ADVANCED BAYESIAN MODELING

BAYESIAN TOOLS FOR INFERENCE:
TOOLS FOR PREDICTION

Posterior Prediction

Suppose you are considering offering your dog-walking service in a new neighborhood of 100 households.

If at least 15 households in this neighborhood are interested, it is worth offering your service there.

What is the posterior probability of this?

You expect that, in this neighborhood, the proportion of households interested in your service is the same as in the community you surveyed.

Letting

$$\tilde{y} \quad = \quad \text{number of interested households in neighborhood}$$

we suppose

$$\tilde{y} \mid \theta \quad \sim \quad \text{Bin}(n = 100, \theta)$$

(This is true provided households are interested independently of each other.)

The distribution of $\tilde{y}$ given the data y is the **posterior predictive distribution**, having density

$$\begin{aligned} p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y)\, d\theta \\ &= \int p(\tilde{y} \mid \theta, y)\, p(\theta \mid y)\, d\theta \\ &= \int p(\tilde{y} \mid \theta)\, p(\theta \mid y)\, d\theta \end{aligned}$$

The last line assumes $\tilde{y}$ is conditionally independent of y given θ, which is true because your survey did not sample this neighborhood.

While an analytical solution for $p(\tilde{y} \mid y)$ is possible in this simple case, it is easier to approximate the desired probability

$$\Pr(\tilde{y} \geq 15 \mid y) \quad = \quad \sum_{\tilde{y}=15}^{100} p(\tilde{y} \mid y)$$

using simulation:

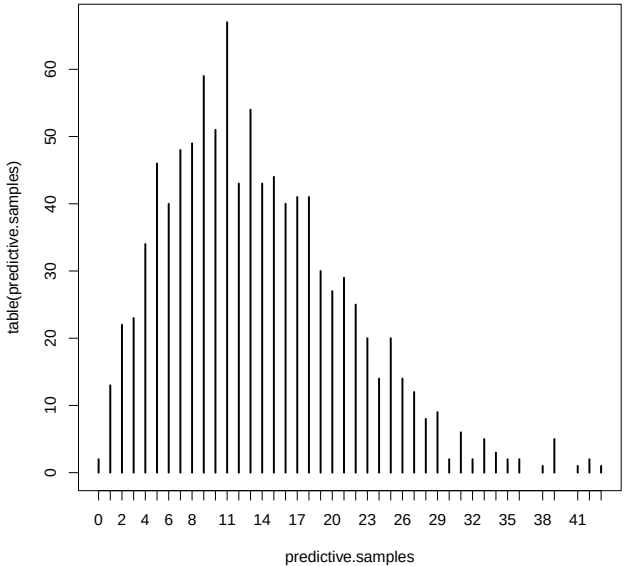
```
> predictive.samples <- rbinom(1000, 100, posterior.samples)

> mean(predictive.samples >= 15)
[1] 0.406
```

So your service has (at least) about a 40% chance of success in this neighborhood.

A simulation-based approximation to the posterior predictive density of $\tilde{y}$:

```
> plot(table(predictive.samples))
```

(Note: This code works only because $\tilde{y}$ is discrete.)


table(predictive.samples)
0
10
20
30
40
50
60
0 2 4 6 8 11 14 17 20 23 26 29 32 35 38 41
predictive.samples

Based on a posterior predictive distribution, a predicted value for $\tilde{y}$ could be

$$\mathrm{E}(\tilde{y} \mid y)$$

and the standard deviation could quantify its uncertainty:

```
> mean(predictive.samples)
[1] 13.651

> sd(predictive.samples)
[1] 7.683504
```

There are also **posterior predictive intervals**, defined similarly to posterior intervals, but for $\tilde{y}$, not θ:

```
> quantile(predictive.samples, c(0.05/2, 1-0.05/2))
 2.5% 97.5%
    2    31
```

The interval $[2, 31]$ is probably a bit conservative, i.e., has posterior predictive probability exceeding 95%, because $\tilde{y}$ is discrete.

Ignoring Data

There may also be **prior probabilities**, **prior expected values**, and **prior intervals**, based on the prior distribution.

Similarly, there may be a **prior predictive distribution** for $\tilde{y}$, based on the prior and the sampling distribution, but ignoring the data.

Though of limited use for inference, these can help choose a prior and possibly assess the Bayesian model.