HW 3

Ishita Dutta

5/25/2022

Start

P1:

(a)

- Ratio Estimation - Since we want to estimate the proportion of time

(b)

- Regression Estimation - When identifying the relationship between variables regression estimation is better unless the origin is passing. However, it seems hard to do the survey of average number of fish caught in August. Therefore, ratio estimation seems good as also.

(c)

- Regression Estimation - When identifying the relationship between variables regression estimation is better unless the origin is passing. In this case surveying the undergraduate students seems not that hard.

(d)

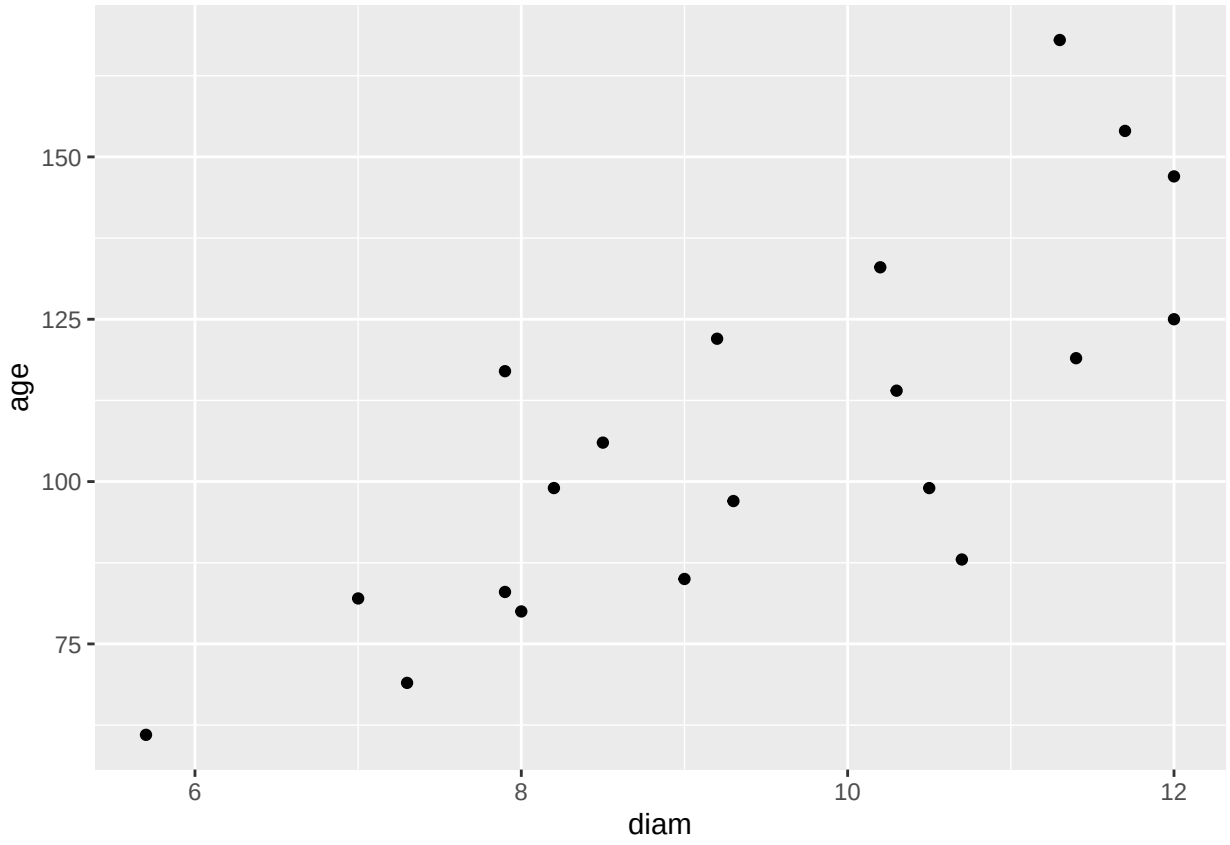
- Regression Estimation

P2:

```
# Noting the data
tree_num = c(1:20)
diam = c(12, 11.4, 7.9, 9, 10.5, 7.9, 7.3, 10.2, 11.7, 11.3, 5.7, 8, 10.3, 12, 9.2, 8.5, 7, 10.7, 9.3, 8
age = c(125, 119, 83, 85, 99, 117, 69, 133, 154, 168, 61, 80, 114, 147, 122, 106, 82, 88, 97, 99)
facval = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
N = 1132
n = 20
tab = data.frame(tree_num, diam, age, facval)
```

a)

```
ggplot(data=tab, mapping=aes(x=diam, y=age)) +
  geom_point()
```



b)

```
mu_diam = sum(diam) / n
mu_age = sum(age)/n
s_diam = sqrt(sum((diam - mu_diam)^2)/(n - 1))
s_age = sqrt(sum((age - mu_age)^2)/(n - 1))
B_hat = mu_age/mu_diam
yrat = B_hat * 10.3     #10.3 is the population diameter.
round(c(10.3, yrat), 1)
```

```
## [1]  10.3 117.6
```

```
se_2 = (sum(age - (diam * B_hat)) ^ 2) / (n - 1)
V_drat = (1 - (n/N)) * ((10.3 / mu_diam)^2) * (se_2  / n)
sqrt(V_drat)
```

```
## [1] 1.503452e-14
```

First output is the coordinates for the ratio estimate. The second output is the SE for y-bar under ratio estimation.

c)

```
s_diam_age = (sum((diam - mu_diam) * (age - mu_age))) / (n - 1)
r = s_diam_age / (s_diam ^ 2)
bhat1 = (r * s_age) / s_diam
bhat0 = mu_age - (bhat1 * mu_diam)
yreg = bhat0  + (bhat1 * 10.3)
round(c(10.3, yreg), 1)
```

```
## [1]  10.3 279.2
```

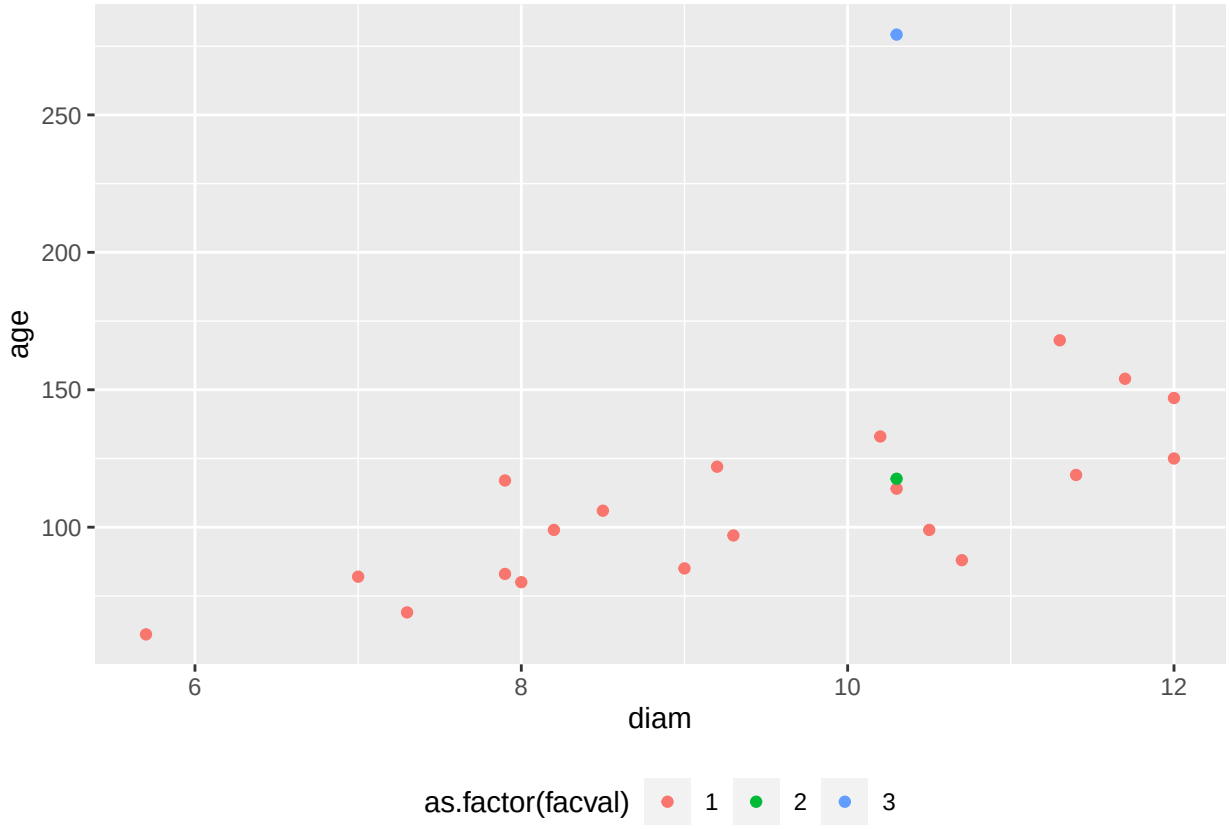
```
di = diam - (yreg + (bhat1 * (age - 10.3)))
mu_di = sum(di) / n
s_di = sqrt(sum((di - mu_di)^2)/(n - 1))
V_dreg = (1 - (n/N)) *  ((s_di ^ 2)/ n)
sqrt(V_dreg)
```

```
## [1] 1219.164
```

First output is the coordinates for the regression estimate. The second output is the SE for y-bar under regression estimation.

d)

```
tab = rbind(tab, c(21, 10.3, yrat, 2), c(22, 10.3, yreg, 3))
ggplot(data=tab, mapping=aes(x=diam, y=age, color = as.factor(facval))) +
  geom_point() +
  theme(legend.position = "bottom")
```



2 is ratio, 3 is regression.

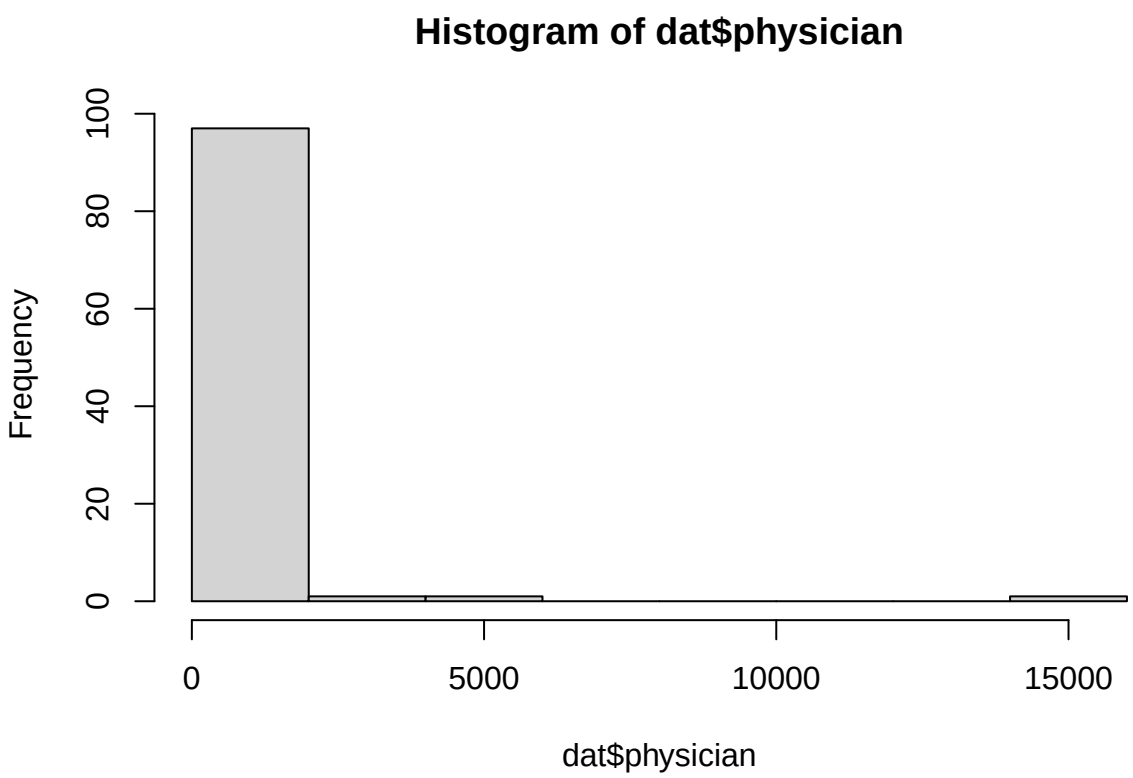
It seems that the ratio estimator will be much closer to the data than the regression estimator (unless I've done something very wrong and missed the calculation)

P3:

(a)

```
dat = read.csv("counties.csv")

hist(dat$physician)
```



(b)

- Total number of physicans

```
#dat
N = 3141
n = 100
mu = mean(dat$physician)
total_p = mu*N
total_p
```

```
## [1] 933411
```

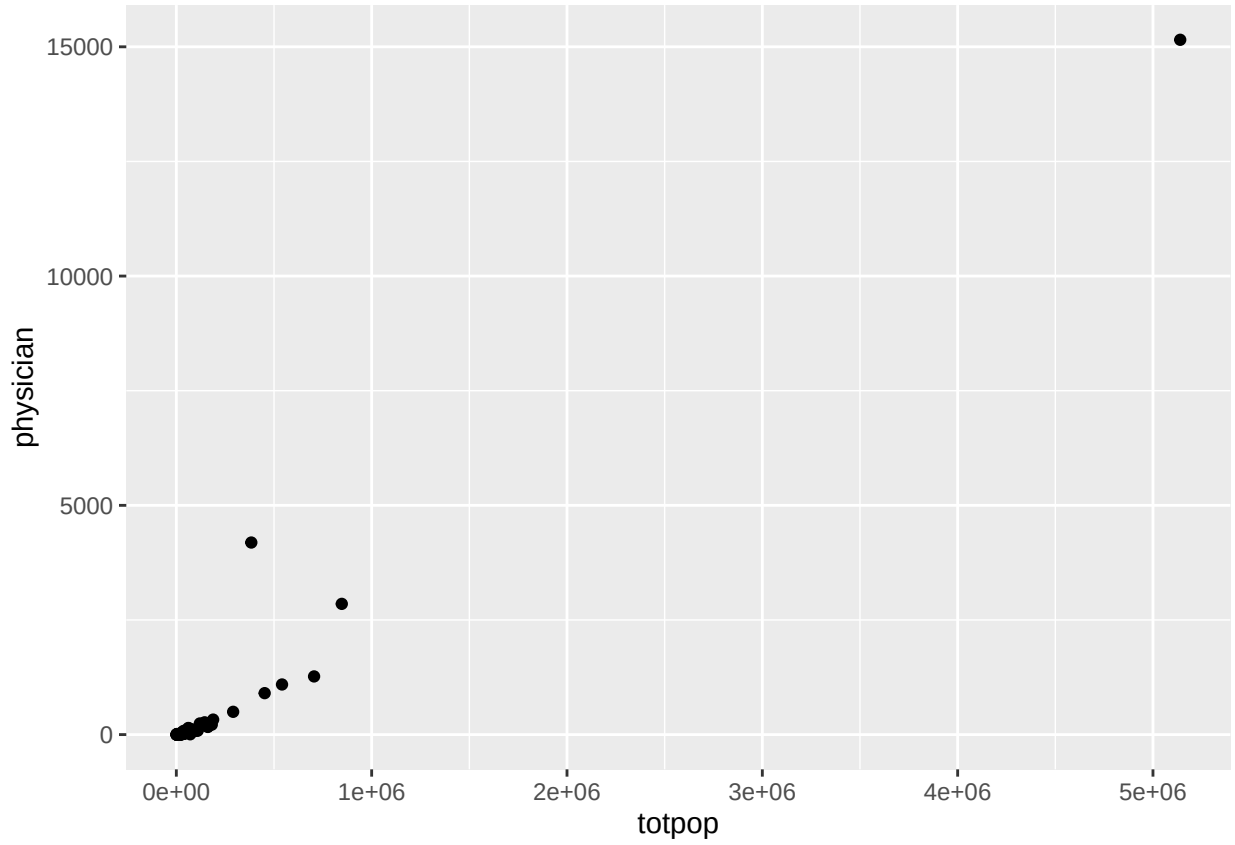
- Standard error of total physicians

```
#standard e
N*sqrt((var(dat$physician)/n)*(1-n/N))
```

```
## [1] 491982.8
```

(c)

```
ggplot(dat,aes(totpop,physician))+
  geom_point()
```



* Just by looking at the plotted graph, it seems like the abline approximately goes through the origin. Therefore observed from the scatter plot, ration estimation seems more appropriate.

(d) need correction

```
#Using ratio estimation
x_bar = mean(dat$totpop)
y_bar = mean(dat$physician)

b_hat = y_bar/x_bar
b_hat
```

```
## [1] 0.002507104
```

```
t_hat_yr  = b_hat * 255077536
t_hat_yr
```

```
## [1] 639506
```

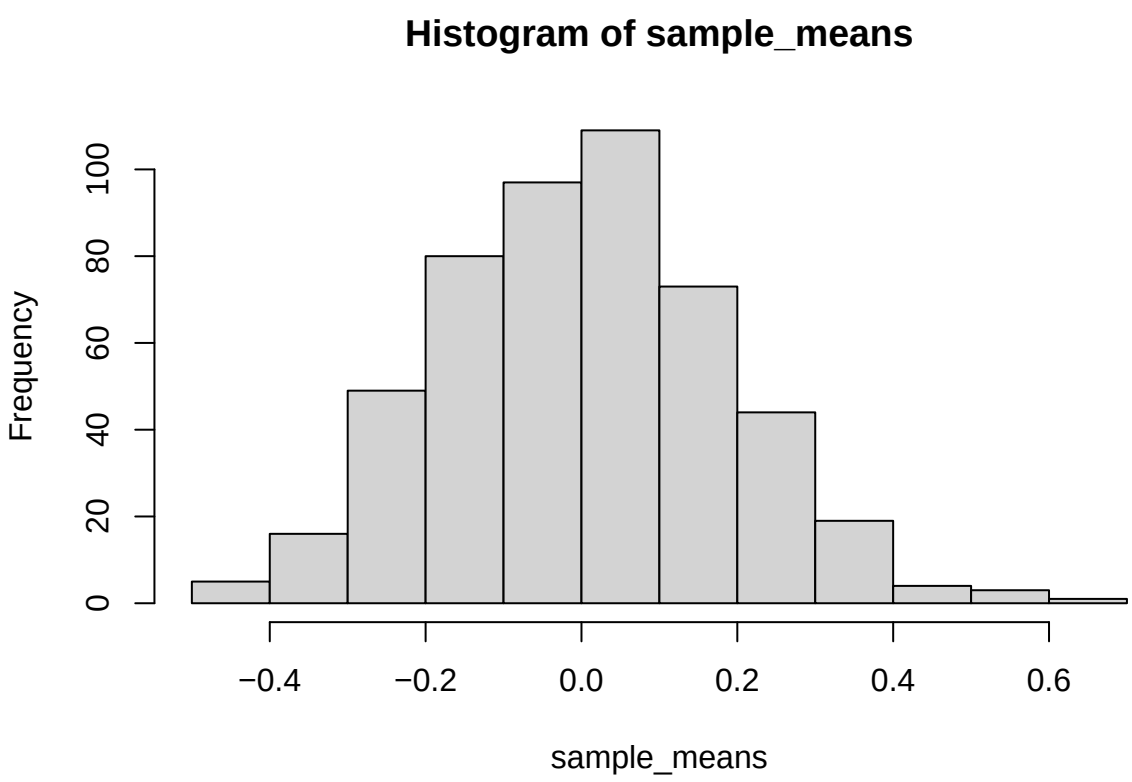
(e)

P4:

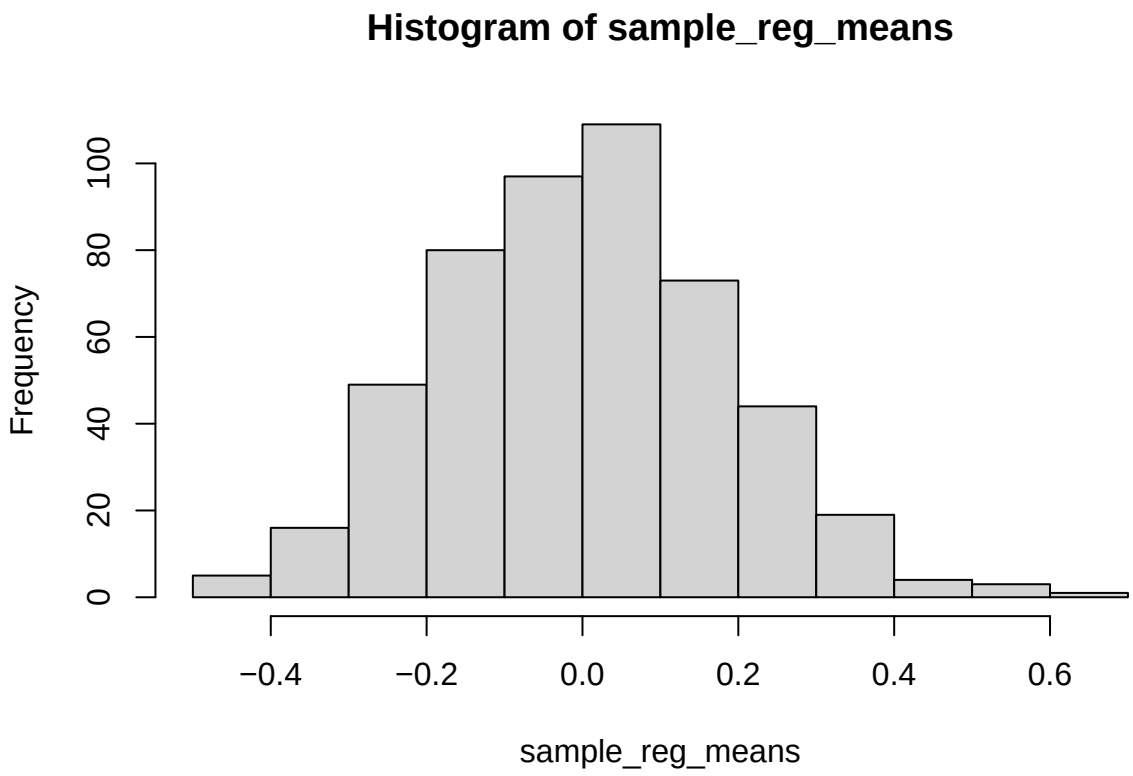
a)

```
N <- 30 # Number of random samples
set.seed(123)
# Target parameters for univariate normal distributions
rho <- 0.5
mu1 <- 0; s1 <- 1
mu2 <- 0; s2 <- 1

set.seed(532389)
# Parameters for bivariate normal distribution
mu <- c(mu1,mu2) # Mean
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2),
           2) # Covariance matrix
#mvrnorm(n = N, mu = mu, Sigma = sigma)
total_samples = c()
sample_means = c()
sample_reg_means = c()
for(i in c(1:500)) {
  sample = data.frame(mvrnorm(n = N, mu = mu, Sigma = sigma))
  colnames(sample) = c('x','y')
  ybar = sum(sample$y) / 30
  xbar = sum(sample$x) / 30
  s_y = sqrt(sum((sample$y - ybar)^2)/(30 - 1))
  s_x = sqrt(sum((sample$x - xbar)^2)/(30 - 1))
  sample_means = c(sample_means, ybar)
  s_reg = (sum((sample$x - xbar) * (sample$y - ybar))) / (30 - 1)
  r = s_reg / (s_x ^ 2)
  bhat1 = (r * s_y) / s_x
  bhat0 = ybar - (bhat1 * xbar)
  yreg = bhat0  + (bhat1 * xbar)
  sample_reg_means = c(sample_reg_means, yreg)
  sample$facval = c(i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i)
  total_samples = rbind(total_samples, sample)
}
hist(sample_means)
```




```
hist(sample_reg_means)
```

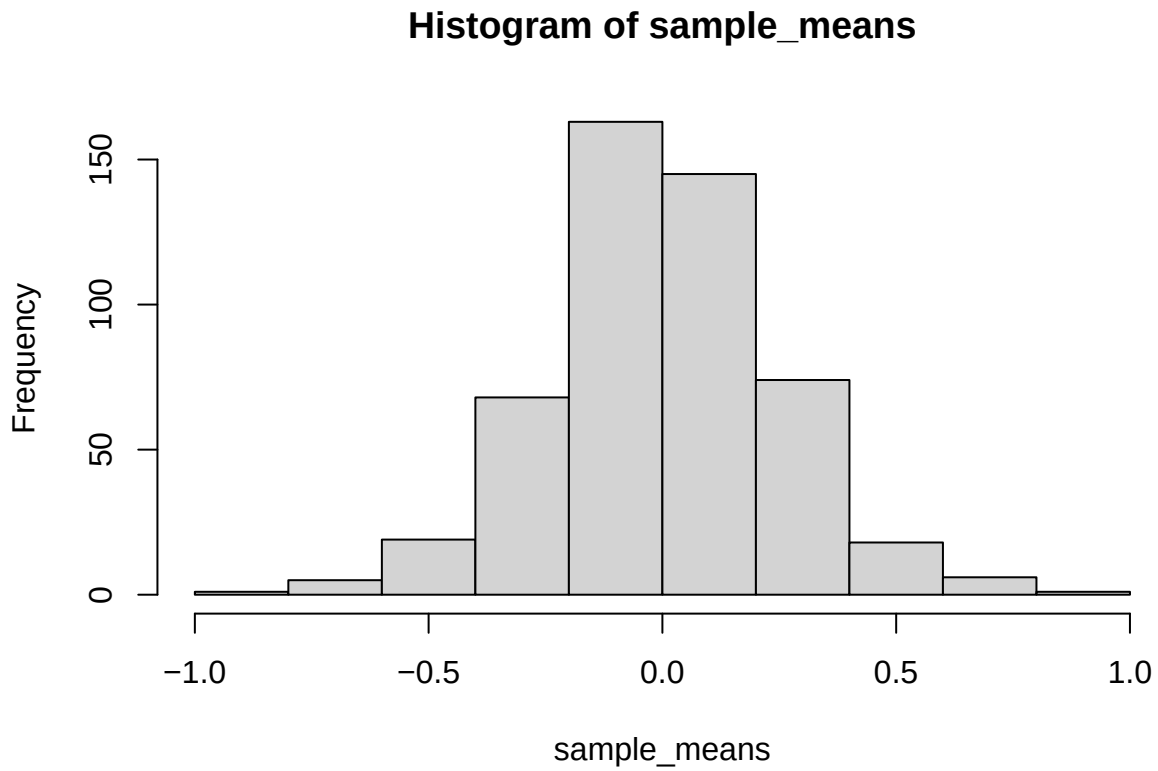


b)

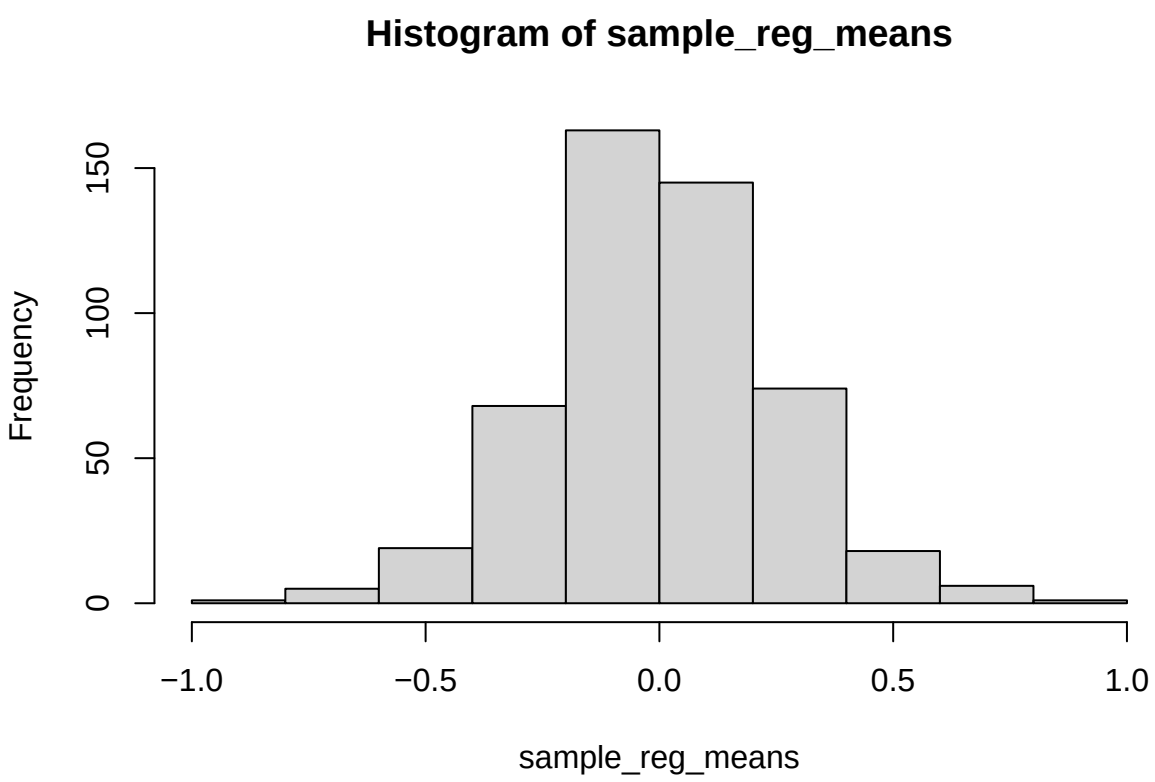
```
N <- 60 # Number of random samples
set.seed(123)
# Target parameters for univariate normal distributions
rho <- 0.5
mu1 <- 0; s1 <- 1
mu2 <- 0; s2 <- 1

set.seed(532389)
# Parameters for bivariate normal distribution
mu <- c(mu1,mu2) # Mean
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2),
           2) # Covariance matrix
#mvrnorm(n = N, mu = mu, Sigma = sigma)
total_samples = c()
sample_means = c()
sample_reg_means = c()
for(i in c(1:500)) {
  sample = data.frame(mvrnorm(n = N, mu = mu, Sigma = sigma))
  colnames(sample) = c('x','y')
  ybar = sum(sample$y) / 30
  xbar = sum(sample$x) / 30
  s_y = sqrt(sum((sample$y - ybar)^2)/(30 - 1))
```

```
  s_x = sqrt(sum((sample$x - xbar)^2)/(30 - 1))
  sample_means = c(sample_means, ybar)
  s_reg = (sum((sample$x - xbar) * (sample$y - ybar))) / (30 - 1)
  r = s_reg / (s_x ^ 2)
  bhat1 = (r * s_y) / s_x
  bhat0 = ybar - (bhat1 * xbar)
  yreg = bhat0  + (bhat1 * xbar)
  sample_reg_means = c(sample_reg_means, yreg)
  sample$facval = c(i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i,i)
  total_samples = rbind(total_samples, sample)
}
hist(sample_means)
```




```
hist(sample_reg_means)
```



P5:

(a)

- council estimates the proportion $\hat{p} = 112/134$ This is correct because out of 157 voters 23 refused so 134 voters answered and 112 people opposed.

(b)

- $\hat{V}(\hat{p}) = 0.83582(1 - 0.83582)/134$ This estimate is wrong because choosing people living in the same area is not SRS. Cluster sampling is needed.

P6:

a)

This is a cluster sample because we are taking every possible sample unit in certain clusters. Our primary sampling units (how we are forming clusters) is our clinics. The clinics chosen in the survey have our secondary sampling units, which are the families who bring children for a health checkup or a sick visit. These are all surveyed from within the chosen clinic. Because all ssu are sampled within a cluster, this is a one-stage cluster sample.

b)

The sampled population is all households who get checkups at a clinic. I would not call this representative, as it does not reflect groups of people who would not necessarily have their children checked up at a clinic, which could be a confounding variable to consider in terms of the data results. This could be because of various groups who might not go to the clinic could be ones with very strong opinions and expressions. An example of families who might not go to a clinic are families without medical insurance.

P7: