BACS-hw3.R

2022-04-11

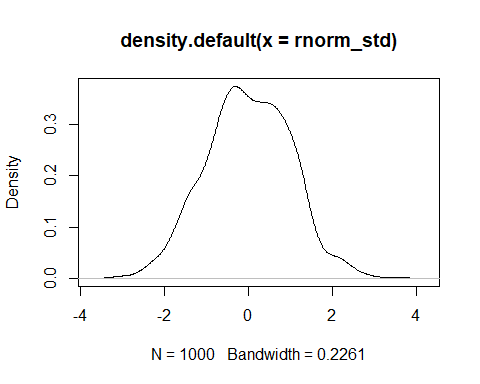
#Question1  
#a  
#1  
set.seed(10)  
rnorm\_dist <- rnorm(1000, mean=940, sd=150)  
  
standardize <- function(x){  
 (x - mean(x)) / sd(x)  
}  
  
rnorm\_std <- standardize(rnorm\_dist)  
mean(rnorm\_std)

## [1] 9.963238e-17

sd(rnorm\_std)

## [1] 1

#2  
plot(density(rnorm\_std))



#3  
standard normal distribution (?

#b  
#1  
setwd("C:/Users/eason/Desktop/清大 BACS/資料/")  
bookings <- read.table("first\_bookings\_datetime\_sample.txt", header=TRUE)  
bookings$datetime[1:9]

## [1] "4/16/2014 17:30" "1/11/2014 20:00" "3/24/2013 12:00" "8/8/2013 12:00"   
## [5] "2/16/2013 18:00" "5/25/2014 15:00" "12/18/2013 19:00" "12/23/2012 12:00"  
## [9] "10/18/2013 20:00"

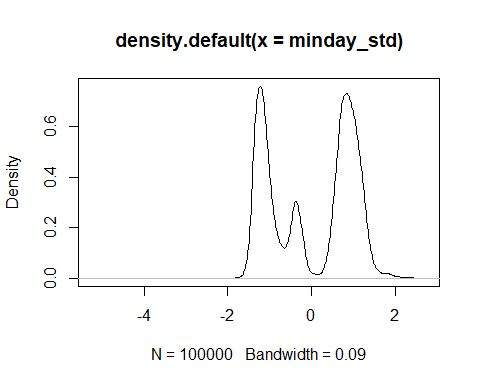
hours <- as.POSIXlt(bookings$datetime, format="%m/%d/%Y %H:%M")$hour  
mins <- as.POSIXlt(bookings$datetime, format="%m/%d/%Y %H:%M")$min  
minday <- hours\*60 + mins  
  
standardize <- function(x){  
 (x - mean(x)) / sd(x)  
}  
  
minday\_std <- standardize(minday)  
mean(minday\_std)

## [1] -4.25589e-17

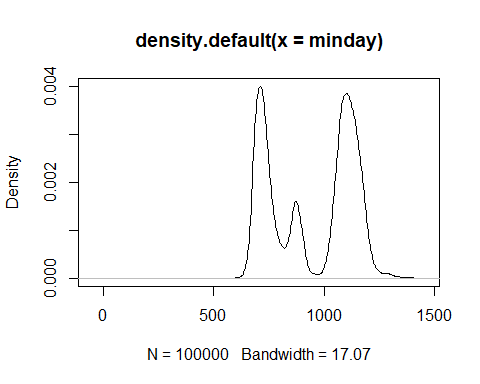
sd(minday\_std)

## [1] 1

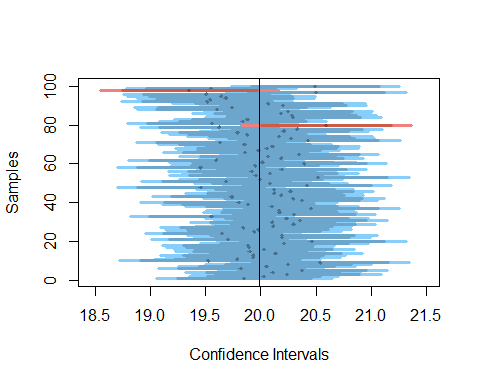
#2  
plot(density(minday\_std))



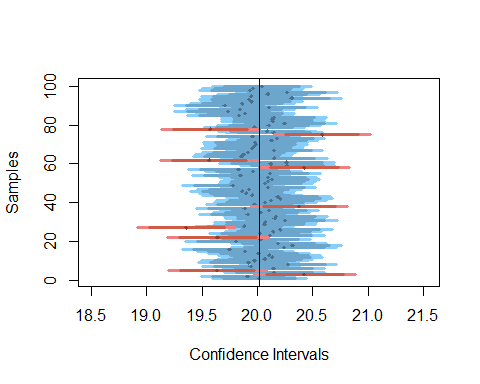
plot(density(minday))



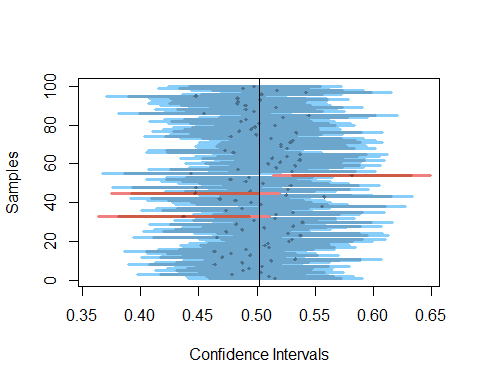
#Question2  
#a  
visualize\_sample\_ci <- function(num\_samples = 100, sample\_size = 100,   
 pop\_size=10000, distr\_func=rnorm, ...) {  
 # Simulate a large population  
 population\_data <- distr\_func(pop\_size, ...)  
 pop\_mean <- mean(population\_data)  
 pop\_sd <- sd(population\_data)  
   
 # Simulate samples  
 samples <- replicate(num\_samples,   
 sample(population\_data, sample\_size, replace=FALSE))  
   
 # Calculate descriptives of samples  
 sample\_means = apply(samples, 2, FUN=mean)  
 sample\_stdevs = apply(samples, 2, FUN=sd)  
 sample\_stderrs <- sample\_stdevs/sqrt(sample\_size)  
 ci95\_low <- sample\_means - sample\_stderrs\*1.96  
 ci95\_high <- sample\_means + sample\_stderrs\*1.96   
 ci99\_low <- sample\_means - sample\_stderrs\*2.58  
 ci99\_high <- sample\_means + sample\_stderrs\*2.58  
   
 # Visualize confidence intervals of all samples  
 plot(NULL, xlim=c(pop\_mean-(pop\_sd/2), pop\_mean+(pop\_sd/2)),   
 ylim=c(1,num\_samples), ylab="Samples", xlab="Confidence Intervals")  
 add\_ci\_segment(ci95\_low, ci95\_high, ci99\_low, ci99\_high,  
 sample\_means, 1:num\_samples, good=TRUE)  
   
 # Visualize samples with CIs that don't include population mean  
 bad = which(((ci95\_low > pop\_mean) | (ci95\_high < pop\_mean)) |  
 ((ci99\_low > pop\_mean) | (ci99\_high < pop\_mean)))  
 add\_ci\_segment(ci95\_low[bad], ci95\_high[bad], ci99\_low[bad], ci99\_high[bad],  
 sample\_means[bad], bad, good=FALSE)  
   
 # Draw true population mean  
 abline(v=mean(population\_data))  
}  
  
add\_ci\_segment <- function(ci95\_low, ci95\_high, ci99\_low, ci99\_high,   
 sample\_means, indices, good=TRUE) {  
 segment\_colors <- list(c("lightcoral", "coral3", "coral4"),  
 c("lightskyblue", "skyblue3", "skyblue4"))  
 color <- segment\_colors[[as.integer(good)+1]]  
   
 segments(ci99\_low, indices, ci99\_high, indices, lwd=3, col=color[1])  
 segments(ci95\_low, indices, ci95\_high, indices, lwd=3, col=color[2])  
 points(sample\_means, indices, pch=18, cex=0.6, col=color[3])  
}  
  
visualize\_sample\_ci(num\_samples = 100, sample\_size = 100, pop\_size = 10000, distr\_func = rnorm, mean = 20, sd = 3)



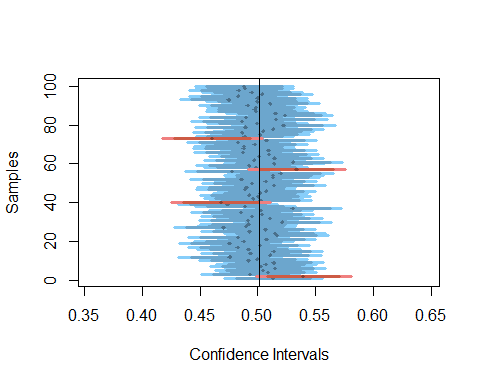
#b  
visualize\_sample\_ci(num\_samples = 100, sample\_size = 300, pop\_size = 10000, distr\_func = rnorm, mean = 20, sd = 3)



#c  
visualize\_sample\_ci(num\_samples = 100, sample\_size = 100, pop\_size = 10000, distr\_func = runif)



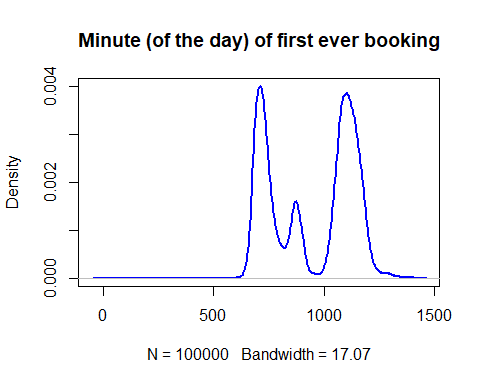
visualize\_sample\_ci(num\_samples = 100, sample\_size = 300, pop\_size = 10000, distr\_func = runif)



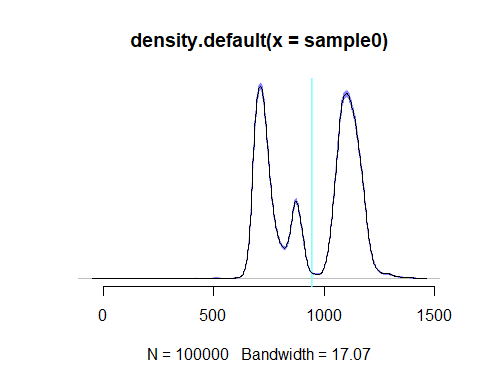
#Question 3  
#a  
setwd("C:/Users/eason/Desktop/清大 BACS/資料/")  
bookings <- read.table("first\_bookings\_datetime\_sample.txt", header=TRUE)  
bookings$datetime[1:9]

## [1] "4/16/2014 17:30" "1/11/2014 20:00" "3/24/2013 12:00" "8/8/2013 12:00"   
## [5] "2/16/2013 18:00" "5/25/2014 15:00" "12/18/2013 19:00" "12/23/2012 12:00"  
## [9] "10/18/2013 20:00"

hours <- as.POSIXlt(bookings$datetime, format="%m/%d/%Y %H:%M")$hour  
mins <- as.POSIXlt(bookings$datetime, format="%m/%d/%Y %H:%M")$min  
minday <- hours\*60 + mins  
plot(density(minday), main="Minute (of the day) of first ever booking", col="blue", lwd=2)



#1  
minday\_mean <- mean(minday)  
minday\_sd <- sd(minday)  
minday\_size <- length(minday)  
minday\_se <- (minday\_sd / sqrt(minday))  
minday\_ci\_95 <- quantile(minday, probs=c(0.025, 0.975))  
  
#2,3 b2的方法也可以 俐落一點  
plot\_resample <- function(sample0){  
 resample <- sample(sample0, length(sample0), replace = TRUE)  
 lines(density(resample), col = rgb(0.5, 0.5, 1, 0.1))  
 resample\_stat <- mean(resample)  
 abline(v = mean(resample), col = rgb(0.5, 1, 1, 0.1))  
 return(resample\_stat)  
}  
  
show\_resample\_width <- function(sample0){  
 num\_bootstraps = 2000  
 plot(density(sample0), lwd = 0, ylab = "", frame.plot = FALSE, yaxt = "n")  
 sample\_means <- replicate(num\_bootstraps, plot\_resample(sample0))  
 lines(density(sample0), lwd = 1, col = "black")  
}  
  
show\_resample\_width(minday)



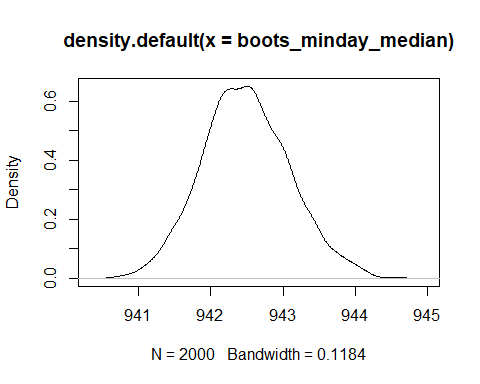
#4  
quantile(show\_resample\_width(minday), probs = c(0.025, 0.975))

## 2.5% 97.5%   
## NA NA

#b  
#1  
minday\_median <- median(minday)  
minday\_median

## [1] 1040

#2  
sample0 <- minday  
resample\_minday\_median <- function(sample0) {  
 resample <- sample(sample0, length(sample0), replace=TRUE)  
 mean(resample)  
 }  
  
boots\_minday\_median <- replicate(2000, resample\_minday\_median(sample0))  
  
plot(density(boots\_minday\_median))



#3  
quantile(boots\_minday\_median, probs = c(0.025, 0.975))

## 2.5% 97.5%   
## 941.3628 943.7383