STA304H1F/1003HF Fall 2018 Assignment

RyanYip

2018骞10鏈<88>25鏃

# Exercise 1

## Load data

certify=read.csv("./certify.csv",header = TRUE)  
summary(certify)

## certify approve speccert wouldyou   
## Min. :0.000 Min. :0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000 1st Qu.:1.000   
## Median :3.000 Median :3.000 Median :5.000 Median :2.000   
## Mean :3.078 Mean :3.252 Mean :3.687 Mean :2.803   
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:5.000   
## Max. :5.000 Max. :5.000 Max. :5.000 Max. :5.000   
##   
## recert subdisc college employ workenv workact   
## Min. :0.000 BI : 940 B: 234 E:3943 A:2005 C: 985   
## 1st Qu.:2.000 BP : 518 M:1515 I: 166 G: 826 E: 862   
## Median :3.000 BE : 501 N: 15 O: 1 I:1716 O: 7   
## Mean :3.275 PE : 448 O: 2 R: 181 O: 1 P:1030   
## 3rd Qu.:5.000 QP : 326 P:2782 S: 219 R:1515   
## Max. :5.000 CM : 310 U: 38 S: 149   
## (Other):1505   
## yearsmem   
## Min. : 0.00   
## 1st Qu.: 5.00   
## Median :12.00   
## Mean :13.62   
## 3rd Qu.:20.00   
## Max. :62.00   
##

\*\* Summary \*\* function show each variable's value in \* Min,1st Qu,Median,3rd qu,Max \*

## (a)

set.seed(20181025) # set seed  
  
#Take a simple random sample without replacement of size 100  
idx=sample(1:dim(certify)[1],size = 100,replace = FALSE)   
sub.certify=certify[idx,]  
  
  
  
#estimate the proportion of respondents who think that ASA should develop some form of certification (YES=1)  
n\_yes=sum(sub.certify$certify==1)  
prob=n\_yes/dim(sub.certify)[1]  
sprintf("the proportion  
of respondents who think that ASA should develop some form of certification is : %f",prob)

## [1] "the proportion\nof respondents who think that ASA should develop some form of certification is : 0.230000"

## (b)

* 1. the sample proportion:p\_hat=x/n
  2. z-score:(p\_hat-p)/sqrt(pq/n),where q=1-p
  3. E=z\_alpha\_2\* sqrt(pq/n)
  4. **alpha** confidence interval: p\_hat-E<p<p\_hat+E,and E=z\_alpha\_2\*sqrt(p\_hat x q\_hat/n)

alpha=1-0.95  
n=dim(sub.certify)[1]  
p\_hat=n\_yes/n  
q\_hat=1-p\_hat  
  
z\_alpha\_2=qnorm(p=1-alpha/2)  
E=z\_alpha\_2\*sqrt(p\_hat\*q\_hat/n)  
  
var\_p\_hat=p\_hat\*q\_hat/n # var(p\_hat),p\_hat variance  
lower=p\_hat-E # lower confidence interval  
upper=p\_hat+E # upper confidence interval

Follow the above analysis,the variance p\_hat is 0.001771,the 95% confidence interval is (0.1475183,0.3124817)

## (c)

We have know that,\*\* E=z\_alpha\_2 x sqrt(p\_hat\*q\_hat/n) \*\*,in this,n(to be estimated) is not know,the p\_hat be estimated in part(a).part(c) tells us that B=5%.

Get: 0.025=E,where p\_hat,q\_hat,z\_alpha we can know,so n=1/(0.025/z\_alpha\_2)\*\*2 x (p\_hat x q\_hat)

B=0.05  
z\_alpha=qnorm(1-B/2)  
  
n\_hat=ceiling(1/(0.025/z\_alpha\_2)\*\*2 \* (p\_hat \* q\_hat))  
sprintf("the sample size n be estimated is about:%f",n\_hat)

## [1] "the sample size n be estimated is about:1089.000000"

So we would need a sample of about 1100 sample.

## (d)

Recalculate parts (a) and (b) using the sample size *n\_hat* 1089 obtained in part (c)

### recaculate part(a)

set.seed(20181025) # set seed  
  
#Take a simple random sample without replacement of size 100  
idx=sample(1:dim(certify)[1],size = n\_hat,replace = FALSE)   
sub.certify=certify[idx,]  
  
  
  
#estimate the proportion of respondents who think that ASA should develop some form of certification (YES=1)  
n\_yes=sum(sub.certify$certify==1)  
prob=n\_yes/dim(sub.certify)[1]  
sprintf("the proportion of respondents who think that ASA should develop some form of certification is : %f",prob)

## [1] "the proportion of respondents who think that ASA should develop some form of certification is : 0.265381"

### recaculate part(b)

alpha=1-0.95  
n=dim(sub.certify)[1]  
p\_hat=n\_yes/n  
q\_hat=1-p\_hat  
  
z\_alpha\_2=qnorm(p=1-alpha/2)  
E=z\_alpha\_2\*sqrt(p\_hat\*q\_hat/n)  
  
var\_p\_hat=p\_hat\*q\_hat/n # var(p\_hat),p\_hat variance  
lower=p\_hat-E # lower confidence interval  
upper=p\_hat+E # upper confidence interval  
sprintf("the confidence interval range is :%f",upper-lower)

## [1] "the confidence interval range is :0.052448"

Follow the above analysis,the variance p\_hat is 1.790210910^{-4},the 95% confidence interval is (0.239157,0.2916052),the confidence interval range is 0.0524482

# Exercise 2

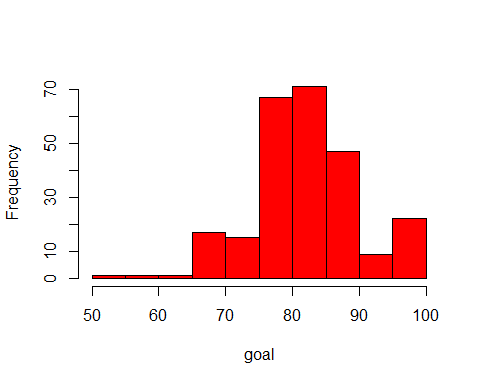
## Load data

goals=read.csv("./goals.csv",header = TRUE)  
summary(goals)

## index goal   
## Min. : 1.0 Min. : 50.00   
## 1st Qu.: 63.5 1st Qu.: 80.00   
## Median :126.0 Median : 85.00   
## Mean :126.0 Mean : 84.12   
## 3rd Qu.:188.5 3rd Qu.: 90.00   
## Max. :251.0 Max. :100.00

## (a)

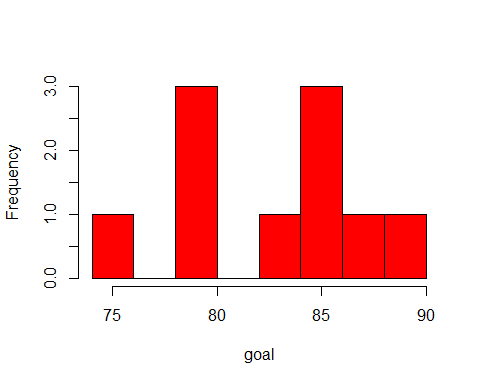
goal\_mean=mean(goals$goal) # cacualate population mean  
goal\_var=var(goals$goal) # caculate population variance  
  
hist(goals$goal,xlab = "goal",main=NULL,col="red") # Draw a histogram



The population mean is 84.1155378,population variance is 64.3665976.and the histogram show us that the center of goal is about 80,proportion of lower goal(50,60) is lower.

## (b)

set.seed(20181026)  
goal=goals$goal  
  
idx\_raw=sample(1:length(goal),size=10,replace = FALSE)  
goal\_1=goal[idx\_raw]# Select a simple random sample of 10 grades.  
  
goal\_1\_mean=mean(goal\_1)  
goal\_1\_var=var(goal\_1)  
  
hist(goal\_1,xlab = "goal",main=NULL,col="red",breaks = 10) # Draw a histogram

 the sample mean is 83.1,sample variance is 19.6555556

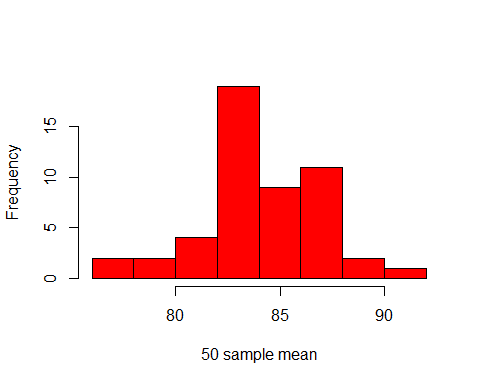
## (c)

#### Repeat part (b) 49 additional times and combine to have 50 sample means

goal\_sample=data.frame(sample\_1=goal\_1)  
idx\_raw=data.frame(idx\_1=idx\_raw)  
RepeatSample<-function(raw=goal\_sample,idx=idx\_raw,size=10,times=49){  
 for(i in 1:times){  
 idx=sample(1:length(goal),size=size,replace = FALSE)  
 sample\_x=data.frame(goal[idx])  
 sample\_idx=data.frame(idx)  
   
 colnames(sample\_x)=paste0("sample\_",i+1)  
 colnames(sample\_idx)=paste0("idx\_",i+1)  
 raw=cbind.data.frame(raw,sample\_x)  
 idx\_raw=cbind.data.frame(idx\_raw,sample\_idx)  
 }  
   
 return(list(raw,idx\_raw))  
}  
  
goal\_sample=RepeatSample(raw = goal\_sample)

#### Draw a histogram to describe your 50 sample means.

goal\_sample\_mean=apply(goal\_sample[[1]],2,mean) # 50 sample mean  
goal\_sample\_var=apply(goal\_sample[[1]],2,var) # 50 sample variance  
  
goal\_sample\_index=goal\_sample[[2]] # the index positions of your 50 samples  
hist(goal\_sample\_mean,xlab = "50 sample mean",main=NULL,col="red") # Draw a histogram



#### Now find the sample mean and sample variance of the 50 sample means

sprintf("the sample mean of the 50 sample means : %f ",mean(goal\_sample\_mean))

## [1] "the sample mean of the 50 sample means : 84.244000 "

sprintf("the sample variance of the 50 sample means : %f ",var(goal\_sample\_mean))

## [1] "the sample variance of the 50 sample means : 8.233943 "

Based on th above result,the mean and variance in

* part(a) : mean :84.1155378--- variance :64.3665976,
* part(b) : mean :83.1--- variance19.6555556
* part(c)(here) : mean : 84.244---variance 8.2339429

Compare with histogram, mean and variance above these,the mean across them have small difference,but the variance differ greatly.Variance(part c) < Variance(part b) < Variance (part c)

## (d)

According to the result in part(c),I consider that the mean and variance of the random sample means are two potential stratification variables for grade goal.In part(c),Comparison shows us that sample mean distribution has more lower variance and centers,which represents more accurate sample. Comparison to simple random sampling,this sample method has more lower variance ,but loss more randomness and representativeness.