Stats 332 Assignment 2

Junqiao Lin

Feb 07, 2019

Question 1.

Do you think politicians are dishonest, or are they like most other people? \

i) This question contains an extreamly ambiguous wording, since it is often unclear what the question is refering to when it's refering to other people (since it can be taken both ways whether an average person is honest or not). The question is also quite open ended since there is different reason why people think politicians are dishonest. \

ii) I would take out the bit about mentioning like most other people, and expand a bit more into detail about specifically what they mean by dishonest. I would also give the interviewr some room to expand on their opinion. \

Many experts note that national spending on social programs is unsustainable. Do you agree that it is wise to cut spending on social programs? \

i) Social program is a bit vague, and adding the first part regarding the expert will cause most people to think of the first social program that does not directly benefit them. Also the first part of the question seems to be derliably bias the towards we should cut spending on social program (since we started the question by mentioning expert) \

ii) I would give a few example of the national spending program when asked the question (i.e. healthcare, housing...) and change the first part of the question to something more moderate. (instead of saying expert note that national spending on social programs is unsustainable).\

Please tell us which social programs should be cut, and why. \

i) Given the board definition of social program, as mention above, people might just give the first thing that comes to mind that does not applied to them. We are also assuming that people respond to Yes to the above question. \

ii) I will provide a list of biggest social program that our government have spend on within the servey. Assuming this question is follow by yes on the previous question.\

Question 2.

a)

```
library(survey)
```

```
Q2.dat<- read.csv("tree.csv")

#initalizing the design for the variable
Q2.srs <- svydesign(ids = ~1, data = Q2.dat, fpc = ~N)
Q2.N <- Q2.dat$N[1]
Q2.diamean <- 10.3
```

a)

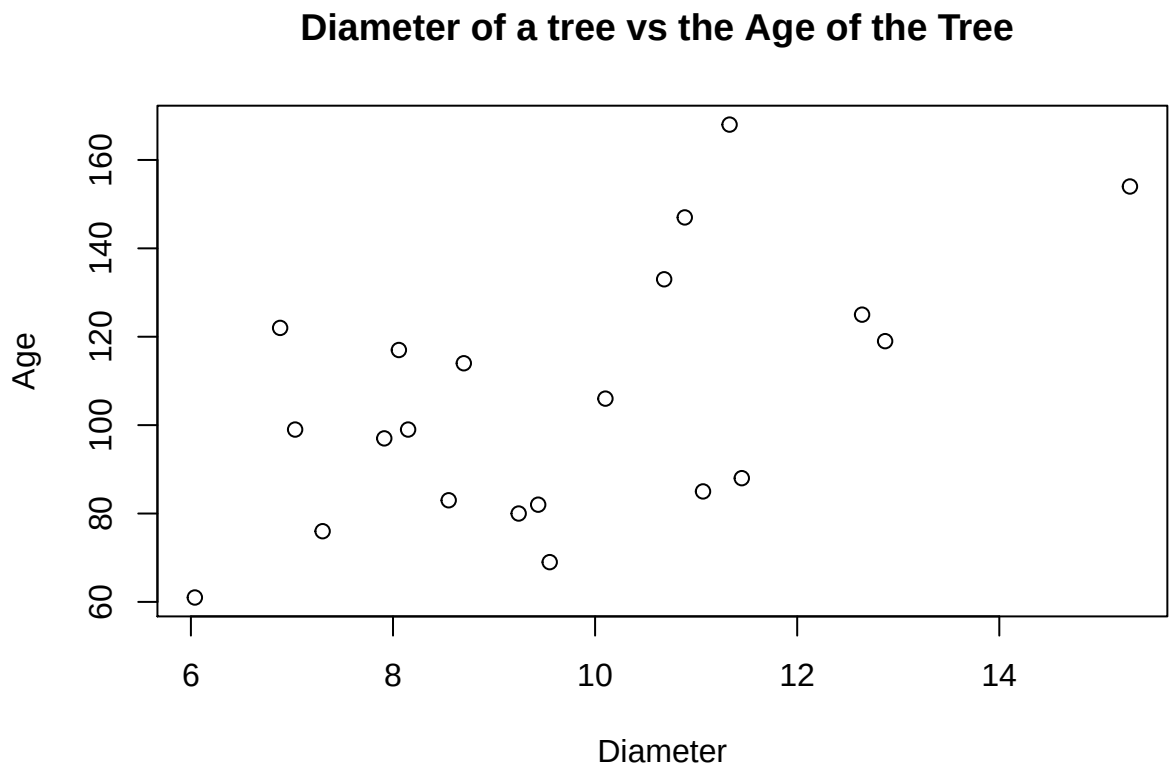
```
svymean(~Age,Q2.srs)
```

```
##      mean     SE
## Age 105.9 6.2222
```

Base on the data, the average age is 105.9 and the Standard Deviation is 6.2222.

b)

```
plot(Q2.dat$Diam, Q2.dat$Age, main="Diameter of a tree vs the Age of the Tree ",
     xlab = "Diameter",
     ylab = "Age")
```



We can see that Age and Diameter seems to have a weak corrlation. In fact if we calculate $\hat{\rho}$:

```r
cor(Q2.dat$Diam, Q2.dat$Age)
```

```
## [1] 0.5629592
```

Which shows that the diameter seems to be weakly corrlated to the age, however $\hat{\rho} \geq 0.5$. Thus we can expect the variance of part a being higher then the variance of part c.

c)

```r
Q2c.ratio <- svyratio(~Age, ~Diam, design=Q2.srs)
predict(Q2c.ratio, total = Q2.diamean)
```

```
## $total
##         Diam
## Age 112.7381
##
## $se
##         Diam
## Age 5.846481
```

Base on ratio estimation, the average age is 112.7381 and the SE is 5.846481.

d)

```r
Q2.diamean <- data.frame(Diam = Q2.diamean)
Q2d.reg <- svyglm(Age~Diam, design=Q2.srs)
predict(Q2d.reg, newdata = Q2.diamean)
```

```
##     link     SE
## 1 110.31 5.3772
```

Base on regression estimation, the average age is 110.31 and the SE is 5.3772.

e) Note if we measure the quality of the estimation base on the lower SE, it is unsurprising to find that the SE for d) is higher then the SE for c) given the result for question 3. Note as analysis in b, since $\hat{\rho} \geq 0.5$, we have the SE from c) to be lower then the SE of a). \

Question 3.

Note under the ratio estimation test, we have $b = \hat{\theta}$ and under regression estimation test we have $b = \frac{\hat{\rho}\hat{\sigma_y}}{\hat{\sigma_x}}$, Hence $\hat{Var}(\hat{\mu_{ratio}}) \geq \hat{Var}(\hat{\mu_{reg}})$ iff the following holds:

$$
\begin{aligned}
\hat{Var}(\hat{\mu_{ratio}}) &\geq \hat{Var}(\hat{\mu_{reg}}) \\
\frac{1}{n}(1-\frac{n}{N})(\hat{\sigma_y^2} + \hat{\theta}^2\hat{\sigma_x}^2 - 2\hat{\theta}\hat{\sigma_x}\hat{\sigma_y}\hat{\rho}) &\geq \frac{1}{n}(1-\frac{n}{N})(\hat{\sigma_y^2} + (\frac{\hat{\rho}\hat{\sigma_y}}{\hat{\sigma_x}})^2\hat{\sigma_x}^2 - 2(\frac{\hat{\rho}\hat{\sigma_y}}{\hat{\sigma_x}})\hat{\sigma_x}\hat{\sigma_y}\hat{\rho}) \\
(\hat{\theta}\hat{\sigma_x})^2 - 2\hat{\theta}\hat{\sigma_x}\hat{\sigma_y}\hat{\rho} &\geq (\frac{\hat{\rho}\hat{\sigma_y}}{\hat{\sigma_x}})^2\hat{\sigma_x}^2 - 2(\frac{\hat{\rho}\hat{\sigma_y}}{\hat{\sigma_x}})\hat{\sigma_x}\hat{\sigma_y}\hat{\rho} \\
(\hat{\theta}\hat{\sigma_x})^2 - 2\hat{\theta}\hat{\sigma_x}\hat{\sigma_y}\hat{\rho} &\geq (\hat{\rho}\hat{\sigma_y})^2 - 2(\hat{\rho}\hat{\sigma_y})^2 \\
(\hat{\theta}\hat{\sigma_x})^2 - 2\hat{\theta}\hat{\sigma_x}\hat{\sigma_y}\hat{\rho} + (\hat{\rho}\hat{\sigma_y})^2 &\geq 0 \\
(\hat{\theta}\hat{\sigma_x} - \hat{\rho}\hat{\sigma_y})^2 &\geq 0
\end{aligned}
$$

Note since $(\hat{\theta}\hat{\sigma_x} - \hat{\rho}\hat{\sigma_y}) \in \mathbb{R}$, $(\hat{\theta}\hat{\sigma_x} - \hat{\rho}\hat{\sigma_y}) \geq 0$ is always true and hence $\hat{Var}(\hat{\mu_{ratio}}) \geq \hat{Var}(\hat{\mu_{reg}})$.

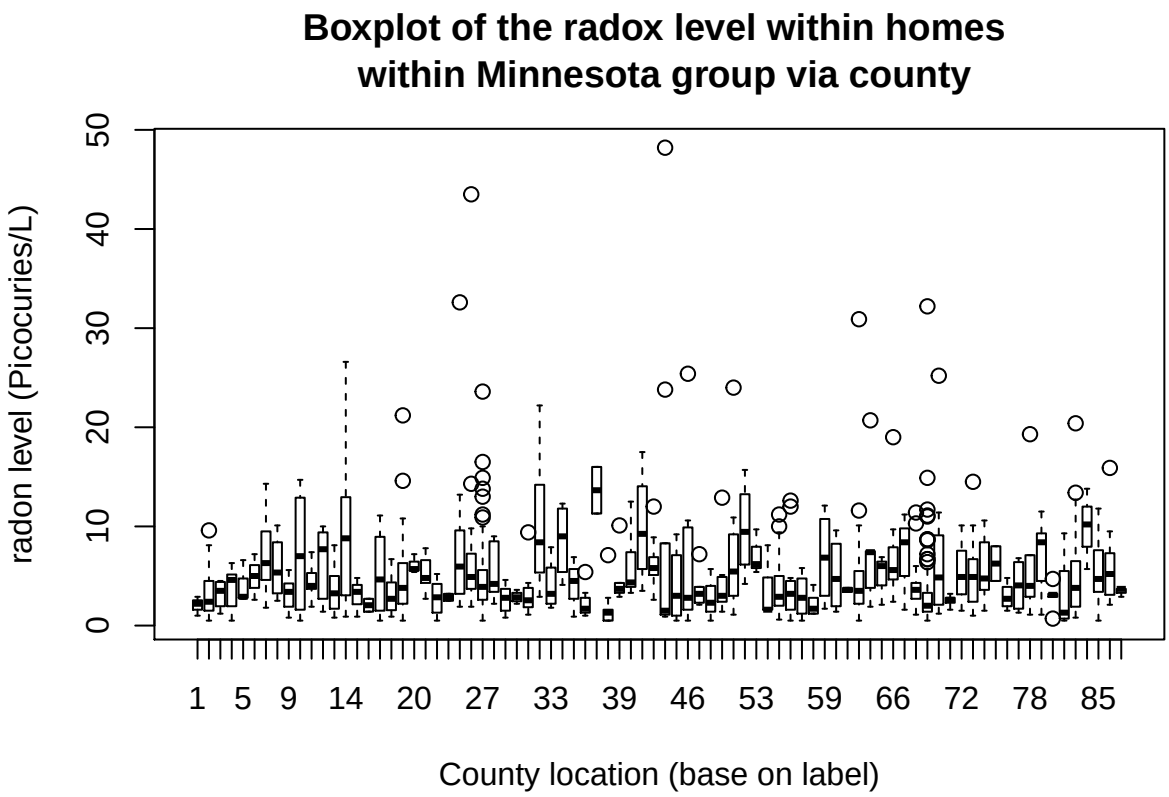
Question 4

Source: https://www.canada.ca/en/health-canada/services/radon.html

a) Since the radon gas usually gets in the house via the break down of uranium in soil and rock. The radon level in a particular house could be dependent on the soil level within the county. Thus it could explain why strata sampling is used, so that a area with more radioactive material wouldn't be overrepersented/underrepersented within the survey (it could happen under SRS).

b)

```
Q4.dat<- read.csv("radon.csv")
boxplot(Q4.dat$radon~Q4.dat$strata, main="Boxplot of the radox level within homes \n within Minnesota g
        xlab="County location (base on label)",
        ylab="radon level (Picocuries/L)")
```



We can see that most county radon level lies bewteen 0-10 (Picociroes per L) With various variance level. There are also a few county with noticiblely higher average and even one county (bewteen 30-40) with an average level of about 15. There is also quite a bit of outliers which have a significant higher level of radon in comparison to the average. \

Given the information above, it reaffrim the decision of using stratisfy sampling due to the difference bewteen mean and variance accross different county. \

c)

```
#initalizing the design for the variable
Q4.srs <- svydesign(id=~1,data = Q4.dat, fpc = ~N,strata=~strata)
svymean(~radon,Q4.srs)
```

```
##         mean     SE
## radon 4.8723 0.1552
```

```
confint(svymean(~radon,Q4.srs))
```

```
##          2.5 %   97.5 %
## radon 4.568129 5.176476
```

Base on the analysis, the average level of radon is 4.8723 pCi/L and the SE is 0.1552. Note this is higher the the recommended level by the USEPA. A reason might be because of the outlier mention above being much higher then standard. The 95 percent CI is (4.568129 5.176476).

d)

```
svymean(~exceed,Q4.srs)
```

```
##           mean     SE
## exceed 0.47052 0.0189
```

```
confint(svymean(~exceed,Q4.srs))
```

```
##            2.5 %   97.5 %
## exceed 0.4333796 0.507662
```

Base on the analysis, appoximately 0.47052 percent of the population within Minnosoda needs to be fixed arrording to USEPA recommendations, with 0.0189 percent SE. The 95 percent CI is (0.4333796,0.507662).