STAT 530: More on MC Approximation and Prediction

Jan 20, 2009

Recall: numerically approximate $E\{g(\theta)|Y=y\}$ by...

$$\bar{g} \; = \; \frac{1}{s}\sum_{i=1}^{s} g\left(\theta^{(i)}\right),$$

where $\{\theta^{(i)}\}_{i=1}^{s}$ are simulated *iid* draws from the posterior distribution. Approximation error $\propto 1/\sqrt{s}$.

Careful with variances:

$(s-1)^{-1}\sum_{i=1}^{s}\{g(\theta^{(i)})-\bar{g}\}^2$ is numerical approximation to $Var(g(\theta)|Y=y)$. Whereas...

$$\sqrt{\frac{(s-1)^{-1}\sum_{i=1}^{s}\{g(\theta^{(i)})-\bar{g}\}^2}{s}}$$

is a ‘simulation SE,’ describing the quality of the numerical approximation of $E(g(\theta)|Y=y)$ by $\bar{\theta}$.

Prediction

And recall $p(\tilde{y}|y_{1:n}) = \int p(\tilde{y}|\theta)p(\theta|y_{1:n})d\theta$.

So simulate $\{(\theta^{(i)}, \tilde{y}^{(i)})\}_{i=1}^{s}$ according to:

- $\theta^{(i)} \sim p\left(\theta|y_{1:n}\right)$
- $\tilde{y}^{(i)} \sim p\left(\tilde{y}|\theta^{(i)}\right)$

Thus $\{\tilde{y}^{(i)}\}_{i=1}^{n}$ represents $p(\tilde{y}|y_{1:n})$.

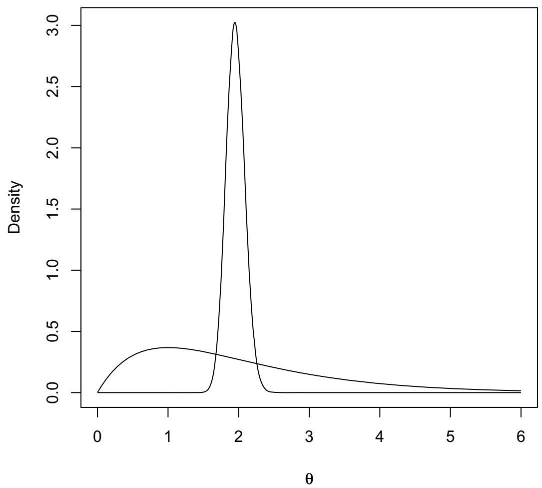
Example from text: Number of children for $n = 111$ 40-year old women without college degrees

# Children	0	1	2	3	4	5	6
Count	20	19	38	20	10	2	2

$$\begin{aligned} Y_1,\ldots,Y_n|\theta &\overset{iid}{\sim} \text{Poisson}(\theta) \\ \theta &\sim \text{Gamma}(2,1) \end{aligned}$$

So that $(\theta|Y_{1:n}=y_{1:n}) \sim \text{Gamma}(2+217, 1+111)$.

Prior to posterior updating



Predictive distribution out of sync with observed data?

```
> s <- 100000
> th.mc <- rgamma(s, 219, 112)
> ypred.mc <- rpois(s, th.mc)

> mean(ypred.mc==1)
[1] 0.27644
> mean(ypred.mc==2)
[1] 0.26898

> sqrt(var(ypred.mc==1)/s)
[1] 0.001414295
> sqrt(var(ypred.mc==2)/s)
[1] 0.001402254
```

Check more formally

Let $t()$ be some statistic, i.e. function of the dataset.

Now consider the predictive distribution over future datasets given the actual dataset (same n).

In terms of $t()$, does the actual dataset 'fit in' with the predictive distribution.

That is, does data generated from the supposed model, under *a posteriori* plausible parameter values, 'look like' the observed data?

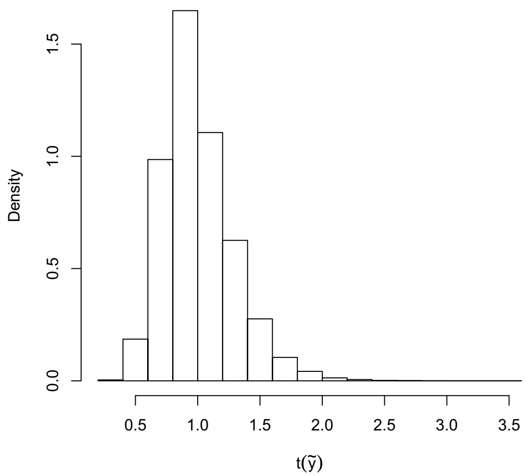
Where does $t(y_{obs})$ lie in the distribution of $t(\tilde{y})|y_{obs}$?

Predictive distribution over $t(\tilde{y}_1^n)$

```
> tfun <- function(y) {sum(y==2)/sum(y==1)}

> preddist <- rep(NA, s)
> for (i in 1:s) {
   preddata <- rpois(111, th.mc[i])
   preddist[i] <- tfun(preddata)
  }

> hist(preddist, xlab=expression(t(tilde(y))),
                 main="",prob=T)
> mean(preddist>=2)
[1] 0.0056
```


Density
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
2.5
3.0
3.5
t(ỹ)