**MSDS 6370 Sampling Statistics**

**Midterm Exam –2018 Spring**

**You are to work independently on this exam.**

**You may NOT consult other people. However, you may use course materials.**

**Rename your answer sheet as Midterm\_YourFullName, attached your SAS output in the associated problems and SAS code as appendix in the problem orders.**

**Submit the answer sheet on 2ds smu🡪Grades🡪Midterm.**

**(Please take SMU Honor code seriously)**

1. **(10 pts)** A SRSWOR is chosen from a population of 400 acres of land. Results are below.

|  |  |
| --- | --- |
| **Price/Acre** | |
|
| **Mean** | 15,050 |
| **Median** | 11,882 |
| **Standard Deviation** | 10,548 |
| **Minimum** | 0 |
| **Maximum** | 53,862 |
| **Sum** | 602,000 |
| **Sample Size** | 40 |

1. What is the standard error of the mean from the sample?
2. Calculate a 95% confidence interval for the population total land value.
3. Suppose you want to retake the sample to achieve a margin of error for the mean of 2000 with a 99% confidence interval. How large a sample size would be needed to achieve this objective?

2. **True/False:** (**20 pts; 2 points for each**)

For each statement, select True or False. Make sure it is clear which one you are choosing. Think Carefully!

2.1 Simple random samples always will give a better estimate of the mean than judgement samples if implemented properly. (TRUE OR FALSE)

2.2 The largest mean for a simple random sample is a subset of the highest values in the population. (TRUE OR FALSE)

2.3 For a yes, no question on a survey you can find the overall proportion of yes answers in the survey by summing the proportions of yes answers in strata. (TRUE OR FALSE)

2.4 Sampling error can be minimized, but never eliminated if the sample does not include the entire population. (TRUE OR FALSE)

2.5 If one observation in a population has a probability of selection that is unknown, then you cannot use weighting to estimate the mean from a probability sample. (TRUE OR FALSE)

2.6 When using the probability of selection to find the mean of a sample, the formulas used include the probabilities of selection in functions in both the numerator and denominator. (TRUE OR FALSE)

2.7 From the Central Limit Theorem we know that the sample mean has a normal distribution. This allows us to construct meaningful confidence intervals for our sample estimates.

(TRUE OR FALSE)

2.8 Proc surveyMeans in SAS is used to select a random sample. (TRUE OR FALSE)

2.9 When using sample data to estimate a population variance you do not have to include the finite population correction. (TRUE OR FALSE)

2.10 The standard error of the mean is the standard deviation of the sample. (TRUE OR FALSE)

3. **(10 points)** The Chicago Fitness Company has five clubs in different parts of the city. The company wants to estimate the total number of customers served in a year by taking an SRSWOR sample of two clubs. Last year, Club1, Club2, Club3, Club4, and Club5 served 1000, 1200, 1400, 1800, and 1600 customers respectively:

(a) (2 points) Fill in the table below showing all samples.

|  |  |  |
| --- | --- | --- |
| **Sample** | **Clubs** | **Total Served** |
| 1 | 1,2 | 5500 |
| 2 | 1,3 | 6000 |
| 3 | 1,4 | 7000 |
| 4 | 1,5 | 6500 |
| 5 | 2,3 | 6500 |
| 6 | 2,4 | 7500 |
| 7 | 2,5 | 7000 |
| 8 | 3,4 | 8000 |
| 9 | 3,5 | 7500 |
| 10 | 4,5 | 8500 |
| Number of Samples | 10 |  |

(b) (3 points) Find the sampling distribution of the estimate of the total number of customers served and fill in the table below.

|  |  |  |
| --- | --- | --- |
| **Total Served** | **# of samples with Value** | **Proportion of samples with value** |
| 5500 | 1 | 0.1 |
| 6000 | 1 | 0.1 |
| 6500 | 2 | 0.2 |
| 7000 | 2 | 0.2 |
| 7500 | 2 | 0.2 |
| 8000 | 1 | 0.1 |
| 8500 | 1 | 0.1 |

(c) (5 points) Suppose clubs 1, 2, and 3 are on the Southside and clubs 4 and 5 are on the Northside of the city.

1. Show the possible samples from taking a sample of one club from the Southside and one from the Northside and calculate an estimate of the population mean using these samples.

|  |  |  |
| --- | --- | --- |
| **Sample** | **Clubs** | **Total Served** |
| 1 | 1,4 | 2800 |
| 2 | 1,5 | 2600 |
| 3 | 2,4 | 3000 |
| 4 | 2,5 | 2800 |
| 5 | 3,4 | 3200 |
| 6 | 3,5 | 3000 |
| Number of Samples | 6 |  |

2. Calculate the standard error of the mean using these samples and show how this standard error is different than the standard error from the simple random sampling without replacement design.

4. **(10 points)** There are 12 peanut farms in a rural county in Florida. 4 farms are owned by a person less than or equal to 40 years of age and 8 are owned by a person older than 40.

A random sample without replacement of size 2 is taken from this population and the proportion older than 40 is estimated from the sample.

1. Calculate the mean and standard deviation of the sampling distribution.
2. Show the distribution of the mean of the sample.

|  |  |  |
| --- | --- | --- |
|  |  | **Proportion of samples with value** |
| 0.0 |  | 0.424 |
| 0.5 |  | 0.485 |
| 1.0 |  | 0.091 |

1. A SRSWOR is taken resulting in the values of 3 people older than forty and 2 people younger than forty. Based on this sample, give an estimate of the proportion of farms owned by people over forty in the population and an estimate of the standard error of this mean.

5. **(20 points)** A school principal wants to estimate the average test score of its students in three tracts. You are given data with test scores from a proportional random sample of fifty students in the school from the three tracts (see sheet Q5Data in the file midtermDatSpring2018.xlsx). The principal thinks the simplest approach is to average all the scores and construct a confidence interval to demonstrate it is a good estimate. The entire first, second, and third tracts in the school contain 55, 80, and 65 students respectively.

When answering the questions below, do not use any statistical software and show your work to demonstrate you performed your calculations manually.

Part (1) (10 points) Show the principal what is the best design to calculate an estimate for the average score of students in the school using the sample data. Use the data and calculate average estimates and standard errors in making your argument.

1.2641

Part (2) (10 points) The principal does some research and says to you he thinks a Neyman allocation may work best for estimating test scores. However, he is not sure how to do this allocation. Using the data you have from the proportional design, recommend a Neyman allocation for another sample.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| stratum |  |  | 50\* | sample size |
| Track 1 | 2,033.4069 | 0.2063 | 10.315 | 10 |
| Track 2 | 4,386.9777 | 0.4452 | 22.26 | 22 |
| Track 3 | 3,434.5212 | 0.3485 | 17.425 | 18 |

**6. Lab exam (30 pts)**

An advertising firm wants to determine how much to emphasize television advertising in a given state county. The county has two towns A & B, and a rural area. The firm only has enough money to sample 100 households. The population data is given in sheet labExamData in midtermDatSpring2018.xlsx. The unit of observation is the household and the variable of interest is hours of TV viewing.

(a) (5 pts) Select a stratified random sample, using proportional allocation. What is the allocation of the sample for this design? Show your computations for determining this allocation.

(b) (5 pts) Using SAS PROC SURVEYSELECT, select a stratified random sample from the population, using the allocation specified in (a). Use a random number seed of 1117 in choosing the sample. Give your SAS code, along with a list of the sample elements. (Hint: To make your life simpler, use the code below to recode place when you choose the sample).

data analDat; set inputData;   
 if place = "A" then strata = 1;  
 if place = "B" then strata = 2;  
 if place = "R" then strata = 3;  
run;

(c) (5 pts) Use SAS PROC SURVEYMEANS with the sample above to compute these statistics for the hours of TV viewing: The estimate of the population mean, the standard error of the estimate, the 95% confidence interval for the mean. Paste your SAS code and output below. (Do not include the graphics in your paste.)

(d) (5 pts) A Neyman allocation of the sample is 38, 33, and 29 for strata A, B, and R respectively. Using SAS PROC SURVEYSELECT, select a stratified random sample from the population, using the Neyman allocation. Use a random number seed of 1118. Give your SAS code.

e) (5 pts) Use SAS PROC SURVEYMEANS with the Neyman sample compute these statistics for the hours of TV viewing: The estimate of the population mean, the standard error of the estimate, the 95% confidence interval for the mean. Paste your SAS code and output below. (Do not include the graphics in your paste.)

f) (5 pts) Comment on whether the Neyman allocation gives better or worse estimates.