STA304H1F/1003HF Fall 2018 Assignment # 3

### 1)

#### (a)

Strata sample sizes are determined by the following equation :

* nh=(Nh/N)\*n

where nh is the sample size for stratum h, Nh is the population size for stratum h, N is total population size, and n is total sample size.And the **nh=5** by caculated

I randomly selected 5 player in each team

#### (b)

The result of b above tell us that,the mean of the variable *logsal = ln(salary)* is 19.9211216,and the a 95% CI is (19.6384173,20.2038258)

#### (c)

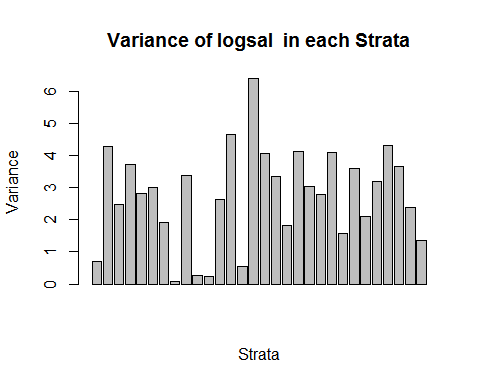
The estimated proportion of players in the data set of stratified sample who are pitchers is 0.5466667,and the 95% CI is (0.4670007,0.6263326)

#### (d)

The *estimated proportion* of players in the data set of a simple random sample who are pitchers is 0.42,and the 95% CI is (0.3410156,0.4989844)

Compared with that of (c),the length of the CI in part (c) is 0.1593318,and the in part (d),the length of the CI is 0.1579687.The proportion in CI of simple random sample is smaller than in stratified sample; and the *estimated porportion* here is smaller than in part(c).

#### (e)

 According to barplot, the difference in variance of *logsal* between each Strata is relatively large,that is ,the fluctuation of *logsal* in each Strata is relatively large.I think optimal allocation would be worthwhile for this problem.it can reduce the cost of Strata sampling beacuse of variance in Statum.

#### (f)

According tho the formule from the text book,we can get the estimated population stratum variance below.

The equation for Neyman allocation can be derived from the equation for optimal allocation by assuming that the direct cost to sample an individual element is equal across strata. Based on Neyman allocation, the best sample size for stratum h would be:

* nh=n\*(Nh\*sigma\_h)/(sum(Ni\*sigma\_h))

where nh is the sample size for stratum h, n is total sample size, Nh is the population size for stratum h, and σh is the standard deviation of stratum h.

So,according to the result above,the estimate the proportional allocation population stratum variances is 0.0146508,and the sample size in each stratum respectively is :3, 7, 5, 6, 5, 5, 5, 1, 6, 2, 2, 5, 7, 2, 7, 6, 6, 5, 7, 5, 5, 6, 4, 6, 5, 6, 7, 6, 5, 4.The optimal allocation stratum variances is 0.0115837,we can know that ,the variance of allocation stratum variances is always smaller than proportional allocation population stratum variances otr equal.

### 2)

## 'data.frame': 251 obs. of 2 variables:  
## $ height : num 158 178 165 165 173 160 190 180 170 178 ...  
## $ handspan: num 19 20.5 21 17.5 13 18 20.3 20.5 22 17 ...

#### (a)

## Parameter 0.1159228   
## Variance Parameter 0   
## Confidence Interval: [ 0.1159228 ; 0.1159228 ]

## A SRS estimator of the population mean handspan is: 17.99

## A ratio estimator of the population mean handspan is: 19.3081

## A regression-based estimator of the population mean handspan is: 19.37006

#### (b)

## SSR error of estimator is : 1.736693

## Ratio error of estimator is : 0.4185967

## Regression error of estimator is: 0.3566356

We can know the error of estimation respectively,and the error of SSR is biggest,Ratio estimator's secondly,and Regression's is smallest.

#### (c)

## The variance of SSR estimator is : 9.158778

## The variance of Ratio estimator is : 1.015144

## The variance of regression-based estimator is : 0.4285674

Accorting to the comparsion,we can know that the variance of SSR is biggest,and there is litte difference between ratio estimator and regression-based estimator.

### 3)

#### (a)

According to the above analysis,we can get the estimated average sales for the week for all supermarkets in the area is 97.9727891,and the a bound on the error of the estimation is (-12852.22,13048.17).This estimation is biased.

#### (b)

## The estimation of total number of boxes of cereal sold by all supermarkets in the area is : 2361.8

Yes,I have enough information to get the estimation.According to the equation:

* mu.sample/mu.total =n /M

So,the estimation can be caculated.

And the a bound on the error of the estimation is (-17446.06,22169.66)

#### Appendix

Code:

### 1)

library(sampling)

baseball<-read.csv("./baseball.csv",head=TRUE)

#summary(baseball)

baseball<-na.omit(baseball) # remove NA sample

baseball<-baseball[!duplicated(baseball),] # remove duplicate sample

#### a)

set.seed(2018)

st<-sampling::strata(data=baseball,stratanames = c("team"),size = rep(5,30),

method="systematic",pik=as.integer(baseball$player))

baseball.strata<-getdata(baseball,st)

#### b)

n<-length(log2(baseball.strata$salary))

mean.salary<-mean(log2(baseball.strata$salary))

salary.var<-var(log2(baseball.strata$salary))

se<-sqrt(salary.var/n)

t.alpha<-qt(1-0.05/2,df=n-1)

# confidence interval

lower.salary<-mean.salary-se\*t.alpha

upper.salary<-mean.salary+se\*t.alpha

#### c)

phat<-sum(baseball.strata$position=="P")/n

p.se<-sqrt(phat\*(1-phat)/n)

z\_alpha<-qnorm(1-0.05/2)

lower.p<-phat-z\_alpha\*p.se

upper.p<-phat+z\_alpha\*p.se

#### d)

idx<-sample(1:dim(baseball)[1],size = n,replace = FALSE) # s simple random sample

baseball.simple<-baseball[idx,]

# repeat (c)

phat.simple<-sum(baseball.simple$position=="P")/n

p.se.simple<-sqrt(phat.simple\*(1-phat.simple)/n)

z\_alpha<-qnorm(1-0.05/2)

lower.p.simple<-phat.simple-z\_alpha\*p.se.simple

upper.p.simple<-phat.simple+z\_alpha\*p.se.simple

#### e)

caculate.var.strata<-function(strata.object){

stratum<-unique(strata.object$team)

variance.set<-unlist(lapply(stratum,function(x){

df<-as.vector(strata.object[strata.object$team==x,"salary"])

if(length(df)==1){

var\_i<-0.0

}else{

var\_i<-var(log2(df))

}

return(var\_i)

}))

df<-data.frame(stratum=stratum,variance=variance.set)

return(df)

}

strata.var<-caculate.var.strata(strata.object = baseball.strata)

barplot(strata.var$variance,main="Variance of logsal in each Strata",xlab = "Strata",ylab = "Variance")

#### f)

Ni<-as.numeric(table(baseball$team)) # number of population size of each stratum

N<-sum(Ni) #total population

ni<-5 #number of sample size of each stratum

L<-30 # Number of stratum

variance.set<-strata.var$variance

V<-unlist(lapply(1:length(variance.set),function(i){

x<-Ni[i]^2\*(1-ni/Ni[i])\*variance.set[i]/ni

}))

V<-sum(V)/(N^2) # population stratum variances

n<-150

Nh<-Ni

N<-sum(Nh)

nh<-n\*(Nh\*sqrt(variance.set))/sum(Nh\*sqrt(variance.set))

st.opt<-sampling::strata(data=baseball,stratanames = c("team"),size =nh,

method="systematic",pik=as.integer(baseball$player))

strata.opt.var<-caculate.var.strata(strata.object = getdata(baseball,st.opt))

variance.opt.set<-strata.opt.var$variance

opt.V<-unlist(lapply(1:length(variance.opt.set),function(i){

x<-Nh[i]^2\*(1-nh[i]/Nh[i])\*variance.opt.set[i]/nh[i]

}))

opt.V<-sum(opt.V)/(N^2) # estimate population stratum variances

### 2)

hh18<-read.csv("./hh18.csv",head=TRUE)

##### a function to calculate ratio estimators

ratio.srs <- function(x, y, opt="Ratio", tauX=NA, N=NA) {

# opt = "Tau" for the total of Y

# opt = "Mu" for the mean of Y

n <- length(x)

if(is.na(N)) {fpc <- 1} else {fpc <- 1-(n/N)}

ratio <- sum(y)/sum(x)

if(is.na(tauX) & is.na(N)) {meanX <- mean(x)} else {meanX <- tauX/N}

var.r <- fpc\*(1/meanX^2)\*(sum((y-ratio\*x)^2)/n\*(n-1))

switch(opt,

"Ratio" = {theta <- ratio

var.theta <- var.r},

"Tau" = {theta <- ratio\*tauX

var.theta <- var.r\*tauX^2},

"Mu" = {theta <- ratio\*meanX

var.theta <- var.r\*meanX^2}

)

B <- 2\*sqrt(var.theta)

cat("Parameter",theta,"\n")

cat("Variance Parameter",var.theta,"\n")

cat("Confidence Interval: ","[",theta-B,";",theta+B,"]","\n")

return(theta)

}

#### a)

seed<-set.seed(2018)

N<-dim(hh18)[1]

n<-10

sample.idx<-sample(1:N,size = n)

sample.hh18<-hh18[sample.idx,]

# SSR estimator

ssr.se<-sqrt(var(sample.hh18$handspan)/(n-1))

t.alpha<-qt(1-0.05/2,df=n-1)

ssr.mean<-mean(sample.hh18$handspan)

lower.ssr<-ssr.mean-ssr.se\*t.alpha

upper.ssr<-ssr.mean+ssr.se\*t.alpha

# Ratio etsimator

rhat<-ratio.srs(hh18$height,hh18$handspan,opt = "Ratio",tauX = sum(hh18$height),N=N)

# regression estimator

LR<-lm(handspan~height,data = hh18)

sample.pred<-predict(LR,newdata = data.frame(height=sample.hh18$height))

lr.mean<-mean(sample.pred)

ratio.mean<-rhat\*mean(sample.hh18$height)

cat("A SRS estimator of the population mean handspan is:",ssr.mean,"\n")

cat("A ratio estimator of the population mean handspan is:",ratio.mean,"\n")

cat("A regression-based estimator of the population mean handspan is:",lr.mean,"\n")

#### b)

ssr.error<-abs(ssr.mean-mean(hh18$handspan))

ratio.error<-abs(ratio.mean-mean(hh18$handspan))

lr.error<-abs(lr.mean-mean(hh18$handspan))

cat("SSR error of estimator is :",ssr.error,"\n")

cat("Ratio error of estimator is :",ratio.error,"\n")

cat("Regression error of estimator is:",lr.error,"\n")

#### c)

ssr.var<-var(sample.hh18$handspan)

ratio.var<-(1-n/N)\*sum((sample.hh18$handspan-rhat\*sample.hh18$height)^2)/(n\*(n-1))

mse<-sum(residuals(LR)^2)/(N-1)

lr.var<-((N-n)/(N\*n))\*mse

cat("The variance of SSR estimator is :",ssr.var,"\n")

cat("The variance of Ratio estimator is :",ratio.var,"\n")

cat("The variance of regression-based estimator is :",lr.var,"\n")

### 3)

#### a)

total.y<-102+90+76+94+120

n<-5 # number of sample in sample city

M<-45+36+20+18+28 # total number of sample city

N<-20

m<-9+7+4+4+6

y.total.hat<-total.y\*M/30

#cat("The estimation average sales for the week for all supermarkets in the area is :",yhat,"\n")

# step 1

mu.r.hat<-sum(45\*102+36\*90+20\*76+18\*94+28\*120)/sum(45+36+20+18+28) # estimate average sales

# step 2

s.r<-sum((45\*(102-mu.r.hat))^2+(36\*(90-mu.r.hat))^2+(20\*(76-mu.r.hat))^2+(18\*(94-mu.r.hat))^2+(28\*(120-mu.r.hat))^2)/(n-1)

# step 3

M.bar<-M/n

# Step 4

sum.x<-45^2\*(1-9/45)\*(20/9)+36^2\*(1-7/36)\*(16/7)+20^2\*(1-4/20)\*(22/4)+18^2\*(1-4/18)\*(26/4)+28^2\*(1-6/28)\*(12/6)

variance.mu.r<-(1-n/N)\*(1/(n\*M.bar^2))\*s.r^2+(1/n\*N\*M.bar^2)\*sum.x

# Step 5

lower.mu.r<-mu.r.hat-2\*sqrt(variance.mu.r)

upper.mu.r<-mu.r.hat+2\*sqrt(variance.mu.r)

#### b)

total.y<-102+90+76+94+120

y.total.hat<-total.y\*M/30 # estimate total size

cat("The estimation of total number of boxes of cereal sold by all supermarkets in the area is :",y.total.hat,"\n")

# Step 1

m.bar<-m/n

M.bar<-M/N

f1<-n/N

f2<-M/N

# Step 2

s.b<-(1/(n-1))\*((45\*102-M\*mu.r.hat/n)^2+(36\*90-M\*mu.r.hat/n)^2+(20\*76-M\*mu.r.hat/n)^2+

(18\*94-M\*mu.r.hat/n)^2+(28\*120-M\*mu.r.hat/n)^2)

# Step 3

s.w<-(20+16+22+26+12)/n

variance.y<-N^2\*(1-f1)/n\*s.b+N^2\*M.bar^2\*(1-f2)/(m.bar\*n)\*s.w

lower.y<-y.total.hat-2\*sqrt(variance.y)

upper.y<-y.total.hat+2\*sqrt(variance.y)