**Exercise 1:** What is a simple random sample? If all sampling units have the same probability of being in the sample, is this always a simple random sample? Why or why not?

**Solution:**
A simple random sample(SRS) is a sampling scheme that comes in two varieties, sampling with replacement and sampling without replacement. In a SRS with replacements, every sample is equally likely and after sampling every sampling unit is returned to the population with the chance of being sampled again. In a SRS without replacement, after sampling, the sampling unit is not returned to the population. The simple random sample can be performed on a variety of populations where sampling units have different probabilities of being sampled ie. Gaussian Distributed, the key point is that every *sample* is equally likely.

**Exercise 2:** Use R to take a simple random sample of size $n = 4$(without replacement) from this population $10, 11, 13, 11, 10, 6, 22, 15, 14, 23$. Please include your output in the homework. Use this sample to compute your estimator of the population mean and find its standard error and a 95 percent confidence interval.

**Solution:**
Note that we are using a SRS without replacement on a relativity small sample, therefore when computing the standard error we must use the finite population correction.

**Code:**

```
> x <-c(10,11,13,11,10,6,22,15,14,23)

> x_sample = sample(x, 4)
    [1] 10 11 22  6

> est_sample_mean = mean(x_sample)
    [1] 12.25

> s2 = var(x)
> std_error = sqrt(((length(x) - length(x_sample))/length(x))
                    *(s2/length(x_sample)))

    [1] 2.075653

> CI_interval =
c(est_sample_mean + 2*std_error, est_sample_mean - 2*std_error)

    [1] 16.401305  8.098695
```

**Exercise 3:** We want to know the total number of moose in an area. Suppose we have divided the region into $N = 200$ quadrat, our guess is that the standard deviation of moose counts is around $s = 3$ moose and we would like a margin of error of less than ± 100 moose. How many sampling units must we visit.

**Solution:**
Taking the formula for the margin of error using standard error with the finite population correction and solving it for $n$ we get,

$$\begin{aligned} n &= \frac{N\sigma^2}{(N-1)\frac{B^2}{4N^2} + \sigma^2} \\ &= \frac{200(3)^2}{(200-1)\frac{100^2}{4(200)^2} + (3)^2} \\ &\approx 83.965 \end{aligned}$$

**Exercise 4:** Suppose we had decided to sample $n = 12$ of the quadrat and got a sample average of 14 moose per quadrat and a sample variance of $s^2 = 125$ square moose per quadrat. Find an estimate of teh total number of moose, along with it's standard error and then construct a 95 percent confidence interval for the total number of moose.

**Solution:**
Given that there are 100 quadrat, and the expected value for the mean of each quadrat is 14 a good estimator for the total population would

$$t = 100 * E(\bar{x}) = 1400.$$

Since the SRS at each quadrat are independent we can similarly estimate the variance,

$$V(t) = 100 * V(\bar{x}) = 12500$$

Computing the standard error,

$$SE = \sqrt(12500) = 111.8$$

Computing the 95% confidence interval for the total population

$$95_{CI} = (1623.6, 1176.4).$$

**Exercise 5:** I wish to estimate the total number of squirrels in a large region. I'll do that by dividing the region into $N = 1500$ transects, each 10m wide and 1km long. I select $n = 10$ of these to visit and count animals. I'll assume that I count everyone in the transect and ignore all animals outside the transect. I get the following counts: 12, 20 ,8, 42, 23, 18, 6, 8, 13, 17.

1. Find an estimator of the total number of squirrels in the entire region and the standard error, along with a 95 percent confidence interval.

   **Solution:**
   We proceed with the same procedure as the last problem. First we need an estimator for the mean number of squirrels in one transect.

   $$\bar{x} = 16.7$$

   Computing the estimate for the total population,

   $$t = 1500\bar{x} = 25050$$

   Computing it's variance,

   $$V(t) = 1500V(\bar{x}) = 165683.3$$

   Computing the SE and 95% confidence interval we get,

   $$SE_T = \sqrt{165683.3} = 407.04$$

   $$95_{CI} = (25864.08, 24235.92)$$

2. If I divide the estimated total by the total area of the region I'll get the density in squirrels. Find a 95 percent confidence interval for this density.

   **Solution:**
   The area of each transect is $1/100$km given the linearly of the expected value our estimate for the population total density is,

   $$t_{density} = \frac{1500\bar{x}}{100} = 250.5$$

   Because of the law of variances we need to square the $1/100$ coefficient when computing the variance of our estimate,

   $$V(t_{density}) = \frac{1500V(\bar{x})}{100^2} \approx 16.7.$$

   Computing the SE and 95% confidence interval,

   $$SE_T = \sqrt{16.7} = 4.07,$$

   $$95_{CI} = (258.6408242.3592).$$

**Exercise 6:** To use the typical estimator ±2 standard errors to find a 95 percent interval for a proportion, you need the sampling distribution of the sample proportion to be close to normal. You can assume this is the case if $n * (estproportion) > 10$ and $n * (1 - estproportion) > 10$. Suppose we are looking for the proportion of spruce trees in a low land forest that have a certain genetic trait. We somehow make a list of $N = 1320$ trees in the area. We take a SRS of size $n = 120$ trees and find tha 13 have the genetic trait.

1. Is this sample size sufficient for us to assume the sampling distribution of $\hat{p}$ is approximately normal? why or why not?

   **Solution:**
   First we need to identify our sample proportion,
   $$\hat{p} = \frac{13}{120}.$$
   Based on the test described above we do have a large enough sample size to assume sampling distribution of $\hat{p}$ is approximately normal,
   $$120 * \frac{13}{120} = 13 > 10.$$

2. Find a 95% confidence interval for the true proportion of trees in the region with the trait.

   **Solution:**
   First we need to compute the standard error,
   $$\hat{p}_{SE} = \sqrt{\frac{1320 - 120}{1320} \frac{\frac{13}{120}(1 - \frac{13}{120})}{120 - 1}} \approx 0.0271.$$
   So we get a 95% confidence interval of,
   $$95_{CI} = (.1625, .0541).$$

**Exercise 7:** I wish to conduct a political poll in a small town. the town has a total of 450 residents ans I can actually take a SRS of them. I would really like a margin of error of, at most, ± .05. What sample size should I take? What sample size would I need to take if I perversely decided to use SRS WITH replacement.

**Solution:**
Assuming the worst case, being that the true proportion is split 50/50 or at least close to it as most political polls are we can compute the sample size without replacement with the following formula,

$$n = \frac{N}{(N-1)B^2+1} = \frac{450}{(450-1)(.05)^2+1} \approx 213.$$

In the case that we take a SRS with replacement we essentially have an infinite population size, and the formula for finding $n$ reduces to,

$$n = \frac{1}{B^2} = \frac{1}{.05^2} = 400.$$

**Exercise 8:** Use bootstrapping on the data in problem 2 to get a 95 percent confidence interval for the population mean.

**Solution:**
**Code:**

```
> x <- c(10,11,13,11,10,6,22,15,14,23)
> n_sim <- 100000 # Number of boostrapped samples
> stored_means <- rep(NA, n_sim) #Storage Vector

> for(i in 1:n_sim){
+     temp_samp <- sample(x, size = 4, replace=TRUE)
+     stored_means[i] <- mean(temp_samp)
+   }

> bootstrapped_mean = mean(stored_means)
> bootstrapped_SE = sd(stored_means)
> bootstrapped_CI = c(bootstrapped_mean + 2*sd(stored_means),
                      bootstrapped_mean - 2*sd(stored_means))

> print(bootstrapped_mean)
    [1] 13.50019
> print(bootstrapped_CI)
    [1] 18.582566  8.417804
```

**Exercise 9:** Take a SRS(without replacement) of size $n = 18$ from the population in the table. Find the 95 percent confidence interval for the total over teh entire area. Later you will compare this estimate with the estimate from a stratified random sample.

**Solution:**
**Code:**

```
> x <- c(1, 3, 4, 5, 3, 3, 0, 7, 5, 0, 1, 4, 1, 4, 2, 3,
         4, 3, 4, 4, 3, 3, 0, 1, 2, 1, 5, 5, 3, 3, 11, 9,
         9, 15, 10, 6, 6, 12, 12, 12, 18, 14, 13, 7, 10,
         13, 3,14,11,11,15,11,13,12,12,13,19,10,11,14,
         13,6,9,7,15,14,9,16,13,12)

> x_sample = sample(x, 18)
> estimated_total = length(x)*mean(x_sample)
    [1] 641.6667

> estimated_total_var = length(x)*var(x_sample)
    [1] 1459.706

> estimated_total_SE = sqrt(estimated_total_var)
    [1] 38.2061

> Confidence_Interval <- c(estimated_total + 2*estimated_total_SE,
                           estimated_total - 2*estimated_total_SE)
    [1] 718.0789 565.2545
```