BST 542 Sampling Theory and Survey Design in Public Health
Homework 3
Due Tuesday, March 5, 2019

Note that the points total 29, but the assignment is worth 10 points. Your homework score will be recorded as your score out of 29, divided by 2.9, rounded to the nearest half point.

1. (11 Points: 2,3,3,3) Researchers are interested in the proportion of drivers in the city who use their cell phone by holding it with their hands while driving (i.e., not using a blue-tooth, or hands-free device). Investigators chose a busy intersection, and observed every sixth car that passed going a particular direction. It was suggested that they look at *every* car and record whether the driver was using the cell phone, but often cars went by too fast and they could misidentify some drivers. So, the plan was to dispatch two investigators, one of whom counted the cars and chose every sixth car and the other looked carefully to see whether the driver was talking on the cell phone. They did this for two hours and observed 300 drivers.

(a) What kind of a sample is this?

This is a systematic sample with k = 6.

(b) Is it a reasonable sampling plan for determining the rate of cell phone usage while driving at that intersection at that time?

For this intersection at this time (time of day and day of the week) this is a reasonable sampling plan. Systematic sampling works poorly when there is a cycle in the data and the period of the cycle agrees with the sampling interval. We wouldn't expect a cycle in the data of cars that pass by.

(c) Is this a reasonable plan for determining the rate of cell phone usage while driving for the city as a whole?

This only covers one intersection at one time, so it may not be representative of the city as a whole.

(d) Can you describe a better sampling plan? Explain.

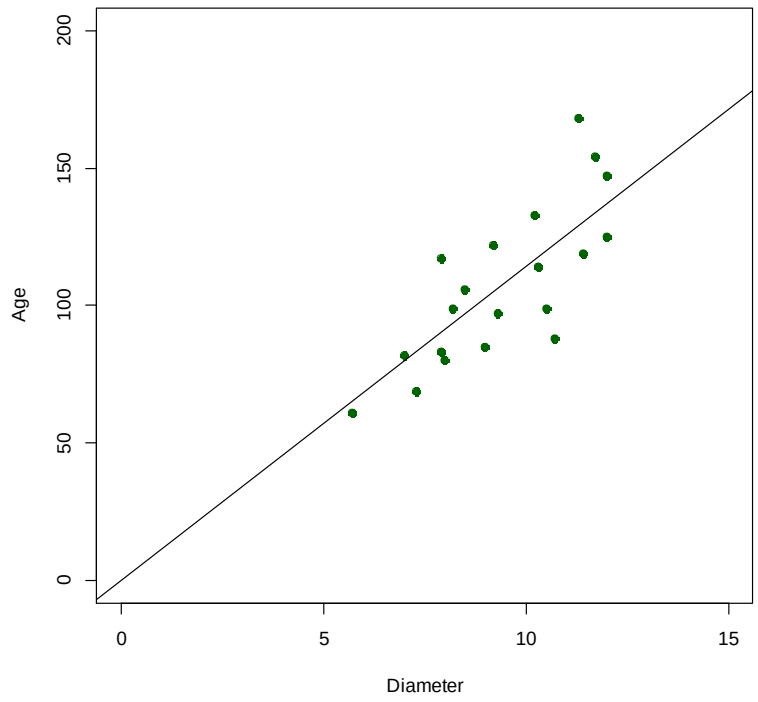
A better plan might be to sample intersections in the city, and sample times of the day and day of the week. Then apply the systematic sample at each selected intersection/time. This would be a systematic sample within a cluster sample.

2. (8 Points: 4,4) Do Problem 4.3 in the textbook. This problem begins "Foresters want to estimate …" Do only parts (a) and (b).

Here is R code for this problem.

```
Diameter = c(12.0, 11.4,  7.9,  9.0, 10.5,  7.9,  7.3, 10.2, 11.7, 11.3,
             5.7,  8.0, 10.3, 12.0,  9.2,  8.5,  7.0, 10.7,  9.3,  8.2 )
Age = c(125, 119,  83,  85,  99, 117,  69, 133, 154, 168,
         61,  80, 114, 147, 122, 106,  82,  88,  97,  99)
x = Diameter
y = Age
N = 1132
n = length(x)
mux = 10.3

ybar = mean(Age)
xbar = mean(Diameter)
Bhat = ybar/xbar
muyhat = Bhat*mux

yhat = Bhat*Diameter

plot( Diameter , Age , pch=19 , col="darkgreen" , xlim=c(0,15) , ylim=c(0,200) )
  abline( a=0 , b=Bhat )

resid = y - yhat
SSE = sum( resid^2 )
se2 = SSE/(n-1)
Vhat = (1-n/N) * (mux/xbar)^2 * (se2/n)
semu = sqrt(Vhat)
print( c(muyhat,semu) )
```

(a) The scatter plot is shown below.



(b) The point estimate for the mean age of trees is 117.620415. The standard error and margin of error are 4.354872 8.709744, respectively.

3. (10 Points) Do Problem 5.7 in the textbook. The problem begins "The new candy Green …". Be sure to estimate the total number of cases and the average number sold per store.

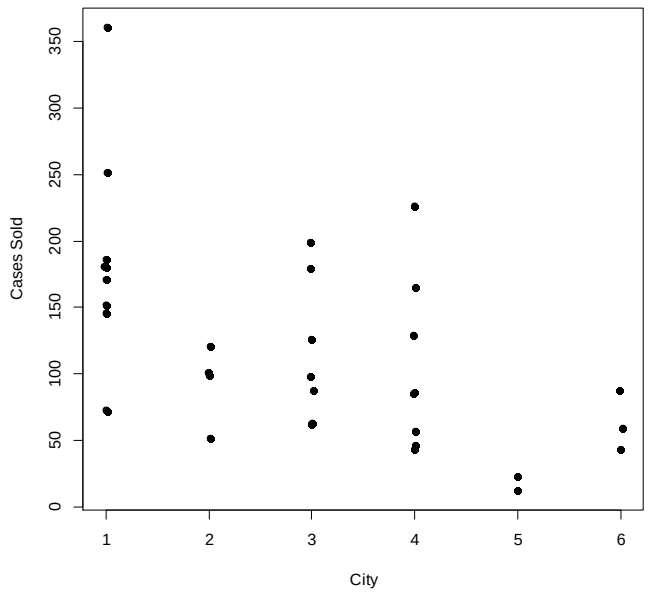
Here is R code for this problem.

```
MM = c(52,19,37,39,8,14)
m  = c(10,4,7,8,2,3)

c1 = c(146,180,251,152,72,181,171,361,73,186)
c2 = c(99,101,52,121)
c3 = c(199,179,98,63,126,87,62)
c4 = c(226,129,57,46,86,43,85,165)
c5 = c(12,23)
c6 = c(87,43,59)
cases = c(c1,c2,c3,c4,c5,c6)
tau = c( sum(c1) , sum(c2) , sum(c3) , sum(c4) , sum(c5) , sum(c6) )

n = 6
N = 45
ybar = rep( 0 , n )
stdev = rep( 0 , n )

ybar[1] = mean(c1);   stdev[1] = sd(c1);
ybar[2] = mean(c2);   stdev[2] = sd(c2);
ybar[3] = mean(c3);   stdev[3] = sd(c3);
ybar[4] = mean(c4);   stdev[4] = sd(c4);
ybar[5] = mean(c5);   stdev[5] = sd(c5);
ybar[6] = mean(c6);   stdev[6] = sd(c6);

summaryStats = cbind( 1:6 , ybar , stdev )
print( summaryStats )

Mbar = mean(MM)
EstimateOfTotal = (N/n)*sum(MM*ybar)
EstimateOfMean = EstimateOfTotal/(N*Mbar)

tauhat = MM*ybar
x1 = (1-m/MM)*MM^2*stdev^2/m
xt = (tauhat - EstimateOfTotal/N)^2
xr = (MM*ybar - MM*EstimateOfMean)^2

s1  = sum(x1)
st2 = sum(xt) / (n-1)
sr2 = sum(xr)

vhatTotal = N^2*(1-n/N)*st2/n  +  (N/n)*sum(x1)
vhatMu = (1/Mbar^2)*(1-n/N)*sr2/n  +  (1/(Mbar^2*N*n))*sum(x1)
```

```
seTotal = sqrt(vhatTotal)
print( c( EstimateOfTotal , seTotal ) )

seMu = sqrt(vhatMu)
print( c( EstimateOfMean , seMu ) )

city = c( rep(1,length(c1)) , rep(2,length(c2)) , rep(3,length(c3)) ,
          rep(4,length(c4)) , rep(5,length(c5)) , rep(6,length(c6)) )
plot( jitter(city,0.1) , cases , pch=19 , xlab="City" , ylab="Cases Sold"
)
```

Here are the summary statistics for each city.

City	ybar	stdev
1	177.3000	83.599641
2	93.2500	29.238958
3	116.2857	54.539632
4	104.6250	64.593095
5	17.5000	7.778175
6	63.0000	22.271057



The point estimate for the total number of cases sold is 152972.22 with a standard error of 56781.08.

The point estimate for the mean number of cases sold per store is 120.68814 with a standard error of 44.26141.