Choose one or more quantitative variable from countyComplete other than the poverty index, and consider the problem of estimating the standard error of the mean for the purpose of expressing a confidence interval.

What is Standard Error?
Standard Error is defined as the standard deviation of the estimate value and the formula to calculate the standard error is as follows –

$$SE = \sigma / \sqrt{n}$$

Where **σ** is standard deviation of the population and n is the number of independent observations from the population or the sample size.

What is a Confidence Interval?
A confidence interval is an interval that will contain a population parameter a specified proportion of the time. The confidence interval can take any number of probabilities, with the most common being 95% or 99%.

Confidence intervals are preferred to point estimates and to interval estimates, because only confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate.
We have two unrealistic ways to compute the sample mean's SE for purposes of expressing a confidence interval:

1. Actually compute the population standard deviation (σ) for our variable over the entire population and divide that by the square root of your sample size.
2. Or sketch the sampling distribution by repeatedly sampling from the population, and use the standard deviation of the resulting distribution.

Now, we will explore the method 1 evaluating the standard error

- Standard Error with population parameter

We are choosing the ‘age_under_18’variable to explore.

Code:

```
#read the countycomplete data into a dataframe
cc = read.csv("countyComplete.csv")
#store the data into a new declared variable age
age = cc$age_under_18
#declare the sample size
n = 100

#create a function for calculating variance
avar <- function(x){sum((x-mean(x))^2)/(length(x))}
```

```
#calculate the population variance and store it in a variable pop_var
pop_var = avar(age)

#calculate the standard deviation of the population
Sd_pop = pop_var ^ 0.5

#calculate the standard error
SE = Sd_pop / (n^0.5)

#creating a random sample of 100 samples from population
x1 = sample(age, 100)

#declaring the point estimate as the sample mean
estimate = mean(x1)
```

Output:

```
> pop_var
[1] 11.38646
> Sd_pop
[1] 3.374383
> SE
[1] 0.3374383
> estimate
[1] 23.115
```

Standard Error of various different sample sizes:

```
> Sd_pop/(300^0.5)
[1] 0.1948201
> Sd_pop/(400^0.5)
[1] 0.1687191
```

The 95% confidence interval for the above sample can be expressed using the formula –

point estimate ± 1.96 x SE

In our case, we would be calculating the 95% confidence interval as following-

23.115 ± 1.96 x 0.3374383

We have calculated the variance using above formula because the the var()

function in R uses n-1 value for calculating the variance by default, whereas the value to be used is ‘n’ which is our sample size.

We are not using the built-in sd() function in R for the same reason above. Instead we are calculating the standard deviation by taking the square root of the variance value which was calculated manually.

Design and execute a series of experiments to compare Kaplan’s resampling method in (his) chapter 5 to the guidelines recommended in OpenIntro chapter 4. Begin with the decision that the records in countyComplete are the population from which you are sampling.

Solution:

1. Open Intro Method
   In the OIS method we assume that we don’t have the population data, even though it is readily available with countycomplete. We pull out one sample from the population and then calculate the standard deviation of that same

Code:

```
##take a sample of size n=100 from population

samp = sample(cc$age_under_18, size = n, replace=TRUE)

##calculate the mean of sample

mean(samp)

#calculate the variance of samp

var_samp = avar(samp)

#Std deviation of sample

sd_samp = var_samp^0.5

#Std error for sample mean

SE_samp = sd_samp / (n^0.5)
```

Output:

```
> mean(samp)
[1] 23.294
> var_samp
[1] 19.61036
> sd_samp
[1] 4.428359
> SE_samp
[1] 0.4428359
```

Here, sample mean from the population is equal to 23.294. Then, we calculate the standard deviation for the standard error formula to be equal to the standard deviation of the sample which is equal to 4.428359. Later, we calculate the standard error to be equal to the 0.44.

The 95% confidence interval for the following estimate is displayed as follows –

$$\mathbf{23.294 \pm 1.96 * 0.44}$$

2. Kaplan Method

In the Kaplan method we assume that we do not have the population data, even though it is readily available with countycomplete data set. We take out one sample from the population and then resample 500 times using the do() function on this sample to create a resampling distribution. We then calculate the confidence interval using the confint() function as recommended by Kaplan.

Code:

```
#take a sample of size= 100 from population
resamp <- deal(cc$age_under_18, 100) #deal function acts similar to the
```

sample() function

```
#resample 500 times using the do() function

trials = do(500) * mean(resample(resamp), replace=TRUE)


#calculate the confidence interval using the confint() function
confint(trials)
```

Output:

```
> confint(trials)
  name    lower   upper level     method estimate
1 mean 22.66023 23.8208  0.95 percentile   23.254
```

For comparing the above two methods, we shall calculate the standard error of the entire population and then compare which of the above methods gave us the values close to the population values.

```
#variance kap
Kap_var = avar(resamp)

#sd kap
sd_Kap= Kap_var^0.5

#SE kap
SE_Kap = sd_Kap/ sqrt(100)
```

Output:

```
#Std error of population
> SE
[1] 0.3374383

#Std error by Kaplan method
> SE_Kap
[1] 0.3177056

#Std error by OIS method
> SE_samp
[1] 0.4428359
```

As we can see, when we compare the standard error values of the two different methods to the actual standard error of the population, we observe that the Kaplan method has a standard error value very close to the population standard error value as compare to the OIS method. Most often we do not have the entire population data and it is not easy to get the sampling distribution. However, we can replicate the sampling distribution method with the Kaplan method. Getting one sample from a population is comparatively easier as opposed to getting the sampling distribution. By using this one sample we can replicate and create our resampling distribution, post which we can calculate the standard deviation for it. Hence, this method gives a value much closer to the population value when compared to taking that one sample that might not really replicate the behaviour of the population. Also, we can see that the confidence interval limits are more accurate for Kaplan method and the sample mean falls perfectly in the limits.

But, in my opinion which one method produces more accurate estimate than the other depends entirely on the data that is available. If we have access to just one sample then first we would check how the data is spread out, and if this sample is distributed somewhat normally then we would not have to perform resampling, and we would be able to calculate the standard error using the OIS method. However, if the data is not normally distributed and it appears slightly skewed, then resampling of the data is needed and we would use the Kaplan method to calculate the standard error.