KU Leuven Summer School
Segment 6
Bayesian Calibration

Paul Gustafson

September 16, 2022

Need a simple sandbox to play in

A stripped-down misclassification problem.

Interested in the prevalance, r, of a disease in a population.

The diagnostic test for the condition is known to have perfect *specificity*, i.e., no false positives.

But there is ambiguity about the *sensitivity*, i.e., could be some false negatives.

The diagnostic test is applied to a random sample of n individuals from the target population.

Generative model (collapsed version)

$Y \mid r, Sn \sim \text{Bin}(n, r)$

$Y^* \mid Y, r, Sn \sim \text{Bin}(Y, Sn)$

- $r \sim \text{Unif}(0, 1)$
- $Sn \sim \text{Beta}(a, b)$
- $(Y^* | r, Sn) \sim \text{Bin}(n, r \times Sn)$

And let's make things quite focussed:

Inputs:

- Dataset (Y^*, n)
- Expert opinion (hyperparameters) (a, b)

Output:

- (general) posterior distribution of (r, Sn)
- (specific) 80% equal-tailed credible interval for r

Computational implementation

Would be easy to implement in JAGS.

But nice to have something **faster,** and less beholden to **diagnostics**, to support **simulation studies**.

Turns out this is a **nearly conjugate** situation

- possible to do *iid* Monte Carlo draws from a decent approximation to the posterior distribution
- possible to correct for the approximation error via **importance weights** that *do not depend on the data*

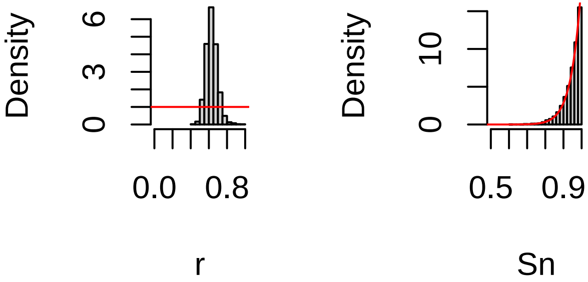
See the appendix for bespoke code, if interested.

Example

Say the data are $(Y^*, n) = (60{,}100)$.

Say the investigator believes that the diagnostic test could be *slightly* insensitive, chooses hyperparameters $(a, b) = (19, 1)$

Marginal posterior distributions of r and Sn, compared to prior



And then focus on the credible interval

From bespoke code, 80% credible interval for r:

this one wider and pushed upward ✓ (relative to)

```
cred.int(60,100, hyp=list(a=19,b=1))
```

```
## [1] 0.558 0.713
```

For comparison, the corresponding interval if the investigator assumes $Sn = 1$

fully conjugate

```
qbeta(c(0.1,0.9), shape1=1+60, shape2=1+(100-60))
```

```
## [1] 0.536 0.660
```

Thought (well simulation) experiment #1A

Frequentist coverage of the 80% credible interval?

At a particular spot in the parameter space.

```
n <- 100; r.tr <- 0.7; sn.tr <- 0.85; m <- 1600

ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   ystr[i] <- rbinom(1, size=n, prob=r.tr*sn.tr)
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr) & (r.tr<intrvl[i,2])
}
```

Experiment #1A, continued

maybe not so

```
head(cbind(r.tr, sn.tr, ystr,intrvl,cover),8)
```

```
##       r.tr sn.tr ystr               cover
## [1,]  0.7  0.85   52 0.477 0.628     0
## [2,]  0.7  0.85   50 0.455 0.608     0
## [3,]  0.7  0.85   51 0.466 0.616     0
## [4,]  0.7  0.85   64 0.598 0.753     1
## [5,]  0.7  0.85   58 0.534 0.691     0
## [6,]  0.7  0.85   57 0.526 0.679     0
## [7,]  0.7  0.85   57 0.527 0.682     0
## [8,]  0.7  0.85   67 0.630 0.785     1
```

surprising - true Sn is in the tail of the prior

```
### frequentist coverage
mean(cover)
```

well below 80%

```
## [1] 0.571
```

```
### average width
mean(intrvl[,2]-intrvl[,1])
```

```
## [1] 0.154
```

Thought experiment #1B

Frequentist coverage of the 80% credible interval?

At a different spot in the parameter space.

```
n <- 100; r.tr <- 0.8; sn.tr <- 0.97; m <- 1600

ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   ystr[i] <- rbinom(1, size=n, prob=r.tr*sn.tr)
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr) & (r.tr<intrvl[i,2])
}
```

Experiment #1B, continued

```
head(cbind(r.tr, sn.tr, ystr,intrvl,cover),8)
```

```
##       r.tr sn.tr ystr             cover
## [1,]  0.8  0.97   83 0.801 0.935     0
## [2,]  0.8  0.97   80 0.767 0.909     1
## [3,]  0.8  0.97   82 0.789 0.928     1
## [4,]  0.8  0.97   74 0.703 0.854     1
## [5,]  0.8  0.97   70 0.660 0.815     1
## [6,]  0.8  0.97   81 0.778 0.918     1
## [7,]  0.8  0.97   75 0.716 0.864     1
## [8,]  0.8  0.97   80 0.767 0.912     1
```

maybe not surprising - true Sn near mode of prior

```
### frequentist coverage
mean(cover)
```

```
## [1] 0.894
```

well above 80%

```
### average length
mean(intrvl[,2]-intrvl[,1])
```

```
## [1] 0.146
```

Thought experiment #2A

i.e. from generative model
f(param) f(data|param)

Repeated sampling of **(parameter,data) pairs**

```
n <- 100; m <- 6400

r.tr <- sn.tr <- ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   r.tr[i] <- runif(1)
   sn.tr[i] <- rbeta(1, shape1=19, shape2=1)

   ystr[i] <- rbinom(1, size=n, prob=r.tr[i]*sn.tr[i])
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr[i]) & (r.tr[i]<intrvl[i,2])
}
```

Thought experiment #2A, continued

```
head(cbind(r.tr, sn.tr, ystr,intrvl,cover),10)
```

80% cred int.

```
##        r.tr sn.tr ystr             cover
##  [1,] 0.260 0.877   16 0.127 0.230     0
##  [2,] 0.508 0.947   37 0.325 0.466     0
##  [3,] 0.142 0.962   13 0.101 0.194     1
##  [4,] 0.806 0.964   80 0.768 0.911     1
##  [5,] 0.150 0.984   16 0.127 0.232     1
##  [6,] 0.639 0.964   63 0.590 0.744     1
##  [7,] 0.194 0.916   21 0.172 0.286     1
##  [8,] 0.854 0.994   83 0.800 0.934     1
##  [9,] 0.402 0.956   38 0.336 0.475     1
## [10,] 0.689 0.978   70 0.662 0.813     1
```

so within simulation error of the nominal 80%

```
mean(cover)
```

```
## [1] 0.793
```

$\pm 2\sqrt{\frac{(.79)(1-.79)}{6400}}$ i.e $\pm .01$

Is this general?

Let $A(Y^*)$ be the credible interval

Under the **generative model:**

jointly random

$$
\begin{aligned}
Pr\{\theta \in A(Y^*)\} &= E\{I_{A(Y^*)}(\theta)\} \\
&= E\{E\{I_{A(Y^*)}(\theta)|Y^*\}\} \\
&= E\{0.8\} \\
&= 0.8
\end{aligned}
$$

by defn of credible interval

very nice calibration property

Thought experiment #2B

Repeated sampling of (parameter,data) pairs *from a different distribution.*

Nature's prior

Investigator's prior

```
n <- 100; m <- 1600

r.tr <- sn.tr <- ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   r.tr[i] <- runif(1)
   sn.tr[i] <- rbeta(1, shape1=15, shape2=5)

   ystr[i] <- rbinom(1, size=n, prob=r.tr[i]*sn.tr[i])
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr[i]) & (r.tr[i]<intrvl[i,2])
}
```

TE #2B, continued

```r
head(cbind(r.tr, sn.tr, ystr,intrvl,cover),10)
```

```
##         r.tr sn.tr ystr               cover
##  [1,] 0.8037 0.724   53 0.4864 0.637     0
##  [2,] 0.3861 0.734   25 0.2103 0.332     0
##  [3,] 0.5380 0.804   47 0.4267 0.574     1
##  [4,] 0.0894 0.848   13 0.0996 0.193     0
##  [5,] 0.6420 0.738   55 0.5061 0.661     1
##  [6,] 0.5261 0.824   39 0.3456 0.489     0
##  [7,] 0.8525 0.882   79 0.7568 0.902     1
##  [8,] 0.2517 0.675   27 0.2293 0.357     1
##  [9,] 0.6042 0.779   46 0.4162 0.563     0
## [10,] 0.5692 0.764   47 0.4238 0.573     1
```

```r
mean(cover)
```

```
## [1] 0.397
```

consequence of mismatch between nature & investigator

Frequentist coverage of the Bayesian interval revisited

Same as 1A and 1B, just with more data now.

Thought Experiment #1A*

doubled

```
n <- 200; r.tr <- 0.7; sn.tr <- 0.85; m <- 1600

ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   ystr[i] <- rbinom(1, size=n, prob=r.tr*sn.tr)
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr) & (r.tr<intrvl[i,2])
}
```

```
mean(cover)
```

```
## [1] 0.429
```

even lower now

```
mean(intrvl[,2]-intrvl[,1])
```

```
## [1] 0.123
```

Thought Experiment #1B*

doubled

```
n <- 200; r.tr <- 0.8; sn.tr <- 0.97; m <- 1600

ystr <- cover <- rep(NA, m)
intrvl <- matrix(NA, m, 2)

for (i in 1:m) {
   ystr[i] <- rbinom(1, size=n, prob=r.tr*sn.tr)
   intrvl[i,] <- cred.int(ystr[i],n, hyp=list(a=19, b=1))
   cover[i] <- (intrvl[i,1]<r.tr) & (r.tr<intrvl[i,2])
}
```

```
mean(cover)
```

```
## [1] 0.922
```

even higher now

```
mean(intrvl[,2]-intrvl[,1])
```

```
## [1] 0.125
```

In textbook problems, what do we expect?

textbook = full information = fully identified

- Frequentist coverage of x-percent credible interval is approximately x, at *every* point in the parameter space.

(Approximately meaning asymptotically.)

(Asymptotically, Bayesian and frequentist match up.)

- Width of interval scales as $\frac{1}{\sqrt{n}}$

In problems like the one here, what happens instead (1 of 3)

- Frequentist coverage of x-percent credible interval *varies widely* across the parameter space
- But no matter what, the average (in a certain sense) of the frequentist coverage is *exactly* x.

recall $0.80 = E\left\{ I_{A(y^*)}(\theta) \right\}$

$$\underset{\text{averaging }\theta\text{ with respect to the prior}}{E}\; E\left\{ \underbrace{I_{A(y^*)}(\theta) \,/\, \theta}_{\text{freq. coverage at }\theta} \right\}$$

In problems like the one here, what happens instead (2 of 3)

- At some places in the parameter space, the frequentist coverage (of the x-percent credible interval) goes to ONE as n goes to infinity.
- And at all the other places, it goes to ZERO
- And we can say something very specific about the set of parameter values for which the limiting frequentist coverage is ONE, namely that

this set has probability x under the prior

In problems like the one here, what happens instead (3 of 3)

$a > 0$
$b > 0$

- The width of the x-percent interval will scale like: $\sqrt{a + \frac{b}{n}}$
- This will not win you friends with your subject-area collaborators, but it is what it is.
- Check this matches up with #1A → #1A*, #1B → #1B*

$\frac{.123}{.154} \approx .8 > \frac{1}{\sqrt{2}} \approx .7$

$\frac{.125}{.146} \approx .86 > \frac{1}{\sqrt{2}}$

Musings about Bayesian coverage (1 of 4)

Clearly having calibration in Bayesian coverage sense is not as strong as having calibration in a frequentist coverage sense

In math terms, say a given interval estimation procedure applied at a given sample size has frequentist coverage $fc(\theta)$, when the parameter value is θ.

Frequentist x-percent confidence interval: $fc(\theta) = x$, for every θ.

Bayesian x-percent credible interval using prior $\pi(\theta)$?

Only full/general guarantee is that $\int fc(\theta)\Pi(\theta)d\theta = x$

Musings (2 of 4)

But in many low-info problems, there **do not exist** interval estimation procedures satisfying $fc(\theta) = x$. (Either for fixed n, or in the large n limit)

In such problems, the Bayesian calibration is the only game in town?

Sidebar: Considerable technical literature (mostly in econometrics) on trying to construct procedures with $fc(\theta) \geq x$, for every θ.

Musings (3 of 4)

What's the narrative to practitioners?

I choose prior π as my pre-data "projection" about the state of the world (i.e., the underlying parameter values).

Post-data, I will be reporting an x-percent credible interval for a (scalar) target parameter.

Then with respect to my joint projection (of parameters and data), there is an x-percent chance I will cover the truth.

Musings (4 of 4)

Or phrased a little differently . . .

I direct a lab that will, over time, study the relationship between **different** exposure, disease pairs

I aspire to specify my prior distribution to correctly reflect the pair-to-pair variation in these associations.

If I meet my aspiration, then, in the long-run, x percent of the x-percent credible intervals the lab reports will contain the truth.

And now that you are primed to think about operating characteristics under repeated sampling of (parameter, data) pairs . . .

There is a sense of best possible estimation, as well as a sense of correct coverage

Let $\pi_{Nature}(\theta)$ be the distribution giving rise to the repeated sampling (along with the distribution of data D given θ)

Amongst **any and all estimators** the minimum mean-squared error (across the repeated sampling) is achieved by the posterior mean of ψ when the investigator's choice of prior distribution matches that of nature.

LAB

Thoughts?

Appendix

```
show(full.pst)
```

```
## function(ystr, n, hyp, m=10000) {
##
##   ### draws from approx posterior in (rstr, sn) parameterization
##   rstr <- rbeta(m, 1+ystr, 1+(n-ystr))
##   sn <- rbeta.trnc(m, rstr, hyp$a-1, hyp$b)
##
##   ### importance weights to correct for approximation
##   impwht <- 1-pbeta(rstr, hyp$a-1, hyp$b)
##   impwht <- m*impwht/sum(impwht)
##
##   ### back to (r, sn) parameterization
##   r <- rstr/sn
##
##   # resample according to the weights,
##   # to get MC representation of actual posterior
##   tmp <- sample(1:m, replace=T, prob=impwht)
##   list(r=r[tmp], sn=sn[tmp])
## }
## <bytecode: 0x000001880a5cde90>
```

Appendix, continued

```r
show(cred.int)
```

```
## function(ystr, n, hyp, m=10000, cr.lev=0.8) {
##
##   rstr <- rbeta(m, 1+ystr, 1+(n-ystr))
##
##   sn <- rbeta.trnc(m, rstr, hyp$a-1, hyp$b)
##
##   r <- rstr/sn
##
##   impwht <- 1-pbeta(rstr, hyp$a-1, hyp$b)
##   impwht <- m*impwht/sum(impwht)
##
##   c(weighted.quantile(r, impwht, (1-cr.lev)/2),
##     weighted.quantile(r, impwht, (1+cr.lev)/2))
## }
```

Appendix, continued

```
show(rbeta.trnc)
```

```
## function(m, lwr, a, b) {
##   qbeta(  runif(m, pbeta(lwr,a, b), 1), a,b)
## }
```