Bootstrap worksheet

Sajith Gowthaman (ek5282))

2/8/2020

Ques 1: a) How many possible bootstrap resamples of these data are there?

```
choose(23,12)

## [1] 1352078
```

1352078 possible bootstrap resamples are there

b)Using R and the sample() function, or a random number table or generator, generate five resamples of the integers from 1 to 12.

```
Org_data<-c(4.94,5.06,4.53,5.07,4.99,5.16,4.38,4.43,4.93,4.72,4.92,4.96)
B_Sample1<-sample(Org_data,1:12,replace = TRUE)
B_Sample2<-sample(Org_data,1:12,replace = TRUE)
B_Sample3<-sample(Org_data,1:12,replace = TRUE)
B_Sample4<-sample(Org_data,1:12,replace = TRUE)
B_Sample5<-sample(Org_data,1:12,replace = TRUE)
```

c)For each of the resamples in b, find the mean of the corresponding elements of the aflatoxin data set. Print out the 5 bootstrap means.

```
mean(Sample1)

## [1] 4.92

mean(Sample2)

## [1] 5.16

mean(Sample3)

## [1] 4.38

mean(Sample4)

## [1] 4.92

mean(Sample5)

## [1] 4.92
```

d)Find the mean of the resample means. Compare this with the mean of the original data set.

```
R_Mean<-c(Sample1,Sample2,Sample3,Sample4,Sample5)
mean(ResampleMean)

## [1] 4.86

mean(org_data)

## [1] 4.840833
```

e)Find the minimum and the maximum of the five resample means. This a crude bootstrap confidence interval on the mean.

```
min(R_Mean)

## [1] 4.38

max(R_Mean)

## [1] 5.16
```

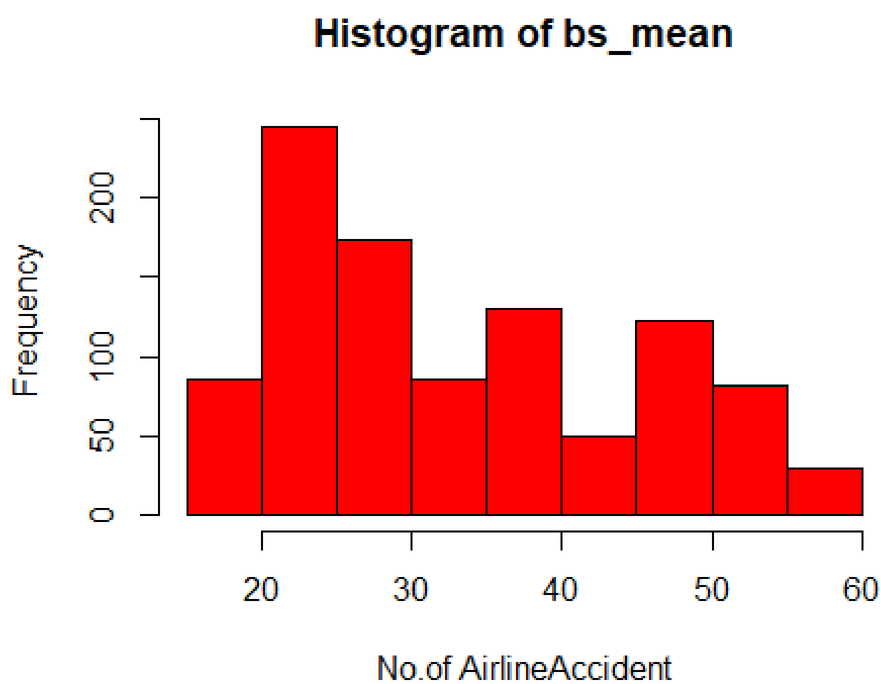
Ques2: a)For the sample data, compute the mean and its standard error and the median

```
Airline_Accident<-c(23, 16, 21, 24, 34, 30,28, 24, 26, 18, 23, 23, 36, 37, 49
, 50, 51, 56, 46, 41, 54, 30, 40,31)
mean(AirlineAccident)

## [1] 33.79167

sd(AirlineAccident)

## [1] 12.06497

median(AirlineAccident)

## [1] 30.5
```

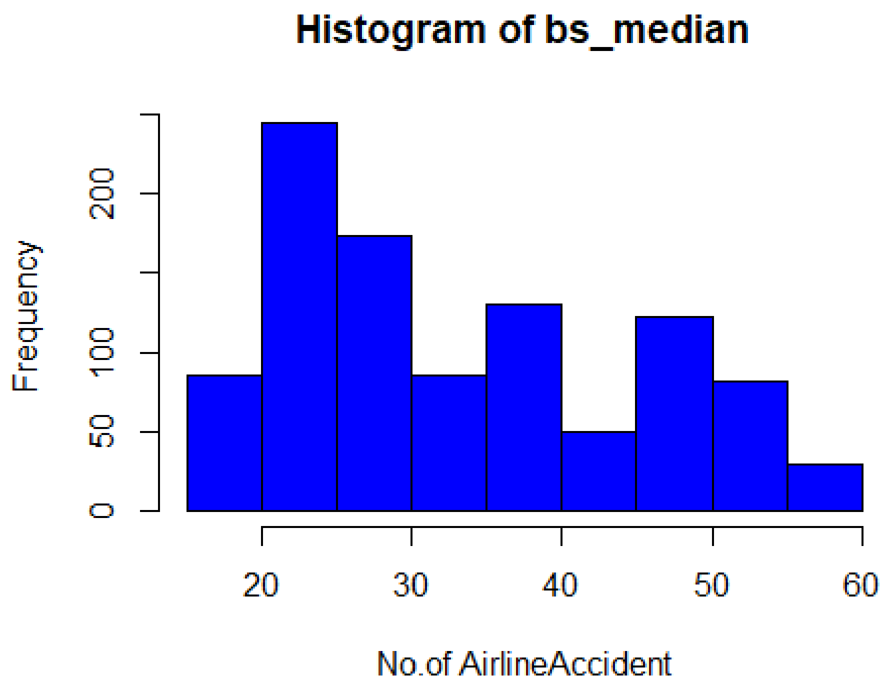
b) Compute bootstrap estimates of the mean, median and 25% trimmed mean with estimates of their standard errors, using B = 1000 resamples.

```
bs_mean=NULL
bs_median=NULL
B=1000
 set.seed(1)
 for (i in 1:B) {
   bs_AirlineAccident=sample(AirlineAccident,1:24,replace = TRUE)
   bs_mean[i]=mean(bs_AirlineAccident)

   bs_median[i]=median(bs_AirlineAccident)
 }
 mean(bs_mean)

## [1] 33.495

 median(bs_median)
```

```
## [1] 30

 mean(bs_mean, trim = .25)

## [1] 31.666

 sd(bs_mean)

## [1] 11.52635

 sd(bs_median)

## [1] 11.52635

par(mfrow=c(1,1))
hist(bs_mean,xlab = "No.of AirlineAccident", ylab = "Frequency",col = "red")
```




```
hist(bs_median,xlab = "No.of AirlineAccident", ylab = "Frequency",col = "blue
")
```



c) Compare parts a and b. How do the estimates compare? # In part a), it returns the average and median of the airline accidents, whereas in part b) apart from the average airline accident and median value, it also shows the 25% trimmed mean of the accident.

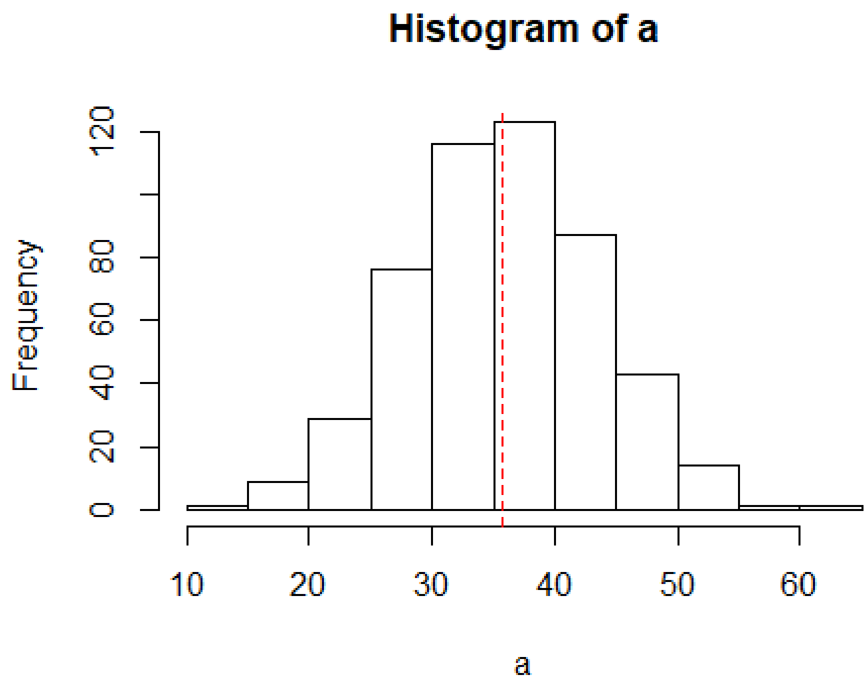
Ques3: Consider a population that has a normal distribution with mean μ = 36, standard deviation σ = 8. a) The sampling distribution of $\bar{X}$ for samples of size 500 will have what distribution, mean and standard error?

ans: For this question, it will be normally distributed with a sample of size 500, mean = 36, and std. error = 8/sq.root(500).

b)Use R to draw a random sample of size 500 from the population. Conduct exploratory data analysis on your sample

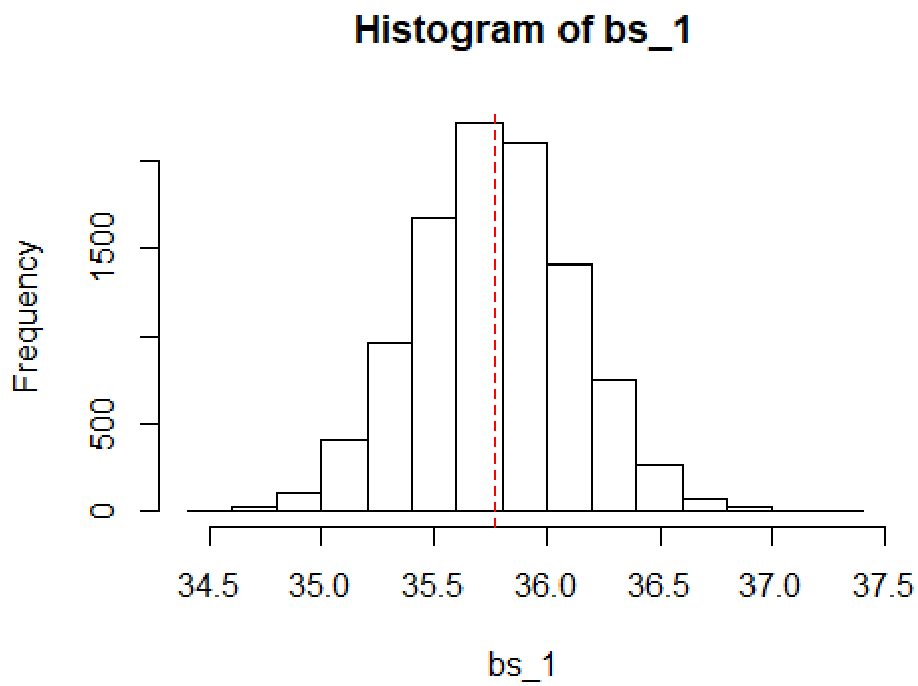
```
set.seed(4)
a=rnorm(500,36,8)
mean(a)

## [1] 35.76682

sd(a,na.rm = FALSE)

## [1] 7.751093

sd(a)/sqrt(500)

## [1] 0.3466394
```

```r
hist(a)
abline(v=mean(a), col="red", lty=2)
```



c)Compute the bootstrap distribution for your sample and note the bootstrap mean and standard error

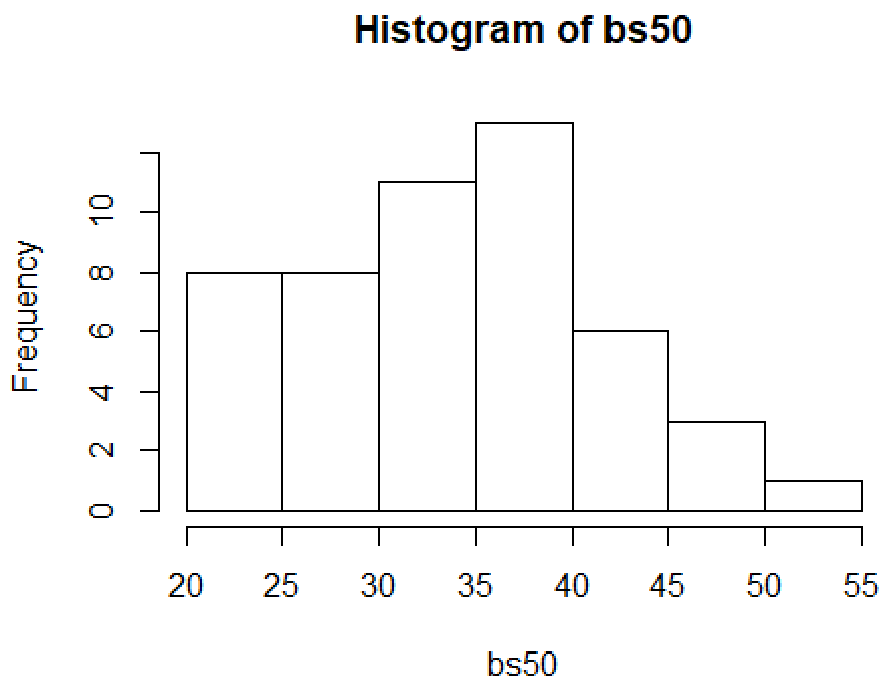
```r
B=10^4
bs_1=NULL

set.seed(1000)

for (i in 1:B) {
  bs1=sample(a,500, replace = TRUE)
  bs_1[i]= mean(bs1)
}
mean(bs_1)

## [1] 35.76719

sd(bs_1)

## [1] 0.3487043

hist(bs_1)
abline(v=mean(bs_1), col="red", lty=2)
```



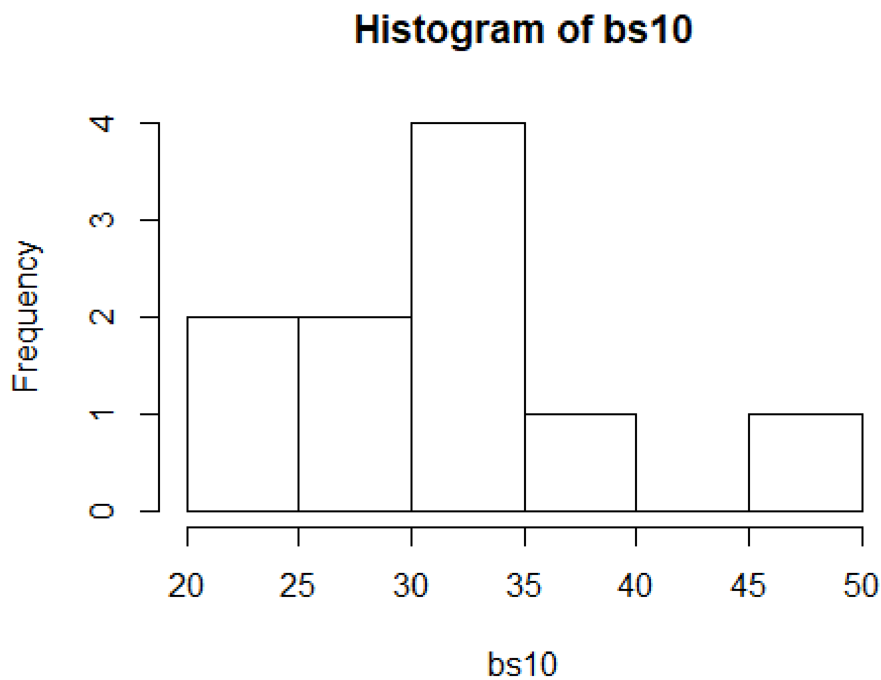
d) #Dist #Mean #std.dev #Population 36 8 #Sampling dist 36 8 #Sample 35.76 7.75 #Bootstrap sample 35.76 0.34

e)Repeat for sample of sizes n = 50 and n = 10. Carefully describe your observations about the effects of sample size on the bootstrap distribution.

```
bs_50=NULL
set.seed(100)
for (i in 1:B) {
bs50=sample(a, 50, replace=T)
bs_50[i]=mean(bs50)
}
mean(bs50)

## [1] 33.92609

sd(bs50)

## [1] 7.831231

hist(bs50)
```




```
bs_10=NULL
set.seed(11)
for (i in 1:B) {
bs10=sample(a, 10, replace=T)
bs_10[i]=mean(bs10)
}
mean(bs10)

## [1] 31.96081

sd(bs10)

## [1] 7.226446

hist(bs10)
```



Ques4: a)

```
(6/8+5/8+5/8+5/8+7/8+4/8)/6

## [1] 0.6666667
```

b)Write out the R code to generate data of 100 parametric bootstrap samples and compute an 80% percentile confidence interval for θ.

```
set.seed(10)
x_binomial=rbinom(100,8,(2/3))
bs_binomial=NULL

for (i in 1:B) {
  bs_sample=sample(x_binomial,100, replace = TRUE)
  bs_binomial[i]=mean(bs_sample)
}

quantile(bs_binomial,c(0.1,0.8))

##  10%  80%
## 5.18 5.43
```

80% percentile confidence interval for θ is between 5.179 and 5.432

Ques 5: a)Propose a parametric approach to answer this question. Mention clearly all assumptions for such an approach

Ans: Lets conduct a hypothesis test and then conduct a T-test to find if there is a significant difference by evaluating the value of P. p-value>0.05. F

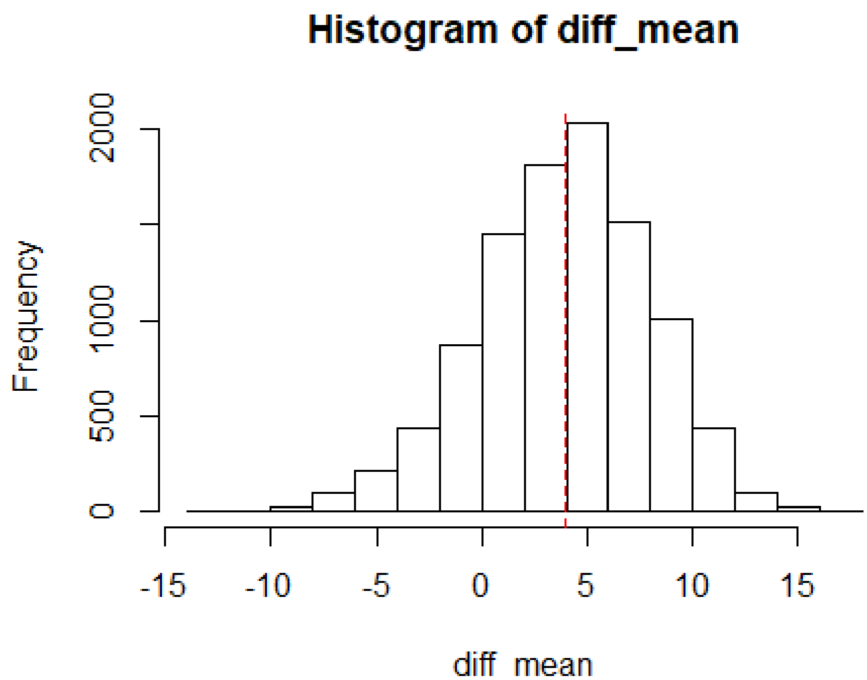
Fail to reject the hypothesis.

```
Full_time<-c(28, 23, 18, 16, 15, 15, 13, 31, 31)
student<-c(9, 11, 14, 14, 16, 19, 37)
t.test(Full_time, student , var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  Full_time and student
## t = 0.95805, df = 14, p-value = 0.3543
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.915511 12.852019
## sample estimates:
## mean of x mean of y
##  21.11111  17.14286
```

b) We fail to reject the hypothesis.

```
B=10^4
diff_mean=numeric(B)
for(i in 1:B)
{
fl.sample=sample(Full_time_prof, 9, T)
gt.sample=sample(Grad_student, 7, T)
diff_mean[i]=mean(fl.sample)- mean(gt.sample)
}
hist(diff_mean)
abline(v=mean(Full_time_prof)-mean(Grad_student), col="red", lty=2)
```




```r
quantile(diff_mean, c(0.025, 0.975))

##      2.5%     97.5%
## -4.555556 11.031746
```