Shoe Sizes Example

Sean van der Merwe

08/03/2022

Don’t copy this - it will be posted on Blackboard after the class.

Introduction

For this example we will start with a simple sample of shoe sizes and try to build a predictive model of shoe sizes for people similar to the ones sampled.

```
"Shoes.xlsx" |> openxlsx::read.xlsx() -> ShoesData
y <- ShoesData$ShoeSize
x <- ShoesData$Gender
table(y)

## y
##    3    4    5    6    7    8   10 10.5   13
##    1    1    1    4    4    1    1    1    1
```

Model

Shoe sizes only come in half sizes, so if we double the sizes we end up with positive whole numbers. So we could use a model for whole numbers, like the Negative Binomial or Discrete Weibull for the double sizes. If we are interested in predicting shoe sales then this model would make a lot of sense.

Alternatively, we could think about the data generating process and realise that shoes go on feet and feet sizes are positive real numbers. People might wear shoes in the closest size to their feet. We could consider the observations as rounded real numbers, or censored, or measured with errors of a certain type.

For today’s example we will use a LogNormal distribution.

Prior

$$\mu \sim N(1.95, \sigma = 0.15)\ ,\ \sigma \sim Gamma(0.001,0.001)$$

Stan model

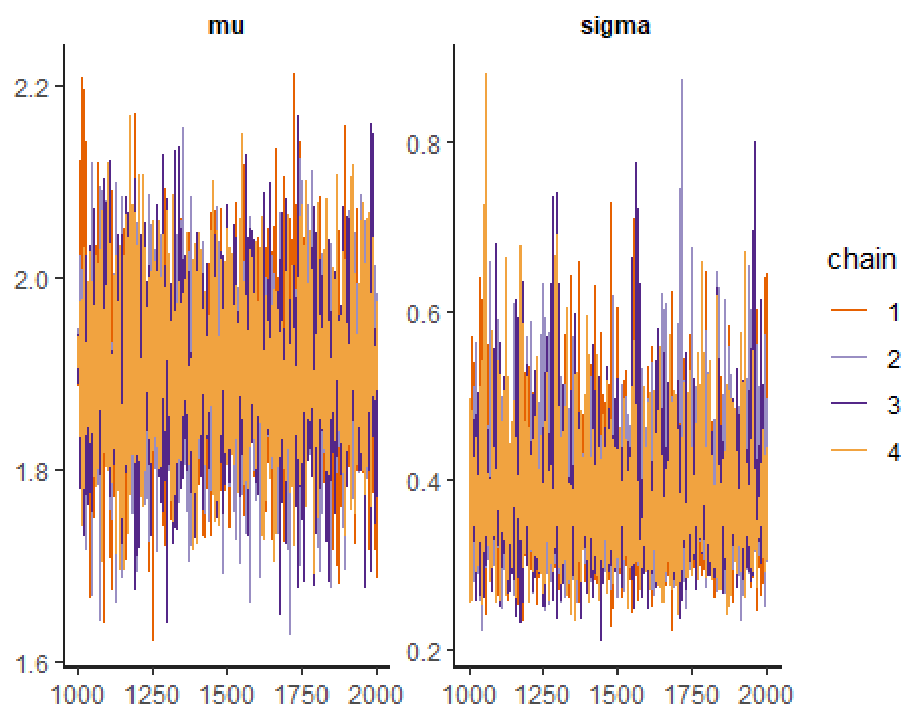
```
data {
  int n;
  real y[n];
}
parameters {
  real mu;
```

```
  real<lower=0> sigma;
}
model {
  y ~ normal(mu, sigma);
  mu ~ normal(1.95, 0.15);
  sigma ~ gamma(0.001, 0.001);
}
```

Run Stan

```
logSizes <- log(ShoesData$ShoeSize)
library(rstan)
shoes1 |> sampling(data = list(y = logSizes, n = length(logSizes))) -> output

output |> traceplot()
```

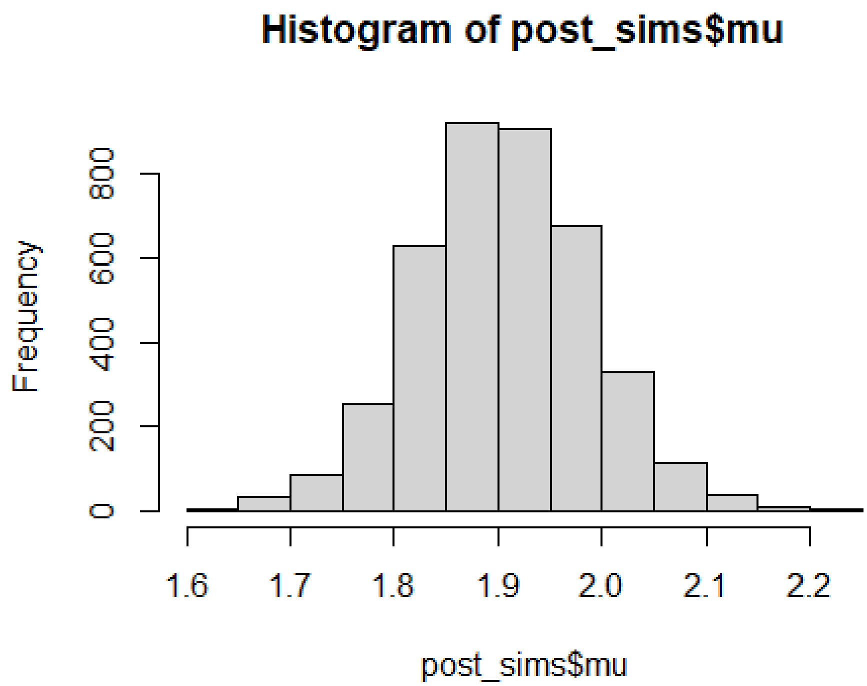



```
output |> extract() -> post_sims
(output |> summary())$summary |> knitr::kable(digits = 3)
```

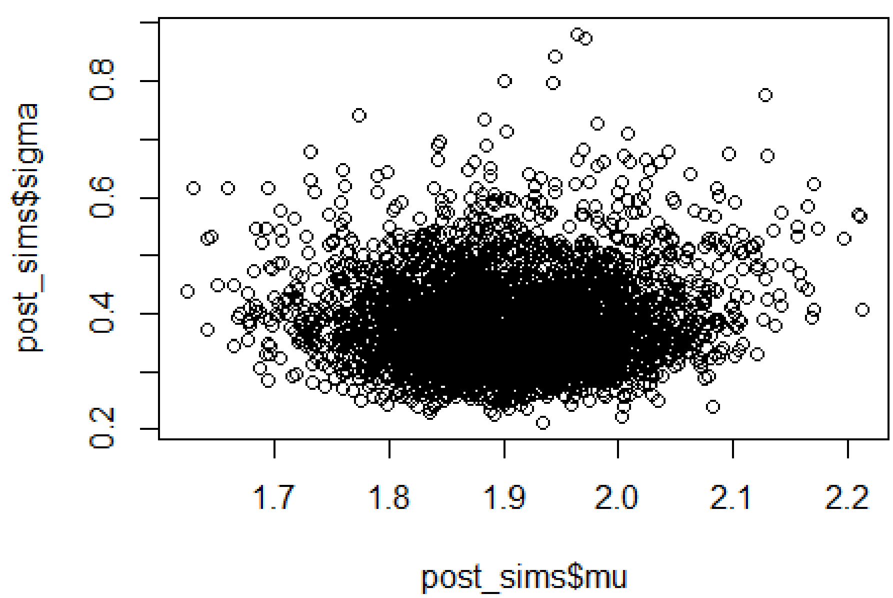
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu	1.905	0.002	0.083	1.739	1.850	1.904	1.961	2.071	2392.455	1.000
sigma	0.383	0.002	0.078	0.267	0.329	0.371	0.423	0.573	2132.323	1.001
lp__	6.960	0.027	1.045	4.052	6.552	7.283	7.703	7.970	1527.975	1.002

Posterior

```
post_sims$mu |> hist()
```




```
plot(post_sims$mu, post_sims$sigma)
```



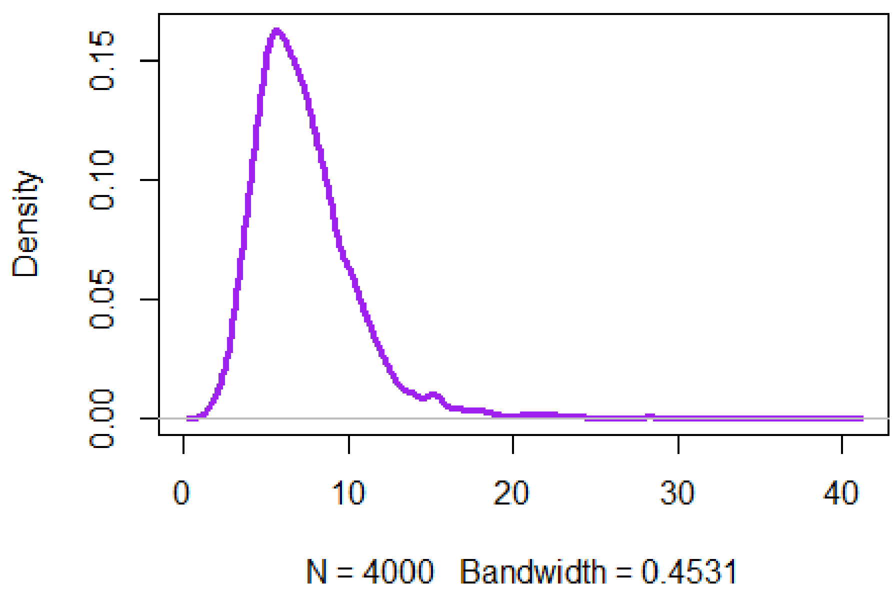

```
(mu_hat <- mean(post_sims$mu))
```

```
## [1] 1.905407
```

Predictive

```
nsims <- length(post_sims$mu)
pred_sims <- exp(rnorm(nsims, post_sims$mu, post_sims$sigma))

pred_sims |> density() |> plot(main ='', lwd=3, col='purple')
```



Probability of a shoe size about 15 given our sample?

```
mean(pred_sims > 15)

## [1] 0.02725
```