ADVANCED BAYESIAN MODELING

Introduction by Example

Flint Water Crisis

The city of Flint, Michigan, changed to a new municipal water source in 2014.

Suspicions arose of high levels of contaminants, including lead.

Officials claimed levels were within regulatory limits, but concerned citizens and scientists initiated their own testing.

Full story: Langkjær-Bain, R. (2017, April). The murky tale of Flint's deceptive water data. *Significance*, *14*(2), 16–21.

Federal Lead and Copper Rule of 1991: Action required if over 10% of homes sampled have lead levels over 15 parts per billion (15 ppb)

Citizen data set (from `http://flintwaterstudy.org`):

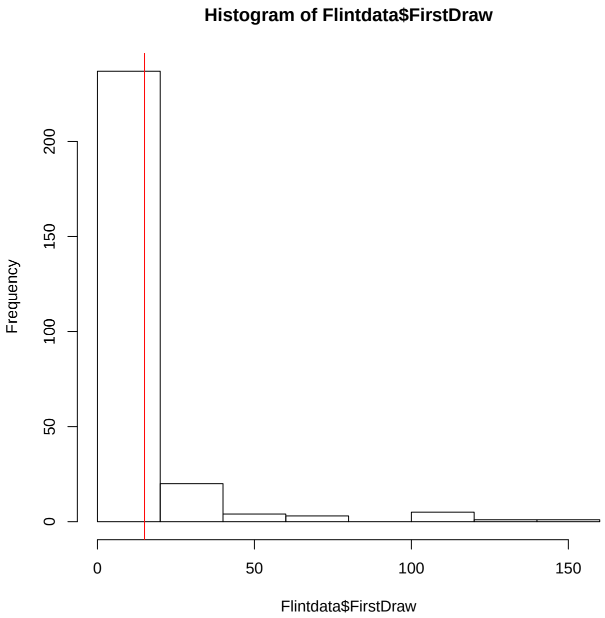
- 271 observations (sampling kits returned)
- Three lead readings (ppb) for each observation:
    - Initial draw
    - After flushing 45 seconds
    - After flushing 2 minutes
- Other variables: house ID number, zip code, ward

Examination of data set in R:

```
> Flintdata <- read.csv("Flintdata.csv", header=TRUE, row.names=1)

> head(Flintdata)
  SampleID ZipCode Ward FirstDraw After45Sec After2Min Notes
1        1   48504    6     0.344      0.226     0.145
2        2   48507    9     8.133     10.770     2.761
3        4   48504    1     1.111      0.110     0.123
4        5   48507    8     8.007      7.446     3.384
5        6   48505    3     1.951      0.048     0.035
6        7   48507    9     7.200      1.400     0.200

> hist(Flintdata$FirstDraw)
> abline(v=15, col="red")
```

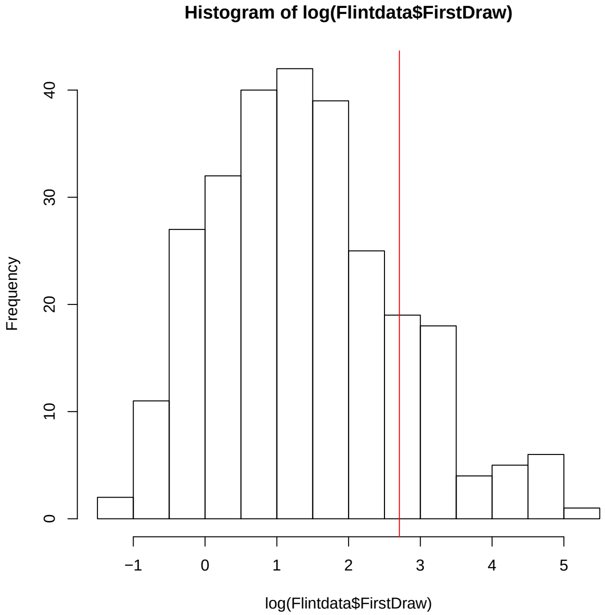

Histogram of Flintdata$FirstDraw
Frequency
0
50
100
150
200
0
50
100
150
Flintdata$FirstDraw

High right skew is apparent – try log scale:

```
> summary(Flintdata$FirstDraw)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.344   1.578   3.521  10.646   9.050 158.000

> summary(log(Flintdata$FirstDraw))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.0671  0.4561  1.2587  1.4029  2.2024  5.0626

> hist(log(Flintdata$FirstDraw))
> abline(v=log(15), col="red")
```


Histogram of log(Flintdata$FirstDraw)
Frequency
0
10
20
30
40
−1
0
1
2
3
4
5
log(Flintdata$FirstDraw)

Much less skew on log scale

Some structure may remain (clustering?) – ignore it for now.

Normal Sampling Distribution

Consider variable y having a normal distribution:

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

$$\mu \quad = \quad \text{mean} \qquad\qquad \sigma^2 \quad = \quad \text{variance}$$

Its density:

$$p(y \mid \mu, \sigma^2) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \qquad\qquad -\infty < y < \infty$$

Typically there are several observations of a variable:

$$y = (y_1, \ldots, y_n)$$

In a normal sample, these observations are (conditionally) independent and from the same normal distribution:

$$y_1, \ldots, y_n \mid \mu, \sigma^2 \sim \text{iid N}(\mu, \sigma^2)$$

$$\text{iid} = \text{independent and identically distributed}$$

$$-\infty < \mu < \infty \qquad \sigma^2 > 0$$

The normal sampling (joint) density:

$$\begin{aligned} p(y \mid \mu, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \mu, \sigma^2) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \\ &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right) \end{aligned}$$

As a function of μ and σ^2, this is the likelihood.

The usual sample mean and sample variance ($n > 1$):

$$\bar{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \qquad\qquad s^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Notice:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \mu)^2 &= \sum_{i=1}^{n}(y_i - \bar{y} + \bar{y} - \mu)^2 \\
&= \sum_{i=1}^{n}(y_i - \bar{y})^2 \;+\; \sum_{i=1}^{n}(\bar{y} - \mu)^2 \\
&= (n-1)s^2 \;+\; n(\mu - \bar{y})^2
\end{aligned}$$

Then the likelihood is

$$
\begin{aligned}
p(y \mid \mu, \sigma^2) &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \left((n-1)s^2 \;+\; n(\mu - \bar{y})^2 \right) \right) \\
&\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2} \right) \exp\left(-\frac{n}{2\sigma^2} (\mu - \bar{y})^2 \right)
\end{aligned}
$$

Note: Only the last factor depends on μ.

Mean-Only Model

For simplicity, begin by <u>assuming σ^2 is a known constant</u>.

Then the only unknown parameter is

$$\theta \;=\; \mu$$

for which the likelihood is

$$p(y \mid \mu) \;\;\propto\;\; \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2\right)$$

(dropping all factors not involving μ).

Note: The proportionality is only in μ (not y or σ^2).