STAT 578: Advanced Bayesian Modeling

Week 2 – Lesson 1

Mean-Only Normal Sample

Fall 2019

Posterior Prediction

The posterior lets us estimate the parameter μ, the population mean.

What if we want some other population feature?

In the Flint example, we might want the percentage of households whose lead measurement exceeds the action level of 15 ppb.

This is the same as the probability that a measurement from a hypothetical newly sampled household exceeds the level.

Predicting a New Observation

Let

$$\tilde{y} \;=\; \text{a newly sampled value from the sampling distribution}$$

(e.g., the log lead level measured in a newly sampled Flint household).

According to our model,

$$\tilde{y} \mid \mu \;\sim\; \mathrm{N}(\mu, \sigma^2)$$

where we continue to regard σ^2 as constant and known.

But μ is unknown, so this is not directly useful.

We need the *posterior predictive* distribution: the distribution of $\tilde{y}$ given data y.

Recall: The posterior predictive density is obtained as

$$p(\tilde{y} \mid y) \;=\; \int p(\tilde{y} \mid \mu, y)\, p(\mu \mid y)\, d\mu$$

The first density in the integrand comes from

$$\tilde{y} \mid \mu, y \;\sim\; \mathrm{N}(\mu, \sigma^2)$$

since $\tilde{y}$ comes from the sampling distribution, but is independent of the data.

The second density is the posterior. For a normal prior, we have

$$\mu \mid y \;\sim\; \mathrm{N}(\mu_n, \tau_n^2)$$

To avoid explicit integration, note that

$$\tilde{y} - \mu \mid \mu, y \;\; \sim \;\; \mathrm{N}(0, \sigma^2)$$

so, conditional on y, $\tilde{y} - \mu$ is normally distributed and independent of μ.

Since

$$\tilde{y} \;\; = \;\; (\tilde{y} - \mu) \; + \; \mu$$

and the sum of two independent normally distributed random variables has a normal distribution,

$$\tilde{y} \mid y \;\; \sim \;\; \mathrm{N}(\mu_n, \sigma^2 + \tau_n^2)$$

Example: Flint Data

The proper prior we considered led to values

$$\mu_n \approx 1.40 \qquad \tau_n^2 \approx 0.0062$$

Combined with the assumption that

$$\sigma^2 = s^2 \approx 1.684$$

we compute $\Pr\big(\tilde{y} > \log(15) \mid y\big)$ to be

```
> pnorm(log(15), mun, sqrt(sigma.2 + tau.2.n), lower.tail=FALSE)
[1] 0.1575152
```

Instead using the flat prior leads to

$$\tilde{y} \mid y \;\sim\; \mathrm{N}(\bar{y}, \sigma^2 + \sigma^2/n)$$

for which $\Pr\big(\tilde{y} > \log(15) \mid y\big)$ becomes

```
> pnorm(log(15), ybar, sqrt(sigma.2 + sigma.2/n), lower.tail=FALSE)
[1] 0.1577242
```

So about 16% of households would have a measurement exceeding the action level. (By law, should be no more than 10%.)

Compare with fraction in the original sample exceeding the level:

```
> mean(Flintdata$FirstDraw > 15)
[1] 0.1660517
```