SYST 664: Assignment 6

Jake Whalen

March 20, 2018

Table of Contents

[Problem 1 1](#_Toc510421579)

[Bottom Concentrations 3](#_Toc510421580)

[Surface Concentrations 3](#_Toc510421581)

[Comparing Concetrations by Depth 4](#_Toc510421582)

[Problem 2 6](#_Toc510421583)

[Problem 3 9](#_Toc510421584)

[Problem 4 10](#_Toc510421585)

[What are your conclusions about the distributions of surface and bottom concentrations? 10](#_Toc510421586)

[Is the assumption of normality reasonable? 11](#_Toc510421587)

[Are the means different for surface and bottom? 13](#_Toc510421588)

[The standard deviations? 14](#_Toc510421589)

# Problem 1

Concentrations of the pollutants aldrin and hexachlorobenzene (HCB) in nanograms per liter were measured in ten surface water samples, ten mid-depth water samples, and ten bottom samples from the Wolf River in Tennessee. The samples were taken downstream from an abandoned dump site previously used by the pesticide industry. The full data set can be found at <http://www.biostat.umn.edu/~lynn/iid/wolf.river.dat>. The code below reads this data set into R.

#URL of data given in assignment  
www <- "http://www.biostat.umn.edu/~lynn/iid/wolf.river.dat"  
  
#Read in the data  
pollutants <- read.table(www, header=TRUE, stringsAsFactors = F)

For this problem, we consider only HCB measurements taken at the bottom and the surface. The question of interest is whether the distribution of HCB concentration depends on the depth at which the measurement was taken. The data for this problem are given below.

pollutants %<>%  
 mutate(id = rep(seq(1, 10, by = 1), 3)) %>%  
 select(-Aldrin) %>%  
 spread(Depth, HCB) %>%  
 select(-id, -Middepth)  
knitr::kable(pollutants)

|  |  |
| --- | --- |
| Bottom | Surface |
| 5.44 | 3.74 |
| 6.88 | 4.61 |
| 5.37 | 4.00 |
| 5.44 | 4.67 |
| 5.03 | 4.87 |
| 6.48 | 5.12 |
| 3.89 | 4.52 |
| 5.85 | 5.29 |
| 6.85 | 5.74 |
| 7.16 | 5.48 |

Assume the observations are independent normal random variables with unknown depth-specific means and and precisions and . Assume independent improper reference priors for the surface and bottom parameters:

This prior can be treated as a normal-gamma prior with = 0, = 0, = -1/2, and = . (This is not a valid normal-gamma distribution, but you can do the usual Bayesian conjugate updating to find the posterior distribution.) Find the joint posterior distribution for the parameters (, , , ).

First I define variables for the prior distributions parameters.

#mu  
prior.center <- 0  
#k  
prior.precision <- 0  
#alpha  
prior.shape <- -1/2  
#beta  
prior.scale <- 'infinity'

Another variable I will define now since it does not change between the depth groupings is the number of observations. The table above shows that at each depth there are 10 new observations.

#Number of Observations  
n <- nrow(pollutants)

### Bottom Concentrations

There are two additional values to go along with the number of observations (sample mean, and Sum of Squared Deviations). All of these values get used next to update the distribution parameters.

#Sample Mean  
bottom.mean <- sum(pollutants$Bottom) / n  
bottom.mean

## [1] 5.839

#Sum of Squared Deviations  
bottom.ssd <- sum((pollutants$Bottom - bottom.mean) ^ 2)  
bottom.ssd

## [1] 9.25329

Now using the values I just calculated I will update the posterior distribution parameters for the measurements recorded near the bottom.

#Center  
bottom.center <- ((prior.precision \* prior.center) + (n \* bottom.mean)) / (prior.precision + n)  
#Precision  
bottom.precision <- prior.precision + n  
#shape  
bottom.shape <- prior.shape + 1/2 \* n  
#scale  
bottom.scale <- (1/2 \* bottom.ssd + (prior.precision \* n \* (bottom.mean - prior.center) ^ 2) / (2 \* (prior.precision + n))) ^ -1  
#Spread  
bottom.spread <- 1 / sqrt(bottom.precision \* bottom.scale \* bottom.shape)  
#Degrees of Freedom  
bottom.dof <- 2 \* bottom.shape

I then use the dt function in R to calculate the density over a range for .

bottom.density <-  
 data.frame('Mean' = seq(-5, 5, length.out = 200)) %>%  
 mutate('Density' = dt(Mean, df = bottom.dof)/bottom.spread,  
 Depth = 'Bottom',  
 Mean = Mean + bottom.center)

### Surface Concentrations

I follow the same steps for the measurements taken near the surface. First I calculate the helpful metrics from the data.

#Sample Mean  
surface.mean <- sum(pollutants$Surface) / n  
surface.mean

## [1] 4.804

#Sum of Squared Deviations  
surface.ssd <- sum((pollutants$Surface - surface.mean) ^ 2)  
surface.ssd

## [1] 3.58824

Then I update the posterior parameters.

#Center  
surface.center <- ((prior.precision \* prior.center) + (n \* surface.mean)) / (prior.precision + n)  
#Precision  
surface.precision <- prior.precision + n  
#shape  
surface.shape <- prior.shape + 1/2 \* n  
#scale  
surface.scale <- (1/2 \* surface.ssd + (prior.precision \* n \* (surface.mean - prior.center) ^ 2) / (2 \* (prior.precision + n))) ^ -1  
#Spread  
surface.spread <- 1 / sqrt(surface.precision \* surface.scale \* surface.shape)  
#Degrees of Freedom  
surface.dof <- 2 \* surface.shape

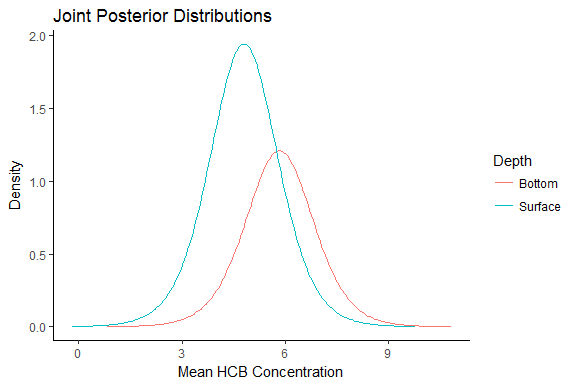
Last I calculate the density and store in a data.frame in R.

surface.density <-  
 data.frame('Mean' = seq(-5, 5, length.out = 200)) %>%  
 mutate('Density' = dt(Mean, df = surface.dof)/surface.spread,  
 Depth = 'Surface',  
 Mean = Mean + surface.center)

### Comparing Concetrations by Depth

Putting both densities together I can then visualize the joint posterior as the individual posteriors for the *bottom* and *surface* depth measurements.

bind\_rows(surface.density, bottom.density) %>%  
 ggplot(aes(x = Mean,  
 y = Density,  
 color = Depth)) +  
 geom\_line(stat = 'identity') +  
 labs(title = 'Joint Posterior Distributions',  
 x = 'Mean HCB Concentration') +  
 theme\_classic()



The next task for the assignment is to find 90% posterior credible intervals (CI) for , , , and . I can accomplish this by doing it in two steps for each depth level. The first step will find the CI for and the second for .

I begin with the bottom depth concentrations. First I find the 0.95 quantile of the t distribution with 9 degrees of freedom.

bottom.t.95 <- qt(0.95, bottom.dof)

Then I use this quantile with the center and spread parameters to calculate the 95% CI for .

bottom.theta.05 <- bottom.center - bottom.t.95 \* bottom.spread  
bottom.theta.95 <- bottom.center + bottom.t.95 \* bottom.spread

The second step solves for the 0.05 and 0.95 quantiles of the gamma(4.5, 0.2161393) distribution. This provides the 90% CI for the parameter.

bottom.P.05 <- qgamma(.05, shape = bottom.shape, scale = bottom.scale)  
bottom.P.95 <- qgamma(.95, shape = bottom.shape, scale = bottom.scale)

I repeat these two steps again for the surface level concentration measurements.

#theta  
surface.t.95 <- qt(0.95, surface.dof)  
surface.theta.05 <- surface.center - surface.t.95 \* surface.spread  
surface.theta.95 <- surface.center + surface.t.95 \* surface.spread  
#P  
surface.P.05 <- qgamma(.05, shape = surface.shape, scale = surface.scale)  
surface.P.95 <- qgamma(.95, shape = surface.shape, scale = surface.scale)

Resulting posterior credible intervals.

|  |  |  |
| --- | --- | --- |
| Parameter | Surface | Bottom |
| Theta | [4.44, 5.17] | [5.25, 6.43] |
| P | [0.93, 4.72] | [0.36, 1.83] |

The 90% CI’s for the parameters do not overlap. The HCB concentration looks like it may be lower at the surface compared to the concentrations levels at the bottom. This is seen in the seperation between the distributions in the graph. It is also obvious when looking at the CI values in the table. The graph helps show the differences in the spread between the two distributions. The distribution specific to the surface concentrations is much tighter and less spread out then the distribution for the bottom concentrations. There may be a difference between the two depths concntrations of the HCB pollutant but more analysis is required in order to support that claim.

# Problem 2

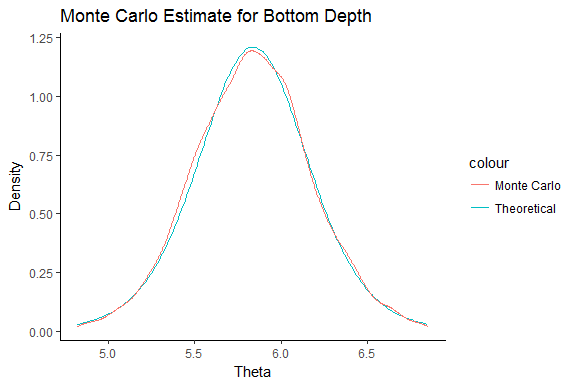
Use direct Monte Carlo to sample 10,000 observations from the joint posterior distribution of (, , , ).

The code below is a function I wrote that takes the code supplied in the slides and makes the monte carlo process reproducible.

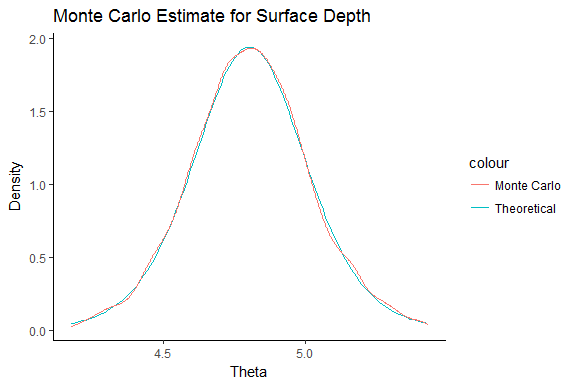
p2\_mc <- function(depth){  
 x <- pollutants[,depth]  
 xbar <- mean(x)  
 n <- length(x)  
 # Assume non-informative prior distribution  
 # Normal-Gamma with mu0=0, k0=0, alpha0=-1/2, beta0=infinity  
 # Posterior hyperparameters mu1, k1, alpha1, beta1  
 mu1 <- xbar  
 k1 <- n  
 alpha1 <- -1/2 + n/2  
 beta1 <- 1/(0.5\*sum((x-xbar)^2))  
 spread1 <- sqrt(1/(k1\*alpha1\*beta1))  
 #Theoretical marginal density for theta  
 thetaVals <- seq(xbar - sd(x), xbar + sd(x), by = 0.01)  
 stdVals <- (thetaVals - mu1)/spread1  
 thetaMargDens <- dt(stdVals,df=2\*alpha1)/spread1  
 #Set simulation sample size  
 numSim <- 10000  
 # Simulate directly from the posterior distribution   
 rhoDirect <- rgamma(numSim,shape=alpha1,scale=beta1)  
 sigmaDirect <- 1/sqrt(rhoDirect)  
 thetaDirect <- rnorm(numSim,mean=xbar,sd=sigmaDirect/sqrt(n))  
 #Plot theoretical and Monte Carlo density functions  
 compare.plot <-  
 ggplot() +  
 geom\_line(data = data.frame('x' = thetaVals,  
 'density' = thetaMargDens),  
 aes(x=x,  
 y=density,   
 color = 'Theoretical')) +  
 geom\_line(data = data.frame('x' = thetaDirect),   
 aes(x=x,   
 color = 'Monte Carlo'),  
 stat = 'density') +  
 labs(title = sprintf('Monte Carlo Estimate for %s Depth', depth),  
 y='Density',  
 x='Theta') +  
 theme\_classic() +  
 xlim(c(xbar - sd(x), xbar + sd(x)))  
 #Return a list with a plot and the monte carlo simulated values  
 res <- list('plot' = compare.plot,  
 'thetaVals' = thetaDirect,  
 'rhoVals' = rhoDirect,  
 'sigmaVals' = sigmaDirect)  
 return(res)  
}

The resulting plots comparing the exact to monte carlo distributions are below.

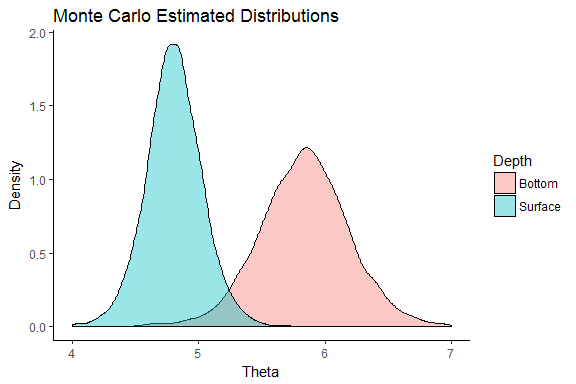
p2\_mc('Bottom')$plot



p2\_mc('Surface')$plot



Another shows the two monte carlo esimates on the same axes.



Use your Monte Carlo samples to estimate 90% posterior credible intervals for , , , and .

mc.surface.theta.ci <- quantile(p2\_mc('Surface')$thetaVals, probs = c(.05, .95))  
mc.surface.rho.ci <- quantile(p2\_mc('Surface')$rhoVals, probs = c(.05, .95))  
  
mc.bottom.theta.ci <- quantile(p2\_mc('Bottom')$thetaVals, probs = c(.05, .95))  
mc.bottom.rho.ci <- quantile(p2\_mc('Bottom')$rhoVals, probs = c(.05, .95))

The table below shows the CI from the monte carlo estimates. I have included the posterior CI values to compare with the results from Problem 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Surface.MC | Surface.Posterior | Bottom.MC | Bottom.Posterior |
| Theta | [4.44, 5.17] | [4.44, 5.17] | [5.25, 6.42] | [5.25, 6.43] |
| P | [0.9, 4.77] | [0.93, 4.72] | [0.37, 1.79] | [0.36, 1.83] |

The 90% CI’s from the monte carlo estimates are very close to the CI’s from the posterior distribution. There are small differences in a few of the values but none of the differences exceeds 0.01. The monte carlo estimates suggest that the HCB concentrations are lower at the surface and more consistent at the surface as well. The concentrations near the bottom have greater spread.

# Problem 3

Use your direct Monte Carlo sample to estimate the probability that the mean bottom concentration is higher than the mean surface concentration and to estimate the probability that the standard deviation of the bottom concentrations is higher than the standard deviation of the surface concentrations. I start by finding the number of instances where the monte carlo estimates for the bottom are higher then the surface. I also count the total number of monte carlo estimates returned (samples).

bottom.theta.higher <- sum(p2\_mc('Bottom')$thetaVals - p2\_mc('Surface')$thetaVals > 0)  
bottom.sigma.higher <- sum(p2\_mc('Bottom')$sigmaVals - p2\_mc('Surface')$sigmaVals > 0)  
samples <- length(p2\_mc('Bottom')$thetaVals)

Then I divide the counts where the bottom estimate was greater then the surface and divide by the total number of estimates to get the probability.

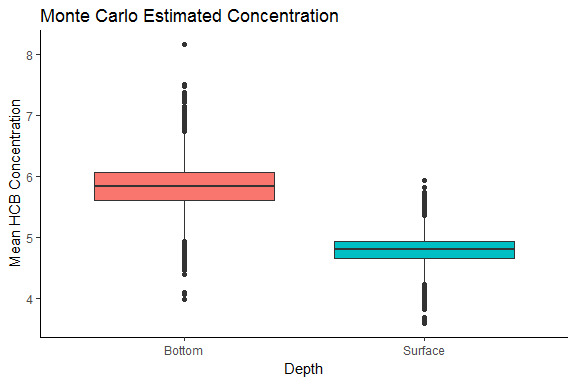
|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Bottom>Surface | #Samples | P(B>S) |
| Theta | 9893 | 10000 | 98.9% |
| StDev | 9122 | 10000 | 91.2% |

Both probabilities and are high. Neither is lower then 0.912.

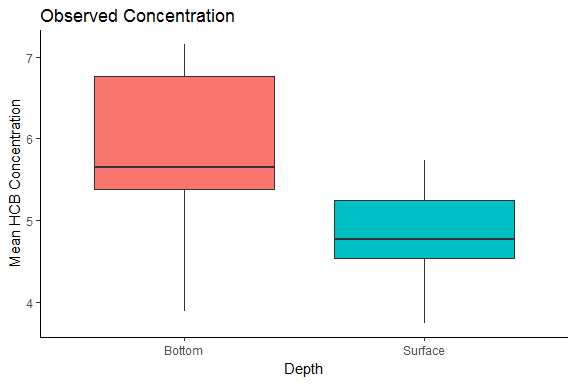
# Problem 4

### What are your conclusions about the distributions of surface and bottom concentrations?

The surface and botom concentrations are not the same. This is seen in the graphs above of the posterior and monte carlo estimated distributions. Another way to look at the two depths difference in concentrations is through a boxplot of the monte carlo values and observed values.



The boxplot of the monte carlo estimates shows that the mean concentrations measured at the surface depth are lower then the HCB mean concentrations measured near the bottom.

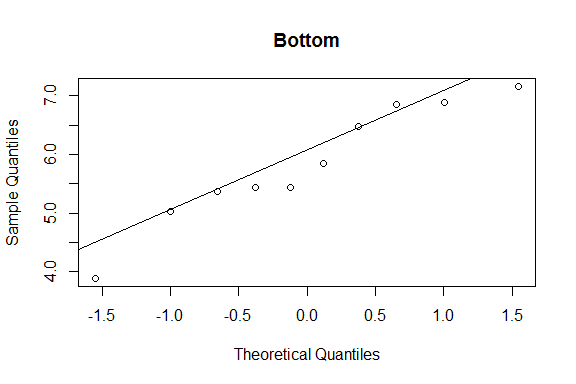


The data agrees with the monte carlo estimates. The two depths have different mean HCB concentration levels.

### Is the assumption of normality reasonable?

I first examine the observations from the bottom depth. A qqplot is a visual way to inspect the data in comparison to a normal distributon.

qqnorm( pollutants$Bottom, main='Bottom' )  
qqline( pollutants$Bottom )



The qqplot for the bottom data looks good. There are no places along the line where the data tails off. Another way to test for a normal distribution is by using the Shapiro-Wilk test of normality.

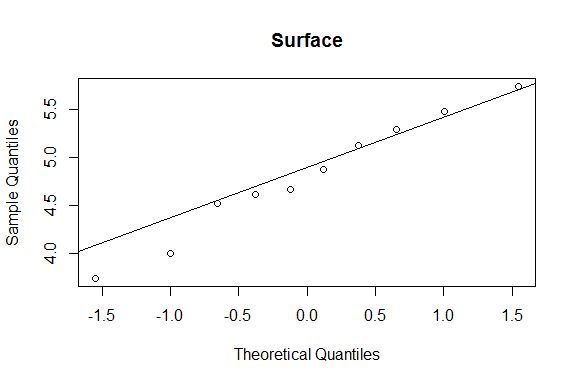
shapiro.test(pollutants$Bottom)

##   
## Shapiro-Wilk normality test  
##   
## data: pollutants$Bottom  
## W = 0.93794, p-value = 0.5303

The p-value from the test is > .05 thus the normal distribution can not be rejected as a fit to this data at the 95% confidence level.

Next I inspect the measurements taken near the surface.

qqnorm(pollutants$Surface, main='Surface')  
qqline( pollutants$Surface )



The qqplot for the bottom data looks good. There are no places along the line where the data tails off. Then I use the Shapiro-Wilk test of normality.

shapiro.test(pollutants$Surface)

##   
## Shapiro-Wilk normality test  
##   
## data: pollutants$Surface  
## W = 0.97454, p-value = 0.9294

The p-value from the test is > .05 thus the normal distribution can not be rejected as a fit to this data at the 95% confidence level.

### Are the means different for surface and bottom?

I used a standard two sample t-test to compare the mean of the two populations. I started by using the observed concentrations.

Surface.Mean.O <- pollutants$Surface  
Bottom.Mean.O <- pollutants$Bottom  
t.test(Surface.Mean.O, Bottom.Mean.O)

##   
## Welch Two Sample t-test  
##   
## data: Surface.Mean.O and Bottom.Mean.O  
## t = -2.74, df = 15.068, p-value = 0.01514  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.8398083 -0.2301917  
## sample estimates:  
## mean of x mean of y   
## 4.804 5.839

The p-value is less then .05 for this test thus we can reject the null hypothesis that the two means are equal.

I then performed a t-test using he estimated theta values from the monte carlo simulation.

Surface.Mean.MC <- p2\_mc('Surface')$thetaVals  
Bottom.Mean.MC <- p2\_mc('Bottom')$thetaVals  
t.test(Surface.Mean.MC, Bottom.Mean.MC)

##   
## Welch Two Sample t-test  
##   
## data: Surface.Mean.MC and Bottom.Mean.MC  
## t = -239.52, df = 16719, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.039951 -1.023068  
## sample estimates:  
## mean of x mean of y   
## 4.806915 5.838424

Again the p-value was lower then .05 so at the 95% confidence level the null hypothesis can be rejected. The true difference between the mean concentrations at the surface and bottom is not equal to 0 based on the observed data and monte carlo simulated results.

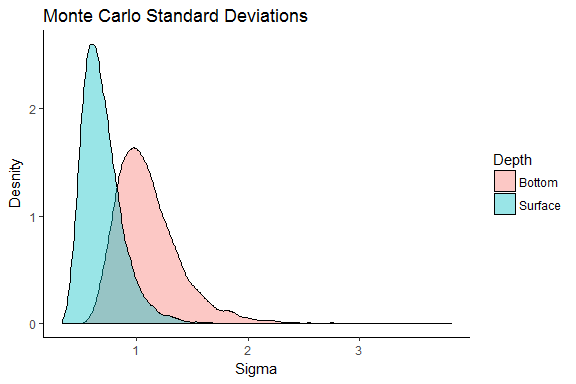
### The standard deviations?

I created two vectors containing the resulting standard deviations from the monte carlo results.

Surface.Sigma.MC <- p2\_mc('Surface')$sigmaVals  
Bottom.Sigma.MC <- p2\_mc('Bottom')$sigmaVals

I visually inspected these by plotting the distribution of standard deviations from the monte carlo simulation.

data.frame('Surface' = Surface.Sigma.MC,  
 'Bottom' = Bottom.Sigma.MC) %>%  
 gather(Depth, Sigma) %>%  
 ggplot() +  
 geom\_density(aes(x=Sigma,  
 fill=Depth),  
 alpha = 0.4) +  
 theme\_classic() +  
 labs(title='Monte Carlo Standard Deviations',  
 y='Desnity')



From the plot of the distributions it looks like the standard deviations are not the same. The standard deviation for the bottom concentrations appear to be larger then the surface standard deviation.

t.test(Surface.Sigma.MC, Bottom.Sigma.MC)

##   
## Welch Two Sample t-test  
##   
## data: Surface.Sigma.MC and Bottom.Sigma.MC  
## t = -115.99, df = 16618, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.4251336 -0.4110032  
## sample estimates:  
## mean of x mean of y   
## 0.6929264 1.1109948

The t-test confirms there is a difference not equal to 0 between the bottom and surface standard deviation values.