Statistics 144 Term Project

Chapter 2
Problem 5:
For the three parts of this problem, I used these formulas to calculate the means and standard errors:

$$\bar{y}_{\mathcal{S}} = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i. \qquad\qquad \text{SE}\,(\bar{y}) = \sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

Note: throughout this project N = 19664 and n = 200

a. The average price a household is willing to pay for cable TV is $10.45 with a standard error of 0.4850345.
b. The average number of TV's in a household in Lockhart City 1.82 with a standard error of 0.07685793.
c. The proportion of houses willing to pay at least $10 for cable service is 0.68 with a standard error of 0.03289902.

Problem 6:
Using the sample size of 200, the estimated average assessed valuation in Lockhart City is $7,3048.96 with a standard error of 548.804 (using the same formulas as in Problem 5).

Formula for the 95% confidence interval: $[\bar{y} - z_{\alpha/2}\text{SE}(\bar{y}), \bar{y} + z_{\alpha/2}\text{SE}(\bar{y})]$

The 95% confidence interval for the average assessed valuation in Lockhart City is (71,973.3, 74,124.62). In other words, we are 95% confident that the mean assessed valuation in Lockhart City would be between $71,973.3 and $74,124.62. This interval includes the known value of $71,117.

Chapter 3
Problem 10:
Formulas for calculating the means and the standard errors:

$$\bar{y}_{\text{str}} = \frac{\hat{t}_{\text{str}}}{N} \qquad\qquad \hat{V}(\bar{y}_{\text{str}}) = \frac{1}{N^2}\hat{V}(\hat{t}_{\text{str}}) = \sum_{h=1}^{H}\left(1-\frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2\frac{s_h^2}{n_h} \qquad\qquad \text{SE}(\bar{y}_{\text{str}}) = \sqrt{\hat{V}(\bar{y}_{\text{str}})}.$$

Note: The stratification variable is number of persons aged 12 or older in the household. Since the household with 5 or more persons aged 12 or older are a small proportion of the population, and we will only sample one of each, this would not give a good estimate of the variance of each stratum. Therefore, I collapsed such households into one stratum, which give a total of five strata. This gives a total of 5 strata.

The stratification order is 4, 2, 3, 1, 5, where 5 actually means 5 or more (persons aged 12 or up).

The Nh's corresponding to the strata in the given order are:
3735, 9195, 4001, 1721, and 1012
And, the nh's corresponding to the strata in the given order are: 38, 94, 41, 18, and 11

When we deal with the question about cable service costs, the simple variances for each stratum are 31.81010, 42.51030, 41.21951, 41.17647, and 26.36364, respectively.

For the number TV sets question, the simple `variances for each stratum are 1.5903272, 1.3754290, 1.2451220, 1.0588235, and 0.4727273, respectively.

In the stratified sampling case, the average price a household is willing to pay for cable TV is approximately $10.67 and the average number of TV's in a household in Lockhart City is 1.95. Compared to the estimates in the simple random sample, these estimates are a little bigger. However, stratified sampling gave better estimates because the standard error for the average cable price is 2.164435e-05, and the standard error for the average number of TV's is 3.649942e-06. Both standard errors are considerably smaller than the ones in SRS case.

Problem 11:
Using the sample of size 200 in chapter 2, the variance in each of the strata for the average price a household is willing to pay for cable TV service are:

44.36572, 34.39478, 53.07843, 38.09211, and 50.27778, respectively.

I used the simple variance formulas $s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2$ for each stratum variance, where y bar is the mean of each stratum.

Problem 12:
Since we are dealing with only urban households, we know what to interview each household would cost $10 and to visit each household would be $3, so the variable cost per household would be $13. We can assume the cost per household is the same for each observation; therefore, we can use the special case of optimal allocation - the Neymann allocation.

Here I assumed that nh is proportional to NhSh, where Sh is specified in problem 11.

Using the Neymann allocation, it appears that we should sample 39 households with 4 adults (aged twelve or up), 85 households with 2 adults, 46 households with 3 adults, 16 households with 1 adult, and 11 households with 5 adults.

This results in the average price a household is willing to pay for cable TV services is approximately $10.12 (R: 10.12376), and the average number of TV sets per household is 1.76 (R: 1.762376).

Problem 14:
Formulas used:

$$e = z_{\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}} \qquad n_0 = \left(\frac{z_{\alpha/2}S}{e}\right)^2 \qquad n = \frac{n_0}{1+\frac{n_0}{N}} = \frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}}$$

Using the variances estimated in question 5 of Chapter 2, we need a sample of at least 226 with simple random sampling to achieve the same precision in estimating the average price a household is wiling to pay as a stratified sample of size 200 using the strata from an optimal allocation.

> ** I used the standard deviation of cable services cost from stratified sampling using Neymann allocation (6.531168) to calculate the precision e. This gives the precision of the stratified sampling approach. Then I used the standard deviation of cable services cost from simple random sampling to achieve the optimal sample size that gives the same precision using the SRS approach. Of course, N and n are the same as in problem 5 in chapter 2.

To achieve the same goal but with strata from a proportional allocation, we would need a sample size of 205.

> ** In this case, I repeated the above process by replacing the standard deviation of cable services cost from stratified sampling using Neymann allocation to that from using proportional allocation (6.303921).

Chapter 4
Problem 16:

Formulas used:

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x}$$

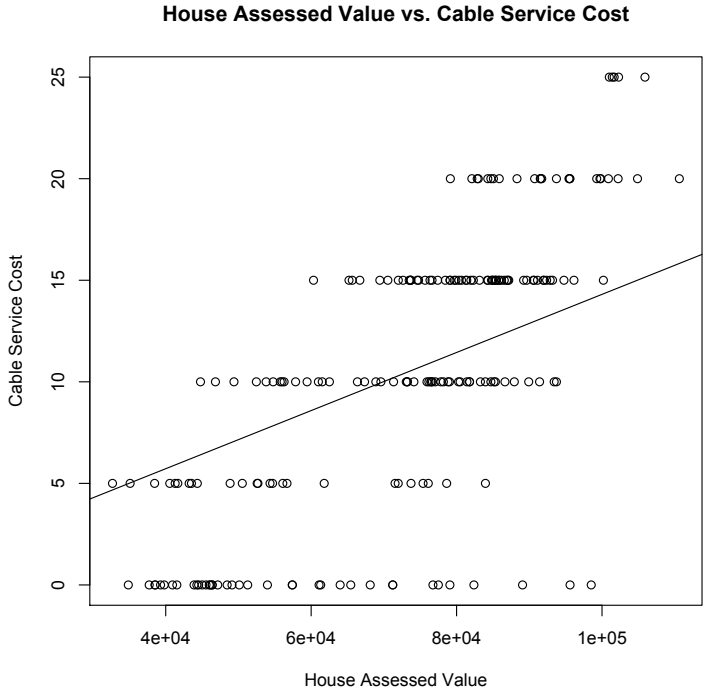
$$\hat{t}_{yr} = \hat{B}t_x$$

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U, \qquad e_i = y_i - \hat{B}x_i, \qquad s_e^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}} e_i^2$$

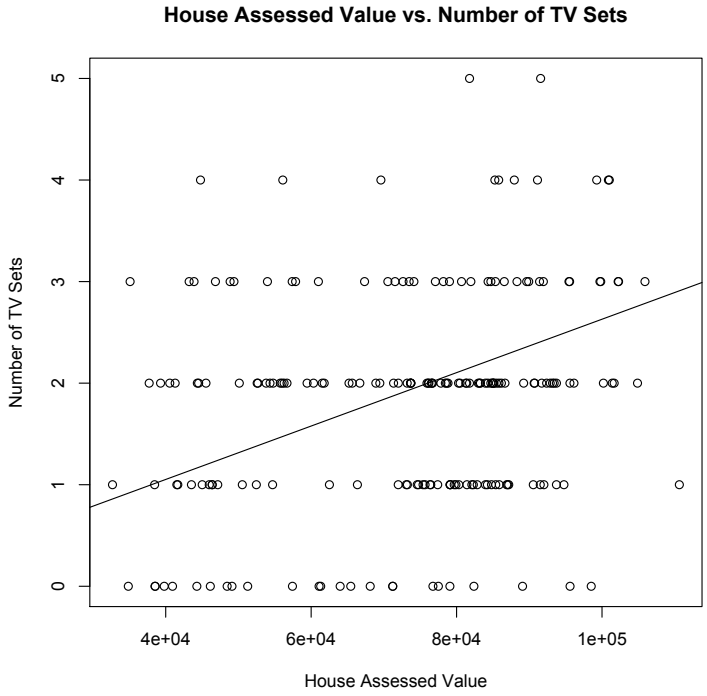
$$\hat{V}(\hat{t}_{yr}) = \hat{V}(t_x\,\hat{B}) = \left(1-\frac{n}{N}\right)\left(\frac{t_x}{\bar{x}}\right)^2\frac{s_e^2}{n}$$

(a) Using the ratio estimate with assessed value of house as the independent variable, the new mean for price a household is willing to pay for cable is approximately $10.17 (R: 10.17357) with a standard error of 0.3641616.

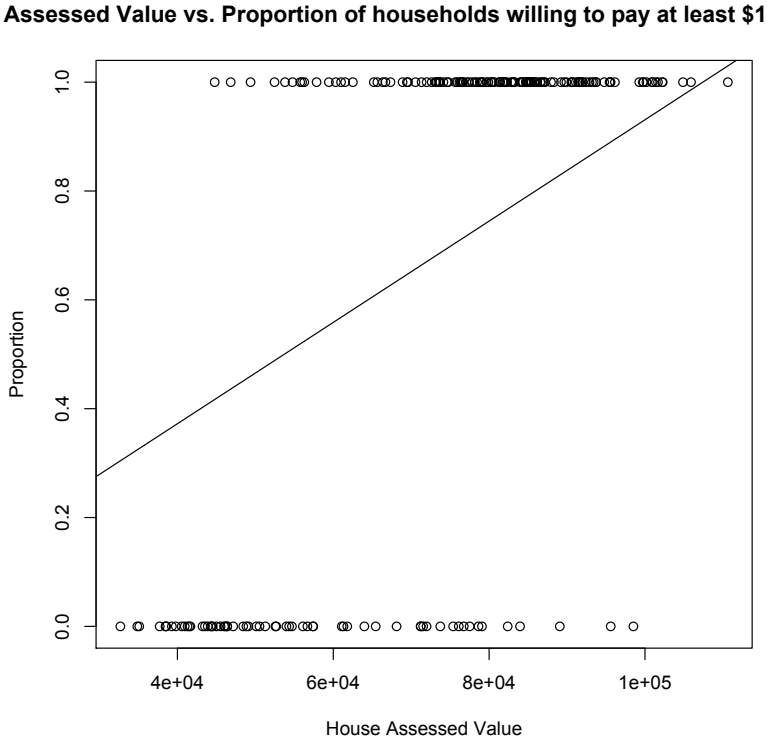
Note: bhat in this case is 0.0001430547, and sample variance of residuals is 28.27116.



(b) The new average number of TV sets per household is 1.869983 with a standard error of 0.07430569. The bhat is 2.629458e-05, and the sample variance of residuals is 1.177062.



(c) The new estimated proportion of households willing to pay at least $10 for cable TV is 0.6620125 with a standard error of 0.02607958. The bhat is 9.308825e-06, and the sample variance for residuals is 0.1449961.



Compared to the results in problem 5 of Chapter 2, these estimates are more precise because the standard error in each case is smaller than corresponding one in problem 5.

According to assumption 5, all other factors being equal, a household with a higher income will tend to be willing to pay more for cable service. We can relate this to the result in part a. From that plot we can see that the high cable TV cost shifts to the upper right of the plot, and the lower cable TV cost tend to concentrate on the lower left hand corner of the plot. In other words, households with higher income tends to be willing to pay more for cable TV services. Also, we can also relate to part c. Here we deal with the proportion, so there are only two values along the y-axis, but we can see that for proportion of 1, there are more scatter points to the right, but for proportion of 0 there are more scatter points to the left. In other words, people with higher income are willing to pay more than $10 for cable TV, whereas people with lower income are willing to pay less than $10 for cable TV.

According to assumption 6, all other factors being equal, a household with a higher income will tend to own more television sets. We can refer to the part b in this case. From the plot, we can see that in each number on the y-axis (the number of TV sets), there are more scatter points to the right hand side of the plot. That means, there people with higher income tends to have TV or more TV's.

Problem 17:
For this problem, I used estimation of domain, the equations I used are:

$$\bar{y}_d = \hat{B} = \frac{\bar{u}}{\bar{x}} = \frac{\hat{t}_u}{\hat{t}_x} \qquad t_x = \sum_{i=1}^{N} x_i = N_d \qquad t_u = \sum_{i=1}^{N} u_i \qquad u_i = y_i x_i = \begin{cases} y_i & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d. \end{cases}$$

$$s_{yd}^2 = \frac{\sum_{i \in \mathcal{S}_d} (y_i - \bar{y}_d)^2}{n_d - 1} \qquad \operatorname{SE}(\bar{y}_d) \approx \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_{yd}^2}{n_d}}.$$

The estimated average number of adults per household in Lockhart City households willing to pay at least $10 for cable service is approximately 2.57 (R: 2.566176), with a standard error of 0.08881707. The coefficient of variance of the estimate is 0.03461066; it is found by dividing the standard error by the mean.

Problem 18:
Having figured out problem 17, we can multiply the mean and standard error by the population size to get the total number of adults in Lockhart City who live in households willing to pay at least $10 for cable services and the corresponding standard error. The total number of adults in Lockhart City who live in households willing to pay at least $10 for cable services is 48495.16 with a standard error of 10165.44. We get the same CV for this problem as the previous one (0.03461066).

Chapter 5
Problem 20:
It costs $6 to visit a rural household and $10 to interview a household, so the variable cost in this study would be $16. It costs $60 to visit a rural district, so $60 is a fixed cost per rural. I have decided to sample 13 rural district, which is within 25% to 50% of the districts. Now we know the cost to visit 13 districts is 13*60 = 780; the cost to sample 100 households is 100*16 = 1600. Then by sampling 10 samples of 100 and calculating the cost for each to estimate our budget, I got $3838 as this project's budget.

By subtracting the fixed cost, we have $3058 to using on sampling individual households, which is enough to visit and interview 191 households.

Then, I checked to see if this survey is within our budget of $3838 by doing:
13*60 + 191*16 = $3836.

This shows that we are barely within budget. Therefore, we can proceed to the next step.

The districts that I will visit are: 12 4 27 6 7 38 35 20 10 19 15 42 39
And the corresponding proportional sample sizes are:
11 13 10 31 18 7 9 7 22 8 19 28 8

Problem 21:
Since this a two-stage, we would use the following equations for unbiased estimation:

$$\hat{t}_{\text{unb}} = \frac{N}{n}\sum_{i\in\mathcal{S}}\hat{t}_i = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\bar{y}_i = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{N}{n}\frac{M_i}{m_i}y_{ij}.$$

$$V(\hat{t}_{\text{unb}}) = N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i},$$

$$s_t^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2$$

$$s_i^2 = \frac{1}{m_i-1}\sum_{j\in\mathcal{S}_i}(y_{ij}-\bar{y}_i)^2$$

From R, we got:
t_hat_unb = 106571.2
Var(t_hat_unb) = 370129442
S^2_t = 3664056
S^2_i =
55.45455 46.47436 79.16667 51.82796 59.55882 30.95238 56.25000 73.80952 66.66667 62.50000 42.69006 70.23810 74.55357
y_bar_hat_unb = 13.4356

and SE(y_bar_hat_unb) = 2.42546

Note: SE(y_bar_hat_unb) = SE(t_hat_unb)/Mo

For ratio estimation, we would use the following equations instead:

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i},$$

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M_i\bar{y}_i - M_i\hat{\bar{y}}_r)^2,$$

Here we get:
Y_bar_hat_r = 12.2974
Sr2 = 17908.72
Var(y_bar_hat_r) = 0.1107744
SE(y_bar_hat_r) = 0.3328279

Problem 22:
For this problem, I will sample another 13 districts using the self-weighted sampling scheme.

The districts I will visit are randomly selected as:
10 42 15 28 18 16 11 6 43 42 40 22 11

We will need equal sample sizes for each district (cluster). Since the fixed cost is the same as in the previous problem, we can sample another 191 individuals; this allows us to sample 14.69231 person per district. Because we have to stay within budget, we would need to round the number down to 14 persons per district.

Psi = Mi/Mo

Mo = 7932, and our Psi's are:
0.01928896 0.02168432 0.01802824 0.05446293 0.03126576 0.01147252 0.01638931 0.01285930 0.03744327 0.01298538 0.03202219 0.04853757 0.01386788

Mi's are: 153 172 143 432 248 91 130 102 297 103 254 385 110

y_bar_i's are: 9.642857 13.214286 11.428571 8.928571 13.928571 15.357143 14.285714 13.214286 15.714286 13.571429 12.500000 12.142857 15.357143

t_hat_i's are: 1475.357 2272.857 1634.286 3857.143 3454.286 1397.500 1857.143 1347.857 4667.143 1397.857 3175.000 4675.000 1689.286

With the above information, we can estimate y_hat_bar_Psi = 13.02198 with a standard error of 2.09579.