This is from problem 67 of chapter 7 : The data set `families` contains information about 43,886 families living in the city of Cyberville. There are 3 family types in the data, coded as 1, 2, 3 respectively. The data set also contains information about the number of people in the family, the number of children, the family income, and the education level of the head of the household. Further details are in the text.

1. Take a simple random sample of 500 families. Estimate the proportion of heads of household who did not receive a high school diploma. Calculate the estimated standard error of this estimate, and form a 95% confidence interval.

2. Now take 100 samples of size 400 each.

   (a) For each sample, find the average family income. **Do not print these and turn them in.**

   (b) Find the average and standard deviation of these 100 estimates and make a histogram of the estimates.

   (c) Superimpose a curve of normal density with the mean and standard deviation of the histogram and comment on the fit.

   (d) Plot the empirical CDF (see section 10.2). On this plot, superimpose the normal cumulative distribution function with mean and sd as defined in (c). Comment on the fit.

   (e) Another way of examining a normal approximation is via normal probability plots (section 9.9). Make such a plot, and comment on what it shows about the approximation.

   (f) For each of the 100 samples, find a 95% confidence interval for the population average income. How many of the intervals actually contain the population target?

3. Take a simple random sample of size 500 and compare the incomes of the three family types by comparing histograms and boxplots (section 10.6).

4. Take simple random samples of size 400 from each of the four regions.

   (a) Compare the incomes of the regions by making parallel boxplots.

   (b) Does there appear to be a difference in education level among the four regions?