Please write legibly and show all of your work, unless otherwise stated. You must attach your computer output to receive points for Part B.

**PART A – METHODOLOGY EXERCISES (60 points)**

1. Search the internet for two sources of information on the living conditions of the homeless for which statistics are given and the sampling techniques used to gather those statistics are described.

   a) Briefly describe your sources and list the web sites where they can be found.
   b) Describe how these sources of information differ in their approach to collecting information.
   c) Describe some of the statistics presented in the sources and the sampling techniques used to gather those statistics.
   d) Evaluate the sampling method in terms of representativeness and generalizability.
   e) Suggest ways to improve the representativeness and generalizability of the studies.

2. Careful studies of the long-term results of medical treatments have become possible for many illnesses in recent years because of the development by managed care organizations (HMOs) of computerized databases on millions of patients. Consider combining databases to generate a massive database on all US residents enrolled in HMOs.

   a) Describe a cluster sampling plan for selecting cases from the databases for such a study.
   b) Evaluate the strengths and weakness of the cluster design that you propose.


**PART B – COMPUTER PROBLEMS (40 points)**

In these problems, you will learn about the effect of sampling and sample size on obtaining estimates. Using SPSS for Windows 10.0, you will extract data from the 1996 General Social Survey (GSS96.sav found on the network at `\storage\projects\sst\sst115-03`) for the variables EDUC (respondent’s highest year of school completed), SEX (respondent’s sex), and SEXHAR (has respondent experienced sexual harassment).

1. You will draw samples of different sizes from the 1996 GSS data set and then calculate the mean and the standard error of the mean for EDUC for the different samples.

   a) Open the 1996 GSS file (GSS96.sav). Obtain the mean and the standard error of the mean for the variable EDUC in the 1996 GSS.
Click on **analyze → descriptive statistics → descriptives** and place **EDUC** in the variable(s) box. In the descriptives window, click on the **options** button and add **S.E. Mean** to the list of statistics generated. Click on **continue**, then **OK** to calculate the statistics. Record the value of the mean and the standard error for **EDUC**.

b) Select a random sample of 10% of the cases. To use the sample procedure, click on **data → select cases**. In the select cases window, choose **random sample of cases**. Click on **sample**. In the **random sample** window, request approximately 10% of all cases. Click on **continue**, then **OK**.

c) Follow the steps in part “a” to request the mean and standard error of the mean for **EDUC**. Record your results.

d) Repeat the steps in part “b”, but select a 65% sample. Request the same statistics on **EDUC**. Record your results.

e) How closely does the mean for **EDUC** from each sample match the mean for the full sample?

f) What happens to the standard error of the mean as you increase the sample size? Explain why.

2. Are women more likely than men to have experienced sexual harassment? Using the 1996 GSS data, examine the relationship between **SEX** and **SEXHAR**.

a) Open the 1996 GSS file. Create a two-way table for the relationship between **SEX** and **SEXHAR**.

Click on **analyze → descriptive statistics → crosstabs**. In the crosstabs window, put **SEXHAR** in the rows of the table and **SEX** in the columns of the table. At the bottom of the window, click on the cells button. In the cells window, choose **column percentages**, then **continue**, then **OK** to generate a two-way table for the relationship.

b) Choose key column percentages to describe the relationship between a respondent’s gender and whether or not he/she has experienced sexual harassment.

c) Select a random sample of 50% of the cases. To use the sample procedure, click on **data → select cases**. In the select cases window, choose **random sample of cases**. Click on **sample**. In the **random sample** window, request approximately 50% of all cases. Click on **continue**, then **OK**.

d) Repeat the steps in part “a” to request a two-way table for the 50% sample. Record your results.

e) Repeat the steps in part “c” to request a random sample of 15% of the cases.

f) Repeat the steps in part “a” to request a two-way table for the 15% sample. Record your results.

g) How closely do the two-way tables from each sample match the table for the full sample? Do the estimates improve as you increase the sample size? Why or why not?