + 1, n_g + 1))$

Bayesian inference for this case

- MLE is $n_b/(n_b + ng)$, but this is not flat-prior posterior mean.

- Posterior means are $(n_b + 1)/(n_b + n_g + 2)$. $\frac{2}{3}$ when $(n_b, n_g) = (1, 0)$, .5 for (1,1), .4 for (1,2), etc.

- With data for multiple families, likelihood is $p^{n_b}(1 - p)^{n_g}$ over all the children in all families.

- Focusing on likelihood avoids the mistake of averaging r across families instead of using sample r across children.

Effective medicine example

- Suppose we thought a priori that the medicine has an observable “effect” x_i in each observation i. This could be, e.g., the difference between some outcome between treated and placebo inidividuals in a matched pair. We assume x_i i.i.d. with density, conditional on mean μ, $N(\mu, 1)$. The known variance is just to make calculations simpler.

- With some probability π the medicine is ineffective, meaning $\mu = 0$, Otherwise $\mu > 0$, with a prior density, e.g., $U(0, B)$.

- The parameter space then consists of 0, which we give base measure 1, and $(0, B)$, on which we use Lebesgue base measure. The likelihood over this space is then proportional to

$$\exp\Bigl(-\tfrac{1}{2}\sum_{1}^{N}(x_i-\mu)^2\Bigr) \propto \exp(-\tfrac{1}{2}N(\bar{x}-\mu)^2)\,.$$

So the posterior density is proportional to

$$\pi\exp(-\tfrac{1}{2}N(\bar{x})^2) \quad \text{on } \mu=0, \quad (1-\pi)B^{-1}\exp(-\tfrac{1}{2}N(\bar{x}-\mu)^2)$$

To get the posterior weight on $\mu > 0$ we have to integrate the posterior density over $(0, B)$. If B is large, or if N is large, the right tail part of the integral will be small, so it will be close to

$$(1-\pi)B^{-1}N^{-1/2}\Phi(N^{1/2}\bar{x})\cdot(2\pi)^{1/2}$$

The test statistic for $\mu = 0$ is $N^{1/2}\bar{x}$, so if, as we have assumed, the sample size was set by making this statistic 2, the integral over the $\mu > 0$ interval is then

$$(1-\pi)B^{-1}N^{-1/2}\Phi(2)\cdot(2\pi)^{1/2}\,.$$

The weight on $\mu = 0$ with $N^{1/2}\bar{x} = 2$ is just πe^{-2}. Since the weight on $\mu = 0$ is constant, while that on the $\mu > 0$ interval decreases at the rate $1/\sqrt{N}$, the odds, for enough large sample sizes, favor $\mu = 0$. Rejection of the $\mu = 0$ null at just the .05 level in a very large sample, becomes evidence in *favor* of $\mu = 0$.

But note that, no matter the sample size, the odds ratio depends strongly on B. This is not a defect of Bayesian inference; it reflects a necessary aspect of inference in this type of problem. To interpret the evidence, we need to assess whether $\bar{x}$ is large or amall relative to what we expected a priori.

General form of the stopping rule principle

- If the stopping rule depends only on the data vector Y, then $p(Y \mid \theta)$ does not depend on the stopping rule. This is easy to see in the "boy bias" example. The probability of a sequence of girls and boys is unaffected by the fact that we know the sequence will stop as soon as a boy appears.

- Another way to see this is to observe that with a stopping rule of the form "stop if $g(Y) > 0$", the sample size T is a function of Y. We could compute a distribution for it from the distribution of Y, but it provides no information about the parameters not already present in $p(Y \mid \theta)$.

Caveats

- That posteriors don't depend on the stopping rule assumes that we observe the likelihood for the full data set.

- For example, if in the medicine example the researcher has actually initiated many trials, but reported only the first one that achieved "significance", there is a selection bias. The correct likelihood has to account for the selection bias.

- If we are told *only* that the researcher has found a t-statistic of plus 2.0, with no information about $\bar{x}$ or T, our conclusions would depend on the stopping rule. Of course if the stopping rule is that the t-statistic alone is reported when it reaches 2.0, the report by itself provides no information at all about μ.

A stopping rule paradox for the AR1

$$y_t = \rho y_{t-1} + \varepsilon_t\,, \quad t = 1, \ldots, T$$

$$\hat{\rho} = \rho + \frac{\sum y_{t-1}\varepsilon_t}{\sum y_{t-1}^2}\,.$$

The OLS estimator $\hat{\rho}$ is the MLE, when we form the likelihood by conditioning on y_0. But, as is well known, it is biased downward when $\rho > 0$, and more biased the closer ρ is to one.

The log likelihood is quadratic in ρ, centered at $\hat{\rho}$, so the flat-prior posterior makes $\hat{\rho}$ the mean, with no "bias correction".

The paradox

The numerator of $\rho - \hat{\rho}$, $\sum y_{t-1}\varepsilon_t$, is a sum of martingale-differences, and is asymptotically normal even when $\rho = 1$ (though in that case it has to be normalized by $1/T$ instead of $1/\sqrt{T}$). The bias comes from the fact that the denominator and numerator are not asymptotically uncorrelated.

However, if we learn that T, rather than being fixed in advance, was determined by adding to the sample until $\sum y_{t-1}^2$ first exceeded an a priori fixed target B, the denominator would no longer be "random" across samples. (There would be slight randomness from the fact that it would always slightly exceed B, but this has negligible effect if B is large.) Then from a frequentist viewpoint $\hat{\rho}$ is unbiased asymptotically, even if $\rho = 1$. This is not even entirely unrealistic: researchers tend not to bother with reporting estimates from samples that they can see are too small to give precise estimates of what they are interested in.

Archers of varying ability

- How can it be that the distributio of $\rho \mid \hat{\rho}$ is symmetric about ρ, while that of $\hat{\rho} \mid \rho$ is skewed downward?

- Consider a line of archers, each shooting at her own target. In any adjacent pair, the one on the left is known to be much more accurate. Her arrows are less likely to miss the target and end up in the grass.

- If they do miss the target, both do so symmetrically: their arrows are equally likely to miss to the right or the left.

Estimating whose arrow it is

- We want an estimator of who shot the arrow, based on data on the location (off target) where the arrow was found.

- Consider two archers in the middle, each of whom has neighbors on each side.

- If we allocate an arrow found between their two targets to whichever archer's target is closer to the arrow, are allocation is unbiased. Conditional on archer A having shot the arrow, it is equally likely that the arrow is mistakenly allocated to the left or to the right of the actual shooter.

- However, common sense (and Bayesian inference) tells us that if the arrow is equidistant between two targets, it is more likely to belong to the less accurate shooter on the right than to the archer on the left.

This explains the AR1 paradox

- The archery example shows that if the accuracy of an unbiased estimator $\hat{\theta}$ increases as θ increases, then after observing $\hat{\theta}$, it makes sense to believe that θ is probably below $\hat{\theta}$, since an upward error from a true θ below $\hat{\theta}$ is more likely than the reverse.

- In the AR1 example, accuracy of $\hat{\rho}$ depends on $\sum y_{t-1}^2$, which is likely to be much larger when ρ is greater than or equal to 1 than when $|\rho| < 1$. This by itself would make us believe ρ is above $\hat{\rho}$. But $\hat{\rho}$ is biased downward, and this exactly offsets the effect of its accuracy increasing with ρ.