Exercise 1: We have to estimate the average house value in an area with $N = 2000$ houses. Normally we would do this though an assessment process, where we take some observed (public) data on the property and then compare each property to 'comparable' properties nearby that have sold recently. This process is fairly accurate (gives a good guess at the potential selling price), but is hard to do. An easier approach is to just look up the assessed values in the property tax database. We actually can combine the two ideas. Let X_i be the easy to get assessed value (we can get this for all properties in the area), and Y_i the difficult to get but more accurate value.

All values are in units of thousands of dollars

The total of X is $\tau_x = 512,400$ in the entire area. We take a sample size of $n = 15$ houses and get the following:

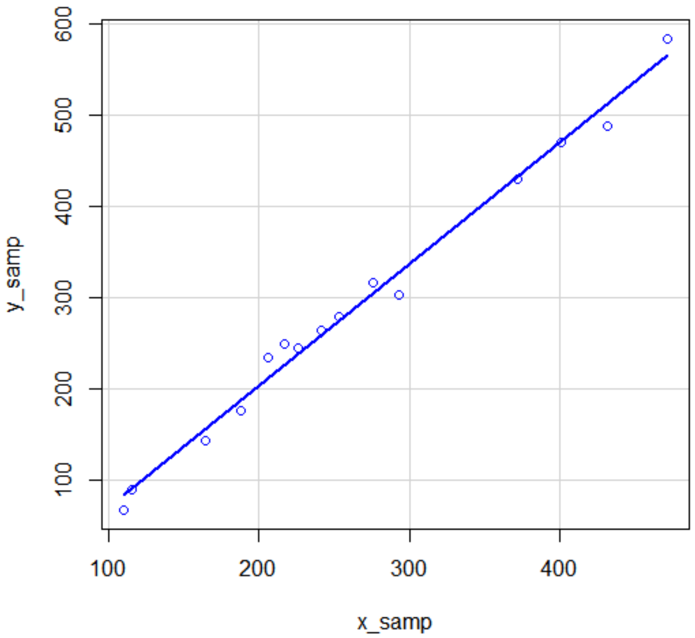
$$X_i = 472, 206, 372, 241, 226, 276, 164, 115, 432, 293, 401, 217, 188, 253, 110$$

$$Y_i = 583, 234, 429, 264, 244, 316, 143, 89, 487, 302, 469, 248, 176, 279, 67$$

a. Plot Y vs X. Does there appear to be a linear relationship?

Solution:
Plotting in r we get the following figure,

Figure 1: Data Scatterplot



The data does look to be very linear, with constant variance. A ratio or regression estimator might work well.

b. Use the regression estimation approach to estimate the true total value of all houses, τ_y with a 95 percent confidence interval.

Solution:
Recall that $\tau_x = 512,400$ and that $N = 2000$ we get a $\mu_x = 512,400/2000$. Fitting a linear model to the data in r, we get the following estimator,

$$\hat{\tau}_y = \beta_0 + \beta_1 \tau_x = -63.717 + 1.333(512,400) = 682965.5$$

Using anova() we can quickly get the MSE them computing the variance we get,
Code:

```
> reg = lm(y_samp ~ x_samp)
Coefficients:
(Intercept)      x_samp
   -63.717        1.333

> regEst = -63.717 +1.333*(512400)
   [1] 682965.5

> anova(reg)
Analysis of Variance Table
Response: y_samp
          Df Sum Sq Mean Sq F value    Pr(>F)
x_samp     1 305986  305986  1178.4 3.846e-14 ***
Residuals 13   3376     260
---

> regEst_var = 2000^2((2000 - 15)/2000)*(260/15)
   [1] 68813333


> CI = c(regEst + 2*sqrt(regEst_var),regEst - 2*sqrt(regEst_var))
   [1] 699556.2 666374.7
```

c. Use the usual SRS chapter 1 estimator using just the Y values to get a 95 percent confidence interval.

Solution:
First we compute the mean house value, then multiply it by the total number of houses $N = 2000$

$$\hat{\tau}_y = \mu_y N = 288.6667(2000) = 577333.4$$

Computing the variance an 95 percent confidence interval,
Code:

```
> SRSEst = mean(y_samp)*2000
    [1] 577333.3
> SRS_SE = sqrt(2000*(2000 - 15)*(var(y_samp)/15))
    [1] 76474.85
> SRS_CI = c(SRSEst + 2*SRS_SE, SRSEst - 2*SRS_SE)
    [1] 730283.0 424383.6
```

This estimator is considerably worse, the standard error is an order of magnitude greater which makes the confidence interval a lot larger. The regression estimator uses information about X and it's relation to Y to reduce the variance, looking at the SRS we can see how much that is leveraged to get a better estimator.

Exercise 2: Consider the following population,

$$\{1, 2, 3, 2, 4, 5, 7, 6, 5, 8, 10, 12, 12, 13, 12, 14, 15, 14, 16, 16, 13, 15\}$$

a. There are two possible 1-in-2 systematic samples. What are these? What is the probability that you will get the first of these samples?

Solution:
The following is the sampling distribution for the population under 1-in-2 systematic sampling,

$$P(X = x) = \left\{ \begin{array}{ll} \frac{1}{2} & \text{x = 1,3,4,7,5,10,12,12,15,16,13} \\ \frac{1}{2} & \text{x = 2,2,5,6,8,12,13,14,14,16,15} \end{array} \right\}$$

b. For each possible sample, compute the estimate of the population mean μ and find its standard error (using the SRS formula we started with in this class for the estimated mean and the standard error). Is this estimator unbiased (how can you tell)? [Note: the estimated mean from a systematic sample, using the SRS estimator, CAN be slightly biased, if N isn't an exact multiple of k.]

Solution:
Computing the population mean for each sample we get, **Code:**

```
> x_one = c(1,3,4,7,5,10,12,12,15,16,13)
> mean(x_one)
    [1] 8.909091
> x_one_SE = sqrt(22^2*((22 - 11)/22)*(var(x_one)/11))
    [1] 24.14125

> x_two = c(2,2,5,6,8,12,13,14,14,16,15)
> mean(x_two)
    [1] 9.727273
> x_two_SE = sqrt(22^2*((22 - 11)/22)*(var(x_two)/11))
    [1] 24.73863
```

Systematic sampling experiences bias when the population is periodic, where the oscillations match the sampling rate. Looking at our population we can see that we have an ordered population which also results in bias in our standard error but it is biased low so we actually have higher accuracy than what is reported.

c. The SRS estimators of the standard error for each sample have been computed in part (b), so this is what you will actually see when you collect your sample (actually, you'll only see one of them...). Do they seem to over-estimate the actual variability of the sample means? How can you tell?

Solution:
Like we illustrated before, since we have an ordered population we get a standard error which is biased low. The estimator for the mean is unbiased, but our estimate of the standard error is larger than the true standard error.

Exercise 3: Taking a close-to-systematic sample: Let's say we have a frame of $N = 100$ sampling units. We think there might be order in the values so we would rather take a systematic sample rather than a SRS (WHY?), but we are afraid of the bias that could occur in variance estimates for systematic samples if there was periodicity in the population.

Here is a way we can compromise. First, divide the frame into five strata each of size $N_i = 20$, with the first stratum consisting of the first 20 units in the frame, etc. Then take a SRS of size $n_i = 2$ from each stratum. This means you sample exactly 10 sampling units.

a. Why is this less prone to problems with periodicity (i.e. safer than a systematic sample) but less likely to have the units clustered in one area in the population (as can happen in a SRS)?

Solution:
This method has no set sampling rate, so it doesn't have the problem of having the sampling rate match the periodicity of the data like you could have with one-in-k systematic sampling. With the described sampling plan we also force the sample to have a good spread of the population, sometimes a SRS will have bad spread.

b. Take a sample as described above from this population:

$pop < -c(8, 19, 31, 39, 51, 67, 54, 75, 87, 101, 103, 116, 107, 116, 113, 137, 148, 133, 150, 153, 165, 171,$

Then compute a 95 percent confidence interval for the population mean μ.

Solution:
Sampling in r we get the following μ and confidence interval,
Code:

```
> Strata_1 = sample(pop[1:20], size = 2, replace=FALSE)
    [1]  54 148
> Strata_2 = sample(pop[21:40], size = 2, replace=FALSE)
    [1] 209 222
> Strata_3 = sample(pop[41:60], size = 2, replace=FALSE)
    [1] 230 256
> Strata_4 = sample(pop[61:80], size = 2, replace=FALSE)
    [1] 202 198
> Strata_5 = sample(pop[81:100], size = 2, replace=FALSE)
    [1] 129  11

> Strata_mean = c(mean(Strata_1), mean(Strata_2),
                  mean(Strata_3), mean(Strata_4),
                  mean(Strata_5))
> W = 1/5

> MeanEST = sum(W*Strata_mean)
    [1] 165.9
```

```
> Strata_Ssquared = c(var(Strata_1), var(Strata_2),
                      var(Strata_3), var(Strata_4), var(Strata_5))
> POPCorrection = (20 - 2)/20
> VarEST = sum(W^2*POPCorrection*(Strata_Ssquared/2))
    [1] 212.589

> CI <- c(MeanEST + 2*sqrt(VarEST), MeanEST - 2*sqrt(VarEST))
    [1] 195.0609 136.7391
```

c. There is an attached paper (it uses Landsat data for a regression estimator, but that's not what I'm interested in at this point): Forest Area Estimation Using Sample Surveys and Landsat MSS and TM data (F. Deppe, 1998), Photogrammetric Engineering and Remote Sensing, vol. 64, No. 4, pp. 285-292). Look at the part (including a figure) describing the Area Frame Sampling Scheme. They treat it like a SRS as far as the estimate and standard error, but examine Fig. 2, where it looks like they are taking a close-to-systematic sample. Describe their sampling approach.

Solution:
From what I could tell they used stratified sampling, but they split the population and sampled the strata in a systematic way like what was described in the previous problem. First they split the area of interest into 36 different strata. In each strata there were 49 different $1km^2$ samples, and in each strata one sample was taken.

Exercise 4: OK, systematic samples are... strange, though often very efficient. For instance, the use of the usual sample average can lead to bias. I'll show you that with a very small population, then show you that there IS an unbiased estimator. (This situation occurs when the population size isn't an exact multiple of k.)

Consider this small population: $\{1, 2, 3, 4, 10\}$, with mean $\mu = 20/5 = 4$. If we take a 1-in-2 sample, we get these two equally-likely samples: $1, 3, 10$ and $2, 4$.

a. Compute the usual average (our unbiased estimator of μ when we used SRS) for each sample. Since the samples are equally-likely, you can average the two averages to get $E(\overline{x})$. Show that this isn't the same as the true population mean. What does this tell you about the sample average as an estimator of μ in this case?

Solution:
From the following r analysis we can see that our systematic samples estimator for μ is not an unbiased estimator since the expected value of the mean sampling distribution does not equal the population mean.
Code:

```
> pop <- c(1, 2, 3, 4, 10)
> Samp_one <- c(1, 3, 10)
> Samp_two <- c(2, 4)
> Sample_means <- c(mean(Samp_one),mean(Samp_two))
    [1] 4.666667 3.000000
> ExpectedValue = sum(.5*Sample_means)
    [1] 3.833333
> mean(pop)
    [1] 4
```

b. There is a ... different... estimator of the population mean. For each sample, you sum up the values, multiply by the number of possible samples (2 in this case), then divide by the population size $N = 5$. In other words, $\hat{\mu} = \frac{k}{N} \sum x_i$. Compute this estimator for each sample. Since the two samples are equally-likely, the average of these two estimators is $E(\hat{\mu})$ Show that this IS an unbiased estimator of the true population mean. (This is based on a type of estimator that we'll see later, called the Horvitz-Thompson estimator).

Solution:
Performing the same analysis but with a new estimator,
Code:

```
> pop <- c(1, 2, 3, 4, 10)
> Samp_one <- c(1, 3, 10)
> Samp_two <- c(2, 4)

> Sample_means <- c((2*sum(Samp_one))/5 ,(2*sum(Samp_two))/5)
    [1] 5.6 2.4
> ExpectedValue = sum(.5*Sample_means)
    [1] 4
> mean(pop)
    [1] 4
```