Bootstrap R-Shizhi Chen

October 6, 2018

1 COURSEWORK 1: Working with Census Data

Shizhi Chen-10307389

1.1 Bootstrap the data and plot the pointwise 95% confidence region around the age histogram

```
In [4]: #Read and organize data
```

```
In [5]: data<-read.csv("Data_AGE_UNIT-2",header=T)
        age<-c(data$Age)
        counts<-c(data$counts)
        d1<-data.frame(id=c(0:100),age,counts)
        d2<-data.frame(id=rep(d1$id,d1$counts),age=rep(d1$age,d1$counts))
        tab1<-as.data.frame(table(d2$age))
        tabf1<-as.data.frame(tab1$Freq)
        rownames(tabf1)<-paste(rep("age",101),c(0:100),sep="_")
        colnames(tabf1)<-c("ocount")
```

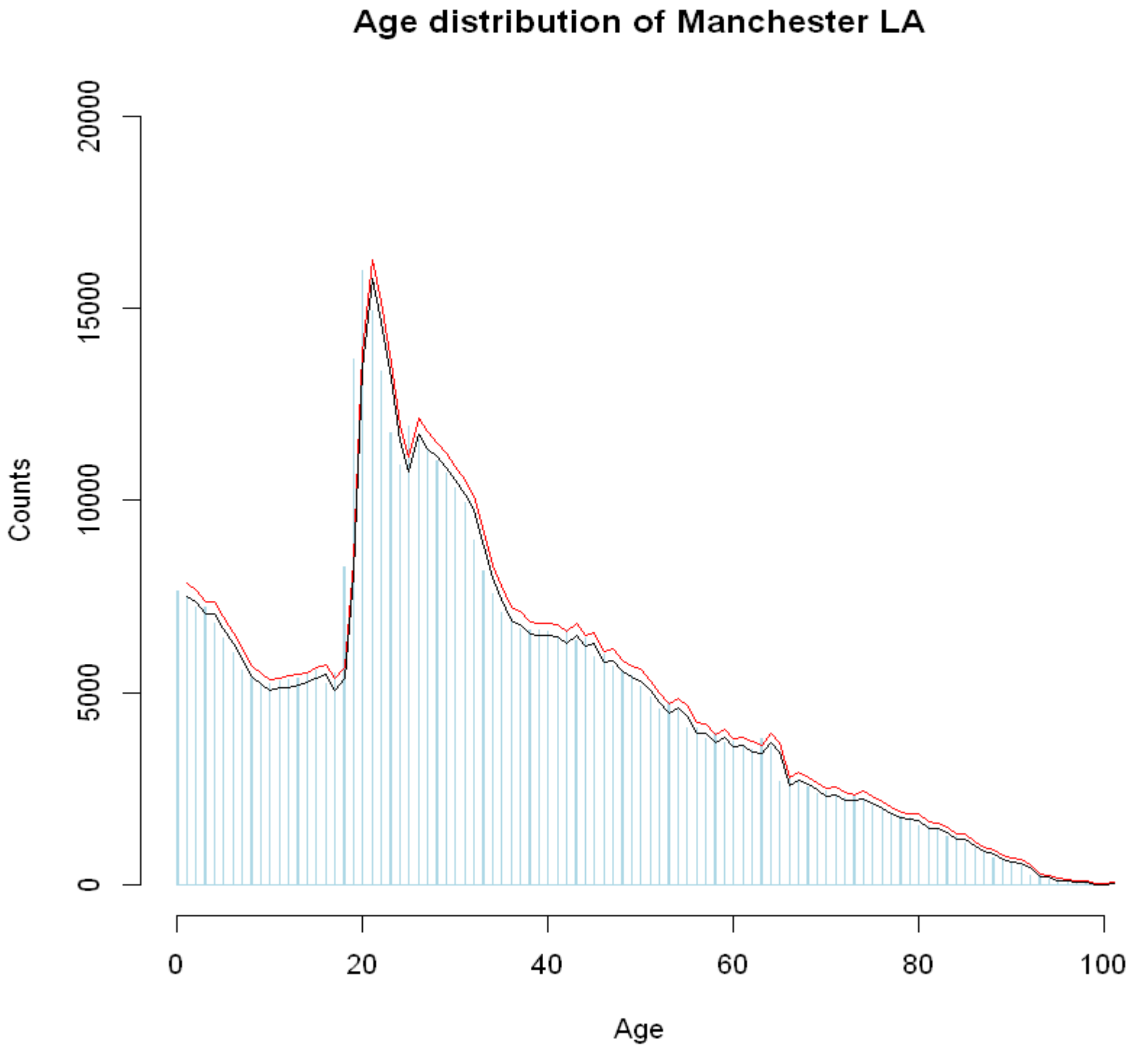
```
In [6]: #Sample m times with replacement (bootstrap)-by using sample function
```

```
In [7]: n<-sum(d1$counts) # calculate the total counts
        m<-200
        boots<-list()
        pre<-list()
        reg<-as.data.frame(matrix(0,nrow=101,ncol=m+1))
        reg[,1]<-c(0:100) #define age range
        for(i in 1:200){
        boots[[i]]<-sample(d2$age,n,replace=T)
        pre[[i]]<-as.data.frame(table(boots[[i]]))
        colnames(pre[[i]])<-c("agegroup","count")
        reg[,i+1]<-pre[[i]]$count
        }
        bootp<-t(reg)[2:201,]
        colnames(bootp)<-paste(rep("age",101),c(0:100),sep="_")
        rownames(bootp)<-paste(rep("bootstrap",200),c(1:200),sep="_")
```

```
In [8]: #set the confidence interval of 95%
In [9]: reg2<-data.frame(pl=rep(0,101),pu=rep(0,101),pm=rep(0,101))
        for(j in 1:101){
        reg2$pl[j]<-quantile(bootp[,j],0.025)
        reg2$pu[j]<-quantile(bootp[,j],0.975)
        }
        rownames(reg2)<-paste(rep("age",101),c(0:100),sep="_")
```

```
In [10]: #draw the original data histogram and the upper and lower confidence interval lines
```

```
In [11]: hist(d2$age,xlab = "Age",ylab="Counts",border = "lightblue", xlim = c(0,100),
              ylim = c(0,20000),breaks =2000,main = "Age distribution of Manchester LA")
         ageboot<-cbind(reg2,tabf1)
         lines(ageboot$pu,col="red") # 97.5% upper line
         lines(ageboot$pl,col="black") # 2.5% lower line
```



1.2 Discuss which ‘wiggles’ are significantly larger than the confidence region

From the figure can be seen that the population between the ages of 18 and 20 is significantly higher than the upper line (The red line) of the 95% confidence interval.

For this phenomenon,people aged around 19 years old could be the migrant population such as a large number of international students who are studying in the bachelor degree in the Manchester′s universities. Thereforethe "wiggles" in 18-20 years old group might be explained due to the undergraduate students are always around 19 years old.

As for the "wiggle" in the 25 years old population, it also considered to be the influence of inter-national students, such as postgraduate students or PhD who are aged around 25 years old. This wiggle is smaller than the previous one and it is reasonable because the number of undergraduates is greater than the postgraduate and PhD students.

In addition, there are some other "wiggles" like the population in 63 years old which are recognized as noise.

1.3 Consider dierent algorithmic approaches to boostrapping and their eiciency

In order to find the“wiggles”or noise in Manchester population distribution, the method of re-sampling was used. In RI use the sample() function to randomly select n observations from the age range 0 to 100, with replacement. This is equivalent to constructing a new bootstrap data set and repeat this process for m times so that we can find the maximum and minimum of each age count from the m sampling process and remove the upper and lower 2.5% values to get the 95% confidence interval. In general,this program needs to run for 1-2 minutes.

Similarly, we can use the np.random.choice function or stairs method (Python) to solve this problem. Howeverin the case of the same number of samples and sampling times it will need to take more than 5mins to figure out the results.