**CSS544 Foundations of Analytics**

**Final Project**

**1. Import Data**

Steps:

1) Select dataset from UCI website and download the *abalone.data* file from the link <http://archive.ics.uci.edu/ml/machine-learning-databases/abalone/> to the local position. (c:/abalone.data)

2) Open the R, import data from local data file and name the table column names:

**mydata<-read.table("c:/abalone.data",**

**sep=",",header=F,col.names=c("SEX","LENGTH",**

**"DIAMETER","HEIGHT","WHOLEWEIGHT","SHUCKEDWEIGHT","VISCERAWEIGHT",**

**"SHELLWEIGHT","RINGS"))**

3) Check the data type for each column and convert the data type to numeric or character if necessary:

Name / Data Type / Measurement Unit / Description   
-----------------------------   
Sex / nominal / -- / M, F, and I (infant)   
Length / continuous / mm / Longest shell measurement   
Diameter / continuous / mm / perpendicular to length   
Height / continuous / mm / with meat in shell   
Whole weight / continuous / grams / whole abalone   
Shucked weight / continuous / grams / weight of meat   
Viscera weight / continuous / grams / gut weight (after bleeding)   
Shell weight / continuous / grams / after being dried   
Rings / integer / -- / +1.5 gives the age in years

**typeof(mydata$LENGTH[mydata$LENGTH!="?"])**

[1] "double"

Thus, the type of the columns from LENGTH to SHELLWEIGHT are numeric. No need to convert.

**typeof(mydata$SEX[mydata$SEX!="?"])**

[1] "integer"

Thus, we need to convert SEX data to character:

**mydata$SEX<-as.character(mydata$SEX)**

4) Save the data as .RData file to local:

**save(mydata,file="c:/ mydata.RData")**

5) when next time use the data, we load the RData file directly.

When other users use this file, just move the “mydata.RData” file in the folder of local position (like c:/)

And then

**load("c:/ mydata.RData")**

**2. Analyze Data**

1) Analysis for categorical and numerical data.

i) Categorical data: analyzing the sex category of abalones in the sample, to see the frequency and proportion of different sex of abalones.

**sex<-mydata$SEX**

**table(sex)**

sex

F I M

1307 1342 1528

Notes: Using table() function to summarize the categorical data (SEX column), showing the frequencies of values in data. It shows that there are 1307 female abalones, 1342 infant abalones, and 1528 male abalones.

**table(sex)/length(sex)**

sex

F I M

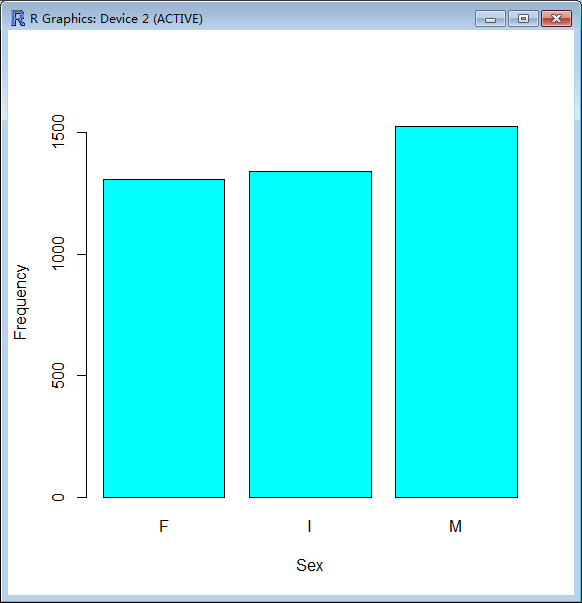
0.3129040 0.3212832 0.3658128

It shows that the proportion of each of the three categories with respect to the total size of dataset. For example, the male abalones take the largest proportion, which is 37% of the whole samples.

**barplot(table(sex),**

**+ col="cyan", ylim=c(0,1600),**

**+ xlab="Sex", ylab="Frequency")**

****

Notes: The frequency of the sex variable is graphically represented using a barplot.

**sexdata<-table(sex)**

**slice.labels<-names(sexdata)**

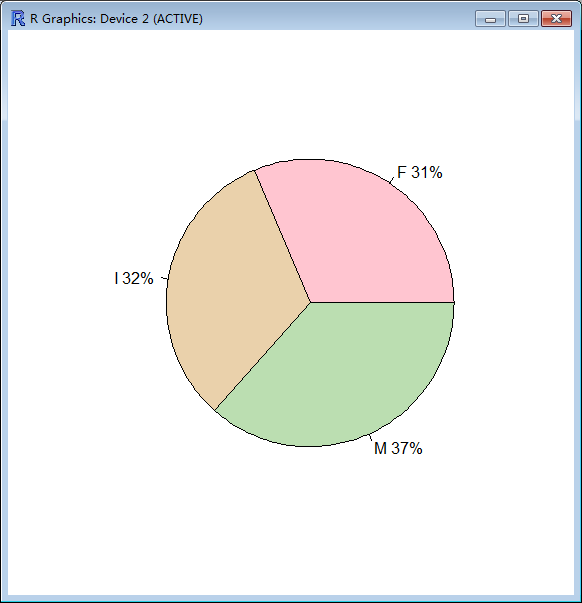
**slice.percents<-round(sexdata/sum(sexdata)\*100)**

**slice.labels<-paste(slice.labels,slice.percents)**

**slice.labels<-paste(slice.labels,"%",sep="")**

**pie(sexdata,labels=slice.labels,**

**col=hcl(c(0,60,120)))**

****

Notes: The pie chart is useful for comparing the proportions of each category. It makes the proportions better look.

ii) Numerical data

Analyzing the Rings column, which predicts the age of abalone. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope.

**rings<-mydata$RINGS**

**mean(rings)**

[1] 9.933684

Notes: The mean of the rings variable is 9.93, which means that the average of abalones’ age is 10 years.

**median(rings)**

[1] 9

**table(rings)**

rings

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

1 1 15 57 115 259 391 568 689 634 487 267 203 126 103 67 58 42 32 26

21 22 23 24 25 26 27 29

14 6 9 2 1 1 2 1

**which(table(rings)==max(table(rings)))**

9

9

Notes: calculate the median and the value with highest frequency.

**range(rings)**

[1] 1 29

**diff(range(rings))**

[1] 28

**var(rings)**

[1] 10.39527

**sd(rings)**

[1] 3.224169

Notes: calculate the range and different of range, variance and standard deviation of the rings variable.

**summary(rings)**

Min. 1st Qu. Median Mean 3rd Qu. Max.

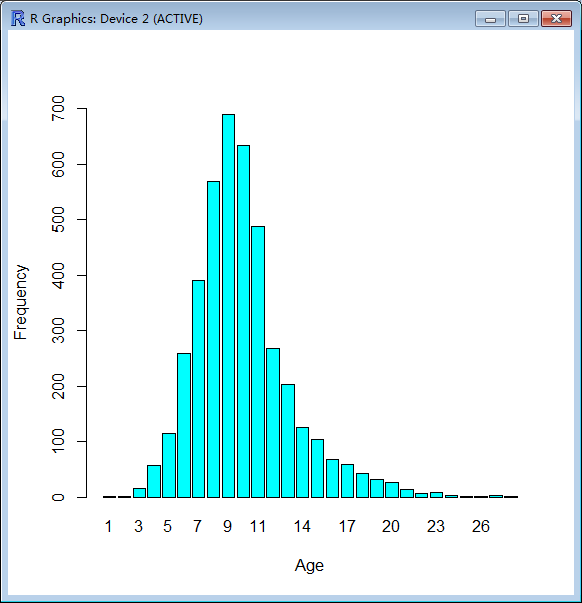
1.000 8.000 9.000 9.934 11.000 29.000

Notes: from the summary five numbers, we can see the maximum value is 29, minimum value is 1, and median value is 9. Which means the oldest abalone is 29 years old, the youngest is just 1 year old. And the median, average, and most frequency abalones’ age is 9 years old.

**barplot(table(rings),**

**+ col="cyan",ylim=c(0,700),**

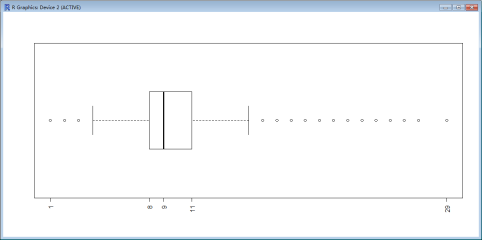
**+ xlab="Age", ylab="Frequency")**



Notes: The barplot can show the frequency graphically. We can see that the 9 age takes the highest frequency. And the Rings value(age value) is Normal Distribution.

**boxplot(rings,horizontal=TRUE,xaxt="n")**

**axis(side=1,at=fivenum(rings),labels=TRUE,las=2)**



Notes: the five numbers can be seen graphically. And there are 16 outliers.

2) Pick one variable with numerical data and examine the distribution of the data.

i) As analyzed above, the Rings value is Normal Distribution.

ii) Pick another numeric data: HEIGHT variable, and examine its distribution:

**height<-mydata$HEIGHT**

**mean(height)**

[1] 0.1395164

**table(height)**

height

0 0.01 0.015 0.02 0.025 0.03 0.035 0.04 0.045 0.05 0.055 0.06

2 1 2 2 5 6 6 13 11 18 25 26

0.065 0.07 0.075 0.08 0.085 0.09 0.095 0.1 0.105 0.11 0.115 0.12

39 47 61 76 74 124 91 145 114 135 133 169

0.125 0.13 0.135 0.14 0.145 0.15 0.155 0.16 0.165 0.17 0.175 0.18

202 169 189 220 182 267 217 205 193 160 211 131

0.185 0.19 0.195 0.2 0.205 0.21 0.215 0.22 0.225 0.23 0.235 0.24

103 103 78 68 45 23 31 17 13 10 6 4

0.25 0.515 1.13

3 1 1

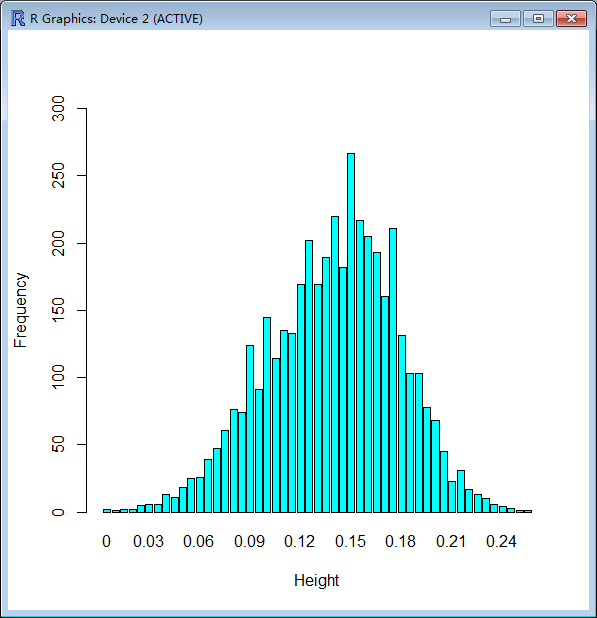
**sd(height)**

[1] 0.04182706

**barplot(table(height),**

**+ col="cyan",ylim=c(0,300),**

**+ xlab="Height", ylab="Frequency")**



Notes: As shown above, the classic bell-shaped curve represents the **normal distribution**. The mean value is 0.14, and standard derivation is 0.042.

3) Draw various random samples of the data and show the applicability of the Central Limit Theorem for this variable.

Take the **Rings variable** analyzed:

i) Select sample of 30 random numbers and 1000 random numebers.

**rings<-mydata$RINGS**

**samplemean<-mean(rings)**

**samplesd<-sd(rings)**

**set.seed(123)**

**par(mfrow = c(1,2))**

**sample<-rnorm(30,mean=samplemean,sd=samplesd)**

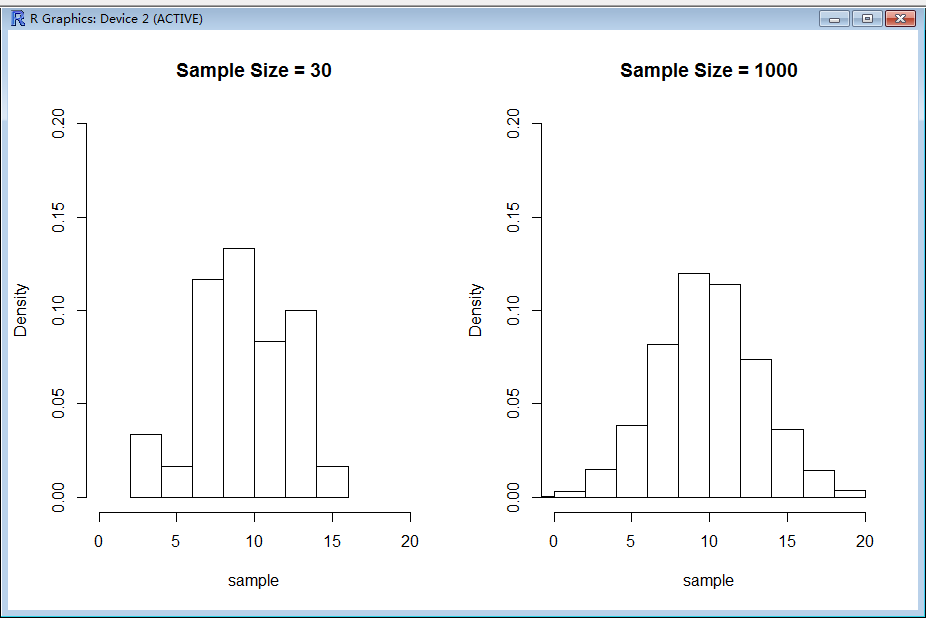
**hist(sample,prob=TRUE,**

**+xlim=c(0,20),ylim=c(0,0.2), main = paste("Sample Size =", 30))**

**sample<-rnorm(1000,mean=samplemean,sd=samplesd)**

**hist(sample,prob=TRUE,**

**+xlim=c(0,20),ylim=c(0,0.2), main = paste("Sample Size =", 1000))**

****

Notes: Draw two samples of data with sample size of 30 and 1000. The above histogram’s shape is approximate shape of normal distribution. It is centered around 9.9(mean).

ii) Applicability of Central Limit Theorem.

**samples<-4000**

**xbar<-numeric(samples)**

**par(mfrow = c(2,2))**

**for (ssize in c(10, 20, 30, 40)) {**

**+for (i in 1:samples) {**

**+ xbar[i] <- mean(rnorm(ssize,**

**+ mean = samplemean, sd = samplesd))**

**+}**

**hist(xbar, prob = TRUE,**

**+xlim=c(5,15), ylim = c(0, 0.8),**

**+main = paste("Sample Size =", ssize))**

**+ cat("Sample Size = ", ssize, " Mean = ", mean(xbar),**

**+ " SD = ", sd(xbar), "\n")**

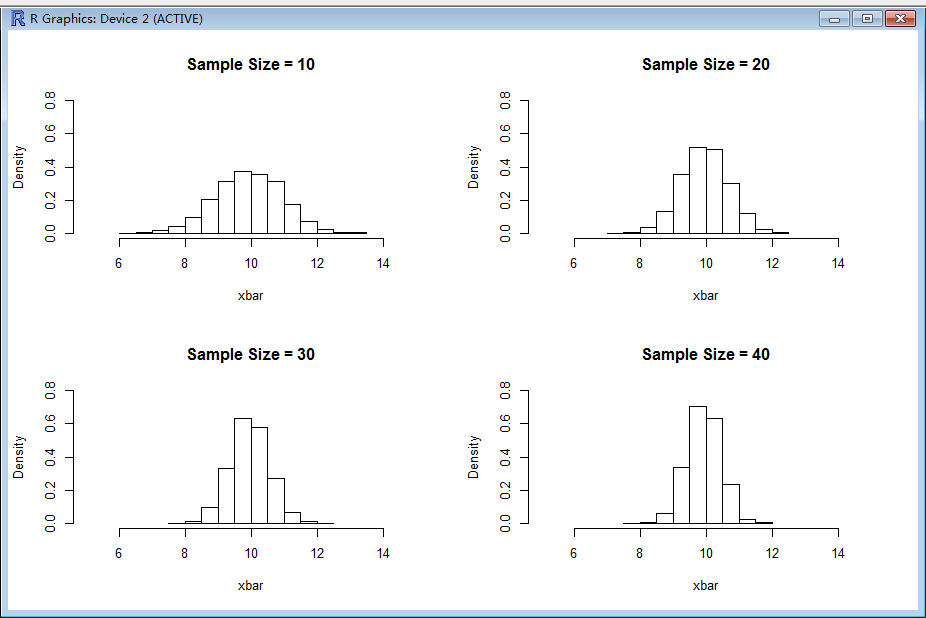
**+}**

Sample Size = 10 Mean = 9.923234 SD = 1.0098

Sample Size = 20 Mean = 9.950863 SD = 0.7147285

Sample Size = 30 Mean = 9.946053 SD = 0.5872742

Sample Size = 40 Mean = 9.929525 SD = 0.507237



Notes: we see from the above charts, the distribution of the sample means for a given sample size of the population has the shape of the normal distribution. And also, the more sample size we set, the spread of the distribution becomes narrower. And the mean for each of the sample mean distribution remains the same as the mean of the data, while the standard deviation decreases as the sample size increases.

4) Show how various sampling methods can be used on the data.

i) Simple Random Sampling

**library(sampling)**

**s<-srswor(200,nrow(mydata))**

**rows<-(1:nrow(mydata))[s!=0]**

**rows<-rep(rows,s[s!=0])**

**rows**

[1] 11 21 41 75 86 101 105 123 160 169 180 190 196

[14] 230 232 237 332 335 342 372 374 396 397 403 436 443

[27] 466 469 478 480 496 507 555 577 622 641 655 667 682

………………

**sample.1<-mydata[rows,]**

**head(sample.1,10)**

SEX LENGTH DIAMETER HEIGHT WHOLEWEIGHT SHUCKEDWEIGHT VISCERAWEIGHT

11 F 0.525 0.380 0.140 0.6065 0.1940 0.1475

21 M 0.355 0.280 0.095 0.2455 0.0955 0.0620

41 F 0.450 0.335 0.105 0.4250 0.1865 0.0910

75 F 0.605 0.450 0.195 1.0980 0.4810 0.2895

86 F 0.570 0.465 0.180 1.2950 0.3390 0.2225

101 I 0.360 0.265 0.095 0.2315 0.1050 0.0460

105 M 0.605 0.470 0.160 1.1735 0.4975 0.2405

123 F 0.515 0.425 0.140 0.7660 0.3040 0.1725

160 F 0.575 0.460 0.165 1.1240 0.2985 0.1785

169 M 0.705 0.560 0.220 1.9810 0.8175 0.3085

SHELLWEIGHT RINGS

11 0.210 14

21 0.075 11

41 0.115 9

75 0.315 13

86 0.440 12

101 0.075 7

105 0.345 12

123 0.255 14

160 0.440 13

169 0.760 14

ii) Systematic Sampling

**N<-nrow(mydata)**

**n<-100**

**k<-ceiling(N/n)**

**k**

[1] 42

**r<-sample(k,1)**

**r**

[1] 23

**s<-seq(r,by=k,length=n)**

**sample.2<-mydata[s,]**

**head(sample.2,10)**

SEX LENGTH DIAMETER HEIGHT WHOLEWEIGHT SHUCKEDWEIGHT VISCERAWEIGHT

23 F 0.565 0.440 0.155 0.9395 0.4275 0.2140

65 M 0.520 0.400 0.120 0.5800 0.2340 0.1315

107 F 0.545 0.430 0.165 0.8020 0.2935 0.1830

149 I 0.175 0.130 0.055 0.0315 0.0105 0.0065

191 M 0.615 0.480 0.180 1.1595 0.4845 0.2165

233 M 0.625 0.505 0.215 1.4455 0.4960 0.2870

275 M 0.630 0.515 0.155 1.2590 0.4105 0.1970

317 F 0.635 0.535 0.190 1.2420 0.5760 0.2475

359 M 0.745 0.585 0.215 2.4990 0.9265 0.4720

401 F 0.485 0.365 0.120 0.5885 0.2700 0.1310

SHELLWEIGHT RINGS

23 0.2700 12

65 0.1850 8

107 0.2800 11

149 0.0125 5

191 0.3250 13

233 0.4350 22

275 0.4100 13

317 0.3900 14

359 0.7000 17

401 0.1750 9

iii) Stratified Sampling

**sex<-mydata$SEX**

**table(sex)**

sex

F I M

1307 1342 1528

**st.1<-strata(mydata,stratanames=c("SEX"),**

**+size=rep(7,3),method="srswor",**

**+description=TRUE)**

Stratum 1

Population total and number of selected units: 1528 7

Stratum 2

Population total and number of selected units: 1307 7

Stratum 3

Population total and number of selected units: 1342 7

Number of strata 3

Total number of selected units 21

**st.1**

SEX ID\_unit Prob Stratum

276 M 276 0.004581152 1

517 M 517 0.004581152 1

1314 M 1314 0.004581152 1

2174 M 2174 0.004581152 1

3092 M 3092 0.004581152 1

3585 M 3585 0.004581152 1

3979 M 3979 0.004581152 1

345 F 345 0.005355777 2

861 F 861 0.005355777 2

1946 F 1946 0.005355777 2

2057 F 2057 0.005355777 2

2160 F 2160 0.005355777 2

3669 F 3669 0.005355777 2

3870 F 3870 0.005355777 2

604 I 604 0.005216095 3

1559 I 1559 0.005216095 3

1880 I 1880 0.005216095 3

2736 I 2736 0.005216095 3

3410 I 3410 0.005216095 3

3642 I 3642 0.005216095 3

3974 I 3974 0.005216095 3

**sample.3<-getdata(mydata,st.1)**

**sample.3**

LENGTH DIAMETER HEIGHT WHOLEWEIGHT SHUCKEDWEIGHT VISCERAWEIGHT

276 0.655 0.540 0.215 1.8440 0.7425 0.3270

517 0.400 0.290 0.115 0.2795 0.1115 0.0575

1314 0.550 0.430 0.150 0.8745 0.4130 0.1905

2174 0.595 0.465 0.125 0.7990 0.3245 0.2000

3092 0.495 0.380 0.120 0.4740 0.1970 0.1065

3585 0.625 0.490 0.165 1.1835 0.5170 0.2375

3979 0.485 0.390 0.120 0.5990 0.2510 0.1345

345 0.460 0.425 0.155 0.7460 0.3005 0.1520

861 0.595 0.475 0.160 1.1405 0.5470 0.2310

1946 0.635 0.505 0.155 1.2895 0.5940 0.3140

2057 0.475 0.380 0.140 0.6890 0.3165 0.1315

2160 0.640 0.510 0.200 1.3905 0.6100 0.3315

3669 0.585 0.465 0.145 0.9855 0.4325 0.2145

3870 0.490 0.385 0.160 0.6560 0.2455 0.1710

604 0.470 0.345 0.115 0.4885 0.2005 0.1080

1559 0.425 0.315 0.080 0.3030 0.1310 0.0585

1880 0.545 0.430 0.140 0.6870 0.2615 0.1405

2736 0.420 0.315 0.110 0.4025 0.1855 0.0830

3410 0.395 0.310 0.095 0.3130 0.1310 0.0720

3642 0.465 0.370 0.115 0.4075 0.1515 0.0935

3974 0.415 0.330 0.100 0.3905 0.1925 0.0755

SHELLWEIGHT RINGS SEX ID\_unit Prob Stratum

276 0.5850 22 M 276 0.004581152 1

517 0.0750 9 M 517 0.004581152 1

1314 0.2480 9 M 1314 0.004581152 1

2174 0.2300 10 M 2174 0.004581152 1

3092 0.1545 10 M 3092 0.004581152 1

3585 0.3900 11 M 3585 0.004581152 1

3979 0.1690 8 M 3979 0.004581152 1

345 0.2400 8 F 345 0.005355777 2

861 0.2710 6 F 861 0.005355777 2

1946 0.3450 11 F 1946 0.005355777 2

2057 0.1955 7 F 2057 0.005355777 2

2160 0.4100 12 F 2160 0.005355777 2

3669 0.2845 10 F 3669 0.005355777 2

3870 0.2050 9 F 3870 0.005355777 2

604 0.1660 11 I 604 0.005216095 3

1559 0.0950 7 I 1559 0.005216095 3

1880 0.2500 9 I 1880 0.005216095 3

2736 0.1015 8 I 2736 0.005216095 3

3410 0.0930 7 I 3410 0.005216095 3

3642 0.1455 9 I 3642 0.005216095 3

3974 0.1025 7 I 3974 0.005216095 3

iv) Cluster Sampling

**table(sex)**

sex

F I M

1307 1342 1528

**cl<-cluster(mydata,c("SEX"),**

**+ size=2,method="srswor")**

**sample.4<-getdata(mydata,cl)**

**table(sample.4$SEX)**

F I M

0 1342 1528

5) For confidence levels of 80 and 90, show the confidence intervals of the mean of the numeric variable for various samples and compare against the population mean.

Take the **LENGTH variable** analyzed for this question. (The LENGTH variable is normal distribution as approved)

i) Take the data from sample.1 analyzed:

**conf<-c(80,90)**

**alpha<-1-conf/100**

**alpha**

[1] 0.2 0.1

**pop.mean<-mean(mydata$LENGTH)**

**pop.sd<-sd(mydata$LENGTH)**

**sample.size<-nrow(sample.1)**

**sd.sample.means<-pop.sd/sqrt(sample.size)**

**sd.sample.means**

[1] 0.008491851

**xbar<-mean(sample.1$LENGTH)**

**xbar**

[1] 0.522525

**for(i in alpha){**

**+ str<-sprintf("%2d%% Conf Level (alpha=%.2f), CI=%.3f-%.3f",**

**+ 100\*(1-i),i,**

**+ xbar-qnorm(1-i/2)\*sd.sample.means,**

**+ xbar+qnorm(1-i/2)\*sd.sample.means)**

**+ cat(str,"\n")**

**+ }**

80% Conf Level (alpha=0.20), CI=0.512-0.533

90% Conf Level (alpha=0.10), CI=0.509-0.536

**pop.mean**

[1] 0.5239921

As can be seen, the two confidence intervals both contain the population mean.

ii) Take the data from sample.2 analyzed:

**sample.size<-nrow(sample.2)**

**sd.sample.means<-pop.sd/sqrt(sample.size)**

**sd.sample.means**

[1] 0.01200929

**xbar<-mean(sample.2$LENGTH[!is.na(sample.2$LENGTH)])**

**xbar**

[1] 0.514596

**for(i in alpha){**

**+ str<-sprintf("%2d%% Conf Level (alpha=%.2f), CI=%.3f-%.3f",**

**+ 100\*(1-i),i,**

**+ xbar-qnorm(1-i/2)\*sd.sample.means,**

**+ xbar+qnorm(1-i/2)\*sd.sample.means)**

**+ cat(str,"\n")**

**+ }**

80% Conf Level (alpha=0.20), CI=0.499-0.530

90% Conf Level (alpha=0.10), CI=0.495-0.534

**pop.mean**

[1] 0.5239921

As can be seen, the two confidence intervals both contain the population mean.

iii) Take the data from sample.3 analyzed:

**sample.size<-nrow(sample.3)**

**sd.sample.means<-pop.sd/sqrt(sample.size)**

**sd.sample.means**

[1] 0.02620642

**xbar<-mean(sample.3$LENGTH[!is.na(sample.3$LENGTH)])**

**xbar**

[1] 0.5152381

**for(i in alpha){**

**+ str<-sprintf("%2d%% Conf Level (alpha=%.2f), CI=%.3f-%.3f",**

**+ 100\*(1-i),i,**

**+ xbar-qnorm(1-i/2)\*sd.sample.means,**

**+ xbar+qnorm(1-i/2)\*sd.sample.means)**

**+ cat(str,"\n")**

**+ }**

80% Conf Level (alpha=0.20), CI=0.482-0.549

90% Conf Level (alpha=0.10), CI=0.472-0.558

**pop.mean**

[1] 0.5239921

As can be seen, the two confidence intervals both contain the population mean.

iv) Take the data from sample.4 analyzed:

**sample.size<-nrow(sample.4)**

**sd.sample.means<-pop.sd/sqrt(sample.size)**

**sd.sample.means**

[1] 0.002333331

**xbar<-mean(sample.4$LENGTH[!is.na(sample.4$LENGTH)])**

**xbar**

[1] 0.5024198

**for(i in alpha){**

**+ str<-sprintf("%2d%% Conf Level (alpha=%.2f), CI=%.3f-%.3f",**

**+ 100\*(1-i),i,**

**+ xbar-qnorm(1-i/2)\*sd.sample.means,**

**+ xbar+qnorm(1-i/2)\*sd.sample.means)**

**+ cat(str,"\n")**

**+ }**

80% Conf Level (alpha=0.20), CI=0.499-0.505

90% Conf Level (alpha=0.10), CI=0.499-0.506

**pop.mean**

[1] 0.5239921

As can be seen, the two confidence intervals do not contain the population mean.