Assignment 8

Jake Whalen

April 10, 2018

Table of Contents

[Problem 1 3](#_Toc511487057)

[Person-Specific Mean 3](#_Toc511487058)

[Common Variance 6](#_Toc511487059)

[Problem 2 9](#_Toc511487060)

[Hyperparameter Estimation 9](#_Toc511487061)

[Person-specific Posterior Distributions 10](#_Toc511487062)

[Problem 3 12](#_Toc511487063)

#Packages  
library(tidyverse)  
library(magrittr)

# Problem 1

Section 18.4 of Gelman, et al. analyzes data on reaction times for 11 non-schizophrenic and 6 schizophrenic subjects. The data set can be found at this URL: <http://www.stat.columbia.edu/~gelman/book/data/schiz.asc>. The first 11 rows are data for the non-schizophrenic subjects. Gelman, et al. assume that the logarithms of the response times for each non-schizophrenic subject are independent and identically distributed normal random variables with person-specific mean and common variance . Discuss whether you think this assumption is reasonable.

I first read the data into R from the URL. I added a column (*condition*) that marked the first 11 rows as having schizophrenia and the last 6 as not having the condition.

#Data URL  
www <- "http://www.stat.columbia.edu/~gelman/book/data/schiz.asc"  
#Read and Wrangle  
ppl <- read.table(www, skip = 5) %>%   
 mutate(condition = rep(c("non-schizophrenics", "schizophrenics"),   
 c(11, 6)))

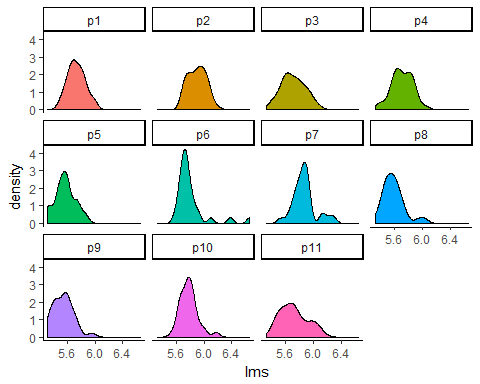
For the purposes of this assignment we are only concerned with the non-schizophrenic subjects reaction times. I filtered the data frame to only include the 11 non-schizophrenic subjects. I then dropped this variable since it was no longer needed and added a variable (*subject*) to differentiate between the 11 subjects. The problem also states we are assuming the log reaction times are normally distributed thus I transformed the given times using the log() equation in R and assigned the results to the variable *lms* in the data frame.

nschiz <- ppl %>%  
 filter(., condition == 'non-schizophrenics') %>%   
 select(-condition) %>%  
 mutate(subject = sprintf("p%s", 1:11)) %>%  
 gather(reaction, ms, -subject) %>%  
 mutate(lms = log(ms))

### Person-Specific Mean

I now have the data ready to check the assumption that the logarithms of the response times for each non-schizophrenic subject are independent and identically distributed normal random variables with person-specific mean and common variance . My first instinct is to plot the density curves using the 11 samples. This plot is shown below.

nschiz %>%  
 mutate(subject = factor(subject, levels = unique(nschiz$subject))) %>%  
 ggplot() +  
 geom\_density(aes(x=lms, fill=subject), show.legend = F) +  
 facet\_wrap(~subject) +  
 theme\_classic()



Subjects p1, p2, p3, p4, and p10 pass the initial eye test for a normal distribution. This is not enough however to carry on with the assumption of normality. A good statistical test to run that checks for normality is the Shapiro-Wilk test. The code and results are displayed below.

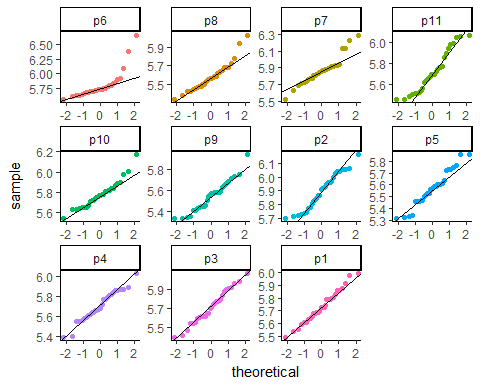
#Run S-W test on all 11 subjects samples and store results  
result.shapiro <- NULL  
for(s in unique(nschiz$subject)){  
 rt <- filter(nschiz, subject == s)$lms  
 st <- shapiro.test(rt)  
 result.shapiro <- rbind(result.shapiro,  
 c(s, st$p.value, st$statistic))  
}  
  
df.sh <-  
 data.frame('subject' = result.shapiro[,1],  
 'PValue' = as.numeric(result.shapiro[,2]),  
 'W' = as.numeric(result.shapiro[,3])) %>%  
 arrange(PValue)

|  |  |  |
| --- | --- | --- |
| Subject | P Value | W Statistic |
| p6 | 0.0000010 | 0.6857910 |
| p8 | 0.0327206 | 0.9233142 |
| p7 | 0.0403218 | 0.9267673 |
| p11 | 0.0485129 | 0.9298023 |
| p10 | 0.0945590 | 0.9406259 |
| p9 | 0.2001269 | 0.9527442 |
| p2 | 0.2502217 | 0.9564187 |
| p5 | 0.3365512 | 0.9614173 |
| p4 | 0.6297270 | 0.9731941 |
| p3 | 0.8338734 | 0.9803152 |
| p1 | 0.9251862 | 0.9843271 |

The first 4 subjects in the table above (p6, p7, p8, p11) have log reaction times that do not seem well fit to a normal distribution (95% confidence level). This is assumed based on the p-values < .05 and the lower W statistics. The rest of the subjects log reaction times do not reject the null hypothesis (95% confidence level) that the data are normally distributed.

Another way to check this assumption is through plots of the empirical and theoretical quantiles (qqplot). The code and chart below shows the results of plotting each subjects sample on a qqplot. I sorted the plots (left to right and down) by the p-values from the Shapiro-wilk test. Thus the less normal subjects are shown first and the more normal later on down the grid.

#Function for finding qqline slope and intercept  
qqplot.data <- function (vec){  
 y <- quantile(vec[!is.na(vec)], c(0.25, 0.75))  
 x <- qnorm(c(0.25, 0.75))  
 slope <- diff(y)/diff(x)  
 int <- y[1L] - slope \* x[1L]  
 return(list('slope'=slope, 'int'=int))  
}  
#Calculating the slope and intercapt for each subjects qqline  
sl.int <- NULL  
for(s in unique(nschiz$subject)){  
 rt <- filter(nschiz, subject == s)$lms  
 if(is.null(sl.int)){  
 sl.int <- data.frame('subject' = s,  
 'slope'=qqplot.data(rt)$slope,  
 'int'=qqplot.data(rt)$int)  
 }  
 else{  
 sl.int <- bind\_rows(sl.int,  
 data.frame('subject' = s,  
 'slope'=qqplot.data(rt)$slope,  
 'int'=qqplot.data(rt)$int))  
 }  
   
}  
#Plotting the qqplots  
ggplot() +  
 stat\_qq(data = nschiz %>%   
 mutate(subject = factor(subject, levels = df.sh$subject)),   
 aes(sample=lms, color = subject),  
 show.legend = F) +  
 geom\_abline(data = sl.int %>%   
 mutate(subject = factor(subject, levels = df.sh$subject)),  
 aes(slope = slope, intercept = int)) +  
 facet\_wrap(~subject, scales = 'free') +  
 theme\_classic()

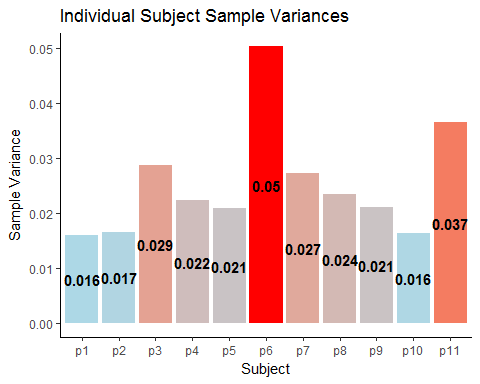


The top row contains the 4 subjects where the Shapiro-wilk test resulted in a rejection of the null hypothesis that the data are from a normal distribution. The plots support those results as seen by the points in the charts trailing off of the respective qqlines. The only other two qqplots that raise some concern are for p10 and p5. In both of these there is some separation at the upper end of the qqline.

The assumption of the data having person-specific means looks reasonable after investigation. While a few of the subjects data do not support this assumption very well the majority do.

### Common Variance

I also need to explore the assumption of the data having a common variance . I start by looking at the sample variances themselves. The question is do they vary drastically or not. The more similar they are the more reasonable the assumption is.

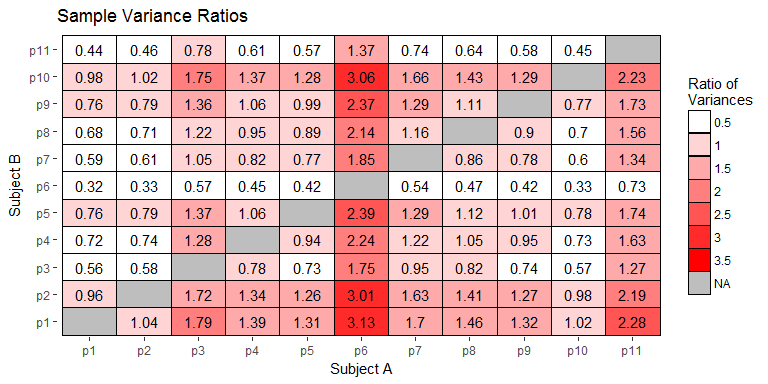


The plot above shows the sample variance for all 11 subjects as the height of the bars. Overall the sample variances seem to be close. Most are in or near the .02 to .03 range. The exception being the subjects p6 and p11. Similar to in the analysis of the means these two subjects do not conform as well as the others to the common variance.

To measure the similarity in the variances I calculated the ratio between every pair of subjects.

#Sample Variance Ratio  
result.var <- NULL  
for(s1 in unique(nschiz$subject)){  
 for(s2 in unique(nschiz$subject)){  
 #Subjects being compared  
 p1 <- filter(nschiz, subject == s1)$lms  
 p2 <- filter(nschiz, subject == s2)$lms  
 #Test  
 res <- var.test(p1, p2, alternative = "two.sided")  
 #Save results  
 result.var <- rbind(result.var,  
 c(s1, s2, res$statistic))  
 }  
}

The heat map below shows these ratios in a grid with color coding to highlight any ratios that are further away from 1. I purposely removed all pairings between the same subject thus the gray diagonal through the map.



A majority of the ratios fall within the 0.5 to 1.5 bins meaning the variances between the subjects are similar (1 is perfect similarity). There are however a large number of ratios that fall far from 1. Looking at the ratios between p6 and other subjects there are multiple ratios greater then 3.

The assumption of a common variance does not seem unreasonable. Subjects p6 and p11 seem to be outliers to a certain extent. The other 9 subjects have reasonably similar variances.

# Problem 2

For this problem, we will assume that the logarithms of the response times for each non- schizophrenic subject are independent and identically distributed normal random variables with person-specific mean . Unlike Gelman, et al., we will assume the precision may also depend on the subject. We will assume that the 11 means and precisions are independent and identically distributed normal-gamma random variables with center , precision multiplier , shape , and scale . To do an empirical Bayes analysis, we need estimates of the hyper parameters , , , and .

### Hyperparameter Estimation

* To estimate the shape and scale , estimate the sample precisions by calculating the sample variances and taking the inverses. Recall that the unknown precision has a gamma distribution with mean and variance . Therefore, we can estimate as the sample variance of precisions divided by the sample means of precisions . Then we can estimate as the sample mean of the divided by the estimate of .

ns.param <-   
 nschiz %>%  
 group\_by(subject) %>%  
 summarise(variance = var(lms)) %>%  
 mutate(precision = 1 / variance) %>%  
 ungroup()  
  
ns.beta <- var(ns.param$precision)/mean(ns.param$precision)  
ns.alpha <- mean(ns.param$precision) / ns.beta

* Estimate the center as the grand mean of all the log reaction times.

ns.mu <- mean(nschiz$lms)

* Estimate the precision multiplier as follows. First, find the sample mean of the log reaction times for each of the 11 subjects. Then calculate the sample variance of these 11 values. Then take the inverse to get the sample precision. Divide this sample precision by the average of to get your estimate of .

ns.k <- nschiz %>%  
 group\_by(subject) %>%  
 summarise(smean = mean(lms)) %>%  
 ungroup() %>%  
 summarise(svar = var(smean)) %>%  
 mutate(sprec = 1/ svar,  
 sk = sprec / mean(ns.param$precision)) %>%  
 select(sk)  
ns.k <- ns.k[[1]]

### Person-specific Posterior Distributions

Using these empirical Bayes estimates of , , , and , find the eleven posterior distributions of . Make a table to show your eleven posterior sets of hyper parameters. The table will have eleven rows, one for each subject, and a column for each of the four posterior hyper parameters.

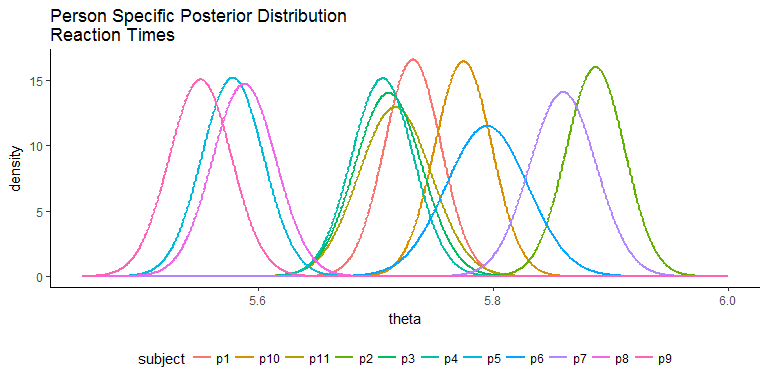
#Calculate Posterior Distributions  
findPost <- function(s){  
 subjSamples <- filter(nschiz, subject == s)$lms  
 nSubj <- nrow(filter(nschiz, subject == s))  
 xbarSubj <- mean(subjSamples)  
 #Alpha  
 aStar <- ns.alpha + nSubj / 2  
 #Beta  
 b.p1 <- 0.5 \* sum((subjSamples - xbarSubj)^2)  
 b.p2 <- ns.k \* nSubj \* (xbarSubj - ns.mu)^2 / (2 \* (ns.k + nSubj))  
 bStar <- (ns.beta^-1 + b.p1 + b.p2)^-1  
 #Mu  
 muStar <- (ns.k \* ns.mu + sum(subjSamples)) / (ns.k + nSubj)  
 #k  
 kStar = ns.k + nSubj  
   
 res <- list("mu" = muStar, "k" = kStar, "alpha" = aStar, "beta" = bStar)  
 return(res)  
}  
  
nsPost <- NULL  
for(s in unique(nschiz$subject)){  
 if(is.null(nsPost)){  
 nsPost <- data.frame(subject = s, findPost(s))  
 }  
 else{  
 nsPost %<>% bind\_rows(., data.frame(subject = s, findPost(s)))  
 }  
}

Table of 11 posterior distributions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| subject | mu | k | alpha | beta |
| p1 | 5.731433 | 31.66414 | 25.19519 | 2.153390 |
| p2 | 5.886999 | 31.66414 | 25.19519 | 2.006858 |
| p3 | 5.710691 | 31.66414 | 25.19519 | 1.544321 |
| p4 | 5.705655 | 31.66414 | 25.19519 | 1.798235 |
| p5 | 5.578023 | 31.66414 | 25.19519 | 1.807194 |
| p6 | 5.794623 | 31.66414 | 25.19519 | 1.036461 |
| p7 | 5.859362 | 31.66414 | 25.19519 | 1.553796 |
| p8 | 5.587903 | 31.66414 | 25.19519 | 1.702975 |
| p9 | 5.550679 | 31.66414 | 25.19519 | 1.776485 |
| p10 | 5.774478 | 31.66414 | 25.19519 | 2.118762 |
| p11 | 5.716373 | 31.66414 | 25.19519 | 1.312583 |

Plot of 11 posterior distributions.

nsNorm <-  
 nsPost %>%  
 mutate(sigma = sqrt(1/(alpha \* beta \* k))) %>%  
 select(mu, sigma) %>%  
 as.matrix()  
nsPostDens <- data.frame(theta = seq(5.45,6,length.out = 1000))  
for(r in 1:nrow(nsNorm)){  
 s <- sprintf("p%s", r)  
 nsPostDens[s] <- dnorm(nsPostDens$theta, nsNorm[r,'mu'], nsNorm[r,'sigma'])  
}  
nsPostDens %>%  
 gather(subject, density, -theta) %>%  
 ggplot() +  
 geom\_line(aes(x=theta, y=density, color=subject), size=1) +  
 theme\_classic() +  
 theme(legend.position = 'bottom') +  
 guides(col = guide\_legend(nrow = 1, byrow = TRUE)) +  
 labs(title='Person Specific Posterior Distribution\nReaction Times')



# Problem 3

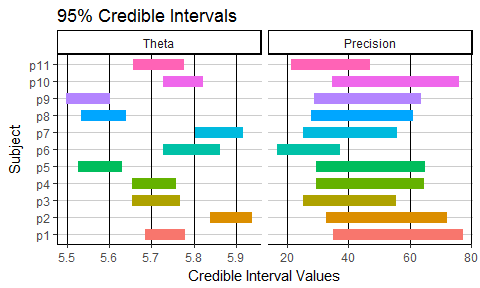
Find 95% credible intervals for each of the eleven means and precisions .

The table below shows the values of the 95% credible intervals.

nsPost %>%  
 mutate(ThetaCI = sprintf("[%s, %s]",  
 round(qnorm(0.025, mu, sqrt(1/(alpha \* beta \* k))), 2),  
 round(qnorm(0.975, mu, sqrt(1/(alpha \* beta \* k))), 2)),  
 PCI = sprintf("[%s, %s]",  
 round(qgamma(0.025, shape=alpha, scale=beta), 2),  
 round(qgamma(0.975, shape=alpha, scale=beta), 2))) %>%  
 select(subject, ThetaCI, PCI) %>%  
 knitr::kable(.)

|  |  |  |
| --- | --- | --- |
| subject | ThetaCI | PCI |
| p1 | [5.68, 5.78] | [35.18, 77.4] |
| p2 | [5.84, 5.94] | [32.78, 72.13] |
| p3 | [5.65, 5.77] | [25.23, 55.51] |
| p4 | [5.65, 5.76] | [29.38, 64.64] |
| p5 | [5.53, 5.63] | [29.52, 64.96] |
| p6 | [5.73, 5.86] | [16.93, 37.25] |
| p7 | [5.8, 5.92] | [25.38, 55.85] |
| p8 | [5.53, 5.64] | [27.82, 61.21] |
| p9 | [5.5, 5.6] | [29.02, 63.85] |
| p10 | [5.73, 5.82] | [34.61, 76.16] |
| p11 | [5.66, 5.78] | [21.44, 47.18] |

Since there are so many values to compare I decided to graph the intervals as line segments. The plot below is the result of this.



The log reaction times vary across subjects. The biggest difference between two subjects mean log reaction times are from p9 to p2. Subject p9 has a CI between 5.5 and 5.6 while p2 has a CI from 5.8 to 5.9. There does seem to be a clustering of subjects in three areas with regards to the CI’s. There are subjects with faster reaction times (lower values) p5, p8, p9. There are subjects with “average” reaction times (middle of the pack) p1, p3, p4, p10, and p11. The third grouping has the slowest reaction times (higher values) p2, p6, and p7.

The precisions for the 11 subjects all fall very close to each other. Subjects p6 and p11 have the precision CI’s least like the others. This is no surprise since they struggled to support the early assumptions of a normal distribution and common variance in Problem 1.