**MSDS 6370 Sampling Statistics Midterm Exam -**

**Summer 2017**

**This exam is due at 11.00p.m. CT on Tuesday, June 20, 2017. You are to work independently on this exam. You may not consult other people. However, you may use course materials. (Please take SMU Honor code seriously)**

1. (10 pts) A SRS is chosen from a population of 1000 households. Results are shown below.

|  |  |
| --- | --- |
| income |  |
|  |  |
| Mean | 60199.95 |
| Standard Error | 6671.014 |
| Median | 47526.5 |
| Standard Deviation | 42191.2 |
| Minimum | 0 |
| Maximum | 215448 |
| Sum | 2407998 |
| Size | 40 |

a) (5 pts) Calculate 90% confidence interval for the population total income.

49,226,131.97, 71,173,768.03)

b) (5 pts) Suppose you wanted to redo the survey above to achieve a moe (margin of error) of $5000 with 90% confidence level. How large a sample size would be needed, if the population from which this sample came has 1000 members?

2. **True/False:** (15 pts; 2.5 points for each) For each statement, select True or False. Make sure it is clear which one you are choosing.

1. The selection probability for each unit from simple random sample of size n is . (where N is the population size and n is the sample size.)

~~True~~ False

1. Measurement error is an example of a nonsampling error

True ~~False~~

1. The selection probability for each unit from systematic sample of size n is . (where , N is the population size and n is the sample size.)

True ~~False~~

1. A statistic is a numerical index that describes some feature of a sample.

True ~~False~~

1. Sampling fraction for each stratum is the same when we use stratified design with proportional allocation. (= sampling fraction in the stratum)

~~True~~ False

1. A stratified design with proportional allocation is an EPSEM (Equal Probability SElection Method) design.

True ~~False~~

3. (15 pts) A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. This number is not directly available on the database, so requires the investigator to examine each record separately.

(i) A frequency table for number for number of refereed publications is given below for an SRS of 50 faculty members.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Refereed Publications | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Faculty members | 28 | 4 | 3 | 4 | 4 | 2 | 1 | 0 | 2 | 1 | 1 |

1. (3 points) Estimate the mean number of publications per faculty member.
2. (2 points) Estimate the proportion of faculty members with no refereed publications.

Estimate of the proportion with no refereed publications = 56%

(ii) However, in the SRS, not all departments were represented. The SRS contained several faculty members from psychology and from chemistry, but none from foreign languages. The following data are from a stratified sample of faculty, using the areas biological sciences, physical sciences, social sciences, and humanities as the strata.

|  |  |  |
| --- | --- | --- |
| **Stratum** | **Number of faculty members in stratum** | **Number of faculty members in sample** |
| Biological Sciences | 102 | 7 |
| Physical Sciences | 310 | 19 |
| Social Sciences | 217 | 13 |
| Humanities | 178 | 11 |
| **Total** | **807** | **50** |

The frequency table for number of publications in the strata is given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of refereed publications | Number of faculty members | | | |
| Biological | Physical | Social | Humanities |
| 0 | 1 | 10 | 9 | 8 |
| 1 | 2 | 2 | 0 | 2 |
| 2 | 0 | 0 | 1 | 0 |
| 3 | 1 | 1 | 0 | 1 |
| 4 | 0 | 2 | 2 | 0 |
| 5 | 2 | 1 | 0 | 0 |
| 6 | 0 | 1 | 1 | 0 |
| 7 | 1 | 0 | 0 | 0 |
| 8 | 0 | 2 | 0 | 0 |

1. (5 points) Estimate the total number of refereed publications by faculty members in the university.
2. (5 points) Estimate the proportion of faculty members with no refereed publications.

Estimate of the proportion with no refereed publications = 56%

4. (10 pts) A stratified sample design was implemented in order to estimate the total $ contributed to the poor in 2016 by all the churches in a community. A stratified design was used, where strata were defined by size of membership. A random sample of 20 churches was selected from each of three strata: Large churches ( = 120), Medium churches ( = 130), Small churches ( = 533). The results were summarized by Excel as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Large* |  | *Medium* |  | *Small* |  |
|  |  |  |  |  |  |
| Mean | 12374.8 | Mean | 4315.7 | Mean | 334.9 |
| Standard Error | 1068.1 | Standard Error | 173.7 | Standard Error | 14.3 |
| Median | 12188.9 | Median | 4260.1 | Median | 338.6 |
| Standard Deviation | 4776.5 | Standard Deviation | 777.0 | Standard Deviation | 63.9 |
| Sample Variance | 22815320.0 | Sample Variance | 603671.0 | Sample Variance | 4078.9 |
| Kurtosis | -0.6 | Kurtosis | 1.4 | Kurtosis | -0.6 |
| Skewness | 0.3 | Skewness | 0.7 | Skewness | 0.0 |
| Range | 16450.2 | Range | 3303.5 | Range | 238.5 |
| Minimum | 5058.9 | Minimum | 3081.4 | Minimum | 211.3 |
| Maximum | 21509.2 | Maximum | 6384.9 | Maximum | 449.8 |
| Sum | 247495.6 | Sum | 86314.2 | Sum | 6698.6 |
| Count | 20 | Count | 20 | Count | 20 |

(a) (5 points) Make an estimate of the mean amount of funds spent on the poor by churches in the community.

By weighting each strata and taking the weighted average of the strata sampling means we can estimate the mean amount of funds spent on the poor by the churches.

To weight I used the number in the strata in the population divided by the number sampled in the strata.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| whi = | 6 |  | 6.5 |  | 26.65 |
|  |  |  |  |  |  |
| ybarstr = | 2841.020051 | |  |  |  |

(b) (5 points) Suppose the same survey is to be done next year with the same budget (i.e., the same sample size is to be used). Should the allocation of sample to strata be kept the same or not? (The objective is still to estimate the total amount of expenditures by all churches in the community.) If kept the same, explain why. If not kept the same, specify a new allocation. (No need to do the calculation. It is enough to explain why we need to go to this new allocation)

The allocation to strata should not be kept the same. Allocating proportionally would improve the estimate but allocating using the Neyman method would improve it further. The sample variance from this year’s sample shows a large difference between strata. Sampling more from the strata with the higher variance and less from the strata with the lower variance would more efficiently use the resources available.

5. (10 pts) Estimate of the proportion of students who are graduates.

|  |  |  |
| --- | --- | --- |
| **Student** | **Stratum** | **Graduate?** |
| Ali | 1 | Y |
| Bucky | 1 | N |
| Judith | 2 | Y |
| Hal | 2 | N |
| Roy | 1 | Y |
| Gideon | 1 | N |
| John | 1 | Y |
| Yusun | 1 | Y |

Step 1. Select 1 student from each stratum.

Step 2. Enumerate all possible samples.

Step 3. Calculate for every sample.

|  |  |  |
| --- | --- | --- |
| Ali | Judith | 1 |
| Bucky | Judith | 0.25 |
| Roy | Judith | 1 |
| Gideon | Judith | 0.25 |
| John | Judith | 1 |
| Yusun | Judith | 1 |
| Ali | Hal | 0.75 |
| Bucky | Hal | 0 |
| Roy | Hal | 0.75 |
| Gideon | Hal | 0 |
| John | Hal | 0.75 |
| Yusun | Hal | 0.75 |

Step 4. Specify the sampling distribution of sample proportions based on this stratified design.

|  |  |
| --- | --- |
| sample proportion | number |
| 0 | 2 |
| 0.25 | 2 |
| 0.75 | 4 |
| 1 | 4 |

Step 5. Calculate the mean of .

|  |  |
| --- | --- |
| mean | 0.625 |

6. (10 pts) A population of 10 balls is made up of 4 white balls and 6 red balls as shown below:

R W W R W R W R R R

A SRS of size 2 is selected from the population and the proportion of red balls in the population is estimated from the sample using as the estimator

.

Find the sampling distribution of .

K = 10/2 = 5

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | MEMBERS | | VALUE |
| 1 | R | R | 1 |
| 2 | W | W | 0 |
| 3 | W | R | 0.5 |
| 4 | R | R | 1 |
| 5 | W | R | 0.5 |

|  |  |  |
| --- | --- | --- |
| STATISTIC VALUE | # SAMPLES | PROPORTION OF SAMPLES |
| 0 | 1 | 0.2 |
| 0.5 | 2 | 0.4 |
| 1 | 2 | 0.4 |

**Lab exam (30 pts) (You can use either SAS or R)**

Consider data from all the public libraries in Texas. The column labeled *circ* gives the circulation (# of books checked out) for each library in a particular year. You will be asked to construct a 95% confidence interval for the mean of a different characteristic (*inq* = # of inquiries answered) for the population of libraries from your chosen sample.

We sorted the libraries into small, medium, and large libraries and will regard those as strata. We show a column labeled 1(Large), 2 (Medium), 3 (Small) to denote which stratum each library belongs to.

Note that this data has been sorted by stratum. This sheet is provided in a csv file named ***libraryinfo.csv***.

Here are the steps you will follow:

(a) (6 points) You will select a stratified random sample of size 40, using proportional allocation. What is the allocation of sample for proportional allocation; i.e., how large a sample will you select from each stratum?

*nS* = 9

*nM* = 9

*nL* = 22

Show your computations for determining this allocation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| n= | 40 | Proportional | | |
| strata | Nh | Nh/N | nh | rounded |
| 1 | 203 | 0.550135501 | 22.00542005 | 22 |
| 2 | 81 | 0.219512195 | 8.780487805 | 9 |
| 3 | 85 | 0.230352304 | 9.214092141 | 9 |
| sum | 369 | 1 | 40 | 40 |

Nh counts the number of libraries in each strata of the population.

Nh/N divides the total population by the number in each strata to find the proportion in each strata.

Nh multiplies the proportion in each strata by the sample size of 40

Rounded rounds that value to a whole number.

(b) (6 pts) Using SAS PROC SURVEYSELECT (or R), select a stratified random sample from the population, using allocation specified in (a). Paste your SAS (or R) code here, along with resulting output.

FILENAME REFFILE '/home/anabbott0/sasuser.v94/libraryinfo.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

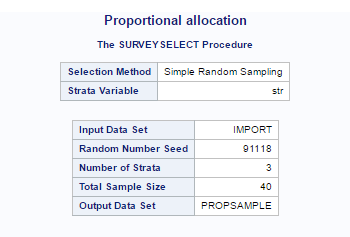
RUN;

proc surveyselect data=WORK.IMPORT method = srs out = propsample

sampsize = (22,9,9) seed=91118;

strata str;

title "Proportional allocation";



(c) (6 pts) **Print out your sample and copy and paste it below**. 3 pts for printing, 3 pts for including code to print.

| **Obs** | **str** | **ID** | **circ** | **inq** | **SelectionProb** | **SamplingWeight** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 1 | 27 | 81348 | 4050 | 0.10837 | 9.22727 |
| **2** | 1 | 90 | 15927 | 450 | 0.10837 | 9.22727 |
| **3** | 1 | 100 | 18911 | 900 | 0.10837 | 9.22727 |
| **4** | 1 | 139 | 26500 | 360 | 0.10837 | 9.22727 |
| **5** | 1 | 157 | 37794 | 892 | 0.10837 | 9.22727 |
| **6** | 1 | 160 | 34735 | 2674 | 0.10837 | 9.22727 |
| **7** | 1 | 204 | 49481 | 2500 | 0.10837 | 9.22727 |
| **8** | 1 | 224 | 10235 | 100 | 0.10837 | 9.22727 |
| **9** | 1 | 247 | 11999 | 0 | 0.10837 | 9.22727 |
| **10** | 1 | 250 | 27837 | 736 | 0.10837 | 9.22727 |
| **11** | 1 | 252 | 19539 | 0 | 0.10837 | 9.22727 |
| **12** | 1 | 263 | 19101 | 540 | 0.10837 | 9.22727 |
| **13** | 1 | 279 | 54182 | 2301 | 0.10837 | 9.22727 |
| **14** | 1 | 289 | 26966 | 1500 | 0.10837 | 9.22727 |
| **15** | 1 | 291 | 18623 | 2685 | 0.10837 | 9.22727 |
| **16** | 1 | 324 | 26384 | 1142 | 0.10837 | 9.22727 |
| **17** | 1 | 326 | 21281 | 650 | 0.10837 | 9.22727 |
| **18** | 1 | 342 | 24705 | 2779 | 0.10837 | 9.22727 |
| **19** | 1 | 350 | 65327 | 1500 | 0.10837 | 9.22727 |
| **20** | 1 | 356 | 17679 | 597 | 0.10837 | 9.22727 |
| **21** | 1 | 358 | 13976 | 205 | 0.10837 | 9.22727 |
| **22** | 1 | 364 | 41682 | 3838 | 0.10837 | 9.22727 |
| **23** | 2 | 3 | 242080 | 37752 | 0.11111 | 9.00000 |
| **24** | 2 | 18 | 146862 | 9599 | 0.11111 | 9.00000 |
| **25** | 2 | 74 | 121190 | 7765 | 0.11111 | 9.00000 |
| **26** | 2 | 107 | 110864 | 6121 | 0.11111 | 9.00000 |
| **27** | 2 | 207 | 242996 | 43378 | 0.11111 | 9.00000 |
| **28** | 2 | 270 | 108798 | 7580 | 0.11111 | 9.00000 |
| **29** | 2 | 297 | 352199 | 38287 | 0.11111 | 9.00000 |
| **30** | 2 | 305 | 826417 | 44256 | 0.11111 | 9.00000 |
| **31** | 2 | 311 | 2271227 | 262141 | 0.11111 | 9.00000 |
| **32** | 3 | 82 | 6245 | 50 | 0.10588 | 9.44444 |
| **33** | 3 | 98 | 5413 | 0 | 0.10588 | 9.44444 |
| **34** | 3 | 128 | 3157 | 295 | 0.10588 | 9.44444 |
| **35** | 3 | 163 | 0 | 0 | 0.10588 | 9.44444 |
| **36** | 3 | 217 | 4867 | 105 | 0.10588 | 9.44444 |
| **37** | 3 | 275 | 1619 | 0 | 0.10588 | 9.44444 |
| **38** | 3 | 325 | 0 | 0 | 0.10588 | 9.44444 |
| **39** | 3 | 351 | 6522 | 39 | 0.10588 | 9.44444 |
| **40** | 3 | 353 | 9175 | 2201 | 0.10588 | 9.44444 |

PROC PRINT DATA = PROPSAMPLE;

RUN;

(d) 12 pts Use SAS PROC SURVEYMEANS (or R) with the sample above to compute these statistics for the number of inquiries: the estimate of population mean, the standard error of the estimate, the 95% confidence interval for the mean. Paste your SAS (or R) code and output below. (This can be included above 5 pts for code, 5 pts for correct CI, mean, and sterr.) (Do not include the graphics in your paste.)

data strsizes;

input str \_total\_;

datalines;

1 203

2 81

3 85

;

run;

proc surveymeans data = propsample MEAN CLM total = strsizes;

var inq;

weight SamplingWeight;

strata str;

title "Proportional allocation";

run;

**Proportional allocation**

**The SURVEYMEANS Procedure**

| **Data Summary** | |
| --- | --- |
| **Number of Strata** | 3 |
| **Number of Observations** | 40 |
| **Sum of Weights** | 369 |

| **Statistics** | | | | |
| --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Std Error of Mean** | **95% CL for Mean** | |
| **inq** | 11972 | 5590.615475 | 644.739282 | 23300.0652 |