Homework 1

Olivia Hackworth

2/6/2019

Normal Distribution with unknown mean

A random sample of n students is drawn from a large population and their weights are measured. The average weight of the n sampled students is $\bar{y} = 71$ kilograms. Assume that the weights in the population are normally distributed with unknown mean θ and known standard deviation 10 kilograms. Suppose your prior distribution for θ is normal with mean 60 and standard deviation 20. (1 kilogram is about 2.20462 pounds)

1. Give your posterior distribution for θ, as a function of n.
2. A new student is sampled at random from teh same population and has a weight of $\tilde{y}$ kilograms. Give the posterior predictive distribution for $\tilde{y}$, as a function of n.
3. For $n = 9$, give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.
4. Do the same for $n = 99$.

1.1

$\bar{y} = 71$
$y|\theta \sim N(\theta, \sigma^2), \sigma^2 = 100kg$
$\theta \sim N(\mu, \tau^2), \tau^2 = 400kg, \mu = 60kg$

$p(\theta|y) \propto p(y|\theta)p(\theta)$

$\propto exp\{\frac{-1}{2\sigma^2}\sum(y_i - \theta)^2\} * exp\{\frac{-1}{2\tau^2}(\theta - \mu)^2\}$
$= exp\{\frac{-1}{2}[\frac{\theta^2}{\tau^2} + \frac{n\theta^2}{\sigma^2} - \frac{2\theta\mu}{\tau^2} - \frac{2\theta\sum y_i}{\sigma^2} + \frac{\sum y_i^2}{\sigma^2} + \frac{\mu^2}{\tau^2}]\}$
$= exp\{\frac{-1}{2}[\theta^2(\frac{1}{\tau^2} + \frac{n}{\sigma^2}) + \theta(\frac{-2\mu}{\tau^2} - \frac{2\sum y_i}{\sigma^2}) + \frac{\sum y_i^2}{\sigma^2} + \frac{\mu^2}{\tau^2}]\}$
Definition: $a = (\frac{1}{\tau^2} + \frac{n}{\sigma^2})$
$b = (\frac{\mu}{\tau^2} + \frac{\sum y_i}{\sigma^2})$
$c = (\frac{\mu^2}{\tau^2} + \frac{\sum y_i^2}{\sigma^2})$
Continuing from above: $= exp\{\frac{-1}{2}[a\theta^2 - 2b\theta + c]\}$
$\propto exp\{\frac{-1}{2}[a\theta^2 - 2b\theta]\}$
$= exp\{\frac{-a}{2}[\theta^2 - 2\frac{b}{a}\theta]\}$
$= exp\{\frac{-a}{2}[\theta^2 - 2\frac{b}{a}\theta + \frac{b^2}{a^2} - \frac{b^2}{a^2}]\}$
$\propto exp\{\frac{-a}{2}[\theta^2 - 2\frac{b}{a}\theta + \frac{b^2}{a^2}]\}$
$= exp\{\frac{-a}{2}[(\theta - \frac{b}{a})^2]\}$

So $\mu_n = \frac{b}{a}$ and $\tau_n^2 = \frac{1}{a}$.

After substituing a and b back in: $\mu_n = \frac{\sigma^2\mu + \tau^2 n\bar{y}}{\sigma^2 + n\tau^2}$ and $\tau_n^2 = [\frac{1}{\tau^2} + \frac{n}{\sigma^2}]^{-1}$

After substituting the known values in, the posterior is:

$\theta|y \sim N(\frac{60+284n}{1+4n}, \frac{100}{1/4+n})$

1.2

As derived in class on 1/28:

$\tilde{y}|y_1...y_n \sim N(\mu_n, \tau_n^2 + \sigma^2)$.

With the information we know, $\tilde{y}|y_1...y_n \sim N(\frac{60+284n}{1+4n}, \frac{100}{1/4+n} + 100)$.

1.3

n = 9

$\theta|y \sim N(70.7027, 10.8108)$

$P(-1.96 < \frac{\theta-70.7027}{\sqrt{10.8108}} < 1.96) = 0.95$
$P(-1.96\sqrt{10.8108} + 70.7027 < \theta < 1.96\sqrt{10.8108} + 70.7027) = 0.95$
$P(64.2583 < \theta < 77.1471) = 0.95$

So the posterior interval is (64.2583,77.1471).

$\tilde{y}|y \sim N(70.7027, 110.8108)$

$P(-1.96 < \frac{\tilde{y}-70.7027}{\sqrt{110.8108}} < 1.96) = 0.95$
$P(-1.96\sqrt{110.8108} + 70.7027 < \tilde{y} < 1.96\sqrt{110.8108} + 70.7027) = 0.95$
$P(50.0704 < \tilde{y} < 91.3350) = 0.95$

So the posterior predictive interval is (50.0704, 91.3350).

1.4

n=99

$\theta|y \sim N(70.9723, 1.0076)$

$P(-1.96 < \frac{\theta-70.9723}{\sqrt{1.0076}} < 1.96) = 0.95$
$P(-1.96\sqrt{1.0076} + 70.9723 < \theta < 1.96\sqrt{1.0076} + 70.9723) = 0.95$
$P(69.0049 < \theta < 72.9397) = 0.95$

So the posterior interval is (69.0049,72.9397).

$\tilde{y}|y \sim N(70.9723, 101.0076)$

$P(-1.96 < \frac{\tilde{y}-70.9723}{\sqrt{101.0076}} < 1.96) = 0.95$
$P(-1.96\sqrt{101.0076} + 70.9723 < \tilde{y} < 1.96\sqrt{101.0076} + 70.9723) = 0.95$
$P(51.2738 < \tilde{y} < 90.6708) = 0.95$

So the posterior predictive interval is (51.2738, 90.6708).

The intervals are narrower in both cases when n=99. This makes sense because n appears in the denominator of the variance calculation, so the variance and thus the width of the interval decreases as the sample size increases.

Discrete Sample Spaces

Suppose there are N cable cars in San Francisco, numbered sequentially from 1 to N. You see a cable car at random; it is numbered 201. You wish to estimate N.

1. Assume your prior distribution on N is geometric with mean 100, ie.

$$p(N) = 0.01 * (0.99)^{N-1}, N = 1, 2, ...$$

What is your posterior distribution for N?

2. What are the posterior mean and standard deviation of N? (Sum the infinite series analytically or approximate them on the computer.)

2.1

```
unnormalized_posterior = function(N){
  p=(1/N)*(.01)*(.99)^(N-1)
  return(p)
}

grid = 201:1e6
un_post_vec = sapply(grid, function(x) unnormalized_posterior(x))

cons = sum(un_post_vec)
cons
```

```
## [1] 0.0004837404
```

The prior for N is $p(N) = (.01)(.99)^{N-1}$
The likelihood is $p(y|N) = \frac{1}{N}$
The normalized posterior is $p(N|y) = \frac{1}{0.0004837} \frac{1}{N} (.01)(.99)^{N-1}$

2.2

```
postmean = function(N){
  m=N*(1/cons)*(1/N)*(.01)*(.99)^(N-1)
  return(m)
}

posteriormean = sum(postmean(201:1e6))
posteriormean
```

```
## [1] 276.9661
```

The posterior mean is 277.

```
postvar = function(N){
  m=(N-posteriormean)^2*(1/cons)*(1/N)*(.01)*(.99)^(N-1)
  return(m)
}

posteriorvar = sum(postvar(201:1e6))
#posteriorvar

posteriorsd = sqrt(posteriorvar)
posteriorsd
```

```
## [1] 79.87247
```

The posterior standard deviation is 80.

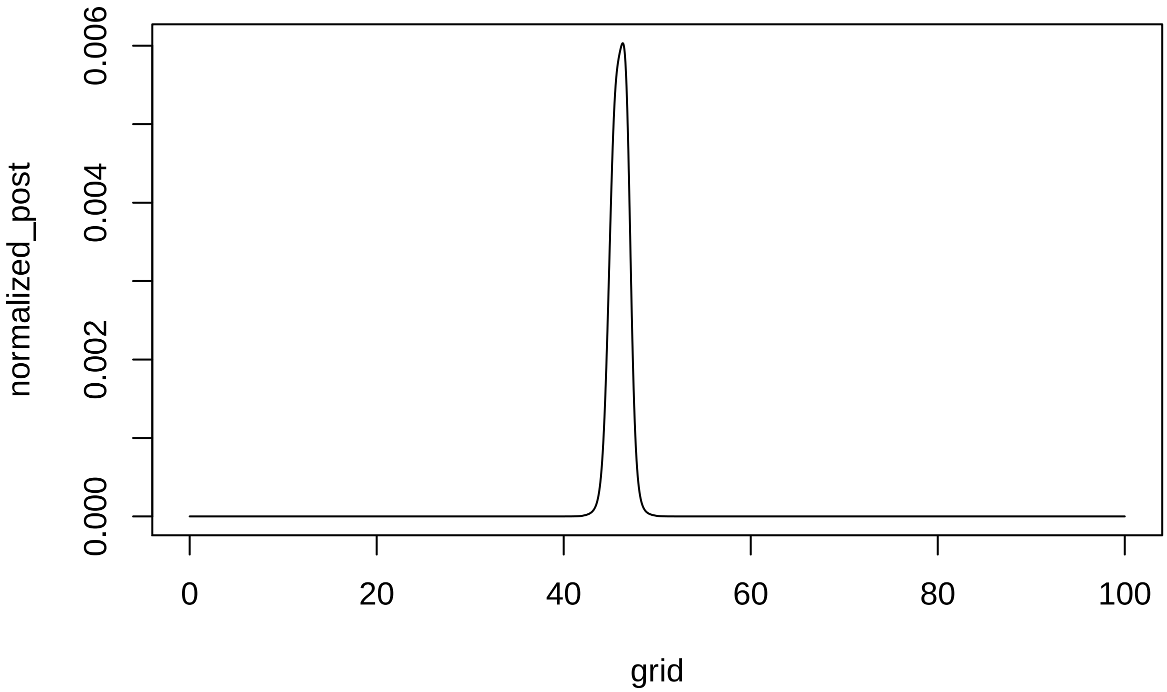
Nonconjugate single parameter model

Suppose $y_1, ..., y_6$ are independent samples from a Cauchy distribution with unknown center θ and known scale 1: $p(y_i|\theta) \propto 1/(1+(y_i-\theta)^2)$. Assume that the prior distribution for θ is uniform on [0,100]. Given the observations $(y_1, ..., y_6) = (42, 44.5, 45.3, 46.8, 47.2, 50)$:

1. Compute the unnormalized posterior density function, $p(\theta)p(y|\theta)$, on a grid of points $\theta = 0, \frac{1}{m}, \frac{2}{m}, ..., 100$, for some large integer m. Using the grid approximation, compute and plot the normalized posterior density function $p(\theta|y)$, as a function of θ.
2. Sample 2000 draws of θ from the posterior density and plot a histogram of the draws.
3. Use the samples of θ to obtain 3000 samples from the predictive distribution of a future observation y_7, and plot a histogram of the predictive draws.

3.1

```r
#Known
scale = 1
y=c(42,44.5,45.3,46.8,47.2,50)

#grid to search over
m = 67
grid = seq(0,m*100,by=1)/m

#posterior distribution - unnormalized
post = function(theta){
  p=exp(-log(100)+sum(dcauchy(y,theta,scale=scale,log=TRUE)))
  return(p)
}

unnormalized_post = sapply(grid,function(x) post(x))
c = sum(unnormalized_post)

#posterior distribution - normalized
normalized_post = (1/c)*unnormalized_post

plot(grid,normalized_post,type="l")
```

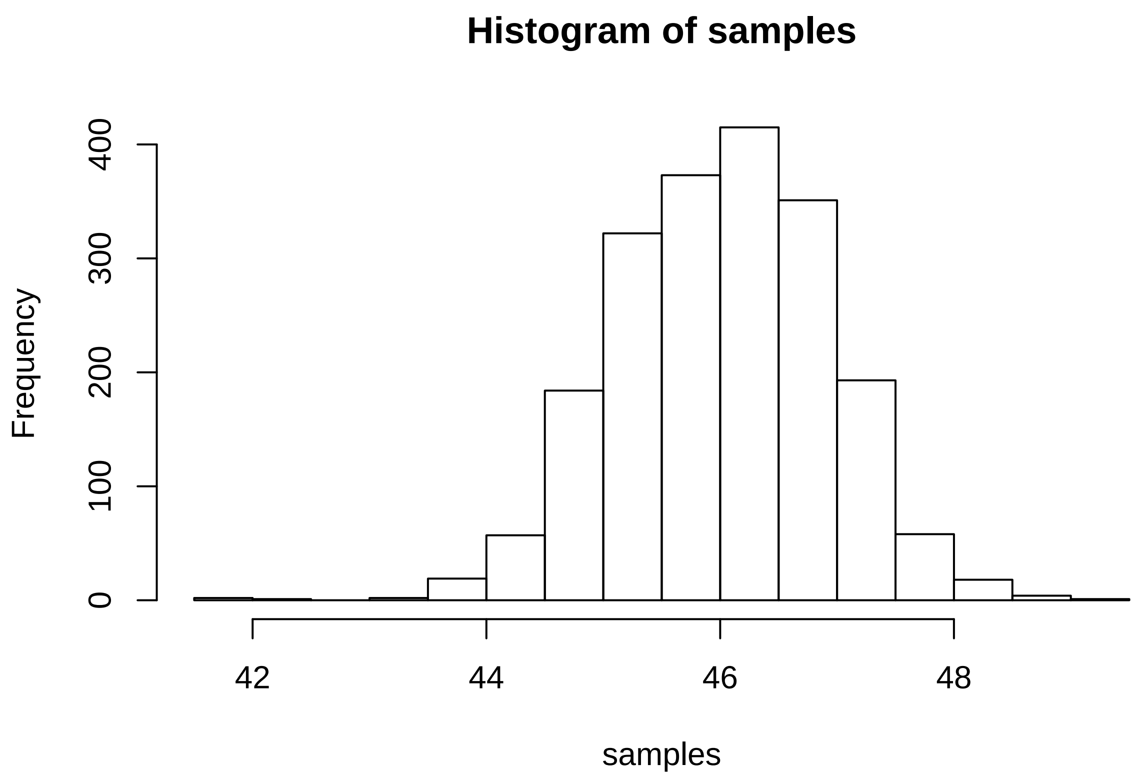


The normalized posterior has a sharp peak around 46 which is the mean of the observed values.

Note: The normalization doesn't take into account the grid size like it should. This mistake is accounted for in future assignments.

3.2

```
samples = sample(grid,size = 2000, replace = T,prob=unnormalized_post)
hist(samples)
```

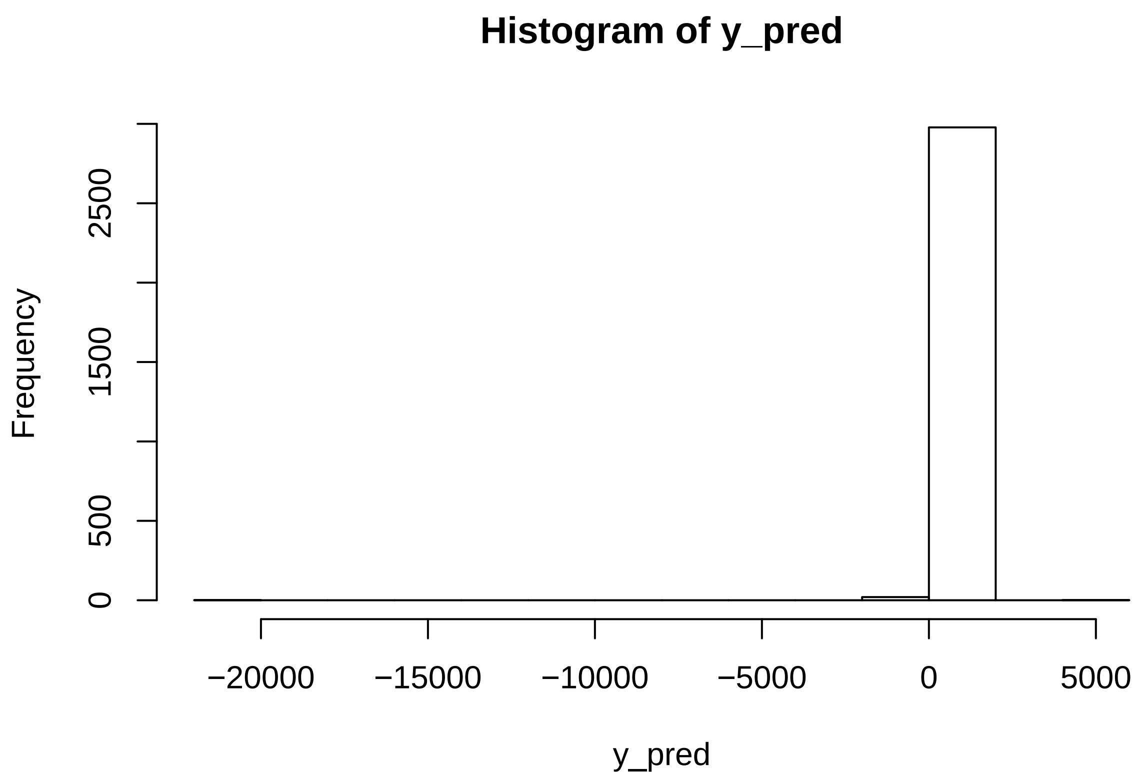



```
y_pred = rcauchy(length(samples),location=samples)
#hist(y_pred)
```

The histogram of samples from the unnormalized posterior is centered around 46. This makes sense based on the data we observed. The mean of the y vector is also very close to 46, so the data is influencing the center of the posterior.

3.3

```
sample2 = sample(grid,size = 3000,replace = T,prob=unnormalized_post)
#hist(samples)
y_pred = rcauchy(length(sample2),location=sample2)
hist(y_pred)
```



This histogram is centered at about 45, but it is very spread out. This makes sense because there is a strong signal around 46 where the next observation is likely to happen, but it is very spread out because the observation comes from a cauchy distribution.